

# **A Multi-agent Large Language Model Framework for Medical Text Summarization and Evaluation: Development and Evaluation Study**

Yuhao Chen, Bo Wen, Farhana Zulkernine

Submitted to: JMIR AI  
on: April 15, 2025

**Disclaimer:** © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

## Table of Contents

---

|                                                          |    |
|----------------------------------------------------------|----|
| Original Manuscript.....                                 | 5  |
| Supplementary Files.....                                 | 23 |
| Figures .....                                            | 24 |
| Figure 1.....                                            | 25 |
| Figure 2.....                                            | 26 |
| Multimedia Appendixes .....                              | 27 |
| Multimedia Appendix 1.....                               | 28 |
| Multimedia Appendix 2.....                               | 28 |
| Multimedia Appendix 3.....                               | 28 |
| Related publication(s) - for reviewers eyes onlies ..... | 29 |
| Related publication(s) - for reviewers eyes only 0.....  | 29 |
| TOC/Feature image for homepages .....                    | 30 |
| TOC/Feature image for homepage 0.....                    | 31 |

# A Multi-agent Large Language Model Framework for Medical Text Summarization and Evaluation: Development and Evaluation Study

Yuhao Chen<sup>1</sup> MSc; Bo Wen<sup>2</sup> PhD; Farhana Zulkernine<sup>3</sup> PhD

<sup>1</sup> School of Computing Queen's University Kingston CA

<sup>2</sup> Digital Health Department IBM (United States) Yorktown Heights US

<sup>3</sup> Queen's University Kingston CA

## Corresponding Author:

Yuhao Chen MSc

School of Computing  
Queen's University  
557 Goodwin Hall, Queen's University  
Kingston  
CA

## Abstract

**Background:** Although Large Language Models (LLMs) show great promises in processing medical text, they are prone to generating incorrect information, commonly referred to as hallucinations. These inaccuracies present a significant risk for clinical applications where precision is critical. Additionally, relying on human experts to review LLM-generated content to ensure accuracy is costly and time-consuming, which sets a barrier against large-scale deployment of LLMs in healthcare settings.

**Objective:** The primary objective of this study is to develop an automatic Artificial Intelligence (AI) system capable of extracting structured information from unstructured medical data and employing advanced reasoning techniques to support reliable clinical decision making. A key aspect of this objective is ensuring that the system incorporates self-verification mechanisms, enabling it to assess the accuracy and reliability of its own outputs. By integrating such mechanisms, we aim to enhance the system's robustness, reduce reliance on human intervention, and improve the overall trustworthiness of AI-driven medical summarization and evaluation.

**Methods:** The proposed framework comprises two layers: a summarization layer and an evaluation layer. The summarization layer employs Llama2-70B and Mistral-7B models to generate concise summaries from unstructured medical data, focusing on tasks such as consumer health question summarization, biomedical answer summarization, and dialog summarization. The evaluation layer uses GPT-4-turbo as a judge, leveraging pairwise comparison strategies and different prompt strategies to evaluate summaries across four dimensions: coherence, consistency, fluency, and relevance. To validate the framework, we compare the judgments generated by the LLMs in the evaluation layer with those provided by medical experts, offering valuable insights into the alignment and reliability of AI-driven evaluations within the medical domain. We also explore a way to handle disagreement among human experts and discuss our methodology in addressing diversity in human perspectives.

**Results:** The study found variability in expert consensus, with average Agreement Rates (ARs) of 19.2% among all experts and 51.6% among groups of three experts. GPT-4 demonstrated alignment with expert judgments, achieving an average AR of 78.44% with at least one expert and comparable performance in cross-validation tests. The enhanced guidance in prompt design improved GPT-4's alignment with expert evaluations, highlighting the importance of effective prompt engineering in auto-evaluation of summarization tasks.

**Conclusions:** This study highlights the potential of LLMs as reliable tools for medical summarization and evaluation, reducing the dependency on human experts. The proposed framework demonstrates scalability and adaptability for clinical applications while addressing key challenges like hallucination and position bias.

(JMIR Preprints 15/04/2025:75932)

DOI: <https://doi.org/10.2196/preprints.75932>

## Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in [JMIR Publications](#)

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in [JMIR Publications](#)

## Original Manuscript

# A Multi-agent LLM Framework for Medical Text Summarization and Evaluation: Development and Evaluation Study

Yuhao Chen<sup>1</sup>; Bo Wen<sup>2</sup>; Farhana Zulkernine<sup>1</sup>

<sup>1</sup>School of Computing, Queen's University, Kingston, Ontario, Canada

<sup>2</sup>Digital Health Department, IBM Research, Yorktown Heights, NY, USA

## Abstract

### Background:

Although Large Language Models (LLMs) show great promises in processing medical text, they are prone to generating incorrect information, commonly referred to as hallucinations. These inaccuracies present a significant risk for clinical applications where precision is critical. Additionally, relying on human experts to review LLM-generated content to ensure accuracy is costly and time-consuming, which sets a barrier against large-scale deployment of LLMs in healthcare settings.

### Objective:

The primary objective of this study is to develop an automatic Artificial Intelligence (AI) system capable of extracting structured information from unstructured medical data and employing advanced reasoning techniques to support reliable clinical decision making. A key aspect of this objective is ensuring that the system incorporates self-verification mechanisms, enabling it to assess the accuracy and reliability of its own outputs. By integrating such mechanisms, we aim to enhance the system's robustness, reduce reliance on human intervention, and improve the overall trustworthiness of AI-driven medical summarization and evaluation.

### Methods:

The proposed framework comprises two layers: a summarization layer and an evaluation layer. The summarization layer employs Llama2-70B and Mistral-7B models to generate concise summaries from unstructured medical data, focusing on tasks such as consumer health question summarization, biomedical answer summarization, and dialog summarization. The evaluation layer uses GPT-4-turbo as a judge, leveraging pairwise comparison strategies and different prompt strategies to evaluate summaries across four dimensions: coherence, consistency, fluency, and relevance. To validate the framework, we compare the judgments generated by the LLMs in the evaluation layer with those provided by medical experts, offering valuable insights into the alignment and reliability of AI-driven evaluations within the medical domain. We also explore a way to handle disagreement among human experts and discuss our methodology in addressing diversity in human perspectives.

### Results:

The study found variability in expert consensus, with average Agreement Rates (ARs) of 19.2% among all experts and 51.6% among groups of three experts. GPT-4 demonstrated alignment with expert judgments, achieving an average AR of 78.44% with at least one expert and comparable performance in cross-validation tests. The enhanced guidance in prompt design improved GPT-4's alignment with expert evaluations, highlighting the importance of effective prompt engineering in auto-evaluation of summarization tasks.

### Conclusions:

This study highlights the potential of LLMs as reliable tools for medical summarization and evaluation, reducing the dependency on human experts. The proposed framework demonstrates scalability and adaptability for clinical applications while addressing key challenges like hallucination and position bias.

**Keywords:** large language model evaluation; medical summarization; multi-agent network; summarization evaluation; LLM as a judge

## Introduction

In the era of big data, healthcare presents unique challenges for researchers, particularly due to the vast amount of unstructured data that can significantly impact patient outcomes [1]. A key challenge is extracting structured information from unstructured medical texts to facilitate search, analysis, and summarization for clinical decision-making [2]. Automatic systems often struggle with information extraction and summarization from medical data due to the length and noise in clinical texts [3], making accurate interpretation challenging.

The rapid advancement of Artificial Intelligence (AI) is transforming the healthcare sector with Large Language Models (LLMs) like GPT-4 [4], which demonstrate impressive capabilities across a range of medical tasks [5-9]. For instance, older models like Med-PaLM and, subsequently, Med-PaLM 2, achieved 67.2% and 86.5% respectively on medical question text summarization tasks with U.S. Medical Licensing Examination (USMLE)-style questions. Newer LLMs like GPT-4, on the other hand, achieved an impressive 90% accuracy on USMLE questions, compared to the 78% average score achieved by human test-takers [5]. GPT-4 also ranked within the top 10% of all participants on German medical licensing exams among medical students [6]. Other models such as Gemini [10][11] and Llama3-OpenBioLLM [12][13], have also shown robust performance in analyzing medical data. Such results highlight LLMs' potential to support medical education and examination systems and aid healthcare practitioners in evidence-based decision support. However, despite these promising developments, LLM applications in healthcare remain primarily experimental. One of the critical limitations is hallucination, where LLMs generate content that is incoherent or factually incorrect [14][15]. This issue poses significant risks in clinical settings, as hallucinated outputs can lead to erroneous diagnoses, inappropriate treatments, mental distress, and reduced trust in AI-driven systems. Conventional similarity-based evaluation metrics, such as ROUGE [16] and BERTSCORE [17] are not able to detect hallucinations. Moreover, these similarity-based automatic evaluation metrics require expert-generated gold standard summaries. Compared to other domains, medical text annotation requires expertise from specialized healthcare professionals, making it significantly more expensive. With the success of the advanced LLM model, there has been a growing interest in developing human-free automatic evaluation frameworks that can leverage LLM's capabilities.

This research is motivated by the growing need for accurate medical summaries, which are critical for clinical reviews, or assessments for insurance and audit processes. The vast volume of daily medical data, such as patient records and lab results, highlights the demand for advanced automatic summarization technologies. Traditional methods, such as manual summarization or rule-based systems, often struggle to efficiently extract relevant information, leading to delays or inaccuracies in clinical workflows.

To address these challenges, this study aims to answer the following primary research question:

“How to build an AI system that can self-evaluate its performance across various medical summarization tasks?” It specifically explores the following key aspects:

1. How well can LLMs generate summaries from medical domain specific data?
2. How do LLMs perform compared to human experts in evaluating the generated summaries in terms of reliability?
3. How do prompt design and position bias influence the reliability of LLM-based evaluations, and how to mitigate the impact?

To answer the above questions, this study focuses on developing an automatic medical question, dialogue, and descriptive text summarization and evaluation system, which serves as a case study to explore how LLMs can improve the extraction of essential insights from complex, unstructured medical data. The key objectives include ensuring medical accuracy, relevance, and proper evaluation of the generated summaries, as a foundational step toward broadening AI applications in healthcare data processing.

## Methods

### Overview

The overall architecture of our proposed system is shown in Fig. 1, which represents a multi-layer LLM agent network composed of two distinct layers. It is inspired by IBM’s SOLOMON (System for Optimizing Language Outputs through Multi-agent Oversight Networks) framework [18].

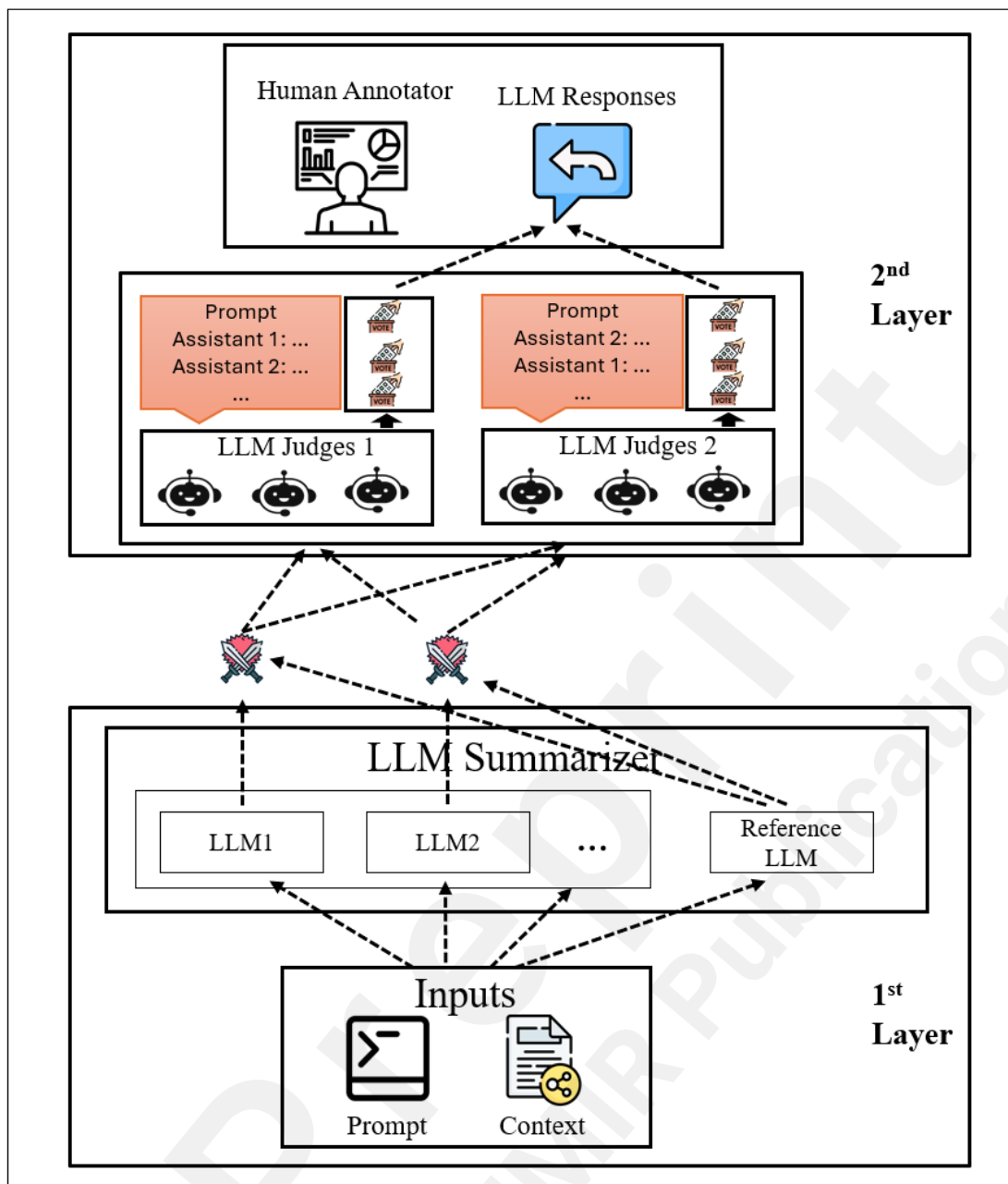


Figure 1. The proposed multi-agent LLM framework with two layers performing summarization and then evaluation.

## Summarization Layer

The network begins with a summarization layer, where LLM summarizers extract key information from unstructured data sources, such as patient inquiries and doctor-patient conversations. This initial layer focuses on transforming lengthy and disorganized medical content into concise summaries that retain essential information for further analysis.

## Evaluation Layer

The second layer, the evaluation layer, deploys LLM judges to assess the quality of the summaries generated by the summarization layer. Rather than using judges to grade each LLM's summaries sequentially, we implement a pairwise comparison strategy where LLM judges compare and evaluate

a pair of outputs to determine the superior one. This approach enables judges to focus on the subtle differences between the two responses, highlighting strengths and weaknesses that might otherwise be overlooked, thereby enhancing the overall precision of the evaluation process.

We adopted the *Referenced Condorcet Assessment (RCA)* [19] approach to optimize our evaluation process. Instead of comparing every model's output against all others, RCA designates one "reference model" and compares each model's output against this reference. This significantly reduces the required number of comparisons, optimizing computational efficiency while maintaining accuracy in identifying high-quality summaries. By establishing a consistent reference point, RCA also ensures consistent evaluation across different datasets. This approach provides reproducibility and improves the overall efficiency of the AI-evaluation layer.

In the evaluation layer, we also implement a *majority voting mechanism* using multiple LLM evaluators to enhance confidence in the decision. A final decision is reached when more than half of the evaluators agree; otherwise, the result is marked as a tie. To address *position bias*, where the order of inputs (summaries) affects output (evaluation)—we employ a *swap-position strategy*. This involves evaluating each summary pair twice with reversed ordering. A summary is declared the winner only if it is preferred in both evaluations; otherwise, the outcome is considered as a tie.

This two-layer approach implements a comprehensive processing pipeline that distills and validates medical information, contributing to more effective and timely healthcare decision-support.

## Model Selection

We select two open-source LLMs to generate summaries in the initial summarization layer. At the time of this study, Llama2 was the state-of-the-art (SOTA) open-source model, making the Llama2-70B-chat-hf version an ideal choice [20]. We also selected Mistral-7B-Instruct-v0.1, which was optimized for faster inference and demonstrated superior performance, surpassing the larger Llama2-13B model in some tasks [21]. By choosing Mistral-7B-Instruct-v0.1 over other Llama2 variants, we benefited from both its strong performance and architectural diversity, reducing the risk of model-specific biases. GPT-3.5 was reported as a strong baseline model in a variety of research [6][8]. Therefore, we selected that as the RCA reference model in this study. For the LLM evaluator, we selected GPT-4-turbo, which was the SOTA LLM reasoning model at the time of this study. A summary of all the LLM models used in this study is given below.

- Llama2-70B-chat-hf (*summarizer*): Llama2 [20] is a collection of transformer decoder-based LLMs, ranging from 7 billion to 70 billion parameters. We use its official chat fine-tuned version, Llama2-70B-chat-hf.
- Mistral-7B-Instruct-v0.1 (*summarizer*): Mistral [21] is a transformer decoder-based LLM with 7 billion parameters, achieving superior performance and efficiency among small-scale LLMs. We select the official instruction fine-tuned version, Mistral-7B-Instruct-v0.1.
- GPT-3.5-turbo (*reference*): GPT-3.5-turbo is another transformer decoder-based LLM with undisclosed architecture and parameters.
- GPT-4-turbo (*evaluator*): GPT-4-turbo is a large-scale, multimodal model based on decoder-based LLM with undisclosed architecture and parameters.

Two key aspects of the proposed architecture and methodology include the robust information processing framework and the adaptability to use any LLM.

## Data Sources

We validate our system for three distinct summarization tasks with different benchmark data sets presenting diverse challenges as described below. These datasets are commonly used in medical text analysis tasks [22].

1. *Task 1 - Consumer Health Question Summarization:* Patients often ask detailed and complex health-related questions, which can be difficult for automated systems to process efficiently. The primary challenge lies in preserving essential medical details while reducing redundancy. MEDIQA-QS [23] and MeQSum [24] datasets contain long and complex consumer health questions. Summaries of such questions can help to analyze service quality, customers' needs and concerns, and thereby, improve customer support.
  - *MEDIQA-QS* comprises de-identified consumer health questions sourced from the U.S. National Library of Medicine (NLM), with gold-standard summaries crafted by medical experts.
  - *MeQSum* is designed for medical problem summarization tasks. It contains 1,000 consumer health questions summarized to facilitate natural language processing (NLP) research specifically for the medical field.
2. *Task 2 - Biomedical Answer Summarization:* Medical queries often receive multiple responses from various sources, requiring aggregation into a coherent summary. The challenge in this task is to extract and generate relevant non-redundant answers based on a given question while ensuring consistency across diverse sources. An important practical use case is to aggregate and summarize multiple relevant responses to a medical query. We address this problem using MEDIQA-ANS [25] and MEDIQA-MAS [23] datasets.
  - *MEDIQA-MAS* consists of collections of responses to specific health-related queries, along with summaries created by experts that integrate information from multiple answers into a single coherent summary.
  - *MEDIQA-ANS* contains health questions asked by consumers, their corresponding answers, and expert-created summaries. The dataset is sourced from reliable organizations such as MedlinePlus and derived from questions submitted to the National Library of Medicine's CHiQA system.
3. *Task 3 - Dialog Summarization:* Online medical advising uses interactive chat or voice to learn customer needs and provides appropriate responses based on expert knowledge. Key challenges in processing the data include handling informal expressions, identifying and filtering relevant medical information, and maintaining the temporal context of symptoms and events. Dialogue summarization is also important to learn about previous information exchanges to provide informed responses. We aim to produce a summary of a patient's medical condition from a dialogue between the patient and the doctor. We use the iCliniq [26] dataset to validate our approach for this task.
  - *iCliniq* is an extensive collection of real-life anonymized medical consultations between patients and physicians. It covers a wide range of medical inquiries and responses, demonstrating a doctor's approach to various health problems.

For each dataset, we randomly selected 50 samples for evaluation. Examples of data samples are shown in Multimedia Appendix 1.

## Prompt Design

In this study, we explore the impact of prompt design on the performance of LLMs across different layers of functionality. Specifically, we distinguish between two key components: (1) the summarizer, which generates summaries based on input data, and (2) the evaluator, which assesses the quality of these summaries.

### *Prompt Design for Summarizer*

For effective benchmarking and comparison with existing research, we adapted the prompts proposed by Jahan et al. [27] for the summarizer, leveraging their established prompt templates as a foundation as illustrated in Multimedia Appendix 2. This approach keeps the prompt design for the summarizer constant, serving as a control variable. By maintaining this consistency, it allows for the analysis of different prompt design approaches for the evaluator, ensuring that any variations in performance can be attributed to the evaluator's prompt design rather than changes in the summarizer's prompt.

### *Prompt Design for Evaluator*

For the evaluator, we introduce two custom prompt design approaches and systematically assess how variations in evaluator prompts influence the quality and reliability of the evaluation process. By comparing the prompts, we aim to identify optimal strategies for leveraging LLMs as evaluators in summarization tasks.

*Prompt-baseline:* Our initial prompt design (Multimedia Appendix 2) is inspired by the work of Zeng et al. [4], which pioneers the use of an LLM as a judge in evaluating the performance of other LLMs in primarily question-answering tasks. To adapt their approach for summarization tasks, we enhanced the prompt to ensure that GPT-4's evaluation focuses on four key dimensions commonly used by humans: *coherence*, *consistency*, *fluency*, and *relevance*. Additionally, we include statements to mitigate positional bias so that the order of summaries does not influence the LLM's judgment.

*Prompt-Enhanced Guidance (EG):* Several aspects of the baseline prompt required improvements, as the instructions in the prompt were broad and ambiguous. Terms such as "coherence," "consistency," "fluency," and "relevance" lack clear operational definitions, which can lead to inconsistencies in evaluation. The GPT-4 evaluator required more specific guidance to apply these criteria consistently. To address the limitations of the prompt-baseline, we developed **Prompt-EG**, a significantly more comprehensive prompt template that provides clearer and more detailed instructions. The update includes detailed descriptions of the four key evaluation criteria and integrates a Chain-of-Thought (CoT) [28] guideline, which guides GPT-4 through a structured, step-by-step evaluation process. Prompt-EG also introduces several new instructions for: 1) balancing brevity and detail to facilitate the selection of a high-quality summary; 2) detecting hallucinations to remove improper content from the generated summary, and 3) ensuring factual consistency, which is the most critical evaluation factor in medical domain. The prompt template is explained in Multimedia Appendix 2.

## Validation by Human Experts

To evaluate the efficacy of our two-layer framework, we designed a comprehensive evaluation protocol involving human experts from the medical domain. This approach enabled us to quantify inter-rater agreement among the medical experts, as well as the concordance between human evaluators and LLM judges. It also provided valuable insights into qualitative assessment based on

the consistency and reliability of generated medical summaries.

For our evaluation, we selected summaries generated by Llama 2 (70B parameters) and GPT-3.5 for their strong performance in medical content generation. After the generation phase, each output from Llama 2 was paired with the corresponding summary from GPT-3.5. To prevent potential evaluation bias, the pairs were anonymized and shuffled. A panel of 4 medical experts then independently evaluated each pair, indicating their preferences based on clinical accuracy, completeness, and clarity.

### ***Variability in Expert Consensus***

Assessing the degree of expert consensus is important as it provides insights into the diversity of expert opinions. A high level of variability in expert judgments suggests the inherent subjectivity in the evaluation of medical summaries. Table 1 highlights the variability in consensus across different expert group sizes. Notably, the average AR where all four experts agree with each other is only 19.2%. However, this rate significantly increases to 51.6% when considering cases where at least three experts agree. This disparity underscores potential differences in expert interpretation or judgment, reflecting the complexity of achieving unanimous agreement in this task.

Table 1. Rate of agreement among experts: Presents the percentage of times all 4 or at least 3 or 2 experts agree in selecting the better summary. The calculation details are provided in Appendix 3 Algorithm 1.2.

| <b>Dataset</b> | <b>All 4 Experts in Agreement</b> | <b>At Least 3 Experts in Agreement</b> | <b>At Least 2 Experts in Agreement</b> |
|----------------|-----------------------------------|----------------------------------------|----------------------------------------|
| MEDIQA-QS      | 12%                               | 36%                                    | 100%                                   |
| MeQSum         | 16%                               | 52%                                    | 100%                                   |
| MEDIQA_ANS     | 28%                               | 68%                                    | 100%                                   |
| MEDIQA_MAS     | 18%                               | 56%                                    | 100%                                   |
| iCliniq        | 22%                               | 46%                                    | 100%                                   |
| <b>Average</b> | <b>19.2%</b>                      | <b>51.6 %</b>                          | <b>100%</b>                            |

Table 2. Average AR between GPT-4 judge with Prompt-EG and each of the 4 experts in preferred summary selection for 5 datasets. The calculation details are provided in Appendix 3 Algorithm 1.4.

| <b>Human Experts</b> | <b>Average AR between GPT-4 and Human Expert</b> |
|----------------------|--------------------------------------------------|
| Expert 1             | 42.18%                                           |
| Expert 2             | 43.79%                                           |
| Expert 3             | 50.54%                                           |
| Expert 4             | 36.65%                                           |

Table 2 reveals significant variations in GPT-4's alignment with individual expert preferences. The highest and lowest agreements of 50.54% and 36.65% with Experts 3 and 4 respectively demonstrate that LLMs inherently encapsulate diverse human perspectives from their training data. This discrepancy underscores that LLMs can access different viewpoints within their knowledge base when handling subjective tasks, particularly through carefully designed prompts and parameter configurations (temperature, top-p, top-k). This observation suggests valuable opportunities for preference customization through advanced prompt engineering.

### ***GPT-4's Evaluation Capability in the Context of Expert Disagreement***

To assess GPT-4's potential as an evaluation tool, we measured its agreement patterns with the medical experts in selecting better summaries. Table 3 shows that GPT-4 achieves an average AR of 78.44% with at least one of the medical experts across datasets. This finding is particularly meaningful when considered alongside the substantial disagreement observed among experts themselves.

Table 3. GPT-4's alignment with at least one expert in preferred summary selection. The calculation details are provided in Appendix 3 Algorithm 1.2.

| <b>Dataset</b> | <b>Rate of Agreement with at least 1 Expert</b> |
|----------------|-------------------------------------------------|
| MEDIQA-QS      | 78.72%                                          |
| MeQSum         | 75.51%                                          |
| MEDIQA_ANS     | 82.00%                                          |
| MEDIQA_MAS     | 76.00%                                          |
| iCliniq        | 80.00%                                          |
| <b>Average</b> | <b>78.44%</b>                                   |

Table 4: ARs among experts, non-experts, and GPT-4 across three simpler medical summarization datasets that do not require in-depth medical knowledge. The evaluations span different datasets and prompt-EG. The calculation details are provided in Appendix 3 Algorithm 1.4.

| <b>Dataset</b> | <b>AR:<br/>Expert vs GPT-4</b> | <b>AR:<br/>Non-expert vs GPT-4</b> | <b>AR:<br/>Expert vs Nonexpert</b> |
|----------------|--------------------------------|------------------------------------|------------------------------------|
| MEDIQA-QS      | 51.06%                         | 57.45%                             | 42.55%                             |
| MeQSum         | 46.94%                         | 75.51%                             | 38.78%                             |
| iCliniq        | 54.00%                         | 60.00%                             | 34.0%                              |
| <b>Average</b> | <b>50.67%</b>                  | <b>64.32%</b>                      | <b>38.44%</b>                      |

In addition to expert evaluations, we conducted non-expert reviews on the shorter and simpler datasets to assess the generalizability, accessibility, and usability of the generated summaries for a broader audience. While medical experts provide critical insights into accuracy and relevance, non-experts offer valuable perspective on clarity and comprehensibility for individuals without specialized medical knowledge. This dual evaluation approach is particularly important in healthcare settings, where patients and caregivers often rely on the summary of medical information. Since the non-experts reported difficulty in identifying factual inaccuracies due to limited medical expertise for the long documents in MEDIQA\_ANS and MEDIQA\_MAS during the review process, we excluded these datasets in this experiment. This way we could demonstrate the general understanding of the non-expert population compared to the experts and GPT-4 in summarization tasks.

Table 4 reveals that GPT-4's judgments align more closely with non-experts (64.32%) than with experts (50.67%) which is expected. This differential alignment can be attributed to GPT-4's training methodology. During fine-tuning, GPT-4 was likely optimized using non-expert human feedback that prioritized general comprehensibility and user satisfaction. However, GPT-4's pretraining on diverse datasets, including medical corpora, provided it with domain knowledge that enables partial alignment with expert judgements. This explains why the AR between GPT-4 and experts (50.67%) exceeds that between experts and non-experts (38.44%), suggesting GPT-4 occupies a middle ground between specialized and general regarding evaluation perspectives.

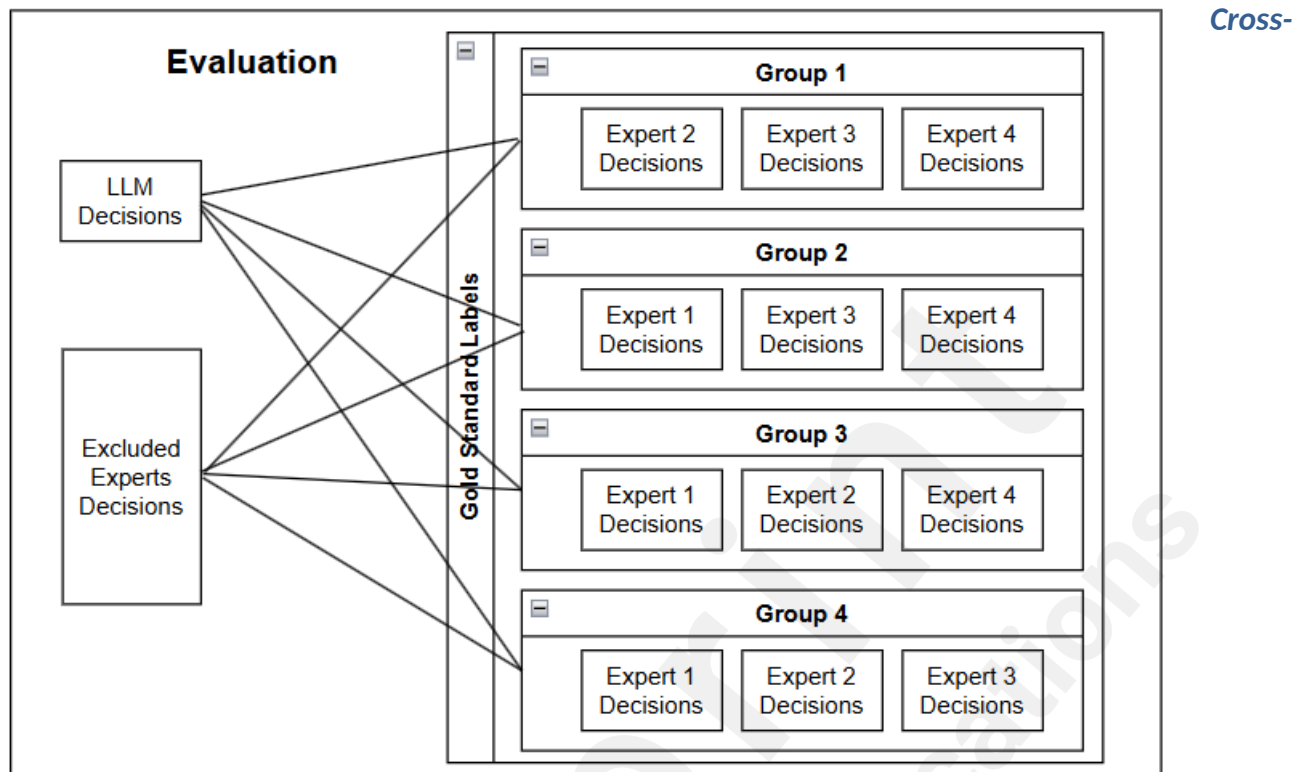


Figure 2. The cross-validation framework. In each iteration, one expert is excluded from the group of four, and the decisions of the remaining three experts are used as the gold standard. GPT-4's decisions and the excluded expert's decisions are then compared against this gold standard, enabling a fair assessment of performance and individual alignment across all groups.

### Validation Framework for Evaluating Judgment Consistency

To establish a fair evaluation of ARs, we employ cross-validation by systematically excluding one expert at a time from the group of four. This process generated four groups, with the majority decision of the remaining three experts serving as the reference or gold standard annotation for each data point. We then compared both GPT-4's decisions and the excluded expert's decisions against this reference, allowing us to assess how individual human evaluators and the LLM perform relative to the same benchmark.

This methodology serves multiple purposes: it evaluates the stability of judgment distributions, identifies potential outliers within the expert group, and provides a consistent framework for comparing human and LLM performance. Table 5 shows that Expert 2 (excluded from Group 2) achieved the highest AR of 55.62% with other experts, while Expert 3 (excluded from Group 3) exhibited the lowest AR of 40.49%. The absence of extreme outliers suggests a natural distribution of expert perspectives rather than anomalous judgments.

Table 5. ARs between GPT-4 and excluded experts in the cross-validation framework. The table shows that the excluded expert achieves a slightly higher average AR (47.38%) compared to GPT-4 (46.95%), highlighting GPT-4's close approximation to human-level performance in selecting better summaries.

| Annotation | Excluded | GPT-4 AR | Llama-3.3 | Mixtral- |
|------------|----------|----------|-----------|----------|
|------------|----------|----------|-----------|----------|

| Group   | Expert's AR   |               | AR     | Large AR |
|---------|---------------|---------------|--------|----------|
| Group 1 | <b>49.87%</b> | 46.02%        | 42.62% | 48.71%   |
| Group 2 | <b>55.62%</b> | 44.33%        | 43.36% | 44.60%   |
| Group 3 | 40.49%        | <b>44.63%</b> | 39.17% | 38.57%   |
| Group 4 | 43.54%        | <b>52.81%</b> | 49.74% | 42.19%   |
| Average | <b>47.38%</b> | 46.95%        | 43.72% | 43.52%   |

Notably, excluded experts achieved an average AR of 47.38%, compared to GPT-4's 46.95%. This narrow gap demonstrates that GPT-4's evaluation capability approximates human-level consistency in this experimental context. We also assessed newer open-source models, including Llama-3.3 and Mistral-Large, finding their performance (43.72% and 43.52% respectively) to be slightly below GPT-4 but still within the range of human variation (lowest 40.49%). These findings suggest that current LLMs can function as supplementary evaluation tools in scenarios where expert resources are limited, though they should not be considered as replacements for human judgment.

## Results

To systematically assess the effectiveness of our proposed framework, we conducted a series of experiments designed to evaluate both the summarization and evaluation layers. These experiments provide a comprehensive assessment of LLMs' summarization capabilities in medical contexts while acknowledging inherent limitations. Examples of GPT-4's judgements are shown in Multimedia Appendix 1.

## Comparative Performance of Llama-2-70B and Mistral-7B Across Datasets

Our first experiment focused on comparing the summarization quality of Llama-2-70B and Mistral-7B across multiple medical summarization datasets. Table 6 presents win rates against GPT-3.5 as a reference point. Llama-2-70B consistently achieved higher win rates than GPT-3.5 across all datasets, whereas Mistral-7B generally performed below GPT-3.5. These comparative results suggest Llama-2-70B's superior performance in medical summarization tasks within our experimental framework.

Table 6. Pairwise win rates (%) of Llama -2-70b and Mistral-7b against GPT-3.5 as a reference across the different medical summarization datasets at the summarization layer of our framework. These summaries are then evaluated by GPT-4 and medical experts. The win rate indicates the percentage of instances where GPT-4 or medical experts determined that a given model produced a superior summary. The "Tie" column represents cases where GPT-4 found no clear preference between the models. For each dataset, the upper win rates show GPT-4's evaluation, while the lower win rates show expert judgments.

| Dataset    | Llama-2-70B Win Rate (%)             |         |        | Mistral-7B Win Rate (%) |         |        |
|------------|--------------------------------------|---------|--------|-------------------------|---------|--------|
|            | GPTs' Judgment vs Experts' Judgement |         |        |                         |         |        |
|            | Llama2                               | GPT-3.5 | Tie    | Mistral                 | GPT-3.5 | Tie    |
| MEDIQA     | 38.30%                               | 25.53%  | 36.17% | 18.00%                  | 30.00%  | 52.00% |
|            | 19.15%                               | 23.40%  | 57.45% | N/A                     | N/A     | N/A    |
| MeQSum     | 38.78%                               | 10.20%  | 51.02% | 14.00%                  | 40.00%  | 46.00% |
|            | 32.65%                               | 12.24%  | 55.10% | N/A                     | N/A     | N/A    |
| MEDIQA-ANS | 38.00%                               | 24.00%  | 38.00% | 34.00%                  | 20.00%  | 46.00% |
|            | 42.00%                               | 26.0%   | 32.00% | N/A                     | N/A     | N/A    |
| MEDIQA-    | 46.00%                               | 28.0%   | 26.00% | 34.00%                  | 44.00%  | 22.00% |

|         |        |        |        |        |        |        |
|---------|--------|--------|--------|--------|--------|--------|
| MAS     | 50.00% | 6.00%  | 44.00% | N/A    | N/A    | N/A    |
| iCliniq | 44.00% | 22.00% | 34.00% | 28.00% | 36.00% | 36.00% |
|         | 34.00% | 12.00% | 54.00% | N/A    | N/A    | N/A    |
| Average | 41.82% | 17.55% | 40.64% | 25.60% | 34.00% | 40.40% |
|         | 35.56% | 15.93% | 48.51% | N/A    | N/A    | N/A    |

## Impact of Prompt Design on Evaluation Consistency

Next, we examined how variations in prompt design affected the consistency of GPT-4's assessments with expert judgements. Table 7 compares ARs between GPT-4 and expert majority decisions across different prompt strategies. The Prompt-EG approach generally outperformed the baseline strategy, achieving higher alignment with expert consensus for most datasets. This result highlights the importance of effective prompt engineering in improving LLMs' evaluation consistency, though the improvements vary by task type and dataset characteristics.

Table 7. ARs between GPT-4 judge and the majority decision of 4 human experts for all datasets, prompt strategies, and summarization models of GPT-3.5 and LLama2-70B in preferred summary selection. The calculation details are provided in Appendix 3 Algorithm 1.4.

| Dataset    | Prompt   | AR            |
|------------|----------|---------------|
| MEDIQA-QS  | Baseline | 42.55%        |
|            | EG       | <b>51.06%</b> |
| MeQSum     | Baseline | 42.86%        |
|            | EG       | <b>46.94%</b> |
| MEDIQA_ANS | Baseline | <b>62.00%</b> |
|            | EG       | 57.99%        |
| MEDIQA_MAS | Baseline | 44.00%        |
|            | EG       | <b>48.00%</b> |
| iCliniq    | Baseline | 48.00%        |
|            | EG       | <b>54.00%</b> |

## Task-Specific Trends and Position Bias

Position bias, where the order of presented summaries influences evaluation outcomes, poses a methodological challenge in LLM-based assessments. To address this, we implemented a swap-position strategy, evaluating summaries in reversed order to detect and mitigate ordering effects. Table 8 illustrates agreement patterns under different sample conditions.

Our analysis revealed task-dependent patterns in performance. For *question summarization* (e.g., *MEDIQA-QS*) and *dialog summarization* (e.g., *iCliniq*), ARs were higher for samples with position bias. Further investigation showed that in these position-biased samples, experts themselves often produced diverse evaluations, resulting in ties or divergent decisions. This diversity suggests that summary quality in these cases was subjectively equal, with no clear superior option. GPT-4's position bias likely stems from sampling different human perspectives from its pretraining knowledge—when summary quality is comparable, the model draws from the diverse human preferences encoded in its parameters, leading to different judgments when presentation order changes. This behavior demonstrates that GPT-4 is inconsistent as it aligns with differing human perspectives that exist within its knowledge base.

In contrast, the *answer summarization task* showed higher ARs for samples without position bias. This pattern likely stems from the more structured nature of answer summarization, which involves extracting answers directly from the provided document, thereby reducing variability in subjective interpretation. This observation is consistent with Table 8's data showing a higher proportion of expert consensus in answer summarization compared to the other tasks.

Table 8. ARs between GPT-4 judge and the majority decision of 4 human experts for all samples, samples with and without position bias during GPT-4 evaluation. The evaluations span different datasets and prompt-EG. The calculation details are provided in Appendix 3 Algorithm 1.4.

| Dataset    | AR (All Samples) | AR (Samples with Position Bias) | AR (Samples without Position Bias) |
|------------|------------------|---------------------------------|------------------------------------|
| MEDIQA-QS  | 51.06%           | <b>100.00%</b>                  | 43.9%                              |
| MeQSum     | 46.94%           | 50.00%                          | 46.34%                             |
| MEDIQA_ANS | 57.99%           | 50.00%                          | 60.52%                             |
| MEDIQA_MAS | 48.00%           | 46.15%                          | 48.65%                             |
| iCliniq    | 54.00%           | <b>70.59%</b>                   | 45.45%                             |

## Discussion

### Key Observations and Implications

This study introduces a multi-agent framework designed to enhance medical summarization and evaluation tasks using LLM. The first layer, the summarization layer, focuses on extracting concise and essential information from unstructured medical data such as patient inquiries and doctor-patient conversations. This transformation enhances data clarity and prepares it for further assessment. The second layer, the evaluation layer, employs LLM judges for pairwise comparison of summaries, leveraging strategies like the *Referenced Condorcet Assessment* to reduce comparison workload and enhance accuracy. A majority-based decision-making approach, coupled with a swap-position method to mitigate position bias, ensures consistent and unbiased evaluations. This two-layer design creates a robust framework for efficient medical summarization and evaluation, contributing to improved healthcare decision-support.

Our findings reveal variable ARs between GPT-4 and medical experts across datasets, reflecting both the inherent subjectivity of summary evaluation and the diversity of expert perspectives. Rather than indicating a failure of the evaluation system, this variation demonstrates that qualitative assessments of summary lack a single "ground truth" even among human experts. GPT-4's ability to align with different expert viewpoints suggests it can access multiple perspectives within its knowledge base—a capability that could be leveraged to customize and personalize evaluation systems that align with specific user preferences or institutional guidelines.

The prompt-EG strategy generally improved alignment with expert consensus, indicating that careful prompt engineering can enhance LLM evaluation consistency. However, the effectiveness varied by task type, with structured tasks like answer summarization showing different patterns than more subjective tasks like dialog summarization. These task-dependent variations underscore the importance of context-specific system design.

Cross-validation results demonstrate that GPT-4's ARs (46.95%) closely approximate those of

individual experts when compared against the same reference standard (47.38%). This finding suggests that current LLMs can achieve human-comparable consistency in evaluation tasks, though with important limitations. The high alignment rate (78.44%) with at least one expert further indicates that GPT-4 can capture perspectives within the natural distribution of human judgment, making it potentially valuable as a complementary evaluation tool rather than a replacement for human expertise.

Our findings should be interpreted in the context of medical summarization tasks—the observed alignment patterns may not generalize to other domains with different subjectivity profiles. The framework's performance demonstrates the feasibility of LLM-assisted evaluation in specialized domains, particularly when human resources are limited or when rapid and scalable assessments are required.

## Limitations

Despite the promising results, several limitations need to be considered. First, our study involved a limited number of medical experts, which may not capture the full diversity of professional opinions in medical summarization tasks. Future work should expand the expert panel to establish more robust benchmarks for human agreement patterns.

Second, we evaluated a limited set of LLMs. As model architectures and training methodologies evolve rapidly, expanding the range of assessed models would provide more comprehensive insights into the capabilities and limitations of different approaches.

Third, while our framework shows promises for evaluation tasks, the quality of generated summaries remains variable. Enhancing summary quality through domain-specific fine-tuning, retrieval-augmented generation (RAG), hybrid human-AI workflows, or leveraging our evaluation system as a reward signal for reinforcement learning approaches to automatically improve the quality of summaries presents important directions for future research.

Finally, our evaluation metrics focus primarily on ARs rather than more nuanced measures for evaluating various qualitative aspects of summaries. Developing multi-dimensional evaluation frameworks that assess factual accuracy, coherence, relevance, and accessibility separately could provide more granular insights into model performance.

## Conclusions

This study presents a multi-agent framework leveraging LLMs for medical summarization and evaluation. By integrating a summarization layer and an evaluation layer, our system addresses key challenges in processing unstructured medical data and ensuring accuracy, coherence, and relevance in generated summaries. The introduction of the *Referenced Condorcet Assessment*, swap position approach, and enhanced prompt designs like Prompt-EG demonstrates how methodological refinements can mitigate common issues such as position bias.

Our findings demonstrate that current LLMs can achieve evaluation consistency comparable to individual human experts, though with important limitations and context dependencies. This suggests the feasibility of LLM-assisted evaluation in specialized domains, particularly when human resources are limited or when rapid, scalable assessment is required.

Future work should focus on expanding expert panels, evaluating diverse LLM architectures,

enhancing summary quality through domain-specific techniques, and developing more nuanced evaluation metrics. By addressing these challenges, we can move toward more robust, transparent, and effective AI systems for medical information processing—systems that complement rather than replace human expertise while maintaining alignment with evolving capabilities in language model technology.

## Conflicts of Interest

None declared

## Multimedia Appendix 1

Examples of data samples for each task, LLMs' summaries and GPT-4's judgments.

## Multimedia Appendix 2

Prompt design for the summarization layer and evaluation layer.

## Multimedia Appendix 3

Detailed algorithms for the calculations.

## References

1. Awrahman BJ, Aziz Fatah C, Hamaamin MY. A Review of the role and challenges of big data in healthcare informatics and analytics. *Comput Intell Neurosci*; 2022 Sep 29; 2022:5317760. [[CrossRef](#)][[Medline](#)]
2. Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data processing and text mining technologies on electronic medical records: a review. *J Healthc Eng*; 2018 Apr 8; 2018:4302425. [[CrossRef](#)][[Medline](#)]
3. Zhuang Y, Zhang J, Li X, Liu C, Yu Y, Dong W, He K. Autonomous international classification of diseases coding using pretrained language models and advanced prompt learning techniques: evaluation of an automated analysis system using medical text. *JMIR Med Inform*; 2025 Jan 6;13: e63020. [[CrossRef](#)][[Medline](#)]
4. Achiam J, et al. GPT-4 technical report. arXiv. Preprint posted online on March 15 2023. [[Free Full text](#)]
5. Brin D., Sorin V, Vaid A. *et al.* Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep* 13;2003;16492. [[CrossRef](#)] [[Medline](#)]
6. Meyer A, Riese J, Streichert T. Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written German medical licensing examination: observational study. *JMIR Med Educ*; 2024;10: e50965 [[CrossRef](#)][[Medline](#)]
7. Flores-Cohaila J, García-Vicente A, Vizcarra-Jiménez S, De la Cruz-Galán J, Gutiérrez-Arratia J, Quiroga Torres B, Taype-Rondan A. Performance of ChatGPT on the peruvian national licensing medical examination: cross-sectional study. *JMIR Med Educ*; 2023;9: e48039 [[CrossRef](#)][[Medline](#)]
8. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. *JMIR Med Educ*; 2023;9: e48002. [[CrossRef](#)][[Medline](#)]
9. Stribling D, Xia Y, Amer M.K. et al. The model student: GPT-4 performance on graduate

- biomedical science exams. *Sci Rep* 14; 2024;5670. [\[CrossRef\]](#) [\[Medline\]](#)
10. Team G, Anil R, Borgeaud S, Alayrac JB, Yu J, Soricut R, Schalkwyk J, et al. Gemini: a family of highly capable multimodal models. *arXiv. Preprint* posted online on Dec 19 2023. [\[Free Full text\]](#)
  11. Grévisse C. LLM-based automatic short answer grading in undergraduate medical education. *BMC Medical Education*; 2024;24(1):1060. [\[CrossRef\]](#)[\[Medline\]](#)
  12. Sankarasubbu M, Pal A. OpenBioLLMs: advancing open-source large language models for healthcare and life sciences; 2024. URL: <https://huggingface.co/blog/aaditya/openbiollm>
  13. Ben Shoham O, Rappoport N. MedConceptsQA: open source medical concepts QA benchmark. *Computers in Biology and Medicine*; 2024; 182:109089. [\[CrossRef\]](#)
  14. Temsah A, Alhasan K, Altamimi I, Jamal A, Al-Eyadhy A, Malki KH, Temsah MH. DeepSeek in healthcare: revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. *Cureus*; 2025 Feb 18;17(2): e79221. [\[CrossRef\]](#) [\[Medline\]](#)
  15. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P. Survey of hallucination in natural language generation. *ACM Computing Surveys*; 2023;55(12); 1-38. [\[CrossRef\]](#)
  16. Lin CY. Rouge: A package for automatic evaluation of summaries. In: *Text summarization branches out*; 2004; 74-81. [\[CrossRef\]](#)
  17. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating text generation with bert. In *International Conference on Learning Representations*; 2020. [\[CrossRef\]](#)[\[Free Full text\]](#)
  18. Wen B, Zhang X. Enhancing reasoning to adapt large language models for domain-specific applications. *Proceedings of the NeurIPS 2024 Workshop on Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*; 2024. [\[CrossRef\]](#)[\[Free Full text\]](#)
  19. Chen Y, Wang Z, Zulkernine F. Comparative analysis of open-source language models in summarizing medical text data. *2024 IEEE International Conference on Digital Health (ICDH)*; 2024. [\[CrossRef\]](#)
  20. Touvron H, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv. Preprint* posted online on July 18 2023. [\[Free Full text\]](#)
  21. Jiang AQ, et al. Mistral 7b. *arXiv. Preprint* posted online on Oct 10 2023. [\[Free Full text\]](#)
  22. Lu W, Wei S, Peng X, Wang YF, Naseem U, Wang S. Medical question summarization with entity-driven contrastive learning. *ACM Trans Asian Low-Resour Lang Inf Process*; 2024;23(4):1–19. [\[CrossRef\]](#)
  23. Ben AA, Mrabet Y, Zhang Y, Shivade C, Langlotz C, Demner-Fushman D. Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*; 2021; 74–85; Online. Association for Computational Linguistics. [\[CrossRef\]](#)
  24. Ben AA, Demner-Fushman D. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; 2019; 2228–2234; Florence, Italy. Association for Computational Linguistics. [\[CrossRef\]](#)
  25. Savary M, Ben Abacha A, Gayen S, et al. Question-driven summarization of answers to consumer health questions. *Sci Data*; 2020; 7:322. [\[CrossRef\]](#)[\[Medline\]](#)
  26. Zeng G, Yang W, Ju Z, Yang Y, Wang S, Zhang R, Zhou M, Zeng J, Dong X, Zhang R, et al. MedDialog: large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2020; 9241–9250; Online. Association for Computational Linguistics. [\[CrossRef\]](#)
  27. Jahan I, Laskar MTR, Peng C, Huang J. Evaluation of ChatGPT on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers. *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*; 2023; 326-336. [\[CrossRef\]](#)

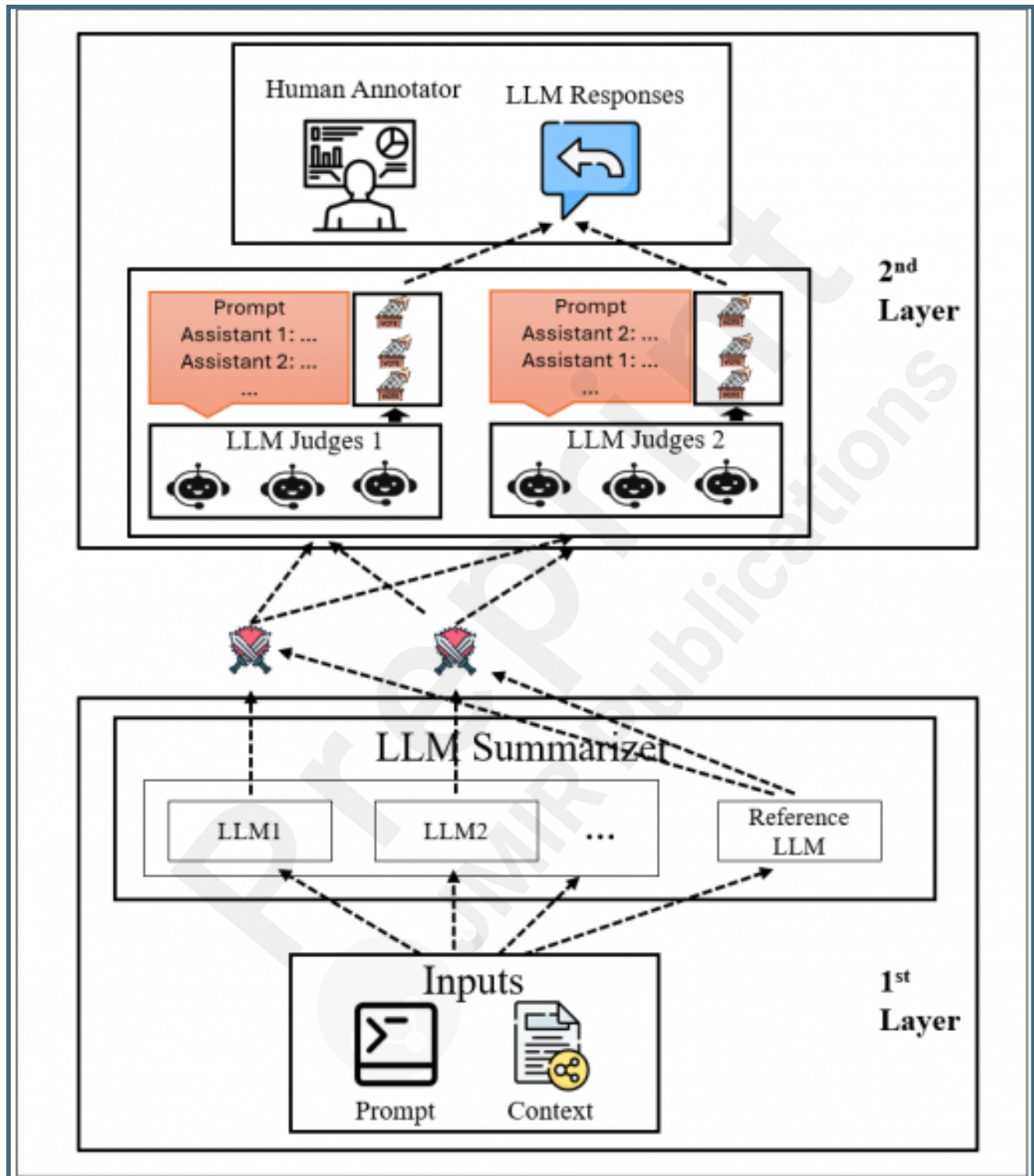
28. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst; 2022; 35:24824–24837. [\[CrossRef\]](#)



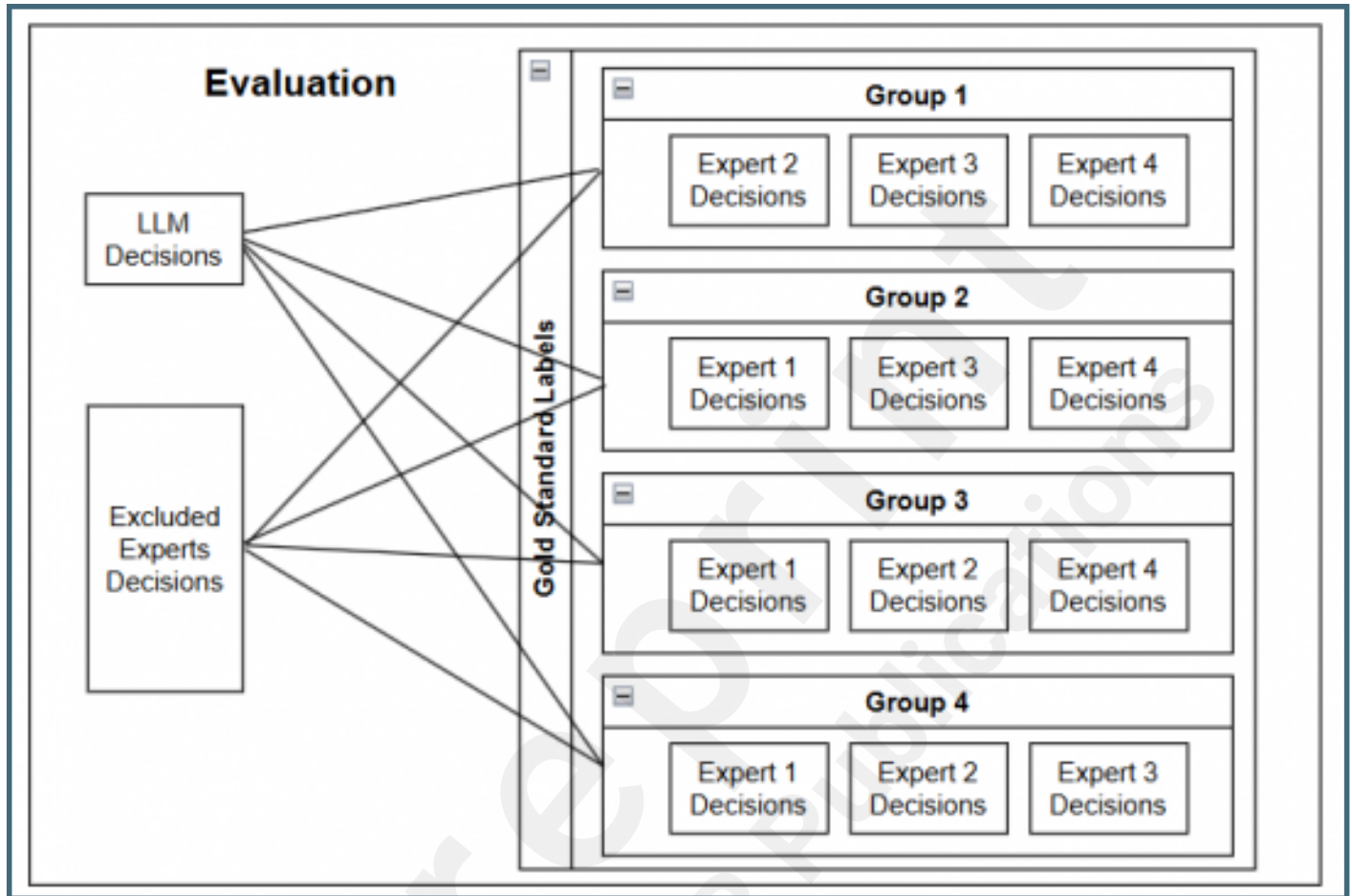
## Supplementary Files

## Figures

The proposed multi-agent LLM framework with two layers performing summarization and then evaluation.



The cross-validation framework. In each iteration, one expert is excluded from the group of four, and the decisions of the remaining three experts are used as the gold standard. GPT-4's decisions and the excluded expert's decisions are then compared against this gold standard, enabling a fair assessment of performance and individual alignment across all groups.



## **Multimedia Appendixes**

Examples of data samples for each task, LLMs' summaries and GPT-4's judgments.

URL: <http://asset.jmir.pub/assets/f7497ec0529de7db31a0fdde765035aa.pdf>

Prompt design for the summarization layer and evaluation layer.

URL: <http://asset.jmir.pub/assets/b407e622a257a84b0bb9fe04551b60b9.pdf>

Detailed algorithms for the calculations.

URL: <http://asset.jmir.pub/assets/5d6e39aeb39ce08d742f89330cbbc01c.pdf>



## Related publication(s) - for reviewers eyes onlies

Previous publication, titled " Comparative analysis of open-source language models in summarizing medical text data". It was published in 2024 IEEE International Conference on Digital Health (ICDH) as a work-in-progress paper.

URL: <http://asset.jmir.pub/assets/2ccbc1625f9e0f98305c5fa47bbc4daa.pdf>

## **TOC/Feature image for homepages**

Multiple AI agents (symbolized as abstract GPT-4 nodes or chat bubbles) analyzing and summarizing patient medical records.

