Original Paper

Large Language Model Synergy for Ensemble Learning in Medical Question Answering: Design and Evaluation Study

Han Yang[1], BSc, MSE; Mingchen Li[2], MS; Huixue Zhou[1], BMed, MBBS; Yongkang Xiao[1], MS; Qian Fang[3], MS; Shuang Zhou[2], PhD; Rui Zhang[2], PhD

[1]Institute for Health Informatics, University of Minnesota, Minneapolis, MN, United States

[2]Division of Computational Health Sciences, Department of Surgery, University of Minnesota, Minneapolis, MN, United States

[3]School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, United States

Corresponding Author:

Rui Zhang, PhD
Division of Computational Health Sciences, Department of Surgery
University of Minnesota
308 Harvard Street SE
Minneapolis, MN, 55455
United States
Phone: 1 612-626-8654
Email: ruizhang@umn.edu

Abstract

Background: Large language models (LLMs) have demonstrated remarkable capabilities in natural language processing tasks, including medical question-answering (QA). However, individual LLMs often exhibit varying performance across different medical QA datasets. We benchmarked individual zero-shot LLMs (GPT-4, Llama2-13B, Vicuna-13B, MedLlama-13B, and MedAlpaca-13B) to assess their baseline performance. Within the benchmark, GPT-4 achieves the best 71% on MedMCQA (medical multiple-choice question answering dataset), Vicuna-13B achieves 89.5% on PubMedQA (a dataset for biomedical question answering), and MedAlpaca-13B achieves the best 70% among all, showing the potential for better performance across different tasks and highlighting the need for strategies that can harness their collective strengths. Ensemble learning methods, combining multiple models to improve overall accuracy and reliability, offer a promising approach to address this challenge.

Objective: To develop and evaluate efficient ensemble learning approaches, we focus on improving performance across 3 medical QA datasets through our proposed two ensemble strategies.

Methods: Our study uses 3 medical QA datasets: PubMedQA (1000 manually labeled and 11,269 test, with yes, no, or maybe answered for each question), MedQA-USMLE (Medical Question Answering dataset based on the United States Medical Licensing Examination; 12,724 English board-style questions; 1272 test, 5 options), and MedMCQA (182,822 training/4183 test questions, 4-option multiple choice). We introduced the LLM-Synergy framework, consisting of two ensemble methods: (1) a Boosting-based Weighted Majority Vote ensemble, refining decision-making by adaptively weighting each LLM and (2) a Cluster-based Dynamic Model Selection ensemble, dynamically selecting optimal LLMs for each query based on question-context embeddings and clustering.

Results: Both ensemble methods outperformed individual LLMs across all 3 datasets. Specifically comparing the best individual LLM, the Boosting-based Majority Weighted Vote achieved accuracies of 35.84% on MedMCQA (+3.81%), 96.21% on PubMedQA (+0.64%), and 37.26% (tie) on MedQA-USMLE. The Cluster-based Dynamic Model Selection yields even higher accuracies of 38.01% (+5.98%) for MedMCQA, 96.36% (+1.09%) for PubMedQA, and 38.13% (+0.87%) for MedQA-USMLE.

Conclusions: The LLM-Synergy framework, using 2 ensemble methods, represents a significant advancement in leveraging LLMs for medical QA tasks. Through effectively combining the strengths of diverse LLMs, this framework provides a flexible and efficient strategy adaptable to current and future challenges in biomedical informatics.

J Med Internet Res 2025;27:e70080; doi: 10.2196/70080

Keywords: large language models; ensemble learning; medical question answering; healthcare AI; GPT-4

Introduction

Question answering (QA) tasks in the medical domain involve a complex process of accurately interpreting and responding to health care–related queries [1]. QA tasks typically encompass two formats: open-ended and structured. In open-ended QA, respondents provide a complete sentence incorporating essential information in response to a question. In structured QA, the question is presented with several options, and the respondent selects the correct option or options by its corresponding identifier. Medical QA systems are designed to provide reliable and precise answers to questions ranging from disease symptoms and treatment options to medical research findings [2]. These systems leverage advanced technologies like natural language processing (NLP) and machine learning approaches to understand and process medical terminology and concepts, making them invaluable tools for health care professionals and patients seeking medical information. The effectiveness of these systems is crucial, as they directly impact health care decision-making and patient care [3-7].

Various models have been used for medical QA. Previously, transformer models like Bidirectional Encoder Representations from Transformers (BERT) [8] played a pivotal role in QA. For instance, He et al [9] infused disease knowledge into a basket of BERT-based models for health QA, demonstrating the viability of disease knowledge infusion in NLP models. Alzubi et al [10] developed another BERT-based model named CoBERT specifically designed for COVID-19–related QA. These BERT-based models have demonstrated solid performance on COVID-19 QA tasks but achieve substantially lower accuracy on structured, domain-specific medical QA benchmarks (eg, $\lesssim$75% vs >90% for GPT-4 on PubMedQA [a dataset for biomedical question answering]), since the expansion of medical corpora and textual resources has necessitated leveraging these large datasets more effectively. This need has been met by the emergence of large language models (LLMs) as a transformative approach to medical QA tasks. Pretrained on extensive and diverse datasets, LLMs like GPT-4 possess a deep understanding of language nuances and medical terminology, enabling them to generate highly accurate and relevant responses to medical queries [11,12]. They represent a significant milestone while dealing with NLP tasks [13], such as text generation [14,15], QA [4,7,14,16-20], Named Entity Recognition [21-25], and so on. Moreover, LLMs include large-sized models like GPT-4 [26] and Llama series (Llama-2, Llama-3, etc) [27-29], as well as some relatively small yet efficient LLMs like Vicuna [30] and Stanford Alpaca. These LLMs, characterized by their vast size, have demonstrated remarkable capabilities in understanding and generating human language across a diverse array of domains [25] as single-model approaches.

Despite these advancements, achieving satisfying performance in medical QA using off-the-shelf LLMs remains underexplored. One primary issue is the phenomenon of "hallucination," where an LLM may generate erroneous answers due to incorrect medical knowledge or reasoning [31-33]. For example, Ji et al [31] reported that Vicuna and Alpaca-L exhibited fact inconsistency rates of 10.4% and 17.6%, respectively, in their study, underscoring the risks in high-stakes medical applications. Besides, though many LLMs are available, determining an optimal LLM with superior performance in arbitrary medical question types remains nontrivial, as the performance may vary significantly with different network architectures, training approaches, training corpora, and test question types [33]. For example, MedAlpaca [34] was fine-tuned on corpora including Wiki-doc–generated QA pairs, while MedLlama was trained on more complex clinical notes (eg, MIMIC datasets). Consequently, MedAlpaca may excel at answering short, fact-based questions such as "What is the normal range for blood pressure?," while MedLlama may be better suited for medical questions involving nuanced patient information, such as "What is the most possible diagnosis for this patient?"

Given that different LLMs have unique characteristics and strengths, ensemble learning, which involves combining diverse models for superior performance, presents a promising approach to alleviate the above issues in medical QA [35]. First, ensemble techniques such as voting [36], boosting [37], and bagging can reduce the impact of hallucinations by filtering out erroneous responses [38]. Specifically, these methods can aggregate outputs from multiple models, effectively neutralizing inaccuracies in individual predictions. Second, ensemble techniques can harness the diverse strengths of the base models by adaptively assigning more weights to the answers from the best-suited base models, thus achieving superior performance on various question types [39,40].

However, it is challenging to effectively ensemble diverse LLMs for medical QA tasks, especially within the medical domain [41]. This integration must consider factors such as the compatibility of different models, the method of aggregating their outputs, and maintaining the interpretability of the responses [42-46]. These considerations are crucial for ensuring that the ensemble not only performs well but also aligns with the stringent requirements of medical applications. There have been only a few studies diving into Ensembling LLMs to achieve better prediction, like LLM-Blender [47], which implements PairRanker and Genfuser as an ensemble framework to generate consistently better responses for a given input. Similarly, the majority voting method proposed by Pitis et al [48] also demonstrates potential. However, these studies primarily focus on open-ended tasks and do not delve into the specifics of medical QA, nor do they include domain-specific LLMs like PMC-LlaMA2 [18] or MedAlpaca [34]. Despite substantial progress, single-model LLMs exhibit inconsistent accuracy across varied medical QA datasets due to their inherent limitations, such as domain-specific knowledge gaps [49], model biases [50], and variability in contextual understanding [51]. While general ensemble approaches exist, current methods have rarely

targeted medical-specific QA tasks explicitly and have not sufficiently addressed the dynamic nature and context-specific demands of biomedical queries. Thus, there remains a critical research gap in developing robust, specialized ensemble strategies specifically tailored to enhance medical QA tasks.

To address these limitations, we introduce LLM-Synergy, a novel ensembling framework tailored for medical QA, with two well-designed meta-learning ensembling methods, providing two innovative approaches combining the strengths of various LLMs, named Boosting-based Weighted Majority Vote Ensemble [48,52,53] and Cluster-based Dynamic Model selection [53,54]. To validate the efficacy of LLM-Synergy and its ensemble methods, we conducted a case study using 3 medical QA datasets: MedMCQA (medical multiple-choice question answering dataset) [55], PubMedQA [56], and MedQA-USMLE (Medical Question Answering dataset based on the United States Medical Licensing Examination) [57].

Our contributions to this study include the following:

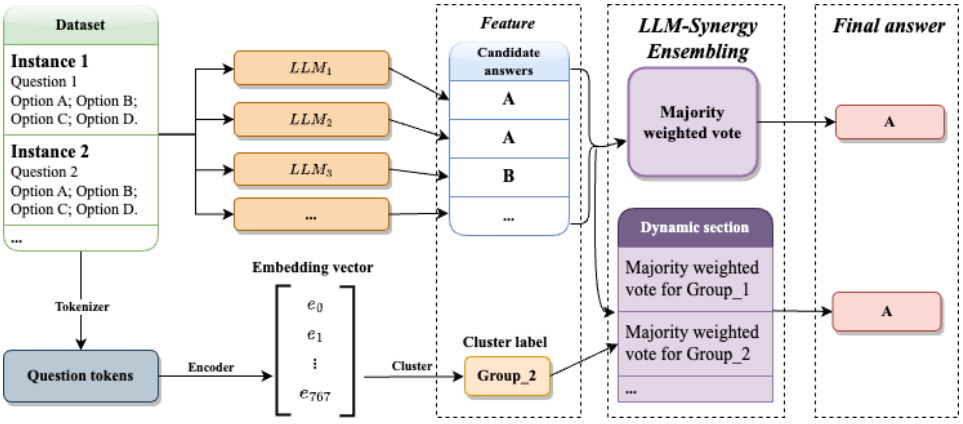
1. We developed innovative LLM ensemble methods, specifically the Boosting-based Weighted Majority Vote Ensemble and Cluster-based Dynamic Model Selection, which offer new approaches, with zero-shot cases, in the medical QA field.
2. We implemented the ensemble methods for LLM methods and improved the performance by 5.98%, 1.09%, and 0.87% compared to the best-performing LLM on 3 medical QA datasets, demonstrating the effectiveness of our ensemble methods. In each case, a tailored ensemble framework was created and adapted to the format of the QA dataset (single-choice or multiple-choice formats).
3. We conducted an error analysis to provide insights and directions for potential future enhancements in the field of medical QA, laying the groundwork for further improvements in this domain.

Methods

Overview

The first step of our methods involves benchmarking leading individual LLMs, including GPT-3.5-turbo, GPT-4, Llama2-13B, Vicuna-13B, MedAlpaca-13B, MedLlama-13B, PMC-Llama-13B, and a random guessing baseline, to evaluate their initial performance on medical QA tasks before applying the LLM-Synergy framework. Within the benchmark, we conduct a sampled test, randomly drawing 200 QA pairs from the 3 medical QA datasets to assess the current capabilities of these LLMs in a medical context as a starting point. This benchmarking serves as a foundational analysis to understand the individual strengths and limitations of each LLM in handling medical QA tasks. Illustrated by Figure 1, which provides an overview of the whole pipeline of LLM-Synergy, following the benchmark assessment, the next phase focuses on the training process of our two proposed ensembling methods within LLM-Synergy: the Boosting-based Weighted Majority Vote Ensemble and the Cluster-based Dynamic Model Selection. The two approaches are designed to combine the unique capabilities of the selected LLMs, aiming to enhance the overall performance of medical QA systems. Moreover, the second method could be regarded as an extensive version of the first one. By implementing these methods, we seek to address the shortcomings of relying on single models and reduce the need for extensive individual model training, thereby creating a more robust and efficient solution for medical QA.

Figure 1. Overview of our large language model synergy framework. LLM: large language model.



Dataset

We used 3 medical QA datasets for our model training and test: MedMCQA [55], PubMedQA [56], and MedQA-USMLE [57].

MedMCQA, released in March 2022 by Pal et al [55], is a comprehensive multiple-choice question dataset derived from mock and past examinations of All India Institute for Medical Sciences and National Eligibility cum Entrance Test for Postgraduate (Pal et al, 2022 [58]), 2 prominent Indian medical entrance exams. It encompasses a training set with 182,822 questions and a test set comprising 4183 questions, covering more than 2400 topics. Each question in this dataset presents 4 answer choices, labeled from A to D.

PubMedQA, introduced in September 2019 by Jin et al [56], is a QA dataset curated from PubMed abstracts. It includes 1000 questions reviewed by experts and 272,500 algorithmically generated QA pairs. The dataset's primary task is to classify research questions into yes, no, or maybe answers, akin to multiple-choice questions. It is divided into three segments: PQA-L with 1k manually labeled pairs, PQA-U with 61.2k unlabeled pairs, and PQA-A featuring 211.3k artificially generated pairs. Here, we only implement a QA process without reasoning, which does not require a corresponding explanation of how the final answer is generated, which may lead to a relatively high accuracy score.

MedQA-USMLE, launched in September 2020 by Jin et al [57], is an innovative dataset of multiple-choice questions tailored to the United States Medical Licensing Exams. This dataset encompasses questions in three languages: English, Simplified Chinese, and Traditional Chinese, with a total of 12,724, 34,251, and 14,123 questions in each respective language. Each question offers 5 choices, ranging from option A to E, sourced from professional medical board examinations. Here, we only experimented with the English QA parts. The detailed LLM prompt can be found in the Multimedia Appendices 1-3.

LLM Benchmark on the 3 Medical QA Datasets

Ahead of implementing our ensembling framework, as a benchmark study, we evaluate the performance of various LLMs on 200 questions from each of the three QA datasets, respectively: PubMedQA, MedQA-USMLE, and MedMCQA, adhering to their specific answer formats. We evaluate our model's performance against several robust baselines relying on the language model (LLM). Table 1 summarizes the 6 LLMs and one random guess predictor, along with the model characteristics including how many parameters within each LLM and a comprehensive description.

Table 2 reports the performance of the selected predictor in answering the 3 medical QA datasets. The benchmark graph illustrates that each predictor exhibits performance levels that exceed random guessing across various medical QA tasks, signaling the inherent capability of the LLMs to understand and process medical queries. Notably, different LLMs demonstrate particular strengths depending on the dataset; for instance, GPT-4 shows a marked proficiency in the MedMCQA tasks, while PMC-Llama -13B stands out in the PubMedQA context. These variations in model performance across tasks provide a solid foundation for the potential enhancement of accuracy through our subsequent ensemble work, suggesting that strategic combinations of these models could capitalize on their respective strengths and mitigate their individual weaknesses. For hyperparameters, we set the temperature to be 0.1 for the least randomness, and the rest, like top_p, top_k, and so on, are all set to default. The detailed LLM prompt can be found in the Multimedia Appendices 1-3.

Table 1. The summarization of 7 large language models running Question Answering models.

QA predictors	Model parameters	Description
GPT-4	1.76 trillion	GPT-4 is a substantial multimodal model designed to respond to questions by providing instructions fed to the GPT-4.
GPT-3.5-turbo	20 billion	Same design as GPT-4; however, GPT-3.5-turbo has fewer parameters than GPT-4.
Llama2-13B	20 billion	Llama 2-13B is part of the Llama 2 series, representing a pretrained generative model. Tuned versions of Llama 2 use supervised fine-tuning and reinforcement learning with human feedback to help generate the answers to given questions.
Vicuna-13B	13 billion	Vicuna-13, similar in size to Llama2-13b, Vicuna is noted for its robustness and adaptability across different types of language processing tasks.
MedAlpaca-13B	13 billion	MedAlpaca-13B is a substantial language model finely tuned for tasks in the medical domain. It stems from Llama (Large Language Model Meta AI) and boasts a significant parameter count of 13billion.
MedLlama-13B	13 billion	MedLlama-13B is initialized from LLaMA-13B and undergoes additional pretraining using a constructed medical corpus derived from 4.8M Pubmed Central papers and Medical Books.
PMC-Llama-13B	13 billion	PMC-Llama-13B is the further tuned version of MedLlama-13B, with the pretrain and instruction-tuning methods.
Random Guess	1	A random guess predictor simply generates a random answer by the equal probability of each option, serving as a reference to compare the large language model–cased predictor.

Table 2. Benchmarking the accuracy of each large language model performs on three medical Question Answering datasets.

LLMs	MedMCQA	PubMedQA	MedQA-USMLE
GPT-4[a]	*.71*[a]	0.59	0.665
GPT-3.5-Turbo	0.49	0.515	0.42

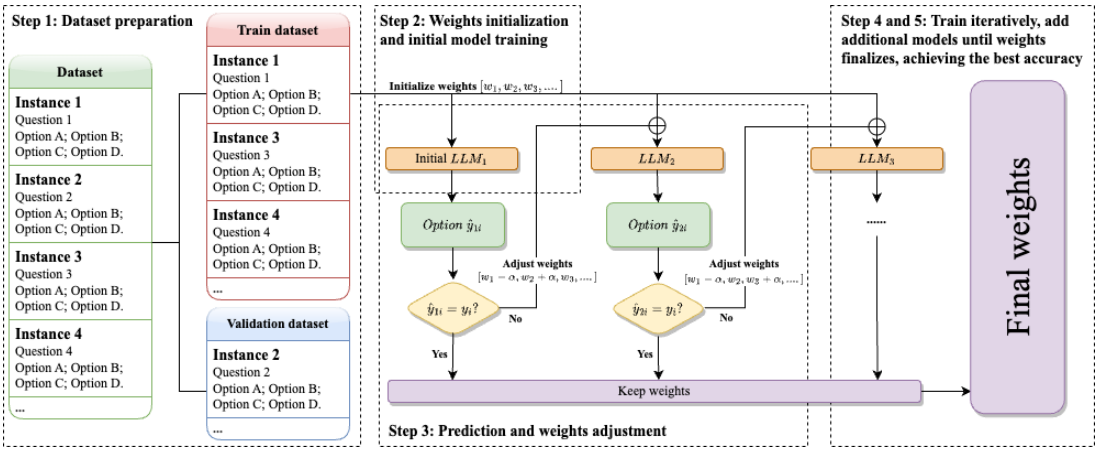
LLMs	MedMCQA	PubMedQA	MedQA-USMLE
Llama2-13B	0.34	0.84	0.24
Vicuna-13B	0.315	*0.*895	0.245
MedAlpaca-13B	0.33	0.695	*0.7*
PMC-Llama-13B	0.335	0.67	0.58
Random Guess	0.23	0.5	0.165

[a]The italicized values indicate the highest performance among each dataset respectively.

Part 1: Boosting-Based Weighted Majority Vote

Figure 2 shows the training process of Boosting-based Weighted Majority Vote Ensemble.

Figure 2. Training process of boosting-based weighted majority vote ensemble. LLM: large language model.



Step 1: Dataset Preparation

For a given medical QA dataset, we stratified split the dataset into a training set and a validation set, with a proportion of 80% and 20%, with a fixed random seed (42 in default in our case), to preserve the same answer-option distributions in two sets. Each QA-pair instance in the dataset consists of a question and single-choice options.

Step 2: Weight Initialization and Initial Model Training

We assigned initial weights to all LLMs with weights (w_1, w_2, I) and initialized the starting status of ensembled LLMs with these weights. There may be different strategies for initialization, and we chose equally weighted initialization, which is the most common way of initialization [52]. For example, in our case, the equal weights of each LLM are assigned as:

$$w_j = \frac{1}{M}, j = 1, 2, ..., M$$

where, M denotes the number of LLMs applied.

And for each training instance i, we define an initial indicator function for prediction correctness:

$$\in_j (i) = \begin{cases} 0, & if\ LLM_j\ correctly\ predicts\ instance\ x_i \\ 1, & if\ LLM_j\ incorrectly\ predicts\ instance\ x_i \end{cases}$$

Then the initial cumulative error, denoted as E_j for M_j, can be computed as:

$$E_j = \frac{\sum_{i=1}^{N} \in_j (i)}{N}$$

where N is the total number of training instances.

Step 3: Prediction and Weight Adjustment

Next, we chose a series of LLMs, denoted as LLM_jeach time, and focused on its wrong prediction: use LLM_j to predict answers for the training set and adjust the weights based on the prediction of instance: If the prediction, $\widehat{y_{1i}}$ for instance i, is incorrect, minus the weight of its corresponding LLM, LLM_j, by an adjustment parameter α_j, i.e. $w_i \leftarrow w_i + \alpha_j$, where α_j is a factor that increases the weight, indicating that the instance needs more attention in the next round of training. The weight adjustment factor α_j for each LLM is calculated as follows:

$$\alpha_j = \frac{1}{2} In\left(\frac{1 - E_j}{E_j}\right)$$

with this adjusting strategy, we shall be able to ensure the models performing better (lower E_j receive larger positive adjustments, whereas poorly performing models (high E_j) obtain smaller or even negative adjustments.

Step 4: Iterative Training With Additional Models

For each subsequent $j-1$ LLM ($LLM_2, LLM_3 \ldots$), repeat the prediction process. If a model correctly predicts the answer, we maintain the current weights. If the prediction $\hat{y}$ is incorrect, the weights were adjusted again. During the process, the implicit cost function guiding these weight updates can be interpreted as:

$$L = \sum_{j=1}^{M} \sum_{i=1}^{N} exp\left(-\hat{y}_{ji} f(x_i)\right)$$

where $f(x_i)$ is the weighted vote prediction of predictions:

$$f(x_i) = \sum_{j=1}^{M} w_j h_j(x_i)$$

where $h_j(x_i)$ is the binary prediction (options chosen for our case) of LLM_j.

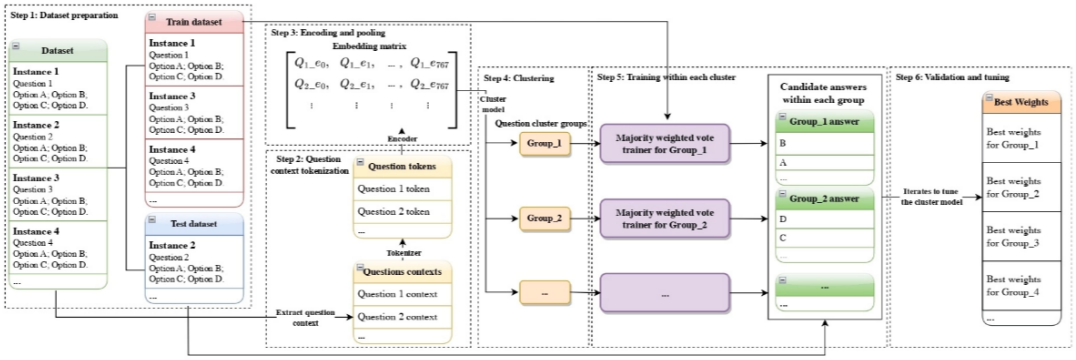
Step 5: Finalize Weights of the Model Ensemble

The iterative weight-update process continues until convergence, defined as the prediction accuracy stabilizing within a predefined small threshold (ie, accuracy change below a certain threshold) over two consecutive iterations. The training and weight adjustments across all LLMs are complete, and the final ensemble model is formed, combining the individual LLMs with the final set of adjusted weights.

Part 2: Cluster-Based Dynamic Model Selection

Figure 3 shows the training process of our second approach: Cluster-Based Dynamic Model Selection, which serves as an extensive approach to the first one.

Figure 3. Training process of Cluster-based dynamic model selection ensemble.



Step 1: Dataset Preparation

Same as Part 1, for a given medical QA dataset, split it into a training set and a validation set with a ratio of 80:20. Each QA-pair instance in the dataset should consist of a question paired with single-choice options.

Step 2: Question Context Tokenization

For each question, we first define the "question context," which refers explicitly to the textual description and details provided within each question instance (excluding the candidate answer options). We then tokenize each question context using a common tokenizer model (in our case, Clinical-BERT [59], which was pretrained on large-scale clinical notes and outperformed general biomedical models, was chosen since it matches our Medical QA situation very well), transforming the textual information into sequences of numerical tokens.

Step 3: Encoding and Pooling

Next, we encode these numerical token sequences using Clinical-BERT (same as step 2) to generate high-dimensional semantic representations. Specifically, we use mean-pooling over Clinical-BERT's final hidden states, obtaining a 768-dimension embedding vector for each question. As a result, all question embeddings collectively form an embedding matrix of size $N \times 768$, where N is the total number of questions.

Step 4: Clustering

Apply a clustering model, KMeans (K-Means Clustering Algorithm) [60] in our experiment specifically, directly to the embedding matrix. This step assigns each question a cluster label, effectively reducing the dimensionality from 768 of an individual question vector to a single cluster group label being 1,2, 3, … K. Thereafter, each question's embedding vector is thus assigned to one of K clusters (labels: 1, 2, 3, …, K), effectively grouping questions that have similar semantic

characteristics. In our experiment, the optimal K was chosen based on the highest mean silhouette coefficient [61] and confirmation by the elbow method [62] on within-cluster sum of squares, to ensure both cluster compactness and stability.

Step 5: Training Within Clusters

For each cluster group identified previously, we implement the Boosting-based Majority Weighted Vote ensemble method (outlined earlier in Part 1). Then for each cluster, the ensemble weights for the LLMs are adjusted specifically to maximize predictive accuracy on questions belonging to that cluster. In other words, each cluster obtains its own optimized combination of LLM weights, tailored precisely to the context characteristics of the cluster.

Step 6: Validation and Hyperparameter Tuning

Finally, we evaluate the ensemble's predictive performance on the validation set by repeating Steps 2 through 5, and iteratively adjust and tune the hyperparameters of the clustering algorithm (K in KMeans in our case), and any other hyperparameters (such as KMeans initialization settings) to achieve the best overall accuracy on the validation set. This ensures the clustering algorithm effectively captures meaningful patterns in question contexts, thereby optimizing our dynamic selection ensemble's final accuracy.

Evaluation

In the evaluation phase of our study, same as others [18,47,55,57], accuracy was used as the primary metric for assessment, given its congruence with the micro F_1-score in our specific experimental context of single-choice or multiple-choice QA tasks. Given the nature of multiple-choice QA datasets, metrics such as Recall and Precision are deemed inappropriate as they are sensitive to changes in the option numbering. For instance, adjusting the order of option labels may result in altering these metric values when the number of options exceeds 2. Based on these considerations, accuracy emerges as a more suitable evaluation metric in our case, providing a robust measure of overall correctness without being affected by variations in option numbering, which ensures a consistent and meaningful assessment of model performance in the specific context of our study.

The test set is distinguished from the training set used for model development and the validation set for hyperparameter optimization that we used within the ensemble training process. This distinct separation ensures an unbiased evaluation of the model's true predictive capabilities. The test sets are derived from the subsets of the MedMCQA, PubMedQA, and MedQA-USMLE datasets, respectively. The MedMCQA test dataset consists of 4183 QA pairs. The PubMedQA test dataset is even more extensive with 11,269 QA pairs, whereas the MedQA-USMLE dataset contains 1272 QA pairs.

Ethical Considerations

Institutional review board exemption was granted because the study made use of publicly available, deidentified datasets and did not involve human subjects research as defined by 45 CFR 46.

Results

In this study, the performance of various LLMs was evaluated against ensemble methods across 3 distinct medical QA datasets, MedMCQA, PubMedQA, and MedQA-USMLE. The performance metric, assumed to be an accuracy score, highlights the differential capabilities of each LLM and our two established predictors of ensemble methods. The detailed result can be seen in Table 3.

For individual models, Llama2-13B demonstrated substantial proficiency in the PubMedQA dataset with an accuracy of 93.09%, indicating a strong alignment with the dataset's characteristics. Conversely, its performance on the MedQA-USMLE dataset was considerably lower at 24.61%, suggesting a potential misalignment with this dataset's attributes or a limitation in handling its complexity. MedLlama-13B showed a similar trend, albeit with marginally lower accuracy figures across the board, peaking at 86.11% for the PubMedQA dataset.

MedAlpaca-13B yielded a divergent performance profile, exhibiting a relatively lower accuracy of 27.97% on PubMedQA, while achieving the highest accuracy among individual LLMs on the MedQA-USMLE dataset at 37.26%. This suggests that MedAlpaca-13B may possess particular strengths in processing the content typified within the MedQA-USMLE exam questions. Vicuna-13B, on the other hand, had the lowest accuracy scores for all datasets, which could indicate a general difficulty with the medical QA task as presented in these datasets.

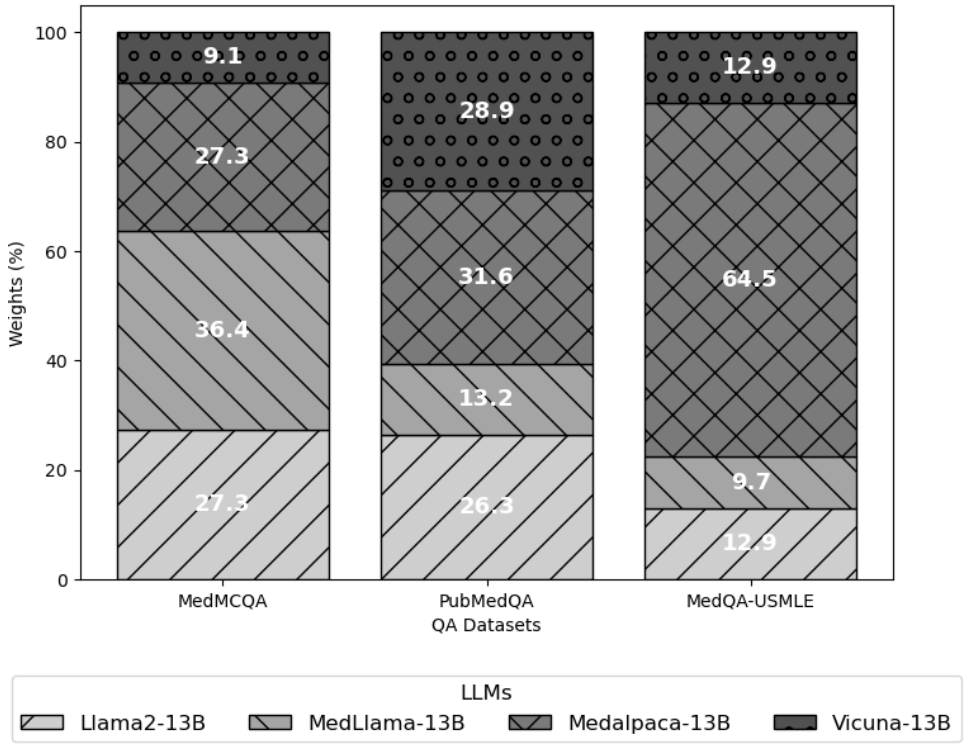
The ensemble approaches, notably the Boosting-based Weighted Majority Vote ensemble and the Cluster-based Dynamic Selection ensemble, were developed to leverage the collective strengths of the individual LLMs. The Boosting-based Weighted Majority Vote Ensemble surpassed the individual model performances on the MedMCQA and MedQA-USMLE datasets and achieved a notable accuracy of 96.21% on PubMedQA, and the detailed weights of component LLMs are disclosed in Figure 4. This enhancement suggests that a static weighted combination of model outputs can capitalize on the diverse expertise of each LLM to improve overall performance. Whereas the Cluster-based Dynamic Selection Ensemble, which introduces a context-aware model selection strategy, further improved upon the Boosting-based Weighted Majority Vote ensemble's performance, achieving the highest accuracy across all datasets: 38.01% on MedMCQA, 96.36% on PubMedQA, and 38.13% on MedQA-USMLE. The specific range of improvement varies from the variation of each LLM. The final cluster numbers are 9 in MedMCQA, 12 in PubMedQA, and 5 in MedQA-USMLE.

Table 3. Test set performance of individual large language model and our ensemble approach.

LLMs	MedMCQA	PubMedQA	MedQA-USMLE
Llama2-13B	32.03	93.09	24.61
MedLlama-13B	31.03	86.11	23.58
MedAlpaca-13B	27.97	95.27	37.26
Vicuna-13B	26.13	93.15	24.14
Boosting-based Weighted Majority Vote	35.84	96.21	37.26
Cluster-based Dynamic Model Selection	*38.01*[a]	*96.36*	*38.13*

[a]The italicized values indicate the highest performance among each dataset respectively.

Figure 4. Weights of the component large language models under each Question Answering dataset.



Discussion

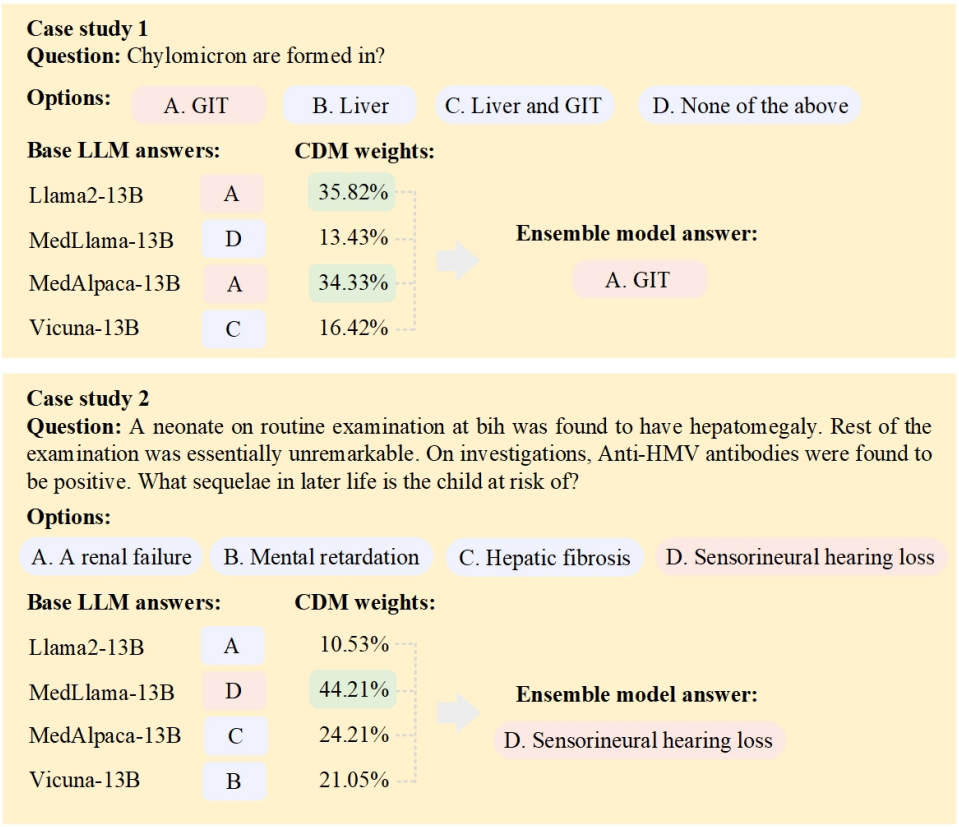
Principal Findings

In this study, we demonstrated that both methods of our proposed LLM-Synergy ensemble framework improved prediction accuracy across three medical QA datasets—MedMCQA, PubMedQA, and MedQA-USMLE—compared to individual LLMs. On MedMCQA, where base models' accuracies ranged narrowly from 26.13% to 32.03%, —Cluster-based Dynamic Model Selection yielded the largest gain (+5.98%) by clustering semantically similar questions and tailoring weights to each cluster's specific context. On PubMedQA, with uniformly high single-model performance (min 86.11%), Boosting-based Weighted Majority Vote already delivered a modest improvement (+0.94%) through global weight adjustments, and Cluster-based Dynamic Model Selection added a further slight boost (+1.15%) by fine-tuning weights per question group. Finally, MedQA-USMLE presented a case where one model (MedAlpaca-13B) dominated; here, Cluster-based Dynamic Model Selection's context-sensitive reweighting prevented over-reliance on that single model and achieved a +0.87%gain over static Boosting-based Weighted Majority Vote.

The effectiveness of this design can be attributed to two key factors: First, the Boosting-based Weighted Majority Vote addresses systematic biases by amplifying the influence of models that consistently perform well across all questions. Second, the Cluster-based Dynamic Model Selection further adapts to question heterogeneity by dynamically identifying the optimal combination of models for each input. Together, these tailored strategies explain the superior performance of LLM-Synergy across varied medical QA formats. Moreover, our analysis reveals that the Cluster-based Dynamic model outperforms the Boosting-based Weighted Majority Vote on the three QA datasets. On MedMCQA, it achieves 38.01% versus 35.84% (+2.17%), on PubMedQA 96.36% versus 96.21% (+0.15%), and on MedQA-USMLE 38.13% versus 37.26% (+0.87%). This improvement likely arises from the Cluster-based Dynamic Model Selection's ability to assign model weights tailored to semantically coherent question clusters, enabling each model to excel in areas where it performs best, rather than relying on a uniform global weighting approach.

To further illustrate how the Cluster-based Dynamic Model Selection adapts model contributions to specific contexts, we visualized weight distributions for several clusters from the MedMCQA dataset. These visualizations demonstrate that assigning greater weights to the most competent LLMs within each question cluster enhances the likelihood of selecting the correct answer, thereby validating the model's superior performance (see Figure 5). For instance, in the fact-based question (ie, the first case in Figure 5), we observed that MedAlpaca and Llama were assigned larger weight values than MedLlama (35.82% and 34.33% versus 13.43%), while in the question involving complex patient information and clinical decision-making (ie, the second case in Figure 5), MedLlama instead obtained larger weights (44.21%) than the others. This discrepancy arises from the distinct training corpora of the base models, which renders these LLMs to have different strengths. Specifically, MedAlpaca and Llama were fine-tuned on corpora including Wiki-doc–generated QA pairs and may excel at answering short, fact-based questions, while MedLlama was trained on more complex clinical notes (eg, MIMIC datasets) and may be better suited for medical questions with nuanced patient information and clinical decision-making.

Figure 5. Case studies of the Cluster-based dynamic model selection on the MedMCQA dataset. The red color denotes the correct answer, and the green color highlights the base LLMs receiving large weights. The results demonstrated that the proposed ensemble model assigned larger weights to the most suitable LLMs within each question cluster, thus enhancing the likelihood of selecting the correct answer. GIT: gastrointestinal tract.



Limitations

LLM-Synergy's effectiveness depends critically on the quality and diversity of its constituent LLMs: if these models share similar biases or blind spots, the ensemble may offer little improvement over any single model. The weighted-majority vote approach, while efficient, uses fixed weights and cannot adapt to question-specific nuances, leading to potential underperformance when individual model strengths vary by context. The cluster-based dynamic selection method addresses this by learning cluster-specific weights, but its success relies on the clustering algorithm's ability to capture relevant question features and on having a sufficiently large, representative validation set for tuning.

Despite avoiding expensive fine-tuning, both methods incur nontrivial computational overhead when integrating multiple LLM inferences and Clinical-BERT embeddings, which may limit real-time deployment in resource-constrained environments. Additionally, both approaches assume that model performance on the validation set reliably predicts test-set behavior; substantial dataset shifts could thus degrade ensemble effectiveness. Finally, the scarcity of direct comparisons with alternative ensemble strategies or traditional fine-tuning on these specific medical QA datasets highlights the need for future work to benchmark and further optimize ensemble configurations.

Computational Cost

Both our ensemble methods operate entirely at inference time—no fine-tuning or retraining of base LLMs is required—resulting in substantially lower overall compute than full model training. We approximate the computational

complexity for the training phase as $O(N \times M \times C_{LLM})$ for method 1, Boosting-based Weighted Majority Vote Ensemble inference, where M is the number of individual LLM, N is the number of training instances (questions to answer), C_{LLM} is the inference cost for each individual LLM. Whereas for method 2, Cluster-based Dynamic Model Selection, inference complexity approximately scales linearly with the number of ensemble models used $O(N \times (M \times C_{LLM} + C_{BERT} + K \times I \times d))$, where K is the chosen number of clusters, I is the number of iterations (typically small, eg, ~20-50), and d is the embedding dimension (768 in our case).

In practice, using our experiments as an example, inference for the ensemble method takes roughly 6-12 seconds per question on a single A100 GPU, compared to 1-2 seconds for an individual model alone. Although multi-second latency may exceed real-time requirements, it remains modest relative to the cost of full model fine-tuning and is easily amortized in batch workflows. Moreover, latency can be reduced via parallelization across GPUs, model distillation into a compact student network, adaptive inference (invoking the full ensemble only when needed), and caching of frequent queries.

Conclusions

The LLM-Synergy framework, with its Boosting-based Weighted Majority Vote and Cluster-based Dynamic Model Selection methods, represents a significant advancement in leveraging LLMs for medical QA tasks and provides an innovative way of efficiently using the development with LLM technologies, customizing for both existing and potentially future challenge tasks in biomedical and health informatics research. Its ability to amalgamate the strengths of multiple models has demonstrated superior accuracy and robustness over individual LLMs. While the framework showcases scalability, flexibility, and adaptability across domains, it also presents opportunities for future enhancements, including increased model diversity, dynamic weighting, and broader domain applications. As such, LLM-Synergy not only addresses current challenges in natural language processing but also sets the stage for continued innovation in artificial intelligence–driven problem-solving.

Acknowledgments

We would like to acknowledge the staff at BPIC of the University of Minnesota and the programmers from RZ's group for their technical support. We also confirm that no generative AI tools (eg, ChatGPT, Bing, and Bard) were used in the ideation, drafting, or editing of this manuscript. All textual content, analyses, and interpretations were produced solely by the listed authors, in full compliance with JMIR Publications' policy on AI usage. This work was supported by the National Institutes of Health National Center for Complementary and Integrative Health (grant number R01AT009457), National Institute on Aging (grant number R01AG078154), and National Cancer Institute (grant number R01CA287413).

Disclaimer

The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

Data Availability

All the dataset used is publicly available. As for the analysis code, the LLM synergy framework used in this study is made publicly available at GitHub.

Authors' Contributions

HY served as the principal author, designing the entire experiment, writing experimental code, and the full manuscript. ML and YX collaborated on the data collection and analysis and also reviewed the paper. HZ assisted in conceiving the study design and reviewed the manuscript. RZ was responsible for auditing the feasibility of the research topic and refining the paper. SZ collaborated on paper drafting, part of the data visualization and creating additional figures. QF created the graphical and workflow diagrams within the manuscript. All authors contributed significantly to the production and proofing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Prompt example for MedMCQA dataset, same for benchmarking and modeling.
[PNG File (Portable Network Graphics File), 138 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Prompt example for PubMedQA dataset, same for benchmarking and modeling.
[PNG File (Portable Network Graphics File), 294 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Prompt example for MedQA-USMLE dataset, English language, same for benchmarking and modeling.
[PNG File (Portable Network Graphics File), 192 KB-Multimedia Appendix 3]

References

1. Athenikos SJ, Han H. Biomedical question answering: a survey. Comput Methods Programs Biomed. Jul 2010;99(1):1-24. [doi: 10.1016/j.cmpb.2009.10.003] [Medline: 19913938]
2. Mitchell JR, Szepietowski P, Howard R, et al. A question-and-answer system to extract data from free-text oncological pathology reports (CancerBERT Network): development study. J Med Internet Res. Mar 23, 2022;24(3):e27210. [doi: 10.2196/27210] [Medline: 35319481]
3. Mutabazi E, Ni J, Tang G, Cao W. A review on medical textual question answering systems based on deep learning approaches. Appl Sci (Basel). Jun 11, 2021;11(12):5456. [doi: 10.3390/app11125456]
4. Jin Q, Yuan Z, Xiong G, et al. Biomedical question answering: a survey of approaches and challenges. ACM Comput Surv. Feb 28, 2023;55(2):1-36. [doi: 10.1145/3490238]
5. Mollá D, Vicedo JL. Question answering in restricted domains: an overview. Comput Linguist Assoc Comput Linguist. Mar 2007;33(1):41-61. [doi: 10.1162/coli.2007.33.1.41]
6. Clark P, Cowhey I, Etzioni O, et al. Think you have solved question answering? try ARC, the AI2 reasoning challenge. arXiv. Preprint posted online on 2018. [doi: 10.48550/ARXIV.1803.05457]
7. Bardhan J, Roberts K, Wang DZ. Question Answering for Electronic Health Records: Scoping Review of Datasets and Models. J Med Internet Res. Oct 30, 2024;26:e53636. [doi: 10.2196/53636] [Medline: 39475821]
8. Devlin J, Chang MW, Lee K, Toutanova K. Pre-training of deep bidirectional transformers for language understanding. Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Jun 2-7, 2019:Association for Computational Linguistics. 4171-4186; Minneapolis, Minnesota. [doi: 10.18653/v1/N19-1423]
9. He Y, Zhu Z, Zhang Y, Chen Q, Caverlee J. Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition. Presented at: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); Nov 16-20, 2020:4604-4614; Online. [doi: 10.18653/v1/2020.emnlp-main.372]
10. Alzubi JA, Jain R, Singh A, Parwekar P, Gupta M. COBERT: COVID-19 Question Answering System using BERT. Arab J Sci Eng. Jun 23, 2021;48(8):1-11. [doi: 10.1007/s13369-021-05810-5] [Medline: 34178569]
11. Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: early experiments with GPT-4. arXiv. Preprint posted online on 2023. [doi: 10.48550/ARXIV.2303.12712]
12. Zhang C, Liu S, Zhou X, et al. Examining the role of large language models in orthopedics: systematic review. J Med Internet Res. Nov 15, 2024;26:e59607. [doi: 10.2196/59607] [Medline: 39546795]
13. Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models. Preprint posted online on Jul 6, 2023. arXiv. [doi: 10.48550/ARXIV.2307.03109]
14. Seo S, Kim K, Yang H. Performance Assessment of Large Language Models in Medical Consultation: Comparative Study. JMIR Med Inform. Feb 12, 2025;13:e64318. [doi: 10.2196/64318] [Medline: 39763114]
15. Tang R, Chuang YN, Hu X. The science of detecting LLM-generated texts. arXiv. Preprint posted online on 2023. [doi: 10.48550/ARXIV.2303.07205]
16. Li M, Zhan Z, Yang H, Xiao Y, Huang J, Zhang R. Benchmarking retrieval-augmented large language models in biomedical NLP: application, robustness, and self-awareness. arXiv. Preprint posted online on 2024. [doi: 10.48550/ARXIV.2405.08151]
17. Tan Y, Min D, Li Y, et al. Can chatgpt replace traditional KBQA models? an in-depth analysis of the question answering performance of the GPT LLM family. arXiv. Preprint posted online on 2023. [doi: 10.48550/ARXIV.2303.07992]
18. Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. PMC-LLaMA: toward building open-source language models for medicine. J Am Med Inform Assoc. Sep 1, 2024;31(9):1833-1843. [doi: 10.1093/jamia/ocae045] [Medline: 38613821]
19. Yan Z, Liu J, Fan Y, et al. Ability of ChatGPT to replace doctors in patient education: cross-sectional comparative analysis of inflammatory bowel disease. J Med Internet Res. Mar 31, 2025;27:e62857. [doi: 10.2196/62857] [Medline: 40163853]
20. He Z, Bhasuran B, Jin Q, et al. Quality of answers of generative large language models versus peer users for interpreting laboratory test results for lay patients: evaluation study. J Med Internet Res. Apr 17, 2024;26:e56655. [doi: 10.2196/56655] [Medline: 38630520]

21. Šuvalov H, Lepson M, Kukk V, et al. Using synthetic health care data to leverage large language models for named entity recognition: development and validation study. J Med Internet Res. Mar 18, 2025;27:e66279. [doi: 10.2196/66279] [Medline: 40101227]
22. Zhou H, Li M, Xiao Y, Yang H, Zhang R. LEAP: LLM instruction-example adaptive prompting framework for biomedical relation extraction. J Am Med Inform Assoc. Sep 1, 2024;31(9):2010-2018. [doi: 10.1093/jamia/ocae147] [Medline: 38904416]
23. Zhou H, Austin R, Lu SC, et al. Complementary and Integrative Health Information in the literature: its lexicon and named entity recognition. J Am Med Inform Assoc. Jan 18, 2024;31(2):426-434. [doi: 10.1093/jamia/ocad216]
24. Li M, Zhou H, Yang H, Zhang R. RT: a Retrieving and Chain-of-Thought framework for few-shot medical named entity recognition. J Am Med Inform Assoc. Sep 1, 2024;31(9):1929-1938. [doi: 10.1093/jamia/ocae095] [Medline: 38708849]
25. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv. Preprint posted online on 2023. [doi: 10.48550/ARXIV.2303.13375]
26. Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. arXiv. Preprint posted online on 2023. [doi: 10.48550/ARXIV.2303.08774]
27. Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. arXiv. Preprint posted online on 2023. [doi: 10.48550/ARXIV.2302.13971]
28. Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models. arXiv. Preprint posted online on 2023. [doi: 10.48550/ARXIV.2307.09288]
29. Grattafiori A, Dubey A, Jauhri A, et al. The llama 3 herd of models. arXiv. Preprint posted online on 2024. [doi: 10.48550/ARXIV.2407.21783]
30. An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality vicuna open-source chatbot impressing GPT-4 90 ChatGPT qual. UC Berkeley Sky Computing Lab. 2023. URL: https://lmsys.org/blog/2023-03-30-vicuna/ [Accessed 2025-06-26]
31. Ji Z, Yu T, Xu Y, Lee N, Ishii E, Fung P. Towards mitigating LLM hallucination via self reflection. Presented at: Findings of the Association for Computational Linguistics. Association for Computational Linguistics. 1827-1843; 2023. [doi: 10.18653/v1/2023.findings-emnlp.123]
32. Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. Interact J Med Res. Jan 28, 2025;14:e59823. [doi: 10.2196/59823] [Medline: 39874574]
33. Aljamaan F, Temsah MH, Altamimi I, et al. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. JMIR Med Inform. Jul 31, 2024;12:e54345. [doi: 10.2196/54345] [Medline: 39083799]
34. Han T, Adams LC, Papaioannou JM, et al. An open-source collection of medical conversational AI models and training data arxiv. arXiv. Preprint posted online on 2023. [doi: 10.48550/ARXIV.2304.08247]
35. Suzuoki S, Hatano K. Reducing hallucinations in large language models: a consensus voting approach using mixture of experts. Preprint posted online on 2024. [doi: 10.36227/techrxiv.171925057.75949684/v1]
36. Ferrario A, Demiray B, Yordanova K, Luo M, Martin M. Social reminiscence in older adults' everyday conversations: automated detection using natural language processing and machine learning. J Med Internet Res. Sep 15, 2020;22(9):e19133. [doi: 10.2196/19133] [Medline: 32866108]
37. Wang J, Chen H, Wang H, et al. A risk prediction model for physical restraints among older Chinese adults in long-term care facilities: machine learning study. J Med Internet Res. Apr 6, 2023;25:e43815. [doi: 10.2196/43815] [Medline: 37023416]
38. Pham DK, Vo BQ. Towards reliable medical question answering: techniques and challenges in mitigating hallucinations in language models. arXiv. Preprint posted online on 2024. [doi: 10.48550/ARXIV.2408.13808]
39. Atf Z, Safavi-Naini SAA, Lewis PR, et al. The challenge of uncertainty quantification of large language models in medicine. arXiv. Preprint posted online on 2025. [doi: 10.48550/ARXIV.2504.05278]
40. Yu F, Wu P, Deng H, et al. A questionnaire-based ensemble learning model to predict the diagnosis of vertigo: model development and validation study. J Med Internet Res. Aug 3, 2022;24(8):e34126. [doi: 10.2196/34126] [Medline: 35921135]
41. Chen X, Zhao Z, Zhang W, et al. EyeGPT for patient inquiries and medical education: development and validation of an ophthalmology large language model. J Med Internet Res. Dec 11, 2024;26:e60063. [doi: 10.2196/60063] [Medline: 39661433]
42. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare. Jan 31, 2022;3(1):1-23. [doi: 10.1145/3458754]
43. Singhal K, Azizi S, Tu T, et al. Publisher correction: large language models encode clinical knowledge. Nature New Biol. Aug 2023;620(7973):E19-E19. [doi: 10.1038/s41586-023-06455-0] [Medline: 37500979]

44. Zhou S, Wang N, Wang L, Liu H, Zhang R. CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. J Am Med Inform Assoc. Jun 14, 2022;29(7):1208-1216. [doi: 10.1093/jamia/ocac040] [Medline: 35333345]
45. Wang B, Xie Q, Pei J, et al. Pre-trained language models in biomedical domain: a systematic survey. ACM Comput Surv. Mar 31, 2024;56(3):1-52. [doi: 10.1145/3611651]
46. Gururangan S, Marasović A, Swayamdipta S. Don't stop pretraining: adapt language models to domains and tasks. Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jul 5-10, 2020:8342-8360; Online. [doi: 10.18653/v1/2020.acl-main.740]
47. Jiang D, Ren X, Lin BY. LLM-blender: ensembling large language models with pairwise ranking and generative fusion proc 61st annu meet assoc comput linguist vol 1. Presented at: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Jul 9-14, 2023:Association for Computational Linguistics. 14165-14178; 2023.[doi: 10.18653/v1/2023.acl-long.792]
48. Pitis S, Zhang MR, Wang A, Ba J. Boosted prompt ensembles for large language models. arXiv. Preprint posted online on 2023. [doi: 10.48550/ARXIV.2304.05970]
49. Preiksaitis C, Ashenburg N, Bunney G, et al. The role of large language models in transforming emergency medicine: scoping review. JMIR Med Inform. May 10, 2024;12:e53787. [doi: 10.2196/53787] [Medline: 38728687]
50. Wilhelm TI, Roos J, Kaczmarczyk R. Large language models for therapy recommendations across 3 clinical specialties: comparative study. J Med Internet Res. Oct 30, 2023;25:e49324. [doi: 10.2196/49324] [Medline: 37902826]
51. Deiner MS, Honcharov V, Li J, Mackey TK, Porco TC, Sarkar U. Large language models can enable inductive thematic analysis of a social media corpus in a single prompt: human validation study. JMIR Infodemiology. Aug 29, 2024;4:e59641. [doi: 10.2196/59641] [Medline: 39207842]
52. Dogan A, Birant D. A weighted majority voting ensemble approach for classification. Presented at: 2019 4th International Conference on Computer Science and Engineering (UBMK). Sep 11-15, 2019:IEEE. 1-6; Samsun, Turkey. [doi: 10.1109/UBMK.2019.8907028]
53. Kolter JZ, Maloof MA. Dynamic weighted majority: an ensemble method for drifting concepts. J Mach Learn Res. 2007;8:2755-2790. [doi: 10.5555/1314498.1390333]
54. Hruschka ER, Covoes TF. Feature selection for cluster analysis: an approach based on the simplified silhouette criterion. Presented at: International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06). Nov 28-30, 2005:IEEE. 32-38; Vienna, Austria. [doi: 10.1109/CIMCA.2005.1631238]
55. Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. In: Flores G, Chen GH, Pollard T, Ho JC, Naumann T, editors. Presented at: Proceedings of the Conference on Health, Inference, and Learning, PMLR. 248-260; 2022.[doi: 10.48550/arXiv.2203.14371]
56. Jin Q, Dhingra B, Liu Z, Cohen W, Lu X. PubMedQA: a dataset for biomedical research question answering. Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Nov 3-7, 2019:2567-2577; Hong Kong, China. [doi: 10.18653/v1/D19-1259]
57. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci (Basel). Jul 12, 2021;11(14):6421. [doi: 10.3390/app11146421]
58. Pal A, Umapathi LK, Sankarasubbu M. Med-halt: medical domain hallucination test for large language models. Presented at: Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL); Dec 6-7, 2023; Singapore. [doi: 10.18653/v1/2023.conll-1.21]
59. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. arXiv. [doi: 10.48550/ARXIV.1904.03323]
60. McQueen J. Some methods for classification and analysis of multivariate observations. Presented at: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California press. 281-297. 1967.
61. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. Nov 1987;20:53-65. [doi: 10.1016/0377-0427(87)90125-7]
62. Thorndike RL. Who belongs in the family? Psychometrika. Dec 1953;18(4):267-276. [doi: 10.1007/BF02289263]

ABBREVIATIONS

BERT: Bidirectional Encoder Representations from Transformers
KMeans: K-Means Clustering Algorithm
LLM: large language model
MedMCQA: medical multiple-choice question answering dataset

MedQA-USMLE: Medical Question Answering dataset based on the United States Medical Licensing Examination
NLP: natural language processing
PubMedQA: a dataset for biomedical question answering
QA: question answering

Edited by Javad Sarvestan, Tiffany Leung; peer-reviewed by Elaheh Moharamkhani, Jiaping Zheng; submitted 14.12.2024; final revised version received 09.05.2025; accepted 12.05.2025; published 14.07.2025

Please cite as:
Yang H, Li M, Zhou H, Xiao Y, Fang Q, Zhou S, Zhang R
Large Language Model Synergy for Ensemble Learning in Medical Question Answering: Design and Evaluation Study
J Med Internet Res 2025;27:e70080
URL: https://www.jmir.org/2025/1/e70080
doi: 10.2196/70080