**Task: Definition Model**

**Prompt:**

I used a language model to define certain word in a sentence. Here are the hypothesis that the model generated, target which is the true definition of the word and sentence which is the sentence on which the word we want to define is:

hypothesis: <hyp>

target: <tgt>

sentence: <src>

Can you determine whether the language model has hallucinated or not? answer it with just a number between 0 and 1 where 1 means it is a complete hallucination and 0 means it is not. don't write anything else.

Results:

|  |  |
| --- | --- |
| P(hallucination) | ChatGPT guess |
| 0.4 | 0.7 |
| 0.8 | 1 |
| 0 | 0.1 |
| 0.2 | 0.8 |
| 1 | 0.7 |

Prompt:

In a Definition Model task, the hypotheses include the language model's prediction, the target is the true definition that we want and the sentence, which is the context in which the word to be defined appears:

hypothesis: <hyp>

target: <tgt>

sentence: <src>

Please provide a numerical score between 0 and 1, where 1 indicates a complete hallucination, and 0 signifies no hallucination. Do not include any additional information beyond the numerical score.

Results:

|  |  |
| --- | --- |
| P(hallucination) | ChatGPT guess |
| 0.4 | 0.8 |
| 0.8 | 0.9 |
| 0 | 1 |
| 0.2 | 0.85 |
| 1 | 0.65 |

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Task: Machine Translation**

**Prompt:**

I used a language model for Machine Translation task. Here are the hypothesis that the model generated, target which is the true translation and sentence which is the sentence that we want to translate:

hypothesis: <hyp>

target: <tgt>

sentence: <src>

Can you determine whether the language model has hallucinated or not? answer it with just a number between 0 and 1 where 1 means it is a complete hallucination and 0 means it is not. don't write anything else.

Results:

|  |  |
| --- | --- |
| P(hallucination) | ChatGPT guess |
| 0.8 | 0.9 |
| 1 | 0.8 |
| 0 | 0.2 |
| 0.6 | 0.9 |
| 0.4 | 0.9 |

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Task: Paraphrase Generating**

**NOTE: Since training data for this task don’t have target, we should only use src and hyp in our prompt.**

**Prompt:**

I used a language model for Paraphrase Generating task. Here are the hypothesis that the model generated and sentence which is the source input for the language model:

hypothesis: <hyp>

sentence: <src>

Can you determine whether the language model has hallucinated or not? answer it with just a number between 0 and 1 where 1 means it is a complete hallucination and 0 means it is not. Don't write anything else.

Results:

|  |  |
| --- | --- |
| P(hallucination) | ChatGPT guess |
| 0.2 | 0.1 |
| 0.6 | 0.8 |
| 0 | 0.1 |
| 0.8 | 0.5 |
| 1 | 0.25 |