Table 1: Accuracy of 3 Methods across 30 Examples

Example	**Method 1: AI-Powered Semantic Similarity**	**Method 2: LLM as a Judge**	**Method 3: Manual Human Evaluation**
1	0.15	0.05	0.00
2	0.88	0.80	0.85
3	0.85	0.75	0.80
4	0.99	1.00	1.00
5	0.60	0.40	0.30
6	0.05	0.10	0.00
7	0.90	0.90	0.95
8	0.98	1.00	1.00
9	0.75	0.70	0.75
10	0.99	1.00	1.00
11	0.80	0.85	0.90
12	0.70	0.60	0.70
13	0.50	0.40	0.50
14	0.99	1.00	1.00
15	0.05	0.00	0.00
16	0.99	1.00	1.00
17	0.40	0.10	0.10
18	0.65	0.50	0.60
19	0.99	1.00	1.00
20	0.30	0.05	0.00
21	0.80	0.70	0.75
22	0.70	0.80	0.85
23	0.10	0.00	0.00
24	0.75	0.75	0.80
25	0.85	0.85	0.90
26	0.10	0.00	0.00
27	0.70	0.65	0.70
28	0.60	0.40	0.50
29	0.99	1.00	1.00
30	0.99	1.00	1.00
Average	**0.675**	**0.640**	**0.655**

1 AI-Powered Semantic Similarity Scoring

This method uses pre-trained sentence embedding models to convert the ground truth and predicted answers into numerical vectors (embeddings). The cosine similarity between these vectors is then calculated, providing a quantitative measure of semantic closeness. A higher score (closer to 1) indicates greater contextual similarity.

2 LLM as a Judge

This approach utilizes another powerful LLM to evaluate the correctness of the predicted answer. A carefully designed prompt instructs the ”judge” LLM to compare the original question, ground truth, and predicted answer, assessing contextual and factual alignment. The output is a accuracy number.

3 Manual Human Evaluation

In this method, a human evaluator manually assesses the predicted answer’s accuracy by comparing it with the ground truth in the context of the original question. The evaluation considers factual correctness, completeness, and relevance. Each prediction is assigned a score based on a defined rubric or subjective judgment, allowing nuanced interpretation of answer quality that automated methods might miss.