南通大学
NANTONG UNIVERSITY

AI4SE的研究问题和方法

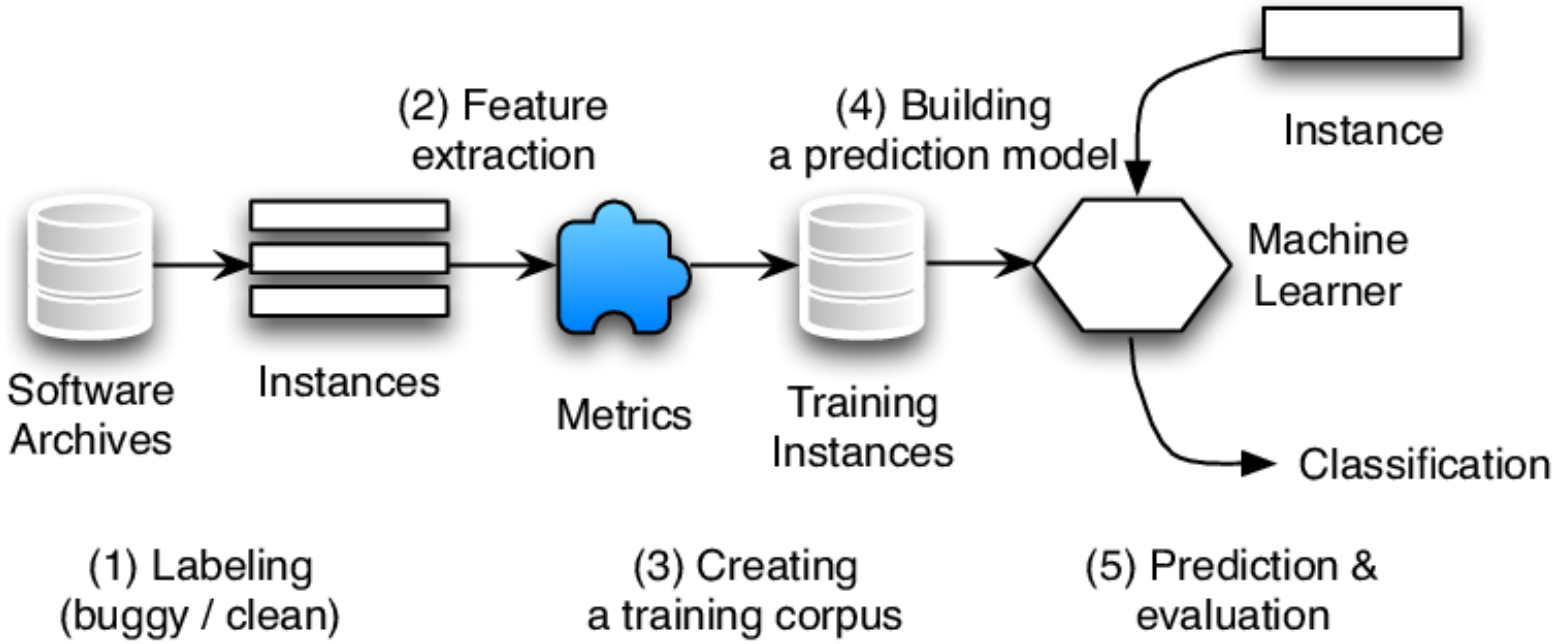
汇报：陈翔
信息科学技术学院
软件工程系

问题类型

- 分类问题
- 生成问题
- 推荐问题
- 问答系统

分类问题

- 软件缺陷预测（Software Defect Prediction）

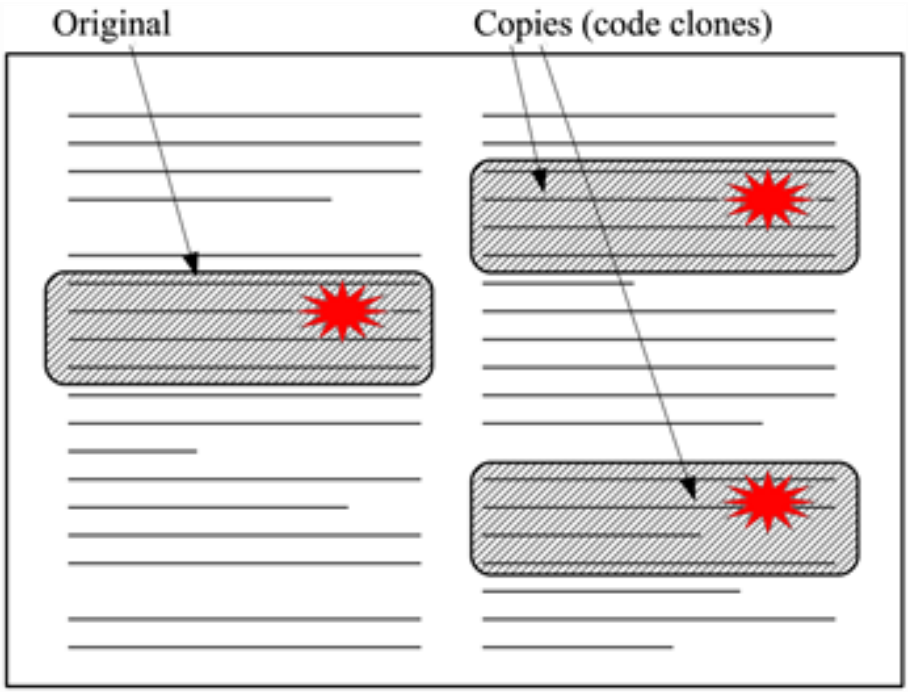


类似问题：

- 即时软件缺陷预测（Just-in-time Defect Prediction）：针对代码提交进行预测
- 漏洞检测/预测（Vulnerability Detection/Prediction）：预测漏洞
- Android Malware Detection：针对Android恶意软件

分类问题

- 重复缺陷报告检测（Duplicate Bug Report Detection）
- 代码克隆检测（ Code Clone Detection）

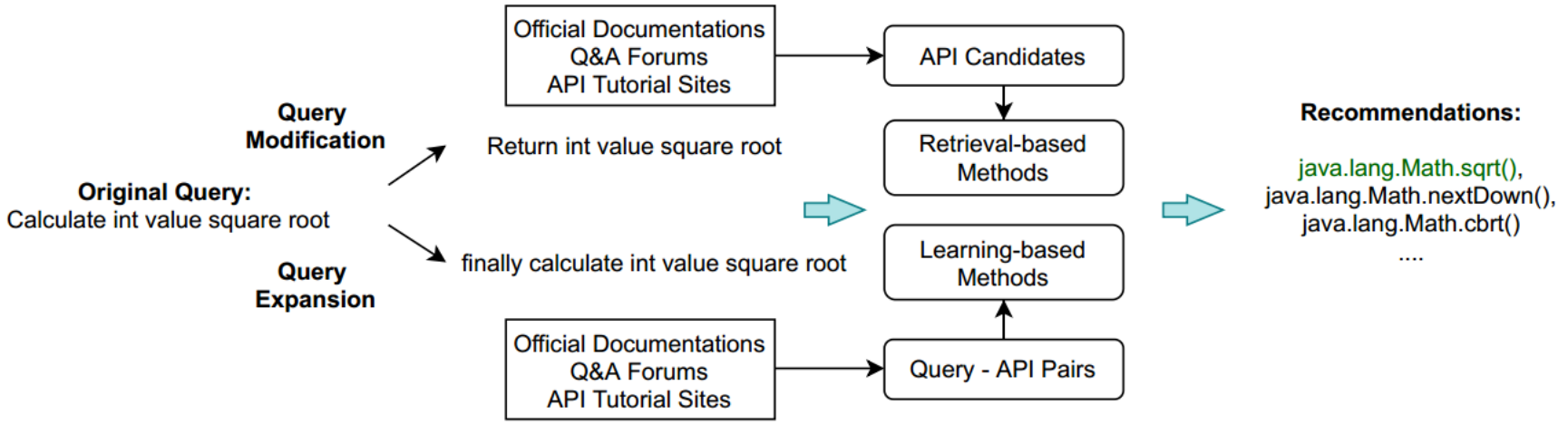


生成问题

- 问题类型
  - 代码到文本：source code summarization、commit message generation、code review comment generation等
  - 文本到代码：通用代码生成、测试用例代码生成、特定领域的代码生成（SQL代码、正则表达式代码、Bash代码）
  - 代码到代码：缺陷/漏洞修复（修复前代码->修复后代码）
  - 文本到文本：查询重构
- 基本的研究思路
  - Seq2seq模型

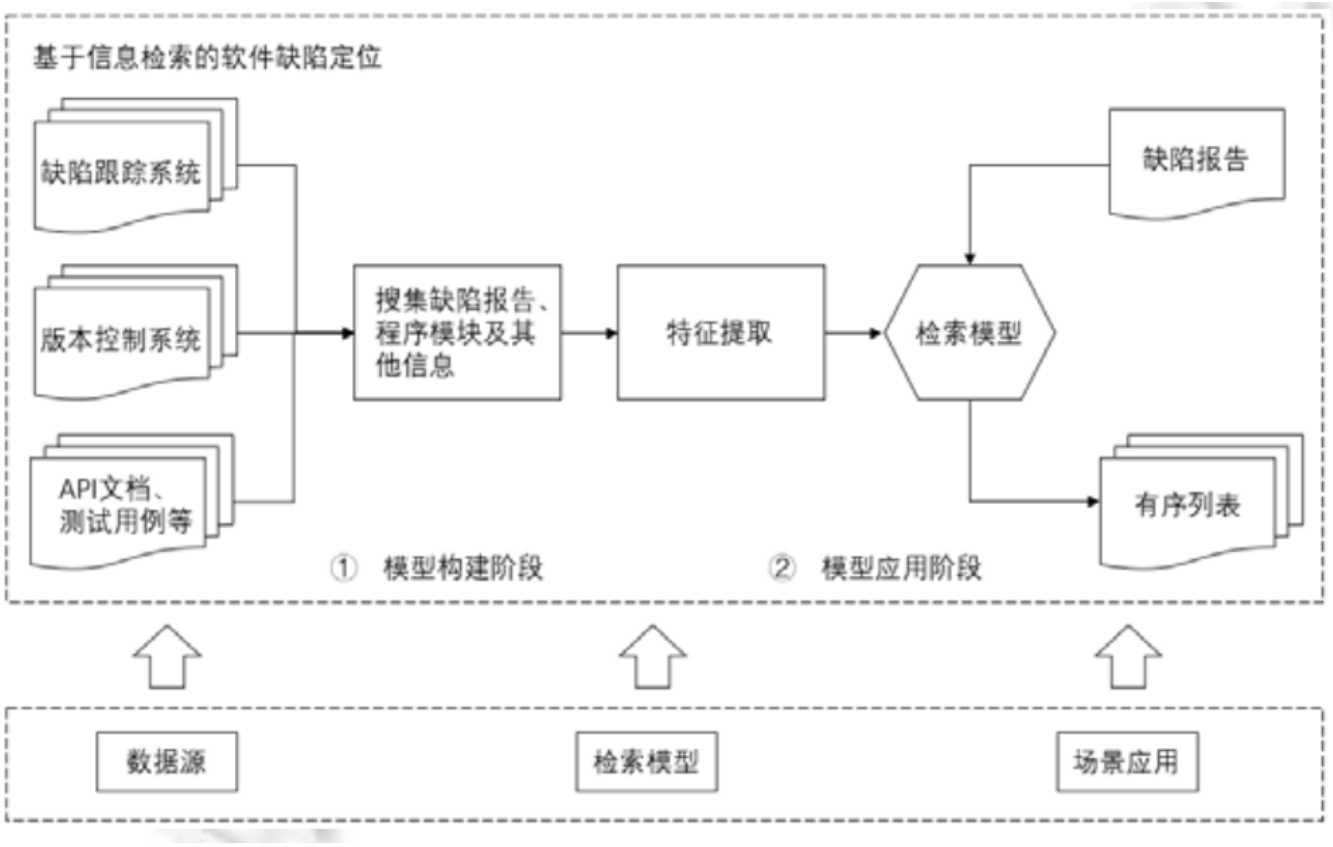
推荐问题

- 代码搜索（基于query，找到需要的API/代码）



推荐问题

- 基于信息检索的缺陷定位（基于bug report，找到项目内的缺陷代码）



问答系统

- 答案生成
  - ShellFusion: Answer Generation for Shell Programming Tasks via Knowledge Fusion ICSE 2022
- APP Review Response
  - Automating App Review Response Generation Based on Contextual Knowledge TOSEM 2022
- 答案总结
  - Answer Summarization for Technical Queries Benchmark and New Approach ASE 2022
- Bug report Summarization
  - Automatic Solution Summarization for Crash Bugs ICSE 2021

研究方法

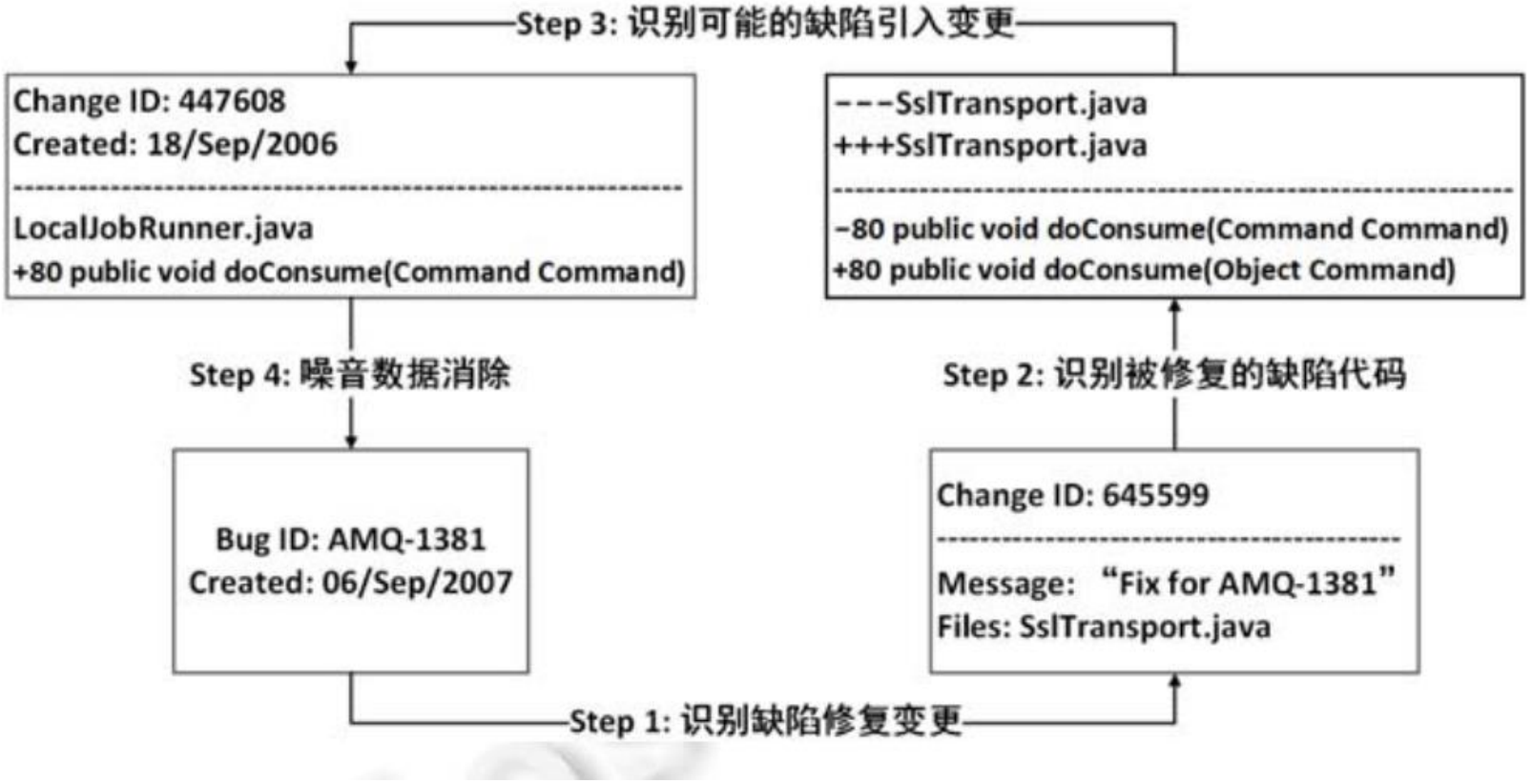
- 数据集角度
- 实验设置角度
- 代码表征学习角度
- 跨编程语言/跨项目问题角度
- 模型构建角度
- 模型评估角度

数据集角度

- 数据集内的噪音识别和移除
  - On the importance of building high-quality training datasets for neural code search ICSE 2022
- 数据预处理
  - 特征选择方法（Feature Selection）
  - 类不平衡方法（Class Imbalanced Learning）
- 数据集的扩充方法（例如保持语义等价的前提下构建新的样本）
  - Bridging pre-trained models and downstream tasks for source code understanding ICSE2022

数据集角度

- 模块标记问题
  - SZZ算法是缺陷预测中的经典问题



实验设置角度

- 数据预处理方法的影响
  - Are We Building on the Rock? On the Importance of Data Preprocessing for Code Summarization FSE 2022
- 数据集的划分方式的影响
  - Impact of Evaluation Methodologies on Code Summarization ACL 2022
- 评测指标的问题
  - CrystalBLEU Precisely and Efficiently Measuring the Similarity of Code ASE 2022
  - On the Use of Evaluation Measures for Defect Prediction Studies ISSTA 2022

实验设置角度

- 超参优化
  - How to Better Distinguish Security Bug Reports (Using Dual Hyperparameter Optimization). EMSE 2021
  - Understanding the automated parameter optimization on transfer learning for cross-project defect prediction: an empirical study. ICSE 2020
- 实现中使用框架的影响
  - 源代码解析工具： Evaluating the Impact of Source Code Parsers on ML4SE Models 2022

代码表征学习角度

- 如何将代码结构信息引入到预训练模型
  - Learning Program Representations with a Tree-Structured Transformer. SANER 2023
- 如何将知识图谱与深度模型进行融合
  - API-Misuse Detection Driven by Fine-Grained API-Constraint Knowledge Graph ASE 2020
- 其他软件制品的表征学习
  - RepresentThemAll: A Universal Learning Representation of Bug Reports ICSE 2022
  - Cc2vec: Distributed representations of code changes ICSE 2020
  - Post2vec: Learning distributed representations of Stack Overflow posts TSE 2021

跨编程语言/跨项目问题角度

- 问题
  - 利用源项目的数据，预测目标项目
  - 利用训练数据较多的编程语言数据，预测低资源（low source）编程语言
- 早期借助迁移学习
  - 跨项目缺陷预测：跨项目缺陷预测问题的综述 计算机学报 2018
- Few-shot learning
  - Cross-Domain Deep Code Search with Few-Shot Meta Learning ICSE 2022
- Zero-shot learning
  - Zero-Shot Program Representation Learning ICPC 2022

模型构建角度

- 预训练+微调
  - Studying the Usage of Text-To-Text Transfer Transformer to Support Code-Related Tasks ICSE 2021
  - An Extensive Study on Pre-trained Models for Program Understanding and Generation ISSTA 2022
- Prompt tuning （pre-train, prompt and predict）
  - No More Fine-Tuning? An Experimental Evaluation of Prompt Tuning in Code Intelligence FSE 2022

模型构建角度

- 基于新型的方法来构建问题
  - 对比学习
    - CLEAR Contrastive Learning for API Recommendation ICSE 2022
  - 对偶学习
    - Leveraging code generation to improve code retrieval and summarization via dual learning WWW 2020
  - 集成学习
    - Ensemble Models for Neural Source Code Summarization of Subroutines. ICSME 2021

模型构建角度

- 基于新型的方法来构建问题
  - 强化学习
    - Improving automatic source code summarization via deep reinforcement learning ASE 2018
  - 图神经网络
    - LineVD Statement-level Vulnerability Detection using Graph Neural Networks MSR 2022
  - 个性化模型
    - Why my code summarization model does not work: Code comment improvement with category prediction TOSEM 2021

模型构建角度

- 基于新型的方法来构建问题
  - 半监督学习
    - FRUGAL: Unlocking Semi-Supervised Learning for Software Analytics. ASE 2021
  - 主动学习
    - Multi-label Classification for Android Malware Based on Active Learning TDSC 2022

模型构建角度

- 简单方法 Vs 复杂的方法
  - Neural-machine-translation-based commit message generation: how far are we? ASE 2018
- 专家特征Vs 深度特征 （数据融合）
  - The Best of Both Worlds: Integrating Semantic Features with Expert Features for Defect Prediction and Localization FSE 2022
- 信息检索方法 Vs 深度学习方法 （方法融合）
  - Automated Assertion Generation via Information Retrieval and Its Integration with Deep Learning ICSE 2022

模型评估角度

- 模型的功能性
- 模型的可解释性
- 模型的鲁棒性
- 模型的安全性
- 模型的公平性
- 模型的部署

模型的功能性

- SE4AI
- 研究角度
  - 覆盖准则
    - NPC: Neuron Path Coverage via Characterizing Decision Logic of Deep Neural Networks TOSEM 2022
  - 测试输入的生成（Fuzzing、符号执行等）
    - DeepHunter: a coverage-guided fuzz testing framework for deep neural networks ISSTA 2019
  - 测试输入优化（借鉴回归测试的工作，例如测试用例选择、排序等）
    - An empirical study on data distribution-aware test selection for deep learning enhancement TOSEM 2022
  - 模型中的缺陷定位
    - DeepLocalize: Fault Localization for Deep Neural Networks. ICSE 2021
  - 模型中的缺陷修复
    - Arachne: Search Based Repair of Deep Neural Networks

模型的功能性

- 测试智能系统
  - 对话系统
    - Natural Test Generation for Precise Testing of Question Answering Software ASE 2022
  - 机器翻译系统
    - Improving Machine Translation Systems via Isotopic Replacement. ICSE 2022
  - 很多工作会使用**蜕变测试**方法，针对性设计出**蜕变关系**

模型的鲁棒性

- Adversarial robustness of deep code comment generation TOSEM 2022
- ReCode Robustness Evaluation of Code Generation Models 2022

模型的安全性

- 针对预训练代码模型
  - Natural Attack for Pre-trained Models of Code. ICSE 2022
- 针对API推荐系统
  - Adversarial Attacks to API Recommender Systems: Time to Wake Up and Smell the Coffee? ASE 2021
- 针对Malware detection
  - Adversarial Machine Learning in Malware Detection: Arms Race between Evasion Attack and Defense 2017

模型的公平性

- 模型中是否存在一些性别歧视等问题
- 模型的非功能性属性
- 最近的一个热门研究方向，可参考的综述
  - Fairness Testing: A Comprehensive Survey and Analysis of Trends 2022
  - Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey 2022
  - Software Fairness: An Analysis and Survey 2022

模型的泛化能力

- 对未看到的数据是否有效
  - Can Neural Clone Detection Generalize to Unseen Functionalities ASE 2021
  - Deep Learning based Vulnerability Detection: Are We There Yet
- 从传统的软件到特定的软件
  - API recommendation for machine learning libraries: how far are we? ESEC/SIGSOFT FSE 2022
  - Code-Based Vulnerability Detection in Node.js Applications: How far are we? ASE 2020

模型的部署

- 将预训练模型压缩成小模型，并可以不会大幅度降低性能
  - Safety and Performance, Why not Both? Bi-Objective Optimized Model Compression toward AI Software Deployment ASE 2022
  - Compressing Pre-trained Models of Code into 3 MB FSE 2022

综述论文

- **熟悉Systematic Literature Review的写法**
  - **去CSUR、TSE、TOSEM、IST、JSS等期刊找相关论文**
  - Deep Learning for Android Malware Defenses: a Systematic Literature Review CSUR 2022
  - Comparing methods for large-scale agile software development: A systematic literature review TSE 2021
  - Opinion mining for software development: a systematic literature review TOSEM 2022

评审AI4SE论文

- 研究问题
  - 是否解决了实际问题
  - 是否重要
  - 是否具有挑战性
- 研究动机
  - 案例
  - 场景

评审AI4SE论文

- 研究方法
  - 针对研究问题，是否有方法定制
  - 给出方法框架图，包含模型的构建和模型的推断
  - 针对定制部分，是否设计消融实验
  - 是否开发了相应工具

评审AI4SE论文

- 实验设置
  - 数据集搜集方法是否合理，如何避免噪音
  - 数据集的划分方式：训练集、验证集、测试集
  - 数据集的评测指标
  - 考虑的基准方法，是否包含了最新方法
  - 模型中参数取值的设定

评审AI4SE论文

- 结果分析
  - 相对基准方法是否有性能显著提升
  - 方法性能提升是否具有显著性
  - 模型的构建和推断时间开销
  - 超参取值对结果的影响
  - Human study，human study使用的方法是否合理
  - 论文的limitation 是什么？选出不work的例子，做原因分析

评审AI4SE论文

- 相关工作
  - 从不同角度来分别总结
  - 每个角度的最后强调一下研究工作的创新之处

评审AI4SE论文

- 共享模型、数据集、代码、实际结果到GitHub
- Readme中提供详细指南