Hierarchical Divide-and-Conquer for Fine-Grained Alignment in LLM-Based Medical Evaluation

Shunfan Zheng[1], **Xiechi Zhang**[1], **Gerard de Melo**[2,3], **Xiaoling Wang**[1], **Linlin Wang**[1*]

[1] East China Normal University
[2] Hasso Plattner Institute
[3] University of Potsdam
{sfzheng, 51255901060}@stu.ecnu.edu.cn, {xlwang, llwang}@cs.ecnu.edu.cn, demelo@uni-potsdam.de

Abstract

In the rapidly evolving landscape of large language models (LLMs) for medical applications, ensuring the reliability and accuracy of these models in clinical settings is paramount. Existing benchmarks often focus on fixed-format tasks like multiple-choice QA, which fail to capture the complexity of real-world clinical diagnostics. Moreover, traditional evaluation metrics and LLM-based evaluators struggle with misalignment, often providing oversimplified assessments that do not adequately reflect human judgment. To address these challenges, we introduce HDCEval, a **H**ierarchical **D**ivide-and-**C**onquer **Eval**uation framework tailored for fine-grained alignment in medical evaluation. HDCEval is built on a set of fine-grained medical evaluation guidelines developed in collaboration with professional doctors, encompassing Patient Question Relevance, Medical Knowledge Correctness, and Expression. The framework decomposes complex evaluation tasks into specialized subtasks, each evaluated by expert models trained through Attribute-Driven Token Optimization (ADTO) on a meticulously curated preference dataset. This hierarchical approach ensures that each aspect of the evaluation is handled with expert precision, leading to a significant improvement in alignment with human evaluators.

Models and supplementary materials: — https://huggingface.co/collections/AAAzsf/hdceval-6762cda19a07c157778aa22d

1 Introduction

With the rapid development of large language models (LLMs) in the medical field, a range of advanced medical LLMs have been developed. However, the reliability and effectiveness of these models must be rigorously evaluated to ensure accurate and safe clinical decisions.

However, existing benchmarks such as MT-Bench (Zheng et al. 2024) and MedBench (Cai et al. 2024) are often limited to tasks in fixed formats such as multiple-choice question answering (QA), as shown in Figure 1, lacking clinical freestyle generation, which does not align with the actual clinical diagnostic process. Moreover, current evaluation metrics fail to provide comprehensive evaluation results, instead offering only simplistic assessments. For instance, traditional n-gram metrics like ROUGE (Lin 2004) and BERT-based semantic similarity metrics (Zhang et al. 2019) yield only a single value, devoid of specific logical explanations.

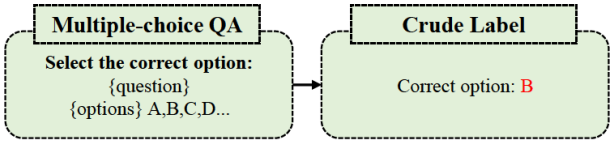


Figure 1: Fixed format task for evaluation.

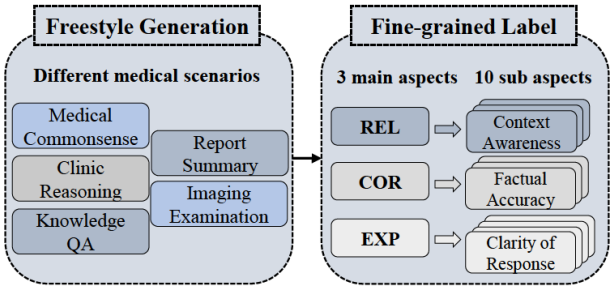


Figure 2: Freestyle fine-grained medical data for evaluation.

LLMs can serve as evaluators (Fu et al. 2024; Kocmi and Federmann 2023) in such freestyle contexts due to their generative capabilities. Unlike traditional metrics, LLMs can offer more nuanced and context-aware assessments by generating detailed feedback and explanations. This allows them to better reflect complex scenarios, such as those found in clinical diagnostics. However, existing LLM evaluators often exhibit misalignment with human evaluators in medical evaluation. For instance, evaluation using GPT-4 or the open-source model PandaLM (Wang et al. 2023c) can inadvertently perpetuate or even amplify existing bias in the training data, leading to skewed and inconsistent assessments (Stureborg, Alikaniotis, and Suhara 2024; Wang et al. 2023b) that may not accurately reflect diverse patient populations or medical scenarios compared to human physicians.

To address the issues above, we first collaborate with professional doctors to propose a set of fine-grained medical evaluation guidelines tailored for detailed medical assessments. These guidelines include three primary aspects: Patient Question Relevance (REL), Medical Knowledge Correctness (COR), and Expression (EXP), each further subdivided into specific sub-aspects.

*Corresponding Author

Based on the guidelines, we introduce HDCEval, a hierarchical divide-and-conquer evaluation framework that consists of two main components. Firstly, the **Divide** component involves a hierarchical decomposition of the evaluation task. The process begins by dividing the complex evaluation into multiple primary tasks. Each primary task is then further subdivided into more detailed subtasks. For each primary task and its corresponding subtasks, we employ a specialized expert model to carry out the evaluation, ensuring precise and expert-aligned assessments. This is in contrast to the BSM method (Saha et al. 2023), which relies on a single, non-specialized model to handle all primary tasks.

In the **Conquer** component, the framework leverages a carefully constructed preference dataset, which is specifically designed to improve alignment with human evaluators. This dataset plays a crucial role in enhancing the performance of each expert model. Based on this dataset, we introduce the Attribute-Driven Token Optimization (ADTO) method for training. This method incorporates reward tokens that guide the optimization of different expert models, ensuring that each model aligns with the specific evaluation criteria of its assigned tasks, thereby enhancing the precision and quality of the overall evaluation. The experimental results demonstrate that HDCEval significantly outperforms existing baseline methods across various medical scenarios. Notably, compared to the PandaLM evaluator, HDCEval achieves an overall improvement in consistency with human evaluations by 23.92%. This highlights the effectiveness of the Hierarchical Divide-and-Conquer Evaluation Framework in aligning model evaluations with expert-level assessments in the medical domain.

Our key contributions can be summarized as follows:

- We propose a comprehensive set of fine-grained medical evaluation guidelines developed in collaboration with professional doctors.
- We introduce HDCEval, a hierarchical divide-and-conquer evaluation framework designed for detailed and accurate medical evaluations, achieving finer-grained evaluation that better aligns with human evaluators.
- We develop and apply the Attribute-Driven Token Optimization (ADTO) strategy, demonstrating that HDCEval surpasses other baselines in accuracy and alignment with human evaluators in freestyle medical contexts.

2 Methodology

Task Formulation

In evaluation tasks, the input x consists of a question q and the model's response r. The goal is to generate an evaluation result E composed of multiple dimensions. In our medical evaluation tasks, the final evaluation consists of m distinct dimensions, denoted as $E = \{E_1, \ldots, E_m\}$, where each E_i represents the assessment of a specific dimension. Each E_i is defined as a tuple

$$E_i = (s_i, p_i), \tag{1}$$

where s_i is the scoring of the response on dimension i, and p_i is the corresponding rationale explaining the reasoning process.

Fine-grained Medical Evaluation Guidelines

Achieving accurate and nuanced evaluations is crucial in clinical diagnostics to ensure patient safety and effective treatment. To address this, we collaborated with medical experts to develop detailed evaluation guidelines specifically designed for medical assessments. These guidelines emphasize three primary aspects:

- **Patient Question Relevance (REL):** This aspect considers how well the medical response addresses the patient's specific questions and concerns. It involves assessing the clarity, directness, and appropriateness of the response in relation to the patient's query.
- **Medical Knowledge Correctness (COR):** This aspect ensures the accuracy of the medical information provided. It involves evaluating whether the response aligns with current medical knowledge, guidelines, and evidence-based practices.
- **Expression (EXP):** This aspect focuses on the clarity and coherence of the response, assessing the language, structure, and presentation of the information to ensure it is easily understandable and professional.

Each primary aspect is further divided into 3-4 sub-aspects to capture the intricacies of medical evaluations thoroughly. For instance, Patient Question Relevance (REL) includes sub-aspects such as Relevance to Patient's Condition (COND), which assesses how directly the response pertains to the patient's specific medical condition. Medical Knowledge Correctness (COR) encompasses sub-aspects like Factual Accuracy (ACC), ensuring the information aligns with current evidence-based practices. These sub-aspects provide a granular framework for evaluation, ensuring comprehensive coverage of each aspect. For each sub-aspect, scores range from 0 to 5, with detailed scoring rules provided in the Technical Appendix within supplementary materials.

Hierarchical Divide-and-Conquer Evaluation Framework

Overview As shown in Figure 3, the Hierarchical Divide-and-Conquer Evaluation Framework tackles medical evaluations by first *dividing* the task into detailed, expert-focused subtasks. Then, it *conquers* these tasks using preference data and Attribute-Driven Token Optimization (ADTO) to refine the model. This method ensures thorough and precise alignment with medical evaluation standards.

Hierarchical Divide Our medical evaluation guidelines are inherently multi-dimensional and strictly constrained, making accurate assessment a challenging task. LLMs often struggle with completing nuanced guideline-based estimation tasks due to their generalized training and lack of fine-tuned specialization.

To address these challenges, we propose a hierarchical divide-and-conquer approach, inspired by BSM (Saha et al. 2023). BSM's methodology demonstrates the efficacy of decomposing complex evaluation tasks into manageable subtasks that can be addressed in parallel. However, BSM's approach relies on a single model for all subtasks, which lim-

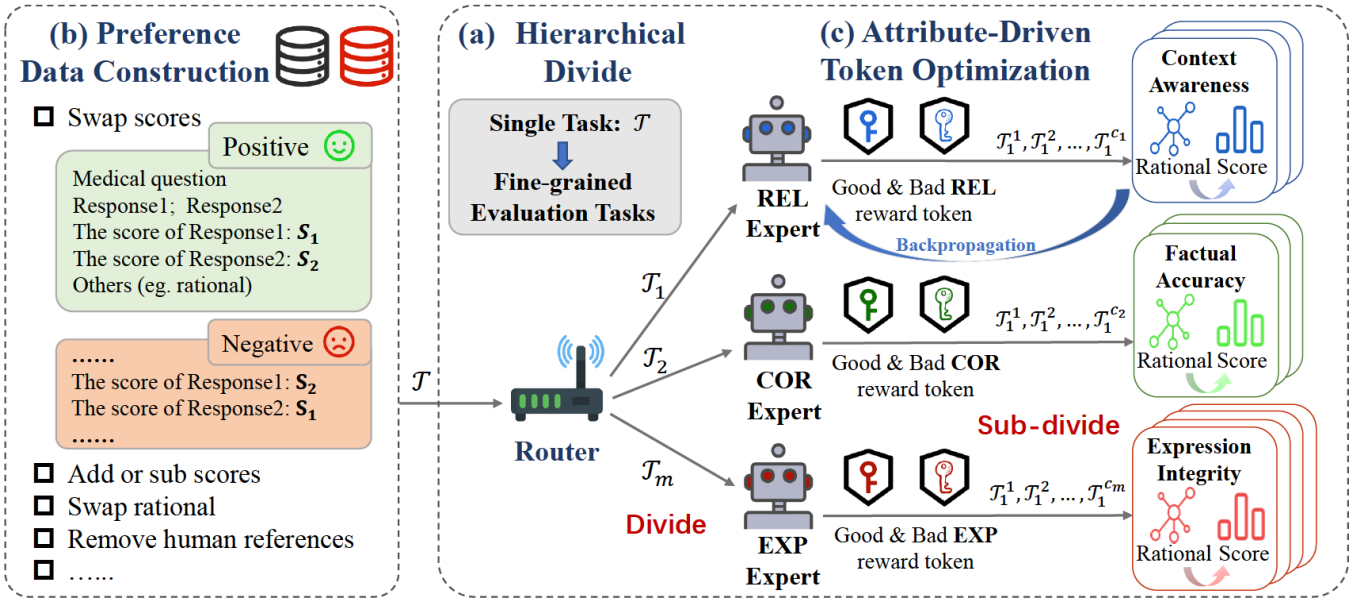


Figure 3: Overview of the Hierarchical Divide-and-Conquer Evaluation Framework. "Hierarchical Divide" represents the *Divide* component, while "Preference Data Construction" and "Attribute-Driven Token Optimization" constitute the *Conquer* component.

its its ability to achieve fine-grained alignment with human evaluators.

In contrast, our framework enhances this approach by using specialized expert models for different aspects of the evaluation. We first decompose the overarching evaluation task $\mathcal{T}$ into n primary evaluation tasks $\mathcal{T}_1, \ldots, \mathcal{T}_n$, each aligned with an expert model. These primary tasks are further subdivided into subtasks to capture the intricacies of the evaluation criteria.

The hierarchical decomposition is structured as follows:

$$\begin{cases} \mathcal{T} = \mathcal{T}_1(x, I_1), \ldots, \mathcal{T}_m(x, I_m) \\ \mathcal{T}_i(x, I_i) = \mathcal{T}_i^1(x, I_i, I_i^1), \ldots, \mathcal{T}_i^c(x, I_i, I_i^{c_i}) \end{cases} \tag{2}$$

Here, I_i represents the instruction of primary evaluation tasks $\mathcal{T}_i$ and I_i^j represents the instruction of subtasks $\mathcal{T}_i^j$.

Each expert model is trained specifically for one primary aspect (including its associated sub-aspects) to ensure fine-grained and accurate alignment with medical evaluation guidelines. This specialization allows for more precise evaluations that closely align with human expertise. Our hierarchical approach effectively manages the complexity of medical evaluations, ensuring a detailed and accurate assessment of each aspect.

Preference Data Construction To enhance model alignment with human evaluators in medical assessments, we develop a preference dataset that specifically targets misalignment and bias. This dataset construction is intricately linked to our fine-grained evaluation guidelines, ensuring that model improvements align with detailed evaluation criteria. The negative samples are constructed from existing positive samples, and this process is represented by the following formula:

- **Swapping Scores of Two Responses:** Let R_1 and R_2 be two responses with scores $S(R_1)$ and $S(R_2)$ from a positive sample. The scores of the corresponding negative sample are swapped:
$$S'(R_1) = S(R_2), S'(R_2) = S(R_1) \tag{3}$$
This method forces the model to determine which response better addresses the patient's query, refining its ability to assess relevance.
- **Simultaneously Adding or Subtracting Scores from Two Responses:** Considering two responses R_1 and R_2 with scores $S(R_1)$ and $S(R_2)$, we adjust the scores by a constant ΔS:
$$S'(R_1) = S(R_1) + \Delta S, \quad S'(R_2) = S(R_2) - \Delta S \tag{4}$$
This technique helps the model differentiate between high-quality and low-quality responses by teaching it to discern changes in accuracy and presentation.
- **Exchanging Rationales of Two Responses:** Let R_1 and R_2 be two responses with corresponding rationales $P(R_1)$ and $P(R_2)$. We swap their rationales:
$$P'(R_1) = P(R_2), \quad P'(R_2) = P(R_1) \tag{5}$$
This method ensures that the model's explanations align with its judgments, thereby reducing logical inconsistencies.
- **Removing Human-Provided Reference Information:** Let E_h represent an evaluation result that includes human-provided reference information I_h. The human-provided information is removed from the evaluation result:
$$E_h' = E_h \setminus I_h \tag{6}$$
This strategy reinforces the importance of human-provided information, allowing the model's outputs to better align with human expectations.

Algorithm 1: Attribute-Driven Token Optimization (ADTO)

Input: $\mathcal{D} = \{(x_j, y_{j,w}, y_{j,l})\}_{j=1}^{N}$, multi-dimensional evaluation data with positive (y_w) and negative (y_l) examples, $\mathcal{M} = \{M_1, M_2, \ldots, M_m\}$, set of fine-grained evaluation models, $\mathcal{R} = \{R_{\text{REL}}, R_{\text{COR}}, R_{\text{EXP}}\}$, set of reward tokens for relevance, correctness, and expression.

Output: Fine-Grained Medical Evaluation Models $\mathcal{M}$.

Initialization: Set model parameters θ_i for each $M_i \in \mathcal{M}$.
for each M_i in $\mathcal{M}$ **do**
  for each training step t **do**
    Sample a mini-batch $(x_j, y_{j,w}, y_{j,l})$ from $\mathcal{D}$
    Determine aspect $a \in \{\text{REL}, \text{COR}, \text{EXP}\}$ for M_i
    Construct reward token $r = R_a$
    Create inputs $z_w = \text{combine}(x_j, r, y_{j,w})$ and $z_l = \text{combine}(x_j, r, y_{j,l})$
    Compute model outputs $o_{j,w} = M_i(z_w; \theta_i)$ and $o_{i,l} = M_i(z_l; \theta_i)$
    Compute loss $\mathcal{L}(o_{j,w}, o_{j,l})$ using Attribute-Driven Token Optimization
    Freeze unrelated layers to stabilize training
    Update other parameters θ_i via gradient descent: $\theta_i \leftarrow \theta_i - \eta \nabla_{\theta_i} \mathcal{L}(o_{j,w}, o_{j,l})$
  end for
end for
return $\mathcal{M}$

Attribute-Driven Token Optimization To further reduce the bias of the evaluator, and improve the alignment between the models and professional physicians, we introduce the Attribute-Driven Token Optimization (ADTO) method.

The ADTO method leverages preference datasets by embedding specific reward tokens within the training data. These tokens represent different aspects of evaluation quality, and guide the model in distinguishing between superior and inferior responses. The integration of reward tokens enables the model to learn from nuanced distinctions that are critical in professional medical assessments.

For each i-th primary aspect evaluation model in our framework, the optimization process is designed to balance the current policy model's responses with those of a reference model. This process is mathematically formulated in Eq. 7 with respect to training model parameters π_θ^i, reference model parameters π_{ref}^i, and hyperparameter β_i as

$$\mathcal{L}_{\text{ADTO}}^i(\pi_\theta^i; \pi_{\text{ref}}^i) = -\mathbb{E}_{(x, y_w^i, y_l^i) \sim \mathcal{D}} \Bigg[\log \sigma \left(\beta_i \log \frac{\pi_\theta^i(y_w^i \mid x, t_w^i, I_i)}{\pi_{\text{ref}}^i(y_w^i \mid x, t_w^i, I_i)} - \beta_i \log \frac{\pi_\theta^i(y_l^i \mid x, t_l^i, I_i)}{\pi_{\text{ref}}^i(y_l^i \mid x, t_l^i, I_i)} \right) \Bigg], \quad (7)$$

where (x, y_w^i, y_l^i) refers to the triplet of (input, good evaluation, bad evaluation), and t_w^i, t_l^i represent different reward tokens in optimization. Here, $\pi_\theta^i(y_w^i \mid x, t_w^i, I_i)$ denotes the cumulative probability of the current policy model generating good responses, while $\pi_{\text{ref}}^i(y_l^i \mid x, t_l^i, I_i)$ represents the cumulative probability of the reference model generating bad responses. σ denotes the sigmoid function. Then, we integrate the optimization processes of all m models within the framework. Further details are specified in Algorithm 1.

The factual knowledge within large language models is often injected into deeper layers. Therefore, to equip the model with more accurate and objective evaluation capabilities while reducing the computational cost, we freeze the first 24 layers of our model and only train the last eight layers.

3 Experiments

Experimental Setup

Our evaluation models are based on the MedLlama2-7B model. We train using a batch size of 128 and a maximum token length of 4,096 on 4 NVIDIA A100-80GB GPUs. To maximize GPU memory usage and accelerate training, we employed the Fully Sharded Data Parallel (Zhao et al. 2023) strategy and the FlashAttention (Dao et al. 2022) algorithm. The learning rates for the instruction tuning and direct preference optimization phases are set to 2×10^{-5} and 5×10^{-7}, respectively. During inference, we use greedy decoding with a temperature of 0 to minimize randomness.

Medical Dataset

Data Source First, we integrate medical questions from different sources including medical meadow wikidoc[1], MedBench (Cai et al. 2024), MedText[2], and MedDialog (Zeng et al. 2020). We perform automated and manual filtering to ensure reliable and safe medical question sources. Then, to diversify the task types of the data and conform to the clinical medical scenario, we divide the data into five specific medical scenarios shown in Figure 2.

Dataset Construction First, we use four different medical models: ChatDoctor, Baize, MedAlpaca, and MedLlama2, to generate responses to the questions. Then, with the assistance of AI, we annotated the 13,452 samples following the Guidelines Instructions. More details about the dataset construction are provided in the Technical Appendix within supplementary materials.

Dataset Validation To validate the effectiveness of our dataset, we use the publicly available MedMCQA dataset (Pal, Umapathi, and Sankarasubbu 2022) as a reference. We evaluated four models on both datasets and calculated their rankings[3]. The results show consistent rankings of the models across the two datasets, with ChatDoctor demonstrating the best performance. Additionally, we find that ChatDoctor exhibits the strongest ability to follow instructions.

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

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

[3] https://github.com/ctllllll/understanding_llm_benchmarks

Medical Scenarios	Pairwise Accuracy (%)			Reference Match (%)			Correlation	
	REL	COR	EXP	REL	COR	EXP	Pearson	ICC
Imaging Examination (Text)								
MedLlama2	61.54*	52.38*	60.91*	60.72*	50.24*	60.86*	0.5484*	0.5893*
PandaLM	54.15*	50.53*	50.53*	55.84*	48.49*	53.71*	0.5604*	0.6141*
ChatGPT	69.23†	64.10†	55.77*	70.12†	64.40†	75.97*	0.5813†	0.6693†
GPT-4	**84.87***	61.54†	71.15*	**80.74***	69.46†	88.41*	0.5898†	0.6849†
Ours (HDCEval)	**84.87***	**79.49***	**75.00***	78.78*	70.03*	**92.98***	**0.6480***	**0.7149***
Clinic Reasoning								
MedLlama2	57.14*	48.83*	56.67*	59.72*	51.43*	56.32*	0.4060*	0.5247*
PandaLM	50.61*	46.12*	47.29*	56.67*	39.25*	51.62*	0.3637*	0.4919*
ChatGPT	64.63†	64.63§	59.69†	66.56†	64.74§	67.25†	0.5483§	0.7101§
GPT-4	78.91*	69.18†	60.20*	78.91†	70.44†	75.13†	0.5599†	**0.7209†**
Ours (HDCEval)	**82.99***	**67.35***	**67.35***	**88.23***	**82.50***	**85.81***	**0.5887***	**0.7209***
Knowledge QA								
MedLlama2	56.67*	47.08*	52.61*	56.67*	50.73*	54.52*	0.5240*	0.6917*
PandaLM	40.53*	39.18*	41.55*	48.65*	40.11*	45.74*	0.5181*	0.6469*
ChatGPT	63.33*	71.11*	58.33*	68.35*	72.01*	70.57†	0.5603*	0.6519*
GPT-4	76.11*	66.11*	61.67*	79.86*	73.67*	73.86*	0.5632*	0.6656*
Ours (HDCEval)	**85.00***	**73.33***	**74.17***	**86.78***	**78.41***	**81.85***	**0.5693***	**0.7073***
Report Summary								
MedLlama2	60.85*	58.86*	61.28*	61.63*	58.36*	62.96*	0.4303*	0.5595*
PandaLM	58.47*	45.20*	62.07*	62.79*	50.19*	63.46*	0.3947*	0.5342*
ChatGPT	72.13†	66.24*	69.68*	77.01*	67.66†	73.72*	0.5864†	0.6936†
GPT-4	74.88*	70.10*	**70.41***	77.06†	68.84*	**75.71***	0.5905*	0.6948*
Ours (HDCEval)	**75.24***	**72.76***	70.03*	**77.62***	**72.31***	73.75*	**0.5913***	**0.7047***
Medical Commonsense								
MedLlama2	58.82*	49.41*	56.73*	61.88*	53.92*	56.13*	0.3609*	0.4923*
PandaLM	57.05*	41.53*	52.71*	60.57*	43.63*	54.12*	0.3256*	0.4507*
ChatGPT	70.59*	68.55*	61.77*	71.41*	72.94*	68.33†	0.5815*	0.7238*
GPT-4	72.55*	68.63*	**79.41***	74.25†	76.91*	69.88*	0.5954*	0.7236*
Ours (HDCEval)	**74.51***	**70.59***	77.94*	**76.58***	**88.35***	**71.50***	**0.6767***	**0.7881***

Table 1: Fine-grained evaluation results. We run models three times and report the average results. * represents a significant difference with our results or significant correlation with human evaluation (t-test, p-value< 0.001), while † and § refer to t-test with $p<0.01$ and $p<0.05$, respectively.

Baselines and Test Set

We selected representative models as baselines, including the closed-source models (ChatGPT and GPT-4) and the open-source models (PandaLM and MedLlama2).

For the test data, we initially extracted 2,994 samples from the constructed dataset to form the test set, with the remaining samples used as the training set. We then hired human doctors to annotate the test data. The annotation process follows the fine-grained medical evaluation guidelines.

Evaluation

The generated evaluation results include both scores and rationales. Therefore, we need to assess these two aspects separately. For the scores, we employ automated metrics, while for the rationales, we rely on evaluations conducted by human doctors.

Human Evaluation The human annotators manually assess whether the rationale from the model matches the rationale from human-provided labels to verify the reasonableness of the model's evaluation results. This process is referred to as **Reference Match**. If the label's rationale indicates an error in the medical knowledge within the response, but the model fails to recognize it, it is considered a mismatch.

Automatic Metrics We use **Pairwise Accuracy** as the primary evaluation metric for the scores. If the relative ranking of the evaluation scores between the two responses generated by the model is consistent with the labels from human doctors, it indicates that the model accurately evaluated the quality of the two responses; otherwise, it does not. Additionally, we use the **Intraclass Correlation Coefficient** (Koo and Li 2016) and **Pearson Correlation Coefficient** (Cohen et al. 2009) to measure the similarity between the model evaluation and human evaluation.

Main Results

Evaluation Metrics Results To demonstrate the capability of HDCEval in fine-grained medical assessment, we arranged for medical experts to evaluate the responses. Table 1 provides the fine-grained evaluation results of HDC-

Method	REL			COR			EXP				AVG
	CONT	COND	CONC	ACC	INFO	UNC	CLAR	LANG	TE	INTE	
HDCEval	**80.91**	**81.82**	**78.83**	**71.65**	**72.18**	**74.27**	**73.48**	**75.14**	**70.84**	**72.14**	**75.13**
$\text{HDCEval}_{\text{no-token}}$	78.57	79.50	78.36	69.93	68.12	68.51	68.40	70.81	67.89	69.98	72.01
$\text{HDCEval}_{\text{no-preference}}$	80.70	80.97	77.94	71.35	69.57	71.29	71.36	69.80	67.13	71.17	73.13

Table 2: Ablation Study on HDCEval Components – Assessing the impact of removing reward tokens and preference data on evaluation accuracy.

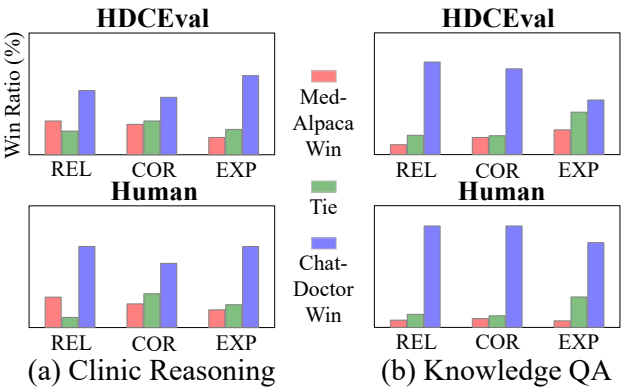


Figure 4: The performance of MedAlpaca and ChatDoctor across multiple medical scenarios is evaluated using HDCEval and compared to human doctors' judgments. "Win" indicates the percentage of cases where a given medical language model outperforms the other, while "Tie" indicates the percentage of cases where both medical LLMs received the same score.

Eval compared to the baselines. From left to right, the results of three different metrics on fine-grained dimensions are included. We observe that HDCEval outperforms other models across multiple scenarios, especially outperforming GPT-4 on reference match and correlation metrics, which reflects better alignment with humans. Regarding pairwise accuracy metrics, it demonstrates a 23.92% improvement compared to PandaLM. From the fine-grained perspective, HDCEval significantly improves evaluation accuracy compared to other baselines in terms of Medical Knowledge Correctness (COR).

Win-Tie-Lose Experiment Results We compiled statistics on the win rates of HDCEval and humans in assessing the quality of response pairs from MedAlpaca and ChatDoctor across different scenarios. The results presented in Figure 4 demonstrate consistent agreement between HDCEval and human evaluators across various medical scenarios. This consensus leads to the conclusion that ChatDoctor is significantly more effective than MedAlpaca.

Double Blind Experiment Results As shown in Figure 5, across the three primary fine-grained evaluation dimensions, human doctors show a preference for HDCEval that is comparable to their preference for GPT-4. In comparison to PandaLM, human doctors consistently favor the evaluation results provided by HDCEval.

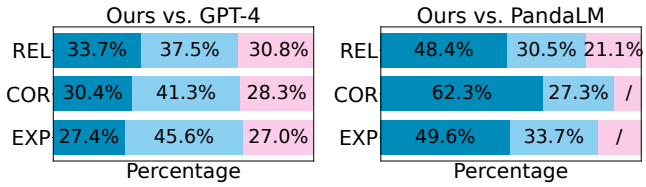


Figure 5: Preferences of human doctors between Our Method, GPT-4, and PandaLM.

Ablation Study

To validate whether our training method can improve assessment ability, we conducted the ablation experiments in Table 2. Removing the reward token weakens the model's perception of good and bad responses, resulting in a 3.75% decrease in evaluation accuracy. When preference data is excluded and only SFT is used for training, the evaluation accuracy drops by 0.5%, as ADTO better utilizes preference data to enhance performance further.

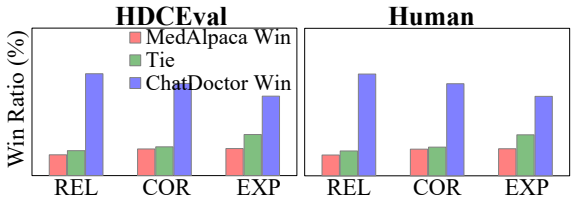
4 Discussion

Effects of Different Input Forms

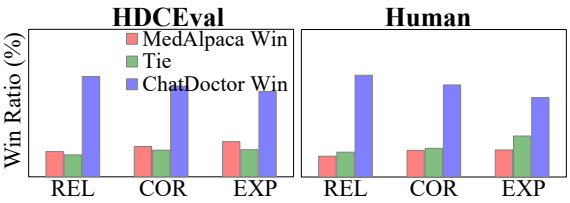
In practical applications, the format of evaluation tasks is not fixed. Therefore, we designed two evaluation tasks to explore HDCEval's generalization ability across different input formats. One task simultaneously evaluates the quality of two responses, while the other task separately evaluates the two responses and then compares the results. The results in Figure 6 indicate that different input formats do not significantly affect the evaluation results of HDCEval.

Exploration of Model Bias

In constructing the preference dataset in Section 2, we employed various strategies to mitigate model biases discussed in previous work (Zheng et al. 2024). For example, swapping the scores of two responses can mitigate position bias. To verify this, we conducted the experiments shown in Table 3, comparing the model bias with and without preference data during training. The results indicate that using preference data improves both position bias and verbosity bias.



(a) Pairwise Evaluation



(b) Single Evaluation

Figure 6: Multi Evaluation Task (Win, Tie, Lose) of HDCEval and Humans.

Bias Type	CONT	COND	CONC
Position Bias			
✓ Preference Data	**28.57**	**26.50**	**29.87**
✗ Preference Data	32.91	31.82	32.91
Verbosity Bias			
✓ Preference Data	**28.17**	**27.43**	**26.29**
✗ Preference Data	32.96	30.14	31.36

Table 3: Model bias for the three sub-aspects of patient question relevance (REL).

Case Study

Question:
What is Dermatophytosis and how can it be distinguished from other illnesses?

Response 1:
Dermatophytosis must be differentiated from other diseases that cause skin lesions, such as other causes of dandruff, seborrheic dermatitis, psoriasis, and other causes of skin rash.

Response 2:
Dermatophytosis is a fungal infection of the skin, hair, and nails. It is caused by several species of fungi, including ...

HDCEval:
Evaluation of Response 1:
Context Awareness (CONT):
Analysis: The response provides a general statement about the need to differentiate Dermatophytosis, but it does not ...
Score: 1
Relevance to Patient's Condition (COND):
Analysis: ....
Score: 1
Addressing Multiple Concerns (CONC):
Analysis: .....
Score: 1
Evaluation of Response 2:
Context Awareness (CONT):
Analysis: This response addresses ...
Score: 4
Relevance to Patient's Condition (COND):
Analysis: .....
Score: 3
Criterion Addressing Multiple Concerns (CONC):
Analysis: .....
Score: 4

Human Annotator:
R1 only extracts the key point of the question "Dermatophytosis", but does not explain what "Dermatophytosis" is, ...

As the above text-box shows, our model evaluates two responses based on detailed criteria sequentially. During the evaluation of each criterion, our model first analyzes each response according to the current criteria and ultimately assigns a score. The evaluation results generated by our model indicate that the first medical LLM's response is inferior to the second medical LLM's response across all three detailed criteria, which is corroborated by the human annotator's evaluation.

5 Related Work

Automated Model Evaluation Many researchers employ machine learning and NLP techniques to automatically evaluate responses from medical large language models. Some traditional metrics such as BLEU (Papineni et al. 2002) and ROUGE (Lin 2004) assess the quality of candidate text by statistically comparing n-grams between candidate and reference texts. However, these metrics are limited to the lexical level, disregarding much of the semantic information (Freitag et al. 2022).

In contrast, using BERT (Devlin et al. 2018) to assess the semantic similarity between candidate and reference embeddings is more reasonable (Zhang et al. 2019; Zhao et al. 2019). However, it can only provide a numerical value and cannot offer more logical explanations (Wang, Cho, and Lewis 2020; Huang et al. 2020), which can lead to a lack of credibility in evaluating medical models and misalignment with humans (Mehri and Eskenazi 2020; Zhong et al. 2022). Furthermore, existing benchmarks such as MT-Bench (Zheng et al. 2024) for evaluating the consistency between LLMs and human preferences, and MedBench (Cai et al. 2024) for medical domain evaluation, often employ fixed-form tasks such as multiple-choice questions, making it challenging to achieve evaluation in freestyle contexts.

LLM-Based Evaluators With the rapid advancement of large language models (LLMs) possessing powerful text comprehension and reasoning capabilities, recent research has seen the emergence of LLM-based evaluators (Fu et al. 2024; Wang et al. 2023a; Chen et al. 2023). They employ LLMs to assess text quality through methods such as prompting. For instance, utilizing models such as ChatGPT and GPT-4 in conjunction with specific prompting templates has enabled automated evaluation with some degree of success (Wu et al. 2023; Nori et al. 2023). However, models like GPT-4 are general-purpose and not specialized for specific evaluation tasks, thus exhibiting certain bias compared to humans (Wang et al. 2023b; Wu and Aji 2023). In contrast, open-source models like PandaLM (Wang et al. 2023c) are dedicated to evaluation tasks, but the medical domain requires rich specialized knowledge, which PandaLM lacks to some extent. In contrast to existing research, our work aims to produce fine-grained evaluation results from LLMs that align well with medical experts.

6 Conclusion

In this paper, we introduce HDCEval, a hierarchical divide-and-conquer evaluation framework specifically designed for evaluating medical language models. By dividing complex evaluation tasks into specialized subtasks and using expert models, HDCEval achieves greater alignment with human judgments and addresses the limitations of existing benchmarks and metrics. Our experiments demonstrate that HDCEval significantly outperforms baseline methods, improving consistency with human evaluations by 23.92%. This framework offers a more accurate, detailed, and reliable approach to assessing medical models, contributing to more effective clinical decision-making.

References

Cai, Y.; Wang, L.; Wang, Y.; de Melo, G.; Zhang, Y.; Wang, Y.; and He, L. 2024. MedBench: A large-scale Chinese benchmark for evaluating medical large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17709–17717.

Chen, Y.; Wang, R.; Jiang, H.; Shi, S.; and Xu, R. 2023. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. *arXiv preprint arXiv:2304.00723*.

Cohen, I.; Huang, Y.; Chen, J.; Benesty, J.; Benesty, J.; Chen, J.; Huang, Y.; and Cohen, I. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, 1–4.

Dao, T.; Fu, D.; Ermon, S.; Rudra, A.; and Ré, C. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. *Advances in Neural Information Processing Systems*, 35: 16344–16359.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Freitag, M.; Rei, R.; Mathur, N.; Lo, C.-k.; Stewart, C.; Avramidis, E.; Kocmi, T.; Foster, G.; Lavie, A.; and Martins, A. F. 2022. Results of WMT22 metrics shared task: Stop using BLEU–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 46–68.

Fu, J.; Ng, S.-K.; Jiang, Z.; and Liu, P. 2024. GPTScore: Evaluate as You Desire. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6556–6576. Mexico City, Mexico: Association for Computational Linguistics.

Huang, L.; Ye, Z.; Qin, J.; Lin, L.; and Liang, X. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. *arXiv preprint arXiv:2010.03994*.

Kocmi, T.; and Federmann, C. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Koo, T. K.; and Li, M. Y. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2): 155–163.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Mehri, S.; and Eskenazi, M. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. *arXiv preprint arXiv:2005.00456*.

Nori, H.; King, N.; McKinney, S. M.; Carignan, D.; and Horvitz, E. 2023. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, 248–260. PMLR.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Saha, S.; Levy, O.; Celikyilmaz, A.; Bansal, M.; Weston, J.; and Li, X. 2023. Branch-solve-merge improves large language model evaluation and generation. *arXiv preprint arXiv:2310.15123*.

Stureborg, R.; Alikaniotis, D.; and Suhara, Y. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.

Wang, A.; Cho, K.; and Lewis, M. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Wang, J.; Liang, Y.; Meng, F.; Sun, Z.; Shi, H.; Li, Z.; Xu, J.; Qu, J.; and Zhou, J. 2023a. Is ChatGPT a good NLG evaluator? A preliminary study. *arXiv preprint arXiv:2303.04048*.

Wang, P.; Li, L.; Chen, L.; Zhu, D.; Lin, B.; Cao, Y.; Liu, Q.; Liu, T.; and Sui, Z. 2023b. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Wang, Y.; Yu, Z.; Zeng, Z.; Yang, L.; Wang, C.; Chen, H.; Jiang, C.; Xie, R.; Wang, J.; Xie, X.; et al. 2023c. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. *arXiv preprint arXiv:2306.05087*.

Wu, M.; and Aji, A. F. 2023. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*.

Wu, Z.; Helaoui, R.; Reforgiato Recupero, D.; and Riboni, D. 2023. Towards Effective Automatic Evaluation of Generated Reflections for Motivational Interviewing. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, 368–373.

Zeng, G.; Yang, W.; Ju, Z.; Yang, Y.; Wang, S.; Zhang, R.; Zhou, M.; Zeng, J.; Dong, X.; Zhang, R.; et al. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9241–9250.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. BERTScore: Evaluating Text Generation with BERT. *arXiv preprint arXiv:1904.09675*.

Zhao, W.; Peyrard, M.; Liu, F.; Gao, Y.; Meyer, C. M.; and Eger, S. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

Zhao, Y.; Gu, A.; Varma, R.; Luo, L.; Huang, C.-C.; Xu, M.; Wright, L.; Shojanazeri, H.; Ott, M.; Shleifer, S.; et al. 2023. PyTorch FSDP: Experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2024. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36.

Zhong, M.; Liu, Y.; Yin, D.; Mao, Y.; Jiao, Y.; Liu, P.; Zhu, C.; Ji, H.; and Han, J. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.