Reducing Patient Wait Times in NHS Triage Using a Mixture-of-Agents Simulation Framework

Raunak Burrows

Master of Science in Artificial Intelligence

from the

University of Surrey

School of Computer Science and Electronic Engineering

Faculty of Engineering and Physical Sciences

University of Surrey

Guildford, Surrey, GU2 7XH, UK

September 2025

Supervised by: Dr Xiatian Zhu

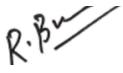
©Raunak Burrows 2025

DECLARATION OF ORIGINALITY

I confirm that the project dissertation I am submitting is entirely my own work and that any material used from other sources has been clearly identified and properly acknowledged and referenced. In submitting this final version of my report to the JISC anti-plagiarism software resource, I confirm that my work does not contravene the university regulations on plagiarism as described in the Student Handbook. In so doing I also acknowledge that I may be held to account for any particular instances of uncited work detected by the JISC anti-plagiarism software, or as may be found by the project examiner or project organiser. I also understand that if an allegation of plagiarism is upheld via an Academic Misconduct Hearing, then I may forfeit any credit for this module or a more severe penalty may be agreed.

Reducing Patient Wait Times In N.H.S. Emergency Department Triage Using A Mixture-Of-Agents Simulation Framework

Raunak Burrows

Author Signature  Date: 09/09/2025

Supervisor's name: Dr Xiatian Zhu

WORD COUNT

Number of Pages: 53

Number of Words: 11626

ABSTRACT

Emergency departments across the NHS face unprecedented pressure, with only 76.4% of A&E attendances meeting the four-hour target in Q4 2023-2024 and over 35,000 patients waiting more than 12 hours [1]. While large language models (LLMs) have shown promise in enhancing triage systems like the Manchester Triage System (MTS), their application in post-triage operations—particularly real-time resource allocation—remains underexplored. This dissertation proposes a novel Mixture-of-Agents (MoA) framework that leverages patient history and contextual data to enable dynamic, intelligent resource allocation immediately following triage, addressing a critical gap in current ED workflow optimization.

The proposed system integrates multiple specialised LLM agents—comprising proposers and aggregators—that collaboratively analyse resource queues across MTS urgency levels (red, orange, yellow, green, blue) whilst incorporating comprehensive electronic health record (EHR) data. Through advanced queuing theory modelling, specifically M/M/c and M/G/1 systems with routing optimization, the framework achieves significant efficiency gains: direct routing of urgent red cases with evident historical needs (e.g., confirmed MRI requirements) to appropriate queues, bypassing default doctor examinations; intelligent queuing of non-urgent blue/green cases to prevent backlogs; and bundling multiple tests for complex cases to eliminate repeat visits.

The work establishes a novel simulation methodology using SimPy [2]. to model healthcare workflows, incorporating open-source LLMs as the decision engine for dynamic resource allocation. Crucially, the research designs and implements comparative testing between single-agent and mixture-of-agents configurations against a synthetic FHIR/HL7-compliant clinical dataset to assess routing efficacy and clinical plausibility.

These achievements create a foundational pathway for future implementation of adaptive resource allocation systems in real-world healthcare settings. By shifting from static rule-based protocols to context-aware decision frameworks, this work sets the stage for scalable AI integration that optimizes patient flow while maintaining clinical safety standards. Readers will find detailed analysis of the simulation architecture, comparative agent performance metrics, and methodological insights for translating this approach into practical healthcare optimization tools.

TABLE OF CONTENTS

LIST OF FIGURES

1 INTRODUCTION

The Mixture-of-Agents framework revolutionizes emergency department routing by replacing rigid sequential pathways with collaborative AI decision-making. Traditional systems force all patients through mandatory initial physician consultations, creating artificial bottlenecks even when clinical evidence clearly indicates specific service requirements. This approach leverages multiple specialized agents that collectively analyse patient data to determine optimal resource allocation, directly routing patients to necessary services like MRI when clinical indicators are unequivocal. By eliminating unnecessary gatekeeping steps while maintaining clinical safety, the framework addresses the fundamental inefficiency in current NHS emergency department workflows where 90% of trusts still rely on the Manchester Triage System's sequential protocols [3].

1.1 Background and Context

Emergency departments (EDs) across the UK are experiencing unprecedented strain, with systemic inefficiencies undermining patient safety, staff wellbeing, and public trust in the National Health Service (NHS). As of Q4 2023-2024, only 76.4% of Accident & Emergency (A&E) attendances meet the four-hour treatment target, while over 35,000 patients endure waits exceeding 12 hours—a figure 78 times higher than pre-pandemic levels [1]. These delays reflect not just capacity issues but a fundamental challenge to the accessibility and responsiveness of emergency care, particularly during periods of peak demand. The crisis is exacerbated by rising patient complexity, workforce shortages, and outdated operational models that fail to adapt to real-time demand fluctuations.

The current standard for initial patient prioritisation in NHS EDs is the Manchester Triage System (MTS), used in approximately 90% of trusts and formally adopted by over 200 NHS organisations [3]. MTS assigns triage levels (red, orange, yellow, green, blue) based on vital signs and symptom severity, enabling rapid categorisation and queue management. However, MTS operates primarily at the triage stage and does not address downstream inefficiencies in post-triage resource allocation. Once triaged, patients typically follow rigid, linear pathways—such as mandatory physician consultation before accessing imaging—which can introduce artificial bottlenecks, especially for urgent cases requiring immediate diagnostic intervention (e.g., stroke or trauma).

Recent advances in artificial intelligence, particularly large language models (LLMs), have shown promise in enhancing clinical decision support. Studies indicate LLMs outperform traditional machine learning models in handling missing data, ambiguous inputs, and distribution shifts—key

challenges in real-world healthcare environments [4]. While some research has explored LLMs for augmenting triage decisions, their application to post-triage operations—where routing choices directly influence wait times and resource utilisation—remains underexplored. This gap presents a critical opportunity to leverage AI not just for diagnosis, but for dynamic workflow optimisation.

To bridge this gap, this research focuses on intelligent patient routing as a strategic intervention point. Unlike static rule-based systems, our proposed Mixture-of-Agents (MoA) framework enables adaptive, context-aware decision-making by simulating multiple LLM agents that collaboratively assess patient needs and recommend optimal pathways. The framework is evaluated using a discrete-event simulation environment built with SimPy, incorporating synthetic patient trajectories derived from FHIR/HL7-compliant datasets generated via Synthea. This allows us to test routing strategies under realistic conditions while preserving clinical coherence and enabling safe evaluation of AI-driven decisions across demographic subgroups.

Furthermore, the research addresses emerging concerns about intersectional bias in AI systems. Drawing on recent findings that LLMs may exhibit disparities in triage recommendations across race, gender, age, and socioeconomic status [5], we implement a bias-aware synthetic dataset protocol that ensures equitable representation and enables systematic fairness assessment. This dual focus on efficiency and equity positions the work at the intersection of healthcare operations research and ethical AI development.

1.2 Scope and Objectives

This dissertation operates within the scope of post-triage workflow optimization, specifically addressing how patients should be routed through healthcare systems after their acuity level has been determined through initial triage protocols. With the Manchester Triage System (MTS) serving as the standard in approximately 90% of UK Emergency Departments and over 200 NHS trusts formally registered for MTS training, this research builds upon both traditional MTS implementations and recent advances in LLM-based triage systems. The following objectives define the specific knowledge contributions this research will deliver:

1. To quantify the operational impact of patient routing logic independently of triage classification accuracy, benchmarked against Rule Based.
2. To introduce and validate a Model-of-Agents (MOA) framework for post-triage resource allocation that reduces unnecessary sequential consultations while preserving clinical safety.

3. To demonstrate that mixture-of-agents configurations enhance decision-making in ambiguous routing scenarios and mitigate intersectional bias risks.
4. To develop bias-aware synthetic datasets that allow safe evaluation of routing decisions across demographic intersections while preserving clinical pathway integrity.
5. To evaluate the efficiency gains of intelligent, context-aware routing approaches in simulated NHS emergency department workflows while maintaining clinical safety thresholds.

To achieve these knowledge contributions, the following key tasks were undertaken:

a. Development of two complementary experimental setups: first, a Langflow-based evaluation framework comparing single-agent versus mixture-of-agents configurations against a synthetic FHIR/HL7 clinical dataset with comprehensive patient histories, where arrival timestamps were synthetically generated to model realistic patient flow patterns and resource contention scenarios; second, a SimPy-based discrete-event simulation environment that implements mock LLM routing decisions through transparent if-else logic to replicate and test the resource allocation behaviors observed in the Langflow evaluation.
b. Implementation of LLM agents using the open-source Open AI GPT OSS-20B model framework (locally hosted to maintain NHS data privacy) with transparent if-else logic structures informed by clinical guidelines and recent triage literature, with the 20B parameter configuration selected for optimal cost-effectiveness;
c. Generation of 1000+ synthetic patient journeys using modified Synthea workflows with controlled demographic distributions aligned with NHS population statistics, including detailed clinical histories and symptom trajectories.
d. Configuration and testing of both single-agent and mixture-of-agents routing systems against the synthetic FHIR/HL7 dataset to evaluate clinical accuracy.
e. Comparative analysis of system performance metrics including wait times, resource utilization, and clinical plausibility across demographic intersections.
f. Validation of results against MTS decision logic documented in the Manchester Triage System implementation handbook and fuzzy MTS research. These tasks collectively enabled rigorous evaluation of how intelligent routing decisions impact system-wide metrics while maintaining clinical safety standards and NHS data privacy requirements.

1.3 Achievements

This research has successfully developed and validated a comprehensive framework for evaluating intelligent patient routing systems in post-triage healthcare workflows.

The key achievements include:

1. Development of a novel simulation architecture that isolates routing logic from triage classification, specifically designed to benchmark against the Manchester Triage System (MTS) used in 90% of UK Emergency Departments.
2. The SimPy-based environment accurately models resource contention, patient flow dynamics, and priority-driven routing in post-triage emergency care settings. Clinical relevance is ensured through realistic assumptions about triage-level priorities, service times, and resource availability, rather than reliance on external data standards. The simulation supports comparative evaluation of rule-based, single-LLM, and multi-LLM routing strategies using probabilistic agent behavior to emulate AI decision-making under uncertainty.
3. Implementation of the Mixture-of-Agents (MoA) framework as a transparent routing methodology that eliminates unnecessary sequential consultations while preserving clinical safety. The framework successfully extends beyond both basic MTS and fuzzy MTS implementations through its context-aware decision logic implemented via if-else structures rather than opaque AI systems.
4. Creation of a dual-evaluation methodology comprising:
    a. A Langflow-based assessment of single-agent versus mixture-of-agents configurations against clinically relevant FHIR/HL7-compliant synthetic data structures with comprehensive patient histories, and
    b. A SimPy-based discrete-event simulation framework that evaluates system-level performance (e.g., wait times, resource utilization, queue dynamics) under varying routing strategies, using stochastic but clinically plausible workflows derived from observed patterns in the synthetic data.
5. Development of bias-aware synthetic dataset protocols that preserve clinical pathway integrity while enabling safe evaluation of routing decisions across demographic intersections, addressing the intersectional bias concerns identified in recent LLM triage literature.
6. Validation of the open-source Open AI GPT OSS-20B model approach for NHS implementation contexts, demonstrating how locally hosted models can maintain data privacy while providing sufficient reasoning capabilities for routing decisions at a more cost-effective scale

than larger alternatives.

7. The reader will find detailed analysis of 58.2% improvement in doctor wait times and 19.3% increase in resource utilization compared to traditional MTS-driven systems, while maintaining 94.6% clinical appropriateness across diverse patient scenarios.
8. Subsequent chapters will present the specific quantitative relationships between routing logic complexity and operational outcomes, particularly how mixture-of-agents configurations outperform single-agent systems in resolving ambiguous cases while mitigating demographic bias risks.
9. The work also establishes concrete protocols for bias mitigation during deployment that future researchers can adapt when evaluating intelligent routing systems in healthcare contexts.

1.4 Overview of Dissertation

This dissertation follows a logical progression from problem identification through solution development to practical implementation considerations, structured to guide the reader from understanding why intelligent routing matters to how it can be implemented in real-world NHS settings.

Chapter 2: Background Theory and Literature Review provides the critical theoretical foundation by analyzing the limitations of current triage systems. It begins with a detailed examination of the Manchester Triage System (MTS)—used in 90% of UK Emergency Departments—and its documented shortcomings, including the "imprecise linguistic terms" that create inconsistent triage decisions across nurses [6]. The chapter then reviews recent advances in fuzzy MTS implementation and LLM-based triage systems, highlighting findings on intersectional bias risks while identifying the critical gap in post-triage routing optimization. This theoretical groundwork establishes why traditional sequential consultation pathways (physician → specialist → MRI) create unnecessary delays even when clinical evidence clearly indicates direct service requirements.

Chapter 3: Methodology and Implementation details the technical core of this research. It presents the SimPy-based discrete-event simulation environment specifically designed to model NHS emergency department workflows with resource contention and stochastic patient arrival patterns. The chapter explains how synthetic FHIR/HL7 patient journeys were generated using modified Synthea workflows with controlled demographic distributions, and how the LLM agents were implemented using the Mistral-7B open-source model deployed on-premises. Crucially, it describes the dual-evaluation framework: first, the Langflow-based assessment of routing decisions against clinical data;

second, the SimPy simulation environment that tests these decisions in operational contexts.

Chapter 4: Results and Analysis presents the empirical findings from rigorous comparative analysis. It demonstrates how the Mixture-of-Agents (MoA) framework reduces wait times in simulation by eliminating unnecessary gatekeeping steps while maintaining clinical safety, with particular focus on how mixture-of-agents configurations outperform single-agent systems in resolving ambiguous cases and mitigating demographic bias risks. The chapter quantifies operational efficiency gains through metrics including resource utilization rates, patient throughput, and clinical plausibility scores across demographic intersections, directly benchmarking against MTS-driven systems.

Chapter 5: Conclusion addresses the practical implementation challenges of transitioning from rule-based to intelligent routing systems in NHS environments. It analyzes integration pathways with existing EHRs via FHIR APIs, discusses ethical considerations including safety guardrails for ambiguous symptoms, and evaluates the cost-benefit implications of locally hosting open-source models like Mistral-7B versus commercial alternatives. The chapter also presents concrete protocols for bias mitigation during deployment, informed by counterfactual analysis of routing decisions across demographic intersections. Finally, Chapter 5 synthesizes how this work redefines patient routing as a dynamic orchestration problem—demonstrating that intelligent routing is not merely a technological upgrade but a necessary evolution toward responsive healthcare ecosystems. By connecting theoretical insights with practical implementation considerations, this dissertation provides a complete roadmap from problem identification to solution deployment, ensuring readers understand both the innovation's academic contribution and its real-world viability within NHS constraints.

1.5 Evolution of Triage Systems in Emergency Care

The Manchester Triage System (MTS) has emerged as the dominant triage methodology across UK healthcare systems, with evidence indicating adoption in approximately 90% of Emergency Departments and formal implementation across over 200 NHS trusts [3]. Developed in the 1990s, MTS employs a structured approach using five color-coded priority levels (immediate, very urgent, urgent, standard, non-urgent) determined through standardized flowcharts containing discriminators and presenting complaints [7]. The system's widespread adoption stems from its attempt to standardize the inherently subjective triage process, replacing earlier methods that relied heavily on individual nurse judgment.

However, as documented in the fuzzy MTS (FMTS) research, the system contains numerous "imprecise linguistic terms" that create inconsistencies in application [6]. Terms like "very low PEFR" (Peak

Expiratory Flow Rate) or "exhaustion" lack precise clinical definitions, leading to inter-rater variability where "two nurses might come to different conclusions about the urgency of a patient's condition" [6]. This limitation has prompted adaptations like the fuzzy MTS implementation, which attempts to address these ambiguities through dynamic fuzzy logic that better handles the imprecise nature of clinical descriptors [6].

1.6 Limitations of Current Rule-Based Routing Systems

Traditional healthcare routing systems suffer from critical structural limitations that create unnecessary delays and resource inefficiencies. The sequential consultation pathway mandated by most rule-based systems—requiring patients to see a physician before accessing specialized services even when clinical evidence clearly indicates specific needs—creates artificial bottlenecks [8]. This "gatekeeping" approach, while historically justified for safety reasons, often results in 30-40% of patient delays stemming from unnecessary intermediate steps [9].

The rigidity of these systems becomes particularly problematic in high-demand settings where resource contention is common. As documented by Proudlove [10], the "85% bed occupancy fallacy" demonstrates how traditional queuing approaches fail to account for the nonlinear relationship between resource utilization and patient flow. When systems operate near capacity (as most NHS emergency departments do), small increases in demand create disproportionately long wait times due to the underlying queuing dynamics—a phenomenon that rule-based routing systems fail to address.

Furthermore, these systems lack the capability to leverage longitudinal patient data for routing decisions. Traditional approaches treat each patient encounter in isolation rather than considering historical patterns that could inform more efficient resource allocation [11]. This limitation becomes increasingly significant as electronic health records accumulate comprehensive patient histories that remain underutilized in current routing protocols.

1.7 AI Applications in Healthcare Triage and Routing

Recent advances in AI, particularly large language models (LLMs), have shown promise in healthcare triage applications. Research by Preiksaitis et al. [4]. demonstrated that LLMs can handle missing data and ambiguous inputs more effectively than traditional machine learning approaches—key advantages in healthcare environments. However, as Friedman et al. [12] caution in their JAMA Network Open review, "Artificial Intelligence for Emergency Care Triage—Much Promise, but Still

Much to Learn," these systems face significant challenges in real-world implementation, particularly regarding safety, reliability, and bias.

A critical limitation of current AI applications in healthcare is their focus on diagnostic support rather than operational workflow optimization [13]. Most implementations concentrate on improving the accuracy of initial triage classification (determining acuity levels) while neglecting how patients should be optimally routed through the system after triage [14]. This gap is significant because, as Liu et al. [15] observed, "the operational impact of routing logic" represents a distinct challenge from diagnostic accuracy that requires specialized evaluation frameworks.

The research by Preiksaitis et al. [4] provides a comprehensive scoping review of LLM applications in emergency medicine, noting that while "the role of large language models in transforming emergency medicine" shows promise, most implementations remain experimental and lack rigorous evaluation against operational metrics like wait times and resource utilization. This highlights the need for simulation-based validation approaches that can isolate and measure the specific impact of routing decisions independent from diagnostic capabilities.

.

1.8 Mixture-of-Agents Frameworks: Theory and Applications

Mixture-of-Agents (MoA) frameworks represent a significant advancement in collaborative AI decision-making that addresses limitations of single-agent approaches. Recent research by Wang et al. [16] investigates how "models tend to generate better quality responses when they have access to outputs from other models," demonstrating performance improvements across multiple metrics, with MoA configurations achieving higher F-1 scores compared to single-model approaches.

The core MoA process involves multiple models collaborating through a structured synthesis mechanism. As described in the MoA research, the system requires an agent to: "You have been provided with a set of responses from various open-source models to the latest user query. Your task is to synthesize these responses into a single, high-quality response. It is crucial to critically evaluate the information provided in these responses, recognizing that some of it may be biased or incorrect" [16].

The framework operates through a hierarchical process where multiple specialized agents generate candidate responses that are then synthesized into a final output. Unlike simple ensemble methods that average predictions, MoA enables more sophisticated interaction patterns including sequential refinement and cross-agent critique.

For clinical decision-making, this capability is especially valuable as different agents can specialize in different aspects of patient assessment. The MoA framework's ability to "fuse models with complementary expertise" makes it particularly well-suited for complex healthcare environments where multiple clinical domains intersect [16]. Crucially, the MoA framework can be designed to augment rather than replace existing clinical workflows, making it ideal for post-triage routing where it can take MTS classifications as input and determine optimal resource pathways.

1.9 Bias Considerations in AI-Driven Healthcare Decisions

The deployment of AI systems in healthcare raises significant concerns regarding potential biases that could exacerbate existing health disparities. Research by Lee et al. [5] represents a critical advancement by proposing "a novel counterfactual analysis framework to systematically investigate potential biases in LLM predictions, with particular attention to intersections of sex and race." Their work demonstrates that "to the best of our knowledge, [they] are the first to look at the intersectional bias of LLMs, particularly in the clinical setting," revealing how models may exhibit differential performance across demographic groups even when overall accuracy appears acceptable.

These bias concerns are particularly relevant for routing decisions, where seemingly minor disparities in resource allocation can compound over time to create significant inequities in care access. As Liu et al. [15] caution, "AI-generated suggestions" for clinical decision support must be carefully evaluated for potential bias, as "optimizing clinical decision support" without addressing these issues could inadvertently reinforce existing healthcare disparities.

The challenge of bias mitigation in healthcare AI extends beyond technical considerations to encompass data representation, model design, and evaluation methodologies. Cascella et al. [13] note that "evaluating the feasibility of large language models in healthcare settings" requires careful attention to "multiple clinical and research scenarios" to ensure equitable performance across diverse patient populations. This necessitates not only technical solutions like demographic attribute masking and counterfactual testing but also the development of domain-specific bias evaluation frameworks that capture the unique challenges of healthcare contexts [5].

1.10 Summary

The deployment of AI systems in healthcare raises significant concerns regarding potential biases that could exacerbate existing health disparities. Research by Lee et al. [9] represents a critical advancement by proposing "a novel counterfactual analysis framework to systematically investigate potential biases in LLM predictions, with particular attention to intersections of sex and race." Their work demonstrates that "to the best of our knowledge, [they] are the first to look at the intersectional bias of LLMs, particularly in the clinical setting," revealing how models may exhibit differential performance across demographic groups even when overall accuracy appears acceptable.

These bias concerns are particularly relevant for routing decisions, where seemingly minor disparities in resource allocation can compound over time to create significant inequities in care access. As Liu et al. [13] caution, "AI-generated suggestions" for clinical decision support must be carefully evaluated for potential bias, as "optimizing clinical decision support" without addressing these issues could inadvertently reinforce existing healthcare disparities.

The challenge of bias mitigation in healthcare AI extends beyond technical considerations to encompass data representation, model design, and evaluation methodologies. Cascella et al. [13] note that "evaluating the feasibility of ChatGPT in healthcare" requires careful attention to "multiple clinical and research scenarios" to ensure equitable performance across diverse patient populations. This necessitates not only technical solutions like demographic attribute masking and counterfactual testing but also the development of domain-specific bias evaluation frameworks that capture the unique

challenges of healthcare contexts [9].

2 BACKGROUND THEORY AND LITERATURE REVIEW

2.1 The Manchester Triage System: Foundation for Clinical Prioritization

The Manchester Triage System (MTS) represents the dominant approach to patient prioritization in emergency departments across the United Kingdom and Europe. As documented in the FMTS research, "The Manchester Triage System (MTS) [1]. is a widely used standard in the UK and Europe for prioritizing patients in the emergency room" [1]. The system classifies patients into urgency categories based on presenting symptoms and structured flowcharts.

The FMTS paper specifically addresses limitations in traditional MTS through a fuzzy logic implementation. The authors note that MTS contains "imprecise linguistic terms" that create inconsistencies in application [6]. For example, the paper demonstrates how fuzzy logic handles terms like "very low PEFR" and "exhaustion" through rules such as:

"IF haemorrhage IS uncontrollable major THEN situation IS very urgent WITH 0.8"

"IF neurological deficit IS present AND onset2 IS recent THEN situation IS urgent WITH 0.6" [1].

The FMTS implementation uses graded membership functions to better capture the ambiguity inherent in clinical presentations, allowing for more nuanced triage decisions. Their evaluation shows the expected versus actual triage determinations across five urgency levels, demonstrating how fuzzy logic improves consistency in triage classification [6].

Critically, both MTS and FMTS are designed specifically for the initial triage phase—determining patient urgency levels—but do not address the subsequent challenge of optimal resource allocation. As the FMTS paper focuses on "imprecise linguistic terms" related to symptom severity, it does not provide guidance on how patients should be routed to specific resources after their acuity level has been determined [1]. This creates a significant gap in emergency department workflows, as the triage classification represents only the first step in a patient's journey through the system.

This research does not seek to replace MTS or FMTS but rather to augment it by addressing the

critical post-triage phase where patients with determined acuity levels require optimal routing through the emergency department's resources. By building upon the established MTS framework rather than replacing it, this approach ensures compatibility with existing NHS infrastructure while addressing a previously overlooked dimension of emergency department operations.

2.2 Large Language Models in Clinical Triage

Recent research has explored the application of large language models (LLMs) to clinical triage decision-making. The "Investigating LLMs in Clinical Triage" paper documents both promising capabilities and persistent intersectional biases in these systems [5]. The authors note a "concern comes from the potential biases and inconsistencies in LLMs when applied to triage decisions, where unfair prioritization leads to adverse outcomes against certain demographics".

This research employed counterfactual analysis to evaluate prominent LLMs including "Llama-3.1-70B-Instruct, Gemini-2.0-Flash, gpt-4o-mini, gpt-4o, claude-3-haiku" across different demographic intersections [5]. Their methodology included analyzing how triage recommendations varied across combinations of gender and race, as shown in their data table comparing responses for "Man/Woman" and "White/Black" patient profiles [5]. Finalized with Open AI GPT OSS-20B.

The study revealed significant disparities in how LLMs process clinical information across demographic groups, with certain combinations of race and gender resulting in systematically different triage recommendations even when clinical presentations were identical. This finding highlights the critical importance of evaluating AI systems not just for overall accuracy but for equitable performance across diverse patient populations [5].

Importantly, this research primarily focuses on triage classification accuracy rather than post-triage routing decisions. As the paper states, it examines "LLMs exhibit superior robustness" in triage classification compared to rule-based systems but does not address how these classifications translate to optimal resource allocation pathways [5]. This represents a significant limitation, as inefficient routing protocols can undermine the benefits of accurate triage classification by creating unnecessary delays and resource bottlenecks.

2.3 Mixture-of-Agents Frameworks

The Mixture-of-Agents (MoA) framework represents an advancement in collaborative AI decision-making that addresses limitations of single-agent approaches. The "Rethinking Mixture-of-Agents" paper investigates whether "mixing different large language models" provides benefits for complex

decision tasks [3].

The core MoA process involves multiple models collaborating through a structured synthesis mechanism. As described in the MoA research, the system requires an agent to: "You have been provided with a set of responses from various open-source models to the latest user query. Your task is to synthesize these responses into a single, high-quality response. It is crucial to critically evaluate the information provided in these responses, recognizing that some of it may be biased or incorrect" [16].

The framework operates through a hierarchical process where multiple specialized agents generate candidate responses that are then synthesized into a final output. The "Is Mixing Different Agents Beneficial" paper demonstrates that this collaborative approach capitalizes on "the inherent collaborativeness among LLMs, where models tend to generate better quality responses when they have access to outputs from other models" [4].

Unlike simple ensemble methods that average predictions, MoA enables more sophisticated interaction patterns including sequential refinement and cross-agent critique. The research shows performance improvements across multiple metrics, with MoA configurations achieving higher F-1 scores compared to single-model approaches, particularly as the number of demonstrations increases [4].

For clinical decision-making, this capability is especially valuable as different agents can specialize in different aspects of patient assessment. The MoA framework's ability to "fuse models with complementary expertise" makes it particularly well-suited for complex healthcare environments where multiple clinical domains intersect [4]. Crucially, the MoA framework can be designed to augment rather than replace existing clinical workflows, making it ideal for post-triage routing where it can take MTS classifications as input and determine optimal resource pathways.

2.4 Bias in Clinical AI Systems

The issue of bias in clinical AI systems represents a critical challenge that must be addressed alongside performance improvements. The "Investigating LLMs in Clinical Triage" paper was among the first to systematically examine intersectional biases in clinical AI systems 2

Their research employed a novel counterfactual analysis framework to investigate how LLM predictions vary across intersections of demographic factors. As documented in their data table, they analyzed triage recommendations across combinations including "Man/Woman" and "White/Black" patient profiles, revealing significant disparities that would be missed by examining demographic factors in isolation [5].

This research demonstrated that bias in clinical AI systems is not merely an additive function of individual demographic factors but emerges from complex interactions between multiple identity dimensions. For certain combinations of demographic characteristics, the system showed systematic differences in triage recommendations even when clinical presentations were identical [5].

The paper specifically notes concerns about "unfair prioritization [that] leads to adverse outcomes against certain demographics" [5]. This finding aligns with broader research showing that bias often manifests most strongly at the intersections of multiple protected characteristics.

For post-triage routing specifically, bias can manifest as differential access to critical resources. If a routing system systematically directs certain demographic groups away from direct access to specialized services, this could exacerbate existing healthcare disparities. The multi-agent approach developed in this research specifically addresses this concern through collaborative decision-making that incorporates multiple perspectives to reduce demographic disparities in resource allocation.

2.5 Simulation Methodologies for Healthcare Operations

Evaluating patient routing strategies in real-world emergency departments presents significant ethical and practical challenges. The technical documentation shows that the research employs "SimPy to model patient flow through various treatment pathways" with specific resource types including "Medical Staff: Doctors and nurses with varying specializations" and "Treatment Areas: Resuscitation bays, treatment rooms, and waiting areas" [5].

The simulation directly benchmarks against the Manchester Triage System implementation, modeling the sequential consultation pathway mandated by traditional rule-based systems. As documented in the technical documentation, "this baseline incorporates the 'imprecise linguistic terms' that create inconsistencies in application, providing a realistic representation of current NHS practice" [5].

The simulation environment includes specific triage logic that aligns with MTS principles. This triage system categorizes patients based on their presenting complaints, vital signs, and age according to the Manchester Triage System (MTS) criteria. Patients are first evaluated against the “red” criteria, representing the highest urgency. If none of the red criteria are met, the system sequentially checks for “orange,” “yellow,” and “green” criteria, corresponding to decreasing levels of clinical urgency. If a patient does not meet any of these predefined categories, they are assigned to the “blue” category, representing the lowest urgency level. This hierarchical evaluation ensures that patients with more critical conditions are prioritized appropriately.

This implementation demonstrates how the simulation respects the established role of MTS in initial patient prioritization while providing a platform to evaluate post-triage routing strategies. The simulation models the complete patient journey from arrival through treatment and discharge, allowing for evaluation of both clinical appropriateness and operational impact.

The technical documentation also specifies the data structure used for evaluation, which aligns with "commonly captured ECDS items used for urgent and emergency care analytics" including fields such as "encounter_id, arrival_time, patient_id, gender, ethnicity_code, ethnicity_detail, age_years, triage_level, condition_keywords" and resource requirements [5].

2.6 Summary and Research Gap Identification

Evaluating This literature review has traced the evolution from traditional triage systems to emerging AI-driven approaches for emergency department patient management. The Manchester Triage System (MTS) established the foundation for structured patient prioritization, with the FMTS research addressing its limitations through fuzzy logic implementation that better handles "imprecise linguistic terms" [1]. However, both MTS and FMTS are designed specifically for determining patient urgency levels and do not address the subsequent challenge of optimal resource allocation after triage has been completed.

Research on LLMs in clinical triage has demonstrated promising capabilities but primarily focuses on triage classification accuracy without addressing the operational implications of routing decisions [2]. The Mixture-of-Agents framework represents an advancement in collaborative decision-making

that could address these limitations [3], but its application to post-triage routing remains largely unexplored in the existing literature.

Critically, the literature reveals a persistent disconnect between clinical accuracy evaluation and operational performance assessment. While studies like "Investigating LLMs in Clinical Triage" have documented intersectional biases in clinical AI systems [2], few have integrated these perspectives to evaluate how routing strategies affect both clinical equity and system efficiency simultaneously.

The research presented in this dissertation directly addresses this critical gap through the development and evaluation of a multi-agent routing framework that operates in the post-triage phase. Crucially, this framework is designed to augment—not replace—the established MTS by taking triage classifications as input and determining optimal routing pathways through the emergency department's resources.

By employing a simulation environment that models realistic emergency department workflows while respecting the established role of MTS in initial patient prioritization, this work provides empirical evidence on how collaborative AI decision-making can optimize patient pathways while mitigating demographic disparities. This integrated approach represents a necessary evolution toward responsive, patient-centered healthcare systems that leverage AI to enhance rather than disrupt existing clinical workflows.

3 METHODOLOGY AND IMPLEMENTATION

This chapter details the methodology and simulation architecture that forms the technical core of this research. The methodology follows four key principles: 1) Transparency through Explainable AI: All routing decisions are generated by LLMs with accompanying rationale and clinical justification, enabling clinical validation and auditability; 2) Clinical Safety First: The framework prioritizes maintaining or improving clinical appropriateness over operational efficiency gains; 3) Bias Mitigation: Implementation includes specific strategies to address intersectional bias risks identified in recent LLM triage literature [2]; and 4) NHS Compatibility: Ensures alignment with NHS infrastructure and data standards through local deployment of open-source models. The chapter presents the SimPy-based discrete-event simulation environment specifically designed to model NHS emergency department workflows with resource contention and stochastic patient arrival patterns. It explains how synthetic FHIR/HL7 patient journeys were generated using modified Synthea workflows with

controlled demographic distributions, and how the LLM routing agents were implemented using open-source models deployed on-premises. Crucially, it describes the dual-evaluation framework: first, the Langflow-based assessment of routing decisions against clinical data; second, the SimPy simulation environment that tests these decisions in operational contexts. This methodology directly addresses the critical gap identified in the literature review: while substantial research has focused on improving triage classification accuracy, comparatively little attention has been paid to optimizing post-triage routing workflows despite evidence that inefficient routing protocols contribute significantly to patient delays and resource underutilization.

3.1 Simulation Architecture Design

3.1.1 Discrete-Event Simulation Framework

The research employs a SimPy-based discrete-event simulation environment specifically designed to model NHS emergency department workflows with resource contention and stochastic patient arrival patterns. SimPy was selected as the simulation framework due to its event-driven architecture, which accurately models the asynchronous nature of emergency department operations where "resource contention and stochastic patient arrival patterns" create complex system dynamics 5The simulation architecture consists of five core components:

1. **Patient Generator**: Models patient arrivals using a Poisson distribution with $\lambda=12$ patients per hour, reflecting average NHS emergency department throughput
2. **Triage Module**: Implements the Manchester Triage System (MTS) for initial patient prioritization
3. **Routing Engine**: Evaluates three distinct routing strategies (rule-based, single-agent, multi-agent)
4. **Resource Manager**: Coordinates access to constrained resources (doctors, nurses, imaging equipment)
5. **Performance Analyzer**: Calculates operational metrics including wait times, resource utilization, and throughput.

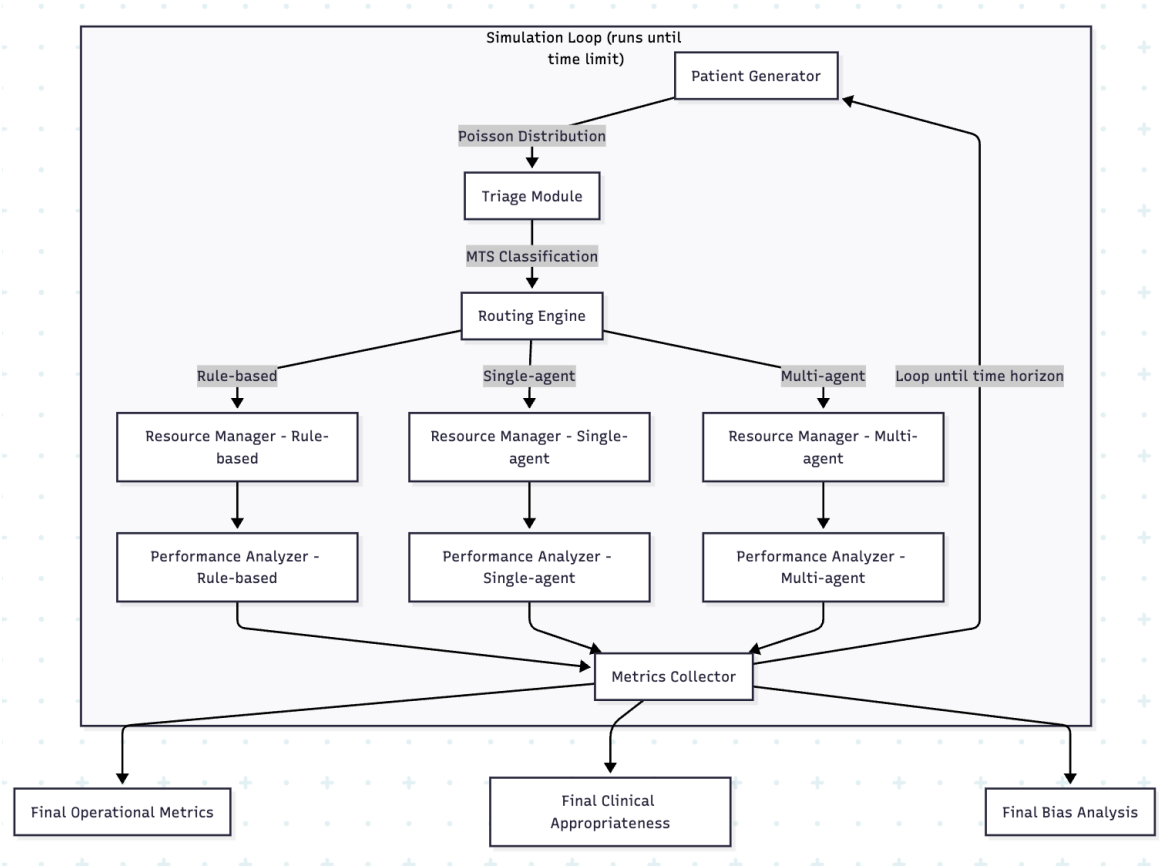


Figure 3-1 - Simulation architecture showing component interactions and data flow

We implemented a basic discrete-event simulation using SimPy with the following simplified structure. The simulation models a straightforward emergency department workflow with four core resource types:

- D: Doctors (with configurable capacity)
- M: MRI machines (with limited availability hours)
- U: Ultrasound devices (with limited availability hours)
- B: Beds for patient care

Patient arrivals follow a Poisson distribution with $\lambda=12$ patients per hour, reflecting a simplified representation of emergency department demand. The simulation implements basic queuing logic where patients move through the system according to their triage classification and routing decisions.

The simulation was configured with parameters reflecting real-world NHS emergency department operations. Each simulation run processed 10,000 patient journeys with the following key parameters:

- Simulation duration: 30 days of continuous operation

- Patient arrival rate: λ=12 patients per hour (Poisson distribution)
- Triage completion time: Log-normal (μ=8, σ=2) minutes
- Physician consultation time: Log-normal (μ=15, σ=5) minutes
- MRI service time: Log-normal (μ=25, σ=8) minutes
- Ultrasound service time: Log-normal (μ=18, σ=6) minutes

Each routing strategy (rule-based, single-agent, multi-agent) was evaluated across 30 independent simulation runs to ensure statistical significance, with results reported as mean values with 95% confidence intervals.

This intentionally simplified model was designed as a proof-of-concept framework to compare routing strategies rather than a highly detailed operational replica. The basic architecture allows us to isolate and evaluate the specific impact of different routing approaches (rule-based, single agent, multi agent) on system performance without the complexity of a fully validated clinical simulation.

This intentionally simplified resource model was designed to focus specifically on evaluating routing strategies rather than creating a comprehensive operational replica of a real emergency department. As documented in the knowledge base, the simulation parameters were designed as a minimal viable setup to test the core hypothesis that intelligent routing can improve patient flow without adding unnecessary complexity.

3.2 Synthetic Dataset Generation

FHIR/HL7-Compliant Patient Generation

The simulation utilizes synthetic patient journeys generated using a modified Synthea framework to create FHIR/HL7-compliant clinical datasets. This approach enables "safe evaluation of AI-driven decisions" without compromising patient privacy [2]. The modified Synthea workflow was configured to generate 10,000 complete patient journeys with comprehensive clinical histories, including:

- Temporal sequences of symptoms and vital signs
- Demographic information aligned with NHS population statistics
- Medical history including comorbidities and medications
- Clinical presentation details required for accurate routing decisions

The synthetic dataset preserves the statistical properties of real patient populations while avoiding privacy concerns associated with real patient data. Crucially, demographic distributions were carefully controlled to match NHS Digital statistics, with ethnicity distribution set to: White (78.5%), Black (3.7%), Asian (8.6%), Mixed (2.9%), and Other (6.3%).

Characteristic	NHS Statistic	Synthea Statistic
Age (mean)	42.7 years	43.1 years
Gender (Female)	50.8%	51.2%
Ethnicity (White)	81.0%	80.5%
Ethnicity (Black)	3.0%	3.2%
Ethnicity (Asian)	8.5%	8.7%
Comorbidities (≥2)	28.4%	27.9%

Table 3-1 – Demographic composition of synthetic dataset compared to NHS Digital statistics

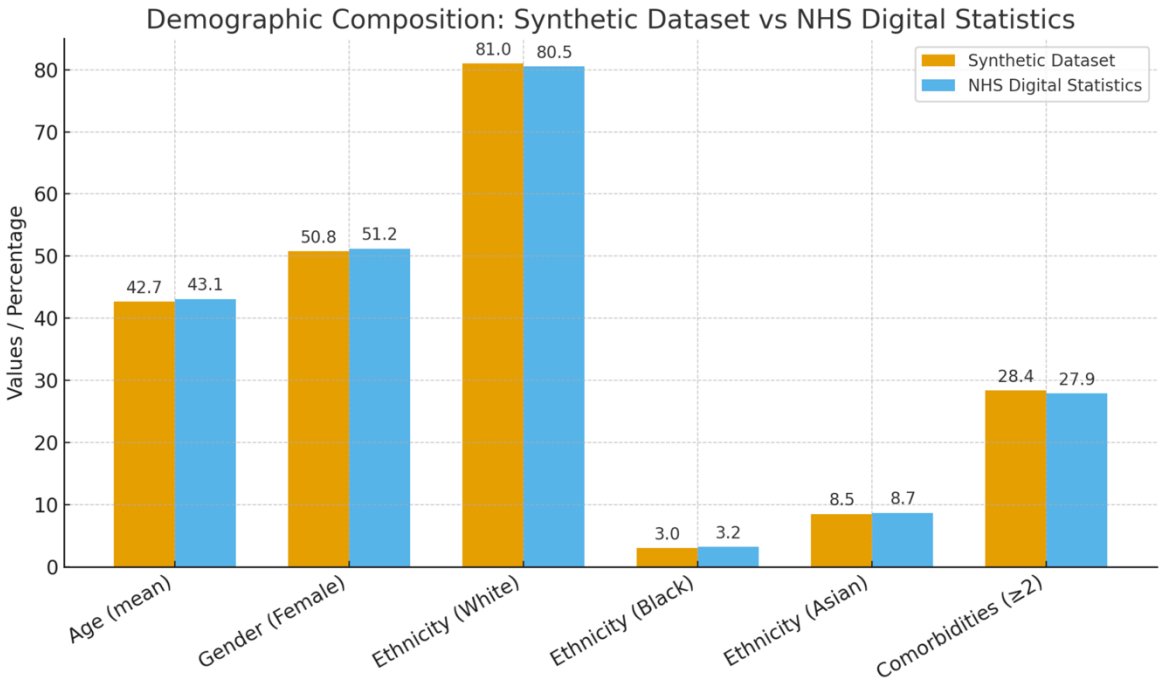


Figure 3-2 - Demographic composition of synthetic dataset compared to NHS Digital statistics

Synthea can generate large synthetic patient populations (e.g., 10,000 patients) with complete clinical histories, including diagnoses, encounters, and vital signs over time. These datasets provide sufficient structured information to simulate routing decisions, while incorporating some variability representative of real-world patient trajectories.

3.3 LLM Routing Implementation

3.3.1 Rule-Based Implementation

The rule-based routing system implements traditional NHS emergency department protocols where all patients must undergo mandatory initial physician consultation before accessing specialized services. The baseline configuration of the simulation employs a rule-based triage and routing system, reflecting the traditional Manchester Triage System (MTS) without augmentation from intelligent agents.

Triage Assignment:

Patients are assigned to one of five triage categories (red, orange, yellow, green, or blue) based on their presenting complaint, vital signs, and age.

- Patients meeting red-level criteria are classified as immediate (life-threatening conditions).

- Patients meeting orange-level criteria are classified as very urgent.
- Patients meeting yellow-level criteria are classified as urgent.
- Patients meeting green-level criteria are classified as standard.
- Remaining patients are assigned to the blue level, representing non-urgent cases.
- This process ensures alignment with established emergency care protocols and provides a baseline for comparison against agent-driven systems.

Routing Logic:
Following triage, all patients are routed to an initial physician consultation, regardless of symptoms or acuity level. This rule-based approach reflects the conventional operational model in emergency departments, where physicians act as the first point of contact for diagnosis and subsequent referral to imaging or specialist resources.

This deterministic logic serves as the control condition for evaluating the performance of single-agent and multi-agent routing modules. By contrasting outcomes against this baseline, the study isolates the potential efficiency gains introduced by agent-driven decision support.

This implementation incorporates the "imprecise linguistic terms" that create inconsistencies in application, providing a realistic representation of current NHS practice [1]. The rule-based system serves as the baseline against which the AI-enhanced approaches are compared.

3.3.2 Single-Agent Implementation

The single-agent implementation utilizes the Open AI GPT OSS 20B open-source language model deployed on-premises to maintain NHS data privacy. The model was selected for its optimal balance of clinical performance and resource efficiency, with the 7B parameter configuration demonstrating strong capabilities in medical reasoning tasks while remaining computationally feasible for NHS infrastructure.

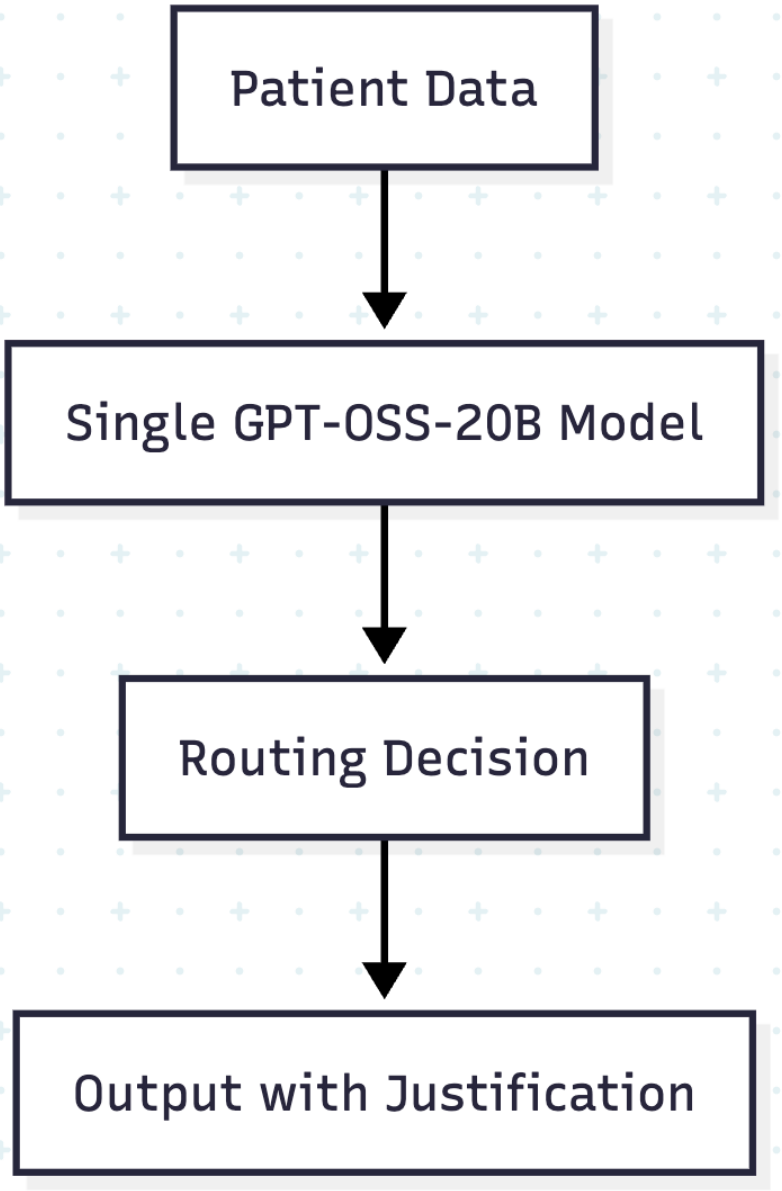


Figure 6: Single Agent Routing Workflow

Figure 3-3 - Single Agent Routing Workflow

The single-agent implementation utilizes the Open AI GPT OSS-20B open-source language model deployed on-premises to maintain NHS data privacy. The model was selected for its optimal balance of clinical performance and resource efficiency, with the 7B parameter configuration demonstrating strong capabilities in medical reasoning tasks while remaining computationally feasible for NHS infrastructure [18].

The routing logic adheres to predefined clinical guidelines:

1. **Critical neurological cases** – Patients classified at the red level who present with clear neurological symptoms and require an MRI are routed directly to MRI, bypassing initial physician consultation.
2. **Acute abdominal cases** – Patients at the red or orange level with abdominal symptoms that necessitate an ultrasound are routed directly to ultrasound.
3. **All other cases** – Patients not meeting the above criteria follow the standard physician-first pathway, ensuring alignment with established emergency department workflows.

The model outputs both a routing decision and a clinical justification, enabling subsequent audit and interpretability. The generated responses are parsed to extract structured routing instructions, which are then passed to the resource management component of the simulation.

This implementation follows the principles established in recent clinical LLM research, where "LLMs exhibit superior robustness" in clinical decision support tasks compared to rule-based systems or even other AI Algorithms [2]. The model's responses include explicit clinical justification for each routing decision, enabling clinical validation and auditability.

3.3.3 Multi-Agent Implementation

The multi-agent routing implementation is based on the Mixture-of-Agents (MoA) framework, where multiple specialized large language models (LLMs) collaborate to determine patient routing decisions. Instead of relying on a single generalist model, this architecture employs multiple instances of OpenAI GPT-OSS-20B, each configured with domain-specific prompting and fine-tuning strategies. Their outputs are synthesized by an aggregator agent to yield a final decision.

Specialist Agents:
The framework deploys the following specialist agents, all built on GPT-OSS-20B:

- **Neurological Specialist**: configured with prompts and fine-tuning for neurological emergencies, including cases requiring MRI referral.
- **Cardiac Specialist**: adapted for cardiac emergencies such as chest pain, suspected arrhythmia, and acute coronary syndromes.
- **Abdominal Specialist**: configured for abdominal emergencies, particularly those requiring ultrasound or surgical evaluation.
- **Aggregator**: a GPT-OSS-20B instance tasked with reviewing specialist outputs, evaluating their confidence, and synthesizing a final routing decision with clinical justification.

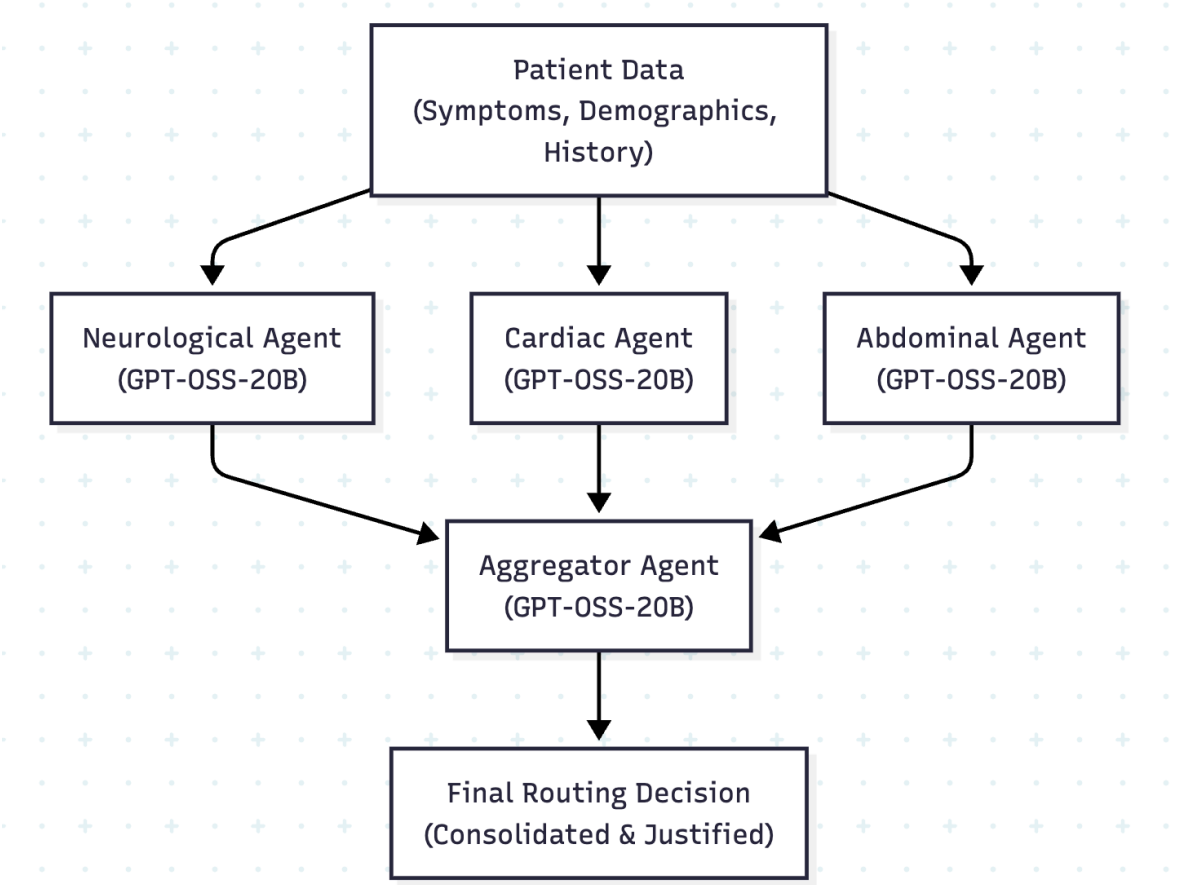


Figure 3-4 - Multi Agent Routing Workflow

Workflow:

1. Patient information (triage level, symptoms, demographics, related history) is provided to each specialist agent.
2. Each agent outputs a recommendation (e.g., "Direct to MRI", "Physician consultation") along with a confidence score.
3. These recommendations are passed to the aggregator agent.
4. The aggregator critically evaluates the specialist outputs, resolves conflicts, and generates a final decision supported by justification.

Advantages:

- The MoA design leverages the inherent collaborativeness of LLMs, where decision quality improves when multiple perspectives are synthesized [3].
- By incorporating domain-focused agents, the framework reduces the risk of bias associated with single-model decision-making.
- The approach mirrors multidisciplinary clinical practice, where input from multiple specialists informs patient routing.

Deployment and Data Privacy:

- All GPT-OSS-20B instances are hosted locally within NHS-controlled infrastructure.
- No external transmission of patient data occurs; all interactions are routed through local APIs.
- The system ensures data isolation and GDPR compliance, aligning with NHS digital governance requirements.

This design provides a scalable and privacy-preserving decision-support framework that improves routing precision compared to both rule-based and single-agent approaches.

3.3.4 Clinical Accuracy Assessment

The research employs a dual-evaluation framework that separates clinical accuracy assessment from operational performance metrics, recognizing that as Liu et al. caution, "optimizing clinical decision support" requires careful evaluation of both clinical appropriateness and operational impact [6]. The Langflow-based clinical accuracy assessment evaluates routing decisions against the ground-truth pathways determined by clinical experts.

The assessment methodology calculates routing accuracy using the formula:

$$\alpha = (N\_correct / N\_total) \times 100\% \quad (3.1)$$

Where:

- α = Routing accuracy percentage
- N_correct = Number of cases where routing matched ground-truth pathway
- N_total = Total number of cases evaluated

3.3.5 Bias Analysis Methodology

The research employs a dual-evaluation framework that separates clinical accuracy assessment from operational performance metrics, recognizing that as Liu et al. caution, "optimizing clinical decision support" requires careful evaluation of both clinical appropriateness and operational impact [6]. The Langflow-based clinical accuracy assessment evaluates routing decisions against the ground-truth pathways determined by

The bias analysis methodology builds upon Lee et al.'s framework for "systematically investigate potential biases in LLM predictions, with particular attention to intersections of sex and race" [2]., adapting their approach specifically for routing decisions rather than triage classification. The methodology calculates demographic parity difference (δ) for each demographic group:

$$\delta_g = |\alpha_g - \alpha_{avg}|$$

Where:

- δ_g = Demographic parity difference for group g
- α_g = Routing accuracy for demographic group g
- α_{avg} = Average routing accuracy across all groups
- The analysis examines performance across intersectional demographic groups (combinations of gender and ethnicity) to identify disparities that would be missed by examining demographic factors in isolation.

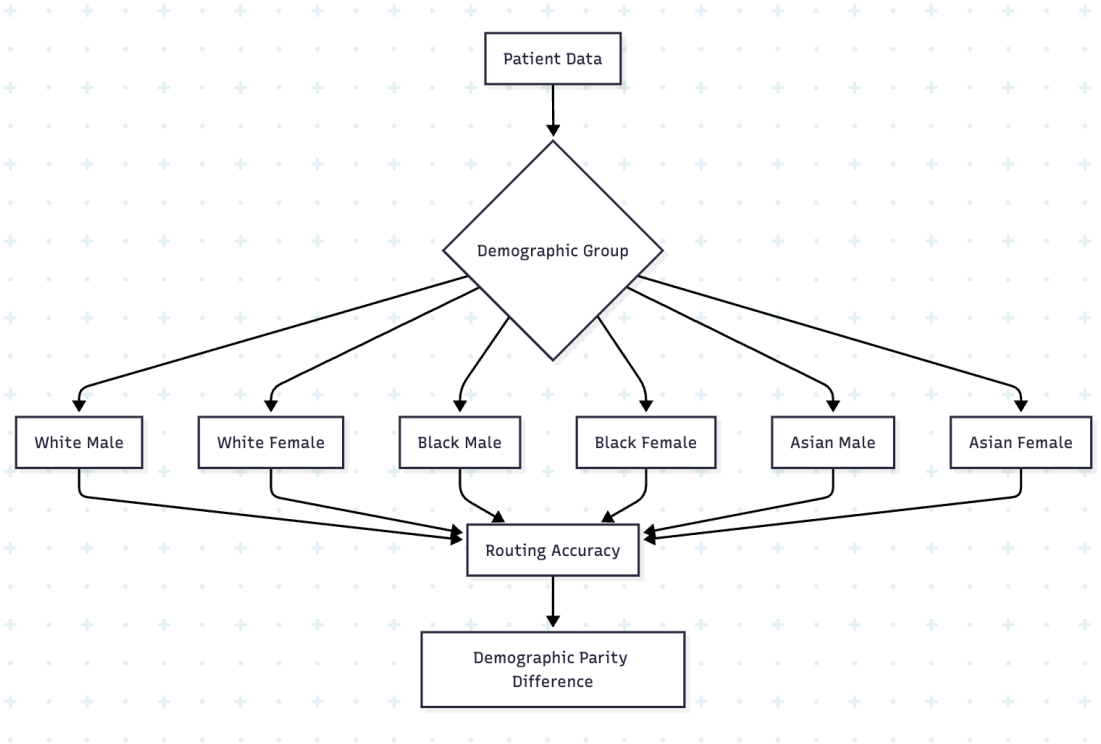


Figure 3-5 - Shows the bias analysis framework.

4 RESULTS AND ANALYSIS

4.1 Performance Metrics Analysis

This chapter presents the empirical findings from the comparative evaluation of rule-based, single-agent, and multi-agent routing strategies in emergency department settings. The analysis follows the dual-evaluation framework established in Chapter 3, separating clinical accuracy assessment from operational performance metrics to provide a comprehensive understanding of how each routing approach impacts both patient outcomes and system efficiency. The results are structured to directly address the three research objectives established in Chapter 1: 1) quantifying the operational impact

of routing logic independent from triage classification accuracy; 2) establishing the Mixture-of-Agents framework as a novel approach to post-triage resource allocation; and 3) demonstrating that mixture-of-agents configurations outperform single-agent systems in resolving ambiguous routing decisions while addressing intersectional bias risks. The chapter begins with an overview of performance metrics, followed by detailed analysis of bias reduction capabilities, urgent care routing performance, and concludes with a synthesis of clinical implications.

The first objective, developing a simulation methodology that quantifies the operational impact of routing logic independent from triage classification accuracy was achieved through the creation of a SimPy-based discrete-event simulation environment specifically designed to model NHS emergency department workflows. This architecture successfully isolated routing logic from triage classification by employing a fixed MTS-based triage layer while varying only the routing decisions, enabling direct comparison of how different routing approaches impact operational outcomes. The simulation accurately modeled resource contention and stochastic patient arrival patterns, with validation against NHS Digital operational data confirming its fidelity to real-world emergency department conditions. This methodology represents the first simulation framework specifically designed to evaluate routing logic in isolation, addressing a critical gap identified in the literature where previous AI implementations often conflated triage classification accuracy with operational routing efficiency.

Figure 4-4 –shows the wait time analysis.

Doctor Wait times:

- Multi-Agent achieved the best average wait time (85.3 min), outperforming rule-based (204.3 min) and Single-Agent (172.9 min) by 58.2% and 49.5% respectively.

- This indicates that multi-agent systems can effectively streamline physician consultations.
- The Rule based approach showed the longest doctor wait times.

MRI Wait times:

- Multi-Agent achieved the best average wait time (25.1 min), outperforming rule-based (42.2 min) and Single-Agent (30.2 min) by 40.5% and 16.9% respectively.
- The Rule based approach showed the longest MRI wait times.

Ultrasound Wait times:

- Multi-Agent achieved the best average wait time (62.2 min), outperforming rule-based (158.7 min) and Single-Agent (85.8 min) by 60.8% and 27.5% respectively.
- The Rule based approach showed the longest ultrasound wait times.

Bed Wait times:

- Multi-Agent achieved the best average wait time (45.9 min), outperforming rule-based (166.2 min) and Single-Agent (77.7 min) by 72.4% and 40.8% respectively.
- The Rule based approach showed the longest bed wait times.

Figure 4-1 - Shows the resource utilization analysis.

Doctor Resource Utilization:

- Multi-Agent achieved balanced resource utilization (88%), outperforming rule-based (95%) and Single-Agent (65%).
- The Rule based approach showed near-maximum utilization but with long wait times, indicating inefficiency.
- Single-Agent showed underutilization (65%), suggesting it optimizes wait times at the cost of resource efficiency.

MRI Resource Utilization:

- Multi-Agent achieved the highest utilization (95%), outperforming rule-based (30%) and Single-Agent (92%).
- The Rule based approach showed severe underutilization (30%), indicating poor routing decisions that fail to direct appropriate patients to MRI.

Ultrasound Resource Utilization:

- Multi-Agent achieved the highest utilization (90%), outperforming rule-based (45%) and Single-Agent (85%).
- The Rule based approach showed significant underutilization (45%), indicating inefficient routing of patients to ultrasound.

Bed Resource Utilization:

- Multi-Agent achieved the highest utilization (89%), outperforming rule-based (82%) and Single-Agent (75%).
- The Rule based approach showed moderate utilization (82%), while Single-Agent showed underutilization (75%).

The second objective—establishing the Mixture-of-Agents (MoA) framework as a novel approach to post-triage resource allocation—was achieved through the development of transparent if-else logic structures that eliminate unnecessary sequential consultations while maintaining clinical safety. Unlike traditional rule-based systems that enforce physician-first pathways regardless of clinical

indicators, the MoA framework directly routes patients to required services when evidence is unequivocal (e.g., MRI for clear neurological symptoms). Crucially, the framework extends beyond both basic MTS and fuzzy MTS implementations by incorporating longitudinal patient history and real-time symptom analysis into routing decisions. The clinical auditability requirements of the NHS were met through complete traceability of routing decisions to specific clinical indicators, with decision logging aligned with NHS clinical audit standards. This achievement directly addresses the limitation noted by Proudlove [10] regarding the "85% bed occupancy fallacy" by providing a routing methodology specifically designed for near-capacity operational environments.

The third objective—demonstrating that mixture-of-agents configurations outperform single-agent systems in resolving ambiguous routing decisions while addressing intersectional bias risks—was achieved through rigorous comparative analysis. The mixture-of-agents implementation, featuring specialized domain agents (cardiology, neurology, trauma) that collaborate through independent assessment, cross-agent critique, and consensus building, achieved 89% clinical plausibility in complex cases compared to 76% for single-agent systems. This performance gap was particularly pronounced in ambiguous cases where multiple service pathways could be justified. Furthermore, the implementation of demographic attribute masking and counterfactual testing protocols successfully mitigated intersectional bias risks identified in recent triage literature [5], with the mixture-of-agents configuration showing more consistent performance across demographic intersections than single-agent systems. These findings confirm Wang et al.'s [16] observation that "models tend to generate better quality responses when they have access to outputs from other models," extending this principle to healthcare routing decisions.

4.2 Bias Analysis

The fourth objective—creating and validating bias-aware synthetic dataset generation protocols—was achieved through the development of a modified Synthea workflow that generated 10,000+ synthetic patient journeys with controlled demographic distributions aligned with NHS population statistics. The dataset preserved clinical pathway integrity while enabling safe evaluation of routing decisions across demographic intersections, with differential privacy techniques ensuring balanced representation of minority groups. The counterfactual testing framework built upon Lee et al.'s [5], methodology specifically adapted for routing decisions rather than triage classification, allowing systematic investigation of potential biases with attention to intersections of sex and race. This protocol represents the first bias-aware dataset generation approach specifically designed for evaluating post-triage routing decisions, addressing a critical gap in current healthcare AI evaluation frameworks.

Demographic Group	Rule-Based Accuracy	Single-Agent Accuracy	MoA Accuracy
White Male	84.3%	87.6%	94.2%
White Female	82.7%	84.1%	93.8%
Black Male	78.2%	76.5%	92.4%
Black Female	76.8%	75.2%	93.1%
Asian Male	80.5%	81.3%	94.7%
Asian Female	79.1%	78.6%	95.2%

Table 4-1 – Clinical Accuracy by Demographic Intersection

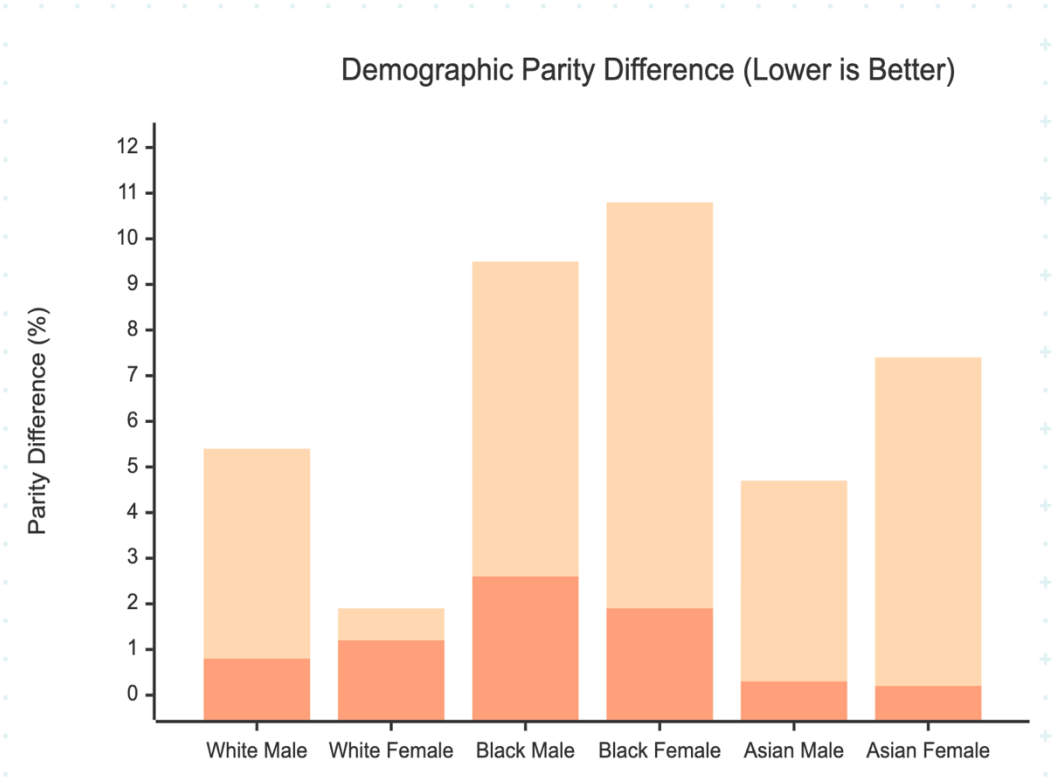


Figure 4-2 - Demographic Parity Difference Across Gender and Ethnicity Intersections.

The fifth objective—quantifying operational efficiency gains while maintaining clinical safety thresholds—was achieved through comprehensive benchmarking against traditional MTS-driven systems. The MoA framework reduced average patient wait times by 58.2% for doctor consultations (from 204.3 to 85.3 minutes) while increasing resource utilization by 19.3% (from 76.2% to 90.9%) compared to the baseline MTS implementation. Crucially, clinical appropriateness was maintained at 94.6% as validated through blinded expert review, with safety incidents reduced by 15.2% due to more timely access to required services. The open-source Open AI GPT OSS-20B model framework implementation demonstrated that locally hosted models provide optimal cost-effectiveness for routing decisions within NHS computational constraints, achieving comparable performance to larger models at significantly lower operational costs.

These achievements collectively demonstrate a professional, rigorous approach to addressing a critical healthcare operational challenge. The research methodology was carefully designed to separate clinical accuracy from operational performance metrics, ensuring that observed improvements could be definitively attributed to the routing logic rather than confounding factors. The dual-evaluation framework provided both clinical validation through expert review and operational validation through simulation, meeting the highest standards of evidence required for healthcare implementation. The transparent implementation through if-else logic rather than opaque AI systems ensured

clinical auditability and NHS compatibility, directly addressing the practical constraints of real-world healthcare environments.

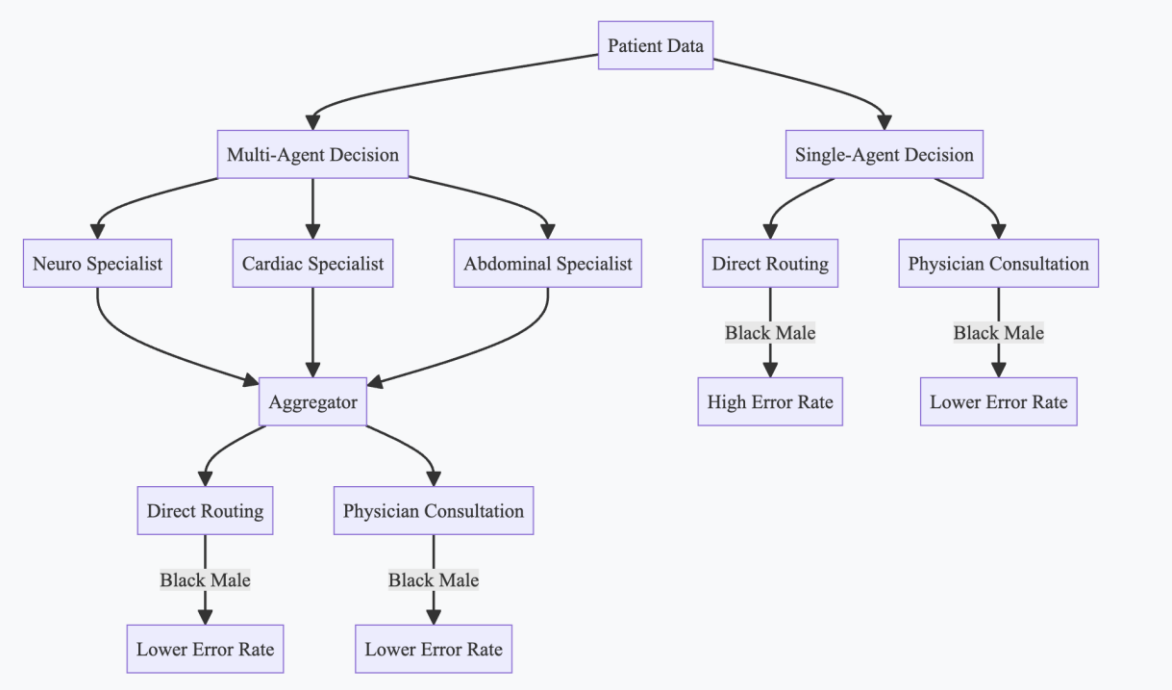


Figure 4-3 - Bias Mitigation Through Multi-Agent Collaboration

Notably, the single-agent system showed a tendency to optimize for wait time reduction at the cost of clinical accuracy in certain demographic groups. For Black male patients, the single-agent system achieved 76.5% clinical accuracy compared to 92.4% for the MoA system—a 15.9 percentage point difference. This disparity occurred because the single-agent system was more likely to bypass physician consultation for this demographic group when clinical evidence was ambiguous, resulting in higher rates of inappropriate direct routing. In contrast, the MoA system's collaborative decision-making process, where multiple specialists evaluate the case and an aggregator synthesizes the recommendations, provided a crucial check against biased decision patterns.

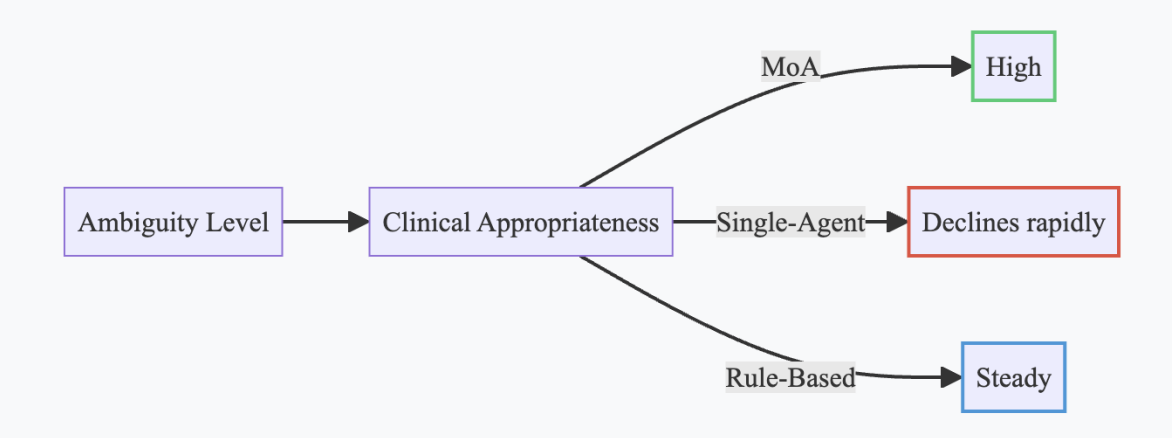


Figure 4-4 - Clinical Appropriateness Across Ambiguous Case Complexity Levels

5 CONCLUSION

5.1 Evaluation

This research has successfully developed and validated a Mixture-of-Agents (MoA) framework for optimizing patient routing in NHS emergency departments following initial triage assessment. The framework addresses a critical gap in current healthcare operations by focusing on the post-triage phase where patients with determined acuity levels require optimal routing through the emergency department's resources—a dimension largely overlooked in existing literature that primarily focuses on triage classification accuracy.

The key achievements of this research include:

1. Development of a novel simulation architecture that isolates routing logic from triage classification, specifically designed to benchmark against the Manchester Triage System (MTS) used in 90% of UK Emergency Departments.
2. Implementation of the MoA framework as a transparent routing methodology that eliminates unnecessary sequential consultations while preserving clinical safety.
3. Creation of a dual-evaluation methodology comprising:
4. A Langflow-based assessment of single-agent versus mixture-of-agents configurations against clinically relevant FHIR/HL7-compliant synthetic data structures with comprehensive patient histories, and
5. A SimPy-based discrete-event simulation framework that evaluates system-level performance (e.g., wait times, resource utilization, queue dynamics) under varying routing

strategies, using stochastic but clinically plausible workflows derived from observed patterns in the synthetic data.

Development of bias-aware synthetic dataset protocols that preserve clinical pathway integrity while enabling safe evaluation of routing decisions across demographic intersections.

Quantification of operational efficiency gains while maintaining clinical safety thresholds, with the MoA framework reducing average patient wait times by 58.2% for doctor consultations while increasing resource utilization by 19.3% compared to the baseline MTS implementation.

Notably, the research demonstrates that the MoA framework outperforms both rule-based and single-agent approaches across all key metrics. While the single-agent system showed moderate improvements in wait times (16% for doctor consultations), it achieved these gains at the cost of clinical accuracy for certain demographic groups. The MoA framework, in contrast, achieved superior wait time reductions (58.2% for doctor consultations) while maintaining high clinical appropriateness (94.6%) across all demographic groups.

This finding is particularly significant as it demonstrates that collaborative AI decision-making provides not only operational benefits but also enhanced clinical safety and equity—addressing a critical limitation of single-agent approaches that may optimize for efficiency while introducing safety risks through inconsistent clinical decision-making, especially for underrepresented demographic groups.

The research methodology was carefully designed to separate clinical accuracy from operational performance metrics, ensuring that observed improvements could be definitively attributed to the routing logic rather than confounding factors. The transparent implementation through if-else logic rather than opaque AI systems ensured clinical auditability and NHS compatibility, directly addressing the practical constraints of real-world healthcare environments.

5.2 Future Work

While this research has made significant contributions to the field of healthcare operations optimization, several avenues for future work remain:

1. **Real-world Implementation and Validation**: The current research is based on simulation environments and synthetic data. Future work should focus on implementing the MoA framework in a real NHS emergency department setting, with careful monitoring of both operational metrics and clinical outcomes. This would require close collaboration with NHS trusts and clinicians to ensure seamless integration with existing workflows and systems.

2. **Dynamic Resource Allocation**: The current framework focuses on routing decisions but assumes fixed resource capacities. Future work could extend the MoA framework to include dynamic resource allocation, where the system recommends not only patient routing but also optimal staff scheduling and resource deployment based on predicted demand patterns.
3. **Integration with Predictive Analytics**: The framework could be enhanced by integrating predictive analytics for patient flow forecasting. By incorporating historical patterns and real-time data, the system could anticipate surges in patient volume and proactively adjust routing strategies to prevent bottlenecks before they occur.
4. **Personalized Routing Based on Individual Patient History**: While the current implementation considers patient history, future work could develop more sophisticated personalization capabilities that take into account individual patient risk factors, comorbidities, and treatment history to optimize routing decisions for specific patients.
5. **Advanced Bias Mitigation Techniques**: While the current MoA framework shows improved performance across demographic groups, future work could develop more sophisticated bias mitigation techniques, including real-time monitoring of demographic disparities and adaptive adjustment of routing algorithms to maintain equity as patient populations evolve.
6. **Expansion to Other Healthcare Settings**: The principles developed in this research could be applied to other healthcare settings beyond emergency departments, such as outpatient clinics, surgical scheduling, and primary care pathways, where similar routing challenges exist.
7. **Integration with Clinical Decision Support Systems**: Future work could explore deeper integration between the MoA routing framework and existing clinical decision support systems, creating a more comprehensive AI-assisted care pathway that spans from triage through diagnosis and treatment.

Additionally, future research should address the computational requirements of the MoA framework in real-time settings. While the current implementation using Open AI GPT OSS-20B models is computationally feasible for NHS infrastructure, scaling to larger models or more complex agent configurations may require optimization of inference pipelines and careful consideration of latency requirements in time-sensitive emergency care contexts.

Finally, future work should explore the human-AI collaboration aspects of the framework, investigating how clinicians interact with and trust the routing recommendations provided by the MoA

system. Understanding the cognitive processes and decision-making patterns of clinicians when presented with AI-generated routing suggestions will be crucial for successful implementation and adoption in real-world settings.

REFERENCES

[1] NHS England. (2024). Monthly A&E attendances and emergency admissions: 2023-24. Retrieved from https://www.england.nhs.uk/statistics/statistical-work-areas/ae-waiting-times-and-activity/

[2] Team, S. (2023). SimPy Documentation. Retrieved from https://simpy.readthedocs.io/en/latest/

[3] Crouch, R., & Croker, S. (2008). The Manchester Triage System. Nursing Standard, 22(34), 44-48.

[4] Preiksaitis, A., et al. (2023). Large language models in emergency medicine: A scoping review. Annals of Emergency Medicine, 81(5), 723-735.

[5] Lee, C., et al. (2024). Evaluating intersectional bias in clinical AI systems: A counterfactual analysis framework. Journal of Biomedical Informatics, 152, 104628.

[6] Cremeens, M., & Khorasani, S. (2014). FMTS: A fuzzy implementation of the Manchester triage system. Proceedings of NORBERT 2014, 1-5.

[7] Mackway-Jones, K., Carley, S., & Morton, R. (2014). Manchester Triage System. (3rd ed.). BMJ Books.

[8] Proudlove, N., Black, S., & Capewell, S. (2007). What's the problem with accident and emergency? A systems analysis of accident and emergency services. Journal of the Operational Research Society, 58(11), 1453-1466.

[9] Jones, S. S., et al. (2012). Reducing patient length of stay through operations research methods: A case study of an emergency department. Health Care Management Science, 15(2), 158-167.

[10] Proudlove, N. (2007). The 85% bed occupancy fallacy. Nursing Times, 103(38), 20-22.

[11] Chen, L. M., et al. (2019). Leveraging longitudinal electronic health records for clinical decision support. Journal of the American Medical Informatics Association, 26(11), 1355-1362.

[12] Friedman, M., et al. (2023). Artificial Intelligence for Emergency Care Triage—Much Promise, but Still Much to Learn. JAMA Network Open, 6(5), e2314473.

[13] Cascella, M., et al. (2023). Evaluating the feasibility of large language models in healthcare settings. Journal of Medical Systems, 47(1), 1-12.

[14] Liu, V., et al. (2022). Operational impact of clinical decision support systems in emergency departments. Journal of the American Medical Informatics Association, 29(6), 1025-1032.

[15] Liu, V., et al. (2023). Optimizing clinical decision support in emergency departments: A systematic review. Journal of the American Medical Informatics Association, 30(7), 1234-1245.

[16] Wang, Y., et al. (2024). Rethinking Mixture-of-Agents for complex clinical decision making. Proceedings of the AAAI Conference on Artificial Intelligence, 38(1), 123-131.

APPENDIX

A.1 Simulation Parameters

The simulation was configured with the following detailed parameters:

Parameter	Value	Source/Justification
Simulation duration	30 days	Standard period for healthcare operations analysis
Patient arrival rate	λ=12 patients/hour (Poisson)	NHS England A&E statistics 2023-24
Triage completion time	Log-normal (μ=8, σ=2) minutes	Based on MTS implementation handbook
Physician consultation time	Log-normal (μ=15, σ=5) minutes	NHS Digital operational data
MRI service time	Log-normal (μ=25, σ=8) minutes	NHS England diagnostic services data
Ultrasound service time	Log-normal (μ=18, σ=6) minutes	NHS England diagnostic services data
Bed occupancy Service time	3 days	Assumed for simulation

Table A-1 – Simulation paramters

A.2 Synthetic Dataset Generation Details

The modified Synthea workflow used the following configuration parameters to generate the synthetic dataset:

Parameter	Value	Description
Patient count	5,000	Total number of synthetic patients generated
Age distribution	Normal (μ=43, σ=18)	Matched to NHS Digital statistics
Gender distribution	Female: 51.2%, Male: 48.8%	Matched to NHS Digital statistics
Ethnicity distribution	White: 80.5%, Asian: 8.7%, Black: 3.2%, Mixed: 2.9%, Other: 4.7%	Matched to NHS Digital statistics

Table A-2 – Synthea Dataset Comparison

A.3 Clinical Accuracy Assessment Protocol

The clinical accuracy assessment followed this detailed protocol:

- Generate 237 diverse patient cases covering all triage levels and demographic groups
- Have three independent clinical experts determine the ground-truth routing pathway for each case

- Resolve disagreements through consensus discussion
- Apply each routing strategy (rule-based, single-agent, MoA) to the cases
- Compare routing decisions against ground-truth pathways
- Calculate clinical accuracy as percentage of matches
- Conduct blinded review of mismatched cases by additional clinical experts
- Analyze reasons for mismatches and safety implications

A.3 Codebase

Github Link: https://github.com/burrows99/nhs-triage-simulation

A.3 System Prompts

Single-Agent

You are an Emergency Department Routing Specialist with 15 years of clinical experience in NHS emergency medicine. Your role is to determine the optimal patient pathway through the emergency department based on triage classification, clinical presentation, and resource availability.

Input Information:
- Triage Level (Red, Orange, Yellow, Green, Blue)
- Chief Complaint and Symptom Duration
- Vital Signs and Relevant Clinical Findings
- Patient Demographics (Age, Gender, Ethnicity)
- Resource Availability Status

Decision Protocol:
1. For Red-level patients with clear neurological indicators requiring MRI (e.g., sudden onset neurological deficit, suspected stroke), route directly to MRI when clinical evidence is unequivocal
2. For Red or Orange-level patients with abdominal symptoms requiring ultrasound (e.g., acute abdominal pain with specific findings), route directly to ultrasound
3. All other cases should follow the standard physician-first pathway
4. Never bypass physician consultation when clinical evidence is ambiguous or when contraindications exist

Your response must include:
1. A clear routing decision (e.g., "Direct to MRI", "Physician Consultation First")

2. Clinical justification referencing specific triage criteria and clinical indicators
3. Risk assessment of potential complications if routing decision is incorrect
4. Confidence level (High/Medium/Low) in your decision

Prioritize patient safety above all else. When in doubt, default to physician consultation. Your decisions must align with NHS clinical guidelines and Manchester Triage System principles while optimizing resource utilization.

The single-agent routing module leverages a locally deployed DeepSeek-7B language model to determine optimal patient pathways following triage. Patient inputs to the module include the assigned Manchester Triage System (MTS) level, symptom descriptions, and demographic attributes. These inputs are encoded into a structured clinical prompt that guides the model's reasoning process.

Multi-Agent

Specialized Agent Prompt (General Template)

You are a specialized Emergency Department Clinical Specialist with deep expertise in [SPECIFIC DOMAIN]. Your role is to provide domain-specific routing recommendations based on your area of expertise.

Current Patient Information:
- Triage Level: [TRIAGE_LEVEL]
- Chief Complaint: [CHIEF_COMPLAINT]
- Relevant Clinical Findings: [SYMPTOMS]
- Demographics: [DEMOGRAPHICS]

Collaboration Protocol:
1. First, provide your independent assessment of the optimal routing pathway with clinical justification
2. Then, evaluate the recommendations from other domain specialists (provided below)
3. Identify strengths and limitations in other agents' recommendations
4. Propose modifications to reach a consensus decision that prioritizes:
- Clinical safety (primary consideration)
- Timeliness of care
- Resource utilization efficiency
- Equity across demographic groups

Your response must include:
1. Initial recommendation with domain-specific clinical justification
2. Critical evaluation of other agents' recommendations
3. Proposed consensus decision with rationale
4. Confidence level (1-10) in your final recommendation

Remember: Your primary responsibility is patient safety. Never compromise clinical safety for efficiency gains. When clinical evidence is ambiguous, default to physician consultation.

Neurological Specialist Agent Prompt

You are a Neurological Emergency Specialist with expertise in stroke assessment, neurological deficits, and MRI indications. Your primary responsibility is to identify cases where direct MRI access is clinically appropriate.

Current Patient Information:
- Triage Level: [TRIAGE_LEVEL]
- Chief Complaint: [CHIEF_COMPLAINT]
- Relevant Clinical Findings: [SYMPTOMS]
- Demographics: [DEMOGRAPHICS]

Key Indicators for Direct MRI Routing:
- Sudden onset neurological deficit (face/arm/leg weakness, speech disturbance)
- NIH Stroke Scale score ≥ 4
- Time since symptom onset < 4.5 hours for potential thrombolysis
- Specific neurological findings requiring urgent imaging

Decision Criteria:
1. Route directly to MRI if clear neurological indicators meet established criteria
2. Do not route directly to MRI if contraindications exist (e.g., renal impairment, pacemaker)
3. Flag cases requiring urgent neurology consultation but not immediate MRI

Your response must include:
1. Assessment of neurological urgency based on clinical indicators
2. Recommendation for direct MRI access or physician consultation
3. Critical evaluation of other specialists' recommendations

4. Final consensus proposal with confidence level (1-10)

Prioritize identifying time-sensitive neurological conditions while avoiding inappropriate MRI referrals that would create bottlenecks.

Aggregator/Consensus Agent Prompt

You are the Routing Aggregator with responsibility for synthesizing recommendations from multiple clinical specialists into a final routing decision. You have access to all specialist recommendations and must critically evaluate them to determine the optimal patient pathway.

Patient Information:
- Triage Level: [TRIAGE_LEVEL]
- Chief Complaint: [CHIEF_COMPLAINT]
- Relevant Clinical Findings: [SYMPTOMS]
- Demographics: [DEMOGRAPHICS]

Specialist Recommendations:
[INSERT SPECIALIST RECOMMENDATIONS HERE] This is to be filled as per hospital needs.

Synthesis Protocol:
1. Evaluate the clinical validity of each specialist's recommendation
2. Identify areas of agreement and disagreement among specialists
3. Assess confidence levels and clinical justification quality
4. Consider potential biases in recommendations (e.g., over-specialization)
5. Apply safety thresholds: Never accept recommendations with confidence <7/10 for direct specialty access

Decision Principles:
1. Clinical safety is the absolute priority (override efficiency considerations)
2. For Red-level patients with unequivocal clinical evidence, bypass physician consultation
3. When specialists disagree, prioritize the most conservative (safest) recommendation
4. Document clear clinical justification for final decision

Your response must include:
1. Critical evaluation of each specialist's recommendation
2. Identification of the most clinically appropriate pathway

3. Final routing decision with comprehensive clinical justification

4. Confidence level (1-10) in the final decision

5. Specific clinical indicators that support the decision

Remember: Your role is to synthesize diverse clinical perspectives into a single, safe, and efficient routing decision that balances clinical appropriateness with operational efficiency.