PQA: Zero-shot Protein Question Answering for Free-form Scientific Enquiry with Language Models

Eli M Carrami[*]
London, UK

Sahand Sharifzadeh
LMU Munich
Munich, Germany

Abstract

Understanding protein structure and function is crucial in biology. However, current computational methods are often task-specific and resource-intensive. To address this, we propose zero-shot Protein Question Answering (PQA), a task designed to answer a wide range of protein-related queries without task-specific training. The success of PQA hinges on high-quality datasets and robust evaluation strategies, both of which are lacking in current research. Existing datasets suffer from biases, noise, and lack of evolutionary context, while current evaluation methods fail to accurately assess model performance. We introduce the Pika framework to overcome these limitations. Pika comprises a curated, debiased dataset tailored for PQA and a biochemically relevant benchmarking strategy. We also propose multimodal large language models as a strong baseline for PQA, leveraging their natural language processing and knowledge. This approach promises a more flexible and efficient way to explore protein properties, advancing protein research. Our comprehensive PQA framework, Pika, including dataset, code, and model checkpoints, is openly accessible on Github, promoting wider research in the field.

1 Introduction

Proteins, essential to biological functions, are complex macromolecules that perform a myriad of cellular roles determined by their complex structures and interactions. As a polymer of amino acids, drawn from a pool of 20 natural ones, the sequence in which these building blocks are arranged dictates the protein's three-dimensional structure, which is critical for its function. Given the critical role of proteins in both fundamental biology and applied biomedical research, a deep understanding of their structures and functions is crucial. Despite significant advances in deciphering proteins' 3D structures, there remains a pressing need for innovative methodologies to facilitate the computational study of their biochemical and functional properties.

Currently, training individual models tailored to specific tasks are the primary approach to computationally studying the biochemical and functional properties of proteins, requiring extensive data collection and training for each unique task. For instance, submissions to the Critical Assessment of Protein Function Annotation algorithms (CAFA) aim to predict the functional annotations such as GO terms for new protein sequences, as required by the multi-year challenge (Function-SIG, 2024). Similarly, several other task-specific models have been developed to predict specific biochemical properties of proteins, such as ligand binding (Wei et al., 2022) or thermal stability (Blaabjerg et al., 2023). To address the limitations of current approaches, we propose to unify protein sequence-related enquiries under a more generic task of zero-shot Protein Question Answering (PQA), where there can be free-form inquiries about known or novel protein sequences.

[*] Corresponding Author: `eli.carrami@gmail.com`

AI for New Drug Modalities at NeurIPS 2024.

The development and assessment of effective zero-shot PQA models critically depend on the availability of high-quality **datasets** and robust **evaluation strategies**, both of which present significant challenges in the current research landscape; unfortunately, existing datasets available for related tasks often exhibit significant limitations. They frequently demonstrate biases towards specific protein families, overlook the critical evolutionary relationships between proteins, and may contain noisy or unreliable annotations. These shortcomings hinder the development of effective PQA models that can generalize across diverse proteins and accurately answer complex queries.

Similarly, the robust evaluation of PQA models necessitates biochemically-relevant benchmarking strategies that focus on the scientific accuracy of model predictions within the context of the specific questions posed. Previous research efforts have not adequately addressed these challenges. They often neglect the impact of evolutionary relationships on information leakage, which can lead to overestimation of model performance. Moreover, they predominantly rely on metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), which have been consistently shown to be inadequate for assessing the accuracy of scientific statements (Mathur et al., 2020). This lack of suitable evaluation tools further impedes progress in the field.

To address these limitations, we have developed the Pika framework, comprised of a curated and debiased dataset specifically designed for the PQA domain, equipped with scientific question and answer (QA) pairs for instructional training as well as a robust and biochemically relevant benchmarking strategy to enable effective evaluation of PQA models. Alongside, we also propose multimodal large language models (LLMs) as a potential solution to PQA and create evolution-aware splits for assessment of their performance. Besides their capability in processing natural questions, LLMs encapsulate a large body of knowledge which could provide further context for the model, enabling it to perform a more flexible and efficient exploration of protein functionalities, bypassing the need for extensive model training and data collection for individual queries.

Finally, comparing the performance of multimodal LLMs with various relevant baselines on our evolution-aware data splits, we show that these models are a promising direction for PQA. In particular, we train and evaluate two multimodal protein-text architectures combining the ESM2 (Lin et al., 2022) protein language model (PLM) with the Phi-2 LLM (Microsoft, 2023), showcasing the seamless integration of protein sequence analysis with natural language processing while highlight the shortcomings of these models in particular when dealing with evolutionary distant proteins. Our results suggest that the strategic adoption of this methodology, especially with more advanced LLMs, has the potential to challenge the current state-of-the-art in task-specific models. It is essential to highlight that, our research, leveraging the robust yet modestly scaled Phi-2 LLM, serves as a robust proof-of-concept to demonstrate the potential of this approach. We hope our work fuels further research and opens new avenues towards performant PQA models.

2 Related Work

Since PQA is distinct from other question-answering tasks, existing scientific QA datasets such as ScienceQA (Lu et al., 2022), PubMedQA (Jin et al., 2019) or SQuAD (Rajpurkar et al., 2016, 2018) are not suitable for PQA training or evaluation. This is because unlike standard QA tasks where the answer to the question can be inferred using logic or existing knowledge, in PQA the answer to the question must be extracted from cross-modality embeddings via the LLM, therefore necessitating specialized protein-text datasets that annotate protein sequences with relevant labels.

Previous studies on integrating text and protein sequences have primarily leveraged large-scale unstructured biomedical text or knowledge graphs to refine protein representations, facilitating downstream tasks like functional classification and sequence generation (e.g. ProtST (Xu et al., 2023), OntoProtein (Zhang et al., 2022)). Alongside, ProteinChat (Guo et al., 2023) has utilized LLMs for knowledge retrieval to facilitate discussions about 3D structure databases, and concurrent to our work, Mol-Instructions (Fang et al., 2024) focuses on improving LLM's understanding of biomolecules (e.g., proteins) by fine-tuning existing LLMs. These recent efforts while introducing valuable protein-text datasets have limitations that make them unsuitable for scientific PQA, as detailed below:

1. **ProtDescribe (Xu et al., 2023):** This dataset contains textual annotations from three fields of uncurated SwissProt entries, risking bias and data leakage due to high sequence similarity and over-representation of certain protein families (Fig. B.1). Also, its use of fields that are not directly inferable from protein sequence alone (e.g., cellular interaction) limits its utility for PQA.

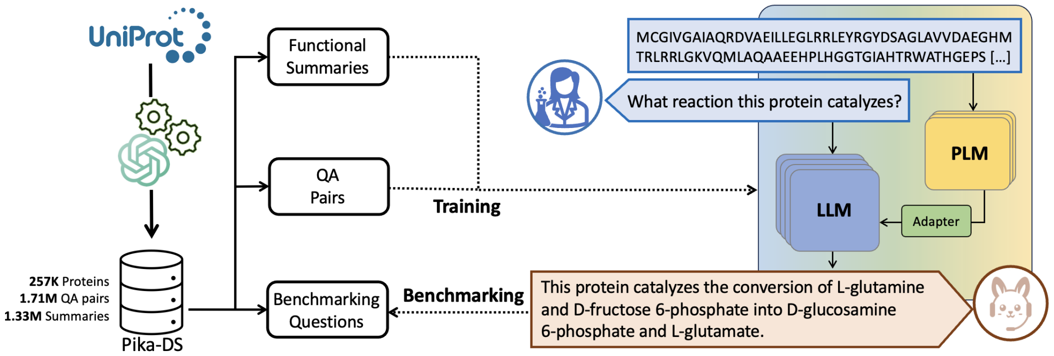
UniProt
Functional Summaries
QA Pairs
Training
257K Proteins
1.71M QA pairs
1.33M Summaries
Pika-DS
Benchmarking Questions
Benchmarking
MCGIVGAIAQRDVAEILLEGLRRLEYRGYDSAGLAVVDAEGHM
TRLRRLGKVQMLAQAAEEHPLHGGTGIAHTRWATHGEPS [...]
What reaction this protein catalyzes?
PLM
LLM
Adapter
This protein catalyzes the conversion of L-glutamine and D-fructose 6-phosphate into D-glucosamine 6-phosphate and L-glutamate.

Figure 1: Schematic of Pika framework. Pika-DS is created from filtered SwissProt entries followed by processing using GPT3.5.

Conversely, it misses out on features like catalytic reactions which are often subject of scientific queries. Crucially, lacking QAs, it fails to directly support PQA assessments. Therefore, while valuable for enhancing protein representations and text-to-protein generation, as intended by the authors, this dataset is not suitable for the scientific PQA.

2. **PDB-QA (Guo et al., 2023) :** With a limited set of 30 predefined QAs on 3D structures from the PDB database, while suitable for knowledge retrieval tasks, its focus on entry-specific details (e.g., submission date, analysis software) renders it ineffective in the context of scientific PQA.
3. **Mol-Instructions (Fang et al., 2024):** This dataset was released concurrent to our work. The protein section of Mol-Instructions represents a relevant set of template-based textual annotations for proteins tailored towards five downstream tasks, including question-answering. The dataset is curated from SwissProt and is debiased using a 90% similarity threshold. This lenient threshold results in an abundance of highly related proteins, which leads to bias and in the absence of an evolution-aware splitting strategy could cause leakage across data splits.

Furthermore, in all previous and concurrent research authors have relied on BLEU or ROUGE metrics for evaluating the scientific accuracy of generate captions or responses. However, these metrics are not suitable for assessing accuracy (Mathur et al., 2020), therefore limiting the scientific scope of past research in PQA domain. As a result, due to the lack of relevant benchmarking strategies as well as biases and noise in existing datasets, the task of free-form zero-shot scientific enquiry of new protein sequences remains unexplored.

3 Pika Framework

Here we detail our dataset, benchmarking and baseline designs for the Pika framework (Fig. 1). All baselines, model architectures and benchmarking were implemented in PyTorch Lightning , and the complete codebase is accessible on github.com/EMCarrami/Pika.

3.1 Pika Dataset

The scientific PQA task is aimed at delivering accurate responses to free-form questions based on an unseen protein sequence. This task emerges from the need for scientific exploration of protein functions via natural language question answering. Therefore, the training and evaluation of multimodal models for the PQA task necessitates comprehensive datasets with scientific textual annotations linked to corresponding protein sequences. We deemed the following three criteria as essential for a specialized PQA dataset:

1. Offers an unbiased representation of known protein sequences, mitigating frequency biases prevalent in existing databases.
2. The dataset ensures that the information associated with each protein sequence is expertly curated, allowing for inference solely based on the protein sequence.
3. Supports relevant benchmarking to assess model performance especially in zero-shot settings.

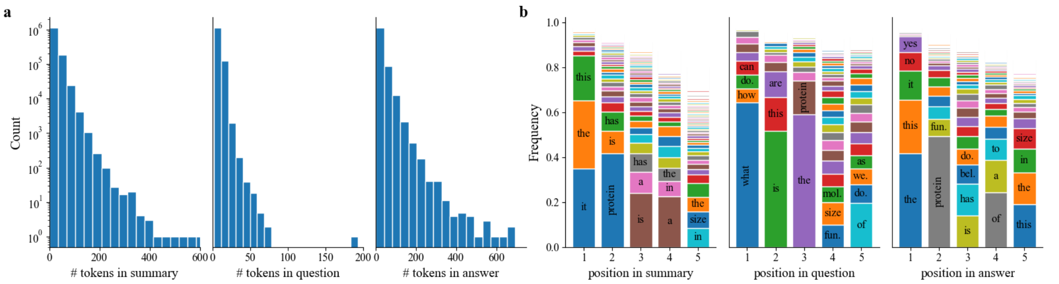

Figure 2: Characteristics of PQA dataset. (a) Distribution of token counts for all examples in Pika-DS. (b) Frequency of words in each position in each section of the dataset. Long words are abbreviated (do.=does, mol.=molecule, fun.=function, we.=weight, bel.=belong).

Since, as discussed in section 2, none of the existing datasets meet all these criteria we created Pika-DS (See sections A.1, A.2, A.3 for details), the first specialized and debiased PQA dataset, accompanied with respective biologically relevant benchmarks for model training and evaluation (summary statistics in Table B.1 and example in Table B.2).

Briefly, we gathered all SwissProt entries from UniProt database (Consortium, 2022) and extracted an expert-curated list of scientific information fields covering a wide range of relevant properties. We then debiased the sequences using a strict 50% similarity threshold. Finally, we employed GPT3.5 API to process the information fields for each protein entry using systematically optimized prompts (Sections A.4 and A.5) to create the Pika-DS's three main components:

- *Summary:* A summary of each protein's functional and biochemical properties, based solely on the provided information excluding the protein's name.
- *QAs:* Several diverse QA pairs for each information field, formatted for LLM training.
- *Metrics:* Single-word answers to a set of predefined scientific questions serving as the ground-truth for our Biochem-Lite benchmarks (Section 3.2.2).

3.1.1 Pika-DS Quality Control

The final Pika-DS comprises 257,167 protein sequences, selected from 185,128 UniRef50 clusters. This dataset is enriched with detailed descriptors for each sequence: a summary statement divided into sentences, multiple scientific QA pairs, and the ground-truth answers for our predefined benchmarking questions. This dataset encapsulates over 105 million protein sequence tokens, approximately 36.4 million tokens in textual summaries, and nearly 47 million tokens in QA pairs (Table B.1). To ensure the high quality of Pika-DS and suitability of generated textual annotations for training multimodal LLMs we assessed various aspects of our dataset.

Human evaluation of GPT3.5 generated annotations: Given the potential limitations of GPT3.5 generated annotations (e.g., hallucination), we conducted a thorough human evaluation of the Pika-DS. For this, 100 randomly selected examples from Pika-DS were manually evaluated by an expert biochemist. Our analysis revealed that out of a total of 1204 summary sentences and QA pairs, 5 were incorrect (0.4%), 21 were of poor quality (1.7%), and 43 were irrelevant (3.6%), resulting in over 94% of the annotations being correct, relevant, and of high quality. Furthermore, only 5 out of 660 metric ground truths were found to be incorrect, yielding a 99.2% accuracy in ground truth metric values. Furthermore, consultations with two expert biologists confirmed that the QA pairs generated by GPT-3.5 in Pika-DS are relevant and scientifically sound, closely matching what they would derive from the provided input fields. They also unanimously agreed that GPT-3.5 questions offer greater diversity compared to their potential queries. These results underscore the effectiveness of our prompt optimization approach for generating high-quality labels using GPT3.5 based on SwissProt information fields and is consistent with a recent study that has found LLM-generated captions can exhibit higher diversity than those created by humans, leading to enhanced training of multimodal LLMs (Sharifzadeh et al., 2024).

Token Count Analysis: In creating the QA pairs using GPT3.5, we instructed the model to cover a broad spectrum of queries and to produce detailed answers suitable for training other LLMs, anticipating elaborate rather than single-word answers. This was confirmed by our analysis of token counts in the summary sentences, questions, and answers of the final dataset. This revealed that while questions are typically shorter (8.5 tokens on average), the summary sentences and answers are longer and more extensive (on average 27 and 16.5 tokens). This presents an adequate number of tokens per example, providing sufficient context for the training of multimodal PQA models (Figure 2a).

Word Frequency Analysis: To ensure the diversity of content in summary statements as well as QA pairs, we examined the frequency of words in the initial five positions of each of these categories. As anticipated, common words such as "the" and "it", in summaries, and interrogatives like "What" and "How" in questions, dominated the first three positions. Similarly, answers often began with "the" or "this" as well as "yes/no" structures, followed by "protein". However, the remaining positions demonstrated a significant lexical diversity. This observation confirmed that our dataset does not exhibit a substantial bias, offering the necessary diversity for effective model training (Figure 2b).

Protein over-representation Analysis: We used pre-computed protein embeddings from UniProt database and visualized them using UMAP (McInnes et al., 2018), highlighting the entries that belong to the top 100 largest clusters. Visual comparison of distribution of protein embeddings before and after our Uniref50-based filtering confirms a strong reduction of over-represented protein families in Pika-DS while maintaining the rich diversity of the dataset (Figure B.2).

3.2 Pika Benchmarks

In this section, we elaborate on our biochemically-focused benchmarking methodologies tailored for evaluating the scientific accuracy of multimodal PQA models.

3.2.1 Motivation & Design Criteria

Conventional linguistic metrics like BLEU and ROUGE, while useful in general linguistic contexts, often fall short in assessing scientific correctness and show poor correlation with human judgment (Mathur et al., 2020). As a result these metrics are inadequate for assessing the performance of multimododal PQA models. Therefore, going beyond standard linguistic evaluations, we designed a purpose-built benchmarking approach incorporating a set of predefined, biochemically-significant questions, that are specifically selected to:

1. Reflect the biochemical properties of proteins, ensuring that a model's accuracy in these responses is indicative of its effectiveness in broader scientific enquiries.
2. Span a spectrum of complexity, from straightforward information extraction with minimal linguistic intricacy to advanced linguistic reasoning for identifying pertinent information.

In consultation with domain experts, we identified five core protein properties at distinct difficulty levels that form the basis for our scientific benchmarking questions: *molecular weight* (mw), *co-factor binding*, *sub-cellular localization*, *protein domains*, and *enzymatic reaction*.

Alongside these benchmarking questions, we also require a robust strategy to assess the correctness of open-ended answers to them. Four possible assessment strategies could be imagined: (1) Exact matching of statements, (2) Keyword comparison, (3) Comparison using stronger LLMs (xLLMs), and (4) Human evaluation. While exact statement matching is useful in some research domains, it lacks the rigour required for assessing scientific accuracy. On the other hand, human evaluation for specialized domains is extremely costly at large scales. However, we can use human supervision to ensure the quality of comparisons conducted by xLLMs. As a result we selected *keyword comparison* and *use of xLLMs* guided by human experts for Pika framework.

3.2.2 Two-Tiered Benchmarking System

The use of xLLMs for response evaluation is more accurate but computationally prohibitive as compared to keyword comparison, which pragmatically balances scientific accuracy with computational efficiency. This balance of performance vs efficiency motivated us to design a two-tiered benchmarking system comprised of a light-weight benchmark employing keyword comparison for efficiency, and a rigorous xLLM-based benchmark for scientific fidelity.

Metric ID	Question	Example Label	Example Response
mw MALE ↓	What is the molecular weight of this protein?	55808	The molecular weight of this protein is 54988 KDa.
exact cofactor ↑	What is a cofactor of this protein?	Zn(2+)	The cofactor of this protein is Zn(2+).
is_enzyme F1 ↑	Can this sequence be considered an enzyme?	True	Yes
location F1 ↑	Where is this protein located?	Membrane	The sub-cellular location of this protein is the cell membrane.
binary Q mean F1 ↑	Is this a membrane protein? Is this a nuclear protein? Is this a mitochondrial protein?	True False False	Yes, it is a single-pass membrane protein. No, it is localized to the cytoplasm. No

Table 1: Questions and example responses for Biochem-Lite (See A.6 for our motivation for each metric). Underlined indicates extracted entities for score calculation. ↑: Higher values better. ↓: Lower values better.

Metric ID	Question	Ground Truth	Example Response
Reaction	What chemical reaction is catalyzed by this protein?	EC = 5.1.1.1, L-alanine → D-alanine	This protein catalyzes the conversion of L-alanine to D-alanine.
Domains	What are the functional domains of this protein?	PLP-binding barrel, Alanine racemase, Alanine racemase C-terminal domain-like, pyridoxal phosphate binding	The functional domains of this protein include the alanine racemase domain, Alanine racemase C-terminal domain-like, Alanine racemase, and PLP-binding barrel.
Cofactors	What are the cofactors of this protein?	pyridoxal 5'-phosphate	The cofactors of this protein are pyridoxal 5'-phosphate and magnesium ions.

Table 2: Questions and example responses for Biochem-ReAct. Presented in order of difficulty, with *Reaction* being the most difficult question for a multimodal LLM to answer based on a protein sequence.

1. **Biochem-Lite:** Pika's light-weight benchmarking involves a set of pre-defined, scientifically relevant questions with simple answers extracted for each protein using GPT3.5 during Pika-DS creation (Table 1). These questions are designed to cover a range of biochemical and functional properties (e.g. binding to cofactros or cellular localization) while evaluating multimodal PQA performance at various levels (i.e., information extraction from protein embeddings, cross-modal information processing by the LLM, or the LLM's ability to generate relevant responses). Although the ground truth answers to these questions are single-worded, for evaluation, the questions are presented in free-form, and the model provides open-ended responses, which is processed via rule-based entity extraction and is scored using adequate metrics as detailed in Section A.7.

2. **Biochem-ReAct:** While light-weight benchmarking questions offer a general overview of model's performance in various scientific aspects, they do not represent real-world scenarios where the scientific fidelity of open-ended responses is vital. Therefore, we also selected three biochemical questions at three distinct difficulty levels surrounding *Reaction & Activity* of the proteins that can be subjects of scientific enquiry into novel protein sequences (Table 2). The complexity of the responses to these questions necessitates assessment using advanced LLMs such as GPT4. To ensure the high reliability of these assessments we performed an iterative prompt optimization with human feedback following the approach described in Section A.4.

Overall, our two-tiered benchmarking system establishes a balanced and robust framework for evaluating multimodal PQA models, ensuring both scientific rigor and computational feasibility.

3.3 Pika Baselines

As discussed in Section 2, previous PQA studies have either been limited to knowledge retrieval or have ignored the evolutionary context of protein sequences. Combined with the use of unsuitable metrics such as BLEU, these studies do not provide relevant benchmarking opportunities for scientific PQA. Furthermore, unlike natural multimodal LLMs (e.g., Vision Question Answering) where human-baselines are possible, in PQA it is not possible to determine the upper-bound of performance by comparison to expert evaluations, because humans are unable to extract any information from protein sequences. Therefore, in the absence of relevant baselines or human evaluation for PQA, we introduce two multimodal PQA-LLMs and define various lower- and upper-bound baselines.

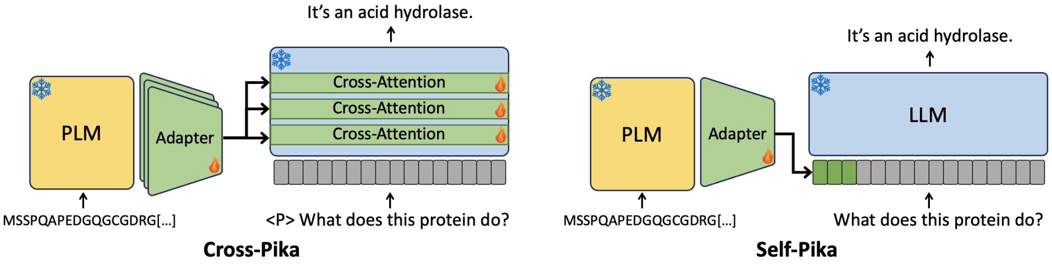

Figure 3: Schematic representation of Cross- and Self-Pika architectures for the scientific PQA task. PLM = Protein Language Model (protein sequence encoder), LLM = Large Language Model. Only the adapter and cross-attention modules (both in green) are trained.

3.3.1 Pika Models

We propose two robust multimodal architectures for PQA tasks, drawing inspiration from successful Vision Language Model (VLM) strategies, replacing the vision-encoder with a protein language model (PLM) (Fig. 3).

1. **Cross-Pika** is inspired from *Flamingo* (Tsimpoukelli et al., 2021) and *Prismer* (Liu et al., 2023). It uses multiple independent learnable adapters, each creating distinct protein latent embeddings for each transformer layer of the LLM. These embeddings are injected into the LLM using a gated cross-attention mechanism before each native self-attention layer. At the cost of increased complexity, this design allows for a nuanced and layer-specific modulation of the LLM, potentially enabling it to process complex protein-related information more effectively, which could aid the model in answering biochemically intricate PQA queries.
2. **Self-Pika** is based on the architecture proposed in *Frozen* (Tsimpoukelli et al., 2021) and uses a single learnable adapter to transform protein sequence embeddings into latent embeddings compatible with language token embeddings. The transformed embeddings, concatenated at the beginning of the LLM's initial token embeddings, allow the protein latent embeddings to influence the LLM's response through its internal self-attention mechanisms. This approach effectively creates soft-prompts conditioned on the input protein sequence. This architecture simplifies the integration process, reducing computational overhead while allowing for the incorporation of essential protein characteristics into the LLM's response.

Considering the diversity in protein lengths, we opted for the Perceiver architecture (Jaegle et al., 2021) as the learnable adapter for both models. This choice standardizes the transformation of protein embeddings into a consistent number of latent embeddings for seamless cross-modality information transfer. In both architectures we keep the pre-trained LLM and PLM frozen, only incorporating trainable "adapter" modules that extract relevant latent representations from PLM's output (protein sequence embeddings) and integrate them into the LLM's transformer architecture, facilitating seamless transfer of information across modalities. In this work we use ESM2 (Lin et al., 2022) as the pre-trained PLM and Phi-2 (Microsoft, 2023) as the pre-trained LLM (In some experiments GPT2 (Radford et al., 2019) was used, where indicated).

3.3.2 Lower-bound Baselines

- **Random Baseline:** Selects a random answer for each question from the pool of relevant answers.
- **LLM Only:** Pre-trained Phi-2 model to determine LLM's base response to Biochem-Lite questions.
- **Pika w/o PLM:** Mirrors Self-Pika's architecture and training, substituting the PLM with simple token embeddings of the protein sequences. This effectively reduces the information to amino acid content only, ignoring the context and order of the sequence.

3.3.3 Upper-bound Baselines

- **MLP:** An MLP on protein sequence embeddings of ESM2 (Details in Section A.8). It defines the upper limit for the information content of ESM2 embeddings, with any performance beyond this by a model potentially indicating generalization due to the knowledge encapsulated in the LLM.

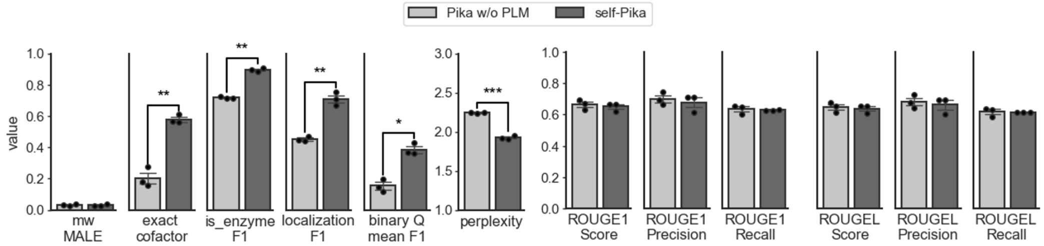

Figure 4: Evaluating the effectiveness of Biochem-Lite vs traditional linguistic metrics for scientific accuracy of PQA. Statistical significance is determined through a one-tailed paired t-test across three randomly seeded data subsets and model training (significance guide: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

- **BLAST:** Considering the evolutionary context of protein sequences, for any queried sequence, we identify the closest related protein sequence in the training data using BLAST (Camacho et al., 2009) and return the respective information as the prediction. This is a known strong upper-bound baseline, as homology is a strong predictor of protein's properties.

4 Experiments & Results

In this section we share our key experimental results. Training details can be found in Sections A.9 and A.10.

4.1 Validation of Biochem-Lite

First, we assessed the reliability of Biochem-Lite metrics in assessing PQA models performance. To We compared a simple Pika model (Self-Pika with ESM2 + GPT2) against its truncated version, Pika w/o PLM (Section 3.3.2), where ESM2 model was replaced with simple protein token embeddings. Considering that the removal of ESM2 should eliminate all functional information, leaving behind only the amino acid content, we expect that the truncated model must perform significantly worse than the original model. Therefore, metrics that fail to show a significant difference between these two would be unsuitable for assessing PQA models. Our results indicate that Biochem-Lite metrics are significantly better in Self-Pika vs Pika w/o PLM, while all ROUGE metrics fail to highlight any differences (Fig. 4). The behaviour of the *mw MALE* metric is expected as the size of a protein only depends on its amino acid content. This finding confirms the inadequacy of traditional linguistic metrics for scientific PQA task, demonstrating the utility and importance of our Biochem-Lite metrics.

4.2 Zero-shot PQA:

We define zero-shot PQA as answering questions about unseen protein sequences. Considering the debiased nature of Pika-DS, this can easily be achieved by splitting the data based on UniRef50 ID of each sequence, thereby ensuring that the validation/test sequences have no more than 50% sequence similarity with any of the training sequences. While this ensures the *unseen sequence* criteria, when considering the evolutionary context of proteins, the 50% similarity threshold is a lenient cutoff. This is because homologous proteins may exhibit as low as 15% sequence similarity (Leander et al., 2022). As a result, on UniRef50-based splits, model performance will be a combination of generalization based on the sequence and model's ability to identify distant evolutionary relations to training data. While such behaviour may be desired for some PQA applications, to extend the scope, we created a more stringent splitting strategy that focuses the evaluation of model's generalization primarily on the basis of proteins sequences without the impact of evolutionary relationships. To achieve this we minimized the the evolutionary connection of validation/test sequences to the training data. This was achieved by first grouping sequences into evolutionary groups (EvoGroups) (See Section A.11 for details) followed by splitting the data based on their EvoGrop assignments, ensuring that sequences in validation and test do not have closely related evolutionary counterparts in the training data. Naturally, we expect that the BLAST baseline should show a strong performance on UniRef50-based splits, while its performance should diminish on EvoGroup-based splits.

Baseline or Model		Biochem-Lite					Biochem-ReAct		
		mw MALE	exact cofactor	is_enzyme F1	location F1	binary mF1	Reaction	Domains	Cofactor
Lower-bound	Random	0.35 (0.01)	0.13 (0.02)	0.49 (0.00)	0.34 (0.02)	-	-	-	-
	Phi-2 only	4.53 (0.01)	0.03 (0.01)	0.34 (0.00)	0.00 (0.00)	0.31 (0.00)	-	-	-
	Pika w/o PLM	**0.02** (0.01)	0.21 (0.01)	0.7 (0.01)	0.47 (0.01)	0.21 (0.08)	0.01	0.32	0.82
Upper-bound	MLP	0.07 (0.01)	-	0.86 (0.01)	0.89 (0.03)	-	-	-	-
	BLAST	-	**0.71** (NA)	0.88 (NA)	**0.93** (NA)	-	**0.94**	**0.89**	**0.99**
Pika Models	Cross-Pika	**0.02** (0.00)	0.47 (0.15)	**0.89** (0.02)	0.77 (0.01)	0.13 (0.02)	0.36	0.60	0.94
	Self-Pika	0.04 (0.02)	0.54 (0.04)	**0.89** (0.01)	0.76 (0.02)	**0.44** (0.04)	0.58	0.79	0.97

Table 3: Performance of Pika models on Uniref50 splits. **Bold** values indicate the best score, underline indicates second best score. Values in ( ) indicate standard deviation. Biochem-Lite results on val set and from three different seeded training. Biochem-ReAct results on test set and a single seed.

4.2.1 Performance on UniRef50 Splits

Table 3 summarizes the benchmarking results of Pika models on a random UniRef50-based split in comparison with various baselines. It is evident that both Pika architectures outperform all lower-bound baselines, indicating successful cross-modality information transfer to the LLM. Notably, our top model surpasses the upper-bound MLP baseline on key Biochem-Lite benchmarking questions, underscoring the potential for generalization via the LLM's knowledge. However, as expected, the strong BLAST baseline outperforms all models in most Biochem-Lite and all Biochem-ReAct metrics. Nevertheless, these results indicate that our Pika models based on the modestly sized Phi-2 LLM, are able to comprehend distant evolutionary relations form the context of sequence embeddings.

Baseline or Model	Biochem-Lite			Biochem-ReAct		
	exact cofactor	is_enzyme F1	location F1	Reaction	Domains	Cofactor
BLAST Baseline	0.21	0.72	0.51	**0.52**	0.49	**0.94**
Self-Pika Model	**0.57**	**0.85**	**0.69**	0.09	**0.54**	0.87

Table 4: Performance of Pika model on EvoGroup splits. **Bold** values indicate the best score. Biochem-Lite results on val set and Biochem-ReAct results on test set both with a single seed.

4.2.2 Performance on EvoGroup Splits

To assess generalization by Pika models, in isolation from the effects of evolutionary relationship of proteins, we performed Self-Pika training on EvoGroup splits and compared the results with BLAST baseline on these splits (Table 4). As expected, our results show a significant reduction in performance of BLAST baseline, specially in more complex metrics such as *Domains* and *Reaction*, and to a lower extent in simpler metrics such as *Cofactor*. This is because, cofactor binding, for instance, relies on small protein motifs that are more conserved across distant evolutionary relations. Crucially, Self-Pika, except on the most complex *Reaction* metric, retained the majority of its performance on EvoGroup splits, allowing it to surpass the performance of BLAST baseline in most metrics. These observations indicate that multimodal PQA models are capable of inferring functional properties both based on the sequence of proteins as well as using distant evolutionary relations.

4.2.3 Correlation of Biochem Metrics

The Self-Pika model demonstrated strong performance across all three Biochem-ReAct questions on the UniRef split, exceeding the *without PLM baseline* by 57%, 47%, and 15% for correctly identifying *Reactions*, *Domains*, and *Cofactors*, respectively, for previously unseen protein sequences. This prompted us to study the correlation of Biochem-Lite metrics to Biochem-ReAct metrics. Comparing Biochem-Lite results against Biochem-ReAct scores for 12 high-performing checkpoints with varied configurations revealed that while traditional linguistic metrics, including perplexity, fall short in predicting the real-world efficacy of PQA models, two Biochem-Lite questions, "exact cofactor recall" and "binary Q mean F1", emerged as strong indicators of multimodal PQA models' real-world performance based on Biochem-ReAct metrics (Fig. B.3).

4.2.4 PQA Learning without Questions

To understand the significance of QA pairs in the training set, we conducted training under two stringent conditions. The first involved using only summary sentences as labels for proteins without any QAs. The second condition included summary sentences and a single Control Question, "Is this a real protein?". During training, we randomly shuffled the tokens within the sequence of half the proteins, setting the expected answer to this question as No (for summary labels all protein sequences remained unchanged). This aims to train the model to understand the task of question answering, with summaries providing the scientific context for learning. Remarkably, introducing the Control Question, even when no other questions were present during training, significantly enhanced performance (Table 5). Additionally, we observed that the performance of the model on the Control Question follows a similar pattern as other metrics (Table B.3). These observation suggest that the current bottleneck in PQA performance is likely the LLM and extending this work to larger LLMs could further improve performance and generalization capabilities.

Training Mode				Biochem-Lite			Biochem-ReAct		
Model	S	Ctrl Q	QA	mw MALE	exact cofactor	binary Q mean F1	Reaction	Domains	Cofactor
self-Pika (S)	✓	✗	✗	4.54	0.01	0.00	-	-	-
self-Pika (S+C)	✓	✓	✗	2.30	0.32	0.42	0.22	0.70	0.95
self-Pika (Q)	✗	✓	✓	**0.04**	0.54	0.44	**0.58**	**0.79**	**0.97**
self-Pika (Q+S)	✓	✓	✓	**0.04**	**0.57**	**0.57**	0.53	0.75	0.96

Table 5: Perfomance of Self-Pika in the absence of QAs during training. **Bold** values indicate best score. Standard deviation for all scores < 0.05 except for those indicated by underline. S = Summary, Ctrl Q = Control Question, QA = QA pairs.

4.2.5 PQA Learning with Novel Proteins

In training multimodal Pika models, we ensured that proteins similar to those in the training dataset were excluded from the test and validation sets, effectively creating a zero-shot scenario for our benchmarking. Given that the ESM2 model was pre-trained on a vast dataset of proteins, there was a potential concern that its embeddings might be influenced by previously encountered sequences. To mitigate this, we utilized the evaluation split reported for ESM2's pre-training, categorizing our benchmarking results into proteins seen and unseen by ESM2. The analysis revealed no significant performance difference between these conditions, affirming that Pika's zero-shot capabilities are not compromised by the prior knowledge embedded in ESM2. This finding underscores the robustness of Pika models in genuine zero-shot PQA tasks, independent of ESM2's pretraining exposure.

4.2.6 Ablation Studies

To identify key contributing factors, we compared model performance across various dimensions. Most notably, reducing the size of the LLM from Phi-2, which has 2.8 billion parameters, to GPT-2 Medium, with 355 million parameters, resulted in a 20% decrease in the key "exact cofactor" metric (Table 6). However, the size of the ESM2 model did not seem to have a significant impact, with performance remaining largely similar between ESM2-S, ESM2-M, and ESM2-L (8M, 65M and 350M parameters, respectively). This observation could be attributed to the richness of representations in the Pre-trained PLMs, potentially highlighting a bottleneck in the LLM or the quality and quantity of the data (Table 6). Concordantly, the comparison of results from the zero-shot experiment also confirms that with a lower volume of data, a tailor-made dataset could yield better results. However, as the dataset size increases—a scenario analogous to the image domain—this factor becomes less crucial. These findings suggest that while LLM capacity significantly influences PQA model performance, the sophistication of sequence embeddings provided by ESMs reaches a point of diminishing returns, underscoring the importance of focusing on LLM enhancements and data quality for future improvements.

LLM	PLM	mw	exact	is_enzyme	location	binary Q	perplexity
		MALE	cofactor	F1	F1	mean F1	
GPT2-M	ESM2-S	0.14	0.22	0.84	0.65	0.26	3.26
	ESM2-M	0.06	0.26	0.86	0.64	0.33	3.12
	ESM2-L	0.07	0.32	0.86	0.70	0.23	3.05
Phi-2	ESM2-S	**0.02**	0.51	0.86	0.71	0.33	1.99
	ESM2-M	**0.02**	**0.56**	0.88	0.75	**0.46**	**1.95**
	ESM2-L	0.04	0.54	**0.89**	**0.76**	0.44	1.98

Table 6: Effect of LLM and PLM size on Self-Pika models' performance. Standard deviation for all scores < 0.05 except for those indicated by underline.

5 Conclusion & Impact

We've established a pioneering framework for zero-shot PQA, presenting a significant step forward in the application of LLMs for scientific enquiry. The introduction of our specialized datasets and biologically relevant benchmarks underpins future explorations at the intersection of computational biology and artificial intelligence. Key insights from our zero-shot evaluations and ablation studies highlight the crucial role of LLMs in multimodal PQA performance, while underscoring the importance of considering evolutionary relation in assessing performance. Since, for Pika-DS we employed GPT3.5 to create synthetic annotations, given the limitations of LLMs in generating large-scale synthetic datasets, we endeavored to minimize the inclusion of harmful content in Pika-DS through prompt optimization and manual evaluation. Nonetheless, due to the dataset's extensive size, there's a slight possibility that unintended harmful content might still be present. Our Pika-based pre-trained models are derived from the publicly accessible and unmoderated Phi-2 LLM. Thus, all cautions, restrictions, and notes associated with Phi-2 (Microsoft, 2023) are applicable to our models. Looking ahead, leveraging larger, more diverse LLMs may offer substantial gains in model generalization, driving us towards the goal of automating accurate scientific enquiry into proteins.

6 Reproducibility & Accessibility

All code and data used for creation of data, model training and baselines are publicly accessible on `www.github.com/EMCarrami/Pika`.

The Pika framework is specifically designed for question answering related to protein sequences. With scientists having identified nearly 0.25 billion protein sequences, and functional annotations available for fewer than a million, our framework offers significant potential for research into these largely unexplored proteins. While our efforts are directed towards scientific research, we recognize the potential risk of misuse by individuals aiming to create or identify harmful substances. We strictly prohibit using our dataset and framework for any illegal activities or actions that could harm individuals or society.

Acknowledgments and Disclosure of Funding

We would like to thank Dr Aida Nematzadeh for critical advice and Maryam Karaminejad Ranjbar for graphic design. We would also like to sincerely thank Relation Therapeutics for providing us with the necessary compute for training the final version of Pika. This work was primarily supported by personal funding and the authors declare no conflict of interest.

References

Blaabjerg, L. M., Kassem, M. M., Good, L. L., Jonsson, N., Cagiada, M., Johansson, K. E., Boomsma, W., Stein, A., and Lindorff-Larsen, K. Rapid protein stability prediction using deep learning representations. *eLife*, 12, 2023. doi: 10.7554/eLife.82593. URL `https://elifesciences.org/articles/82593`.

Camacho, C., Coulouris, G., Avagyan, V., et al. BLAST+: architecture and applications. *BMC Bioinformatics*, 10:421, 2009. doi: 10.1186/1471-2105-10-421. Published 2009 Dec 15.

Consortium, T. U. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 11 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac1052. URL `https://doi.org/10.1093/nar/gkac1052`.

Fang, Y., Liang, X., Zhang, N., Liu, K., Huang, R., Chen, Z., Fan, X., and Chen, H. Mol-instructions: A large-scale biomolecular instruction dataset for large language models, 2024.

Function-SIG. The critical assessment of protein function annotation algorithms (cafa). `https://biofunctionprediction.org/cafa/`, 2024.

Guo, H., Huo, M., Zhang, R., and Xie, P. Proteinchat: Towards achieving chatgpt-like functionalities on protein 3d structures. *TechRxiv:23120606.v1*, 2023.

Jaegle, A., Borgeaud, S., Alayrac, J.-B., Recasens, A., Goh, G., Terzić, K., Dehghani, M., Metzler, D., Kumar, K., Sifre, L., et al. Perceiver: General perception with iterative attention. *arXiv preprint arXiv:2103.03206*, 2021.

Jiang, Z., Yang, M. Y. R., Tsirlin, M., Tang, R., and Lin, J. Less is more: Parameter-free text classification with gzip, 2022.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W., and Lu, X. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

Leander, M., Liu, Z., Cui, Q., and Raman, S. Deep mutational scanning and machine learning reveal structural and molecular rules governing allosteric hotspots in homologous proteins. *eLife*, 11: e79932, oct 2022. ISSN 2050-084X. doi: 10.7554/eLife.79932. URL `https://doi.org/10.7554/eLife.79932`.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL Workshop: Text Summarization Braches Out 2004*, pp. 10, 01 2004.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

Liu, S., Fan, L., Johns, E., Yu, Z., Xiao, C., and Anandkumar, A. Prismer: A vision-language model with an ensemble of experts. *arXiv preprint arXiv:2303.02506*, 2023.

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Mathur, N. et al. Re-evaluating automatic summarization with bleu and rouge. *Conference Paper*, 2020.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Microsoft. Microsoft/phi-2 model card. `https://huggingface.co/microsoft/phi-2`, 2023. Version 7e10f3ea09c0ebd373aebc73bc6e6ca58204628d.

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

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., and Finn, R. D. HMMER web server: 2018 update. *Nucleic Acids Research*, 46(W1):W200–W204, 06 2018. ISSN 0305-1048. doi: 10.1093/nar/gky448. URL `https://doi.org/10.1093/nar/gky448`.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019. URL https://openai.com/blog/better-language-models/.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text, 2016.

Rajpurkar, P., Jia, R., and Liang, P. Know what you don't know: Unanswerable questions for squad, 2018.

Sharifzadeh, S., Kaplanis, C., Pathak, S., Kumaran, D., Ilic, A., Mitrovic, J., Blundell, C., and Banino, A. Synth^2: Boosting visual-language models with synthetic captions and image embeddings, 2024.

Tsimpoukelli, M., Menick, J. L., Cabi, S., Eslami, S., Vinyals, O., and Hill, F. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34: 200–212, 2021.

Wei, B., Zhang, Y., and Gong, X. Deeplpi: a novel deep learning-based model for protein–ligand interaction prediction for drug repurposing. *Scientific Reports*, 12:18200, 2022. doi: 10.1038/s41598-022-23014-1. URL https://www.nature.com/articles/s41598-022-23014-1.

Xu, M., Yuan, X., Miret, S., and Tang, J. Protst: Multi-modality learning of protein sequences and biomedical texts. *arXiv preprint arXiv:2301.12040*, 2023.

Zhang, N., Bi, Z., Liang, X., Cheng, S., Hong, H., Deng, S., Lian, J., Zhang, Q., and Chen, H. Ontoprotein: Protein pretraining with gene ontology embedding. *arXiv preprint arXiv:2201.11147*, 2022.

A Supplementary Methods

A.1 Creation of Pika-DS

Pika-DS was prepared in three phases:

1. We utilized the SwissProt database as a foundational resource. SwissProt, the reviewed section of UniProt, features approximately 570,000 detailed annotations of protein sequences spanning all domains of life and viruses. For each entry, we extracted an expert-curated list of scientific information covering a wide range of subjects including evolutionary, biochemical and functional properties (Section A.2). Removing proteins shorter than 30 amino acids, we obtained 3.7 million information fields for over 548,000 protein sequences.
2. UniProt also offers similarity-based clustering of its entries at various sequence similarity thresholds of 100%, 90%, and 50%, known as UniRef clusters. For instance, protein sequences that share a minimum of 50% pairwise sequence similarity will be assigned the same UniRef50 cluster. Recognizing a strong bias towards more commonly studied protein families in SwissProt (Figure B.1), we limited Pika-DS to a maximum of two most informative sequences per UniRef50 cluster using a custom algorithm (Section A.3). This resulted in a debiased set of over 257,000 protein sequences and 1.17 million information fields.
3. Lastly, we used GPT3.5 API to process the information fields for each protein entry using systematically optimized prompts (Sections A.4 and A.5) to create the Pika-DS's three main components (example in Table B.2):
   - *Summary:* A summary of each protein's functional and biochemical properties, based solely on the provided information excluding the protein's name.
   - *QAs:* Several diverse QA pairs for each information field, formatted for LLM training.
   - *Metrics:* Single-word answers to a set of predefined scientific questions serving as the ground-truth for our Biochem-Lite benchmarks (Section 3.2.2).

A.2 Data collection from UniProt

We retrieved XML files of all SwissProt entries longer than 30 amino acids, with the cut-off date of 14-08-2023, using UniProt's API. Since these entries contained extensive extraneous information, like author names, submission dates, etc., we employed a rule-based pre-processing approach to extract fields relevant to each protein's functional and biochemical characteristics. This selection, guided by an expert biochemist, included *sequence* (molecular mass and length), *organism* (top three taxonomic levels), *catalytic activity* (including EC number), *biophysicochemical properties* (pH and temperature dependence), *cofactor*, *subunit* (excluding fields containing "interact", "associate", or "complex"), *subcellular location* (excluding isoforms), and functional domains: *GO* (only molecular function, omitting biological process and cellular component), *Gene3D*, and *SUPFAM*.

A.3 Debiasing SwissProt Entries Using UniRef50 Clusters

To mitigate bias in SwissProt entries, we employed UniRef50 clusters, sourced from the UniProt FTP's idmapping file. First, we merged clusters representing isoforms with their corresponding main protein clusters to consolidate isoform-driven redundancy. Specifically, if the isoform of a protein was a member of another Uniref50 cluster, we merged the isoform cluster into the main protein's cluster. This step aimed to tighten clustering criteria, avoiding oversampling due to isoforms. Next, within each merged cluster, sequences shorter than 25% of their cluster's median length were excluded to ensure a focus on sufficiently representative sequences. This filtering is due the fact that some Uniprot sequences have incomplete sequences. Finally, following the methodology outlined by Jiang et al. (Jiang et al., 2022), we calculated gzip information for each entry. Initially, the entry with the highest gzip information was identified. Subsequently, we assessed the additional gzip information provided by each remaining entry relative to this top candidate. The entry offering the highest additional gzip information was selected for inclusion. Throughout this selection process, cluster representatives were given priority in the event of a tie, ensuring the most informative representatives were chosen. This debiasing process was designed to refine the Pika-DS by leveraging UniRef50 clusters, enhancing the dataset's quality and representativeness for downstream applications.

A.4 GPT3.5 Prompt Optimization Strategy

To ensure the high quality of GPT3.5 generated information and QA pairs, we performed a systematic prompt optimization focusing on GPT3.5's adherence to using only the given data, summarizing complex biological reactions accurately, and avoiding speculative or ambiguous language. This was performed in an iterative process where at each iteration, GPT3.5 was provided with the prompt and all the extracted information for 50 randomly selected protein sequences. Next, 100 outputs from each of the summary statements, QA pairs and evaluation metrics were evaluated against the input information of the respective proteins by an expert biochemist and the number of incorrect, poor quality and irrelevant (correct but unrelated to the protein's function) statements or QA pairs were noted. At each step, the prompt was modified to address the most problematic issues, with a target of no more than one incorrect, two poor quality and two irrelevant values in each category (optimised prompt in Section A.5).

A.5 GPT3.5 Prompts Used for Creation of Pika-DS

Summarising and QAs: The following instructions were used to create summary statements and QA annotations based on information fields collected from SwissProt:

You will receive details about a specific protein. Perform the following tasks and print each result in a new line:

1) Provide a factual summary, without using the protein's name with a maximum of 500 words. Your summary must accurately and scientifically describe the functional, biochemical and structural properties of this protein based only on the provided information. Ensure that the summary follows a natural and scientific flow, starting with general information such as structure, localization and taxonomy before detailing functional and biochemical properties. Ensure that all key points are covered and DON'T provide any extra information than what is stated in the input.

2) For each type of information provided, create a question-and-answer pair to elucidate an aspect of the protein's functionality or biochemical properties without using the protein's name. Phrase your questions and answers such that they will be suitable for training a language model. DON'T enumerate or label the questions and print each question and its answer pair in the same line.

- For all tasks if the input contains large group of properties only provide the canonical and crucial information rather than enumerating every single entry. - Where applicable, summarise enzymatic reactions into one or two of the generic classes of the activities. - DON'T use any of your knowledge to add additional context or information. DON'T add any speculative or unwarranted information that is not specifically provided. - AVOID using generic phrases or embellished terms such as 'highly', 'several', 'diverse' and 'various'. - Exactly follow the output format provided below to ensure consistency. Final output format: summary: [YOUR SUMMARY] QA pairs: 1) [YOUR QUESTION] [YOUR ANSWER] 2) [YOUR QUESTION] [YOUR ANSWER] [...] n) [YOUR QUESTION] [YOUR ANSWER]

Biochem-Lite Ground truth: The following instructions were used to collect ground-truth values for Biochem-Lite questions based on summary statements and information fields:

You will receive details about a specific protein. Provide a single word answer to the following questions. Print each question and your answer in the same new line. If the question does not apply to the protein, ignore the question. 1) Is this protein localized to the cell membrane? 2) Is this a membrane protein? 3) Is this protein localized to nucleus? 4) Is this protein localized to mitochondria? 5) Does this protein bind to DNA? 6) Does this protein bind to RNA? 7) Is this protein an enzyme? 8) What are co-factors of this protein as a comma separated list?

- Exactly follow the output format provided below to ensure consistency. Final output format: 1) [MY QUESTION] [SINGLE WORD ANSWER e.g. YES/NO/UNKNOWN] 2) [MY QUESTION] [SINGLE WORD ANSWER e.g. YES/NO/UNKNOWN] [...] 7) [MY QUESTION] [SINGLE WORD ANSWER e.g. YES/NO/UNKNOWN] 8) [MY QUESTION] [Comma separated list of co-factors e.g. FAD,FMN/UNKNOWN]

A.6 Biochem-Lite Metric Questions and Motivation

mw MALE

- **Question**: What is the molecular weight of this protein?
- **Score:** Mean Absolute Error of log values
- **Presence in training:** in exact form
- **Motivation:** Proteins are large polymers of amino acids. The molecular weight of a protein depends not only on the number of amino acids but also on the types of amino acids in the sequence.

Cofactor recall

- **Question**: What is a cofactor of this protein?
- **Score**: If exact match, score=1; else 0 (average across examples)
- **Presence in training:** in exact form
- **Motivation**: Many proteins require other small molecular entities to function. These could be ions, or small molecules, or a combination of these, such as a Heme-Fe. Proteins often bind to their cofactors via motifs, which are short 3D structures formed due to the presence of specific sequences. Identifying the cofactor of a protein requires the extraction and analysis of these sequence motifs from the sequence embeddings.

is_enzyme F1

- **Question**: Can this sequence be considered an enzyme?
- **Score**: F1 score
- **Presence in training:** in similar form
- **Motivation**: The motivation for this question is similar to that of cofactor recall, focusing on the functional characterization of proteins. Enzymatic activity is a crucial aspect of protein function, requiring specific structural or sequence features for identification.

Location F1

- **Question**: Where is this protein located?
- **Score**: F1 score
- **Presence in training**: in exact form
- **Motivation**: Proteins must be localized to specific compartments within cells to perform their functions correctly. This localization is directed by specific targeting sequences or 3D structures. Extracting this information can be challenging, and the identification of the destination for a protein can vary depending on the organism and context.

Binary localization average F1

- **Question**: Is this a membrane protein? Is this a nuclear protein? Is this a mitochondrial protein?
- **Score**: Mean of F1 scores of all 3 questions
- **Presence in training:** never in this form
- **Motivation**: This question is related to the previous one but focuses on specific locations in a binary format. While the information extraction process is similar, the LLM needs to perform reasoning to find the answer.

A.7 Metric equations used in Biochem-Lite

- *mw MALE*: The mean-absolute log error (MALE) of the predicted molecular weight (MW) is:

$$MALE = \frac{1}{N}\sum_{i=1}^{N} |\log_{10}(\hat{MW}_i/MW_i)| \quad (1)$$

where $\hat{MW}_i$ is the predicted MW, MW_i is the ground truth MW, and N is number of examples.

- *exact cofactor*: The score is computed as:

$$\text{Score}_{\text{exact}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\exists w \in R_i : w \in GT_i) \tag{2}$$

  where R_i is the set of words in response i, GT_i is the set of ground truth cofactor words for protein i, and $\mathbb{1}$ is the indicator function that is 1 for exactly one match and 0 otherwise.

- *location F1*: The F1 score is the harmonic mean of Precision and Recall. The predicted class assignments for the calculation of Precision and Recall are based on the exclusive presence of correct labels (*membrane*, *nucleus*, or *mitochondrion*) in the generated response. Responses lacking or containing multiple labels receive the *none* class label.

- *is_enzyme F1*: The F1 score is computed similar to *location F1*. Responses are classified as True or False based on the exclusive presence of *yes* or *no*, respectively. Any deviation is labeled as *none*.

- *binary Q mean F1*: We have:

$$F1_{\text{binary Q mean}} = \frac{1}{3}\sum_{q=1}^{3} F1_q \tag{3}$$

  where $F1_q$ is the F1 score for each binary question (class assignments similar to *is_enzyme F1*).

A.8 MLP Baseline

We use a 3-layer MLP with GELU activation on protein sequence embeddings from the Pre-trained ESM2. Where possible, we convert each light-weight benchmarking question into a classification or value prediction task. More specifically we performed regression for log values of mw for mw MALE and performed classification for *is_enzyme* and *location* questions. The MLP is then trained to predict the value or correct class labels using mean squared error or cross entropy loss, respectively. This baseline sets an upper limit for the information content inherent in ESM2 embeddings.

A.9 Training

We devised a simple training strategy for PQA models, keeping both the PLM and LLM frozen and optimizing for the causal language model loss with AdamW. Hyperparameters were optimised following a greedy search as detailed in section A.10. All training were performed on a single A100-80GB or H100-80GB Nvidia GPU. Unless otherwise stated, both the QA and summary statement section of the Pika-DS was used for the training of all Pika models. Examples were split to train, validation and test set based on the UniRef50 cluster or EvoGroup of their respective protein sequences in 94.5%, 0.05% and 5% ratios, respectively.

A.10 Hyperparameter Optimization

Greedy hyper parameter search was performed for both architectures, monitoring the is_enzyme metric. For all experiments, training was performed on 25000 protein sequences and metrics were computed on 250 unseen proteins with gpt2-medium model as the LLM and esm2_t12_35M model as the protein sequence encoder. The sweep was performed, in order, on optimizer's weight decay [0, 1e-4, 1e-2] and learning rate [1e-5, 1e-4, 1e-3], batch size [2, 4, 8], Perceiver latent size [32, 64, 100] and the number of Perceiver layers [1, 2, 4]. The final set of hyper parameters were as follows: learning rate: 1e-4, weight decay: 1e-4, batch size: 8, Perceiver latent size: 100, and number of Perceiver layers: 1 (Cross-Pika architecture), 4 (Self-Pika architecture).

A.11 Creation of EvoGroups

EvoGroups were decided by starting from a random sequence in the data, identifying all its related sequences using JackHammer (Potter et al., 2018) with a very lenient cut-off threshold for e-E-score of 1.0 to ensure a broad grouping of related proteins. For each randmoly selected sequence, all related sequences were marked as belonging to the same EvoGroup and were removed from the remainder of the dataset for further iterations with newly selected sequences. The process was repeated until all sequences were assigned an EvoGroup.

B Supplementary Figures & Tables

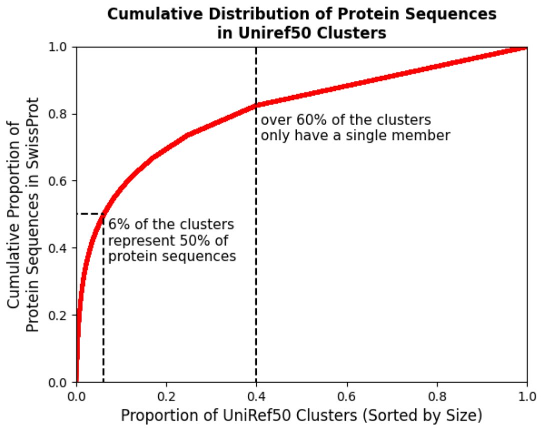

Figure B.1: Over-representation bias in SwissProt database.

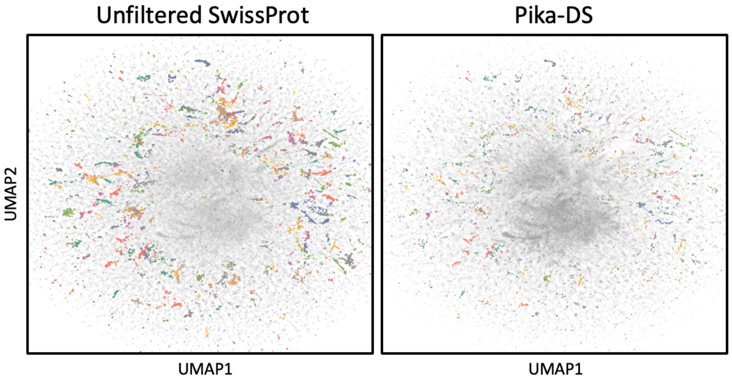

Figure B.2: Comparison of sequence bias in SwissProt database, before and after filtering. Members of the top100 largest UniRef50 clusters are colored. The strong overrepresentation of highly studied protein groups is apparent in the unfiltered plot, while the Pika-DS shows a significant reduction in this bias

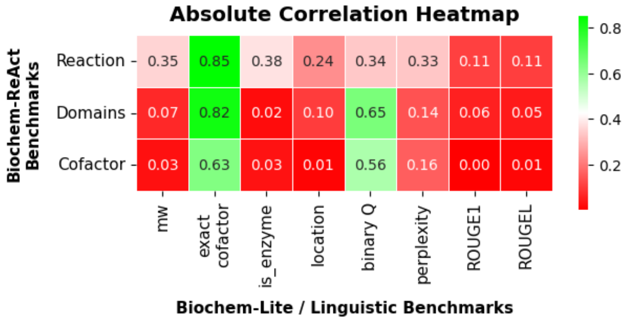

Figure B.3: Cross-correlation of Biochem-ReAct Benchmarks with Biochem-Lite/Linguistic Benchmarks. Absolute values.

	# total	# unique	# words	# tokens
Proteins	257 K	252 K	-	105.3 M
Info fields	1.17 M	319 K	10.8 M	29.5 M
Summaries	1.33 M	861 K	22.5 M	36.4 M
Questions	1.71 M	98 K	15.7 M	14.9 M
Answers	1.71 M	783 K	18.4 M	31.8 M

Table B.1: Pika-DS size and token counts. Values for Summaries represent the number of sentences.

uniref	UniRef50_Q9HUN4	
sequence	MRPLVATVDLTALRHNYLLAKQCAPQRKAFAVVKANAYGHGAPEAVTALREIADGFAVACLEEAE[...]	
fields	catalytic activity: EC = 5.1.1.1, L-alanine = D-alanine	
	cofactor: pyridoxal 5’-phosphate	
	functional domains: alanine racemase activity, Alanine racemase, Alanine racemase C-terminal domain-like, PLP-binding barrel, pyridoxal phosphate binding	
	taxonomy: Bacteria, Pseudomonadota, Gammaproteobacteria	
	protein size: 358 aa, 38485 KDa	
summary	This protein is an alanine racemase enzyme found in various bacteria, specifically in the Pseudomonadota phylum of the Gammaproteobacteria class	
	It is composed of 358 amino acids with a molecular weight of 38,485 KDa	
	The protein exhibits a catalytic activity with an Enzyme Commission (EC) number of 5.1.1.1, where it converts L-alanine to D-alanine	
	Its enzymatic function is facilitated by a cofactor known as pyridoxal 5’-phosphate	
	The protein contains several functional domains, including the Alanine racemase activity domain, the Alanine racemase C-terminal domain-like, the PLP-binding barrel domain, and the pyridoxal phosphate binding domain.	
qa	What is the catalytic activity of this protein? The catalytic activity of this protein is the conversion of L-alanine to D-alanine.	
	What is the cofactor associated with this protein? The cofactor associated with this protein is pyridoxal 5’-phosphate.	
	Which phylum does this protein belong to? This protein belongs to the Pseudomonadota phylum.	
	What is the molecular weight of this protein? The molecular weight of this protein is 38,485 KDa.	
	How many amino acids are present in this protein? This protein is composed of 358 amino acids.	
	What are the functional domains of this protein? The functional domains of this protein include the Alanine racemase activity domain, the Alanine racemase C-terminal domain-like, the PLP-binding barrel domain, and the pyridoxal phosphate binding domain.’,	
	What class of bacteria is this protein found in? This protein is found in the Gammaproteobacteria class.	
	Can this protein act as an enzyme? Yes	
metrics	in_membrane	False
	in_nucleus	False
	in_mitochondria	False
	is_enzyme	True
	cofactor	pyridoxal 5’-phosphate

Table B.2: An example entry in Pika-DS representing Uniprto ID A4VQM5

	Random	LLM only	self-Pika w/o PLM	self-Pika
F1 Score	0.49	0.00	0.43	0.99
Accuracy	0.50	0.00	0.51	0.99

Table B.3: Performance on Control Question.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification:

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For larger models, due to excessive computational expenses, experiments are performed once.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.