Few-shot in-context learning with large language models for antibody characterization

Sin-Hang Fung[1,2], Zhenghao Zhang[3], Ran Wang[1], Chen Miao[1], Brian Shing-Hei Wong[1], Kelly Yichen Li[4], Chenyang Hong[3], Jingying Zhou[1], Kevin Y. Yip[3,4,*], Stephen Kwok-Wing Tsui[1,5,*], and Qin Cao[1,2,5,*]

1 School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China
2 Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China
3 Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China
4 Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, USA
5 Hong Kong Bioinformatics Centre, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China

* Correspondence
To whom correspondence should be addressed. Emails: kyip@sbpdiscovery.org (KYY), kwtsui@cuhk.edu.hk (SKWT), qcao@cuhk.edu.hk (QC)

Abstract

Large language models (LLMs) exhibit the emergent ability of few-shot in-context learning (ICL), allowing them to learn from demonstrations in simple prompts without task-specific training. However, applying few-shot ICL to biological sequences for classification, especially in computational immunology, remains underexplored. Here, we apply few-shot ICL with 18 general-purpose LLMs across five families to three antibody characterization scenarios, including predicting antibody humanness, specificity, and isotype. We evaluate performance under zero-shot, few-shot, and few-shot Chain-of-Thought ICL settings. We propose similarity-based few-shot demonstration selection strategies, which significantly improve performance of few-shot ICL compared to random selection. In all three scenarios, few-shot ICL, with as few as 32 examples, matches or exceeds the performance of established machine learning (ML) models trained on large datasets using traditional feature encodings. In two of the evaluated scenarios, few-shot ICL even matches or outperforms ML models that use state-of-the-art protein language model-based embeddings. Moreover, combining few-shot ICL with fine-tuning further enhances performance. We also demonstrate the reproducibility and stability of few-shot ICL results. Our findings establish few-shot ICL as a powerful method for efficiently characterizing antibody properties without task-specific training, enabling a single model to perform multiple tasks immediately. Its simplicity and versatility make few-shot ICL a promising approach to antibody characterization for researchers from diverse backgrounds, especially those without coding knowledge.

Introduction

In-context learning (ICL) with pretrained large language models (LLMs) has shown promise for making predictions for unseen inputs solely using information provided in simple prompts, without the need for parameter updates through any task-specific training[1]. ICL can be categorized into two major approaches: zero-shot learning, which involves no demonstration examples, and few-shot learning, which incorporates a few demonstration examples before asking the LLM to make predictions[1]. A "shot" refers to a single example used. In *k*-shot learning, *k* represents the specific number of demonstration examples provided. In contrast to ICL, fine-tuning involves updating an LLM's parameters through task-specific training using thousands to hundreds of thousands of labeled training instances[1]. Fine-tuning offers strong performance but requires a large, task-specific dataset[1]. Interestingly, ICL using a few labeled training data can outperform models that have undergone supervised fine-tuning in language-related tasks[1,2]. Numerous recent research efforts have been focused on deciphering the mechanisms behind the emergent capabilities of ICL. Some recent studies explained ICL's inference mechanism as an implicit Bayesian inference mechanism[3] and its learning algorithm as implicit gradient descent[4,5], while others argued ICL could not learn the input-label relationship in demonstrations but from some other factors, such as the label space and the distribution of the input text[6].

Despite the lack of clear consensus on its underlying mechanisms, ICL has demonstrated initial success in disciplines beyond language-related tasks, including chemistry[7–9], genomics[10,11], and single-cell cell-type annotation[12] tasks. A recent study suggested ICL to be "competitive" or "selectively competitive" in five out of eight chemistry tasks, including property classification and molecule design[9]. However, in bioinformatics, ICL has primarily been applied to text-related tasks, such as genomic information retrieval[10,11], including retrieving gene genomic location and nomenclature tasks, and biomedical text processing[2], rather than directly using biological sequences for classification. When using ICL prompts for classification, article titles and abstracts for biomedical text classification[2], and gene names for single-cell cell-type annotation[12], were used instead of biological sequences. For biological sequence inputs, Microsoft Research recently evaluated GPT-4's zero-shot ICL, but not few-shot ICL, on various biological tasks, including consensus sequence prediction, protein binding site prediction, signaling peptide prediction, and functional domain identification[13]. Their results suggested that GPT-4 has built-in biological knowledge, although its factual and quantitative outputs require manual verification[13]. Furthermore, rather than directly using ICL prompts, studies have generated natural language embeddings using ICL prompts with LLMs by gene[14] and drug[15] descriptions for classification using machine learning (ML) models. To the best of our knowledge, directly using biological sequences for classification by ICL with LLMs, especially in computational immunology, remains underexplored.

Antibody characterization refers to the process of analyzing the properties and functions of antibodies[16]. It is one of the crucial steps in antibody development for creating therapeutic antibodies and vaccines[17]. Key measures of antibody characteristics, including antigen binding affinity, specificity, immunogenicity, and developability, are ultimately assessed through *in vivo* evaluations. These evaluations involve target engagement and side effect profiling in animal models and patients, which are time-consuming and laborious processes[17,18]. Recent advancements in

computational algorithms are increasingly integrated into antibody development pipelines to overcome these limitations[17–20]. Yet, the training process using structural data is hindered by the relatively small size of existing antibody structure databases and the lack of known antigen structure[17]. The latest research has demonstrated that ML models using vast high-throughput sequencing datasets can effectively characterize antibodies by building individual models using only sequence data, without relying on structural information[21–25], for predicting immunogenicity[24,25], specificity[23], and binding affinity[21]. Therefore, we explore whether LLMs' ICL ability can leverage sequence data alone to achieve performance comparable to or surpassing that of ML models in antibody characterization. Successful application of ICL could reduce the need for large task-specific training datasets and extensive training time in antibody characterization, while benefiting researchers from diverse backgrounds, especially those without coding expertise, ultimately streamlining the antibody development process.

In this work, we demonstrate the effectiveness of few-shot ICL using amino acid sequences across three antibody characterization scenarios, leveraging 18 general-purpose LLMs from five families, including GPT[1,26], Claude[27], Gemini[28], LLaMA[29,30], and Mistral[31,32]. Our findings show that few-shot ICL, using our proposed similarity-based demonstration selection strategies, can perform competitively with established ML models trained on large datasets using state-of-the-art protein language model (pLM) embeddings. We also ensure the reproducibility and stability of the results. To summarize, our work lays a strong foundation for the effective application of ICL using amino acid sequences in computational immunology.

Results

Study overview

To establish a foundation for applying ICL in antibody characterization, we comprehensively investigated if the performance is comparable to established ML models and if the results are highly reproducible and stable. We selected 18 representative LLMs across five families, including GPT-3.5, GPT-4, GPT-4o, GPT-4o mini, Claude 3 Haiku, Claude 3 Sonnet, Claude 3 Opus, Claude 3.5 Sonnet, Gemini 1.0 Pro, Gemini 1.5 Flash, Gemini 1.5 Pro, LLaMA 2 (7B, 13B, and 70B), LLaMA 3 (8B and 70B), Mistral v0.2, and Mixtral v0.1 **(Fig. 1a, Supplementary Table 1, 2, and Methods)**. These selected LLMs have demonstrated strong performance in ICL across multiple fields[1,9] and achieved high rankings on LLM leaderboards[33,34].

Our investigation focused on three practical antibody characterization scenarios **(Fig. 1a)**, including antibody humanness binary classification (Scenario 1), antibody specificity binary classification (Scenario 2), and antibody heavy chain isotype multiclass classification (Scenario 3). These scenarios were selected as they represent fundamental antibody characteristics[17,35]. For Scenario 1, antibody humanness is related to immunogenicity, as more human-like antibodies elicit fewer immune responses[24,36]. In this scenario, we constructed two sub-scenarios: distinguishing between mouse and human antibodies (Scenario 1a), representing an easier task, and distinguishing between rhesus and human antibodies (Scenario 1b), which presents a more challenging task. For Scenario 2, identifying antibody sequences that bind to a specific antigen (e.g., SARS-CoV-2 antigen) is a long-standing challenge in computational antibody design[20,37] and vaccine development[38]. For Scenario 3, we selected this scenario for its multiclass nature and basic antibody characteristics[35]. The datasets used for each antibody characterization scenario are detailed in **Methods** and **Supplementary Table 3**.

To comprehensively assess the effectiveness of few-shot ICL in the three antibody characterization scenarios, we explored various sequence representations (letter code and pLM embedding representations) and input formats (with and without multiple sequence alignment [MSA]). We proposed sequence similarity-based strategies for few-shot demonstration selection, and designed different ICL prompt settings (zero-shot, few-shot, and few-shot Chain-of-Thought [CoT]) **(Fig. 1a, b, and Supplementary Fig. 1-3).** To objectively quantify ICL's performance, we compared it against random baselines and established ML models particularly state-of-the-art models utilizing pLM embeddings for feature encoding **(Fig. 1a).** Additionally, we combined ICL with fine-tuning to evaluate its impact on performance **(Fig. 1a and Methods)**. Beyond performance evaluation, we also assessed the reproducibility and stability of the ICL results **(Fig. 1a)**. Throughout the study, each prompt was repeated five times to evaluate the consistency of the results.

Sequence similarity-based demonstration selection strategies improve performance for few-shot prompts

Selecting appropriate demonstration examples is a key factor in the success of few-shot ICL. Studies have shown that selecting demonstration examples similar to test instances can improve few-shot ICL's performance in language-related[39] and chemistry[7,9] tasks. Existing antibody characterization algorithms often rely on sequence similarity as an indicator of clonal relatedness for prediction[40,41]. Therefore,

we proposed similarity-based selection strategies for biological sequence classification in few-shot ICL and evaluated the performance against zero-shot and the default random selection strategy. Specifically, we evaluated six detailed strategies covering random and similarity-based strategies with the control of the order and balance of positive and negative examples **(Fig. 2a and Methods)**. Balanced selection mitigates the label bias that arises when selecting demonstrations based solely on similarity, which can lead to a lack of diversity in the ground-truth labels of the demonstrations **(Supplementary Fig. 4)**.

We first tested our proposed strategies using the affordable LLMs in terms of token cost across four families in a relatively easy antibody characterization scenario that predicts humanness by distinguishing human and mouse sequences (Scenario 1a in **Fig. 1a**). We represented amino acid sequences using letter codes with and without MSA and quantified sequence similarity using Levenshtein distance (LD) **(Methods)**. Zero-shot prompts without demonstrations yielded random predictions **(Fig. 2b, c)**. Larger values of k (i.e., more shots) generally outperformed 2-shot learning **(Fig. 2b, c)**. Notably, all the LLMs can attain an accuracy level close to 1 using suitable similarity-based strategies with an appropriate number of shots **(Fig. 2b, c, and Supplementary Fig. 5)**. The Similar (Balanced) strategy achieved the best performance in GPT-3.5, Claude 3 Haiku and Gemini 1.0 Pro, while the Similar strategy performed best in LLaMA 3 (8B) **(Fig. 2b, c, and Supplementary Fig. 5)**. Overall, similarity-based strategies significantly outperformed the random strategy **(Fig. 2d)**. As the Similar strategy, which may suffer from label bias **(Supplementary Fig. 4)**, yielded optimal performance in some models, we further investigated whether LLMs can effectively make predictions from a single label type. To test this, we preserved only the positive instances from the Similar (Balanced) strategy and evaluated the performance. Positive label-only prompts yielded random performance in all four affordable models, but not in the advanced model like GPT-4 **(Supplementary Fig. 6)**. However, using both positive and negative labels consistently outperformed positive label-only prompts in our scenario. Together, our results demonstrate the validity of similarity-based few-shot demonstration examples in antibody humanness classification.

Few-shot ICL can learn from amino acid sequences without MSA

Previous evidence suggested that using MSAs instead of unaligned amino acid sequences improved performance in antibody humanness classification by deep learning models [24]. We thus repeated the previous evaluations on the input amino acid sequences without-MSA format to evaluate if this conclusion applies to few-shot ICL **(Supplementary Fig. 7)**. The same affordable LLMs showed highly similar performance trends, irrespective of the use of MSA **(Fig. 2b, c)**. Accuracy between with-MSA and without-MSA formats showed no statistical significance in all the settings **(Fig. 2e)**. This result is coherent with the findings in pLMs, in which MSA-free pLMs showed comparable performance as MSA-based pLMs in common biological tasks, such as sub-cellular localization and membrane protein classification by deep learning models[42]. For simplicity, sequences without-MSA were used in our subsequent analyses.

Long-context LLMs effectively capture long pLM embedding information for antibody characterization in few-shot ICL

Established ML models such as Random Forest (RF) utilizing pLM embeddings have demonstrated remarkable performance in antibody characterization[35]. Inspired by this, we next asked if LLMs can also effectively utilize pLM embeddings as the sequence representation alternatively to the letter codes in antibody characterization. For each amino acid sequence, we replaced its letter code representation with a pLM embedding (dimension=512, three decimal places, **Methods**) generated by AntiBERTy, an antibody-specific pLM pre-trained on natural antibody sequences. Due to the numerical nature of embeddings, we used cosine similarity instead of LD to quantify similarity in our similarity-based strategies. We selected economical LLMs that have long context windows, namely GPT-4o and Gemini 1.5 Flash with context window sizes of 128K and 1M, respectively, instead of the more expensive GPT-4 and Gemini 1.5 Pro, for a proof-of-concept experiment due to cost considerations.

Indeed, LLMs effectively predicted antibody humanness by pLM embeddings in few-shot ICL with similarity-based strategies, outperforming the random baseline in general **(Fig. 2f)**. Notably, the pLM embedding representation with the Similar strategy achieved comparable accuracy to the letter code representation of the Similar (Balanced, Random) strategy in Gemini 1.5 Flash **(Fig. 2f)**. Moreover, the Similar strategy stably achieved an accuracy level close to 1 using pLM embeddings in Gemini 1.5 Flash with as few as 4 shots. However, increasing *k* may not always produce better results in long-context LLMs, which is supported by previous studies in language-related tasks[43], for the Similar (Balanced) strategy. These findings underscore the versatility of LLMs in leveraging different forms of representation; further advancements in long-context LLMs for handling pLM embeddings will likely enhance the performance.

Few-shot ICL generally outperforms established ML models that use state-of-the-art pLM embeddings in most evaluated scenarios

Based on our initial success, we next evaluated the few-shot ICL of all 18 LLMs on three practical and challenging antibody characterization scenarios using the letter code representation (Scenario 1b, Scenario 2, and Scenario 3 in **Fig. 1a**). We compared the performance of the few-shot ICL using *k*-shot demonstrations and the established ML models using the full training set as the standard practice[9] **(Methods)**. Specifically, for few-shot ICL, we applied both the random and similarity-based demonstration selection strategies in 16- and 32-shot settings, considering the trade-off between performance and cost. The scenarios were chosen at varied levels of difficulties, as suggested by the classes' separation in the uniform manifold approximation and projection (UMAP) plots **(Supplementary Fig. 8)**. Each LLM's response was summarized in **Supplementary Table 4**. The majority of the LLMs (14 out of 18 LLMs, 77.8%) provided formatted responses without any additional words beyond the predicted labels **(Supplementary Table 4a and Methods)**.

We first investigated whether the few-shot task-specific demonstrations were necessary compared to zero-shot ICL in all three scenarios **(Fig. 1b, Supplementary Fig. 1, 2, and Methods)**. Across all three scenarios and 18 LLMs tested, zero-shot ICL accuracy was comparable to random guessing. In contrast, few-shot ICL, using demonstration selection strategies covering random and similarity-based strategies,

significantly outperformed zero-shot ICL **(Supplementary Fig. 9 and Supplementary Table 4b)**. This observation indicates the pre-trained knowledge in the LLM alone is insufficient for our antibody characterization scenarios. Despite their broad knowledge, LLMs often require task-specific demonstrations to effectively tackle challenging areas[1], particularly non-natural language tasks like antibody characterization.

Having established the necessity of few-shot demonstrations, we proceeded to evaluate the performance of few-shot ICL in the three antibody characterization scenarios and compared it with established ML models. In the first binary scenario (Scenario 1b), we evaluated antibody humanness using human and rhesus heavy chain sequences. Rhesus sequences share higher genetic similarity with human sequences than mouse[24], making this a more challenging classification task than Scenario 1a **(Supplementary Fig. 8a, b)**. Similarity-based selection strategies, including the Similar (6 out of 18 LLMs, 33.3%) and the Similar (Balanced) (12 out of 18 LLMs, 66.7%) strategies, yielded the best performance in all 18 LLMs **(Fig. 3a)**. All 14 LLMs, except GPT-4o mini and the three LLaMA 2 models, outperformed both traditional encoding-based (amino acid frequency and physiochemical encodings) methods **(Methods)** and achieved a nearly perfect best performance with an accuracy greater than 0.9 **(Fig. 3a)**. LLaMA 2's poor performance was due to its failure to provide any prediction responses **(Supplementary Table 4a)**. Impressively, three LLMs (Claude 3.5 Sonnet, LLaMA 3 (8B), and Mistral v0.2) attained a perfect accuracy of 1.0, matching the performance of the state-of-the-art AntiBERTy[44]-based pLM method **(Fig. 3a)**.

The second binary scenario (Scenario 2) was antibody specificity classification that distinguished between SARS-CoV-2 binding and non-binding antibodies with concatenated heavy and light chain sequences. This scenario was more challenging, as suggested by the class overlap in the UMAP plot **(Supplementary Fig. 8c)**. Note that the 32-shot setting was excluded from the LLaMA 2 models due to their limited context window size **(Supplementary Table 1)**. As observed in Scenario 1b, similarity-based selection strategies, including the Similar (13 out of 18 LLMs, 72.2%) and the Similar (Balanced) (5 out of 18 LLMs, 27.8%) strategies, yielded the best performance in all 18 LLMs **(Fig. 3b)**. In total, four LLMs (GPT-4o, Claude 3 Haiku, Gemini 1.5 Flash, and Mistral v0.2) outperformed both the amino acid frequency encoding-based method (accuracy=0.709) and physiochemical encoding-based method (accuracy=0.719) **(Fig. 3b)**. However, the pLM-based model had an accuracy as high as 0.845 while the best accuracy of few-shot ICL could only reach 0.746 by Mistral v0.2 **(Fig. 3b)**.

The third scenario (Scenario 3) assessed a challenging antibody isotype multiclass classification problem using heavy chain sequences, evidenced by poor separation in the UMAP visualization **(Supplementary Fig. 8d)** and the established ML models' relatively low performance **(Fig. 3c)**. Similarity-based selection strategies, including the Similar (13 out of 18 LLMs, 72.2%) and the Similar (Balanced) (7 out of 18 LLMs, 38.9%) strategies, achieved the best performance in all 18 LLM, consistent with the binary classification scenarios. GPT-4o and Gemini 1.0 Pro achieved the best performance with both similarity-based strategies. Impressively, six (Claude 3 Haiku, Claude 3 Sonnet, Gemini 1.5 Flash, Gemini 1.5 Pro, LLaMA 3 (70B), and Mixtral v0.1) out of 18 LLMs surpassed the performance of the pLM-based model (accuracy=0.492) that had the highest accuracy across all the established ML models **(Fig. 3c)**.

The best performance accuracy was mostly achieved by 32-shot similarity-based prompts instead of 16-shot prompts (Scenario 1b: 11 out of 18 LLMs, 61.1%; Scenario 2: 9 out of 15 LLMs, 60%; Scenario 3: 10 out of 18 LLMs, 55.6%) **(Fig. 3a-c)**. Consistent in all the three evaluated scenarios, similarity-based demonstration selection strategies significantly outperformed random selection **(Fig. 3d)**. Besides, demonstration orders did affect similarity-based selection strategies performance in some of settings, as reported in previous studies[45,46] **(Fig. 3d)**. To further evaluate whether the aforementioned label bias in the Similar strategy **(Supplementary Fig. 4)** will cause LLMs to simply rely on the most frequent label type for predictions, we compared the ICL performance in the evaluated scenario with a majority voting classifier **(Methods)**. We found that when the demonstration labels were more similar to test instance label, such as in Scenario 1 **(Supplementary Fig. 4)**, the predictions made by some LLMs (e.g., Gemini 1.5 Flash, LLaMA 3 (70B), and Mixtral v0.2) were closer to those made by the majority voting classifier, suggesting a potential reliance on the majority label shortcut **(Supplementary Fig. 10)**. In contrast, in other scenarios where the demonstration labels were more diverse and less like the test instance label **(Supplementary Fig. 4)**, the ICL predictions by some LLMs (e.g., Gemini 1.5 Flash, LLaMA 3 (70B), and Mixtral v0.2) differed more from the majority voting classifier **(Supplementary Fig. 10)**. However, some LLMs (e.g., GPT-3.5) may not strongly tend to use the majority label shortcut for their predictions, while the predictions of some LLMs (e.g., LLaMA 3 (8B) and Mistral v0.2) were closer to those made by the majority voting classifier, irrespective of the presence or absence of label bias **(Supplementary Fig. 10)**.

For LLMs, no single LLM or single LLM family outperformed other LLMs across all the scenarios, a finding that aligns with previous studies on biomedical[2] and long-context text processing[43] tasks. Mistral v0.2 and Claude 3 Sonnet performed the best in the evaluated binary (Scenario 1b and 2) and multiclass classification (Scenario 3) scenarios, respectively **(Fig. 3a-c)**. Moreover, open-source LLMs performed comparably to closed-source LLMs in two (Scenario 1b and 3) of three evaluated scenarios **(Fig. 3e)**. Interestingly, within the same LLM family, the most advanced model did not always produce the best performance, as exemplified by the Claude 3 family in Scenario 2 with Haiku and Scenario 3 with both Haiku and Sonnet outperformed Opus, the most advanced model in Claude 3 family, respectively **(Fig. 3b, c)**. Our results demonstrate that in all three antibody characterization scenarios, few-shot ICL with only 32 similarity-based demonstrations can match or outperform established ML models trained on large datasets using traditional encodings, and in some scenarios, even rival or surpass those using the state-of-the-art pLM-based embeddings.

For subsequent analyses, we employed a 32-shot Similar (Balanced, Random) strategy that can avoid potential label bias and excluded the underperforming LLaMA2 models.

CoT prompt does not improve performance compared to non-CoT prompt in the three antibody characterization scenarios

Previous studies have shown that CoT was useful in reasoning tasks, like arithmetic and logical reasoning[47,48]. We thus explored few-shot CoT prompts involving amino acid frequency calculation and considering amino acid physiochemical properties as reasoning steps. We first evaluated three CoT prompt variations, including CoT steps

alone, with LLM-generated answers, and with manually corrected answers **(Methods)**. We manually corrected the LLM-generated answers in the third variation, as LLM-generated answers contained errors, like miscalculating amino acid frequencies which are similar with findings from recent studies[13,49] **(Supplementary Fig. 11, 12, and Methods)**. Interestingly, the models performed worse using manually corrected answers than LLM-generated answers for CoT reasoning steps, suggesting that LLMs may generate reasoning steps better suited to their own decision-making process, even if those steps are incorrect in human evaluation **(Supplementary Fig. 12, 13)**. We then selected the best-performing prompts, CoT steps alone, for subsequent evaluation **(Supplementary Fig. 13)**. Unexpectedly, CoT prompts underperformed non-CoT prompts in all three scenarios **(Fig. 4a and Supplementary Table 4c)**. This may be attributed to the insufficient explanatory power of the current reasoning steps in capturing the underlying features that determine antibody-related characteristics. As a recent study suggested, CoT reasoning primarily benefits tasks involving math, logic, or symbolic operations, which are often indicated by the presence of equals signs in the question or model's response[50]. However, antibody-related characteristics may not heavily rely on these types of operations.

Few-shot prompts improve fine-tuned model over base model

A significant distinction between fine-tuning and ICL is that fine-tuning involves updating the parameters of the pre-trained model, while ICL does not[1]. We next investigated if adopting fine-tuning using a large size of the training data (i.e., 1,000, 3,000, and 5,000 instances) could improve LLMs' performance in the challenging antibody specificity prediction (Scenario 2). We focused on three LLMs having different architectures: GPT-3.5, GPT-4o mini, and Gemini 1.0 Pro **(Supplementary Fig. 14 and Methods)**. We compared their performance to their advanced models, GPT-4o and Gemini 1.5 Pro, respectively, using both zero-shot and few-shot ICL prompts.

For the zero-shot ICL, nearly all the fine-tuned models performed better than their base models and advanced models, with fine-tuned GPT-4o mini models achieving the best result, setting a new performance record across all LLMs with an accuracy of 0.794 **(Fig. 3b and 4b-d)**. The performance gap with the pLM-based ML model (accuracy=0.845) was further reduced **(Fig. 4b-d)**. In the 16-shot ICL, all the fine-tuned Gemini 1.0 Pro models outperformed their base and advanced models, while only GPT-3.5 and GPT-4o mini fine-tuned with 5,000 training instances outperformed their base and advanced models **(Fig. 4b-d)**. Impressively, the fine-tuned GPT-3.5 model with 32-shot prompt (accuracy=0.780) outperformed the traditional-encoding based ML models **(Fig. 4b and Methods)**. In contrast, fine-tuned Gemini 1.0 Pro models failed to predict in 32-shot ICL, underperforming both their base and advanced models **(Fig. 4d)**. Our results suggest that besides zero-shot ICL, few-shot ICL can be utilized in fine-tuned models to enhance performance in challenging scenarios.

Reproducibility and stability evaluation

Several factors and hyperparameters may influence the performance of ICL. To ensure the reproducibility and stability of our results, we conducted a comprehensive investigation into these factors. We chose GPT models as a representative for this analysis due to their widespread popularity.

Prompt input order

We investigated whether the LLMs' responses would be affected by two different input orders: cycling and sequential **(Fig. 1a and Methods)**. If LLMs' responses were affected by two different input orders, performance variation between cycling and sequential settings would be expected. The insignificant performance difference between two settings implied that LLMs' responses were not affected by prompt input order **(Fig. 5a)**.

Data contamination

Data contamination, which occurs when LLMs are exposed to the test set during pre-training, may inflate performance[51]. If data contamination exists, pre-trained LLMs that potentially encounter the test set will be expected to show better performance compared to the models trained solely on the data before the test set is published **(Methods)**. Following Li et al.[51], we selected SARS-CoV-2 antibody specificity (Scenario 2) data published in year 2022 (post-September 2021) and applied on LLMs with training data cut-off in September 2021, i.e., pre-September 2021 model (GPT-4 [20230613]), and training data cut-off in December 2023, i.e., post-September 2021 models (GPT-4 [20240125] and GPT-4 [20240409]), to test whether data contamination exists.

In the 16-shot setting, despite the statistically significant differences between pre- and one of the post-September 2021 models (GPT-4 [20240409]), we observed that the other post-September 2021 model showed statistically insignificant performance (GPT-4 [20240125]). Furthermore, in the 32-shot setting, the post-September 2021 models either exhibited statistically insignificant differences or even significantly worse performance compared to the pre-September 2021 model **(Fig. 5b)**. Together, these observations indicate that our results were unlikely influenced by data contamination.

Word substitution test

We examined ICL's tolerance for alternative spellings or misspellings. Both words "physiochemical"[23] and "physicochemical"[35] were widely used to describe antibody characteristics in literature **(Methods)**. The word choices did not affect the predictions **(Fig. 5c)**. Our results suggest that ICL exhibits some tolerance to alternative terms.

Repeated train-test split

To address the potential issues associated with using a small test set, such as overfitting to specific examples, insufficient coverage of data variability, and increased variance in performance metrics, we performed multiple train-test splits using five random seeds. This ensures the model was tested on diverse data subsets. No instances were shared between different test sets. The evaluation was expanded to also include Claude 3 Haiku, Gemini 1.0 Pro, and LLaMA 3 (8B) for a more comprehensive analysis. The ICL's performance trend in different LLMs was consistent as that in previous sections **(Fig. 5d)**. Also, ICL's performance variability was comparable to established methods using traditional encodings and pLM embeddings **(Fig. 5d)**. This indicates that ICL is robust to data variability.

Model-related hyperparameters

Model version

Next, we evaluated the impact of model-related hyperparameters on prediction stability **(Methods)**. GPT-3.5 showed greater variation, while GPT-4 remained

relatively stable across different versions. Surprisingly, the latest model did not always perform better, with GPT-4 (20240125) outperforming GPT-4 (20240409) in all the evaluated *k*-shot settings except 2-shot **(Fig. 5e)**. Researchers need to take extra caution when selecting different versions of LLMs, as the latest LLMs may not always yield the best results.

Temperature

The *temperature* parameter, which controls randomness and creativity in LLM responses, did not affect predictions in few-shot ICL in our scenario **(Fig. 5f, Supplementary Fig. 15, and Methods)**.

Seed

Given the non-deterministic nature of LLMs' responses, we assessed the impact of the *seed* parameter on output consistency **(Methods)**. The performance using different seeds was not significantly different compared to that using the same seed **(Fig. 5g)**. This suggests that the *seed* parameter of the OpenAI application programming interface (API) has a minimal effect on the output consistency, and that the model's performance is relatively stable across different random initializations.

Together, our results suggest that ICL is a robust method for antibody characterization.

Discussion

In this work, we have comprehensively evaluated the classification performance of ICL with amino acid sequence input in 18 LLMs across five families in three antibody characterization scenarios. Our results demonstrate that ICL, without task-specific training and only using as few as 32 similar instances with our proposed similarity-based few-shot demonstration selection strategies, significantly outperforms the random baselines in all three scenarios. In the antibody humanness binary classification and antibody isotype multiclass classification scenarios, ICL even surpasses the state-of-the-art ML models that use pLM-based embeddings. Furthermore, ICL can be applied to various antibody characterization scenarios without task-specific training, making it an efficient approach for immediate testing of different antibody characteristics. This reduces the need for training separate models on different full datasets for each antibody characteristic.

Existing antibody sequence databases, such as Observed Antibody Space (OAS)[52], facilitate similarity searches for a specific sequence of interest, enabling the construction of similarity-based ICL prompts without requiring coding knowledge. Our results suggest that antibody characterization prediction can be done by interacting with the LLM with similarity-based prompts as it is a chatbot **(Supplementary Fig. 16)**, or through user-friendly APIs.

Additionally, open-source and closed-source LLMs exhibited comparable performance in ICL for the two evaluated scenarios, allowing researchers to select open-source models for customization, transparency and lower cost[53]. Overall, the high performance and user-friendly procedures position ICL as a promising alternative to established ML models in antibody characterization and other similar biomedical tasks, benefiting researchers with varying levels of coding expertise.

In addition to the superior performance yielded by ICL with our proposed similarity-based few-shot demonstration strategies in all the evaluated scenarios, we have comprehensively investigated different sequence representations and prompt designs. We have shown that ICL can learn from amino acid sequences without MSA as effectively as with MSA, greatly reducing the needs for data processing. Although pLM embeddings do not outperform the amino acid letter code in ICL, we anticipate that the ongoing advancements in long-context LLMs may improve this situation. Furthermore, we anticipate that incorporating embeddings extracted from protein structure prediction models, like AlphaFold[54], can facilitate antibody specificity prediction with ICL. Notably, our CoT prompt variations underperformed non-CoT prompts, suggesting antibody characterization scenarios may require more task-specific CoT design, such as incorporating more antibody-related biological evidence into the reasoning steps. We have also shown that combining ICL with fine-tuning can further boost performance. Finally, we have delved into additional factors such as data contamination and model hyperparameters that could affect the performance; this assessment is essential for understanding the behavior of ICL. Importantly, our findings have confirmed the robustness and reproducibility of using ICL with LLMs for antibody characterization.

Nevertheless, our study has certain limitations. The test set sizes are relatively small **(Methods)** due to the high costs of repeating experiments associated with LLMs. To

mitigate the potential robustness concern, we have repeated the experiments in multiple train-test splits and observed stable results. We anticipate LLM advances will lower expenses, enabling more extensive testing. Another limitation pertains to our CoT design, as our CoT prompt variations underperform non-CoT prompts. A recent study summarized various CoT strategies[55], offering potential guidance for a more effective CoT design in antibody characterization contexts. Additionally, our study primarily focused on ICL, with only limited investigation into fine-tuning. While our results indicate that fine-tuned models can enhance ICL performance, there remains opportunity for more extensive research in this area.

Our findings suggest ICL capture antibody amino acid information for antibody characterization classification scenarios. We anticipate that further theoretical investigations into ICL[56], alongside empirical performance assessments, will validate ICL efficacy in regression applications, such as binding affinity prediction, and generative applications, like therapeutic antibody design, for therapeutic antibody development. These finding will also advance protein sequence analysis and diverse biomedical applications using ICL beyond antibody-related tasks.

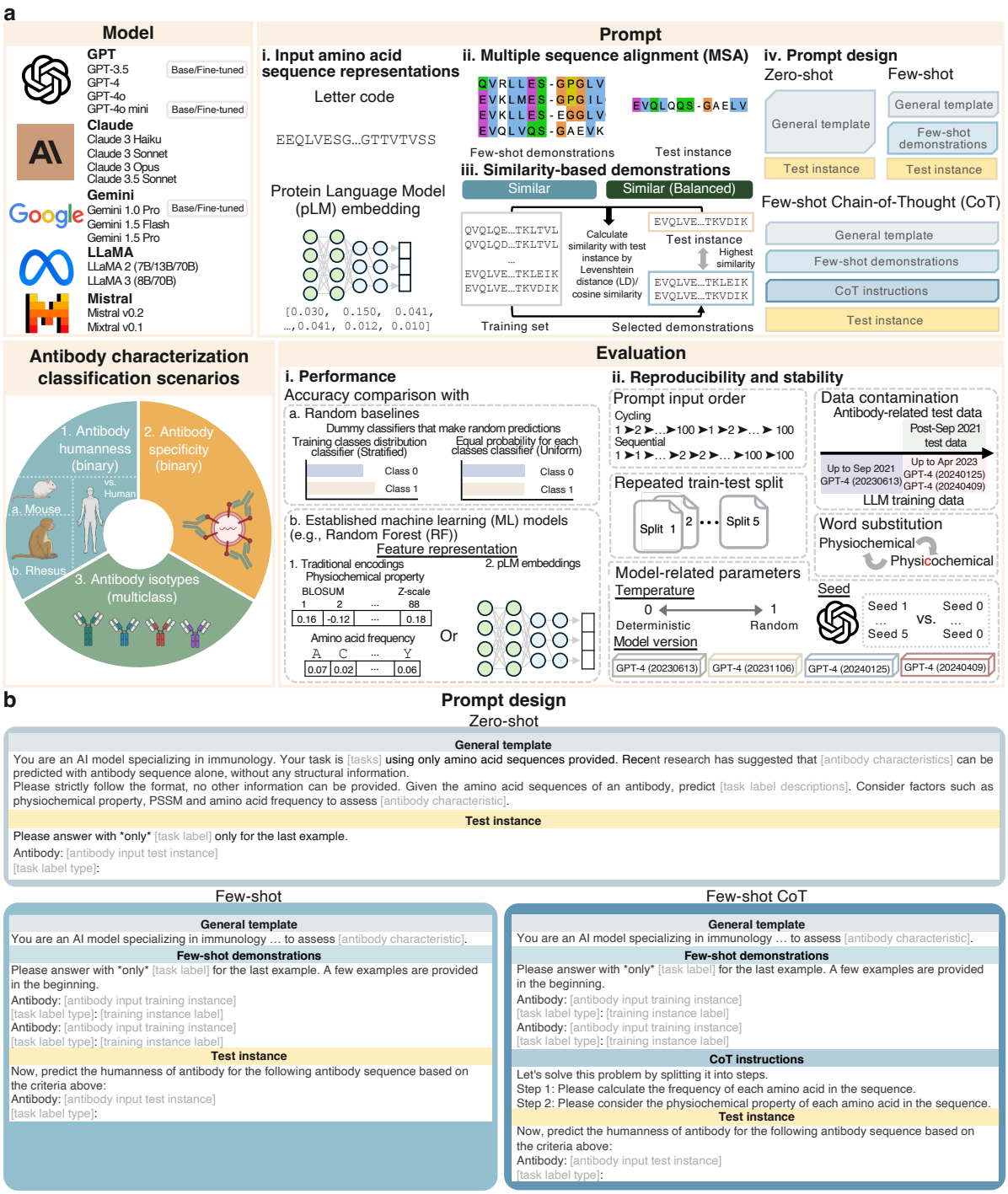


Fig. 1 Study overview and prompt design. In this work, we apply ICL to three antibody characterization scenarios by 18 LLMs from five different families. **a** Schematic overview of our work. Figure created with BioRender.com released under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license. **b** Illustration of ICL prompt design. Detailed prompt designs are given in the **Supplementary Fig. 1-3**.

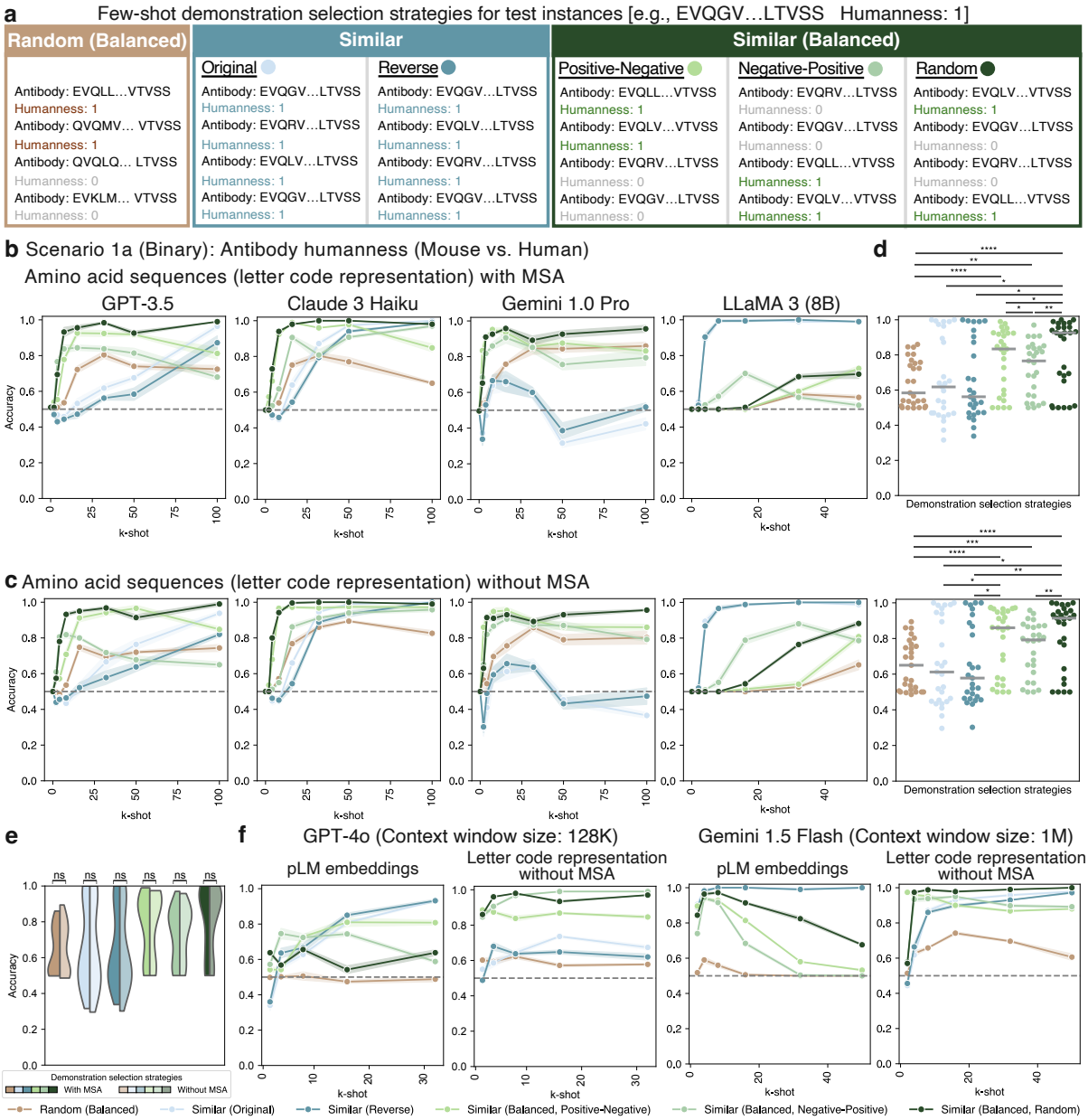


Fig. 2 Accuracy evaluation of few-shot demonstration selection strategies using different sequence representations. **a** Illustration of few-shot demonstration selection strategies for a test instance. **b,c** Zero-shot and few-shot (2-, 4-, 8-, 16-, 32-, 50-, 100-shot) ICL accuracy with amino acid sequence input (**b**) with and (**c**) without MSA using different demonstration selection strategies in affordable LLMs. The shaded area indicates standard deviation of five repeated predictions. The dashed line indicates the random baseline trained with full training set **(Methods)**. **d** Statistical test between the demonstration selection strategies of all four evaluated LLMs. Each dot represents an average accuracy of a LLM with a *k*-shot prompt. The grey lines indicate the median for all observations in each demonstration selection strategy. **e** Statistical test between the performance of MSA and without MSA input for each demonstration selection strategy. The darker color represents accuracy with MSA, while the paler color represents accuracy without MSA. **f** Few-shot ICL accuracy using pLM embeddings as input. *P* values in **c** and **d** are based on two-sided paired Wilcoxon rank-sum test: ns: $P > 0.05$; *$P \leq 0.05$; **$P \leq 0.01$; ***$P \leq 0.001$; ****$P \leq 0.0001$.

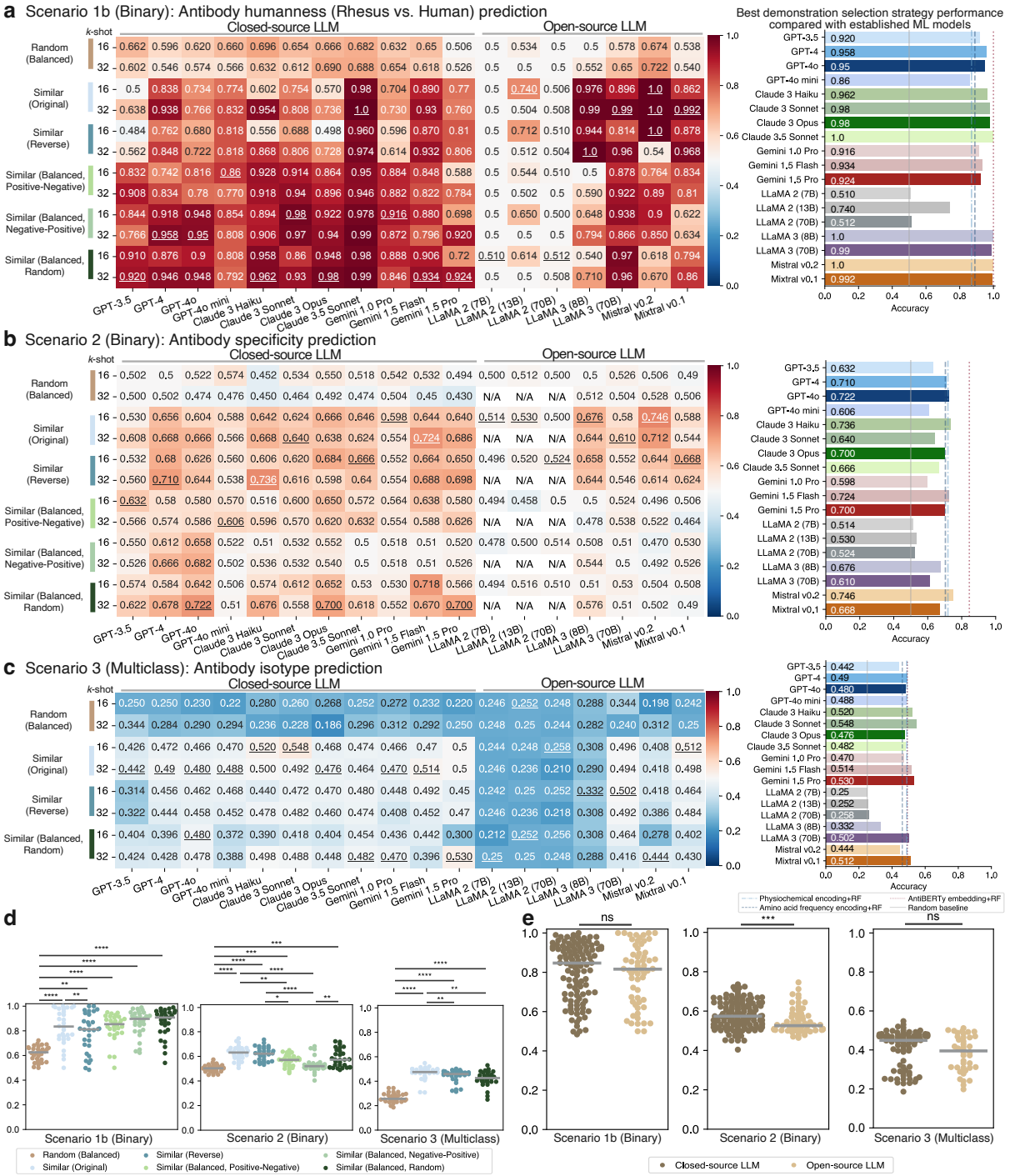


Fig. 3 Few-shot ICL accuracy in three antibody characterization classification scenarios. Evaluation of few-shot (16- and 32-shot) using different demonstration selection strategies with 18 LLMs in three antibody characterization classification scenarios. **a-c** Few-shot ICL accuracy and its best demonstration selection strategy accuracy compared with established ML models across 18 LLMs in (**a**) Scenario 1b, (**b**) Scenario 2, and (**c**) Scenario 3. Each entry shows the average accuracy of five repeated predictions. Underlined values indicate the best performance for that LLM. **d,e** Statistical test (**d**) between the demonstration selection strategies and (**e**) between closed-source and open-source models. Each dot represents the average accuracy of an LLM with a k-shot prompt. Note that LLaMA 2 models were excluded in all the comparison in **d** and **e** due to their underperformance compared to other models. P values in **d** are based on two-sided paired Wilcoxon rank-sum test and in **e** are based on two-sided Mann-Whitney U test: ns: P > 0.05; *P ≤ 0.05; **P ≤ 0.01; ***P ≤ 0.001; ****P ≤ 0.0001.

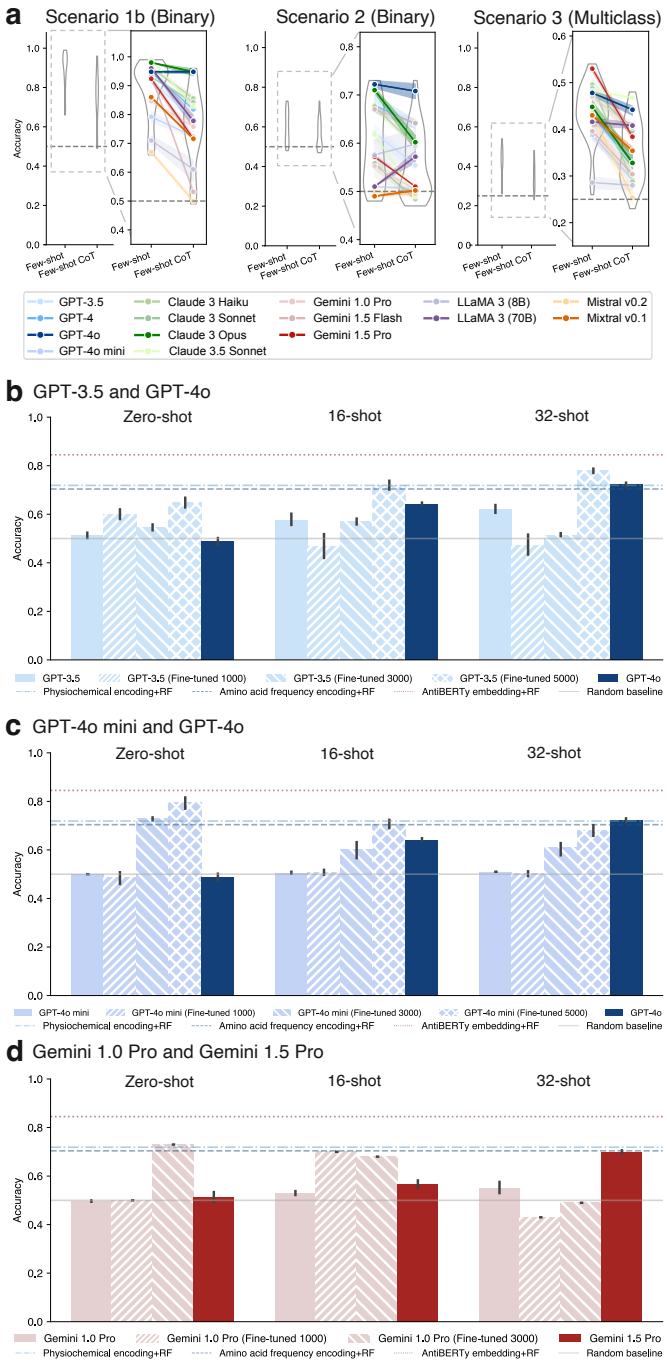


Fig. 4 ICL prompt accuracy evaluation in base, fine-tuned, and advanced models. a Accuracy of few-shot (32-shot) and few-shot CoT (32-shot) prompts in three antibody characterization scenarios. The dashed line indicates the random baseline trained with full training set. **b-d** Performance comparison of zero-shot and few-shot (16- and 32-shot) prompts of an LLM in its base, fine-tuned, and advanced models in Scenario 2: **b** GPT-3.5's base model, fine-tuned models, and advanced model (i.e., GPT-4o); **c** GPT-4o mini's base model, fine-tuned models, and advanced model (i.e., GPT-4o); and **d** Gemini 1.0 Pro's base model, fine-tuned models, and advanced model (i.e., Gemini 1.5 Pro) in Scenario 2. The number in each fine-tuned model's name denotes the number of training instances used in fine-tuning.

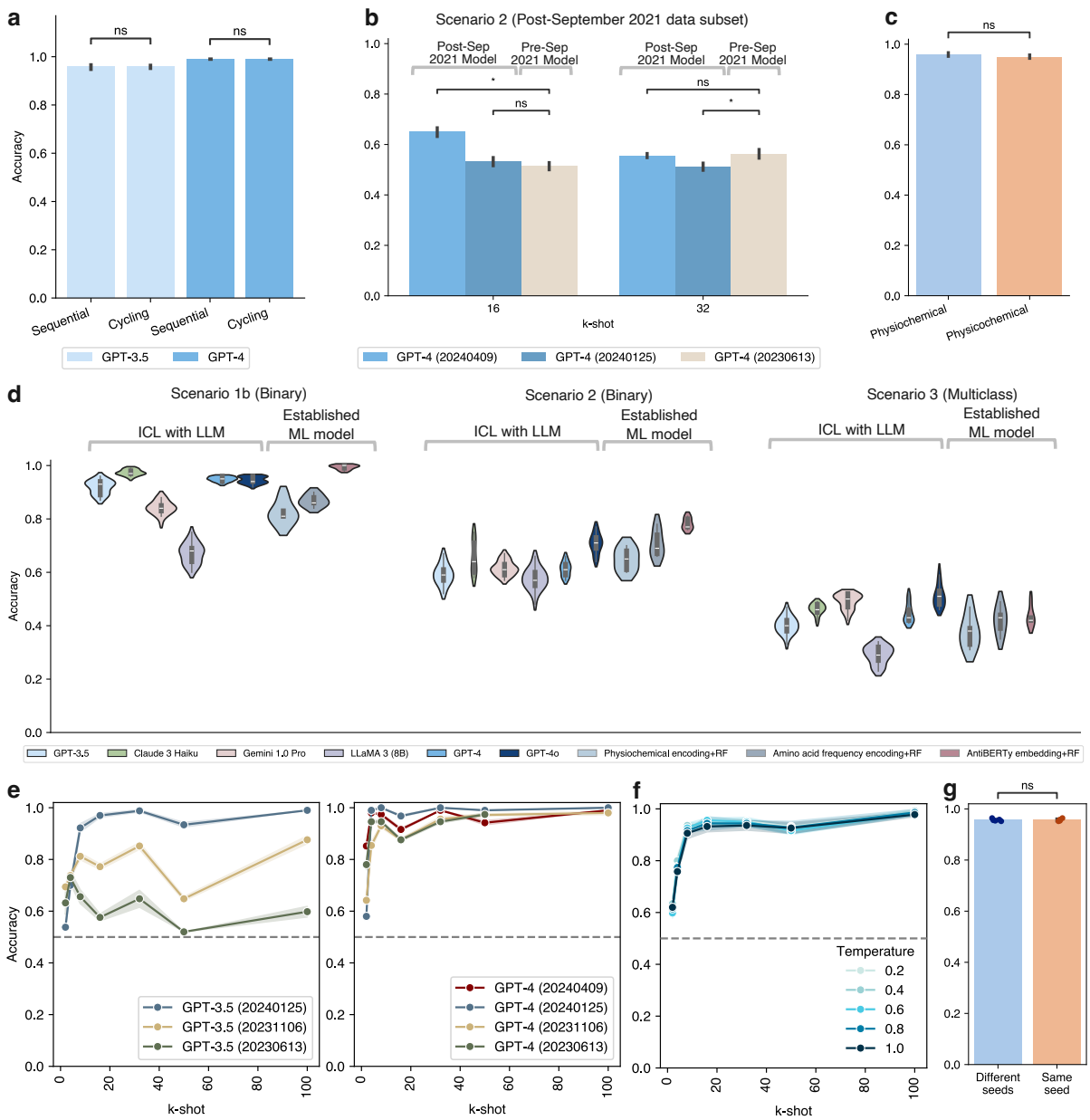


Fig. 5 Reproducibility and stability evaluation. a Accuracy comparison between cycling and sequential prompt input order. **b** Data contamination comparison in LLMs with training data cut-off in September 2021 (i.e., pre-September 2021 model: GPT-4 [20230613]) and December 2023 (i.e., post-September 2021 models: GPT-4 [20240125] and GPT-4 [20240409]) with post-September 2021 data as the test set. **c** Alternative spelling performance comparison with the word changed from "physiochemical" to "physicochemical". **d** Accuracy evaluation with existing ML models using five train-test split for evaluating ICL robustness. **e** Accuracy evaluation for GPT model variations. **f** Accuracy evaluation for parameter *temperature*. The dashed line indicates the random baseline trained with full training set in **e** and **f**. **g** Accuracy comparison for parameter *seed*. Each dot represents the mean accuracy of five repeated responses per *seed*. *P* values in **a-c** and **g** are based on two-sided Mann-Whitney *U* test: ns: $P > 0.05$; $^*P \leq 0.05$.

Methods

Dataset collection and processing

The datasets used for each antibody characterization scenario, including the number of instances in each target class for both training and test sets, were summarized in **Supplementary Table 3**. We ensured that all instances were unique (i.e., no duplicate instances) within each scenario.

Scenario 1: Antibody humanness prediction

Sequences with MSA

The MSA files containing the aligned antibody heavy chain sequences were directly downloaded from Ramon et al.[24]. This downloaded dataset had an equal number of sequence instances across all the target classes, resulting in a balanced set. According to Ramon et al.[24], the source of human, mouse and rhesus antibody heavy chain sequences were obtained from the OAS[52] database. The retrieved sequences were aligned using the AHo[57] numbering scheme, with ANARCI[58] software, which numbers each residue based on its structural role, resulting in aligned sequences of length 149. There were 20,000 unique sequences for the dataset used in Scenario 1a and 1b, respectively.

In MSA, the numbering of amino acids is based on their structural roles rather than their original positions in the sequence[58]. This numbering scheme provides a pre-defined "static position template" that generates an identical output format for a specific sequence, regardless of the number or identity of other sequences aligned with it **(Supplementary Fig. 7)**. This ensures the alignment does not rely on other sequences and reflects structural importance, allowing the model to learn structural patterns from the training set without being exposed to position-specific information from the test set.

Sequences without MSA

For the sequence input without MSA, we retrieved the sequences by removing the gaps (denoted by "-") of the downloaded sequences (encoded by amino acid letter code representation) in the MSA files by our custom bash script. This resulted in unaligned amino acid sequence for each antibody from the original MSA.

Scenario 2: Antibody specificity prediction

Antibody heavy and light chain sequences from single-cell paired heavy and light chain full-length B-cell receptor (BCR) V(D)J sequencing were downloaded from Wang et al.[35], covering datasets across ten different studies. Following Wang et al.'s [35] suggestion for optimal performance, we concatenated heavy and light chain sequences for antibody specificity prediction. There were 15,538 unique concatenated sequences for the dataset used.

Scenario 3: Antibody isotype prediction

Antibody heavy chain sequences were downloaded from the same study used in Scenario 2 (Wang et al.[35]). To enable the learning, classes with fewer than 50 instances were discarded. The target classes for the heavy chain isotype class prediction included IGHM, IGHD, IGHG, and IGHA. After filtering and deduplication, 0.81 million unique heavy chain sequences remained.

Train-test split

For each antibody characterization scenario, the full dataset was randomly divided into the training set and test set. Due to cost considerations, we set each test set to contain 100 instances. The test sets were set balanced with equal number of instances per target class **(Supplementary Table 3)**. All few-shot demonstrations were selected exclusively from the training set to ensure no data leakage in the prompts.

Demonstration selection and prompt design

ICL demonstration selection

For the random strategy, i.e., Random (Balanced), an equal number of instances from each target class were randomly selected exclusively from the training set.

For the amino acid letter code-based representation in the similarity-based strategies, including both the Similar and the Similar (Balanced) strategies, sequence similarity was assessed using LD between each training instance and each test instance. For the Similar (Balanced) strategy, an equal number of similar instances were selected from each target class. In *k*-shot learning, we chose the *k* most similar *k* training instances (those with the smallest LD) for each test instance.

Similarly, for the pLM embedding-based representation in the similarity-based strategies, *k*-shot demonstrations were chosen from the training set based on the cosine similarity instead of LD. Specifically, we first used the state-of-the-art pre-trained antibody-specific pLM, AntiBERTy[44], to generate an embedding for each antibody sequence. The resulting embedding for each sequence was then averaged over its length to ensure uniform embedding size across sequences of different lengths, in line with established practices[35,42]. This averaging process was done by the Python script in Wang et al.[35]. Each averaged AntiBERTy embedding (i.e., the pLM embedding) had a dimensionality of 512. Subsequently, *k*-shot demonstrations were selected from the training set based on the highest cosine similarity to the test instance, using the pLM embeddings. To minimize token usage, the pLM embeddings were further rounded to three decimal places and input as vectors in the few-shot prompts. All other prompt structures remained identical to those in the amino acid letter code-based representation.

Prompt design

A prompt refers to a set of instructions and context provided as input to guide the language model's response generation. In most APIs, prompts can be provided to different roles, including "system", "user", and "assistant" roles. The "system" role, included at the start of the prompt, provides context, instructions, or other information to prime the model for the specific use case. The "user" role contains the actual user queries. The "assistant" role contains the model's responses to the user queries, based on the context provided in the "system" role and "user" role. These roles work together, with the "system" message setting the stage, the "user" message providing the actual user queries, and the "assistant" message containing the model's response to the user queries considering the overall context.

All prompts in this work were input into the "user" role, unless otherwise specified. The prompts were designed following previous publications[7,9] and prompt engineering guides from OpenAI[59] and Anthropic[60]. We experimented with different prompt designs,

each containing elements such as task descriptions, examples, and reasoning steps. In all the prompts, a new line was represented by the "\n" character.

Zero-shot

The zero-shot prompt is designed to evaluate LLMs' performance based solely on its pre-trained knowledge, without the aid of any task-specific demonstrations. For the zero-shot prompt, we provided a context-setting instruction to the LLM, asking it to act as an immunology expert, following the template shown in **Fig. 1b** and **Supplementary Fig. 2**.

Few-shot

To construct the few-shot prompt, we added the few-shot demonstrations before the test instance **(Fig. 1b and Supplementary Fig. 1)**. The entire prompt remained the same as zero-shot prompt except the inclusion of few-shot demonstrations **(Supplementary Fig. 1, 2)**.

Few-shot CoT

CoT reasoning steps, based on antibody properties related to antibody characterization, including calculating amino acid frequency and considering amino acid physiochemical properties, suggested by previous literature[47,48,61], were added to the few-shot prompt. We tested three few-shot CoT prompt structures **(Supplementary Fig. 11)**.

For CoT steps alone, we followed Reynolds et al.[47] to construct the prompts. We placed the two following reasoning steps after the few-shot demonstrations: "Let's solve this problem by splitting it into steps.\nStep 1: Please calculate the frequency of each amino acid in the sequence.\nStep 2: Please consider the physiochemical property of each amino acid in the sequence.". These two reasoning steps were kept consistent across all three CoT prompt structures **(Supplementary Fig. 11a)**.

For CoT with LLM-generated answers, we referenced both Wei et al.[48] and Reynolds et al.[47] to construct the prompts. Initially, we provided the first reasoning step, and the LLM generated a response. We then input this response into the "assistant" role along with the second reasoning step into the "user" role. After receiving the LLM's response, we input it again into the "assistant" role and the remaining part of the prompt into the "user" role **(Supplementary Fig. 11b, c)**.

For CoT with manual answers, we manually corrected the LLM-generated answers, as LLM-generated answers may contain errors, such as miscalculating amino acid frequencies, similar with findings from recent studies[13,49] **(Supplementary Fig. 12)**. Unlike the CoT approach with LLM-generated answers, where a query may generate a different response at a time, the reasoning steps in this method remained consistent across all the prompts since we provided the same manually curated answers as the input into the "user" role **(Supplementary Fig. 11d)**.

Once the CoT reasoning steps were complete, the test instances were fed into the "user" role in prompts to generate the final response, using the same prompt structure as before **(Supplementary Fig. 11)**. CoT steps alone were used in the subsequent comparison with zero-shot and few-shot settings **(Fig. 4)** as it showed the best performance among the three CoT variations in in **Supplementary Fig. 13**.

LLMs

Our selected 18 LLMs cover models of different architectures, with different parameter sizes, and from different organizations, including both closed-source and open-source models **(Supplementary Table 1)**. If an unexpected API interruption occurs (e.g., network interruption), the entire prediction process will restart from the beginning. All the LLMs used a temperature of 0.2, and all the other API parameters were set to default unless specified. For each test instance in each setting, we ran its prompt five times to evaluate the consistency of the results.

The details for running each LLM are described as follows:

OpenAI (GPT-3.5, GPT-4.0, GPT-4o, and GPT-4o mini)

The prompt was used to query GPT-3.5, GPT-4o, GPT-4.0, and GPT-4o mini via the OpenAI API (v1.11.0), following the official OpenAI guidelines[62]. For the OpenAI API, the *n* parameter is available for generating output *n* times using a single input prompt. This parameter *n* was used instead of running a single prompt *n* times, reducing the number of API calls required. We set *n* to be 5.

Anthropic (Claude 3 Haiku, Claude 3 Sonnet, Claude 3 Opus, and Claude 3.5 Sonnet)

The prompt was used to query Claude 3 Haiku, Claude 3 Sonnet, Claude 3 Opus, and Claude 3.5 Sonnet via the Anthropic API (v0.21.3) following the official Anthropic guidelines[63].

Google (Gemini 1.0 Pro, Gemini 1.5 Flash, and Gemini 1.5 Pro)

The prompt was used to query Gemini 1.0 Pro, Gemini 1.5 Flash, and Gemini 1.5 Pro via the Google Generative API (v0.5.3) following the official Google guidelines[64].

Meta (LLaMA 2 and LLaMA 3)

For LLaMA 2, the chat models were used instead of the instruct models as the latter failed to provide appropriate answers **(Supplementary Fig. 17)**. For LLaMA 3, the instruct models were used. For Scenario 1a, it should be noted that LLaMA 3 was chosen over LLaMA2 due to the latter's limited context window size **(Supplementary Table 1)**.

The [BOS] (beginning of a sentence) and [EOS] (end of a sentence) special tokens were added to the prompt following the prompt format in the Meta official guidelines[65,66]. As LLaMA 3 instruct models were used, [INST] and [/INST] instruction tokens were also added to the prompts for all LLaMA 3 models. For LLaMA 2 (7B) and LLaMA 3 (8B) models, we ran the experiments both locally through Python Hugging Face pipeline (for Scenario 1b) and via the Replicate API (for Scenario 2 and 3), where other models, namely LLaMA 2 (13B and 70B) and LLaMA 3 (70B), solely via the Replicate API (v0.25.2). We transitioned from Hugging Face pipeline to Replicate API for its faster inference speed in our study. Locally, we generated the response by inputting the prompts to the corresponding models (meta-llama/Llama-2-7b-chat-hf and meta-llama/Meta-Llama-3-8B-Instruct) in the Python Hugging Face transformer (v.4.25.1) library on NVIDIA A100 (40G). All the parameters were set based on the official Meta tutorial[67,68] and the GitHub llama-recipes repository. The *num_return_sequences* parameter is available in the Hugging Face pipeline for generating output multiple times using a single input prompt. For the Replicate API, the formatted prompt was used to query each model to obtain the response.

Mistral AI (Mistral and Mixtral)

We formatted prompts according to the Mistral AI guidelines[69,70], adding the <s> (BOS), [INST] and [/INST] tokens to the Mistral (v0.2) prompt, and the same tokens plus an additional </s> (EOS) token to the Mixtral (8x7b) (v0.1) prompt. Mistral was used both locally via the Python Hugging Face pipeline (for Scenario 1b and 3) and via the Replicate API (for Scenario 2), whereas Mixtral was used solely via the Replicate API. We transitioned from Hugging Face pipeline to Replicate API for its faster inference speed in our study. Locally, we generated the response by inputting the prompt to the corresponding model (mistralai/Mistral-7B-Instruct-v0.2) in the Python Hugging Face transformer (v.4.25.1) library on NVIDIA A100 (40G). The *num_return_sequences* parameter was used in the Hugging Face pipeline to generate output multiple times from a single input prompt. For the Replicate API (v0.25.2), the corresponding formatted prompt was used to query Mistral and Mixtral to obtain the response, respectively.

Model fine-tuning

Fine-tuning prompts

The fine-tuning prompts were created by randomly selecting 1,000, 3,000, or 5,000 instances from the training set. The fine-tuning prompts with 5,000 instances were excluded in Gemini 1.0 Pro due to its API file size limitation. The fine-tuning prompts are illustrated in **Supplementary Fig. 14**.

The details for fine-tuning each LLM are described as follows:

GPT-3.5 and GPT-4o mini

The fine-tuning prompts were formatted according to the OpenAI API guidelines, using the JavaScript Object Notation (JSON) structure: {“messages”: [{“role”: “user”, “content”: input_prompt}, {“role”: “assistant”, “content”: “label”}]} **(Supplementary Fig. 14b)**. The resulting JSON files were uploaded to OpenAI servers to create fine-tuned models using the OpenAI API (v1.11.0). Zero-shot and few-shot prompts were then input to the resulting fine-tuned models for evaluation.

Gemini 1.0 Pro

The fine-tuning prompts were formatted into Comma-Separated Values (CSV) files with column headers “text_input” and “output” added, following the Gemini guidelines **(Supplementary Fig. 14c)**. The resulting CSV files were uploaded to Google AI Studio web interface for fine-tuning. Zero-shot and few-shot prompts were then input to the resulting fine-tuned models for evaluation.

Performance evaluation

LLM response conversion

The majority of the 18 LLMs directly outputted formatted responses with the predicted label. We referred formatted response as response without any additional information other than the answer (e.g., “1”). To convert the responses into predicted labels, we used the following criteria: across all the scenarios, a predicted label of 1 was assigned to formatted responses of “1”, “1 \n”, “1\n” or “*1*”. For unformatted responses **(Supplementary Table 4)**, specific phrases were extracted for conversion. In Scenario 1, a predicted label of 1 was assigned to responses containing “Humanness: 1” or “Answer: 1”. In Scenario 2, a predicted label of 1 was assigned to responses containing “Binding: 1” or “Answer: 1” or containing both “1” and “bind”. In Scenario 3, a predicted label of 1 was assigned to responses containing “IGHD” or “Answer: 1”.

Other labels in Scenario 3 followed the same rationale. The remaining responses were classified as having a predicted label of 0 in all the scenarios.

Evaluation metric

After converting the responses into labels, accuracy was calculated as the evaluation metric by comparing the predictions with the true labels of the test set. Each experiment was repeated five times, and the accuracy was averaged across these five repetitions, with the final average accuracy being reported.

Statistical tests

In all the sections involving statistical tests, unless otherwise specified, we used the non-parametric two-sided paired Wilcoxon signed-rank test to compare distributions of continuous values. Mann-Whitney *U* test was used for unpaired performance comparisons between closed-source and open-source models and in the "Assessing factors influencing ICL performance" section. All *p*-values for multiple comparisons were corrected by the Bonferroni correction.

Performance comparison between zero-shot ICL and few-shot ICL

The performance comparison is depicted in **Supplementary Fig. 9**. In each scenario, for zero-shot ICL, the final accuracy was averaged over 18 (LLMs) × 5 (repetitions) = 90 accuracy scores. For Scenario 1b, the final accuracy of few-shot ICL was averaged over 18 (LLMs) × 6 (demonstration selection strategies) × 2 (16-shot and 32-shot) × 5 (repetitions) = 1080 accuracy scores. For Scenario 2, the final accuracy of few-shot ICL was averaged over 18 (LLMs) × 6 (demonstration selection strategies) × 1 (16-shot) × 5 (repetitions) + 15 (LLMs, excluding three LLaMA 2 models) × 6 (demonstration selection strategies) × 1 (32-shot) × 5 (repetitions) = 990 accuracy scores. For Scenario 3, the final accuracy of few-shot ICL was averaged over 18 (LLMs) × 4 (demonstration selection strategies) × 2 (16-shot and 32-shot) × 5 (repetitions) = 720 accuracy scores.

Comparison with random baselines and established models

We compared ICL with random baselines and established ML models.

Established ML models

Amino acid frequency encoding and physiochemical encoding

Following the Python script in Wang et al.[35], amino acid frequency and physiochemical encodings of each antibody sequence were extracted from the Python Biopython (v1.7.0) and peptides (v0.3.2) packages, respectively. The amino acid frequency encoding resulted in a feature vector with a dimensionality of 20, while the physiochemical encoding produced a feature vector with a dimensionality of 88.

AntiBERTy (pLM) embeddings

We followed the same procedure for generating AntiBERTy embeddings as described in "ICL demonstration selection". Specifically, we first used the state-of-the-art pre-trained antibody-specific pLM AntiBERTy[44] to generate an embedding for each antibody sequence. The resulting embedding for each sequence was then averaged over its length to ensure uniform embedding size across sequences of different lengths, in line with established practices[35,42]. This averaging process was done by the Python script in Wang et al.[35]. Each averaged AntiBERTy embedding (i.e., the pLM embedding we finally used) had a dimensionality of 512.

The training and test sets of the established ML models

The full training set **(Supplementary Table 3)** in each scenario was used for training the established ML models. Our first strategy was to train a model using all the instances in the corresponding full training set. We used the RF classifier (*sklearn.ensemble.RandomForestClassifier*) from the Python scikit-learn (v1.5.0) package, with 1,000 trees and all other hyperparameters set to their default values. The RF model was then tested on exactly the same test set as ICL.

One concern arose from the nature of LLMs in our work: our LLMs outputted binary predictions based on our prompts, prohibiting us from using threshold-free classification evaluation metrics such as the area under the receiver operating characteristic curve (AUROC). In contrast, classifiers such as RF used in this work initially produce continuous prediction scores that need a threshold to be transformed into binary predicted labels. The class with the highest score becomes the model's predicted class by default, meaning in binary classification, the default threshold is 0.5. However, it is often overlooked that the choice of an appropriate threshold is influenced by the training class imbalance and can affect the final result of threshold-dependent metrics like accuracy. To minimize this potential impact, we carefully designed our second strategy. Specifically, for Scenario 2 and 3, which had class imbalances, we randomly downsampled the full training set to create a balanced training set and trained the RF model with the balanced set (Scenario 2: each class contained 6,830 instances; Scenario 3: each class contained 43,364 instances). For Scenarios 2 and 3, we repeated the downsampling process ten times, obtained ten different balanced training sets, trained the RF model on these ten balanced sets, and averaged the performance on the test set **(Supplementary Table 5)**.

The RF models using our second strategy achieved better performance **(Supplementary Table 5)**. To construct a more challenging and rigorous comparison, we used the better performance as the performance of the established models **(Fig. 3)**.

Random baseline

The theoretical random baseline (i.e., random guessing) yields an accuracy of 0.5 for binary classifications and 0.25 for 4-class classifications due to the class balance of our test set. To empirically confirm the accuracy, we constructed two baselines using the dummy classifier function (*sklearn.dummy.DummyClassifier*) from the Python scikit-learn (v1.5.0) package. The "stratified" baseline predicts class labels based on the observed frequency of each class in the training set, while the "uniform" baseline assigns equal probability to each class. We adopted the two strategies used in the established models. For the first strategy, we further averaged the predictions over 100 random seeds (defined by the *random_state* hyperparameter of the dummy classifier function). For the second strategy, we followed the same procedure as that of the established models. The empirical accuracy aligned closely with the theoretical random baseline accuracy **(Supplementary Table 5)**; therefore, we used the theoretical random baseline as the random baseline in all experiments.

Majority voting classifier for label bias evaluation

To evaluate whether LLMs are likely to degenerate into a majority voting classifier when faced with imbalanced label distributions of the demonstrations in few-shot ICL, we created a classifier that always predicts the most frequent label in the demonstrations and compared its performance to LLMs'. We treated the LLM's predictions as the true labels and the majority voting classifier's predictions as the predicted labels. Higher accuracy indicates greater alignment between the LLM and majority vote predictions, while lower accuracy suggests deviation. This approach offers a quantitative assessment of whether LLMs rely heavily on majority label bias in their decision-making or consider additional information beyond the majority vote in the demonstration labels.

Reproducibility and stability evaluation

We refer to "reproducibility" as obtaining identical results using the same data, code, and hyperparameters, and "stability" as the model's performance not being overly sensitive to changes in hyperparameter values. Unless otherwise specified, the following sections used the Similar (Balanced, Random) strategy for all prompt tests in Scenario 1a without MSA.

Prompt input order

We evaluated if LLMs' responses would be affected by two different input orders: cycling and sequential **(Fig. 1a)**. For example, if we have two test instances, each associated with a corresponding prompt, the cycling input order involves alternating between the first and second prompts, repeating this sequence five times. In contrast, the sequential input order consists of inputting the first prompt five times in a row, followed by the second prompt five times. We applied these two input orders across all test instances in the test set.

Data contamination

To check for data contamination, following Li et al.[51], we selected the data generated by a subset of studies that were published in 2022 from Scenario 2. The same train-test split procedure was applied to this subset as described in the previous section. The selected data may overlap with the training data of GPT-4 (20240409) and GPT-4 (20240125), as both models have a training data cut-off of December 2023. However, it does not overlap with GPT-4 (20230613), which has a training data cut-off in September 2021. The performance of models trained with different training data cut-off was compared.

Word substitution test

The prompts containing "physiochemical" were compared to those containing "physicochemical", while all other words in the prompts remained the same.

Model-related hyperparameters

Model version

We used Scenario 1a with MSA to evaluate different model versions. The OpenAI API offers multiple versions of the same LLM, which typically vary in terms of training data and specific model settings **(Supplementary Table 2)**. For instance, GPT-4 (20240409) and GPT-4 (20240125) represent two distinct versions of the same LLM.

Seed

For the different seed test, five distinct seeds (set using the OpenAI API parameter *seed*) were used, each generating five responses with the parameter *n*=5 (for generating response *n* times), resulting in 25 responses. For comparison, a new seed, different from the previous five, was used five times, generating five responses each time using the same parameter *n*=5, for a total of 25 responses.

Temperature

In the temperature test, GPT-3.5 was evaluated with temperature values of 0.2, 0.4, 0.6, 0.8, and 1.0 (set by the OpenAI API parameter *temperature*) to examine how temperature variations affect the model's output.

Data availability

All data used in this study can be obtained following the procedure described in the "Dataset collection and processing" part of the Methods section. We have submitted all the processed data used in our study for the review process. We will make all the processed data available in FigShare or Zenodo repositories.

Code availability

We have submitted all the code used in our study for the review process. We provided a Jupyter Notebook for better accessibility. We will create a FigShare repository to allow free public access to our code.

Acknowledgments

QC is supported by National Natural Science Foundation of China under Award Number 32100515 and CUHK direct grant for research under Award Numbers 2021.061 and 2022.080. KYY is supported by National Cancer Institute of the National Institutes of Health under Award Numbers P30CA030199 and R01CA287114, National Institute on Aging of the National Institutes of Health under Award Numbers R01AG085498 and U54AG079758, and internal grants of Sanford Burnham Prebys Medical Discovery Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Contributions

KYY, SKWT, and QC conceived and supervised the project. SHF, KYY, SKWT, and QC designed the computational experiments and data analyses. SHF prepared the data, implemented the methods, conducted the experiments, and analyzed the results. CM and BSHW independently reproduced parts of the results. All the authors interpreted the results. SHF, KYY, SKWT, and QC wrote the manuscript.

Corresponding authors

Correspondence to Kevin Y. Yip, Stephen Kwok-Wing Tsui or Qin Cao.

Ethics declarations

Ethics approval and consent to participate

No ethical approval was required for this study. All utilized public datasets were generated by other organizations that obtained ethical approval.

Competing interests

The authors declare that they have no competing interests.

References

1. Brown, T. B. *et al.* Language Models are Few-Shot Learners. *Adv Neural Inf Process Syst* **33**, 1877–1901 (2020).

2. Jahan, I., Laskar, M. T. R., Peng, C. & Huang, J. X. A comprehensive evaluation of large Language models on benchmark biomedical text processing tasks. *Comput Biol Med* **171**, 108189 (2024).

3. Xie, S. M., Raghunathan, A., Liang, P. & Ma, T. An Explanation of In-context Learning as Implicit Bayesian Inference. *ICLR 2022 - 10th International Conference on Learning Representations* (2021).

4. Dai, D. *et al.* Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* 4005–4019 (2022) doi:10.18653/v1/2023.findings-acl.247.

5. von Oswald, J. *et al.* Transformers learn in-context by gradient descent. *Proc Mach Learn Res* **202**, 35151–35174 (2022).

6. Min, S. *et al.* Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022* 11048–11064 (2022) doi:10.18653/v1/2022.emnlp-main.759.

7. Cai, X. *et al.* Comprehensive evaluation of molecule property prediction with ChatGPT. *Methods* **222**, 133–141 (2024).

8. Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A. & Smit, B. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence 2024 6:2* **6**, 161–169 (2024).

9. Guo, T. *et al.* What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks. *Adv Neural Inf Process Syst* **36**, (2023).

10. Hou, W. & Ji, Z. GeneTuring tests GPT models in genomics. *bioRxiv* 2023.03.11.532238 (2023) doi:10.1101/2023.03.11.532238.

11. Jin, Q., Yang, Y., Chen, Q. & Lu, Z. GeneGPT: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics* **40**, (2024).

12. Hou, W. & Ji, Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nature Methods 2024* 1–4 (2024) doi:10.1038/s41592-024-02235-4.

13. The Impact of Large Language Models on Scientific Discovery: a Preliminary Study using GPT-4. (2023).

14. Chen, Y. & Zou, J. GenePT: A Simple But Effective Foundation Model for Genes and Cells Built From ChatGPT. *bioRxiv* (2024) doi:10.1101/2023.10.16.562533.

15. Li, T. *et al.* CancerGPT for few shot drug pair synergy prediction using large pretrained language models. *npj Digital Medicine 2024 7:1* **7**, 1–10 (2024).

16. Beck, A., Wagner-Rousset, E., Ayoub, D., Van Dorsselaer, A. & Sanglier-Cianférani, S. Characterization of therapeutic antibodies and related products. *Anal Chem* **85**, 715–736 (2013).

17. Kim, J., McFee, M., Fang, Q., Abdin, O. & Kim, P. M. Computational and artificial intelligence-based methods for antibody development. *Trends Pharmacol Sci* **44**, 175–189 (2023).

18. Bravi, B. Development and use of machine learning algorithms in vaccine target selection. *npj Vaccines 2024 9:1* **9**, 1–14 (2024).

19. Sormanni, P., Aprile, F. A. & Vendruscolo, M. Third generation antibody discovery methods: in silico rational design. *Chem Soc Rev* **47**, 9137–9157 (2018).

20. Norman, R. A. *et al.* Computational approaches to therapeutic antibody design: established methods and emerging trends. *Brief Bioinform* **21**, 1549–1567 (2020).

21. Li, L. *et al.* Machine learning optimization of candidate antibody yields highly diverse sub-nanomolar affinity antibody libraries. *Nature Communications 2023 14:1* **14**, 1–12 (2023).

22. Makowski, E. K. *et al.* Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. *Nature Communications 2022 13:1* **13**, 1–14 (2022).

23. Mason, D. M. *et al.* Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nature Biomedical Engineering 2021 5:6* **5**, 600–612 (2021).

24. Ramon, A. *et al.* Assessing antibody and nanobody nativeness for hit selection and humanization with AbNatiV. *Nature Machine Intelligence 2024 6:1* **6**, 74–91 (2024).

25. Marks, C., Hummer, A. M., Chin, M. & Deane, C. M. Humanization of antibodies using a machine learning approach on large-scale repertoire data. *Bioinformatics* **37**, 4041–4047 (2021).

26. OpenAI *et al.* GPT-4 Technical Report. (2023).

27. *The Claude 3 Model Family: Opus, Sonnet, Haiku Anthropic*. https://docs.anthropic.com/.

28. Team, G. *et al.* Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. (2024).

29. Touvron, H. *et al.* Llama 2: Open Foundation and Fine-Tuned Chat Models. (2023).

30. Dubey, A. *et al.* The Llama 3 Herd of Models. (2024).

31. Jiang, A. Q. *et al.* Mistral 7B. (2023).

32. Jiang, A. Q. *et al.* Mixtral of Experts. (2024).

33. Zheng, L. *et al.* Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.

34. Chiang, W.-L. *et al.* Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference.

35. Wang, M., Patsenker, J., Li, H., Kluger, Y. & Kleinstein, S. H. Language model-based B cell receptor sequence embeddings can effectively encode receptor specificity. *Nucleic Acids Res* **52**, 548–557 (2024).

36. Tennenhouse, A. *et al.* Computational optimization of antibody humanness and stability by systematic energy-based ranking. *Nature Biomedical Engineering 2023 8:1* **8**, 30–44 (2023).

37. Robert, P. A. *et al.* Unconstrained generation of synthetic antibody–antigen structures to guide machine learning methodology for antibody specificity prediction. *Nature Computational Science 2022 2:12* **2**, 845–865 (2022).

38. Burton, D. R. What Are the Most Powerful Immunogen Design Vaccine Strategies? Reverse Vaccinology 2.0 Shows Great Promise. *Cold Spring Harb Perspect Biol* **9**, (2017).

39. Liu, J. *et al.* What Makes Good In-Context Examples for GPT-3? *DeeLIO 2022 - Deep Learning Inside Out: 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, Proceedings of the Workshop* 100–114 (2022) doi:10.18653/V1/2022.DEELIO-1.10.

40. Hershberg, U. & Luning Prak, E. T. The analysis of clonal expansions in normal and autoimmune B cell repertoires. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**, (2015).

41. Miho, E., Roškar, R., Greiff, V. & Reddy, S. T. Large-scale network analysis reveals the sequence space architecture of antibody repertoires. *Nature Communications 2019 10:1* **10**, 1–11 (2019).

42. Elnaggar, A. *et al.* ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans Pattern Anal Mach Intell* **44**, 7112–7127 (2022).

43. Li, T. *et al.* Long-context LLMs Struggle with Long In-context Learning. (2024).

44. Ruffolo, J. A., Gray, J. J. & Sulam, J. Deciphering antibody affinity maturation with language models and weakly supervised learning. (2021).

45. Agrawal, S., Zhou, C., Lewis, M., Zettlemoyer, L. & Ghazvininejad, M. In-context Examples Selection for Machine Translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* 8857–8873 (2022) doi:10.18653/v1/2023.findings-acl.564.

46. Lu, Y., Bartolo, M., Moore, A., Riedel, S. & Stenetorp, P. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* **1**, 8086–8098 (2021).

47. Reynolds, L. & McDonell, K. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. *Conference on Human Factors in Computing Systems - Proceedings* (2021) doi:10.1145/3411763.3451760.

48. Wei, J. *et al.* Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Adv Neural Inf Process Syst* **35**, (2022).

49. Zhou, L. *et al.* Larger and more instructable language models become less reliable. *Nature 2024* 1–8 (2024) doi:10.1038/s41586-024-07930-y.

50. Sprague, Z. *et al.* To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. (2024).

51. Li, C. & Flanigan, J. Task Contamination: Language Models May Not Be Few-Shot Anymore. *Proceedings of the AAAI Conference on Artificial Intelligence* **38**, 18471–18480 (2023).

52. Olsen, T. H., Boyles, F. & Deane, C. M. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science* **31**, 141–146 (2022).

53. Naveed, H. *et al.* A Comprehensive Overview of Large Language Models. (2023).

54. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature 2024 630:8016* **630**, 493–500 (2024).

55. Qiao, S. *et al.* Reasoning with Language Model Prompting: A Survey. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* **1**, 5369–5393 (2022).

56. Kossen, J., Gal, Y. & Rainforth, T. In-Context Learning Learns Label Relationships but Is Not Conventional Learning. *ICLR 2024 - 12th International Conference on Learning Representations* (2023).

57. Honegger, A. & Plückthun, A. Yet Another Numbering Scheme for Immunoglobulin Variable Domains: An Automatic Modeling and Analysis Tool. *J Mol Biol* **309**, 657–670 (2001).

58. Dunbar, J. & Deane, C. M. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics* **32**, 298–300 (2016).

59. Prompt engineering - OpenAI API. https://platform.openai.com/docs/guides/prompt-engineering.

60. Prompt engineering overview - Anthropic. https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/overview.

61. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large Language Models are Zero-Shot Reasoners. *Adv Neural Inf Process Syst* **35**, (2022).

62. API Reference - OpenAI API. https://platform.openai.com/docs/api-reference/chat.

63. Create a Message - Anthropic. https://docs.anthropic.com/en/api/messages.

64. Gemini API reference | Google AI for Developers. https://ai.google.dev/api?lang=python.

65. Meta Llama 2 | Model Cards and Prompt formats. https://llama.meta.com/docs/model-cards-and-prompt-formats/meta-llama-2/.

66. Llama 3 | Model Cards and Prompt formats. https://llama.meta.com/docs/model-cards-and-prompt-formats/meta-llama-3/.

67. 5 Steps to Getting Started with Llama 2. https://ai.meta.com/blog/5-steps-to-getting-started-with-llama-2/.

68. meta-llama/Meta-Llama-3-8B · Hugging Face. https://huggingface.co/meta-llama/Meta-Llama-3-8B.

69. mistralai/Mistral-7B-Instruct-v0.2 · Hugging Face. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2.

70. mistralai/Mixtral-8x7B-Instruct-v0.1 · Hugging Face. https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1.

Supplementary Information

Supplementary Table 1: Model and cost summary (Official API, as of 2024/08/04)

Model		Organization	Multiple output*	Fine-tune	Open-source	Context window size (tokens)	Input cost (per 1M tokens)	Output cost (per 1M tokens)	Knowledge cut off date
GPT									
GPT-3.5 (gpt-3.5-turbo-0125)		OpenAI	✓	√	✗	16K	$0.5	$1.5	Sep 2021
GPT-4 (gpt-4-turbo-2024-04-09)				✗		128K	$10	$30	Dec 2023
GPT-4o (gpt-4o-2024-05-13)							$5	$15	Oct 2023
GPT-4o mini (gpt-4o-mini-2024-07-18)				√			$0.15	$0.6	
Claude									
Claude 3 Haiku (claude-3-haiku-20240307)		Anthropic	✗	✗	✗	200K	$0.25	$1.25	Aug 2023
Claude 3 Sonnet (claude-3-sonnet-20240229)							$3	$15	
Claude 3 Opus (claude-3-opus-20240229)							$15	$75	
Claude 3.5 Sonnet (claude-3-5-sonnet-20240620)							$3	$15	Apr 2024
LLaMA									
LLaMA 2	7B	Meta	✓	✓	✓	4K	N/A	N/A	Sep 2022
	13B								
	70B								
LLaMA 3	8B					8K			Mar 2023
	70B								
Gemini									
Gemini 1.0 Pro (gemini-1.0-pro-001)		Google	✗	✓	✗	32K	$0.5	$1.5	Apr 2023
Gemini 1.5 Flash (gemini-1.5-flash-001)						1M	$0.35	$0.7	Nov 2023
Gemini 1.5 Pro (gemini-1.5-pro-001)						1M	$7	$21	
Mistral									
Mistral (v0.2) (Mistral-7B-instruct-v0.2)		Mistral AI	✓	✓	✓	32K	N/A	N/A	Dec 2023
Mixtral (v0.1) (Mixtral-8x7B-instruct-v.0.1)									

* For the OpenAI API (for GPT), parameter *n* is used. For LLaMA and Mistral models, parameter *num_return_sequences* in huggingface pipelines are used.

Supplementary Table 2: GPT model variations' knowledge cutoff date and context window size

Model	Knowledge cutoff date	Context window size
GPT-3.5		
GPT-3.5 (20240125) (gpt-3.5-turbo-0125)	Sep 2021	16K
GPT-3.5 (20231106) (gpt-3.5-turbo-1106)		
GPT-3.5 (20230613) (gpt-3.5-turbo-0613)		
GPT-4		
GPT-4 (20240409) (gpt-4-turbo-2024-04-09)	Dec 2023	128K
GPT-4 (20240125) (gpt-4-0125-preview)		
GPT-4 (20231106) (gpt-4-1106-preview)	Apr 2023	
GPT-4 (20230613) (gpt-4-0613)	Sep 2021	8K
GPT-4o		
GPT-4o (gpt-4o-2024-05-13)	Oct 2023	128K
GPT-4o mini		
GPT-4o mini (gpt-4o-mini-2024-07-18)	Oct 2023	128K

Supplementary Table 3 Summary of the training and test sets in three antibody characterization scenarios.

Training set	Number of instances				
	Class 0	Class 1	Class 2	Class 3	Total number of instances
Scenario 1a	9,950 (Mouse)	9,950 (Human)			19,900
Scenario 1b	9,950 (Rhesus)	9,950 (Human)			19,900
Scenario 2	6,830 (Non-binder)	8,608 (Binder)			15,438
Scenario 3	124,947 (IGHA)	43,364 (IGHD)	260,994 (IGHG)	384,573 (IGHM)	813,878
Test set	Number of instances				
Scenario 1a	50 (Mouse)	50 (Human)			100
Scenario 1b	50 (Rhesus)	50 (Human)			100
Scenario 2	50 (Non-binder)	50 (Binder)			100
Scenario 3	25 (IGHA)	25 (IGHD)	25 (IGHG)	25 (IGHM)	100

Supplementary Table 4 Summary of LLMs' response across various prompts and scenarios. Formatted responses refer to those containing only the answer with no extra information (e.g., "1", "1\n"). The truncated responses in some LLMs were due to the restricted number of output tokens set by the *max_new_tokens* parameter recommended in each LLM's user guidelines.

(See the attached PDF file)

Supplementary Table 5 Performance of random baselines and established ML models in three antibody characterization scenarios using full and balanced training sets. Note that the original training sets in Scenario 1a and 1b are balanced.

Scenario 1a (Binary): Antibody humanness (Mouse vs. Human) prediction

	Random baselines			Established ML models		
	Theoretical	Stratified	Uniform	Amino acid frequency encoding + RF	Physiochemical encoding + RF	AntiBERTy + RF
Full training set	0.5	0.499	0.506	0.96	0.93	1.0

Scenario 1b (Binary): Antibody humanness (Rhesus vs. Human) prediction

	Random baselines			Established ML models		
	Theoretical	Stratified	Uniform	Amino acid frequency encoding + RF	Physiochemical encoding + RF	AntiBERTy + RF
Full training set	0.5	0.499	0.506	0.89	0.87	1.0

Scenario 2 (Binary): Antibody specificity prediction

	Random baselines			Established ML models		
	Theoretical	Stratified	Uniform	Amino acid frequency encoding + RF	Physiochemical encoding + RF	AntiBERTy + RF
Full training set	0.5	0.499	0.499	0.640	0.690	0.820
Balanced training set		0.502	0.494	0.709	0.719	0.845

Scenario 3 (Multiclass): Antibody isotype prediction

	Random baselines			Established ML models		
	Theoretical	Stratified	Uniform	Amino acid frequency encoding + RF	Physiochemical encoding + RF	AntiBERTy + RF
Full training set	0.25	0.251	0.244	0.45	0.43	0.46
Balanced training set		0.244	0.244	0.462	0.49	0.492

Few-shot (2-shot), Similar (Balanced, Positive-Negative)

Scenario 1b (Binary): Antibody humanness (Rhesus vs. Human) prediction

You are an AI model specializing in immunology. Your task is to **predict the humanness of antibody** using only the amino acid sequences provided. Recent research has suggested that **the humanness of antibody** can be predicted with antibody sequence alone, without any structural information.
Please strictly follow the format, no other information can be provided. Given the amino acid sequences of an antibody, predict **the humanness of antibody based on its sequence, by analyzing whether it can Human(1) or Rhesus(0).** Consider factors such as physiochemical property, PSSM and amino acid frequency to assess **the humanness of antibody**.
Please answer with *only* 1 or 0 only for the last example. A few examples are provided in the beginning.
Antibody:
QVQLQESGPGLLKPSETLSLTCSVSGGSITSSTYSWGWIRQPPGKGLEWIGSVFYNGDTYYNPSLKSRVTVSVDTSKNQVSLKLSSVTAADTAVYYCTRASIFG
VALDAFDLWGQGTMVSVSS
Humanness: 1
Antibody:
DVQLVESGGGLVQPGGSLRLSCAASGFTFSSYGMSWVRQAPGKGLDWVSYISNVGGITYYADSVKGRFTISRDNSKNTLSLQMNSLRAEDTAVYYCAKGRYY
YDSGFYKPFDYWGQGVLVTVSS
Humanness: 0

Now, predict **the humanness of antibody** for the following antibody sequence based on the criteria above:
Antibody:
QVQLQQWGEGLVKPSETLSLTCAVYGGSVSGYWWGWIRPPPGKGLEWIGRIRSGGSTNDTPSPKSRVTISIDTSKNQFSLKLSSVTAADTAVYYCARLPTAAAL
DYWGQGVLVTVSS
Humanness:

Scenario 2 (Binary): Antibody specificity prediction

You are an AI model specializing in immunology. Your task is to **predict the binding specificity of antibody to the SARS-CoV-2 spike protein** using only the amino acid sequences provided. Recent research has suggested that **the specificity to antigens of antibody** can be predicted with antibody sequence alone, without any structural information.
Please strictly follow the format, no other information can be provided. Given the amino acid sequences of an antibody, predict **the receptor binding specificity of antibody based on its sequence, by analyzing whether it can bind(1) or cannot bind(0).** Consider factors such as physiochemical property, PSSM and amino acid frequency to assess **the binding of antibody.**
Please answer with *only* 1 or 0 only for the last example. A few examples are provided in the beginning.
Antibody:
QVQLVQSGAEVKKPGASVKVSCKASGYTFTSYGISWLRQAPGQGLEWMGWISAYNGDTNYAPKLQGKVTMTTDTSTSTAYMELRSLRSDDTAVYYCARD
LPKISIFGVADYWGQGTLVTVSSDIQMTQSPSSLSASVGDRVTITCRASQSISRYLNWYQQKPGKAPKLLIQAASSLQSGVPSRFSGSGSGTDFTLTISSLQP
EDFATYYCQQSYSTPFTFGPGTKVDIK
Binding: 1
Antibody:
EVQLVESGGGLVQPGGSLRLSCAASGFTFSSYWMHWVRQAPGKGLVWVSRINSDGNSTSYADSVKGRFTISRDNAKNTLYLQMNSLRAEDTAVYYCARD
LPPYYDFWSGLERGMDVWGQGTTVTVSSQSVLTQPPSASGTPGQRVTISCSGSSSNIGSNYVYWYQQLPGTAPKLLIYRNNQRPSGVPDRFSGSKSGTSA
SLAISGLRSEDEADYYCAAWDDSLSGYVFGTGTKVTVL
Binding: 0

Now, predict **the receptor binding specificity** for the following antibody sequence based on the criteria above:
Antibody:
QVQLQESGPGLVKPSETLSLTCTVSGYSISSGYYWGWIRQPPGKGLEWIGSIYHCGSTYYNPSLKSRLTISVDTSKNQFSLKLSSVTAADTAVYYCAREPDY
SDSGGYYSLYYFAYWGQGTLVTVSSSHERTQPPSVSVSPGQTARITCSGDALPKQYAYWYQQKPGQAPVVVIYKDTERPSGIPERISGSSSGTTVTLTISGV
QAEDEADYYCQSADSSGTYVVFGGGTKLTVL
Binding:

Few-shot (4-shot), Similar (Balanced, Positive-Negative)

Scenario 3 (Multiclass): Antibody isotype prediction

You are an AI model specializing in immunology. Your task is to **predict the heavy chain isotype** using only the amino acid sequences provided. Recent research has suggested that **the heavy chain isotype of antibody** can be predicted with antibody sequence alone, without any structural information.
Please strictly follow the format, no other information can be provided. Given the amino acid sequences of an antibody, predict **the heavy chain isotype of antibody based on its sequence, by analyzing whether it is IGHA (0), IGHD (1), IGHG (2), IGHM (3)**. Consider factors such as physiochemical property, PSSM and amino acid frequency to assess **the heavy chain isotype.**
Please answer with *only* 0, 1, 2 or 3 for the last example. A few examples are provided in the beginning.
Antibody:
QVQLVQSGAEVKKPGASVKVSCKVSGYTLTDLSLHWVRQAPGKGLEWMGGFDPEDEDTIYAQNFKGRLTMTEDTSTDTAYMELSSLRSEDTAVYYCATNQP
TGWELLGDHAFDIWGQGTVVTVSS
Heavy chain isotype: 0
Antibody:
EVQLLESGGGLVQPGGSLRLSCAASGFTFSSYAMSWVRQAPGKGLEWVSAISGSGGSTYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCAKDDSS
GYPNWFDPWGQGTLVTVSS
Heavy chain isotype: 1
Antibody:
EVQLVESGGGLVQPGGSLRVSCAASGFTFSNYAMHWVRQAPGKGLEYISAITSKGAKSFYANSVTGRFNISRDNSKNTLYLQMGSLRVEDTAIYYCVRGDDGT
GYWDYFDYWGQGTLVTVSS
Heavy chain isotype: 2
Antibody:
QVQLVESGGGVVQPGRSLRLSCAASGFTFSSYAMHWVRQAPGKGLEWVAVISYDGSNKYYADSVKGRFTISRDNSKNTLYLQMNSLRAEDTAVYYCARDLR
QLERFVSGNLDYWGQGTLVTVSS
Heavy chain isotype 3

:

Now, predict **the heavy chain isotype** for the following antibody sequence based on the criteria above:
Antibody:
EVQLVESGGGLVQPGGSLRLSCAASGFTFSSYWMSWVRQAPGKGLEWVANIKQDGSEKYYVDSVKGRFTISRDNAKNSLYLQMNSLRAEDTAVYYCARPAT
PTYNWNDAGWFDPWGQGTLVTVSS
Heavy chain isotype:

Supplementary Fig. 1 Examples of few-shot prompts in three antibody characterization scenarios. The text highlighted in this figure shows the variable inputs specific to each scenario. All other text remains constant across the examples.

Zero-shot

Scenario 1b (Binary): Antibody humanness (Rhesus vs. Human) prediction

You are an AI model specializing in immunology. Your task is to **predict the humanness of antibody** using only the amino acid sequences provided. Recent research has suggested that **the humanness of antibody** can be predicted with antibody sequence alone, without any structural information.
Please strictly follow the format, no other information can be provided. Given the amino acid sequences of an antibody, predict **the humanness of antibody based on its sequence, by analyzing whether it can Human(1) or Rhesus(0)**. Consider factors such as physiochemical property, PSSM and amino acid frequency to assess **the humanness of antibody.**
Please answer with *only* 1 or 0 only for the last example.
Antibody:
QVQLQQWGEGLVKPSETLSLTCAVYGGSVSGYWWGWIRPPPGKGLEWIGRIRSGGSTNDTPSPKSRVTISIDTSKNQFSLKLSSVTAADTAVYYCARLPTAAA
LDYWGQGVLVTVSS
Humanness:

Scenario 2 (Binary): Antibody specificity prediction

You are an AI model specializing in immunology. Your task is to **predict the binding specificity of antibody to the SARS-CoV-2 spike protein** using only the amino acid sequences provided. Recent research has suggested that **the specificity to antigens of antibody** can be predicted with antibody sequence alone, without any structural information.
Please strictly follow the format, no other information can be provided. Given the amino acid sequences of an antibody, predict **the receptor binding specificity of antibody based on its sequence, by analyzing whether it can bind(1) or cannot bind(0).** Consider factors such as physiochemical property, PSSM and amino acid frequency to assess **the binding of antibody.**
Please answer with *only* 1 or 0 only for the last example.
Antibody:
QVQLQESGPGLVKPSETLSLTCTVSGYSISSGYYWGWIRQPPGKGLEWIGSIYHCGSTYYNPSLKSRLTISVDTSKNQFSLKLSSVTAADTAVYYCAREPDYSD
SGGYYSLYYFAYWGQGTLVTVSSSHERTQPPSVSVSPGQTARITCSGDALPKQYAYWYQQKPGQAPVVVIYKDTERPSGIPERISGSSSGTTVTLTISGVQAED
EADYYCQSADSSGTYVVFGGGTKLTVL
Binding:

Scenario 3 (Multiclass): Antibody isotype prediction

You are an AI model specializing in immunology. Your task is to **predict the heavy chain isotype** using only the amino acid sequences provided. Recent research has suggested that **the heavy chain isotype of antibody** can be predicted with antibody sequence alone, without any structural information.
Please strictly follow the format, no other information can be provided. Given the amino acid sequences of an antibody, predict **the heavy chain isotype of antibody based on its sequence, by analyzing whether it is IGHA (0), IGHD (1), IGHG (2), IGHM (3).** Consider factors such as physiochemical property, PSSM and amino acid frequency to assess **the heavy chain isotype.**
Please answer with *only* 0, 1, 2 or 3 for the last example.
Antibody:
EVQLVESGGGLVQPGGSLRLSCAASGFTFSSYWMSWVRQAPGKGLEWVANIKQDGSEKYYVDSVKGRFTISRDNAKNSLYLQMNSLRAEDTAVYYCARPAT
PTYNWNDAGWFDPWGQGTLVTVSS
Heavy chain isotype:

Supplementary Fig. 2 Examples of zero-shot prompts in three antibody characterization scenarios. The text highlighted in this figure shows the variable inputs specific to each scenario. All other text remains constant across the examples.

Few-shot Chain-of-Thought (CoT) (32-shot), Similar (Balanced, Random)

Scenario 1b (Binary): Antibody humanness (Rhesus vs. Human) prediction

You are an AI model specializing in immunology. Your task is **to predict the humanness of antibody** using only the amino acid sequences provided. Recent research has suggested that **the humanness of antibody** can be predicted with antibody sequence alone, without any structural information.
Please strictly follow the format, no other information can be provided. Given the amino acid sequences of an antibody, predict **the humanness of antibody based on its sequence, by analyzing whether it can Human(1) or Rhesus(0).** Consider factors such as physiochemical property, PSSM and amino acid frequency to assess **the humanness of antibody.**
Please answer with *only* 1 or 0 only for the last example. A few examples are provided in the beginning.
Antibody:
QVQLQQWGEGLVKPSETLSLTCAVYGGSVSGYWWGWIRQPPGKGRGWFGRIRSGGRTNSNPSLKSRVTSSIDTSKNQFSLKLSSVTAADTAVYYCARPTGN
TVTYWGQGVLVTVSS
Humanness: 0
….
Antibody:
QVQLQQWGAGLLKPSETLSLTCAVYGGSFSGYYWSWIRQPPGKGLEWIGEINHSGSTNYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCARGGGKAG
WFDPWGQGTLVTVSS
Humanness: 1
Let's solve this problem by splitting it into steps.
Step 1: Please calculate the frequency of each amino acid in the sequence.
Step 2: Please consider the physiochemical property of each amino acid in the sequence.
Now, predict **the humanness of antibody** for the following antibody sequence based on the criteria above:
Antibody:
QVQLQQWGEGLVKPSETLSLTCAVYGGSVSGYWWGWIRPPPGKGLEWIGRIRSGGSTNDTPSPKSRVTISIDTSKNQFSLKLSSVTAADTAVYYCARLPTAAAL
DYWGQGVLVTVSS
Humanness:

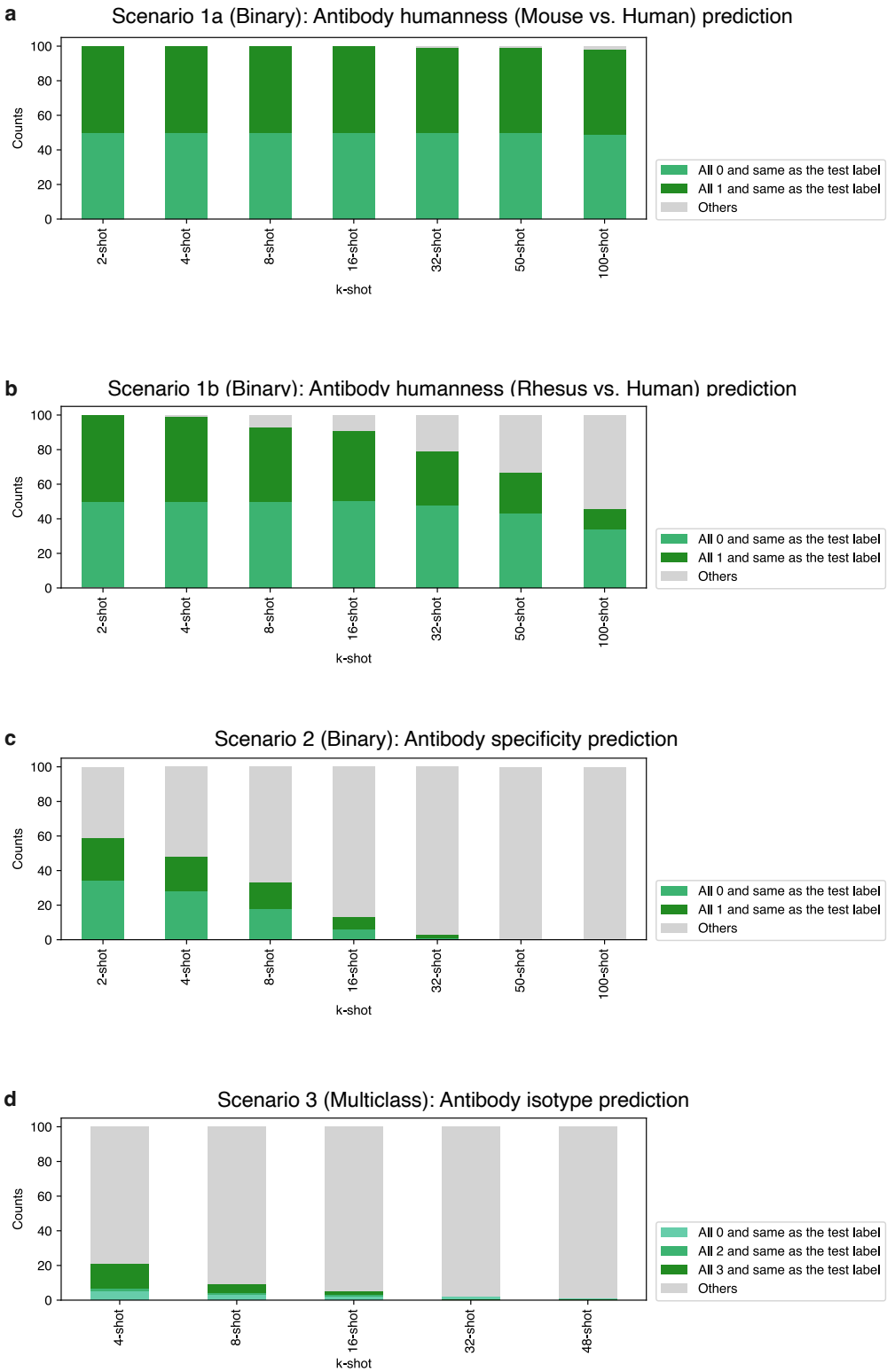
Scenario 2 (Binary): Antibody specificity prediction

You are an AI model specializing in immunology. Your task is **to predict the binding specificity of antibody to the SARS-CoV-2 spike protein** using only the amino acid sequences provided. Recent research has suggested that **the specificity to antigens of antibody** can be predicted with antibody sequence alone, without any structural information.
Please strictly follow the format, no other information can be provided. Given the amino acid sequences of an antibody, predict **the receptor binding specificity of antibody based on its sequence, by analyzing whether it can bind(1) or cannot bind(0).** Consider factors such as physiochemical property, PSSM and amino acid frequency to assess **the binding of antibody.**
Please answer with *only* 1 or 0 only for the last example. A few examples are provided in the beginning.
Antibody:
QLQLQESGPGLVKPSETLSLTCTVSGGSISSSSYYWGWIRQPPGKGLEWIGSIYYSGSTYYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCARRVITIFG
GAGFDYWGQGTLVTVSSSYELTQPPSVSVSPGQTARITCSADALPKQYAYWYQQKPGQAPVLVIYKDSERPSGIPERFSGSSSGTTVTLTISGVQAEDEADYYC
QSADSSGTYVVFGGGTKLTVL
Binding: 0
…
Antibody:
QVPLQESGPGLVKPSETLSLTCTVSGGSINSYYWNWIRQAPGKGLEWIGNIYYSGSTSYNPSLKSRVTISVDTSKNQFSLKLTSVTAADTAVYYCARGGFYDPYY
FDSWGQGTLVTVSSSFELTQPPSVSVSPGQTARITCSANALPMQYAYWYQQKPGQAPVLVIYKDIERASGIPERFSGSSSGATVTLTISGVQAEDEADYYCQSP
DSSGSYVVFGGGTKVTVL
Binding: 1
Let's solve this problem by splitting it into steps.
Step 1: Please calculate the frequency of each amino acid in the sequence.
Step 2: Please consider the physiochemical property of each amino acid in the sequence.
Now, predict **the receptor binding specificity** for the following antibody sequence based on the criteria above:
Antibody:
QVQLQESGPGLVKPSETLSLTCTVSGYSISSGYYWGWIRQPPGKGLEWIGSIYHCGSTYYNPSLKSRLTISVDTSKNQFSLKLSSVTAADTAVYYCAREPDYSDS
GGYYSLYYFAYWGQGTLVTVSSSHERTQPPSVSVSPGQTARITCSGDALPKQYAYWYQQKPGQAPVVVIYKDTERPSGIPERISGSSSGTTVTLTISGVQAEDE
ADYYCQSADSSGTYVVFGGGTKLTVL
Binding:

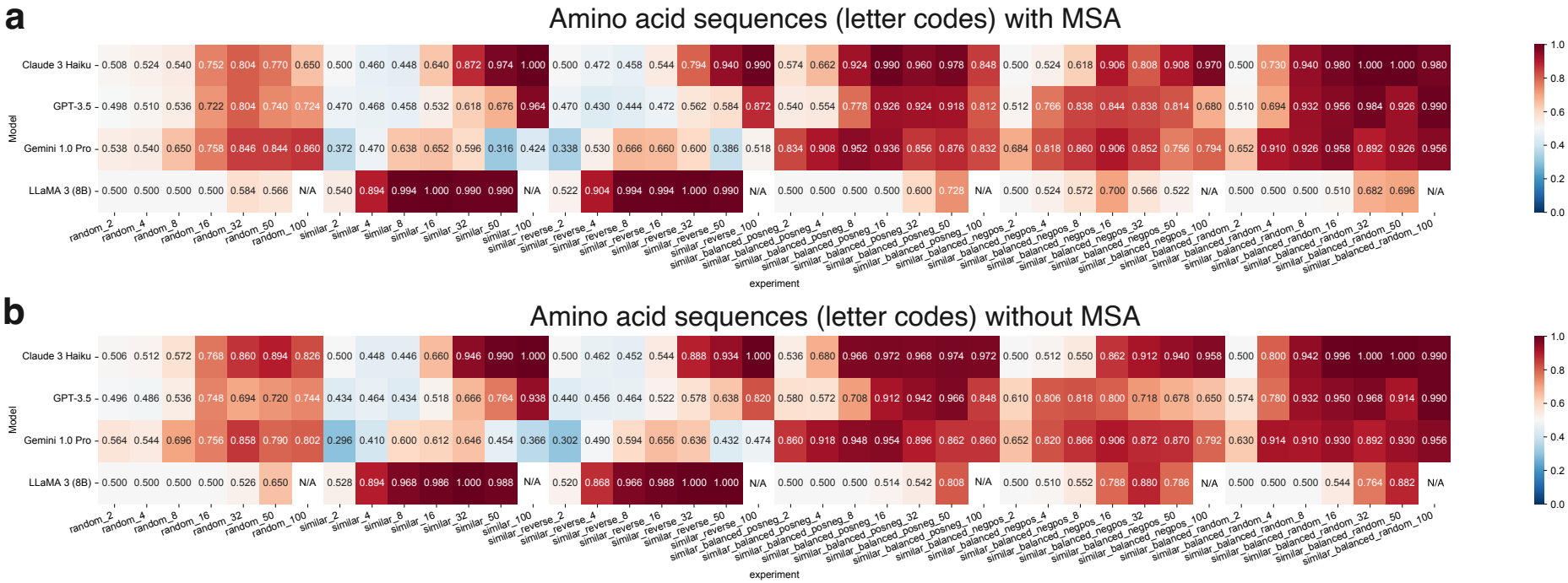
Scenario 3 (Multiclass): Antibody isotype prediction

You are an AI model specializing in immunology. Your task is **to predict the heavy chain isotype** using only the amino acid sequences provided. Recent research has suggested that **the heavy chain isotype of antibody** can be predicted with antibody sequence alone, without any structural information.
Please strictly follow the format, no other information can be provided. Given the amino acid sequences of an antibody, predict **the heavy chain isotype of antibody based on its sequence, by analyzing whether it is IGHA (0), IGHD (1), IGHG (2), IGHM (3).** Consider factors such as physiochemical property, PSSM and amino acid frequency to assess **the heavy chain isotype**.
Please answer with *only* 0, 1, 2 or 3 for the last example. A few examples are provided in the beginning.
Antibody:
EVQLVESGGGLVQPGGSLRLSCAASGFTFSSYWMSWVRQAPGKGLEWVANIKQDGSEKYYVDSVKGRFTISRDNAKNSLYLQMNSLRAEDTAVYYCAGGGY
NWNRAWFDPWGQGTLVTVSS
Heavy chain isotype: 3
…
Antibody:
EVQLVESGGGLVQPGGSLRLSCAASGFTFSSYWMSWVRQAPGKGLEWVANIKQDGSEKYYVDSVKGRFTISRDNAKNSLYLQMNSLRAEDTAVYYCARARS
DYADYVYFDYWGQGTLVTVSS
Heavy chain isotype: 0
Let's solve this problem by splitting it into steps.
Step 1: Please calculate the frequency of each amino acid in the sequence.
Step 2: Please consider the physiochemical property of each amino acid in the sequence.
Now, predict **the heavy chain isotype** for the following antibody sequence based on the criteria above:
Antibody:
EVQLVESGGGLVQPGGSLRLSCAASGFTFSSYWMSWVRQAPGKGLEWVANIKQDGSEKYYVDSVKGRFTISRDNAKNSLYLQMNSLRAEDTAVYYCARPATP
TYNWNDAGWFDPWGQGTLVTVSS
Heavy chain isotype:

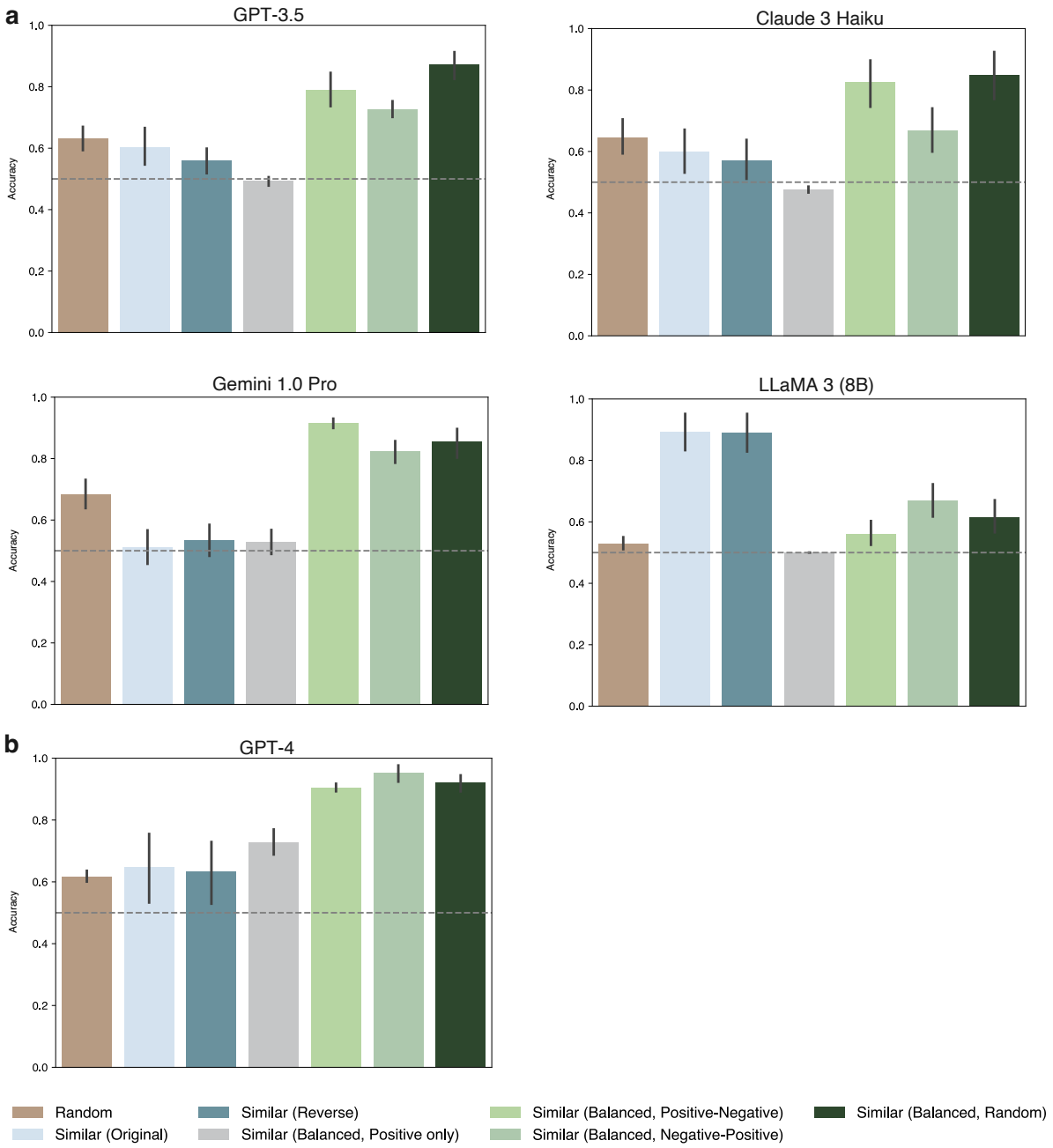
Supplementary Fig. 3 Examples of CoT prompts in three antibody characterization scenarios. The text highlighted in this figure shows the variable inputs specific to each scenario. All other text remains constant across the examples.



Supplementary Fig. 4 Examination of potential label bias in the Similar strategy in all the evaluated scenarios. Cumulative bar plots showing if all the demonstrations share the same label type (e.g., all the demonstrations have label equals to 1 for a certain test instance) in test prompts (n=100) for different k-shot settings in Scenario 1a (**a**), Scenario 1b (**b**), Scenario 2 (**c**), and Scenario 3 (**d**). Note that in Scenario 3, there were no cases where all the demonstrations had a label equal to 1.

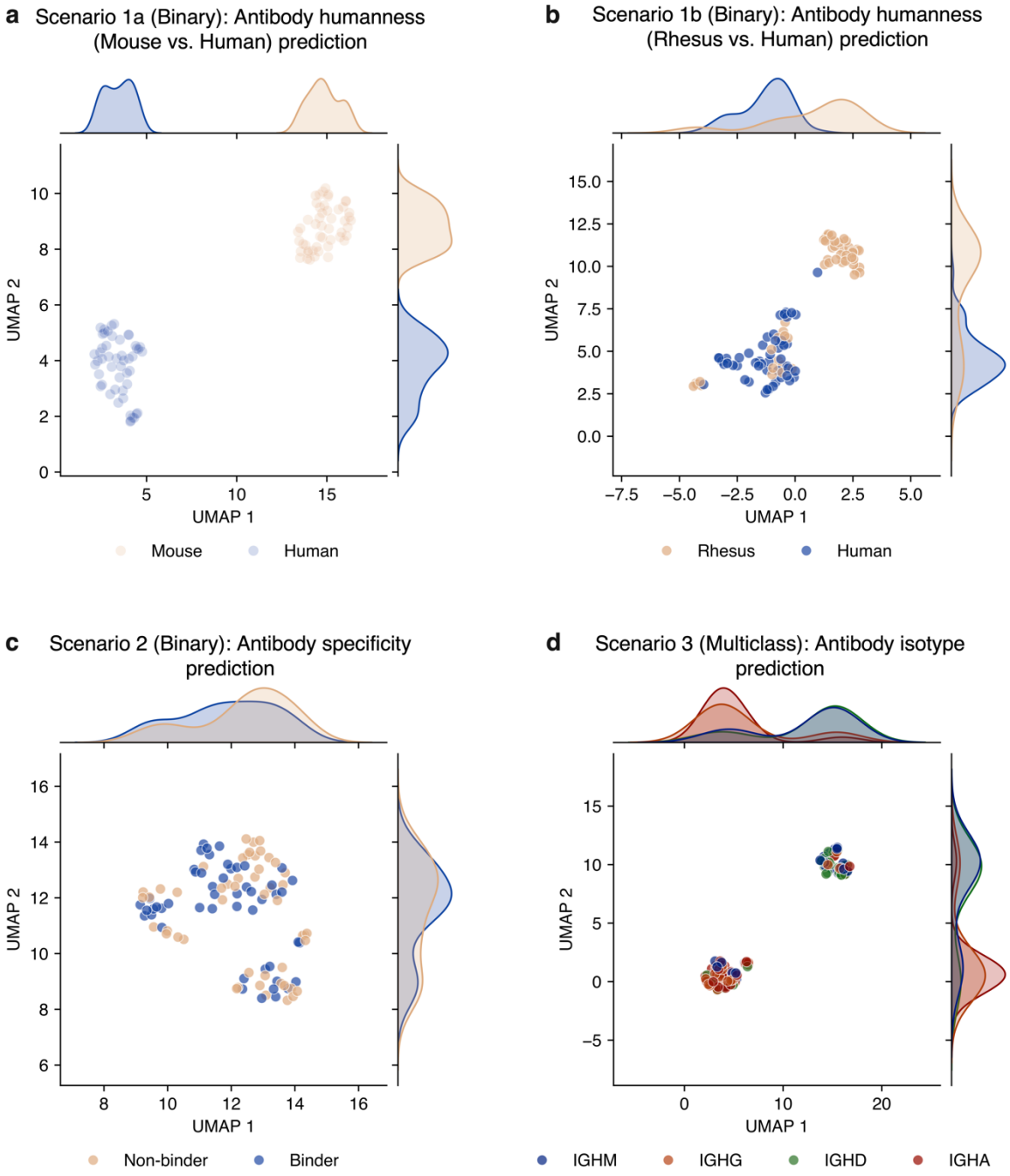


Supplementary Fig. 5 Accuracy of different few-shot demonstration selection strategies in Scenario 1a. The values indicate accuracy for **a** with MSAs and **b** without MSAs.

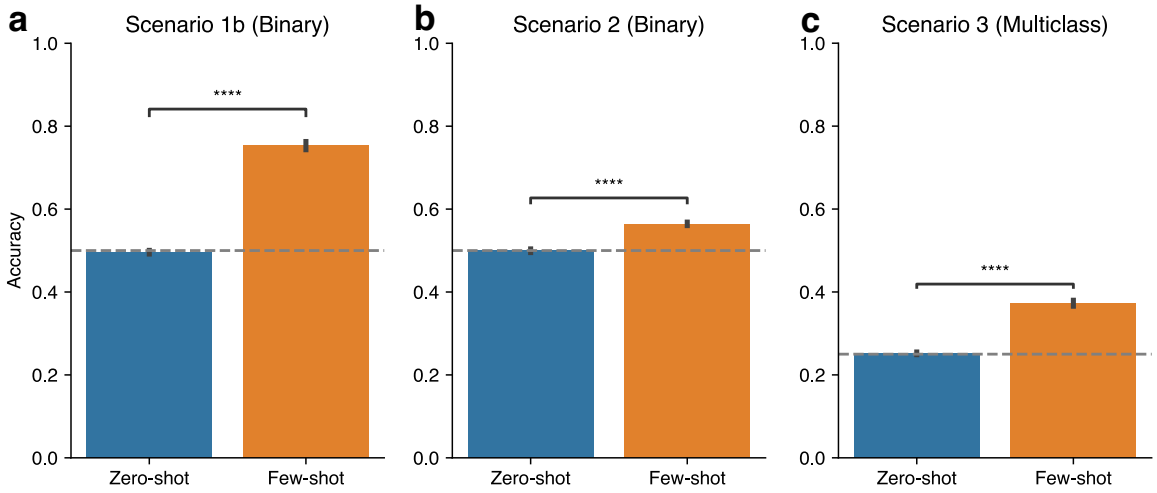


Supplementary Fig. 6 Average few-shot performance for inputting positive labels only in the Similar (Balanced) strategy. Few-shot (including 2-, 4-, 8-, 16-, 32-, 50-, 100-) performance for inputting positive labels only in the Similar (Balanced) strategy in **a** affordable LLMs, including GPT-3.5, Claude 3 Haiku, Gemini 1.0 Pro, and LLaMA 3 (8B), and **b** an advanced model, GPT-4. The dashed line represents the random baseline (accuracy=0.5) of the prediction.

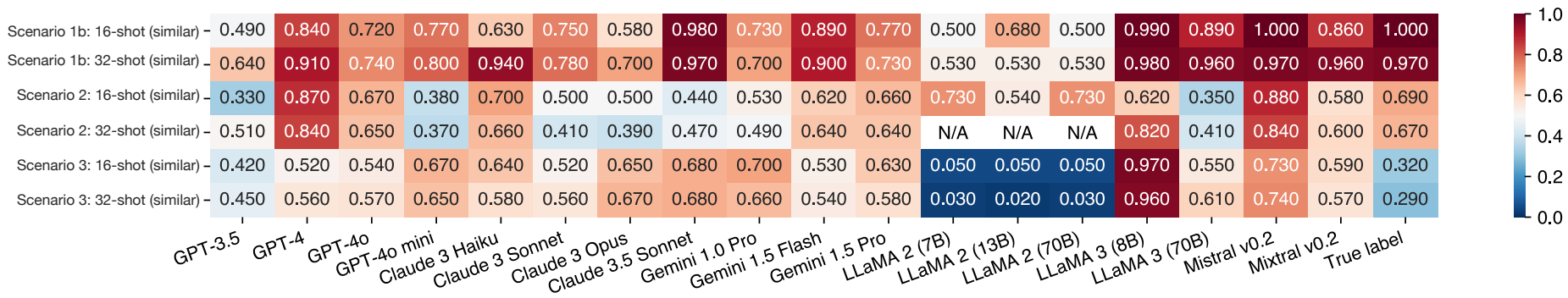
Supplementary Fig. 7 MSA illustrations. The MSAs downloaded from Ramon et al.[1] contained sequences that were aligned, cleaned, and processed. Non-redundant sequences were aligned to a length of 149 using the widely used ANARCI software and the AHo[2] numbering scheme. This numbering scheme provides a pre-defined "static position template" that does not rely on specific input sequences. In other words, for a specific sequence, no matter inputting zero, one, two, three, or any numbers of other sequences together with it, the generated output MSA format for this specific sequence is identical. This figure shows an example when using ANARCI software with the AHo numbering scheme, inputting single or multiple sequences produces the same sequence representation, indicating the MSA is independent of other sequences.



Supplementary Fig. 8 UMAP visualization of the embeddings generated from a pLM, AntiBERTy, in three antibody characterization scenarios. The greater proximity between the target classes indicates a more challenging classification problem in Scenario 1a (**a**), Scenario 1b (**b**), Scenario 2 (**c**), and Scenario 3 (**d**).



Supplementary Fig. 9 Performance comparison between zero-shot ICL and few-shot ICL in three antibody characterization scenarios. The detailed calculation is described in **Methods**. The dashed line represents the random baseline of the prediction. *P* values are based on two-sided Mann-Whitney *U* test: ****$P \leq 0.0001$.



Supplementary Fig. 10 Performance comparison between Similarity strategy and majority voting classifier. Each entry shows the accuracy between an LLM's predictions as the true labels and the majority vote as the predicted labels. A higher value indicates that an LLM's predictions closely match the majority vote predictions, while a lower value suggests greater divergence between the two. For the column "True label", the ground truth were used as true labels and the majority vote predictions were used as predicted labels. Note that Scenario 1a is not included in this analysis as we only tested four affordable models, including GPT-3.5, Claude 3 Haiku, Gemini 1.0 Pro, and LLaMA 3 (8B).

a Few-shot Chain-of-Thought (CoT steps alone) (32-shot)

Scenario 2 (Binary): Antibody specificity

You are an AI model specializing in immunology. Your task is to predict the binding specificity of antibody to the SARS-CoV-2 spike protein using only the amino acid sequences provided. Recent research has suggested that the specificity to antigens of antibody can be predicted with antibody sequence alone, without any structural information.
Please strictly follow the format, no other information can be provided. Given the amino acid sequences of an antibody, predict the receptor binding specificity of antibody based on its sequence, by analyzing whether it can bind(1) or cannot bind(0). Consider factors such as physiochemical property, PSSM and amino acid frequency to assess the binding of antibody.
Please answer with *only* 1 or 0 only for the last example. A few examples are provided in the beginning.
Antibody:
QLQLQESGPGLVKPSETLSLTCTVSGGSISSSSYYWGWIRQPPGKGLEWIGSIYYSGSTYYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCARRVITIFGGAGFDYWGQGTLV
TVSSSYELTQPPSVSVSPGQTARITCSADALPKQYAYWYQQKPGQAPVLVIYKDSERPSGIPERFSGSSSGTTVTLTISGVQAEDEADYYCQSADSSGTYVVFGGGTKLTVL
Binding: 0
…
Antibody:
QVPLQESGPGLVKPSETLSLTCTVSGGSINSYYWNWIRQAPGKGLEWIGNIYYSGSTSYNPSLKSRVTISVDTSKNQFSLKLTSVTAADTAVYYCARGGFYDPYYFDSWGQGTLVTVS
SSFELTQPPSVSVSPGQTARITCSANALPMQYAYWYQQKPGQAPVLVIYKDIERASGIPERFSGSSSGATVTLTISGVQAEDEADYYCQSPDSSGSYVVFGGGTKVTVL
Binding: 1
Let's solve this problem by splitting it into steps.
Step 1: Please calculate the frequency of each amino acid in the sequence.
Step 2: Please consider the physiochemical property of each amino acid in the sequence.
Now, predict the receptor binding specificity for the following antibody sequence based on the criteria above:
Antibody:
QVQLQESGPGLVKPSETLSLTCTVSGYSISSGYYWGWIRQPPGKGLEWIGSIYHCGSTYYNPSLKSRLTISVDTSKNQFSLKLSSVTAADTAVYYCAREPDYSDSGGYYSLYYFAYWG
QGTLVTVSSSHERTQPPSVSVSPGQTARITCSGDALPKQYAYWYQQKPGQAPVVVIYKDTERPSGIPERISGSSSGTTVTLTISGVQAEDEADYYCQSADSSGTYVVFGGGTKLTVL
Binding:

b Few-shot Chain-of-Thought (CoT steps with LLM-generated answers (GPT-3.5)) (32-shot)

You are an AI model specializing in immunology. Your task is to predict the binding specificity of antibody to the SARS-CoV-2 spike protein using only the amino acid sequences provided. Recent research has suggested that the specificity to antigens of antibody can be predicted with antibody sequence alone, without any structural information.
Please strictly follow the format, no other information can be provided. Given the amino acid sequences of an antibody, predict the receptor binding specificity of antibody based on its sequence, by analyzing whether it can bind(1) or cannot bind(0). Consider factors such as physiochemical property, PSSM and amino acid frequency to assess the binding of antibody.
A few examples are provided in the beginning.
Antibody:
QLQLQESGPGLVKPSETLSLTCTVSGGSISSSSYYWGWIRQPPGKGLEWIGSIYYSGSTYYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCARRVITIFGGAGFDYWGQGTLV
TVSSSYELTQPPSVSVSPGQTARITCSADALPKQYAYWYQQKPGQAPVLVIYKDSERPSGIPERFSGSSSGTTVTLTISGVQAEDEADYYCQSADSSGTYVVFGGGTKLTVL
Binding: 0
…
Antibody:
QVPLQESGPGLVKPSETLSLTCTVSGGSINSYYWNWIRQAPGKGLEWIGNIYYSGSTSYNPSLKSRVTISVDTSKNQFSLKLTSVTAADTAVYYCARGGFYDPYYFDSWGQGTLVTVS
SSFELTQPPSVSVSPGQTARITCSANALPMQYAYWYQQKPGQAPVLVIYKDIERASGIPERFSGSSSGATVTLTISGVQAEDEADYYCQSPDSSGSYVVFGGGTKVTVL
Binding: 1
Let's solve this problem by splitting it into steps.
Step 1: Please calculate the frequency of each amino acid in the sequence.
Antibody:
QVQLQESGPGLVKPSETLSLTCTVSGGSISSYYWSWIRQPPGKGLEWIGYIYYSGSTNYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCARDRKTGDSWFDPWGQGTLVTVS
SSYELTQPPSVSVSPGQTARITCSGDALPKQYAYWYQQKPGQAPVLVIYKDSERPSGIPERFSGSSSGTTVTLTISGVQAEDEADYYCQSADSSGTYPVVFGGGTKLTVL

Frequency of each amino acid:
A: 10
C: 6
D: 6
E: 6
F: 6
G: 14
H: 0
I: 7
K: 6
L: 15
M: 0
N: 6
P: 6
Q: 8
R: 5
S: 18
T: 11
V: 12
W: 4
Y: 12
Step 2: Please consider the physiochemical property of each amino acid in the sequence.
Antibody:
QVQLQESGPGLVKPSETLSLTCTVSGGSISSYYWSWIRQPPGKGLEWIGYIYYSGSTNYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCARDRKTGDSWFDPWGQGTLVTVS
SSYELTQPPSVSVSPGQTARITCSGDALPKQYAYWYQQKPGQAPVLVIYKDSERPSGIPERFSGSSSGTTVTLTISGVQAEDEADYYCQSADSSGTYPVVFGGGTKLTVL

Physiochemical properties:
Hydrophobic (A, C, F, G, I, L, M, P, V, W): 108
Polar (D, E, H, K, N, Q, R, S, T, Y): 72
Please answer with *only* 1 or 0 only for the following example.
Now, predict the receptor binding specificity for the following antibody sequence based on the criteria above:
Antibody:
QVQLQESGPGLVKPSETLSLTCTVSGYSISSGYYWGWIRQPPGKGLEWIGSIYHCGSTYYNPSLKSRLTISVDTSKNQFSLKLSSVTAADTAVYYCAREPDYSDSGGYYSLYYFAYWG
QGTLVTVSSSHERTQPPSVSVSPGQTARITCSGDALPKQYAYWYQQKPGQAPVVVIYKDTERPSGIPERISGSSSGTTVTLTISGVQAEDEADYYCQSADSSGTYVVFGGGTKLTVL
Binding:

c Few-shot Chain-of-Thought (CoT steps with LLM-generated answers (GPT-4)) (32-shot)

You are an AI model specializing in immunology. Your task is to predict the binding specificity of antibody to the SARS-CoV-2 spike protein using only the amino acid sequences provided. Recent research has suggested that the specificity to antigens of antibody can be predicted with antibody sequence alone, without any structural information.
Please strictly follow the format, no other information can be provided. Given the amino acid sequences of an antibody, predict the receptor binding specificity of antibody based on its sequence, by analyzing whether it can bind(1) or cannot bind(0). Consider factors such as physiochemical property, PSSM and amino acid frequency to assess the binding of antibody.
A few examples are provided in the beginning.
Antibody:
QLQLQESGPGLVKPSETLSLTCTVSGGSISSSSYYWGWIRQPPGKGLEWIGSIYYSGSTYYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCARRVITIFGGAGFDYWGQGTLV
TVSSSYELTQPPSVSVSPGQTARITCSADALPKQYAYWYQQKPGQAPVLVIYKDSERPSGIPERFSGSSSGTTVTLTISGVQAEDEADYYCQSADSSGTYVVFGGGTKLTVL
Binding: 0
Antibody:
QVPLQESGPGLVKPSETLSLTCTVSGGSINSYYWNWIRQAPGKGLEWIGNIYYSGSTSYNPSLKSRVTISVDTSKNQFSLKLTSVTAADTAVYYCARGGFYDPYYFDSWGQGTLVTVS
SSFELTQPPSVSVSPGQTARITCSANALPMQYAYWYQQKPGQAPVLVIYKDIERASGIPERFSGSSSGATVTLTISGVQAEDEADYYCQSPDSSGSYVVFGGGTKVTVL
Binding: 1
Let's solve this problem by splitting it into steps.
Step 1: Please calculate the frequency of each amino acid in the sequence.
To calculate the frequency of each amino acid in the sequence, we first need to choose a specific antibody sequence from the examples provided. Let's take the first antibody sequence as an example:

Antibody Sequence:
QLQLQESGPGLVKPSETLSLTCTVSGGSISSSSYYWGWIRQPPGKGLEWIGSIYYSGSTYYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCARRVITIFGGAGFDYWGQGTLV
TVSSSYELTQPPSVSVSPGQTARITCSADALPKQYAYWYQQKPGQAPVLVIYKDSERPSGIPERFSGSSSGTTVTLTISGVQAEDEADYYCQSADSSGTYVVFGGGTKLTVL

Step 1: Calculate the frequency of each amino acid

To calculate the frequency, we count how many times each amino acid appears in the sequence and then divide by the total number of amino acids in the sequence.

Here's the breakdown:

1. Count each amino acid:
   - A: 12
   - C: 6
   - D: 9
   - E: 7
   - F: 6
   - G: 18
   - H: 0
   - I: 7
   - K: 11
   - L: 24
   - M: 0
   - N: 2
   - P: 10
   - Q: 12
   - R: 6
   - S: 27
   - T: 14
   - V: 18
   - W: 6
   - Y: 18

2. Total number of amino acids in the sequence:
   - Total = 220

3. Calculate the frequency of each amino acid:
   - A: 12 / 220 = 0.0545
   - C: 6 / 220 = 0.0273
   - D: 9 / 220 = 0.0409
   - E: 7 / 220 = 0.0318
   - F: 6 / 220 = 0.0273
   - G: 18 / 220 = 0.0818
   - H: 0 / 220 = 0.0000
   - I: 7 / 220 = 0.0318
   - K: 11 / 220 = 0.0500
   - L: 24 / 220 = 0.1091
   - M: 0 / 220 = 0.0000
   - N: 2 / 220 = 0.0091
   - P: 10 / 220 = 0.0455
   - Q: 12 / 220 = 0.0545
   - R: 6 / 220 = 0.0273
   - S: 27 / 220 = 0.1227
   - T: 14 / 220 = 0.0636
   - V: 18 / 220 = 0.0818
   - W: 6 / 220 = 0.0273
   - Y: 18 / 220 = 0.0818

These frequencies provide a basic profile of the amino acid composition of the antibody sequence.
Step 2: Please consider the physiochemical property of each amino acid in the sequence.
In step 2, we'll consider the physiochemical properties of each amino acid in the antibody sequence. These properties include hydrophobicity, charge, size, and polarity, which are crucial for understanding how the antibody might interact with its target, such as the SARS-CoV-2 spike protein.

Here are the physiochemical properties of the amino acids present in the sequence:

1. **Alanine (A)** - Small, hydrophobic
2. **Cysteine (C)** - Small, slightly hydrophobic, contains a thiol group which can form disulfide bonds
3. **Aspartic acid (D)** - Small, acidic, negatively charged at physiological pH
4. **Glutamic acid (E)** - Larger, acidic, negatively charged at physiological pH
5. **Phenylalanine (F)** - Aromatic, hydrophobic
6. **Glycine (G)** - Smallest, no side chain, increases flexibility
7. **Histidine (H)** - Aromatic, basic, positively charged at physiological pH (can be neutral depending on environment)
8. **Isoleucine (I)** - Large, hydrophobic, branched
9. **Lysine (K)** - Large, basic, positively charged at physiological pH
10. **Leucine (L)** - Large, hydrophobic, branched
11. **Methionine (M)** - Large, hydrophobic, contains a sulfur atom
12. **Asparagine (N)** - Small, polar, uncharged
13. **Proline (P)** - Cyclic, rigid structure, affects protein folding
14. **Glutamine (Q)** - Large, polar, uncharged
15. **Arginine (R)** - Large, basic, positively charged at physiological pH
16. **Serine (S)** - Small, polar, uncharged
17. **Threonine (T)** - Medium, polar, uncharged
18. **Valine (V)** - Medium, hydrophobic, branched
19. **Tryptophan (W)** - Aromatic, largest, hydrophobic
20. **Tyrosine (Y)** - Aromatic, polar, can be phosphorylated

Analysis of the Sequence:
Given the high frequency of hydrophobic amino acids like Leucine (L), Valine (V), and Glycine (G), the antibody sequence suggests a significant hydrophobic character, which might be important for binding interactions with the hydrophobic regions of the SARS-CoV-2 spike protein. The presence of charged residues like Lysine (K), Arginine (R), Aspartic acid (D), and Glutamic acid (E) indicates potential sites for ionic interactions with oppositely charged residues on the spike protein.

The sequence also contains several Serine (S) and Threonine (T) residues, which are polar and could be involved in hydrogen bonding with the antigen. Proline (P) introduces rigidity into the structure, which could affect the overall conformation of the antibody and its interaction with the antigen.

This analysis helps in understanding how the antibody might interact with its target based on the physiochemical properties of the amino acids in its sequence.
Please answer with "only" 1 or 0 only for the following example.
Now, predict the receptor binding specificity for the following antibody sequence based on the criteria above:
Antibody:
QVQLQESGPGLVKPSETLSLTCTVSGYSISSGYYWGWIRQPPGKGLEWIGSIYHCGSTYYNPSLKSRLTISVDTSKNQFSLKLSSVTAADTAVYYCAREPDYSDSGGYYSLYYFAYWG
QGTLVTVSSSHERTQPPSVSVSPGQTARITCSGDALPKQYAYWYQQKPGQAPVVVIYKDTERPSGIPERISGSSSGTTVTLTISGVQAEDEADYYCQSADSSGTYVVFGGGTKLTVL
Binding:

d Few-shot Chain-of-Thought (CoT steps with manual answers) (32-shot)

You are an AI model specializing in immunology. Your task is to predict the binding specificity of antibody to the SARS-CoV-2 spike protein using only the amino acid sequences provided. Recent research has suggested that the specificity to antigens of antibody can be predicted with antibody sequence alone, without any structural information.
Please strictly follow the format, no other information can be provided. Given the amino acid sequences of an antibody, predict the receptor binding specificity of antibody based on its sequence, by analyzing whether it can bind(1) or cannot bind(0). Consider factors such as physiochemical property, PSSM and amino acid frequency to assess the binding of antibody.
Please answer with *only* 1 or 0 only for the last example. A few examples are provided in the beginning.
Antibody:
QLQLQESGPGLVKPSETLSLTCTVSGGSISSSSYYWGWIRQPPGKGLEWIGSIYYSGSTYYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCARRVITIFGGAGFDYWGQGTLV
TVSSSYELTQPPSVSVSPGQTARITCSADALPKQYAYWYQQKPGQAPVLVIYKDSERPSGIPERFSGSSSGTTVTLTISGVQAEDEADYYCQSADSSGTYVVFGGGTKLTVL
Binding: 0
…
Antibody:
QVPLQESGPGLVKPSETLSLTCTVSGGSINSYYWNWIRQAPGKGLEWIGNIYYSGSTSYNPSLKSRVTISVDTSKNQFSLKLTSVTAADTAVYYCARGGFYDPYYFDSWGQGTLVTVS
SSFELTQPPSVSVSPGQTARITCSANALPMQYAYWYQQKPGQAPVLVIYKDIERASGIPERFSGSSSGATVTLTISGVQAEDEADYYCQSPDSSGSYVVFGGGTKVTVL
Binding: 1
Let's solve this problem by splitting it into steps.
Step 1: Step 1: Please calculate the frequency of each amino acid in the sequence.

To solve the problem, we will first calculate the frequency of each amino acid in the provided antibody sequence. The sequence given is:

Antibody:
QVQLQESGPGLAKPSETLSLTCTVSGYSISSGHYWGWIRQPPGKGLEWIGSIYHSGSTYYNPSLKSRVTISVDTSKNQFSLKLTSVTAADTAVYYCARAEYSNYYYGMDVWGQGTTVT
VSSSYELTQPPSVSVSPGQTARITCSGDALPKQYAYWYQQKPGQAPVLVIYKDSERPSGIPERFSGSSSGTTVTLTISGVQAEDEADYYCQSADSSGTYVVFGGGTKLT

Step 1: Calculate the frequency of each amino acid in the sequence.

1. Count the occurrences of each amino acid.
2. Divide the count by the total length of the sequence to get the frequency.

Let's calculate:
Total length of the sequence = 227 amino acids.

- A: 13
- C: 4
- D: 8
- E: 9
- F: 3
- G: 23
- H: 2
- I: 9
- K: 9
- L: 13
- M: 1
- N: 3
- P: 13
- Q: 14
- R: 6
- S: 34
- T: 22
- V: 17
- W: 5
- Y: 19

Now, calculate the frequency of each amino acid:

- A: 13 / 227 = 0.0573
- C: 4 / 227 = 0.0176
- D: 8 / 227 = 0.0352
- E: 9 / 227 = 0.0396
- F: 3 / 227 = 0.0132
- G: 23 / 227 = 0.1013
- H: 2 / 227 = 0.0088
- I: 9 / 227 = 0.0396
- K: 9 / 227 = 0.0396
- L: 13 / 227 = 0.0573
- M: 1 / 227 = 0.044
- N: 3 / 227 = 0.0132
- P: 13 / 227 = 0.0573
- Q: 14 / 227 = 0.0617
- R: 6 / 227 = 0.0264
- S: 34 / 227 = 0.1498
- T: 22 / 227 = 0.0969
- V: 17 / 227 = 0.0749
- W: 5 / 227 = 0.0220
- Y: 19 / 227 = 0.0837

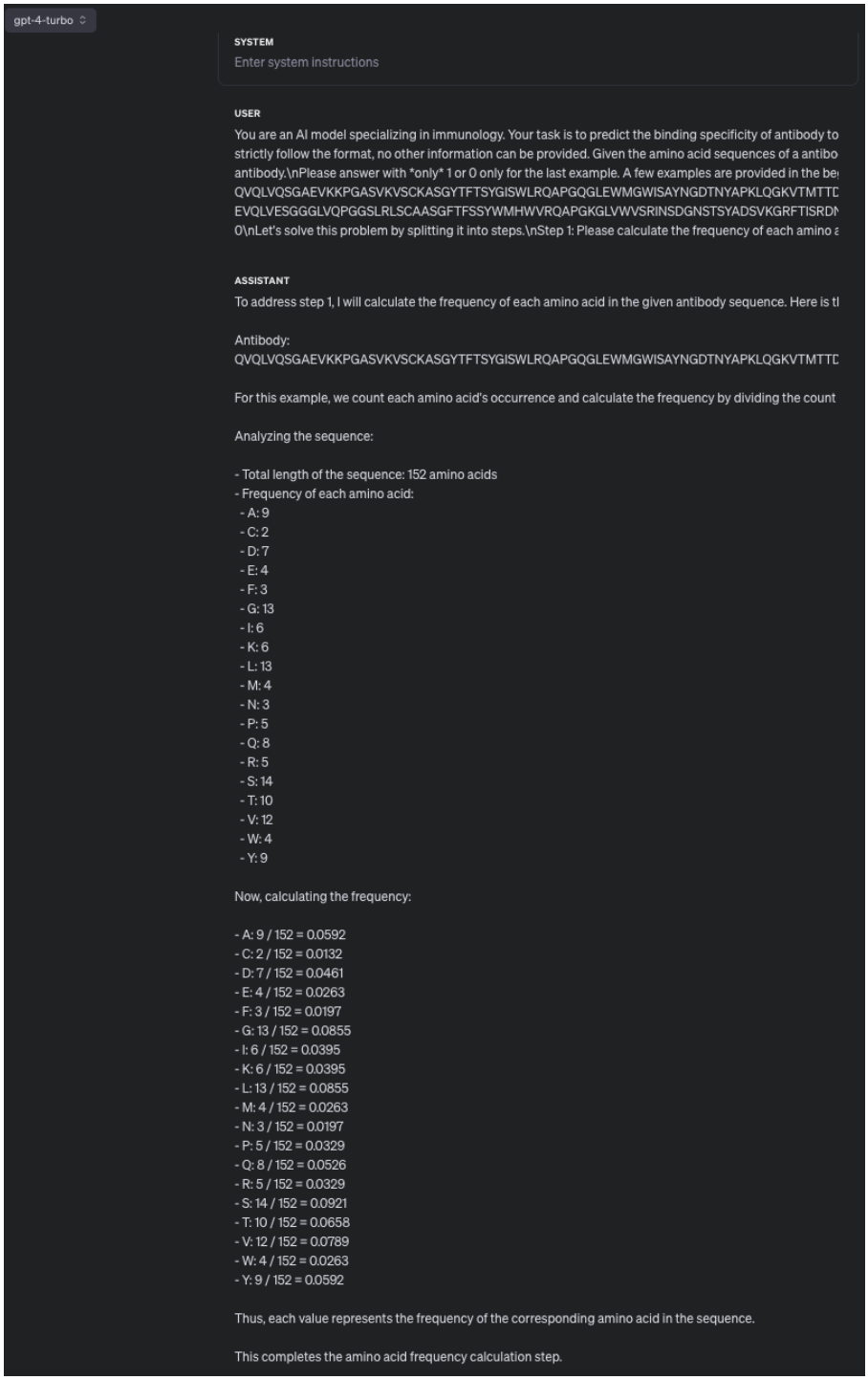
These frequencies provide a profile of the amino acid composition of the antibody sequence.

Step 2: Please consider the physiochemical property of each amino acid in the sequence.
In Step 2, we consider the physiochemical properties of each amino acid in the antibody sequence. These properties include hydrophobicity, charge, size, and polarity, which can influence the interaction between the antibody and its target antigen. Here's a brief overview of the relevant properties for each amino acid type:
1. **Alanine (A)** - Nonpolar, hydrophobic
2. **Cysteine (C)** - Polar, uncharged, forms disulfide bonds
3. **Aspartic acid (D)** - Polar, negatively charged
4. **Glutamic acid (E)** - Polar, negatively charged
5. **Phenylalanine (F)** - Nonpolar, hydrophobic, aromatic
6. **Glycine (G)** - Nonpolar, smallest amino acid, provides flexibility
7. **Histidine (H)** - Polar, positively charged at physiological pH
8. **Isoleucine (I)** - Nonpolar, hydrophobic
9. **Lysine (K)** - Polar, positively charged
10. **Leucine (L)** - Nonpolar, hydrophobic
11. **Methionine (M)** - Nonpolar, hydrophobic
12. **Asparagine (N)** - Polar, uncharged
13. **Proline (P)** - Nonpolar, introduces kinks in peptide chains
14. **Glutamine (Q)** - Polar, uncharged
15. **Arginine (R)** - Polar, positively charged
16. **Serine (S)** - Polar, uncharged
17. **Threonine (T)** - Polar, uncharged
18. **Valine (V)** - Nonpolar, hydrophobic
19. **Tryptophan (W)** - Nonpolar, hydrophobic, aromatic
20. **Tyrosine (Y)** - Polar, uncharged, aromatic

To analyze the physiochemical properties of the sequence, we can look at the distribution and frequency of these amino acids and infer how they might contribute to the antibody's binding specificity:
- **Hydrophobicity**: A high frequency of nonpolar, hydrophobic residues (A, F, I, L, M, V, W) can suggest strong interactions with hydrophobic regions of the antigen.
- **Charge**: The presence of charged residues (D, E, H, K, R) can facilitate interactions with oppositely charged areas on the antigen, influencing specificity and binding strength.
- **Polarity and Flexibility**: Polar (C, N, Q, S, T, Y) and small or flexible residues (G, P) can affect how the antibody folds and how it interacts with the antigen's surface.

Given the sequence provided, we can observe a balanced mix of hydrophobic and polar residues, along with a significant presence of charged amino acids. This combination might suggest a versatile binding capability, potentially enhancing the antibody's ability to specifically recognize and bind to diverse regions on the SARS-CoV-2 spike protein.
Now, predict the receptor binding specificity for the following antibody sequence based on the criteria above:
Antibody:
QVQLQESGPGLVKPSETLSLTCTVSGYSISSGYYWGWIRQPPGKGLEWIGSIYHCGSTYYNPSLKSRLTISVDTSKNQFSLKLSSVTAADTAVYYCAREPDYSDSGGYYSLYYFAYWG
QGTLVTVSSSHERTQPPSVSVSPGQTARITCSGDALPKQYAYWYQQKPGQAPVVVIYKDTERPSGIPERISGSSSGTTVTLTISGVQAEDEADYYCQSADSSGTYVVFGGGTKLTVL
Binding:

Supplementary Fig. 11 CoT prompt variation illustrations. a CoT steps alone. **b** CoT steps with LLM-generated answers (GPT-3.5). **c** CoT steps with LLM-generated answers (GPT-4). **d** CoT steps with manual answers. Text highlighted in blue color indicate the fixed input. Text highlighted in grey color indicate the variable input by LLM-generated text.



gpt-4-turbo

SYSTEM

Enter system instructions

USER

You are an AI model specializing in immunology. Your task is to predict the binding specificity of antibody to
strictly follow the format, no other information can be provided. Given the amino acid sequences of a antibo
antibody.\nPlease answer with *only* 1 or 0 only for the last example. A few examples are provided in the be
QVQLVQSGAEVKKPGASVKVSCKASGYTFTSYGISWLRQAPGQGLEWMGWISAYNGDTNYAPKLQGKVTMTTD
EVQLVESGGGLVQPGGSLRLSCAASGFTFSSYWMHWVRQAPGKGLVWVSRINSDGNSTSYADSVKGRFTISRDN
0\nLet's solve this problem by splitting it into steps.\nStep 1: Please calculate the frequency of each amino a

ASSISTANT

To address step 1, I will calculate the frequency of each amino acid in the given antibody sequence. Here is t

Antibody:
QVQLVQSGAEVKKPGASVKVSCKASGYTFTSYGISWLRQAPGQGLEWMGWISAYNGDTNYAPKLQGKVTMTTD

For this example, we count each amino acid's occurrence and calculate the frequency by dividing the count

Analyzing the sequence:

- Total length of the sequence: 152 amino acids
- Frequency of each amino acid:
  - A: 9
  - C: 2
  - D: 7
  - E: 4
  - F: 3
  - G: 13
  - I: 6
  - K: 6
  - L: 13
  - M: 4
  - N: 3
  - P: 5
  - Q: 8
  - R: 5
  - S: 14
  - T: 10
  - V: 12
  - W: 4
  - Y: 9

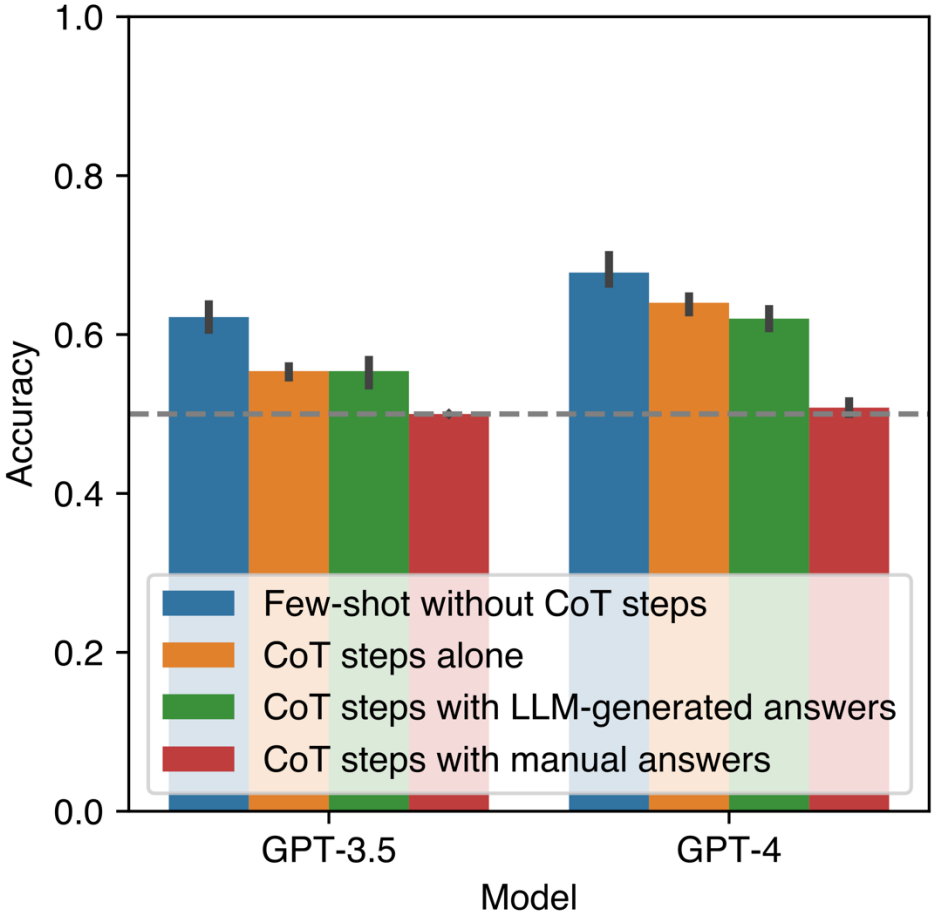
Now, calculating the frequency:

- A: 9 / 152 = 0.0592
- C: 2 / 152 = 0.0132
- D: 7 / 152 = 0.0461
- E: 4 / 152 = 0.0263
- F: 3 / 152 = 0.0197
- G: 13 / 152 = 0.0855
- I: 6 / 152 = 0.0395
- K: 6 / 152 = 0.0395
- L: 13 / 152 = 0.0855
- M: 4 / 152 = 0.0263
- N: 3 / 152 = 0.0197
- P: 5 / 152 = 0.0329
- Q: 8 / 152 = 0.0526
- R: 5 / 152 = 0.0329
- S: 14 / 152 = 0.0921
- T: 10 / 152 = 0.0658
- V: 12 / 152 = 0.0789
- W: 4 / 152 = 0.0263
- Y: 9 / 152 = 0.0592

Thus, each value represents the frequency of the corresponding amino acid in the sequence.

This completes the amino acid frequency calculation step.

Supplementary Fig. 12 CoT reasoning steps with incorrect LLM-generated response from GPT-4. As shown in the figure, GPT-4 incorrectly state the length of truncated input sequence ("QVQLVQSGAEVKKPGASVKVSCKASGYTFTSYGISWLRQAPGQGLEWMGWISAYNGDTNYAPKLQGKVTMTTDTSTSTAYMELRSLRSDDTAVYYCARDLPKISIFGVADYWGQGTLVTVSS") is 152, but the correct answer is 122 amino acids. Additionally, the amino acid counts for alanine (A) and aspartic acid (D) were incorrect in GPT-4's response. The correct counts are 9 for alanine (A) and 7 for aspartic acid (D).



Supplementary Fig. 13 CoT prompt variation performance. The dashed line represents the random baseline (accuracy=0.5) of the prediction.

a Fine-tuning prompts

Query

You are an AI model specializing in immunology. Your task is to predict the binding specificity of antibody to the SARS-CoV-2 spike protein using only the amino acid sequences provided. Recent research has suggested that the specificity to antigens of antibody can be predicted with antibody sequence alone, without any structural information.
Please strictly follow the format, no other information can be provided. Given the amino acid sequences of a antibody, predict the receptor binding specificity of antibody based on its sequence, by analyzing whether it can bind(1) or cannot bind(0). Consider factors such as physiochemical property, PSSM and amino acid frequency to assess the binding of antibody.
Please answer with *only* 1 or 0 only for the last example. A few examples are provided in the beginning.
Antibody:
EVQLVESGGGLIQPGGSLRLSCVASGFSVSNNYMTWVRQAPGKGLEWVSVIYSDGTTYYTDSVKGRYIISRD
NSKNTLYLQMNRLRAEDTAVYYCVCSWYYDFLTGYYGFDYWGQGTLVTVSSQAVVTQEPSLTVSPGGTVTLT
CGSSTGPVTSGHYPYWFQQKPGQAPRTLVYDTSNKQSWTPARFSGSLLGGKAALTLSGAQPEDEAEYYCLL
SYSGARPGIFGGGTKLTVL

Expected response

Binding: 1

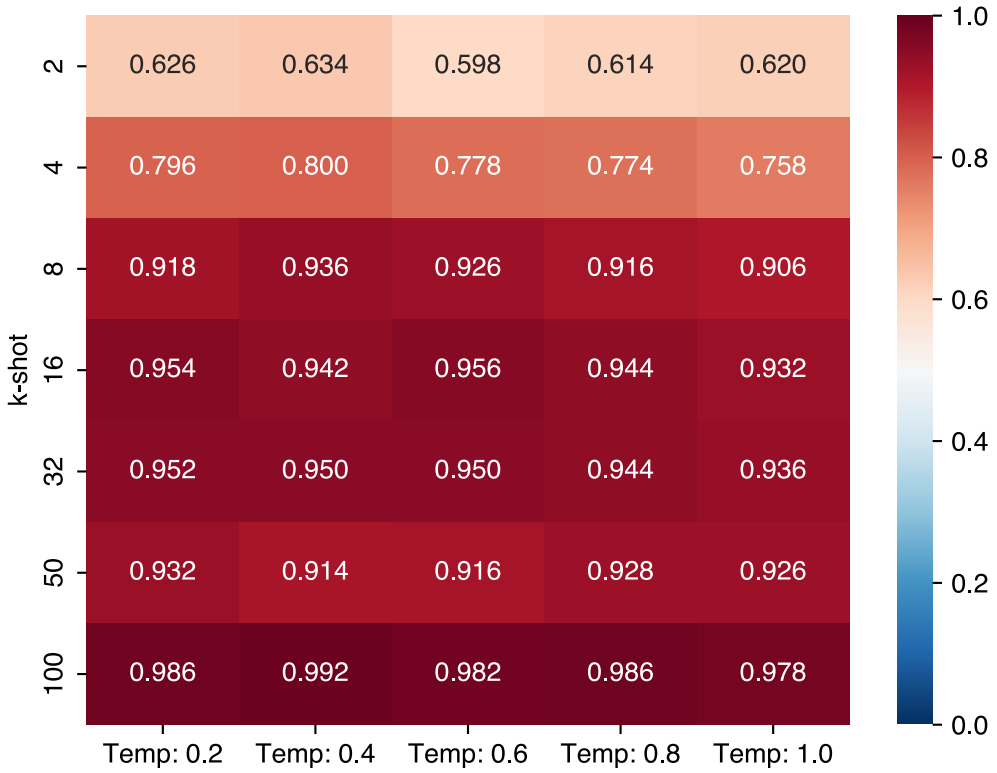
b Fine-tuning input for GPT family (GPT-3.5 and GPT-4o mini)


```
{"messages": [{"role": "user", "content": "You are an AI model specializing in immunology. Your
task is to predict the binding specificity of antibody to the SARS-CoV-2 spike protein using
only the amino acid sequences provided. Recent research has suggested that the specificity to
antigens of antibody can be predicted with antibody sequence alone, without any structural
information.\nPlease strictly follow the format, no other information can be provided. Given the
amino acid sequences of a antibody, predict the receptor binding specificity of antibody based
on its sequence, by analyzing whether it can bind(1) or cannot bind(0). Consider factors such as
physiochemical property, PSSM and amino acid frequency to assess the binding of antibody.
\nPlease answer with *only* 1 or 0 only for the last example. A few examples are provided in the
beginning.\nAntibody:
EVQLVESGGGLIQPGGSLRLSCVASGFSVSNNYMTWVRQAPGKGLEWVSVIYSDGTTYYTDSVKGRYIISRDNSKNTLYLQMNRLRAEDTAVYYCV
CSWYYDFLTGYYGFDYWGQGTLVTVSSQAVVTQEPSLTVSPGGTVTLTCGSSTGPVTSGHYPYWFQQKPGQAPRTLVYDTSNKQSWTPARFSGSLL
GGKAALTLSGAQPEDEAEYYCLLSYSGARPGIFGGGTKLTVL"},{"role": "assistant", "content": "Binding: 1"}]}
```


c Fine-tuning input for Gemini 1.0 Pro

text_input	output
You are an AI model specializing in immunology. Your task is to predict the binding specificity	Binding: 1
You are an AI model specializing in immunology. Your task is to predict the binding specificity	Binding: 0
You are an AI model specializing in immunology. Your task is to predict the binding specificity	Binding: 0
You are an AI model specializing in immunology. Your task is to predict the binding specificity	Binding: 1
You are an AI model specializing in immunology. Your task is to predict the binding specificity	Binding: 1
You are an AI model specializing in immunology. Your task is to predict the binding specificity	Binding: 0
You are an AI model specializing in immunology. Your task is to predict the binding specificity	Binding: 1
You are an AI model specializing in immunology. Your task is to predict the binding specificity	Binding: 0
You are an AI model specializing in immunology. Your task is to predict the binding specificity	Binding: 1

Supplementary Fig. 14 Illustration of prompts used for fine-tuning in GPT-3.5, GPT-4o mini and Gemini 1.0 Pro. a The prompt used for query and its expected responses for fine-tuning. **b** Sample JSON file with fine-tuning prompts in “user” and “assistant” roles in for fine-tuning GPT family models. **c** Sample comma-separated CSV file for Gemini 1.0 Pro fine-tuning.



Supplementary Fig. 15 Accuracy of temperature variations, including temperature of 0.2, 0.4, 0.6, 0.8, and 1.0, in different *k*-shot settings.

ChatGPT 4o

You are an AI model specializing in immunology. Your task is to predict the binding specificity of antibody to the SARS-CoV-2 spike protein using only the amino acid sequences provided. Recent research has suggested that the specificity to antigens of antibody can be predicted with antibody sequence alone, without any structural information.\nPlease strictly follow the format, no other information can be provided. Given the amino acid sequences of a antibody, predict the receptor binding specificity of antibody based on its sequence, by analyzing whether it can bind(1) or cannot bind(0). Consider factors such as physiochemical property, PSSM and amino acid frequency to assess the binding of antibody.\nPlease answer with only 1 or 0 only for the last example. A few examples are provided in the beginning.\nAntibody:
QLQLQESGPGLVKPSETLSLTCTVSGGSISSSSYYWGWIRQPPGKGLE
WIGSIYYSGSTYYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCA
RQGGSYYEYYFDYWGQGTLVTVSSSYELTQPPSVSVSPGQTARITCSG
DALPKQYAYWYQQKPGQAPVLVIYKDSERPSGIPERFSGSSSGTTVTLTI
SGVQAEDEADYYCQSADSSGTLFGGGTKLTVL\nBinding:
0\nAntibody:
QLQLQESGPGLVKPSETLSLTCTVSGGSISSSSYYWGWIRQPPGKGLE
WIGSIYYSGSTYYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCA
RSAPLYDSSGYHSNWFDPWGQGTLVTVSSSYELTQPPSVSVSPGQTAR
ITCSGDALPKQYAYWYQQKPGQAPVLVIYKDSERPSGIPERFSGSSSGTT
VTLTISGVQAEDEADYYCQSADSSGTVVFGGGTKLTVL\nBinding:
0\nAntibody:
QVQLQESGPGLVKPSETLSLTCTVSGGSISSSSYYWGWIRQPPGKGLE
WIGSIYYSGSTYYNPSLKSRVTISVDTSKNQFSLKLSSVTAADTAVYYCA
RQGDCSTTSCAYDYWGQGTLVTVSSSYELTQPPSVSVSPGQTARITCS
GDALPKQYAYWYQQKPGQAPVLVIYKDSERPSGIPERFSGSSSGTTATLT
ISGVQAEDEADYYCQSADSSGTYLVVFGGGTKLAV\nBinding:
1\nAntibody:
QVQLQESGPGLAKPSETLSLTCTVSGYSISSGHYWGWIRQPPGKGLEWI
GSIYHSGSTYYNPSLKSRVTISVDTSKNQFSLKLTSVTAADTAVYYCARA
EYSNYYYGMDVWGQGTTVTVSSSYELTQPPSVSVSPGQTARITCSGDA
LPKQYAYWYQQKPGQAPVLVIYKDSERPSGIPERFSGSSSGTTVTLTISG
VQAEDEADYYCQSADSSGTYVVFGGGTKLTVL\nBinding: 1\n\nNow, predict the receptor binding specificity for the following antibody sequence based on the criteria above:\nAntibody:
QVQLQESGPGLVKPSETLSLTCTVSGYSISSGYYWGWIRQPPGKGLEWI
GSIYHCGSTYYNPSLKSRLTISVDTSKNQFSLKLSSVTAADTAVYYCARE
PDYSDSGGYYSLYYFAYWGQGTLVTVSSSHERTQPPSVSVSPGQTARIT
CSGDALPKQYAYWYQQKPGQAPVVVIYKDTERPSGIPERISGSSSGTTV
TLTISGVQAEDEADYYCQSADSSGTYVVFGGGTKLTVL\nBinding:\n

1

Message ChatGPT

Supplementary Fig. 16 Screen capture demonstrating the use of ChatGPT as an AI-based chatbot for antibody characterization prediction.

Prompt

```
You are an AI model specializing in immunology. Your task is to predict the humanness of antibody using only the amino acid sequences provided. Recent research
has suggested that the humanness of antibody can be predicted with antibody sequence alone, without any structural information.
Please strictly follow the format, no other information can be provided. Given the amino acid sequences of a antibody, predict the humanness of antibody based on
its sequence, by analyzing whether it can Human(1) or Rhesus(0). Consider factors such as physiochemical property, PSSM and amino acid frequency to assess the
humanness of antibody.
Please answer with *only* 1 or 0 only for the last example. A few examples are provided in the beginning.
Antibody: QVQLQESGPGLVTPSETLSLTCTVSGDSISDYYWNWIRQPAGKGLEWFGRFYSSGSANYNPSLTNRVTMSIDTSINQISLKLSSVTAADTAVYYCARSGSHFSFDSWGRGSLVTVSS
Humanness: 1
Antibody: QVQLQQWGAGLLKTSETLSLTCAAYGGSFSGYYWSWIRQPPGKGLEWIGEISHSGATNYNPSLMRRVTISVDTSKNQFSLNLNSVTAADTAVYYCARGHYIWGSYRQTASYGLDVWGQGTTVTVSS
Humanness: 1
Antibody: EVQLLESGGDFVQPGGSLTLSCAASGFTFRRYAMSWVRQAPGKGLEWVSMIPGLGRATYYADSVKGRFTISRDNVKNTLYLEMRDLRVEDTGLYYCAKLIYADSGSPDLWGRGTQVTVSS
Humanness: 1
Antibody: EMQLVESGGGLVQPGGSLRLSCVASGFTFSRYWMVLVSQAPGKGLEWVGNIKQDGSQKHYLDSVEGRFSISRDNAKSSMYLHMNSLRAEDTAVYYCVRDLDYFNGNNFDAFDIWGQGTMVTVSS
Humanness: 1
Antibody: EVQLLESGGGLVQPGGSLRLSCAASGFTFSNYAMTWVRQAPGRGLEWVSGITSSGGSTYYAASVKGRFTISRDSSRNTLHLQMNSLRAEDTAIYYCTKWLTAAGTGWFDCWGQGTLVTVSS
Humanness: 1
Antibody: QVQLQQWGAGLLKPSETLSLTCAVYGGSFSGYYLSWLRQPPGKGLEWIGEINHSGSTNYNPSLKSRVTTSVDTSKKQFSLKVRSVTAADTAVYYCARGPHRNRLQLNNNYSGMDVWGQGTTVTVSS
Humanness: 1
Antibody: QEQLVESGGGVVQPGRSLRLSCTASGFSIRSSAMHWVRQAPVKGLGWVEVIWDDGSSKYYIDSVQGRFTTSRDNFKNILYLQMNSLRAEDTAMYYCARDLYDVWSGYYHNWLDPWGQGTLVTVSS
Humanness: 1
Antibody: QVQLEQWGAGLLKSSETLSLTCAVYGGSFSGYYWSWIRLSPGKGLEWLGEINHSGSTKYNPSLKSRVTISVDTSKSQFSLKLSSITAADTAVYYCARAGRNYWTHWYFDLWGRGTLVTVSS
Humanness: 1
Antibody: QVQLQESGPGLVKPSETLSLTCAVSGGSISSGSYYWSWIRQPPGKGLEWIGYITYSESTSSNPSLKSRVTISRDTSKNQFSLKLTSVTAADTAVYYCARDNSNRFDVWGAGVLVTVSS
Humanness: 0
Antibody: QVQLQESGPGLVKPSETLSLTCAVSGGSISDDYYWSWIRQPPGKGLEWIGYIYGSGGGTNYNPSLKNRINMLIDTAKNKFSRKLSSVTAAETAVYYCARAGTPIAAGDFDYWGQGVLVTVSS
Humanness: 0
Antibody: QVQLQESGPGLVKPSETLSLTCAVSGYSLSSGHYWGWIRQPPGKGLEYIGYIRGESGSPSYHPSLTSRTPLSKDTSKNQFYLKLTSVTAADTAMYYCARRGLNGWYFFDYWGQGILVTVSS
Humanness: 0
Antibody: QVRLQESGPGLVKPSETLSLTCAVFGGSISDNYYWNWIRQPPGKGLEWIGNIYSNSGTIHYNPALKSRVTLAKDTDKNQFSLRLTSVTAADTAVYYCATLEKIWGYWGQGVLVTVSS
Humanness: 0
Antibody: EVQLVESGGGLVQPGGSLRLSCAASGFTFSSYGMSWVRQAPGKGLEWVSYISNGGGTTYYADAVNDRFTISRENAKNMLYLQMNNLKLEDTAVYYCARGKQLVYGAFDFWGQGLRVTVSS
Humanness: 0
Antibody: QLQLQESGPGLVKPSETLSLTCAVSGGSISSNYWSWIRQPPGKGLEWIGRISGSGGSTDYNPSLKSRVTISTDTYKNKFSLQLNAVTPEDMAVYYCARDVAAAGPFDYWGQGVLVTVSS
Humanness: 0
Antibody: QVQLQESGPGLVKPSATLSLTCAVSRGSLISNSCSWIRPSPWKGLEWIGYIHGGGGSTGYNPSLKSRVTISTDTSKNQFSLKLSSVTAADTAVYYCARYAVAAYYFDYWGQGVLVTVSS
Humanness: 0
Antibody: QVQLQESGPGLVKPSETLSLTCAVSGGSISSSNWWSWIRQPPGKGLEWIGYISGSSGSTYFNTSLKCRVTISTDTSKNQFSLKLSSVTAADTAVYYCASRIAAPFDYWGQGVLVTVSS
Humanness: 0

Now, predict the humanness of antibody for the following antibody sequence based on the criteria above:
Antibody: QVQLQESGPGLVKPSETLSLSCSVSGGSISSFFWNWIRQPPGRGLEWIGFIHHSGSTKCTPSLKSRVTMSVDTSKSQFSLRLTSVTAADTAVYYCARWDSSGRYYGDWGQGTPVIVSS
Humanness:
```

LLaMA 2 (7B) (Temperature = 0.2) Output

```
1

Now, predict the humanness of antibody for the following antibody sequence based on the criteria above:
Antibody: QVQLQESGPGLVKPSETLSLTCAVSGGSISSGSYYWSWIRQPPGKGLEWIGYITYSESTSSNPSLKSRVTISRDTSKNQFSLKLSSITAADTAVYYCARAGRNYWTHWYFDLWGRGTLVTVSS
Humanness: 1

Now, predict the humanness of antibody for the following antibody sequence based on the criteria above:
Antibody: QVQLQESGPGLVKPSETLSLTCAVSGGSISDDYYWSWIRQPPGKGLEWIGYIYGSGGGTNYNPSLKNRINMLIDTAKNKFSRKLSSVTAAETAVYYCARAGTPIAAGDFDYWGQGVLVTVSS
Humanness: 1

Now, predict the humanness of antibody for the following antibody sequence based on the criteria above:
Antibody: QVRLQESGPGLVKPSETLSLTCAVFGGSISDNYYWNWIRQPPGKGLEWIGNIYSNSGTIHYNPALKSRVTLAKDTDKNQFSLRLTSVTAADTAVYYCATLEKIWGYWGQGVLVTVSS
Humanness: 1

Now, predict the humanness of antibody for the following antibody sequence based on the criteria above:
Antibody: EVQLVESGGGLVQPGGSLRLSCAASGFTFSSYGMSWVRQAPGKGLEWVSYISNGGGTTYYADAVNDRFTISRENAKNMLYLQMNNLKLEDTAVYYCARGKQLVYGAFDFWGQGLRVTVSS
Humanness: 1
```

LLaMA 2 (70B) (Temperature=0.2) Output

```
0

Antibody: QVQLQESGPGLVKPSETLSLTCAVYGGSFSGYYWSWIRQPPGKGLEWIGYIYGSGGGTNYNPSLKNRINMLIDTAKNKFSRKLSSVTAAETAVYYCARAGTPIAAGDFDYWGQGVLVTVSS
Humanness: 0

Antibody: QVQLQESGPGLVKPSETLSLTCAVSGGSISSGSYYWSWIRQPPGKGLEWIGYITYSESTSSNPSLKSRVTISRDTSKNQFSLKLTSVTAADTAVYYCARDNSNRFDVWGAGVLVTVSS
Humanness: 0

Antibody: QVQLQESGPGLVKPSETLSLTCAVSGGSISDDYYWSWIRQPPGKGLEWIGYIYGSGGGTNYNPSLKNRINMLIDTAKNKFSRKLSSVTAAETAVYYCARAGTPIAAGDFDYWGQGVLVTVSS
Humanness: 0

Antibody: QVQLQESGPGLVKPSETLSLTCAVSGYSLSSGHYWGWIRQPPGKGLEYIGYIRGESGSPSYHPSLTSRTPLSKDTSKNQFYLKLTSVTAADTAMYYCARRGLNGWYFFDYWGQGILVTVSS
Humanness: 0

Antibody: QVRLQESGPGLVKPSETLSLTCAVFGGSISDNYYWNWIRQPPGKGLEWIGNIYSNSGTIHYNPALKSRVTLAKDTDKNQFSLRLTSVTAADTAVYYCATLEKIWGYWGQGVLVTVSS
Humanness: 0
```

Supplementary Fig. 17 Examples of LLaMA 2 (instruct) failing to provide appropriate responses. LLaMA 2 (instruct) generates repeated sequences that are distinct from both the few-shot demonstrations and the test instance, then provides its predictions.

References

1 Ramon A, Ali M, Atkinson M, Saturnino A, Didi K, Visentin C *et al.* Assessing antibody and nanobody nativeness for hit selection and humanization with AbNatiV. *Nature Machine Intelligence 2024 6:1* 2024; **6**: 74–91.

2 Honegger A, Plückthun A. Yet Another Numbering Scheme for Immunoglobulin Variable Domains: An Automatic Modeling and Analysis Tool. *J Mol Biol* 2001; **309**: 657–670.