| Level                | Туре         | Metric                          | R1     |          |        |                   |   |   | R2     |          |        |               |        |   | Diff (R1 - R2)    |    |    |               |    | Final Score |                   |        |          |          |           |          |             |             |
|----------------------|--------------|---------------------------------|--------|----------|--------|-------------------|---|---|--------|----------|--------|---------------|--------|---|-------------------|----|----|---------------|----|-------------|-------------------|--------|----------|----------|-----------|----------|-------------|-------------|
|                      |              |                                 | Finan  | nce Dash | board  | ard MVP Predictor |   |   | Finar  | nce Dasi | nboard | MVP Predictor |        |   | Finance Dashboard |    |    | MVP Predictor |    |             | Finance Dashboard |        |          | M        | /P Predic | ctor     |             |             |
|                      |              |                                 | Α      | В        | С      | Α                 | В | С | Α      | В        | С      | Α             | В      | С | Α                 | В  | С  | Α             | В  | С           | Α                 | В      | С        | Α        | В         | С        | Average     | SD          |
|                      | BuggyCode    | Summary Accuracy                | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 4        | 5      | 5             | 5      | 5 | 0                 | 1  | 0  | 0             | 0  | 0           | 5                 | 4.5    | 5        | 5        | 5         | 5        | 4.75        | 0.226133508 |
|                      |              | Summary Clarity                 | 5      | 4        | 5      | 5                 | 5 | 5 | 5      | 5        | 5      | 5             | 5      | 5 | 0                 | -1 | 0  | 0             | 0  | 0           | 5                 | 4.5    | 5        | 5        | 5         | 5        | 4.875       | 0.24618298  |
|                      |              | Segmentation Granularity        | 5      | 4        | 5      | 3                 | 5 | 5 | 4      | 5        | 5      | 5             | 5      | 5 | 1                 | -1 | 0  | -2            | 0  | 0           | 4.5               | 4.5    | 5        | 3.5      | 5         | 5        | 4.666666667 |             |
|                      |              | Summary Accuracy                | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 4        | 4      | 5             | 5      | 5 | 0                 | 1  | 1  | 0             | 0  | 0           | 5                 | 4.5    | 4.5      | 5        | 5         | 5        |             |             |
| Unstructured, Low    | GroundTruth  | Summary Clarity                 | 5      | 4        | 5      | 5                 | 5 | 5 | 5      | 5        | 4      | 5             | 4      | 5 | 0                 | -1 | 1  | 0             | 1  | 0           | 5                 | 4.5    | 4.5      | 5        | 4.5       | 5        |             |             |
|                      |              | Segmentation Granularity        | 5      | 4        | 5      | 3                 | 5 | 5 | 4      | 5        | 4      | 4             | 4      | 5 | 1                 | -1 | 1  | -1            | 1  | 0           | 4.5               | 4.5    | 4.5      | 3.5      | 4.5       | 5        |             |             |
|                      | Summary Diff | Diff Faithfulness               | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 4      | 5             | 5      | 5 | 0                 | 0  | 1  | 0             | 0  | 0           | 5                 | 5      | 4.5      | 5        | 5         | 5        | 4.016666667 | 0.204124145 |
|                      |              | Diff Completeness               | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 4      | 5             | 5      | 5 | 0                 | 0  | 1  | 0             | 0  | 0           | 5                 | 5      | 4.5      | 5        | 5         | 5        |             | 0.204124145 |
|                      |              |                                 | 5      | 5        | 5      | 4                 | 5 | - | -      | -        |        |               |        | - | -                 | -  |    | -             | 0  | -           |                   | 5      | 5        | 4.5      | -         | 5        |             |             |
| Unstructured, Medium |              | Diff Salience                   | -      | -        |        |                   | - | 5 | 5      | 5        | 5      | 5             | 4      | 5 | 0                 | 0  | 0  | -1            | 1  | 0           | 5                 | -      | + -      |          | 4.5       | -        | 4.833333333 |             |
|                      | BuggyCode    | Summary Accuracy                | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 4      | 5             | 5      | 5 | 0                 | 0  | 1  | 0             | 0  | 0           | 5                 | 5      | 4.5      | 5        | 5         | 5        | 4.916666667 | 0.1946247   |
|                      |              | Summary Clarity                 | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 5      | 5             | 4      | 5 | 0                 | 0  | 0  | 0             | 1  | 0           | 5                 | 5      | 5        | 5        | 4.5       | 5        | 4.916666667 | 0.2461829   |
|                      |              | Segmentation Granularity        | 5      | 5        | 4      | 5                 | 5 | 5 | 5      | 5        | 5      | 5             | 5      | 5 | 0                 | 0  | -1 | 0             | 0  | 0           | 5                 | 5      | 4.5      | 5        | 5         | 5        | 4.833333333 | 0.310790780 |
|                      | GroundTruth  | Summary Accuracy                | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 4      | 5             | 5      | 5 | 0                 | 0  | 1  | 0             | 0  | 0           | 5                 | 5      | 4.5      | 5        | 5         | 5        |             |             |
|                      |              | Summary Clarity                 | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 4      | 4             | 4      | 5 | 0                 | 0  | 1  | 1             | 1  | 0           | 5                 | 5      | 4.5      | 4.5      | 4.5       | 5        |             |             |
|                      |              | Segmentation Granularity        | 5      | 5        | 4      | 5                 | 5 | 5 | 5      | 5        | 4      | 5             | 5      | 5 | 0                 | 0  | 0  | 0             | 0  | 0           | 5                 | 5      | 4        | 5        | 5         | 5        |             |             |
|                      | Summary Diff | Diff Faithfulness               | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 4      | 5             | 5      | 5 | 0                 | 0  | 1  | 0             | 0  | 0           | 5                 | 5      | 4.5      | 5        | 5         | 5        | 4.916666667 | 0.20412414  |
|                      |              | Diff Completeness               | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 5      | 4             | 5      | 5 | 0                 | 0  | 0  | 1             | 0  | 0           | 5                 | 5      | 5        | 4.5      | 5         | 5        | 4.916666667 | 0.20412414  |
|                      |              | Diff Salience                   | 5      | 5        | 5      | 4                 | 4 | 4 | 3      | 4        | 4      | 4             | 4      | 4 | 2                 | 1  | 1  | 0             | 0  | 0           | 4                 | 4.5    | 4.5      | 4        | 4         | 4        | 4.166666667 | 0.258198889 |
| Unstructured, High   | BuggyCode    | Summary Accuracy                | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 5      | 5             | 5      | 5 | 0                 | 0  | 0  | 0             | 0  | 0           | 5                 | 5      | 5        | 5        | 5         | 5        |             | 0.144337567 |
|                      |              | Summary Clarity                 | 4      | 5        | 4      | 5                 | 5 | 5 | 4      | 3        | 4      | 5             | 5      | 4 | 0                 | 2  | 0  | 0             | 0  | 1           | 4                 | 4.5    | 4        | 5        | 5         | 4.5      | 4.708333333 |             |
|                      |              | Segmentation Granularity        | 3      | 5        | 3      | 5                 | 4 | 3 | 5      | 4        | 3      | 3             | 4      | 5 | -2                | 1  | 0  | 2             | 0  | -2          | 3                 | 4.5    | 3        | 4.5      | 4         | 3.5      | 4.083333333 |             |
|                      |              | Summary Accuracy                | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 4      | 5             | 5      | 5 | 0                 | 0  | 1  | 0             | 0  | 0           | 5                 | 5      | 4.5      | 5        | 5         | 5        | 4.003333333 | 0.74230330  |
|                      | GroundTruth  |                                 | 5      | 5        | 4      | 5                 | 5 | 5 | -      |          | -      |               |        |   | 4                 |    |    |               | -  | _           |                   | 4.5    | 4.5      | 4.5      | 4.5       | 4.5      |             |             |
|                      |              | Summary Clarity                 |        |          |        |                   |   |   | 4      | 3        | 4      | 3             | 4      | 4 | 1                 | 2  | 0  | 2             | 1  | 1           | 4.5               |        |          |          |           |          |             |             |
|                      |              | Segmentation Granularity        | 3      | 5        | 3      | 5                 | 4 | 3 | 3      | 4        | 3      | 5             | 5      | 5 | 0                 | 1  | 0  | 0             | -1 | -2          | 3                 | 4.5    | 3        | 5        | 4.5       | 4        |             |             |
|                      | Summary Diff | Diff Faithfulness               | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 5      | 5             | 5      | 5 | 0                 | 0  | 0  | 0             | 0  | 0           | 5                 | 5      | 5        | 5        | 5         | 5        | 5           |             |
|                      |              | Diff Completeness               | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 4        | 4      | 5             | 4      | 5 | 0                 | 1  | 1  | 0             | 1  | 0           | 5                 | 4.5    | 4.5      | 5        | 4.5       | 5        |             | 0.273861278 |
|                      |              | Diff Salience                   | 3      | 5        | 5      | 3                 | 5 | 2 | 2      | 5        | 4      | 4             | 4      | 3 | 1                 | 0  | 1  | -1            | 1  | -1          | 2.5               | 5      | 4.5      | 3.5      | 4.5       | 2.5      | 3.75        |             |
| Structured, Low      | BuggyCode    | Summary Accuracy                | 5      | 5        | 5      | 5                 | 5 | 5 | 4      | 5        | 4      | 5             | 5      | 4 | 1                 | 0  | 1  | 0             | 0  | 1           | 4.5               | 5      | 4.5      | 5        | 5         | 4.5      | 4.708333333 | 0.22613350  |
|                      |              | Summary Clarity                 | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 5      | 5             | 5      | 5 | 0                 | 0  | 0  | 0             | 0  | 0           | 5                 | 5      | 5        | 5        | 5         | 5        | 5           |             |
|                      |              | Segmentation Granularity        | 5      | 4        | 5      | 5                 | 5 | 4 | 4      | 5        | 4      | 5             | 5      | 5 | 1                 | -1 | 1  | 0             | 0  | -1          | 4.5               | 4.5    | 4.5      | 5        | 5         | 4.5      | 4.833333333 | 0.26111648  |
|                      | GroundTruth  | Summary Accuracy                | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 5      | 5             | 5      | 5 | 0                 | 0  | 0  | 0             | 0  | 0           | 5                 | 5      | 5        | 5        | 5         | 5        |             |             |
|                      |              | Summary Clarity                 | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 5      | 5             | 5      | 5 | 0                 | 0  | 0  | 0             | 0  | 0           | 5                 | 5      | 5        | 5        | 5         | 5        |             |             |
|                      |              | Segmentation Granularity        | 5      | 4        | 5      | 5                 | 5 | 4 | 5      | 5        | 5      | 5             | 5      | 5 | 0                 | -1 | 0  | 0             | 0  | -1          | 5                 | 4.5    | 5        | 5        | 5         | 4.5      |             |             |
|                      | Summary Diff | Diff Faithfulness               | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 5      | 5             | 5      | 5 | 0                 | 0  | 0  | 0             | 0  | 0           | 5                 | 5      | 5        | 5        | 5         | 5        | 5           |             |
|                      |              | Diff Completeness               | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 5      | 4             | 5      | 4 | 0                 | 0  | 0  | 1             | 0  | 1           | 5                 | 5      | 5        | 4.5      | 5         | 4.5      | 4.833333333 | 0.25819888  |
|                      |              | Diff Salience                   | 4      | 5        | 5      | 5                 | 5 | 5 | 4      | 5        | 5      | 4             | 5      | 5 | 0                 | 0  | 0  | 1             | 0  | 0           | 4                 | 5      | 5        | 4.5      | 5         | 5        |             | 0.418330013 |
| Structured, Medium   | BuggyCode    | Summary Accuracy                | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 5      | 4             | 5      | 5 | 0                 | 0  | 0  | 1             | 0  | 0           | 5                 | 5      | 5        | 4.5      | 5         | 5        | 4.916666667 | 0.19462473  |
|                      |              | Summary Clarity                 | 5      | 5        | 5      | 5                 | 5 | 5 | 4      | 5        | 5      | 5             | 4      | 5 | 1                 | 0  | 0  | 0             | 1  | 0           | 4.5               | 5      | 5        | 5        | 4.5       | 5        | 4.875       |             |
|                      |              | Segmentation Granularity        | 4      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        |        |               | 5      |   | -1                | 0  | 0  | 0             | 0  | 0           | 4.5               | 5      | 5        | 5        | 5         | 5        |             |             |
|                      |              |                                 | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 5      | 5             | 5      | 5 | -1                | 0  | 0  | 1             | 0  | 0           | 5                 | 5      | 5        | 4.5      | 5         | 5        | 4.958333333 | 0.19462473  |
|                      | GroundTruth  | Summary Accuracy                |        |          |        |                   |   |   | -      | -        |        |               | -      | - | -                 | -  |    | -             | -  |             |                   |        |          |          |           |          |             |             |
|                      |              | Summary Clarity                 | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 5      | 5             | 5      | 5 | 0                 | 0  | 0  | 0             | 0  | 0           | 5                 | 5      | 5        | 5        | 5         | 5        |             |             |
|                      |              | Segmentation Granularity        | 4      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 5      | 5             | 5      | 5 | -1                | 0  | 0  | 0             | 0  | 0           | 4.5               | 5      | 5        | 5        | 5         | 5        |             |             |
|                      | Summary Diff | Diff Faithfulness               | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 5      | 5             | 5      | 4 | 0                 | 0  | 0  | 0             | 0  | 1           | 5                 | 5      | 5        | 5        | 5         | 4.5      | 4.916666667 | 0.20412414  |
|                      |              | Diff Completeness               | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 5      | 5             | 5      | 5 | 0                 | 0  | 0  | 0             | 0  | 0           | 5                 | 5      | 5        | 5        | 5         | 5        | 5           |             |
|                      |              | Diff Salience                   | 3      | 5        | 5      | 5                 | 5 | 4 | 3      | 5        | 5      | 4             | 4      | 5 | 0                 | 0  | 0  | 1             | 1  | -1          | 3                 | 5      | 5        | 4.5      | 4.5       | 4.5      | 4.416666667 | 0.73598007  |
| Structured, High     | BuggyCode    | Summary Accuracy                | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 5      | 5             | 5      | 5 | 0                 | 0  | 0  | 0             | 0  | 0           | 5                 | 5      | 5        | 5        | 5         | 5        | 4.916666667 |             |
|                      |              | Summary Clarity                 | 5      | 5        | 5      | 5                 | 5 | 5 | 4      | 4        | 5      | 4             | 4      | 5 | 1                 | 1  | 0  | 1             | 1  | 0           | 4.5               | 4.5    | 5        | 4.5      | 4.5       | 5        | 4.833333333 | 0.2461829   |
|                      |              | Segmentation Granularity        | 4      | 5        | 5      | 4                 | 5 | 5 | 5      | 5        | 5      | 5             | 5      | 5 | -1                | 0  | 0  | -1            | 0  | 0           | 4.5               | 5      | 5        | 4.5      | 5         | 5        | 4.75        | 0.33709993  |
|                      | GroundTruth  | Summary Accuracy                | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 5      | 5             | 5      | 5 | 0                 | 0  | 0  | 0             | 0  | 0           | 5                 | 5      | 5        | 5        | 5         | 5        |             |             |
|                      |              | Summary Clarity                 | 5      | 5        | 5      | 5                 | 5 | 5 | 4      | 4        | 4      | 5             | 4      | 5 | 1                 | 1  | 1  | 0             | 1  | 0           | 4.5               | 4.5    | 4.5      | 5        | 4.5       | 5        |             |             |
|                      |              | Segmentation Granularity        | 4      | 5        | 5      | 4                 | 4 | 5 | 4      | 5        | 5      | 5             | 5      | 5 | 0                 | 0  | 0  | -1            | -1 | 0           | 4                 | 5      | 5        | 4.5      | 4.5       | 5        |             |             |
|                      |              | Diff Faithfulness               | 5      | 5        | 5      | 5                 | 5 | 5 | 5      | 5        | 5      | 5             | 5      | 5 | 0                 | 0  | 0  | 0             | 0  | 0           | 5                 | 5      | 5        | 5        | 5         | 5        | 5           |             |
|                      | Summary Diff |                                 |        |          |        |                   |   |   |        |          |        |               |        |   |                   | -  |    |               |    |             |                   |        |          |          |           |          |             |             |
|                      |              |                                 |        |          |        |                   |   |   | -      |          |        |               | -      | - | U                 | -  | -  | -             | -  | _           |                   |        |          |          |           |          | -           | 1.18321595  |
|                      |              | Diff Completeness Diff Salience | 5<br>3 | 5        | 5<br>5 | 5<br>5            | 5 | 5 | 5<br>2 | 5        | 5<br>4 | 5<br>4        | 5<br>5 | 5 | 1                 | 0  | 0  | 0             | 0  | -1          | 2.5               | 5<br>5 | 5<br>4.5 | 5<br>4.5 | 5<br>5    | 5<br>2.5 |             | 5<br>4      |