The Optimization Paradox in Clinical AI Multi-Agent Systems

Suhana Bedi
suhana@stanford.edu
Stanford School of Medicine
Stanford, CA, USA

Iddah Mlauzi
iddah@stanford.edu
Department of Computer Science,
Stanford University
Stanford, CA, USA

Daniel Shin
dshin@stanford.edu
Department of Computer Science,
Stanford University
Stanford, CA, USA

Sanmi Koyejo
sanmi@stanford.edu
Department of Computer Science,
Stanford University
Stanford, CA, USA

Nigam H. Shah
nigam@stanford.edu
Department of Medicine, Stanford
School of Medicine
Stanford, CA, USA

Abstract

Multi-agent artificial intelligence systems are increasingly deployed in clinical settings, yet the relationship between component-level optimization and system-wide performance remains poorly understood. We evaluated this relationship using 2,400 real patient cases from the MIMIC-CDM dataset across four abdominal pathologies (appendicitis, pancreatitis, cholecystitis, diverticulitis), decomposing clinical diagnosis into information gathering, interpretation, and differential diagnosis. We evaluated single agent systems (one model performing all tasks) against multi-agent systems (specialized models for each task) using comprehensive metrics spanning diagnostic outcomes, process adherence, and cost efficiency. Our results reveal a paradox: while multi-agent systems generally outperformed single agents, the component-optimized or *Best of Breed* system with superior components and excellent process metrics (85.5% information accuracy) significantly underperformed in diagnostic accuracy (67.7% vs. 77.4% for a top multi-agent system). This finding underscores that successful integration of AI in healthcare requires not just component level optimization but also attention to information flow and compatibility between agents. Our findings highlight the need for end to end system validation rather than relying on component metrics alone.

Keywords

Multi-agent systems, Clinical decision support, Artificial intelligence, Healthcare AI, System integration

1 Introduction

Artificial intelligence (AI) is rapidly transforming healthcare across diagnosis [1], treatment planning [2], and patient management [3]. As AI systems grow in complexity, the focus has shifted from single-model solutions toward networks of specialized models ("agents") [4] that collaboratively handle different aspects of patient care. Recent studies, including Google DeepMind's AMIE [5], have demonstrated agent-based systems exceeding primary care physicians' performance in randomized clinical settings, while frameworks like MASH [6] and CRAFT-MD [7] have explored both the potential and pitfalls of multi-agent approaches.

Multi-agent AI systems mirror interdisciplinary healthcare teams, where specialists such as radiologists, pathologists, and physicians collaborate to synthesize comprehensive diagnoses. This modular approach can improve interpretability [8], simplify troubleshooting, and enable task-specific optimization [9]. However, a critical challenge arises from interactions among individually optimized agents [10]. We term this the **Optimization Paradox:** the phenomenon where excellent performance at the individual agent or component level does not necessarily translate to high overall system performance. This misalignment between individual and system-level effectiveness poses risks to patient safety and clinician trust.

This study addresses the Optimization Paradox within clinical decision support systems by examining three essential components of the diagnostic process: information gathering (ordering appropriate clinical tests), interpretation (analyzing lab results), and differential diagnosis (identifying potential medical conditions) (Figure 1). We compare multi-agent systems, where specialized agents manage each task, to single-agent systems, where one model performs all tasks. Our evaluation uses the MIMIC-CDM dataset comprising 2,400 real patient cases across four common abdominal pathologies [11].

We introduce clinically relevant evaluations extending beyond diagnostic accuracy to include process metrics (appropriate test ordering and accurate lab value interpretation) and cost efficiency metrics (clinical resource utilization and computational demands). Our findings reveal that while certain multi-agent systems achieve impressive process-level performance, this does not always translate into clinical effectiveness. The component-optimized or *Best of Breed* system exemplifies this paradox: despite achieving 85.5% accuracy in lab interpretation, its overall diagnostic accuracy (67.7%) was significantly lower than a top performing multi-agent system (77.4%; McNemar's test, $p < 0.001$) without component optimization. This 10-percentage accuracy drop poses clinically significant risks, potentially increasing misdiagnoses and compromising patient outcomes when AI systems are deployed solely based on component-level validation [12].

Our study underscores the necessity of rigorous, end-to-end validation of AI systems prior to clinical implementation, emphasizing that effective patient outcomes depend on careful system-wide integration rather than isolated component excellence.

2 Methods

This study utilized the MIMIC-CDM dataset, a curated subset of MIMIC-IV containing 2,400 real patient cases [11]. We examined

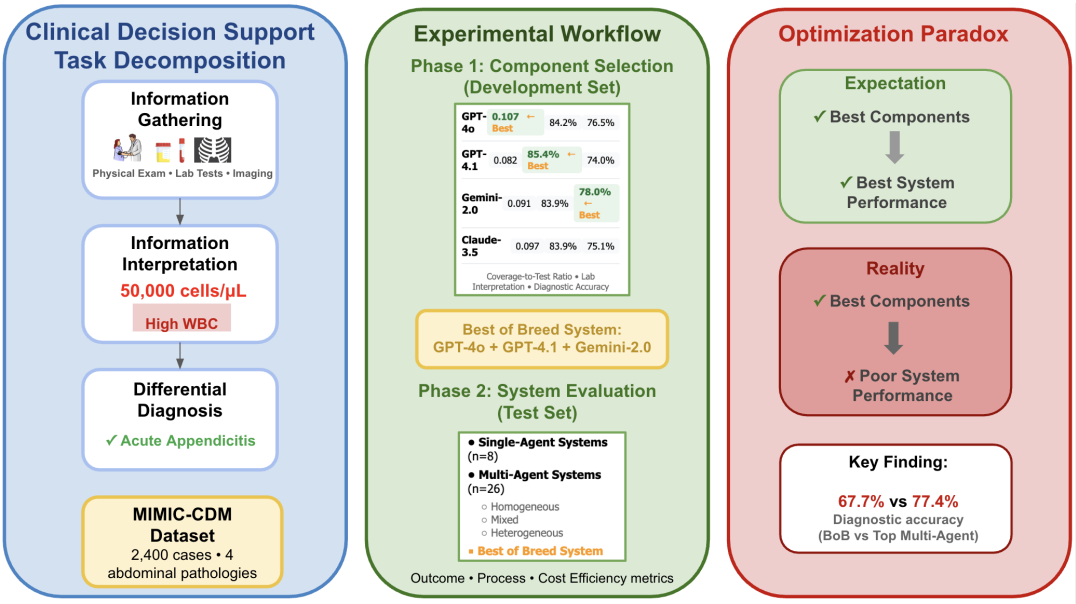

Figure 1: Overview of experimental methodology and the Optimization Paradox. The Clinical Decision Support Task Decomposition (left) breaks the diagnostic process into three specialized components. The Experimental Workflow (center) shows Phase 1 component selection and Phase 2 system comparison. The Optimization Paradox (right) illustrates the counterintuitive finding: while the *Best of Breed* system was constructed from top-performing components, it achieved poor diagnostic accuracy compared to alternative systems.

four prevalent abdominal pathologies: appendicitis (957 cases), pancreatitis (538 cases), cholecystitis (648 cases), and diverticulitis (257 cases). These conditions were selected based on their high emergency department prevalence [13, 14], diagnostic complexity due to overlapping presentations [15], and comprehensive clinical data availability within MIMIC-CDM. All patients presented with acute abdominal pain and received one of these four diagnoses.

Each case encompasses complete clinical data including patient history, physical examination findings, laboratory results (138,788 values from 480 unique tests), imaging reports (5,959 reports: abdominal CT, ultrasound, X-ray), and procedural information. All data was de-identified with primary diagnoses masked to prevent pattern matching.

2.1 Decomposition of the Clinical Decision Support Task

We adapted MIMIC-CDM's framework, decomposing the diagnostic workflow into three tasks: (1) **Information Gathering**—requesting relevant clinical data including physical examination, laboratory tests, and imaging; (2) **Information Interpretation**—processing raw clinical data and classifying results relative to reference ranges; and (3) **Differential Diagnosis**—synthesizing information to generate ranked diagnostic possibilities through clinical reasoning.

We compared single-agent systems performing all tasks end-to-end versus multi-agent systems with specialized task division. A *Retriever LLM* processed information requests and retrieved specified tests from patient records, maintaining data integrity. GPT-4o was selected for cost-effectiveness and 100% retrieval accuracy.

2.2 Data Splits

We reserved 20 cases (5 per pathology) as a pilot set for prompt development and pipeline testing, excluding them from all evaluations. The remaining 2,380 cases were stratified by pathology and split equally: 1,190 cases for development (Phase 1 component selection to construct the *Best of Breed* system) and 1,190 cases for held-out testing (Phase 2 final performance evaluation). No training or fine-tuning was performed.

2.3 Model Implementation

We created agents using LLMs from multiple families including GPT (GPT-4o, GPT-4.1), Claude (Claude-3.5-Sonnet), Gemini (Gemini-1.5-Pro, Gemini-2.0-Flash), Llama (Llama-3.3-70b), and reasoning models (o3-mini, DeepSeek-R1). API calls to all models were made through a secure cloud computing environment, ensuring patient data remained within the institutional environment and maintaining full HIPAA compliance.

2.4 Evaluation Metrics

We assessed performance using a set of metrics spanning diagnostic outcomes, process adherence, and cost efficiency. These metrics were designed to capture the quality of final diagnostic decisions, the clinical appropriateness of the decision-making process, and resource utilization throughout the workflow.

2.4.1 ***Outcome Metrics*** We evaluated diagnostic accuracy across multiple dimensions to assess the quality of clinical reasoning:

- **Overall accuracy:** Micro-averaged (treating each case equally) and macro-averaged (treating each pathology equally) accuracy across all patients.
- **Disease-specific accuracy:** Individual accuracy for each of the four target pathologies (appendicitis, pancreatitis, cholecystitis, diverticulitis).
- **Top-k accuracy:** Frequency with which the correct diagnosis appeared in the top 1, 3, or 5 positions of the ranked differential diagnosis list, reflecting real-world scenarios where clinicians consider multiple possibilities.

2.4.2 ***Process Metrics*** We assessed adherence to established clinical guidelines for information gathering and interpretation, which are essential for evidence-based medical practice:

Information gathering was evaluated along four key dimensions:

- **Coverage:** We identified recommended physical examination maneuvers, laboratory test categories, and imaging modalities for each pathology based on published clinical guidelines [16–23]. Coverage scores assessed the proportion of recommended categories requested:

$$\text{Coverage Score} = \frac{N_{\text{lab}} + N_{\text{img}} + N_{\text{maneuver}}}{N_{\text{lab\_rec}} + N_{\text{img\_rec}} + N_{\text{maneuver\_rec}}}$$

  where N represents covered categories and N_{rec} represents total recommended categories for laboratory tests, imaging, and physical examination, respectively. High coverage indicates comprehensive, guideline-concordant information gathering.
- **Guideline adherence for physical examination:** Clinical guidelines recommend physical examination as the initial diagnostic step for acute abdominal symptoms. We measured the percentage of cases where agents correctly ordered physical examination first.
- **Average number of tests per patient:** Average number of diagnostic tests (laboratory, imaging, and physical examination maneuvers) requested per patient, providing insight into resource utilization patterns.
- **Coverage-to-test ratio:** Balances comprehensive test assessment with efficient resource utilization:

$$\text{Coverage-to-test ratio} = \frac{\text{Coverage Score}}{\text{Average tests per patient}}$$

  This metric rewards high guideline adherence with minimal test ordering, balancing diagnostic necessity against cost and patient burden. We examined the relationship between this ratio and diagnostic accuracy using Spearman correlation analysis.

Information interpretation was measured as the percentage of lab values correctly classified as "high," "normal," or "low" relative to reference ranges—a fundamental clinical skill requiring basic numerical literacy.

2.4.3 Cost Efficiency Metrics

- **Computational cost:** Estimated cost based on token usage and publicly available API pricing for each model.
- **Clinical resource cost:** Average reimbursement cost for all laboratory tests ordered per patient, providing a realistic estimate of healthcare expenditure associated with each system's diagnostic approach.

2.5 Experiments

2.5.1 ***Phase 1: Component Selection and Best of Breed Construction*** Individual LLMs performed all three tasks end-to-end on the development set (n=1,190). We selected the top-performing agent for each task to construct the *Best-of-Breed (BoB)* system:

- **BoB Information Gathering Agent:** Selected for highest coverage-to-test ratio with coverage score >0.5, preventing spuriously high ratios from low coverage.
- **BoB Information Interpretation Agent:** Selected for highest laboratory interpretation accuracy.
- **BoB Differential Diagnosis Agent:** Selected for highest diagnostic accuracy (micro-averaged/top-1).

2.5.2 ***Phase 2: System Performance Comparison*** All systems were evaluated on the held-out test set (n=1,190) to assess real-world performance and investigate the optimization paradox. The evaluation included:

- **Single-Agent Systems:** Individual LLMs performing all three tasks end-to-end, serving as baselines for comparison.
- **Multi-Agent Systems:** Three specialized agents coordinated by a basic orchestrator using conditional logic to route tasks appropriately (Figure 2). The orchestrator implemented a defined workflow with explicit handoffs: (1) Information Gathering Agent requests tests, (2) Retriever LLM fetches requested data, (3) Information Interpretation Agent processes lab results, and (4) Differential Diagnosis Agent generates ranked diagnoses. Complete implementation details are provided in the supplementary material. These systems are categorized by their model backbone composition:
  - **Homogeneous:** All three agents use the same backbone (e.g., GPT-4o/GPT-4o/GPT-4o)
  - **Mixed:** Two agents share a backbone, one differs (e.g., GPT-4o/GPT-4o/Gemini-Flash)
  - **Heterogeneous:** All three agents use distinct backbones (e.g., GPT-4o/Claude-3.5/Gemini-Flash)
- **Best of Breed System:** A heterogeneous multi-agent system constructed using the top-performing agent from Phase 1 for each specialized task. Given the computational complexity of evaluating all possible heterogeneous combinations, we constructed additional systems by selecting the next best performing LLMs for each component, ensuring meaningful diversity.

2.6 Statistical Analysis

Primary Metrics: We used win rate as our primary performance metric, defined as the percentage of pairwise comparisons where one system type outperforms another. Win rates are robust for small sample sizes and less sensitive to outliers.

Statistical Tests: We applied Mann-Whitney U tests for group comparisons (e.g., single-agent vs. multi-agent systems) and McNemar's test for paired comparisons between specific systems (e.g.,

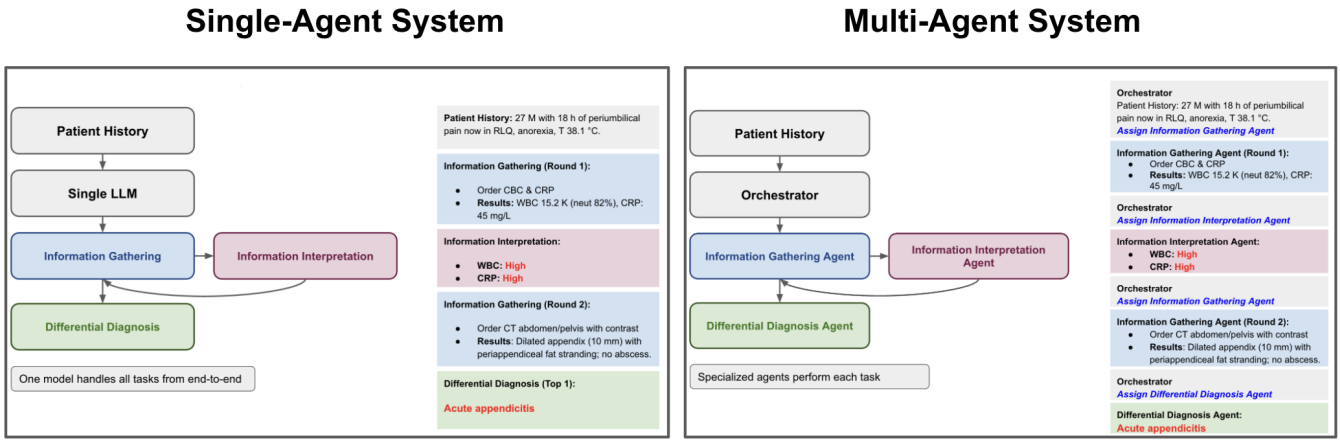

Figure 2: System architectures and workflows demonstrated with an acute appendicitis case. **Left panel:** Single-agent system where one LLM handles the complete clinical workflow from patient history through information gathering, interpretation, and differential diagnosis. **Right panel:** Multi-agent system where an orchestrator coordinates specialized agents for each task.

Best of Breed vs. other systems). We report test statistics, p-values, and 95% confidence intervals.

Effect Size Analysis: We complemented statistical tests with effect size measures to quantify the magnitude of performance differences:

- **Cohen's d:** For comparing different system types (e.g., single-agent vs. multi-agent)
- **Glass's Delta:** Used for comparing a specific agent system (e.g., *Best of Breed*) against a reference group (e.g., all other multi-agent systems), calculated as: $d = \frac{\text{Mean}_{\text{specific}} - \text{Mean}_{\text{reference}}}{\text{SD}_{\text{reference}}}$

Effect sizes were interpreted using standard conventions: small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$) effects.

LLM	Info Gathering	Info Interp	Diag Acc
Gemini-2.0-Flash	0.091	83.9	**77.98**
GPT-4o	**0.107**	84.2	76.47
Claude-3.5-Sonnet	0.097	83.9	75.13
Llama-3.3-70b	0.088	75.9	74.96
DeepSeek-R1	0.086	71.9	74.87
GPT-4.1	0.082	**85.4**	73.95
Gemini-1.5-Pro	0.082	79.8	72.77
o3-mini	0.081	81.6	61.6

Table 1: Task-level performance for single-agents with the best value for each task highlighted in **bold**. Coverage-to-test ratio measures information gathering, lab interpretation accuracy reflects information interpretation, and micro-averaged diagnostic accuracy represents differential diagnosis capability.

3 Results

3.1 Phase 1: Component Selection and Best of Breed Construction

We evaluated individual LLMs on all three tasks using the development set (n=1,190). The top performers for each task formed our *Best of Breed* system (Table 1).

- **Information Gathering:** GPT-4o achieved the highest coverage-to-test ratio (0.107) with good average coverage per patient (0.64), making it an optimal information gathering agent.
- **Information Interpretation:** GPT-4.1 led in laboratory interpretation tasks with 85.4% accuracy in classifying lab values relative to reference ranges.
- **Differential Diagnosis:** Gemini-2.0-Flash demonstrated superior micro-averaged diagnostic accuracy (78%).

Based on these evaluations, we constructed the *Best of Breed* system using GPT-4o for information gathering, GPT-4.1 for information interpretation, and Gemini-2.0-Flash for differential diagnosis.

3.2 Phase 2: System Performance Comparison

A total of 8 single-agent and 26 multi-agent systems were evaluated on the held-out test set (Table 2).

3.2.1 ***Multi-Agent vs Single-Agent Systems*** Multi-agent systems (n=26) significantly outperformed single-agent systems (n=8) on process metrics, winning 78.4% of pairwise comparisons for information gathering (p = 0.015), 87.5% for information interpretation (p = 0.0008), and 76.9% for computational cost (p = 0.022). However, multi-agent system advantages were modest and non-significant for diagnostic accuracy (52.9% win rate) and clinical resource costs (60.1% win rate). Thus, multi-agent systems enhance process quality and computational efficiency but show limited improvement in clinical outcomes. Notably, the coverage-to-test ratio showed no significant correlation with diagnostic accuracy (Spearman's $\rho = -0.057$, $p = 0.748$), suggesting the disconnect between process metrics and outcomes. All raw numbers can be found in Appendix 4.4.

Metric Category	Multi-Agent Systems vs. Single-Agent Systems (Win Rate for Multi-Agents)	BoB System vs. Other Multi-Agent Systems (Win Rate for BoB)	BoB System vs. Single-Agent Systems (Win Rate for BoB)
Information Gathering	**78.4%** (Cohen's d = 1.06)	**80.0%** (Glass's Δ = 1.29)	**100%** (Glass's Δ = 1.66)
Information Interpretation	**87.5%** (Cohen's d = 1.79)	**76.0%** (Glass's Δ = 0.60)	**100%** (Glass's Δ = 1.44)
Diagnosis Accuracy	52.9% (Cohen's d = 0.05)	**8.0%** (Glass's Δ = −1.04)	**12.5%** (Glass's Δ = −1.15)
Computational Cost	**76.9%** (Cohen's d = 0.96)	**80.0%** (Glass's Δ = 0.81)	**87.5%** (Glass's Δ = 0.92)
Clinical Resource Cost	60.1% (Cohen's d = 0.31)	**84.0%** (Glass's Δ = 1.46)	**87.5%** (Glass's Δ = 1.31)

Table 2: Win rates (%) between multi-agent, BoB, and single-agent systems across multiple evaluation metrics. Coverage-to-test ratio measures information gathering, lab interpretation accuracy reflects information interpretation , and micro-averaged diagnostic accuracy represents differential diagnosis capability. Effect sizes are reported using Cohen's d for parametric comparisons and Glass's Δ for non-parametric comparisons (**Bold indicates large or medium effect size, |Cohen's d| $\geq$ 0.5 or |Glass's Δ| $\geq$ 0.5**).

3.2.2 ***Best of Breed vs Multi-Agent Systems: The Optimization Paradox*** The Optimization Paradox emerged when comparing the *Best of Breed* system against other multi-agent systems (n=25). While BoB achieved high win rates across process metrics including information gathering (80.0%), information interpretation (76.0%), computational cost (80.0%), and clinical resource cost (84.0%), it underperformed in diagnostic accuracy with only an 8.0% win rate.

Direct comparison with the top-performing multi-agent system revealed that BoB's diagnostic accuracy (67.65%) was significantly lower than the baseline (77.39%), representing a 9.75% decrease (95% CI: 7.11% to 12.38%; McNemar's test, $p < 0.0001$). This contrast between strong component performance and poor diagnostic outcomes demonstrates the Optimization Paradox: optimizing individual components can undermine overall system performance.

3.2.3 ***Best of Breed vs Single-Agent Systems*** The Optimization Paradox became even more pronounced when comparing the *Best of Breed* system against single-agent systems (n=8). While BoB achieved perfect win rates across process metrics including information gathering (100%) and information interpretation (100%), along with strong performance in cost efficiency metrics including computational cost (87.5% win rate) and clinical resource cost (87.5% win rate), it critically underperformed in diagnostic accuracy with only a 12.5% win rate.

Direct comparison with the top-performing single-agent system revealed that BoB's diagnostic accuracy (67.65%) was significantly lower than the baseline (75.63%), representing a 7.98% decrease (95% CI: 5.39% to 10.57%; McNemar's test, $p < 0.0001$). This contrast between strong operational metrics and poor diagnostic outcomes further confirms that optimizing individual components can undermine overall system performance.

3.2.4 ***Model Backbone Effects on Diagnostic Performance*** The Optimization Paradox in our *Best of Breed* system prompted investigation into whether diagnostic performance relates to model diversity within multi-agent systems. We compared diagnostic accuracy across homogeneous systems (n=7, all agents from same backbone), mixed systems (n=12, two agents from one backbone), and heterogeneous systems (n=6, all agents from different backbones, including BoB). We excluded one extreme outlier (DeepSeek system with 54% accuracy) from the homogeneous group analysis.

Homogeneous systems (median = 74.29%) showed no significant difference from mixed systems (median = 74.75%; $p = 0.967$, Cohen's d = 0.019). However, heterogeneous systems (median = 71.22%) demonstrated lower performance than both homogeneous ($p = 0.073$, Cohen's d = 1.41) and mixed systems ($p = 0.039$, Cohen's d = 1.17). While p-values did not reach Bonferroni-corrected significance ($\alpha = 0.0167$), the large effect sizes suggest heterogeneous compositions face inherent diagnostic challenges, potentially explaining the Optimization Paradox.

3.2.5 ***Why Best of Breed Fails: Information Flow Breakdown*** To understand why the *Best of Breed* system failed despite superior component metrics, we conducted systematic error analysis on all 1,190 test cases, comparing failure patterns against the top-performing multi-agent system (Table 3).

Failure Type	Best of Breed	Top Multi-Agent System
Overall Performance		
Hallucinated Test Results	165 (13.87%)	5 (0.42%)
Unauthorized Test Ordering	165 (13.87%)	9 (0.76%)
Insufficient Info Gathering	84 (7.06%)	23 (1.93%)
Head-to-Head Comparison		
Hallucinated Test Results	81 (46.55%)	0 (0.00%)
Insufficient Info Gathering	63 (36.21%)	0 (0.00%)
Other Failures	30 (17.24%)	0 (0.00%)

Table 3: System failure comparison showing overall performance (1,190 cases) and head-to-head analysis of 174 cases where the top-performing multi-agent system succeeded but Best of Breed failed.

Information Gathering Failures: BoB's information gathering agent (GPT-4o) exhibited critical failure patterns, including insufficient information gathering in 7.06% of cases where the agent concluded test gathering despite missing essential diagnostic information, particularly imaging tests in pancreatitis cases.

Diagnosis Agent Failures: These information deficits triggered compensatory behaviors in BoB's diagnosis agent (Gemini-2.0-Flash). When faced with insufficient data, the agent violated protocol by attempting unauthorized test ordering in 13.87% of cases,

and then hallucinated test results at the same rate. This represents a serious safety failure in clinical decision-making.

Top-Performing Multi-Agent System Success: In contrast, the top-performing multi-agent system (using Gemini-2.0-Flash for information gathering and interpretation and GPT-4o for diagnosis) demonstrated superior information flow management with lower failure rates: only 1.93% insufficient information gathering, 0.76% unauthorized test ordering, and 0.42% hallucinated results, representing a **33-fold** reduction in hallucination compared to BoB.

Head-to-Head Analysis: Direct comparison of 174 cases where the top-performing multi-agent system succeeded but BoB failed revealed that nearly half (46.55%) involved test result hallucination, while 36.21% showed insufficient information gathering. These findings demonstrate that the Optimization Paradox stems from fundamental agent compatibility issues rather than individual component deficiencies.

These results show that component-level metrics cannot capture agent interactions. Individually superior agents may create systematic coordination failures when combined, undermining overall performance despite strong standalone capabilities.

4 Discussion

Our study reveals a striking **Optimization Paradox** in multi-agent systems: the *Best of Breed* system, built from top-performing components, excelled in process and cost efficiency metrics yet achieved only 67.7% diagnostic accuracy. This level of performance, coupled with test result hallucination in 13.87% of cases, is clinically unacceptable and represents a serious safety hazard, potentially leading to delayed or incorrect treatment, unnecessary procedures, and adverse outcomes [12]. This paradox demonstrates that successful multi-agent systems require not just component optimization but careful attention to information flow between agents.

Several factors explain this surprising outcome. First, our process metrics captured whether agents followed guidelines but missed diagnostic relevance. Thus, the *Best of Breed* system efficiently followed clinical guidelines, but these metrics failed to assess whether collected data matched the diagnostic agent's specific requirements.

Second, the diagnostic agent showed poor adaptability when processing information from unfamiliar upstream partners. During Phase 1 evaluation, it performed well on its specialized task. When combined with other top-performing agents in the *Best of Breed* system in Phase 2, the information flow and formatting patterns were disrupted, causing systematic diagnostic failures. This highlights that multi-agent systems require holistic evaluation rather than component-level optimization.

Our backbone composition analysis reveals the underlying mechanism: while mixed systems (two identical + one different model) performed as well as homogeneous systems, heterogeneous systems like *Best of Breed* showed significant degradation. The coordination challenges likely stem from fundamental differences in how model backbones process and communicate information. Each backbone (GPT, Claude, Gemini) has distinct training approaches, prompt sensitivity patterns, and output formatting preferences. When these different "communication styles" interact, information may be lost or misinterpreted during handoffs. Our error analysis confirms this mechanism: in the 174 cases where the top-performing system succeeded but *Best of Breed* failed, nearly half (46.55%) involved dangerous test result hallucination, while the compatible agents showed no hallucination failures.

These technical findings have important practical implications for AI deployment. Healthcare organizations should exercise caution when adopting modular AI solutions, as component metrics poorly predict integrated performance. Procuring best-in-class point solutions for each task while expecting seamless integration can create nominally efficient but ultimately ineffective and potentially unsafe pipelines. End-to-end validation against clinical outcomes is essential before deployment. In addition, regulatory frameworks should require system-level performance evidence rather than relying solely on component accuracy metrics.

Several limitations warrant consideration. The dataset (2,400 cases from a single academic center) limits generalizability across institutions and patient populations, and lacks external validation. Our focus on four abdominal pathologies may not generalize to other clinical domains. Additionally, our component selection prioritized single metrics per task rather than multi-dimensional optimization strategies, and our multi-agent system used basic orchestration without iterative reasoning or dynamic agent communication.

Future work should develop process metrics that better correlate with clinical outcomes, investigate selection methods that optimize for system-level performance rather than isolated component excellence, and explore dynamic agent architectures capable of iterative reasoning and self-correction. Most importantly, external validation across diverse clinical settings is needed to establish the generalizability of the Optimization Paradox.

Supplementary Material

Additional details including prompt development, guideline recommended tests, and orchestrator implementation are provided in the supplementary material document.

References

[1] Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu, Yawen Guo, Zaifu Zhan, Sirui Ding, Jiashuo Wang, Kaishuai Xu, Yi Fang, Liqiao Xia, Jeremy Yeung, Daochen Zha, Genevieve B. Melton, Mingquan Lin, and Rui Zhang. Large language models for disease diagnosis: A scoping review, 2024.

[2] Theresa Isabelle Wilhelm, Jonas Roos, and Robert Kaczmarczyk. Large language models for therapy recommendations across 3 clinical specialties: Comparative study. *Journal of Medical Internet Research*, 25:e49324, Oct 2023.

[3] Jin Rui Edmund Neo, Joon Sin Ser, and San San Tay. Use of large language model-based chatbots in managing the rehabilitation concerns and education needs of outpatient stroke survivors and caregivers. *Frontiers in Digital Health*, 6:1395501, May 2024.

[4] Ranjan Sapkota, Konstantinos I. Roumeliotis, and Manoj Karkee. Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges, 2025.

[5] Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, and Vivek Natarajan. Towards conversational diagnostic artificial intelligence. *Nature*, April 2025. Published online 09 April 2025.

[6] Michael Moritz, Eric Topol, and Pranav Rajpurkar. Coordinated ai agents for advancing healthcare. *Nature Biomedical Engineering*, 2025. Commentary.

[7] Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. CRAFT-MD: A conversational evaluation framework for comprehensive assessment of clinical LLMs. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*,

2024.
[8] Rachel Gordon. Ai agents help explain other ai systems, January 2024. MIT News, Massachusetts Institute of Technology.
[9] Raphael Shu, Yi Zhang, Michelle Yuan, Nilaksh Das, and Monica Sunkara. Unlocking complex problem-solving with multi-agent collaboration on amazon bedrock. https://aws.amazon.com/blogs/machine-learning/unlocking-complex-problem-solving-with-multi-agent-collaboration-on-amazon-bedrock/, January 2025. AWS Machine Learning Blog.
[10] Lance B. Eliot. Multi-agent ai orchestration shaping up but here's why it might not be fully shipshape. *Forbes*, November 2024. Innovation, AI.
[11] Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30:2613–2622, July 2024.
[12] Christopher R. Macaluso and Robert M. McNamara. Evaluation and management of acute abdominal pain in the emergency department. *International Journal of General Medicine*, 5:789–797, 2012.
[13] Salomone Di Saverio, Mauro Podda, Belinda De Simone, Marco Ceresoli, Goran Augustin, Antonio Gori, Luca Ansaloni, Marja Boermeester, Massimo Sartelli, Federico Coccolini, and Fausto Catena. Diagnosis and treatment of acute appendicitis: 2020 update of the wses jerusalem guidelines. *World Journal of Emergency Surgery*, 15(1):27, 2020.
[14] Gianfranco Cervellin, Riccardo Mora, Andrea Ticinesi, Tiziana Meschi, Ivan Comelli, Fausto Catena, and Giuseppe Lippi. Epidemiology and outcomes of acute abdominal pain in a large urban emergency department: retrospective analysis of 5,340 cases. *Emergency and Critical Care Medicine*, 4(19):1–8, 2016. Available online October 15, 2016.
[15] Adriaan C. van Breda Vriesman and Julien B. C. M. Puylaert. Mimics of appendicitis: Alternative nonsurgical diagnoses with sonography and ct. *AJR American Journal of Roentgenology*, 186(4):1103–1112, 2006.
[16] Salomone Di Saverio, Mauro Podda, Belinda De Simone, Marco Ceresoli, Goran Augustin, Antonio Gori, Luca Ansaloni, Marja Boermeester, Massimo Sartelli, Federico Coccolini, and Fausto Catena. Diagnosis and treatment of acute appendicitis: 2020 update of the wses jerusalem guidelines. *World Journal of Emergency Surgery*, 15:27, 2020.
[17] Michele Pisano, Nadia Allievi, Kurinchi Gurusamy, Giovanni Borzellino, Carlos A. Gomes, Andrew W. Kirkpatrick, and et al. 2020 world society of emergency surgery updated guidelines for the diagnosis and treatment of acute calculus cholecystitis. *World Journal of Emergency Surgery*, 15:61, 2020.
[18] John F. Hall, Patricia L. Roberts, Ramon Ricciardi, Thomas E. Read, Benjamin R. Davis, Paul W. Marcello, and Scott R. Steele. The american society of colon and rectal surgeons clinical practice guidelines for the treatment of left-sided colonic diverticulitis. *Diseases of the Colon & Rectum*, 63(5):728–747, 2020.
[19] Ari Leppäniemi, Matti Tolonen, Antonio Tarasconi, Gabriele Anania, Helena Segovia-Lohse, Edoardo Gamberini, and et al. 2019 wses guidelines for the management of severe acute pancreatitis. *World Journal of Emergency Surgery*, 14:27, 2019.
[20] Steven H Yale, Halil Tekiner, and Elizabeth S Yale. Signs and syndromes in acute appendicitis: A pathophysiologic approach. *World Journal of Gastrointestinal Surgery*, 14(7):727–730, 2022.
[21] Sajad Ahmad Salati, Khalid Alkhalifah, and Abdul Majeed Salem Almousa. Eponymous signs of acute cholecystitis – a review. *Saudi Medical Journal Students*, 3(2):44–55, 2023.
[22] Anne F Peery, Temitope O Keku, Joseph A Galanko, and Robert S Sandler. Colonic diverticulosis is not associated with painful abdominal symptoms in a us population. *Gastro Hep Advances*, 1(1):15–22, 2022.
[23] Scott et al. Tenner. American college of gastroenterology guideline: management of acute pancreatitis. *The American Journal of Gastroenterology*, 108(9), 2013.

Supplementary Material

4.1 Prompt Development

Single-agent:

```
# SYSTEM PROMPT
"""
You are a medical-AI assistant helping a physician diagnose and treat patients.
**Always follow the exact output formats below.**
-------------------------------------------------------------------------
FORMAT 1  (when you still need more information)
Thought: <your reasoning about what information is needed and why>

[If the immediately-preceding message is a Tool output that contains
laboratory values, INSERT the next section exactly once.]

Lab Interpretation: {
    "test_name": {"value": <number>, "interpretation": "high/normal/low"},
    ...
}
Action: <one of: Physical Examination |
Laboratory Tests | Imaging>
Action Input: <comma-separated list of specific tests, imaging studies or
physical exam maneuver you are requesting.>
IMPORTANT: You can only request one action type at a time. Do not combine
multiple action types.
-------------------------------------------------------------------------
FORMAT 2  (when you are ready to give the final answer)
Thought: <your complete clinical reasoning>
**Final Diagnosis (ranked):**
1. <most likely diagnosis>
2. <second most likely diagnosis>
3. <third most likely diagnosis>
4. <fourth most likely diagnosis>
5. <fifth most likely diagnosis>
Treatment: <detailed evidence-based treatment plan>

**IMPORTANT: After providing FORMAT 2, your task is COMPLETE. Do NOT request
any further actions or tools.  FORMAT 2 is the FINAL output.
Once you provide FORMAT 2, the conversation ENDS.**
-------------------------------------------------------------------------
HARD RULES  (read carefully)

1. **Mandatory Lab Interpretation**
   · If the last message you received is a Tool output with lab data, you MUST
   include the "Lab Interpretation" JSON block.
   · If you omit it, your answer will be rejected and you will be asked
   to try again.

2. JSON validity
   · The Lab Interpretation block must be valid JSON (double quotes,
   no trailing commas).
   · Include both the numeric value and the interpretation ("high", "normal",
   or "low") for every test you mention.

3. Do NOT mix elements from different formats.

4. "Action Input" is **only** for naming new tests or imaging studies you want
to order.  Never place results or interpretations there.

5. **Action Input Content:** The "Action Input" field should ONLY contain
a comma-separated list  of test names, imaging studies, or
physical exam maneuvers. Do NOT include any thoughts, reasoning, interpretations,
or other text in the "Action Input" field.

6. **STOP AFTER FORMAT 2:** Once you have provided FORMAT 2
(Final Diagnosis and Treatment), you MUST stop. Do NOT ask for any more
information or tools after FORMAT 2.

7. Stop asking for additional information
when you are confident enough to provide FORMAT 2.
-------------------------------------------------------------------------
EXAMPLES

Lab Interpretation: {
    "WBC":  {"value": 12.5, "interpretation": "high"},
    "CRP":  {"value": 5.0,  "interpretation": "normal"}
}
Action: Laboratory Tests
Action Input: Serum Lipase, Abdominal Ultrasound
Action: Physical Examination
Action Input: McBurney's Point Tenderness
"""
```

Figure 3: Prompts used for the single-agent systems.

Multi-agent:

```
# INFORMATION GATHERING AGENT

INFO_GATHERING_PROMPT = """\
You are a medical-AI assistant helping a physician COLLECT information that
will later be used to diagnose and treat the patient.
**Always follow the exact output formats below.**
--------------------------------------------------------------------------
FORMAT 1  (when you still need more information)

Thought: <your reasoning about what information is needed and why>
Action: <one of: Physical Examination | Laboratory Tests | Imaging>
Action Input: <comma-separated list of specific tests, imaging studies or
physical exam maneuver you are requesting.>

IMPORTANT: You can only request one action type at a time. Do not combine
multiple action types.
--------------------------------------------------------------------------
FORMAT 2  (when you are done collecting information)

Thought: <your complete clinical reasoning>
Action: done
Action Input: ""

**IMPORTANT: After providing FORMAT 2, your task is COMPLETE. Do NOT request
any further actions or tools.
FORMAT 2 is the FINAL output. Once you provide FORMAT 2, conversation ENDS.**

--------------------------------------------------------------------------
HARD RULES  (read carefully)

1. Do NOT mix elements from different formats.

2. “Action Input” is **only** for naming new tests or imaging studies
you want to order.  Never place results or interpretations there.

3. **Action Input Content:** The "Action Input" field should ONLY contain
a comma-separated list of test names, imaging studies, or physical exams.
Do NOT include any thoughts, reasoning, interpretations, or other text
in the "Action Input" field.

4. **STOP AFTER FORMAT 2:** Once you have provided FORMAT 2, you MUST stop.
Do NOT ask for any more information or tools after FORMAT 2.

5. Stop asking for additional information when you are confident enough
to provide FORMAT 2.
"""

# INFORMATION INTERPRETATION AGENT
INTERPRETATION_PROMPT = """\
You are a medical-AI assistant helping a physician interpret lab results
that have already been retrieved.
**Always follow the exact output formats below.**

--------------------------------------------------------------------------
FORMAT  (interpret the lab panel you just received)

[If the immediately-preceding message is a Tool output that contains lab
values, INSERT the next section exactly once.]

Lab Interpretation: {
    "test_name": {"value": <number>, "interpretation": "high/normal/low"},
    ...
}

**IMPORTANT: After providing this FORMAT, your task is COMPLETE.
Do NOT request any further actions or tools.  This FORMAT is the FINAL output.
Once you provide this FORMAT, the conversation ENDS.**
HARD RULES  (read carefully)

1. **Mandatory Lab Interpretation**
   • If the last message you received is a Tool output with lab data,
   you MUST include the “Lab Interpretation” JSON block
   • If you omit it, your answer will be rejected and you will be asked
     to try again.

2. JSON validity
   • The Lab Interpretation block must be valid JSON (double quotes,
     no trailing commas).
   • Include both the numeric value and the interpretation
     (“high”, “normal”, or “low”) for every test you mention.

3. Do NOT mix elements from different formats.
"""
```

Figure 4: Prompts used for the multi-agent systems.

Retriever LLM:

```
LABS_MATCHER_PROMPT = """
Available laboratory tests and their results: {available_tests}.
Requested tests: {requested_tests}.

Please retrieve and return the results for the requested tests.
Return each test name along with its corresponding result.
If a test is not available, state that.
Respond in natural language
"""

IMAGING_MATCHER_PROMPT = """
Available imaging studies: {available_imaging}.
Requested imaging: {requested_imaging}.

Please retrieve and return the full report only
for the imaging study that best matches the
requested imaging from the available list.
If the requested imaging is not available, state that.
Do not propose or mention any additional or alternative tests or imaging.
Return the study name along with the full report.
Respond in natural language.
"""
```

Figure 5: Prompts used for the retriever LLM.

4.2 Guideline recommended tests

Pathology	Physical Exam Maneuver	Synonyms
Appendicitis	McBurney’s Point Tenderness	mcburney, mcburney’s, mcburney point, mcburney’s point, point of mcburney, mcburney tenderness, right iliac tenderness, tenderness at mcburney, tenderness at mcburney’s point
Cholecystitis	Murphy’s Sign	murphy, murphy’s, murphy sign, murphy’s sign, inspiratory arrest, halted inspiration, interruption of breath, breath catching, respiratory arrest with palpation
Diverticulitis	Left Lower Quadrant Tenderness	left lower quadrant, llq, sigmoid, sigmoid tenderness, tenderness over sigmoid, left iliac fossa, lif, left-sided abdominal tenderness, sigmoid colon tenderness
Pancreatitis	Epigastric Tenderness	epigastric, epigastrium, upper abdominal, mid-upper abdomen, central upper abdomen, transabdominal tenderness, midline upper abdomen, central abdominal tenderness, mid-epigastric

Table 4: Physical Examinations and Synonyms by Pathology

Pathology	Recommended Tests
Appendicitis	**Inflammation:** • White Blood Cell Count (WBC) • C-Reactive Protein (CRP)
Cholecystitis	**Inflammation:** • White Blood Cell Count (WBC) • C-Reactive Protein (CRP) **Assess the risk of Chronic Bile Duct Stones (CBDS):** • Alanine Transaminase (ALT) • Aspartate Transaminase (AST) • Alkaline Phosphatase (ALP) • Gamma Glutamyltransferase (GGT) • Bilirubin
Diverticulitis	**Inflammation:** • White Blood Cell Count (WBC) • C-Reactive Protein (CRP) (predicts severity)
Pancreatitis	**Serum pancreatic enzyme:** • Lipase • Amylase **Other:** • C-Reactive Protein (CRP) • Hematocrit • Blood Urea Nitrogen (BUN) • Procalcitonin • serum triglyceride and calcium levels (in absence of gallstones or significant alcohol use)

Table 5: Recommended Lab Tests by Pathology

4.3 Orchestrator Implementation

The orchestrator coordinated specialized agents through a sequential workflow built on LangGraph. Each agent received the patient's clinical context and complete conversation history, enabling informed decision-making at each step. The Information Gathering Agent iteratively requested clinical tests until signaling completion with "Action: done" or reaching the 10-turn limit, at which point control passed to subsequent agents. Data retrieval was handled through the RetrieveResults tool, which processed three types of requests: physical examinations, laboratory tests, and imaging studies. When agents requested specific tests using natural language (e.g., "Complete blood count, C-reactive protein"), GPT-4o served as a retriever to identify and return the relevant patient data from clinical records.

Example Workflow:

```
1. Information Gathering Agent: "Action: Physical Examination
   Action Input: Abdominal tenderness, McBurney's point"

2. RetrieveResults Tool: Returns physical exam findings

3. Information Gathering Agent: "Action: Laboratory Tests
   Action Input: Complete blood count, C-reactive protein"

4. RetrieveResults Tool: Returns lab values

5. Information Interpretation Agent: Processes lab results as
   {"WBC": {"value": 15000, "interpretation": "high"}}

6. Information Gathering Agent: "Action: done"

7. Differential Diagnosis Agent: Generates final ranked diagnosis
```

Figure 6: Sequential processing of a patient case demonstrating the flow from information gathering through data retrieval, interpretation, and final diagnosis for acute abdominal pain.

The system incorporated robust error handling, including format validation with single retry attempts for malformed outputs and 60-second API timeouts with exponential backoff for rate limiting. When requested tests were unavailable, the workflow continued with accessible data rather than terminating. Token usage was tracked separately for each agent phase to enable precise cost calculations across heterogeneous multi-agent configurations.

4.4 Test Set Results

Agent System	Micro Avg Accuracy	Macro Avg Accuracy	Appendicitis Accuracy	Pancreatitis Accuracy	Cholecystitis Accuracy	Diverticulitis Accuracy	Top3 Acc	Top5 Acc
multi_gemini-flash_gemini-flash_gpt	77.39	73.75	93.58	67.64	68.45	65.32	86.05	88.57
multi_gemini-flash_gpt_gpt	77.31	74.69	93.38	66.55	66.25	72.58	87.48	88.91
multi_o3-mini_o3-mini_o3-mini	76.97	73.69	93.8	64.36	68.04	68.55	85.88	88.49
multi_claude_gpt_gpt	76.22	72.94	92.95	69.09	62.78	66.94	85.13	87.48
multi_gemini-flash_gemini-flash_gemini-flash	75.88	71.98	93.63	57.76	68.77	67.74	84.79	87.98
single_gemini-flash_gpt	75.63	71.71	89.62	63.18	70.35	63.71	83.87	87.65
multi_claude_claude_gpt	75.63	72.8	92.52	65.09	63.41	70.16	85.55	87.39
multi_llama_gpt_gpt	75.38	72.87	88.89	65.09	68.14	69.35	85.38	88.32
multi_gemini_gemini_gemini	74.96	71.53	87.23	73.55	64.67	60.66	85.21	88.91
multi_gemini_gpt_gpt	74.79	72.01	88.44	72.73	62.66	64.23	84.29	87.73
multi_gemini_gemini_gpt	74.71	72.09	88.46	73.45	60.88	65.57	84.87	87.14
single_gpt-4.1_gpt	74.37	71.08	89.1	71.74	60.57	62.9	83.19	85.55
multi_claude_claude_claude	74.29	70.89	91.53	64.98	59.31	67.74	86.55	88.74
single_gpt_gpt	73.95	71.01	87.82	73.19	59.31	63.71	82.61	84.62
single_llama_gpt	73.95	70.54	87.82	66.3	65.93	62.1	84.79	87.98
single_deepseek_gpt	73.95	70.99	85.68	72.46	63.72	62.1	84.37	86.13
multi_llama_llama_gpt	73.7	70.6	88.22	67.39	63.09	63.71	84.96	87.56
multi_claude_gpt-4.1_gemini-flash	73.19	69.25	91.74	58.12	61.83	65.32	81.43	84.37
multi_llama_gpt-4.1_gemini-flash	72.77	68.54	87.08	61.01	67.19	58.87	82.18	86.13
multi_llama_llama_llama	72.44	69.98	85.47	62.18	66.14	66.13	84.2	88.07
multi_gpt_gpt_claude	71.93	68.35	87.08	64.62	59.62	62.1	83.36	86.89
multi_gpt_gpt-4.1_llama	71.85	69.06	85.04	62.18	65.3	63.71	82.52	86.3
multi_gpt-4.1_gpt-4.1_gpt-4.1	71.76	68.93	85.87	69.09	58.68	62.1	82.94	85.97
single_gemini_gpt	71.09	67.21	81.36	77.98	57.1	52.42	83.03	88.32
multi_gpt_gpt_gpt	71.01	68.36	85.47	58.91	63.72	65.32	81.18	84.29
multi_gpt_claude_claude	70.59	66.8	86.65	61.37	58.68	60.48	82.77	86.55
multi_gpt_gpt-4.1_claude	70.59	67.12	86.44	59.21	59.94	62.9	83.19	87.23
single_claude_gpt	70.59	67.34	85.17	64.98	57.1	62.1	83.11	86.97
multi_gpt-4.1_gpt_gpt	69.33	67.16	83.76	67.64	53	64.23	80.34	83.03
multi_gpt-4.1_gpt-4.1_gpt	69.24	66.9	81.84	70.55	53.94	61.29	80.08	83.19
multi_gpt_gpt-4.1_gemini-flash	67.65	64.6	81.14	54.15	61.83	61.29	76.05	79.16
multi_gpt_claude_gemini-flash	67.39	63.74	80.08	53.43	64.98	56.45	75.97	79.16
single_o3-mini_gpt	63.45	58.76	77.78	57.25	56.47	43.55	74.12	78.57
multi_deepseek_deepseek_deepseek	53.95	55.96	64.3	57.92	51.18	50.43	61.76	63.61

Table 6: Outcome metrics for the single and multi-agent systems on the test set

Agent System	Physical Exam First	Physical Exam Any	Avg Tools	Avg Labs	Avg Img	Avg Physical Exam	Coverage	Coverage-Test Ratio	Lab Interp
multi_gemini-flash_gemini-flash_gpt	72.27	78.15	6.47	4.25	1.33	3.43	0.71	0.11	82.81
multi_gemini-flash_gpt_gpt	71.09	77.56	6.40	4.22	1.37	3.32	0.71	0.11	85.78
multi_o3-mini_o3-mini_o3-mini	59.41	80.76	5.39	3.01	1.56	2.81	0.65	0.12	85.76
multi_claude_gpt_gpt	89.16	94.20	8.27	5.98	1.35	6.23	0.80	0.10	85.43
multi_gemini-flash_gemini-flash_gemini-flash	72.27	78.49	6.44	4.23	1.31	3.46	0.71	0.11	81.68
single_gemini-flash_gpt	49.75	58.15	7.13	5.23	1.28	2.46	0.67	0.09	80.67
multi_claude_claude_gpt	88.99	93.95	8.18	5.88	1.36	6.18	0.79	0.10	85.47
multi_llama_gpt_gpt	50.17	73.78	8.59	5.83	2.02	2.98	0.81	0.09	87.07
multi_gemini_gemini_gemini	44.12	46.64	7.15	5.60	1.08	3.59	0.62	0.09	77.76
multi_gemini_gpt_gpt	44.71	46.97	7.20	5.67	1.05	3.57	0.62	0.09	85.34
multi_gemini_gemini_gpt	43.95	45.97	7.09	5.58	1.04	3.51	0.63	0.09	78.17
single_gpt-4.1_gpt	63.70	64.03	7.06	5.39	1.00	4.73	0.59	0.08	83.99
multi_claude_claude_claude	88.91	94.79	8.25	5.98	1.33	6.27	0.79	0.10	85.22
single_gpt_gpt	49.92	50.08	4.25	2.98	1.03	0.95	0.52	0.12	83.36
single_llama_gpt	16.55	77.90	9.39	6.67	1.92	2.65	0.82	0.09	75.38
single_deepseek_gpt	51.26	64.71	6.78	4.36	1.11	3.50	0.60	0.09	72.63
multi_llama_llama_gpt	50.42	75.46	8.65	5.90	1.99	3.03	0.81	0.09	84.80
multi_claude_gpt-4.1_gemini-flash	88.99	94.71	8.26	5.96	1.35	6.23	0.80	0.10	83.94
multi_llama_gpt-4.1_gemini-flash	50.59	75.71	8.97	6.13	2.04	3.26	0.82	0.09	85.78
multi_llama_llama_llama	54.87	80.08	8.84	6.02	2.01	3.24	0.83	0.09	84.98
multi_gpt_gpt_claude	52.35	52.77	3.61	1.69	1.34	2.36	0.46	0.13	85.50
multi_gpt_gpt-4.1_llama	52.69	52.77	3.51	1.66	1.32	2.30	0.46	0.13	85.40
multi_gpt-4.1_gpt-4.1_gpt-4.1	87.82	88.40	6.91	5.03	0.96	6.88	0.59	0.09	83.54
gemini_claude	29.66	30.34	6.29	5.37	0.61	2.27	0.53	0.08	79.23
multi_gpt_gpt_gpt	52.52	52.94	3.49	1.66	1.29	2.35	0.46	0.13	84.99
multi_gpt_claude_claude	50.84	51.01	3.55	1.70	1.29	2.33	0.46	0.13	82.64
multi_gpt_gpt-4.1_claude	51.26	51.68	3.29	1.60	1.13	2.29	0.45	0.14	86.28
single_claude_gpt	69.16	74.37	7.19	5.52	0.93	4.62	0.69	0.10	82.75
multi_gpt-4.1_gpt_gpt	87.90	88.40	6.77	4.96	0.90	6.74	0.60	0.09	85.71
multi_gpt-4.1_gpt-4.1_gpt	88.66	88.74	6.74	4.88	0.93	6.96	0.59	0.09	83.40
multi_gpt_gpt-4.1_gemini-flash	53.70	54.03	3.62	1.70	1.32	2.47	0.46	0.13	85.50
multi_gpt_claude_gemini-flash	52.10	52.35	3.41	1.59	1.25	2.34	0.45	0.13	82.71
single_o3-mini_gpt	1.93	1.93	0.03	0.01	0.00	0.03	0.00	0.03	73.24
multi_deepseek_deepseek_deepseek	81.76	85.88	7.56	4.86	1.45	4.96	0.76	0.10	82.76

Table 7: Process metrics for the single and multi-agent systems on the test set

Agent System	Computational Cost	Lab Cost
multi_gemini-flash_gemini-flash_gpt	16.26	49.56
multi_gemini-flash_gpt_gpt	21.29	53.63
multi_o3-mini_o3-mini_o3-mini	14.99	43.43
multi_claude_gpt_gpt	37.03	86.43
multi_gemini-flash_gemini-flash_gemini-flash	12.72	48.38
single_gemini-flash_gpt	20.53	62.81
multi_claude_claude_gpt	35.78	87.04
multi_llama_gpt_gpt	20.89	58.62
multi_gemini_gemini_gemini	19.62	80.14
multi_gemini_gpt_gpt	23.34	99.05
multi_gemini_gemini_gpt	21.20	94.38
single_gpt-4.1_gpt	42.21	115.76
multi_claude_claude_claude	38.25	86.97
single_gpt_gpt	36.45	21.66
single_llama_gpt	26.12	71.71
single_deepseek_gpt	40.51	62.74
multi_llama_llama_gpt	15.77	61.74
multi_claude_gpt-4.1_gemini-flash	34.56	84.35
multi_llama_gpt-4.1_gemini-flash	22.90	69.52
multi_llama_llama_llama	12.06	57.29
multi_gpt_gpt_claude	18.85	14.14
multi_gpt_gpt-4.1_llama	15.62	13.43
multi_gpt-4.1_gpt-4.1_gpt-4.1	25.20	90.55
single_gemini_gpt	27.19	106.11
multi_gpt_gpt_gpt	16.40	14.09
multi_gpt_claude_claude	18.47	14.64
multi_gpt_gpt-4.1_claude	20.21	13.78
single_claude_gpt	57.37	85.21
multi_gpt-4.1_gpt_gpt	24.47	78.27
multi_gpt-4.1_gpt-4.1_gpt	24.06	77.02
multi_gpt_gpt-4.1_gemini-flash	15.75	14.36
multi_gpt_claude_gemini-flash	15.22	14.91
single_o3-mini_gpt	0.61	0.05
multi_deepseek_deepseek_deepseek	18.52	44.65

Table 8: Cost Efficiency metrics for the single and multi-agent systems on the test set