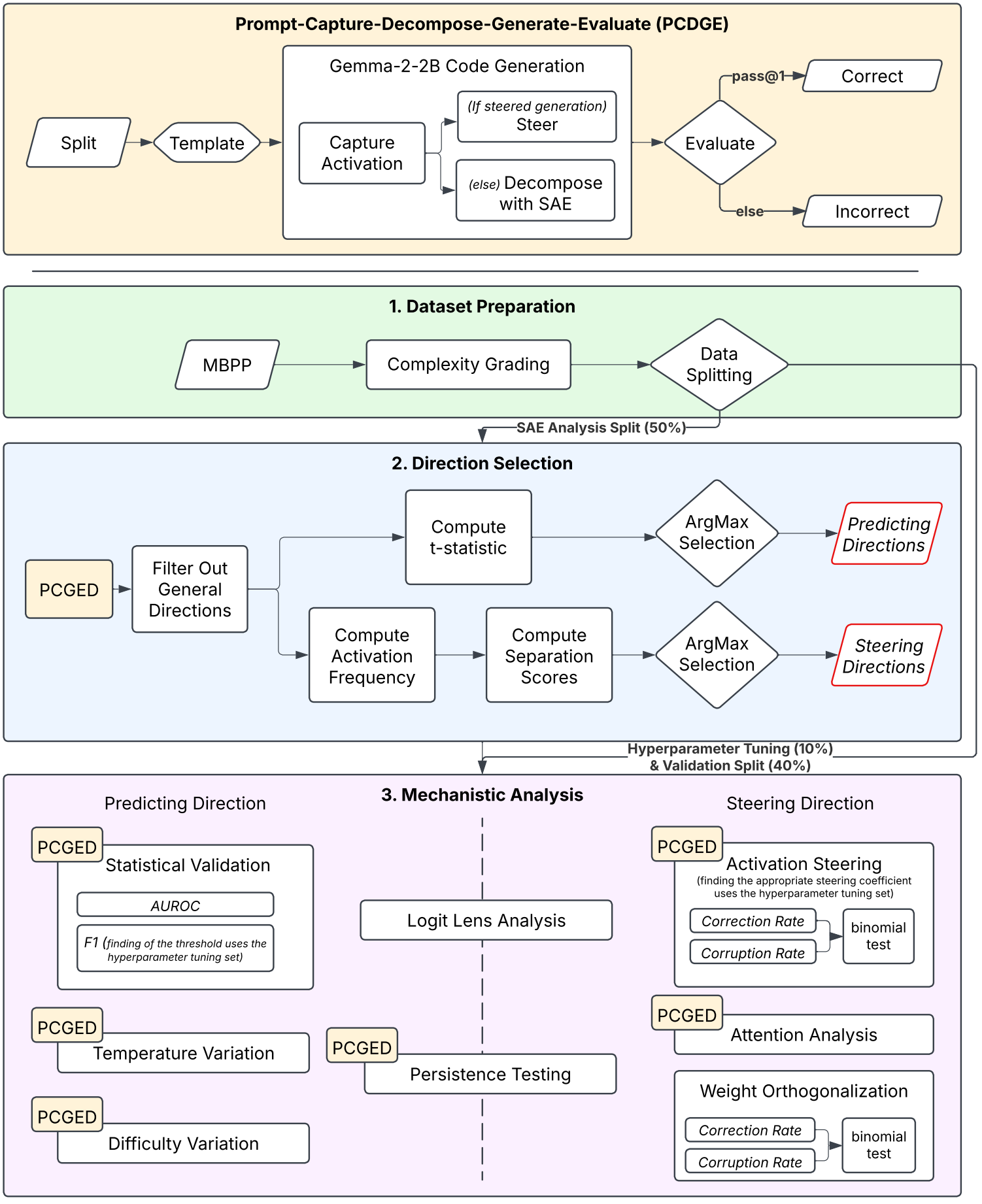
Prompt-Capture-Decompose-Generate-Evaluate (PCDGE)
Split
Template
Gemma-2-2B Code Generation
Capture Activation
(If steered generation) Steer
(else) Decompose with SAE
Evaluate
pass@1
Correct
else
Incorrect
1. Dataset Preparation
MBPP
Complexity Grading
Data Splitting
SAE Analysis Split (50%)
2. Direction Selection
PCGED
Filter Out General Directions
Compute t-statistic
ArgMax Selection
Predicting Directions
Compute Activation Frequency
Compute Separation Scores
ArgMax Selection
Steering Directions
Hyperparameter Tuning (10%) & Validation Split (40%)
3. Mechanistic Analysis
Predicting Direction
PCGED
Statistical Validation
AUROC
F1 (finding of the threshold uses the hyperparameter tuning set)
PCGED
Temperature Variation
PCGED
Difficulty Variation
Logit Lens Analysis
PCGED
Persistence Testing
Steering Direction
PCGED
Activation Steering
(finding the appropriate steering coefficient uses the hyperparameter tuning set)
Correction Rate
Corruption Rate
binomial test
PCGED
Attention Analysis
Weight Orthogonalization
Correction Rate
Corruption Rate
binomial test