ML MODEL TESTING RESULTS PROGRESS

Model Testing 1 (21/2/26)

```
--- Core performance metrics ---

[Test 1] Accuracy: 79.34%

[Test 2] Weighted F1-score: 79.27%

[Test 3] Classification Report:
              precision    recall  f1-score   support

     Comment       1.00      0.94      0.97        17
        Like       0.60      0.61      0.61        49
        Play       0.90      0.94      0.92        50
      Search       1.00      1.00      1.00        47
   Subscribe       0.60      0.58      0.59        50

    accuracy                           0.79       213
   macro avg       0.82      0.81      0.82       213
weighted avg       0.79      0.79      0.79       213

[Test 4] Confusion Matrix:
[[16  0  1  0  0]
 [ 0 30  2  0 17]
 [ 0  1 47  0  2]
 [ 0  0  0 47  0]
 [ 0 19  2  0 29]]

[Test 5] ROC-AUC Score:
ROC-AUC Score (multi-class, OVR): 0.9557

[Test 6] Per-class accuracy:
Comment: 94.12%
Like: 61.22%
Play: 94.00%
Search: 100.00%
Subscribe: 58.00%
```

```
--- Model robustness & reliability ---

[Test 10] Overfitting check:

Training Accuracy: 100.00%
Testing Accuracy:  79.34%

[Test 11] Class distribution:
Total samples: 645
action
Like          150
Play          150
Subscribe     150
Search        143
Comment        52
Name: count, dtype: int64

[Test 12]Prediction confidence (first 10 rows):
Sample 0: 0.910
Sample 1: 0.605
Sample 2: 0.970
Sample 3: 0.940
Sample 4: 0.670
Sample 5: 1.000
Sample 6: 0.495
Sample 7: 0.600
Sample 8: 0.835
Sample 9: 1.000

[Test 13] Noise sensitivity test:
Noise 0.5% -> Accuracy: 81.22%
Noise 1.0% -> Accuracy: 77.46%
Noise 5.0% -> Accuracy: 78.87%
```

```
[Test 14] Stability test:

Stability scores: [0.8169014084507042, 0.82629
10798122066, 0.8262910798122066, 0.81690140845
07042, 0.784037558685446]
Mean stability: 0.8140845070422535
```

```
--- Feature analysis ---

[Test 7] Top Feature Importances:
duration              0.264287
std_iat               0.111429
pk_count              0.107639
avg_inbound_size      0.103818
total_bytes           0.099190
avg_pkt_size          0.083606
avg_iat               0.071522
outbound_ratio        0.048938
avg_outbound_size     0.046687
std_pkt_size          0.040553
max_pkt_size          0.022331
dtype: float64

[Test 8] Permutation Importance:
duration              2.638498e-01
std_iat               5.164319e-02
avg_outbound_size     1.408451e-02
avg_iat               7.042254e-03
pk_count              2.816901e-03
avg_inbound_size      1.110223e-17
total_bytes          -1.408451e-03
outbound_ratio       -4.694836e-03
std_pkt_size         -5.164319e-03
avg_pkt_size         -5.164319e-03
max_pkt_size         -7.981221e-03
dtype: float64

[Test 9] Feature ablation results:
Baseline accuracy: 79.34%
Without pk_count: 79.34%
Without total_bytes: 75.12%
Without avg_pkt_size: 78.40%
Without std_pkt_size: 76.53%
Without max_pkt_size: 78.40%
Without avg_iat: 78.87%
Without std_iat: 77.46%
Without duration: 77.93%
```

```
--- Error & unseen label analysis ---

[Test 15] Error analysis:

Misclassified samples:
     pk_count  total_bytes  ...       true       pred
380      2079      2275493  ...  Subscribe       Like
86       4981      5571646  ...       Like  Subscribe
429      2322      2512597  ...  Subscribe       Like
372      1839      1949806  ...  Subscribe       Like
129      3728      4117888  ...       Like  Subscribe
416      2576      2786165  ...  Subscribe       Like
395      5168      5824475  ...  Subscribe       Like
432      6049      6915035  ...  Subscribe       Like
83       5571      6361361  ...       Like  Subscribe
366      7274      8205666  ...  Subscribe       Like
402      2084      2161600  ...  Subscribe       Like
220      1572      6319154  ...       Play  Subscribe
373      3021      3204798  ...  Subscribe       Like
61       4129      4476640  ...       Like       Play
87       2644      2916359  ...       Like  Subscribe
8         872      1900734  ...       Like  Subscribe
149      1465      1552587  ...       Like  Subscribe
110      2866      3123472  ...       Like  Subscribe
423      6349      7213319  ...  Subscribe       Like
70       1727      1775297  ...       Like  Subscribe

[20 rows x 13 columns]
Total misclassified: 44

[Test 16] Testing model handling of unseen labels:

Unknown label handling test passed.
Model correctly raises error for unseen class:
Error: y contains previously unseen labels: 'new_action'

[test 17] General Sanity check (first 5 rows of features):
     pk_count  total_bytes  ...  avg_outbound_size  avg_inbound_size
485      7342      8039065  ...         235.863911       1229.147717
449      4782      5405458  ...         188.973638       1257.520057
460     10287     11589139  ...         158.495955       1258.782234
488      4328      4424436  ...         332.895161       1165.391183
345       741      1907284  ...         634.190476       3573.554192

[5 rows x 11 columns]
```

Confusion Matrix

ROC Curves (One-vs-Rest)

Comment (AUC = 1.00)
Like (AUC = 0.91)
Play (AUC = 0.98)
Search (AUC = 1.00)
Subscribe (AUC = 0.89)

Feature Importance

Permutation Importance

Prediction Confidence Distribution