Mini Project 3

Robert Richter

11.02.2025

Step 1: Data Preparation

- Similar to Task 2
- Usage of Decision Tree Classifier (from *sklearn*)
- Word count for explanation complexity

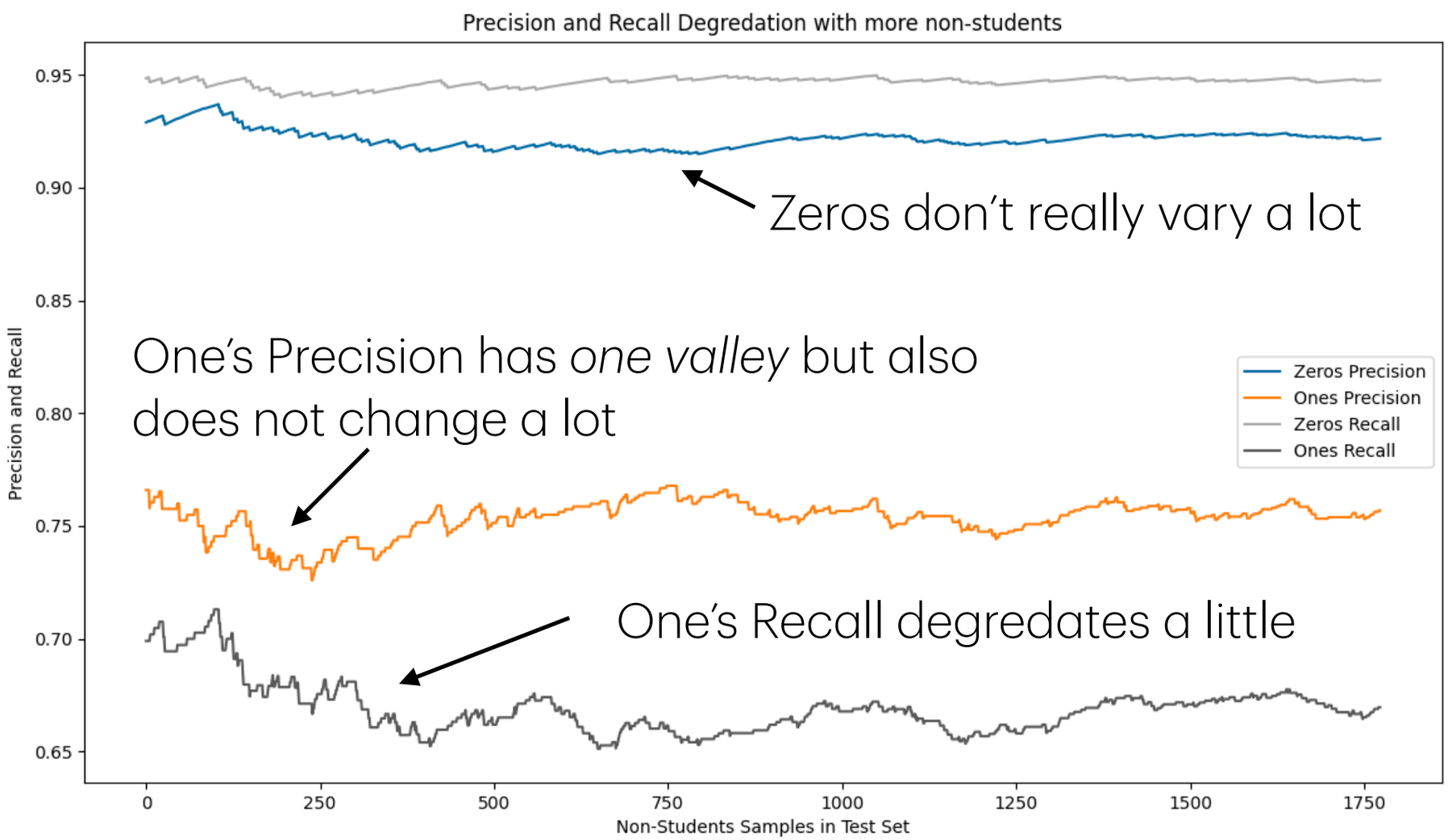
Attempt 1: Training on purely Graduates

Recall and Precision Results

- Training Set: Graduates
- Testing Set: Under-Graduates

	precision	recall	f1-score	support
0	0.93	0.95	0.94	427
1	0.77	0.72	0.74	103
accuracy			0.90	530
macro avg	0.85	0.83	0.84	530
weighted avg	0.90	0.90	0.90	530

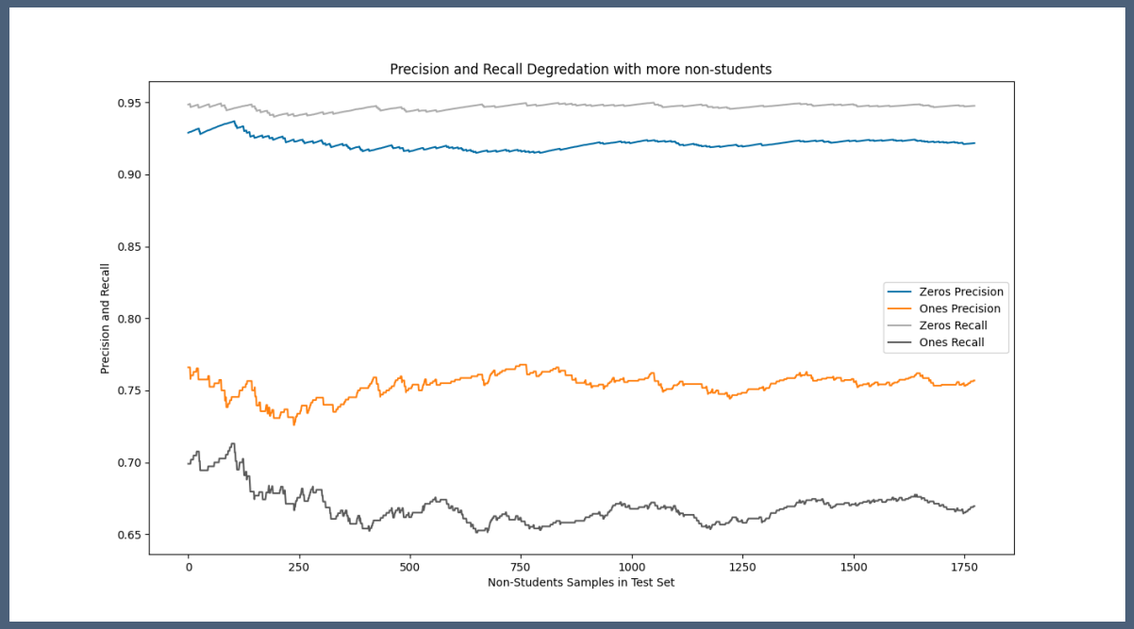
Attempt 1: Degradation with purely Graduates



Question 1.1

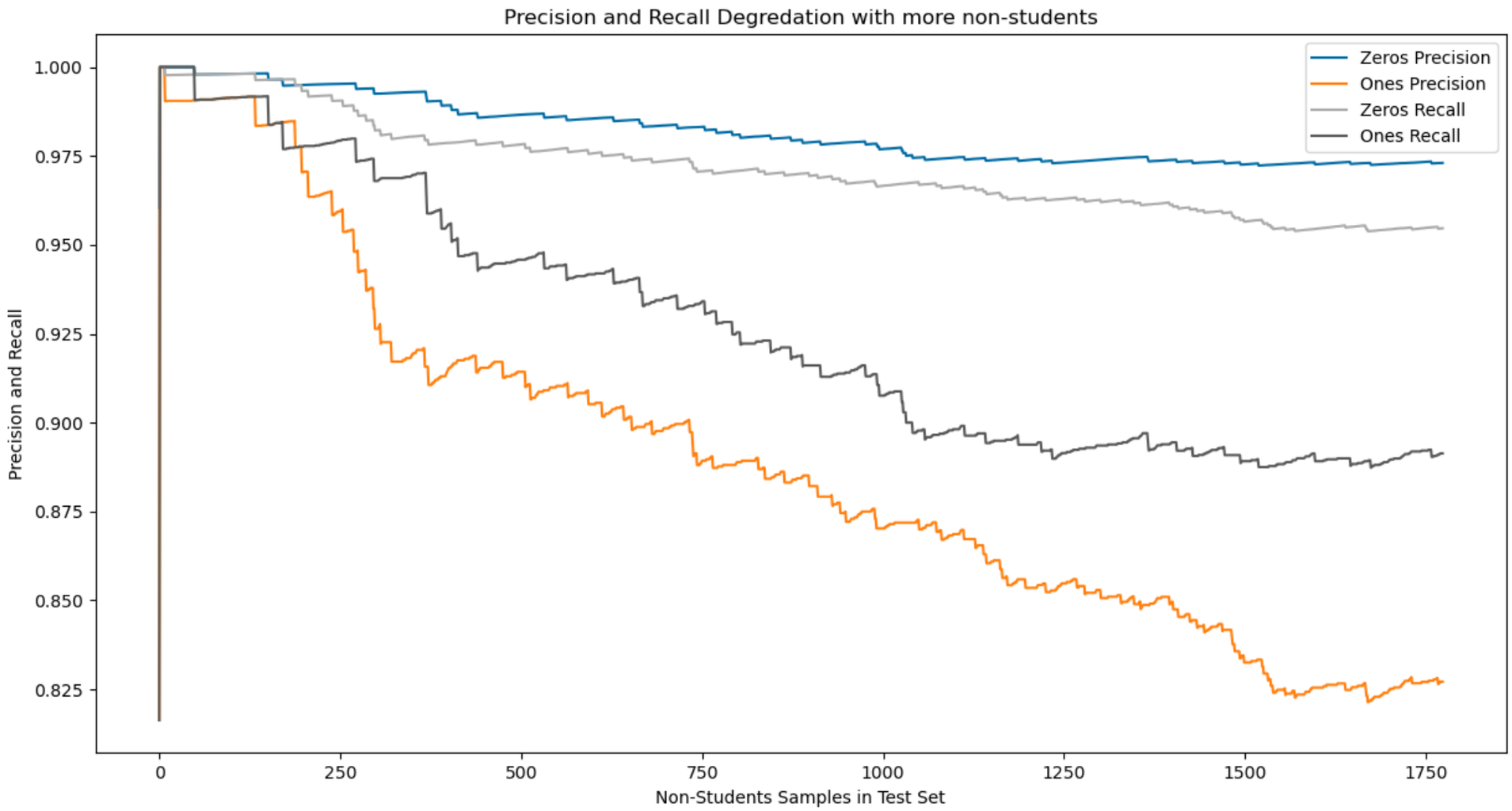
For the impact of 5% and 10% loss on precision and recall, what is the min-number of non-students added on average to the holdout set?

- Zero are stable in precision and recall (i.e., no degradation)
- One's Precision highest degradation about 4% at 250 additional samples
- One's Recall highest degradation about 7% between 100 samples and 400 samples
-



Attempt 2: Training on Under-Graduates

Recall and Precision Results

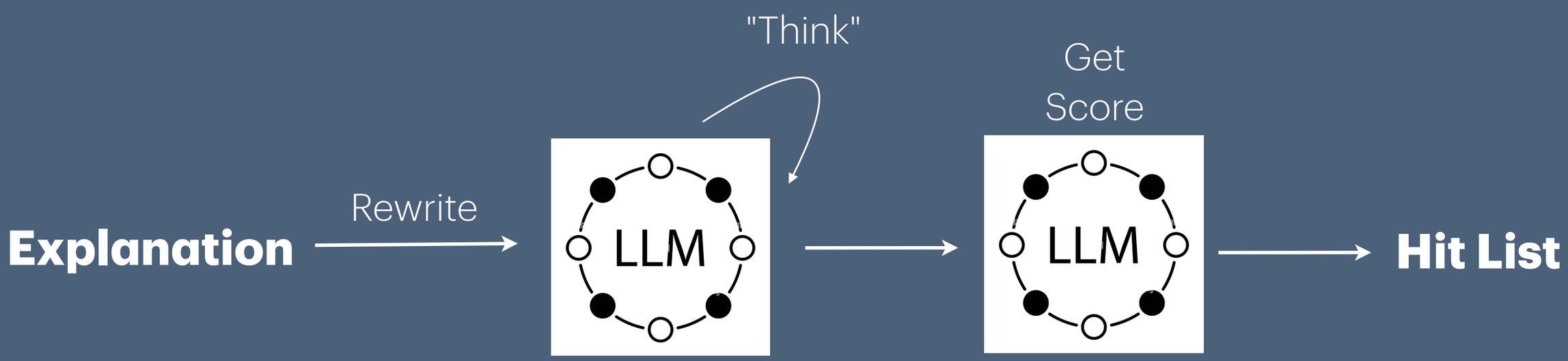


Task 2: Necessary and Sufficient Explanations

How to find the value of an answer?

Metric?

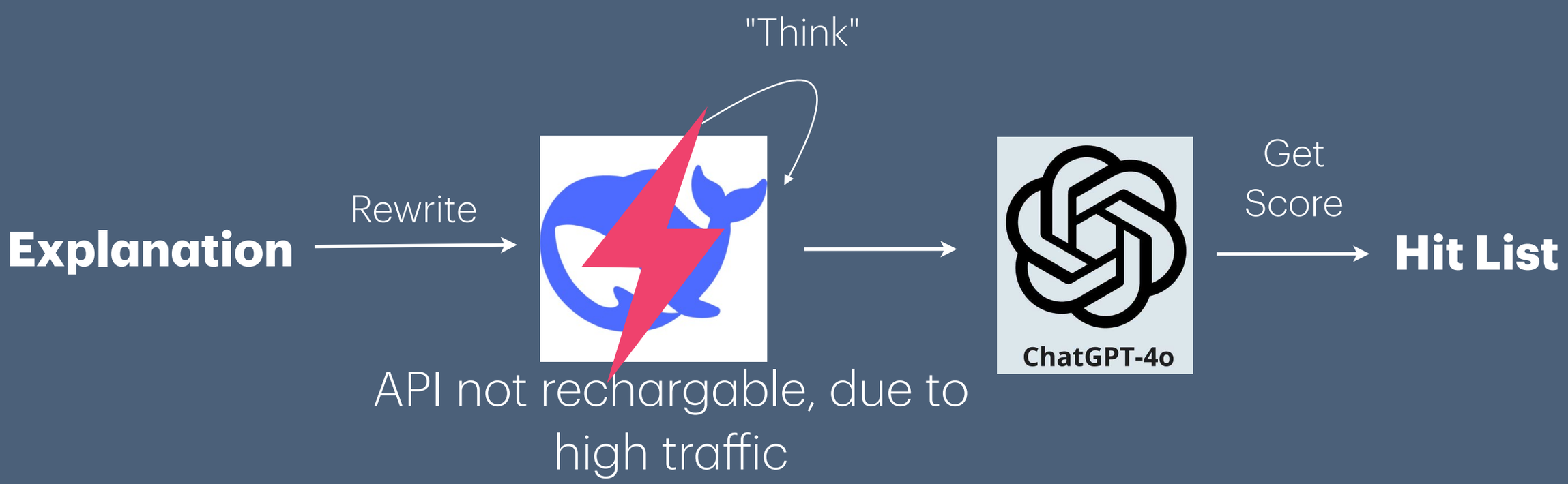
- There are a bunch of metrics out there
- But we need to consider too many variables to draw a conclusion on how good an explanation is
- Only humans can do this (by thinking about it and talking)
- Here: LLMs instead of Humans
- Scores are values between 0 and 100 (0: useless explanation, 100: perfect explanation)



How to find the value of an answer?

Metric?

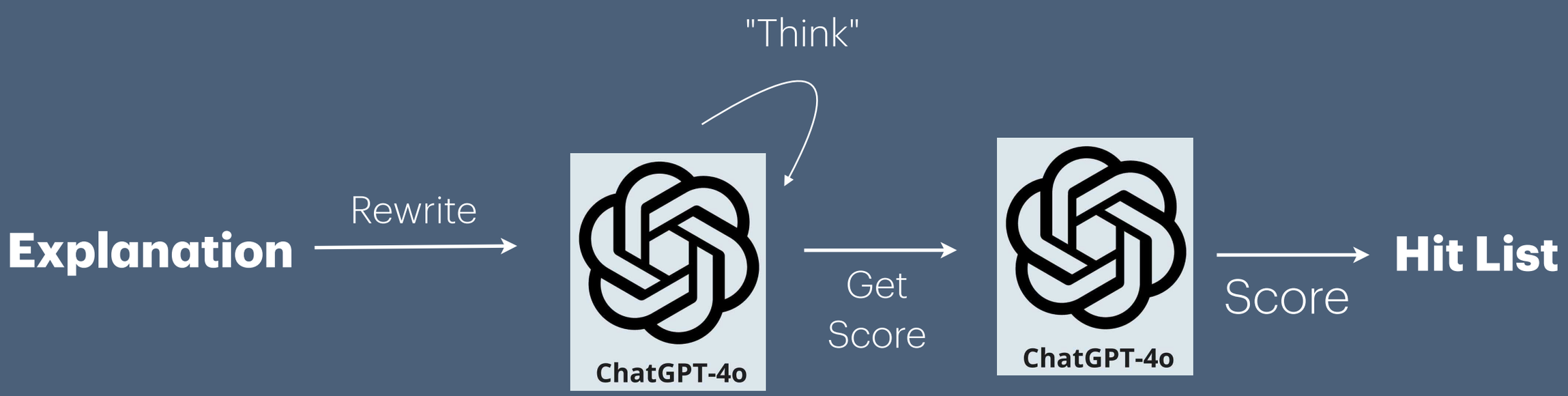
- There are a bunch of metrics out there
- But we need to consider too many variables to draw a conclusion on how good an explanation is
- Only humans can do this (by thinking about it and talking)
- Here: LLMs instead of Humans
- Scores are values between 0 and 100 (0: bad explanation, 100: perfect explanation)



How to find the value of an answer?

Metric to measure quality

- There are a bunch of metrics out there
- But we need to consider too many variables to draw a conclusion on how good an explanation is ==> But humans can do this (by thinking about it and talking)
- Here: LLMs instead of Human
- Idea: 1. Rewrite explanation ("Thinking") 2. Score explanation between 0 and 100 (0: bad explanation, 100: perfect explanation) ==> Ground Truth



Question 2: Results

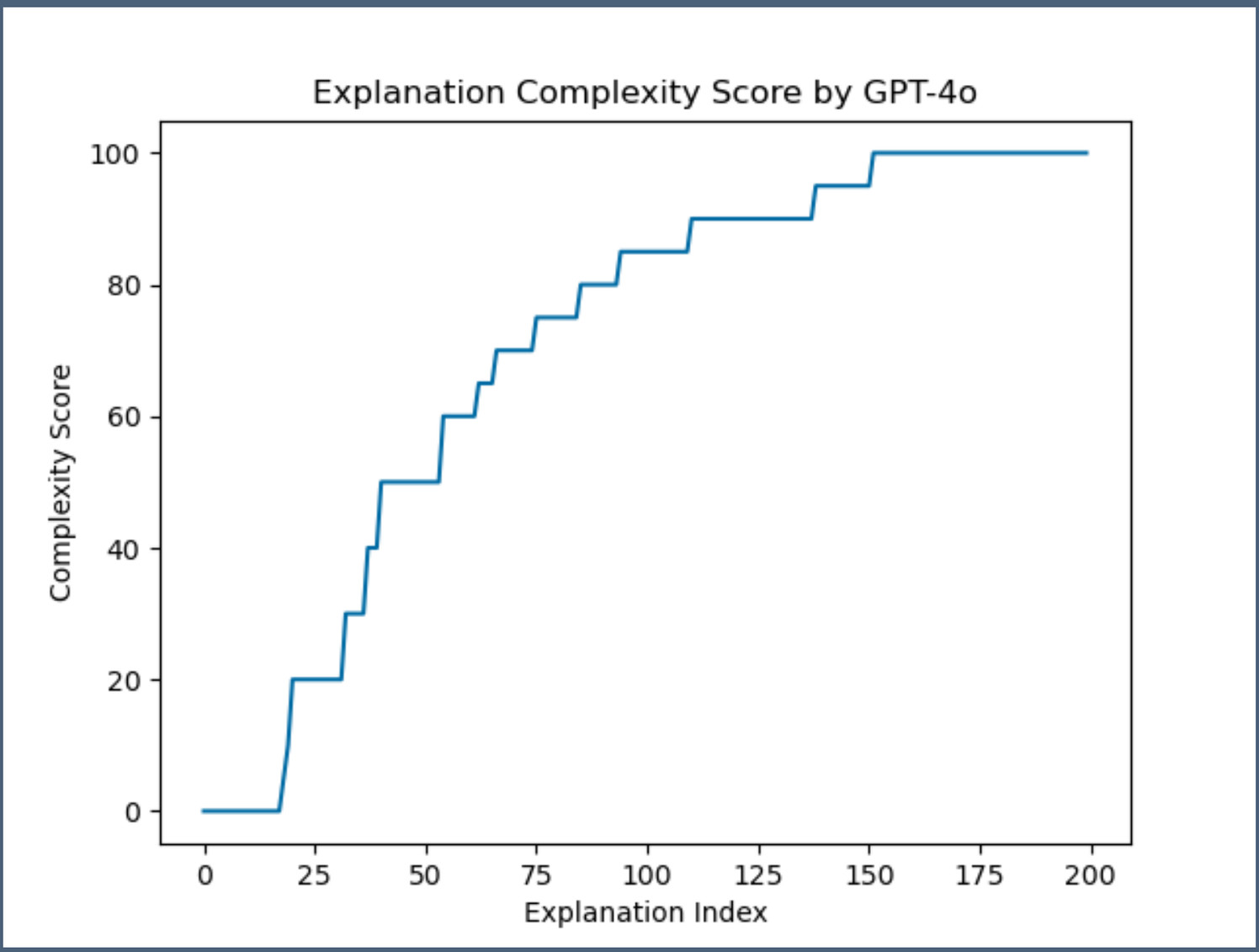
- Very verbose intermediate results (wouldn't humans be too?)

The time should not be negative value

The time should not be negative value --> The bug report indicates an issue where a time-related value, which presumably should always be positive or zero, is instead being allowed to be negative. This could indicate a logic or validation error in the code where time inputs or calculations are not being properly constrained to non-negative values. It's important to check where this time value is being set or calculated and ensure that appropriate checks and validation mechanisms are in place to prevent it from becoming negative.

Question 2: Results

Scoring of individual Bug Reports (only for HIT01_8)



HIT01_8

- quality of bug reports distributed in radical function

Question 2: Merged Bug Reports

Scores of merged Explanations

- Merging started with the Top-Rated Bug Reports
- Score remained steady at 90-100 for merging 2 to 50 bug reports
- Outcome of merging the Top-50 Bug Reports:

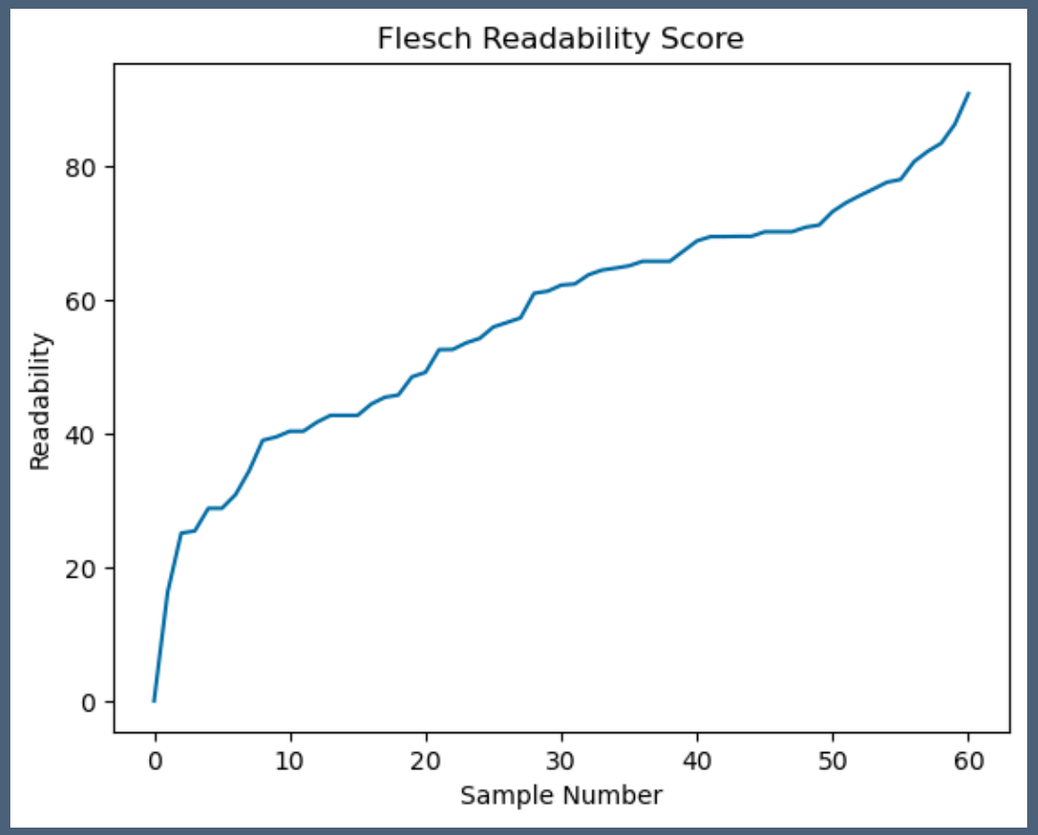
The predominant issue across these bug reports is the occurrence of an `IllegalArgumentException` being thrown due to an out-of-range `minutesOffset` value in a method that presumably handles time zone offset settings, such as `DateTimeZone.forOffsetHoursMinutes()`. The exception arises on line 280 because the code on line 279 checks if `minutesOffset` is less than 0 or greater than 59. As a key detail, passing a negative integer (e.g., -15) for `minutesOffset` triggers this exception. Various explanations suggest that based on the comments or documentation for the method, inputs for `minutesOffset` should allow negative values up to -59 if accompanied by negative hours, suggesting a mismatch between the method's implementation and its intended specification. Furthermore, numerous explanations highlight this inconsistency and suggest modifying the conditional check to accommodate the correct range of valid input values for `minutesOffset`. Additionally, there is mention of `NullPointerException` and `StringIndexOutOfBoundsException` in explanations, but these relate to different contexts or code portions not directly linked to the `minutesOffset` issue.

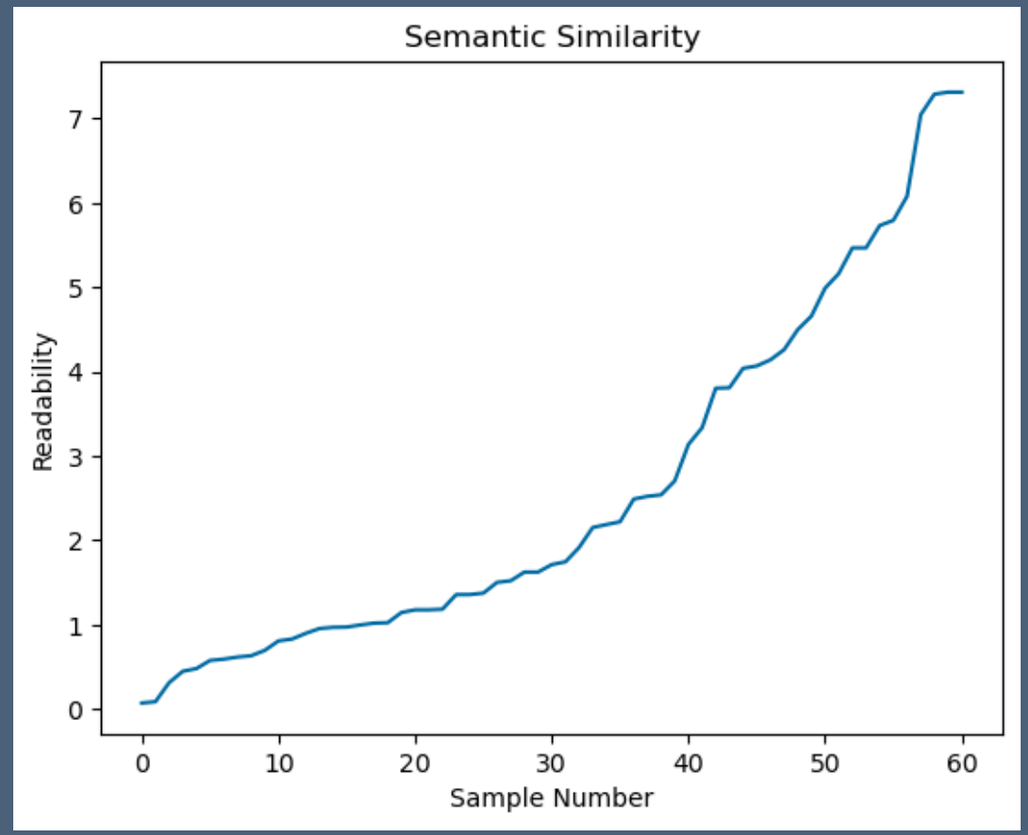
Question 3: Diversity?

- More numerical approach
- Usage of Entropy per Feature and Jaccard Similarity

```
                                entropy   jaccard
Answer.ID                      7.662591  1.000000
FailingMethod                  7.668480  0.000000
Question.ID                    7.658442  0.808081
Answer.duration                7.024122  1.000000
Answer.confidence              7.683782  0.555960
Answer.difficulty              7.759347  0.754747
Answer.option                  6.824738  0.491313
Answer.order                   7.765796  0.457172
Code.LOC                       7.339980  0.484848
Code.complexity                7.588799  0.565657
Worker.score                   7.835589  0.626263
Worker.yearsOfExperience       7.462690  0.936566
Worker.age                     7.815515  0.956970
Worker.gender                  6.442120  0.368485
Worker.whereLearnedToCode      7.677123  0.794747
Worker.country                 7.067054  0.788889
Worker.programmingLanguage     7.625334  0.689091
Answer.explanationComplexity   7.494021  0.975152
Mean entropy: 7.4664179773076995
Mean jaccard: 0.6807744107744108
```

Best Readability + Semantic Similarity to Ground Thruth (only for HIT01_8 due to computational constraints)





- Semantic Similarity using *BAAI/bge-reranker-large*
- Sweetspot (argmax(3 * similarity + 0.05 * flesch):

The value of minutes offset does not have valid argument as a result this method will not be called as and argument exception will be displayed.

How to achieve diverse datasets?

Formulas

$$X : \text{All Bug Descriptions}$$
$$T(X) = \{(f, s) | f = \text{flesch}(x), s = \text{similarity}(x), x \in X\}$$
$$\forall T' \subseteq T(X) : \text{argmax}(\text{jaccard\_diversity}(T'))$$

How to calculate this? Naive: Calculate for all subsets (2**n subsets)

Therefore: Use vastly different values from curve of readability and semantic similarity. Check those for their results on Jaccard Diversity.