WESTLAKE UNIVERSITY | SCHOOL OF ENGINEERING

Weekly Report

Wang Tianyi

October 29, 2024

Outline

Outline

Logical Evaluation of Reasoning Paths

- Define the reasoning chain $T = \{T_1, T_2, \ldots, T_m\}$, where each T_i represents an inference step.
- Use a Natural Language Inference (NLI) model to evaluate the logical relationship between adjacent steps T_i and T_{i+1}.
- Possible relations:

$$\mathrm{NLI}(T_i, T_{i+1}) = \begin{cases} \text{Entailment} & \text{Reasoning logic is rational} \\ \text{Neutral} & \text{Reasoning logic is uncertain} \\ \text{Contradiction} & \text{Reasoning logic contradicts} \end{cases}$$

```
> (base) wangtianyi@wangtianyideMacBook-Air rag_gpu_v1.3 % python test.py
 Please enter your research question (or type 'exit' to quit): Question: In an experiment observing ligand binding patterns on Gold Nano-Bipyramids (Au
 NBPs) with low ligand coverage, what was the predominant observation regarding the ligand binding behavior?\n\nGiven options:\nA. Ligands bind predomi
 nantly to sharp tips due to high surface energy, leading to growth redirection.\nB. Ligands bind indiscriminately across the entire surface due to uni
 form surface energy, leading to isotropic growth.\nC. Ligands fail to bind to the bipyramids as the surface energy is too low to facilitate attachment
 , resulting in no growth change.\nD. Ligands bind preferentially to the base of the bipyramids due to low surface energy, resulting in elongated growt
 h.
 Generated Query: ligand binding patterns Gold Nano-Bipyramids low ligand coverage predominant observation behavior

 Consolidated Reasoning Chain:
 - Reasoning Step: The consolidated context consistently indicates that in an experiment with Gold Nano-Bipyramids (AuNBPs) and low ligand coverage, li
 gands bind predominantly to sharp tips due to high surface energy, leading to growth redirection.
 - Reasoning Step: This observation suggests that the ligand binding behavior is influenced by the high surface energy present at the sharp tips of the
  Gold Nano-Bipyramids (AuNBPs).
 - Reasoning Step: The ligands' preference for binding to the sharp tips rather than the base or other areas implies a specific affinity for regions wi
 th high surface energy.
 - Reasoning Step: The consistent pattern of ligands binding to sharp tips aligns with the concept of ligands being attracted to sites with increased s
 urface energy for growth redirection.
 - Reasoning Step: Therefore, based on the evidence provided in the articles, the predominant observation regarding ligand binding behavior on Gold Nan
 o-Bipyramids (AuNBPs) with low ligand coverage is that ligands bind predominantly to sharp tips due to high surface energy, leading to growth redirect
 ion.
```

NLI Model Judgment and Scoring

Scoring System

Each pair (T_i, T_{i+1}) receives a score based on the NLI result:

$$S(T_i, T_{i+1}) = \begin{cases} 1 & \text{if Entailment} \\ 0 & \text{if Neutral} \\ -1 & \text{if Contradiction} \end{cases}$$

Average Logic Score L

The overall logical consistency of the reasoning chain is given by:

$$L = \frac{1}{m-1} \sum_{i=1}^{m-1} S(T_i, T_{i+1})$$

where $L \approx 1$ indicates high logical consistency, and $L \approx -1$ suggests logical contradictions.

Interpretation of the Logic Score

- $L \approx 1$: The reasoning chain is coherent and logically consistent.
- $L \approx 0$: Logic is uncertain across steps, with possible logical jumps.
- $L \approx -1$: Significant logical contradictions are present in the reasoning chain.

Note: In the actual code, I set a threshold where $L \geq 0.5$ is considered *reasonable*, and $L \leq -0.5$ is considered *contradiction*.

```
Logical Consistency Scores for Each Step Pair:
Step 1 to Step 2 - Score: 1
Step 2 to Step 3 - Score: -1
Step 3 to Step 4 - Score: 1
Step 4 to Step 5 - Score: 1

Final Average Logical Consistency Score: 0.50
```

Summary Standardizer

- In initial testing, it was observed that GPT sometimes generates final summaries that do not strictly follow the prompt-imposed format.
- This inconsistency in format led to issues in post-processing, particularly in applying regex-based answer extraction.
- As a result, benchmark tests were affected since the system could not reliably match answers in improperly formatted summaries.

Standardization Details

Standardization Criteria

The Standardizer focuses on specific criteria, including:

- **Consistency in terminology:** Ensuring uniform terminology across the summary.
- **Structural clarity:**Adjusting sentence order and paragraph breaks to improve readability.
- **Formatted correctly:**Especially Ans: such a clear answer to the part

Outcome of Summary Format Correction

- After implementing the Standardizer mechanism, the consistency of summary format significantly improved.
- This allowed the regex patterns to correctly match and extract answers, leading to more reliable benchmark test results.
- Overall, the Standardizer step enhanced the robustness and accuracy of the system's performance in tests.

```
(base) wangtianyi@wangtianyideMacBook-Air rag_gpu_v1.3 % python test_standardize_summary.py
Original Summary:

   Literature Summary: Studies on the growth of AuNBPs (Gold Nano-Bipyramids) under varying conditions suggest that ligand binding behavior is influe
nced by surface energy. The summaries indicate that binding occurs more prominently at high-energy tips, although some abstracts do not confirm this e
ffect under all conditions. Observations across experiments lack direct evidence for specific structural changes under low ligand concentrations. None
theless, the consensus supports a binding preference at the tips of the bipyramids.

Ans: The correct option seems to be closest to B, given the evidence presented, though it's based on general trends rather than specific confirmation.

References:
1. No citation available for some referenced studies
2. Some studies referenced were inconclusive


Standardized Summary:
 Literature Summary: Studies on the growth of AuNBPs (Gold Nano-Bipyramids) under varying conditions suggest that ligand binding behavior is influence
d by surface energy. The summaries indicate that binding occurs more prominently at high-energy tips, although some abstracts do not confirm this effe
ct under all conditions. Observations across experiments lack direct evidence for specific structural changes under low ligand concentrations. Nonethe
less, the consensus supports a binding preference at the tips of the bipyramids.

Ans: B, given the evidence presented, though it's based on general trends rather than specific confirmation.

References:
1. No citation available for some referenced studies
2. Some studies referenced were inconclusive
```

Outline

Benchmark Testing Automation

1. Developed an automated script to run benchmark tests.
2. Implemented rules to validate answer formats for consistency.
3. Recorded accuracy metrics in real-time.

- **Challenge**: Encountered issues with inconsistent output formats.
- **Solution**: Enhanced prompt design and used regex for answer extraction.

```
Progress: 95/775 (12.26%)
Final Accuracy: 51.58% (49/95)
[96/775] Question 96 - Incorrect
Logical Consistency of Reasoning Chain: reasonable
Progress: 96/775 (12.39%)
Final Accuracy: 51.04% (49/96)
[97/775] Question 97 - Incorrect
Logical Consistency of Reasoning Chain: reasonable
Progress: 97/775 (12.52%)
Final Accuracy: 50.52% (49/97)
[98/775] Question 98 - Correct
Logical Consistency of Reasoning Chain: neutral
Progress: 98/775 (12.65%)
Final Accuracy: 51.02% (50/98)
[99/775] Question 99 - Incorrect
Logical Consistency of Reasoning Chain: neutral
Progress: 99/775 (12.77%)
Final Accuracy: 50.51% (50/99)
[100/775] Question 100 - Incorrect
Logical Consistency of Reasoning Chain: neutral
Progress: 100/775 (12.90%)
Final Accuracy: 50.00% (50/100)
```

Logging and Progress Tracking

Key Features

1. Added logging for each answer generation step.
2. Implemented progress tracking to avoid data loss.
3. Enhanced fault tolerance for long-running experiments.