

FCS 南京大学徐宝文等，以代码为中心的生成式软件工程任务的数据准备和质量：系统性文献综述

作者：writer 来源：科学网

本文原地址：<https://www.iikx.com/news/progress/41951.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

FCS

南

京大

学徐宝文

等，以代码为中心的生成式软件工程任务的数据准备和质量：系统性文献综述。论文标题：Data preparation and quality for code-centric generative software engineering tasks: a systematic literature review

期刊：Frontiers of Computer Science

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

发表时间：20 Jun 2025

DOI：10.1007/s11704-025-41376-3

微信链接：点击此处阅读微信文章

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✉

State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

Received December 17, 2024; accepted June 16, 2025

E-mail: fengyang@nju.edu.cn; bwxu@nju.edu.cn

© The Author(s) 2025. This article is published with open access at link.springer.com and journal.hep.com.cn

引用格式：

Shihao WENG, Yang FENG, Yining YIN, Zhenlun ZHANG, Baowen XU. Data preparation and quality for code-centric generative software engineering tasks: a systematic literature review. *Front. Comput. Sci.*, 2026, 20 (9) : 2009203

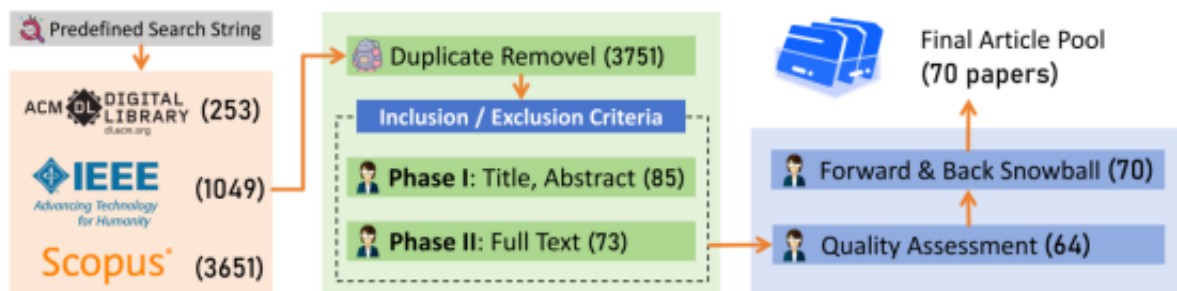
阅读原文：



问题概述

针对生成式软件工程任务中数据准备与质量保障问题，南京大学徐宝文教授团队撰写了该系统综述论文。

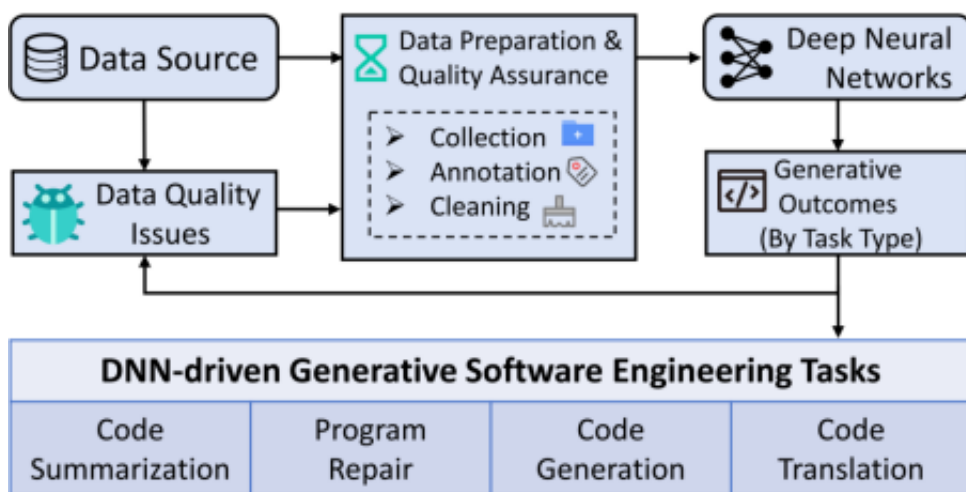
聚焦于代码摘要、程序修复、代码生成与翻译等任务中，深度神经网络模型的性能高度依赖于训练数据的质量。文章通过系统性文献回顾，梳理了数据构建流程、常见质量问题及已有应对策略，填补了该领域缺乏系统综述的空白。



技术步骤

作者从70篇相关研究中提取数据，围绕三个核心问题展开：一是数据集构建的常见流程，包括任务需求、数据采集与注释方法；二是常见数据质量问题，如噪声、分布不均、缺乏多样性等，并分析其成因与影响；三是应对策略，包括数据清洗、增强、标准化构建方法和多语言联合训练等。

。



实验结果

研究发现，数据质量直接影响生成模型的泛化能力与稳定性。通过对现有方法的整理与评估，论文为构建高质量数据集提供了系统性建议，有助于提升生成式软件工程模型在实际场景中的可靠性与表现。该综述为后续研究者提供了数据构建与质量保障的全面参考。

GSETs	Data quality issue	Freq.	Ratio
CS	Noisy data.	3	20.0%
	Information redundancy.	1	6.7%
	Insufficient data collection.	1	6.7%
	Uneven data distribution.	2	13.3%
	Data scarcity.	2	13.3%
	Lack of diversity.	3	20.0%
	Lack of standardization.	1	6.7%
	Inaccurate associations.	1	6.7%
PR	Unrelated commits.	4	14.8%
	Outdated Datasets.	2	7.4%
	The rarity of manually validated.	4	14.8%
	Repeated use of benchmarks.	2	7.4%
	Data leakage.	1	3.7%
	Insufficient novelty and diversity.	2	7.40%
CG	Insufficient granularity.	1	3.70%
	Functionality correctness overemphasis.	1	4.2%
	Lacking real-world coding issues.	2	8.3%
	Scarcity and lack of diversity.	2	8.3%
	Lack tests.	3	12.5%
	Unreasonable format and style.	2	8.3%
CT	Vague problem descriptions.	1	4.2%
	Insufficient programming languages.	2	14.3%
	Scarcity of parallel data.	4	28.6%
	Insufficiency of diversity.	3	21.4%
	Uneven distribution.	1	7.1%
	Poor code quality and execution.	3	21.4%

期刊简介

Frontiers of Computer Science (FCS) 是由教育部主管、高等教育出版社和北京航空航天大学共同主办，南京大学支持，SpringerNature 公司海外发行的英文学术期刊。本刊于 2007 年创刊，月刊，全球发行。主要刊登计算机科学领域具有创新性的综述论文、研究论文等。本刊主编为周志华院士，执行主编为周昆教授。编委会及青年 AE

团队由国内外知名学者及优秀青年学者组成。本刊被 SCI、Ei、DBLP、INSPEC、SCOPUS 和中国科学引文数据库 (CSCD) 核心库等收录，为 CCF 推荐 B 类期刊；两次入选中国科技期刊国际影响力提升计划；入选第 4 届中国国际化精品科技期刊；两次入选中国科技期刊卓越行动计划（一期梯队、二期领军）。

来源：Frontiers of Computer Science

更多 科学进展 请访问 <https://www.iikx.com/news/progress/>

本文版权归原作者所有，请勿用于商业用途，[爱科学iikx.com](http://www.iikx.com)转发