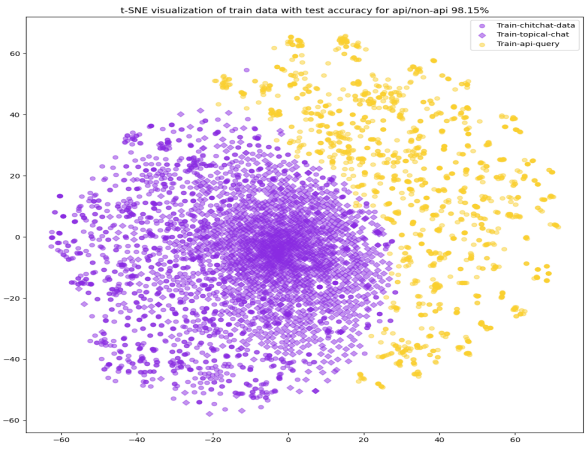
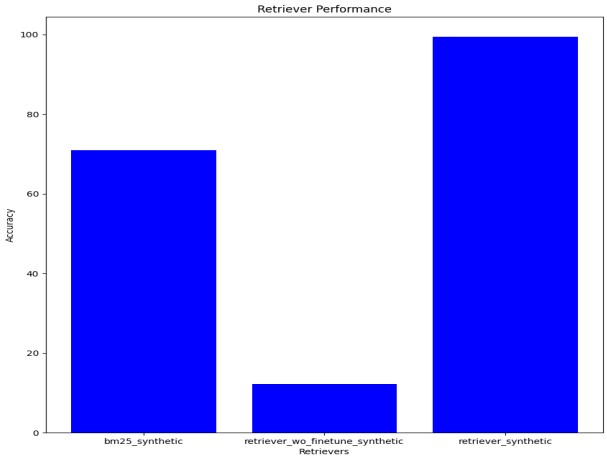
Performance Report



Chitchat classification model performance



Retriever training performance

Section 1: Chitchat Classification Analysis

We assessed the effectiveness of BioMANIA in each stage of the ChatBot, starting from the user initiating the conversation to the end of the conversation. We first investigated BioMANIA's performance in distinguishing between user queries related to API-encoded data analysis and casual chit-chat. We randomly sampled an equal number of chit-chat samples from two existing chit-chat datasets, Topical-Chat and Qnamaker. The number of samples for each cluster is 1588. As shown in figure 1., the TF-IDF embeddings of chit-chat samples and API-encoded instructions distinctly separate in the t-SNE projection. The notable separation allows a simple nearest centroid classifier to achieve a 98.15% classification accuracy on the test data.

Section 2: API Retriever Training Insights

After initiating an API call, we examined BioMANIA's performance in identifying the top-3 candidate APIs by utilizing a API Retriever fine-tuned from the BERT-BASE model. We assessed the retrieval performance by comparing it to two baseline methods: the untuned BERT-BASE model and BM25, a commonly used deterministic relevance measure for document comparison. The fine-tuned API Retriever attains an accuracy of 99.45% for synthetic instructions in shortlisting the target API within the top 3 candidates, while the untuned BERT-BASE model reaches 12.19%. In contrast, the BM25-based API Retriever, which lacks access to training data, achieves an accuracy of 70.91% for synthetic instructions, respectively. The sample size of the splitted training dataset and validation dataset are [1443, 361].

Section 3: GPT Model Performance Comparison

Lastly, we examined BioMANIA's performance in predicting the target API from the shortlist. We assessed the prediction performance from multiple settings, considering factors such as whether to use the LLM for prediction or not, whether to use the API Retriever or not, whether to formulate the problem as a prediction task or a sequence generation task (by generating the API from scratch), and whether to use GPT-3.5 or GPT-4. The performance of in-context API classification model based on GPT or synthetic instructions in predicting the target API from the top 3 candidates is presented as below. It's worth noting that a noticeable portion of misclassifications is caused by the presence of ambiguous APIs by design, which BioMANIA identifies during ChatBot generation.

task	model_name	accuracy	total	filtered_accuracy	filter_total	top_k	test_val
./gpt/5-shot-classify	gpt-3.5-turbo-16k	0.6813	182	0.7126	174	3	human annotate
./gpt/5-shot-classify	gpt-3.5-turbo-16k	0.7582	182	0.7797	177	3	synthetic
./gpt/5-shot-classify	gpt-3.5-turbo-16k	0.6593	182	0.6977	172	5	human annotate
./gpt/5-shot-classify	gpt-3.5-turbo-16k	0.7527	182	0.7919	173	5	synthetic
./gpt/5-shot-classify	gpt-3.5-turbo-16k	0.6703	182	0.7052	173	10	human annotate
./gpt/5-shot-classify	gpt-3.5-turbo-16k	0.7582	182	0.8023	172	10	synthetic
./gpt/5-shot-classify	gpt-4	0.7802	182	0.8114	175	3	human annotate
./gpt/5-shot-classify	gpt-4	0.8516	182	0.8908	174	3	synthetic
./gpt/5-shot-classify	gpt-4	0.7912	182	0.8324	173	5	human annotate
./gpt/5-shot-classify	gpt-4	0.8571	182	0.8966	174	5	synthetic
./gpt/5-shot-classify	gpt-4	0.8077	182	0.8497	173	10	human annotate
./gpt/5-shot-classify	gpt-4	0.8516	182	0.8857	175	10	synthetic