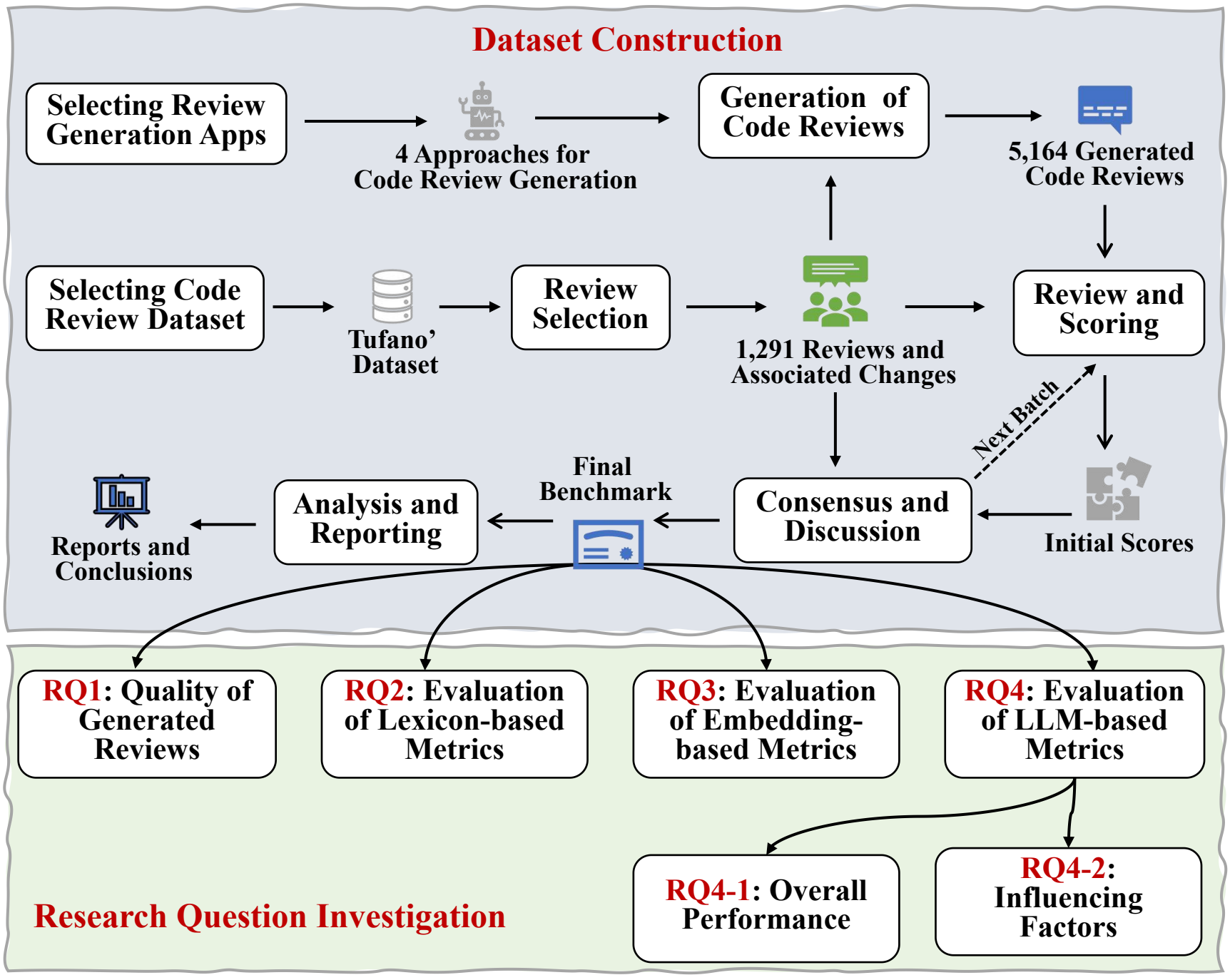
Dataset Construction
Selecting Review Generation Apps
4 Approaches for Code Review Generation
Generation of Code Reviews
5,164 Generated Code Reviews
Selecting Code Review Dataset
Tufano' Dataset
Review Selection
1,291 Reviews and Associated Changes
Review and Scoring
Next Batch
Initial Scores
Consensus and Discussion
Final Benchmark
Analysis and Reporting
Reports and Conclusions
RQ1: Quality of Generated Reviews
RQ2: Evaluation of Lexicon-based Metrics
RQ3: Evaluation of Embedding-based Metrics
RQ4: Evaluation of LLM-based Metrics
RQ4-1: Overall Performance
RQ4-2: Influencing Factors
Research Question Investigation