

**A Sample Calculation for Accident Date for GPT-3.5 (n=81)**

Incorrect Accident Date Extract for Daily Star =15 out of 15

Incorrect Accident Date Extract for Dhaka Tribune =32 out of 32

Incorrect Accident Date Extract for Prothom Alo =33 out of 34

So, Total Incorrect Accident Date Extract =80 out of 81

% Accuracy for Accident Date (GPT-3.5) = = 1.23%

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Report Categorization Task Evaluation** | | | | | | | | |
| **LLM** |  | **Input Data** | | | | **Results** | | |
| **Newspaper** | **TP** | **TN** | **FP** | **FN** | **Precision** | **Recall** | **F1-score** |
| GPT 4 | Daily Star | 15 | 25 | 0 | 0 | 100% | 100% | 100% |
| Dhaka Tribune | 31 | 9 | 0 | 0 |
| Prothom Alo | 34 | 6 | 0 | 0 |
| Total | 80 | 40 | 0 | 0 |
| GPT 3.5 | Daily Star | 15 | 9 | 16 | 0 | 74.07% | 100% | 85.11% |
| Dhaka Tribune | 31 | 1 | 8 | 0 |
| Prothom Alo | 34 | 2 | 4 | 0 |
| Total | 80 | 12 | 28 | 0 |
| Llama3 | Daily Star | 15 | 25 | 0 | 0 | 98.75% | 98.75% | 98.75% |
| Dhaka Tribune | 31 | 8 | 1 | 0 |
| Prothom Alo | 33 | 6 | 0 | 1 |
| Total | 79 | 39 | 1 | 1 |

**A Sample Calculation for Report Categorization Task Evaluation For GPT-3.5**

From Table,

Total True Positive (TP) = 80

Total True Negative (TN) = 12

Total False Positive (FP) = 28

Total False Negative (FN) = 0

So,

Precision, P=

Recall, R=

F1-score=

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data Extraction Task Evaluation** | | | | | | | | |
| **LLM** |  | **Input Data** | | | | **Results** | | |
| **Newspaper** | **TP** | **TN** | **FP** | **FN** | **Precision** | **Recall** | **F1-score** |
| GPT 4 | Daily Star | 109 | 9 | 2 | 0 | 97.90% | 99.47% | 98.68% |
| Dhaka Tribune | 207 | 37 | 3 | 1 |
| Prothom Alo | 244 | 28 | 7 | 2 |
| Total | 560 | 74 | 12 | 3 |
| GPT 3.5 | Daily Star | 70 | 11 | 4 | 35 | 94.33% | 70.11% | 80.44% |
| Dhaka Tribune | 148 | 48 | 10 | 42 |
| Prothom Alo | 148 | 37 | 8 | 79 |
| Total | 366 | 96 | 22 | 156 |
| Llama3 | Daily Star | 102 | 6 | 10 | 2 | 88.97% | 98.83% | 93.64% |
| Dhaka Tribune | 195 | 32 | 21 | 0 |
| Prothom Alo | 211 | 17 | 32 | 4 |
| Total | 508 | 55 | 63 | 6 |

**A Sample Calculation for Data Extraction Task Evaluation for Llama3**

From Table,

Total True Positive (TP) = 508

Total True Negative (TN) = 55

Total False Positive (FP) = 63

Total False Negative (FN) = 6

So,

Precision, P=

Recall, R=

F1-score=