EPIBENCH: CAN LLMS UNDERSTAND EPITOPES FOR ANTIBODY DRUG DISCOVERY?

Anonymous authors
Paper under double-blind review

ABSTRACT

Epitopes determine where antibodies bind antigens and shape downstream therapeutic properties such as functional blockade and escape resistance, making epitope understanding central to antibody drug discovery. Although large language models (LLMs) have shown strong biomedical reasoning ability, it remains unclear whether they can infer epitope information directly from antigen and antibody sequences. Existing epitope resources typically focus on isolated prediction tasks or rely on specialized structural settings, while general protein benchmarks do not evaluate epitope-centered decisions across the antibody development workflow. To address this gap, we introduce EpiBench, a closed-book, sequence-based, and automatically scorable benchmark for evaluating epitope reasoning in LLMs. EpiBench contains 1,609 curated samples grounded in structural antibody–antigen contacts, curated functional B-cell assays, and deep mutational scanning escape measurements. It covers five connected tasks: targetable region discovery, antibody-conditioned epitope identification, epitope binning, functional epitope assessment, and antibody escape assessment, with controlled sampling to reduce shortcut-based evaluation artifacts. We evaluate nine general-purpose LLMs and analyze their behavior through task-specific baselines, antigen length stratification, explicit-reasoning comparison, and failure-mode inspection. The results show that current LLMs capture partial epitope-related signals but remain limited in antibody-specific sequence grounding, long-context residue localization, and biologically grounded reasoning. Therefore, EpiBench provides a diagnostic testbed for measuring and improving sequence-aware biomedical LLMs toward reliable LLM-assisted antibody discovery.

1 INTRODUCTION

The discovery of therapeutic antibodies is central to modern biomedicine and increasingly important for precision treatment across diverse diseases. The therapeutic effect of an antibody depends not only on whether it recognizes an antigen but also on which antigen region it engages. This binding region is known as the epitope, and it shapes target engagement, functional blockade, immune escape, and therapeutic specificity. Understanding antibody antigen binding through epitopes is therefore a fundamental problem in antibody drug design and optimization. In recent years, task specific deep learning methods have been developed to predict epitopes from antigen sequences or protein structures (Clifford et al., 2022; Ponomarenko et al., 2008; Liu et al., 2024). These specialized models provide evidence that epitope related signals are computationally predictable, yet they are usually optimized for a narrow prediction target and do not directly address broader epitope centered decisions in antibody drug development. In parallel, protein language models such as ESM have expanded protein sequence modeling toward general representation learning, showing that large scale self supervised training can capture structural and functional regularities from biological sequences (Rives et al., 2021; Lin et al., 2023). Together, these advances suggest that antigen and antibody sequences contain learnable signals for reasoning about antibody antigen interfaces.

Meanwhile, LLMs have demonstrated strong understanding and reasoning abilities across biomedical tasks (Singhal et al., 2023), but their ability to infer epitope information from antigen and antibody sequences remains unclear. This motivates a central question: **Can LLMs infer epitope information from antigen and antibody sequences and use it to support antibody drug design?** Addressing this question requires moving beyond isolated epitope prediction and evaluating whether

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

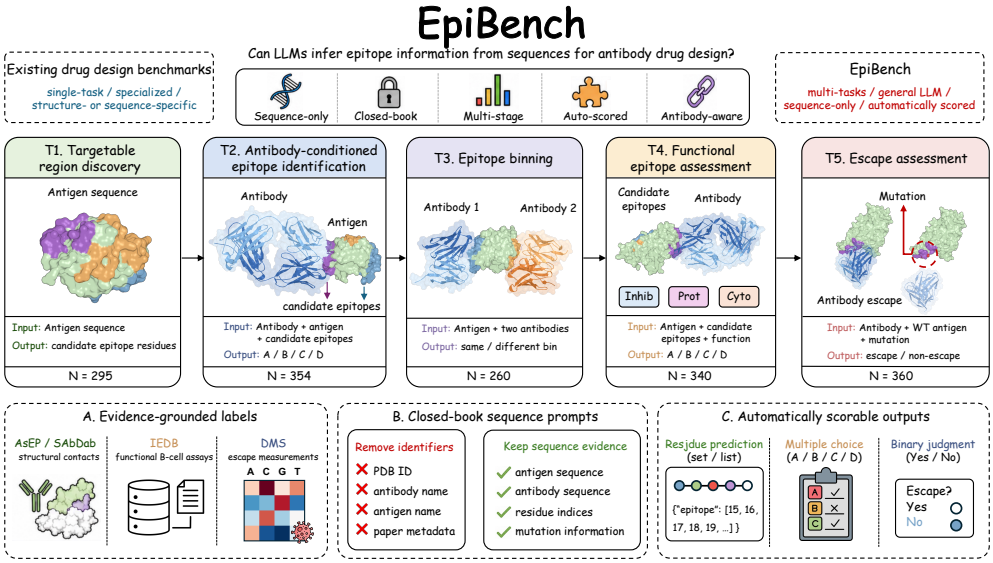


Figure 1: Overview of EpiBench. The benchmark evaluates whether LLMs can infer epitope information from antigen and antibody sequences through a closed-book, sequence-based pipeline covering targetable region discovery, antibody-conditioned epitope identification, epitope binning, functional epitope assessment, and escape assessment. Labels are grounded in structural contacts, curated functional assays, and deep mutational scanning evidence, while prompts remove identifiers and retain sequence-level evidence for automatically scorable evaluation.

LLMs can support a sequence driven decision chain that includes binding site localization, antibody grouping, functional interpretation, and escape assessment.

To answer this question, we introduce EpiBench, a sequence based benchmark for evaluating epitope reasoning in antibody drug discovery. EpiBench is designed to simulate the workflow of antibody drug development, where epitope information supports multiple decisions across discovery, characterization, and optimization. It covers five connected decision tasks, including antibody free epitope prediction, antibody conditioned epitope identification, antibody epitope binning, functional epitope assessment, and antibody escape assessment. The benchmark contains 1,609 curated samples constructed with controlled sampling strategies to improve diversity, balance answer distributions, and reduce shortcut based evaluation artifacts. According to the biological objective of each task, we formulate the samples as question answering, multiple choice, and binary judgment problems, and design task adapted evaluation metrics for systematic assessment. We summarize our main contributions below.

- We introduce EpiBench, a standardized benchmark for evaluating LLMs on epitope reasoning in antibody drug discovery. EpiBench is organized around the practical workflow of antibody development and covers five key stages spanning epitope localization, antibody conditioned recognition, epitope binning, functional assessment, and escape analysis.

- We conduct a comprehensive evaluation of nine LLMs and compare them with task-specific epitope models where directly applicable. The results show that current LLMs capture partial epitope-related signals, but remain limited in long-context residue localization, antibody-specific sequence grounding, and tasks requiring indirect functional or escape reasoning.

- We provide a fine grained analysis of model behavior across tasks, revealing systematic failure modes in antibody antigen interface reasoning. These findings clarify current limitations of LLMs for epitope centered drug discovery and suggest directions for future biological sequence reasoning models.

2 Related Work

2.1 Task Specific Epitope Prediction Models

Computational prediction of antibody antigen interfaces has traditionally been studied through task specific epitope prediction models. Early sequence based methods predicted linear B cell epitopes from antigen primary sequences using physicochemical propensity scores, local sequence windows, recurrent neural networks, and kernel based classifiers (Emini et al., 1985; Kolaskar & Tongaonkar, 1990; Larsen et al., 2006; Saha & Raghava, 2006; El-Manzalawy et al., 2008). More recent methods further incorporate deep learning and protein language representations to improve epitope prediction under antigen only or antibody conditioned settings (Clifford et al., 2022; Liu et al., 2024). These studies provide direct evidence that epitope related signals can be learned from biological sequences, but their learning objectives are usually tied to a single prediction target and a fixed input format.

Complementary studies predict conformational epitopes by using three dimensional antigen structures, solvent exposure, surface geometry, residue neighborhoods, or sequence derived approximations of spatial binding sites (Haste Andersen et al., 2006; Ponomarenko et al., 2008; Sun et al., 2009; Ansari & Raghava, 2010). However, most existing methods are designed around a single prediction endpoint, which leaves multi task epitope reasoning in antibody drug design insufficiently examined. EpiBench addresses this evaluation gap by organizing the assessment around five key stages, including epitope localization, antibody conditioned recognition, epitope binning, functional assessment, and escape analysis.

2.2 Protein Language Models and Sequence Understanding Benchmarks

Protein language models have reframed protein sequence analysis as self supervised representation learning over amino acid sequences. Models such as ProtTrans and ESM train Transformer architectures on large protein corpora and show that learned representations encode structural, functional, and evolutionary information (Elnaggar et al., 2022; Rives et al., 2021). ESMFold further demonstrates sequence to structure prediction from language model representations, while autoregressive models such as ProGen and ProGen2 extend language modeling to controllable protein sequence generation (Lin et al., 2023; Madani et al., 2023; Nijkamp et al., 2022). Recent biomolecular LLM studies further connect biological sequences with textual instructions, as shown by Mol-Instructions and ProtLLM (Fang et al., 2023; Zhuo et al., 2024). Together, these works support the view that protein sequences contain rich learnable signals that can be exploited by language modeling.

Protein sequence benchmarks have provided standardized settings for measuring these modeling capabilities. TAPE evaluates transfer learning across structural and functional protein tasks, PEER expands protein sequence understanding to function, localization, structure, protein protein interaction, and protein ligand interaction prediction, and fitness oriented benchmarks such as FLIP and ProteinGym evaluate few shot fitness prediction and variant effect modeling (Rao et al., 2019; Xu et al., 2022; Mollon et al., 2025; Notin et al., 2022). These benchmarks are essential for assessing general protein sequence representations, but they primarily focus on broad biological properties, fitness landscapes, or variant effects. They do not directly examine epitope centered reasoning for antibody antigen recognition, where models must relate binding site localization, antibody grouping, functional interpretation, and escape assessment within a drug development context. EpiBench focuses on this missing setting by evaluating epitope reasoning across tasks aligned with antibody drug discovery.

3 EpiBench

3.1 Design Principles

Biomedical LLM evaluation has largely centered on textual scientific knowledge, including biomedical question answering, literature search, evidence retrieval, and experimental report interpretation. These settings are important for measuring how models access and organize biomedical information, but they do not directly assess whether language models can reason from biological sequences. EpiBench complements this line of evaluation by constructing a sequence centered benchmark in which models must infer epitope related information from antigen and antibody sequences and use it in molecular decision tasks. This setting directly links scientific reasoning with biological

Table 1: Comparison with representative epitope, antibody, and protein benchmarks shows that EpiBench uniquely combines sequence-accessible epitope-centered antibody–antigen reasoning, multi-source evidence, drug-discovery objectives, LLM-ready formulation, and shortcut-controlled automatic evaluation.

	Dataset				Task		Evaluation			
	Epitope centered	Ab–Ag context	Sequence input	Multi-source evidence	Multi-objective	Drug discovery	LLM-ready	Standard protocol	Auto scoring	Bias control
ABCPred (Saha & Raghava, 2006)	✓	✗	✓	✗	✗	✗	✗	✓	✓	✗
BCPred (El-Manzalawy et al., 2008)	✓	✗	✓	✗	✗	✗	✗	✓	✓	✗
BepiPred-3.0 eval set (Clifford et al., 2022)	✓	✗	✓	✓	✗	✗	✗	✓	✓	✗
AsEP (Liu et al., 2024)	✓	✓	✗	✗	✗	✓	✗	✓	✓	✓
CHIMERA-Bench (Ahmed et al., 2026)	✓	✓	✗	✗	✓	✓	✗	✓	✓	✓
TDC immune tasks (Huang et al., 2022)	✓	✗	✓	✓	✓	✓	✗	✓	✓	✗
TDC developability (Huang et al., 2022)	✗	✗	✓	✗	✗	✓	✗	✓	✓	✗
Ginkgo AbDev (van Niekerk et al., 2026)	✗	✗	✓	✗	✓	✓	✗	✓	✓	✓
FLAb (Chungyoun et al., 2024)	✗	✗	✓	✓	✓	✓	✗	✓	✓	✗
AbBiBench (Zhao et al., 2025)	✗	✓	✓	✓	✓	✓	✗	✓	✓	✓
TAPE (Rao et al., 2019)	✗	✗	✓	✓	✓	✗	✗	✓	✓	✓
PEER (Xu et al., 2022)	✗	✗	✓	✓	✓	✗	✗	✓	✓	✓
ProteinGLUE (Capel et al., 2022)	✗	✗	✓	✓	✓	✗	✗	✓	✓	✗
PETA (Tan et al., 2024)	✗	✗	✓	✓	✓	✗	✗	✓	✓	✓
ProteinGym (Notin et al., 2023)	✗	✗	✓	✓	✓	✓	✗	✓	✓	✓
FLIP/FLIP2 (Dallago et al., 2021; Didi et al., 2026)	✗	✗	✓	✓	✓	✓	✗	✓	✓	✓
ATOM3D (Townshend et al., 2020)	✗	✗	✗	✓	✓	✓	✗	✓	✓	✓
EpiBench (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

sequence understanding and provides a controlled testbed for examining the potential of general purpose LLMs in complex sequence driven scientific problems.

The construction of EpiBench follows five principles: (1) aligning tasks with key stages of antibody development, from candidate epitope discovery to escape assessment; (2) deriving labels from experimentally grounded sources, including structural antibody antigen complexes, curated functional B cell assays, and deep mutational scanning measurements (Liu et al., 2024; Dunbar et al., 2014; Vita et al., 2025; Cao et al., 2022); (3) using a closed book sequence formulation that removes identifiers such as PDB IDs, antigen names, antibody names, accession numbers, native numbering schemes, and publication metadata; (4) adopting task adapted answer formats, including residue prediction, multiple choice, and binary judgment; and (5) controlling answer distributions, candidate orders, distractors, and matched pairs to reduce shortcut based evaluation artifacts.

3.2 Benchmark Task Design

EpiBench converts epitope centered decisions in antibody drug discovery into task specific prompts with automatically scorable outputs. This section describes the motivation and high level construction strategy for each task, while detailed data filtering, label construction, and sampling procedures are provided in Appendix B.1.

Task1: Antibody Free Epitope Prediction. This task evaluates whether a model can identify antigen regions that are likely to be recognized before a specific antibody candidate is available. This setting reflects early target and epitope assessment, where researchers need to prioritize potentially targetable regions on an antigen sequence. We operationalize the task by asking the model to predict a limited set of candidate epitope residues from a position numbered antigen sequence. The reference label is constructed from the union of structurally observed antibody contact residues on the same antigen, which supports recall oriented evaluation while acknowledging that known epitopes may be incomplete (Liu et al., 2024; Dunbar et al., 2014).

Task2: Antibody Conditioned Epitope Identification. This task evaluates whether a model can associate a specific antibody sequence with its corresponding binding region on an antigen. This task reflects mechanism characterization for candidate antibodies, where the key question is not whether a region is generally epitope like, but which region is recognized by the given antibody. We formulate the task as a multiple choice problem over candidate epitopes, using the true structural epitope as the answer and constructing distractors from other real epitopes on the same antigen or matched synthetic regions. This design keeps the task automatically scorable while preserving antibody specific reasoning.

Task3: Antibody Epitope Binning. This task evaluates whether a model can determine whether two antibodies bind overlapping antigen regions. This task corresponds to antibody panel organization, where epitope redundancy and diversity affect lead selection, combination design, and downstream characterization. We formulate binning as a binary judgment over two antibodies against the

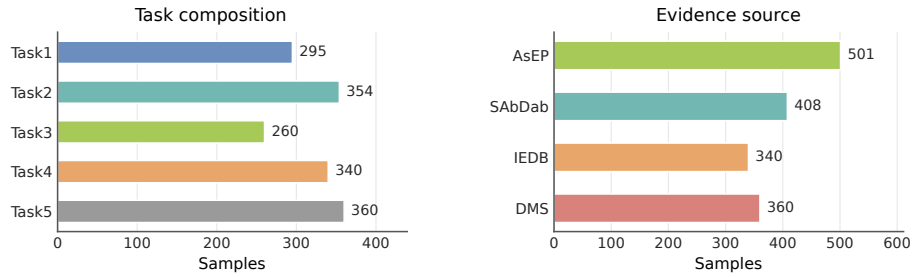


Figure 2: Dataset composition of EpiBench. The two panels report sample counts across the five benchmark tasks and four evidence sources. T1–T5 denote targetable region discovery, antibody-conditioned epitope identification, epitope binning, functional epitope assessment, and antibody escape assessment.

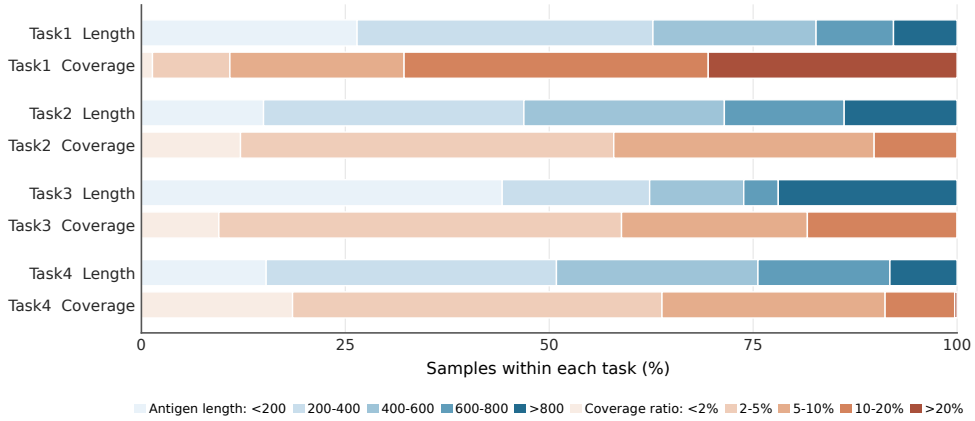


Figure 3: Antigen length and epitope coverage distributions in EpiBench. For each of Tasks 1–4, the upper stacked bar shows the antigen length distribution and the lower stacked bar shows the epitope coverage ratio distribution. Task 5 is excluded because it evaluates mutation-specific escape on a fixed antigen window and does not define an explicit epitope residue set for coverage analysis.

same antigen, with labels derived from structural epitope overlap. To reduce shortcuts based only on antibody sequence similarity, positive and negative pairs are sampled with matched antibody similarity profiles when possible.

Task4: Functional Epitope Assessment. This task evaluates whether a model can distinguish epitopes associated with functional antibody outcomes from epitopes that only show binding evidence or lack functional support. This task reflects the distinction between antigen recognition and therapeutic relevance, since not every binding site leads to neutralization, inhibition, protection, or effector mediated activity. We formulate the task as a multiple choice problem over same antigen candidate epitopes, where the answer is supported by curated functional B cell assay evidence and distractors are selected from binding only or property matched candidate regions (Vita et al., 2025). This design encourages models to reason about functional consequence rather than epitope likelihood alone.

Task5: Antibody Escape Assessment. This task evaluates whether an antigen point mutation disrupts recognition by a particular antibody. This task reflects lead optimization and resistance risk assessment, where the same mutation may have different effects across antibodies depending on their binding interfaces. We formulate escape assessment as a binary judgment from antibody variable domain sequences, a wild type antigen sequence, and a single antigen mutation. Labels are derived from deep mutational scanning measurements, and mutation level matching is used to reduce shortcuts based only on mutation identity (Cao et al., 2022). The prompt removes antigen and antibody identifiers so that the task emphasizes sequence conditioned escape reasoning.

Table 2: Overall zero-shot performance of nine general-purpose LLMs on EpiBench (0–100 scale), with random and specialized-model references shown at the bottom. Task 1 reports RegionRecall@50 (RegR@50) and AUROC, while Tasks 2–5 report accuracy. Within LLM rows, **bold** and underlined values denote the best and second-best results, respectively. N/A indicates that the corresponding reference score is not applicable or no directly matched specialized model is available.

	T1: Region Discovery		T2: Epitope Identification	T3: Epitope Binning	T4: Functional Assessment	T5: Escape Risk
	RegR@50 ↑	AUROC ↑	Acc. ↑	Acc. ↑	Acc. ↑	Acc. ↑
Reasoning Models						
Gemini-3-Flash	**56.9**	**54.6**	28.2	53.5	**35.6**	46.2
GPT-5.5	49.7	52.3	31.9	**67.3**	33.2	52.2
Qwen3.7-Plus	47.8	52.0	24.0	58.9	25.0	48.6
GLM-5.2	45.0	51.4	24.3	56.9	24.1	50.0
DeepSeek-V4-Pro	31.8	51.3	30.8	55.0	30.3	48.3
Kimi-K2.6	36.7	51.1	25.4	51.9	27.1	41.4
Nonreasoning Models						
GPT-4o	38.5	50.5	25.4	51.2	23.8	52.2
Grok 4.20	40.8	50.9	**33.6**	65.8	30.0	**54.7**
Qwen3-235B	40.5	50.3	29.4	49.2	24.4	46.7
Random baseline	*N/A*	*50.0*	*25.0*	*50.0*	*25.0*	*50.0*
Specialized model	60.4	58.0	40.2	70.8	*N/A*	*N/A*

3.3 Dataset Statistics

Figures 2 and 3 summarize the composition and sequence scale of EpiBench. The benchmark contains 1,609 samples across five tasks, with each task contributing between 15% and 25% of the full dataset. The samples are derived from four source datasets, including AsEP (Liu et al., 2024), SAbDab (Dunbar et al., 2014), DMS escape measurements (Cao et al., 2022), and IEDB (Vita et al., 2025). We report sequence length and epitope coverage statistics only for Tasks 1 to 4 because Task5 is designed for mutation specific escape assessment, does not require full length antigen input, and does not define an explicit epitope residue set for coverage analysis. Overall, EpiBench includes a representative collection of antigen sequences and task relevant evidence, enabling systematic evaluation of LLMs on epitope reasoning.

4 Experiments

4.1 Overall Model Performance

Table 2 reports the performance of nine general-purpose LLMs across the five EpiBench tasks. We select models according to two dimensions, weight accessibility and reasoning capability. The resulting set covers closed-source reasoning models (Gemini-3-Flash, GPT-5.5, and Qwen3.7-Plus), closed-source nonreasoning models (GPT-4o and Grok 4.20), open-weight reasoning models (GLM-5.2, DeepSeek-V4-Pro, and Kimi-K2.6), and an open-weight nonreasoning model (Qwen3-235B). All models are evaluated under a unified zero-shot protocol that permits explicit intermediate reasoning. Detailed model and prompt configurations are provided in Appendix C. To provide additional context, the bottom rows of Table 2 include random baselines and specialized epitope models where directly applicable. The specialized comparison uses BepiPred-3.0 for Task 1 and EpiPred-derived baselines for Tasks 2–3, while Tasks 4–5 are marked as N/A because we did not identify established sequence-based models that directly match EpiBench's functional epitope assessment and antibody-specific escape settings. Based on these evaluations, we make three observations.

(1) Current LLMs capture coarse epitope signals but remain weak at precise interface discrimination. On Task 1, the clear improvement in RegionRecall@50 shows that some models can prioritize broadly relevant antigen regions, whereas AUROC remains close to random, indicating that they do not reliably distinguish epitope from non-epitope residues across the sequence. Performance is similarly limited on Tasks 2 and 5, where success requires conditioning the prediction on a particular antibody and resolving antibody-specific binding or escape effects. These results suggest that recognizing generic antigen properties is more accessible than grounding a specific antibody–antigen interaction from sequence alone.

(2) Task formulation and biological dependency strongly shape the apparent difficulty. Task 3 shows the largest gain over its random baseline, but it asks for a binary comparison between two antibodies and may admit pairwise sequence-similarity cues without requiring explicit localization of the binding interface. By contrast, Task 4 requires connecting an epitope to downstream functional evidence, and Task 5 further requires predicting the effect of a single mutation for a particular antibody. The weaker results on these tasks reflect the increasing need to infer latent structural contacts and functional consequences that are not directly expressed in the input sequences.

(3) Neither a single model nor reasoning capability yields a uniform advantage. Gemini-3-Flash is strongest on region discovery and functional assessment, GPT-5.5 performs most consistently and leads epitope binning, while the nonreasoning Grok 4.20 leads epitope identification and escape assessment. This task-dependent ranking indicates that generic language-model reasoning does not by itself overcome the central bottlenecks of epitope understanding: long-range sequence grounding, residue-level correspondence, and antibody-specific structural inference. Overall, current LLMs provide partial competence on individual stages, but do not yet support the full epitope-centered antibody development workflow reliably.

4.2 Antigen Length Sensitivity

Therapeutic antibody discovery involves antigens with diverse sequence scales, ranging from short domains to longer multi-domain proteins. This diversity raises a practical question: can LLMs understand epitope-related information from antigen sequences of different lengths, rather than only from compact inputs? To examine this question, we stratify Tasks 1–4 by antigen sequence length and report the results in Figure 4. Task 5 is not included because it evaluates mutation-level escape on a fixed antigen window instead of variable-length full-antigen inputs, so antigen length is not a meaningful axis for that task.

Task 1 exhibits the strongest dependence on antigen length. The average RegR@50 decreases from 81.3 for antigens shorter than 200 residues to 12.8 for antigens longer than 800 residues. This task asks the model to propose a small set of epitope residues before any antibody sequence is given, which is biologically close to identifying potentially targetable surface regions on an antigen. As the antigen becomes longer, the number of non-epitope residues, structural domains, and plausible surface patches increases, while the prediction budget remains fixed. The observed decline therefore reflects the increasing difficulty of recognizing sparse epitope-like signals against a larger antigen background. It also suggests that current LLMs have limited ability to preserve fine-grained residue position information when the relevant region occupies only a small fraction of a long sequence.

Tasks 2–4 provide a complementary view of how LLMs handle antigen sequences once the question supplies additional biological structure. Their performance changes are weaker and non-monotonic across length bins, and Task 3 does not decline on longer antigens. This suggests that LLMs can sometimes retain enough sequence-level information for candidate-based or comparative decisions even when the antigen is long. However, stable performance across length bins should not be interpreted as robust epitope understanding. In Task 2, the model must connect antibody sequence features with the correct candidate epitope; in Task 3, it must compare whether two antibodies are likely to target overlapping regions; and in Task 4, it must associate a candidate epitope with functional antibody evidence. These tasks require the model to organize sequence information around antibody specificity, epitope equivalence, and functional consequence. The length-stratified results therefore suggest that current LLMs are not limited only by raw input length; their broader challenge is to convert antigen and antibody sequences into biologically grounded epitope-level decisions.

4.3 Effect of Explicit Reasoning

Explicit reasoning prompts such as chain-of-thought (CoT) are commonly used to improve LLM performance, but their benefits are not guaranteed across tasks. Prior studies show that CoT effectiveness depends on task and prompt structure, that intermediate rationales may partly act as task specification rather than a universally useful reasoning mechanism, and that models vary in how strongly they condition on generated rationales (Madaan et al., 2023; Lanham et al., 2023). Since the main EpiBench evaluation uses a unified zero-shot protocol that allows explicit intermediate reasoning, we compare it with an implicit-reasoning setting to quantify how much this reasoning format contributes when models interpret antigen and antibody sequences. In the implicit setting, models are asked to provide the final answer without generating intermediate reasoning steps.

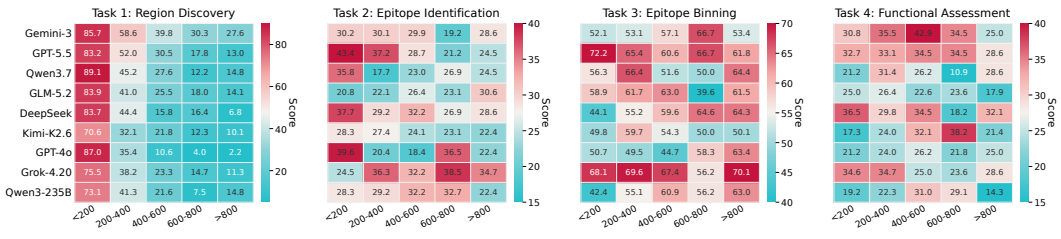


Figure 4: Performance of nine LLMs across antigen sequence length bins on EpiBench Tasks 1–4. Rows denote models and columns denote antigen length bins. Task 1 reports RegR@50, and Tasks 2–4 report accuracy. Each panel uses a task-specific color scale to emphasize within-task variation across sequence lengths.

Table 3: Comparison of model performance between CoT prompting and the direct-answer setting. Δ = Explicit thinking process − Implicit thinking process, and Avg. is the unweighted mean across tasks.

	T1: Region Discovery		T2: Epitope Identification		T3: Epitope Binning		T4: Functional Assessment		T5: Escape Risk		Avg.	
Model	**Implicit**	Δ	**Implicit**	Δ	**Implicit**	Δ	**Implicit**	Δ	**Implicit**	Δ	**Implicit**	Δ
Gemini-3-Flash	48.9	+8.0	29.7	-1.4	62.7	-9.2	31.2	+4.4	47.8	-1.6	44.0	+0.0
GPT-5.5	43.5	+6.2	28.2	+3.7	60.0	+7.3	28.5	+4.7	46.9	+5.3	41.4	+5.4
Qwen3.7-Plus	42.5	+5.3	31.1	-7.1	65.8	-6.9	25.0	+0.0	50.3	-1.7	42.9	-2.1
GLM-5.2	41.8	+3.1	26.8	-2.5	58.8	-1.9	31.5	-7.4	52.5	-2.5	42.3	-2.2
DeepSeek-V4-Pro	42.0	-10.2	25.7	+5.1	51.9	+3.1	26.2	+4.1	55.6	-7.2	40.3	-1.0
Kimi-K2.6	33.4	+3.2	27.1	-1.7	53.8	-1.9	25.3	+1.8	34.4	+6.9	34.8	+1.7
GPT-4o	21.0	+17.5	28.2	-2.8	59.2	-8.1	26.5	-2.6	53.6	-1.4	37.7	+0.5
Grok 4.20	41.3	-0.6	29.1	+4.5	60.8	+5.0	30.0	+0.0	50.0	+4.7	42.2	+2.7
Qwen3-235B	35.2	+5.3	29.1	+0.3	60.8	-11.5	22.6	+1.8	53.6	-6.9	40.3	-2.2

Table 3 shows that moving from implicit to explicit reasoning has model-dependent and task-dependent effects. GPT-5.5 is the most sensitive to this comparison, with explicit reasoning exceeding implicit reasoning by 5.4 points on average, whereas several other models show negligible or negative average differences. The clearest positive differences appear in Task 1, where most models perform better under the main explicit-reasoning protocol. This suggests that explicit reasoning can help organize the open-ended selection of sparse epitope-like regions from a full antigen sequence. However, the effect is much less consistent for Tasks 2–5. These tasks require the model to connect antibody sequences, candidate epitopes, functional annotations, or escape mutations to the underlying epitope relationship, and changing whether reasoning is expressed explicitly does not reliably change these sequence-grounded decisions. Overall, the comparison indicates that explicit reasoning can support some models and tasks, but EpiBench performance is not explained by prompting format alone; the central challenge remains biologically grounded sequence understanding.

4.4 Failure Mode Analysis

The preceding results suggest that current LLMs still have substantial room to improve in epitope-centered sequence understanding, but aggregate scores alone do not indicate which aspects of the reasoning process require improvement. We therefore analyze the reasoning traces of parseable but incorrect predictions under the main explicit-reasoning protocol. This analysis focuses on model behavior after a valid answer has been produced, separating reasoning-level failures from output-format failures.

We use a rule-based, multi-label taxonomy to identify broad reasoning patterns that appear in incorrect traces. The categories include claims of memorized biological knowledge, generic physicochemical reasoning, fixation on sequence motifs or functional sites, CDR-sequence or germline-based shortcuts, and explicit fallback behavior such as guessing or choosing the safest option. Because one trace can contain several of these patterns, Figure 5 reports each task as a normalized distribution over pattern occurrences rather than mutually exclusive error classes.

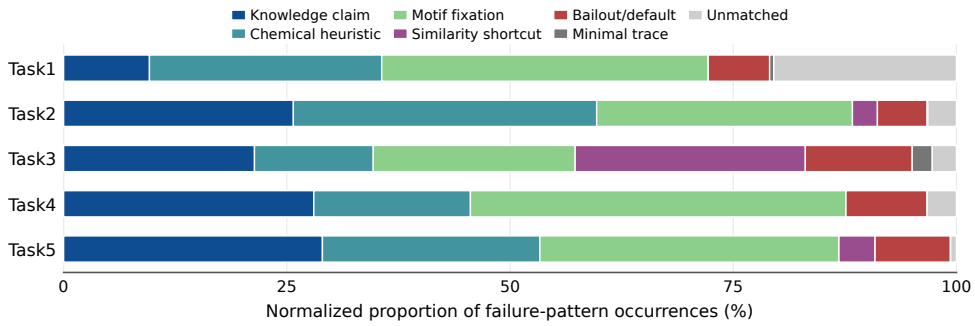


Figure 5: Failure-pattern composition across EpiBench tasks. We analyze parseable but incorrect CoT predictions from models with accessible reasoning traces and tag each trace with broad reasoning-pattern indicators. Because one trace may contain multiple patterns, each bar reports the normalized proportion of pattern occurrences within a task; unmatched denotes incorrect traces that do not trigger any predefined pattern.

The resulting patterns suggest that many incorrect answers are supported by plausible but weakly grounded biological rationales. Motif fixation is prominent in Tasks 1, 4, and 5, where models often emphasize recognizable sequence or functional-site cues even when these cues are insufficient to determine the correct epitope region, functional epitope, or escape effect. Knowledge claims and physicochemical heuristics are also frequent, especially in antibody-conditioned tasks, indicating that models often appeal to remembered antibody or antigen associations and generic residue-level interaction principles. Task 3 shows a distinct pattern: similarity shortcuts contribute a large fraction of failures, consistent with the task requiring epitope-bin equivalence rather than simple CDR-sequence or germline-based similarity. These observations do not imply that the detected patterns are the only causes of failure, but they show that incorrect LLM reasoning often relies on indirect biological proxies instead of consistently grounding predictions in antibody-specific epitope relationships.

5 Discussion

EpiBench occupies a distinct evaluation niche for biomedical LLMs: it asks whether general-purpose models can use antigen and antibody sequences to support epitope-centered decisions in antibody drug discovery. Unlike single-task epitope prediction datasets or broad protein benchmarks, EpiBench connects epitope localization, antibody-conditioned recognition, epitope binning, functional assessment, and escape assessment within a unified closed-book and automatically scorable framework. Our experiments show that current LLMs capture partial epitope-related signals, but still lag behind task-specific sequence models where such comparisons are available, show limited robustness across long antigen contexts, benefit inconsistently from explicit reasoning, and often rely on indirect biological proxies in incorrect reasoning traces. These findings position EpiBench not merely as a leaderboard, but as a diagnostic benchmark for identifying which aspects of antibody–antigen sequence understanding need improvement. By grounding evaluation in experimentally supported epitope evidence while preserving an LLM-ready formulation, EpiBench provides a practical testbed for developing and comparing sequence-aware biomedical LLMs, and for measuring progress toward reliable LLM-assisted antibody discovery.

AI use statement

In this work, we used generative AI tools to help identify related literature and to improve the grammar, clarity, and readability of the manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Mansoor Ahmed, Nadeem Taj, Imdad Ullah Khan, Hemanth Venkateswara, and Murray Patterson. CHIMERA-bench: A benchmark dataset for epitope-specific antibody design. In *ICLR 2026*

Workshop on Generative and Experimental Perspectives for Biomolecular Design, 2026. URL https://openreview.net/forum?id=PyZvVIJbSy.

Hifzur Rahman Ansari and Gajendra P. S. Raghava. Identification of conformational B-cell epitopes in an antigen from its primary sequence. *Immunome Research*, 6:6, 2010. doi: 10.1186/1745-7580-6-6.

Yunlong Cao, Ayijiang Yisimayi, Fanchong Jian, Weiliang Song, Tianhe Xiao, Lei Wang, Shuo Du, Jing Wang, Qianqian Li, Xiaosu Chen, et al. BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection. *Nature*, 608:593–602, 2022. doi: 10.1038/s41586-022-04980-y.

Henriette Capel, Robin Weiler, Maurits Dijkstra, Reinier Vleugels, Peter Bloem, and K. Anton Feenstra. ProteinGLUE multi-task benchmark suite for self-supervised protein modeling. *Scientific Reports*, 12:16047, 2022. doi: 10.1038/s41598-022-19608-4.

Michael Chungyoun, Jeffrey Ruffolo, and Jeffrey J. Gray. FLAb: Benchmarking deep learning methods for antibody fitness prediction. *bioRxiv*, 2024. doi: 10.1101/2024.01.13.575504.

Joakim Nøddeskov Clifford, Magnus Haraldson Høie, Sebastian Deleuran, Bjoern Peters, Morten Nielsen, and Paolo Marcatili. Bepipred-3.0: Improved B-cell epitope prediction using protein language models. *Protein Science*, 31(12):e4497, 2022. doi: 10.1002/pro.4497.

Christian Dallago, Jody Mou, Kadina E. Johnston, Bruce J. Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K. Yang. FLIP: Benchmark tasks in fitness landscape inference for proteins. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Kieran Didi, Sarah Alamdari, Alex X. Lu, Bruce Wittmann, Kadina E. Johnston, Ava P. Amini, Ali Madani, Maya Czeneszew, Christian Dallago, and Kevin K. Yang. FLIP2: Expanding protein fitness landscape benchmarks for real-world machine learning applications. *Proceedings of the International Conference on Machine Learning*, 2026.

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M. Deane. SAbDab: the structural antibody database. *Nucleic Acids Research*, 42(D1):D1140–D1146, 2014. doi: 10.1093/nar/gkt1043.

Yasser El-Manzalawy, Drena Dobbs, and Vasant Honavar. Predicting linear B-cell epitopes using string kernels. *Journal of Molecular Recognition*, 21(4):243–255, 2008. doi: 10.1002/jmr.893.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

E. A. Emini, J. V. Hughes, D. S. Perlow, and J. Boger. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *Journal of Virology*, 55(3):836–839, 1985. doi: 10.1128/jvi.55.3.836-839.1985.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-Instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*, 2023. doi: 10.48550/arXiv.2306.08018.

Pernille Haste Andersen, Morten Nielsen, and Ole Lund. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Science*, 15(11):2558–2567, 2006. doi: 10.1110/ps.062405906.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Artificial intelligence foundation for therapeutic science. *Nature Chemical Biology*, 18(10):1033–1036, 2022. doi: 10.1038/s41589-022-01131-2.

A. S. Kolaskar and P. C. Tongaonkar. A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Letters*, 276(1-2):172–174, 1990. doi: 10.1016/0014-5793(90)80535-Q.

Tamera Lanham et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. doi: 10.48550/arXiv.2307.13702.

Jens Erik Pontoppidan Larsen, Ole Lund, and Morten Nielsen. Improved method for predicting linear B-cell epitopes. *Immunome Research*, 2:2, 2006. doi: 10.1186/1745-7580-2-2.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Chunan Liu, Lilian Denzler, Yihong Chen, Andrew Martin, and Brooks Paige. AsEP: Benchmarking deep learning methods for antibody-specific epitope prediction. *arXiv preprint arXiv:2407.18184*, 2024. doi: 10.48550/arXiv.2407.18184.

Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. What makes chain-of-thought prompting effective? a counterfactual study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1448–1535. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.101.

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41:1099–1106, 2023. doi: 10.1038/s41587-022-01618-2.

Manuel F. Mollon, Joaquin Gonzalez-Rodriguez, Alicia Lozano-Diez, Daniel Ramos, and Doroteo T. Toledano. Exploring large protein language models in constrained evaluation scenarios within the FLIP benchmark. *arXiv preprint arXiv:2501.18223*, 2025. doi: 10.48550/arXiv.2501.18223.

Erik Nijkamp, Jeffrey Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the boundaries of protein language models. *arXiv preprint arXiv:2206.13517*, 2022. doi: 10.48550/arXiv.2206.13517.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan Gomez, Debora S. Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. *arXiv preprint arXiv:2205.13760*, 2022. doi: 10.48550/arXiv.2205.13760.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora Marks. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64331–64379, 2023.

Julia Ponomarenko, Huynh-Hoa Bui, Wei Li, Nicholas Fusseder, Philip E. Bourne, Alessandro Sette, and Bjoern Peters. ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics*, 9:514, 2008. doi: 10.1186/1471-2105-9-514.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems*, 2019.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118.

Sudipto Saha and Gajendra P. S. Raghava. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins: Structure, Function, and Bioinformatics*, 65(1):40–48, 2006. doi: 10.1002/prot.21078.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023. doi: 10.1038/s41586-023-06291-2.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017. doi: 10.1038/nbt.3988.

Jing Sun et al. SEPPA: a computational server for spatial epitope prediction of protein antigens. *Nucleic Acids Research*, 37(Web Server issue):W612–W616, 2009. doi: 10.1093/nar/gkp417.

Yang Tan, Mingchen Li, Ziyi Zhou, Pan Tan, Huiqun Yu, Guisheng Fan, and Liang Hong. PETA: evaluating the impact of protein transfer learning with sub-word tokenization on downstream applications. *Journal of Cheminformatics*, 16:92, 2024. doi: 10.1186/s13321-024-00884-3.

Raphael J. L. Townshend, Martin Vögele, Patricia Suriana, Alexander Derry, Alexander Powers, Yianni Laloudakis, Sidhika Balachandar, Brandon Anderson, Stephan Eismann, Risi Kondor, Russ B. Altman, and Ron O. Dror. ATOM3D: Tasks on molecules in three dimensions. *arXiv preprint arXiv:2012.04035*, 2020.

Lood van Niekerk, Joshua Moller, Seth Ritter, Porfirio Quintero-Cadena, Rich Cohen, Georgia Channing, Michael Chungyoun, Laura Rand, Alexander Smith, Aanal Bhatt, et al. 2025 Ginkgo Datapoints antibody developability competition outcomes: limited model performance and a call for data standardization. *mAbs*, 18(1):2634216, 2026. doi: 10.1080/19420862.2026.2634216.

Randi Vita, Nina Blazeska, Daniel Marrama, IEDB Curation Team Members, Sebastian Duesing, Jason Bennett, Jason Greenbaum, Marcus De Almeida Mendes, Jarjapu Mahita, Daniel K. Wheeler, Jason R. Cantrell, James A. Overton, Darren A. Natale, Alessandro Sette, and Bjoern Peters. The immune epitope database (IEDB): 2024 update. *Nucleic Acids Research*, 53(D1):D436–D443, 2025. doi: 10.1093/nar/gkae1092.

Katharina Waury, Stefan Lelieveld, Sanne Abeln, and Henk-Jan van den Ham. Comparison of sequence- and structure-based antibody clustering approaches on simulated repertoire sequencing data. *PLOS Computational Biology*, 21:e1013057, 2025. doi: 10.1371/journal.pcbi.1013057.

Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Chang Ma, Runcheng Liu, and Jian Tang. PEER: A comprehensive and multi-task benchmark for protein sequence understanding. *arXiv preprint arXiv:2206.02096*, 2022. doi: 10.48550/arXiv.2206.02096.

Xinyan Zhao, Yi-Ching Tang, Akshita Singh, Victor J. Cantu, KwanHo An, Junseok Lee, Adam E. Stogsdill, Ashwin Kumar Ramesh, Zhiqiang An, Xiaoqian Jiang, and Yejin Kim. Benchmark for antibody binding affinity maturation and design. *arXiv preprint arXiv:2506.04235*, 2025.

Le Zhuo, Zewen Chi, Minghao Xu, Heyan Huang, Heqi Zheng, Conghui He, Xian-Ling Mao, and Wentao Zhang. ProtLLM: An interleaved protein-language LLM with protein-as-word pre-training. *arXiv preprint arXiv:2403.07920*, 2024. doi: 10.48550/arXiv.2403.07920.

A Biological Concepts and Task Formalization

A.1 Epitope-related Biological Concepts

EpiBench evaluates sequence based reasoning about antigen–antibody recognition, so we first clarify the biological concepts used throughout the benchmark. An antigen is the molecular target recognized by an antibody, and an antibody binds an antigen through its variable domains, especially the complementarity determining regions (CDRs). The antigen residues recognized by an antibody are called the epitope, while the antibody residues involved in recognition are often referred to as the paratope. In structural settings, an epitope can be represented as a set of antigen residues that are spatially close to antibody CDR residues; in sequence based assay settings, it can also appear as a linear peptide segment with experimentally measured binding or functional evidence.

Epitopes are central to several downstream decisions in antibody discovery. A targetable antigen region denotes an antigen region repeatedly or plausibly engaged by antibodies and is useful for early antigen characterization. Antibody conditioned epitope identification asks which region is recognized by a particular antibody, rather than which region is generally targetable. Epitope binning groups antibodies according to whether they bind the same or strongly overlapping antigen regions, which supports panel selection and mechanism comparison. A functional epitope is an epitope

associated with a downstream biological effect, such as inhibition, protection, or cytotoxicity, rather than binding evidence alone. Finally, an escape mutation is an antigen mutation that reduces or disrupts recognition by a given antibody. These concepts motivate the five tasks in EpiBench.

A.2 Formal Definition of EpiBench Tasks

EpiBench converts the above biological concepts into five automatically scorable sequence based tasks. Each task removes direct identifiers and provides only the sequence level evidence needed for the corresponding decision. Table 4 summarizes the biological question, model input, and expected output for each task.

Table 4: Formal definition of the five EpiBench tasks. Each task represents one epitope centered decision in antibody drug discovery and is formulated with sequence based inputs and automatically scorable outputs.

Task	**Biological question**	**Input**	**Output**
Task1	Which antigen residues form targetable regions?	Antigen sequence	Epitope residue indices
Task2	Which candidate epitope is bound by a given antibody?	Antigen sequence, antibody sequences, and candidate epitope regions	One candidate epitope
Task3	Do two antibodies belong to the same epitope bin?	Antigen sequence and two antibody sequence pairs	Same or different bin
Task4	Which candidate epitope is associated with the specified functional effect?	Antigen sequence, candidate epitope regions, and functional context	One functional epitope
Task5	Does an antigen mutation cause escape from a given antibody?	Antibody sequences, wild type antigen region, and mutation	Escape or non escape

The tasks differ in output format because they correspond to different experimental decisions. Task1 requires residue level localization over an antigen sequence. Tasks 2 and 4 are formulated as multiple choice decisions because the biological setting often involves selecting a binding or functional region from candidate epitopes. Tasks 3 and 5 are formulated as binary judgments because epitope binning and escape assessment naturally ask whether a relation holds for a specific antibody pair or antibody–mutation pair. This formulation keeps the benchmark close to practical antibody discovery questions while allowing deterministic scoring.

B Dataset Construction and Quality Control

B.1 Source Datasets

EpiBench is constructed from experimentally grounded resources that provide complementary evidence for epitope centered antibody discovery tasks. Structural antibody–antigen complexes provide residue level contact labels for Tasks 1 to 3, curated B cell assay records provide functional epitope evidence for Task4, and deep mutational scanning measurements provide mutation specific escape labels for Task5. Table 5 summarizes the source datasets and their roles in EpiBench.

Table 5: Source datasets used to construct EpiBench. Each source contributes a distinct experimental evidence type for one or more benchmark tasks.

Source	Evidence type	Tasks	Role in EpiBench
AsEP (Liu et al., 2024)	Curated antibody–antigen structural contacts	Tasks 1–3	Provides antibody-specific epitope footprints and contact based residue labels
SAbDab (Dunbar et al., 2014)	Antibody–antigen complex structures	Tasks 1–3	Expands the structural record pool beyond curated AsEP examples
IEDB (Vita et al., 2025)	Curated B cell epitope assay records	Task4	Provides functional and binding-only epitope evidence for same-antigen candidate construction
UniProt-linked antigen sequences	Parent protein sequences for linear peptides	Task4	Maps IEDB peptide epitopes back to antigen sequence coordinates
DMS escape measurements (Cao et al., 2022)	Antibody-specific mutation escape scores	Task5	Provides escape and non-escape labels for antibody–mutation pairs

B.2 Task-specific Construction Procedures

B.2.1 Task1: Targetable Region Discovery

Task1 is built from the shared structural record pool used by Tasks 1 to 3, containing 3346 antibody antigen complexes from curated AsEP records and newly processed SAbDab structures (Liu et al., 2024; Dunbar et al., 2014). Following the contact based definition used in AsEP, an epitope residue is defined as an antigen surface residue whose heavy atoms are within 4.5 Å of antibody CDR heavy atoms (Liu et al., 2024). We retain standard protein antigens with valid mapped epitopes, cluster antigen sequences with MMseqs2 at 95% identity and 0.8 coverage (Steinegger & Söding, 2017), merge domain and full length variants using a 0.9 identity and 0.8 coverage search criterion, and cluster antibodies at 95% identity using concatenated heavy and light chain variable domains. For each antigen group, the longest antigen is used as the reference, one representative complex is selected for each distinct antibody, and all representative footprints are mapped to the reference sequence. The Task1 label is the union of mapped footprints, with patch labels obtained by grouping overlapping footprints, while antigen groups with fewer than two distinct antibodies or excessive surface coverage are removed. The final set contains 295 antigens, and Algorithm 1 summarizes the shared structural construction procedure.

Algorithm 1 Targetable Region Discovery Dataset Construction

Require: Structural records $\mathcal{R}$, where each record r contains antigen A_r, antibody $B_r = H_r \| L_r$, epitope E_r, and antigen surface S_r.
Ensure: Task1 set $\mathcal{D}_1 = \{(A_{\text{ref}}, U, P)\}$ with union epitope U and patch labels P.

$\mathcal{R}' \leftarrow \text{Filter}(\mathcal{R}; |A_r| \geq 50, E_r \neq \emptyset)$.
$\mathcal{G} \leftarrow \text{ClusterAntigens}(\mathcal{R}'; (0.95, 0.8), (0.9, 0.8))$.
$\mathcal{D}_1 \leftarrow \emptyset$.
for antigen group $G \in \mathcal{G}$ **do**
  $A_{\text{ref}} \leftarrow \text{LongestAntigen}(G)$.
  $\mathcal{C}_{\text{ab}} \leftarrow \text{ClusterAntibodies}(G; B_r, 0.95)$.
  if $|\mathcal{C}_{\text{ab}}| < 2$ **then continue**.
  $Q \leftarrow \{\text{BestFootprint}(C) : C \in \mathcal{C}_{\text{ab}}\}$.
  $(\widetilde{E}_r, \widetilde{S}_r) \leftarrow \text{MapToRef}(E_r, S_r, A_{\text{ref}})$ for $r \in Q$.
  $U \leftarrow \bigcup_{r \in Q} \widetilde{E}_r$, $S \leftarrow \bigcup_{r \in Q} \widetilde{S}_r$.
  if $|U|/|S| > 0.5$ **then continue**.
  $P \leftarrow \text{BuildPatches}(\{\widetilde{E}_r : r \in Q\}; \text{OvCoeff} \geq 0.5)$.
  $\mathcal{D}_1 \leftarrow \mathcal{D}_1 \cup \{(A_{\text{ref}}, U, P)\}$.
end for
return $\mathcal{D}_1$

B.2.2 TASK2: ANTIBODY-CONDITIONED EPITOPE IDENTIFICATION

Task2 reuses the structural records and reference mapped antibody footprints from Task1, but keeps each representative antibody footprint as an antibody specific label rather than collapsing footprints into an antigen level union. Footprints with fewer than three residues are removed, and overlapping footprints on the same antigen are grouped into candidate regions using an overlap coefficient threshold of 0.5. We construct multiple choice items in two tiers: Tier A uses antigen groups with at least four real candidate regions and samples distractors from other real regions on the same antigen, while Tier B uses the largest valid footprint as the answer and samples non overlapping synthetic distractors matched by size, span, and segmented shape. Candidate orders are shuffled and answer letters are balanced, yielding 354 examples with 254 real distractor items and 100 synthetic distractor items.

Table 6: De-identification and shortcut-control strategies used during EpiBench construction.

Control strategy	**Applied tasks**	**Purpose**
Remove PDB IDs, antibody names, antigen names, accession numbers, variant names, and publication metadata	Tasks 1–5	Reduce direct lookup from prompt identifiers or source-record metadata
Use sequence-level prompts with position-numbered antigens, antibody VH/VL sequences, CDRs, candidate residues, or mutations as needed	Tasks 1–5	Keep only the evidence required for the target biological decision
Cluster antigen sequences and merge near-duplicate antigen records during structural task construction	Tasks 1–3	Reduce redundancy from closely related structural records
Cluster antibodies or remove near-duplicate antibody pairs using variable-domain or CDR similarity	Tasks 1–3	Reduce repeated antibody-derived shortcuts
Randomize and balance candidate option orders	Tasks 2, 4	Reduce answer-position bias in multiple-choice questions
Match same-bin and different-bin antibody pairs within CDR-similarity strata	Task3	Reduce shortcuts based only on antibody sequence similarity
Match escape and non-escape examples at the mutation level	Task5	Reduce shortcuts based only on mutation identity
Balance binary labels and multiple-choice answer letters where applicable	Tasks 2–5	Reduce class-prior and option-prior biases

B.2.3 TASK3: EPITOPE BINNING

Task3 is constructed by grouping structural records with exactly identical antigen sequences and comparing antibody epitope residue sets in the same coordinate system. For each antibody pair, we compute the epitope Jaccard index $J = |E_a \cap E_b|/|E_a \cup E_b|$ and label pairs as same bin when $J \geq 0.75$ following the structure based binning criterion in Waury et al. (2025); pairs with no shared residues are labeled as different bin, and partially overlapping pairs are discarded. Near duplicate antibodies are removed by CDR sequence similarity, and candidate pairs are sampled within CDR similarity quantile bins so that same and different pairs have matched antibody similarity profiles. The released set contains 260 balanced pairs, with 130 same bin and 130 different bin examples, and is stratified by antigen length tertiles.

B.2.4 TASK4: FUNCTIONAL EPITOPE ASSESSMENT

Task4 is constructed from IEDB B cell assay records with linear peptide epitopes and UniProt linked parent antigens (Vita et al., 2025). We retain peptides of length 6 to 40 that contain only standard amino acids, occur uniquely in the antigen sequence, and belong to antigens no longer than 1200 residues. Assay names are mapped into functional mechanisms, including inhibition, protection, and cytotoxicity, and each epitope is assigned its strongest evidence type among functional positive, functional negative, and binding only. Each multiple choice item uses a functional positive epitope as the answer and selects distractors from same antigen functional negative or binding only epitopes when available, otherwise using non overlapping synthetic windows matched by length and sequence properties. Antigens are clustered to limit redundancy, answer letters are balanced, and the final set contains 340 items with mechanism categories inhibition, protection, and cytotoxicity.

B.2.5 TASK5: ESCAPE ASSESSMENT

Task5 is built from deep mutational scanning escape measurements for SARS CoV 2 RBD antibodies (Cao et al., 2022). For each antibody, raw escape scores are normalized by the maximum escape value observed for that antibody, and mutations are labeled as escape when the normalized score is at least 0.50 and non escape when it is at most 0.05, with intermediate cases discarded. To reduce shortcuts from mutation identity, we retain only mutations that have both escape and non escape antibody examples and sample one example from each class for the same mutation. The antigen input is rendered as the RBD window with local one based numbering, while native identifiers, antibody names, antigen names, and variant names are removed from prompts. The final set contains 360 balanced examples across five wild type residue classes, with 180 escape and 180 non escape labels.

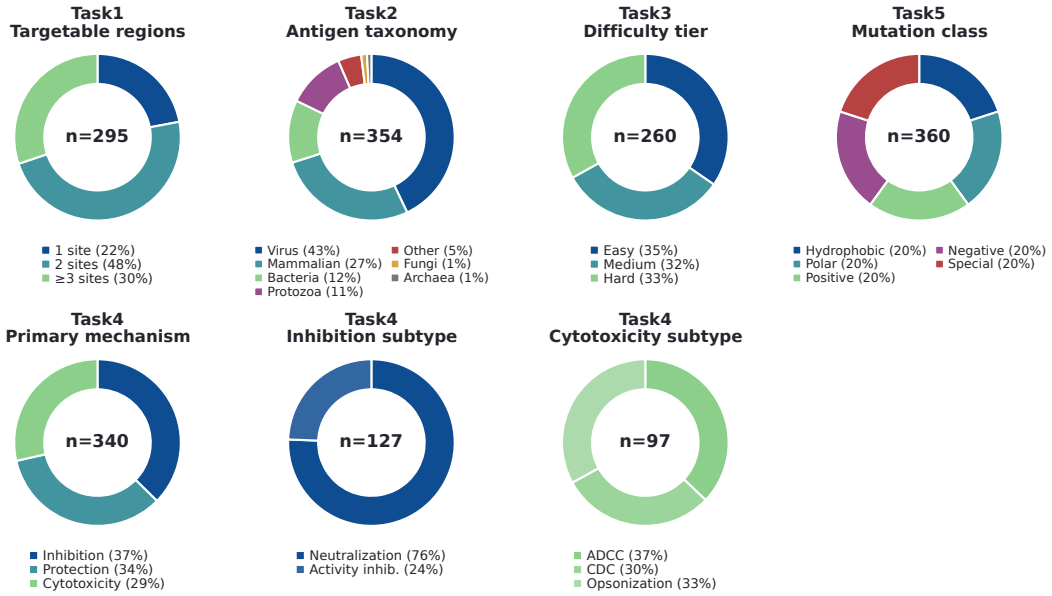


Figure 6: Fine grained subcategory distributions in EpiBench. Each task is assigned a primary subcategory for stratified evaluation. Task4 also includes secondary mechanism distributions for inhibition and cytotoxicity, while protection is retained at the primary mechanism level.

B.3 De-identification and Leakage Control

EpiBench is designed as a closed book, sequence based benchmark, so the construction process removes direct identifiers and reduces shortcut signals that could make the tasks solvable without epitope reasoning. We do not claim that any benchmark can fully eliminate pretraining memorization of public biological sequences. Instead, our controls target prompt level leakage and distributional shortcuts: the rendered prompts exclude record identifiers and metadata, while task specific sampling reduces answer position, class prior, antibody similarity, and mutation identity biases. Table 6 summarizes these controls.

B.4 Dataset Statistics and Subcategory Definitions

To support stratified evaluation, each EpiBench task is assigned one primary subcategory label along an objective and task relevant axis. Task1 is stratified by the number of targetable antigen regions, computed from overlap based epitope patches. Task2 is stratified by antigen taxonomic category, which captures source organism diversity. Task3 is stratified by difficulty tier using antigen sequence length tertiles, reflecting the larger search space induced by longer antigens. Task4 is stratified by the primary functional mechanism of the correct epitope, including inhibition, protection, and cytotoxicity. Task5 is stratified by the physicochemical class of the wild type residue being mutated. These subcategory labels are used for reporting model behavior across biological and difficulty conditions, and are not included in the model prompts or answer labels.

Our construction increases subcategory diversity where possible, but does not force strict uniformity across all tasks because some categories are rare after experimental evidence filtering and quality control. Task5 is intentionally balanced across the five mutation residue classes, whereas Tasks 1 to 4 retain the natural distribution of high confidence samples after filtering. This design preserves data reliability while enabling post hoc analysis of whether model performance is concentrated in majority categories. For Task4, we additionally record secondary mechanisms for inhibition and cytotoxicity, while protection is kept at the primary mechanism level because no reliable secondary mechanism partition is used in this benchmark. The resulting fine grained distributions are shown in Figure 6.

B.5 Answer Distribution and Choice Randomization

For tasks with discrete answer formats, we balance answer distributions to reduce position and class-prior shortcuts. Tasks 2 and 4 randomize candidate order and maintain nearly uniform A/B/C/D answer proportions, while Tasks 3 and 5 contain balanced binary labels; Task1 is excluded because it is a residue-level prediction task without fixed answer options. The resulting distributions are shown in Figure 7.

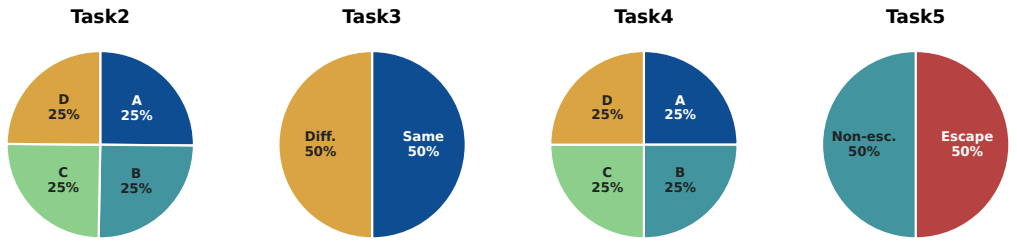


Figure 7: Answer distributions for EpiBench tasks with discrete outputs. Tasks 2 and 4 use four-way multiple-choice answers, while Tasks 3 and 5 use binary labels. Task1 is excluded because it is a residue-level prediction task without fixed answer options.

C Prompt Templates and Qualitative Examples

C.1 Prompt Templates

EpiBench uses a consistent prompt interface across tasks: a task-specific biological role, a closed-book sequence input, a concise task question, and a deterministic JSON answer format. The explicit-reasoning setting used in the main experiments asks models to expose intermediate reasoning before the final JSON line; the direct-answer setting in Section 4.3 keeps the same task definition and

output schema but replaces the reasoning instruction with "Respond with ONLY a JSON object and no other text."

System Prompt with Explicit Reasoning (CoT)

Role: You are an expert in epitope-centered antibody drug discovery. Your task is to answer the question by analyzing only the provided antigen and antibody sequence information.

Question: $question$

Sequence evidence: $sequence$

Choices: $choices$

Instruction: Think step-by-step using biological sequence evidence. Do not copy or restate long sequences in the response; quote only short substrings if needed.

Output format: End with a JSON object and no extra text after it: {"answer": "$answer$"}.

System Prompt without Explicit Reasoning (Non-CoT)

Role: You are an expert in epitope-centered antibody drug discovery. Your task is to answer the question by analyzing only the provided antigen and antibody sequence information.

Question: $question$

Sequence evidence: $sequence$

Choices: $choices$

Instruction: Provide the final answer directly. Do not include intermediate reasoning or explanatory text.

Output format: Output only a JSON object: {"answer": "$answer$"}.

C.2 Task-specific Prompt Examples

The examples below are shortened excerpts derived from the rendered EpiBench CoT user prompts after de-identification. Except for truncating long antigen or antibody sequences and omitting residue-number columns for readability, the field names, candidate lists, and question wording follow the prompts used in evaluation.

Example: Targetable Region Discovery

Antigen (1310 residues; numbering omitted):

`MFVFLVLLPLVSSQCVNLTTRTQLPPAYTN...SHPQFEK`

Question: Predict AT MOST 50 residue POSITIONS on this antigen most likely to be part of an antibody epitope (1-based positions, ranked most-likely first; no more than 50). Reason briefly, then give your final answer in the required JSON format with all positions in the array.

Example: Antibody-conditioned Epitope Identification

Antibody heavy chain (VH, variable domain):

`QVQLVQSGAEVKKPGSSVKVSCKASGGTFSSYP...CARGIASAGTPDYFFY`

CDR1: GGTFSSYP CDR2: IIPLFGTA CDR3: ARGIASAGTPDYFFY

Antibody light chain (VL, variable domain):

`EIVLTQSPGTLSLSPGERATLSCRASQSISSSY...CQQYGSSPYT`

CDR1: QSISSSY CDR2: GAS CDR3: QQYGSSPYT

Antigen (894 residues; numbering omitted):

`TQVCTGTDMKLRLPASPETH...KIEWHESRHHHHH`

Candidate epitopes (each token = residue+position on the antigen):
(A) P557, E558, D560, Q561, K569, D570, P571, P572, F573, P579, Y588, M589, I591, K593, D596, A600, Q602, P603, C604, P605
(B) E373, Q376, E379, D395, S396, L397, P398, D399, S401, V402, Q404, G427, S429, W430, R434, T450, H451, R477, E481, V483, E485, G486, L487, P501, G502, T504, Q505
(C) L654, G655, G656, P657, S686, H687, E688, D689, P690, E691, K741, N744, K745, A746, L747, P748, A749, P750
(D) T669, L670, M671, I672, S673, H729, Q730, L733, E801, G804, V841, S843, M847, A850, L851, H852, N853, H854, Y855, T856, Q857, K858, S859, L860, S861

Question: Which candidate epitope does THIS antibody bind? Reason step-by-step, then give your final answer in the required JSON format.

Example: Epitope Binning

Antigen (365 residues; numbering omitted):

`DNLWVTVYYGVPVWKDADTT...NWRSELYKYKVVQIE`

Antibody A — heavy chain (H):
`HVQLVQSGGGVKKIGAAVRISCEVSGYNFMDQF...ARGPSGENYPFHY`

CDR1: GYNFMDQF CDR2: MNPIYGQV CDR3: ARGPSGENYPFHY

Antibody A — light chain (L): `LTQPASMSASPGQSVTISCSGTRHIIS...YICNTYEF`

CDR1: RHIIS CDR2: DDD CDR3: NTYEF

Antibody B — heavy chain (H):
`QVQLVQSGGQMKKPGESMRISCRASGYEFIDCT...TRGKNCDYNWDFEH`

CDR1: GYEFIDCT CDR2: LKPRGGAV CDR3: TRGKNCDYNWDFEH

Antibody B — light chain (L): `VLTQSPGTLSLSPGETAIISCRTSQYGS...CQQYEF`

CDR1: QYGS CDR2: SGS CDR3: QQYEF

Question: Do antibody A and antibody B bind overlapping epitope regions on this antigen? Reason step-by-step, then give your final answer in the required JSON format.

Example: Functional Epitope Assessment

Antigen sequence (824 residues; numbering omitted):

`MEFIPTQTFYNRRYQPRPWT...RTRNSDPEHGGSTV`

Desired antibody-related functional outcome: a functional antibody response (neutralization, inhibition of activity, protection, or antibody-mediated cytotoxicity)

Candidate B-cell epitopes (peptides from this antigen):
A. residues 567-584: VPRNAELGDRKGKIHIPF
B. residues 326-335: STKDNFNVYK
C. residues 12-21: RRYQPRPWTP
D. residues 300-309: EDNVMRPGYY

Question: Which candidate epitope is experimentally associated with the desired functional antibody response? Reason step-by-step, then give your final answer in the required JSON format.

Example: Escape Assessment

Antibody heavy chain (VH, variable domain):

`EVQLVQSGAEVKKPGASVKVSCKASGYTFTGYY...CARAAPFYDFWSGYSYFDY`
CDR1: GYTFTGYY CDR2: INPISSGT CDR3: ARAAPFYDFWSGYSYFDY

Antibody light chain (VL, variable domain):

`EIVMMQSPGTLSLSPGERATLSCRASQSVSSSY...CQQYGSSAWT`
CDR1: QSVSSSY CDR2: GAS CDR3: QQYGSSAWT

Wild-type antigen (201 residues; numbering omitted):

`NITNLCPFGEVFNATRFASV...PLQSYGFQPTNGVGYQPYRV...PATVCGPKKST`

Mutation:
position 176: Q -¿ A

Question: Will this antigen mutation cause the antibody to lose binding? Reason step-by-step, then give your final answer in the required JSON format.

C.3 Representative Error Case

The case below shows a real incorrect Task5 response from DeepSeek-V4-Pro. We do not repeat the task prompt; instead, we show the stored model trace in layers, with only long sequence-counting spans shortened by ellipses. The source record contains a local RBD mutation at position 154 that corresponds to full-spike E484H, and the hidden gold label is `escape`.

Case Study: Coordinate Misalignment in Escape Assessment

Sample: Task5, `task5_0161` **Model**: DeepSeek-V4-Pro **Parsed answer**: `non_escape`
Gold label: `escape`

Layer 1: Antigen recognition. *Model output*: "The antigen is presumably the SARS-CoV-2 spike protein receptor-binding domain (RBD) based on the sequence? Let's check: The sequence provided is 201 residues, starting with `NITNLCPFGE...`. This matches the SARS-CoV-2 RBD sequence."

Summary: The model correctly recognizes the biological context as a SARS-CoV-2 RBD-derived sequence.

Layer 2: Local-to-global coordinate mapping. *Model output*: "Position 154 is indeed E (glutamic acid). Mutation E154H. ... **position 154 in this numbering corresponds to full spike position 472**? Actually: 1 -¿ 319, 2 -¿ 320, ... so n -¿ 318+n. So **154 -¿ 318+154 = 472**."

Summary: The failure begins when the model assumes an incorrect RBD offset and converts the local mutation into **`E472H`** rather than the source-record `E484H`.

Layer 3: Speculative antibody identification. *Model output*: "The heavy chain CDR3 sequence `ARDAGKSPLSYWYFDL` is quite distinctive. ... Another possibility: This might be the antibody 'C105' or 'C144'? ... I think the antibody 'C135' (or 'C144') might have a similar CDR3."

Summary: The model then uses memorized antibody associations as supporting evidence, although the evaluation input is de-identified and the source metadata does not provide these names to the model.

Layer 4: Escape decision from the wrong site. *Model output*: "I recall the C144 escape map. ... The escape mutations are at E484, F490, and maybe S494. **I don't think E472 is a hit**. ... Therefore, the answer should be **`non_escape`**."

Summary: The model recalls that E484 can be escape-associated, but because it has already mapped the mutation to E472, it reaches the opposite decision.

Evaluation: The output is parsed deterministically as {`"answer": "non_escape"`} and compared against the gold label {`"answer": "escape"`}.

Error cause: An incorrect local-to-global coordinate mapping leads the model to evaluate `E472H` instead of the true `E484H` mutation.

This example illustrates why EpiBench evaluates sequence-grounded answers rather than only the plausibility of generated explanations. The response contains domain-relevant biological knowledge, but the final answer is determined by a concrete sequence-coordinate error.

D SUPPLEMENTARY EXPERIMENTAL ANALYSES

D.1 PERFORMANCE ACROSS TASK SUBCATEGORIES

EpiBench contains controlled heterogeneity within each task, including target-region size, antigen taxonomy, binning difficulty, functional mechanism, and mutation chemistry. We therefore analyze whether the overall results are driven by a narrow subset of examples or persist across the main task-defined subcategories. Figure 8 reports model performance across these subcategories while keeping the model order fixed.

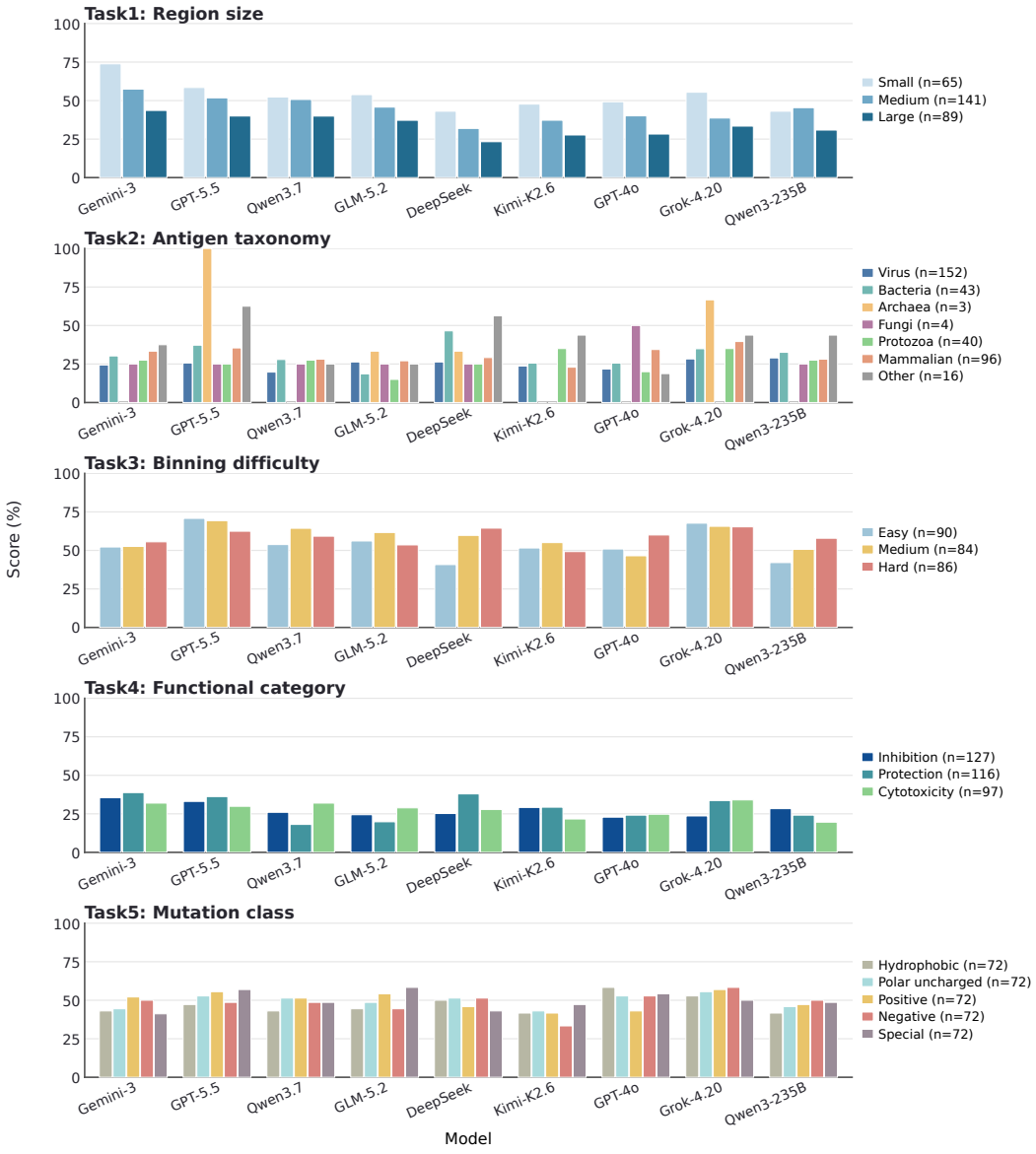


Figure 8: Subcategory-wise model performance across EpiBench. Each panel corresponds to one task, the x-axis lists evaluated LLMs, and grouped bars denote the task-specific main subcategories. Scores are reported using the same task metrics as the main evaluation and shown as percentages. Task4 is summarized by primary functional categories. Rare Task2 taxonomy categories such as Archaea and Fungi contain few examples and should be interpreted cautiously.

The subcategory trends provide complementary evidence to the aggregate results in Section 4. Task1 shows a consistent decrease from small to large target regions for most models, suggesting that residue-level region discovery becomes harder as the target set becomes more complex. Task2

varies across antigen taxonomy, although rare categories should not be over-interpreted. Task3 does not follow a simple monotonic pattern over the predefined difficulty groups, indicating that epitope binning depends on more than this coarse difficulty label. Task4 remains low across inhibition, protection, and cytotoxicity, and Task5 is comparatively similar across mutation chemistry classes. Overall, the subcategory analysis suggests that EpiBench difficulty is not caused by a single dominant subset, but by task-specific forms of sequence-grounded reasoning.

D.2 Answer-position Balance and Option-wise Accuracy

For tasks with discrete outputs, we examine whether model performance is affected by answer-position or label preference rather than the biological content of the prompt. EpiBench therefore balances the gold answer distribution for multiple-choice and binary tasks, and we further compute option-wise accuracy by aggregating the nine main LLM runs after deduplicating predictions by final test sample ID.

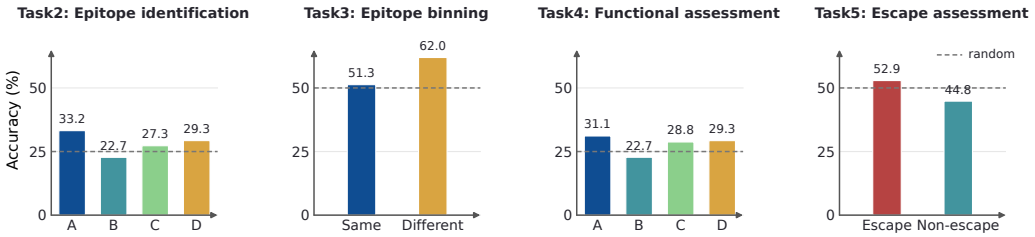


Figure 9: Option-wise accuracy under balanced answer distributions. Bars report aggregated accuracy over the nine evaluated LLMs for each gold answer option or binary label, and dashed lines denote random baselines. The results show no strong systematic preference for a specific answer position at the aggregate level, supporting the fairness of the discrete-choice evaluation.

The option-wise results are broadly balanced across answer positions, especially for the four-way multiple-choice tasks. This indicates that the reported scores are not explained by a simple shortcut such as consistently favoring one option letter or one binary label. The remaining variation is therefore better interpreted as model behavior on the underlying sequence-based decision, rather than an artifact of an imbalanced answer key.