Mini project 2

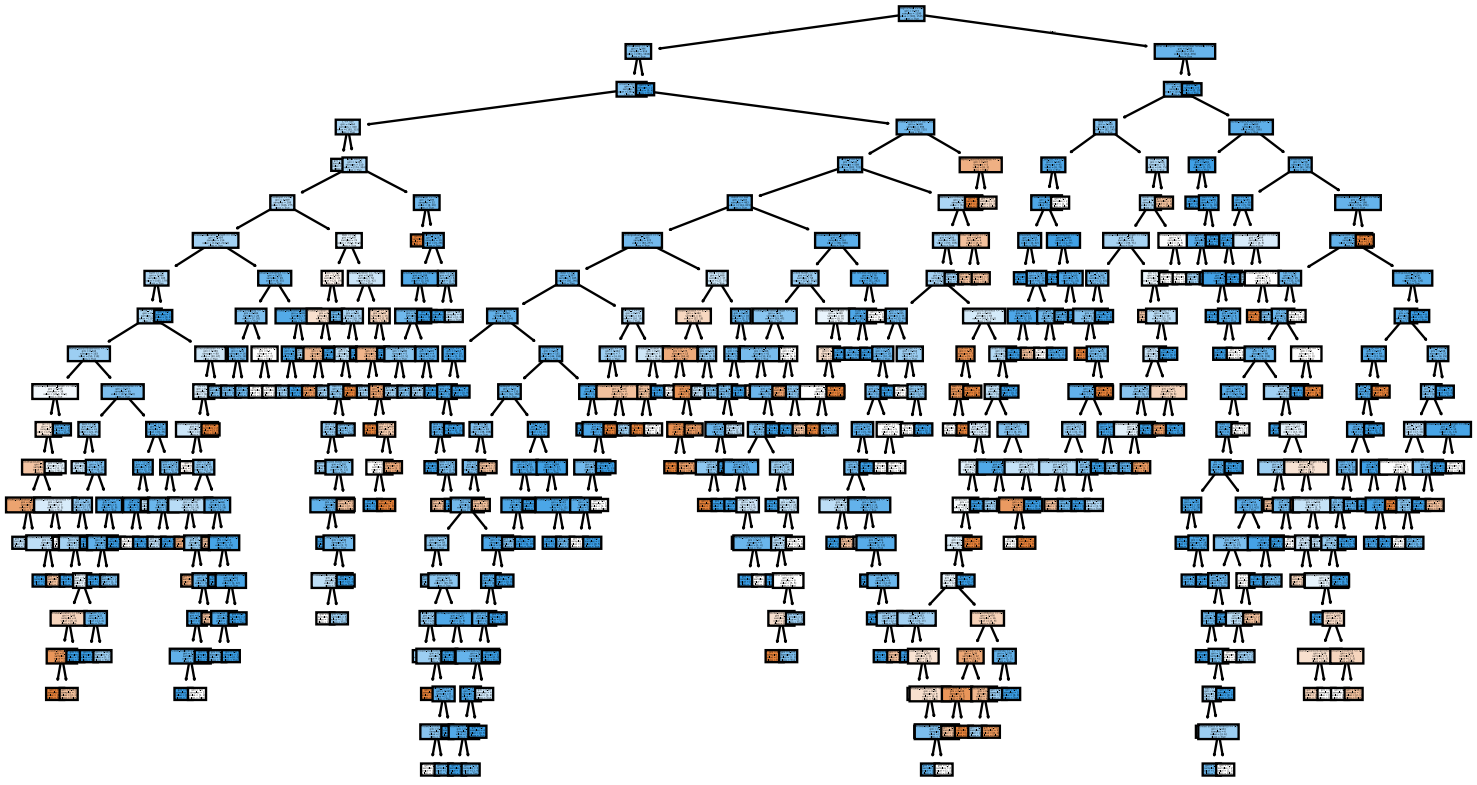
Consolidated bug reports

Henok Lachmann and Noel Bastubbe

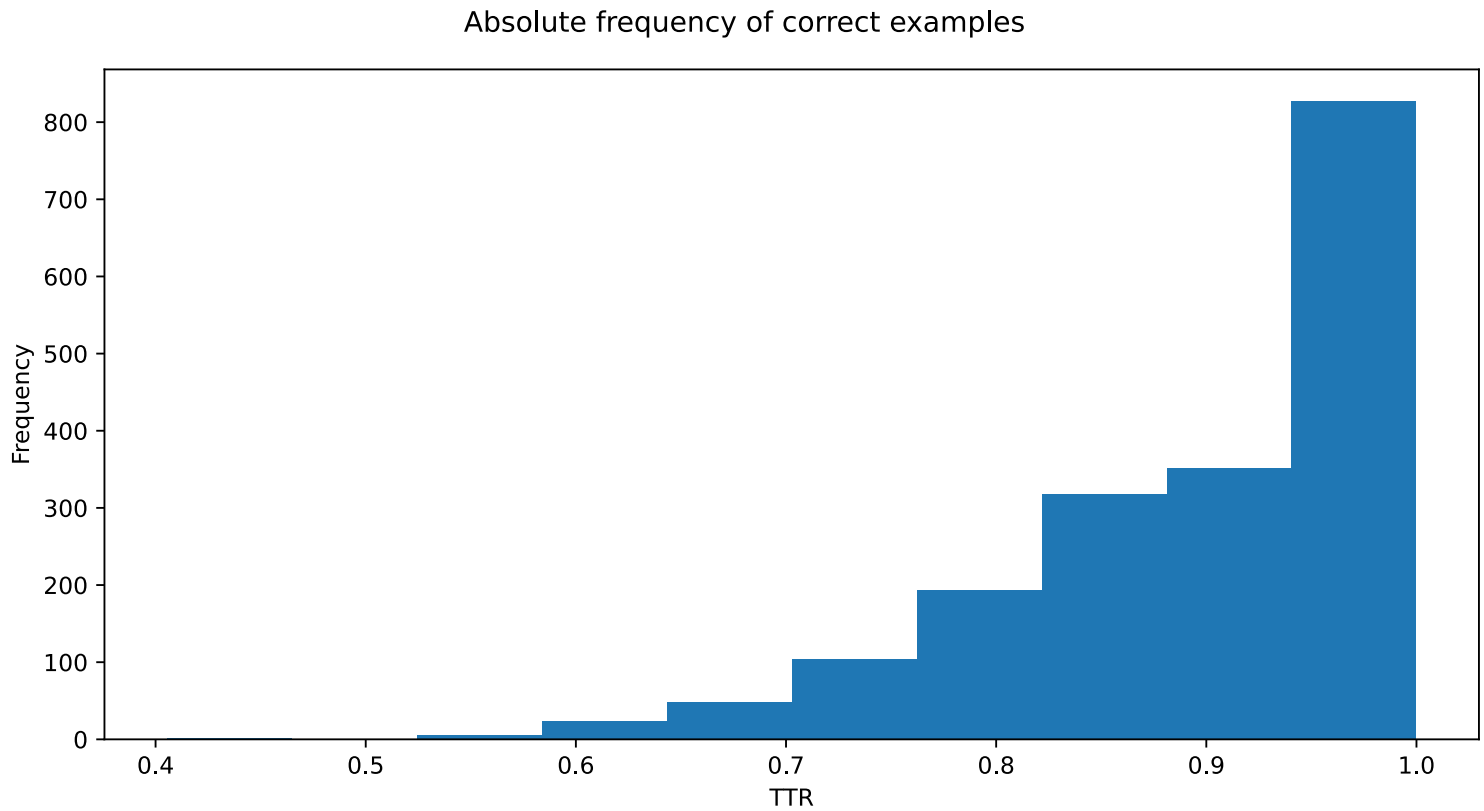
Train model

- first decision tree model, then random forest
- 5-fold cross validation on parameter grid
  - max_depth
  - min_samples_split
  - min_samples_leaf
  - max_features
  - n_estimators
- one hot encoding for categorical and multi features
- Best overall *f*1 score: 0.8421

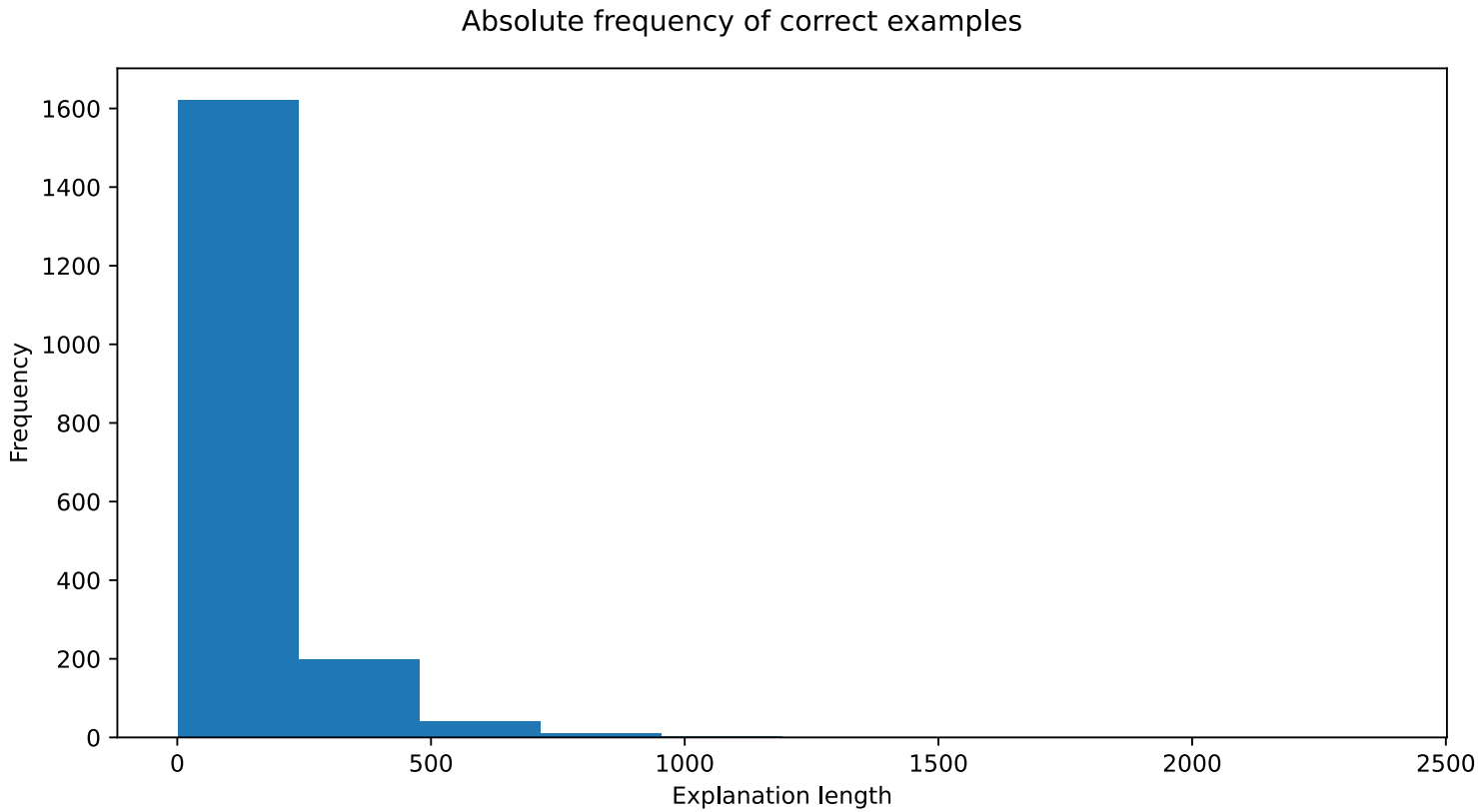
Categorize answers



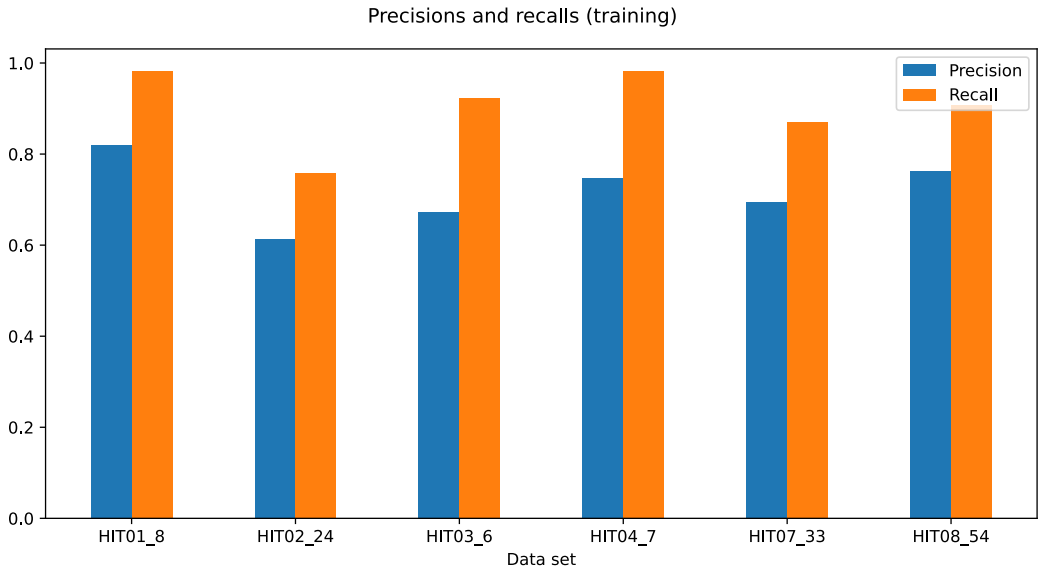
Categorize answers

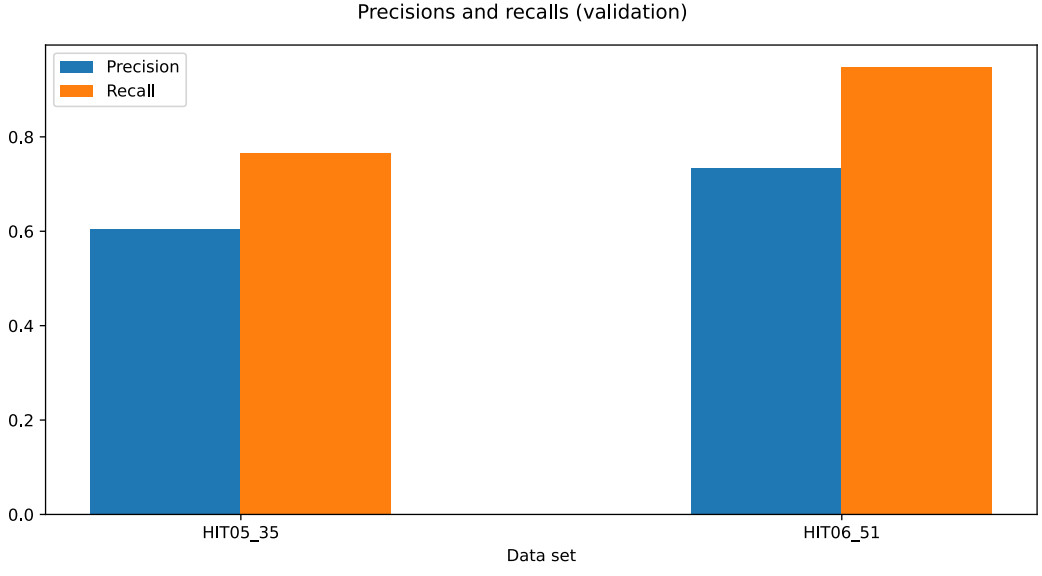


Categorize answers

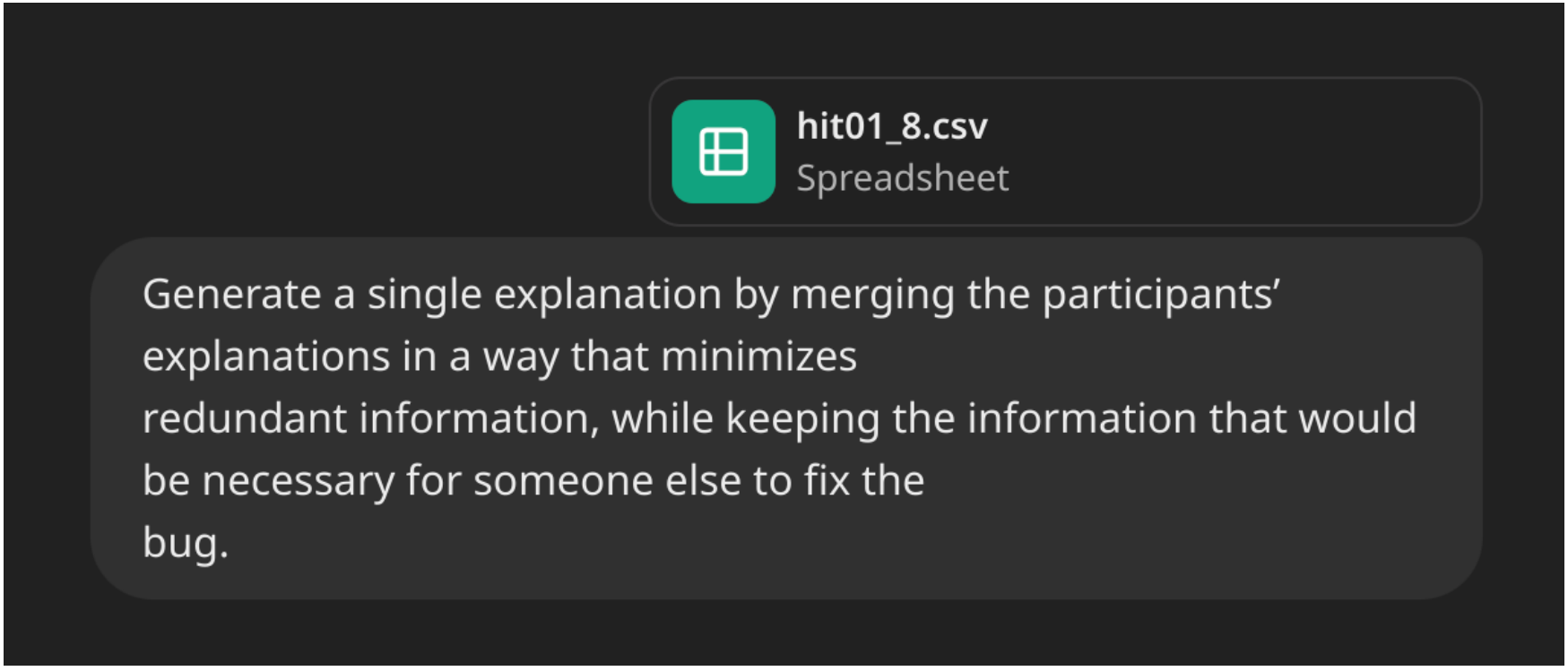


Categorize answers





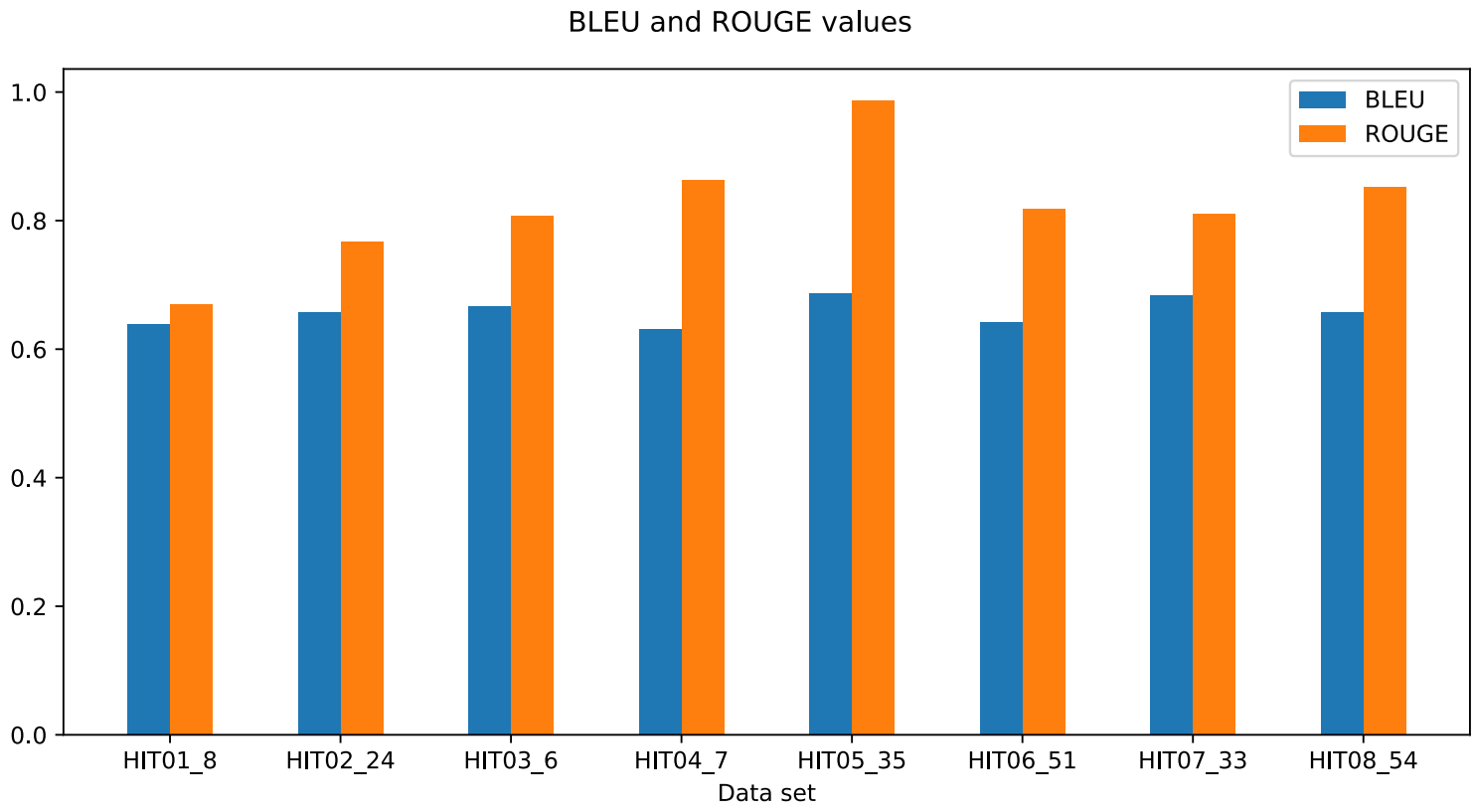
Compare original with synthetic explanations



Compare original with synthetic explanations

"The issue likely stems from an incorrect implementation of the input parsing logic in the code. This could be due to a misalignment between the expected and actual input formats or missing validation checks. The function responsible for handling the input might not properly account for edge cases or invalid data, leading to unexpected behavior or errors during execution. Additionally, there might be an overlooked dependency or an outdated library causing compatibility issues. Reviewing the relevant section of the code and adding appropriate logging can help pinpoint the exact cause of the bug. It's also crucial to ensure that all input data meets the specified criteria before processing."

Compare original with synthetic explanations



Quality of the data and consolidated explanations

- explanation of the LLM long and detailed
- after reprompting 2-line summaries of explanations generated
- explanation is good if it leads a developer to the fix

Keeping the classifier up-to-date

- there might be distribution shift in the occurrence of frequent bugs
- development landscape is transforming rather quickly (new languages, IDEs, idioms, patterns)

Testing the output of the classifier and the LLM

- *BLEU* and *ROUGE* were not suited to judge the quality of the LLM explanations
- tried word and characterwise tokenization
- need semantic equivalence, not lexicographic similarity

Debugging the integration between the classifier and the LLM

- good classifier acts as a first gate
- when *f*1 score is sufficient, feed explanations into LLM for consolidation
- eases search for causes of errors or inaccuracies