

# GenomeQA: Benchmarking General Large Language Models for Genome Sequence Understanding

Weicai Long<sup>1,†</sup>, Yusen Hou<sup>1,†</sup>, Junning Feng<sup>1</sup>, Houcheng Su<sup>1</sup>,  
Shuo Yang<sup>2</sup>, Donglin Xie<sup>3</sup>, Yanlin Zhang<sup>1,\*</sup>,

<sup>1</sup>Hong Kong University of Science and Technology (Guangzhou),

<sup>2</sup>The University of Hong Kong, <sup>3</sup>Peking University

Correspondence\*: yanlinzhang@hkust-gz.edu.cn

## Abstract

Large Language Models (LLMs) are increasingly adopted as conversational assistants in genomics, where they are mainly used to reason over biological knowledge, annotations, and analysis outputs through natural language interfaces. However, existing benchmarks either focus on specialized DNA models trained for sequence prediction or evaluate biological knowledge using text-only questions, leaving the behavior of general-purpose LLMs when directly exposed to raw genome sequences underexplored. We introduce GenomeQA, a benchmark designed to provide a controlled evaluation setting for general-purpose LLMs on sequence-based genome inference tasks. GenomeQA comprises 5,200 samples drawn from multiple biological databases, with sequence lengths ranging from 6 to 1,000 base pairs (bp), spanning six task families: Enhancer and Promoter Identification, Splice Site Identification, Taxonomic Classification, Histone Mark Prediction, Transcription Factor Binding Site Prediction, and TF Motif Prediction. Across six frontier LLMs, we find that models often outperform random baselines, particularly on tasks driven by local sequence cues such as GC content and short motifs, while performance degrades on tasks that require more indirect or multi-step inference over sequence patterns. GenomeQA establishes a diagnostic benchmark for studying and improving the use of general-purpose LLMs on raw genomic sequences<sup>1</sup>.

## 1 Introduction

Genome analysis has long relied on specialized sequence models and task-specific pipelines. Recent years have seen rapid progress in DNA foundation models that are trained directly on nucleotide sequences, as well as emerging DNA-Text systems

that couple a dedicated DNA encoder with a large language model (Zhou et al., 2024; Dalla-Torre et al., 2025; Brixi et al., 2025; Schiff et al., 2024; de Almeida et al., 2025; Fallahpour et al., 2025). These approaches achieve strong supervised performance, but they typically require task-specific heads, probing, or additional adaptation to support downstream applications. In parallel, large language models have become widely used as conversational assistants across scientific domains, including chemistry (Hao et al., 2025; M. Bran et al., 2024), physics (Arora et al., 2023; Xu et al., 2025), and medicine (Singhal et al., 2023; Zheng et al., 2025). In genomics, their most common role is to reason over derived information such as annotations for genes and variants, functional summaries, experimental metadata, literature, and results from existing bioinformatics tools through natural language interfaces. However, direct interaction with raw genomic sequences poses challenges that differ fundamentally from standard natural language processing. DNA lacks human-interpretable semantic units such as words or grammar, exhibits long-range dependencies, and encodes biological signals in highly degenerate and context-dependent patterns (Cheng et al., 2025b; Benegas et al., 2025). As a result, it remains unclear how general-purpose LLMs behave when they are directly exposed to nucleotide sequences, and whether their responses reflect non-trivial sequence-level cues or are driven primarily by superficial heuristics.

Despite the increasing presence of LLMs in genomics studies, there is currently no standardized benchmark that evaluates how general-purpose language models perform when directly exposed to raw genomic sequences. Existing genome benchmarks (Marin et al., 2024; Grešová et al., 2023; Patel et al., 2024) primarily target trainable DNA-specific models and assess representation quality under fine-tuning or probing. In contrast, example benchmarks (Chen et al., 2025; Queen et al.,

<sup>†</sup>Co-first authors

<sup>1</sup>Data and code are available at <https://github.com/ai4nucleome/GenomeQA>

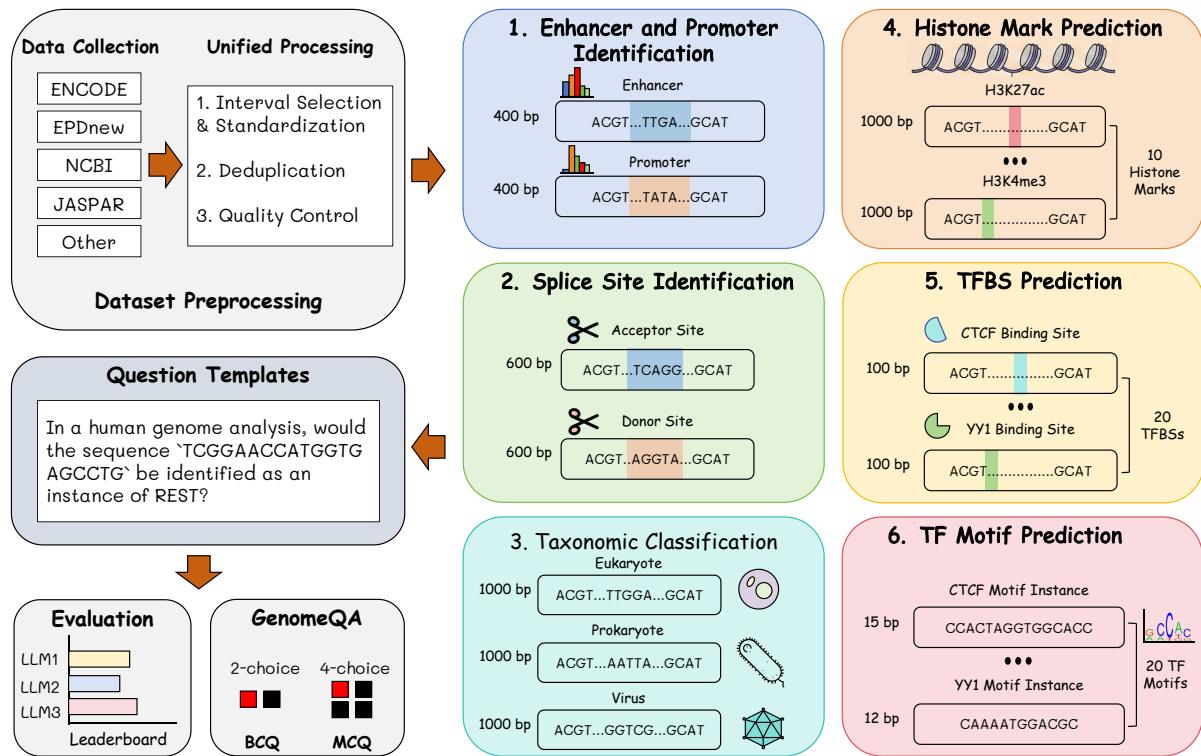


Figure 1: Overview of GenomeQA. The pipeline consists of Data collection, processing, construction and LLM evaluation.

2025; Yin et al., 2025) evaluate biological knowledge using text-only questions without requiring sequence-level analysis. Consequently, a practically important evaluation setting remains underexplored: a general-purpose LLM receiving natural language questions together with raw nucleotide sequences and producing answers under a fixed instruction-following protocol.

We take a step toward filling this gap by introducing GenomeQA, a question-answering benchmark designed to provide a controlled evaluation of general-purpose LLMs on sequence-based genome inference tasks. As shown in Figure 1, GenomeQA comprises 5,200 samples across six representative task families: Enhancer and Promoter Identification, Splice Site Identification, Taxonomic Classification, Histone Mark Prediction, Transcription Factor Binding Site (TFBS) Prediction, and Transcription Factor (TF) Motif Prediction. Each instance is formatted as either a Binary Choice Question (BCQ) for validity judgment or a four-option Multiple Choice Question (MCQ). To ensure consistent evaluation across models and tasks, we use a single system prompt derived from a small pilot study; the prompt is fixed throughout all experiments and provides domain-relevant guidance for

analyzing sequence signals such as motifs and base composition.

Our contributions are summarized as follows:

- We introduce GenomeQA, a benchmark that provides a controlled evaluation setting for assessing how general-purpose LLMs perform on sequence-based genome inference tasks. The benchmark comprises 5,200 curated samples spanning six biologically grounded task families.
- We conduct a comprehensive evaluation of six frontier LLMs, establishing baseline performance on raw DNA sequences and showing that current models can exploit certain local sequence signals (e.g., GC content and short motifs) but struggle with tasks requiring more complex or indirect inference.
- We present a fine-grained analysis of failure modes such as Sequence Motif Over-reliance and Character Fidelity Loss, highlighting systematic error patterns and directions for future research.

## 2 Related Work

### 2.1 Genome-Specific Models

Recent DNA foundation models treat genome sequences as language. They pretrain on billions of nucleotides and add task-specific heads for downstream applications such as regulatory elements prediction and splice site identification. Typical models include DNABERT-2 (Zhou et al., 2024), Nucleotide Transformer (Dalla-Torre et al., 2025), HyenaDNA (Nguyen et al., 2023), Genos (Lin et al., 2025), GENA-LM (Fishman et al., 2025) and Evo (Nguyen et al., 2024). Recent multimodal approaches couple a pretrained DNA foundation model with a general-purpose large language model (LLM), such as ChatNT (de Almeida et al., 2025), BioReason (Fallahpour et al., 2025) and Omni-DNA (Li et al., 2025). Empirically, these models achieve strong performance and can match or even surpass state-of-the-art results on standard genome understanding benchmarks. Meanwhile, some researchers argue that an alternative route is to repurpose general LLMs into DNA-LLMs directly (Cheng et al., 2025a), e.g., by adapting tokenization and training objectives so that the LLM can model DNA sequences without an explicit separate DNA encoder. Overall, despite their different interfaces, these methods share a key property: they rely on models that are specifically trained or substantially adapted on DNA sequences.

### 2.2 Genome Benchmarks

Existing benchmarks for genome modeling primarily target models that are specifically designed or adapted for genome data. Sequence-based benchmarks, such as BEND (Marin et al., 2024), DNA-LongBench (Cheng et al., 2025b), DART-Eval (Patel et al., 2024), and the Genomics Long-range Benchmark (Kao et al., 2024), are mainly constructed to evaluate genome foundation models. These benchmarks typically consist of raw DNA sequences paired with predefined prediction tasks, and are intended to measure a model’s ability to learn transferable sequence representations for downstream biological applications. Another line of benchmarks focuses on biomedical or clinical knowledge assessment through natural language questions, such as CMExam (Liu et al., 2023), Bio-Benchmark (Jiang et al., 2025), and EHRXQA (Bae et al., 2023). While these benchmarks evaluate language understanding and domain knowledge, they do not require models to directly process or

reason over raw genomic sequences. More recently, several benchmarks and evaluation protocols have emerged in the context of multimodal genome-language systems, including Lab-Bench (Laurent et al., 2024) and the task suites used in ChatNT and BioReason. These benchmarks partially involve sequence-based question answering, but they rely on a dedicated genome encoder to transform DNA sequences into latent representations before passing them to a large language model. As a result, the evaluated capability is that of an integrated multimodal system, rather than the intrinsic ability of a general-purpose LLM to interpret DNA sequences. In contrast, GenomeQA is designed to assess general-purpose LLMs in a setting where raw DNA sequences are provided directly as input. Each instance contains real genomic sequences with task-relevant signals, and all tasks are reformulated into a unified natural language question-answering format. GenomeQA does not assume a genome-specific encoder or additional training on DNA data, and is intended to support controlled measurement of LLMs performance on sequence-based genome inference tasks.

## 3 GenomeQA Construction

### 3.1 Design Principles

An overview of the dataset construction pipeline is shown in Figure 1. GenomeQA is built through a systematic process that transforms curated genome annotations into a unified evaluation framework. (1) We first identify six fundamental task families by collecting high-quality annotations from established databases and repositories, including ENCODE (Consortium, 2012), EPDnew (Dreos et al., 2013), NCBI (National Center for Biotechnology Information, 1988–2025), JASPAR (Rauluseviciute et al., 2023), as well as downstream tasks used in the Nucleotide Transformer (NT) benchmark (Dalla-Torre et al., 2025). A central design principle is the preservation of biological hierarchy: tasks are grouped to contrast genomic elements at the same functional level (e.g., enhancers versus promoters), rather than mixing signals across disparate biological scales. (2) All selected tasks are processed through a unified three-stage workflow to ensure consistency and data quality. First, we perform interval selection and length standardization to calibrate sequence windows, ensuring that each sequence sufficiently covers the motifs or regulatory signals required for the task. Sequence ex-

Task	Source	Seq. Len	Label	BCQ		MCQ	
				Counts	Avg. Len	Counts	Avg. Len
Enhancer and Promoter Identification	EPDnew,SCREEN	400	2	500	176	500	70
Splice Site Identification	NT downstream tasks	600	3	500	254	500	94
Taxonomic Classification	NCBI RefSeq	1000	3	500	402	500	146
Histone Mark Prediction	NT downstream tasks	1000	10	500	419	500	176
TFBS Prediction	ENCODE, JASPAR	100	20	500	61	500	31
TF Motif Prediction	ENCODE, JASPAR	6-20	20	100	23	100	16

Table 1: The statistics of GenomeQA suite, where Source indicates the data source, Seq. Len refs to the length of DNA sequence, and Label denotes the number of label sets. Avg. Len represents the average lengths of the questions. We report the number of tokens after tokenization using the Llama-4 tokenizer.

traction is conducted using Bedtools (Quinlan and Hall, 2010). Second, we remove overlapping samples in six tasks to reduce ambiguity and improve label reliability. Third, a quality control filter is applied to exclude sequences containing ambiguous nucleotide bases. (3) This process concludes with a question formulation stage, where each validated DNA sequence is instantiated into standardized natural language templates. Example question templates are provided in Appendix A.1. All tasks are presented in either Binary Choice Question (BCQ) or Multiple Choice Question (MCQ) formats. By constraining the answer space and standardizing the question structure, GenomeQA provides a controlled evaluation setting for assessing models’ ability to reason over raw DNA sequences.

### 3.2 Task Families

**Enhancer and Promoter Identification.** This task focuses on distinguishing cis-regulatory elements in the human genome, specifically promoters and enhancers. Promoter candidates are sourced from EPDnew(Dreos et al., 2013), covering regions from 299 base pairs (bp) upstream to 100 bp downstream of the transcription start site. Enhancer candidates are selected from the ENCODE SCREEN database(Consortium, 2012), with each region centered and resized to match the length of promoter sequences. To ensure that each 400-bp segment corresponds to a single regulatory element, we remove overlapping regions both within and across datasets. The label sets and label statistics are provided in Section A.2.

**Splice Site Identification.** This task focuses on human splice acceptor and donor sites. We aggregate positive 600 bp windows from the NT downstream tasks into a unified pool. After that, we relabel windows overlapping by more than 350 bp as containing both elements, while dis-

carding those with shorter overlaps. For questions about the presence of splice sites, we introduce composition-matched negatives by generating dinucleotide-preserving shuffled controls(Jiang et al., 2008) at construction period, forcing models to rely on higher-order motifs rather than simple base composition cues. The label sets and label statistics are provided in A.3.

**Taxonomic Classification.** This task evaluates whether LLMs can recover broad taxonomic groups from sequence alone. We sample 1 kbp fragments from NCBI RefSeq(O’Leary et al., 2015) assemblies representing eukaryote, prokaryote, and virus. Each fragment is assigned a high-level label with approximately balanced sampling across three groups. The complete species list and label statistics are provided in A.4.

**Histone Mark Prediction.** This task focuses on the identification of specific histone modifications from genome sequences in human K562 cells. We utilize 10 distinct histone marks from NT downstream tasks, retaining only positive 1 kbp windows. To ensure unique classification, we remove overlapping windows so that each region is associated with exactly one label. Beyond identifying individual marks, we further assess the model’s understanding of functional chromatin states by categorizing some marks as either open(e.g., H3K4me3, H3K27ac, H3K9ac) or repressive(e.g., H3K9me3, H3K27me3). This multi-dimensional annotation enables us to construct questions ranging from specific mark classification to broader chromatin accessibility that test whether the model perceives the underlying functional similarities between different histone modifications. The full histone mark list and label statistics are provided in A.5.

**TFBS Prediction.** We design this task to evaluate whether LLMs can accurately identify the specific transcription factor binding sites (TFBS)

present in a 100 bp genome window. We select 20 transcription factors (TF) with distinct motifs and source their ChIP-seq peaks from ENCODE. For each TF, we select peak intervals by placing a 100 bp window around the summit and then remove any sample pairs that overlap by more than 70 bp. To obtain precise labels, we scan the remaining sequences with FIMO(Grant et al., 2011) using JASPAR position weight matrices (PWM), recording all factors with motif instances in each window. This multi-label dataset enables questions regarding the presence of specific TFs. Additionally, we use CTCF as a canonical architectural protein that organizes chromatin loops and topologically associating domain (TAD) boundaries, constructing questions that test whether models understand its link to 3D genome architecture without naming the factor explicitly. The complete TF list and label statistics are provided in A.6.

**TF Motif Identification.** The final task focuses on short motif instances underlying the TFBS prediction task. Using the FIMO results described above, we collect all motif instances for the same 20 transcription factors and deduplicate them, yielding segments from 6 to 20 bp, each associated with a single TF label. Although the questions are linguistically simple, this task probes whether LLMs encode any recognizable representation of canonical transcription factor motifs. The label sets and label statistics are provided in A.7.

### 3.3 Dataset statistics

Table 1 provides a comprehensive overview of GenomeQA, detailing the biological focus, data sources, sequence lengths, label set sizes, volumes and average length of question instances for each task. Most tasks contribute 500 BCQs and 500 MCQs, while the simpler motif task includes 100 of each. Sequence lengths range from short motifs (6–20 bp) to medium regulatory windows (100–400 bp) and large genome contexts (1 kbp), requiring models to process both local patterns and broader organizational structures. Figure 2 displays the distribution of correct answers across option positions for all tasks. These distributions are nearly uniform, confirming that answer keys are balanced and free from position bias. Together, these metrics indicate that GenomeQA is structurally balanced, ensuring that evaluation results reflect genuine sequence understanding rather than dataset artifacts.

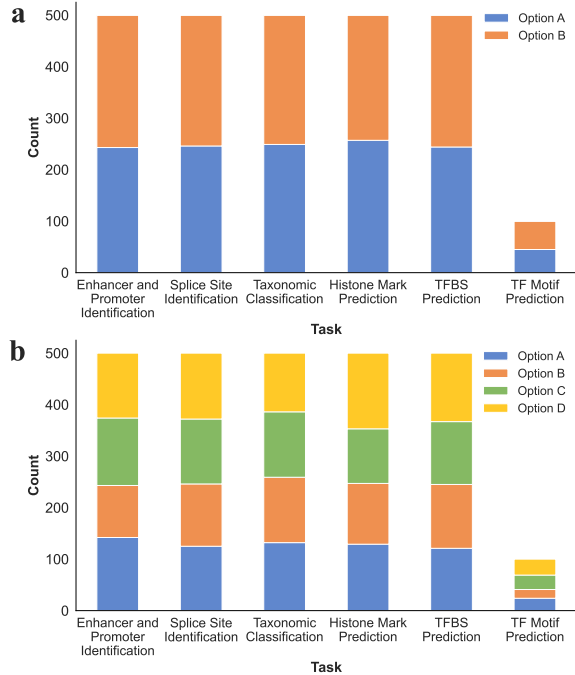


Figure 2: Distribution of Options in BCQ (a) and MCQ (b) of GenomeQA.

## 4 Experiments and Analysis

### 4.1 Experimental setup

We describe the baseline models, evaluated models, prompt configuration, and evaluation metrics below.

**Baseline Models.** We include three supervised baselines: a GC-content classifier, a 3-mer frequency linear classifier, and DNABERT-2 fine-tuned separately for each task family. These baselines provide both simple sequence-statistic references and a representative genome foundation model baseline beyond random guessing. To ensure fair comparison, all three baselines are trained on a shared supervised dataset aligned with the six GenomeQA task families using the same train/validation splits. We report these models as task-specific reference points, since they require supervised adaptation. Full dataset statistics and training details are provided in Appendix D.

**Evaluated Models.** We evaluate six state-of-the-art general large language models to assess their capabilities on genome data. The model set includes proprietary frontier models: Claude-Sonnet-4.5(Anthropic, 2025)<sup>2</sup>, GPT-5.1(OpenAI, 2025)<sup>3</sup>, Gemini-3-Pro(Google, 2025)<sup>4</sup>, Grok-

<sup>2</sup>Developed by Anthropic.

<sup>3</sup>Developed by OpenAI.

<sup>4</sup>Developed by Google. We use Gemini-3-Pro-Preview.

4.1(xAI, 2025)<sup>5</sup>, Llama-4(Meta, 2025)<sup>6</sup>, Qwen3-Max(Alibaba, 2025)<sup>7</sup>. We enable thinking mode whenever supported for all models to maximize their potential for complex biological deduction.

**Prompt Settings.** We use a single fixed system prompt across all tasks and models, with details provided in Appendix B. No task-specific few-shot examples are included. The prompt provides domain-relevant guidance and a standardized output format for analyzing sequence signals (e.g., motifs and base composition), ensuring consistency across models while minimizing variation due to prompt design.

**Metrics.** We report both classification accuracy and weighted-F1. Since the answer options are randomly permuted and correspond to raw DNA sequences rather than stable semantic labels, we evaluate predictions at the option-letter level. Full weighted-F1 results are provided in Appendix E.

**Statistical Uncertainty and Significance.** To assess whether the observed performance differences are statistically reliable, we report confidence intervals and significance tests for each task setting. Full results are provided in Appendix F.

## 4.2 Main Results

Table 2 and Figure 3 summarize the performance of the evaluated models on GenomeQA under BCQ and MCQ settings. We highlight three observations. **(1) Frontier LLMs outperform random baselines but show substantial performance variation across tasks.** Among the evaluated models, Gemini-3-Pro achieves the highest average accuracy (66.27% on BCQ and 60.87% on MCQ), with Claude-Sonnet-4.5, GPT-5.1, and Grok-4.1 forming a closely clustered second tier. In contrast, Llama-4 and Qwen3-Max exhibit lower overall accuracy (e.g., Qwen3-Max achieves 56.87% on BCQ and 41.38% on MCQ). These results indicate that, even for the strongest models, performance on GenomeQA remains weak and inconsistent across tasks. **(2) Model performance correlates with task complexity and the depth of reasoning required for each task.** As shown in Figure 3, Large language models achieve respectable accuracy on Enhancer and Promoter Identification, Taxonomic Classification, and TF Motif Prediction as these questions are formatted as direct pattern

<sup>5</sup>Developed by xAI. We use Grok-4.1-Fast.

<sup>6</sup>Developed by Meta. We use Llama-4-Maverick-17B-128E-Instruct.

<sup>7</sup>Developed by Alibaba. We use Qwen3-Max-Preview.

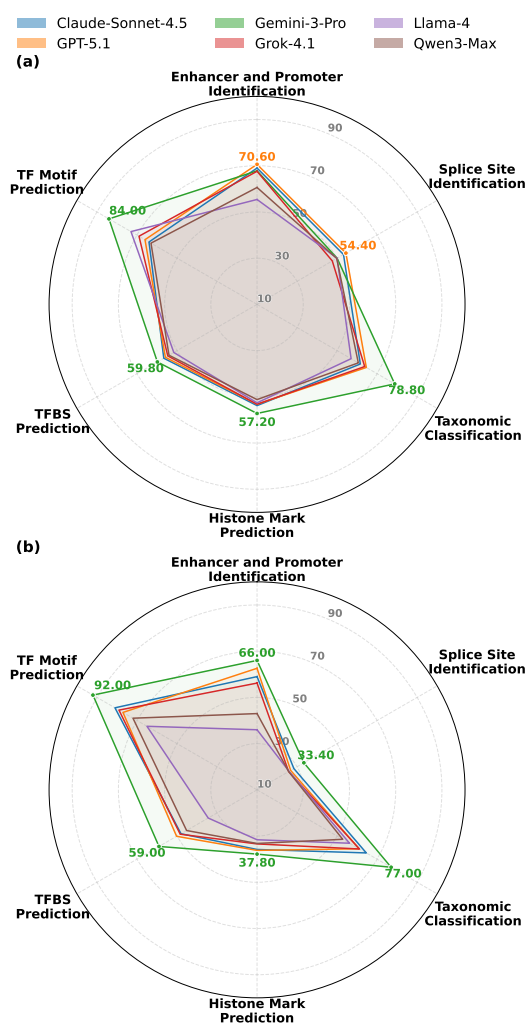


Figure 3: Performance comparison of frontier LLMs across the six GenomeQA tasks under (a) BCQ and (b) MCQ settings. Scores are reported as task-wise accuracy (%). Numerical annotations indicate the highest accuracy achieved for each task.

recognition. In contrast, performance drops significantly on Splice Site Identification, Histone Mark Prediction, and TFBS Prediction. These difficult tasks involve more complex genome signals such as the long-range patterns of histone marks. Furthermore, the framework incorporates indirect reasoning to increase the complexity of these specific tasks. The consistently low accuracy across these categories proves that current LLMs struggle to execute indirect reasoning when they encounter intricate genome data. We also evaluate balanced and oversampled subsets separately and the main model rank and qualitative conclusions remain unchanged. Detailed results are reported in Appendix G. **(3) Multiple choice formats enhance relative discrimination.** The random baselines are 50.00% for BCQ and 25.00% for MCQ, with empirical

Model	Thinking	Enhancer and Promoter Identification	Splice Site Identification	Taxonomic Classification	Histone Mark Prediction	TFBS Prediction	TF Motif Prediction	Avg.
<b>BCQ</b>								
Claude-Sonnet-4.5	Y	69.00	53.20	61.60	53.80	56.60	64.00	59.70
GPT-5.1	Y	70.60	54.40	64.60	53.20	55.80	66.00	60.77
Gemini-3-Pro	Y	67.80	50.00	78.80	<u>57.20</u>	<u>59.80</u>	<b>84.00</b>	66.27
Grok-4.1	Y	67.60	47.60	63.80	53.20	54.60	69.00	59.30
Llama-4	N	55.40	50.00	57.00	52.60	51.60	<u>73.00</u>	56.60
Qwen3-Max	Y	60.60	49.80	60.60	51.20	54.00	63.00	56.53
GC-classifier	N	74.60	<u>56.00</u>	70.00	54.80	52.60	51.00	59.83
3mer-classifier	N	<u>81.80</u>	55.00	<u>94.80</u>	56.60	58.60	70.00	<u>69.47</u>
DNABERT2	N	<b>91.20</b>	<b>71.80</b>	<b>99.20</b>	<b>65.60</b>	<b>60.60</b>	67.00	<b>75.90</b>
Random	N			50.00				
<b>MCQ</b>								
Claude-Sonnet-4.5	Y	59.00	28.40	64.60	35.80	48.00	<u>81.00</u>	52.80
GPT-5.1	Y	62.60	26.80	61.00	36.20	50.20	77.00	52.30
Gemini-3-Pro	Y	66.00	<u>33.40</u>	77.00	<u>37.80</u>	<u>59.00</u>	<b>92.00</b>	<u>60.87</u>
Grok-4.1	Y	56.20	25.80	61.20	33.40	48.40	79.00	50.67
Llama-4	N	36.00	26.00	56.20	31.60	34.40	65.00	41.53
Qwen3-Max	Y	43.00	25.80	52.80	33.20	45.20	72.00	45.33
GC-classifier	N	57.00	25.20	54.40	35.40	41.60	37.00	41.77
3mer-classifier	N	<u>69.40</u>	28.20	<u>91.00</u>	35.60	51.20	57.00	55.40
DNABERT2	N	<b>82.80</b>	<b>52.80</b>	<b>98.40</b>	<b>48.20</b>	<b>66.60</b>	54.00	<b>67.13</b>
Random	N			25.00				

Table 2: Overall results of LLMs on GenomeQA. Thinking denotes whether the model utilizes Chain-of-Thought reasoning. The table reports the classification accuracy (%) for each subtask and Avg. denotes the average accuracy. The best and second-best performances in each task are highlighted in bold and underlined, respectively.

validation provided in Appendix C. Although absolute accuracy is naturally lower in the multiple choice setting, the relative improvement over the baseline is substantially higher. This trend occurs as the format shifts the task from absolute verification to comparative ranking. Unlike isolated binary decisions, the provided options serve as contextual anchors that narrow the search space, enabling models to evaluate relative likelihoods among candidates. Consequently, the performance gain over chance is approximately two-fold higher than in the binary setting. This indicates that the comparative structure effectively leverages the probabilistic ranking capabilities of models to reduce classification noise more robustly than direct verification.

### 4.3 Impact of Thinking Process

To examine the benefits of the explicit reasoning process, we evaluate selected LLMs equipped with the thinking mode. The goal is to assess their ability to perform step-by-step deductions and improve performance on complex genome tasks. We compare the performance of GPT-5.1 and Qwen3-Max by enabling and disabling their thinking features across both binary and multiple-choice settings. As shown in Table 3, the integration of the thinking process leads to consistent performance gains.

model	Thinking	BCQ	MCQ
GPT-5.1	✓	60.77	52.30
	✗	58.03	43.97
Qwen3-Max	✓	56.53	45.33
	✗	54.57	40.80

Table 3: Performance comparison of LLMs with and without the thinking process on GenomeQA. The table reports the accuracy percentages on BCQ and MCQ.

Specifically, GPT-5.1 achieves a significant improvement in the multiple-choice setting, where the accuracy increases from 43.97% to 52.30%. This represents a notable enhancement in its ability to filter out distractors. Qwen3-Max also exhibits improvements, although the margins are smaller compared to GPT-5.1. For instance, its binary classification accuracy rises from 54.57% to 56.53%. The disparity in gains between the two models suggests that the effectiveness of the thinking mode depends heavily on the underlying domain knowledge of the base model. These results underscore the importance of enabling thinking capabilities to handle the intricate reasoning required for genome sequence analysis.

#### 4.4 Impact of Implicit Target Inference

We design a controlled comparison to evaluate model performance on questions that require an additional inference step beyond direct sequence recognition. We focus on CTCF-related instances in the TFBS Prediction task. CTCF is a canonical transcription factor whose binding sites are strongly associated with higher-order chromatin organization, including chromatin loops and topologically associating domains (TADs). As a result, questions about 3D genome structure can implicitly point to CTCF, which must then be linked back to sequence-level evidence. Specifically, we construct two question variants over the same underlying sequence set. In the *direct* setting, the question explicitly names the target (CTCF) and asks whether the sequence contains a CTCF binding site, which can be answered by direct pattern recognition. In the *indirect* setting, the question does not mention CTCF and instead asks whether the sequence is associated with the formation of chromatin loops or TAD boundaries. Answering this variant requires a multi-step mapping: (i) infer the relevant regulatory factor implied by the functional description (CTCF), and (ii) evaluate whether the input sequence contains sequence patterns consistent with that factor. As shown in Table 4, making the target explicit substantially improves accuracy across models. For example, in the multiple-choice setting, Claude-Sonnet-4.5 and GPT-5.1 increase from 27.11% to 63.86%, and Gemini-3-Pro increases from 44.58% to 67.47%. In contrast, performance in the inference setting is often close to the random baseline. These results suggest that the additional target-inference step is a major source of difficulty in this setting.

#### 4.5 Failure Case Study

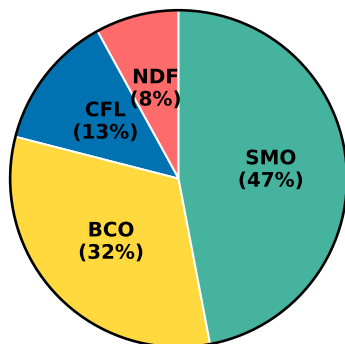


Figure 4: Distribution of failed cases.

To gain a deeper understanding of the limitations

Model	BCQ		MCQ	
	w/	w/o	w/	w/o
Claude-Sonnet-4.5	68.07	54.22	63.86	27.11
GPT-5.1	68.07	52.41	63.86	27.11
Gemini-3-Pro	74.70	60.24	67.47	44.58
Grok-4.1	61.45	54.82	58.43	27.11
Llama-4	58.43	50.00	37.95	24.10
Qwen3-Max	67.47	45.78	67.45	24.10

Table 4: Impact of implicit target inference on TFBS prediction task. The table reports accuracy percentages where w/ denotes the target (CTCF) is explicitly named, allowing direct pattern recognition and w/o denotes the implied target setting.

of current LLMs in genome analysis, we analyze 200 error samples produced by Gemini-3-Pro. As shown in Figure 4, our qualitative analysis categorizes these failures into four distinct types, revealing the cognitive disconnect between general LLM and the rigorous requirements of genomics. Details are provided in Appendix I.

##### Sequence Motif Over-reliance (SMO, 47%).

Failures occur when models rely on general sequence elements while neglecting specific details. For instance, in Histone Mark Prediction, the model incorrectly classifies an open *Alu* repeat as closed (Su et al., 2014). It simply applies the general rule that transposable elements are repressed and overlooks the high GC content of this specific element.

##### Base Composition Over-reliance (BCO, 32%).

Failures occur when models rely on statistical summaries while ignoring structural patterns. For instance, in Taxonomic Classification, the model incorrectly identifies a virus as a prokaryote. It uses the high GC content as a shortcut and ignores the specific gene organization that actually points to a virus.

##### Character Fidelity Loss (CFL, 13%).

Models frequently lose character-level fidelity in long sequences, leading to the fabrication of non-existent sub-sequences to support their claims. In Enhancer and Promoter Identification, the model hallucinates a specific motif sequence as evidence that does not actually exist within the input sequence.

##### Noise Distinction Failure (NDF, 8%).

Failures occur when models fail to recognize meaningless patterns in shuffled negative samples. For instance, in Splice Site Identification, the model analyzes a randomized control sequence. It fails to detect the

random nature of the input and performs a pseudo-analysis to incorrectly classify it as a splice site.

## 5 Conclusion

We introduce GenomeQA, a benchmark designed to support systematic evaluation of general-purpose large language models on sequence-based genome inference tasks. By curating biologically grounded tasks across multiple genomic contexts and adopting standardized question formats, GenomeQA provides a controlled evaluation setting for analyzing model performance on raw DNA sequences. Our experimental results show that current LLMs can leverage certain local sequence signals but exhibit substantial performance variation across tasks, particularly when targets are implicit or require additional inference steps. These findings highlight both the current limitations of general-purpose LLMs in this setting and the need for more reliable evaluation tools. We hope GenomeQA will serve as a useful diagnostic resource for future research on genome-aware language modeling and the integration of LLMs with genomic data.

## Limitations

We acknowledge three primary limitations of this work. **(1) Benchmark scale.** Evaluating multiple frontier LLMs under thinking-enabled settings incurs substantial computational cost. For example, Claude-Sonnet requires on the order of minutes to process a single sample, and Qwen3-Max can take even longer. To balance inference cost with task coverage, we curate a high-quality but moderately sized dataset. As a result, while GenomeQA is suitable for systematic evaluation, its current scale is not intended for full-parameter fine-tuning of large models. **(2) Task scope.** GenomeQA focuses on foundational sequence-based tasks such as motif recognition and coarse taxonomic classification. More complex biological problems, including variant effect prediction, and gene expression modeling, are not covered. These tasks typically require much longer sequences that exceed current context window limits and depend on additional modalities such as chromatin accessibility, histone modifications, and three-dimensional genome structure. Extending GenomeQA to incorporate such multi-omics signals is an important direction for future work. **(3) Confidential Intervals.** A further limitation is that our reported uncertainty is based on single-run evaluations. While Wilson intervals and

exact binomial tests quantify finite-sample uncertainty over benchmark questions, they do not capture cross-run variation arising from API stochasticity or backend nondeterminism. Repeated-run evaluation is an important direction for future work.

## Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 32500550), the Ministry of Human Resources and Social Security of the People's Republic of China (No. Y20250128) and a Guangdong Provincial Project (2024QN11N085).

## Ethical considerations

Our GenomeQA benchmark is built upon publicly available genome databases and established biological datasets, which are open for academic and research purposes. Additionally, we have rigorously reviewed the benchmark to ensure strict adherence to ethical guidelines. The dataset consists of de-identified public reference sequences and contains no personally identifiable information or private human genetic data. We have also verified that the benchmark focuses solely on fundamental biological understanding and contains no content related to biosecurity risks, harmful pathogens, or inappropriate material.

## References

- Alibaba. 2025. [Qwen3-max: Just scale it](#). Accessed: 2025-09-24.
- Anthropic. 2025. [Introducing claude sonnet 4.5](#). Accessed: 2025-09-29.
- Daman Arora, Himanshu Singh, and 1 others. 2023. Have llms advanced enough? a challenging problem solving benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7527–7543.
- Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, and 1 others. 2023. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Advances in Neural Information Processing Systems*, 36:3867–3880.
- Gonzalo Benegas, Carlos Albors, Alan J Aw, Chengzhong Ye, and Yun S Song. 2025. A dna language model based on multispecies alignment predicts the effects of genome-wide variants. *Nature Biotechnology*, pages 1–6.

- 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.
- Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, and 1 others. 2025. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature communications*, 16(1):3280.
- Wenduo Cheng, Junhong Shen, Mikhail Khodak, Jian Ma, and Ameet Talwalkar. 2025a. **L2g: Repurposing language models for genomics tasks**. *Transactions on Machine Learning Research*.
- Wenduo Cheng, Zhenqiao Song, Yang Zhang, Shike Wang, Danqing Wang, Muyu Yang, Lei Li, and Jian Ma. 2025b. Dnalongbench: a benchmark suite for long-range dna prediction tasks. *Nature Communications*, 16(1):10108.
- The ENCODE Project Consortium. 2012. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74.
- 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.
- Bernardo P de Almeida, Guillaume Richard, Hugo Dalla-Torre, Christopher Blum, Lorenz Hexemer, Priyanka Pandey, Stefan Laurent, Chandana Rajesh, Marie Lopez, Alexandre Laterre, and 1 others. 2025. A multimodal conversational agent for dna, rna and protein tasks. *Nature Machine Intelligence*, pages 1–14.
- René Dreos, Giovanna Ambrosini, Rouayda Cavin Périer, and Philipp Bucher. 2013. Epd and epdnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic acids research*, 41(D1):D157–D164.
- Adibvafa Fallahpour, Andrew Magnuson, Purav Gupta, Shihao Ma, Jack Naimer, Arnav Shah, Haonan Duan, Omar Ibrahim, Hani Goodarzi, Chris J. Maddison, and BO WANG. 2025. Bioreason: Incentivizing multimodal biological reasoning within a DNA-LLM model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Veniamin Fishman, Yuri Kuratov, Aleksei Shmelev, Maxim Petrov, Dmitry Penzar, Denis Shepelin, Nikolay Chekanov, Olga Kardymon, and Mikhail Burtsev. 2025. Gena-lm: a family of open-source foundational dna language models for long sequences. *Nucleic Acids Research*, 53(2):gkae1310.
- Google. 2025. **Gemini 3 pro: Best for complex tasks and bringing creative concepts to life**. Accessed: 2025-11-18.
- Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. 2011. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018.
- 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.
- Li Hao, He CAO, Bin Feng, Daniel Shao, Xiangru Tang, Zhiyuan Yan, Yonghong Tian, Li Yuan, and Yu Li. 2025. Beyond chemical QA: Evaluating LLM’s chemical reasoning with modular chemical operations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Jiyue Jiang, Pengan Chen, Jiuming Wang, Dongchen He, Ziqin Wei, Liang Hong, Licheng Zong, Sheng Wang, Qinze Yu, Zixian Ma, and 1 others. 2025. Benchmarking large language models on multiple tasks in bioinformatics nlp with prompting. *arXiv preprint arXiv:2503.04013*.
- Minghui Jiang, James Anderson, Joel Gillespie, and Martin Mayne. 2008. ushuffle: A useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*, 9(1):192.
- Chia Hsiang Kao, Evan Trop, McKinley Polen, Yair Schiff, Bernardo P de Almeida, Aaron Gokaslan, Thomas PIERROT, and Volodymyr Kuleshov. 2024. Advancing dna language models: The genomics long-range benchmark. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*.
- Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Sidharth Narayanan, Manvitha Ponnampati, Andrew D White, and Samuel G Rodrigues. 2024. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*.
- Zehui Li, Vallijah Subasri, Yifei Shen, Dongsheng Li, Wentao Gu, Guy-Bart Stan, Yiren Zhao, and Caihua Shan. 2025. Omni-DNA: A genomic model supporting sequence understanding, long-context, and textual annotation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Adi Lin, Bin Xie, Cheng Ye, Cheng Wang, Duoyuan Chen, Ercheng Wang, Fanfeng Lu, Guirong Xue, Haiqiang Zhang, Jiajie Zhan, Jianfeng Zhang, Jiangshuan Pang, Jianqiang Liang, Jiawei Lin, Jiabin Ma, Jie Hu, Jing Ma, Jinni Dong, Jiongzheng Li, and 43 others. 2025. Genos: A human-centric genomic foundation model. *GigaScience*, page giae132.
- Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu

- You, Zhenhua Guo, LEI ZHU, and Michael Lingzhi Li. 2023. Benchmarking large language models on cmexam - a comprehensive chinese medical exam dataset. In *Advances in Neural Information Processing Systems*, volume 36, pages 52430–52452. Curran Associates, Inc.
- Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535.
- Frederikke Isa Marin, Felix Teufel, Marc Horlacher, Dennis Madsen, Dennis Pultz, Ole Winther, and Wouter Boomsma. 2024. BEND: Benchmarking DNA language models on biologically meaningful tasks. In *The Twelfth International Conference on Learning Representations*.
- Meta. 2025. [The llama 4 herd: The beginning of a new era of natively multimodal ai innovation](#). Accessed: 2025-04-05.
- National Center for Biotechnology Information. 1988–2025. National Center for Biotechnology Information (NCBI). <https://www.ncbi.nlm.nih.gov/>.
- Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel H. King, Garyk Brixi, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian L. Hie. 2024. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336.
- Eric Nguyen, Michael Poli, Marjan Faizi, Armin W Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton M. Rabideau, Yoshua Bengio, Stefano Ermon, Christopher Re, and Stephen Baccus. 2023. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Nuala A. O’Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufu, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, and 36 others. 2015. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745.
- OpenAI. 2025. [Gpt-5.1: A smarter, more conversational chatgpt](#). Accessed: 2025-11-12.
- Aman Patel, Arpita Singhal, Austin Wang, Anusri Pampari, Maya Kasowski, and Anshul Kundaje. 2024. DART-eval: A comprehensive DNA language model evaluation benchmark on regulatory DNA. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Owen Queen, Harrison G Zhang, and James Zou. 2025. CGBench: Benchmarking language model scientific reasoning for clinical genetics research. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Aaron R. Quinlan and Ira M. Hall. 2010. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.
- Ieva Rauluseviciute, Rafael Riudavets-Puig, Romain Blanc-Mathieu, Jaime A Castro-Mondragon, Katalin Ferenc, Vipin Kumar, Roza Berhanu Lemma, Jérémy Lucas, Jeanne Chèneby, Damir Baranasic, Aziz Khan, Oriol Fornes, Sveinung Gundersen, Morten Johansen, Eivind Hovig, Boris Lenhard, Albin Sandelin, Wyeth W Wasserman, François Parcy, and Anthony Mathelier. 2023. Jaspar 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 52(D1):D174–D182.
- Yair Schiff, Chia Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. 2024. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. In *Forty-first International Conference on Machine Learning*.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Ming Su, Dali Han, Jerome Boyd-Kirkup, Xiaoming Yu, and Jing-Dong J Han. 2014. Evolution of alu elements toward enhancers. *Cell reports*, 7(2):376–385.
- xAI. 2025. [Grok 4.1 fast and agent tools api](#). Accessed: 2025-11-19.
- Yinggan Xu, Hana Kimlee, Yijia Xiao, and Di Luo. 2025. Advancing ai-scientist understanding: Making llm think like a physicist with interpretable reasoning. *arXiv preprint arXiv:2504.01911*.
- Ming Yin, Yuanhao Qu, Dyllan Liu, Ling Yang, Le Cong, and Mengdi Wang. 2025. Genome-bench: A scientific reasoning benchmark from real-world expert discussions. *bioRxiv*, pages 2025–06.
- Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh T. N. Nguyen, Lauren T. May, Geoffrey I. Webb, and Shirui Pan. 2025. Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence*, 7(3):437–447.
- Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana V Davuluri, and Han Liu. 2024. DNABERT-2: Efficient foundation model and benchmark for multi-species genomes. In *The Twelfth International Conference on Learning Representations*.

## A Details for GenomeQA construction

### A.1 Question Templates

As described in Section 3.1, we formulate two distinct question types: Binary Choice Questions (BCQ) and Multiple Choice Questions (MCQ). For each type, we curate a set of templates that were generally applied consistently across most tasks. While the template structure remains largely consistent across most tasks, we introduce specialized templates for the Splice Site Prediction, Histone Mark Prediction, and Transcription Factor Binding Site (TFBS) Prediction tasks to slightly increase the complexity of the tasks. Comprehensive examples of these templates are presented in Table 5.

### A.2 Enhancer and Promoter Identification

#### A.2.1 Label Sets and Entity List

For the Enhancer and Promoter Identification task, the label set is comprised of enhancer and promoter.

#### A.2.2 Label Statistics

In this task, each sequence corresponds to a single, mutually exclusive label. We analyze the distribution of question types to verify dataset balance, as described in Figure 5(a) and Figure 5(b). In the Binary Choice Questions, the distribution is nearly symmetrical with 257 questions targeting enhancers and 243 targeting promoters. Similarly, the Multiple Choice Questions maintain this balanced approach featuring 244 enhancer-related queries and 256 promoter-related queries. This uniform distribution ensures that the model is evaluated equally on its ability to identify both regulatory elements.

### A.3 Splice Site Identification

#### A.3.1 Label Sets and Entity List

For the Splice Site Identification task, the label set is comprised of acceptor, donor and both patterns.

#### A.3.2 Label Statistics

In this task, each sequence corresponds to a single splice site category. In the Binary Choice Questions, the distribution remains relatively comparable across the three classes, with 176 questions targeting acceptors, 188 targeting donors, and 136 focusing on dual-site sequences. This balance extends to the Multiple Choice Questions, which feature an almost perfect split: 172 acceptor, 163

donor, and 165 both-related queries. Such a distribution guarantees that the performance metrics reflect a holistic understanding of all splice site configurations rather than a bias toward a specific type. The details are provided in Figure 5(c) and Figure 5(d).

### A.4 Taxonomic Classification

#### A.4.1 Label Sets and Entity List

In this task, the label set encompasses Eukaryote, Prokaryote, and Virus categories. We source eukaryotic sequences from *Homo sapiens*, *Mus musculus*, and *Sus scrofa*. For Prokaryotes, we select three common bacterial species. To address the shorter length of viral DNA, we utilize 9 viral species to ensure a sufficient number of candidate samples. Table 6 provides detailed specifications.

#### A.4.2 Label Statistics

In this task, each sequence corresponds to a single taxonomic category, classifying inputs as Eukaryote, Prokaryote, or Virus. We analyze the distribution of question types to verify dataset balance, as described in Figure 5(e) and Figure 5(f). In the Binary Choice Questions, the dataset exhibits a well-proportioned composition across the three domains, with 170 questions targeting eukaryotes, 150 targeting prokaryotes, and 180 focusing on viral sequences. This balanced structure is mirrored in the Multiple Choice Questions, which feature a comparable distribution: 177 eukaryote, 165 prokaryote, and 158 virus-related queries. Such uniform coverage ensures that the model is rigorously evaluated on its ability to distinguish genome signatures across diverse biological domains without bias.

### A.5 Histone Mark Prediction

#### A.5.1 Label Sets and Entity List

We utilize all 10 histone marks from the downstream Nucleotide Transformer dataset for the recognition task. Building on this, we select five histone marks to construct the chromatin accessibility task, consisting of three marks from open regions and two from closed regions. The "Undefined" represents the normal usage. Table 7 contains detailed information.

#### A.5.2 Label Statistics

In this task, each sequence corresponds to a single histone modification mark. The detailed information are provided in Figure 5(g) and Figure

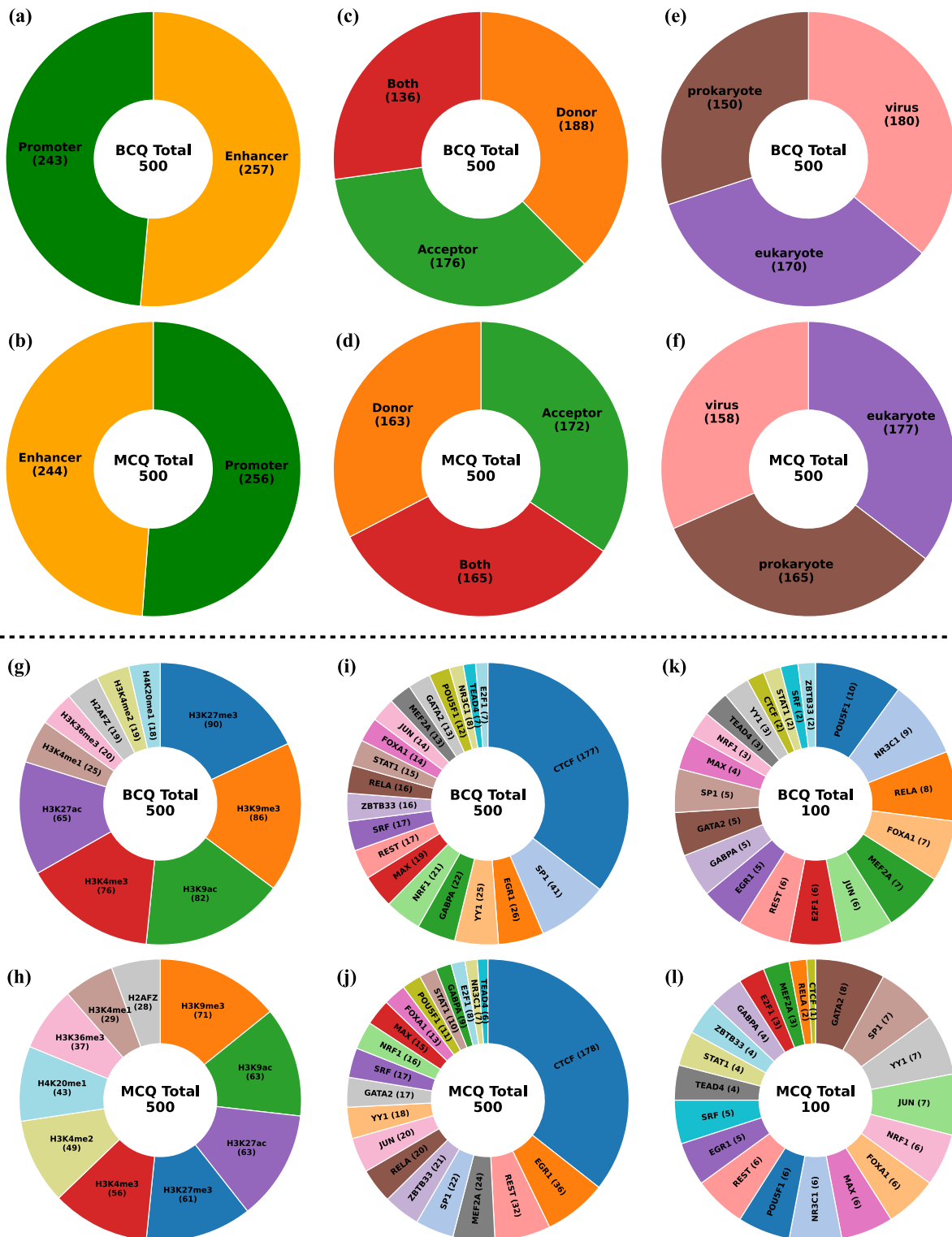


Figure 5: Label statistics for all tasks. The charts display the distribution of question types for BCQ and MCQ. (a)-(b): Enhancer and Promoter Identification. (c)-(d): Splice Site Identification. (e)-(f): Taxonomic Classification. (g)-(h): Histone Mark Prediction. (i)-(j): TFBS Prediction. (k)-(l): TF Motif Prediction.

Type	Examples	Note
BCQ	Does the Human DNA sequence {seq} contain any {target}?	{seq} represents a sequence, {target} is the query label
	Does this Human DNA sequence contain authentic {target}? Sequence: {seq}	Specifically used in Splice-Site Prediction
	Consider the chromatin state in K562 cells. Is this Human DNA sequence located in an {state} region? Sequence: {seq}	Specifically used in Histone Mark Prediction
MCQ	We are looking for a sequence bound by the master regulator of chromatin looping and insulation. Is this sequence a match? {seq}	Specifically used in TFBS prediction
	Which of the following Human DNA sequences contains a {target}?	{target} is the query label
	There is an exception in this group. While three sequences contain {target1}, one actually harbors a {target2}. Can you find it?	{target1} and {target2} represent query labels
	Determine the chromatin accessibility state of this K562 sequence: {seq}	Specifically used in Histone Mark Prediction
	Which of the following Human DNA sequence is associated with the formation of chromatin loops and TAD boundaries?	Specifically used in TFBS prediction

Table 5: Question Templates in GenomeQA.

Taxonomy	Species	NCBI Accession
Eukaryote	Homo sapiens	GCF_000001405.40
	Mus musculus	GCF_000001635.27
	Sus scrofa	GCF_000003025.6
Prokaryote	Escherichia coli	GCF_000005845.2
	Staphylococcus aureus	GCF_000013425.1
	Bacillus subtilis	GCF_000009045.1
Virus	Bacteriophage phiX174	NC_001422.1
	Simian virus 40	NC_001669.1
	Parvovirus B19	NC_000883.2
	Human papillomavirus type 16	NC_001526.4
	Human adenovirus 5	NC_001405.1
	Enterobacteria phage lambda	NC_001416.1
	Human alphaherpesvirus 1	NC_001806.2
	Pandoravirus salinus	NC_022098.1
	Megavirus chilensis	NC_016072.1
	Acanthamoeba polyphaga mimivirus	NC_014649.1

Table 6: Species in Taxonomic Classification task.

Chrom. State	Selected Histone Marks
Open	H3K4me3, H3K9ac, H3K27ac
Close	H3K9me3, H3K27me3
Undefined	H2AFZ, H3K4me1
	H3K4me2, H3K36me3, H4K20me1

Table 7: Histone marks in the Histone Mark Prediction task. Chrom. State indicates the chromatin state.

3D Genome	Selected TFs
Related	CTCF
Unrelated	SP1, GABPA, E2F1, TEAD4, NRF1
Undefined	REST, EGR1, YY1, ZBTB33, FOXA1, JUN, MAX
	SRF, MEF2A, RELA, STAT1, NR3C1, POU5F1, GATA2

Table 8: TFs in TFBS Prediction.

5(h). Unlike the strictly uniform distributions in other tasks, the data here reveals an intentional skew in both Binary and Multiple Choice Questions. This deviation arises from the introduction of a Chromatin Accessibility reasoning dimension. To construct questions specifically focusing on open versus closed chromatin states, we prioritized a subset of five markers: three associated with open regions (H3K4me3, H3K27ac, H3K9ac) and two with closed regions (H3K9me3, H3K27me3). Consequently, questions targeting these five labels appear significantly more frequently than those targeting the remaining markers in the set, reflecting a design choice to stress-test the model’s understanding of chromatin structure.

## A.6 TFBS Prediction

### A.6.1 Label Sets and Entity List

As described in Section 3.2, we select 20 transcription factors with distinct motif patterns from the JASPAR database, as Figure 6 illustrates. On this basis, we choose six factors to construct the 3D genome structure correlation task. Among these, CTCF associates with 3D architecture through its role in forming TADs and loops. The label Undefined indicates standard usage without contributing to additional problem dimensions. Table 8 provides detailed specifications.

### A.6.2 Label Statistics

In this task, while genome sequences naturally contain multiple binding motifs (multi-label), each question interrogates the presence of a specific target Transcription Factor selected from a set of 20. The details are provided in Figure 5(i) and Figure 5(j). Similar to the Histone Mark task, the data ex-

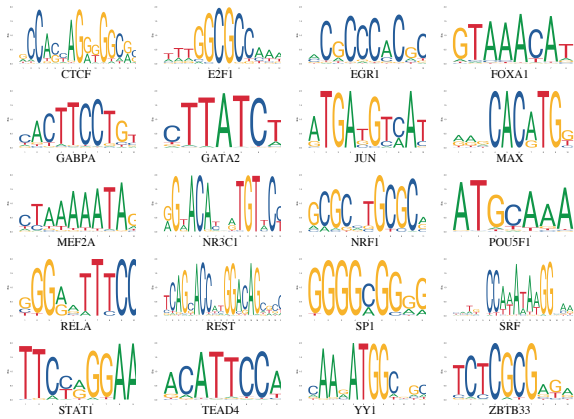


Figure 6: Motifs in TFBS Prediction.

hibits a deliberate distributional skew across both Binary and Multiple Choice Questions. To evaluate the model’s understanding of chromatin architecture, we significantly oversampled six specific TFs: CTCF, a key architectural protein facilitating TADs and loops, and SP1, GABPA, E2F1, TEAD4, and NRF1. Consequently, questions targeting these six factors dominate the statistics compared to the remaining 14 labels, prioritizing the assessment of structural regulatory logic over a uniform distribution.

## A.7 TF Motif Identification

### A.7.1 Label Sets and Entity List

We use the same label set as in the TFBS task. In this task, we do not differentiate between whether a factor is associated with 3D genome structure.

### A.7.2 Label Statistics

In this task, each sequence corresponds to a single dominant motif selected from a set of 20 transcription factors. The statistics are shown in Figure 5(k) and Figure 5(l). In the Binary Choice Questions, the sampling is uniformly distributed across the 20 categories, ensuring no single motif dominates the evaluation. Similarly, the Multiple Choice Questions maintain this balanced approach, featuring an equitable distribution of target queries across the entire label set. This uniform coverage guarantees that the model’s ability to recognize short consensus patterns is tested fairly across diverse biological contexts without class bias.

## B System Prompt Optimization

To ensure that the evaluation accurately reflects the genome reasoning capabilities of Large Language Models (LLMs) rather than their sensitiv-

ity to prompt engineering, we conduct an ablation study to derive an optimized system prompt. This prompt aims to maximize the models’ ability to interpret DNA sequences by providing clear domain constraints and reasoning protocols.

### B.1 Development Set Construction

To balance computational cost with statistical representativeness, we construct a small-scale development set comprising 90 samples. We employ a stratified sampling strategy by randomly selecting three samples for each unique combination of problem type, task category, and specific problem dimension. This selection spans binary and multiple-choice formats across tasks such as TFBS prediction and splice site identification, including specific dimensions like 3D genome structure correlation. This compact yet diverse dataset serves as the testbed for iterative prompt refinement.

### B.2 Optimization Strategy

The optimization process follows a two-stage iteration starting with a baseline prompt generated by Gemini 3 Pro, as Figure 7 illustrates. This initial text provides a general persona of a computational biology expert and standard instructions to analyze sequences. We conduct manual error analysis on the model outputs and observe that models frequently hallucinate constraints or refuse to answer due to a perceived lack of evidence. To address this, we manually refine the text into the Optimized System Prompt presented in Figure 8. Key enhancements include explicitly instructing models to choose the most likely option based on probabilistic signals, defining six specific biological domains to narrow the search space, and mandating a structured analysis of motif organization. This final requirement specifically aids in distinguishing real sequences from dinucleotide-preserved shuffled controls.

### B.3 Performance Validation

We evaluate the effectiveness of the optimized prompt against the baseline across six LLMs on the development set. As the ablation studies in Table 9 and Table 10 demonstrate, the optimized prompt consistently matches or outperforms the baseline version. Across all six evaluated models, the optimized prompt yields performance gains in at least three out of six tasks per model. The improvement is particularly notable in tasks requiring subtle pattern recognition, where the explicit

## Base Prompt

You are an expert in Computational Biology and Bioinformatics. Your task is to solve problems related to genome sequence analysis.

You will be provided with a specific question and a set of options. The DNA sequences may appear in the question description or within the options themselves.

Please follow these instructions:

1. Identify the specific biological task (e.g., distinguishing enhancer vs. promoter).
2. Briefly analyze the key sequence features or biological context relevant to the question. Keep this analysis concise (1-3 sentences) to justify your choice.
3. Select the most accurate option letter (e.g., A, B, C, D) .

Strictly follow the following output format:

### Analysis

[Concise analysis in 1-3 sentences]

### Answer

Answer: [Option Letter]

Figure 7: The Base Prompt used as a baseline. It contains only basic task descriptions and formatting instructions without domain-specific reasoning guidance.

instruction to avoid hedging forces models to leverage weak biological signals for decision-making.

## C Empirical Baseline Verification

To validate the reliability of using theoretical probability as a performance floor, we calculate empirical random baselines by simulating random guessing across all evaluation tasks. We compare these empirical values against the standard theoretical expectations of 50.00% for Binary Choice Questions (BCQ) and 25.00% for Multiple Choice Questions (MCQ).

As presented in Table 11, the empirical results align closely with theoretical expectations. The average empirical accuracy across all tasks is 51.17% for BCQ and 25.23% for MCQ, exhibiting only negligible deviations from the theoretical values. The overall consistency confirms that the theoretical baselines serve as accurate and fair proxies for zero-knowledge performance in our main analysis.

## D Additional Baseline Details

### D.1 Supervised Genomic Baselines

We include three additional baselines to complement the zero-shot evaluation of frontier LLMs: (i) a GC-content classifier, (ii) a 3-mer frequency linear classifier, and (iii) DNABERT-2 fine-tuned separately for each task family.

**GC-content classifier.** For each sequence, we compute the GC fraction as a single scalar feature. We fit Gaussian Naive Bayes for Tasks 1/2/3/4/6

and independent per-label Gaussian models for the multi-label TFBS task (Task 5).

**3-mer classifier.** We represent each sequence using overlapping 3-mer frequency features and train a logistic regression classifier. For TFBS, we use one-vs-rest binary classifiers for each label.

**DNABERT-2.** We use DNABERT-2 as a representative genome foundation model and fine-tune one model per task family. Each model is trained for 5 epochs with 50-step warm-up to a peak learning rate of  $3 \times 10^{-5}$ , followed by cosine decay.

### D.2 Shared Fine-tuning Dataset

All three baselines are trained on the same supervised train/validation splits aligned with the six GenomeQA task families. Table 12 summarizes the dataset sizes and task types.

## E Additional Evaluation Metrics

To complement accuracy, we report option-level weighted F1 for both BCQ and MCQ settings. We treat the predicted choice letter as the class label (A/B for BCQ; A/B/C/D for MCQ) and compute weighted F1 over these option classes. As shown in Table 13, weighted F1 exhibits the same overall trends and model rankings as accuracy.

## F Statistical Uncertainty Analysis

For each model  $\times$  task  $\times$  format setting, we treat each question as an independent Bernoulli trial and compute a Wilson 95% confidence interval for

Model	BCQ (33 samples)														No. Task Wins
	Enhancer and Promoter Identification		Splice Site Identification		Taxonomic Classification		Histone Mark Prediction		TFBS Prediction		TF Motif Prediction		Avg.		
	Base	Opt.	Base	Opt.	Base	Opt.	Base	Opt.	Base	Opt.	Base	Opt.	Base	Opt.	
Claude-Sonnet-4.5	100.00	66.67	44.44	22.22	33.33	<b>100.00</b>	33.33	<b>50.00</b>	66.67	55.56	66.67	<b>66.67</b>	57.41	<b>60.19</b>	3
GPT-5.1	100.00	<b>100.00</b>	88.89	44.44	100.00	66.67	66.67	<b>66.67</b>	55.56	<b>66.67</b>	33.33	<b>100.00</b>	74.07	<b>74.07</b>	4
Gemini3-Pro	66.67	<b>100.00</b>	44.44	<b>55.56</b>	100.00	<b>100.00</b>	50.00	33.33	77.78	<b>77.78</b>	100.00	<b>100.00</b>	73.15	<b>77.78</b>	5
Grok-4.1	100.00	<b>100.00</b>	33.33	<b>66.67</b>	66.67	<b>66.67</b>	33.33	<b>50.00</b>	66.67	<b>66.67</b>	33.33	<b>66.67</b>	55.56	<b>69.44</b>	6
Llama4	66.67	<b>66.67</b>	55.56	<b>55.56</b>	66.67	<b>66.67</b>	83.33	<b>83.33</b>	77.78	66.67	0.00	<b>66.67</b>	58.33	<b>67.59</b>	5
Qwen3-Max	0.00	<b>100.00</b>	66.67	<b>77.78</b>	66.67	<b>66.67</b>	50.00	<b>83.33</b>	66.67	55.56	100.00	66.67	58.33	<b>75.00</b>	4

Table 9: Ablation study on Binary Choice Questions. Performance comparison between the initial and refined prompts on the 33-sample BCQ development set. Base denotes the Base System Prompt and Opt. denotes the Optimized System Prompt. Avg. represents the average accuracy across all six tasks. No. Task Wins indicates the number of tasks where the optimized prompt achieved equal or higher accuracy than the base prompt. Bold values highlight instances where the optimized prompt resulted in improved or equivalent performance compared to the baseline.

Model	MCQ (57 samples)														No. Task Wins
	Enhancer and Promoter Identification		Splice Site Identification		Taxonomic Classification		Histone Mark Prediction		TFBS Prediction		TF Motif Prediction		Avg.		
	Base	Opt.	Base	Opt.	Base	Opt.	Base	Opt.	Base	Opt.	Base	Opt.	Base	Opt.	
Claude-Sonnet-4.5	66.67	<b>77.78</b>	44.44	33.33	77.78	<b>88.89</b>	40.00	<b>66.67</b>	44.44	<b>66.67</b>	83.33	<b>83.33</b>	59.44	<b>69.44</b>	5
GPT-5.1	44.44	<b>77.78</b>	11.11	<b>22.22</b>	66.67	<b>66.67</b>	40.00	<b>53.33</b>	44.44	<b>55.56</b>	100.00	83.33	51.11	<b>59.81</b>	5
Gemini3-Pro	66.67	55.56	22.22	<b>44.44</b>	100.00	88.89	40.00	<b>46.67</b>	22.22	<b>44.44</b>	100.00	<b>100.00</b>	58.52	<b>63.33</b>	4
Grok-4.1	44.44	<b>44.44</b>	33.33	22.22	55.56	44.44	33.33	<b>40.00</b>	77.78	66.67	100.00	<b>100.00</b>	57.41	52.96	3
Llama4	33.33	11.11	11.11	<b>11.11</b>	77.78	55.56	46.67	<b>46.67</b>	55.56	<b>66.67</b>	66.67	<b>83.33</b>	48.52	45.74	4
Qwen3-Max	55.56	44.44	11.11	<b>11.11</b>	66.67	<b>77.78</b>	33.33	<b>40.00</b>	77.78	55.56	83.33	<b>83.33</b>	54.63	52.04	4

Table 10: Ablation study on Multiple Choice Questions. Performance comparison on the 57-sample MCQ development set. Base denotes the Base System Prompt and Opt. denotes the Optimized System Prompt. Avg. represents the average accuracy across all six tasks. No. Task Wins indicates the number of tasks where the optimized prompt achieved equal or higher accuracy than the base prompt. Bold formatting signifies that the optimized prompt achieved performance parity or improvement over the base prompt.

Task	BCQ	MCQ
Enhancer and Promoter Identification	51.40	25.80
Splice Site Identification	52.00	24.00
Taxonomic Classification	50.80	28.20
Histone Mark Prediction	49.60	23.00
TFBS Prediction	49.20	22.40
TF Motif Prediction	54.00	28.00
Avg	51.17	25.23

Table 11: Empirical random baseline accuracy (%) across different tasks in GenomeQA.

accuracy. We test whether performance exceeds random chance using a one-sided exact binomial test, with chance levels of 0.50 for BCQ and 0.25 for MCQ. We then apply Benjamini-Hochberg correction across all tested settings.

Table 14 reports per-task summary confidence intervals and the number of models that significantly outperform chance after correction. Table 15 re-

Task	Train	Val	Labels	Task Type
Enhancer and Promoter Identification	10000	2500	2	Multi-class
Splice Site Identification	11384	2846	4	Multi-class
Taxonomic Classification	8479	2120	3	Multi-class
Histone Mark Prediction	20000	5000	10	Multi-class
TFBS Prediction	12050	3010	20	Multi-label
TF Motif Prediction	7060	1765	20	Multi-class

Table 12: The statistics of finetuning dataset.

ports overall average accuracies together with 95% confidence intervals to contextualize the stability of the model ordering. These analyses quantify finite-sample uncertainty under a single evaluation run and should not be interpreted as repeated-run variance estimates.

## G Balanced and Oversampled Slice Analysis

Histone Mark Prediction and TFBS Prediction include intentionally oversampled stress labels in the full benchmark. To verify that the main conclusions are not driven solely by this distributional

Model	Thinking	Enhancer and Promoter Identification	Splice Site Identification	Taxonomic Classification	Histone Mark Prediction	TFBS Prediction	TF Motif Prediction	Avg.
<b>BCQ</b>								
Claude-Sonnet-4.5	Y	69.01	53.11	61.57	53.74	56.50	64.09	59.67
GPT-5.1	Y	<b>70.59</b>	<b>54.37</b>	64.60	53.16	55.74	66.05	60.75
Gemini-3-Pro	Y	67.81	50.00	<b>78.77</b>	<b>57.20</b>	<b>59.80</b>	<b>84.04</b>	<b>66.27</b>
Grok-4.1	Y	67.57	47.59	63.76	53.13	54.51	69.08	59.27
Llama-4	N	55.03	49.58	56.37	51.96	51.41	71.06	55.90
Qwen3-Max	Y	60.50	49.75	60.60	51.11	53.71	63.03	56.45
<b>MCQ</b>								
Claude-Sonnet-4.5	Y	59.00	28.32	64.62	35.64	47.98	81.04	52.77
GPT-5.1	Y	62.64	26.75	60.95	36.25	50.07	76.92	52.26
Gemini-3-Pro	Y	<b>66.05</b>	<b>33.42</b>	<b>77.01</b>	<b>37.73</b>	<b>58.95</b>	<b>92.08</b>	<b>60.87</b>
Grok-4.1	Y	56.18	25.74	61.17	33.01	48.44	78.71	50.54
Llama-4	N	36.22	26.07	56.13	30.13	33.77	65.03	41.23
Qwen3-Max	Y	43.04	25.72	52.72	32.89	45.19	71.88	45.24

Table 13: Overall weighted-F1 results of LLMs on GenomeQA.

Task	BCQ Avg. Acc (95% CI)	BCQ: Models Beating Chance	MCQ Avg. Acc (95% CI)	MCQ: Models Beating Chance
Enhancer and Promoter Identification	65.2% [63.4, 66.9]	6 / 6	53.8% [52.0, 55.6]	6 / 6
Splice Site Identification	50.8% [49.0, 52.6]	1 / 6	27.7% [26.1, 29.3]	2 / 6
Taxonomic Classification	64.4% [62.7, 66.1]	6 / 6	62.1% [60.4, 63.9]	6 / 6
Histone Mark Prediction	53.5% [51.7, 55.3]	2 / 6	34.7% [33.0, 36.4]	6 / 6
TFBS Prediction	55.4% [53.6, 57.2]	5 / 6	47.5% [45.8, 49.3]	6 / 6
TF Motif Prediction	69.8% [66.0, 73.4]	6 / 6	77.7% [74.2, 80.8]	6 / 6

Table 14: Per-task confidence intervals and significance against chance. For each task under BCQ and MCQ, we report the mean accuracy across models with Wilson 95% confidence intervals, and the number of models whose performance significantly exceeds the random baseline (0.50 for BCQ; 0.25 for MCQ) under a one-sided exact binomial test with Benjamini-Hochberg correction.

Model	Avg. Acc	95% CI
Gemini	63.60	[62.20, 64.90]
GPT-5.1	56.50	[55.00, 58.10]
Claude	56.30	[54.80, 57.80]
Grok	55.00	[53.50, 56.50]
Qwen3	50.90	[49.40, 52.50]
Llama4	49.10	[47.50, 50.60]

Table 15: Overall model accuracies with Wilson 95% confidence intervals. The intervals quantify finite-sample uncertainty under the observed single evaluation run and are used to contextualize the overall model ordering, rather than to claim pairwise ranking significance.

skew, we separately report results on a balanced slice and an oversampled (stress) slice for both BCQ and MCQ. As shown in Tables 16, the overall model ordering remains broadly stable across the two slices. In TFBS Prediction, the oversampled stress subset is substantially harder, especially for questions involving CTCF/3D-genome-related signals, whereas for Histone Mark Prediction the stress subset can be easier because it emphasizes coarser chromatin-state patterns.

## H Details of Thinking Mode

To improve reproducibility, we explicitly report how thinking mode is controlled for each provider, together with aggregate token usage statistics. As shown in Table 17, all models are evaluated under the same decoding settings; thinking is controlled via provider-side API toggles rather than prompt modifications. For models accessed through OpenRouter, thinking is enabled with `{"reasoning": {"enabled": true}}` (default medium effort) and disabled in the ablation setting with `{"reasoning": {"effort": "none"}}`. For Qwen3-Max accessed through the Alibaba API, we control reasoning with `{"enable_thinking": true}` or `false`. We do not impose an explicit reasoning-token budget.

## I Case Study

Figure 9 presents examples of four failure modes in genome analysis: Sequence Motif Over-reliance (SMO), Base Composition Over-reliance (BCO), Character Fidelity Loss (CFL), and Noise Distinction Failure (NDF). Each example includes the in-

put context, model output, and an analysis of the error. These cases highlight distinct cognitive gaps in general LLMs: prioritizing general sequence elements over specific details (SMO), using statistical shortcuts like GC content while ignoring structural patterns (BCO), losing character-level fidelity to fabricate non-existent sub-sequences (CFL), and the sycophantic tendency to rationalize random noise as valid biological signals (NDF). These examples illustrate the systematic error patterns in current LLMs and underscore the need for domain-aligned reasoning capabilities.

Model	Histone Mark Prediction (Balanced)	TFBS Prediction (Balanced)	Avg.	Histone Mark Prediction (Oversampled)	TFBS Prediction (Oversampled)	Avg.
BCQ						
Claude-Sonnet-4.5	52.24	57.78	55.01	54.85	54.22	54.54
GPT-5.1	53.23	57.49	55.36	53.18	52.41	52.80
Gemini3-Pro	53.73	59.58	<b>56.66</b>	59.53	60.24	<b>59.89</b>
Grok-4.1	56.72	54.49	55.61	50.84	54.82	52.83
Llama4	51.24	52.40	51.82	53.51	50.00	51.76
Qwen3-Max	48.26	58.08	53.17	53.18	45.78	49.48
MCQ						
Claude-Sonnet-4.5	32.00	58.38	45.19	47.20	27.11	37.16
GPT-5.1	32.27	61.68	46.98	48.00	27.11	37.56
Gemini3-Pro	33.07	66.17	<b>49.62</b>	52.00	44.58	<b>48.29</b>
Grok-4.1	30.13	58.98	44.56	43.20	27.11	35.16
Llama4	29.60	39.52	34.56	37.60	24.10	30.85
Qwen3-Max	30.93	55.69	43.31	40.00	24.10	32.05

Table 16: BCQ and MCQ results on balanced and oversampled slices for the two task families with intentional label skew: Histone Mark Prediction and TFBS Prediction. The balanced slice reduces label-frequency skew, whereas the oversampled slice preserves the stress-test setting used in the main benchmark.

Model	Provider (API)	Thinking	Total Token Usage: Mean (Min - Max)
Claude-Sonnet-4.5	OpenRouter	On	5,750 (1,512 - 30,657)
GPT-5.1	OpenRouter	On	2,777 (955 - 13,706)
Gemini-3-Pro	OpenRouter	On	5,403 (1,414 - 30,021)
Grok-4.1	OpenRouter	On	6,224 (1,410 - 28,408)
Qwen3-Max	Alibaba	On	8,623 (1,485 - 152,558)
Llama-4	OpenRouter	Not supported	1,606 (913 - 10,419)
GPT-5.1 (Ablation)	OpenRouter	Off	1,574 (912 - 3,239)
Qwen3-Max (Ablation)	Alibaba	Off	1,600 (931 - 3,331)

Table 17: Thinking mode configuration and resource usage

## Optimized Prompt

You are an expert in Computational Biology, Regulatory Genomics, and Bioinformatics. Your goal is to infer the correct answer by analyzing nucleotide sequences with probabilistic reasoning and biological domain knowledge, not by exploiting superficial text patterns or option positions.

Questions implicitly belong to one of six domains (you must infer which):

1. Regulatory identity: enhancer vs promoter in human.
2. Splice-site analysis: acceptor vs donor vs both, and distinguishing real sites from dinucleotide-preserved shuffled controls.
3. Taxonomy: classifying sequences as eukaryote, prokaryote, or virus.
4. Epigenetics: histone mark identity and chromatin state (open/accessible vs closed/repressive) in human K562 cells.
5. TFBS prediction: locating binding sites of specific transcription factors within longer human sequences.
6. Motif identification: recognizing short (6-20 bp) consensus patterns.

Please adhere to the following Reasoning Protocol:

### 1. Infer task & question type

From the wording, infer the biological domain and whether the question is: binary verification, multi-class label selection, sequence selection, odd-one-out, or real-vs-shuffled discrimination.

### 2. Analyze sequence patterns

Assume all sequences are in functional 5'→3' orientation unless stated otherwise. Do NOT perform reverse-complement analysis unless explicitly asked. Focus on informative patterns:

- Short motifs and motif-like patterns, their approximate positions and local context.
- Base composition, GC content, k-mer and codon-like usage, periodicity, and sequence complexity.
- Differences between options (for selection / odd-one-out), such as one sequence having a stronger or more coherent pattern than the others.

Treat any examples you know (e.g., canonical splice motifs, promoter features, TF consensus-like sites, species-specific composition) as \*hints\*, not as a closed list: you are encouraged to use any additional recurrent or statistically distinctive patterns you detect in the sequences, even if they are not classic textbook motifs.

### 3. Make a probabilistic decision

Do NOT require perfect consensus matches; biological signals are degenerate and noisy.

- Interpret strict verbs (“is this X?”, “confirm”) as asking which label is \*more likely\*.
- For real vs dinucleotide-preserved shuffled/background comparisons, remember that low-level composition is similar; rely more on motif organization, local structure, and plausibly functional subpatterns.
- Do not hedge in your analysis; always choose the single most biologically plausible option.

Output rules:

- Base your decision strictly on nucleotide patterns and biological meaning of the labels, not on dataset biases or text-only artifacts.
- Do NOT copy or restate any full DNA sequence in your analysis; refer only to features (motifs, composition, relative differences, etc.).
- For all questions, there is exactly one best answer. Always pick a single option letter.

Output format (strictly follow):

### Analysis

1-3 concise sentences explaining (i) the inferred task/domain and (ii) the key sequence-based or comparative reasoning that makes your chosen option most probable.

### Answer

Answer: [Single option letter, e.g., A / B / C / D]

Figure 8: The Optimized Prompt designed for the LLM. It includes explicit role definition, domain constraints, a step-by-step reasoning protocol, and strict output formatting rules.

## Sequence Motif Over-reliance

### Histone Mark Prediction (ID: 428)

#### Input

**Q:** Sequence:

AAGG...**GCAGTGGCTC**...AATT  
Analyze the Human DNA sequence in K562 cells. Is this sequence found in an open/accessible environment?

**Label:** Yes

#### Model Output

**Analysis:** ...Alu Repeat Element: A large portion of the sequence GCAGTGGCTC ... **While Alu elements are CpG-rich internally, they are** predominantly methylated and associated with heterochromatin or **repressed regions** ...

**Answer:** No

**Comment:** The model applies a general repression heuristic, ignoring the Alu element's high GC content that suggests accessibility.

## Base Composition Over-reliance

### Taxonomic Classification (ID: 139)

#### Input

**Q:** Determine the category of the organism associated with this DNA sequence: **CACG...GCCATCCG...CTCG**.

**Label:** Virus

#### Model Output

**Analysis:** ...The provided DNA sequence displays an **extremely high GC content (visually estimated >70%)**. This composition is a hallmark of high-GC Gram-positive **bacteria**, specifically the phylum Actinobacteria...

**Answer:** Prokaryote

**Comment:** The model uses high GC content as a shortcut for bacterial classification, ignoring specific gene organization that points to a viral origin.

## Character Fidelity Loss

### Enhancer and Promoter Identification (ID: 442)

#### Input

**Q:** Is there any enhancer or promoter in the following Human DNA sequence?

Sequence: **TTTG...AGGGAGCGGAGCC...CCAG**.

**Label:** Enhancer

#### Model Output

**Analysis:** ...The sequence contains multiple GC-rich motifs (e.g., '**GGGCGG**', '**CCGCCC**') resembling Sp1 binding sites, which are common in TATA-less **promoters**...

**Answer:** Promoter

**Comment:** Count("GGGCGG") = 0, Count("CCGCCC") = 0. The model hallucinates specific motif sequences that are entirely absent from the input to support its classification.

## Noise Distinction Failure

### Splice Site Identification (ID: 417)

#### Input

**Sample Type:** Dinucleotide-preserving Shuffled Sequence

**Q:** Does this Human DNA sequence contain **authentic** splice sites?

Sequence: **GTGT...TATT**

**Label:** No

#### Model Output

**Analysis:** ...it contains a strong donor splice site motif...The preservation of these complex, position-specific motifs **strongly indicates this is a real genomic sequence** rather than a shuffled control...

**Answer:** Yes

**Comment:** The input is a shuffled control. The model fails to recognize the meaningless patterns in the shuffled sequence, ignoring the stochastic nature that indicates it is random noise.

Figure 9: Case Study. Examples of failure modes in GenomeQA along with input details and model responses. Green: Ground truth labels, specific input characteristics provided by the benchmark. Red: Incorrect reasoning, or fabricated evidence generated by the model.