Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

ASE Project 2

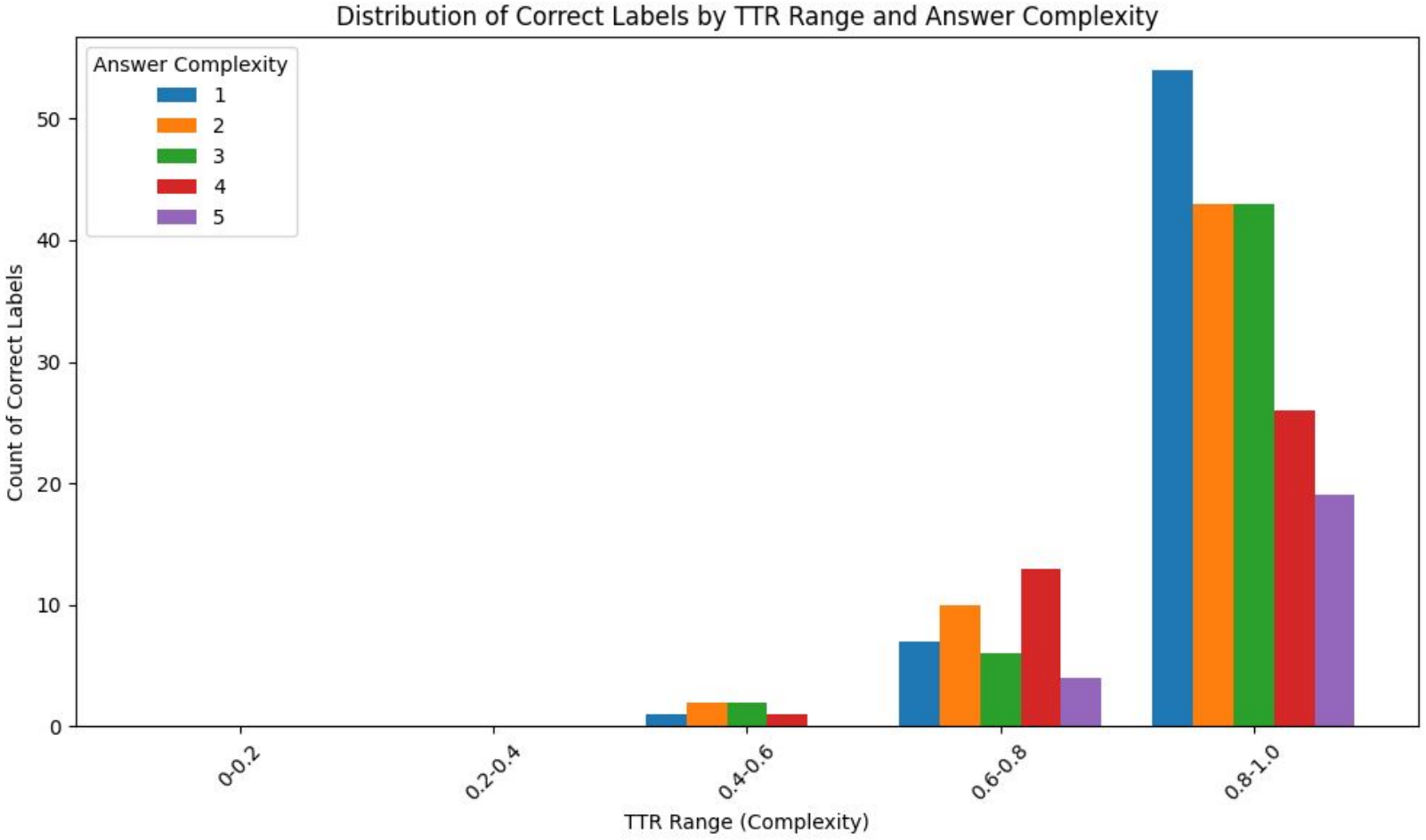
Lamin & Ibrahim
10.12.2024

Task 1

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Decision Tree Performance.

```
Cross-validated Precision: [0.2962963  0.125      0.2        0.34482759 0.26923077]
Cross-validated Recall: [0.0952381  0.0952381  0.07142857 0.0952381  0.08333333]
Accuracy score:  0.73125
-----
FailingMethod: HIT01_8
Precision: 0.18
Recall: 0.05
-----
FailingMethod: HIT02_24
Precision: 0.60
Recall: 0.07
-----
Length of holdout set:  320
Holdout Precision: 0.31
Holdout Recall: 0.06
```


Distribution of Correct Labels by TTR Range and Answer Complexity
Answer Complexity
1
2
3
4
5
Count of Correct Labels
0
10
20
30
40
50
0-0.2
0.2-0.4
0.4-0.6
0.6-0.8
0.8-1.0
TTR Range (Complexity)

Prompting

Self-Made Prompt

- Basic instructions were provided, as outlined in the project slides.
- Additional prompting was required to achieve a result.

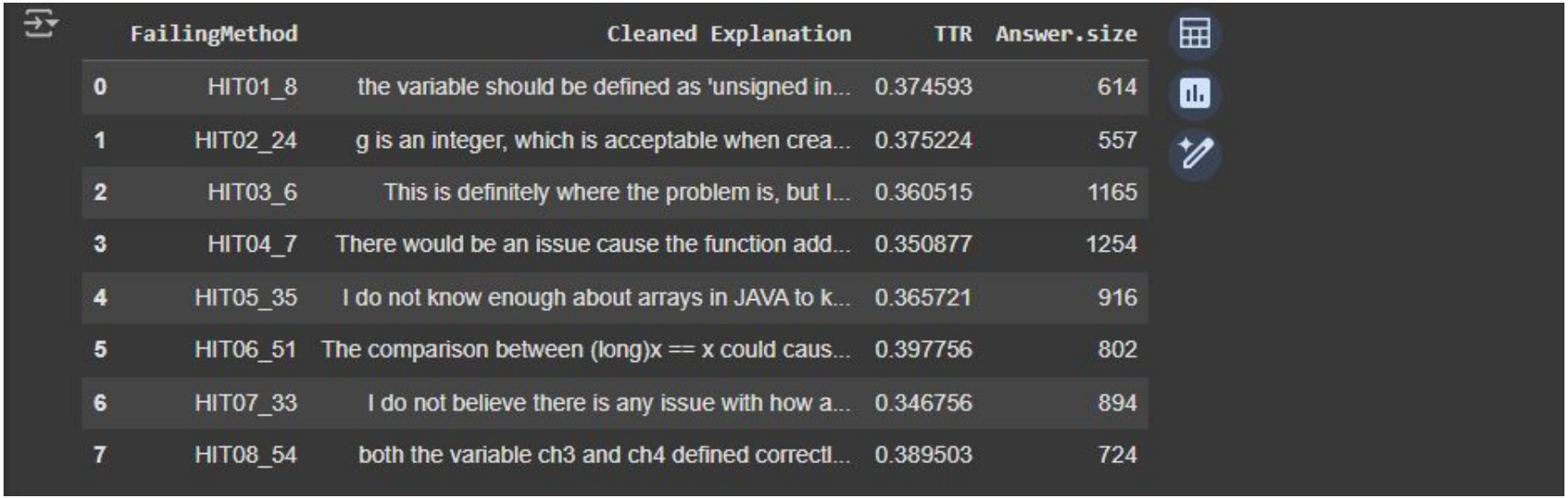
LLM-Generated Prompt

- Included specific commands to eliminate redundant information.
- Required less additional prompting compared to the first approach.

LLM-Generated Prompt w/ Roleplay

- The LLM assumed the role of a Software Engineer/Data Scientist working in a company.
- The effort required was similar to the second approach.

LLM Prompting Approach 1



	FailingMethod	Cleaned Explanation	TTR	Answer.size
0	HIT01_8	the variable should be defined as 'unsigned in...	0.374593	614
1	HIT02_24	g is an integer, which is acceptable when crea...	0.375224	557
2	HIT03_6	This is definitely where the problem is, but I...	0.360515	1165
3	HIT04_7	There would be an issue cause the function add...	0.350877	1254
4	HIT05_35	I do not know enough about arrays in JAVA to k...	0.365721	916
5	HIT06_51	The comparison between (long)x == x could caus...	0.397756	802
6	HIT07_33	I do not believe there is any issue with how a...	0.346756	894
7	HIT08_54	both the variable ch3 and ch4 defined correctl...	0.389503	724

Metrics / Failing Method

	FailingMethod	ROGUE1	ROGUE2	ROGUEL	ROGUELsum	BLEU
0	HIT01_8	0.197467	0.188434	0.115622	0.115622	0.000187
1	HIT02_24	0.299141	0.283718	0.163802	0.163802	0.010667
2	HIT03_6	0.223827	0.213549	0.112195	0.112195	0.000646
3	HIT04_7	0.143587	0.135971	0.078544	0.078544	0.000005
4	HIT05_35	0.301356	0.287069	0.150678	0.150678	0.009524

Specification 1

Data Quality

- Validation of integrity by checking for outliers and duplicates.
- Possible Over-fitting due to data leakage.
- Could be improved with multiple column combination

Demographic change adaptation

- Static models struggle with evolving user demographics or bug types

Solutions:

- Periodically retrain classifiers with updated data.
- Utilize active learning to identify and integrate new patterns.
- Monitor performance metrics across demographic subgroups.

Specification 2

Explanation Quality Estimate

- It is important to measure the clarity, relevance and correctness of the LLM-Generated explanation (consolation content)

Clarity,correctness could be cross checked with:

- Human reviews incorporated.
- Usage of standard metric evaluation ie. BLUE & ROUGE-L to ensure quality
- Collection of feedback for continuous improvement