Geneverse: A Collection of Open-source Multimodal Large Language Models for Genomic and Proteomic Research

Tianyu Liu
Yale University
tianyu.liu@yale.edu

Yijia Xiao
UCLA
yijia.xiao@cs.ucla.edu

Xiao Luo
UCLA
xiao.luo@cs.ucla.edu

Hua Xu
Yale University
hua.xu@yale.edu

W. Jim Zheng
UT Health
Wenjin.J.Zheng@uth.tmc.edu

Hongyu Zhao[*]
Yale University
hongyu.zhao@yale.edu

Abstract

The applications of large language models (LLMs) are promising for biomedical and healthcare research. Despite the availability of open-source LLMs trained using a wide range of biomedical data, current research on the applications of LLMs to genomics and proteomics is still limited. To fill this gap, we propose a collection of finetuned LLMs and multimodal LLMs (MLLMs), known as Geneverse, for three novel tasks in genomic and proteomic research. The models in Geneverse are trained and evaluated based on domain-specific datasets, and we use advanced parameter-efficient finetuning techniques to achieve the model adaptation for tasks including the generation of descriptions for gene functions, protein function inference from its structure, and marker gene selection from spatial transcriptomic data. We demonstrate that adapted LLMs and MLLMs perform well for these tasks and may outperform closed-source large-scale models based on our evaluations focusing on both truthfulness and structural correctness. All of the training strategies and base models we used are freely accessible. Our codes can be found at `https://github.com/HelloWorldLTY/Geneverse`.

1 Introduction

Foundation Models (FMs) (Bommasani et al., 2021) have attracted great attention recently because of their superb functionality for handling multiple downstream tasks by adaption, especially in the landscape of Natural Language Processing (NLP) (Wu et al., 2023b; Zhao et al., 2023). In this area, we focus on one specific type of FMs, known as generative Large Language Models (LLMs) (Zhao et al., 2023). LLMs have demonstrated their strong ability to perform various tasks and domains in NLP (Bommasani et al., 2021; Zhao et al., 2023), and they have revolutionized the approaches of productivity improvement with machine intelligence (Byun et al., 2023) as well. Powerful LLMs including ChatGPT (Achiam et al., 2023) and Bard (Team et al., 2023) exhibit the ability to understand and reason as a human, and they are capable of complicated problem-solving tasks. However, since these well-trained models are both closed-source, the emergence of open-source LLM development is also important. Powerful open-source LLMs including LLaMA2 (Touvron et al., 2023) and Mistral (Jiang et al., 2023) have already become the base models of many LLMs for scientific research (Labrak et al., 2024; Wu et al., 2024a).

Beyond the success of LLMs, we are also interested in extending the input of LLMs with datasets from different modalities, including images (Liu et al., 2024) and scientific graphs (Wei et al., 2023). Such extensions of LLMs are meaningful because different modalities offer different aspects of a good representation (Wu et al., 2023a; Huh et al., 2024). Therefore, Multimodal LLMs (MLLMs) are emerging as a new research hotspot (Wu et al., 2023c). Models like Gemini (closed-source) (Team et al., 2023), GPT-4 (closed-source) (Achiam et al., 2023) and LLaVA (open-source) (Liu et al., 2024) allow the combination of images and texts as input, thus making MLLMs understand non-text data and generate responses accordingly (Liu et al., 2023c). The performance of these models demonstrates that treating images as a special language can extend the capabilities of LLMs to solve more tasks including image-based conversation, image description, and complex reasoning (Li et al., 2024; Meskó, 2023).

Although we have seen successful examples related to the development of LLMs and MLLMs, the application of LLMs towards healthcare and biomedical research presents both opportunities and challenges (He et al., 2023; Zhou et al., 2023). First, different from other areas, biomedical research highlights strict requirements for

---

[*] Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 4819–4836
November 12-16, 2024 ©2024 Association for Computational Linguistics

the factualness or truthfulness of model outputs (Labrak et al., 2024), as well as the need to validate outputs by clinical or biological experiments (Thirunavukarasu et al., 2023). When LLMs are used to assist physicians in the diagnosis of patients, an incorrect diagnosis of a patient's condition can dramatically affect the process and even cause the patient to miss the optimal time for treatment. Second, the output of LLMs should comply with established scientific knowledge (Arora and Goyal, 2023). Third, the performance of finetuned open-source LLMs has not outperformed large or proprietary models (Labrak et al., 2024) for such research. Regarding the opportunities, most of the LLMs for biomedical research focus on precision healthcare (Wu et al., 2024a) and are finetuned with information from the corpus of biomedical articles (Luo et al., 2023), while the LLMs or MLLMs for genomic or proteomic research are not well-discussed and remain to be explored. Therefore, we aim to contribute such FMs to accelerate research related to the Central Dogma (Shapiro, 2009) at the molecular level (Brownlee et al., 2023).

To fill the gap in this area, we propose a collection of LLMs and MLLMs, named as Geneverse[1], for solving genomic and proteomic problems. Our collection mainly focuses on the functions of genes and proteins, as well as their interactions. Overall, our contributions can be summarized as follows:

1. By performing a comprehensive benchmarking analysis for different base models finetuned with the gene functional description data, we created a set of LLMs with different scales for genomic research. Our benchmark results included evaluations for both the correctness of structure and the truthfulness. We selected the best base models and created an Artificial Intelligence (AI) assistant.

2. By leveraging the training datasets from both official databases as well as synthetic data from advanced LLMs, we proposed a new adaptation framework for biomedical research and built a set of powerful LLMs based on the best base model for generating the summary of gene functions and protein functions after benchmarking.

3. We introduced MLLMs in genomic and proteomic research by incorporating the structural information from proteins and expression information from genes, as well as their functional descriptions, as input. We then finetuned MLLMs for two tasks including protein identification and marker gene identification.

4. We explored the possibility of handling more tasks with finetuned FMs as well as improving the performances of models with more advanced techniques, which offered guidance for future biomedical AI research.

2 Related Work

2.1 LLMs and MLLMs for general purpose

General purpose models including GPT-4 (Achiam et al., 2023), Gemini (Team et al., 2023) and Claude (Enis and Hopkins, 2024) have demonstrated exceptional performances across various tasks in language processing, under both zero-shot and few-shots learning frameworks. However, all of the models above are closed-source. Meanwhile, there are also attractive open-source LLMs including LLaMA series (Touvron et al., 2023; Meta, 2024), Mistral (Jiang et al., 2023) and Gemma (Google, 2024), which also demonstrate comparable performances with proprietary models under certain tasks. Furthermore, the capacities of LLMs with multimodal information (e.g., audio, images, and DNA sequences, etc.) as inputs/outputs, known as MLLMs, have also shown remarkable performances under many multi-modal tasks. However, the models that serve as generalists often fail to answer certain domain-specific questions, for example, in the biological domain.

2.2 LLMs and MLLMs for biomedical research

The adaptation of LLMs and MLLMs to address tasks in biology and medicine has attracted research communities. This topic studies how to leverage these models' abilities to address specific biomedical problems. Encoder-based models including BioBERT (Lee et al., 2020) and MedBERT (Rasmy et al., 2021) have been designed for answering biomedical questions. Recently, decoder-based models including BioGPT (Luo et al., 2022), BioMedLM (Bolton et al., 2024), ClinicalGPT (Wang et al., 2023a), BioMedGPT (Luo et al., 2023), MEDITRON (Chen et al., 2023), BioMistral (Labrak et al., 2024) and Med-PALM 2 (Qian et al., 2024) are pre-trained or fine-tuned with biomedical databases and used for answering questions in genomics, proteomics, and clinical informatics, etc. Other methods including GeneGPT (Jin et al., 2024) and scELMo (Liu et al., 2023b) utilize tool

---

[1] The name is inspired by the combination of GeneGPT and universe.

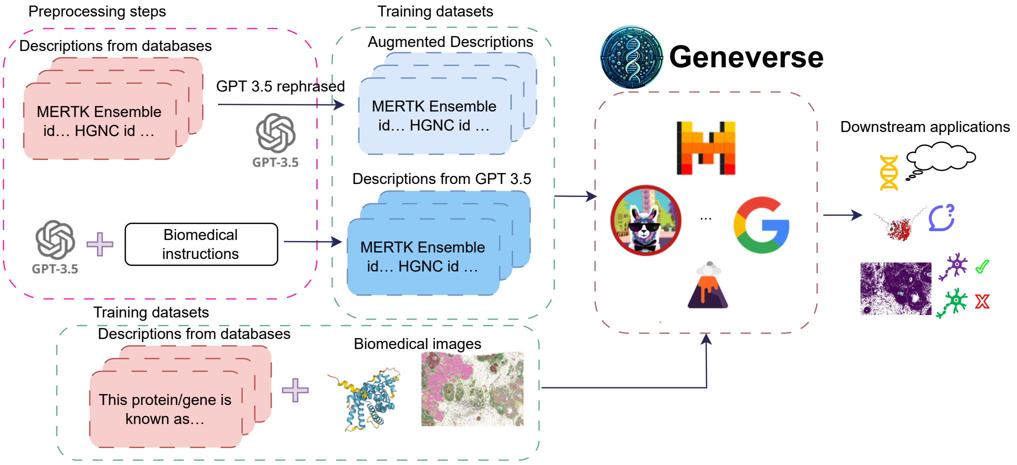

Preprocessing steps
Descriptions from databases
MERTK Ensemble id… HGNC id …
GPT 3.5 rephrased
GPT-3.5
GPT-3.5
Biomedical instructions
Training datasets
Augmented Descriptions
MERTK Ensemble id… HGNC id …
Descriptions from GPT 3.5
MERTK Ensemble id… HGNC id …
Geneverse
…
Downstream applications
Training datasets
Descriptions from databases
This protein/gene is known as…
Biomedical images

Figure 1: The landscape of Geneverse. To generate LLMs for genomic and proteomic analysis, we incorporate the training datasets from rephrased descriptions for gene functions as well as synthetic descriptions from GPT 3.5. We then adjust the base model with different strategies and select the best candidate. To generate MLLMs for genomic and proteomic analysis, we incorporate the training datasets from known databases, including both descriptions and corresponding images. We then finetune the base model with different strategies and select the best candidate. The logo of Geneverse is generated by DALLE (OpenAI, 2024).

augmentation or prompt engineering to address similar questions without extra training. By jointly modelling biomedical images and texts, methods including LLaVA-Med (Li et al., 2024) and Med-MLLM (Liu et al., 2023a), have been designed to address multimodal problems. However, some of the models limit the accessibility while others do not show satisfactory performances in our targeted tasks. Therefore, we still need a better foundation model for healthcare and biology research.

3 Methods

3.1 Problem statement

In this work, we focus on an LLM or MLLM $\mathcal{M}$, which accepts the prompt $P(T, I, m)$ as input, where T represents the text description, I represents the image and m represents the given task. For an LLM, we set I as an empty item. For an MLLM, we use both T and I. For each prompt, we have a corresponding model output as $O = \mathcal{M}(P(T, I, m))$. For a validation dataset, we compare the difference between O and the ground truth information G to evaluate model performance. The more similar O and G are (The higher their similarity score is), the better the model is.

3.2 Overview of Geneverse

Our model collection Geneverse is based on a group of pre-training open-source foundation models including LLaMA2-7B, LLaMA2-13B, Mistral-7B, Gemma-7B (Google, 2024), LLaMAPro-8B (Wu et al., 2024b), LLaMA3-8B (Meta, 2024) and LLaVA-7B. The first six models are designed for one text-based task, while the last one is designed for two image-text-based tasks. We finetuned these models based on both Parameter-Efficient FineTuning (PEFT) technology and full-parameter finetuning technology, and the finetuning setting is consistent with the Supervised Instruct Finetuning (SIFT) approach. The PEFT technology we used is Low-rank Adaption (LoRA) (Hu et al., 2022). Our training datasets are constructed based on augmented datasets from the National Center of Biotechnology Information (NCBI) (Sayers et al., 2021) and the generated synthetic datasets from GPT 3.5 with designed instructions. In the inference process, we did not use the sampling mode and thus the outputs were not affected by random seeds. The training mechanism is summarized in Figure 1. We include the details of problem settings and training design in the following subsections.

3.3 Supervised finetuning process of LLM

To finetune an LLM as an AI assistant for genomic and proteomic information queries, we prepare the training dataset with both real data and synthetic data with a data augmentation policy. To collect the real data, we use scripts to access the gene information from NCBI pages and rephrase the content of each gene with GPT 3.5 to normal sequence. To collect the synthetic data, we use the same prompt to ask GPT 3.5 to generate responses for protein-encoding genes. By combining two different datasets together, we finalize our finetuning datasets. During the finetuning process, we search the optimized hyper-parameter settings and follow the default evaluation settings of Stanford-alpaca (Taori et al., 2023) and Alpaca-lora (Wang, 2023).

3.4 Supervised finetuning process of MLLM

To finetune an MLLM as a multimodal AI assistant for genomic and proteomic information query, we prepare the training datasets for two representative tasks in biology. Learning a protein's function based on its structure (the extra modality) and identifying marker genes for certain cell types based on the locations of cells (the extra modality) is difficult but meaningful (Bernhofer et al., 2021; Yuan et al., 2024). Therefore, we consider the protein function classification task (known as the protein task) for protein structures and the maker genes identification task for spatial transcriptomics (known as the marker gene task). For the protein task, we downloaded the protein structure images from the DeepMind Alphafold2 website (Jumper et al., 2021) with the fixed capture angle. We load the image information using Pymol (Yuan et al., 2017) and construct the instruction tuning dataset. We then finetune LLaVA with LoRA based on our datasets. We search the optimized hyper-parameter settings and use the default evaluation settings from LLaVA v1.5. For the marker gene task, the pipeline is the same but we utilize different image datasets (Lin et al., 2020) as well as different instructions.

3.5 Post-processing steps

To tackle the two problems of the outputs of LLMs, known as the incorrect inference of numerical features of genes and redundant descriptions (also known as degeneration) of genes, we design two approaches as post-processing methods to improve the quality of model outputs.

Our post-processing methods are based on the idea of tool-augmented design (Li et al., 2023). To handle the first problem, we collect the gene ID information from public databases (Sayers et al., 2021) and replace the information in the outputs of LLMs with the numerical information from the database. To handle the second problem, since the redundant information is related to the aliases of genes, we delete the content related to aliases in the model outputs and insert the aliases based on the correct information from our database.

3.6 Evaluation

In the evaluation process, we consider two different aspects for evaluation, known as grader for truthfulness and grader for structural correctness. The former grader works for evaluating whether the outputs from LLMs or MLLMs match the fact, while the latter grader works for evaluating whether the outputs from LLMs or MLLMs match the structure requirements from the prompt. Higher scores mean better model performance for each grader.

We first consider the grader for the truthfulness. For the gene description task, we assign 1 for the outputs of LLMs which match the major gene function and gene name, and 0 otherwise. For the protein task, we compute the length of the largest common string between the output of MLLMs and the correct result and divide it by the length of the correct result as the metric. For the marker gene task, we assign 1 for the outputs of MLLMs which match the ground truth marker gene-cell type relation, and 0 otherwise. We finally average the scores of different samples as the final score.

We then consider the grader for structural correctness. For all of the tasks we evaluate, we assign 1 for the outputs if they match the structural requirements from the prompt, and 0 otherwise. We finally average the scores of different samples as the final score.

In order to validate our proposed metric, we also considered the use of BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002) and ROUGEk (k Grams Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) scores as additional evaluations. BLEU score is based on the overlap value of $n-grams$ for two strings. ROUGEk score considers the overlap of k grams between the reference text and generated text, and here $k = 1$. In order to evaluate the quality of embeddings from generated texts, we computed the Normalized Mutual Information (NMI) score (Pedregosa et al., 2011) between Leiden clusters (Traag et al., 2019) and observed labels. All metrics are in [0,1], and higher values mean better model performances.

4 Results

4.1 Benchmarking LLMs for summarizing of gene functions

We first performed benchmarking analysis by fixing the training datasets to find the best base model and the best training strategy. Our task is to generate the summary of gene functions based on

Model	Factual Score	Structural Score	Average	BLEU	ROUGE1	Finetuning?
LLaMA2-7B (LoRA)	0.850	1.000	0.925	**0.395**	0.580	Y
LLaMA2-7B (full)	0.900	0.900	0.900	0.345	0.546	Y
LLaMA2-7B (RAG)	0.650	0.800	0.725	0.155	0.324	N
LLaMA2-13B (LoRA)	0.900	1.000	0.950	0.374	**0.600**	Y
LLaMA2-13B (RAG)	0.600	1.000	0.800	0.196	0.358	N
Mistral-7B (LoRA)	0.950	1.000	**0.975**	0.349	0.592	Y
Mistral-7B (RAG)	0.650	0.050	0.350	0.127	0.269	N
Gemma-7B (LoRA)	0.950	1.000	**0.975**	**0.425**	**0.613**	Y
Gemma-7B (RAG)	0.350	1.000	0.675	0.073	0.183	N
LLaMAPro-8B (LoRA)	0.950	1.000	**0.975**	**0.387**	**0.627**	Y
LLaMAPro-8B (RAG)	0.000	0.000	0.000	0.000	0.000	N
LLaMA3-8B (LoRA)	0.900	1.000	0.950	0.324	0.541	Y
LLaMA3-8B (RAG)	0.250	0.400	0.325	0.011	0.196	N
GPT 3.5	0.900	1.000	0.950	0.123	0.306	N
GPT 3.5 (RAG)	0.850	1.000	0.925	0.144	0.312	N
GPT 4	0.850	1.000	0.925	0.118	0.309	N
BioMedLM	0.000	0.000	0.000	0.046	0.193	N
bioGPT	0.000	0.000	0.000	0.041	0.186	N
GeneGPT	0.750	1.000	0.875	0.141	0.304	N
BioMistral	0.000	0.000	0.000	0.000	0.000	N

Table 1: Evaluations for the task: Gene Function Description. We boldfaced the top three methods ranked by Average scores, BLEU scores and ROUGE1 scores.

the prompt only containing the task description. Out of $\sim$ 20,000 genes (Dolgin, 2017) in total, we randomly selected 20 genes referred from (Jin et al., 2024) and their observed descriptions to construct the evaluation datasets. Since our prompt also contains requirements for the structure of model outputs, we consider a novel evaluation of the output structure as well as the truthfulness of the contents. The illustration of evaluators is shown in Extended Data Figure 1. The criteria for computing the accuracy are summarized in the Methods section. We used accuracy as a metric to evaluate these two types of metrics. We also included the BLEU score as an additional metric. The methods we included in this section for benchmarking are LLaMA2-7B (LoRA), LLaMA2-7B (full), LLaMA2-7B (RAG), LLaMA2-13B (LoRA), LLaMA2-13B (full), Mistral-7B (LoRA), Mistral-7B (RAG), Gemma-7B (LoRA), Gemma-7B (RAG), LLaMAPro-8B (LoRA), LLaMAPro-8B (RAG), LLaMA3-8B (LoRA) and LLaMA3-8B (RAG). Here RAG represents the retrieval-augmented generation technology (Lewis et al., 2020). We also considered including state-of-the-art models for general models and more biomedical-focused models, including GPT 3.5 (Brown et al., 2020), GPT 3.5 (RAG), GPT 4 (Achiam et al., 2023), BioMedLM (Bolton et al., 2024), bioGPT (Luo et al., 2022), GeneGPT (Jin et al., 2024), and BioMistral (Labrak et al., 2024). Details of baselines are included in Appendix A. Our evaluations are shown in Table 1.

Based on our evaluation results, we found that Geneverse based on finetuning Mistral-7B, Gemma-7B, and LLaMAPro achieved the best results by averaging the scores from both content level and structure level. The latter two models also had better performances evaluated by BLEU and ROUGE1 scores. We computed the Pearson correlation between BLEU (ROUGE1) scores and average scores, which showed significant positive correlations (coefficient=0.76, p-value=9.5e-5 for BLEU, and coefficient=0.80, p-value=1.6e-5 for ROUGE1), which lends support of our proposed metric. We note that finetuned Mistral-7B outperformed finetuned models from the LLaMA family with different scales, but also outperformed large-scale general LLMs including GPT 3.5 and GPT 4. On the other hand, other methods including bioGPT and BioMistral did not generate meaningful descriptions for gene functions. Meanwhile, methods based on RAG also did not satisfy our requirements and the generated outputs sometimes have multiple paragraphs. Therefore, the structure score of Mistral-7B (RAG) is low. Our conclusions also agree with recent research (Wu et al., 2024c) challenging the capacity of RAG in addressing biomedical problems. Therefore, we still need to finetune models for domain-specific tasks.

Moreover, the model outputs under different cases all failed in summarizing the numerical information of genes including Ensemble id (Howe et al., 2021) and HUGO gene nomenclature committee (HGNC) id (Povey et al., 2001). To tackle this problem, we have developed a retrieval database to replace the incorrect information of model output as an important post-processing step. By adding this component, our generated summary is precise in describing both the functional and identifiable information of genes. The details of the post-processing method are included in the Methods section. The outputs of different models for evaluations are in Supplementary File 1.

4.2 LLM finetuning with diverse datasets

As shown in Figure 1, we also used the outputs generated by GPT-3.5 to enrich the data diversity for model training. To ensure the correctness of training datasets, we only focused on protein-encoding genes for the outputs of GPT 3.5. While for non-coding genes and pseudogenes, we only used the descriptions from NCBI. There are two advantages brought by such a data augmentation approach. The first advantage is that GPT 3.5 has specific knowledge of certain genes, thus helping our AI assistant answer questions related to genes that are not collected in the NCBI database. The other advantage is that GPT 3.5's output is more like human language and has more comprehensive functional information to enhance the quality of the model output. A comparison of outputs for models trained with different datasets is shown in Extended Data Figure 4. From this figure, we can find that introducing more data can help LLMs generate more informative summaries for the gene GLI1 by reducing redundant information. Details of model outputs are summarized in Supplementary File 1.

To visualize the outputs of Geneverse after finetuning, we utilize the embeddings layer from OpenAI to transfer the descriptions of genes into numerical vectors (as gene embeddings) and visualize the gene embeddings using Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018). In Figure 2 (a), we label the genes using gene functional information from (Theodoris et al., 2023). We can find that genes with similar functions display shared patterns. Details of labels are summarized in Extended Data Figure 5. Furthermore, the Gene Ontology Enrichment Analysis (GOEA) (Ashburner et al., 2000; Aleksander et al., 2023; Fang et al., 2023) results are summarized in Extended Data Figure 2, which include top 10 pathways ranked by $-log($Adjusted P-value$)$ (P-value corrected by Bonferroni method) for the top 10 clusters ranked by the number of genes. We find many pathways related to important biological activities such as RNA transcription and metabolism. Therefore, our sentences cover the similarities and differences of the genes' functions.

4.3 Finetuning an MLLM for genomic and proteomic application

By incorporating the image information into the model training process, we can construct domain-specific MLLMs for several downstream tasks. We designed two tasks and demonstrated the potential of MLLMs for analyzing proteins and spatial transcriptomics by finetuning LLaVA on these two novel tasks, shown in Extended Data Figure 3. The first task is known as protein classification, which means we intend to identify the specific protein with its image. The images of protein structure for training and testing come from the databases of AlphaFold2 (Jumper et al., 2021). Using visual representation of proteins can highlight important structural motifs (El Khoury et al., 2023), domains (Petsko and Ringe, 2004), and interactions (Hugo et al., 2013) for understanding protein function discoveries. Using image representation of proteins to analyze their structural properties and functions is well discussed (Roy et al., 2011; S Bernardes, 2013; Whisstock and Lesk, 2003; Bernhofer et al., 2021). There exist uniqueness and advantages of analyzing proteins with their corresponding images (2D information) rather than 3D structures, which include reducing the cost of experiments (Sara et al., 2021), improving the visualization of functional information (Roy et al., 2010), enriching prior information for prediction (McGuffin et al., 2019), and others. The second task is known as spatial marker gene identification, which means that for a gene g, we intend to know whether it is a maker gene of cell type c based on the spatial transcriptomic data. Understanding the marker genes is important for analyzing cell-cell interaction in the context of spatial transcriptomic data (Wang et al., 2023b). The spatial transcriptomic data for training and testing come from (Lin et al., 2020) collected from human breast tissue. Both tasks have not been discussed previously and are important for biomedical research. As for baselines, we included LLaVA-7B, MoE-LLaVA, GPT-4, and Gemini. Our finetuned version of LLaVA is known as LLaVA-7B (LoRA). Details of baselines are included in Appendix A.

Model	Factual score	Structural score	Average	Finetuning?
LLaVA-7B	0	1	0.5	N
MoE-LLaVA	0	1	0.5	N
GPT-4v	0	0	0	N
GPT-4o	0	0	0	N
Gemini	0	0	0	N
LLaVA-7B (LoRA)	0.29	1	0.645	Y

Table 2: Evaluations for the task: Protein Classification.

For the first task, we compared the three baseline models with our finetuned model, and the results are summarized in Table 2. Based on this table, we found that the baselines without finetuning cannot generate correct labels for proteins from images. Moreover, the output of GPT-4 does not even follow the structure requirements from our prompts, which questioned the contributions of large-scale models for this task. However, if we finetuned LLaVA based on the training dataset, the factual score increases while the high structural score is also preserved. Therefore, leveraging domain-specific knowledge is an important step to handle this classification problem. However, only partial protein names were matched for every sample, thus the finetuning process still needs further investigation for performance improvement.

For the second task, we compared the three baseline models with our finetuned model, and the results are summarized in Table 3. According to this table, the results from GPT-4 are slightly better

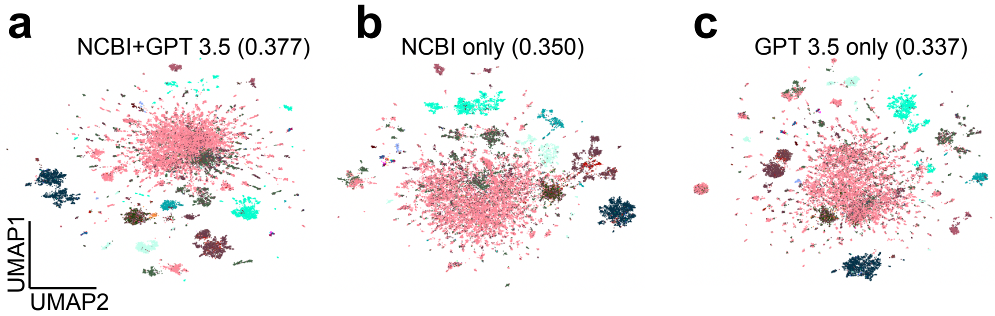


Figure 2: UMAPs for the gene embeddings colored by gene functional information. Panels (a)-(c) represent the outputs of LLMs trained based on datasets from NCBI+GPT 3.5, NCBI only and GPT 3.5. We report the NMI score of each embeddings followed by their sources.

Model	Factual Score	Structural Score	Average	Finetuning?
LLaVA-7B	0.5	0.425	0.4625	N
MoE-LLaVA	0.5	1	0.75	N
GPT-4v	0.575	0.125	0.35	N
GPT-4o	0	0	0	N
Gemini	0.6	0.025	0.3125	N
LLaVA-7B (LoRA)	0.9	1	0.95	Y

Table 3: Evaluations for the task: Marker Genes.

than random guesses, and all the baseline models did not show a strong ability to identify marker genes for different cell types given the spatial figures colored by gene expression levels. However, LLaVA-7B and GPT-4 did not perform well in following the structural requirements. Moreover, prior knowledge in the training process might help in this prediction process. For example, all the methods predicted gene CD4 as a marker gene of CD4 T cell, which matched the biological knowledge, and the reason is from the existing information of the pre-training text databases. Moreover, our finetuned LLaVA outperformed other models again in the factual scores, which had 90% accuracy. In addition, the outputs of our finetuned LLaVA also followed the requirements of structure existing in the prompts. Therefore, finetuning an MLLM with image information could boost the ability of models to learn about specific biological problems, especially for tasks related to spatial transcriptomics.

4.4 Sensitivity analysis

In this section, we analyzed the factors that may affect the finetuning process, including hyper-parameters and data ablation. The results are summarized in Figure 3. We utilized default hyper parameters to finetune different LLMs and MLLMs, and selected the best base model used in this section for sensitivity analysis.

We first discuss the sensitivity of LLM training. For hyper-parameter tuning, we considered tuning the number of epochs and cut-off length (it means we use different lengths to truncate one sentence). For data ablation, we considered three conditions: 1. only NCBI data (NCBI), 2. only GPT 3.5 data (GPT 3.5), and 3. the combination of NCBI and GPT data (NCBI+GPT 3.5). Extended Data Figure 6 shows the results of data ablation, which suggests that integrating both NCBI and GPT 3.5 leads to the best performance of LLM in this task. The LLM based on GPT 3.5 to finetune does not accurately summarise the function of non-protein-encoding genes, as shown in Supplementary File 1. Moreover, LLMs trained based on the combination of datasets also generated a better representation. Comparing with outputs of LLMs trained based on the second condition (shown in Figure 2 (b)) and the third condition (shown in Figure 2 (c)), embeddings from our current design also have the highest NMI score. Figure 3 (a) shows the relation between epochs and model performance, which suggests that more epochs lead to better model performance in the description generation task. Figure 3 (b) shows the relation between cut-off length and model performance, which suggests that cut-off length does not affect the performance of LLM in this task.

We then discuss the sensitivity of MLLM training. Based on our observations in the training process of LLMs, we only considered hyper-parameter tuning of the number of epochs. The relation between the number of epochs and model performance is illustrated in Figures 3 (c) and (d) for the protein task and in Figure for the gene task. According to Figure 3 (c), increasing the number of epochs leads to a model performance drop in the training process of MLLMs for the protein task, reflected in the average score and the truthful score. However, according to Figure 3 (d), increasing the number

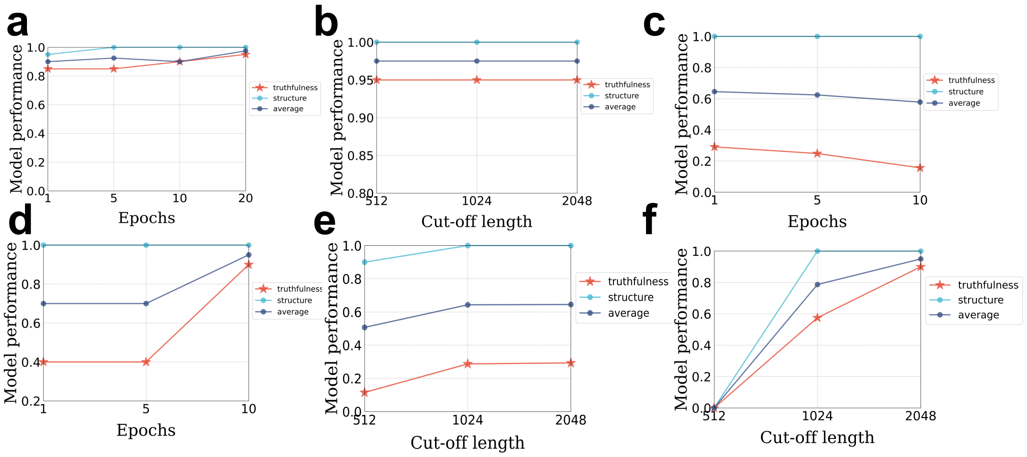


Figure 3: Results of sensitivity analysis for the training of different models. (a) The relation between the number of epochs and model performance of LLMs. (b) The relation between the number of cut-off length and model performance of LLMs. (c) The relation between the number of epochs and model performance of MLLMs for the protein task. (d) The relation between the number of epochs and model performance of MLLMs for the gene task. (e) The relation between the number of cut-off length and model performance of MLLMs for the protein task. (f) The relation between the number of cut-off length and model performance of MLLMs for the gene task.

of epochs leads to better model performance in the training process of MLLMs for the gene task, reflected in the average score and the score for truthfulness. Similarly, conclusions based on cut-off length for MLLMs are presented in Figures 3 (e) and (f). Therefore, task variability affects the relationship between hyper-parameters and model performance, possibly due to the complexity of biological data and questions.

We also illustrate the relation between top-k candidates (It means the number of top samples retrieved from the reference data ranked by vector similarity) and model performance for RAG-based LLMs in Extended Data Figure 7. This figure shows that increasing top-k candidates leads to a model performance drop, especially for the score of structure. The RAG-based LLM using the largest top-k generated results free of structure, which was not suitable for the application of a useful AI assistant. Moreover, we considered using Lora+ (Hayou et al., 2024) to replace LoRA in the training process, but the average score did not change. The running time of the finetuning process is summarized in Supplementary File 2.

5 Discussion

Foundation models, especially LLMs, are powerful tools for scientific research. In this manuscript, we introduced Geneverse, a collection of LLMs and MLLMs tailored for genomic and proteomic tasks. We compared different strategies and base models for generating accurate function descriptions of genes based on the adapted LLMs. Moreover, we also explored the applications of MLLMs for a protein-level task and a gene-level task, as an example of leveraging the multimodal information. The models in Geneverse demonstrated better performance in the selected tasks than closed-source models, thus advocating the contributions of open-source base models for scientific research. We also discussed the factors affecting model adaptation, which advanced this field by offering suggestions.

Our future work aims to increase the generation quality of outputs from Geneverse with more advanced techniques. We will also investigate why RAG-based LLMs fail and improve them. Moreover, we will collaborate with other institutes to access more high-quality datasets for model adaptation. The next-generation multimodal AI assistant for genomics and proteomics will come soon.

6 Limitations

First, finetuning an FM or LLM requires a large amount of computing resources. Based on our testing results, we need at least one H100 or equivalent GPU cores to perform finetuning with LoRA. Secondly, to utilize methods based on closed-source large models, including GPT series and GeneGPT, we need to use the OpenAI API to call the method, thus we need additional resources for deploying such a method. Finally, a good domain-specific FM should align with the version of state-of-the-art base models. Therefore, we should keep our model updated by aligning the best version of Geneverse with the best open-source base model. Moreover, due to the limitation of computing resources, we only considered 7B-level and 13B-level models in the current version. In the future, we need to explore LLMs or MLLMs with larger scales. Finally, our evaluations for the domain-specific FMs only focus on three tasks and more tasks should be included in the future.

7 Ethics Statement

The users are solely responsible for the content they generate with models in Geneverse, and there are no mechanisms in place for addressing harmful, unfaithful, biased, and toxic content disclosure. Any modifications of the models should be released under different version numbers to keep track of the original models related to this manuscript.

The target of current Geneverse only serves for academic research. The users cannot use it for other purposes. Finally, we are not responsible for any effects of the use of the model.

8 Codes and Reproductivities

The codes and datasets we used in this manuscript can be found in `https://github.com/HelloWorldLTY/Geneverse`. We used the MIT licenses. All of the model weights can be accessed via requests. We used two NVIDIA H100 GPU cores for model training, while the maximum system memory for training is 100 GB.

9 Acknowledgements

This project is supported in part by NIH grants R01 GM134005, U24 HG012108, 1R01AG066749, and DoD W81XH2210162, and OpenAI's Researcher Access Program.

10 Conflict of interests

No authors declared conflict of interests.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, et al. 2023. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031.

Sanjeev Arora and Anirudh Goyal. 2023. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936*.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Michael Bernhofer, Christian Dallago, Tim Karl, Venkata Satagopam, Michael Heinzinger, Maria Littmann, Tobias Olenyi, Jiajun Qiu, Konstantin Schütze, Guy Yachdav, et al. 2021. Predictprotein-predicting protein structure and function for 29 years. *Nucleic acids research*, 49(W1):W535–W540.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, et al. 2024. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Alexander EI Brownlee, James Callan, Karine Even-Mendoza, Alina Geiger, Carol Hanna, Justyna Petke, Federica Sarro, and Dominik Sobania. 2023. Enhancing genetic improvement mutations using large language models. In *International Symposium on Search Based Software Engineering*, pages 153–159. Springer.

Courtni Byun, Piper Vasicek, and Kevin Seppi. 2023. Dispensing with humans in human-computer interaction research. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–26.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Elie Dolgin. 2017. The most popular genes in the human genome. *Nature*, 551(7681):427–432.

Gilbert El Khoury, Wael Azzam, and Joseph Rebehmed. 2023. Pyprotif: a pymol plugin to retrieve and visualize protein motifs for structural studies. *Amino Acids*, 55(10):1429–1436.

Maxim Enis and Mark Hopkins. 2024. From llm to nmt: Advancing low-resource machine translation with claude. *Preprint*, arXiv:2404.13813.

Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. 2023. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics*, 39(1):btac757.

Google. 2024. Gemma. `https://huggingface.co/google/gemma-7b`.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. 2024. Lora+: Efficient low rank adaptation of large models. *arXiv preprint arXiv:2402.12354*.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*.

Kevin L Howe, Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, Jyothish Bhai, et al. 2021. Ensembl 2021. *Nucleic acids research*, 49(D1):D884–D891.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Willy Hugo, Wing-Kin Sung, and See-Kiong Ng. 2013. Discovering interacting domains and motifs in protein–protein interactions. *Data Mining for Systems Biology: Methods and Protocols*, pages 9–20.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2024. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2):btae075.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023. API-bank: A comprehensive benchmark for tool-augmented LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3102–3116, Singapore. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yingxin Lin, Yue Cao, Hani Jieun Kim, Agus Salim, Terence P Speed, David M Lin, Pengyi Yang, and Jean Yee Hwa Yang. 2020. scclassify: sample size estimation and multiscale classification of cells using single and multiple reference. *Molecular systems biology*, 16(6):e9389.

Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, et al. 2023a. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine*, 6(1):226.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Tianyu Liu, Tianqi Chen, Wangjie Zheng, Xiao Luo, and Hongyu Zhao. 2023b. scelmo: Embeddings from language models are good learners for single-cell data analysis. *bioRxiv*, pages 2023–12.

Zhengliang Liu, Aoxiao Zhong, Yiwei Li, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, Peng Shu, Cheng Chen, Sekeun Kim, et al. 2023c. Radiology-gpt: A large language model for radiology. *arXiv preprint arXiv:2306.08666*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *Preprint*, arXiv:2308.09442.

Liam J McGuffin, Recep Adiyaman, Ali HA Maghrabi, Ahmad N Shuid, Danielle A Brackenridge, John O Nealon, and Limcy S Philomina. 2019. Intfold: an integrated web resource for high performance protein structure and function prediction. *Nucleic acids research*, 47(W1):W408–W413.

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

Bertalan Meskó. 2023. The impact of multimodal large language models on health care's future. *Journal of Medical Internet Research*, 25:e52865.

Meta. 2024. Llama3.

OpenAI. 2024. Dalle. `https://help.openai.com/en/articles/6640875-how-should-i-credit-dall-e-in-my-work`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Gregory A Petsko and Dagmar Ringe. 2004. *Protein structure and function*. New Science Press.

Sue Povey, Ruth Lovering, Elspeth Bruford, Mathew Wright, Michael Lush, and Hester Wain. 2001. The hugo gene nomenclature committee (hgnc). *Human genetics*, 109:678–680.

Jili Qian, Zhengyu Jin, Quan Zhang, Guoqing Cai, and Beichang Liu. 2024. A liver cancer question-answering system based on next-generation intelligence and the large model med-palm 2. *International Journal of Computer Science and Information Technology*, 2(1):28–35.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86.

Ambrish Roy, Alper Kucukural, and Yang Zhang. 2010. I-tasser: a unified platform for automated protein structure and function prediction. *Nature protocols*, 5(4):725–738.

Ambrish Roy, Dong Xu, Jonathan Poisson, and Yang Zhang. 2011. A protocol for computer-based protein structure and function prediction. *Journal of visualized experiments: JoVE*, (57).

Juliana S Bernardes. 2013. A review of protein function prediction under machine learning perspective. *Recent patents on biotechnology*, 7(2):122–141.

Samia Tasnim Sara, Md Mehedi Hasan, Ahsan Ahmad, and Swakkhar Shatabda. 2021. Convolutional neural networks with image representation of amino acid sequences for protein function prediction. *Computational Biology and Chemistry*, 92:107494.

Eric W Sayers, Jeffrey Beck, Evan E Bolton, Devon Bourexis, James R Brister, Kathi Canese, Donald C Comeau, Kathryn Funk, Sunghwan Kim, William Klimke, et al. 2021. Database resources of the national center for biotechnology information. *Nucleic acids research*, 49(D1):D10.

James A Shapiro. 2009. Revisiting the central dogma in the 21st century. *Annals of the New York Academy of Sciences*, 1178(1):6–28.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. 2023. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. 2019. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233.

Eric J. Wang. 2023. Alpaca-lora. `https://github.com/tloen/alpaca-lora`.

Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. 2023a. Clinicalgpt: large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968*.

Xinyi Wang, Axel A Almet, and Qing Nie. 2023b. The promising application of cell-cell interaction analysis in cancer from single-cell and spatial transcriptomics. In *Seminars in cancer biology*. Elsevier.

Lanning Wei, Zhiqiang He, Huan Zhao, and Quanming Yao. 2023. Unleashing the power of graph learning through llm-based autonomous agents. *arXiv preprint arXiv:2309.04565*.

James C Whisstock and Arthur M Lesk. 2003. Prediction of protein function from protein sequence and structure. *Quarterly reviews of biophysics*, 36(3):307–340.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024a. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045.

Chengyue Wu, Yukang Gan, Yixiao Ge, Zeyu Lu, Jiahao Wang, Ye Feng, Ping Luo, and Ying Shan. 2024b. Llama pro: Progressive llama with block expansion. *arXiv preprint arXiv:2401.02415*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. 2023a. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F Wong, and Lidia S Chao. 2023b. A survey on llm-gernerated text detection: Necessity, methods, and future directions. *arXiv preprint arXiv:2310.14724*.

Kevin Wu, Eric Wu, and James Zou. 2024c. How faithful are rag models? quantifying the tug-of-war between rag and llms' internal prior. *arXiv preprint arXiv:2404.10198*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023c. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

Shuguang Yuan, HC Stephen Chan, and Zhenquan Hu. 2017. Using pymol as a platform for computational drug design. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 7(2):e1298.

Zhiyuan Yuan, Fangyuan Zhao, Senlin Lin, Yu Zhao, Jianhua Yao, Yan Cui, Xiao-Yong Zhang, and Yi Zhao. 2024. Benchmarking spatial clustering methods with spatially resolved transcriptomics data. *Nature Methods*, pages 1–11.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Hongjian Zhou, Boyang Gu, Xinyu Zou, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Xian Wu, et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

A Explanations of baseline models.

Here we highlight the summary of different selected open-source LLMs and MLLMs used in this manuscript for either baselines or base models.

For LLMs, we have the following models:

- LLaMA2: LLaMA2 is a collection of pretrained and finetuned LLMs for chatting and other downstream tasks. The base model we choose for finetuning is from LLaMA2-Chat, which is optimized for dialogue use cases.
- LLaMA3: LLaMA3 is a collection of pretrained and finetuned LLMs for chatting and other downstream tasks. The base model we choose for finetuning is from LLaMA3-Instruct, which is optimized for dialogue use cases.
- Mistral: Mistral is an open-source LLM that is based on a different framework compared with LLaMA2. It can also be finetuned into different cases and we also use its chat mode as the base model.
- Gemma: Gemma is also an open-source LLM which is the state-of-the-art 7B-base model. It can also be finetuned into different cases and we also use its chat mode as the base model.
- LLaMAPro: LLaMAPro is also an open-source LLM modified based on LLaMA2. It offers extra blocks for finetuning and thus this approach can reduce the catastrophic forgetting of a finetuned LLM.
- GPT 3.5: GPT 3.5 is a generative pretraining model finetuned for the chatting task. We used the OpenAI API to access this model.
- GPT 4: GPT 4 is a generative pretraining model finetuned for the chatting task. We used the OpenAI API to access this model.

GPT 4 also has the ability to handle image data as input.

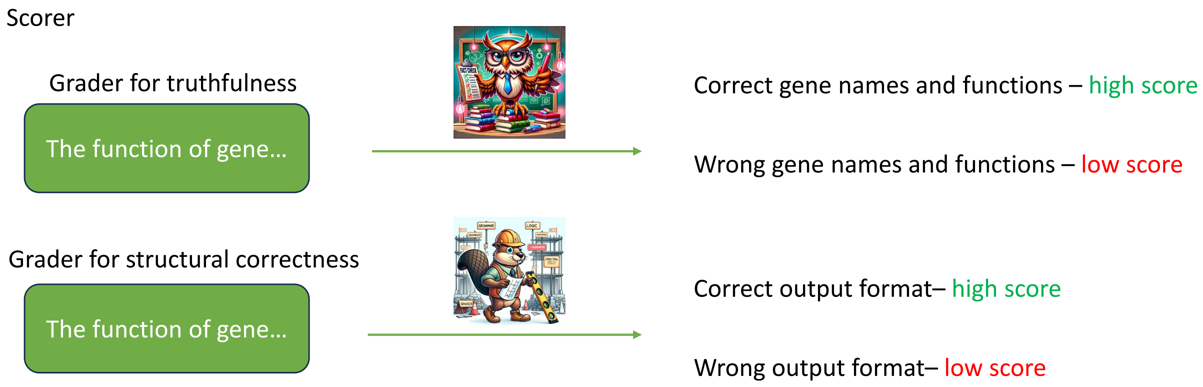
- bioGPT: bioGPT is a finetuned GPT-2 model based on medical data. It is designed for the applications of LLMs in biomedical cases.
- BioMedLM: BioMedLM is a finetuned GPT-2 model based on PubMed data. It is designed for the applications of LLMs in biomedical cases, especially for question-answering (QA) problems.
- GeneGPT: GeneGPT is a tool-augmented method for tasks related to medical queries. GeneGPT is based on searching documents to enrich the information in the prompts. We use the OpenAI API to access the base model of this method.
- bioMistral: bioMistral is a finetuned Mistral model based on medical data. It is designed for the applications of LLMs in biomedical cases.

For MLLMs, we have the following models:

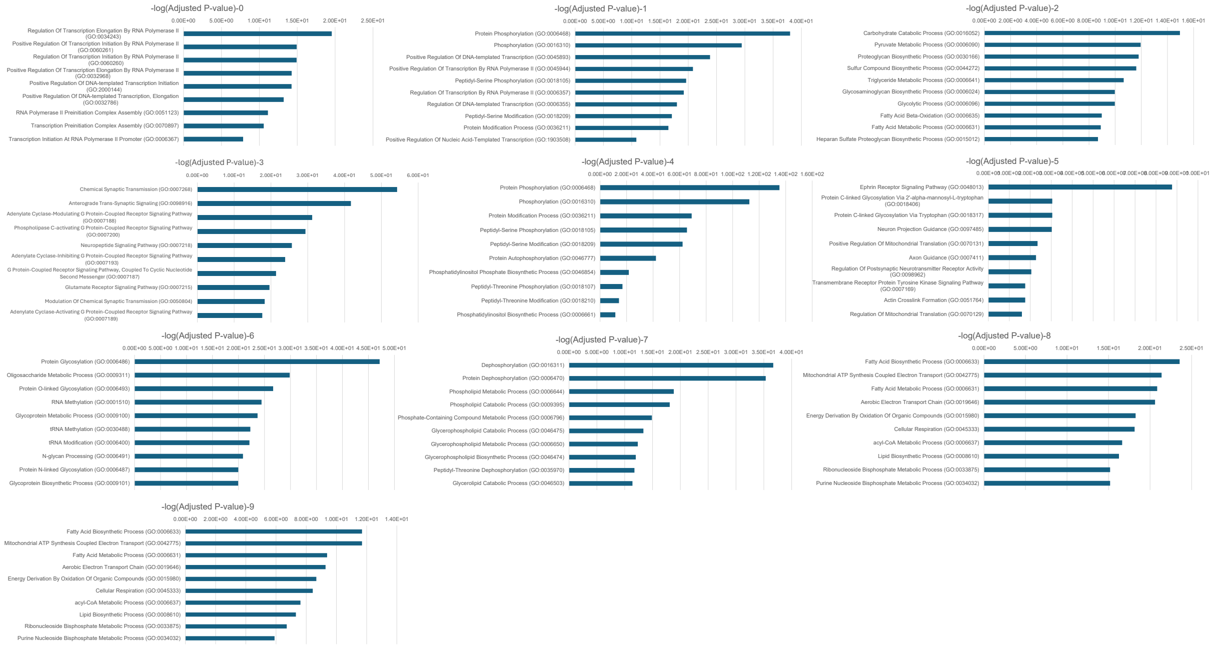
- LLaVA: LLaVA is a collection of pre-trained and finetuned MLLMs for chatting and other downstream tasks. The base model we choose for finetuning is from LLaVA-v1.5, which is recommended in the tutorials of LLaVA.
- MoE-LLaVA: MoE-LLaVA is the mixture-of-experts (MOE) version LLaVA. It utilizes different experts in the fine-tuning stages with the same inputs from embeddings of images and embeddings of words. We access this model based on the Huggingface Space resource.
- Gemini: Gemini is a closed-source MLLM for various tasks. We use an interactive platform from Google to access this model.
- GPT-4: Introduced before. Here the GPT-4 model includes GPT-4v and GPT-4o, which represent their abilities to handle image-based data as prompts. GPT-4o is the most advanced model from OpenAI.

B Supplementary figures

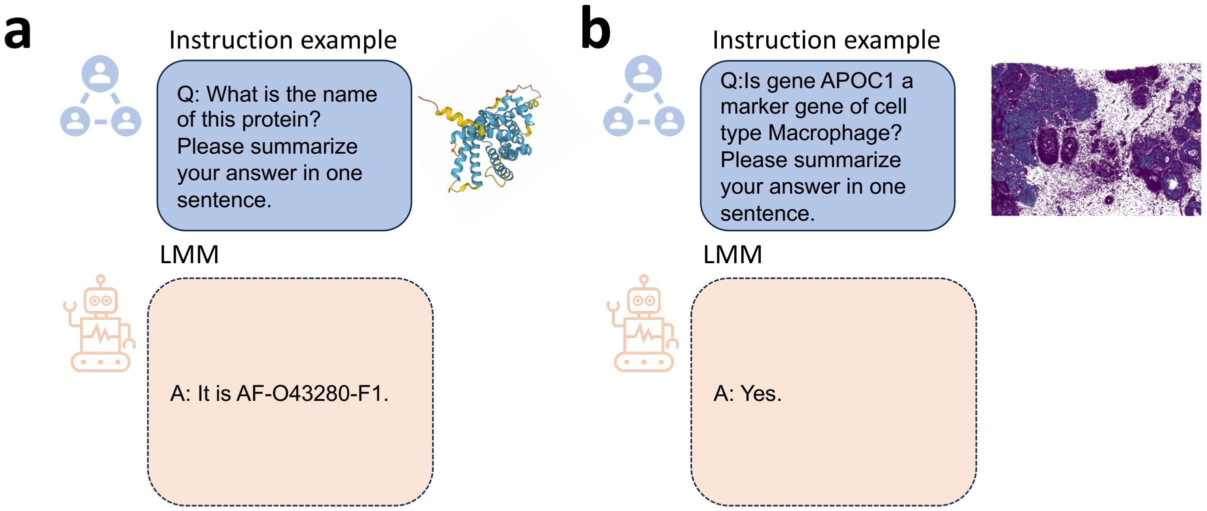
This section contains all the supplementary figures, starting from next page.


Scorer
Grader for truthfulness
The function of gene…
Correct gene names and functions – high score
Wrong gene names and functions – low score
Grader for structural correctness
The function of gene…
Correct output format– high score
Wrong output format– low score

Extended Data Fig. 1: Definition of different evaluators or scorers. For the scorer focusing on truthfulness, we evaluate the matching level of model outputs for the description of gene properties and gene functions. For the scorer focusing on structural correctness, we evaluate the correctness of the structure of model outputs by comparing them to the limitations in the prompt. The logos of scorers are generated by DALLE (OpenAI, 2024).



Extended Data Fig. 2: Figures of GOEA results. Each figure represents top 10 pathways in one cluster, and the pathways are ranked by $-\log(\text{Adjusted P-value})$.



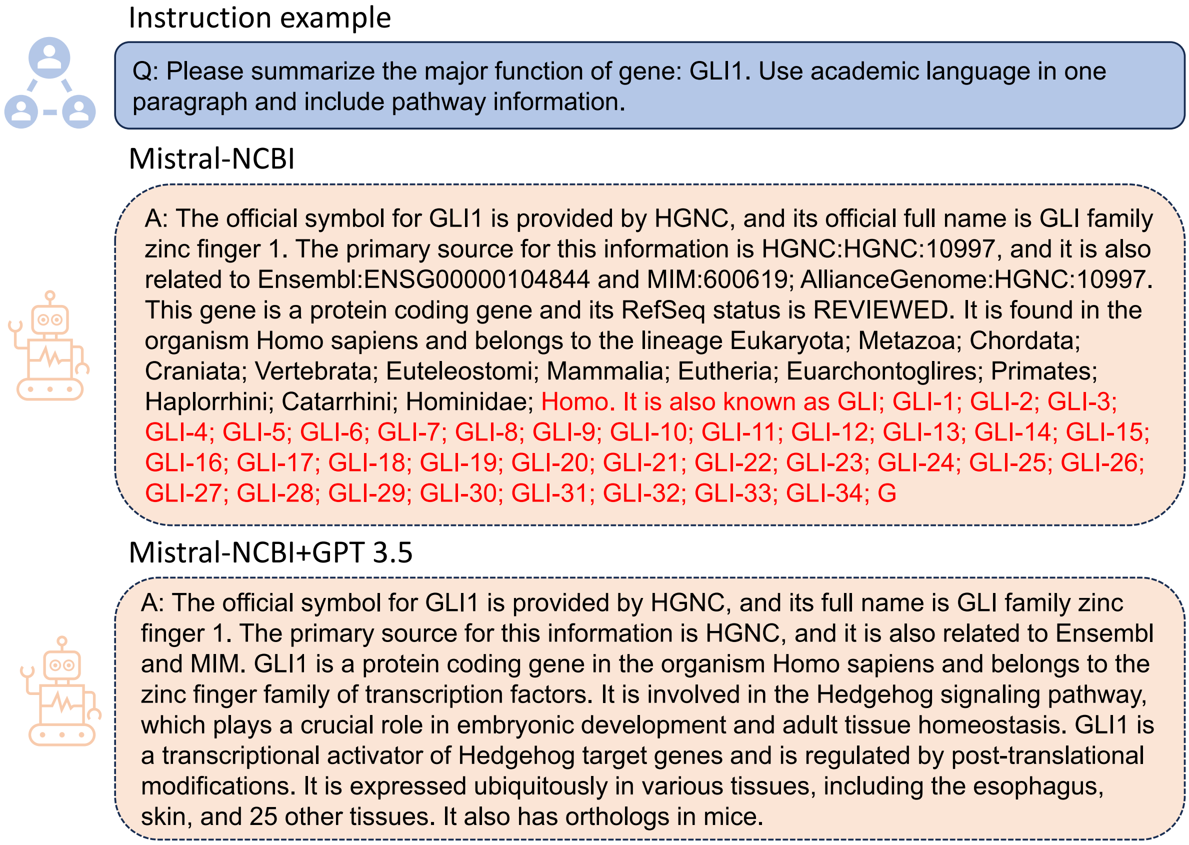
Q: What is the name of this protein? Please summarize your answer in one sentence.

A: It is AF-O043280-F1.

Q:Is gene APOC1 a marker gene of cell type Macrophage? Please summarize your answer in one sentence.

A: Yes.

Extended Data Fig. 3: Figures to illustrate our proposed two tasks. (a) The QA setting of protein function inference task. (b) The QA setting of marker gene identification task.



Q: Please summarize the major function of gene: GLI1. Use academic language in one paragraph and include pathway information.

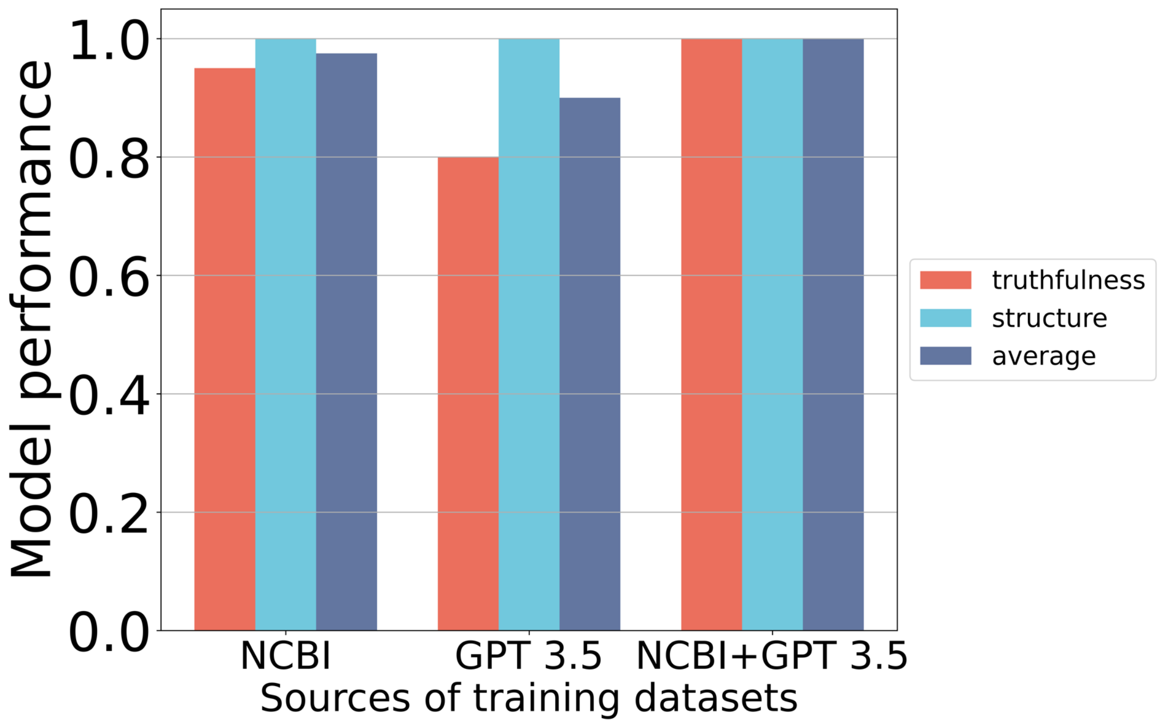
A: The official symbol for GLI1 is provided by HGNC, and its official full name is GLI family zinc finger 1. The primary source for this information is HGNC:HGNC:10997, and it is also related to Ensembl:ENSG00000104844 and MIM:600619; AllianceGenome:HGNC:10997. This gene is a protein coding gene and its RefSeq status is REVIEWED. It is found in the organism Homo sapiens and belongs to the lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo. It is also known as GLI; GLI-1; GLI-2; GLI-3; GLI-4; GLI-5; GLI-6; GLI-7; GLI-8; GLI-9; GLI-10; GLI-11; GLI-12; GLI-13; GLI-14; GLI-15; GLI-16; GLI-17; GLI-18; GLI-19; GLI-20; GLI-21; GLI-22; GLI-23; GLI-24; GLI-25; GLI-26; GLI-27; GLI-28; GLI-29; GLI-30; GLI-31; GLI-32; GLI-33; GLI-34; G

A: The official symbol for GLI1 is provided by HGNC, and its full name is GLI family zinc finger 1. The primary source for this information is HGNC, and it is also related to Ensembl and MIM. GLI1 is a protein coding gene in the organism Homo sapiens and belongs to the zinc finger family of transcription factors. It is involved in the Hedgehog signaling pathway, which plays a crucial role in embryonic development and adult tissue homeostasis. GLI1 is a transcriptional activator of Hedgehog target genes and is regulated by post-translational modifications. It is expressed ubiquitously in various tissues, including the esophagus, skin, and 25 other tissues. It also has orthologs in mice.

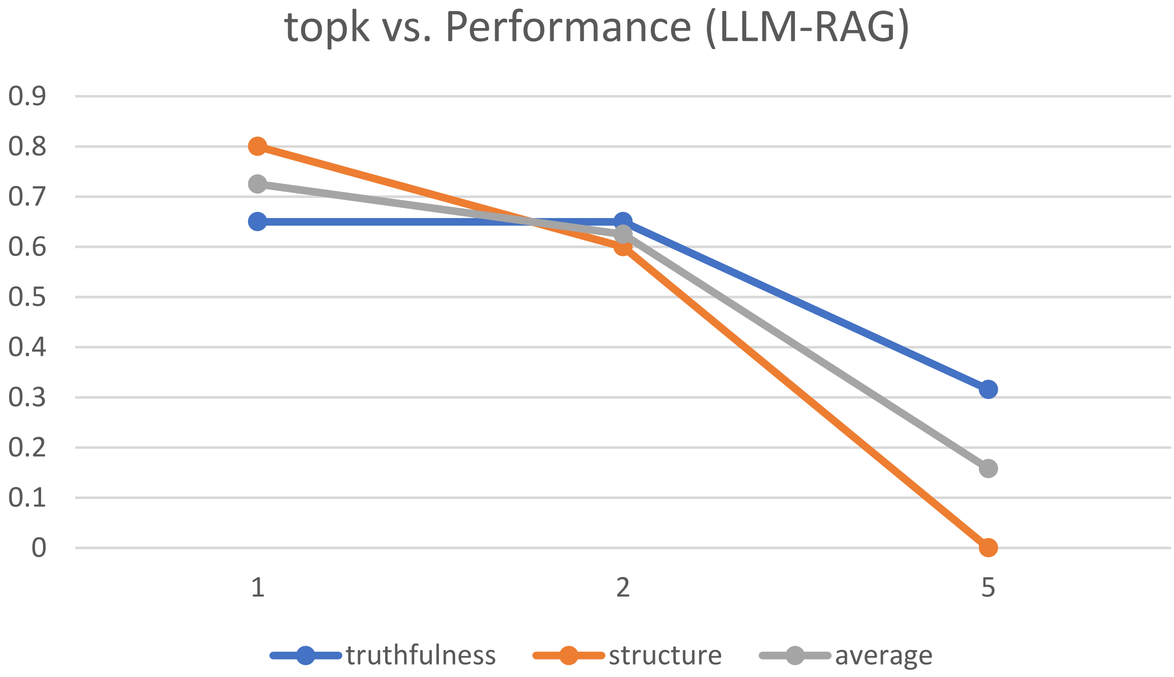
Extended Data Fig. 4: An example to illustrate the contribution of data augmentation. We compared the description of gene GLI1 from two models trained with different datasets but from the same instruction/prompt. We highlighted the redundant part in red.

- ● IG_C_gene
- ● IG_C_pseudogene
- ● IG_D_gene
- ● IG_J_gene
- ● IG_J_pseudogene
- ● IG_V_gene
- ● IG_V_pseudogene
- ● TEC
- ● TEC+misc_RNA
- ● TEC+protein_coding
- ● TEC+pseudogene
- ● TR_C_gene
- ● TR_D_gene
- ● TR_J_gene
- ● TR_J_pseudogene
- ● TR_V_gene
- ● TR_V_pseudogene
- ● Unknown
- ● antisense
- ● antisense+lncRNA
- ● antisense+miRNA
- ● antisense+misc_RNA
- ● antisense+processed_transcript
- ● antisense+protein_coding
- ● lincRNA
- ● lincRNA+lncRNA
- ● lincRNA+miRNA
- ● lincRNA+misc_RNA
- ● lincRNA+processed_transcript
- ● lincRNA+processed_transcript+pseudogene
- ● lincRNA+protein_coding
- ● lincRNA+snoRNA
- ● lincRNA+transcribed_processed_pseudogene
- ● lncRNA
- ● lncRNA+misc_RNA
- ● lncRNA+processed_transcript
- ● lncRNA+protein_coding
- ● miRNA
- ● miRNA+processed_transcript
- ● miRNA+pseudogene
- ● miRNA+sense_intronic
- ● misc_RNA
- ● misc_RNA+processed_transcript
- ● misc_RNA+snRNA
- ● polymorphic_pseudogene
- ● polymorphic_pseudogene+protein_coding
- ● polymorphic_pseudogene+pseudogene
- ● processed_pseudogene
- ● processed_pseudogene+protein_coding
- ● processed_pseudogene+pseudogene
- ● processed_transcript
- ● processed_transcript+protein_coding
- ● processed_transcript+pseudogene
- ● processed_transcript+scaRNA+snoRNA
- ● processed_transcript+sense_intronic
- ● processed_transcript+sense_overlapping
- ● processed_transcript+snoRNA
- ● protein_coding
- ● protein_coding+pseudogene
- ● protein_coding+unprocessed_pseudogene
- ● pseudogene
- ● pseudogene+snRNA
- ● pseudogene+transcribed_processed_pseudogene
- ● pseudogene+transcribed_unprocessed_pseudogene
- ● pseudogene+unprocessed_pseudogene
- ● rRNA
- ● rRNA+rRNA_pseudogene
- ● rRNA_pseudogene
- ● scaRNA+snoRNA
- ● sense_intronic
- ● sense_intronic+snoRNA
- ● sense_overlapping
- ● snRNA
- ● snoRNA
- ● transcribed_processed_pseudogene
- ● transcribed_unitary_pseudogene
- ● transcribed_unprocessed_pseudogene
- ● unitary_pseudogene
- ● unprocessed_pseudogene

Extended Data Fig. 5: Annotation of gene functions by different colors.



Extended Data Fig. 6: The results of data ablation for the training process.



Extended Data Fig. 7: Relation between top-k candidates and model performance.