AI-Assisted Evaluation Report

Definitions

Precision: The fraction of predicted positive instances that are truly positive. It measures the accuracy of the positive predictions.

Recall: The fraction of actual positive instances that are correctly predicted. It measures the ability to identify all positive instances.

F1-Score: The harmonic mean of precision and recall, providing a balance between the two-especially useful when the class distribution is imbalanced.

1. Overall Accuracy

MUT correctly classified 29 out of 46 documents.

Overall Accuracy: 63.04%

2. Precision, Recall, and F1-Score (Per Document Type)

Best Performance: W2 (F1-score: 0.84)

Worst Performance: None of the above (F1-score: 0.00)

Detailed Metrics per Document Type:

Invoice: Precision = 0.62, Recall = 0.38, F1-Score = 0.48, Support = 13.0

Receipt: Precision = 0.54, Recall = 0.78, F1-Score = 0.64, Support = 9.0

W2: Precision = 0.89, Recall = 0.80, F1-Score = 0.84, Support = 10.0

Bank-statement: Precision = 0.56, Recall = 0.82, F1-Score = 0.67, Support = 11.0

None of the above: Precision = 0.00, Recall = 0.00, F1-Score = 0.00, Support = 3.0

3. Misclassification Insights

Bank-statement: Recall = 81.82% (Correct: 9 of 11), Misclassified: 2, Most common misclassification: Invoice

Invoice: Recall = 38.46% (Correct: 5 of 13), Misclassified: 8, Most common misclassification: Receipt

None of the above: Recall = 0.00% (Correct: 0 of 3), Misclassified: 3, Most common misclassification: Bank-statement

Receipt: Recall = 77.78% (Correct: 7 of 9), Misclassified: 2, Most common misclassification: Invoice

W2: Recall = 80.00% (Correct: 8 of 10), Misclassified: 2, Most common misclassification: Receipt

4. Confidence Score Analysis

Bank-statement: Average MUT Confidence = 0.42, Ground Truth Confidence = 0.95, Drift = 0.53

Invoice: Average MUT Confidence = 0.27, Ground Truth Confidence = 0.95, Drift = 0.68

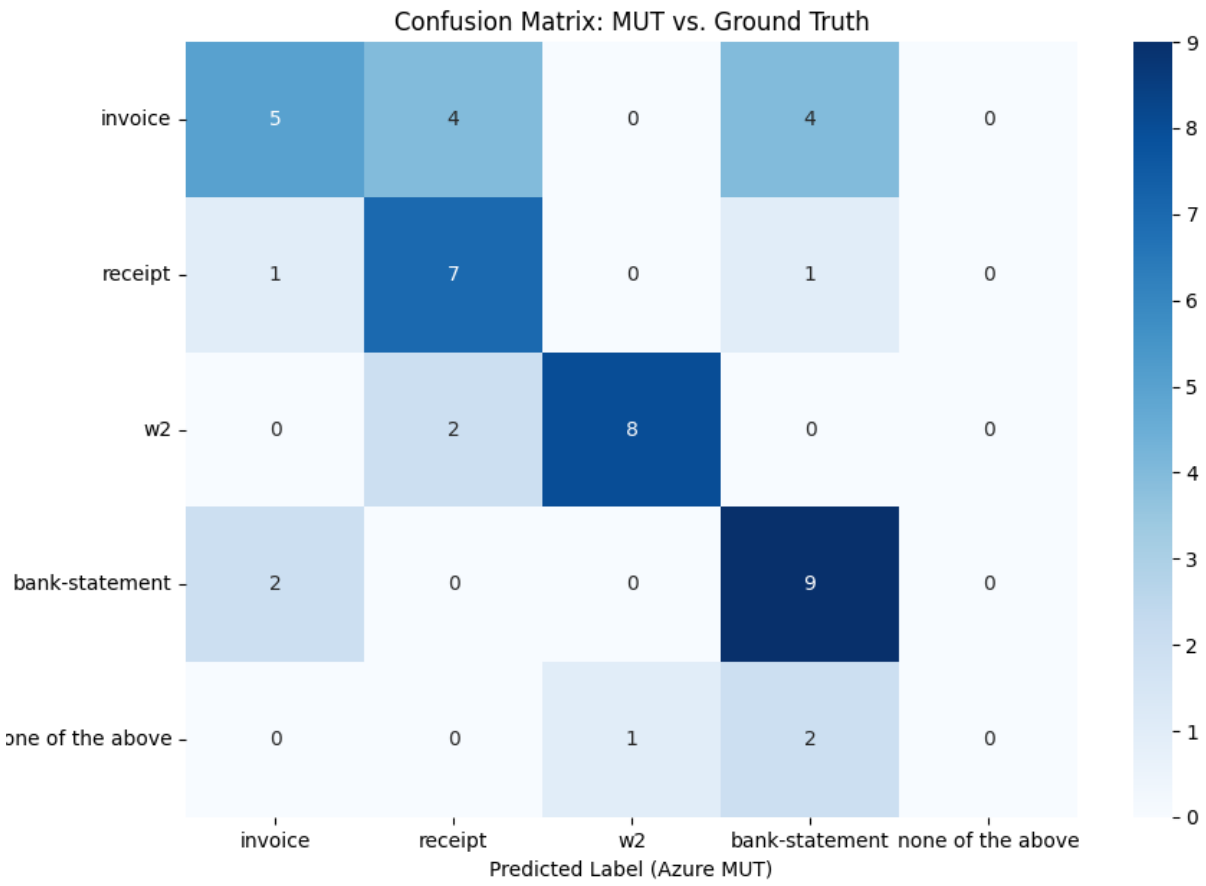
None of the above: Average MUT Confidence = 0.39, Ground Truth Confidence = 0.95, Drift = 0.56

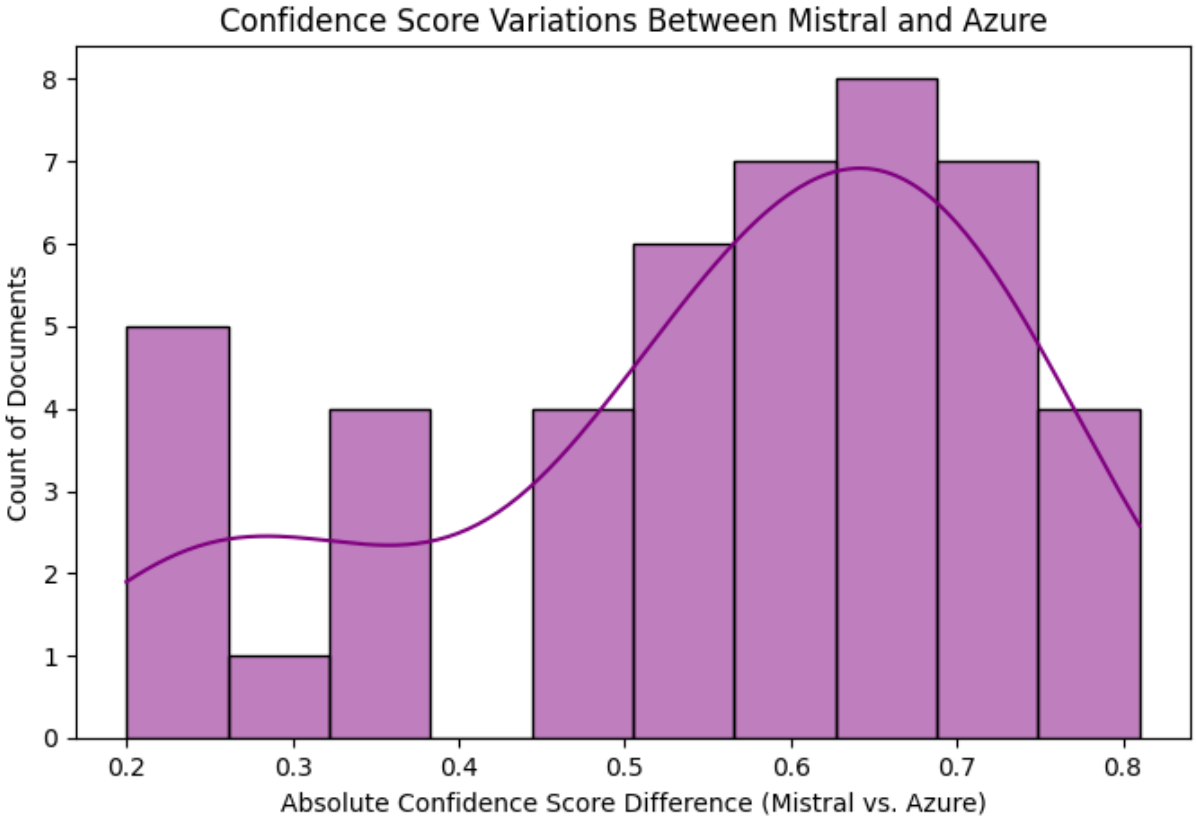
Receipt: Average MUT Confidence = 0.47, Ground Truth Confidence = 0.95, Drift = 0.48

W2: Average MUT Confidence = 0.50, Ground Truth Confidence = 0.98, Drift = 0.48

Overall, there are 9 low-confidence MUT predictions (less than or equal to 0.25). The highest average MUT confidence is observed for W2.

Visual Insights


Confusion Matrix: MUT vs. Ground Truth
invoice
receipt
w2
bank-statement
one of the above
5
4
0
4
0
1
7
0
1
0
0
2
8
0
0
2
0
0
9
0
0
0
1
2
0
9
8
7
6
5
4
3
2
1
0
invoice
receipt
w2
bank-statement
none of the above
Predicted Label (Azure MUT)



Understanding the Charts: Ideal vs. Actual

1. Understanding the Confusion Matrix

What is an 'Ideal' Confusion Matrix?

An ideal confusion matrix would have high values on the diagonal (correct classifications) and low values elsewhere (misclassifications).

Ideal Example (Perfect Classifier):

		Predicted			
		Invoice	Receipt	W2	Bank
Actual	Invoice	10	0	0	0
	Receipt	0	10	0	0
	W2	0	0	10	0
	Bank	0	0	0	10

Perfect Model - All correct classifications are along the diagonal, with no misclassifications.

What Does Our Actual Confusion Matrix Show?

	Invoice	Receipt	W2	Bank-statement	None of the above

```
                        ------------------------------------------------------
Actual |           Invoice        5        4   0              4                  0
Actual |           Receipt        1        7   0              1                  0
Actual |                W2        0        2   8              0                  0
Actual |    Bank-statement        2        0   0              9                  0
Actual | None of the above        0        0   1              2                  0
```

Dynamic Insights: Compare the counts along the diagonal with the off-diagonal values to understand misclassifications.

2. Understanding the Confidence Score Chart

What Would an 'Ideal' Confidence Score Chart Look Like?

Ideally, Mistral and Azure should have very similar confidence scores for each document. The histogram bars should be concentrated around 0.0 - 0.2 (small differences).

Ideal Confidence Score Chart:

```
|*************
|*****************
|*******************
|*************
|*****
|*
----------------------------
  0.0   0.2   0.4   0.6   0.8
  (Small Confidence Differences - Good Agreement)
```

What Does Our Actual Confidence Score Chart Show?

```
|
|******
|**********
|***************
|
----------------------------
  0.0   0.2   0.4   0.6   0.8
```

Dynamic Insights: Larger bars toward higher differences indicate greater disagreement between models.