PhageBench: Can LLMs Understand Raw Bacteriophage Genomes?

Anonymous ACL submission

Abstract

Bacteriophages, often referred to as the dark matter of the biosphere, play a critical role in regulating microbial ecosystems and in antibiotic alternatives. Thus, accurate interpretation of their genomes holds significant scientific and practical value. While general-purpose Large Language Models (LLMs) excel at understanding biological texts, their ability to directly interpret raw nucleotide sequences and perform biological reasoning remains underexplored. To address this, we introduce PhageBench, the first benchmark designed to evaluate phage genome understanding by mirroring the workflow of bioinformatics experts. The dataset contains 5,600 high-quality samples covering five core tasks across three stages: Screening, Quality Control, and Phenotype Annotation. Our evaluation of eight LLMs reveals that general-purpose reasoning models significantly outperform random baselines in phage contig identification and host prediction, demonstrating promising potential for genomic understanding. However, they exhibit significant limitations in complex reasoning tasks involving long-range dependencies and fine-grained functional localization. These findings highlight the necessity of developing next-generation models with enhanced reasoning capabilities for biological sequences[1].

1 Introduction

The intersection of artificial intelligence and biology has shifted from analyzing biological texts to decoding, the language of life, DNA sequences. Recent Genomic Foundation Models (GFMs) such as, Evo (Nguyen et al., 2024; Brixi et al., 2025) and Caduceus(Schiff et al., 2024), have demonstrated that deep learning architectures can capture the complex syntax and semantics of DNA. These models can predict gene expression (Hou et al., 2024; Avsec et al., 2025) and even design whole genome of functional bacteriophages (phages) (King et al., 2025). However, these specialized models require extensive pre-training on DNA sequence corpora, similar to how general LLMs are trained on large text corpora. In parallel, general-purpose Large Language Models (LLMs) have exhibited strong reasoning abilities and performance on various biology-related tasks (Qu et al., 2025; Swanson et al., 2025; Wang et al., 2025). This raises a critical question: ***Can general-purpose LLMs understand raw DNA sequences without specific pre-training?***

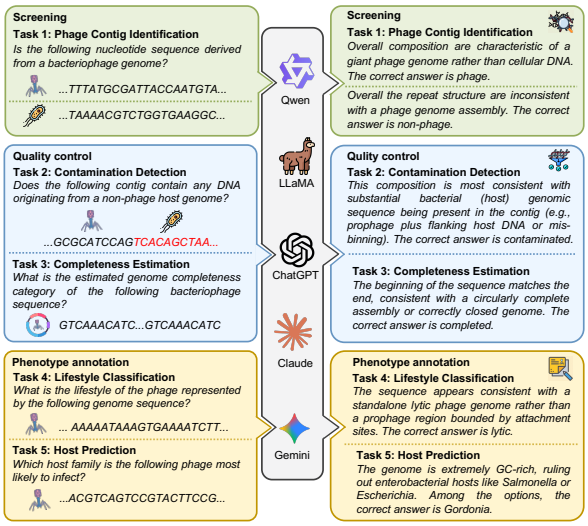


Figure 1: PhageBench challenges general-purpose LLMs to perform five phage analysis tasks, processing raw nucleotide sequence inputs to generate reasoning-based phenotypic predictions.

To address this question, we propose a new challenge for the natural language processing (NLP) community: phage genome annotation. We select phages, viruses that infect bacteria, as the testbed for three reasons. First, the length of a typical phage genome ranges from approximately 3 to 150 kilobase pairs (kb) (Dion et al., 2020), which

[1] Our PhageBench dataset and evaluation code are available at https://anonymous.4open.science/r/PhageBench-D428

aligns well with the context windows of modern LLMs. This makes phages an ideal subject for evaluating non-natural langauge long-context reasoning. Second, unlike eukaryotic genomes which contain long non-coding introns, phage genomes exhibit a dense, linear organization where genes are arranged in a structured syntax similar to natural language (Mavrich and Hatfull, 2017). Previous studies indicate that this structural conservation allows for functional prediction based on genomic position even when sequence homology is low (Grigson et al., 2025). This allows us to test whether LLMs can infer function in the absence of explicit sequence matches. Third, phages are the most abundant biological entities in the biosphere and play critical roles in regulating microbial ecosystems (Rohwer, 2003), food engineering (Prasad et al., 2025), and phage therapy as alternatives to antibiotics (Strathdee et al., 2023; Skurnik et al., 2025). However, the vast majority of phage sequences identified in metagenomic data remain unannotated dark matter lacking basic taxonomic and phenotypic information. This annotation bottleneck limits our ability to harness phages for therapeutic and biotechnological applications.

Therefore, achieving efficient and accurate phage genome annotation is of significant scientific and practical importance. To systematically investigate the ability of LLMs to understand phage genomes, we introduce PhageBench, a multi-task benchmark designed to mirror the complete workflow of expert phage analysis. PhageBench follows a standard structure across three stages: Screening (phage contigs identification from environmental noise), Quality Control (contamination detection and completeness estimation), and Phenotype Annotation (lifestyle prediction and host prediction) (Fig. 1). This design reflects how biologists progressively filter and characterize sequences, moving from simple identification to complex biological inference. The benchmark contains over 5,600 samples across five tasks, with careful controls for sequence length distribution and class balance to ensure models cannot rely on statistical biases.

We evaluate eight advanced LLMs on PhageBench, including reasoning models. Our analysis reveals that general-purpose reasoning LLMs demonstrate promising potential in phage identification and host prediction tasks without domain-specific pre-training. However, they exhibit limitations in modeling long-range dependencies, leading to declines in performance in completeness estimation tasks involving global structures as sequence length increases. Furthermore, although these models are capable of applying accurate biological logic during reasoning, they struggle to precisely identify functional gene fragments within raw nucleotide sequences spanning tens of thousands of base pairs. To summarize, our contributions are:

- We first introduce raw phage genome annotation as a challenging testbed for evaluating the reasoning capabilities of general-purpose LLMs. This task requires models to process long-context sequences and perform zero-shot logical reasoning, providing a proxy for measuring LLM capabilities in long-range and complex syntactic analysis in non-natural language domains.
- We construct and release PhageBench, a comprehensive dataset containing 5,600 high-quality samples. PhageBench mirrors the real workflow of bioinformatics experts, covering three critical stages in phage genome annotation.
- By applying general-purpose LLMs to raw phage genome annotation, PhageBench provides a novel analysis pathway for genomics that does not depend on sequence homology. Our evaluation demonstrates the potential of LLMs for understanding raw genomic sequences, offering preliminary evidence for the feasibility of using general reasoning engines to accelerate the discovery, annotation, and therapeutic development of unknown phages in metagenomic data.

2 Related Work

2.1 Phage Genome Analysis

The computational analysis of phage genomes has evolved from alignment-based sequence matching to feature-learning approaches. Traditional methods, such as BLAST (Altschul et al., 1990), rely on homology search against reference databases like GenBank (Benson et al., 2018). While accurate for known sequences, these methods fail to identify sequences with low homology. To address this, learning-based approaches have been developed (Camargo et al., 2024; Peng et al., 2024). VirSorter2 (Guo et al., 2021) integrates multiple classifiers to capture diverse viral signals, while

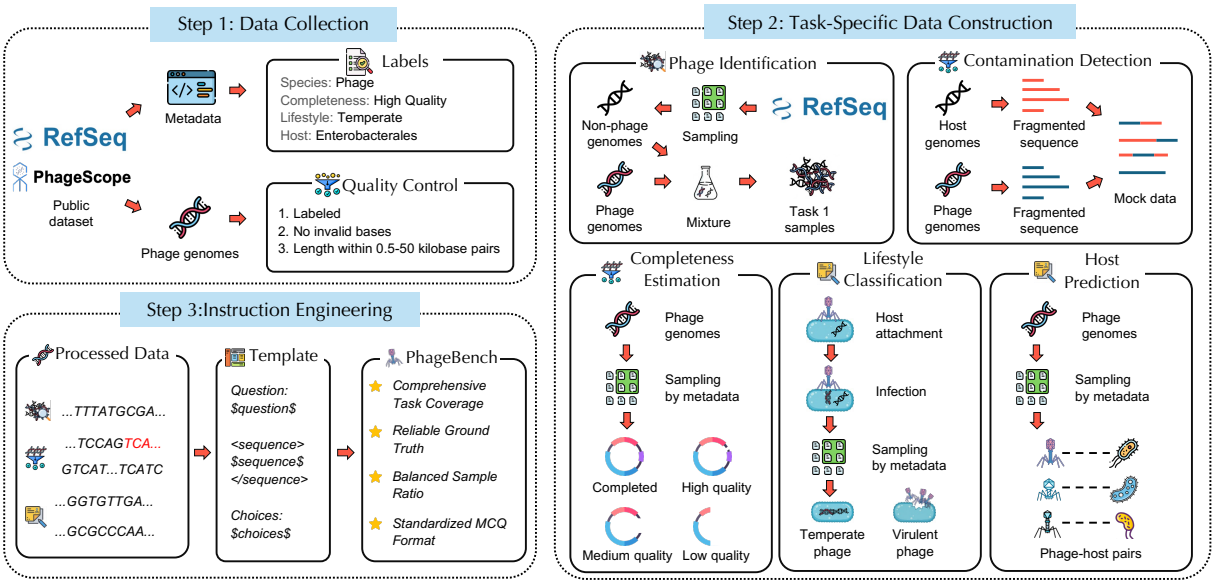


Figure 2: The overview of data construction of PhageBench. We filter raw sequences from public repositories, apply task-specific processing rules, and format the resulting samples into standardized instructions for evaluation. For detailed procedures for each task, please refer to Appendix B.1.

CheckV (Nayfach et al., 2021a) assesses genome completeness based on gene content and terminal repeats. However, these discriminative models typically output binary classifications without biological reasoning and often specialized for a single task. Our work investigates whether general-purpose LLMs can bridge this gap by leveraging their inherent reasoning capabilities to analyze raw genomic sequences directly.

2.2 Genomic Foundation Models and Benchmarks

The intersection of AI and genomics has spurred the development of specialized GFMs (Zhou et al., 2023; Nguyen et al., 2023; Dalla-Torre et al., 2025). Models such as DNABERT (Ji et al., 2021) apply Transformer architectures to DNA sequences. Recent generative models, such as Evo, have demonstrated the ability to design whole phage genomes (King et al., 2025). However, these specialized models require resource-intensive pre-training on massive genomic corpora. While some efforts, such as L2G (Cheng et al., 2024), explore repurposing general LLMs for genomics, they still rely on extensive fine-tuning.

Existing genomic benchmarks, such as GenomicBenchmarks (Grešová et al., 2023) and DNA-LongBench (Cheng et al., 2025), are primarily designed to evaluate GFMs and predominantly focus on mammalian genomic features, such as chromatin accessibility and complex regulatory mechanisms. These tasks differ significantly from the syntactic gene organization logic of prokaryotic phages. On the other hand, benchmarks designed for general LLMs, such as GeneTuring (Shang et al., 2025b), focus on knowledge retrieval rather than reasoning on raw genomic sequences. PhageBench prioritizes the understanding and reasoning of raw phage genome sequences. To the best of our knowledge, this is the first study to evaluate general-purpose LLMs on authentic whole-genome analysis tasks, thereby deepening our understanding of the capabilities of LLMs within the life sciences.

3 The PhageBench Dataset

3.1 Underlying Principles

While existing benchmarks in biology primarily focus on natural language-based tasks, such as question answering, information retrieval, or text summarization, PhageBench establishes a comprehensive evaluation dataset that integrates biological reasoning with the direct understanding and annotation of raw phage genomic sequences. Aiming to provide a rigorous standard for assessing off-the-shelf models, PhageBench examines the potential of general-purpose NLP technologies in driving scientific discovery within biology.

The construction of PhageBench is guided by four core design principles: (1) encompassing three distinct analysis stages mapped into five core tasks

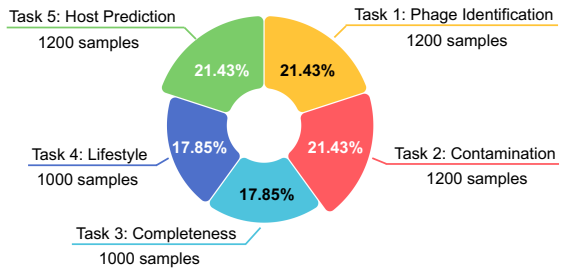


Figure 3: The compositional structure of PhageBench. The sizes of the constructed instruction datasets are labeled below the horizontal lines, and the percentages on the pie charts represent the proportion of data within each major category.

to cover the full spectrum of phage genome analysis; (2) ensuring all sequences are derived from authoritative databases to serve as high-confidence ground truth; (3) controlling the ratio of positive to negative samples; and (4) adopting a standardized multiple-choice question format.

3.2 Tasks Construction

Phage genome analysis requires identifying entities from complex background sequences, assessing their quality, and ultimately inferring their biological functions. PhageBench mirrors this path through three distinct stages: Screening, Quality Control, and Phenotype Annotation. The overall data construction pipeline is illustrated in Fig. 2. We collected raw sequences and metadata from authoritative databases, applied rigorous task-specific processing rules, and used only complete-level phage genomes for all tasks except completeness estimation task. Detailed construction procedures are provided in Appendix B.1.

Stage 1: Screening. Metagenomic sequencing has revolutionized our access to total genetic material, yet it inherently yields a complex admixture of sequences derived from bacteria, archaea, eukaryotes, and viruses. Consequently, the foundational step in phage annotation is the accurate extraction of phage signals from this multispecies background.

Task 1: Phage Contig Identification functions as a binary classification challenge designed to distinguish phage sequences from a heterogeneous array of non-phage genomic DNA. This task evaluates the capacity of a model to discriminate specific phage signatures from diverse biological confounders, including non-phage viruses, protozoa, plasmids, fungi, bacteria, and archaea.

Stage 2: Quality Control. Even within identified phage contigs, sequencing and assembly processes may yield fragmented genomes lacking essential genomic regions or chimeric sequence containing host DNA. These quality issues can mislead downstream analyses. Therefore, this stage evaluates the capacity of LLMs to assess both sequence purity and structural integrity.

Task 2: Contamination Detection focuses on the identification of host genomic fragments that have been erroneously assembled with phage sequences. Unlike the previous screening stage, this task operates on the premise that the input sequences are phage-associated and challenges the model to discern whether a given contig is pure or contaminated by host-derived genetic sequences.

Task 3: Completeness Estimation requires the model to evaluate the assembly quality of a phage genome, addressing the prevalence of partial viral sequences in metagenomic data due to insufficient sequencing depth. This task is formulated as a four-way classification problem where the model must categorize sequences into COMPLETE (representing full-length genomes), HIGH-QUALITY, MEDIUM-QUALITY, and LOW-QUALITY tiers based on their genomic integrity and the presence of features such as terminal repeats.

Stage 3: Phenotype Annotation. The final stage demands high-level biological reasoning to infer functional traits and ecological roles directly from genomic sequences. Distinguishing between virulent phages, which drive the lytic cycle, and temperate phages, capable of genomic integration, is essential for applications ranging from antibiotic alternatives to genome engineering. Furthermore, accurate host prediction remains a cornerstone for the effective deployment of phage therapy and biocontrol strategies.

Task 4: Lifestyle Classification requires the model to predict the survival strategy of a phage by distinguishing between virulent and temperate lifestyles. This binary classification task evaluates the ability of an LLM to recognize specific genomic signatures associated with integration machinery, such as integrases and attachment sites, versus those indicative of exclusive lytic cycles.

Task 5: Host Prediction challenges the model to identify the specific bacterial host of a phage across three taxonomic levels, including Order, Family, and Genus. This multi-class classification problem requires the decoding of subtle interaction signals, such as codon usage bias or receptor-binding

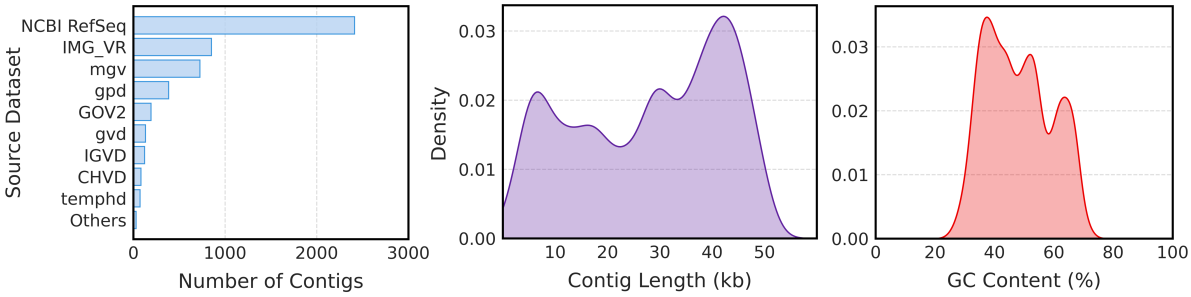


Figure 4: Statistical characteristics of the genomic sequences within the PhageBench dataset. The left panel displays the distribution of sequence sources. The central panel illustrates the probability density of contig lengths. The right panel depicts the distribution of GC content.

protein domains, to infer ecological relationships directly from raw genomic sequences without reliance on sequence homology.

Instruction Engineering. To better evaluate the ability of LLMs to understand and annotate raw phage genomic sequences, we designed task-specific instructions for each task category. To reduce data bias and ensure evaluation balance, all samples are formatted as multiple-choice questions: binary classification tasks (Tasks 1, 2, and 4) use two options, while multi-class classification tasks (Tasks 3 and 5) use four options. We randomly shuffled the option order for each sample to prevent position bias.

3.3 Dataset Statistics

Fig. 3 and Fig. 4 summarize the statistical properties of tasks in PhageBench. The dataset comprises a total of 5,600 samples across five tasks: Tasks 1, 2, and 5 each contain 1,200 samples, while Tasks 3 and 4 each contain 1,000 samples. All phage genomic sequences range from 500 bp to 50 kb in length, with a mean length of 28.15 kb and a median of 29.82 kb. The GC content across all phage sequences has a mean of 48.06% and a median of 47.00%, reflecting typical phage genomic composition. These statistics, along with the diverse distribution of sequence sources illustrated in Fig. 4, demonstrate that PhageBench encompasses a representative collection of phage genomic sequences suitable for evaluating LLM reasoning capabilities. More detailed statistics are provided in Appendix B.2.

4 Evaluation

We evaluate eight advanced LLMs spanning four categories based on weight accessibility and reasoning capability. *Open-weight, non-reasoning* models include LLaMA-4 (Meta, 2025) and Qwen3-235b (Yang et al., 2025). *Closed-weight, non-reasoning* models include GPT-4o-mini (OpenAI, 2024). *Open-weight, reasoning* models include GPT-OSS-120b (Agarwal et al., 2025). *Closed-weight, reasoning* models include GPT-5.2 (OpenAI, 2025), Gemini-3-flash (Google, 2025), Claude-sonnet-4.5 (Anthropic, 2025), and Qwen3-max (Alibaba, 2025). This selection covers a diverse range of model architectures and capabilities, enabling comprehensive evaluation of LLMs' performance on raw phage genome understanding and analysis.

All tasks are evaluated under a zero-shot chain of thought setting without in-context examples, allowing us to assess the models' inherent biological sequence reasoning capabilities in realistic usage scenarios. For models with built-in reasoning capabilities, we enable these features by default.

We use classification accuracy as the primary evaluation metric. Since PhageBench maintains strict class balance across all tasks, accuracy provides an unbiased measure of model performance. Additional details on experimental settings are provided in Appendix C.

4.1 LLMs' Performance on Raw Phage Genome Understanding

Table 1 presents the comprehensive performance of evaluated models across the five PhageBench tasks. The results indicate that while LLMs possess promising capabilities for decoding raw phage sequences, this domain remains a formidable challenge. Gemini-3-flash achieves the highest average accuracy at 56.33%, significantly surpassing the random baseline, followed by GPT-5.2. In contrast, non-reasoning models largely cluster around the random baseline. We observe two consistent trends.

Table 1: Model performance accuracy (%) on PhageBench tasks under zero shot CoT setting. Bold: the best performance. Underlined: the second performance. Avg.: the average across all tasks. T: task.

Model	T1: Phage Identification	T2: Contamination	T3: Completeness	T4: Lifestyle	T5: Host Prediction	Avg.
Non-reasoning						
GPT-4o-mini	50.00	50.08	24.50	48.40	27.75	40.15
LLaMA 4	53.25	49.58	27.10	54.50	28.58	42.60
Qwen3-235b	51.58	49.33	25.40	52.00	28.58	41.38
Reasoning						
GPT-OSS-120b	50.08	46.58	24.20	49.50	41.00	42.27
GPT-5.2	70.75	54.17	**43.10**	51.50	35.25	50.95
Gemini-3-flash	**70.83**	**58.83**	32.10	**57.40**	**62.50**	**56.33**
Claude-sonnet-4.5	57.92	51.83	35.50	50.00	44.17	47.88
Qwen3-max	57.42	52.58	33.80	49.90	34.75	45.69
Random	50.00	50.00	25.00	50.00	25.00	40.00

First, reasoning-enhanced models outperform their non-reasoning counterparts, suggesting that the ability to generate intermediate logical steps facilitates phage genome comprehension. Second, closed-weight models maintain a distinct performance advantage over open-weight alternatives, highlighting a capability gap that remains to be bridged in raw phage genome understanding.

Proficiency in Recognizing Statistical Genomic Features. Model performance is strongest when tasks rely on distinguishing broad compositional signatures. In phage contig identification task, Gemini-3-flash and GPT-5.2 achieve approximately 70% accuracy. Similarly, in host prediction, Gemini-3-flash reaches 62.50%, more than double the random baseline. These results imply that current LLMs can successfully discern the characteristic biases, such as codon usage or nucleotide composition, that differentiate phages from background noise or define host compatibility. This suggests models effectively treat DNA as a structured language with recognizable local syntax.

Competence in Global Structural Assessment. Moving beyond statistical features, completeness estimation evaluates the capacity to process genomic information at a global scale. Top models reliably exceed the 25% random baseline, with GPT-5.2 reaching 43.10%, indicating a fundamental grasp of structural integrity. However, performance remains lower than in identification tasks. This intermediate performance likely reflects the challenge of long-range dependency, as determining completeness necessitates verifying distal structural features, such as direct terminal repeats (DTR) at the sequence boundaries. While LLMs exhibit awareness of these global structures, maintaining such extended coherence remains a bottleneck compared to statistical pattern recognizing.

Challenges in Fine-grained Inference. Conversely, models encounter significant obstacles when tasks demand disentangling highly similar sequences or identifying sparse functional motifs. Contamination detection proves difficult because phage and host sequences share high statistical similarity driven by long-term coevolution, rendering simple statistical features matching insufficient. Lifestyle classification presents a profound challenge in functional inference. Unlike taxonomic classification, differentiating virulent from temperate phages requires identifying specific elements, such as integrases, within long sequences. Current models struggle to perform this level of precise retrieval and reasoning directly from raw nucleotides.

4.2 Performance Across Sequence Length

Phage genomes exhibit significant diversity in length. This biological variation raises a critical question for practical application: *at what genome scale can LLMs reliably interpret raw genomic sequences?* To answer this, we analyzed model accuracy across length bins ranging from 0.5kb to 50kb for Tasks 1–4 (Fig. 5), excluding Task 5 due to sampling limitations.

The results reveal distinct behavioral patterns rooted in the nature of each task. For phage identification, top models display remarkable stability across all length bins. This length-invariance suggests that the discriminative features are syntactic

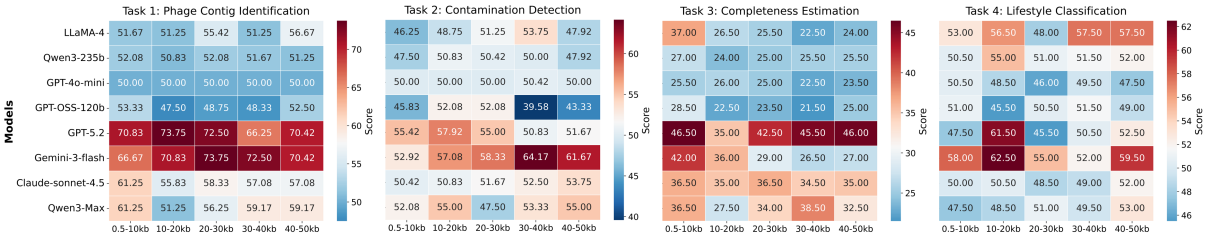


Figure 5: Performance heatmaps across sequence length bins for Tasks 1–4. Each cell represents the accuracy of a specific model on sequences within a particular length range.

Table 2: Comparison of model accuracy (%) between zero-shot and zero-shot with CoT modes. Grey cells indicate the $\Delta = Acc_{CoT} - Acc_{0-shot}$, which is the improvement of CoT. Positive values indicate CoT improvement. Bold values denote the best 0-shot performance per task.

Model	T1: Phage Identification		T2: Contamination		T3: Completeness		T4: Lifestyle		T5: Host Prediction		Average	
	0-shot	Δ	0-shot	Δ	0-shot	Δ	0-shot	Δ	0-shot	Δ	0-shot	Δ
GPT-4o-mini	50.00	0.00	50.00	+0.08	27.90	−3.40	48.20	+0.20	26.75	+1.00	40.57	−0.42
LLaMA-4	51.75	+1.50	51.33	−1.75	25.60	+1.50	50.60	+3.90	27.83	+0.75	41.42	+1.18
Qwen3-235b	50.00	+1.58	50.42	−1.09	25.00	+0.40	50.60	+1.40	28.42	+0.16	40.89	+0.49
GPT-5.2	68.50	+2.25	51.42	+2.75	**36.60**	+6.50	50.50	+1.00	29.17	+6.08	47.24	+3.71
Gemini-3-flash	**70.25**	+0.58	**56.42**	+2.41	30.20	+1.90	**58.20**	−0.80	**64.58**	−2.08	**55.93**	+0.40
Qwen3-Max	50.08	+7.34	50.33	+2.25	24.90	+8.90	48.70	+1.20	27.58	+7.17	40.32	+5.37

in nature and remain easily detectable regardless of context size. Conversely, completeness estimation exposes the limitations of global structural processing, where accuracy declines notably as sequences approach 50kb. This degradation reflects a failure in long-range dependency, as current architectures struggle to link distal features required to verify genome closure in extended contexts.

Complex inference tasks exhibit divergent responses to increased context. Contamination detection benefits from longer sequences, as extended inputs allow for evidence accumulation of heterogeneous signals that reveal host-derived fragments. In contrast, lifestyle classification remains near-random across all lengths. This indicates a semantic bottleneck rather than a contextual one, implying that the primary challenge lies in the precise retrieval of sparse functional markers which is not ameliorated simply by providing additional nucleotide context.

4.3 Impact of Test-Time Scaling

While our main evaluation shown in Table 1 utilizes CoT prompting to maximize model capabilities, recent studies suggest that explicit reasoning is not universally beneficial across all domains (Wang et al., 2023; Turpin et al., 2023). To quantify the extent to which test-time computational scaling specifically enhances raw genomic understanding, we conducted an ablation study comparing CoT against a direct zero-shot baseline.

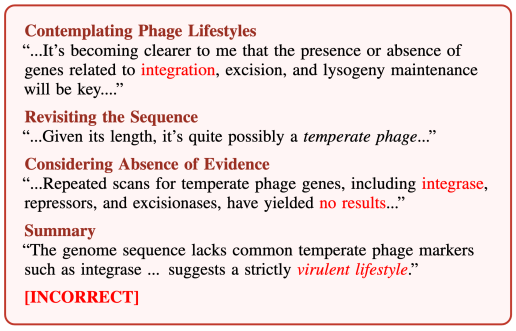
As shown in Table 2, allocating additional test-time computation generally serves as a powerful amplifier for decoding phage genomes. The majority of models exhibit positive performance gains when reasoning is activated. This confirms that for biological sequences, thinking time can effectively translate into performance gains.

However, we observe distinct performance regressions in specific LLMs, such as Gemini-3-flash on host prediction. In these specific cases, the model's intuitive alignment with genomic patterns proves more robust than its generated reasoning chains, highlighting the nuance required when applying test-time scaling to biological data.

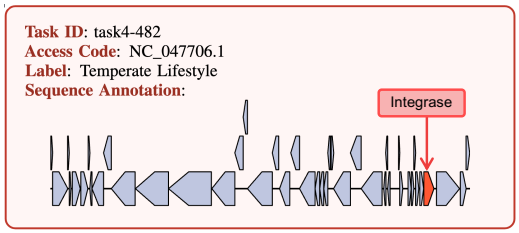
4.4 Success and Failure Analysis

To understand the factors underlying model performance, we analyzed their accuracy stratified by ground-truth labels (see Appendix C.3), and examined the chain-of-thought reasoning traces of Gemini-3-flash.

Our analysis reveals that leading models possess genuine pattern recognition capabilities for genomic sequences. In phage identification, Gemini-3-flash achieves 99.0% accuracy on true phages. Crucially, its performance on non-phage samples tracks with biological similarity: accuracy is high



(A) Model's Hallucinated Reasoning Chain.



(B) Ground Truth Label.

Figure 6: A case study of hallucinated reasoning in Lifestyle Classification. (A) While the model correctly identifies integrase as a key marker, it incorrectly asserts its absence in the sequence.(B) The ground truth annotation confirms the presence of integrase.

for eukaryotic protozoa (69.0%) but drops to 10.0% for prokaryotic plasmids, which share genomic features with phages. The reasoning traces frequently cite nucleotide composition, GC content distributions, and the presence of open reading frames as discriminative features. This implies the model successfully discriminates broad biological domains but struggles with fine-grained structural mimics.

However, these models are fundamentally constrained by a disconnect between abstract biological knowledge and precise sequence grounding. The most common errors made by Gemini-3-flash involve explicit claims about gene absence or existence contrary to fact. As visualized in Fig. 6, the model generates biologically plausible reasoning without the capacity to verify specific claims against the input sequence. Further qualitative analysis is provided in Appendix D.

5 Findings and Open Directions

While general-purpose LLMs demonstrate promising capabilities in genomic interpretation, our evaluation uncovers significant deficiencies in fine-grained sequence reasoning and long-range dependency modeling. To accelerate scientific discovery in phage genomics, we call for closer collaboration between the NLP and Biology communities. We highlight two critical directions for future research.

Enhancing Sequence-Grounded Reasoning Capabilities. Our analysis suggests a fundamental disconnect between the model's abstract biological knowledge and its perception of raw nucleotide sequences. To bridge this gap, we propose two strategies. The first is Multi-modal Alignment. By fine-tuning LLMs on paired raw genomes and functional annotations, models can learn to map sequence motifs directly to biological phenotypes, moving beyond simple statistical correlations to semantic understanding. The second strategy involves Tool-Augmented Reasoning. Recognizing that visual scanning of long sequences is an inherent weakness of token-based models, future frameworks may adopt an *Extract-then-Infer* workflow. By granting LLMs access to external tools, models can verify their intuitive hypotheses against rigorous computational results.

Optimizing Global Dependency Modeling. The degradation of performance in completeness estimation tasks as sequence length increases highlights that LLMs struggle to maintain the global coherence necessary to correlate distal features, such as direct terminal repeats. Since determining genome completeness often relies on specific structural rules (e.g., checking if the 5' and 3' ends match), prompts that explicitly instruct the model to first extract the terminal sequences and then perform a local comparison may effectively bypass the interference of intermediate context.

6 Conclusion

This study introduces PhageBench, a new benchmark designed to advance the phage genome annotation. By establishing a rigorous evaluation protocol, we assessed advanced LLMs and confirmed their latent capabilities for phage genome analysis. However, significant performance bottlenecks remain before LLMs can effectively drive scientific discovery in phage genomics, particularly regarding the critical challenge of grounding abstract biological knowledge in concrete sequence evidence. These findings offer novel insights for extending the utility of LLMs within the life sciences and establish a basis for future optimization. We hope that PhageBench will foster innovative collaboration between the NLP and life science communities, accelerating the discovery and application of phage dark matter in real-world scenarios.

Limitations

Our study has three primary limitations regarding data scope, tokenization mechanisms, and model transparency.

Data. To balance computational costs with current context window capabilities, PhageBench restricts genomic sequences to a maximum length of 50kb. While this covers a significant portion of the phage population, it inherently excludes larger phage entities, such as jumbo phages (genomes exceeding 200kb), which often harbor complex metabolic genes and distinct evolutionary strategies. Consequently, our findings may not fully generalize to these more extensive and complex genomic architectures.

Tokenization. We evaluated models using their default tokenization strategies, which utilize Byte Pair Encoding (BPE) optimized for natural language. This approach is likely suboptimal for genomic data, as BPE tends to aggregate nucleotide subsequences based on statistical frequency rather than biological significance, potentially obscuring functional motifs. Future work could explore alternative preprocessing strategies, such as inserting spaces to enforce single-base tokenization or adopting k-mer based aggregation, to better align input representations with biological logic.

Models. The inclusion of closed-source models introduces the unavoidable risk of data contamination. Since the pre-training corpora of these models are not public, we cannot definitively rule out the possibility that sequences from PhageBench were present in their training data. However, the frequent occurrence of hallucinatory reasoning observed in our error analysis suggests that even if models have encountered these sequences, they rely heavily on imperfect recall or probabilistic guessing rather than robust memorization.

Ethical Considerations

The PhageBench dataset presented in this study is constructed entirely from publicly available genomic repositories. We confirm that no proprietary, confidential, or personally identifiable information is involved in this work, and all data collection procedures strictly adhere to established community norms and best practices for genomic data usage.

Regarding biosafety, phages are viruses that specifically infect and replicate within bacteria; they do not infect human cells and are generally considered to have a high safety profile. Furthermore, the primary objective of this benchmark is to evaluate the capability of LLMs to understand and reason over existing biological sequences. This study does not involve the generative design of novel biological entities or the modification of functional pathogens.

However, we acknowledge the broader ethical implications and potential dual-use risks associated with the intersection of artificial intelligence and biology. We urge all users utilizing PhageBench to strictly adhere to ethical guidelines and biosafety regulations governing computational biology to ensure the responsible development of AI technologies.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Alibaba. 2025. Qwen3-max: Just scale it. Accessed: 2025-09-24.

Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

Anthropic. 2025. Introducing claude sonnet 4.5. Accessed: 2025-09-29.

Žiga Avsec, Natasha Latysheva, Jun Cheng, Guido Novati, Kyle R Taylor, Tom Ward, Clare Bycroft, Lauren Nicolaisen, Eirini Arvaniti, Joshua Pan, and 1 others. 2025. Alphagenome: advancing regulatory variant effect prediction with a unified dna sequence model. *bioRxiv*, pages 2025–06.

Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, James Ostell, Kim D Pruitt, and Eric W Sayers. 2018. Genbank. *Nucleic acids research*, 46(D1):D41–D47.

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, and 1 others. 2025. Genome modeling and design across all domains of life with evo 2. *BioRxiv*, pages 2025–02.

Antonio Pedro Camargo, Simon Roux, Frederik Schulz, Michal Babinski, Yan Xu, Bin Hu, Patrick SG Chain, Stephen Nayfach, and Nikos C Kyrpides. 2024. Identification of mobile genetic elements with genomad. *Nature biotechnology*, 42(8):1303–1312.

Wenduo Cheng, Junhong Shen, Mikhail Khodak, Jian Ma, and Ameet Talwalkar. 2024. L2g: Repurposing language models for genomics tasks. *bioRxiv*.

Wenduo Cheng, Zhenqiao Song, Yang Zhang, Shike Wang, Danqing Wang, Muyu Yang, Lei Li, and Jian Ma. 2025. Dnalongbench: a benchmark suite for long-range dna prediction tasks. *bioRxiv*.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, and 1 others. 2025. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297.

Moïra B Dion, Frank Oechslin, and Sylvain Moineau. 2020. Phage diversity, genomics and phylogeny. *Nature Reviews Microbiology*, 18(3):125–138.

Google. 2025. Gemini 3 flash: Best for frontier intelligence at speed. Accessed: 2025-12-17.

Ann C Gregory, Olivier Zablocki, Ahmed A Zayed, Allison Howell, Benjamin Bolduc, and Matthew B Sullivan. 2020. The gut virome database reveals age-dependent patterns of virome diversity in the human gut. *Cell host & microbe*, 28(5):724–740.

Ann C Gregory, Ahmed A Zayed, Nádia Conceição-Neto, Ben Temperton, Ben Bolduc, Adriana Alberti, Mathieu Ardyna, Ksenia Arkhipova, Margaux Carmichael, Corinne Cruaud, and 1 others. 2019. Marine dna viral macro-and microdiversity from pole to pole. *Cell*, 177(5):1109–1123.

Katarína Grešová, Vlastimil Martinek, David Čechák, Petr Šimeček, and Panagiotis Alexiou. 2023. Genomic benchmarks: a collection of datasets for genomic sequence classification. *BMC Genomic Data*, 24(1):25.

Susanna R Grigson, George Bouras, Bhavya Papudeshi, Vijini Mallawaarachchi, Michael R Roach, Przemyslaw Decewicz, and Robert A Edwards. 2025. Synteny-aware functional annotation of bacteriophage genomes with phynteny. *bioRxiv*, pages 2025–07.

Jiarong Guo, Ben Bolduc, Ahmed A Zayed, Arvind Varsani, Guillermo Dominguez-Huerta, Tom O Delmont, Akbar Adjie Pratama, M Consuelo Gazitúa, Dean Vik, Matthew B Sullivan, and 1 others. 2021. Virsorter2: a multi-classifier, expert-guided approach to detect diverse dna and rna viruses. *Microbiome*, 9(1):37.

Xin Hou, Yong He, Pan Fang, Shi-Qiang Mei, Zan Xu, Wei-Chen Wu, Jun-Hua Tian, Shun Zhang, Zhen-Yu Zeng, Qin-Yu Gou, and 1 others. 2024. Using artificial intelligence to document the hidden rna virosphere. *Cell*, 187(24):6929–6942.

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.

Samuel H King, Claudia L Driscoll, David B Li, Daniel Guo, Aditi T Merchant, Garyk Brixi, Max E Wilkinson, and Brian L Hie. 2025. Generative design of novel bacteriophages with genome language models. *BioRxiv*, pages 2025–09.

Travis N Mavrich and Graham F Hatfull. 2017. Bacteriophage evolution differs by host, lifestyle and genome. *Nature Microbiology*, 2(9):1–9.

Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. Accessed: 2025-04-05.

Stephen Nayfach, Antonio Pedro Camargo, Frederik Schulz, Emiley Eloe-Fadrosh, Simon Roux, and Nikos C Kyrpides. 2021a. Checkv assesses the quality and completeness of metagenome-assembled viral genomes. *Nature biotechnology*, 39(5):578–585.

Stephen Nayfach, David Páez-Espino, Lee Call, Soo Jen Low, Hila Sberro, Natalia N Ivanova, Amy D Proal, Michael A Fischbach, Ami S Bhatt, Philip Hugenholtz, and 1 others. 2021b. Metagenomic compendium of 189,680 dna viruses from the human gut microbiome. *Nature microbiology*, 6(7):960–970.

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, and 1 others. 2024. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, and 1 others. 2023. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36:43177–43201.

Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, and 1 others. 2016. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence. Accessed: 2024-07-18.

OpenAI. 2025. Introducing gpt-5.2. Accessed: 2025-12-11.

Cheng Peng, Jiayu Shang, Jiaojiao Guan, and Yanni Sun. 2025. Viralqc: A tool for assessing completeness and contamination of predicted viral contigs. *arXiv preprint arXiv:2504.05790*.

Cheng Peng, Jiayu Shang, Jiaojiao Guan, Donglin Wang, and Yanni Sun. 2024. Viralm: empowering virus discovery through the genome foundation model. *Bioinformatics*, 40(12):btae704.

Akansha Prasad, Shadman Khan, Fatima Arshad, Hareet Sidhu, Kyle Jackson, Roderick MacLachlan, Ekaterina Kvitka, Veronica Grignano, Hannah Mann, Carlos DM Filipe, and 1 others. 2025. Bacteriophage-loaded microneedle patches for targeted and minimally disruptive foodborne pathogen decontamination. *Science Advances*, 11(44):eadx6918.

Yuanhao Qu, Kaixuan Huang, Ming Yin, Kanghong Zhan, Dyllan Liu, Di Yin, Henry C Cousins, William A Johnson, Xiaotong Wang, Mihir Shah, and 1 others. 2025. Crispr-gpt for agentic automation of gene-editing experiments. *Nature Biomedical Engineering*, pages 1–14.

Forest Rohwer. 2003. Global phage diversity. *Cell*, 113(2):141.

Daniel A Russell and Graham F Hatfull. 2017. Phagesdb: the actinobacteriophage database. *Bioinformatics*, 33(5):784–786.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. 2024. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *Proceedings of machine learning research*, 235:43632.

Jiayu Shang, Cheng Peng, Jiaojiao Guan, Dehan Cai, Donglin Wang, and Yanni Sun. 2025a. From genomic signals to prediction tools: a critical feature analysis and rigorous benchmark for phage–host prediction. *Briefings in Bioinformatics*, 26(6):bbaf626.

Xinyi Shang, Xu Liao, Zhicheng Ji, and Wenpin Hou. 2025b. Benchmarking large language models for genomic knowledge with geneturing. *Briefings in Bioinformatics*, 26(5):bbaf492.

Mikael Skurnik, Sivan Alkalay-Oren, Maarten Boon, Martha Clokie, Thomas Sicheritz-Pontén, Krystyna Dąbrowska, Graham F Hatfull, Ronen Hazan, Matti Jalasvuori, Saija Kiljunen, and 1 others. 2025. Phage therapy. *Nature Reviews Methods Primers*, 5(1):9.

Steffanie A Strathdee, Graham F Hatfull, Vivek K Mutalik, and Robert T Schooley. 2023. Phage therapy: From biological mechanisms to future directions. *Cell*, 186(1):17–31.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. 2025. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, 646(8085):716–723.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 2717–2739.

Ruo Han Wang, Shuo Yang, Zhixuan Liu, Yuanzheng Zhang, Xueying Wang, Zixin Xu, Jianping Wang, and Shuai Cheng Li. 2024. Phagescope: a well-annotated bacteriophage database with automatic analyses and visualizations. *Nucleic acids research*, 52(D1):D756–D761.

Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping Day, Christina Ross, Robert Leaman, and Zhiyong Lu. 2025. Geneagent: self-verification language agent for gene-set analysis using domain databases. *Nature Methods*, 22(8):1677–1685.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. 2023. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*.

A Broader Impact

This work extends beyond specific phage genome annotation tasks, holding significant implications for both the artificial intelligence and computational biology communities. By introducing PhageBench, we push the frontier of Large Language Models from processing scientific literature to decoding the fundamental language of life itself, highlighting the urgent need for models that adhere to strict biological rules rather than relying on probabilistic approximation. On a societal level, efficient genome annotation is critical for addressing the antibiotic resistance crisis through phage therapy, which is currently bottlenecked by uncharacterized phage dark matter. Our study explores the potential of LLMs to streamline the screening and characterization of these entities, offering a perspective on how AI might assist in shortening the timeline from environmental discovery to clinical application. Ultimately, we hope this benchmark serves as a catalyst for interdisciplinary collaboration, demystifying genomic data for AI researchers while objectively demonstrating the current capabilities and limitations of LLMs to the biological community.

B PhageBench Benchmark Details

B.1 Detailed Data Construction

B.1.1 Source Data Collection and Preprocessing

The foundation of the PhageBench dataset is built upon high-quality genomic sequences derived from authoritative repositories. We utilized PhageScope (Wang et al., 2024) and NCBI RefSeq (O'Leary et al., 2016) as the primary reservoir for phage genomes. PhageScope is a comprehensive database that aggregates sequences from multiple public archives, including GenBank (Benson et al., 2018), RefSeq (O'Leary et al., 2016), PhagesDB (Russell and Hatfull, 2017), GOV2 (Gregory et al., 2019), GVD (Gregory et al., 2020), and MGV (Nayfach et al., 2021b), while providing systematic annotations generated by fifteen SOTA bioinformatics tools. This integration ensures that our positive samples represent a diverse and high-confidence collection of phage genomes. For the construction of negative samples required in phage contig identification task, we sourced non-phage sequences directly from the NCBI RefSeq database. We applied a strict length filtering criterion. Only sequences with clear labels and lengths ranging from 0.5 to 50 kilobases were retained. Furthermore, we performed a quality control step to exclude sequences containing invalid bases, such as ambiguous 'N' characters, ensuring that the model input consists solely of high-quality raw nucleotide sequences.

B.1.2 Construction of Phage Contig Identification

To evaluate the capability of LLMs in distinguishing phage sequences from environmental noise, we constructed a binary classification dataset with a balanced distribution of positive and negative samples. The negative samples were curated from the NCBI RefSeq database, spanning six distinct biological categories: non-phage viruses, protozoa, plasmids, fungi, bacteria, and archaea. We implemented a category-specific processing strategy to handle the significant length variations among these biological entities. For non-phage viruses, protozoa, and plasmids, we utilized their complete genomic sequences, as their natural lengths generally fall within our target distribution. In contrast, for fungi, bacteria, and archaea, which typically possess much larger genomes, we applied a random cropping strategy to generate fragmented sequences. This process ensured that the length distribution of the negative fragments strictly matched that of the positive phage samples. To prevent the model from distinguishing classes based solely on sequence length, we maintained a strict 1:1 ratio between phage and non-phage samples, and within the negative set, the six biological sources were also balanced equally to ensure diversity.

B.1.3 Mock Data Generation for Contamination Detection

Unlike random noise insertion, we followed previous work (Peng et al., 2025) to simulate realistic biological contamination events. We selected host fragments based on the ground-truth host metadata provided by PhageScope, ensuring that the contaminant sequences correspond to the actual biological hosts of the phages. We defined the contamination ratio (r) as the length of the host fragment divided by the total sequence length. To assess model sensitivity, we generated chimeric samples at three specific contamination ratios: 12.5%, 25%, and 50%. We employed three distinct insertion strategies, Prefix, Suffix and Internal, to simulate different assembly error patterns as shown in Fig. 2. The generation process is formally described in Algorithm 1.

Algorithm 1 Mock Data Generation

Require: Phage Sequence S_p, Host Sequence S_h, Contamination Ratio r, Mode M
Ensure: Chimeric Sequence $S_{chimera}$

```
Calculate target total length L_total ← Length(S_p)/(1 − r)
Calculate host fragment length L_host ← L_total × r
Extract host fragment F_h ← S_h[random : random + L_host]
if M is PREFIX then
    S_chimera ← F_h + S_p
else if M is SUFFIX then
    S_chimera ← S_p + F_h
else if M is INTERNAL then
    Select random split point k where 0 < k < Length(S_p)
    S_chimera ← S_p[0 : k] + F_h + S_p[k : end]
end if
return S_chimera
```

Table 3: The specific instructions designed for each task in PhageBench.

Task	Question Template
Phage Contig Identification	Is the following nucleotide sequence derived from a bacteriophage genome?
Contamination Classification	This sequence is a phage-associated contig. Does it contain any DNA originating from a non-phage host genome?
Completeness Classification	What is the estimated genome completeness category of the following bacteriophage sequence?
Lifestyle Classification	What is the most likely lifestyle of the bacteriophage represented by the following genome sequence?
Host Prediction	Which host family/genus/order is the following bacteriophage most likely to infect?

B.1.4 Construction of Completeness Estimation

For genome completeness estimation, we extracted high-confidence completeness labels directly from the PhageScope database, which uses the standard tool CheckV (Nayfach et al., 2021a) for quality assessment. The samples were categorized into four classes: Complete, High-quality, Medium-quality, and Low-quality. To ensure a rigorous evaluation, we applied strict length binning and class balancing. Specifically, we ensured an equal number of samples for each of the four completeness categories within every defined length interval.

B.1.5 Construction of Phenotype Annotation

The phenotype annotation stage consists of lifestyle classification and host prediction. For lifestyle classification, we extracted lifestyle labels directly from PhageScope. For host prediction, we followed previous work (Shang et al., 2025a) to use the RefSeq Virus–Host Database (RefSeq-VHDB) as the data source. To construct the multiple-choice questions for this task, we implemented a hardness-aware distractor generation strategy. The incorrect options, or distractors, were randomly selected from different hosts within the same taxonomic level as the correct answer. For instance, if the ground truth host belongs to a specific Family, the distractors were sampled from other bacterial Families. This design challenges the model to distinguish between biologically comparable entities rather than identifying easy out-of-distribution targets.

B.1.6 Instruction Design and MCQ Construction

To transform the biological data into a format suitable for LLM evaluation, we employed an instruction engineering process. Each raw nucleotide sequence was encapsulated within a task-specific prompt template designed to simulate a bioinformatics query; the specific instructions and templates for each task are presented in Table 3. To standardize the evaluation and eliminate potential position bias, the order of the multiple-choice options was randomly shuffled for every sample. Finally, we required all models to generate outputs in a structured JSON format. This standardized output format facilitates automated parsing and ensures the accuracy of the quantitative evaluation.

B.2 Detailed Dataset Statistics

B.2.1 General Statistical Characteristics

The PhageBench dataset is constructed to ensure robust statistical properties that reflect real-world biological diversity while maintaining computational feasibility for LLMs. As illustrated in Fig. 4, the dataset aggregates genomic sequences from a wide array of sources, ranging from environmental metagenomes found in databases such as GOV2 and GVD to high-quality isolates from NCBI RefSeq. This diversity ensures that LLMs are evaluated across various ecological niches and sequencing protocols. Regarding sequence properties, we strictly controlled the length of genomic segments to fall within the range of 0.5 to 50 kb. The resulting dataset exhibits a mean sequence length of approximately 28.15 kb and a median of 29.82 kb. Furthermore, the GC content of the sequences spans a wide dynamic range with a mean of 48.06%, indicating that the dataset covers phages adapted to diverse host environments.

B.2.2 Fine-Grained Class Distribution

To prevent LLMs from exploiting statistical shortcuts or class priors, we enforced a rigorous class balancing strategy across all tasks. Fig. 7 and Fig. 8 present the detailed breakdown of sample distri-

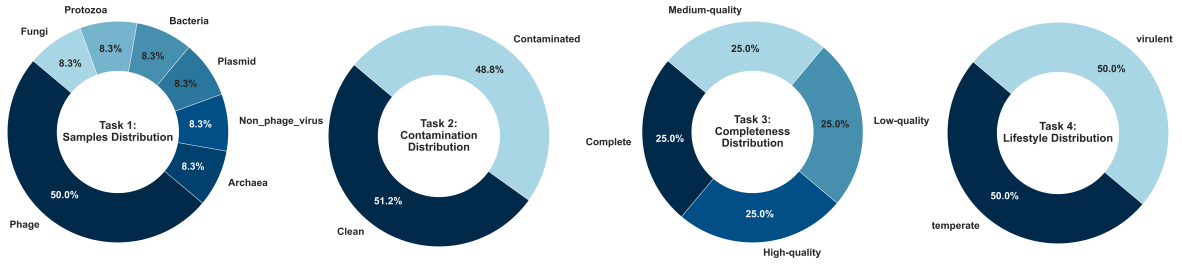


Figure 7: Detailed class and subclass distributions across the task 1–4. The pie charts demonstrate the strict balancing strategy applied to primary classes and quality tiers ensuring statistical fairness.

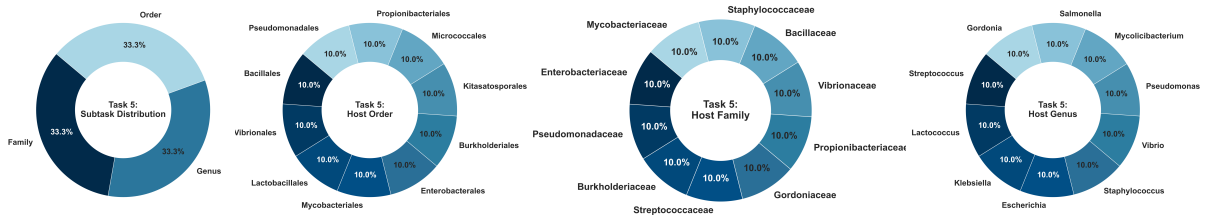


Figure 8: Detailed class distributions of task 5. The pie charts demonstrate the strict balancing strategy applied to fine-grained taxonomic categories.

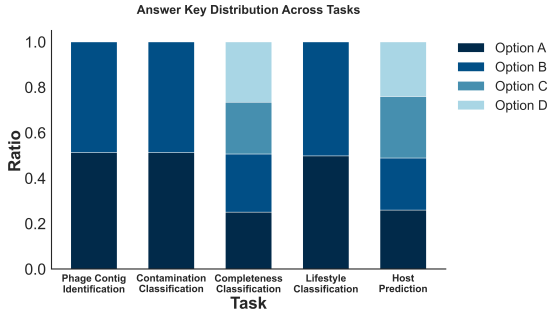


Figure 9: Distribution of correct answer keys across all tasks. The bar charts verify that the correct options (A, B, C, D) are uniformly distributed, confirming the effectiveness of the option shuffling strategy in mitigating position bias.

butions for each task. For phage contig identification, we maintained a strict 1:1 ratio between positive phage samples and negative non-phage samples. Crucially, the negative class is further stratified into six distinct biological categories, including bacteria, archaea, plasmids, fungi, protozoa, and non-phage viruses, which are uniformly distributed, with each category constituting approximately 8.3% of the total dataset. This granular balance ensures that the model cannot simply learn to distinguish phages from a single dominant background noise.

Similarly, for contamination detection, the dataset is balanced between clean and contaminated sequences. Completeness estimation task features a uniform distribution across four quality tiers, with Complete, High-quality, Medium-quality, and Low-quality samples each representing exactly 25.0% of the data. For lifestyle classification, we achieved a balanced split between virulent and temperate phages. Finally, host prediction task demonstrates our commitment to taxonomic fairness. The samples are equally divided across three taxonomic levels (Order, Family, Genus), and within each level, we ensured that specific host taxa are sampled evenly, as depicted in the taxonomic rings in Fig. 8. This design forces the model to perform genuine sequence-based reasoning rather than relying on the prevalence of common host species.

B.2.3 Answer Key Distribution

Beyond data sampling, the design of the evaluation prompt is critical for ensuring fairness, particularly for LLMs that may exhibit position bias. To mitigate this, we implemented a stochastic shuffling mechanism for the multiple-choice options in every sample. Fig. 9 illustrates the distribution of the correct answer keys (Option A, B, C, or D) across all five tasks. For binary classification tasks, the correct answers are equiprobably distributed between Option A and Option B. For multi-class tasks, the correct answers are uniformly distributed across all four options, with each option accounting for approximately 25% of the total. This uniform distribution confirms that our randomization strategy

effectively eliminated position bias, ensuring that the reported accuracy reflects the genomic understanding capabilities of LLMs rather than artifacts of prompt engineering.

B.3 Example Samples

In this section, we provide examples from PhageBench to illustrate the standardized input-output format used in our evaluation. Each example demonstrates how a raw nucleotide sequence is encapsulated within a task-specific instruction template, paired with shuffled multiple-choice options.

Example: Phage Contig Identification

Question: Is the following nucleotide sequence derived from a bacteriophage genome?

Sequence: TTTATGCGATTACCAACACGTGT...

Choices:
(A) It is not from a bacteriophage genome.
(B) It is from a bacteriophage genome.

Example: Contamination Classification

Question: This sequence is a phage-associated contig. Does it contain any DNA originating from a non-phage host genome?

Sequence: GAATCTGCCAAGAAGCAACGTC...

Choices:
(A) No, the sequence does not contain host contamination.
(B) Yes, the sequence contains host contamination."

Example: Completeness Classification

Question: What is the estimated genome completeness category of the following bacteriophage sequence?

Sequence: GGCGACGCGACAGTCATGGCATG...

Choices:
(A) High-quality.
(B) Medium-quality.
(C) Low-quality.
(D) Complete.

Example: Lifestyle Classification

Question: What is the most likely lifestyle of the bacteriophage represented by the following genome sequence?

Sequence: ACGGTCGCCGGTGTTGGTGGCTG...

Choices:
(A) Virulent (lytic).
(B) Temperate (lysogenic-capable)."

Example: Host Prediction

Question: Which host family is the following bacteriophage most likely to infect?

Sequence: CCCTCACCCTGAACGCGCCCAATT...

Choices:
(A) Enterobacteriaceae.
(B) Propionibacteriaceae.
(C) Vibrionaceae.
(D) Staphylococcaceae.

C Evaluation Details

C.1 Implementation Details

All models are accessed via the OpenRouter API, which provides unified access to diverse LLMs without requiring local computational resources. For models that support temperature configuration, we set the sampling temperature to 0.1 to ensure reproducibility. For models under chain-of-thought or thinking mode, we configure the maximum reasoning length to 2,048 tokens and set the thinking effort to medium. We include a random baseline for reference, representing the expected accuracy under uniform random guessing: 50% for binary classification tasks (Tasks 1, 2, and 4) and 25% for four-way classification tasks (Tasks 3 and 5).

For the zero-shot baselines, we excluded GPT-OSS-120b due to the inability to disable its reasoning mode, and Claude-Sonnet-4.5 due to computational cost constraints.

C.2 Evaluation Prompts

To ensure a rigorous and standardized assessment of Large Language Model capabilities across diverse genomic tasks, we designed a unified prompt engineering framework. All evaluation prompts share a consistent structural foundation, establishing a domain-expert persona and enforcing strict constraints on output formatting. Specifically, we explicitly instructed models to refrain from restating the lengthy raw nucleotide sequences in their re-

Prompt Used for Question Answering under Zero Shot+CoT

```
You are an expert specializing in phage genomics.
Your task is to answer single-choice questions by analyzing the given nucleotide sequence and the question.
Based on the evidence presented in the nucleotide sequence, use logical reasoning and your best knowledge of biological principles to determine your answer.
Do not copy or restate the full DNA sequence in your response and reasoning; at most quote short substrings less than 30 characters if needed.
Think step-by-step.

Question: $question$
<sequence>
$sequence$
</sequence>
Choices: $choices$

Format your answer in the following format, with no extra text before or after:
{"analysis": "Your analysis within 1-3 concise sentences for the given sequence and question.",
"answer": "X"}
where X is one of: $options$.
```

Figure 10: The zero shot with CoT prompt.

sponses to prevent context window exhaustion and required all outputs to be structured in a machine-parseable JSON format. This design facilitates automated quantitative evaluation and ensures that performance metrics reflect genuine reasoning ability rather than parsing errors.

To quantify the impact of explicit reasoning strategies on genomic understanding, we implemented two distinct prompt variations corresponding to our experimental settings. The first variation, the Zero-shot CoT prompt (Fig. 10), incorporates the "Think step-by-step" directive and mandates an "analysis" field within the output JSON. This structure forces the model to articulate its intermediate logical steps before concluding with a prediction, allowing us to inspect the biological validity of its reasoning trace. The second variation, the standard Zero-shot prompt (Fig. 11), serves as a direct inference baseline. It removes the reasoning triggers and the analysis field, requiring the model to output only the final prediction..

C.3 Deep Analysis of Results by Labels

To further investigate the underlying behavioral patterns of LLMs, we analyzed model performance stratified by ground-truth labels as illustrated in Fig. 12.

In phage contig identification task, we observe a distinct performance disparity across biological domains. Models achieve high accuracy when distinguishing phages from eukaryotic sequences such as protozoa and fungi. However, performance drops to near-random levels when distinguishing phages from plasmids. This contrast implies a reliance on broad statistical features rather than syntactic understanding. Plasmids and phages are both prokaryotic mobile genetic elements that share similar nucleotide composition and GC content. The inability of models to differentiate these statistically similar entities suggests they fail to identify the specific structural gene modules, such as capsid or tail assembly genes, that definitively characterize a phage genome.

The contamination detection task further highlights the sensitivity of models to signal strength. We observe a positive correlation where accuracy improves as the contamination ratio increases from 12.5% to 50%. Notably, models frequently fail to identify contamination at the 12.5% level. This indicates a high threshold for detecting heterogeneity. Since host and phage sequences often possess similar statistical properties, models likely smooth over the 12.5% contamination as natural viral variation. This reflects a limitation in checking for functional internal consistency, as the models cannot semantically identify the intrusion of non-phage metabolic genes within the viral context.

For completeness estimation, most models ex-

Prompt Used for Question Answering under Zero Shot

```
You are an expert specializing in phage genomics.

Your task is to answer single-choice questions by analyzing the given nucleotide sequence and the question.

Based on the evidence presented in the nucleotide sequence, use your best knowledge of biological principles to determine your answer.

Do not copy or restate the full DNA sequence in your response; at most quote short substrings less than 30 characters if needed.

Question: $question$

<sequence>
$sequence$
</sequence>

Choices: $choices$

Format your answer in the following format, with no extra text before or after:

{"answer": "X"}

where X is one of: $options$.
```

Figure 11: The zero shot prompt.

hibit moderate performance. This reflects the difficulty of associating distal features like terminal repeats across long sequence contexts. Among the evaluated architectures, GPT-5.2 demonstrates superior performance in this category, suggesting a relatively stronger capacity for maintaining global structural coherence compared to other reasoning models.

In the lifestyle classification task, the classification of lifestyle often hinges on the presence of sparse functional markers such as integrase genes. The lower performance in this category points to a bottleneck in semantic retrieval where models struggle to precisely locate over singular gene events buried within extended genomic sequences.

Finally, the Host Prediction task reveals that model performance does not decline linearly with increasing taxonomic resolution. The evaluation results indicate that accuracy peaks at the Family level, followed by the Order level, while the Genus level remains the most challenging category for current models. Across all three taxonomic tiers, Gemini-3-flash consistently outperforms other architectures and achieves state-of-the-art results, highlighting its robust capacity for identifying host-specific genomic signatures relative to the other evaluated LLMs.

D Qualitative Results

To provide a deeper understanding of how general-purpose LLMs interpret raw phage genome sequences, we conducted a qualitative analysis of the reasoning traces generated by Gemini-3-flash. This section examines specific examples across the five PhageBench tasks. The analysis is categorized into success cases (Fig. 13 and 17) and failure cases (Fig. 14-16).

D.1 Analysis of Success Cases

In phage contig identification task, Gemini-3-flash demonstrates a robust ability to recognize the fundamental syntax of phage genomes. As illustrated in Fig. 13, the model's reasoning trace explicitly identifies characteristic biological features inherent to phages rather than relying on simple memorization. The model correctly cites high gene density and modular genomic organization as primary evidence. Furthermore, it detects specific regulatory elements and structural gene modules typical of the *Siphoviridae* or *Myoviridae* families. This suggests that the model can successfully parse raw nucleotide sequences to extract higher-order syntactic structures that distinguish phage contig from other biological backgrounds.

Similarly, the model exhibits strong performance in host prediction by leveraging statistical genomic signatures. In Fig. 17, the model accurately predicts *Pseudomonadaceae* as the host family. The

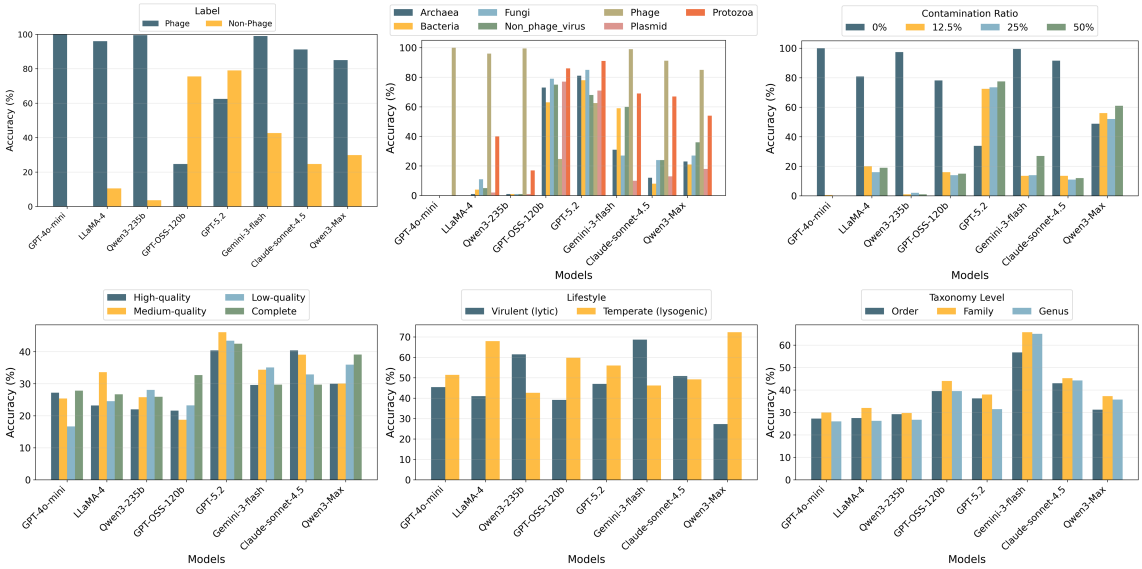


Figure 12: Detailed performance analysis of LLMs on PhageBench tasks. (Top) Phage Identification results separated by binary classes and fine-grained biological sources, followed by Contamination Detection across varying contamination ratios; (Bottom) Completeness Estimation across quality tiers, Lifestyle Classification phenotypes, and Host Prediction at three taxonomic resolutions.

reasoning trace reveals that the model does not merely guess but actively computes compositional metrics, explicitly noting that the sequence possesses a high GC content of approximately 60–65%. It then correctly associates this statistical bias with the genomic characteristics of the *Pseudomonas* genus. This indicates that the model is capable of grounding its predictions in concrete statistical properties of the input sequence.

D.2 Analysis of Failure Cases

Despite these successes, the model shows significant limitations when processing heterogeneous sequences, as observed in Fig. 14. The model failed to detect the contamination and incorrectly classified the sequence as a pure phage contig. The reasoning trace asserts that no significant regions of non-phage host DNA were identified. This failure suggests that the model tends to homogenize the input context; once it identifies strong phage signals, it may overlook or suppress conflicting signals from the host segments. This insensitivity to sequence heterogeneity highlights a deficiency in fine-grained sequence segmentation within mixed biological contexts.

The limitations of the model are further exacerbated by hallucinatory reasoning in tasks requiring long-range dependency analysis. As shown in Fig. 15, the model incorrectly classified a high-quality linear fragment as a complete genome. Crucially, the reasoning trace contains a fabrication of structural evidence, where the model claims to detect repetitive terminal sequences indicative of circularity. Since the input sequence was a linear contig lacking these physical ends, this represents a structural hallucination.

Finally, in the lifestyle classification task (Fig. 16), the model incorrectly predicted a virulent lifestyle for a temperate phage by erroneously identifying the sequence as belonging to the well-known virulent bacteriophage *PhiX174*. Based on this false identification, the model logically deduced a lytic lifestyle, ignoring the actual sequence evidence that might suggest a temperate nature, such as integration modules. This indicates that the model's internal knowledge base can sometimes override specific sequence evidence, leading to knowledge-driven hallucinations.

Phage Contig Identification

Task ID: task1-983

Instruction: Is the following nucleotide sequence derived from a bacteriophage genome?

Sequence: TTATTATTTTATCTTCGATGAATGTGGCGACGGCATGATTCTTGCCGATTTTCAGCGGCGCGCCCT...

Choices: (A) It is not from a bacteriophage genome. (B) It is from a bacteriophage genome.

Gemini-3-flash Output:

Examining Sequence for Phage

I'm now diving into the sequence, which is quite hefty. I'm focusing on identifying telltale signs of a bacteriophage genome, such as looking for typical phage-specific genes or features that could indicate its origins. I've also begun checking for patterns that might suggest its bacterial host.

Scrutinizing Sequence Further

Okay, I'm now zeroing in on the finer details. I've been meticulously examining the sequence in the beginning. I'm noting potential regulatory elements and looking for any repeating motifs, which could hint at phage-specific features. I'm also observing gene density as I go and making a mental note of all unusual GC content. I'm focusing on identifying any specific phage-like sequences within the genome.

Uncovering Sequence Motifs

I'm now intently searching for patterns in this sequence. Specific phage-like sequences are my target. I'm noting potential regulatory elements and repeating motifs, as well as gene density and unusual GC content. I'm focusing on trying to identify any sequence that would be specific to this phage.

Decoding Potential Signals

My focus is on pinpointing potential regulatory elements and their significance. I am meticulously analyzing the sequence for distinctive patterns and signals. Currently, I'm carefully examining the sequence for unusual patterns or motifs. I'm keen to find any sequence that could provide clues to the phage's life cycle.

Interpreting Sequence Data

I'm now focusing on this new data. I've done a first pass, and I'm looking for any open reading frames and unusual base compositions. I have been carefully examining the sequence for potential start and stop codons. So far, the GC content seems relatively stable overall. I'm keen to identify any sequences that might hint at promoter or terminator regions. I am currently examining the sequence for any repeats or unusual motifs.

Assessing GC-Rich Regions

My current focus is on these GC-rich areas I've just highlighted. I'm keen to determine if they hold any clues. I've been carefully analyzing the sequence for potential regulatory elements and their significance. I'm looking for any unique patterns. I'm noting potential start and stop codons. I'm currently scrutinizing the sequences to see if there are any specific phage-like features within.

Analysis: The sequence contains characteristic bacteriophage genomic features, including *high gene density*, *modular organization*, and specific regulatory elements like rho-independent terminators and phage-like promoter regions. Additionally, the presence of genes encoding structural proteins and DNA replication machinery typical of the Siphoviridae or Myoviridae families confirms its origin.

Answer: The correct answer is (B)

[Correct] ✅

Ground Truth Infomation:

Accession Code: IMGVR_UViG_3300045988_013037|3300045988|Ga0495776_004326

Contig GC Content: 53.79%

Contig Length: 32229 bp

Sequence Source: IMG_VR

Label: (B)

Figure 13: Case study of successful Phage Contig Identification.

Contamination Detection

Task ID: task2-96

Instruction: This sequence is a phage-associated contig. Does it contain any DNA originating from a non-phage host genome?

Sequence: ACGTCGACAGCGTCACCCTGGGGCAGAACGGCGGCGAGCTGATGCTCAACCTCGCAGGCCTG...

Choices: (A) No, the sequence does not contain host contamination. (B) Yes, the sequence contains host contamination.

Gemini-3-flash Output:

Scanning for Contamination
I'm currently focused on the possibility of host DNA contamination within the phage-associated contig. Given the contig's length, the initial step involves scanning for features common to phages, like structural proteins associated with the capsid, tail, and baseplate. I'll need to develop methods of characterizing these.

Detecting Contamination Markers
I've moved on to identifying potential host features within the contig. I'm looking for metabolic genes like *rRNA*, *tRNA*, and housekeeping genes, as well as gene clusters related to primary metabolism. I'm also searching for CRISPR arrays and prophage integration sites to further classify any non-phage content.

Assessing Gene Predictions
I'm now diving into the sequence itself, scrutinizing the gene predictions and looking for those mentioned previously: structural proteins, lysis genes, replication and regulatory components. I'm also starting the hunt for potential host genes, focusing on those involved in metabolism. So far, the contig shows promise for a host of phage-associated genes.

Identifying Host DNA
I'm now deep-diving into the sequence to find any hints of host contamination. I'm focusing on features common to host genomes, hoping for a match. I'm prioritizing the scanning of housekeeping genes, metabolic pathways, and prophage integration sites, looking for any features not typical of phages.

Assessing Sequence Data
I'm now examining the full sequence, and confirming the presence of previously-mentioned genes. I am now trying to find matches with potential phage-related genes, such as those that encode for lysis proteins and structural proteins. I'm also now seeking matches for host genes in order to assess potential contamination.

Evaluating DNA Sequence
I'm now deep into the sequence data, with an initial scan complete. I am currently cross-referencing my results with the initial gene predictions. So far, the data supports the presence of potential phage and host genes. Now, I'm focusing on those markers that might indicate contamination.

Analysis: The sequence contains numerous genes encoding characteristic phage structural proteins, such as tail fibers, capsid proteins, and baseplate components, as well as phage-specific replication and lysis machinery. No significant regions of non-phage host genomic DNA, such as primary metabolic pathways or housekeeping genes, were identified within the contig.

Answer: The correct answer is (A).

[Incorrect] ❌

Ground Truth Infomation:

Phage:

- **Accession Code:** NC_049437.1
- **Sequence Source:** NCBI RefSeq

Host:

- **Accession Code:** GCF_000006945.2
- **Sequence Source:** NCBI RefSeq

Contamination Ratio: 50%

Contig GC Content: 48.61%

Contig Length: 6875 bp

Label: (B)

Figure 14: Case study of failure in Contamination Detection.

Completeness Estimation

Task ID: task3-869

Instruction: What is the estimated genome completeness category of the following bacteriophage sequence?

Sequence: AGTGAAATACCTCCCTTTTGTGGATTAGTCTATTTGCCGACTTTTTGTGTTGGTGGTGAGTGTTGT...

Choices: (A) High-quality. (B) Complete. (C) Medium-quality. (D) Low-quality.

Gemini-3-flash Output:

Estimating Genome Completeness
I've been analyzing the nucleotide sequence, focusing on features indicative of genome completeness. Initial assessment suggests I need to evaluate the presence of essential genes and the absence of gaps or ambiguous regions within the sequence. My next step is to correlate these findings with a scoring system to assign a completeness category from the provided choices.

Examining Sequence Features
I'm now diving into the bacteriophage sequence itself. I've begun identifying potential direct terminal repeats and looking for any indications of circularity, since the metadata isn't available. Simultaneously, I'll be searching for essential genes like capsid and tail proteins to bolster the completeness assessment.

Inferring Completeness Now
I'm now immersed in the sequence, searching for DTRs to check for circularity, or some other end-to-end alignment. I'm also actively hunting for key genes. The goal is to develop a scoring system that can confidently place this phage genome into one of the categories.

Inferring Completeness Now
I've been analyzing the nucleotide sequence, focusing on features indicative of genome completeness. I'm actively looking for DTRs to check for circularity. Essential gene markers are on my list, but I haven't found any yet. I'll need a scoring system to sort this into one of the four categories.

Refining the Approach
I'm now prioritizing the sequence analysis. I'm focusing on key gene markers for genome completeness and scanning for indicators of circularization, to determine a more precise result. I can now start to think about the scoring system that I'll need to use to select the best option.

Scoring and Categorizing
I'm now formulating a scoring system to correlate observed features with the assigned categories. I'll need to weigh the presence of DTRs, essential genes, and the genome size. The absence of specific markers will factor into the assessment. I think I have a good idea of how I can sort this genome.

Analysis: The sequence contains a large number of characteristic bacteriophage genes (terminase, portal, capsid, tail proteins) and shows significant length and complexity. The presence of repetitive terminal sequences and a high density of essential phage functional modules suggests it is a complete genome.

Answer: The correct answer is (B).

[Incorrect] ❌

Ground Truth Infomation:

Accession Code: NC_027362.1

Sequence Source: NCBI RefSeq

Contig GC Content: 53.96%

Contig Length: 29394 bp

Label: (A)

Figure 15: Case study of structural hallucination in Completeness Estimation.

Lifestyle Classification

Task ID: task4-190

Instruction: What is the most likely lifestyle of the bacteriophage represented by the following genome sequence?

Sequence: ATGATGCAATTTAAAAATAAAGTGAAAATCTTAGGCGCTAAGGCTGTTGATTTTAAAACTGACGAT...

Choices: (A) Temperate (lysogenic-capable). (B) Virulent (lytic).

Gemini-3-flash Output:

Examining Phage Genomes
I'm focused on the phage genome data now, specifically trying to identify markers pointing to its lifestyle. I'm prioritizing genes indicative of lysogeny; integrase and excisionase are high on my list, alongside repressor proteins like cI. Finding *attP* sites is crucial for assessing potential for lysogenic behavior.

Identifying Lifestyle Markers
I've begun to scan the sequence, looking for telltale signs of the phage's preferred lifestyle. My search now includes key genes like those related to lysogeny, such as integrase or recombinase. Initial findings include a few promising motifs. I'm focusing on parsing the sequence and analyzing further.

Searching for Motifs
I'm now zeroing in on the sequence provided, specifically looking for common phage gene patterns. Initial keyword searches are underway, and I'm cross-referencing this data with known marker genes. The sequence contains a potential *attP* site, but I'm still analyzing its significance. Further scrutiny is needed to assess its importance.

Identifying Key Motifs
I'm now zeroing in on the sequence; it's revealing potential repressor elements like a zinc finger motif, suggesting a temperate lifestyle. Then I had a realization-- the sequence strongly resembles a small, single-stranded DNA phage, specifically something like PhiX174. That could significantly alter the analysis.

Re-examining PhiX174 Data
The initial sequence analysis suggested a possible temperate lifestyle. However, closer inspection of the sequence revealed significant homology to the *virulent* PhiX174. Scanning for key PhiX174 genes is underway. I've pinpointed segments of the sequence matching PhiX174's genome. This strongly points to a lytic lifestyle.

Confirming PhiX174 Identification
The sequence analysis now confirms that the phage genome strongly resembles that of PhiX174, a *virulent* phage. I've pinpointed characteristic regions of its genome and identified the Gene A sequence. This confirms the lytic nature of the phage, making answer (B) the correct choice.

Reaching a Conclusion
I have methodically reviewed the genome sequence and matched it with established markers. It's unequivocally the PhiX174 genome, a *virulent* phage. I'm now certain of my conclusion: the presence of key genes confirms its lytic lifestyle and makes answer (B) the correct choice.

Analysis: The provided sequence corresponds to the genome of bacteriophage PhiX174, a well-known member of the Microviridae family. This phage is strictly virulent, utilizing a lytic cycle for replication and lacking genes for integration (integrase) or lysogenic maintenance.

Answer: The correct answer is (B).

[Incorrect]

Ground Truth Infomation:

Accession Code: IMGVR_UViG_643348502_000001|643348502|643348581|2104978-2114268

Sequence Source: IMG_VR

Contig GC Content: 37.85%

Contig Length: 9291 bp

Label: (A)

Figure 16: Case study of failure in Lifestyle Classification.

Host Prediction

Task ID: task5-368

Instruction: Which host family is the following bacteriophage most likely to infect?

Sequence: ATATAGAACTGGGCGTCAGTCCGTACTTCCGGGCCAGGGCGTGATGGTTTCGTCCGTTGAACTC...

Choices: (A) Pseudomonadaceae. (B) Streptococcaceae. (C) Mycobacteriaceae. (D) Enterobacteriaceae.

Gemini-3-flash Output:

Examining Phage Sequences
I'm now focusing on the sequence length; it's significant. The approach will involve scrutinizing the sequence for telltale signs – specific genes, recognizable patterns, and possibly the GC content – that might point to the host. I'm prioritizing the identification of common phage signatures.

Searching for Key Markers
I'm now prioritizing the identification of common phage signatures. The sequence is quite long. My next step will be to scan for open reading frames (ORFs). I'm also planning to investigate the GC content to see if it provides any clues.

Scanning for Host Indicators
I'm now prioritizing the identification of common phage signatures. I'm focusing on key genes, motifs and GC content that might indicate the host. I'm currently scanning for open reading frames (ORFs) to identify potential protein-coding regions. I'll examine the GC content soon.

Assessing Sequence Patterns
Now I'm looking for patterns, or specific genes. I'm focusing on common phage genes like terminase and tail fibers to pinpoint the host family. I'm also going to begin looking for open reading frames (ORFs) to get a handle on potential protein-coding regions. Then I can examine GC content.

Pinpointing Host Targets
I'm now prioritizing the identification of common phage signatures. I'm focusing on common phage genes and also looking for open reading frames. The next step is to examine GC content. I also plan on scanning for specific sequence patterns to further refine my approach.

Assessing GC Content
I'm now focusing on specific patterns and GC content. After identifying some open reading frames (ORFs), I'm now examining the GC content of a segment of the sequence. I'm hoping it will provide information about the host organism.

Analysis: The nucleotide sequence exhibits a high GC content (approximately 60-65%), which is characteristic of the Pseudomonadaceae family. Additionally, the presence of specific genomic features and the overall organization are consistent with bacteriophages that infect Pseudomonas species.

Answer: The correct answer is (A).

[Correct] ✅

Ground Truth Infomation:

Accession Code: NC_074747.1

Sequence Source: NCBI RefSeq

Contig GC Content: 62.82%

Contig Length: 41303 bp

Label: (A)

Figure 17: Case study of successful Host Prediction.