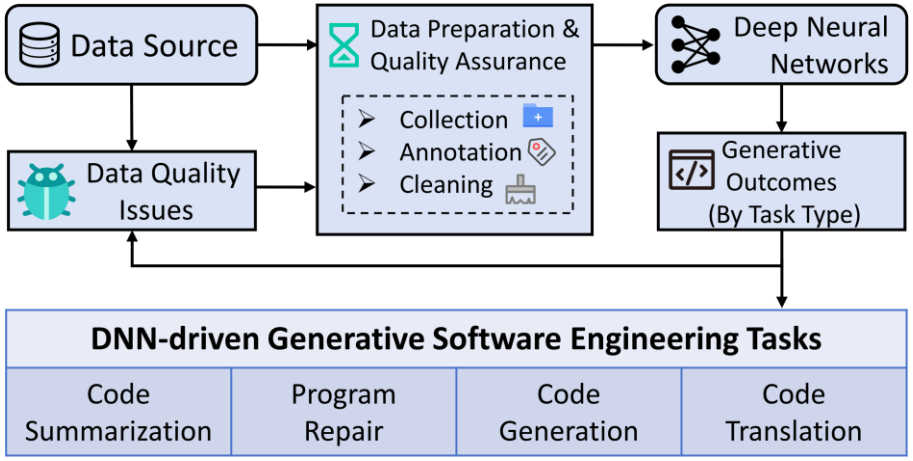
Data preparation and quality for code-centric generative software engineering tasks: a systematic literature review

Shihao WENG, Yang FENG, Yining YIN, Zhenlun ZHANG, Baowen XU

Frontiers of Computer Science, DOI: 10.1007/s11704-025-41376-3

Problems & Ideas

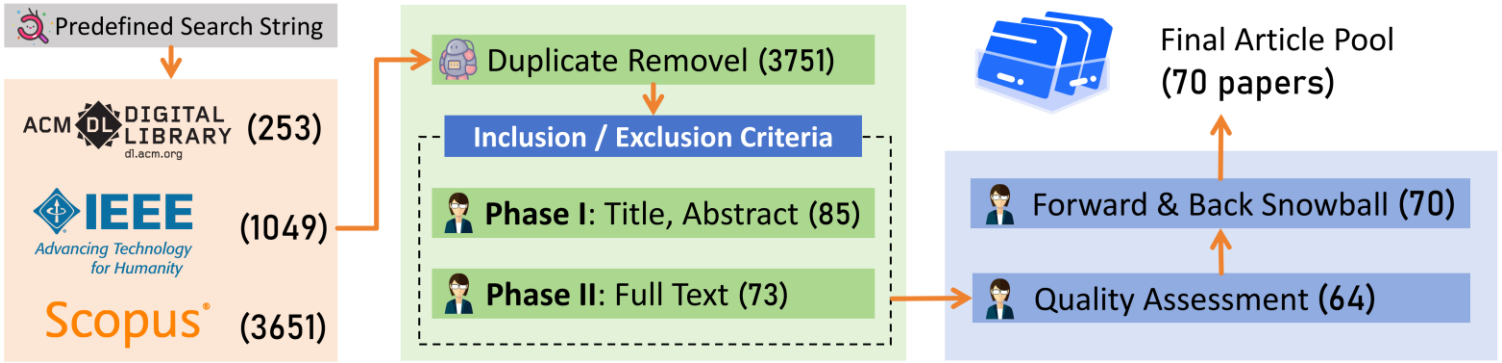
- Challenges in Dataset Construction for Generative Software Engineering:
  - DNN models in generative software engineering tasks (GSETs) rely heavily on dataset quality.
  - Current datasets suffer from noise, imbalance, lack of diversity, outdated data, and inconsistent annotations.
- Ideas: A systematic review framework that analyzes dataset construction, identifies quality issues, and summarizes solutions to improve data reliability in GSETs.



Conceptual framework of DNN-driven Generative Software Engineering Tasks.

Main Contributions

- Contributions:
  - This paper is the first literature review to analyze and summarize the processes of dataset construction and their quality issues in code-centric GSETs;
  - This paper provides a detailed analysis of data quality issues and their impact on model perfor mance, evaluates solutions proposed in the litera ture, and guides the direction for future research;
  - This paper offers practical recommendations for subsequent researchers on building high-quality datasets, including how to select and clean data, and how to annotate and validate datasets, to enhance the performance of DNN models in GSETs.



The process of paper search and selection.