Benchmark Studio

LLM MODEL PERFORMANCE COMPARISON ANALYSIS

Report Milestone 3

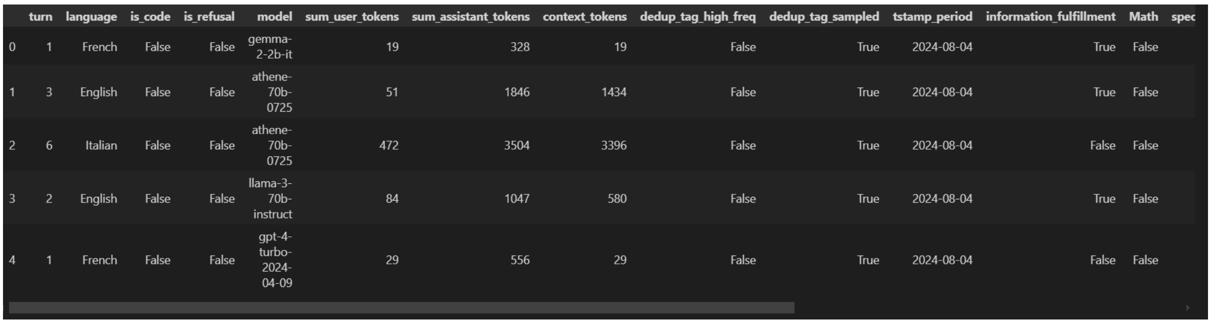
Report by:

Akshara Sri Lakshmipathy - akla8196@colorado.edu

Harish Nandhan Shanmugam - hash1366@colorado.edu

Shivaraj Senthil Rajan - shse1502@colorado.edu

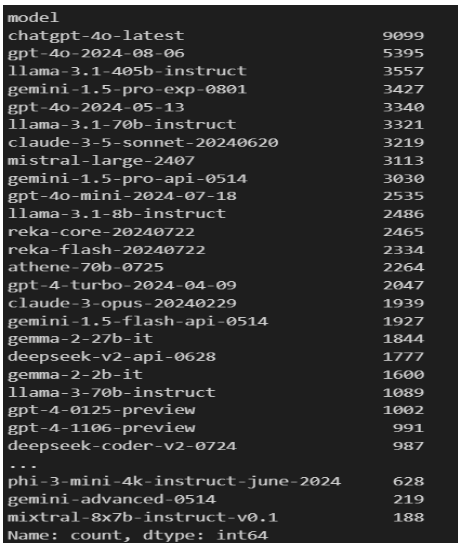
In this we imported a CSV file (battles_data_cleaned.csv) into a DataFrame and then displayed the first few rows.



	turn	language	is_code	is_refusal	model	sum_user_tokens	sum_assistant_tokens	context_tokens	dedup_tag_high_freq	dedup_tag_sampled	tstamp_period	information_fulfillment	Math	spec
0	1	French	False	False	gemma-2-2b-it	19	328	19	False	True	2024-08-04	True	False	
1	3	English	False	False	athene-70b-0725	51	1846	1434	False	True	2024-08-04	True	False	
2	6	Italian	False	False	athene-70b-0725	472	3504	3396	False	True	2024-08-04	False	False	
3	2	English	False	False	llama-3-70b-instruct	84	1047	580	False	True	2024-08-04	True	False	
4	1	French	False	False	gpt-4-turbo-2024-04-09	29	556	29	False	True	2024-08-04	False	False	

HANDLING CLASS IMBALANCE

It calculates the frequency of models in the dataset, filter them and keep only with at least 1000 occurrences and create a new filtered DataFrame containing only these models.

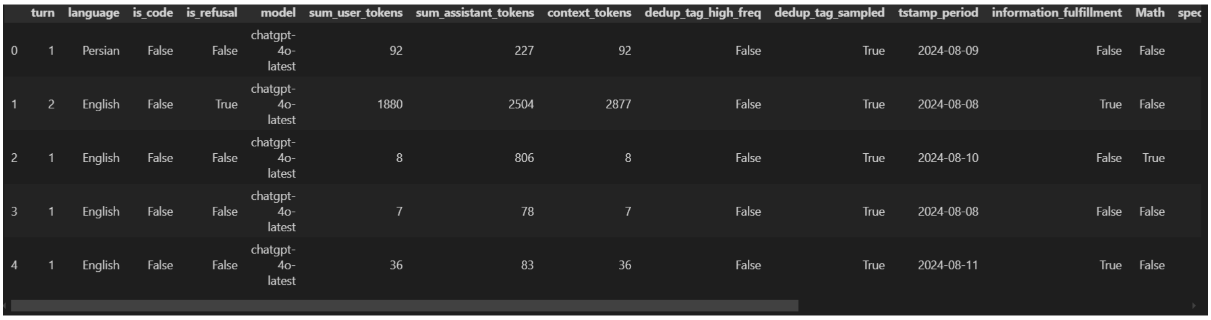


```
model
chatgpt-4o-latest                  9099
gpt-4o-2024-08-06                  5395
llama-3.1-405b-instruct            3557
gemini-1.5-pro-exp-0801            3427
gpt-4o-2024-05-13                  3340
llama-3.1-70b-instruct             3321
claude-3-5-sonnet-20240620         3219
mistral-large-2407                 3113
gemini-1.5-pro-api-0514            3030
gpt-4o-mini-2024-07-18             2535
llama-3.1-8b-instruct              2486
reka-core-20240722                 2465
reka-flash-20240722                2334
athene-70b-0725                    2264
gpt-4-turbo-2024-04-09             2047
claude-3-opus-20240229             1939
gemini-1.5-flash-api-0514          1927
gemma-2-27b-it                     1844
deepseek-v2-api-0628               1777
gemma-2-2b-it                      1600
llama-3-70b-instruct               1089
gpt-4-0125-preview                 1002
gpt-4-1106-preview                  991
deepseek-coder-v2-0724              987
...
phi-3-mini-4k-instruct-june-2024    628
gemini-advanced-0514                219
mixtral-8x7b-instruct-v0.1          188
Name: count, dtype: int64
```

It calculates the value counts of the 'model' column in the filtered data and prints the number of models.

```
22
```

To balance the dataset we can control the number of samples for each unique model. It filters models that have at least 1000 samples, and then resamples each model to ensure an equal representation of samples either by undersampling or oversampling. The code first saves the balanced dataset (balanced_data) to a CSV file and printed the first few rows of the data are displayed to confirm the dataset's successful import in resampled_data.head()



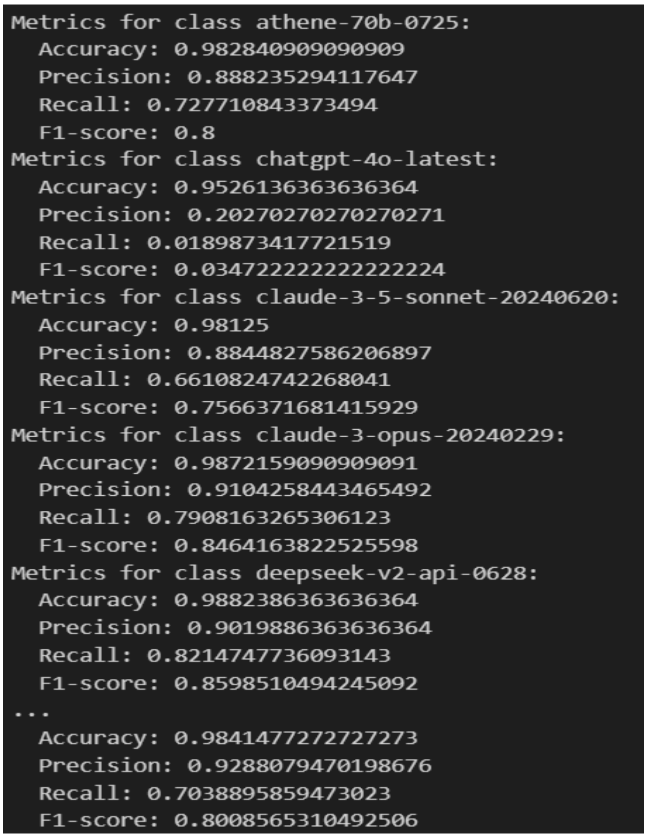
	turn	language	is_code	is_refusal	model	sum_user_tokens	sum_assistant_tokens	context_tokens	dedup_tag_high_freq	dedup_tag_sampled	tstamp_period	information_fulfillment	Math	spec
0	1	Persian	False	False	chatgpt-4o-latest	92	227	92	False	True	2024-08-09	False	False	
1	2	English	False	True	chatgpt-4o-latest	1880	2504	2877	False	True	2024-08-08	True	False	
2	1	English	False	False	chatgpt-4o-latest	8	806	8	False	True	2024-08-10	False	True	
3	1	English	False	False	chatgpt-4o-latest	7	78	7	False	True	2024-08-08	False	False	
4	1	English	False	False	chatgpt-4o-latest	36	83	36	False	True	2024-08-11	True	False	

After this the code initializes and applies a LabelEncoder to the 'language' column from categorical text to numeric values. A LabelBinarizer is used to binarize the target labels, which helps transform the target variable into a format compatible with the multiclass classification. And then code splits the dataset into training and testing sets (80% training, 20% testing).

MODEL IMPLEMENTATION

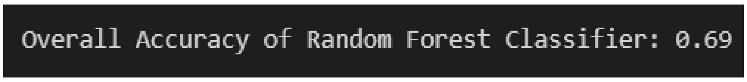
1. RANDOM FOREST

A **Random Forest** binary classifier is trained, resulting in multiple classifiers. After training, the classifiers are used to collect the probability scores of each instance in the test set for each class and the class with the highest probability is then selected for each test instance where it calculates and displays metrics—accuracy, precision, recall, and F1-score—for each of the binary classifiers on the test set



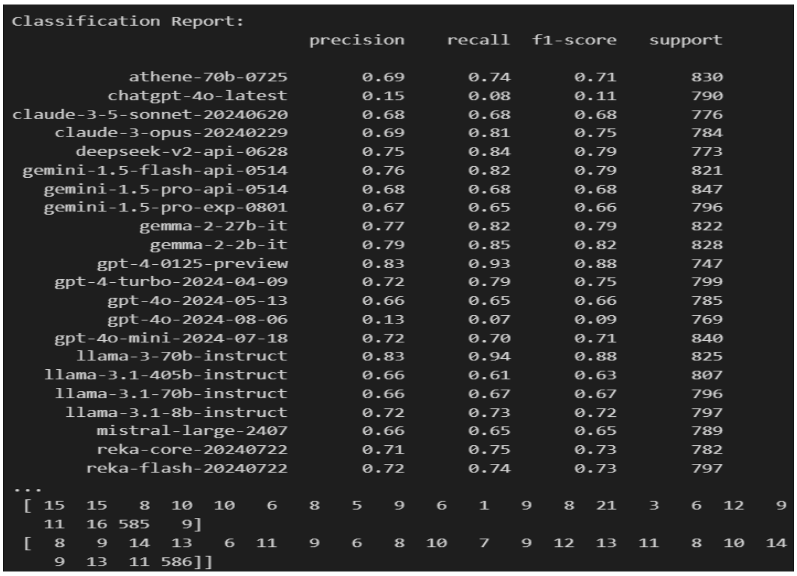
```
Metrics for class athene-70b-0725:
  Accuracy: 0.9828409090909
  Precision: 0.888235294117647
  Recall: 0.727710843373494
  F1-score: 0.8
Metrics for class chatgpt-4o-latest:
  Accuracy: 0.9526136363636364
  Precision: 0.20270270270270271
  Recall: 0.0189873417721519
  F1-score: 0.034722222222222224
Metrics for class claude-3-5-sonnet-20240620:
  Accuracy: 0.98125
  Precision: 0.8844827586206897
  Recall: 0.6610824742268041
  F1-score: 0.7566371681415929
Metrics for class claude-3-opus-20240229:
  Accuracy: 0.9872159090909091
  Precision: 0.9104258443465492
  Recall: 0.7908163265306123
  F1-score: 0.8464163822525598
Metrics for class deepseek-v2-api-0628:
  Accuracy: 0.9882386363636364
  Precision: 0.9019886363636364
  Recall: 0.8214747736093143
  F1-score: 0.8598510494245092
...
  Accuracy: 0.9841477272727273
  Precision: 0.9288079470198676
  Recall: 0.7038895859473023
  F1-score: 0.8008565310492506
```

Printed the overall accuracy of the Random Forest classifier across all classes



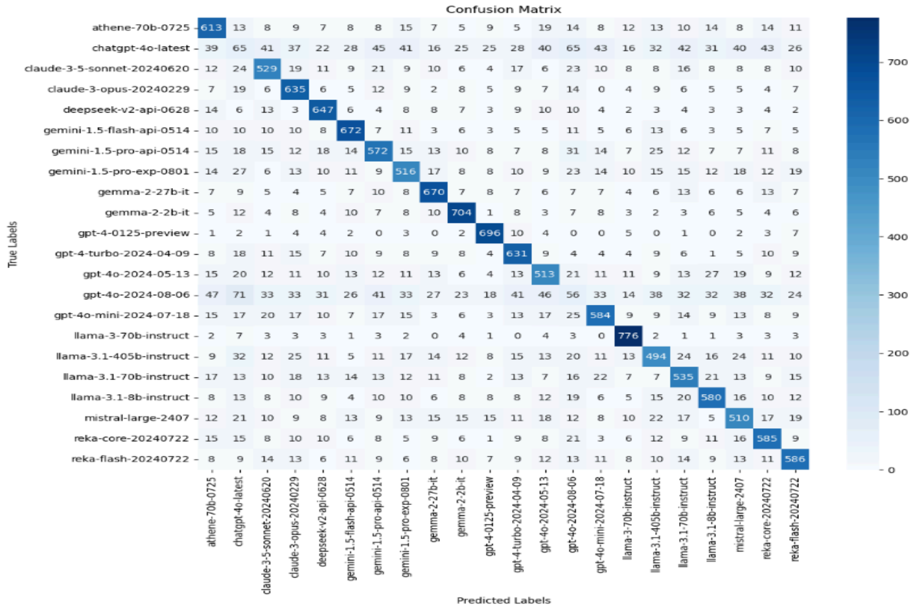
```
Overall Accuracy of Random Forest Classifier: 0.69
```

gpt-4-0125-preview and llama-3-70b-instruct showing high precision, recall, and F1-scores, while models like chatgpt-4o-latest underperformed. Most of the models have moderate performance, with scores ranging from 0.60 to 0.80.



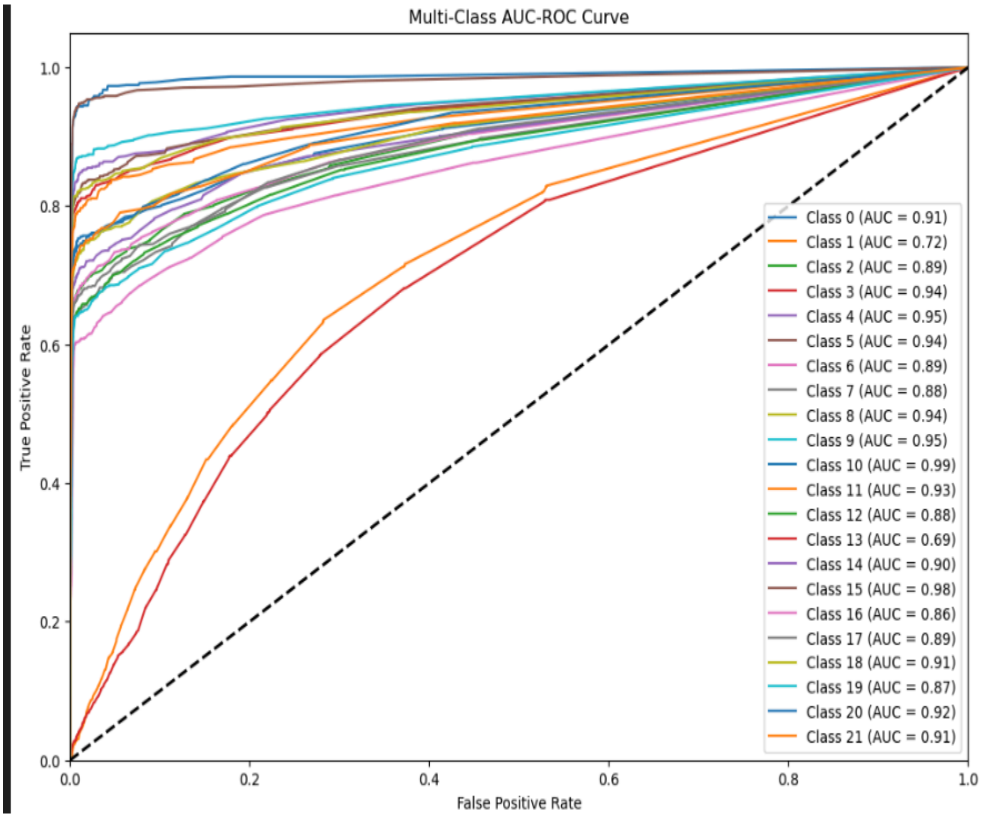
```
Classification Report:
                            precision    recall  f1-score   support

           athene-70b-0725       0.69      0.74      0.71       830
         chatgpt-4o-latest       0.15      0.08      0.11       790
claude-3-5-sonnet-20240620       0.68      0.68      0.68       776
    claude-3-opus-20240229       0.69      0.81      0.75       784
      deepseek-v2-api-0628       0.75      0.84      0.79       773
 gemini-1.5-flash-api-0514       0.76      0.82      0.79       821
   gemini-1.5-pro-api-0514       0.68      0.68      0.68       847
   gemini-1.5-pro-exp-0801       0.67      0.65      0.66       796
            gemma-2-27b-it       0.77      0.82      0.79       822
             gemma-2-2b-it       0.79      0.85      0.82       828
        gpt-4-0125-preview       0.83      0.93      0.88       747
    gpt-4-turbo-2024-04-09       0.72      0.79      0.75       799
         gpt-4o-2024-05-13       0.66      0.65      0.66       785
         gpt-4o-2024-08-06       0.13      0.07      0.09       769
    gpt-4o-mini-2024-07-18       0.72      0.70      0.71       840
      llama-3-70b-instruct       0.83      0.94      0.88       825
   llama-3.1-405b-instruct       0.66      0.61      0.63       807
    llama-3.1-70b-instruct       0.66      0.67      0.67       796
     llama-3.1-8b-instruct       0.72      0.73      0.72       797
        mistral-large-2407       0.66      0.65      0.65       789
         reka-core-20240722       0.71      0.75      0.73       782
        reka-flash-20240722       0.72      0.74      0.73       797
...
 [ 15  15   8  10  10   6   8   5   9   6   1   9   8  21   3   6  12   9
   11  16 585   9]
 [  8   9  14  13   6  11   9   6   8  10   7   9  12  13  11   8  10  14
    9  13  11 586]]
```

The matrix provides a view of how well each model is performing to the others in predicting different classes.



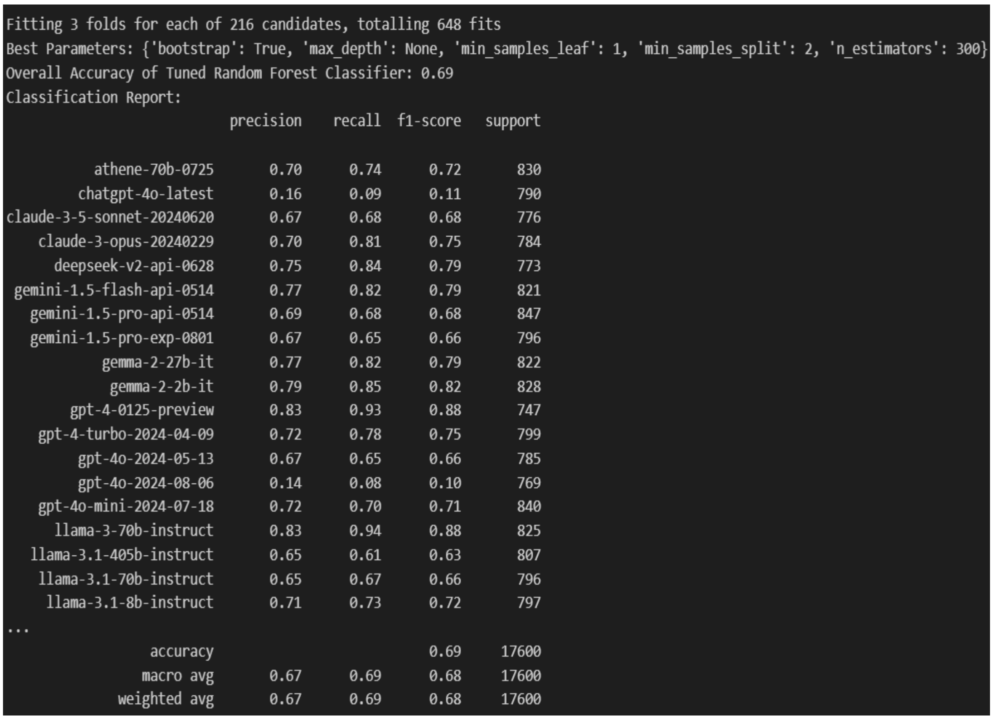
	athene-70b-0725	chatgpt-4o-latest	claude-3-5-sonnet-20240620	claude-3-opus-20240229	deepseek-v2-api-0628	gemini-1.5-flash-api-0514	gemini-1.5-pro-api-0514	gemini-1.5-pro-exp-0801	gemma-2-27b-it	gemma-2-2b-it	gpt-4-0125-preview	gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	gpt-4o-2024-08-06	gpt-4o-mini-2024-07-18	llama-3-70b-instruct	llama-3.1-405b-instruct	llama-3.1-70b-instruct	llama-3.1-8b-instruct	mistral-large-2407	reka-core-20240722	reka-flash-20240722
athene-70b-0725	613	13	8	9	7	8	8	15	7	5	9	5	19	14	8	12	13	10	14	8	14	11
chatgpt-4o-latest	39	65	41	37	22	28	45	41	16	25	25	28	40	65	43	16	32	42	31	40	43	26
claude-3-5-sonnet-20240620	12	24	529	19	11	9	21	9	10	6	4	17	6	23	10	8	8	16	8	8	8	10
claude-3-opus-20240229	7	19	6	635	6	5	12	9	2	8	5	9	7	14	0	4	9	6	5	5	4	7
deepseek-v2-api-0628	14	6	13	3	647	6	4	8	8	7	3	9	10	10	4	2	3	4	3	3	4	2
gemini-1.5-flash-api-0514	10	10	10	10	8	672	7	11	3	6	3	5	5	11	5	6	13	6	3	5	7	5
gemini-1.5-pro-api-0514	15	18	15	12	18	14	572	15	13	10	8	7	8	31	14	7	25	12	7	7	11	8
gemini-1.5-pro-exp-0801	14	27	6	13	10	11	9	516	17	8	8	10	9	23	14	10	15	15	12	18	12	19
gemma-2-27b-it	7	9	5	4	5	7	10	8	670	7	8	7	6	7	7	4	6	13	6	6	13	7
gemma-2-2b-it	5	12	4	8	4	10	7	8	10	704	1	8	3	7	8	3	2	3	6	5	4	6
gpt-4-0125-preview	1	2	1	4	4	2	0	3	0	2	696	10	4	0	0	5	0	1	0	2	3	7
gpt-4-turbo-2024-04-09	8	18	11	15	7	10	9	8	9	8	4	631	9	4	4	4	9	6	1	5	10	9
gpt-4o-2024-05-13	15	20	12	11	10	13	12	11	13	6	4	13	513	21	11	11	9	13	27	19	9	12
gpt-4o-2024-08-06	47	71	33	33	31	26	41	33	27	23	18	41	46	56	33	14	38	32	32	38	32	24
gpt-4o-mini-2024-07-18	15	17	20	17	10	7	17	15	3	6	3	13	17	25	584	9	9	14	9	13	8	9
llama-3-70b-instruct	2	7	3	3	3	1	3	2	0	4	1	0	4	3	0	776	2	1	1	3	3	3
llama-3.1-405b-instruct	9	32	12	25	11	5	11	17	14	12	8	15	13	20	11	13	494	24	16	24	11	10
llama-3.1-70b-instruct	17	13	10	18	13	14	13	12	11	8	2	13	7	16	22	7	7	535	21	13	9	15
llama-3.1-8b-instruct	8	13	8	10	9	4	10	10	6	8	8	8	12	19	6	5	15	20	580	16	10	12
mistral-large-2407	12	21	10	9	8	13	9	13	15	15	15	11	18	12	8	10	22	17	5	510	17	19
reka-core-20240722	15	15	8	10	10	6	8	5	9	6	1	9	8	21	3	6	12	9	11	16	585	9
reka-flash-20240722	8	9	14	13	6	11	9	6	8	10	7	9	12	13	11	8	10	14	9	13	11	586

The area under each curve (AUC) is an indicator of how well the model can distinguish between classes, with values closer to 1 indicating better performance.



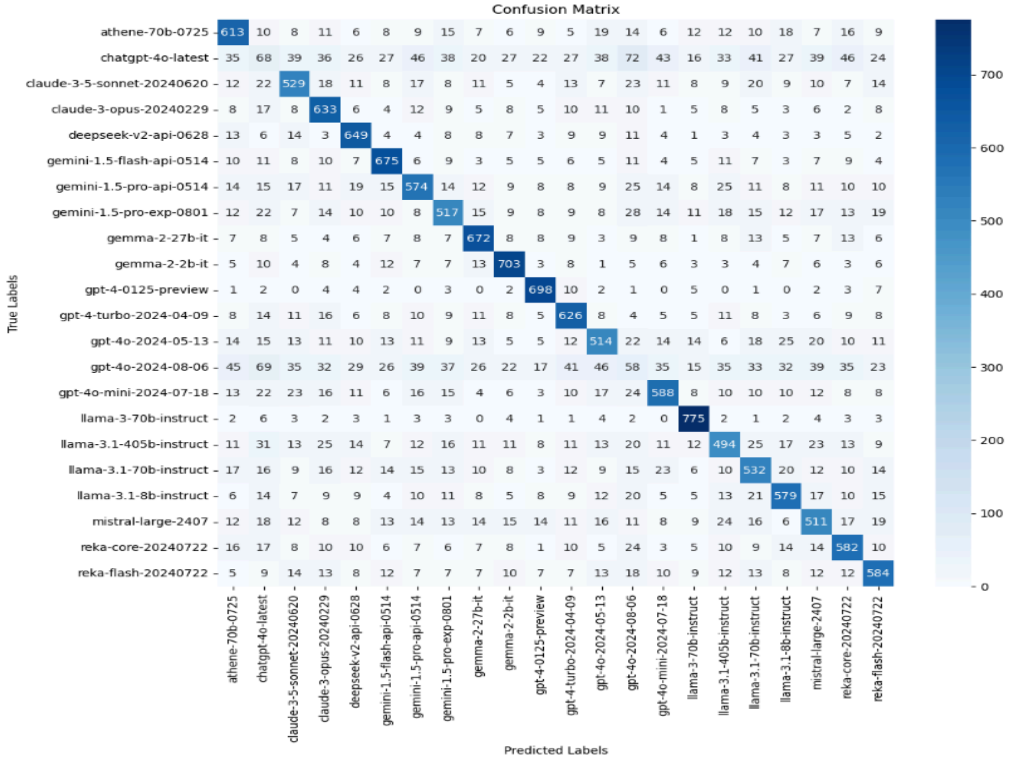
The Random Forest Classifier undergoes hyperparameter tuning using GridSearchCV to find the optimal parameters, which are used to train a One-vs-Rest classification model. After training, predictions are made by selecting the class with the highest probability for each instance. The evaluation involves calculating overall accuracy, classification report for precision, recall, and F1-score, and visualizing the confusion matrix. Additionally, AUC-ROC curves are plotted.

The Random Forest model achieved an overall accuracy of 0.69, with gpt-4-0125-preview and llama-3-70b-instruct performing F1-scores above 0.88. Models like chatgpt-4o-latest had low F1-scores around 0.11.



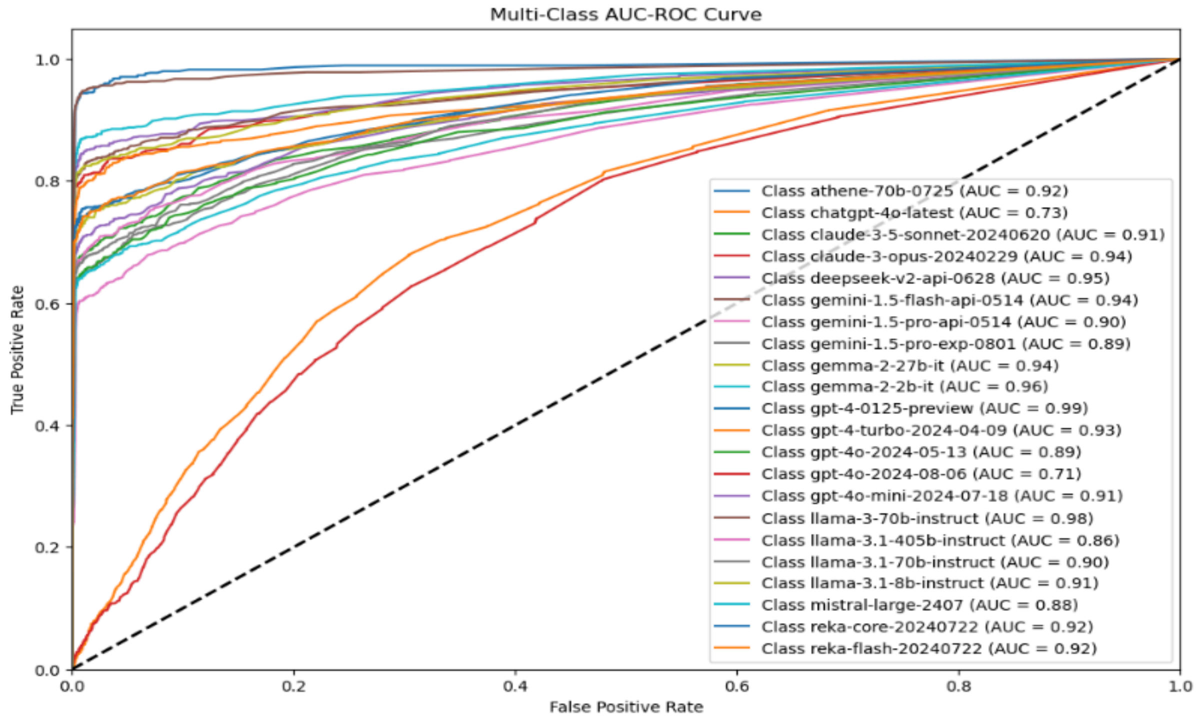
```
Fitting 3 folds for each of 216 candidates, totalling 648 fits
Best Parameters: {'bootstrap': True, 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}
Overall Accuracy of Tuned Random Forest Classifier: 0.69
Classification Report:
                            precision    recall  f1-score   support

           athene-70b-0725       0.70      0.74      0.72       830
         chatgpt-4o-latest       0.16      0.09      0.11       790
claude-3-5-sonnet-20240620       0.67      0.68      0.68       776
    claude-3-opus-20240229       0.70      0.81      0.75       784
        deepseek-v2-api-0628       0.75      0.84      0.79       773
  gemini-1.5-flash-api-0514       0.77      0.82      0.79       821
    gemini-1.5-pro-api-0514       0.69      0.68      0.68       847
    gemini-1.5-pro-exp-0801       0.67      0.65      0.66       796
             gemma-2-27b-it       0.77      0.82      0.79       822
              gemma-2-2b-it       0.79      0.85      0.82       828
         gpt-4-0125-preview       0.83      0.93      0.88       747
     gpt-4-turbo-2024-04-09       0.72      0.78      0.75       799
          gpt-4o-2024-05-13       0.67      0.65      0.66       785
          gpt-4o-2024-08-06       0.14      0.08      0.10       769
     gpt-4o-mini-2024-07-18       0.72      0.70      0.71       840
        llama-3-70b-instruct       0.83      0.94      0.88       825
     llama-3.1-405b-instruct       0.65      0.61      0.63       807
      llama-3.1-70b-instruct       0.65      0.67      0.66       796
       llama-3.1-8b-instruct       0.71      0.73      0.72       797
...
                  accuracy                           0.69     17600
                 macro avg       0.67      0.69      0.68     17600
              weighted avg       0.67      0.69      0.68     17600
```

High values on the diagonal elements indicate a high number of correct predictions, while off-diagonal elements show the misclassifications. For example, model 'deepseek-v2-api-0628' has 649 correct predictions for one of its classes, while it has misclassified other classes a few times as shown by the non-diagonal elements.



	athene-70b-0725	chatgpt-4o-latest	claude-3-5-sonnet-20240620	claude-3-opus-20240229	deepseek-v2-api-0628	gemini-1.5-flash-api-0514	gemini-1.5-pro-api-0514	gemini-1.5-pro-exp-0801	gemma-2-27b-it	gemma-2-2b-it	gpt-4-0125-preview	gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	gpt-4o-2024-08-06	gpt-4o-mini-2024-07-18	llama-3-70b-instruct	llama-3.1-405b-instruct	llama-3.1-70b-instruct	llama-3.1-8b-instruct	mistral-large-2407	reka-core-20240722	reka-flash-20240722
athene-70b-0725	613	10	8	11	6	8	9	15	7	6	9	5	19	14	6	12	12	10	18	7	16	9
chatgpt-4o-latest	35	68	39	36	26	27	46	38	20	27	22	27	38	72	43	16	33	41	27	39	46	24
claude-3-5-sonnet-20240620	12	22	529	18	11	8	17	8	11	5	4	13	7	23	11	8	9	20	9	10	7	14
claude-3-opus-20240229	8	17	8	633	6	4	12	9	5	8	5	10	11	10	1	5	8	5	3	6	2	8
deepseek-v2-api-0628	13	6	14	3	649	4	4	8	8	7	3	9	9	11	4	1	3	4	3	3	5	2
gemini-1.5-flash-api-0514	10	11	8	10	7	675	6	9	3	5	5	6	5	11	4	5	11	7	3	7	9	4
gemini-1.5-pro-api-0514	14	15	17	11	19	15	574	14	12	9	8	8	9	25	14	8	25	11	8	11	10	10
gemini-1.5-pro-exp-0801	12	22	7	14	10	10	8	517	15	9	8	9	8	28	14	11	18	15	12	17	13	19
gemma-2-27b-it	7	8	5	4	6	7	8	7	672	8	8	9	3	9	8	1	8	13	5	7	13	6
gemma-2-2b-it	5	10	4	8	4	12	7	7	13	703	3	8	1	5	6	3	3	4	7	6	3	6
gpt-4-0125-preview	1	2	0	4	4	2	0	3	0	2	698	10	2	1	0	5	0	1	0	2	3	7
gpt-4-turbo-2024-04-09	8	14	11	16	6	8	10	9	11	8	5	626	8	4	5	5	11	8	3	6	9	8
gpt-4o-2024-05-13	14	15	13	11	10	13	11	9	13	5	5	12	514	22	14	14	6	18	25	20	10	11
gpt-4o-2024-08-06	45	69	35	32	29	26	39	37	26	22	17	41	46	58	35	15	35	33	32	39	35	23
gpt-4o-mini-2024-07-18	13	22	23	16	11	6	16	15	4	6	3	10	17	24	588	8	10	10	10	12	8	8
llama-3-70b-instruct	2	6	3	2	3	1	3	3	0	4	1	1	4	2	0	775	2	1	2	4	3	3
llama-3.1-405b-instruct	11	31	13	25	14	7	12	16	11	11	8	11	13	20	11	12	494	25	17	23	13	9
llama-3.1-70b-instruct	17	16	9	16	12	14	15	13	10	8	3	12	9	15	23	6	10	532	20	12	10	14
llama-3.1-8b-instruct	6	14	7	9	9	4	10	11	8	5	8	9	12	20	5	5	13	21	579	17	10	15
mistral-large-2407	12	18	12	8	8	13	14	13	14	15	14	11	16	11	8	9	24	16	6	511	17	19
reka-core-20240722	16	17	8	10	10	6	7	6	7	8	1	10	5	24	3	5	10	9	14	14	582	10
reka-flash-20240722	5	9	14	13	8	12	7	7	7	10	7	7	13	18	10	9	12	13	8	12	12	584

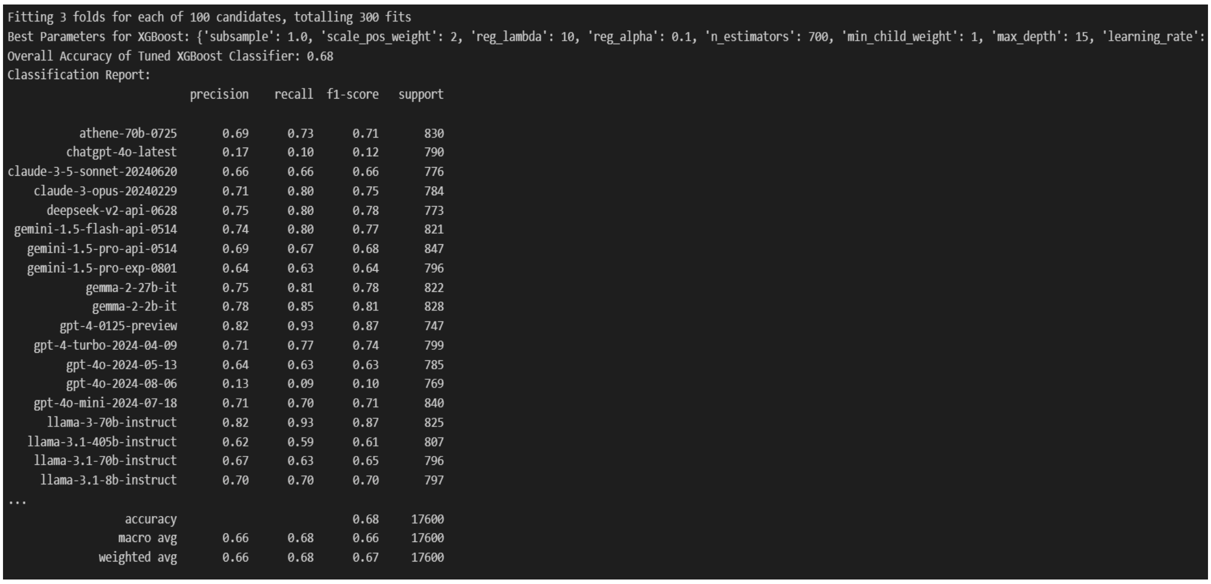
Higher AUC values, closer to 1, indicate better model performance.



2. XG BOOST

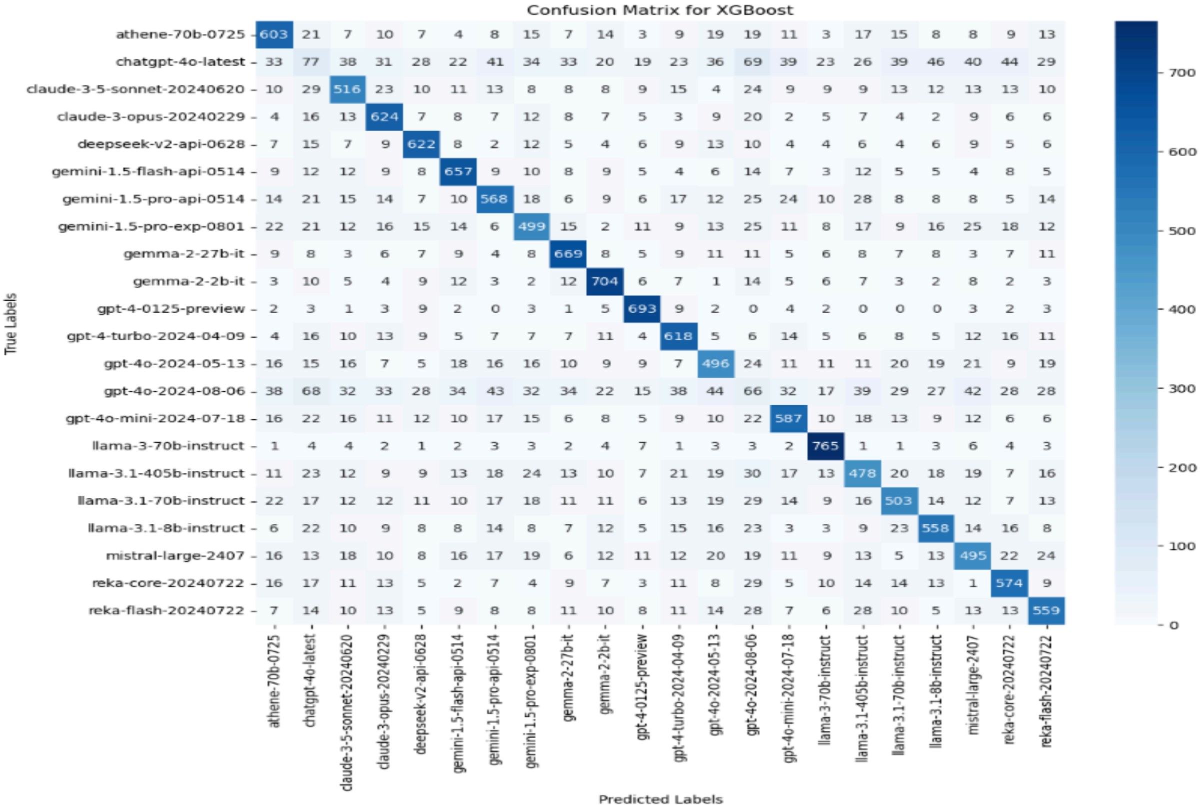
The **XGBoost** classifier achieved an overall accuracy of 0.68.

We have used XGBoost for a multi-class classification using a One-vs-Rest approach and involves the following steps: Data Loading and Preprocessing, Hyperparameter Tuning using RandomizedSearchCV to tune hyperparameters and selecting the best parameters. Then it trains separate One-vs-Rest XGBoost classifiers for each class. Making predictions which involves calculating overall accuracy, classification report for precision, recall, and F1-score, and visualizing the confusion matrix. Additionally, AUC-ROC curves are plotted.



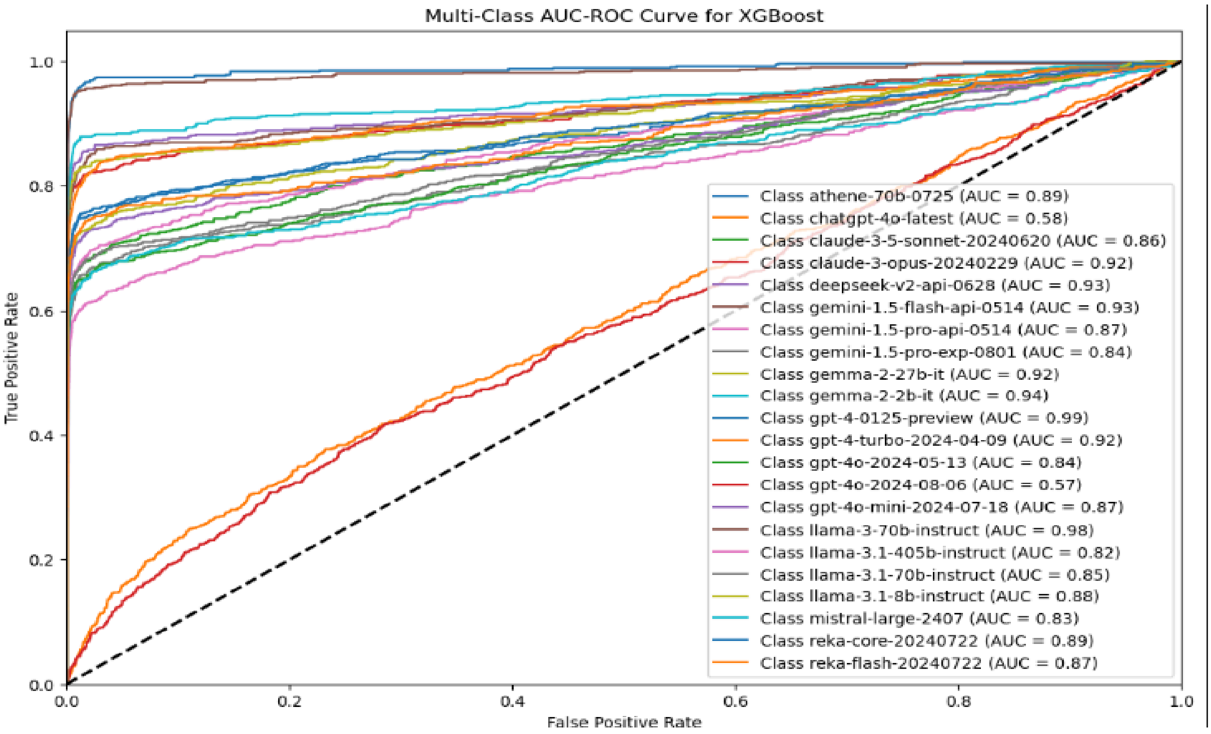
```
Fitting 3 folds for each of 100 candidates, totalling 300 fits
Best Parameters for XGBoost: {'subsample': 1.0, 'scale_pos_weight': 2, 'reg_lambda': 10, 'reg_alpha': 0.1, 'n_estimators': 700, 'min_child_weight': 1, 'max_depth': 15, 'learning_rate':
Overall Accuracy of Tuned XGBoost Classifier: 0.68
Classification Report:
                            precision    recall  f1-score   support

           athene-70b-0725       0.69      0.73      0.71       830
         chatgpt-4o-latest       0.17      0.10      0.12       790
claude-3-5-sonnet-20240620       0.66      0.66      0.66       776
    claude-3-opus-20240229       0.71      0.80      0.75       784
       deepseek-v2-api-0628       0.75      0.80      0.78       773
 gemini-1.5-flash-api-0514       0.74      0.80      0.77       821
   gemini-1.5-pro-api-0514       0.69      0.67      0.68       847
   gemini-1.5-pro-exp-0801       0.64      0.63      0.64       796
            gemma-2-27b-it       0.75      0.81      0.78       822
             gemma-2-2b-it       0.78      0.85      0.81       828
        gpt-4-0125-preview       0.82      0.93      0.87       747
    gpt-4-turbo-2024-04-09       0.71      0.77      0.74       799
         gpt-4o-2024-05-13       0.64      0.63      0.63       785
         gpt-4o-2024-08-06       0.13      0.09      0.10       769
    gpt-4o-mini-2024-07-18       0.71      0.70      0.71       840
      llama-3-70b-instruct       0.82      0.93      0.87       825
   llama-3.1-405b-instruct       0.62      0.59      0.61       807
    llama-3.1-70b-instruct       0.67      0.63      0.65       796
     llama-3.1-8b-instruct       0.70      0.70      0.70       797
...
                  accuracy                           0.68     17600
                 macro avg       0.66      0.68      0.66     17600
              weighted avg       0.66      0.68      0.67     17600
```

The diagonal values, which show the highest numbers (e.g., 603 for the first class), indicate correct predictions.



True Labels \ Predicted Labels	athene-70b-0725	chatgpt-4o-latest	claude-3-5-sonnet-20240620	claude-3-opus-20240229	deepseek-v2-api-0628	gemini-1.5-flash-api-0514	gemini-1.5-pro-api-0514	gemini-1.5-pro-exp-0801	gemma-2-27b-it	gemma-2-2b-it	gpt-4-0125-preview	gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	gpt-4o-2024-08-06	gpt-4o-mini-2024-07-18	llama-3-70b-instruct	llama-3.1-405b-instruct	llama-3.1-70b-instruct	llama-3.1-8b-instruct	mistral-large-2407	reka-core-20240722	reka-flash-20240722
athene-70b-0725	603	21	7	10	7	4	8	15	7	14	3	9	19	19	11	3	17	15	8	8	9	13
chatgpt-4o-latest	33	77	38	31	28	22	41	34	33	20	19	23	36	69	39	23	26	39	46	40	44	29
claude-3-5-sonnet-20240620	10	29	516	23	10	11	13	8	8	8	9	15	4	24	9	9	9	13	12	13	13	10
claude-3-opus-20240229	4	16	13	624	7	8	7	12	8	7	5	3	9	20	2	5	7	4	2	9	6	6
deepseek-v2-api-0628	7	15	7	9	622	8	2	12	5	4	6	9	13	10	4	4	6	4	6	9	5	6
gemini-1.5-flash-api-0514	9	12	12	9	8	657	9	10	8	9	5	4	6	14	7	3	12	5	5	4	8	5
gemini-1.5-pro-api-0514	14	21	15	14	7	10	568	18	6	9	6	17	12	25	24	10	28	8	8	8	5	14
gemini-1.5-pro-exp-0801	22	21	12	16	15	14	6	499	15	2	11	9	13	25	11	8	17	9	16	25	18	12
gemma-2-27b-it	9	8	3	6	7	9	4	8	669	8	5	9	11	11	5	6	8	7	8	3	7	11
gemma-2-2b-it	3	10	5	4	9	12	3	2	12	704	6	7	1	14	5	6	7	3	2	8	2	3
gpt-4-0125-preview	2	3	1	3	9	2	0	3	1	5	693	9	2	0	4	2	0	0	0	3	2	3
gpt-4-turbo-2024-04-09	4	16	10	13	9	5	7	7	7	11	4	618	5	6	14	5	6	8	5	12	16	11
gpt-4o-2024-05-13	16	15	16	7	5	18	16	16	10	9	9	7	496	24	11	11	11	20	19	21	9	19
gpt-4o-2024-08-06	38	68	32	33	28	34	43	32	34	22	15	38	44	66	32	17	39	29	27	42	28	28
gpt-4o-mini-2024-07-18	16	22	16	11	12	10	17	15	6	8	5	9	10	22	587	10	18	13	9	12	6	6
llama-3-70b-instruct	1	4	4	2	1	2	3	3	2	4	7	1	3	3	2	765	1	1	3	6	4	3
llama-3.1-405b-instruct	11	23	12	9	9	13	18	24	13	10	7	21	19	30	17	13	478	20	18	19	7	16
llama-3.1-70b-instruct	22	17	12	12	11	10	17	18	11	11	6	13	19	29	14	9	16	503	14	12	7	13
llama-3.1-8b-instruct	6	22	10	9	8	8	14	8	7	12	5	15	16	23	3	3	9	23	558	14	16	8
mistral-large-2407	16	13	18	10	8	16	17	19	6	12	11	12	20	19	11	9	13	5	13	495	22	24
reka-core-20240722	16	17	11	13	5	2	7	4	9	7	3	11	8	29	5	10	14	14	13	1	574	9
reka-flash-20240722	7	14	10	13	5	9	8	8	11	10	8	11	14	28	7	6	28	10	5	13	13	559

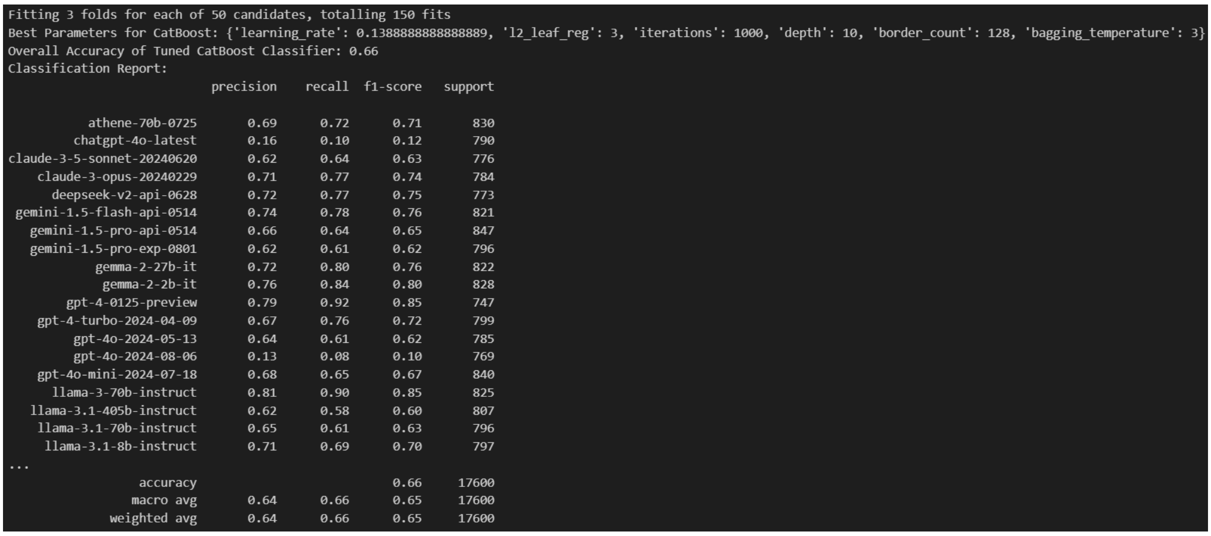
The curves closer to the top-left corner of the plot indicate better performance, with AUC values closer to 1.0



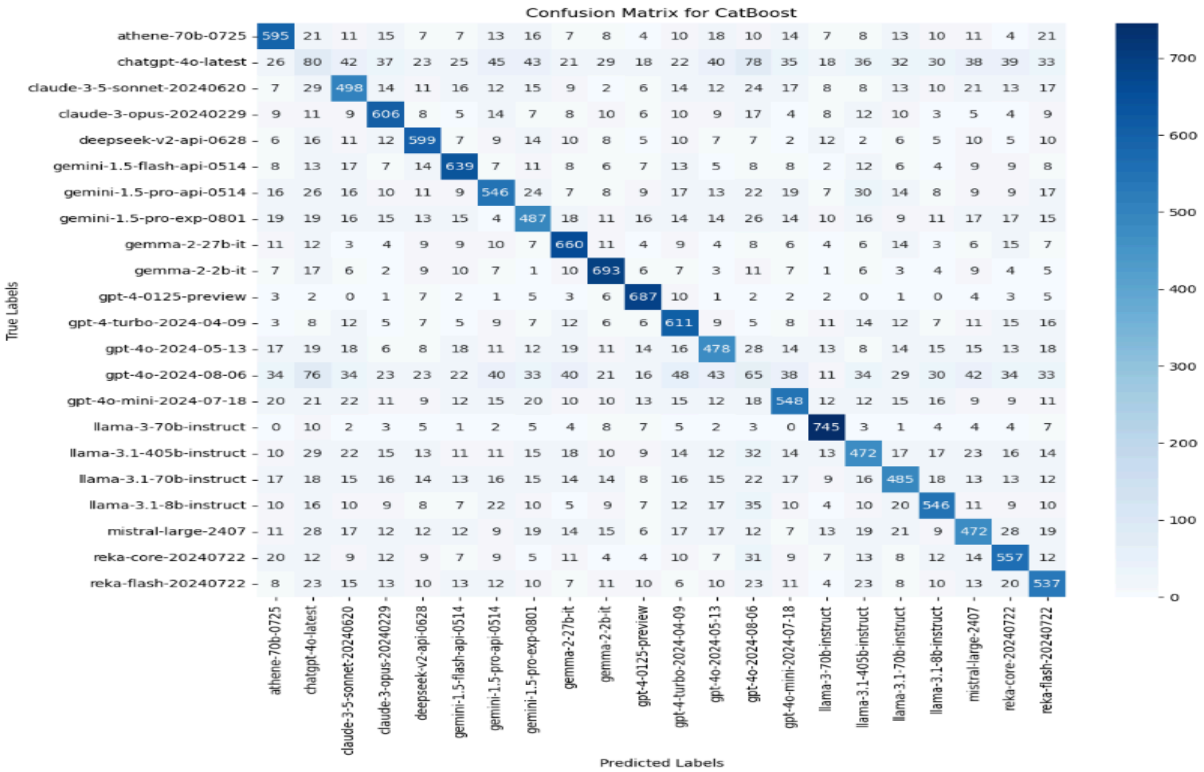
3. CATBOOST

The **CatBoost** classifier achieved an overall accuracy of 0.66.

Here we have used CatBoost and involve the following steps: Data Loading and Preprocessing, Hyperparameter Tuning using RandomizedSearchCV to tune hyperparameters and selecting the best parameters. Then it trains separate CatBoost classifiers One-vs-Rest approach. Making predictions which involves calculating overall accuracy, classification report for precision, recall, and F1-score, and visualizing the confusion matrix. Additionally, AUC-ROC curves are plotted.

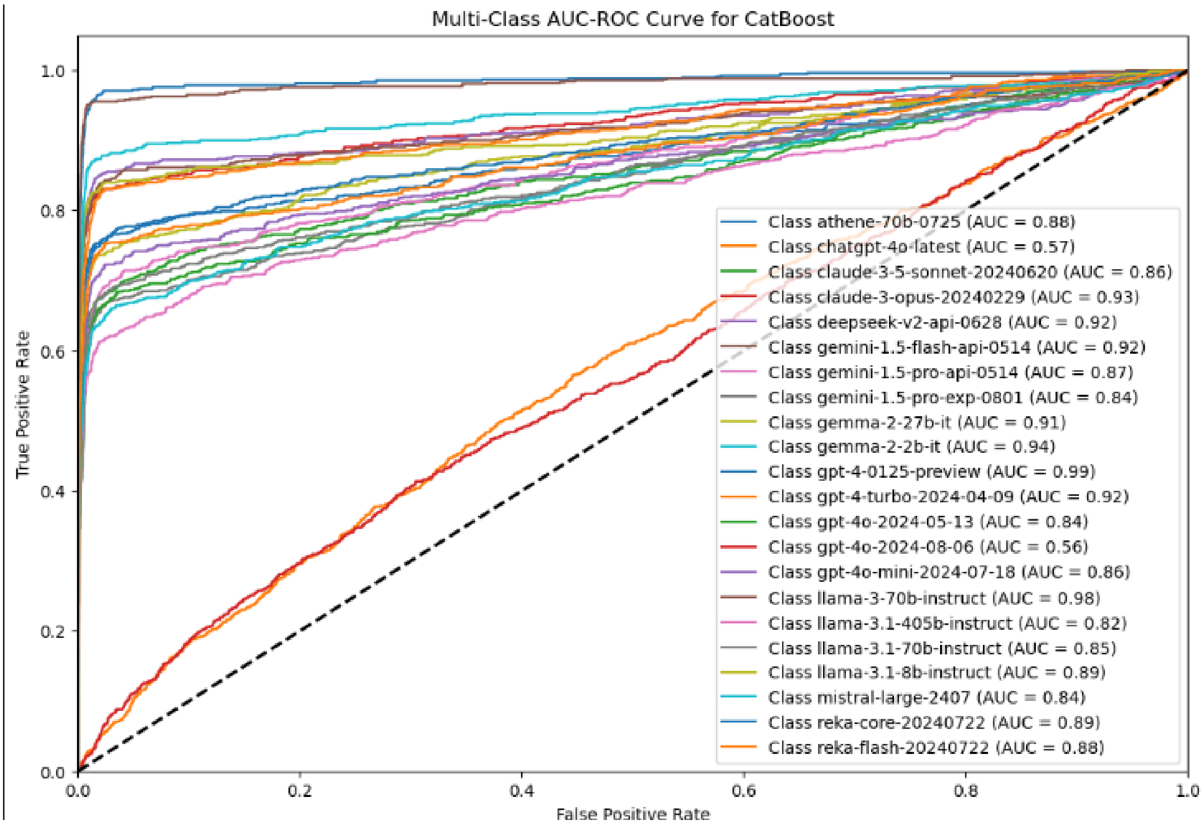


```
Fitting 3 folds for each of 50 candidates, totalling 150 fits
Best Parameters for CatBoost: {'learning_rate': 0.1388888888888889, 'l2_leaf_reg': 3, 'iterations': 1000, 'depth': 10, 'border_count': 128, 'bagging_temperature': 3}
Overall Accuracy of Tuned CatBoost Classifier: 0.66
Classification Report:
                            precision    recall  f1-score   support

           athene-70b-0725       0.69      0.72      0.71       830
         chatgpt-4o-latest       0.16      0.10      0.12       790
claude-3-5-sonnet-20240620       0.62      0.64      0.63       776
    claude-3-opus-20240229       0.71      0.77      0.74       784
      deepseek-v2-api-0628       0.72      0.77      0.75       773
 gemini-1.5-flash-api-0514       0.74      0.78      0.76       821
   gemini-1.5-pro-api-0514       0.66      0.64      0.65       847
   gemini-1.5-pro-exp-0801       0.62      0.61      0.62       796
            gemma-2-27b-it       0.72      0.80      0.76       822
             gemma-2-2b-it       0.76      0.84      0.80       828
        gpt-4-0125-preview       0.79      0.92      0.85       747
    gpt-4-turbo-2024-04-09       0.67      0.76      0.72       799
         gpt-4o-2024-05-13       0.64      0.61      0.62       785
         gpt-4o-2024-08-06       0.13      0.08      0.10       769
    gpt-4o-mini-2024-07-18       0.68      0.65      0.67       840
      llama-3-70b-instruct       0.81      0.90      0.85       825
   llama-3.1-405b-instruct       0.62      0.58      0.60       807
    llama-3.1-70b-instruct       0.65      0.61      0.63       796
     llama-3.1-8b-instruct       0.71      0.69      0.70       797
...
                  accuracy                           0.66     17600
                 macro avg       0.64      0.66      0.65     17600
              weighted avg       0.64      0.66      0.65     17600
```



True Labels \ Predicted Labels	athene-70b-0725	chatgpt-4o-latest	claude-3-5-sonnet-20240620	claude-3-opus-20240229	deepseek-v2-api-0628	gemini-1.5-flash-api-0514	gemini-1.5-pro-api-0514	gemini-1.5-pro-exp-0801	gemma-2-27b-it	gemma-2-2b-it	gpt-4-0125-preview	gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	gpt-4o-2024-08-06	gpt-4o-mini-2024-07-18	llama-3-70b-instruct	llama-3.1-405b-instruct	llama-3.1-70b-instruct	llama-3.1-8b-instruct	mistral-large-2407	reka-core-20240722	reka-flash-20240722
athene-70b-0725	595	21	11	15	7	7	13	16	7	8	4	10	18	10	14	7	8	13	10	11	4	21
chatgpt-4o-latest	26	80	42	37	23	25	45	43	21	29	18	22	40	78	35	18	36	32	30	38	39	33
claude-3-5-sonnet-20240620	7	29	498	14	11	16	12	15	9	2	6	14	12	24	17	8	8	13	10	21	13	17
claude-3-opus-20240229	9	11	9	606	8	5	14	7	8	10	6	10	9	17	4	8	12	10	3	5	4	9
deepseek-v2-api-0628	6	16	11	12	599	7	9	14	10	8	5	10	7	7	2	12	2	6	5	10	5	10
gemini-1.5-flash-api-0514	8	13	17	7	14	639	7	11	8	6	7	13	5	8	8	2	12	6	4	9	9	8
gemini-1.5-pro-api-0514	16	26	16	10	11	9	546	24	7	8	9	17	13	22	19	7	30	14	8	9	9	17
gemini-1.5-pro-exp-0801	19	19	16	15	13	15	4	487	18	11	16	14	14	26	14	10	16	9	11	17	17	15
gemma-2-27b-it	11	12	3	4	9	9	10	7	660	11	4	9	4	8	6	4	6	14	3	6	15	7
gemma-2-2b-it	7	17	6	2	9	10	7	1	10	693	6	7	3	11	7	1	6	3	4	9	4	5
gpt-4-0125-preview	3	2	0	1	7	2	1	5	3	6	687	10	1	2	2	2	0	1	0	4	3	5
gpt-4-turbo-2024-04-09	3	8	12	5	7	5	9	7	12	6	6	611	9	5	8	11	14	12	7	11	15	16
gpt-4o-2024-05-13	17	19	18	6	8	18	11	12	19	11	14	16	478	28	14	13	8	14	15	15	13	18
gpt-4o-2024-08-06	34	76	34	23	23	22	40	33	40	21	16	48	43	65	38	11	34	29	30	42	34	33
gpt-4o-mini-2024-07-18	20	21	22	11	9	12	15	20	10	10	13	15	12	18	548	12	12	15	16	9	9	11
llama-3-70b-instruct	0	10	2	3	5	1	2	5	4	8	7	5	2	3	0	745	3	1	4	4	4	7
llama-3.1-405b-instruct	10	29	22	15	13	11	11	15	18	10	9	14	12	32	14	13	472	17	17	23	16	14
llama-3.1-70b-instruct	17	18	15	16	14	13	16	15	14	14	8	16	15	22	17	9	16	485	18	13	13	12
llama-3.1-8b-instruct	10	16	10	9	8	7	22	10	5	9	7	12	17	35	10	4	10	20	546	11	9	10
mistral-large-2407	11	28	17	12	12	12	9	19	14	15	6	17	17	12	7	13	19	21	9	472	28	19
reka-core-20240722	20	12	9	12	9	7	9	5	11	4	4	10	7	31	9	7	13	8	12	14	557	12
reka-flash-20240722	8	23	15	13	10	13	12	10	7	11	10	6	10	23	11	4	23	8	10	13	20	537

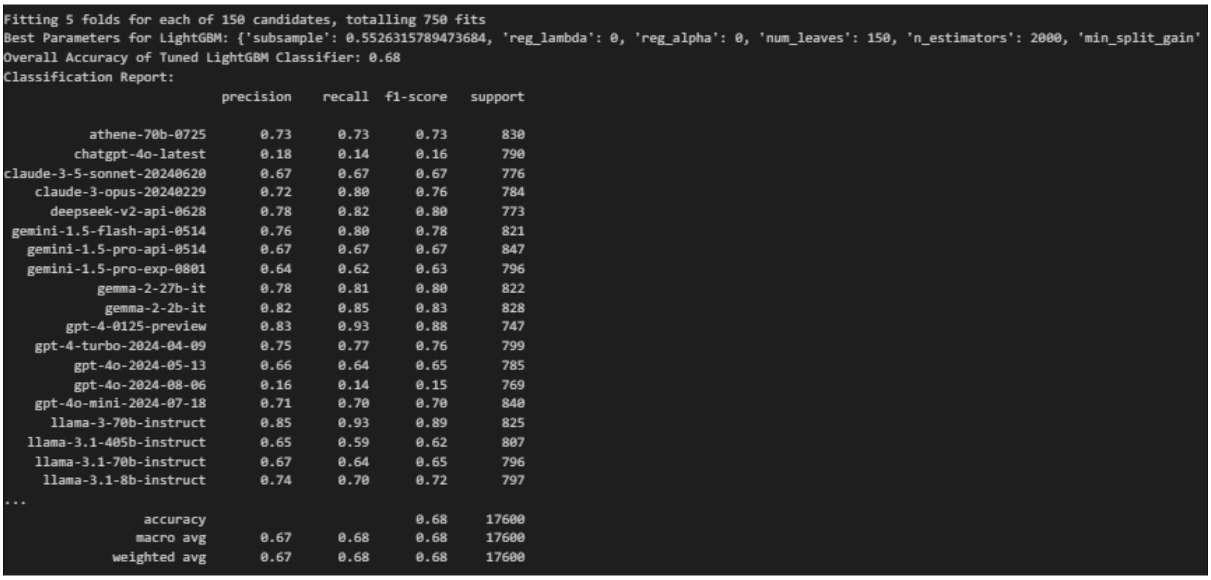
The Area Under the Curve (AUC) where values close to 1 indicate high ability and values closer to 0.5 suggest no better performance.



4. LIGHTGBM

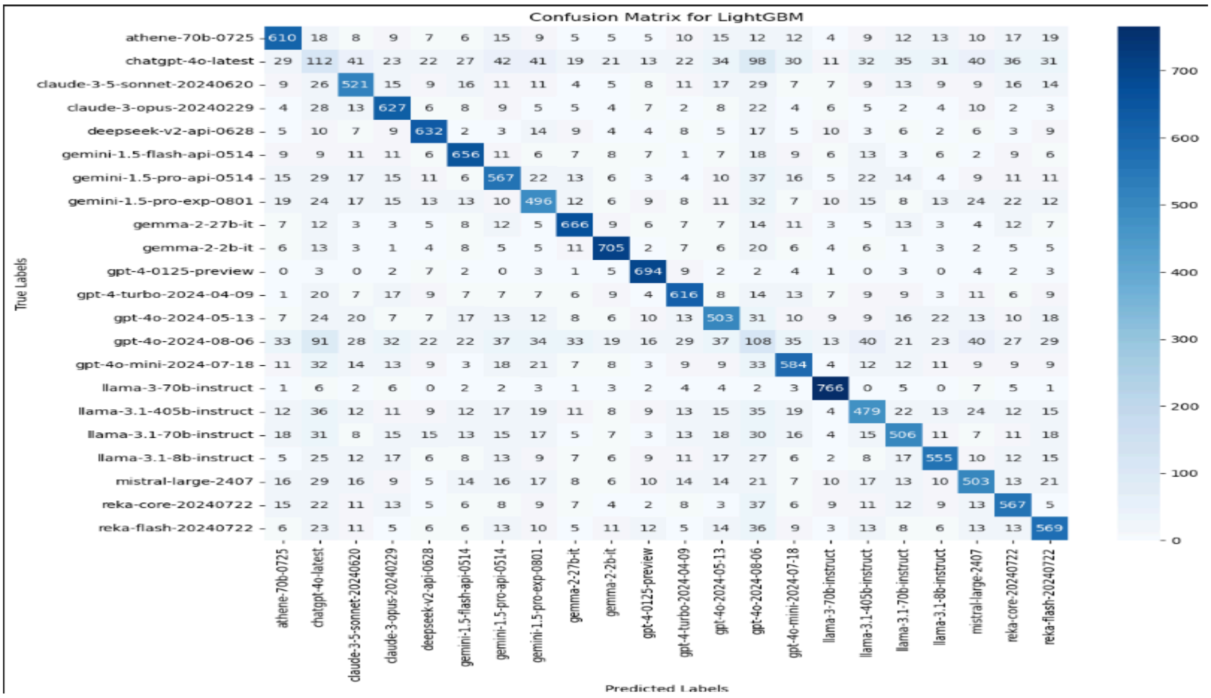
The **LightGBM** classifier achieved an overall accuracy of 0.68.

The code uses LightGBM for a multi-class classification through a One-vs-Rest approach and involves the following steps: Data Loading and Preprocessing, Hyperparameter Tuning using RandomizedSearchCV with 150 iterations and 5-fold cross-validation to tune hyperparameters and selecting the best parameters. Then using the One-vs-Rest method, separate LightGBM classifiers and train it. Evaluating the results with confusion matrices, classification reports, and AUC-ROC curves.



```
Fitting 5 folds for each of 150 candidates, totalling 750 fits
Best Parameters for LightGBM: {'subsample': 0.5526315789473684, 'reg_lambda': 0, 'reg_alpha': 0, 'num_leaves': 150, 'n_estimators': 2000, 'min_split_gain'
Overall Accuracy of Tuned LightGBM Classifier: 0.68
Classification Report:
                            precision    recall  f1-score   support

           athene-70b-0725       0.73      0.73      0.73       830
        chatgpt-4o-latest       0.18      0.14      0.16       790
claude-3-5-sonnet-20240620       0.67      0.67      0.67       776
    claude-3-opus-20240229       0.72      0.80      0.76       784
      deepseek-v2-api-0628       0.78      0.82      0.80       773
 gemini-1.5-flash-api-0514       0.76      0.80      0.78       821
   gemini-1.5-pro-api-0514       0.67      0.67      0.67       847
   gemini-1.5-pro-exp-0801       0.64      0.62      0.63       796
            gemma-2-27b-it       0.78      0.81      0.80       822
             gemma-2-2b-it       0.82      0.85      0.83       828
       gpt-4-0125-preview       0.83      0.93      0.88       747
    gpt-4-turbo-2024-04-09       0.75      0.77      0.76       799
        gpt-4o-2024-05-13       0.66      0.64      0.65       785
        gpt-4o-2024-08-06       0.16      0.14      0.15       769
    gpt-4o-mini-2024-07-18       0.71      0.70      0.70       840
      llama-3-70b-instruct       0.85      0.93      0.89       825
   llama-3.1-405b-instruct       0.65      0.59      0.62       807
    llama-3.1-70b-instruct       0.67      0.64      0.65       796
     llama-3.1-8b-instruct       0.74      0.70      0.72       797
...
                  accuracy                           0.68     17600
                 macro avg       0.67      0.68      0.68     17600
              weighted avg       0.67      0.68      0.68     17600
```

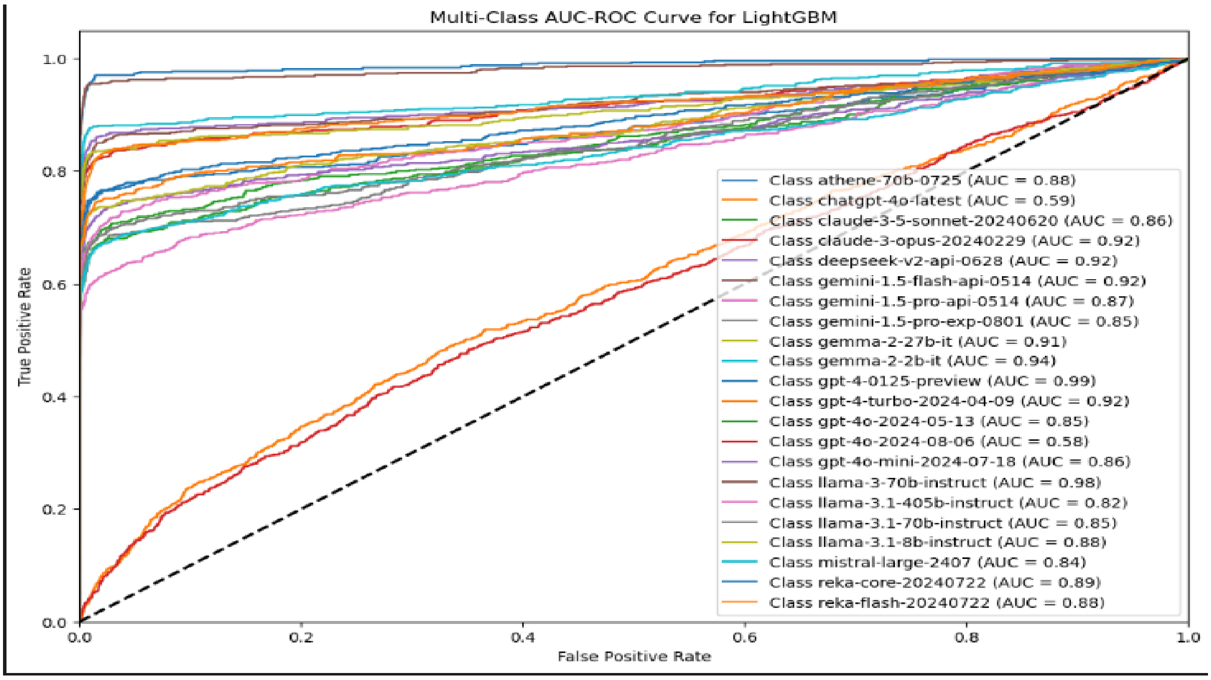
High diagonal values suggest that the classifier performs well in predicting those classes, whereas high off-diagonal values indicate areas where the classifier confuses one class for another



Confusion Matrix for LightGBM

	athene-70b-0725	chatgpt-4o-latest	claude-3-5-sonnet-20240620	claude-3-opus-20240229	deepseek-v2-api-0628	gemini-1.5-flash-api-0514	gemini-1.5-pro-api-0514	gemini-1.5-pro-exp-0801	gemma-2-27b-it	gemma-2-2b-it	gpt-4-0125-preview	gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	gpt-4o-2024-08-06	gpt-4o-mini-2024-07-18	llama-3-70b-instruct	llama-3.1-405b-instruct	llama-3.1-70b-instruct	llama-3.1-8b-instruct	mistral-large-2407	reka-core-20240722	reka-flash-20240722
athene-70b-0725	610	18	8	9	7	6	15	9	5	5	5	10	15	12	12	4	9	12	13	10	17	19
chatgpt-4o-latest	29	112	41	23	22	27	42	41	19	21	13	22	34	98	30	11	32	35	31	40	36	31
claude-3-5-sonnet-20240620	9	26	521	15	9	16	11	11	4	5	8	11	17	29	7	7	9	13	9	9	16	14
claude-3-opus-20240229	4	28	13	627	6	8	9	5	5	4	7	2	8	22	4	6	5	2	4	10	2	3
deepseek-v2-api-0628	5	10	7	9	632	2	3	14	9	4	4	8	5	17	5	10	3	6	2	6	3	9
gemini-1.5-flash-api-0514	9	9	11	11	6	656	11	6	7	8	7	1	7	18	9	6	13	3	6	2	9	6
gemini-1.5-pro-api-0514	15	29	17	15	11	6	567	22	13	6	3	4	10	37	16	5	22	14	4	9	11	11
gemini-1.5-pro-exp-0801	19	24	17	15	13	13	10	496	12	6	9	8	11	32	7	10	15	8	13	24	22	12
gemma-2-27b-it	7	12	3	3	5	8	12	5	666	9	6	7	7	14	11	3	5	13	3	4	12	7
gemma-2-2b-it	6	13	3	1	4	8	5	5	11	705	2	7	6	20	6	4	6	1	3	2	5	5
gpt-4-0125-preview	0	3	0	2	7	2	0	3	1	5	694	9	2	2	4	1	0	3	0	4	2	3
gpt-4-turbo-2024-04-09	1	20	7	17	9	7	7	7	6	9	4	616	8	14	13	7	9	9	3	11	6	9
gpt-4o-2024-05-13	7	24	20	7	7	17	13	12	8	6	10	13	503	31	10	9	9	16	22	13	10	18
gpt-4o-2024-08-06	33	91	28	32	22	22	37	34	33	19	16	29	37	108	35	13	40	21	23	40	27	29
gpt-4o-mini-2024-07-18	11	32	14	13	9	3	18	21	7	8	3	9	9	33	584	4	12	12	11	9	9	9
llama-3-70b-instruct	1	6	2	6	0	2	2	3	1	3	2	4	4	2	3	766	0	5	0	7	5	1
llama-3.1-405b-instruct	12	36	12	11	9	12	17	19	11	8	9	13	15	35	19	4	479	22	13	24	12	15
llama-3.1-70b-instruct	18	31	8	15	15	13	15	17	5	7	3	13	18	30	16	4	15	506	11	7	11	18
llama-3.1-8b-instruct	5	25	12	17	6	8	13	9	7	6	9	11	17	27	6	2	8	17	555	10	12	15
mistral-large-2407	16	29	16	9	5	14	16	17	8	6	10	14	14	21	7	10	17	13	10	503	13	21
reka-core-20240722	15	22	11	13	5	6	8	9	7	4	2	8	3	37	6	9	11	12	9	13	567	5
reka-flash-20240722	6	23	11	5	6	6	13	10	5	11	12	5	14	36	9	3	13	8	6	13	13	569

Values close to 1 indicating excellent ability and values near 0.5 suggesting no better accuracy.

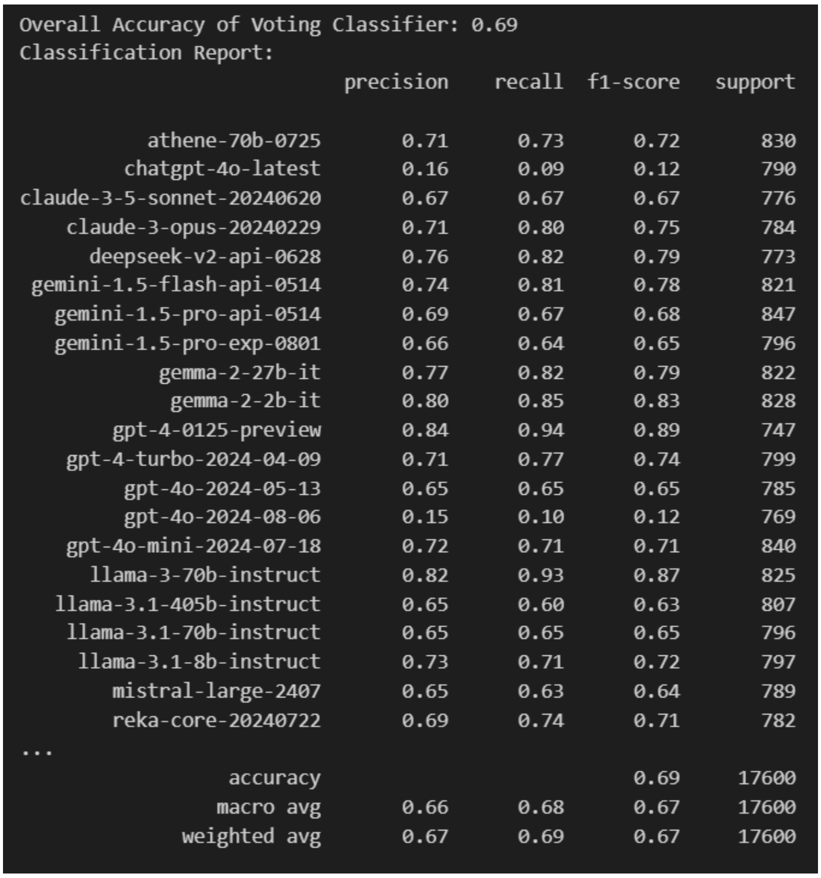


Multi-Class AUC-ROC Curve for LightGBM

5. VOTING

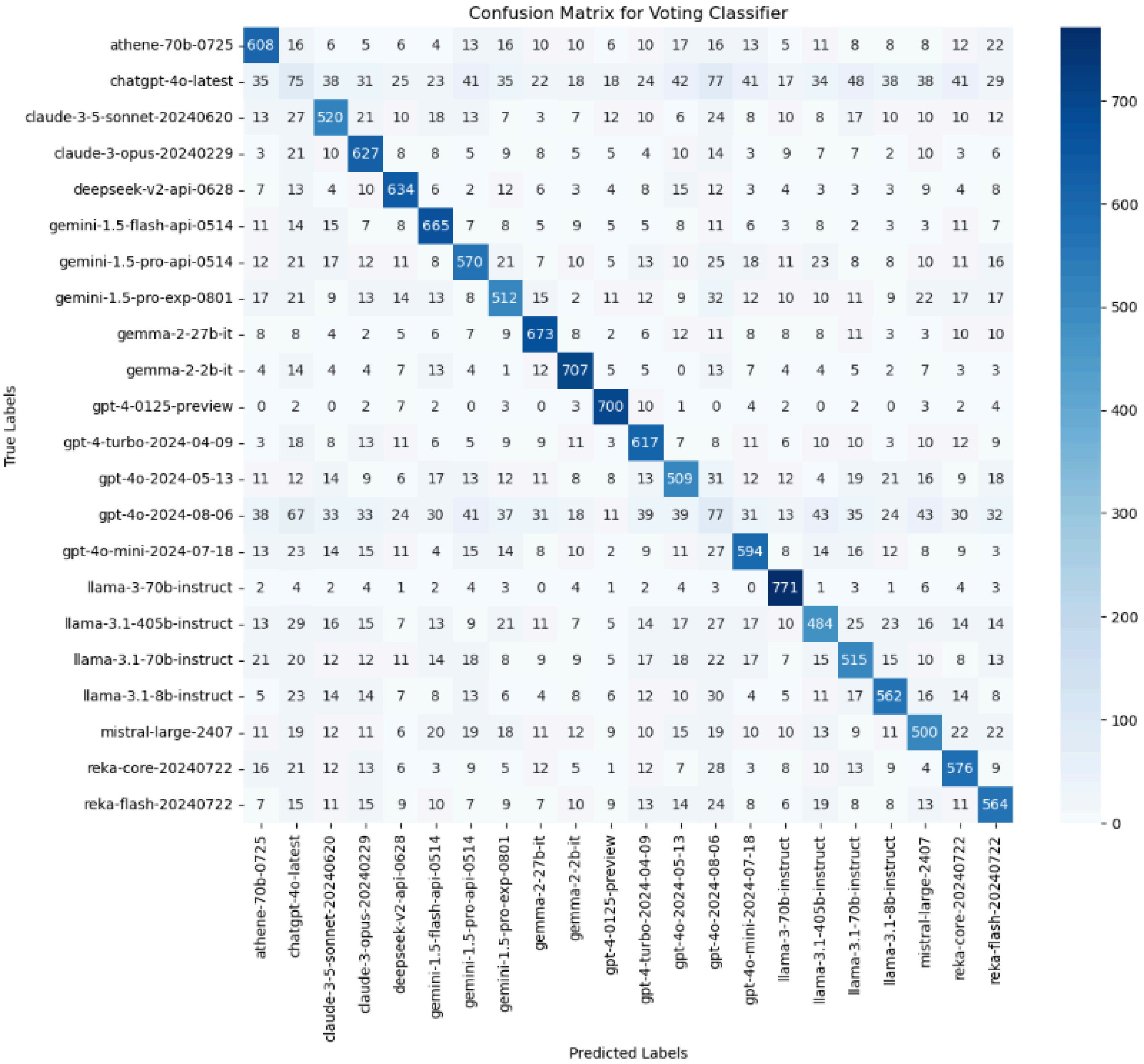
The **Voting** classifier achieved an overall accuracy of 0.69.

We handled a multi-class classification by utilizing a **Voting Classifier** by combining LightGBM, CatBoost, and XGBoost through a One-vs-Rest (OvR) strategy. and involves the following steps: Data Loading and Preprocessing, Model Initialization: where it initializes three models LightGBM, CatBoost, and XGBoost with hyperparameters which constructs a Voting Classifier using soft voting from all three models. Then Train the One-vs-Rest Model with multiple Voting Classifiers in a One-vs-Rest. Evaluating the results with confusion matrices and classification reports.



```
Overall Accuracy of Voting Classifier: 0.69
Classification Report:
                            precision    recall  f1-score   support

            athene-70b-0725       0.71      0.73      0.72       830
          chatgpt-4o-latest       0.16      0.09      0.12       790
claude-3-5-sonnet-20240620       0.67      0.67      0.67       776
    claude-3-opus-20240229       0.71      0.80      0.75       784
      deepseek-v2-api-0628       0.76      0.82      0.79       773
 gemini-1.5-flash-api-0514       0.74      0.81      0.78       821
   gemini-1.5-pro-api-0514       0.69      0.67      0.68       847
   gemini-1.5-pro-exp-0801       0.66      0.64      0.65       796
             gemma-2-27b-it       0.77      0.82      0.79       822
              gemma-2-2b-it       0.80      0.85      0.83       828
         gpt-4-0125-preview       0.84      0.94      0.89       747
     gpt-4-turbo-2024-04-09       0.71      0.77      0.74       799
          gpt-4o-2024-05-13       0.65      0.65      0.65       785
          gpt-4o-2024-08-06       0.15      0.10      0.12       769
     gpt-4o-mini-2024-07-18       0.72      0.71      0.71       840
      llama-3-70b-instruct       0.82      0.93      0.87       825
    llama-3.1-405b-instruct       0.65      0.60      0.63       807
     llama-3.1-70b-instruct       0.65      0.65      0.65       796
      llama-3.1-8b-instruct       0.73      0.71      0.72       797
         mistral-large-2407       0.65      0.63      0.64       789
         reka-core-20240722       0.69      0.74      0.71       782
...
                   accuracy                           0.69     17600
                  macro avg       0.66      0.68      0.67     17600
               weighted avg       0.67      0.69      0.67     17600
```

The numbers on the matrix diagonal (e.g., 608 for the first class) indicate correct predictions for each class, while off-diagonal numbers represent misclassifications.



Confusion Matrix for Voting Classifier

True Labels \ Predicted Labels	athene-70b-0725	chatgpt-4o-latest	claude-3-5-sonnet-20240620	claude-3-opus-20240229	deepseek-v2-api-0628	gemini-1.5-flash-api-0514	gemini-1.5-pro-api-0514	gemini-1.5-pro-exp-0801	gemma-2-27b-it	gemma-2-2b-it	gpt-4-0125-preview	gpt-4-turbo-2024-04-09	gpt-4o-2024-05-13	gpt-4o-2024-08-06	gpt-4o-mini-2024-07-18	llama-3-70b-instruct	llama-3.1-405b-instruct	llama-3.1-70b-instruct	llama-3.1-8b-instruct	mistral-large-2407	reka-core-20240722	reka-flash-20240722
athene-70b-0725	608	16	6	5	6	4	13	16	10	10	6	10	17	16	13	5	11	8	8	8	12	22
chatgpt-4o-latest	35	75	38	31	25	23	41	35	22	18	18	24	42	77	41	17	34	48	38	38	41	29
claude-3-5-sonnet-20240620	13	27	520	21	10	18	13	7	3	7	12	10	6	24	8	10	8	17	10	10	10	12
claude-3-opus-20240229	3	21	10	627	8	8	5	9	8	5	5	4	10	14	3	9	7	7	2	10	3	6
deepseek-v2-api-0628	7	13	4	10	634	6	2	12	6	3	4	8	15	12	3	4	3	3	3	9	4	8
gemini-1.5-flash-api-0514	11	14	15	7	8	665	7	8	5	9	5	5	8	11	6	3	8	2	3	3	11	7
gemini-1.5-pro-api-0514	12	21	17	12	11	8	570	21	7	10	5	13	10	25	18	11	23	8	8	10	11	16
gemini-1.5-pro-exp-0801	17	21	9	13	14	13	8	512	15	2	11	12	9	32	12	10	10	11	9	22	17	17
gemma-2-27b-it	8	8	4	2	5	6	7	9	673	8	2	6	12	11	8	8	8	11	3	3	10	10
gemma-2-2b-it	4	14	4	4	7	13	4	1	12	707	5	5	0	13	7	4	4	5	2	7	3	3
gpt-4-0125-preview	0	2	0	2	7	2	0	3	0	3	700	10	1	0	4	2	0	2	0	3	2	4
gpt-4-turbo-2024-04-09	3	18	8	13	11	6	5	9	9	11	3	617	7	8	11	6	10	10	3	10	12	9
gpt-4o-2024-05-13	11	12	14	9	6	17	13	12	11	8	8	13	509	31	12	12	4	19	21	16	9	18
gpt-4o-2024-08-06	38	67	33	33	24	30	41	37	31	18	11	39	39	77	31	13	43	35	24	43	30	32
gpt-4o-mini-2024-07-18	13	23	14	15	11	4	15	14	8	10	2	9	11	27	594	8	14	16	12	8	9	3
llama-3-70b-instruct	2	4	2	4	1	2	4	3	0	4	1	2	4	3	0	771	1	3	1	6	4	3
llama-3.1-405b-instruct	13	29	16	15	7	13	9	21	11	7	5	14	17	27	17	10	484	25	23	16	14	14
llama-3.1-70b-instruct	21	20	12	12	11	14	18	8	9	9	5	17	18	22	17	7	15	515	15	10	8	13
llama-3.1-8b-instruct	5	23	14	14	7	8	13	6	4	8	6	12	10	30	4	5	11	17	562	16	14	8
mistral-large-2407	11	19	12	11	6	20	19	18	11	12	9	10	15	19	10	10	13	9	11	500	22	22
reka-core-20240722	16	21	12	13	6	3	9	5	12	5	1	12	7	28	3	8	10	13	9	4	576	9
reka-flash-20240722	7	15	11	15	9	10	7	9	7	10	9	13	14	24	8	6	19	8	8	13	11	564

ACCURACY ACHIEVED BY EACH MODEL

MODEL NAME	ACCURACY
RANDOM FOREST	0.69
XG BOOST	0.68
CATBOOST	0.66
LIGHTGBM	0.68
VOTING	0.69