中山大學 软件工程学院

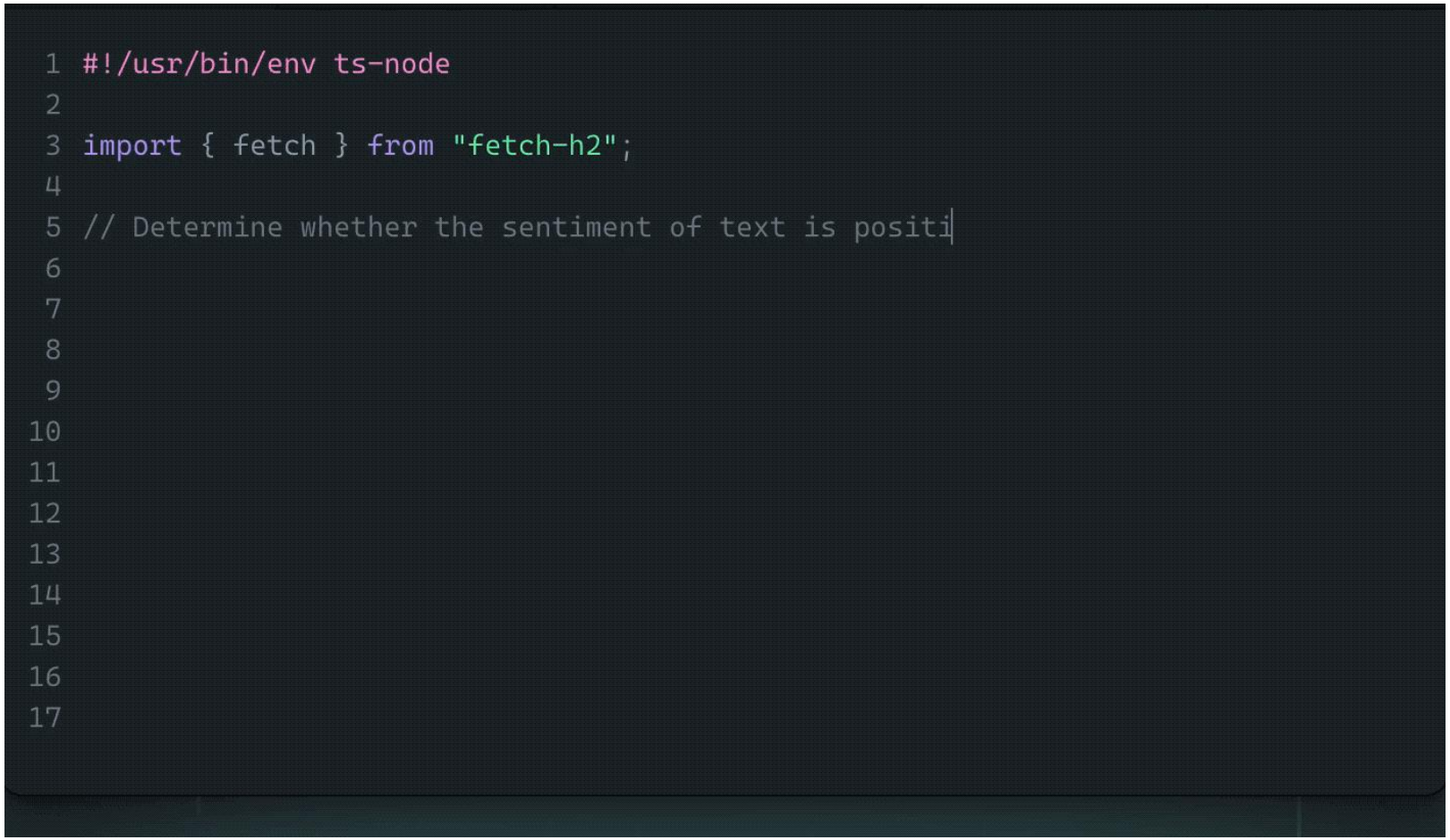
SUN YAT-SEN UNIVERSITY SCHOOL OF SOFTWARE ENGINEERING

代码智能：从任务特定模型到通用大模型的探索之路

王焱林

wangylin36@mail.sysu.edu.cn

GitHub Copilot



https://segmentfault.com/a/1190000040931798

“你已经是一个成熟的AI了，该学会自己补全代码了”

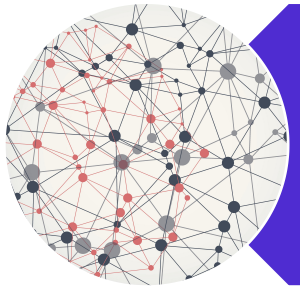
“用AI来写AI代码，“卷死”其他程序员？”

“Copilot在写代码，我在摸鱼。”

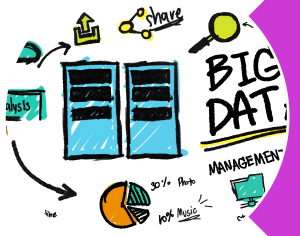
“麻瓜也可以写程序”

“愚蠢的人类，快走开——我的代码我自己写！”

背景



大数据时代下，人工智能的方法特别是深度学习技术在很多领域(如计算机视觉，自然语言处理等）取得了巨大成功.



于此同时，大量的高质量的开源代码涌现出来，例如，截止2022年6月1日，Github上被赞数量超过1千的开源项目超过3万个，包含超千万行代码。

越来越多的研究开始探索使用人工智能的方法，智能化地对大规模的代码数据进行分析和建模，获得代码表征，提升软件工程技术的性能。

目录

目录

背景

- 现代社会中，软件无处不在，涉及到教育、医疗、娱乐和公共安全。
- 由于软件的复杂性，软件的开发、运行和维护耗时耗力。

代码检索技术

给定自然语言查询，自动从代码库检索语义相关的代码片段。

代码补全技术

代码补全技术会基于对已有的上下文的理解，给出下一个代码变量或者下一行代码的一些参考。

代码摘要生成技术

代码摘要生成技术会用简洁的自然语言总结代码片段所执行的功能。

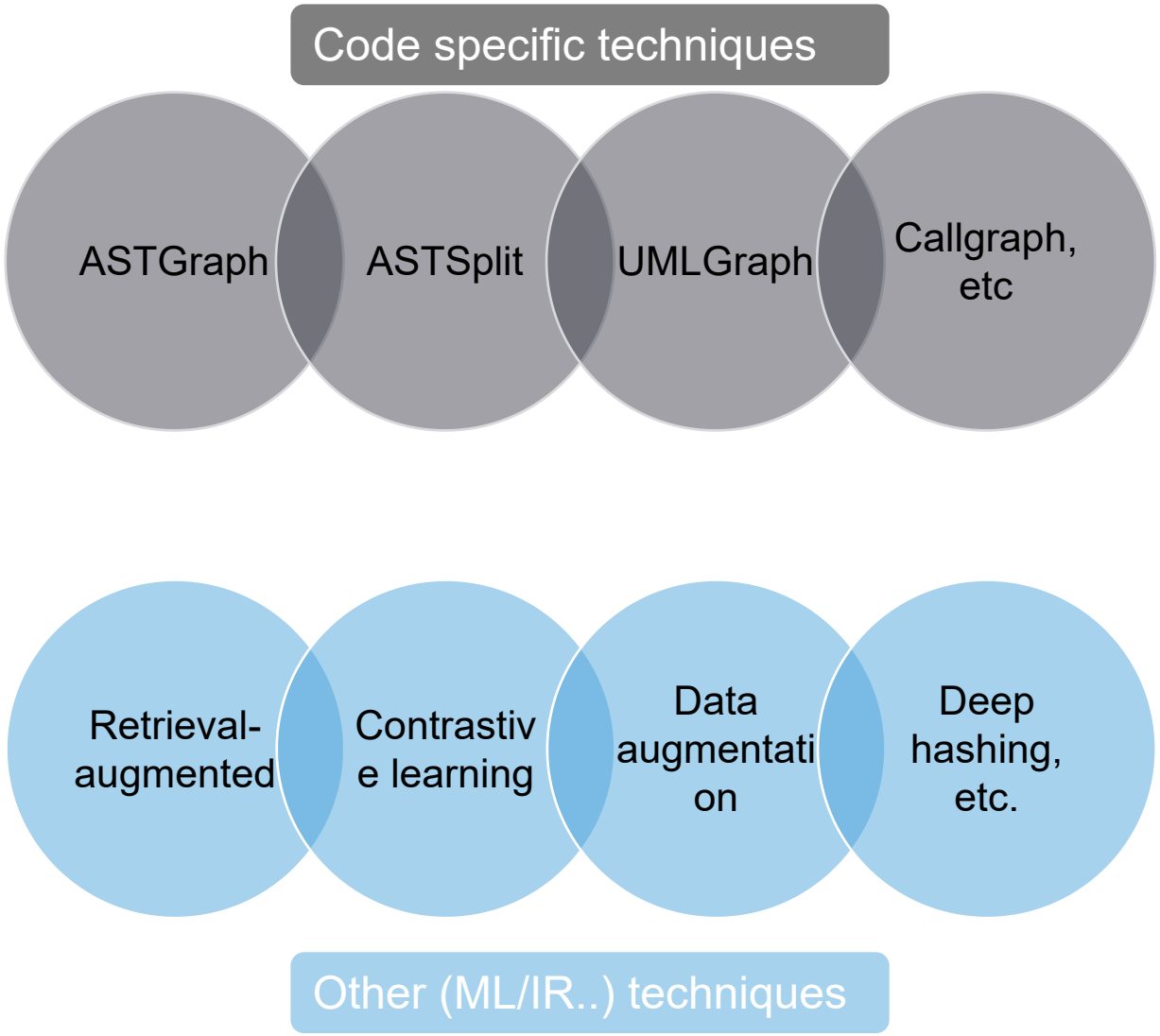
测试用例生成技术

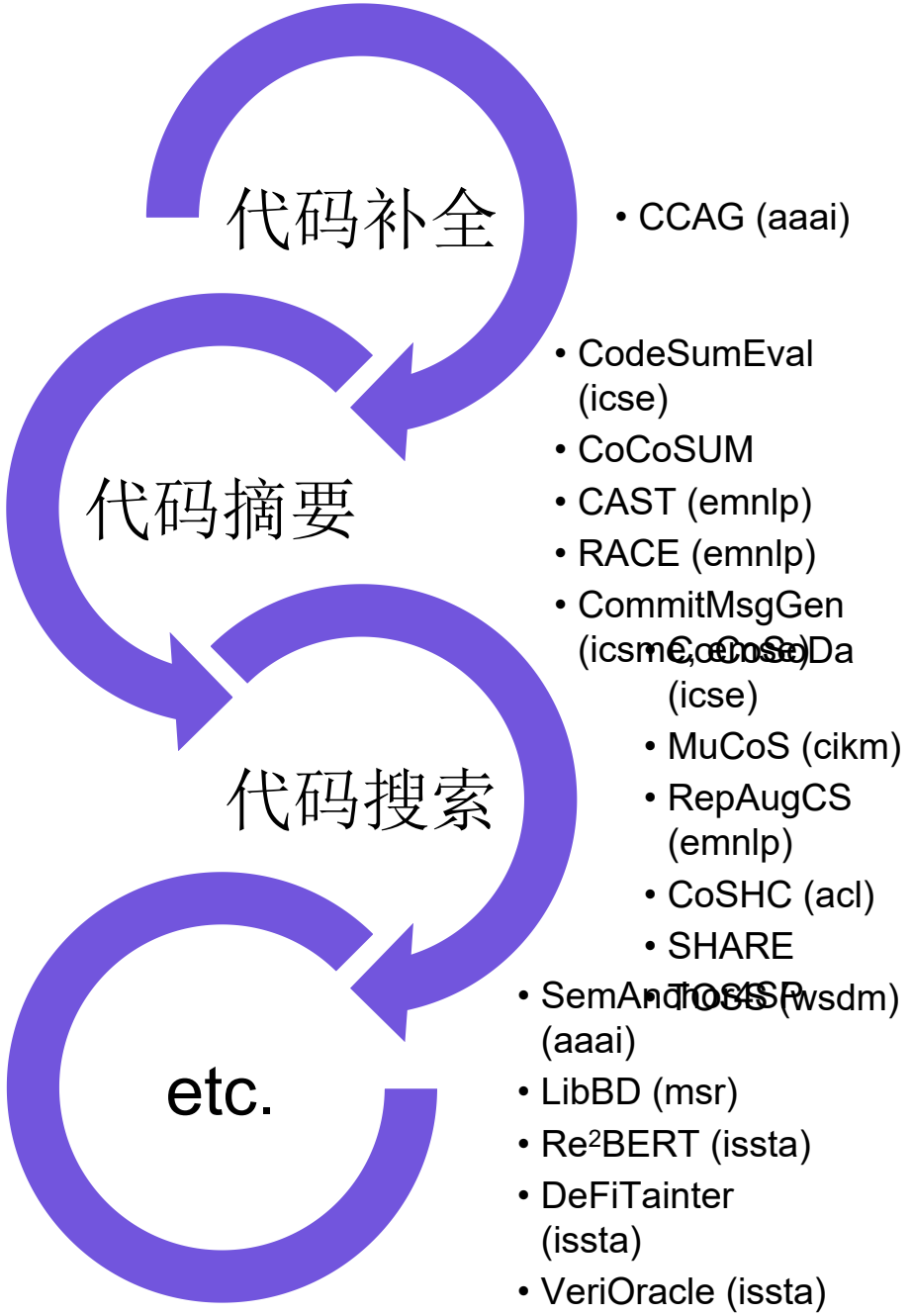
测试用例生成系统会自动生成一些测试用例测试我们实现的代码是否符合需求。

.....

这些技术的核心在于对代码的分析和建模。代码的特性例如丰富的结构、严格的语法、复杂的依赖关系和灵活的实现方式给代码的分析和建模带来诸多挑战。

任务特定模型上的探索之路

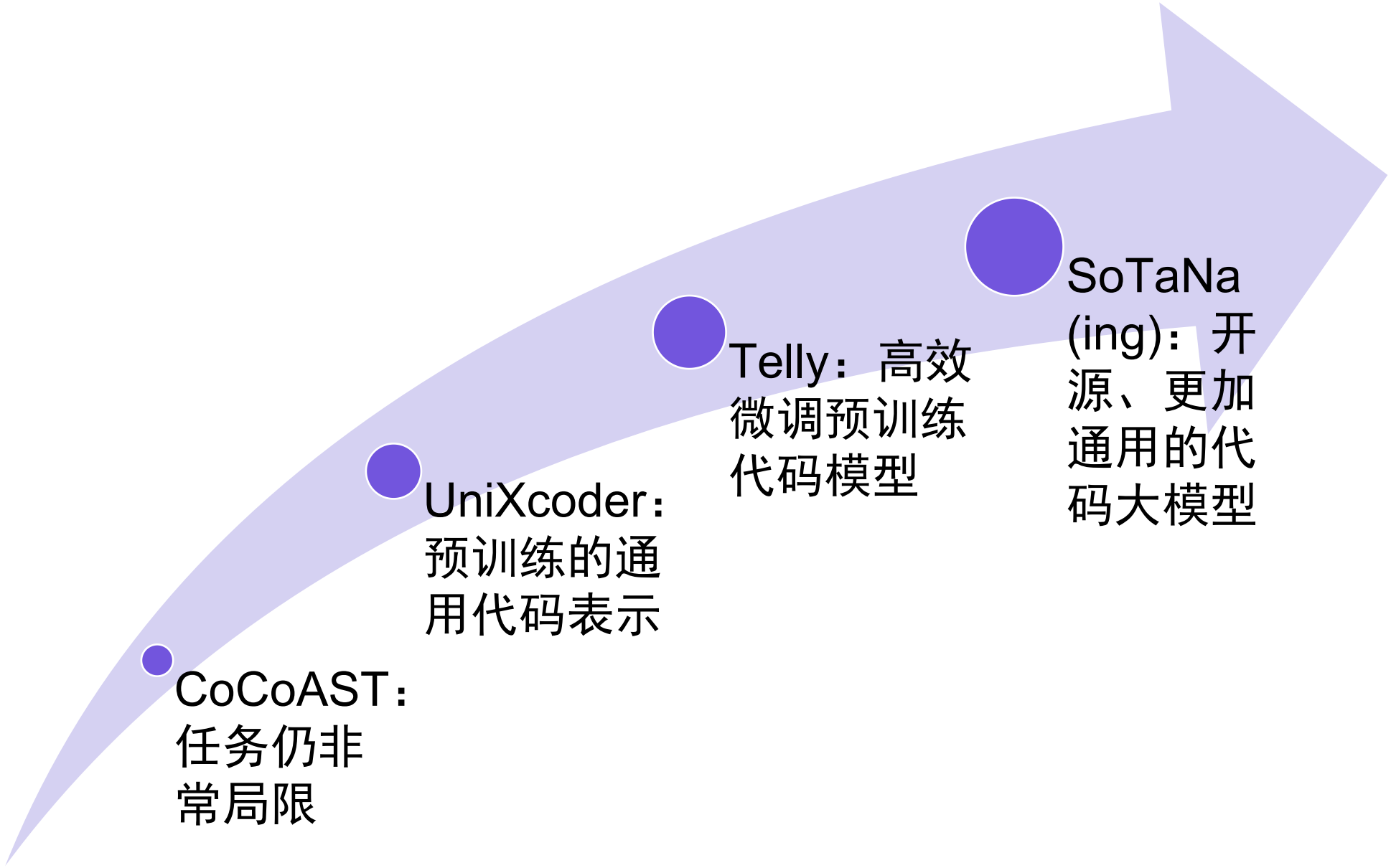




- CCAG (aaai)
- CodeSumEval (icse)
- CoCoSUM
- CAST (emnlp)
- RACE (emnlp)
- CommitMsgGen (icsme, emse)
- CoSeDa (icse)
- MuCoS (cikm)
- RepAugCS (emnlp)
- CoSHC (acl)
- SHARE (wsdm)
- SemAnTOSS (aaai)
- LibBD (msr)
- Re^{2}BERT (issta)
- DeFiTainter (issta)
- VeriOracle (issta)

目录

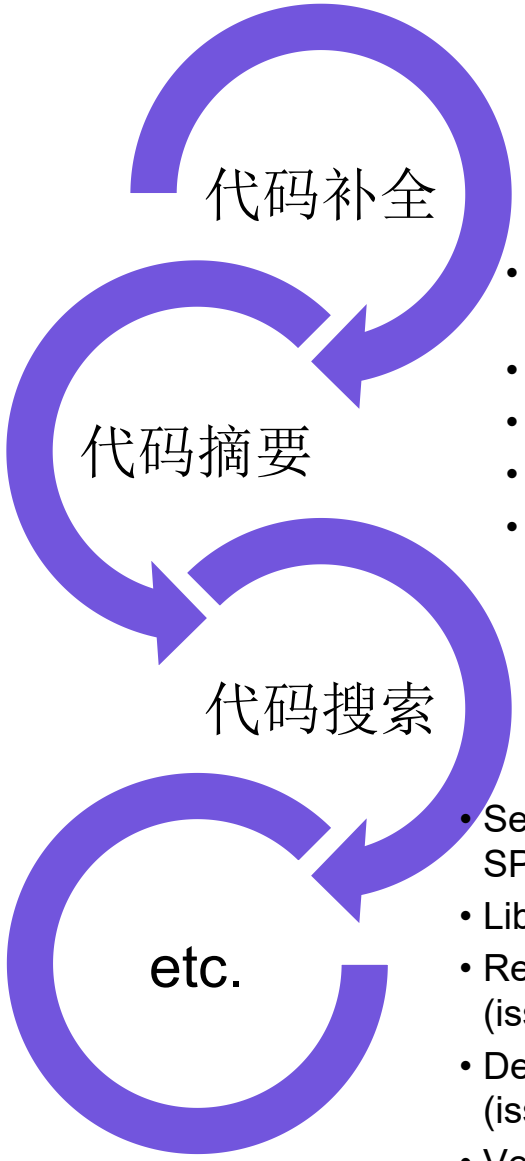
软工通用模型的探索之路



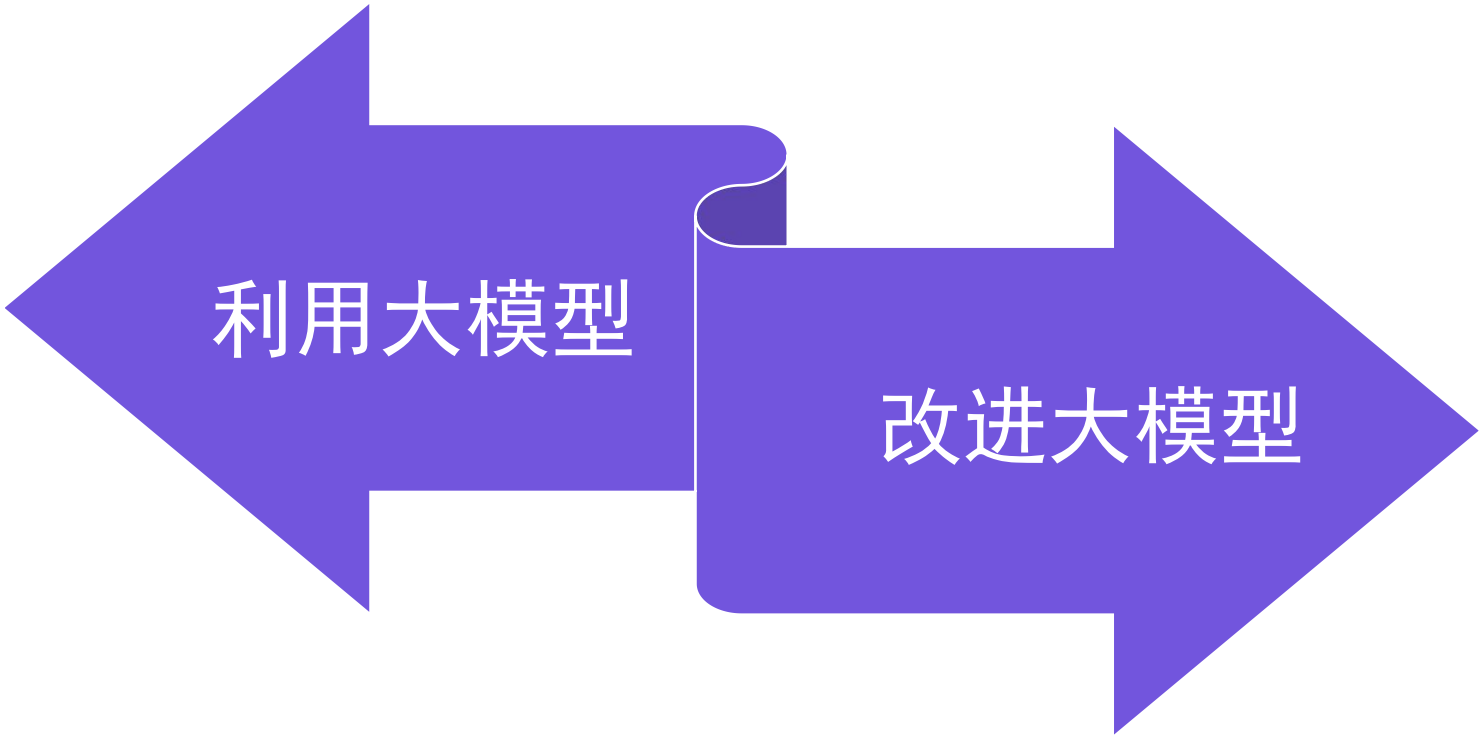
大模型时代，以前在任务特定模型上的探索还有意义吗？

目录

软件工程(incl.任务特定方法)+大模型的求索之路



- CCAG (aaai)
- CodeSumEval (icse)
- CoCoSUM
- CAST (emnlp)
- RACE (emnlp)
- CommitMsgGen (icsme, emse)
- CoCoSoDa (icse)
- MuCoS (cikm)
- RepAugCS (emnlp)
- CoSHC (acl)
- SHARE
- TOSS (wsdm)
- SemAnchor4SP(aaai)
- LibBD (msr)
- Re²BERT (issta)
- DeFiTainter (issta)
- VeriOracle (issta)



风险

[Submitted on 20 Aug 2021 (v1), revised 23 Aug 2021 (this version, v2), latest version 16 Dec 2021 (v3)]

An Empirical Cybersecurity Evaluation of GitHub Copilot's Code Contributions

Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, Ramesh Karri

There is burgeoning interest in designing AI-based systems to assist humans in designing computing systems, including tools that automatically generate computer code. The most notable of these comes in the form of the first self-described `AI pair programmer', GitHub Copilot, a language model trained over open-source GitHub code. However, code often contains bugs - and so, given the vast quantity of unvetted code that Copilot has processed, it is certain that the language model will have learned from exploitable, buggy code. This raises concerns on the security of Copilot's code contributions. In this work, we systematically investigate the prevalence and conditions that can cause GitHub Copilot to recommend insecure code. To perform this analysis we prompt Copilot to generate code in scenarios relevant to high-risk CWEs (e.g. those from MITRE's "Top 25" list). We explore Copilot's performance on three distinct code generation axes -- examining how it performs given diversity of weaknesses, diversity of prompts, and diversity of domains. In total, we produce 89 different scenarios for Copilot to complete, producing 1,692 programs. Of these, we found approximately 40% to be vulnerable.

研究人员总共为Copilot生成了89个不同的场景，生成了1692个程序，发现40%程序存在安全漏洞。

https://arxiv.org/abs/2108.09293v2?utm_source=labnotes.org

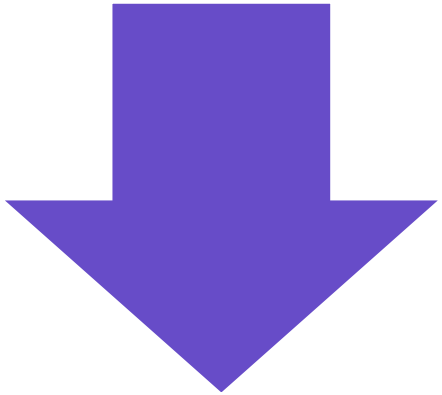
GitHub Copilot 给我补了一张谁的身份证上来？？？

翻译推文

21年8月6日, 22:25 · Twitter Web App

166 转推　**39** 引用推文　**875** 喜欢

软件工程(incl.任务特定方法)+大模型的求索之路



利用大模型

- ChatGPT漏洞检测
- ChatGPT代码搜索
- Evaluation metrics
- Benchmark -> 影响力

改进大模型

- 多模态代码表示及对齐
  提升LLM推理能力
- 探索Code-Tailored
  Encoder大模型
- Security for LLMs