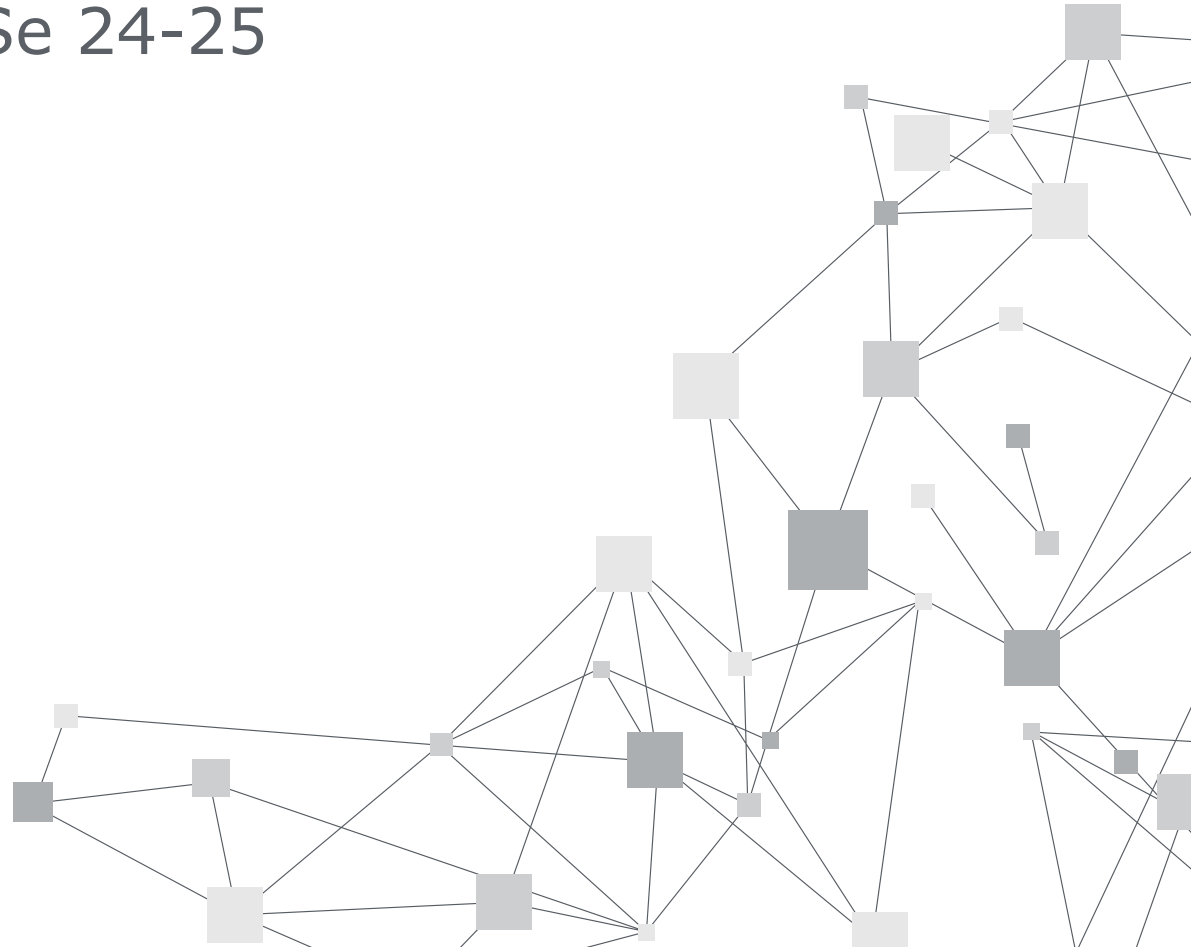
Mini-Project-3

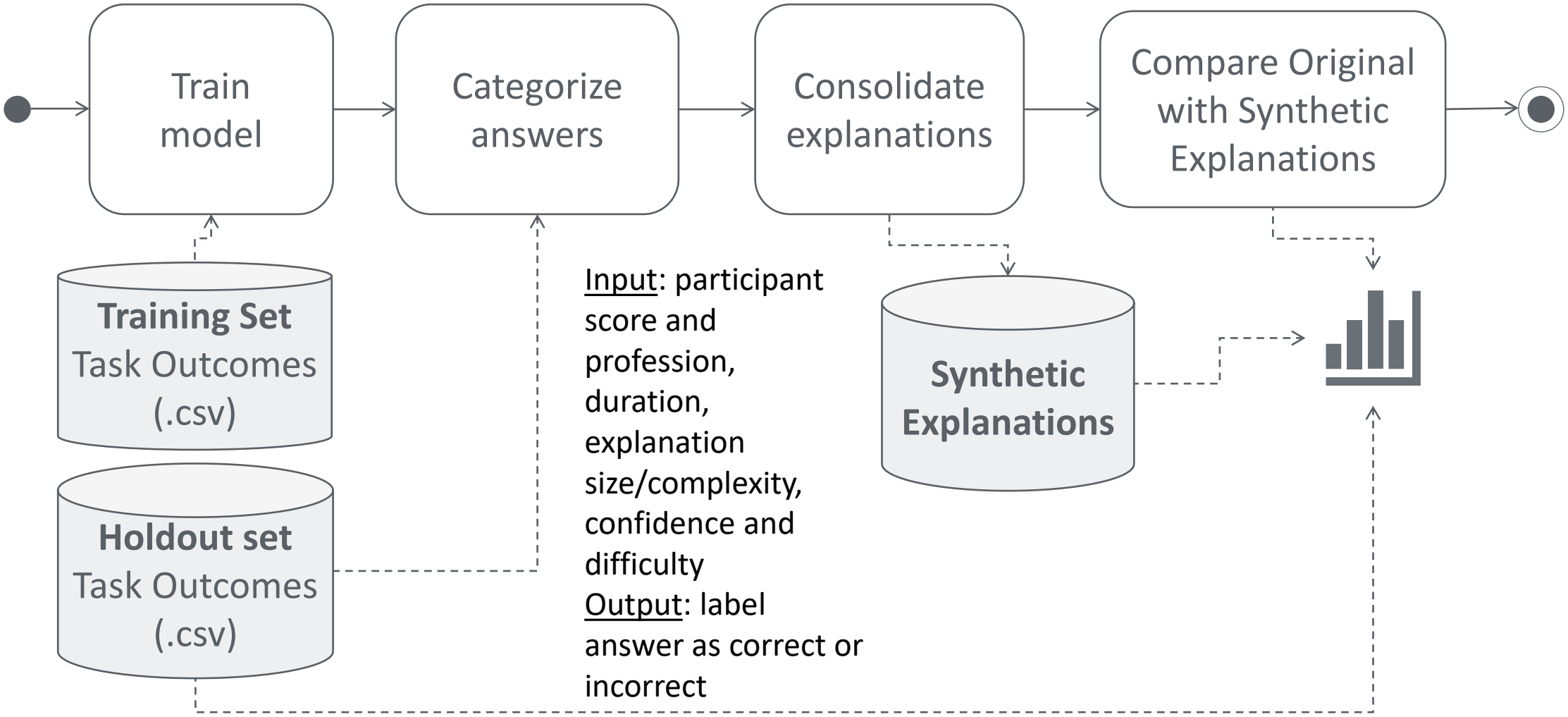
Advanced Software Engineering WiSe 24-25

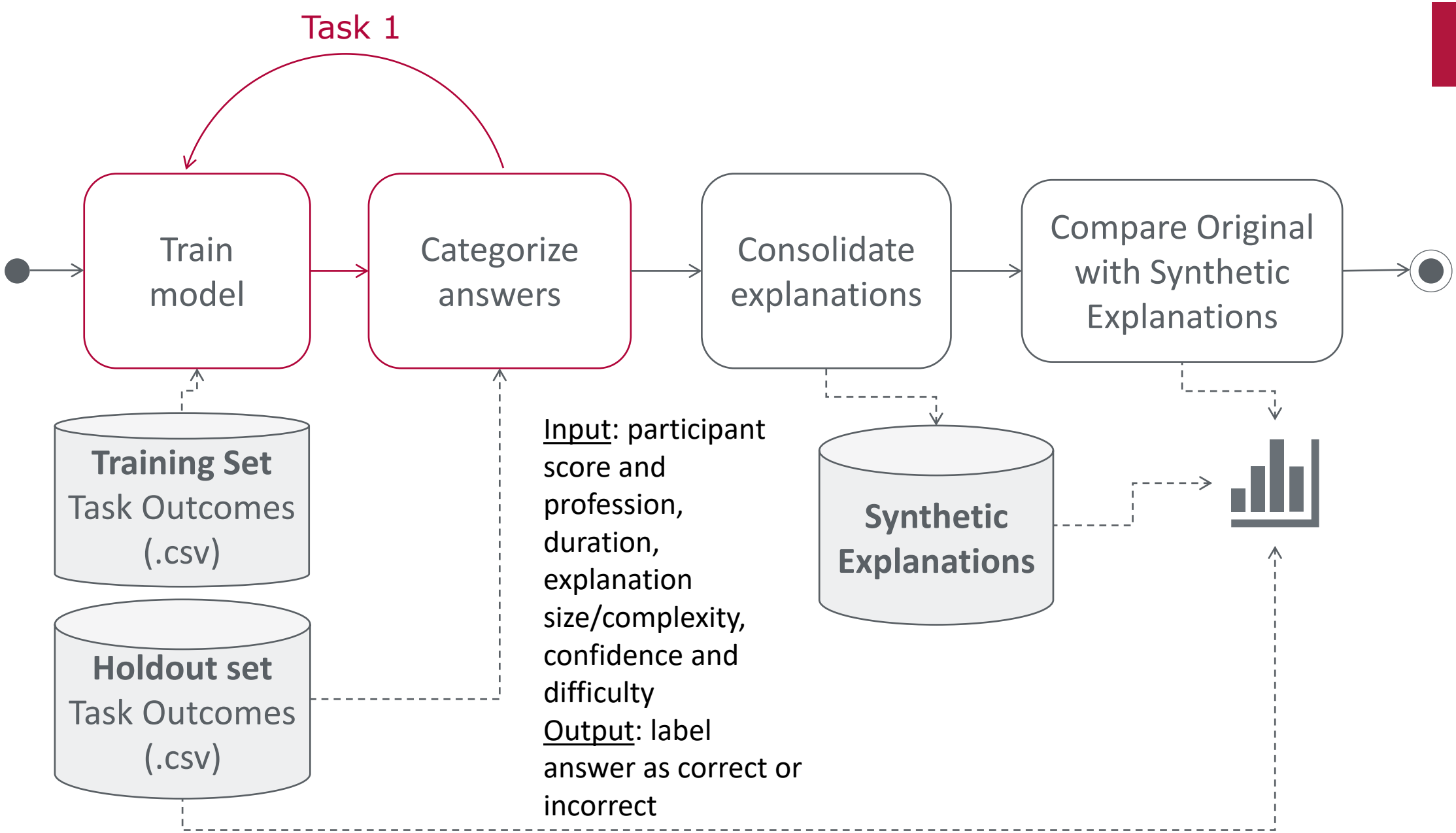
Dr. Sona Ghahremani

Results due **Feb. 11, 2025**
(in form of a report)



Consolidating bug report explanations




Task 1
Train model
Categorize answers
Consolidate explanations
Compare Original with Synthetic Explanations
Training Set
Task Outcomes (.csv)
Holdout set
Task Outcomes (.csv)
Input: participant score and profession, duration, explanation size/complexity, confidence and difficulty
Output: label answer as correct or incorrect
Synthetic Explanations

Task 1 : Distribution shifts

Goal: The goal of this task is to measure the impact of changes (distribution shifts) in one input feature (”Participant’s Profession”)

Prepare data: Split the data set in training and holdout set for “Students”

- “Students” are “Graduates” and “Undergraduates”

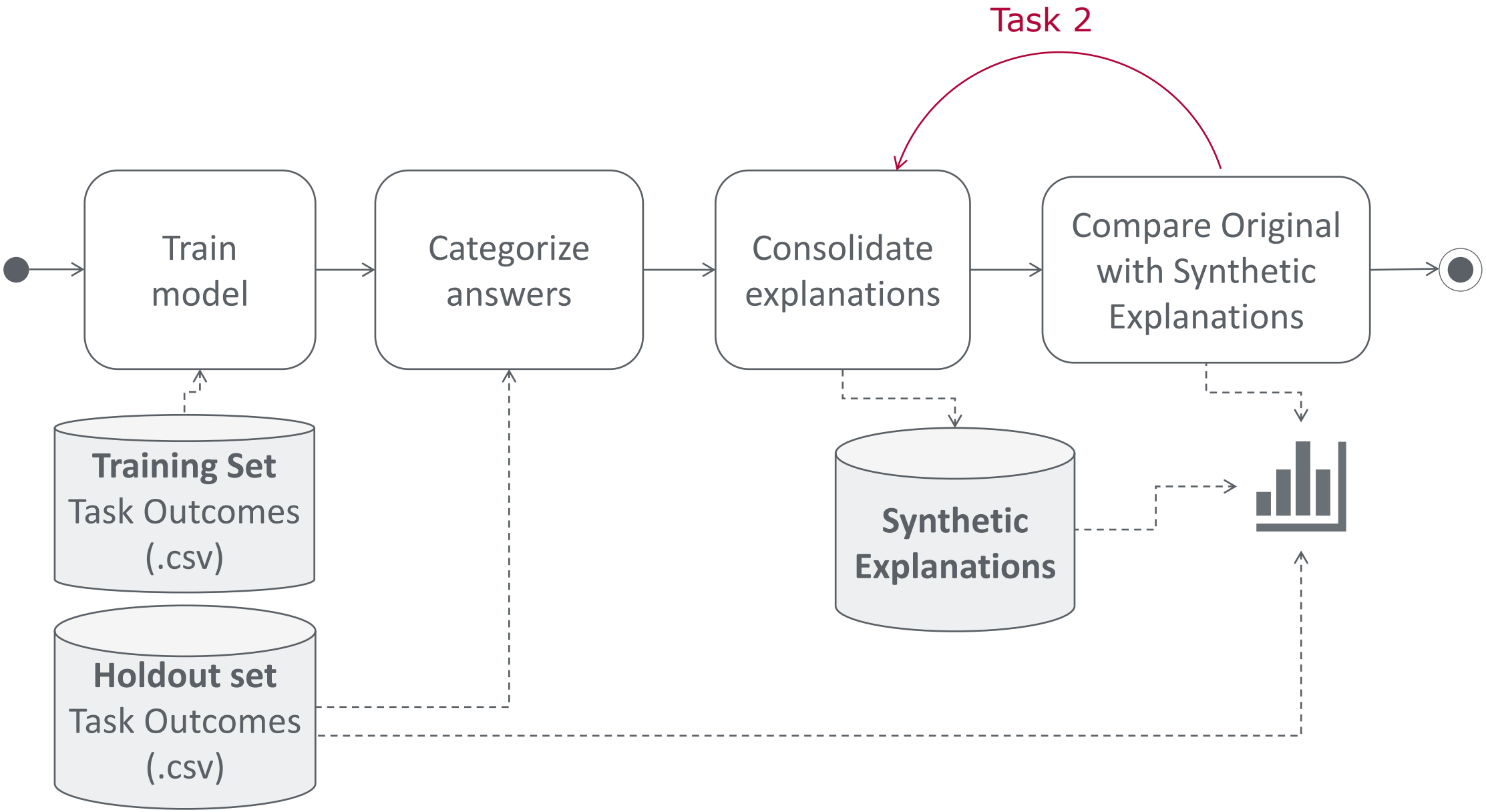
Train the Model: Build the classifier for “Students. Report the precision and recall of the Student model on the holdout set

Measure Impact: Gradually add at random “Non-Students” to the holdout set and report on the degradation of the classifier’s precision and recall

- The Non-Students are: "Hobbyist", "Professional", "Programmer", and "Other”

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?

Question 1.2: What is the min number of “Non-Students” to train a model that produces similar outcome to the model trained on mixed data (from mini project 2)?


Task 2
Train model
Categorize answers
Consolidate explanations
Compare Original with Synthetic Explanations
Training Set
Task Outcomes (.csv)
Holdout set
Task Outcomes (.csv)
Synthetic Explanations

Task 2 : Necessary and Sufficient Explanations

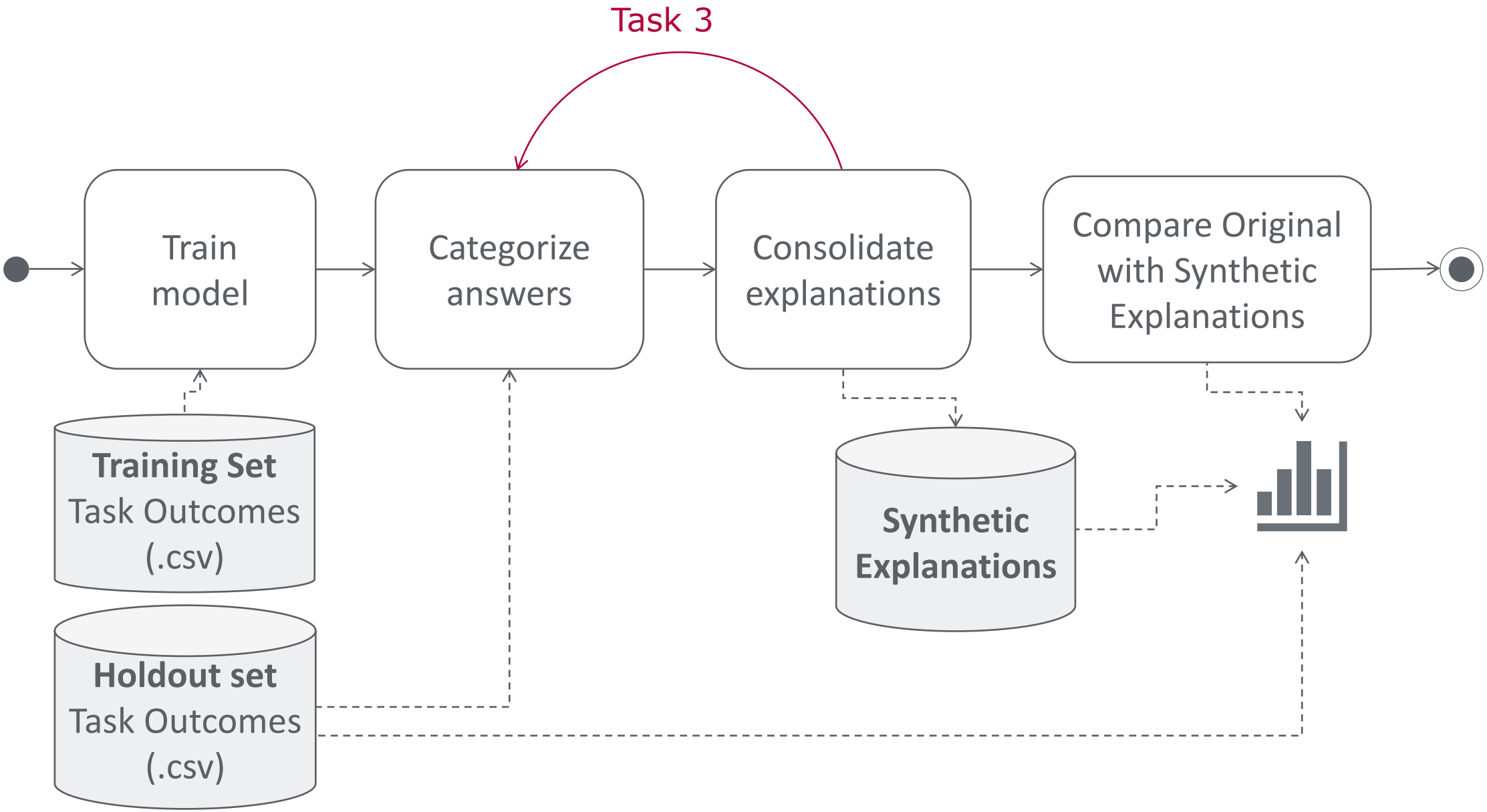
Goal: The goal of this task is to select a minimal number of explanations that after merging (by the LLM) will satisfy your defined thresholds of readability and semantic similarity to your ground truth

Criteria: Readability, Semantic Similarity to ground truth (choose a metric and justify)

Generate Ground truth: Note that you will need to produce your own ground truth of explanation that contains all necessary and sufficient information to understand and fix the bug. Use raw data to generate your ground truth

Question 2.1: Define thresholds on readability and semantic similarity to reason about if and how many explanation(s) need to be added to the final consolidated explanations

- Report on your findings


Task 3
Train model
Categorize answers
Consolidate explanations
Compare Original with Synthetic Explanations
Training Set
Task Outcomes (.csv)
Holdout set
Task Outcomes (.csv)
Synthetic Explanations

Task 3: Diverse Explanations

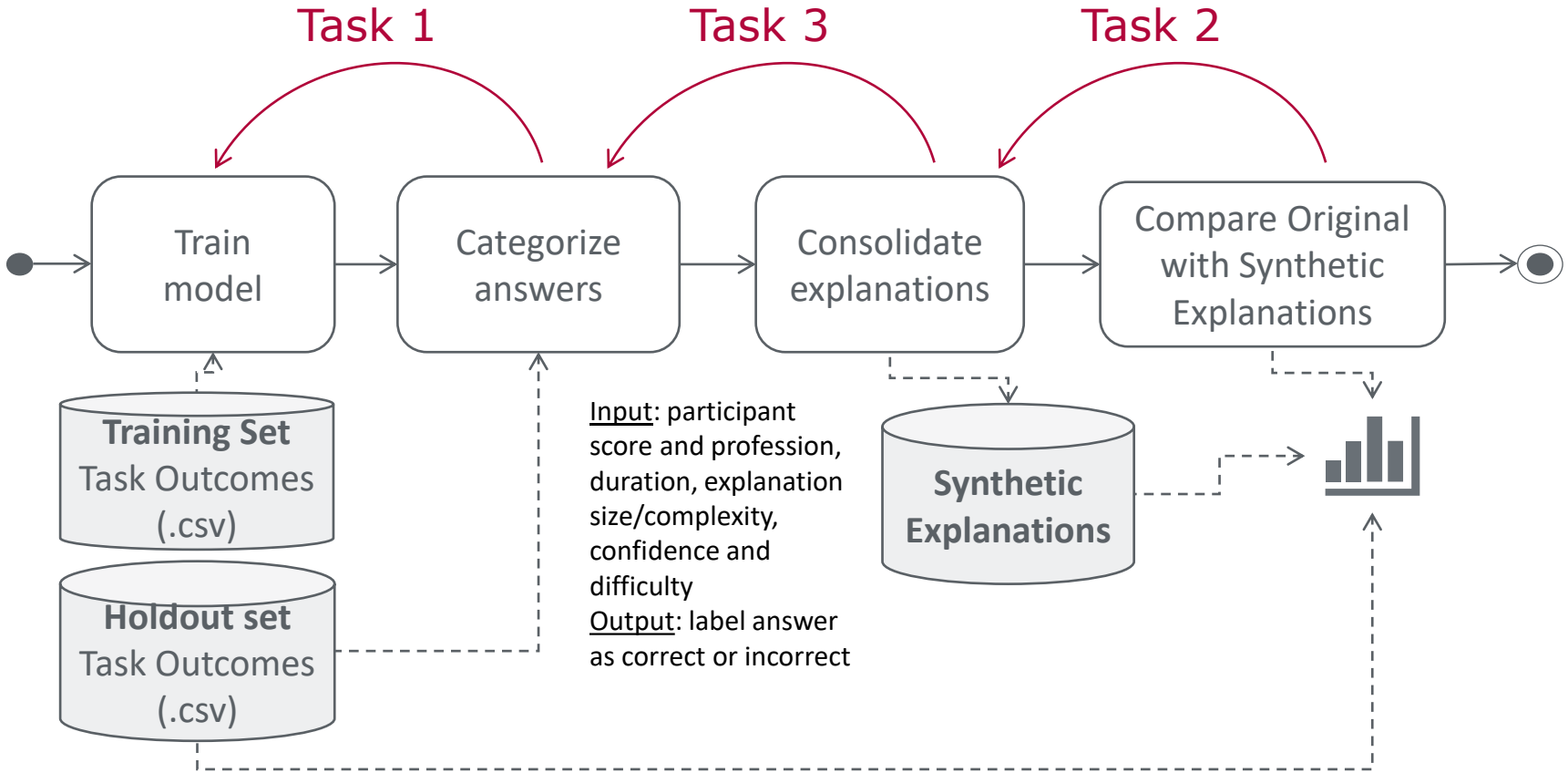
Goal: The goal of this task is to select a diverse set of explanations that are still necessary and sufficient in terms of allowing someone to understand and fix a bug. Use readability and semantic similarity to your ground truth (produced in Task-2) to find a set of explanations that are maximally diverse.

Objectives: High readability, high semantic similarity, and high diversity of the demographics/answer attributes

Question 3.1 : how would you measure diversity? E.g., entropy of each feature

Question 3.2 : what is the max readability and semantic similarity independent of the diversity?

Question 3.3 : what is the max diversity for (previously achieved) max semantic similarity ? (compromising readability)


Task 1
Task 3
Task 2
Train model
Categorize answers
Consolidate explanations
Compare Original with Synthetic Explanations
Training Set
Task Outcomes (.csv)
Holdout set
Task Outcomes (.csv)
Input: participant score and profession, duration, explanation size/complexity, confidence and difficulty
Output: label answer as correct or incorrect
Synthetic Explanations

End