Benchmarking Large Language Models on Multiple Tasks in Bioinformatics NLP with Prompting

Anonymous ACL submission

Abstract

Large language models (LLMs) have become important tools in solving biological problems, offering improvements in accuracy and adaptability over conventional methods. Several benchmarks have been proposed to evaluate the performance of these LLMs. However, current benchmarks can hardly evaluate the performance of these models across diverse tasks effectively. In this paper, we introduce a comprehensive prompting-based benchmarking framework, termed Bio-benchmark, which includes 30 key bioinformatics tasks covering areas such as proteins, RNA, drugs, electronic health records, and traditional Chinese medicine. Using this benchmark, we evaluate six mainstream LLMs, including GPT-4o and Llama-3.1-70b, etc., using 0-shot and few-shot Chain-of-Thought (CoT) settings without fine-tuning to reveal their intrinsic capabilities. To improve the efficiency of our evaluations, we demonstrate BioFinder, a new tool for extracting answers from LLM responses, which increases extraction accuracy by over 40% compared to existing methods. Our benchmark results show the biological tasks suitable for current LLMs and identify specific areas requiring enhancement. Furthermore, we propose targeted prompt engineering strategies for optimizing LLM performance in these contexts. Based on these findings, we provide recommendations for the development of more robust LLMs tailored for various biological applications. This work offers a comprehensive evaluation framework and robust tools to support the application of LLMs in bioinformatics.

1 Introduction

Computational methods have become essential for solving many biological problems, such as protein folding (Jumper et al., 2021), function annotation (Singh, 2024), and designing new biomolecules (Lisanza et al., 2024). With the rise of language models in natural language processing (NLP) (Brown, 2020; Jiang et al., 2023), especially LLMs (Touvron et al., 2023; Bubeck et al., 2023; Guo et al., 2025), these powerful tools can now be applied to biology as well, especially models with over 1 billion parameters, such as ProLLaMA (Lv et al., 2024), Evo (Nguyen et al., 2024; Brixi et al., 2025), and Med-PaLM 2 (Singhal et al., 2023). In addition to handling biomedical text data such as electronic health record (EHR) and traditional Chinese medicine (TCM) question-answering (QA) systems, LLMs are also effective in analyzing biological targets like proteins and nucleic acids due to their similarity to natural language sequences. Utilizing LLMs in biological tasks has resulted in significant advancements over traditional methods, including improved accuracy, better generalization, and enhanced learning from fewer samples (Zhang et al., 2024b).

A key strategy for leveraging LLMs in biology is prompting, which involves designing specific input formats to guide the model's output (Liu et al., 2023). Prompting offers several benefits: it requires less computational power and data compared to fine-tuning, enables rapid adaptation to various tasks, and fully utilizes the model's existing knowledge. Additionally, prompting facilitates the resolution of complex biological questions by framing them in a way that LLMs can effectively interpret and answer.

Many specialized LLMs have been developed for different biological tasks (Wang et al., 2024a; Lv et al., 2024; Bolton et al., 2024), but there is currently no thorough investigation of which biological task is better suited for which LLM architecture. Benchmarking is a rigorous method for assessing the compatibility and performance of an LLM architecture across a variety of tasks (McIntosh et al., 2024; Jiang et al., 2024; Ali et al., 2024). However, current benchmark methods for LLMs on biological tasks face several challenges. First, most benchmarks designed for small-scale mod-

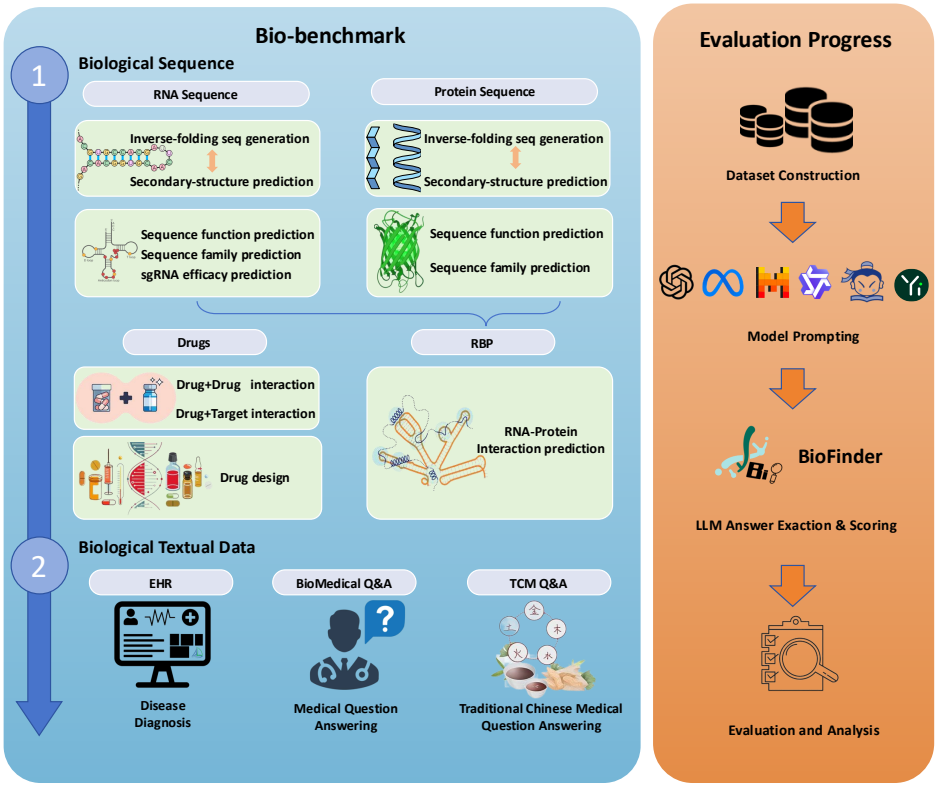

Figure 1: Overview of the paper. Bio-benchmark is divided into sequence (Protein, RNA, RBP, Drug) and text (EHR, Medical, TCM) data (left). The process is to use six representative LLMs to generate answers for a total of 30 subtasks in seven domains in biological information through 0-shot and few-shot after the benchmark is built. After extracting the answers generated by LLMs using our proposed BioFinder, we evaluate and analyze the extracted answers of different LLMs (right). **The experiment results table is in Appendix Table 2, Figure 2, 3, and 7.**

els are inadequate for evaluating LLMs due to their expanded hypothesis space. Second, existing validation datasets often contain overlapping data, highlighting the necessity for high-quality, clean benchmark sets. Third, despite the growing interest in developing LLMs for diverse tasks, there are limited benchmarks tailored to evaluate these comprehensive models.

Inspired by these needs, this paper introduces a comprehensive prompting-based benchmarking scheme for evaluating the performance of LLMs on various biological tasks. We present evaluation methods for LLMs using either biological sequence data or biomedicine-related text inputs, and apply these benchmarks to currently available LLMs, with an emphasis on general LLMs such as GPT-4o. Our assessment specifically focuses on benchmarking vanilla LLMs without fine-tuning under zero-shot or few-shot CoT (Wei et al., 2022) conditions to test the intrinsic abilities of LLMs, which can also better inform their performance after fine-tuning. Moreover, as LLMs often embed the key answer within the generated text output and there are currently no effective methods for extracting biological sequences and texts (Gu et al., 2022), we propose a novel answer extraction method, BioFinder, to retrieve key answers from responses and enhance the reliability of our LLM benchmark. Based on our comprehensive benchmark results, we analyze and summarize the biological tasks that are suitable for current LLMs to solve and provide architectural recommendations to guide the development of future LLMs for diverse biological tasks.

Our main contributions are as follows: (1) we propose a bioinformatics benchmark (Bio-benchmark) including 30 important tasks related to protein, RNA, RNA-binding protein (RBP), drug, electronic health record, medical QA and traditional Chinese medicine. We further test these tasks in zero-shot and few-shot (with CoT) settings across six mainstream vanilla LLMs without fine-tuning: GPT-4o (OpenAI, 2024), Qwen-2.5-72b (Hui et al., 2024), Llama-3.1-70b (Dubey

et al., 2024), Mistral-large-2 (Mistral-AI, 2024), Yi-1.5-34b (0.1AI, 2024), and InternLM-2.5-20b (Cai et al., 2024); (2) we propose an answer extraction tool (BioFinder) for LLMs on bioinformatics tasks to accurately extract answers from LLMs responses, which exceeds the accuracy of existing methods by over 40%; (3) we evaluate and analyze the answers extracted by xFinder (Yu et al., 2024) and BioFinder on different tasks, identify which benchmarked tasks are well-suited for current state-of-the-art LLMs, and propose a prompt-based approach to enhance LLM performance on tasks that are currently less effectively addressed.

2 Benchmark construction

Protein. Datasets for predicting protein secondary structure and function are developed using data from the Protein Data Bank (PDB), involving deduplication and length restrictions to manage computational load. The protein secondary structure dataset was diversified through random sampling based on sequence lengths. For species prediction, sampling is guided by sequence distributions across species and family sequence counts, ensuring a balanced representation of families. These preprocessing steps are critical for building robust models that accurately predict protein functionalities or species, advancing research in this field.

We involve four subtasks in the protein section, including protein family sequence design, protein species prediction, protein inverse-folding design, and protein structure prediction. For instance, targeting on protein inverse-folding design task, given the protein secondary structure "CCSHHHHHHHHHHHHHHHHHHHHHCCCCC", we prompt the language model to design the protein sequence whose secondary structure matched the given one. Protein structure prediction tasks are used to predict the secondary structure for a given specific protein sequence. For protein family sequence design, we aim to generate protein sequences belonging to a specific protein family based on the protein family ID, like "PF01713". For protein species classification, we utilize a language model to predict the category of each protein based on its sequence and the predefined protein species categories.

RNA. The dataset construction for RNA secondary structure prediction, inverse folding, and functional prediction involved meticulous preprocessing across multiple sources. The bpRNA dataset provide the foundation for RNA structure and folding benchmarks, where sequences underwent deduplication, removal of entries exceeding 1024 nucleotides, and random sampling based on sequence length distributions to ensure diversity and representativeness.

For RNA functional prediction, data from RNA-Central are used, with samples select based on the distribution of sequences across different functions and families. This approach ensure a balanced representation of various RNA families and functions, facilitating the development of robust models for accurately predicting RNA characteristics and behaviors in molecular biology research.

For the RNA section, we involve 5 subtasks, including RNA family sequence design, RNA function prediction, RNA inverse-folding design, RNA structure prediction, sgRNA efficiency prediction. The first four tasks are consistent with those in the protein section, while the last involve predicting the efficiency of a given sgRNA, such as "AAAAAAAAUUGGAGGCCUUUCCCCUGGGCA", with an example efficiency prediction of 90%.

RNA-binding protein. To rigorously define a benchmark set for RNA-protein interaction-related tasks, we first obtain a set of RNA sequences experimentally verified to interact with various RNA-binding proteins. We then categorize these proteins by their binding domains, such as RNA recognition motif (RRM) and zinc finger. For the sampled protein of each binding domain, we curate balanced positive and negative RNA sequences binding to this protein. Notably, the same RNA-binding protein can demonstrate varying behavior in different species, so the species information is also included in the prompt benchmark. The classification benchmark is then constructed accordingly by prompting the model with the protein name, species, and RNA sequence information.

In this section, we involve RNA-binding protein prediction tasks. Given a specific RNA sequence like "AGAAGGUGUGAGAUUAAUGGAUGGGGUAGCUGACG", we prompt the language model to decide whether it can bind to a specific protein like "Pp_0237" or not.

Drug. In constructing a comprehensive drug benchmark for machine learning-guided drug design, three essential tasks are addressed: antibiotic design, drug-drug interactions, and drug-target interactions. These tasks are critical as they encompass the full spectrum of drug development, from

discovery through to clinical application, and directly impact the success rate, efficiency, and safety of pharmaceuticals. For antibiotic design, the data includes novel antibiotic structures to address the urgent need for treatments against antimicrobial resistance and to test the predictive capabilities of LLMs. In the area of drug-drug interactions, the dataset is carefully curated to include 86 different interaction types, ensuring a broad representation of potential clinical scenarios to enhance medication safety and efficacy. For drug-target interactions, the selected data ensures diversity and specificity with non-overlapping drug structures and targets with low sequence similarity, streamlining the pathway from experimental screening to clinical application.

We focus on three essential tasks: drug-drug interaction prediction, drug-target interaction prediction, and drug design. For example, in the task of drug design, we prompt the model to determine if a molecule like "Cc1c(Br)c(=O)c(C(=O)N2CC(C)S(=O)C(C)C2)cn1C1CC1" can have potent efficacy on A. baumannii (a bacteria). And the example answer should be "Not potent".

Electronic health record. EHR data is scarce, particularly within the MIMIC database (Johnson et al., 2023), which predominantly contains data from ICUs and emergency departments, and is encumbered by privacy-related access restrictions. To address this, three diverse benchmarks have been selected for utilizing EHR data effectively: predicting diagnostic outcomes using patient information and test results, analyzing patient data incrementally to devise treatment plans, and generating medical reports from doctor-patient dialogues. These benchmarks are essential for improving health outcomes and the efficiency of medical services.

General medical question-answering. The selected data for medical knowledge benchmarks emphasize accuracy and are derived from expert-verified exam questions. These benchmarks are geographically diverse: MedQA sources from exams in Mainland China, Taiwan, and the USA; MedMCQA from Indian exams; and HeadQA from Spanish exams. They cover a wide range of topics, with MMCU including 13 subcategories like medical imaging and infectious diseases, while HeadQA spans six areas such as medicine and psychology. Some datasets simulate real-world medical consultations requiring long-text responses to open-ended questions, although most benchmarks use multiple-choice formats due to limited alternatives, reflecting a compromise in developing practical, 0-shot style medical question answering.

Questiong-Answering in traditional Chinese medicine. High-quality benchmarks for TCM are rare, primarily consisting of multiple-choice formats that often mix Eastern and Western medical concepts, making it difficult to isolate pure TCM content. The CMB/CMMLU datasets fill this gap by incorporating questions from ancient texts and traditional TCM, expressed in Classical Chinese, which tests models' understanding of historical language nuances. Furthermore, the TCM SD dataset uses real clinical TCM cases, where questions focus on predicting diseases and syndromes from patient information using specific TCM terminologies, ensuring a practical evaluation of models' clinical capabilities in TCM.

3 Evaluation methods

3.1 Problem definition

Given a set of questions $Q = \{q_i \mid q_i \in \Sigma^*\}$, our goal is to evaluate the performance of LLMs on various tasks. Since we adopt an unconstrained output format using the CoT method, accurately extracting the standard answer a_i from the generated text $y_i = \text{LLM}(q_i)$ is challenging; traditional RegEx-based methods struggle to balance false positives and negatives. To address this, we utilize the BioFinder framework (Section A.1) with an answer extraction function $E : \Sigma^* \to \Sigma^*$ to extract the key answer $k_i = E(y_i)$ from the model output y_i.

We divide the evaluation tasks into objective and subjective evaluations. For **objective evaluation tasks** ($t_i \in T_{\text{obj}}$), such as multiple-choice questions and character matching with a definitive standard answer a_i, we define an evaluation function $M_{\text{obj}} : \Sigma^* \times \Sigma^* \times T_{\text{obj}} \to \mathbb{R}$ to compare the extracted answer k_i with the standard answer a_i and compute the performance of LLM on task t_i:

$$\text{Performance}_{\text{obj}} = M_{\text{obj}}(k_i, a_i, t_i). \quad (1)$$

For **subjective evaluation tasks** ($t_i \in T_{\text{subj}}$), such as long-text generation, open-ended question answering, and case analysis, we evaluate the quality of the model's output through the following:

Similarity calculation: Define a similarity function $S : \Sigma^* \times \Sigma^* \to \mathbb{R}$ to compute the textual and semantic similarity between the generated answer

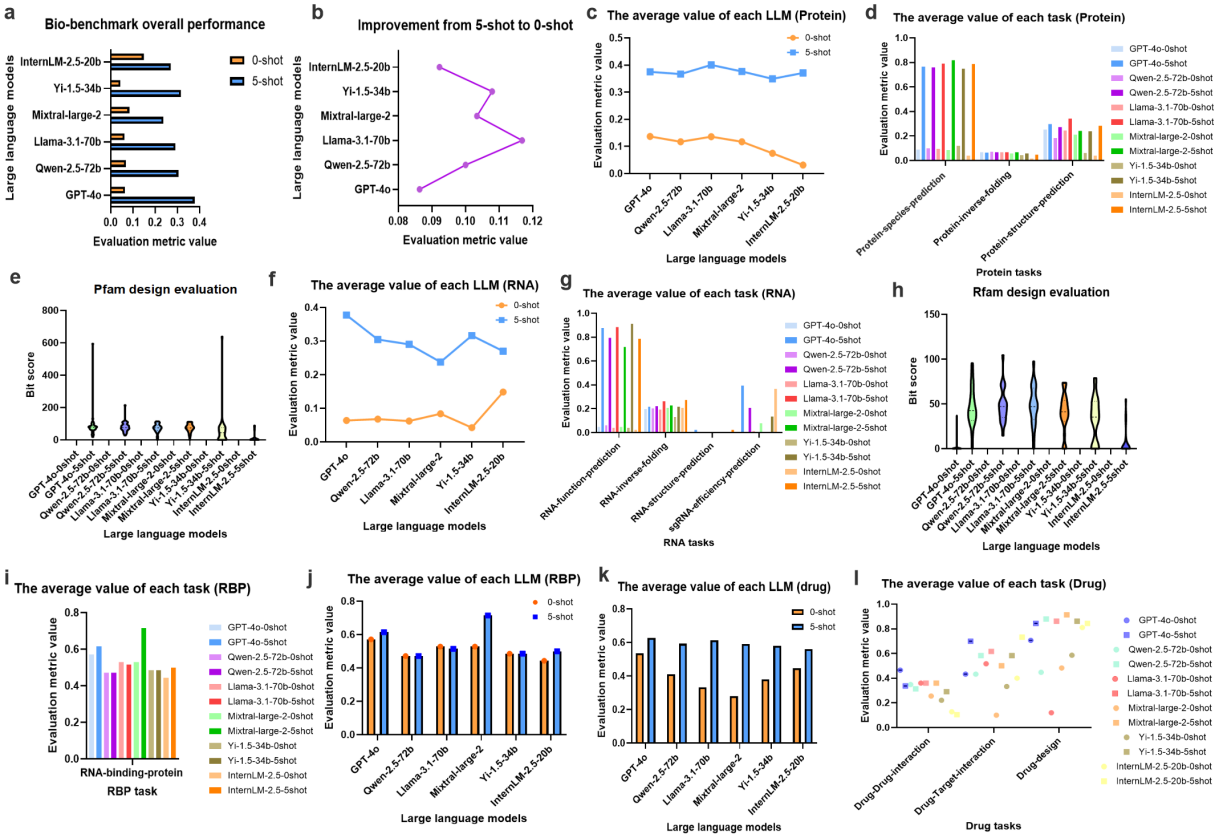

Figure 2: The performance of sequence and drug benchmarks in the Bio-benchmark. **a, b** represent overall performance. **c, d, e** indicate performance on protein benchmark. **f, g, h** show performance on RNA benchmark. **i, j** reflect performance on RBP benchmark. **k, l** demonstrate performance on drug benchmark.

y_i and the reference answer a_i:

$$s_i = S(y_i, a_i). \tag{2}$$

Expertise assessment: Use advanced LLMs like GPT-4o to assess the professional quality of the generated content y_i, obtaining a quality score q_i. **Logical consistency judgment**: Use the trained BioFinder to perform Natural Language Inference (NLI), defining a logical relation function:

$$N : \Sigma^* \times \Sigma^* \rightarrow \{\text{Entailment}, \text{Contradiction}, \text{Neutral}\} \tag{3}$$

to conduct fine-grained analysis of the logical relationship between y_i and a_i:

$$n_i = N(y_i, a_i). \tag{4}$$

Through the BioFinder framework, we can efficiently and accurately extract key answers from LLM-generated texts without constraining the output format, enabling a comprehensive evaluation of the model's capabilities and generation quality.

3.2 Fine-grained textual evaluation

To evaluate the quality of the text generation, we employ metrics including similarity, expertise, and logical consistency. Specifically, inspired by FACTSOCRE and OLAPH (Min et al., 2023; Jeong et al., 2024), we compare the LLM's response $\hat{P}$ with reference statements categorized as *Must-Have* (MH) and *Nice-to-Have* (NH), assigned weights w_{MH} and w_{NH}, respectively, with $w_{\text{MH}} > w_{\text{NH}}$. Using the BioFinder framework and **Natural Language Inference** (NLI), we classify the relationship between the LLM's response and each reference statement as Entailed, Contradicted, or Neutral. Based on this classification, we define the following metrics:

Comprehensiveness measures the extent to which the model's response correctly includes the reference statements, defined as:

$$\text{Comprehensiveness}(\hat{P}) = \frac{\sum_{x \in \text{Entailed}} w(x)}{\sum_{x \in S} w(x)}, \tag{5}$$

where $S = \text{MH} \cup \text{NH}$ represents the set of all reference statements, Entailed is the subset of statements classified as entailed by BioFinder, and $w(x)$

is the weight of statement x.

Hallucination Rate evaluates the proportion of contradictory information in the LLM's response relative to the reference statements, reflecting the degree of misinformation, defined as:

$$\text{Hallucination rate}(\hat{P}) = \frac{\sum_{x \in \text{Contradicted}} w(x)}{\sum_{x \in S} w(x)}, \tag{6}$$

where Contradicted is the subset of statements classified as contradicted by BioFinder.

Omission rate measures the extent to which the model's response omits reference statements (especially must-have information), defined as:

$$\text{Omission Rate}(\hat{P}) = \frac{\sum_{x \in \text{Neutral}} w(x)}{\sum_{x \in S} w(x)}, \tag{7}$$

where Neutral is the subset of statements classified as neutral by BioFinder.

To assess the stability of the LLM's performance across the dataset, we calculate the variability of the metrics and define **Consistency** as:

$$\text{Consistency} = 1 - \frac{\sigma_{\text{Comp}} + \sigma_{\text{Hall}} + \sigma_{\text{Omit}}}{\sum \sigma_{\text{max}}}, \tag{8}$$

where σ_{Comp}, σ_{Hall}, σ_{Omit} are the standard deviations of Comprehensiveness, Hallucination Rate, and Omission Rate, respectively, and σ_{max} is the sum of maximum possible standard deviations.

We combine the above metrics to define the model's overall performance score:

$$\begin{aligned}\text{Overall Score} = \alpha\overline{\text{Comp}} - \beta\overline{\text{Hall}} \\ - \gamma\overline{\text{Omit}} + \delta\text{Consistency},\end{aligned} \tag{9}$$

where $\overline{\text{Comp}}$, $\overline{\text{Hall}}$, $\overline{\text{Omit}}$ are the average values of Comprehensiveness, Hallucination Rate, and Omission Rate, respectively, and α, β, γ, δ are weighting coefficients satisfying $\alpha+\beta+\gamma+\delta = 1$.

4 Experiment and results

For BioFinder setting, baselines, evaluation, and subjective task evaluation, please see Appendix A.1. For Bio-benchmark setting, baselines, evaluation, etc. please see Appendix A.2.

4.1 Experiment Setting

We conduct experiments on the Bio-benchmark datasets. We use Openai-API and 8*A100-80G GPUs to infer LLMs with LM-Deploy and Huggingface backend (Contributors, 2023a; hug, 2024). We employ sampling hyperparameters with top-p set to 1.0 and a temperature of 0.2 for generation (Specific prompts in the Appendix D). We use BioFinder to extract the answers of objective evaluation and execute NLI tasks.

4.2 Evaluation of Different Tasks

For more detailed Bio-benchmark experimental evaluation metrics, please see Appendix A.2.3.

1. Protein: Pfam design (bit score), protein species prediction (Acc.), protein inverse folding (recovery rate), protein structure prediction (recovery rate). **2. RNA:** RNA function prediction (Acc.), RNA inverse folding (recovery rate), RNA structure prediction (F1), Rfam design (bit score), sgRNA efficiency prediction (Acc.). **3. RBP:** RNA-binding protein (Acc.). **4. Drug:** Drug-Drug interaction (Acc.), drug-target interaction (Acc.), drug design (Acc.). **5. EHR:** Agentclinic (Acc.), CMB-Clinic (fluency, relevance, completeness, medical proficiency), IMCS-MRG (Rouge-L). **6. Medical:** HeadQA (Acc.), MedLFQA-HealthQA (COMPREHENSIVENESS), MedLFQA-KQA (COMPREHENSIVENESS), MedLFQA-LiveQA (COMPREHENSIVENESS), MedLFQA-MedicationQA (COMPREHENSIVENESS), MedMCQA (Acc.), MedQA-CN (Acc.), MedQA-TW (Acc.), MedQA-US (Acc.), MMCU (Acc.). **7. TCM:** CMB-Exam (Acc.), CMMLU-TCM (Acc.), MedLFQA-TCM (Acc.), TCMSD (Acc.).

5 Analysis

We observe that prompt formats significantly affect LLM performance in biological sequence inference. Specifically, continuous bio-sequences versus those separated by spaces or newline characters lead to different tokenizations due to LLMs using Byte Pair Encoding (BPE). BPE tokenizers often treat several consecutive uppercase letters as a single token, which impairs the ability of LLMs to understand and generate bio-sequences. Figure 4 showed that separating sequences with newline characters achieves over three times the alignment accuracy compared to continuous inputs.

In addition, from Figure 2, Figure 3 and Table 2, we can obtain the following results.

5.1 Protein benchmark

In the protein species prediction, few-shot significantly improved accuracy across all LLMs. Yi-1.5-34b experience at least a sixfold increase, while

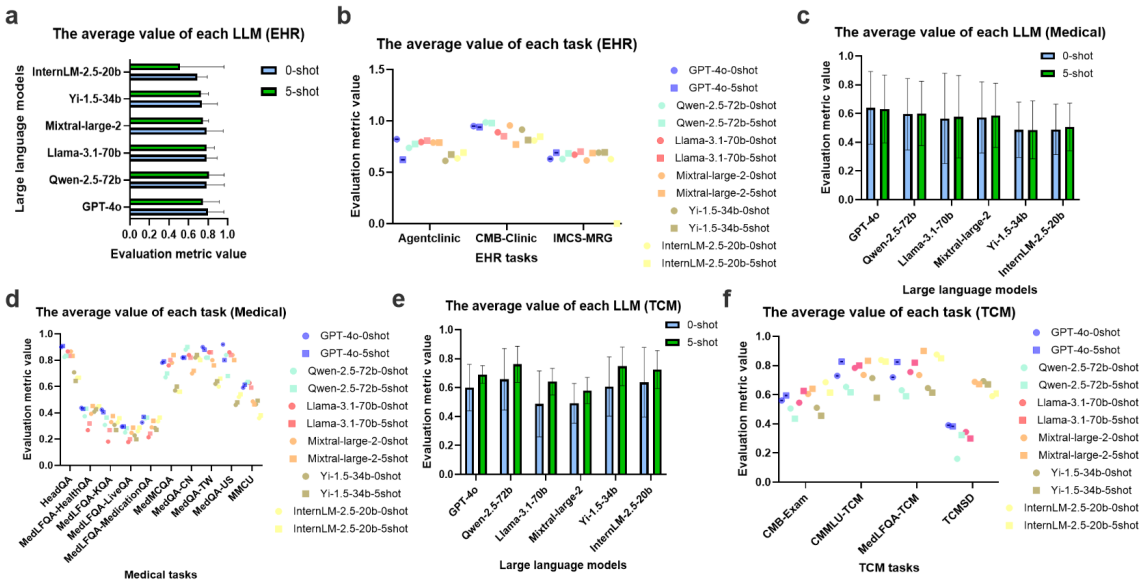

Figure 3: The performance of biotext benchmarks in the Bio-benchmark. **a, b** represent performance on EHR benchmark. **c, d** indicate performance on Medical-QA benchmark. **e, f** show performance on TCM benchmark.

	RegEx	GPT-4o	BioFinder	RoBERTa-large-mnli	GTE-large-en-v1.5
MCQ Matching	77.5	65.8	**95.5**	-	-
Text Matching	74.8	80.5	**94.3**	-	-
Numerical Matching	68.1	67.0	**95.5**	-	-
Sequence Extraction	68.0	38.5	**93.5**	-	-
Overall Objective Tasks	72.1	62.9	**94.7**	-	-
NLI Tasks	-	59.9	**89.8**	27.0	23.4

Table 1: Comparison of performance on different tasks between RegEx (OpenCompass), GPT-4o, RoBERTa-large-mnli, GTE-large-en-v1.5, and BioFinder.

InternLM-2.5-20b witness a maximum nearly twentyfold increase. Mistral-large-2 lead with 82% accuracy, followed by Llama-3.1-70b at 79%.

For inverse protein folding and protein structure prediction, we measure the sequence recovery rate. The latter task show notably better performance, with accuracy approximately four times higher than the former. Notably, in a five-shot setting, Llama-3.1 70b achieve the highest score of 34% in protein structure prediction.

In addition, we evaluate the proficiency of LLMs in generating protein sequences from specific Pfam IDs. Initially, all LLMs score zero in 0-shot tests, indicating their inability to generate accurate sequences based solely on Pfam IDs. However, with 10-shot prompting, LLMs substantially improve, achieving bit scores over 50%, with GPT-4o reaching the highest at 87%.

5.2 RNA benchmark

We evaluate large language models on RNA sequence processing tasks. For RNA function prediction, similar to protein species prediction, few-shot (5-shot) prompting significantly enhance model performance over 0-shot, with Llama-3.1-70b 5-shot achieving the highest accuracy of 89%.

In RNA-inverse folding and structure prediction tasks, results are generally disappointing. However, RNA inverse folding outperforms structure prediction tasks, with Llama-3.1-70b 5-shot leading with a recovery rate of 26%. Few-shot prompting notably improves outcomes in both RNA tasks.

We also examine sgRNA efficiency prediction, crucial for CRISPR gene editing success. GPT-4o 5-shot and InternLM-2.5-20b 0-shot both score above 30%. Notably, InternLM-2.5-20b performed better in 0-shot than in 5-shot, where it scores zero, suggesting no benefit from extra examples.

Lastly, we evaluate RNA generation from specific Rfam IDs, using bit scores to judge sequence quality. 10-shot prompting consistently outperforms 0-shot across all LLMs, though the average score of 40.78% is lower than the 72.63% achieved in the protein section under the same conditions.

5.3 RBP Benchmark

The RNA binding protein task is used to evaluate the ability of LLMs to predict whether a given RNA sequence can bind to a specific protein. These LLMs achieved an average accuracy of 53%, with the highest scores being 71% by Mistral-large-2 (5-shot) and 61% by GPT-4o (5-shot). In this task, the few-shot prompting strategy slightly improve performance compared to 0-shot prompting.

5.4 Drug Benchmark

According to Figure 5, we conduct experiments on drug sequences similar to those performed on biological sequences. By counting the number of atoms in the drug sequences, we discover that the prompt format significantly impacts the LLMs' drug-sequence generation. In the drug design task, large language models are tested on their ability to predict the efficacy of molecular entities (SMILES strings) against specific bacteria. Using 5-shot prompting, all LLMs perform well, achieving accuracy scores above 80%, with Mistral-large-2 5-shot topping at 91%. Notably, 5-shot prompting boost the performance of Llama-3.1-70b significantly, from 12% in 0-shot to 86%.

In the Drug-Drug and Drug-Target interaction tasks, LLMs evaluate potential interactions. Drug-Drug interaction predictions are unsatisfactory across all models, with the best score being 0.47 by GPT-4o in 0-shot, which decreases to 0.34 with 5-shot prompting. Conversely, Drug-Target interaction predictions show better results; InternLM-2.5-20b 5-shot and GPT-4o 5-shot achieve the highest scores of 73% and 70%, respectively. 5-shot prompting significantly enhances performance, increasing the accuracy from 62% in 0-shot to 37%.

5.5 EHR Benchmark

In the EHR benchmark, LLMs show high accuracy across three medical reasoning subtasks: AgentClinic, CMB-Clinic, and IMCS-MRG. For the AgentClinic task, average metric values are 73.2% (0-shot) and 72.7% (5-shot), with GPT-4o achieving a high of 82.24% in 0-shot.

Performance in the CMB-Clinic task was excellent, with average metric values of 0.917 in 0-shot and 0.867 in five-shot settings using CoT reasoning. This performance closely matches reference answers, highlighting the potential of LLMs in medical case diagnosis.

It is noted that for the CMB-Clinic task, the accuracy of GPT-4o slightly decreased in the five-shot setting comparing with the 0-shot setting. This decrease is attributed to interference from example questions and the inclusion of irrelevant information in few-shot prompting, potentially leading to reduced accuracy when longer inputs do not improve reasoning outcomes.

5.6 Medical Benchmark

For the Medical-QA Benchmark, which includes seven datasets, the average performance of LLMs in 0-shot and 5-shot settings is similar. However, the overall performance varied significantly across different datasets. LLMs perform well on multiple-choice problems including HeadQA, MedMCQA, MedQA-CN, MedQA-TW, and MedQA-USA, with average 0-shot metric values of 80.6%, 68.7%, 81.3%, 76.7%, and 74.0%, respectively. Medical QA tasks are particularly well-suited for LLMs, which align effectively with the domain.

Furthermore, there is little to no improvement with five-shot prompting, and in some cases, performance decreased. The reason can be that biomedical QA is a task that aligns well with LLMs, achieving high 0-shot performance. The additional five-shot examples provide minimal benefit.

5.7 TCM Benchmark

Overall, LLMs perform well on the TCM benchmark. Out of the four subtasks, only TCMSD show relatively lower performance with an average metric value of 31.7%. Notably, there is a significant improvement in five-shot compared to 0-shot settings across various subtasks. The average metric value improves from 31.7% to 65.3%. This is because TCM is less commonly considered during model training, and the use of prompts can enhance performance. We infer that if LLMs are fine-tuned with the related corpus, the performance could improve further.

6 Conclusion

This study introduces a benchmarking framework to evaluate LLM performance on bioinformatics tasks. By leveraging BioFinder, we accurately extract key answers from LLM outputs, boosting precision. Our findings show robust performance across protein, RNA, and drug design tasks, particularly in few-shot settings. Future work may refine prompt engineering for improved efficiency and accuracy.

Limitations

Despite the comprehensive nature of the Bio-benchmark and the novel approach of the BioFinder tool, our study has several limitations that warrant consideration. Firstly, the benchmark primarily assesses the intrinsic capabilities of LLMs using 0-shot and few-shot settings without fine-tuning. While this approach provides insights into the raw abilities of these models, it may not fully reflect their performance in practical, real-world applications where fine-tuning on specific tasks can significantly enhance effectiveness.

Secondly, while we cover a broad range of biological tasks, the selection of these tasks and their corresponding datasets might still not be representative of all possible bioinformatics challenges. The diversity and complexity of biological data are such that even a comprehensive benchmark like ours cannot encapsulate all potential tasks and scenarios. This limitation could affect the generalizability of our findings and recommendations.

Furthermore, the performance of the BioFinder tool, although superior to existing methods by a significant margin, is heavily dependent on the quality and structure of the data fed into the LLMs. In scenarios where the LLM outputs are ambiguous or overly complex, the extraction accuracy of BioFinder might be compromised. This limitation underscores the ongoing challenge in the field of bioinformatics to develop methods that are robust across a wide variety of data types and quality.

Additionally, our benchmark does not address the computational and financial costs associated with deploying large language models. The use of such models, particularly in 0-shot and few-shot scenarios, can be resource-intensive. This aspect could limit the accessibility of our proposed methods for researchers with limited resources.

Lastly, while our study indicates promising areas for the application of LLMs in bioinformatics, it also reveals that certain tasks remain challenging for these models. Developing LLM architectures or training strategies that can tackle these hard-to-model aspects of biological data remains a key area for future research. The ongoing development of LLMs should focus on enhancing adaptability and accuracy in these less tractable areas to broaden the utility of LLMs in bioinformatics.

Ethics Statement

This paper does not involve an ethics statement.

References

2024. Hugging face inference api and endpoints. https://huggingface.co/docs/inference-endpoints/index.

0.1AI. 2024. 01-ai/Yi-1.5-34B-Chat · Hugging Face — huggingface.co. https://huggingface.co/01-ai/Yi-1.5-34B-Chat. [Accessed 15-10-2024].

Murtaza Ali, Prerna Rao, Yifan Mai, and Benjamin Xie. 2024. Using benchmarking infrastructure to evaluate llm performance on cs concept inventories: Challenges, opportunities, and critiques. In *Proceedings of the 2024 ACM Conference on International Computing Education Research-Volume 1*, pages 452–468.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, et al. 2024. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Advances in Neural Information Processing Systems*, 36.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, et al. 2024. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*.

Garyk Brixi, Matthew G Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A Gonzalez, Samuel H King, David B Li, Aditi T Merchant, Mohsen Naghipourfar, Eric Nguyen, Chiara Ricci-Tam, David W Romero, Gwanggyu Sun, Ali Taghibakshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K Wang, Etowah Adams, Stephen A Baccus, Steven Dillmann, Stefano Ermon, Daniel Guo, Rajesh Ilango, Ken Janik, Amy X Lu, Reshma Mehta, Mohammad R.K. Mofrad, Madelena Y Ng, Jaspreet Pannu, Christopher Re, Jonathan C Schmok, John St. John, Jeremy Sullivan, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Tom McGrath, Kimberly Powell, Dave P. Burke, Hani Goodarzi, Patrick D Hsu, and Brian Hie. 2025. Genome modeling and design across all domains of life with evo 2. *bioRxiv*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Qijie Chen, Haotong Sun, Haoyang Liu, Yinghui Jiang, Ting Ran, Xurui Jin, Xianglu Xiao, Zhimin Lin, Hongming Chen, and Zhangmin Niu. 2023. An extensive benchmark study on biomedical text generation and mining with chatgpt. *Bioinformatics*, 39(9):btad557.

Shan Chen, Yingya Li, Sheng Lu, Hoang Van, Hugo JWL Aerts, Guergana K Savova, and Danielle S Bitterman. 2024. Evaluating the chatgpt family of models for biomedical reasoning and classification. *Journal of the American Medical Informatics Association*, 31(4):940–948.

LMDeploy Contributors. 2023a. Lmdeploy: A toolkit for compressing, deploying, and serving llm. https://github.com/InternLM/lmdeploy.

OpenCompass Contributors. 2023b. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

OpenCompass Contributors. 2023c. Opencompass: A universal evaluation platform for foundation models. *GitHub repository*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Carl Edwards, Aakanksha Naik, Tushar Khot, Martin Burke, Heng Ji, and Tom Hope. 2023. Synergpt: In-context learning for personalized drug synergy prediction and drug design. *arXiv preprint arXiv:2307.11694*.

Yanjun Gao, Skatje Myers, Shan Chen, Dmitriy Dligach, Timothy A Miller, Danielle Bitterman, Matthew Churpek, and Majid Afshar. 2024. When raw data prevails: Are large language model embeddings effective in numerical data representation for medical machine learning applications? *arXiv preprint arXiv:2408.11854*.

Jiasheng Gu, Hongyu Zhao, Hanzi Xu, Liangyu Nie, Hongyuan Mei, and Wenpeng Yin. 2022. Robustness of learning from task instructions. *arXiv preprint arXiv:2212.03813*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Ahmed Abdeen Hamed, Tamer E Fandy, and Xindong Wu. 2024. Accelerating complex disease treatment through network medicine and genai: A case study on drug repurposing for breast cancer. *arXiv preprint arXiv:2406.13106*.

Chaoqun He, Renjie Luo, Shengding Hu, Yuanqian Zhao, Jie Zhou, Hanghao Wu, Jiajie Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2024a. Ultraeval: A lightweight platform for flexible and comprehensive evaluation for llms. *arXiv preprint arXiv:2404.07584*.

Yunzhen He, Hiroaki Yamagiwa, and Hidetoshi Shimodaira. 2024b. Shimo lab at" discharge me!": Discharge summarization by prompt-driven concatenation of electronic health record sections. *arXiv preprint arXiv:2406.18094*.

Wenpin Hou and Zhicheng Ji. 2024. Assessing gpt-4 for cell type annotation in single-cell rna-seq analysis. *Nature Methods*, pages 1–4.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Hui Huang, Yingqi Qu, Jing Liu, Muyun Yang, and Tiejun Zhao. 2024. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers. *arXiv preprint arXiv:2403.02839*.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Minbyul Jeong, Hyeon Hwang, Chanwoong Yoon, Taewhoo Lee, and Jaewoo Kang. 2024. Olaph: Improving factuality in biomedical long-form question answering. *arXiv preprint arXiv:2405.12701*.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. 2021. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120.

Jiyue Jiang, Pengan Chen, Liheng Chen, Sheng Wang, Qinghang Bao, Lingpeng Kong, Yu Li, and Chuan Wu. 2024. How well do llms handle cantonese? benchmarking cantonese capabilities of large language models. *arXiv preprint arXiv:2408.16756*.

Jiyue Jiang, Sheng Wang, Qintong Li, Lingpeng Kong, and Chuan Wu. 2023. A cognitive stimulation dialogue system with multi-source knowledge fusion for elders with cognitive impairment. *arXiv preprint arXiv:2305.08200*.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589.

Deqian Kong, Yuhao Huang, Jianwen Xie, Edouardo Honig, Ming Xu, Shuanghong Xue, Pei Lin, Sanping Zhou, Sheng Zhong, Nanning Zheng, et al. 2024. Dual-space optimization: Improved molecule sequence design by latent prompt transformer. *arXiv preprint arXiv:2402.17179*.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023a. Huatuo-26m, a large-scale chinese medical qa dataset. *Preprint*, arXiv:2305.01526.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Zhongshen Li, Junru Jin, Wentao Long, and Leyi Wei. 2023b. Plpmpro: Enhancing promoter sequence prediction with prompt-learning based pre-trained language model. *Computers in Biology and Medicine*, 164:107260.

Sidney Lyayuga Lisanza, Jacob Merle Gershon, Samuel WK Tipps, Jeremiah Nelson Sims, Lucas Arnoldt, Samuel J Hendel, Miriam K Simma, Ge Liu, Muna Yase, Hongwei Wu, et al. 2024. Multistate and functional protein design using rosettafold sequence space diffusion. *Nature biotechnology*, pages 1–11.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, 2021. *URL https://arxiv. org/abs/2107.13586*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Jiaxi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian. 2024. Prollama: A protein large language model for multi-task protein language processing. *arXiv preprint arXiv:2402.16445*.

Timothy R McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Paul Watters, and Malka N Halgamuge. 2024. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *arXiv preprint arXiv:2402.09880*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Mistral-AI. 2024. migtissera/Tess-3-Mistral-Large-2-123B · Hugging Face — huggingface.co. https://huggingface.co/migtissera/Tess-3-Mistral-Large-2-123B. [Accessed 15-10-2024].

Eric P Nawrocki and Sean R Eddy. 2013. Infernal 1.1: 100-fold faster rna homology searches. *Bioinformatics*, 29(22):2933–2935.

Eric P Nawrocki, Diana L Kolbe, and Sean R Eddy. 2009. Infernal 1.0: inference of rna alignments. *Bioinformatics*, 25(10):1335–1337.

Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel H. King, Garyk Brixi, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian L. Hie. 2024. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336.

Pascal Notin, Aaron W Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, Nathan Rollins, Ada Shaw, Ruben Weitzman, Jonathan Frazer, et al. 2023. Proteingym: Large-scale benchmarks for protein design and fitness prediction. *bioRxiv*.

OpenAI. 2024. Gpt-4o. https://openai.com/index/hello-gpt-4o/. [Accessed 15-10-2024].

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

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Yuchen Ren, Zhiyuan Chen, Lifeng Qiao, Hongtai Jing, Yuchen Cai, Sheng Xu, Peng Ye, Xinzhu Ma, Siqi Sun, Hongliang Yan, et al. 2024. Beacon: Benchmark for comprehensive rna tasks and language models. *arXiv preprint arXiv:2406.10391*.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2022. Leveraging large language models for multiple choice question answering. arxiv 2022. *arXiv preprint arXiv:2210.12353*.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. *arXiv preprint arXiv:2402.16786*.

Frederic Runge, Karim Farid, Jorg KH Franke, and Frank Hutter. 2024. Rnabench: A comprehensive library for in silico rna modelling. *bioRxiv*, pages 2024–01.

Arunima Singh. 2024. Llms predict protein phases. *Nature Methods*, pages 1–1.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Sarvesh Soni, Surabhi Datta, and Kirk Roberts. 2023. quehry: a question answering system to query electronic health records. *Journal of the American Medical Informatics Association*, 30(6):1091–1102.

Yang Tan, Mingchen Li, Zijie Huang, Huiqun Yu, and Guisheng Fan. 2023. Medchatzh: a better medical adviser learns from better instructions. *arXiv preprint arXiv:2309.01114*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ning Wang, Jiang Bian, Yuchen Li, Xuhong Li, Shahid Mumtaz, Linghe Kong, and Haoyi Xiong. 2024a. Multi-purpose rna language modelling with motif-aware pretraining and type-guided fine-tuning. *Nature Machine Intelligence*, pages 1–10.

Sheng Wang, Liheng Chen, Pengan Chen, Jingwei Dong, Boyang Xue, Jiyue Jiang, Lingpeng Kong, and Chuan Wu. 2024b. Mos: Unleashing parameter efficiency of low-rank adaptation with mixture of shards. *arXiv preprint arXiv:2410.00938*.

Sheng Wang, Liheng Chen, Jiyue Jiang, Boyang Xue, Lingpeng Kong, and Chuan Wu. 2024c. Lora meets dropout under a unified framework. *arXiv preprint arXiv:2403.00812*.

Sheng Wang, Boyang Xue, Jiacheng Ye, Jiyue Jiang, Liheng Chen, Lingpeng Kong, and Chuan Wu. 2024d. Prolora: Partial rotation empowers more parameter-efficient lora. *arXiv preprint arXiv:2402.16902*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. 2023a. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*.

Zeyuan Wang, Qiang Zhang, Keyan Ding, Ming Qin, Xiang Zhuang, Xiaotong Li, and Huajun Chen. 2023b. Instructprotein: Aligning human and protein language via knowledge instruction. *arXiv preprint arXiv:2310.03269*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Peng Wu, Huabin Du, Yingchao Yan, Tzong-Yi Lee, Chen Bai, and Song Wu. 2024a. Guided diffusion for molecular generation with interaction prompt. *Briefings in Bioinformatics*, 25(3):bbae174.

Xidong Wu, Yiming Zeng, Arun Das, Sumin Jo, Tinghe Zhang, Parth Patel, Jianqiu Zhang, Shou-Jiang Gao, Dexter Pratt, Yu-Chiao Chiu, et al. 2024b. regulogpt: Harnessing gpt for knowledge graph construction of molecular regulatory pathways. *bioRxiv*.

Xi Xiao, Wentao Wang, Jiacheng Xie, Lijing Zhu, Gaofei Chen, Zhengji Li, Tianyang Wang, and Min Xu. 2024. Hgtdp-dta: Hybrid graph-transformer with dynamic prompt for drug-target binding affinity prediction. *arXiv preprint arXiv:2406.17697*.

Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Ma Chang, Runcheng Liu, and Jian Tang. 2022. Peer: a comprehensive and multi-task benchmark for protein sequence understanding. *Advances in Neural Information Processing Systems*, 35:35156–35173.

Fei Ye, Zaixiang Zheng, Dongyu Xue, Yuning Shen, Lihao Wang, Yiming Ma, Yan Wang, Xinyou Wang, Xiangxin Zhou, and Quanquan Gu. 2024. Proteinbench: A holistic evaluation of protein foundation models. *arXiv preprint arXiv:2409.06744*.

Jian Ye, Scott McGinnis, and Thomas L Madden. 2006. Blast: improvements for better sequence analysis. *Nucleic acids research*, 34(suppl_2):W6–W9.

Li Yizhen, Huang Shaohan, Qi Jiaxing, Quan Lei, Han Dongran, and Luan Zhongzhi. 2024. Exploring the comprehension of chatgpt in traditional chinese medicine knowledge. *arXiv preprint arXiv:2403.09164*.

Qingchen Yu, Zifan Zheng, Shichao Song, Zhiyu Li, Feiyu Xiong, Bo Tang, and Ding Chen. 2024. xfinder: Robust and pinpoint answer extraction for large language models. *arXiv preprint arXiv:2405.11874*.

Zhao Yukun, Yan Lingyong, Sun Weiwei, Xing Guoliang, Wang Shuaiqiang, Meng Chong, Cheng Zhicong, Ren Zhaochun, and Yin Dawei. 2024. Improving the robustness of large language models via consistency alignment. *arXiv preprint arXiv:2403.14221*.

Ke Zhang, Yimiao Feng, and Jie Zheng. 2024a. Prompt-based generation of natural language explanations of synthetic lethality for cancer drug discovery. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13131–13142.

Qiang Zhang, Keyang Ding, Tianwen Lyv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, et al. 2024b. Scientific large language models: A survey on biological & chemical domains. *arXiv preprint arXiv:2401.14656*.

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

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

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

Shuxian Zou, Shentong Mo, Hui Li, Xingyi Cheng, Le Song, and Eric Xing. 2023. Linker-tuning: Optimizing continuous prompts for heterodimeric protein prediction.

A Appendix

A.1 Experiments (BioFinder)

A.1.1 Experiment Setting

As outlined in Section 3.1, evaluating LLMs presents unique challenges. Although the xFinder framework demonstrated optimal performance on diverse tasks like multiple-choice questions, short text matching, and mathematical extraction, it fell short on the Bio-benchmark dataset (Yu et al., 2024). To bridge this gap, we adapt xFinder-llama38it[1] into BioFinder, enhancing its capability for biological sequence extraction and long-text Natural Language Inference (NLI) tasks without reference answers.

For the biological sequence extraction, we manually annotate 1,428 samples from six models in Bio-benchmark, focusing on 0-shot and five-shot outputs, and iteratively retrained using GPT4o and RegEx on misextracted sequences (Contributors, 2023b). The NLI task involve collecting 13,698 samples, applying multi-round reasoning with GPT4o, and refining outputs through self-consistency checks and manual reviews (Wang et al., 2022).

Training uses advanced fine-tuning methodologies such as LoRA, PRoLoRA, QLoRA, and MoS (Hu et al., 2021; Wang et al., 2024c,d,b; Dettmers et al., 2024), specifically employing the Xtuner framework with QLoRA. Prompts are strategically designed for each task to enhance the model's adherence to instructions.

A.1.2 Baselines and Evaluation

Objective Answer Extraction For objective tasks such as multiple-choice questions, short text matching, and mathematical answer extraction, we fine-tune xFinder to enhance its bio-sequence extraction capabilities while retaining all its original functionalities. Table 2 compares the performance of BioFinder, regular expressions, and GPT-4. BioFinder achieved a 93.5% extraction accuracy

[1] https://huggingface.co/IAAR-Shanghai/xFinder-llama38it

in bio-sequence extraction tasks, exceeding regular expressions by over 30%, and delivering twice the extraction quality of GPT-4 using the same prompt. Across all objective tasks, BioFinder's average extraction accuracy reached 94.7%, comparable to single-round human review, outperforming regular expressions at 72.1% and GPT-4 at 62.9%.

Subjective Task Evaluation We evaluate subjective tasks by comparing BioFinder with GPT-4[2], RoBERTa-large-mnli[3], and GTE-large-en-v1.5[4]. The comparison involve mapping model outputs and reference answers into a shared embedding space and assessing cosine similarity. BioFinder outperform the others on a medical NLI test, achieving 89.8% accuracy compared to 59.9% for GPT-4, 27.0% for RoBERTa-large-mnli, and 23.4% for GTE-large-en-v1.5, demonstrating state-of-the-art performance.

RoBERTa-large-mnli, not initially design for long sequences (>512 tokens), required text segmentation and a bagging approach to mitigate semantic interruptions. In addition, embedding models are found to be ineffective for NLI tasks, often misclassifying antonym sentences as similar, contrary to their expected 'Contradiction' classification in NLI tasks, as detailed in Appendix B.

A.2 Experiments (Bio-Benchmark)

A.2.1 Experiment Setting

We conduct experiments on the Bio-benchmark datasets. We use Openai-API and 8*A100-80G GPUs to infer LLMs with LM-Deploy and Huggingface backend (Contributors, 2023a; hug, 2024). We employ sampling hyperparameters with top-p set to 1.0 and a temperature of 0.2 for generation (Specific prompts in the Appendix D). We use BioFinder to extract the answers of objective evaluation and execute NLI tasks.

A.2.2 Evaluation metrics

To evaluate the performance of Bio-benchmark tasks on LLMs, we use the following metrics: (1) Bit score (Ye et al., 2006; Nawrocki et al., 2009; Nawrocki and Eddy, 2013): Measures the alignment quality of sequences, normalizing the raw alignment score by database size and randomness. (2) Recovery rate: Indicates the proportion of relevant instances accurately retrieved from the total relevant instances available. (3) Accuracy (Acc.): Evaluates the proportion of correct predictions made out of the total predictions. (4) **BERTScore** (Zhang et al., 2019): We adopt BERTScore to compare the similarity between embeddings of a generated sentence and the reference sentence. (5) Rouge-L: Evaluates the similarity of text based on the longest common subsequence, focusing on sequence order and content overlap in summaries or translations. (6) GPT evaluation metrics (fluency, relevance, completeness, medicine proficiency).

A.2.3 Evaluation of Different Tasks

For more detailed Bio-benchmark experimental evaluation metrics, please see Appendix A.2.3.

1. Protein: Pfam design (bit score), protein species prediction (Acc.), protein inverse folding (recovery rate), protein structure prediction (recovery rate). **2. RNA:** RNA function prediction (Acc.), RNA inverse folding (recovery rate), RNA structure prediction (F1), Rfam design (bit score), sgRNA efficiency prediction (Acc.). **3. RBP:** RNA-binding protein (Acc.). **4. Drug:** Drug-Drug interaction (Acc.), drug-target interaction (Acc.), drug design (Acc.). **5. EHR:** Agentclinic (Acc.), CMB-Clinic (fluency, relevance, completeness, medical proficiency), IMCS-MRG (Rouge-L). **6. Medical:** HeadQA (Acc.), MedLFQA-HealthQA (COMPREHENSIVENESS), MedLFQA-KQA (COMPREHENSIVENESS), MedLFQA-LiveQA (COMPREHENSIVENESS), MedLFQA-MedicationQA (COMPREHENSIVENESS), MedMCQA (Acc.), MedQA-CN (Acc.), MedQA-TW (Acc.), MedQA-US (Acc.), MMCU (Acc.). **7. TCM:** CMB-Exam (Acc.), CMMLU-TCM (Acc.), MedLFQA-TCM (Acc.), TCMSD (Acc.).

A.3 Evaluation Tools

- **Rouge-l:** from rouge_metric import PyRouge

A.4 Results

A.5 Related Work

A.5.1 Bioinformatics Benchmark

With the emergence of LLMs for various biological tasks, there is also a growing number of benchmark methods designed for different inputs, including RNA, protein, and biomedicine-related texts. Protein is one of the first biological fields that LLM has been employed with multiple established benchmarks, including protein fitness prediction (Notin

[2] https://openai.com/index/gpt-4/

[3] https://huggingface.co/FacebookAI/roberta-large-mnli

[4] https://huggingface.co/Alibaba-NLP/gte-large-en-v1.5

Task	Subtask	Count	GPT-4o		InternLM-2.5 20b		Llama-3.1 70b		Mistral-large-2		Qwen 2.5-72b		Yi-1.5 34b	
			0-shot	5-shot	0-shot	5-shot	0-shot	5-shot	0-shot	5-shot	0-shot	5-shot	0-shot	5-shot
Protein	Protein-species-prediction	200	9.00	76.50	4.00	78.50	9.50	79.00	8.50	82.00	10.00	76.00	12.00	75.00
Protein	Protein-inverse-folding	264	6.97	6.29	1.46	4.64	6.69	6.88	5.65	6.70	7.02	6.95	4.48	5.77
Protein	Protein-structure-prediction	264	25.09	29.79	3.96	28.21	24.63	34.31	21.02	24.23	18.11	27.13	5.99	23.99
RBP	RNA-binding-protein	70	57.14	61.43	44.29	50.00	52.86	51.43	52.86	71.43	47.14	47.14	48.57	48.57
RNA	RNA-function-prediction	280	4.64	87.86	2.14	78.57	3.93	88.57	4.64	71.79	6.07	79.29	3.93	91.07
RNA	RNA-inverse-folding	200	19.76	21.71	20.64	27.19	19.50	26.25	20.62	22.90	20.28	21.92	12.99	21.82
RNA	RNA-structure-prediction	200	0.50	2.14	0.00	2.26	0.07	0.75	0.65	0.10	0.43	0.07	0.09	0.37
RNA	sgRNA-efficiency-prediction	300	0.67	39.33	36.67	0.00	1.33	0.67	7.67	0.33	0.33	20.67	0.00	13.33
Drug	Drug-Drug-interaction	86	46.51	33.72	12.79	10.47	36.05	36.05	25.58	36.05	34.88	31.40	22.09	29.07
Drug	Drug-Target-interaction	60	43.33	70.00	40.00	73.33	51.67	61.67	10.00	50.00	43.33	58.33	33.33	58.33
Drug	Drug-design	58	70.69	84.48	81.03	84.48	12.07	86.21	48.28	91.38	44.83	87.93	58.62	86.21
EHR	Agentclinic	214	82.24	62.15	63.55	69.16	79.44	80.84	78.97	78.97	73.83	77.57	61.21	67.29
EHR	CMB-Clinic	74	94.93	93.92	80.95	84.66	88.90	85.14	95.54	77.08	98.56	97.97	91.55	81.35
EHR	IMCS-MRG	200	63.02	69.02	62.82	0.00	67.21	70.21	61.56	68.64	62.98	68.48	69.20	69.40
Medical	HeadQA	120	90.00	90.83	66.67	66.67	86.67	83.33	86.67	83.33	82.50	83.33	70.83	64.17
Medical	MedLFQA-HealthQA	50	43.76	43.03	44.27	45.03	26.75	31.84	39.19	45.24	37.37	43.20	40.81	42.67
Medical	MedLFQA-KQA	50	42.34	40.41	32.73	35.89	18.12	26.76	32.93	37.55	30.82	36.25	30.77	26.88
Medical	MedLFQA-LiveQA	50	29.57	29.60	26.05	28.69	17.76	19.95	24.80	28.67	25.07	28.25	22.82	19.85
Medical	MedLFQA-MedicationQA	50	32.70	36.94	33.54	36.31	21.45	24.25	28.99	33.84	35.82	36.26	28.51	27.08
Medical	MedMCQA	100	78.00	79.00	56.00	56.00	76.00	80.00	78.00	84.00	67.00	70.00	57.00	60.00
Medical	MedQA-CN	50	82.00	82.00	82.00	80.00	84.00	82.00	70.00	72.00	88.00	90.00	82.00	84.00
Medical	MedQA-TW	50	90.00	88.00	60.00	66.00	86.00	86.00	80.00	76.00	82.00	82.00	62.00	64.00
Medical	MedQA-US	50	92.00	80.00	52.00	54.00	86.00	84.00	84.00	80.00	84.00	68.00	46.00	48.00
Medical	MMCU	142	59.15	61.27	35.92	38.03	62.68	59.15	48.59	46.48	62.68	63.38	46.48	49.30
TCM	CMB-Exam	200	56.00	60.50	62.50	61.50	50.50	51.00	43.50	45.50	59.50	64.00	54.50	68.50
TCM	CMMLU-TCM	185	72.97	73.51	80.00	82.70	65.41	71.35	61.62	57.84	82.70	83.24	78.38	83.78
TCM	MedLFQA-TCM	200	72.00	73.50	82.00	85.00	63.00	64.50	59.00	61.50	82.50	90.00	75.50	87.50
TCM	TCMSD	200	39.00	68.75	30.00	60.75	16.00	69.25	32.25	67.00	38.25	67.25	34.50	59.00

Table 2: Performance of Various LLMs on Bioinformatics Tasks

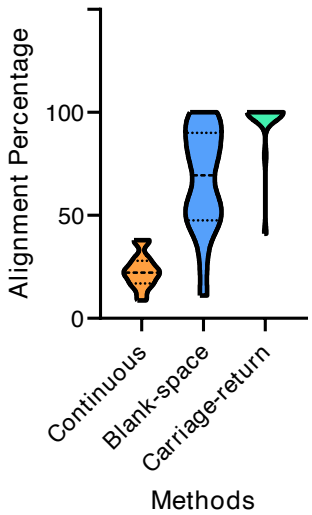

Figure 4: The proportion of Alignment by different methods.

et al., 2023), protein design (Ye et al., 2024), and multi-task ones (Xu et al., 2022). As for the more recently studied RNA LLMs, benchmarks have been created to include various structure-, function-, and engineering-related tasks (Runge et al., 2024; Ren et al., 2024). For LLMs working with biomedical textual data, there have also been various benchmark datasets, focusing on EHR (Bae et al., 2024; Soni et al., 2023), medicine QA (Jin et al., 2021; Pal et al., 2022; Jin et al., 2019), and specifically traditional Chinese medicine QA (Tan et al., 2023; Li et al., 2023a).

A.5.2 Answer Quality Assessment

Bio-benchmark dataset encompasses diverse tasks with complex answer formats, including objective tasks like Multiple-Choice Questions (MCQA) (Robinson et al., 2022; He et al., 2024a), short text matching, mathematical inference, and biological sequence prediction, as well as subjective tasks such as long text generation and open-ended QA. Consequently, the answer formats generated by LLMs are highly complex and varied.

LLMs' outputs are often uncontrollable; the output of different LLMs could differ significantly even with the same prompt (Yukun et al., 2024; Gu et al., 2022). Traditional evaluation frameworks attempt to constrain outputs through prompts, but this can negatively impact generation accuracy and reasoning ability (Röttger et al., 2024). For complex tasks, LLMs' instruction-following ability decreases, leading to answers that do not adhere to specific formats (Asai et al., 2023).

Traditional evaluation frameworks like LM Eval Harness and OpenCompass typically use regular expressions (RegEx) or Judge models to extract answers (Gao et al., 2024; Contributors, 2023c;

	SMILES	GT_C	GT_N	GT_O	Pred_C	Pred_N	Pred_O
1	OC(C1=CC=CC=C1)(C1=CC=CC=C1)C12CC[N+](CCOCC3=CC=CC=C3)(CC1)CC2	29	1	2	29	1	2
2	[H][C@@]12C[C@@H](C)[C@H](C(=O)CO)[C@@]1(C)C[C@H](O)[C@@]1(F)[C@@]2([H])CCC2=CC(=O)C=C[C@]12C	22	0	4	20	0	3
3	CC(C)C1=C(C(=O)NC2=CC=CC=C2)C(=C(N1CC[C@@H](O)C[C@@H](O)CC(O)=O)C1=CC=C(F)C=C1)C1=CC=CC=C1	33	2	5	29	2	5
4	C[C@H](CC1=CC2=C(N(CCCO)CC2)C(=C1)C(N)=O)NCCOC1=CC=CC=C1OCC(F)(F)F	25	3	4	25	3	4
5	CCN[C@@H]1C[C@H](N)[C@@H](O[C@H]2OC(CN)=CC[C@H]2N)[C@H](O)[C@H]1O[C@H]1OC[C@](C)(O)[C@H](NC)[C@H]1O	21	5	7	21	5	9
6	CCCOC1=C(C=C(C=C1)S(=O)(=O)NCCC1CCCN1C)C1=NC(=O)C2=C(N1)C(CCC)=NN2C	25	6	4	24	6	4
7	CC\\C(=C(/C1=CC=CC=C1)C1=CC=C(OCCN(C)C)C=C1)C1=CC=CC=C1	26	1	1	24	1	1
8	OC1=C(CC2=C(O)C3=C(OC2=O)C=CC=C3)C(=O)OC2=C1C=CC=C2	19	0	6	21	0	5

Figure 5: The proportion of Alignment by different methods.

Dubois et al., 2024; Zheng et al., 2023), aiming to cover various output formats. This approach requires significant effort in designing RegEx patterns and struggles to balance false positives and omissions (Zhu et al., 2023; Wang et al., 2023a). Moreover, the generality and accuracy of judge models are significantly inferior to advanced LLMs or human evaluation (Huang et al., 2024), and the cost of using advanced LLMs or human evaluators on large datasets is prohibitively high. To address these challenges, xFinder proposed a framework that uses LLMs for inference, achieving excellent results in answer extraction (Yu et al., 2024). However, this framework's capability and applicability in bioinformatics need enhancement; it cannot extract answers from data without reference answers and currently cannot assess the quality of long-text responses.

A.5.3 Bioinformatics Prompting

Biological Sequences In bioinformatics, prompting techniques have been applied to both protein and nucleic acid sequences. For DNA sequences, template-based prompts generate natural language explanations of gene interactions, benefiting various aspects like the interpretability of synthetic lethality predictions for drug discovery (Petroni et al., 2019; Liu et al., 2021; Zhang et al., 2024a). Soft prompts integrate DNA sequences into tunable templates (Li et al., 2023b), improving the performance of models like DNABERT (Ji et al., 2021) in recognizing promoter sequences. For RNA sequences, CoT and repeated prompt have been applied on RNA-seq-based cell type annotation with GPT-3 and GPT-4 (Hou and Ji, 2024). For protein sequences, interaction prompts (Wu et al., 2024a) and continuous prompts (Li and Liang, 2021; Zou et al., 2023) aid in predicting protein structures and interactions by embedding task-specific information directly into the input sequences. Additionally, protein CoT has also been applied in predicting nonphysical protein-protein interaction (Wang et al., 2023b).

Drugs Prompting methods are also utilized in drug-related tasks, including drug-target binding affinity prediction: dynamic prompt generation captures unique interactions between drugs and their targets by integrating context-specific prompts with molecular features (Xiao et al., 2024). Additionally, latent vector prompts enhance molecular design by incorporating continuous prompts into transformer models through cross-attention mechanisms (Kong et al., 2024). In-context learning strategies, such as those used for predicting synergistic drug combinations, leverage masking and graph representations to improve personalized drug synergy predictions (Edwards et al., 2023).

Biological Textual Data Prompting techniques extend to biological textual data, including EHR and biomedical QA systems. In EHRs, explanatory prompts streamline clinical documentation by contextualizing each section, reducing the clinicians' workload (He et al., 2024b). Prompt engineering methods, including paraphrasing and persona-based instructions, enhance the effectiveness of LLM embeddings for medical diagnostics and prognostics (Gao et al., 2024; Chen et al., 2023). In biomedical QA systems, strategies like CoT (Chen et al., 2024; Wu et al., 2024b) and Graph of Thoughts (GoT) (Hamed et al., 2024) can help improve reasoning abilities and reduce misinformation in responses. Specialized prompts for TCM QA systems incorporate domain-specific knowledge and examples, optimizing model performance in specialized medical contexts (Yizhen et al., 2024).

B Appendix B

We evaluated the cosine similarities among various sentences embedded using the GTE-large-

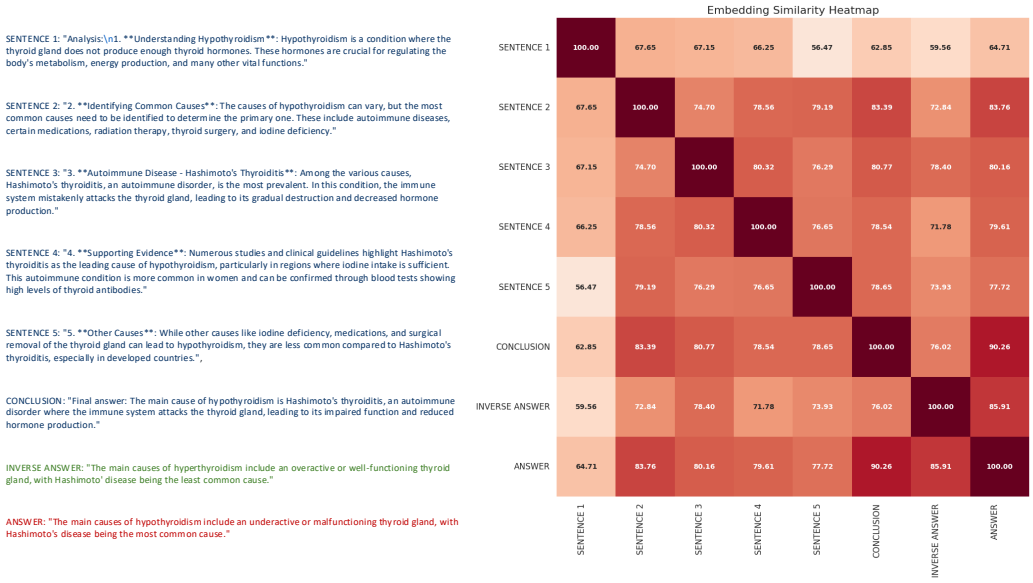

Figure 6: Cosine similarities between sentence embeddings using the GTE-large-en-v1.5 model. SENTENCE1-CONCLUSION is the LLM's response, ANSWER is the reference answer, and INVERSE ANSWER is the negation of the reference answer. The high similarity between the reference answer and its inverse suggests limitations in the embedding model's ability to distinguish logical negations.

en-v1.5 model. Specifically, SENTENCE1-CONCLUSION represents the LLM's response, ANSWER denotes the reference answer, and INVERSE ANSWER is the negation of the reference answer. As illustrated in Figure 6, when the reference answer is negated, the similarity between the reference answer and its inverse is second only to the similarity between the CONCLUSION and the reference answer. This observation indicates that the embedding model does not effectively capture the logical and semantic relationships between sentences.

C Overall evaluation metric values

C.1 Bioinformatics Sequence

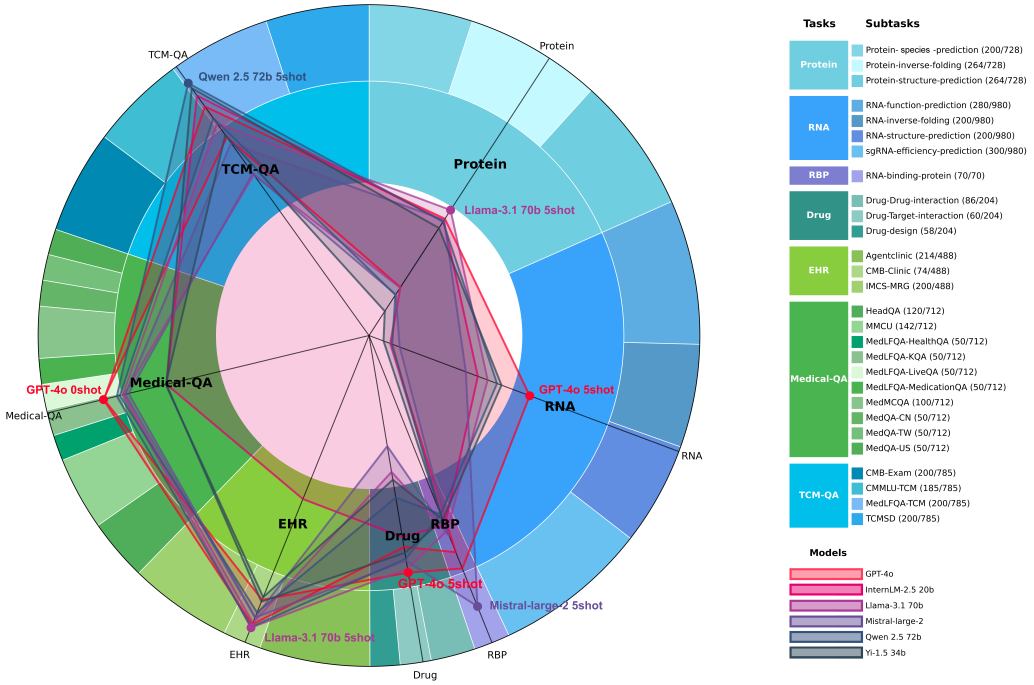

Figure 7: Overall evaluation metric values

D Prompt Template in Inference

You are a leading expert in {category}. Carefully analyze the following question and provide a step-by-step step solution.

Question: [**QUESTION**]

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]

2. [Second key point or step]

3. [Third key point or step]

(Add more steps if necessary)

Final answer: [Your concise and accurate answer based on the analysis above]

Ensure your analysis is thorough and your final answer is precise and directly addresses the question.

You are a leading expert in drug design. Carefully analyze the following question and provide a step-by-step step solution.

Example 1-5:

Question: [*EXAMPLE_QUESTION*]

Answer: [*EXAMPLE_ANSWER*]

Question: [**QUESTION**]

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]

2. [Second key point or step]

3. [Third key point or step]

(Add more steps if necessary)

Final answer: [Your concise and accurate answer based on the analysis above]

Ensure your analysis is thorough and your final answer is precise and directly addresses the question.

You are a leading expert in pharmacology. Carefully analyze the following question about drug-drug interactions and provide a step-by-step step solution.

Example 1-5:

Question: *[EXAMPLE_QUESTION]*

Answer: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]

2. [Second key point or step]

3. [Third key point or step]

(Add more steps if necessary)

Final answer: [Your concise and accurate answer based on the analysis above]

Ensure your analysis is thorough and your final answer is precise and directly addresses the question.

You are a leading expert in molecular biology. Carefully analyze the following question about drug-target interactions and provide a step-by-step step solution.

Example 1-5:

Question: *[EXAMPLE_QUESTION]*

Answer: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]

2. [Second key point or step]

3. [Third key point or step]

(Add more steps if necessary)

Final answer (1 or 0):[Your concise and accurate answer based on the analysis above; output 1 if there is binding affinity, or 0 if there is not]

Ensure your analysis is thorough and your final answer is precise and directly addresses the question.

You are a leading expert in protein biology. Carefully analyze the following sequence and predict its function by providing a step-by-step step solution.

Example 1-5:

Sequence: *[EXAMPLE_QUESTION]*

Function: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]

2. [Second key point or step]

3. [Third key point or step]

(Add more steps if necessary)

Predicted function: [Your concise and accurate prediction based on the analysis above]

Ensure your analysis is thorough and your final answer is precise and directly addresses the sequence.

You are a leading expert in structural biology. Carefully analyze the following protein secondary structure notation and generate the corresponding amino acid sequence by providing a step-by-step step solution.

Example 1-5:

Secondary Structure: *[EXAMPLE_QUESTION]*

Sequence: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]

2. [Second key point or step]

3. [Third key point or step]

(Add more steps if necessary)

Generated amino acid sequence: [Your generated sequence based on the analysis above]

Ensure your analysis is thorough and your generated sequence is precise and directly addresses the secondary structure.

You are a leading expert in protein structure. Carefully analyze the following amino acid sequence and predict its secondary structure using the sst8 notation by providing a step-by-step step solution.

Example 1-5:

Protein Sequence: *[EXAMPLE_QUESTION]*

Secondary Structure: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]

2. [Second key point or step]

3. [Third key point or step]

(Add more steps if necessary)

Predicted secondary structure (sst8 notation): [Your prediction based on the analysis above]

Ensure your analysis is thorough and your predicted structure is precise and directly addresses the sequence.

You are a leading expert in RNA biology. Carefully analyze the following RNA sequence and predict if it can bind to the specified protein by providing a step-by-step step solution.

Example 1-5:

Sequence: *[EXAMPLE_QUESTION]*

Binding Affinity: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]

2. [Second key point or step]

3. [Third key point or step]

(Add more steps if necessary)

Binding prediction (true/false): [Your prediction based on the analysis above]

Ensure your analysis is thorough and your prediction is precise and directly addresses the sequence.

You are a leading expert in RNA biology. Carefully analyze the following RNA sequence and predict its function by providing a step-by-step step solution.

Example 1-5:

Sequence: *[EXAMPLE_QUESTION]*

Function: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]

2. [Second key point or step]

3. [Third key point or step]

(Add more steps if necessary)

Predicted function: [Your prediction based on the analysis above]

Ensure your analysis is thorough and your predicted function is precise and directly addresses the sequence.

You are a leading expert in RNA structure. Carefully analyze the following dot-bracket notation of the RNA secondary structure and generate the corresponding RNA sequence by providing a step-by-step step solution.

Example 1-5:

Secondary Structure: *[EXAMPLE_QUESTION]*

RNA Sequence: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]

2. [Second key point or step]

3. [Third key point or step]

(Add more steps if necessary)

Generated RNA sequence: [Your generated sequence based on the analysis above]

Ensure your analysis is thorough and your generated sequence is precise and directly addresses the secondary structure.

You are a leading expert in RNA folding. Carefully analyze the following RNA sequence and predict its secondary structure using dot-bracket notation by providing a step-by-step step solution.

Example 1-5:

RNA Sequence: *[EXAMPLE_QUESTION]*

Secondary Structure: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]

2. [Second key point or step]

3. [Third key point or step]

(Add more steps if necessary)

Predicted secondary structure (dot-bracket notation): [Your prediction based on the analysis above]

Ensure your analysis is thorough and your predicted structure is precise and directly addresses the sequence.

You are a leading expert in CRISPR technology. Carefully analyze the following sgRNA sequence and predict its efficiency by providing a step-by-step step solution.

Example 1-5:

Sequence: *[EXAMPLE_QUESTION]*

Efficiency (%): *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]

2. [Second key point or step]

3. [Third key point or step]

(Add more steps if necessary)

Predicted efficiency (%): [Your prediction as a percentage between 0 and 100 based on the analysis above]

Ensure your analysis is thorough and your predicted efficiency is precise and directly addresses the sequence.

You are a medical expert. Analyze the following patient information, physical examination findings, and test results, then provide a diagnosis.

Example 1-5:

Patient Information: *[EXAMPLE_QUESTION]*

Diagnosis: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Provide a detailed analysis and your final diagnosis. Your response should be structured as follows:

Analysis:

1. [Key findings from patient history]
2. [Significant physical examination results]
3. [Relevant test results]
4. [Differential diagnosis considerations]

Final Diagnosis: [Your concise and accurate diagnosis based on the analysis above]

Ensure your analysis is thorough and your final diagnosis is precise and directly addresses the patient's presentation.

You are a medical expert. Read the following patient information and answer the questions.

Example 1-5:

Case: *[EXAMPLE_QUESTION]*

Answer: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Provide detailed answers.

You are a medical expert. Based on the following patient description and doctor-patient conversation, generate a diagnostic report.

Example 1-5:

Patient Description and Conversation: *[EXAMPLE_QUESTION]*

Diagnostic Report: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Provide the diagnostic report in the following format:

{Provide the report format as per the data}

Answer the following multiple-choice question.

Example 1-5:

Medical question: *[EXAMPLE_QUESTION]*

Correct answer: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Provide the correct option.

Answer the following multiple-choice question related to Traditional Chinese Medicine.

Example 1-5:

Medical question: *[EXAMPLE_QUESTION]*

Correct answer: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Provide the correct option.

Answer the following multiple-choice question.

Example 1-5:

Medical question: *[EXAMPLE_QUESTION]*

Correct answer: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Provide the correct option.

Using Traditional Chinese Medicine diagnostics, determine the patient's disease and syndrome based on the following information.

Example 1-5:

Medical question: *[EXAMPLE_QUESTION]*

Correct answer: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Provide your answer in the following format:

Disease:

Syndrome:

You are a leading expert in medical sciences. Carefully analyze the following multiple-choice question and provide the correct answer along with your reasoning.

Example 1-5:

Medical Question: *[EXAMPLE_QUESTION]*

Correct Answer: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]
2. [Second key point or step]
3. [Third key point or step]

(Add more steps if necessary)

Final answer: [Your concise and accurate answer based on the analysis above]

Ensure your analysis is thorough and your final answer is precise and directly addresses the question.

You are a medical expert. Carefully analyze the following health-related question and provide a detailed answer.

Example 1-5:

Health Question: *[EXAMPLE_QUESTION]*

Answer: *{free_form_answer}*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]

2. [Second key point or step]

3. [Third key point or step]

(Add more steps if necessary)

Final answer: [Your detailed answer based on the analysis above]

Ensure your analysis is thorough and your final answer is comprehensive and directly addresses the question.

You are a medical expert. Carefully analyze the following knowledge-based question and provide a detailed answer.

Example 1-5:

Knowledge Question: *[EXAMPLE_QUESTION]*

Answer: *{free_form_answer}*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]

2. [Second key point or step]

3. [Third key point or step]

(Add more steps if necessary)

Final answer: [Your detailed answer based on the analysis above]

Ensure your analysis is thorough and your final answer is comprehensive and directly addresses the question.

You are a medical expert. Carefully analyze the following live QA question and provide a comprehensive answer.

Example 1-5:

Live QA Question: *[EXAMPLE_QUESTION]*

Answer: *{free_form_answer}*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]

2. [Second key point or step]

3. [Third key point or step]

(Add more steps if necessary)

Final answer: [Your comprehensive answer based on the analysis above]

Ensure your analysis is thorough and your final answer is detailed and directly addresses the question.

You are a medical expert. Carefully analyze the following question about medication and provide a detailed answer.

Example 1-5:

Medication Question: *[EXAMPLE_QUESTION]*

Answer: *{free_form_answer}*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]

2. [Second key point or step]

3. [Third key point or step]

(Add more steps if necessary)

Final answer: [Your detailed answer based on the analysis above]

Ensure your analysis is thorough and your final answer is comprehensive and directly addresses the question.

You are a leading expert in medical sciences. Carefully analyze the following multiple-choice question and provide the correct answer along with your reasoning.

Example 1-5:

Medical Question: *[EXAMPLE_QUESTION]*

Correct Answer: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]
2. [Second key point or step]
3. [Third key point or step]

(Add more steps if necessary)

Final answer: [Your concise and accurate answer based on the analysis above]

Ensure your analysis is thorough and your final answer is precise and directly addresses the question.

You are a leading expert in Chinese medical sciences. Carefully analyze the following multiple-choice question and provide the correct answer along with your reasoning.

Example 1-5:

Medical Question (CN): *[EXAMPLE_QUESTION]*

Correct Answer: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]
2. [Second key point or step]
3. [Third key point or step]

(Add more steps if necessary)

Final answer: [Your concise and accurate answer based on the analysis above]

Ensure your analysis is thorough and your final answer is precise and directly addresses the question.

You are a leading expert in Taiwanese medical sciences. Carefully analyze the following multiple-choice question and provide the correct answer along with your reasoning.

Example 1-5:

Medical Question (TW): *[EXAMPLE_QUESTION]*

Correct Answer: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]
2. [Second key point or step]
3. [Third key point or step]

(Add more steps if necessary)

Final answer: [Your concise and accurate answer based on the analysis above]

Ensure your analysis is thorough and your final answer is precise and directly addresses the question.

You are a leading expert in US medical sciences. Carefully analyze the following multiple-choice question and provide the correct answer along with your reasoning.

Example 1-5:

Medical Question (US): *[EXAMPLE_QUESTION]*

Correct Answer: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]
2. [Second key point or step]
3. [Third key point or step]

(Add more steps if necessary)

Final answer: [Your concise and accurate answer based on the analysis above]

Ensure your analysis is thorough and your final answer is precise and directly addresses the question.

You are a leading expert in fundamental medical sciences. Carefully analyze the following multiple-choice question and provide the correct answer along with your reasoning.

Example 1-5:

MMCU Medical Question: *[EXAMPLE_QUESTION]*

Correct Answer: *[EXAMPLE_ANSWER]*

Question: **[QUESTION]**

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]

2. [Second key point or step]

3. [Third key point or step]

(Add more steps if necessary)

Final answer: [Your concise and accurate answer based on the analysis above]

Ensure your analysis is thorough and your final answer is precise and directly addresses the question.

You are a leading expert in protein. Carefully analyze the following question and provide a step-by-step step solution.

Question: [**QUESTION**]

The example sequences of this family are:

Example1-10:

[Example Protein Sequence]

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]
2. [Second key point or step]
3. [Third key point or step]

(Add more steps if necessary)

Final answer: [Your concise and accurate answer based on the analysis above]

Ensure your analysis is thorough and your final answer is precise and directly addresses the question.

You are a leading expert in RNA. Carefully analyze the following question and provide a step-by-step step solution.

Question: [**QUESTION**]

The example sequences of this family are:

Example1-10:

[Example RNA Sequence]

Your response should be structured as follows:

Analysis:

1. [First key point or step in your reasoning]
2. [Second key point or step]
3. [Third key point or step]

(Add more steps if necessary)

Final answer: [Your concise and accurate answer based on the analysis above]

Ensure your analysis is thorough and your final answer is precise and directly addresses the question.