Revisiting Benchmark and Assessment: An Agent-based Exploratory Dynamic Evaluation Framework for LLMs

Anonymous ACL submission

Abstract

While various vertical domain large language models (LLMs) have been developed, the challenge of automatically evaluating their performance across different domains remains significant. Current benchmark-based evaluation methods exhibit rigid, aimless interactions and rely on pre-collected static datasets that are costly to build, inflexible across domains, and misaligned with practical user needs. To address this issue, we revisit the evaluation components and introduce two concepts: **Benchmark+**, which extends traditional question-answer benchmark into a more flexible "strategy-criterion" format; and **Assessment+**, which enhances the interaction process, enabling deeper exploration and supporting both quantitative metrics and qualitative insights. These concepts capture the nuanced behaviors of LLMs through richer, multi-turn interactions. We propose an agent-based evaluation framework called TESTAGENT, which implements these concepts through retrieval augmented generation and reinforcement learning. Experiments on tasks ranging from constructing vertical domain evaluation to activating existing benchmarks demonstrate the effectiveness of TESTAGENT across various scenarios. We believe this work offers an interesting perspective on automatic evaluation for LLMs.

1 Introduction

Large language models (LLMs) (Brown et al., 2020; Bubeck et al., 2023) have demonstrated remarkable capabilities across a wide range of tasks, reshaping various industries and leading to the development of numerous domain-specific LLMs and applications (Bommasani et al., 2021; Huang et al., 2024). Nonetheless, due to the auto-regressive generation manner, LLMs exhibit high variability in the output. Effective evaluation can provide a better understanding of their capabilities and limitations (Peng et al., 2024; Guo et al., 2023).

Defining what constitutes "effective evaluation" requires us to listen to the voices of those involved in the LLMs industry (See Appendix D for details). To meet the evaluation challenge and understand practical demands and bottlenecks, we conducted online surveys, on-site visits, and document reviews on 50 companies in the LLMs industry.

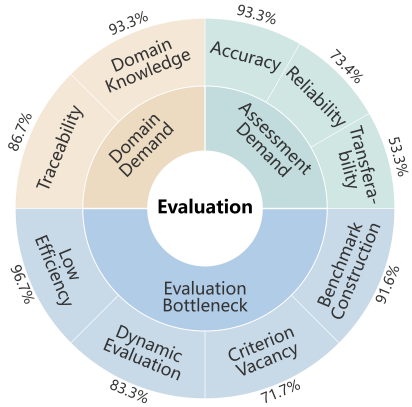

Figure 1: Statistical results of evaluation demand and bottleneck in industry.

As shown in Figure 1, 46 out of 50 companies (93.3%) with LLMs deployments focus on their products' domain expertise and knowledge capabilities, underscoring the crucial need to evaluate specialized knowledge in the LLMs industry. In terms of assessment demand, LLMs companies primarily concentrate on the accuracy, reliability, and transferability of their products. We further examined the current state of LLMs evaluation supply. Specifically, over 90% of evaluation organizations face the difficulty of efficiently constructing vertical domain benchmarks and ensuring datasets accurately reflect real-world scenarios. Additionally, 83.3% of companies emphasize the importance of dynamic evaluation techniques, and 71.7% highlight the lack of tailored criteria for different datasets and features. These bottlenecks hinder the evaluation infrastructure from meeting indus-

try demands, affecting product development and promotion. The findings confirm the urgency of efficient benchmark construction and assessment methods adapted to applications.

Existing evaluation methods are primarily based on benchmarks (Chang et al., 2024; Peng et al., 2024). These methods require LLMs to answer the questions in a predefined dataset and score the responses via predetermined metrics to quantify their capabilities. However, these approaches face certain limitations. Firstly, benchmarks for vertical domains are difficult to construct in advance, while building them manually on the spot can be costly (Raju et al., 2024). Secondly, traditional benchmarks typically present static question-answer formats, often involving only one (Huang et al., 2023; Hendrycks et al., 2020; Rajpurkar et al., 2016) or a few pre-collected rounds of dialogue (Zheng et al., 2023). These are far removed from real-world scenarios and fail to explore the model's capabilities fully (Bai et al., 2024). Finally, evaluation metrics are typically employed to provide numerical values that facilitate direct comparisons among similar models. Nevertheless, in vertical domains, the absence of comparable models creates uncertainty for testers. Testers are concerned with utilizing the metrics to uncover potential issues deeply rather than barely obtaining a quantitative number.

In this paper, we revisit two key aspects of LLMs evaluation: benchmark and assessment. Our contributions are threefold:

- **Benchmark+**: We introduce a generalized definition of benchmark, expanding the traditional "question-answer" pair into a broader "strategy-criterion" format. This positions conventional benchmarks as a specific case within this extended structure, allowing for a more comprehensive and nuanced evaluation.

- **Assessment+**: We enhance the assessment process by adopting a dynamic approach, shifting from static execution to purposeful exploration based on "Benchmark+". This approach facilitates a deeper investigation into the subtle and dynamic performance of LLMs, resulting in both quantitative and qualitative insights.

- **TESTAGENT**: We propose an agent-based evaluation framework that autonomously constructs "Benchmark+" using retrieval-augmented generation (RAG) and performs "Assessment+" through reinforcement learning (RL). This framework enables in-depth, dynamic evaluations and can also activate static benchmarks, allowing for interactive assessments based on existing benchmarks.

2 Methodology

In this section, we reflect on the evaluation components and introduce the improved concepts called Benchmark+ and Assessment+ in Section 2.1 and Section 2.2, respectively. These novel concepts are implemented through TESTAGENT framework, as illustrated in Figure 2. Furthermore, methods for activating existing static benchmarks are explored in Section 2.3.

2.1 Revisiting Benchmark

Commonly used benchmarks are typically composed of static "question-answer" pairs (Q, A), structured as: $\mathcal{B} := \{(Q^i, A^i)\}_{i=1}^N$, where N represents the total number of questions in the benchmark and i denotes the i-th pair. A wide range of benchmarks have been developed to evaluate the performance of LLMs on general tasks. These benchmarks typically require the target LLM to generate short responses to predefined questions, such as multiple-choice answers or brief text, following a relatively fixed evaluation process. However, general benchmarks are insufficient for vertical domains, as they fail to address the specific needs of the vertical areas. Furthermore, scenarios in vertical domains cannot be effectively assessed through a limited number of fixed interactions. For example, in medical consultation scenarios, developers require LLMs to handle multiple rounds of interactions with the user while maintaining a stable consultation strategy for each condition.

To address these challenges, we expand the form of benchmarks to "strategy-criterion" pairs, enhancing the benchmark's dynamism. Specifically, we define the benchmark as follows:

$$\mathcal{B}^*(D; T; M; K) := \{(S^i(\cdot), C^i(\cdot))\}_{i=1}^N,$$

where

$$\begin{aligned} S^i &= \{S_j^i \mid j = 1, 2, \cdots, n_i\}, \\ S_j^i &= \begin{cases} K(D; T) & \text{if } j = 0 \\ \Pr(\cdot | H_{j-1}^i; D; T) & \text{if } j > 0 \end{cases}, \\ C^i &= K(D; T; S_0^i). \end{aligned}$$

Here, S^i represents the strategy corresponding to the i-th element in $\mathcal{B}^*$. The strategy S^i is an abstraction of questions sequence, comprising n_i specific

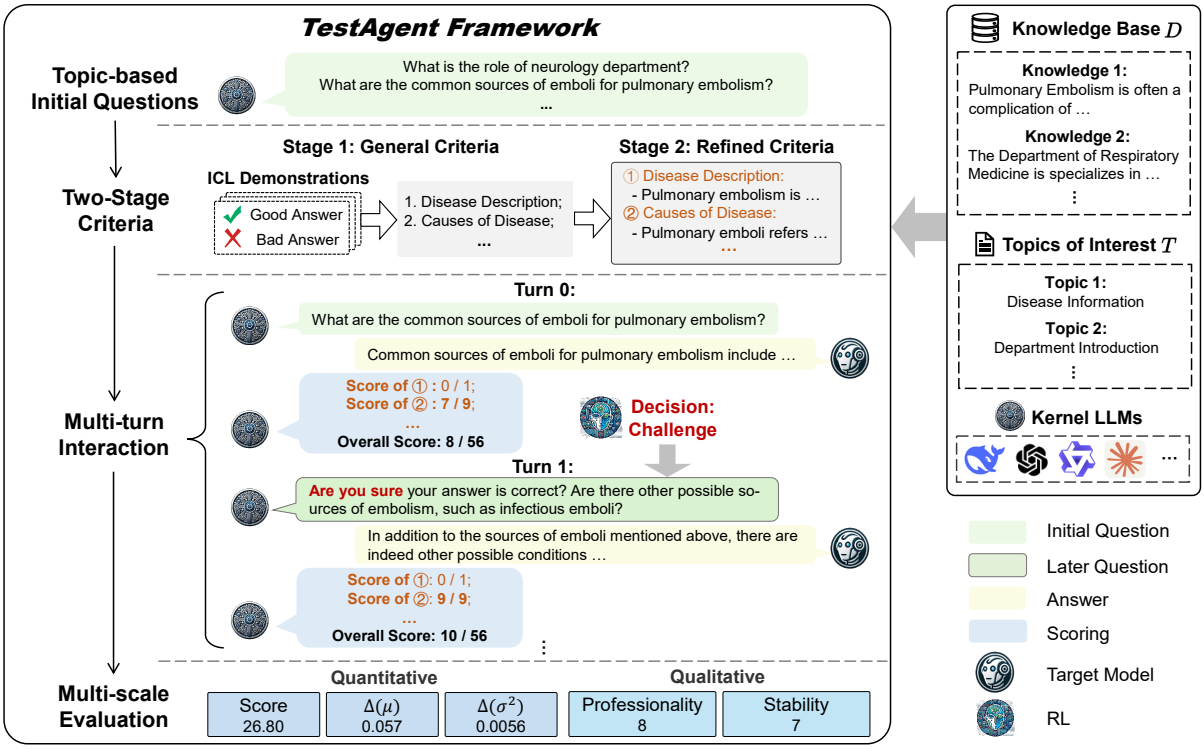

Figure 2: Illustration of our agent-based framework. TESTAGENT first generates initial questions and general topic-level criteria based on the knowledge base and interested topics. It subsequently refines the criteria into more specific question-level criteria. Reinforcement Learning then guides the subsequent multi-turn interactions, determining whether to challenge or follow up with additional questions. Multi-scale quantitative and qualitative evaluations are ultimately derived from these interactions.

question distributions S_j^i that are generated based on the interactions with the target model M. H_j^i denotes the interaction history up to time j, consisting of tuples of the form:

$$H_j^i = \{((Q_k^i, A_k^i, K(C^i; A_k^i))\}_{k=0}^{j}. \quad (1)$$

Here, question Q_k^i sampled from S_k^i, answer A_k^i generated from M, and $K(C^i; A_k^i)$ denotes corresponding evaluation.

Initially, the strategy S_0^i is generated by feeding the domain knowledge D and the user's topics of interest T into the evaluation model K (referred to as the kernel model). Subsequently, the step-wise question distribution $S_j^i (j \neq 0)$ is derived based on H_{j-1}^i, D, and T.

Note that when questions are predetermined and the criteria C are the ground truth, the strategy becomes deterministic, eliminating the need for dynamic generation. In this case, B^* degenerates into B, indicating that the existing benchmark B is a special case of B^*. In the following section, we detail the process of automatically constructing such benchmarks B^* in vertical domains. The complete pipeline is depicted in Figure 3.

2.1.1 Topic-based Initial Question Generation

In different domain scenarios, there may be expectations for LLMs to handle specific topics, such as "disease consultation" and "department introduction" topics in the medical scenarios. These specific demands are crucial for generating test cases that align with user needs in practical applications. Consequently, the current research question can be summarized as follows:

> **RQ1:** What should we discuss?

We employ RAG technology to address **RQ1**. First, a domain-specific knowledge base D will be fed into the kernel model K for understanding, allowing for the extraction of key knowledge points to form an RAG system. Then, we use prompt engineering to frame the user's topics of interest T, generating questions Q_0 that are more aligned with user concerns. These questions not only include direct inquiries but also encompass contextual scenarios, similar to role-playing, making them closer to real downstream task scenarios. The detailed prompt can be found in the Appendix G. Note that Q_0 only determines the initial question in a

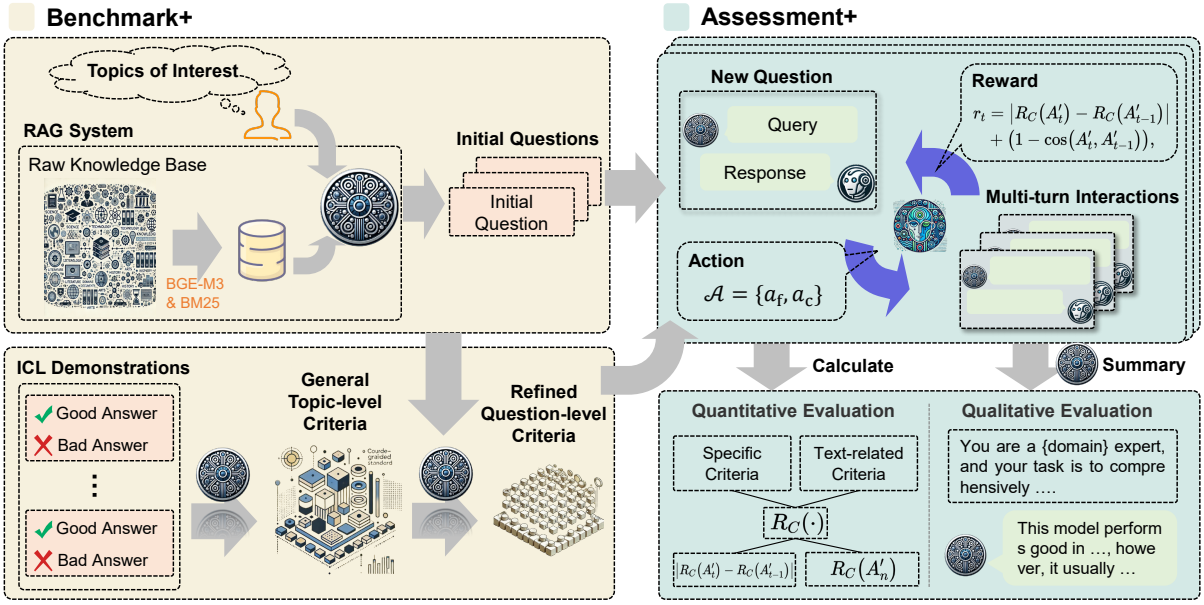

Figure 3: The pipeline of Benchmark+ and Assessment+ implementation. The left section depicts the construction of Benchmark+. It first automatically generates initial questions built upon the RAG system. Good and bad answer examples are provided to establish general criteria through in-context learning, and these are further refined into question-specific criteria for each generated question. The right section shows the Assessment+ pipeline, which incorporates RL-driven exploration and multi-scale evaluation. The RL-based interactions introduce a dynamic element to Benchmark+, enabling more adaptive and comprehensive assessments.

sequence, while subsequent questions need to be generated through the strategy. We will elaborate on the details in Section 2.2.1.

2.1.2 Two-Stage Criteria Generation

We also need to establish evaluation criteria. In other words, we must address the following research questions:

> **RQ2:** How should the question be evaluated?

We construct evaluation criteria C from a general topic to a question-refined level. It first locates the category of the topic, then searches and extracts the relevant documents for the question, thus helping to identify user-specific goals when interacting with the model. The motivation behind this procedure is to enhance the controllability of criterion generation and ensure that the question-relevant criteria align with the patterns of specific topics.

Reasonable criteria should include two aspects: the first relates to content-specific criteria, denoted as G, while the second pertains to text-related criteria E, such as fluency and whether the tone is human-like. Therefore, the pattern of the criteria can be formalized as follows:

$$C(D; T; S_0) \rightarrow \{G, E\}.$$

Stage 1: General topic-level Criteria. To establish general criteria on a given topic T, we design an in-context learning (ICL) prompt paradigm for kernel model K. For each topic, three demonstrations (including the question, good answer, and bad answer) are incorporated into the prompt. K is instructed to summarize the answer pattern for the topic. By presenting the domain topic and a few demonstrations to K, this approach extracts essential information, thereby ensuring the criterion is topic-specific and provides a more precise direction for refining the criteria further.

Stage 2: Refined question-level Criteria. With the general criteria established, the next goal is to compile refined, question-level criteria for automatic evaluation. Building on the general criteria, we prompt K to fill question-specific information into criteria using content retrieved by RAG. Specifically, first, RAG invokes the retriever to obtain documents corresponding to the question Q_0 from the domain knowledge base D. This retrieved content is then utilized in the prompt to assist K in filling in concrete details within the general criteria. Eventually, this paradigm develops clear and traceable criteria for various questions grounded in structure-uniformed topic general criteria C and relevant information sources. The example of our

prompt is shown in the Appendix G.

Metric. Based on the question and criteria, we can now evaluate the answer from the target model M. To achieve quantitative evaluation, we introduce an evaluation metric score R_C, defined as

$$R_C(A) := K(C; A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(G_i, A) + \frac{1}{m}\sum_{j=1}^{m} L(E_j, A),$$

where n is the total number of key information in criteria, m is the full marks of the textual characteristics index, G_i and E_j denotes the i-th content-specific criteria and the j-th text-related criteria, separately. A represents the answer from the target model M. $\mathbb{1}(x, y)$ is the indicator function, returning 1 if y includes x correctly and 0 otherwise, $L(x, y)$ denotes the integer score (from 1 to 10), evaluating how well y satisfies x based on the relevant context.

In practice, questions are generated dynamically following the strategy S^i, while the criteria remain consistent within an interaction sequence i. For each turn, kernel model K scores the answer from perspective in $\{G, E\} \in C$. Benefiting from the provision of refined criteria in a specific aspect before the evaluation, the model's scoring demonstrates reliability as shown in Appendix B.

2.2 Revisiting Assessment

General benchmarks assess target LLM's capabilities using specific metrics, such as accuracy in multiple-choice questions (MMLU (Hendrycks et al., 2020)), success rates in planning tasks (MINT (Wang et al., 2024b)), and general semantic aspects like coherence and relevance. Although these metrics are widely used, the target LLM M interacts with each test case only once or a few times, however, in real-world, users often often have multiple interactions with the LLMs around a topic. Therefore, these metrics fail to capture the deeper performance of M. In this section, we aim to implement multi-turn dialogues in assessment with a stronger emphasis on exploration and obtaining rich dialogue logs. Based on these dialogues, more evaluation metrics that fully capture the dynamic information can be developed.

2.2.1 RL-based Strategy Generator

As mentioned above, we attempt to address the following question:

> **RQ3:** How can we make the testing process more exploratory?

Initial questions have been generated in Section 2.1.1. The next step is to generate follow-up questions. Recall the definition of interaction history in (1): $H_n = \{(Q_k, A_k, K(C; A_k))\}_{k=0}^{n}$, where each question Q_{j+1} is sampled from the distribution $S_{j+1} = \Pr(\cdot|H_j; D; T)$, $(j = 0, \cdots, n)$, based on its interaction history. This process can be formulated as a sequential decision-making problem and trained using reinforcement learning (RL) with a policy π_θ, such that $Q_{j+1} := \pi_\theta(H_j; D; T)$.

However, treating raw text questions Q_{j+1} as actions would result in an excessively large action space. To facilitate the RL training, we attempt to generalize Q_{j+1} through a finite set of actions $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$, and define the policy as $\pi_\theta : \mathcal{H} \to \mathcal{A}$, where $\mathcal{H}$ is the set of all possible histories. As a result, the question is derived from the action space $\mathcal{A}$, knowledge base space $\mathcal{D}$ and topic space $\mathcal{T}$ via the kernel model mapping $k : \mathcal{A} \times \mathcal{D} \times \mathcal{T} \to \mathcal{Q}$, where $\mathcal{Q}$ is the space of possible queries. The overall function f is expressed as:

$$Q_{j+1} := f(H_j; D; T) = K(\pi_\theta(H_j); D; T),$$

which indicates that Q_{j+1} is generated by first selecting an action a through $\pi_\theta(H_j)$ and subsequently applying a to the kernel model K.

In practical scenarios, users may ask follow-up questions to target LLM M when additional information is needed. Conversely, if the M's responses are contradictory or incorrect, a natural behavior is to challenge M by asking why. To this end, we train an RL model to choose whether to ask follow-up questions or challenge the answers from the target model M. Now, we present the design of the state space, action space, and reward.

State. The state space is comprised of the following information:

1. **Question embedding.** To better capture the question type and the difficulty, which can influence decision-making, the initial question is embedded into the state representation instead of using the raw text directly.

2. **Score.** The score and its variation help assess the degree of knowledge acquisition. Therefore, we incorporate the current score, score variation, and the cosine similarity between consecutive responses into the state space.

Action. The action space $\mathcal{A}$ includes asking follow-up questions (a_{f}) and challenge further (a_{c}).
Reward. The reward is defined as

$$r_t = \left|R_C(A'_t) - R_C(A'_{t-1})\right| + \left(1 - \cos(A'_t, A'_{t-1})\right),$$

where A'_t and A'_{t-1} are the summarized current and previous responses, facilitating the comparison of information change. Intuitively, a large reward is assigned when more information is obtained from the target model or when the model produces contradictions, while a smaller reward is given in other cases. The goal is to extract the maximum information embedded within the target model, enabling rapid exploration of the model's boundaries.

We use the Proximal Policy Optimization (PPO) (Schulman et al., 2017) as the RL algorithm. The trained policy provides guidance for multi-turn interactive evaluation, addressing **RQ3**. RL policy offers dynamic alternative evaluation in a realistic paradigm rather than the static question-answer format of common benchmarks, enabling a more comprehensive assessment of model performance.

2.2.2 Qualitative and Quantitative Evaluation

The dynamic interaction with the target LLM M can be implemented with the strategy generator. Compared to existing evaluation procedures, such dynamic exploration allows us to delve deeper into a specific topic, thus performing more comprehensive qualitative and quantitative evaluations. Now, we need to address the following research question:

> **RQ4:** What insights can we derive from these records?

For quantitative evaluations, the last-turn score and the score fluctuation after the challenge question serve as a basis for assessing the professionalism and stability of the target LLM. To perform a deeper qualitative evaluation, we leverage the capabilities of the kernel model K to conduct a comprehensive evaluation of the target model, examining stability and professionalism based on historical dialogues. Thus, we address **RQ4** by providing a thorough evaluation that spans both structured metrics and unstructured performance aspects.

2.3 Activating Existing Benchmarks

For vertical domain scenarios, our approach has automated the processes of constructing test datasets, interaction, and evaluation. However, we still need to address the final research question:

> **RQ5:** How can we introduce dynamism to the existing benchmarks?

Existing benchmarks usually employ predefined metrics such as Exact Match, F1-Score, and so on to assess model accuracy. However, the flexible and varied expressions of LLMs make it challenging for these metrics to assess the alignment between model-generated long texts and ground-truth answers. These metrics may fail to capture nuanced correspondences and can even result in misjudgments. More importantly, the static dataset deviates from real-world scenarios, where users often pose multiple questions and engage in follow-up inquiries within the topic. Additionally, the benchmark lacks the ability to challenge responses, making it difficult to evaluate the stability of the LLMs's performance.

Dynamic evaluation via activating existing benchmarks can be achieved with minimal adjustments to our framework. Benefiting from the existing question-answer pairs in the benchmark, there is no need to generate additional questions or evaluation criteria. We still adopt the interactive method mentioned above, while the reward is determined by comparing model-generated answers with ground-truth answers in the benchmark to assess the correctness of the model's responses.

Specifically, we use the classic reading comprehension benchmark, SQuAD (Rajpurkar et al., 2016), as an example to demonstrate our dynamic extension process. In SQuAD, given a context paragraph, multiple independent questions and corresponding answers are provided. We autonomously select whether to ask a follow-up question or challenge the given answer, thereby enabling the integration of existing questions and the dynamic extension of the benchmark.

3 Experiments

In this section, we conduct experiments aimed at thoroughly evaluating the TestAgent. We explore both automatic evaluation in vertical domains from scratch and the activation of existing benchmarks in general domains. First, we conducted a meta-evaluation of the key components of TestAgent to validate its effectiveness. Subsequently, we report the main evaluation results across diverse models—comprising both open-source and closed-source options—spanning three different domains, demonstrating the practicality of TestAgent in

Table 1: Main Evaluation. Quantitative metrics: *Score* is the sum of the last turn's score across 50 conversations, with a maximum score of 50; $\Delta\ (\mu)$ and $\Delta\ (\sigma^2)$ denote the mean and the variance of score change after a challenge, respectively. Qualitative metrics: *Act.* denotes the accuracy; *Usr.* denotes the satisfaction of User needs; *Clar.* denotes the expressions Clarity; *Coh.* is the answer Coherence of the model; *Ques.* and *Chall.* measure the ability to handle follow-up questions and challenges, respectively. For each qualitative perspective, the Overall is its comprehensive evaluation. Scores for these qualitative metrics range from 1 to 10.

Metrics	**Quantitative**			**Professionalism (1-10)**					**Stability (1-10)**				
	Score ↑	$\Delta\ (\mu)$	$\Delta\ (\sigma^2)$	*Act.*	*Usr.*	*Clar.*	Avg.	Overall	*Coh.*	*Ques.*	*Chall.*	Avg.	Overall
Medical													
GPT-4o	**26.80**	0.057	0.0056	**7**	**6**	**9**	**7.3**	**7**	**8**	**7**	**7**	**7.3**	**7**
DeepSeek	26.68	0.067	0.0074	6	5	8	6.3	6	7	6	6	6.3	6
ZhongJing	20.67	0.056	0.0038	5	5	8	6.0	6	6	6	6	6.0	6
BianQue	14.09	0.056	0.0042	3	4	5	4.0	4	5	4	5	4.7	5
DoctorGLM	16.58	0.061	0.0051	3	4	6	4.3	4	4	3	4	3.7	4
Government													
GPT-4o	23.39	0.062	0.0037	4	5	6	5.0	5	5	6	5	5.3	5
DeepSeek	**23.42**	0.051	0.0050	**4**	**5**	**6**	**5.0**	**5**	**5**	**6**	**6**	**5.7**	**6**
Reading Comprehension													
GPT-4o	**48.93**	0.0047	0.0006	**10**	**9**	**9**	**9.3**	**9**	**10**	**10**	**10**	**10.0**	**10**
DeepSeek	48.70	0.0078	0.0014	10	9	9	9.3	9	10	9	9	9.3	9

assessing LLMs-powered applications. Additionally, an ablation study on the RL module is conducted, highlighting its trainability and the effectiveness of the designed state space.

3.1 Experiments Setup

Models and Domains. We consider government and medical tasks in the vertical domain and reading comprehension tasks in the general domain. To perform evaluation, we adopt DeepSeek-V2.5 (DeepSeek-AI, 2024) as the kernel model. The target model we assess in the medical domain including ZhongJing (Yang et al., 2024), Bianque (Chen et al., 2023), DoctorGLM (Xiong et al., 2023), DeepSeek-V2.5 and GPT-4o (Achiam et al., 2023). For the government and the reading comprehension tasks, we evaluate GPT-4o and DeepSeek-V2.5. In the RL module, the policy and value network are both 3-layer MLPs with 256-128 hidden states, trained for 6000 steps.

Knowledge Bases. In the medical domain, we collect common disease documents from OpenKG (Peng et al., 2021) and department introductions from CareGPT (Wang et al., 2023) to build the knowledge base. The disease consultation and department introduction topics are introduced. In the government domain, we extract the knowledge base from Chinese official fund management center websites (Center, 2024). It is classified into 8 fund-specific topics, which include individual, unit, establish, amend, draw, loan, repayment and inquiry. The raw knowledge is in TXT and JSON forms. We employ a hybrid approach that combines the BGE-M3 embedding model (Chen et al., 2024) with BM25. We adopt BCE-reranker (NetEase Youdao, 2023) to act on the retrieved documents.

3.2 Meta-Evaluation

We conducted meta-evaluations for both the benchmarks generated by TESTAGENT and the scoring of the kernel model. In different scenarios, we submitted interaction records to human experts for assessment, where they provided ratings on various performance dimensions. The benchmark+ and the scoring demonstrate a robust correlation with human preferences, underscoring the efficacy of automatic evaluation. Details are in Appendix B.

3.3 Domain Evaluation

We report the main result of vertical and general domain evaluations from quantitative and qualitative perspectives in Table 1. The quantitative metrics encompass three key aspects: last-turn score R_C summation (*Score*), as well as the mean $\Delta\ (\mu)$ and variance $\Delta\ (\sigma^2)$ of score fluctuations observed after the challenges. Meanwhile, the qualitative assessment

of model performance is based on two key dimensions: professionalism and stability. The professionalism dimension evaluates the model through factual accuracy (*Act.*), satisfaction of user needs (*Usr.*), and clarity of expression (*Clar.*). The stability dimension assesses the model's consistency (*Coh.*), its robustness in responding to challenges (*Chall.*), and its ability to handle follow-up questions (*Ques.*) effectively. The concrete description for the qualitative assessment is provided in Appendix G.

3.3.1 Vertical Domain

For different vertical domains, benchmarks can be rapidly constructed and dynamically evaluated by building a domain-specific knowledge base through the TESTAGENT framework. We examine two representative cases: the highly specialized domain of government affairs and the comparatively less specialized domain of medical healthcare.

Medical Domain As demonstrated in the first category of Table 1, GPT-4o achieves the highest *Score* (26.80) in the medical domain, indicating its superior domain expertise compared to other models. DeepSeek-V2.5 follows closely with a *Score* of 26.68, while BianQue and DoctorGLM underperform, with *Scores* of 14.09 and 16.58, respectively. Regarding professionalism, GPT-4o excels with an overall score of 7, reflecting relatively high factual accuracy, user need satisfaction, and clarity. However, BianQue and DoctorGLM struggle with professionalism and stability, indicating limitations in handling domain-specific queries and multi-turn interactions. Although ZhongJing, BianQue, and DoctorGLM are specialized medical models, their smaller model sizes and limited training data result in suboptimal performance in realistic, multi-turn medical dialogue scenarios.

Government Domain. In the government domain, DeepSeek-V2.5 marginally outperforms GPT-4o (23.42 vs. 23.39). Its variance (0.0050) in score changes is slightly higher than GPT-4o (0.0037), suggesting that DeepSeek-V2.5 may provide more diverse answers within this domain. Their professionalism and stability perform similarly, with DeepSeek-V2.5 showing a slight edge in handling challenges, suggesting it is more robust in government-related scenarios. This result aligns with expectations, as DeepSeek-V2.5 is recognized for its superior performance in Chinese language tasks, evidenced by its strong results in AlignBench (Liu et al., 2024), and the government knowledge base is built on the Chinese government platforms.

3.3.2 General Domain

In the reading comprehension domain, GPT-4o attains the highest *Score* of 48.93 and demonstrates significantly strong professionalism and stability, achieving an overall score of 9 and 10, separately. Specifically, it excels in factual accuracy, clarity, and strong coherence across multiple turns and robustness in handling follow-up questions and challenges. The minimal fluctuations ($\Delta\ (\mu) = 0.0047$, $\Delta\ (\sigma^2) = 0.0006$) when facing challenges, along with high scores in the *Chall.* metric, further indicating its stable responses. DeepSeek-V2.5 also performs well, with *Score* of 48.70 and overall professional and stable scores of 9. Both models perform exceptionally in this reading comprehension task compared to their performance in vertical domains, highlighting the models' effectiveness in broader, non-domain-specific tasks.

3.4 Ablation Study

The ablation study on the RL module is provided in Appendix A. The result emphasizes the role of the RL module and state space design in improving model performance across multiple domains, reinforcing the importance of RL for multi-turn decision-making. Additionally, the analysis of traditional evaluation metrics in Appendix C reveals the limitations of static benchmarks, reinforcing the superiority of our more intelligent evaluation. Case studies in Appendix F exemplify real-world applications, illustrating the nuanced interactions and decision points that arise during conversations.

4 Conclusion

In this paper, we revisit and extend two key components of existing LLMs evaluation, resulting in Benchmark+ and Assessment+. We introduced an agent-based evaluation framework called TESTAGENT to implement these concepts. It can automatically construct "strategy-criterion" pairs around domain knowledge, which form the extended benchmark, and autonomously perform exploratory interactions and assessment. Our experimental results demonstrate the framework's versatility and effectiveness. TESTAGENT streamlines benchmark generation and offers comprehensive LLMs assessment across domains, which provides insights into LLMs performance and helps developers build more reliable, domain-specific applications.

5 Limitaions

Concerning the scope of the domain, we currently handle text-based data for the benchmark construction, excluding the multi-modal data. Note that our proposed framework can still be applied when utilizing an off-the-shelf multi-modal large model as the kernel model, which sheds light on the more general multi-modal domain exploration.

Our method, while insightful, operates under the hypothesis that LLMs can accurately evaluate the model performance. The scoring of each response depends on the kernel model, which may raise concerns about the reliability of the metrics. Our approach generates criteria using a RAG-based process before scoring, rather than relying solely on direct scoring, facilitating the scoring precision improvement of the kernel model. To further enhance the credibility of the metrics, we will release a hybrid evaluation strategy that combines rule-based and semi-rule-based evaluation methods.

Currently, our action space in RL is limited to two types of interaction actions, which may constrain the diversity of interactions. A promising direction is to explore more varied interaction types that simulate real-world scenarios. In future work, we will explore more interaction forms, and delve into a deeper and more comprehensive evaluation of model performance.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *Preprint*, arXiv:2402.14762.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *Preprint*, arXiv:2108.07258.

T.B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Askell Amanda, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Henighan Tom, Rewon Child, A. Ramesh, DanielM. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, EricJ. Sigler, Mateusz Litwin, Scott Gray, Chess Benjamin, Jack Clark, Christopher Berner, McCandlish Sam, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *arXiv: Computation and Language,arXiv: Computation and Language*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Shanghai Provident Fund Management Center. 2024. Shanghai housing provident fund. https://shgjj.com/.

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.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2318–2335, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Yirong Chen, Zhenyu Wang, Xiaofen Xing, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, Xiangmin Xu, et al. 2023. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *Preprint*, arXiv:2310.15896.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *Preprint*, arXiv:2310.19736.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *Advances in Neural Information Processing Systems*.

Yining Huang, Keke Tang, and Meilian Chen. 2024. A comprehensive survey on evaluating large language model applications in the medical industry. *Preprint*, arXiv:2404.15777.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Weize Kong, Spurthi Amba Hombaiah, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Prewrite: Prompt rewriting with reinforcement learning. *Preprint*, arXiv:2401.08189.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *Preprint*, arXiv:2401.16745.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *Preprint*, arXiv:2306.09212.

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

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Andrew Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Xiaotao Gu, Hongning Wang, Jing Zhang, Minlie Huang, Yuxiao Dong, and Jie Tang. 2024. AlignBench: Benchmarking Chinese alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Inc. NetEase Youdao. 2023. Bcembedding: Bilingual and crosslingual embedding for rag. https://github.com/netease-youdao/BCEmbedding.

Chen Peng, Jun Zhang, Yanchao Xu, and Jianan Yang. 2021. Diseasekg. http://data.openkg.cn/dataset/disease-information.

Ji-Lun Peng, Sijia Cheng, Egil Diau, Yung-Yu Shih, Po-Heng Chen, Yen-Ting Lin, and Yun-Nung Chen. 2024. A survey of useful llm evaluation. *Preprint*, arXiv:2406.00936.

Zhenting Qi, Xiaoyu Tan, Shaojie Shi, Chao Qu, Yinghui Xu, and Yuan Qi. 2023. Pillow: Enhancing efficient instruction fine-tuning via prompt matching. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 471–482.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Ravi Raju, Swayambhoo Jain, Bo Li, Jonathan Li, and Urmish Thakker. 2024. Constructing domain-specific evaluation sets for llm-as-a-judge. *Preprint*, arXiv:2408.08808.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, and Adrià Garriga-Alonso. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Hao Sun, Alihan Hüyük, and Mihaela van der Schaar. 2023a. Query-dependent prompt evaluation and optimization with offline inverse rl. In *The Twelfth International Conference on Learning Representations*.

Yuchong Sun, Che Liu, Jinwen Huang, Ruihua Song, Fuzheng Zhang, Di Zhang, Zhongyuan Wang, and Kun Gai. 2023b. Parrot: Enhancing multi-turn chat models by learning to ask questions. *Preprint*, arXiv:2310.07301.

Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024a. A user-centric benchmark for evaluating large language models. *Preprint*, arXiv:2404.13940.

Rongsheng Wang, Ruizhe Zhou, Haoming Chen, Yapeng Wang, and Tao Tan. 2023. Caregpt: Medical llm, open source driven for a healthy future. https://github.com/WangRongsheng/CareGPT.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2024b. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. In *International Conference on Learning Representations*, pages 1–35.

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

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19368–19376.

Tianjun Zhang, Xuezhi Wang, Denny Zhou, Dale Schuurmans, and Joseph E Gonzalez. 2023. Tempera: Test-time prompt editing via reinforcement learning. In *The Eleventh International Conference on Learning Representations*.

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. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

A Ablation Study

We investigate the impact of the trainable decision-making RL module and its components in TestAgent through an ablation study, with the results shown in Table 2. Specifically, we compare three variants of TestAgent: TestAgent (Skeleton), which refers to the model framework before RL training; TestAgent (noE), where the question embedding is removed from the state space; and TestAgent (Vanilla) is our framework without RL module, which decide the follow-up question or challenge merely with prompt.

The evaluation metrics focus on the average return of a conversation after RL training, the model challenge completion rate (Challenge Ratio), and the model follow-up question completion rate (Question Ratio). Specifically, the model is expected to challenge when the score change is small, indicating insufficient new information, and to ask follow-up questions when the score change is large, signifying relevant information has been obtained. The actual number of challenges (or follow-up questions) divided by the expected number yields the corresponding ratios.

Table 2: Ablation Study of RL components.

Metrics	Embed	Train	Return	Challenge Ratio	Question Ratio
Government					
`TestAgent`	✔	✔	**0.64**	**0.66**	0.61
`TestAgent (Skeleton)`	✔	✗	0.46	0.60	0.52
`TestAgent (noE)`	✗	✔	0.61	0.58	0.56
`TestAgent (Vanilla)`	-	-	-	0.00	**1.00**
Medical					
`TestAgent`	✔	✔	**0.65**	**0.67**	**0.69**
`TestAgent (Skeleton)`	✔	✗	0.56	0.45	0.64
Reading Comprehension					
`TestAgent`	✔	✔	**0.72**	**0.56**	**0.56**
`TestAgent (Skeleton)`	✔	✗	0.57	0.44	0.52

RL Module. The results demonstrate that reinforcement learning (RL) training enhances the evaluation performance of TestAgent. In the government domain, the fully implemented TestAgent achieves the highest return of 0.64, outperforming both TestAgent (Skeleton), which achieves 0.46, and TestAgent (noState), which reaches 0.61. This trend is consistent across the medical and reading comprehension domains, where TestAgent also records higher scores after RL training. Additionally, the Challenge and Question Ratios show notable improvements as a result of RL training. It is important to note, however, that TestAgent (Vanilla), which relies solely on the LLMs without the RL module, demonstrates a perfect score in the Question Ratios. However, this performance stems from mechanical behavior, as it consistently adopts a follow-up questioning strategy (Challenge Ratio = 1) without adapting to the quality of the model's responses. This rigid approach prevents TestAgent (Vanilla) from exploring the performance boundaries of the target model, let alone providing a comprehensive evaluation of stability. These findings highlight the crucial role of reinforcement learning in facilitating more effective decision-making processes, enabling the system to flexibly determine when to question or challenge the target model during multi-turn conversations, thereby offering a deeper and more nuanced evaluation.

State Space. We explore the impact of state space design on RL performance using the government domain as an example. The structure of the state space plays an important role in training efficacy. Removing the question embedding from the state space, as in TestAgent (noE), leads to a notable performance decline. The return drops from 0.64 in the fully implemented TestAgent to 0.61 in TestAgent (noE). Similarly, the Challenge and Question ratios also show reduced performance without state-space design. This highlights that incorporating question embeddings into the state space is essential

for maintaining coherence and robustness when handling complex, multi-turn interactions.

B Meta-Evaluation

Meta-Evaluation on Benchmark+. We randomly sample 50 sets of conversations from each domain, and five human experts independently rate the generated questions and criteria across the following five dimensions on a scale from 1 to 5: *Relevancy* evaluates whether the generated questions are domain-appropriate and aligned with the specific topic under analysis. *Scenarios* evaluates how well the questions are grounded in realistic user contexts, reflecting practical, situation-based use cases. *Coherence* focuses on the thematic consistency across multiple turns in a dialogue. *Completeness* gauges whether the generated criteria cover the key aspects to answer the generated questions. Lastly, *Logic* measures the logical structure and reasoning within the generated criteria.

Table 3: Alignment of generated Questions and Criteria with Human Evaluation. Five human evaluators scored the generated content on a scale of 1 to 5. For each metric, we report the mean score (μ) and variance (σ^2), representing the evaluators' assessment of the content quality across different dimensions.

Metrics	**Questions**						**Criteria**			
	Relevancy		*Scenarios*		*Coherence*		*Completeness*		*Logic*	
	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
Human	4.93	0.0014	4.80	0.013	4.90	0.0035	4.79	0.014	4.85	0.0076

The average scores and variance for each metric are reported in Table 3. The results demonstrate that the generated questions are highly relevant to the domains under consideration, with an average *Relevancy* score of 4.93 and a minimal variance of 0.0014. Furthermore, the questions are well-aligned with real-user scenarios and remain strong coherence with the specific topics in multi-turn interaction. Regarding evaluation criteria, the *Completeness* score of 4.79 indicates TESTAGENT's capacity to extract and formulate relevant criteria that effectively cover key aspects required to address the generated questions. The *Logic* score of 4.85, accompanied by a variance of 0.0076, further confirms the logical structure of the criteria. The consistently high average scores across all metrics, combined with the small variances, indicate the reliability and precision of the TESTAGENT system in generating both questions and criteria that are well-aligned with human judgment, reinforcing the system's robustness and effectiveness in dynamic benchmark generation.

Meta-Evaluation on Kernel Model. We conducted experiments using GPT-4o and DeepSeek-V2.5 as kernel models, collecting 120 topics in total (60 from GPT-4o and 60 from DeepSeek-V2.5), resulting in 600 rounds of dialogue (5 rounds per interaction). Three human experts independently evaluated the scores given by kernel models for these dialogues, assigning a score of 1 if they agreed with the model's rating and 0 otherwise. We measured the Inter-Annotator Agreement (IAA) by averaging Cohen's Kappa coefficients for each annotator pair in Table 5, yielding an average IAA of 0.820. This indicates the near-perfect agreement among the annotators. The final evaluation results are shown in Table 4.

Table 4: Model Scoring Performance. μ and σ^2 denote the average score and the variance of human satisfaction, respectively.

Model	μ	σ^2
GPT-4o	0.974	0.025
DeepSeek-V2.5	0.957	0.040

Table 5: Cohen's Kappa (κ) for Annotators.

Annotator Pair	κ
A + B	0.792
B + C	0.812
A + C	0.850

From the results, we observe that GPT-4o achieves high human satisfaction (μ=0.974) with stable performance (σ^2=0.025), i.e., it produces more accurate scores for the response. DeepSeek-V2.5 performs slightly lower, with a satisfaction score of 0.957. Overall, both kernel models demonstrate relatively high reliability, confirming the validity of our evaluation approach.

C Difference Assessment

We compute Pearson, Spearman, and Kendall-Tau correlation coefficients to gauge the difference between TESTAGENT's scores and conventional metrics ROUGE (Lin, 2004). As shown in Table 6, TESTAGENT exhibits a notable difference from the ROUGE metric, highlighting its effectiveness in evaluating dynamically generated, open-ended conversations. ROUGE metric often errs when measuring the similarity between long sentences and standard answers, While TESTAGENT demonstrates a robust capability in handling lengthy texts, accurately assigning scores to reflect correctness. This explains the observed differences in their correlation coefficients.

Table 6: Pearson, Spearman, Kendall-Tau correlation coefficients of the score assigned by TESTAGENT with ROUGE and their variance.

Metrics	Pearson	Spearman	Kendall-Tau	variance
ROUGE	0.677	0.669	0.628	0.136

D Evaluation Subjects

Human. All of our experts are researchers with a minimum of a master's degree in fields related to artificial intelligence. They voluntarily participated in evaluating our work and reached a consensus on the intended use of the data.
For the meta-evaluation of Benchmark+, experts were provided with the following guidelines: "You are presented with multi-turn questions and corresponding evaluation criteria generated for the {domain} domain. Please them on a scale from 1 to 5 across the following five dimensions: 1. Relevancy: Assess whether the generated questions are appropriate for the domain and align with the specific topic being analyzed. 2. Scenarios: Evaluate how well the questions reflect practical, real-world situations and user contexts relevant to the domain. 3. Coherence: Consider the consistency of the questions across multiple turns within a dialogue, ensuring they maintain a clear thematic connection. 4. Completeness: Determine if the criteria sufficiently cover all the key aspects needed to address the generated questions. 5. Logic: Evaluate the logical flow and reasoning presented in the criteria, ensuring they are structured in a sound and coherent manner. Please use this scale to rate each dimension, ensuring your feedback is clear and consistent across all evaluations."
For the meta-evaluation of the kernel model, each expert received the following instructions: "You are presented with several conversations between a target model and a kernel model, where the kernel model scores the target model. Please assess the reasonableness and stability of the kernel model's scoring and assign a score of 1 if you are satisfied with the scoring, or 0 if you are not."

Industry. The LLMs industry chain is divided into three layers: the infrastructure layer, the model and tools layer, and the application layer. Our survey covers 16 companies in the infrastructure layer, with leading enterprises including Shanghai Supercomputer Center, Tencent Yangtze River Delta Artificial Intelligence Advanced Computing Center, Alibaba Cloud East China Intelligent Computing Center, and Huawei Software Technologies Co., Ltd. The model and tools layer includes 20 companies, with key players such as Shanghai SenseTime Intelligent Technology Co., Ltd., Shanghai Artificial Intelligence Laboratory, Shanghai Xiyu Technology Co., Ltd., and Huawei Software Technologies Co., Ltd. The application layer consists of 34 companies, with leading enterprises including Transwarp Technology (Shanghai) Co., Ltd., iFLYTEK Co., Ltd., Shanghai Kuanyu Digital Technology Co., Ltd., Alibaba Network Technology Co., Ltd., and Silicon Intelligence Technology Co., Ltd.

E Related Works

General Evaluation. Various benchmarks have been proposed to evaluate the broad abilities of large language models (LLMs). For knowledge, MMLU (Hendrycks et al., 2020), C-Eval(Huang et al.,

2023) and CMMLU (Li et al., 2023) are commonly used, while comprehension can be tested through TriviaQA (Joshi et al., 2017) and SQuAD (Rajpurkar et al., 2016). Mathematical reasoning abilities are assessed with GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021), and coding skills are evaluated through the HumanEval (Chen et al., 2021) dataset, which focuses on code generation tasks. BIG-Bench (Srivastava et al., 2022) is a broad benchmark for general tasks, covering diverse topics, such as linguistics, childhood development, math, common-sense reasoning, biology, physics, social bias, software development, etc. While the above benchmarks provide a comprehensive assessment of LLMs performance in general domains, they are still confined to predefined domains. In evaluating vertical domain systems, it faces limitations including the high cost of dataset construction and criterion development. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) has demonstrated a strong ability to provide up-to-date and helpful external knowledge, making it widely adopted across various tasks to support generative AI in producing high-quality outputs (Fan et al., 2024). In this paper, we leverage RAG to retrieve and utilize domain-specific knowledge bases, thereby enabling RAG to contribute to the automatic construction of tailored benchmarks. Our approach enables the agile design of benchmarks, effectively addressing the challenges encountered in vertical domains.

Multi-turn Evaluation. MT-bench (Zheng et al., 2023) is the first benchmark build for multi-turn instruction-following evaluation, which contains two-turn questions for each dialogue. MT-Bench++ (Sun et al., 2023b) extent MT-bench to an eight-turn evaluation dataset with six manually written instructions. To better capture the complexity of real-world multi-turn dialogue scenarios, MT-bench-101 (Bai et al., 2024) formulated a three-tier hierarchical ability taxonomy and used GPT-4o to collect questions focused on these abilities. Similarly, MT-Eval (Kwan et al., 2024) constructs multi-turn queries by augmenting existing datasets and creating new examples with GPT-4o. However, these benchmarks overlook real-world user interactions. To address this, Wang et al. (2024a) collected authentic feedback from real users, forming a multi-cultural, multi-LLMs user-reported scenario dataset. MINT (Wang et al., 2024b) was developed to evaluate LLMs' task-solving abilities in multi-turn interactions, using tools and natural language feedback. Despite these efforts, such approaches do not fully reflect practical multi-turn dialogues, where users engage in a question-challenge manner. In this work, we focus on automatically generating follow-up questions that mimic real user interactions, enabling dynamic benchmark creation and a comprehensive exploration of the model's capabilities.

RL for LLMs. Recent research has explored the use of reinforcement learning (RL) in prompt optimization for LLMs. RLPROMPT (Deng et al., 2022) presents an efficient approach to optimize discrete prompts without the need for gradient information. TEMPERA (Zhang et al., 2023) introduces a novel action space for flexible prompt editing, adapting to different queries. PILLOW (Qi et al., 2023) utilizes RL to enhance the performance of Low-Rank Adaptation by intelligently selecting and concatenating prompts from a user-defined pool, reducing computational costs. PRewrite (Kong et al., 2024) automates the rewriting of prompts via RL for improved performance on downstream tasks. Prompt-OIRL (Sun et al., 2023a) utilizes offline inverse reinforcement learning to optimize prompts economically, demonstrating the potential of RL in prompt engineering for LLMs. Inspired by such ideas, we frame the user-centered benchmark construction as a question-challenge decision with an RL-based framework to perform automatic prompt selection in multi-turn question generation.

F Case Study in Medical Domain

The following cases illustrate the evaluation process in detail, demonstrating that TestAgent can effectively generate benchmarks for vertical domains and perform automatic assessment that meets realistic needs.

Initial Questions for Disease Topic

1. What are the common symptoms of decompression sickness?
2. What are the main clinical manifestations of SARS?
3. What are the emergency treatment methods for allergic shock?

4. What are the arterial blood gas analysis criteria for respiratory failure?
5. What are the common symptoms of tuberculous empyema?
6. How can the occurrence of septic shock be prevented?
7. What are the main causes of sick building syndrome?
8. How can the occurrence of respiratory foreign bodies in children be prevented?

General Criterion for Disease Topic

1. **Disease Name:**
- Clearly state the full name of the disease, including common aliases.
2. **Disease Description:**
- Provide a concise and detailed definition or background information, covering causes, symptoms, and related important features.
3. **Disease Classification:**
- Mention the medical classification of the disease, such as internal medicine, respiratory medicine, pediatrics, etc.
4. **Preventive Measures:**
- List feasible and effective preventive methods.
5. **Causes of Illness:**
- Clarify the main pathogenic factors and related triggers.
6. **Pathogenesis:**
- Briefly explain the biological mechanism or pathological process of the disease.
7. **Symptoms:**
- List specific and common symptoms or signs.
8. **Insurance Status:**
- Indicate whether the disease is covered by health insurance.
9. **Incidence Probability:**
- Provide data on the incidence or probability of occurrence of the disease.
10. **Susceptible Population:**
- Identify high-risk groups or those prone to the disease.
11. **Transmission Method:**
- Explain the main transmission routes of the disease; if non-contagious, specify as such.
12. **Comorbid Conditions:**
- Mention other health problems or complications that may coexist with this disease.
13. **Consultation Department:**
- Recommend the medical department patients should visit.
14. **Treatment Methods:**
- Describe the main treatment approaches, including both pharmacological and non-pharmacological treatments.
15. **Treatment Duration:**
- Provide an approximate time range for treatment.
16. **Cure Probability:**
- Provide data on the cure rate for the disease.
17. **Common Medications:**
- List common medications used to treat this disease.
18. **Treatment Costs:**
- Provide an estimated cost range for treating this disease.
19. **Diagnostic Tests:**
- List possible examinations and tests needed for diagnosing this disease.
20. **Recommended Diet:**
- Provide a list of foods suggested for patients.

21. **Prohibited Foods:**
- Provide a list of foods patients should avoid.
22. **Recommended Recipes:**
- Provide one or two suitable recipes or food combinations for patients.
23. **Recommended Medications:**
- List specific brands or manufacturers of recommended medications.
24. **Detailed Medication Information:**
- Provide detailed information on some recommended medications, including drug names and manufacturers.

Refined Criterion for Question: "What are the emergency treatment methods for allergic shock?"

1. **Disease Name:**
- Clearly state the full name of the disease: Anaphylactic Shock
2. **Disease Description:**
- Provide a concise and detailed definition or background information: Anaphylactic shock is a systemic rapid allergic reaction primarily characterized by acute peripheral circulation insufficiency, triggered by specific allergens that are generally harmless to the human body in patients with an allergic constitution.
- Cover causes, symptoms, and related important features: Causes include antibiotics, local anesthetics, vitamins, etc.; symptoms include skin itching, hives, difficulty breathing, chest tightness, coughing, abdominal pain, nausea, vomiting, dizziness, and pallor.
3. **Disease Classification:**
- Mention the medical classification of the disease: Medical Encyclopedia, Emergency Medicine
4. **Preventive Measures:**
- List feasible and effective preventive methods: Consult detailed allergy history before medication, minimize unnecessary injections, prefer oral formulations, and observe allergic patients for 15–20 minutes after injections. Conduct skin prick tests to avoid using medications that may yield positive results.
5. **Causes of Illness:**
- Clarify the main pathogenic factors and related triggers: Endocrine substances, enzymes, pollen extracts, food, antisera, occupational exposure proteins, insect venoms, etc.
6. **Pathogenesis:**
- Briefly explain the biological mechanism or pathological process: External antigenic substances entering the body can stimulate the immune system to produce corresponding antibodies. The production of IgE varies significantly among individuals due to genetic differences. These specific IgE have a strong affinity for target cells in the skin, bronchi, blood vessel walls, etc. When the same antigen contacts a sensitized individual again, it can trigger a widespread Type I hypersensitivity reaction.
7. **Symptoms:**
- List specific and common symptoms or signs: Pale lips, fainting, altered consciousness, hives, consciousness disorders, palpitations, hypotension, etc.
8. **Insurance Status:**
- Indicate whether the disease is covered by health insurance: No
9. **Incidence Probability:**
- Provide data on the incidence or probability of occurrence of the disease: 0.001%
10. **Susceptible Population:**
- Identify high-risk groups or those prone to the disease: Individuals with an allergic constitution
11. **Transmission Method:**
- Explain the main transmission routes of the disease; if non-contagious, specify as such: Non-contagious

12. **Comorbid Conditions:**
- Mention other health problems or complications that may coexist with this disease: Edema
13. **Consultation Department:**
- Recommend the medical department patients should visit: Emergency Department
14. **Treatment Methods:**
- Describe the main treatment approach: Immediate on-site rescue upon occurrence
15. **Treatment Duration:**
- Provide an approximate time range for treatment: Depends on the specific situation
16. **Cure Probability:**
- Provide data on the cure rate for the disease: 80%
17. **Common Medications:**
- List common medications used to treat this disease: Diphenhydramine hydrochloride injection, Naloxone hydrochloride injection, Epinephrine hydrochloride injection, Dobutamine glucose injection, Hydrocortisone injection, Dextrose 40 injection, Sodium hydrocortisone succinate for injection
18. **Treatment Costs:**
- Provide an estimated cost range for treating this disease: Approximately (5000-10000 yuan) at a city-level top-tier hospital
19. **Diagnostic Tests:**
- List possible examinations and tests needed for diagnosing this disease: Electrocardiogram, Free heparin time, Complete blood count, Blood biochemical panel, Renal function tests, Intradermal enzyme tests, Heparin-antithrombin IV complex, Streptomycin intradermal test, Central venous pressure measurement (CVP)
20. **Recommended Diet:**
- Provide a list of foods suggested for patients: Eggs, quail eggs, sesame seeds, pumpkin seeds
21. **Prohibited Foods:**
- Provide a list of foods patients should avoid: Honey, white lentils, crabs, lard (rendered fat)
22. **Recommended Recipes:**
- Provide one or two suitable recipes or food combinations for patients: None
23. **Recommended Medications:**
- List specific brands or manufacturers of recommended medications: Suicheng Pharmaceutical Diphenhydramine hydrochloride injection, Fuan Pharmaceutical Group Sodium hydrocortisone succinate for injection, Shiyao Silver Lake Epinephrine hydrochloride injection, Wuzhong Pharmaceutical Naloxone hydrochloride injection, Jinyao Pharmaceutical Epinephrine hydrochloride injection, Beijing Yongkang Pharmaceutical Epinephrine hydrochloride injection, Sunuo Naloxone hydrochloride injection, Haiwangfuyuan Dobutamine glucose injection, Aierhaitai Naloxone hydrochloride injection, Disainuo Pharmaceutical Naloxone hydrochloride injection, Changzheng Fumin Jingshan Dextrose 40 injection, Jiangsu Yuanheng Dextrose 40 injection, Yunchong Pharmaceutical Dextrose 40 injection, Tianfeng Pharmaceutical Dextrose 40 injection, Yuandong Pharmaceutical Naloxone hydrochloride injection, Beijing Yongkang Pharmaceutical Diphenhydramine hydrochloride injection, Lijun Pharmaceutical Hydrocortisone injection, Tianyi Qinkun Dextrose 40 injection, Baiyunshan Mingxing Epinephrine hydrochloride injection, Yue Pharmaceutical Dextrose 40 injection, Dikan Changjiang Pharmaceutical Epinephrine hydrochloride injection, Chengdu Tiantai Mountain Pharmaceutical Naloxone hydrochloride injection, Yonghe Pharmaceutical Dextrose 40 injection
24. **Detailed Medication Information:**
- Provide detailed information on some recommended medications, including drug names and manufacturers: Suicheng Pharmaceutical Diphenhydramine hydrochloride injection, Fuan Pharmaceutical Group Sodium hydrocortisone succinate for injection, Shiyao Silver Lake Epinephrine hydrochloride injection, Wuzhong Pharmaceutical Naloxone hydrochloride injection, Jinyao Pharmaceutical Epinephrine hydrochloride injection, Beijing Yongkang Pharmaceutical Epinephrine hydrochloride

injection, Sunuo Naloxone hydrochloride injection, Haiwangfuyuan Dobutamine glucose injection, Aierhaitai Naloxone hydrochloride injection, Disainuo Pharmaceutical Naloxone hydrochloride injection, Changzheng Fumin Jingshan Dextrose 40 injection, Jiangsu Yuanheng Dextrose 40 injection, Yunchong Pharmaceutical Dextrose 40 injection, Tianfeng Pharmaceutical Dextrose 40 injection, Yuandong Pharmaceutical Naloxone hydrochloride injection, Beijing Yongkang Pharmaceutical Diphenhydramine hydrochloride injection, Lijun Pharmaceutical Hydrocortisone injection, Tianyi Qinkun Dextrose 40 injection, Baiyunshan Mingxing Epinephrine hydrochloride injection, Yue Pharmaceutical Dextrose 40 injection, Dikan Changjiang Pharmaceutical Epinephrine hydrochloride injection, Chengdu Tiantai Mountain Pharmaceutical Naloxone hydrochloride injection, Yonghe Pharmaceutical Dextrose 40 injection.

Below, we present the 6-turn interaction starting from the initial question "What are the emergency treatment methods for allergic shock?"

——————- Turn 0 ——————-

Initial Question: What are the emergency treatment methods for allergic shock?

Full marks: 1. Disease Name: 1 point; 2. Disease Description: 2 points; 3. Disease Classification: 1 point; 4. Preventive Measures: 1 point; 5. Causes of Disease: 1 point; 6. Pathogenesis: 1 point; 7. Symptoms: 1 point; 8. Insurance Status: 1 point; 9. Incidence Probability: 1 point; 10. Vulnerable Population: 1 point; 11. Transmission Method: 1 point; 12. Comorbidities: 1 point; 13. Consultation Department: 1 point; 14. Treatment Method: 1 point; 15. Treatment Duration: 1 point; 16. Cure Probability: 1 point; 17. Common Medications: 1 point; 18. Treatment Cost: 1 point; 19. Examination Items: 1 point; 20. Suitable Diet: 1 point; 21. Forbidden Diet: 1 point; 22. Recommended Recipes: 1 point; 23. Recommended Medications: 1 point; 24. Medication Details: 1 point.

Answer: Allergic shock is a severe allergic reaction that usually requires emergency treatment. Here are the common emergency treatment methods for allergic shock: Stay calm, have the patient lie flat, loosen the collar and belt, and ensure the airway is clear. Immediately administer intravenous epinephrine to dilate blood vessels and increase blood pressure. Monitor vital signs such as blood pressure, heart rate, and respiration, and record the time for the doctor's diagnosis and treatment. If the patient has difficulty breathing, perform artificial respiration and initiate cardiopulmonary resuscitation. Transport the patient to the hospital for treatment, including fluid and blood transfusion support. It is important to note that if the patient experiences severe allergic reactions, such as difficulty breathing, rapid heartbeat, or confusion, call emergency services or take them to the hospital immediately.

Concluded Answer: Allergic shock is a severe allergic reaction that requires emergency treatment. Common emergency treatment methods include: Stay calm, have the patient lie flat, loosen the collar and belt, and ensure the airway is clear. Immediately administer intravenous epinephrine to dilate blood vessels and increase blood pressure. Monitor the patient's blood pressure, heart rate, and respiration, and record the time for the doctor's diagnosis and treatment. If the patient has difficulty breathing, perform artificial respiration and initiate cardiopulmonary resuscitation. Transport the patient to the hospital for treatment, including fluid and blood transfusion support. It is important to note that if the patient experiences severe allergic reactions, such as difficulty breathing, rapid heartbeat, or confusion, call emergency services or take them to the hospital immediately.

Score: 1. Disease Name: 1/1; 2. Disease Description: 0/2; 3. Disease Classification: 0/1; 4. Preventive Measures: 0/1; 5. Causes of Disease: 0/1; 6. Pathogenesis: 0/1; 7. Symptoms: 0/1; 8. Insurance Status: 0/1; 9. Incidence Probability: 0/1; 10. Vulnerable Population: 0/1; 11. Transmission Method: 0/1; 12. Comorbidities: 0/1; 13. Consultation Department: 0/1; 14. Treatment Method: 1/1; 15. Treatment Duration: 0/1; 16. Cure Probability: 0/1; 17. Common

Medications: 0/1; 18. Treatment Cost: 0/1; 19. Examination Items: 0/1; 20. Suitable Diet: 0/1; 21. Forbidden Diet: 0/1; 22. Recommended Recipes: 0/1; 23. Recommended Medications: 0/1; 24. Medication Details: 0/1.
Total Score: 2/25.
————————- Turn 1 ————————-
Follow-up Question: What are the causes of allergic shock?
Answer: The causes of allergic shock may be due to allergic reactions, which can include: drug allergies (certain medications such as antibiotics, hormones, anticoagulants, non-steroidal anti-inflammatory drugs, etc.), food allergies (such as eggs, peanuts, milk, etc.), environmental factors (such as climate changes, air pressure changes, pollen, etc.), and psychological factors (such as emotional fluctuations, anxiety, stress, etc.). In summary, allergic shock is a serious allergic reaction that requires emergency treatment.
Concluded Answer: Allergic shock is a severe allergic reaction that usually requires emergency treatment. Common emergency treatment methods include: remain calm, have the patient lie flat, loosen the collar and belt to ensure the airway is clear; immediately administer intravenous epinephrine to dilate blood vessels and increase blood pressure; monitor vital signs such as blood pressure, heart rate, and respiration, and record the time for the doctor's diagnosis and treatment. If the patient has difficulty breathing, perform artificial respiration and initiate cardiopulmonary resuscitation. Transport the patient to the hospital for treatment, including fluid and blood transfusion support. It is important to note that if the patient experiences severe allergic reactions, such as difficulty breathing, rapid heartbeat, or confusion, call emergency services or take them to the hospital immediately.
Score: 1. Disease Name: 1/1; 2. Disease Description: 0/2; 3. Disease Classification: 0/1; 4. Preventive Measures: 0/1; 5. Causes of Disease: 1/1; 6. Pathogenesis: 0/1; 7. Symptoms: 0/1; 8. Insurance Status: 0/1; 9. Incidence Probability: 0/1; 10. Vulnerable Population: 0/1; 11. Transmission Method: 0/1; 12. Comorbidities: 0/1; 13. Consultation Department: 0/1; 14. Treatment Method: 1/1; 15. Treatment Duration: 0/1; 16. Cure Probability: 0/1; 17. Common Medications: 0/1; 18. Treatment Cost: 0/1; 19. Examination Items: 0/1; 20. Suitable Diet: 0/1; 21. Forbidden Diet: 0/1; 22. Recommended Recipes: 0/1; 23. Recommended Medications: 0/1; 24. Medication Details: 0/1.
Total Score: 3/25.
————————- Turn 2 ————————-
Follow-up Question: What is the pathogenesis of allergic shock?
Answer: The pathogenesis of allergic shock is quite complex and involves multiple factors. Some common mechanisms include: 1. Allergic reactions: Allergens are a significant cause of allergic shock, which may include certain drugs, foods, and pollen. 2. Immune system abnormalities: Allergic reactions can be caused by immune system dysfunction, such as allergic asthma and allergic rhinitis. 3. Vascular spasms: Allergic reactions can also lead to vascular spasms, contributing to conditions like allergic shock and allergic dermatitis. 4. Neural reflexes: Allergic reactions may involve responses from the nervous system, as seen in allergic shock and allergic asthma. In summary, the pathogenesis of allergic shock is complex and requires comprehensive analysis and diagnosis based on specific circumstances.
Concluded Answer: Allergic shock is a severe allergic reaction that usually requires emergency treatment. Common emergency treatment methods include: remain calm, have the patient lie flat, loosen the collar and belt to ensure the airway is clear; immediately administer intravenous epinephrine to dilate blood vessels and increase blood pressure; monitor vital signs such as blood pressure, heart rate, and respiration, and record the time for the doctor's diagnosis and treatment. If the patient has difficulty breathing, perform artificial respiration and initiate cardiopulmonary resuscitation. Transport the patient to the hospital for treatment, including fluid and blood transfusion support. It is important to note that if the patient experiences severe allergic reactions, such

as difficulty breathing, rapid heartbeat, or confusion, call emergency services or take them to the hospital immediately.
Score: 1. Disease Name: 1/1; 2. Disease Description: 0/2; 3. Disease Classification: 0/1; 4. Preventive Measures: 0/1; 5. Causes of Disease: 1/1; 6. Pathogenesis: 0.5/1; 7. Symptoms: 0/1; 8. Insurance Status: 0/1; 9. Incidence Probability: 0/1; 10. Vulnerable Population: 0/1; 11. Transmission Method: 0/1; 12. Comorbidities: 0/1; 13. Consultation Department: 0/1; 14. Treatment Method: 1/1; 15. Treatment Duration: 0/1; 16. Cure Probability: 0/1; 17. Common Medications: 0/1; 18. Treatment Cost: 0/1; 19. Examination Items: 0/1; 20. Suitable Diet: 0/1; 21. Forbidden Diet: 0/1; 22. Recommended Recipes: 0/1; 23. Recommended Medications: 0/1; 24. Medication Details: 0/1.
Total Score: 3.5/25.
————- Turn 3 ————-
Challenge: Are vascular spasms and neural reflexes the main mechanisms of pathogenesis in allergic shock?
Answer: Yes, the pathogenesis of allergic shock is quite complex and involves multiple factors. Specifically, vascular spasms and neural reflexes are among the main mechanisms of allergic shock. The mechanisms include: 1. Allergic reactions: Allergens are a significant cause, which may include certain drugs, foods, and pollen. 2. Immune system abnormalities: Allergic reactions can result from immune system dysfunction, such as allergic asthma and allergic rhinitis. 3. Vascular spasms: Allergic reactions may lead to vascular spasms, contributing to conditions like allergic shock and allergic dermatitis. 4. Neural reflexes: Responses from the nervous system may also play a role, as seen in allergic shock and allergic asthma. In summary, the pathogenesis of allergic shock is complex and requires comprehensive analysis and diagnosis based on specific circumstances.
Concluded Answer: Allergic shock is a severe allergic reaction that usually requires emergency treatment. Common emergency treatment methods include: remain calm, have the patient lie flat, loosen the collar and belt to ensure the airway is clear; immediately administer intravenous epinephrine to dilate blood vessels and increase blood pressure; monitor vital signs such as blood pressure, heart rate, and respiration, and record the time for the doctor's diagnosis and treatment. If the patient has difficulty breathing, perform artificial respiration and initiate cardiopulmonary resuscitation. Transport the patient to the hospital for treatment, including fluid and blood transfusion support. It is important to note that if the patient experiences severe allergic reactions, such as difficulty breathing, rapid heartbeat, or confusion, call emergency services or take them to the hospital immediately.
Evaluation: 1. Disease Name: 1/1; 2. Disease Description: 0/2; 3. Disease Classification: 0/1; 4. Preventive Measures: 0/1; 5. Causes of Disease: 1/1; 6. Pathogenesis: 0.5/1; 7. Symptoms: 0/1; 8. Insurance Status: 0/1; 9. Incidence Probability: 0/1; 10. Vulnerable Population: 0/1; 11. Transmission Method: 0/1; 12. Comorbidities: 0/1; 13. Consultation Department: 0/1; 14. Treatment Method: 1/1; 15. Treatment Duration: 0/1; 16. Cure Probability: 0/1; 17. Common Medications: 0/1; 18. Treatment Cost: 0/1; 19. Examination Items: 0/1; 20. Suitable Diet: 0/1; 21. Forbidden Diet: 0/1; 22. Recommended Recipes: 0/1; 23. Recommended Medications: 0/1; 24. Medication Details: 0/1.
Total Score: 3.5/25.
————- Turn 4 ————-
Follow-up Question: What are the symptoms of allergic shock?
Answer: The main symptoms of allergic shock include: difficulty breathing, rapid heartbeat, confusion, facial and limb cyanosis, nausea, vomiting, diarrhea, dizziness, flushing, sweating, abnormal heart function (increased or decreased blood pressure), respiratory symptoms (like allergic asthma and rhinitis), and neurological symptoms (such as confusion, seizures, and coma). It is important to note that the causes of allergic shock can be diverse and require comprehensive analysis and diagnosis.

Concluded Answer: Allergic shock is a severe allergic reaction that requires emergency treatment. Common emergency treatment methods include: keeping the patient calm, having them lie flat, loosening their collar and belt to maintain airway clearance; immediately administering intravenous epinephrine to dilate blood vessels and raise blood pressure; monitoring vital signs such as blood pressure, heart rate, and respiration while recording the time for proper medical assessment. If the patient experiences difficulty breathing, perform artificial respiration and initiate cardiopulmonary resuscitation. The patient should be transported to a hospital for supportive treatment, including fluid and blood transfusions. If severe symptoms like difficulty breathing, rapid heartbeat, or confusion occur, call emergency services or take the patient to the hospital immediately.
Evaluation: 1. Disease Name: 1/1; 2. Disease Description: 1.5/2; 3. Disease Classification: 0/1; 4. Preventive Measures: 0/1; 5. Causes of Disease: 1/1; 6. Pathogenesis: 0.5/1; 7. Symptoms: 1/1; 8. Insurance Status: 0/1; 9. Incidence Probability: 0/1; 10. Vulnerable Population: 0/1; 11. Transmission Method: 0/1; 12. Comorbidities: 0/1; 13. Consultation Department: 0/1; 14. Treatment Method: 1/1; 15. Treatment Duration: 0/1; 16. Cure Probability: 0/1; 17. Common Medications: 0/1; 18. Treatment Cost: 0/1; 19. Examination Items: 0/1; 20. Suitable Diet: 0/1; 21. Forbidden Diet: 0/1; 22. Recommended Recipes: 0/1; 23. Recommended Medications: 0/1; 24. Medication Details: 0/1.
Total Score: 6/25.
————- Turn 5 ————-
Follow-up Question: What are the specific physiological processes involved in the pathogenesis of allergic shock besides allergic reactions and immune system abnormalities?
Answer: The pathogenesis of allergic shock is complex and requires comprehensive analysis and diagnosis. Common mechanisms include: 1. Allergic reactions triggered by allergens such as drugs, food, or pollen; 2. Abnormal immune responses; 3. Vascular spasms, which may lead to conditions like allergic shock and dermatitis; 4. Neural reflexes related to the nervous system; 5. Neurological symptoms that can arise from the nervous system's influence; and 6. Respiratory symptoms associated with allergic reactions. Each case needs individual assessment.
Concluded Answer: Allergic shock is a serious allergic reaction that usually necessitates emergency treatment. Common emergency interventions include keeping the patient calm, having them lie flat, loosening clothing to ensure airway clearance, and immediately administering intravenous epinephrine to dilate blood vessels and elevate blood pressure. Vital signs should be monitored and recorded for further medical evaluation. If breathing difficulties arise, perform artificial respiration and initiate cardiopulmonary resuscitation. Transport the patient to the hospital for additional treatments like fluid and blood transfusion. It is crucial to act immediately if severe symptoms like difficulty breathing, rapid heartbeat, or confusion occur.
Evaluation: 1. Disease Name: 1/1; 2. Disease Description: 1.5/2; 3. Disease Classification: 0/1; 4. Preventive Measures: 0/1; 5. Causes of Disease: 1/1; 6. Pathogenesis: 0.5/1; 7. Symptoms: 1/1; 8. Insurance Status: 0/1; 9. Incidence Probability: 0/1; 10. Vulnerable Population: 0/1; 11. Transmission Method: 0/1; 12. Comorbidities: 0/1; 13. Consultation Department: 0/1; 14. Treatment Method: 1/1; 15. Treatment Duration: 0/1; 16. Cure Probability: 0/1; 17. Common Medications: 0/1; 18. Treatment Cost: 0/1; 19. Examination Items: 0/1; 20. Suitable Diet: 0/1; 21. Forbidden Diet: 0/1; 22. Recommended Recipes: 0/1; 23. Recommended Medications: 0/1; 24. Medication Details: 0/1.
Total Score: 6/25.

G Prompt

Prompt for General Domain-level Criteria

You are an expert in {Domain}. Below are some questions, reference answers, and apprentice responses for the {Topic} scenario: {Examples}.
Please summarize the shortcomings of the apprentice's response in one sentence. Finally, provide a brief list of criteria that others can use to evaluate responses to similar questions.

Prompt for Refined Question-level Criteria

You are an expert in {Domain}. Below is the knowledge base for the question {Question}.
Knowledge Base: {Retrieved Context}.
Your task is to refine the following response criteria based on the knowledge base, ignoring any aspects not covered in it.
Criteria: {General Criteria}.
Please refine the criteria related to answering the question based on the contents of the knowledge base, clearly listing the specific elements that the response should include in order to evaluate its performance. Do not simplify the content; only list the refined criteria without explanations or responses.

Prompt for Scoring

You are an expert in {Domain}. Below is the response from the model to be evaluated. Your task is to assess the model's capability. The evaluation criteria are as follows: {Refined Criteria}.
Model Response: {Answer}.
Please first list the criteria and the maximum score for each, with the maximum score being the number of key pieces of information for that criterion. Then, evaluate and score the response based on each criterion, providing the score/max score for each. Scoring can be in decimal form. Finally, present the total score/max score as a numerical value. Do not include greetings. Your output format should be:
{"Criteria and Full Marks": "", "Evaluation": ""}.

Prompt for Initial Questions

You are a user of the {Domain} model. Please generate different user questions for the {Topic} scenario based on the context to evaluate the model's understanding of {Domain}. Ensure that the questions are fact-based, can be answered from the context, and are logically structured. List the questions directly without providing answers.
Relevant context: {Retrieved Context}.
Useful User Questions:

Prompt for Follow-up Questions

Please simulate a user of the {Domain} model. Based on the dialogue history and the scores assigned to the model's responses according to the criteria, generate a follow-up question targeting the areas with low scores in order to assess the model's understanding of {Domain}. The follow-up question should be answerable based on the criteria. Your follow-up should relate to the dialogue history.
Criteria: {Refined Criteria}.
Dialogue: {History}.
Score: {Score}.
Instructions: Your follow-up should aim to reference the model's responses while maintaining coherence in the dialogue and enhancing its interconnectedness. Please note that it is important not to disclose the content of the criteria. Do not include greetings; simply provide the follow-up

question. Your question should be answerable based on the criteria.

Prompt for Challenges

You are a user of the {Domain} model. Below is the dialogue between you and the model, along with the evaluation criteria. Please pose a challenging question to the model in the form of "Are you sure your answer is correct? "
Criteria: {Refined Criteria}.
Dialogue: {History}.
Instructions: Your challenge should relate to the previous question and the criteria. It should be based on your historical questions while maintaining coherence in the dialogue and enhancing its interconnectedness. Do not include greetings; simply provide your challenge. Ensure that your challenge can be answered based on the evaluation criteria.

Prompt for Qualitative Evaluation

You are a {domain} expert, and your task is to provide a comprehensive evaluation of the model's capabilities. You need to assess the model's responses in terms of professionalism and stability:

Professionalism Evaluation:
1. Factual Accuracy: Is the information provided by the model accurate and consistent with established standards? You may refer to the score given.
2. Meeting User Needs: Does the response meet the purpose and needs expressed by the user? Does it provide a complete and appropriate answer to the question?
3. Clarity: Is the response clear and easy to understand? Does it use simple language and structure to allow the user to easily grasp the information?

Stability Evaluation:
1. Consistency: Does the model maintain consistency in the information it provides across multiple responses?
2. Handling of Challenges: Does the model maintain the correctness of its answers when faced with a challenge?
3. Handling of Follow-up Questions: Is the model able to provide further information when prompted with follow-up questions and address the evaluator's inquiries?
Note that longer responses are not always better. Short and concise responses that meet the above requirements are preferred.

Evaluation Procedure: When you begin your evaluation, you need to follow the below procedure:
1. Evaluate the AI assistant's response based on the different dimensions. After evaluating each dimension, assign a score between 1 and 10 for that dimension.
2. Finally, based on your evaluations of each dimension, provide an overall score for the model's professionalism and stability, respectively, on a 1-10 scale.
3. Your scoring should be as strict as possible and adhere to the following scoring rules: Overall, the higher the quality of the model's response, the higher the score.
-If the model's response contains irrelevant information, fundamental factual errors, or harmful content, the total score must be 1 to 2 points.
-If the model's response contains no severe errors and is mostly harmless but of poor quality and does not meet user needs, the total score should be 3 to 4 points.
-If the model's response generally meets the user's needs but performs poorly in some dimensions and is of medium quality, the total score should be 5 to 6 points.
-If the model's response is close in quality to the reference answer and performs well in all

dimensions, the total score should be 7 to 8 points.
-Only if the model's response significantly exceeds the reference answer in quality, fully addressing the user's problem and all needs, and performing close to perfect across all dimensions, should the total score be 9 to 10 points.

Remember: You must provide an evaluation and explanation before assigning a score. After explaining each dimension, include the score for that dimension. Then, at the end of your response, return all your scores in the following dictionary format (including brackets), ensuring that all your scores are integers: {{'Dimension 1': score, 'Dimension 2': score, ..., 'Overall Score': score}}. For example: {{'Consistency': 9, 'Handling of Challenges': 6, ..., 'Stability Overall Score': 7}}, {{'Factual Accuracy': 9, 'Meeting User Needs': 6, ..., 'Professionalism Overall Score': 7}}.

Below is the conversation history and scoring, where the evaluator chooses to ask follow-up questions or challenge:
Conversation History and Scores: {History and Scores}
Evaluation Criteria: {criteria}