Report2

David Zhu rz2718

August 2025

1 Understanding Level Evaluation Report

1.1 Objective

This experiment evaluates two approaches for determining a student's understanding level in a Socratic AI tutoring system:

1. **Version 1 (V1)**: GPT provides the understanding score directly based on the student's answer.
2. **Version 2 (V2)**: GPT first generates a reference answer, then the student's answer is compared against it in terms of accuracy, completeness, and reasoning.

1.2 Methodology

V1: Direct GPT Scoring The prompt included the question, the concept, and the student's answer, asking GPT to return a JSON object containing `"understanding_score"` in the range [0, 1]. No reference answer was used.

V2: Reference Answer Comparison This method:

1. Generates a high-quality reference answer using GPT.
2. Compares the student's answer to the reference along:
   - **Accuracy**
   - **Completeness**
   - **Reasoning**
3. Produces an `"understanding_level"` score and qualitative feedback.

Question	Student Answer	V1 Score	V2 Score
What is gradient descent?	"11111111"	0.0	0.0
What is gradient descent?	"Gradient descent is gradient in calculus"	0.1	0.0
What is gradient descent?	"Gradient descent in ML/DL is the iterative process of tuning parameters by moving opposite to the gradient."	0.9	0.9
What is SuperGLUE?	"SuperGLUE consists of a diverse set of tasks that require advanced understanding."	0.8 (too high)	0.5 (correctly penalized for missing

Table 1: Comparison of V1 direct GPT scoring and V2 reference-answer comparison.

1.3 Test Cases and Results

1.4 Analysis

- V1 often overestimates understanding when answers contain partial or related keywords.
- V2 penalizes incomplete answers appropriately, especially when key reference points are missing.
- Example: For the SuperGLUE question, V1 returned 0.8 for an incomplete answer, whereas V2 correctly assigned 0.5 and identified missing elements.

1.5 Conclusion

The reference-answer comparison (V2) provides more reliable and objective understanding level scores, particularly in distinguishing partial knowledge from full mastery. For production, V2 is recommended, potentially combined with embedding-based semantic similarity for further robustness.

A Full JSON Outputs from Colab Tests

A.1 Version 1 (Direct GPT Scoring)


```
{'understanding_score': 0.0}
{'understanding_score': 0.1}
{'understanding_score': 0.9}
{
 "understanding_score": 0.8
}
```


A.2 Version 2 (Reference Answer Comparison)


```
{
 "understanding_level": 0.5,
 "response_quality": "low",
 "key_points_covered": [
   "diverse set of tasks",
   "require advanced language understanding"
 ],
 "missing_elements": [
   "benchmark purpose",
   "performance metrics",
   "continual improvement"
 ],
 "suggested_follow_up": "Can you provide more details on how SuperGLUE evaluates the perform
 "feedback": "The student's answer partially matches the reference by mentioning the diverse
 "grasp_adjustment": 0.0
}

{
 "understanding_level": 0.5,
 "response_quality": "low",
 "key_points_covered": [
   "tasks are diverse"
 ],
 "missing_elements": [
   "definition of SuperGLUE",
   "increased difficulty",
   "higher benchmark scores",
   "continuous improvement"
 ],
 "suggested_follow_up": "Can you provide more details about the purpose and evolution of Sup
 "feedback": "The student's answer only partially matches the reference answer. They correct
 "grasp_adjustment": 0.0
}
```