AutoMedEval: Harnessing Language Models for Automatic Medical Capability Evaluation

Xiechi Zhang[1], Zetian Ouyang[1], Linlin Wang[1]*, Gerard de Melo[2], Zhu Cao[3], Xiaoling Wang[1], Ya Zhang[4,5], Yanfeng Wang[4,5], Liang He[1]
[1]East China Normal University, [2]Hasso Plattner Institute/University of Potsdam
[3]Tongji University, [4]Shanghai Jiao Tong University, [5]Shanghai AI Laboratory
{51255901060,51265901102}@stu.ecnu.edu.cn,
{llwang,xlwang,lhe}@cs.ecnu.edu.cn, gdm@demelo.org
caozhu55@gmail.com, {ya_zhang,wangyanfeng}@sjtu.edu.cn

Abstract

With the proliferation of large language models (LLMs) in the medical domain, there is increasing demand for improved evaluation techniques to assess their capabilities. However, traditional metrics like F1 and ROUGE, which rely on token overlaps to measure quality, significantly overlook the importance of medical terminology. While human evaluation tends to be more reliable, it can be very costly and may as well suffer from inaccuracies due to limits in human expertise and motivation. Although there are some evaluation methods based on LLMs, their usability in the medical field is limited due to their proprietary nature or lack of expertise. To tackle these challenges, we present AutoMedEval, an open-sourced automatic evaluation model with 13B parameters specifically engineered to measure the question-answering proficiency of medical LLMs. The overarching objective of AutoMedEval is to assess the quality of responses produced by diverse models, aspiring to significantly reduce the dependence on human evaluation. Specifically, we propose a hierarchical training method involving curriculum instruction tuning and an iterative knowledge introspection mechanism, enabling AutoMedEval to acquire professional medical assessment capabilities with limited instructional data. Human evaluations indicate that AutoMedEval surpasses other baselines in terms of correlation with human judgments.

1 Introduction

The emergence of increasingly powerful large language models (LLMs) has sparked significant advances across a range of real-world applications, including the medical field. The advent of numerous medical LLMs, e.g., the Med-PaLM series of models (Singhal et al., 2023, 2025; Tu et al., 2024), MedAlpaca (Han et al., 2023), MedLLaMA (Wu et al., 2023), and DoctorGLM (Xiong et al., 2023), highlights the need for reliable and comprehensive evaluation systems to assess and compare their performance. To date, comprehensively evaluating the capabilities of various medical LLMs remains a formidable challenge, due to the required medical expert knowledge and huge workload (Singhal et al., 2023; Chang et al., 2024).

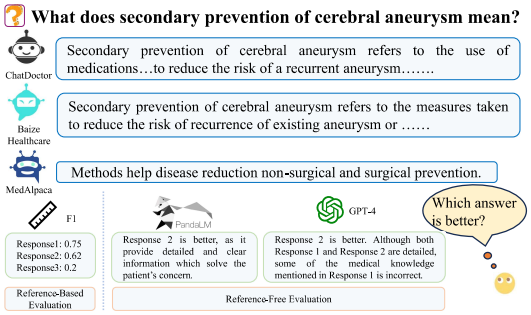

Figure 1: Typical example of medical LLMs' responses evaluation.

Human evaluation is often used to gauge the efficacy of medical LLMs, but this method is labor-intensive, time-consuming, and impractical on a large scale, especially when highly sought medical experts are needed (Xu et al., 2023b). As Figure 1 shows, traditional automatic evaluation methods, such as F1 and ROUGE (Lin, 2004), tend to focus predominantly on lexical matches, frequently overlooking semantic nuances. While metrics based on pre-trained language models like BERTScore can assess at the sentence level, they do not align well with human judgment (Liu et al., 2023; Nie et al., 2024). Benchmarks like USMLE (Jin et al., 2021) can be used to assess the capabilities of medical LLMs, but they cannot perform open-ended generation evaluations under common clinical settings. Although proprietary models like GPT-4 (Nori et al., 2023) can provide detailed assessments, they are hampered by a lack of transparency and potential information leakage. Furthermore, GPT-4's performance on certain medical tasks is notably lower than that of human doctors (Lai et al., 2023), suggesting a low correlation with human judgment.

*Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6272–6285
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Leveraging open-sourced LLMs as evaluators represents an innovative and promising approach, already implemented in general evaluation domains. Although open-sourced evaluation models, such as PandaLM-7B (Wang et al., 2024) and Auto-J (Li et al., 2023a) have been proposed, these models are designed for general scenarios. They often fall short in the medical domain, which necessitates specialized professional knowledge, making human evaluation essential but impractical at scale. This highlights the urgent need for an automatic, open-sourced evaluation model equipped with specialized medical expertise. However, incorporating accurate medical knowledge into the model and enabling it to perform detailed, human-like evaluations is challenging, especially when lacking large-scale, high-quality training data.

We propose AutoMedEval, a comprehensive evaluation model designed to provide detailed, human-aligned assessments, specifically intended to assist medical model developers in comparing the performance of different medical models. Detailed construction steps are as follows: (1) We first construct evaluation instructions from an existing medical QA dataset, developed using dynamic knowledge completion chains and validated by physicians through a double-checking procedure to ensure relevance for model training. (2) Based on the curated instructions, we leverage a hierarchical training approach to develop a novel LLM-based evaluation model, which is built upon the MedLLaMA-13B[1] model.

Hierarchical training approach includes curriculum instruction tuning and iterative knowledge introspection phases, allowing the model to align more closely with human evaluation, even in the absence of large-scale high-quality data. ① The curriculum instruction tuning phase comprises three stages designed to imbue AutoMedEval with essential medical knowledge, enabling a deeper understanding of the evaluation task. ② The iterative knowledge introspection phase empowers AutoMedEval to continuously refine its evaluation accuracy by integrating revision suggestions derived from collaborative feedback between the model and physicians, establishing it as a reliable tool for evaluating medical LLMs' performance.

To sum up, our key contributions are as follows:

- By leveraging a medical vector database with dynamic knowledge completions chains, we meticulously curate a high-quality medical instruction dataset. This dataset, rigorously validated by experienced physicians, serves as a robust foundation for the training of automatic evaluation models in the medical field.
- We propose an automated evaluation model called AutoMedEval, which is trained using a hierarchical training method and can introduce detailed, human-correlated evaluation of medical models.
- Human evaluation and double-blind experiments are performed, which prove the effectiveness of the AutoMedEval model and the hierarchical training method.

2 Task Formulation

This task consists of two sub-tasks: (i) instruction dataset construction and (ii) training of an automated evaluation model. Given a QA dataset T which consists of QA pairs (q_i, a_i) such that for each question q_i, each medical model j in a set S is used to generate a response $r_{i,j}$, leading to a combined $(q_i, r_{i,1}, ..., r_{i,|S|}, a_i)$ for evaluation. After evaluating using a retrieval augmented LLM (GPT-4 and ChatGPT), we obtain the evaluation of each instance, which consists of a rationale text e_i and a score s_i. Then the evaluation content and inputs are combined as instruction datasets in the format of $(q_i, r_{i,1}, ..., r_{i,|S|}, a_i, e_i, s_i)$. The evaluation model M trained using the instruction dataset should be able to rank each response $r_{i,j}$ and determine which medical model in S is the best-performing one.

3 Methodology

Overview As shown in the left part of Figure 2, we first introduce the construction of a double checked instruction dataset based on vector databases and LLMs (Section 3.1). Subsequently, as depicted in the middle part of Figure 2, we present a hierarchical training strategy, including the Curriculum Instruction Tuning phase (Section 3.2) and the Iterative Knowledge Introspection phase (Section 3.3). It is designed to effectively integrate the medical knowledge and evaluative criteria from the instruction dataset into the model and calibrate the model's evaluation with human standards.

3.1 Retrieval Augmented Instruction Dataset

As illustrated in the left part of Figure 2, to facilitate training automatic evaluation models in the medi-

[1]https://huggingface.co/chaoyi-wu/MedLLaMA_13B

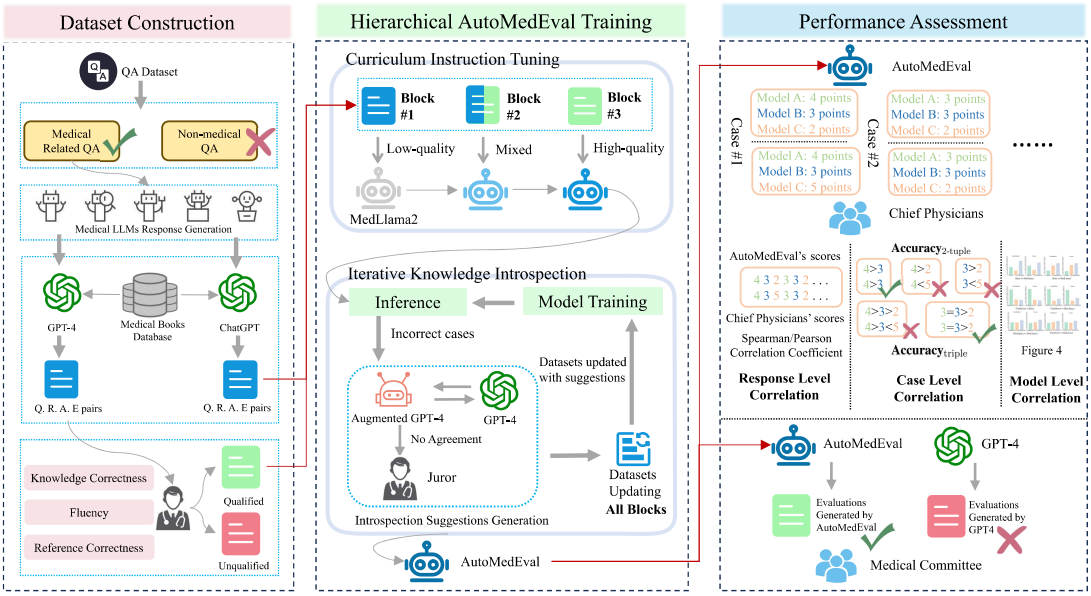

Figure 2: Creation of our instruction dataset and automatic evaluation model AutoMedEval.

Strategy	Data Block	Data Quality	Target	Level
Curriculum Instruction Tuning	Block #1	Low	Evaluation Pattern	Lv. 1
	Block #2	Mixed	From Pattern to Quality	Lv. 2
	Block #3	High	Evaluation Quality	Lv. 3
Iterative Knowledge Introspection	All Blocks	Mixed	Evaluation Calibration	Lv. 4

Table 1: Different stages of hierarchical training.

cal field, we construct an instruction dataset based on 10K open-domain medical question answers pairs T sourced from medical-meadow-wikidoc[2] dataset. Specifically, we screen out a set F_1 of 126 cases with medically irrelevant content using pattern matching (example can be seen in Appendix Table 4), leaving 9,874 cases ($D = T \backslash F_1$) including topics shown in Appendix Figure 7. Then we exploit medical LLMs, including ChatDoctor (Li et al., 2023b) and Baize Healthcare (Xu et al., 2023a), to generate responses, and create "*patient's question–responses–answer*" $(q_i, r_{i,1}, r_{i,2}, a_i)$ tuples as the inputs to GPT-4 for evaluation.

To craft domain-specific instruction data, we adopt the common practice (Wang et al., 2024) of distilling data from GPT-4 as evaluative evidence. We compile a collection of books and manuals shown in Appendix Table 6, encompassing medical knowledge and evaluation guidelines, to build a vector database. Subsequently, we utilize a customized prompt (Appendix Figure 8) with one-shot prompting to optimize GPT-4's adaptation and propose a novel dynamic knowledge completion chain (Algorithm 1) to boost the reasoning ability of GPT-4. Specifically, the instruction enables GPT-4 to generate a query starting with "**[Question]**" when uncertain about professional nouns or evaluation, and then our retrieval-augmented architecture will vectorize the query to provision corresponding knowledge from the vector database. The extracted evidence aids in forming a dynamic completion chain that guides GPT-4 to deliver more credible high-quality medical evaluations.

[2]https://huggingface.co/datasets/medalpaca/medical_meadow_wikidoc

Algorithm 1 Knowledge Completion Chain

Input: $\mathcal{D}$ = Selected medical-meadow-wikidoc data.
$\mathcal{T}$ = Maximum number of questions.
Output: E = Dataset with synthesized evaluations.
for $i \leftarrow 1$ to $|\mathcal{D}|$ **do**
  $d \leftarrow \mathcal{D}_i$
  for $t \leftarrow 1$ to $\mathcal{T}$ **do**
    $e \leftarrow \text{GPT}(d)$
    if "Question" in e **then**
      $k_{i,t} \leftarrow \text{Query}(\text{DB}_{\text{Pinecone}}, e)$
      $d \leftarrow (d, e, k_{i,t})$
    else
      $E \leftarrow E \cup (\mathcal{D}_i, e)$
      break
    end if
  end for
end for

Reliablity Verification After collecting the evaluations, we engage two chief physicians to scrutinize the evaluation content of each case about (1) the correctness of medical knowledge involved when analyzing each model's response. (2) whether there are instances of misattribution when analyzing each model's response. (3) the fluency of language and the ease of understanding during the analysis. Any instruction data failing to satisfy the above criteria are regarded as invalid and discarded. After verification, the proportion of sam-

ples that meet the standard for each sub-aspect of the evaluation content in set E is 94.06%, 90.71%, and 94.04%, corresponding to the previous criteria. Ultimately, a set F_2 of 305 unqualified cases is excluded, leaving a set comprising 9,569 entries ($R = E \backslash F_2$) and an example of the instruction is shown in the Appendix Figure 9.

3.2 Curriculum Instruction Tuning

To imitate the lack of high-quality tuning data, we draw instead on ChatGPT to generate 9,569 evaluations S on the same data. Then we use a classifier (details can be found in Appendix B) to obtain 4,788 high-quality instructions $R^{'}$ from GPT-4 and 3,823 relatively high-quality instructions $S^{'}$ from ChatGPT, with no intersection between the two sets of data, to train the model.

As depicted in the middle part of Figure 2, we use MedLLaMA as the foundation model. Inspired by the concept of cognitive synergy (Goertzel, 2017), we first train the AutoMedEval model using the curriculum instruction tuning strategy. As shown in the top part of Table 1, the curriculum learning strategy comprises three-stage instruction tuning which incorporates critical medical and evaluation knowledge into the MedLLaMA model. Specifically, we randomly select 1,911 evaluations $S_1^{'}$ from $S^{'}$ as the instructions for curriculum #1, which aims to help the model recognize evaluation patterns. Then we choose 2,394 evaluations $R_3^{'}$ from $R^{'}$ as curriculum #3 for steering the model to focus on the evaluation quality. Finally, we combine the remaining evaluations ($R^{'} - R_3^{'} + S^{'} - S_1^{'}$) to form the instructions for curriculum #2, which ensures the model's transformation from patterns to quality. Therefore, MedLLaMA is trained sequentially with #1, #2, and #3, and the training objective is formulated for model parameters $\widehat{\theta}$:

$$\widehat{\theta^*} = \underset{\widehat{\theta}}{\operatorname{argmax}} \sum_{j=1}^{N} \log p\left(Y^j \mid X^j; I^j, \phi\right) \quad (1)$$

where $I = \{I_c; I_m; I_h\}$ refers to curriculum-based instructions that belong to curriculum #1, #2, and #3, respectively. X^j and Y^j denote the inputs and generated evalution of the j-th training instance. ϕ denotes the remaining parameters of the model.

3.3 Iterative Knowledge Introspection

Subsequently, we introduce a technique shown in the Algorithm 2 to mitigate AutoMedEval's incorrect evaluation. One potential reason for such inaccuracies could be that AutoMedEval harbors misconceptions about certain elements of evaluation knowledge. Inspired by the process of humans refining initial drafts with feedback (Flower and Hayes, 1981) and cognitive introspection (Chen et al., 2023; Wang and Zhao, 2024), we propose an iterative knowledge introspection approach to enhance the multi-step medical reasoning accuracy of our AutoMedEval model and calibrate it with human standards. This approach ultimately pushes the upper limit of its capabilities through AI-doctor collaboration. Specifically, we utilize our retrieval-augmented GPT-4 to generate revision suggestions and employ the standard GPT-4 as a judge to assess these suggestions. As shown in Figure 3, if a consensus is not reached after up to three rounds of discussion, one chief physician will serve as the jury to make the final decision.These suggestions are then used to update original instruction datasets and iteratively fine-tune our model for knowledge introspection. Note that we generate revision suggestions exclusively for those training instances that were inaccurately predicted by the model from the previous iteration. The probability for the refined output $\widehat{Y}^j$ is formulated as

$$P_\theta = \sum_{Y^j} p\left(\widehat{Y}^j \mid X^j, Y^j, S; I^j, \phi\right) p\left(Y^j \mid X^j; I^j, \phi\right) \quad (2)$$

where S refers to the revision suggestions from the AI-doctor collaborative feedback.

Algorithm 2 Iterative Knowledge Introspection

Input: $\mathcal{R}$=$R^{'} + S^{'}$, $\mathcal{M}$=Previous trained model.
$\mathcal{T}$=Number of iterations.
Output: Trained AutoMedEval model $\mathcal{M}$.
for $t \leftarrow 1$ to $\mathcal{T}$ **do**
  $\mathcal{G} \leftarrow \mathcal{M}(\mathcal{R})$
  $\mathcal{I} \leftarrow$ Evaluate($\mathcal{G}$) // $\mathcal{I}$: incorrect cases
  for $i \leftarrow 1$ to $|\mathcal{I}|$ **do**
    $s_i \leftarrow$ Agent-Human Conversation($\mathcal{I}_i$)
    $\mathcal{I}_i \leftarrow (\mathcal{I}_i, s_i)$
  end for
  $\mathcal{R}^{'} \leftarrow$ Update($\mathcal{R}, \mathcal{I}$)
  $\mathcal{M} \leftarrow$ Train($\mathcal{R}^{'}, \mathcal{M}$)
end for

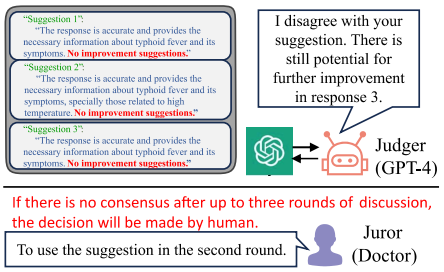

Figure 3: Collaborative Knowledge Introspection

Evaluation Models	Rationale Evaluation		Score Evaluation			
	BERTScore	BARTScore	Spearman	Pearson	$\text{Accuracy}_{\text{2-tuple}}$	$\text{Accuracy}_{\text{triple}}$
Closed-sourced models						
text-davinci-003	88.47*	-2.7589*	0.3267*	0.5459*	47.92*	17.11*
ChatGPT (gpt-3.5-turbo-0125)	94.58*	-1.7301†	0.4856*	0.5216*	64.38*	32.12*
GPT-4 (gpt-4-turbo-2024-04-09)	N/A	N/A	0.5128*	0.5689*	67.98*	35.42*
Gemini (gemini-2.0-flash)	93.67*	-2.0143*	0.5674*	0.6082*	71.68*	42.63*
Open-sourced models						
Vicuna-7B	86.94*	-2.7438*	0.2669*	0.2474*	56.48*	30.17*
Vicuna-13B	87.86*	-2.7609*	0.4766*	0.4908*	67.64*	41.87*
DeepSeek-R1-Distill-Llama-8B	94.02*	-1.9054*	0.5527*	0.6148*	71.06*	41.52*
MedLLaMA	85.84*	-2.7528*	0.3223*	0.4848*	55.18*	26.23*
PandaLM	N/A	N/A	0.4311*	0.4196*	67.27*	36.62*
Ours (AutoMedEval)	**94.72**	-1.7136	**0.6314***	**0.6854***	**74.61**	**48.65**
Ablation Study						
w/o Knowledge Completion Chain	93.98*	-1.7139	0.4998*	0.5510*	63.76*	27.38*
w/o Curriculum Instruction Tuning	94.69†	-1.7159	0.5375*	0.6026*	67.61*	38.88*
w/o Iterative Knowledge Introspection	94.62*	-1.6758*	0.5864*	0.6461*	70.62*	44.61*
Vanilla Instruction Tuning						
5,000 GPT-4 instructions w/ MedLLaMA	94.13*	-1.7208§	0.5614*	0.6251*	66.02*	31.88*
9,000 GPT-4 instructions w/ MedLLaMA	94.67*	**-1.6299***	0.5152*	0.5832*	67.96*	35.13*

Table 2: Comparison of results on different models. We run models three times and report the average results. * represents a significant difference with our results or significant correlation with human annotation (t-test, p-value$<$ 0.001), while † and § refer to t-test with $p<0.01$ and $p<0.05$, respectively. And the intraclass correlation coefficient (ICC) and Krippendorff's alpha among five doctors is 0.712 and 0.725.

4 Experiments

4.1 Baselines and Test Set

Baselines for our experiments include models ranging from closed-sourced models (text-davinci-003, ChatGPT, GPT-4, Gemini) to open-sourced ones (PandaLM, Vicuna-7B/13B, Deepseek). The test set mainly includes 958 entries from Medical Meadow Wikidoc dataset and an additional 172 entries from MedText[3] dataset. As for medical models evaluated, in addition to the ChatDoctor and Baize models mentioned earlier, MedAlpaca (Han et al., 2023) and MedLlama2[4], which are not utilized in the initial response gathering phase, are employed to generate responses for the questions in the test set and then evaluated by AutoMedEval.

4.2 Human Annotation and Judgement

For evaluating AutoMedEval's capability, we engage five doctors to annotate the quality of responses generated by medical models in the aspects of *Relevancy*, *Fluency*, and *Knowledge Correctness* and judge the evaluation generated by AutoMedEval in the aspects of *Reference*, *Fluency*, and *Knowledge Correctness*. Then the average score of the metrics mentioned above is defined as the final score of each response or evaluation content. After annotation, we calculate the intraclass correlation coefficient (ICC) and Krippendorff's alpha of five doctors to validate the annotation reliability. More annotation details can be found in Appendix D.

[3]https://huggingface.co/datasets/BI55/MedText

[4]https://huggingface.co/llSourcell/medllama2_7b

4.3 Evaluation Methods

Score Evaluation As depicted in the right part of Figure 2, we employ $\textbf{Accuracy}_{\text{2-tuple}}$ and $\textbf{Accuracy}_{\text{triple}}$ to assess the correlation at the case level. $\textbf{Accuracy}_{\text{2-tuple}}$ measures the consistency between the relative magnitude of AutoMedEval's scoring and annotated scores while $\textbf{Accuracy}_{\text{triple}}$ measures alignment over three medical LLMs. Besides, we use both **Spearman** and **Pearson** metrics to measure the correlation between AutoMedEval and humans at the response level.

Rationale Evaluation We use two semantic evaluation metrics, **BERTScore** (Zhang et al., 2019) and **BARTScore** (Yuan et al., 2021), to assist evaluation by using evaluations generated by GPT-4 as the reference answer.

Double-blind Preference A double-blind preference experiment is conducted to evaluate the practical applicability of AutoMedEval. Medical master candidates are instructed to make preference selections between the evaluation produced by AutoMedEval and the assessment results derived from GPT-4. More details can be seen in Appendix E.

4.4 Main Results

4.4.1 Score & Rationale Evaluation Results

To demonstrate the capability of AutoMedEval in evaluating medical models, we conducted a human-centered evaluation with the participation of chief

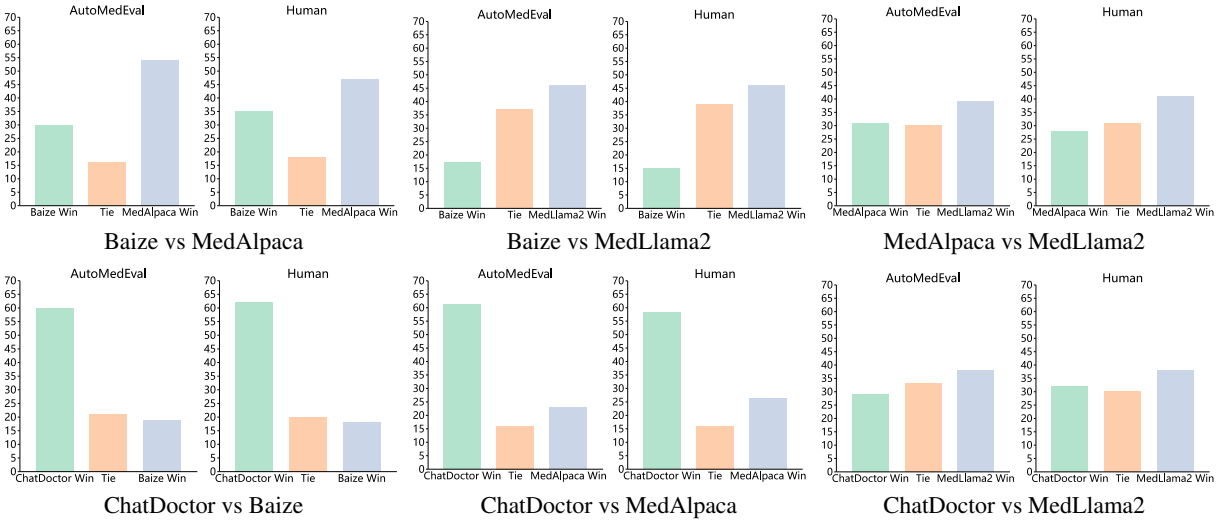

Figure 4: AutoMedEval and chief physicians' judgment on every two medical LLMs' performance. "Win" means the ratio of cases where the current medical LLM outperforms another, while "Tie" means both medical LLMs' scores are the same.

physicians. The main results in Table 2 provide a comparison of AutoMedEval with representative baselines. We observe that AutoMedEval outperforms all other evaluation models, particularly surpassing GPT-4 by a significant margin across all metrics. Compared to PandaLM, an automatic evaluator in general domain, our AutoMedEval model achieves a relative improvement of 10.9% on $\text{Accuracy}_{\text{2-tuple}}$.

4.4.2 Human-AutoMedEval Correlation

We draw a comparative visualization in Figure 4 that delineates the win rate comparisons between every pair of medical LLMs, as adjudicated by the chief physicians or the AutoMedEval model. A detailed examination of the win rate distributions within the evaluative outputs from AutoMedEval with those derived from human assessments reveals a notable congruence. This alignment suggests that AutoMedEval possesses the capability to discern the superior model with a degree of precision comparable to that of human evaluators, thereby underscoring the robust consistency between AutoMedEval and human annotators. Note that during the response generation phase, we use pairs of medical models to generate responses separately for each case. As shown in Appendix C, a sampling strategy is applied, causing the decoded response to vary for the same question under different cases, even for the same model. Therefore, performance transitivity does not apply to the results in Figure 4. Additionally, as illustrated in Figure 5 (a), the distribution of scores assigned by annotators to the evaluation content generated by AutoMedEval further demonstrates the experts' level of endorsement for it.

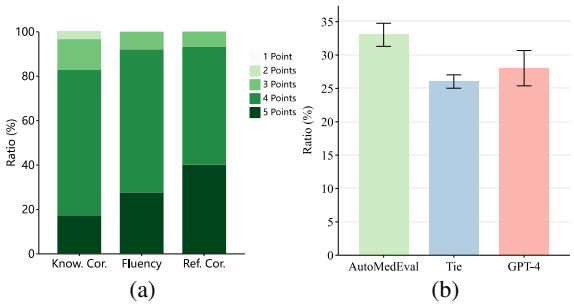

Figure 5: (a) Human assessment results on AutoMedEval's evaluation content. Know. Cor. represents Knowledge Correctness and Ref. Cor. means Reference Correctness. (b) Double-blind preference experiment results.

4.4.3 Double-blind Experiment Results

As illustrated in Figure 5 (b), the evaluation outcomes generated by AutoMedEval garnered greater endorsement among medical professionals when contrasted with those produced by GPT-4. The error bars denote that despite the variance in preference selections among the three medical experts, the proportion of preferences accorded to AutoMedEval consistently surpassed those allocated to GPT-4.

4.5 Quantitative Analysis

4.5.1 Ablation Studies

To quantify the contributions of various components in AutoMedEval, we conducte ablation studies with three simplified architectures in Table 2. We observe that all components contribute signifi-

cant improvements. For example, the ablation of knowledge completion chain results in a relative 14.5% degradation in $\text{Accuracy}_{\text{2-tuple}}$. Additionally, the removal of iterative knowledge introspection leads to a relative 5.3% and 8.3% decrease in $\text{Accuracy}_{\text{2-tuple}}$ and $\text{Accuracy}_{\text{triple}}$ respectively.

Iteration Times	0	1	2
$\textbf{Accuracy}_{\text{triple}}$	44.61	47.13	48.65
$\textbf{Accuracy}_{\text{2-tuple}}$	70.80	73.01	74.61

Table 3: Iteration results on the test set.

4.5.2 Effect of Varying Iterative Rounds

To investigate the impact of varying iterative rounds in knowledge introspection training, we conduct comparative experiments by training AutoMedEval with different iterations of knowledge introspection. As Table 3 depicts, an additional iteration of knowledge introspection can enhance the performance of AutoMedEval. For instance, after two iterations, our model outperforms curriculum instruction tuning by a relative 9.1% improvement on $\text{Accuracy}_{\text{triple}}$. We conduct mathematical modeling to determine when AutoMedEval achieves its best performance and find that AutoMedEval will outperform curriculum instruction tuning with a relative improvement of 17% on the $\text{Accuracy}_{\text{triple}}$ metric and there will not be any further growth after six iterations. Details of mathematical modeling are given in Appendix F.

4.5.3 Effectiveness under Limited Instructions

The results in Table 2 also confirm the effectiveness of AutoMedEval using limited high-quality instructions. We observe that transitioning from 5,000 GPT-4 generated instructions to 9,000 instructions results in a significant improvement in $\text{Accuracy}_{\text{triple}}$ for MedLLaMA, indicating that increasing the number of high-quality instructions for tuning can enhance overall performance. Interestingly, AutoMedEval, utilizing 5,000 high-quality and 4,000 low-quality instructions, outperforms MedLLaMA with 9,000 high-quality instances by a relative improvement of 38.5%. This demonstrates AutoMedEval's ability to address the scarcity of high-quality instructions by utilizing lower-quality data through our proposed instruction tuning.

4.6 Qualitative Analysis

4.6.1 Case Study on Evaluations

As the case Figure 6 shows, during evaluation, our model first attempts to interpret the patient's question, and then evaluates each response according to the question. Eventually, it will assign scores to each response. In the given example, AutoMedEval first comprehends the patient's question and then evaluates each response about whether it explains the secondary prevention of cerebral aneurysm. This also entails comparing the content of different responses, such as the detailed measures mentioned in the first response, which are lacking in the second response. Thus, our model gives 5 points to ChatDoctor's response, 4 points to Baize Healthcare's response, and 1 point to Retrieval's response, which appears reasonable and accurate.

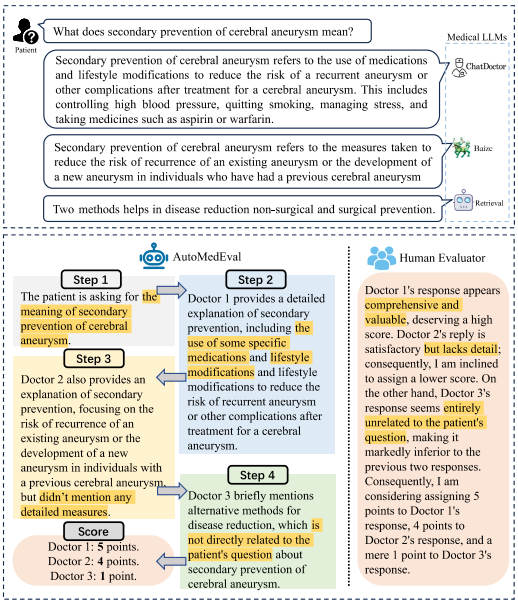

Figure 6: The rationale and score from AutoMedEval (left) and human annotator (right) for one example.

4.6.2 Quality Inspection for Instructions

For additional inspection of the data, we invoked LDA to model the topics in the remaining 9,569 instructions. We observe that the grouped topics are indeed highly task-related, including *disease, medicine, diagnostics, treatment, prevention measures* and *medical books*.

Additionally, we conducted manual reviews of all non-consensus cases to check revision suggestions after receiving collaborative feedback. We observed that many instructions with incorrect medical knowledge were corrected, providing our knowledge introspection stage with more precise evidence. For the example in Figure 3, we obtained a new revision suggestion as *"The response should include information about the symptoms of typhoid fever related to high temperature."*

5 Discussions

5.1 GPT-4 vs AutoMedEval

With the development of LLMs, traditional metrics such as BLEU and ROUGE are no longer suitable for assessing the capabilities of LLMs in the NLG domain. Although proprietary models like GPT-4 have been proven to possess a certain level of general evaluation capability ((Wang et al., 2023), (Li et al., 2024)), ablation studies on GPT-4 reveal that its assessment ability in domains requiring specialized knowledge needs further enhancement. Moreover, due to their proprietary nature, models like GPT-4 are not suitable for use in professional fields with high privacy requirements, such as the medical field. In contrast, the AutoMedEval model we proposed, is an open-source evaluation model built upon a medical LLM backbone, which can effectively advance the development of evaluation models in the medical domain.

5.2 Limitation of AutoMedEval

We randomly sampled 20 cases to explore the limitations of AutoMedEval's medical knowledge. While it is evident that all evaluation content from AutoMedEval involves a certain level of medical knowledge, some inaccuracies can be categorized as follows: (1) incomplete expressions, (2) outdated information, (3) bias and hallucination, (4) unsupported ratings.

6 Related Work

Evaluation methodologies for medical LLMs generally fall into two categories: automatic and human evaluation. For automatic evaluation, fixed pre-defined metrics like precision, recall, and F1-score have long been used for information extraction (Chinchor and Sundheim, 1993; Chen et al., 2024b). The perplexity (Manning and Schutze, 1999) metric was introduced to score the fluency of a model's output sentences. Comparisons with reference outputs can be obtained using BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). Most traditional automated metrics, however, assess the effectiveness of models solely at the lexical level, which is inadequate for more complex generation tasks, due to their failure to consider semantics and poor alignment with human judgments.

Human evaluations can be adjusted based on specific needs and expertise, but they may be prone to errors due to human limitations, especially in knowledge-intensive fields like medicine. Typically, human evaluation is used to support automatic evaluation results (Belz and Reiter, 2006). While human evaluation can be applied to entire datasets (Xu et al., 2023b), it is impractical at large scales due to the significant resources required, including time and money.

While LLMs continue to advance at a rapid pace, progress on automated evaluation methods to assess their performance has lagged behind. Although ChatGPT, GPT-4, and Gemini can aid automated assessments (Nori et al., 2023; Li et al., 2024; Wang et al., 2023), they remain suboptimal options due to their proprietary nature and lack of reproducibility. Open-source evaluation models such as PandaLM and Auto-J are meant for generic tasks (Wang et al., 2024) and obtain unsatisfactory results in the information-rich medical domain. Though some benchmarks have been proposed (Pal et al., 2022; Jin et al., 2021), it remains challenging to evaluate LLM's open-ended QA performance with benchmarks (Zheng et al., 2023; Chen et al., 2024a).

7 Conclusion

In this paper, we propose an instructions dataset and an automated evaluation model AutoMedEval that can facilitate the automatic evaluation of LLM in the medical field. Specifically, we use medical question-answer data and a novel dynamic knowledge completion chain method to collect evaluation results from the ChatGPT and GPT-4 models, which are then verified by chief physicians.

We further propose a hierarchical training strategy that including two techniques to train the model when lacking high-quality instruction data. One of these is curriculum learning, which uses lower-quality data and high-quality data to progressively bootstrap the model to gain evaluation capabilities. Another technique is iterative knowledge introspection, i.e., using training data to obtain extra suggestions, which helps the model align well with human judgment. Both training techniques benefit the model's performance in a series of comparative and ablation experiments.

Although AutoMedEval, trained on our new instruction dataset, demonstrates good performance, there is still some gap from practical applications. We will focus on collecting additional instruction data to develop superior evaluation models.

License

The dataset, models, and tools used in this paper are open-sourced or permitted to be used for scientific research. The AutoMedEval model in the paper should only be used for research purposes.

Ethics Statement and Limitations

We caution that the AutoMedEval model is primarily designed to help advance research and assess the performance of newly developed medical large language models in fundamental research settings. It is not intended to prove a medical model's suitability or effectiveness for genuine real-world deployment. As is the case for other automated text evaluation methods, the evaluations that our framework produces are merely shown to exhibit correlations with human judgments. However, despite a high correlation, such models can nevertheless make important mistakes. These obtained correlations are instead intended to help advance the state-of-the-art in research on medical language models.

Another point that we wish to emphasize is the important role of data in medical language models. Special care was taken in this work to ensure that the data utilized is taken solely from open-source platforms that do not contain personal information.

References

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of nlg systems. In *11th conference of the european chapter of the association for computational linguistics*, pages 313–320.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Kedi Chen, Qin Chen, Jie Zhou, Yishen He, and Liang He. 2024a. Diahalu: A dialogue-level hallucination evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 9057–9079. Association for Computational Linguistics.

Kedi Chen, Jie Zhou, Qin Chen, Shunyu Liu, and Liang He. 2024b. A regularization-based transfer learning method for information extraction via instructed graph decoder. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/-COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 1472–1485. ELRA and ICCL.

Liting Chen, Lu Wang, Hang Dong, Yali Du, Jie Yan, Fangkai Yang, Shuang Li, Pu Zhao, Si Qin, Saravan Rajmohan, et al. 2023. Introspective tips: Large language model for in-context decision making. *arXiv preprint arXiv:2305.11598*.

Nancy Chinchor and Beth M Sundheim. 1993. Muc-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*.

Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College composition and communication*, 32(4):365–387.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Ben Goertzel. 2017. Toward a formal model of cognitive synergy. *arXiv preprint arXiv:1703.04361*.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

U Hin Lai, Keng Sam Wu, Ting-Yu Hsu, and Jessie Kai Ching Kan. 2023. Evaluating the performance of chatgpt-4 on the united kingdom medical licensing assessment. *Frontiers in Medicine*, 10.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023a. Generative judge for evaluating alignment. *Preprint*, arXiv:2310.05470.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023b. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. 2024. Leveraging large language models for nlg evaluation: A survey. *arXiv preprint arXiv:2401.07103*.

Chin Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

Allen Nie, Yuhui Zhang, Atharva Shailesh Amdekar, Chris Piech, Tatsunori B Hashimoto, and Tobias Gerstenberg. 2024. Moca: Measuring human-language model alignment on causal and moral judgment tasks. *Advances in Neural Information Processing Systems*, 36.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

S Rajbhandari, J Rasley, O Ruwase, and Y He. 2019. Zero: memory optimization towards training a trillion parameter models. arxiv e-prints arxiv: 11910.02054 (2019).

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, pages 1–9.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. 2024. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, et al. 2024. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. In *ICLR*.

Yuqing Wang and Yun Zhao. 2024. Metacognitive prompting improves understanding in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1914–1926.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023a. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6268–6278.

Jie Xu, Lu Lu, Sen Yang, Bilin Liang, Xinwei Peng, Jiali Pang, Jinru Ding, Xiaoming Shi, Lingrui Yang, Huan Song, Kang Li, Xin Sun, and Shaoting Zhang. 2023b. Medgpteval: A dataset and benchmark to evaluate responses of large language models in medicine. *Preprint*, arXiv:2305.07340.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Example of Filtered Data

An example data item filtered through pattern matching string "rephras" is shown in Table 4.

Composition	Content
instruction	Answer this question truthfully
input	Could you please provide me with the text that needs to be **rephras**ed? As "What is human DNA?" is already in proper English.
output	The prehistory period dates from around 7 x 10 6 b2k to about 7,000 b2k.

Table 4: Example data filtered through pattern matching.

B Details of the Classifier

We use the verified GPT4 evaluations and the ChatGPT evaluations to train an evaluation quality classifier. Specifically, we initially label 200 cases with obvious reference or medical knowledge errors as "low quality" and mark 200 verified evaluations generated by GPT-4 as "high quality". Subsequently, a contrastive learning framework SimCSE (Gao et al., 2021) is leveraged to derive the embeddings for each instruction. Ultimately, we use embedding-label pairs to train a support vector machine for classification and employ grid search to optimize the classifier's parameters. The quality classifier's accuracy on the 100 test cases is 91% and is employed to distinguish the high-quality portions among the augmented instruction dataset.

C Experimental Settings

We train our 13B AutoMedEval model using 8 NVIDIA A100 80GB GPUs and leverage DeepSpeed (Rasley et al., 2020) ZeRO-Stage 3 (Rajbhandari et al., 2019) for optimization. We use AdamW optimization with the WarmupDecay learning rate scheduler and set the learning rate to be 2×10^{-5} with a warmup ratio of 0.03 for stability. Our model undergoes 5 training epochs with a batch size of 2 per GPU and a gradient accumulation step of 8. We adopt Flash attention (Dao et al., 2022; Dao, 2023) for efficient memory usage, thereby allowing a maximum input size of 2,048 tokens. For ChatDoctor, Baize Healthcare and other medical models during response generation, we set the temperature to 0.5 and the maximum number of new tokens as 200, and apply a sampling strategy with top k and top p as 50 and 1, respectively.

D Annotation Details

The expert team was composed of five doctors from different departments of tertiary hospitals consisting of three attending doctors and two chief physicians. An example of annoation results is shown in Table 7. For annotation, we developed and deployed a simple website that predefined all rating rules and requires the doctors to click on the relevant options. Before annotating, the following assessment standards were given to and read by each doctor:

Evaluation Metrics:

- *Relevancy assesses how well generated responses match the corresponding questions.*
- *Fluency evaluates naturalness and human-like quality of responses or AutoMedEval's evaluations.*
- *Knowledge Correctness is medical knowledge accuracy in responses or evaluations.*
- *Reference Correctness assesses whether there are instances of misattribution in AutoMedEval's evaluation when quoting the medical LLMs.*

Two tasks need to be completed:
(1) After each response, three scores need to be assigned, corresponding to the following three aspects:

- *Assess the relevance of each response to the question (whether it answers the question asked).*
- *Evaluate the correctness of medical knowledge contained in each response (appropriate use of terminology).*
- *Assess the fluency of language and the ease of understanding of each response.*

(2) At the end of each sample, rate the following three aspects related to the "assessment":

- *Evaluate the correctness of medical domain knowledge applied when analyzing each doctor's response.*
- *Assess whether there are instances of misattribution when analyzing each doctor's response.*
- *Evaluate the fluency of language and the ease of understanding during the analysis.*

Note: Please do not refer to the scores in the "assessment" section before completing the first part of the task.

After annotation is completed, we perform an averaging operation on the scores assigned by all annotators for each annotation item to derive the final rating.

E Detail of Double Blind Preference Experiment

We invited three medical master candidates as the medical committee to participate in a double-blind preference experiment, with 87 randomly selected test samples. Before annotation, they were instructed to read the following guidelines:

After reading two evaluation results, you need to choose which one is better according the metrics listed below:

- *Knowledge Correctness: Evaluate the correctness of medical domain knowledge applied when analyzing each doctor's response.*
- *Reference Correctness: Assess whether there are instances of misattribution when analyzing each doctor's response (e.g., mistaking the content of the second doctor's response for the first doctor's response or unpractical information).*
- *Fluency: Evaluate the fluency of language and the ease of understanding during the analysis.*

F Mathematical Modeling

We utilize a variation of the sigmoid function to mimic the $P_t^{1,2,3}$ function as shown in Eq. (3), Eq. (4), and Eq. (5), which represents the probability each predicted score matches the ground truth.

$$P_t^1 = \frac{a_1}{1 + e^{-0.453t-2.83}} \quad (3)$$

$$P_t^2 = \frac{a_2}{1 + e^{-0.453t-2.83}} \quad (4)$$

$$P_t^3 = \frac{a_3}{1 + e^{-0.453t-2.83}} \quad (5)$$

The accuracy of each sample is obtained by multiplying the three accuracy functions mentioned above, resulting in:

$$f(t) = \frac{a_1 * a_2 * a_3}{(1 + e^{-0.453t-2.83})^3}. \quad (6)$$

Iteration Times	$\text{Accuracy}_{\text{triple}}$
Iter 0	44.61
Iter 1	47.13
Iter 2	48.65

Table 5: Iteration results on the test set.

Using the data of iteration results in Table 5, we calculate the final accuracy estimation as follows.

$$f(t) = \frac{1 * 0.9 * 0.586}{(1 + e^{-0.453t-2.83})^3} = \frac{0.5274}{(1 + e^{-0.453t-2.83})^3} \quad (7)$$

When the iteration count is set to 6, the accuracy reaches 52.21%.

G Instruction Examples

Figures 8 and 9 depict the instruction utilized to obtain evaluation data from GPT-4 and an example instruction exploited to tune the model, respectively.

H Example of Incorrect Evaluation

An incorrect evaluation case with reference evaluation is depicted in Table 8.

Figure 7: Topics covered by the dataset.

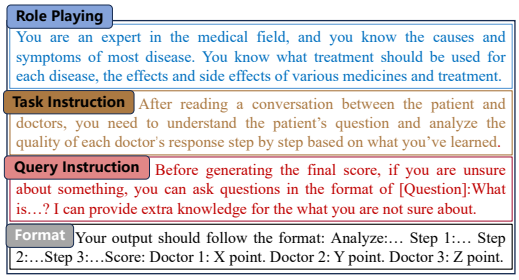

Figure 8: Instruction with diverse factors for GPT-4.

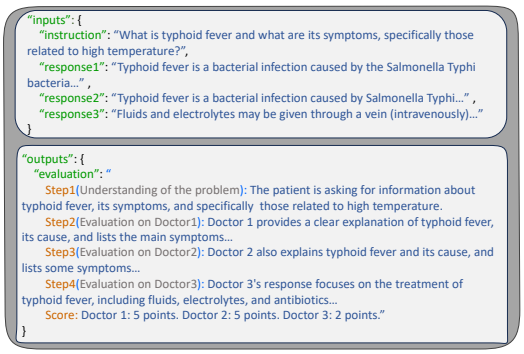

Figure 9: Example instruction for tuning AutoMedEval.

Medical Knowledge	Number of Pages
PFENNINGER & FOWLOER'S Procedures for Primary Care	1,763
DECISION MAKING in Medicine	753
TEXTBOOK OF Physical Diagnosis HISTORY AND EXAMINATION	965
Primer of DIAGNOSTIC IMAGING	808
THORACIC IMAGING	501
SEIDEL'S GUIDE TO PHYSICAL EXAMINATION AN INTERPROFESSIONAL APPROACH	676
Differential Diagnosis OF Common Complaints	703
Diagnostic Imaging Interventional Procedures	909
DIAGNOSTIC ULTRASOUND	2,249
Roberts and Hedges' CLINICAL PROCEDURES in Emergency Medicine and Acute Care	1,786
Mosby's Manual of Diagnostic and Laboratory Tests	1,201
HENRY'S Clinical Diagnosis AND Management By Laboratory Methods	1,956
Grainger & Allison's DIAGNOSTIC RADIOLOGY	2,218
Mosby's Diagnostic and Laboratory Test Reference	1,095
Evaluation Knowledge	**Number of Pages**
Matrix specification of Core Clinical Conditions for the Physician Associate by category of level of competence	127
Professional Practice Evaluation Policy	18
SoP Consolidated Version	202

Table 6: Details of the books and manuals used to build the vector database.

Composition	Content
Question	What information is available on Verapamil and Trandolapril?
Response 1 (ChatDoctor)	Verapamil and Trandolapril are both medications used to treat high blood pressure, also known as hypertension. Verapamil is a calcium channel blocker that helps to relax the blood vessels and reduce blood pressure by preventing the constriction of the blood vessels. Trandolapril is an ACE inhibitor which works by blocking the enzyme responsible for converting angiotensin I into a more potent form, thus reducing blood pressure. Both medications can be taken alone or in combination with other medicines. Common side effects of both drugs include dizziness, headache, fatigue, cough, and nausea. It is important to inform your doctor if you experience any serious side effects such as difficulty breathing, swelling, rash, or fever.
Annotator	Relevancy: 4 points. Knowledge Correctness: 4 points. Fluency: 4 points.
Response 2 (Baize Healthcare)	Verapamil is a calcium channel blocker medication used to treat high blood pressure, chest pain, and certain types of irregular heartbeats. Trandolapril is an ACE inhibitor used to treat high blood pressure. Both medications have potential side effects and should be taken under the guidance of a healthcare provider.
Annotator	Relevancy: 3 points. Knowledge Correctness: 2 points. Fluency: 2 points.
Resposne 3 (MedAlpaca)	The combination of verapamil and trandolapril is used to treat high blood pressure. It is a combination of two medications. It decreases certain chemicals that tighten the blood vessels, so blood flows more smoothly. It also relaxes your blood vessels so your heart does not have to pump as hard. This medication is sometimes prescribed for other uses; ask your doctor or pharmacist for more information.
Annotator	Relevancy: 2 points. Knowledge Correctness: 2 points. Fluency: 2 points.
Evaluation (AutoMedEval)	Analyze: Step 1: The patient is asking for information on Verapamil and Trandolapril. Step 2: Doctor 1 provides a comprehensive response, explaining the drug of both medications, their mechanism of action, and common side effects. The response also mentions the importance of informing a doctor if oftencing serious side effects. Step 3: Doctor 2 also provides information on the uses of Verapamil and Trandolapril, but the response is less detailed than Doctor 1's. The side effects mentioned are related to the medications, and the response emphasizes the importance of prescriptions under the care of a healthcare provider. Step 4: Doctor 3 explains the purpose of the combination between Verapamil and Trandolapril, but the response is cut off and does not provide complete information on the medications or their side effects. Score: Doctor 1: 5 points. Doctor 2: 4 points. Doctor 3: 2 points.
Annotator	Knowledge Correctness: 4 points. Reference: 4 points. Fluency: 4 points.

Table 7: An example of human annotation results.

Composition	Content
Question	What is a cavernous angioma?
Response ChatDoctor	A cavernous angioma is a type of vascular malformation, which are abnormal collections of blood vessels that form in the body and can cause problems such as bleeding, pain, and vision loss. They usually occur in the head or neck but can also be found elsewhere in the body.
Response Baize Healthcare	A cavernous angioma is a type of vascular tumor that forms in the brain. It is also known as a cavernous malformation or a cavernoma.
Resposne Reference	Cavernous angioma appears very similar on magnetic resonance imaging (MRI) with calcified neoplastic and hemorrhagic lesions, most especially renal cell carcinoma, melanoma, pleomorphic xanthoastrocytomas, and oligodendrogliomas.
Evaluation	Analyze: Step 1: The patient is asking for information about cavernous angioma. Step 2: Doctor 1 provides a general description of cavernous angioma, mentioning that it is a type of vascular malformation and can cause problems. However, it does not specifically address cavernous angioma. Step 3: Doctor 2 provides a more accurate description of cavernous angioma, stating that it is a type of vascular tumor that forms in the brain and is also known as a cavernous malformation or a cavernoma. Step 4: Doctor 3 provides information on the appearance of cavernous angioma on MRI, but does not provide a clear definition or explanation of what cavernous angioma is. Score: Doctor 1: 2 points. Doctor 2: 5 points. Doctor 3: 1 point.
Reference Evaluation	Analyze: Step 1: The patient is asking for information about cavernous angioma. Step 2: Doctor 1 provides a definition of cavernous angioma, mentioning that it is a benign tumor of the blood vessels, and usually appears in the brain or spine. They also mention the cause and some possible symptoms. Step 3: Doctor 2 also provides a definition of cavernous angioma, but focuses more on its occurrence as a type of brain tumor due to abnormal blood vessel growth. Step 4: Doctor 3 provides more detailed information about cavernous angioma, including its prevalence, the age range when symptoms usually manifest, and the fact that it can occur de novo, not just congenitally. Score: Doctor 1: 5 points. Doctor 2: 3 points. Doctor 3: 5 points.

Table 8: An incorrect evaluation case with reference evaluation.