面向Coding Agent 的可验证任务合成与反馈校准

From Benchmark Diagnosis to Verifiable Data Scaling

陈国鑫

中国人民大学

Outlines

AI Box

- Why Beyond SWE-Bench?
- BeyondSWE: Seeing the Boundary
- Data Scaling: Learning from Verifiable Tasks
- Towards Long-Horizon Code Agents

Why Beyond SWE-Bench?

AI Box

- 真实场景经常需要外部知识、版本约束与全局改动

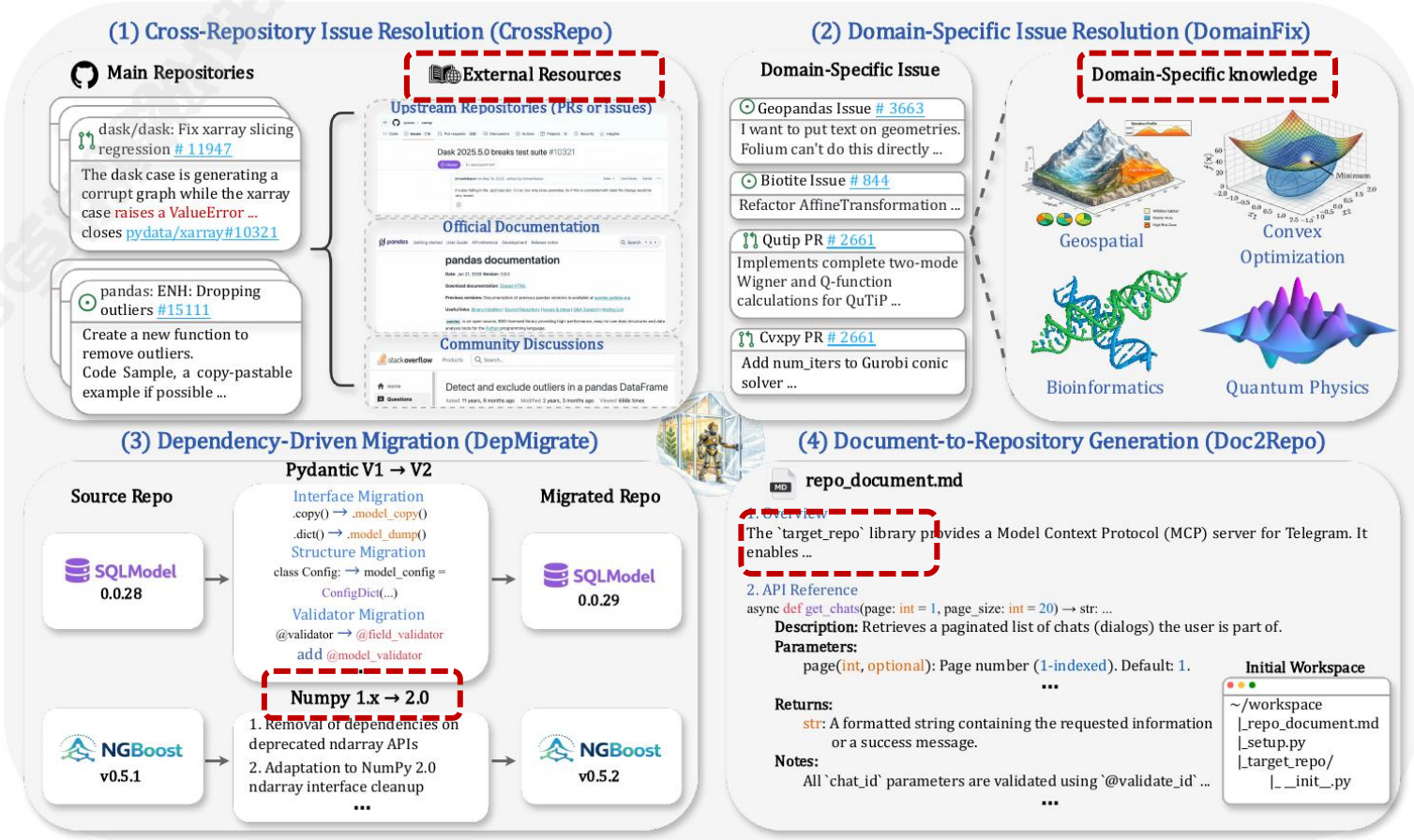
Benchmark	Scope		Statistics		
	Resol.	Knowledge	#Repo	#Files	#Lines
SWE-bench-Verified	Local Func	Within Repo	12	1.3	11.6
SWE-bench-Live	Local Func	Within Repo	223	2.7	65.1
SWE-bench Pro	Local Func	Within Repo	41	4.1	107.4
CrossRepo	Local Func	Cross Repo	67	4.1	190.7
DomainFix	Local Func	Domain	12	4.2	157.6
DepMigrate	Global Repo	Official Docs	120	8.4	281.6
Doc2Repo	Global Repo	Human Spec	50	26.8	3528.4
BeyondSWE	**Mix**	**Mix**	**246**	**10.9**	**1039.6**

Table 1. Comparison with existing SWE benchmarks

BeyondSWE: Seeing the Boundary

AI Box

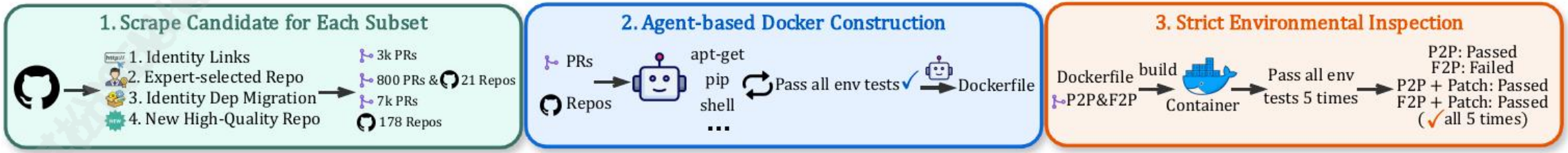
- Knowledge scope × resolution scope：外部知识 + 全局修改



避免环境噪声

AI Box

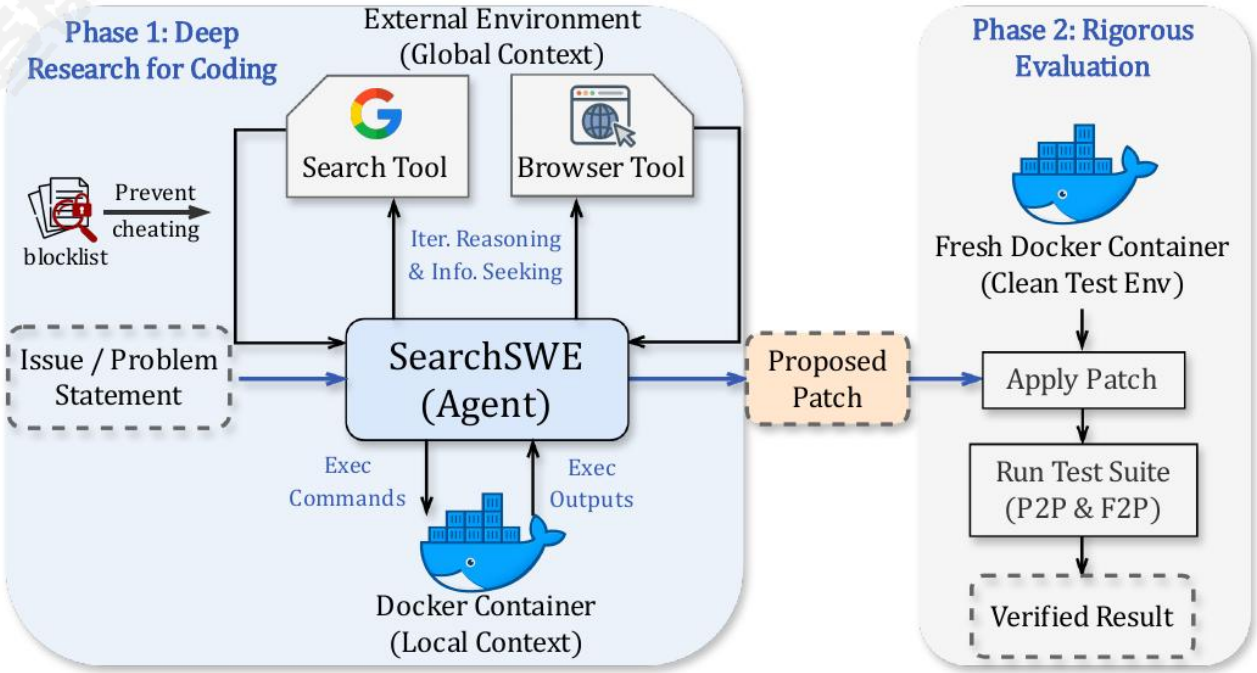
- Docker 构建 + 五次稳定检查 + fresh container 验证
  - pytest的执行具有随机性 (固定了测试顺序）
  - Docker本身测试具有不稳定性（去除边界案例）
  - agent可能修改test文件（确保pytest不被修改+fresh container验证）



Deep Search for Coding

AI Box

- 黑盒：Codex
- 白盒：SearchSWE: 唯一变量是 Search/Fetch, blocklist 防止作弊



实验结果：仍未饱和

AI Box

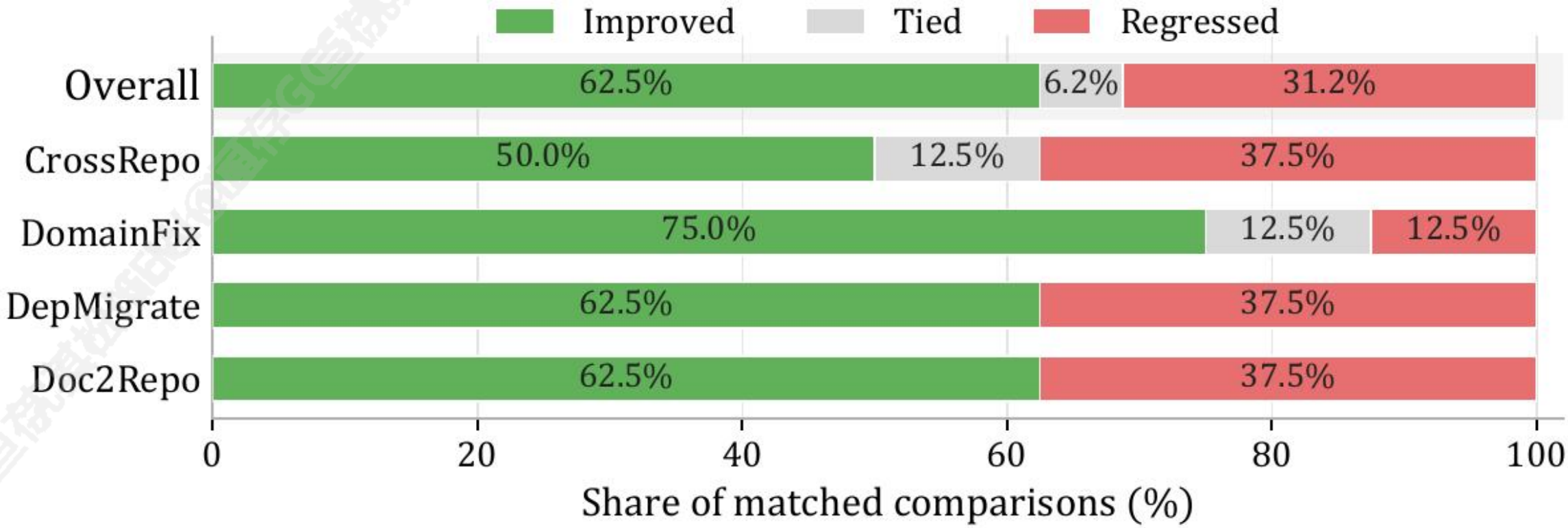
➢ Finding 1: OpenHands 46.12； SearchSWE 48.73； Codex 56.65

Model	CrossRepo %Resolved	DomainFix %Resolved	DepMigrate %Resolved	Doc2Repo Pass Rate	Doc2Repo #(Alm.) Corr.	AVG
Codex						
GPT-5.4 (xhigh) w/ SearchSWE Prompt	**55.17** +8.9	**61.11** +19.4	**48.59** +4.2	**61.74** +0.1	**(7) / 2**	**56.65** +8.2
GPT-5.4 (xhigh) w/ Default Prompt	46.23	41.67	44.38	61.64	**(7) / 2**	48.48
OpenHands						
DeepSeek-V4-Pro(Max) (DeepSeek-AI, 2026)	44.00	**38.89**	**44.38**	**57.20**	(5) / 1	**46.12**
GLM-5 (Zeng et al., 2026)	**44.67**	33.33	42.13	56.76	(7) / 3	44.22
Qwen3.5-Plus (Qwen Team, 2026)	41.50	**38.89**	41.01	52.41	(3) / 1	43.45
Gemini 3 Pro (Google, 2025b)	41.50	31.94	41.81	52.03	**(8) / 2**	41.82
GPT-5.4 (OpenAI, 2026)	43.00	27.78	37.50	56.30	(5) / 3	41.15
Kimi-K2.5 (Team et al., 2026)	40.50	34.72	39.89	51.36	(3) / 1	41.62
Seed-Coder-2.0(Seed et al., 2025)	39.50	24.29	33.15	55.54	(5) / 2	38.12
MiniMax-M2.5 (MiniMax, 2026)	40.00	25.00	37.64	46.57	(5) / 1	37.30
SearchSWE						
DeepSeek-V4-Pro(Max) (DeepSeek-AI, 2026)	**48.50** +4.5	**43.06** +4.2	**47.19** +2.8	56.16 -1.0	(5) / 2	**48.73** +2.6
GLM-5 (Zeng et al., 2026)	43.88 -0.8	35.94 +2.6	47.13 +5.0	**60.05** +3.3	(7) / 3	46.75 +2.5
Qwen3.5-Plus (Qwen Team, 2026)	41.50	34.72 -4.2	39.89 -1.1	54.90 +2.5	(3) / 1	42.75 -0.7
Gemini 3 Pro (Google, 2025b)	41.12 -0.4	39.44 +7.5	44.07 +2.3	50.73 -1.3	(4) / 2	43.84 +2.0
GPT-5.4 (OpenAI, 2026)	45.00 +2.0	31.94 +4.2	38.76 +1.3	58.35 +2.1	**(7) / 4**	43.51 +2.4
Kimi-K2.5 (Team et al., 2026)	43.00 +2.5	34.72	37.64 -2.3	52.22 +0.9	(5) / 2	41.90 +0.3
Seed-Coder-2.0 (Seed et al., 2025)	43.15 +3.7	30.43 +6.1	35.80 +2.7	49.67 -5.9	(3) / 1	39.76 +1.6
MiniMax-M2.5 (MiniMax, 2026)	39.00 -1.0	31.94 +6.9	37.08 -0.6	50.66 +4.1	(3) / 0	39.67 +2.4

Finding 2: Search 有用，但不稳定

AI Box

- 20/32 提升；31.2% 回退；DomainFix 最受益；
- search和coding各自发展迅猛，但融合的能力无法自然涌现

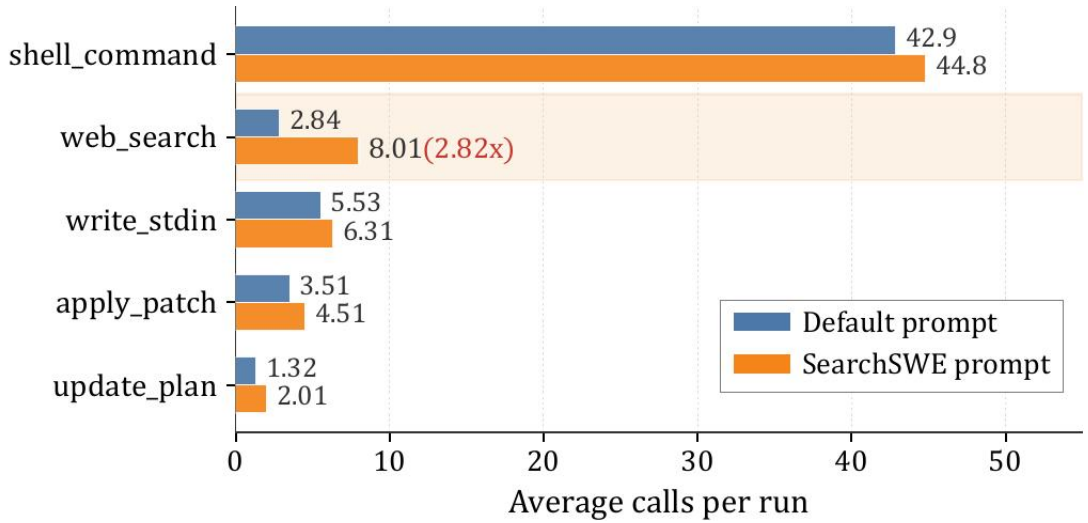


Finding 3: 对于codex，仍需要显式引导使用搜索

AI Box

- web_search 2.84→8.01；coding action 基本相近

Model	CrossRepo %Resolved	DomainFix %Resolved	DepMigrate %Resolved	Doc2Repo Pass Rate	Doc2Repo #(Alm.) Corr.	AVG
Codex						
GPT-5.4 (xhigh) w/ SearchSWE Prompt	**55.17** +8.9	**61.11** +19.4	**48.59** +4.2	**61.74** +0.1	**(7) / 2**	**56.65** +8.2
GPT-5.4 (xhigh) w/ Default Prompt	46.23	41.67	44.38	61.64	**(7) / 2**	48.48



Finding 4: 更多搜索 ≠ 更高收益

AI Box

- 关键是检索到的证据是否精确 （类似原来search 任务）
- 是否能被转成局部可执行改动（search转换为code代码）

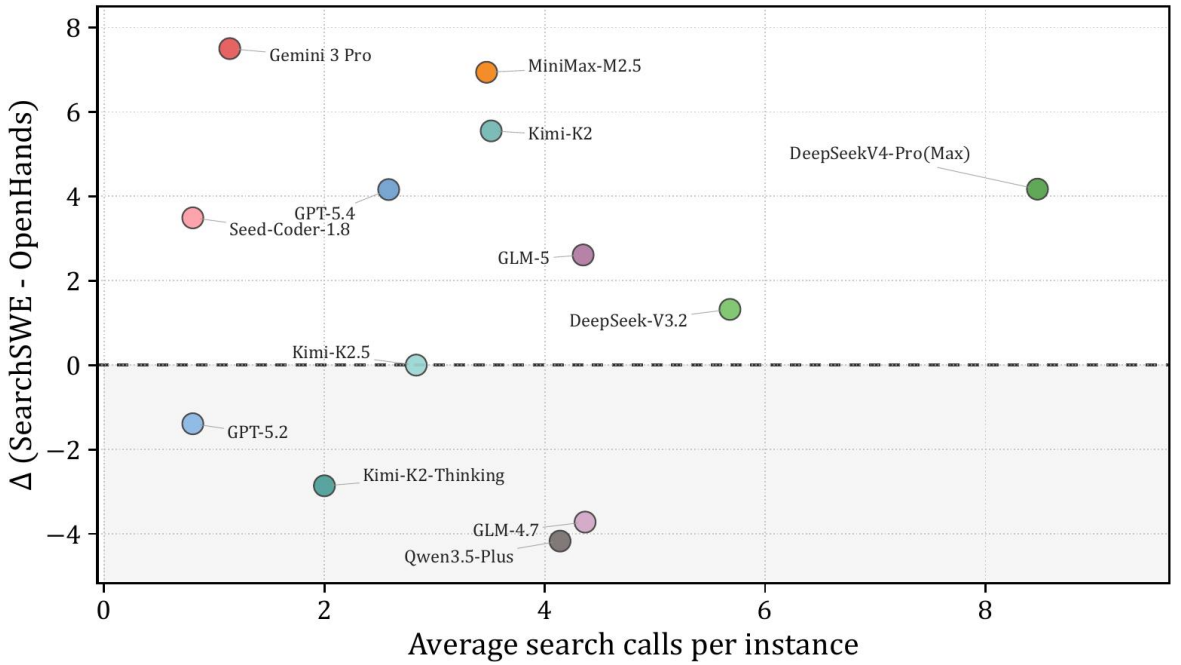


Figure 6a: Search effort vs gain on DomainFix

Finding 5: Doc2Repo任务里搜索会变钝

AI Box

➢Doc2Repo 的主要约束来自开发文档；外部搜索容易转移注意力

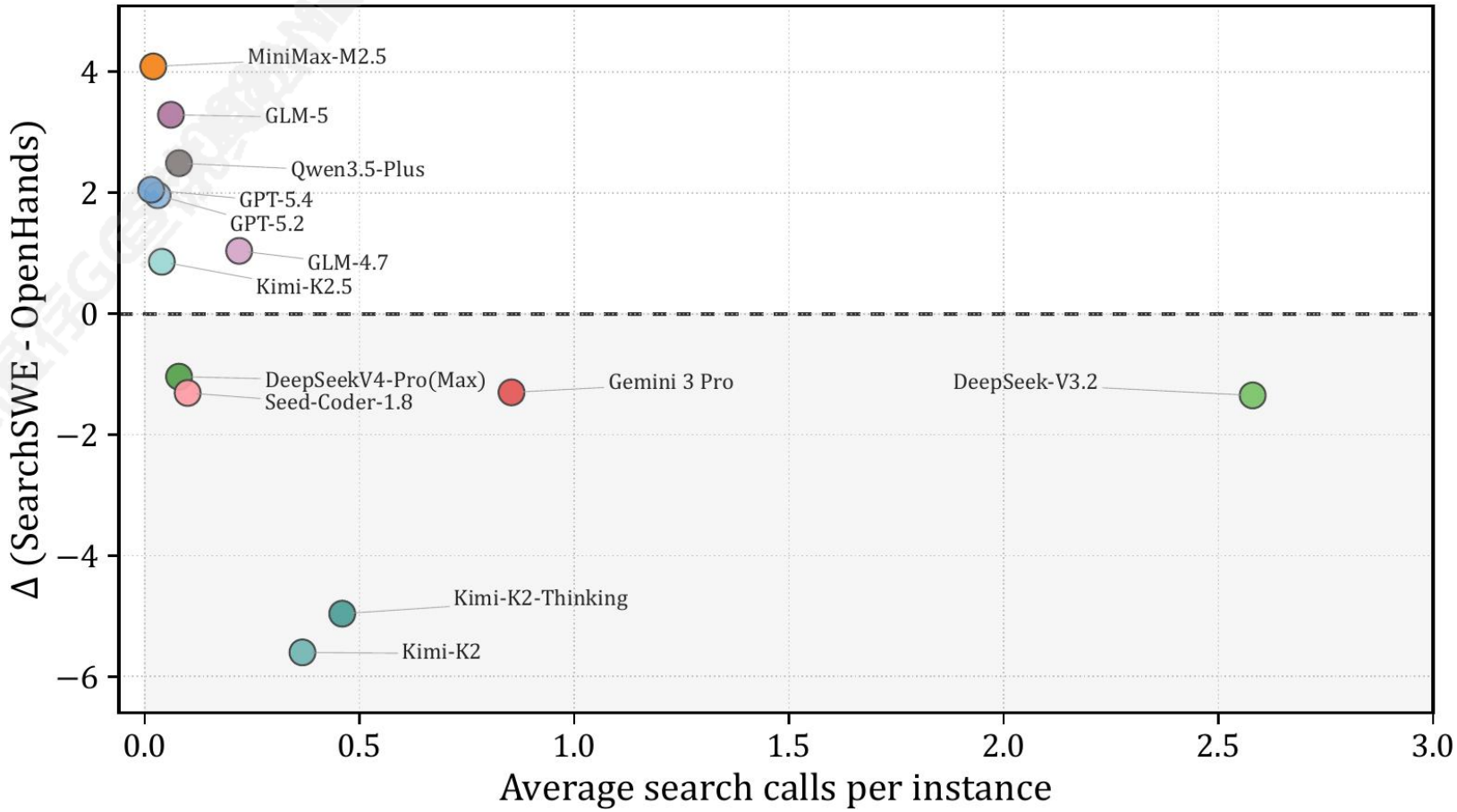


Figure 6b: Search effort vs gain on Doc2Repo

Finding 6: 预算更大不等于更有效

AI Box

- 高 token 轨迹常消耗在重复探索、噪声搜索和无效修补上

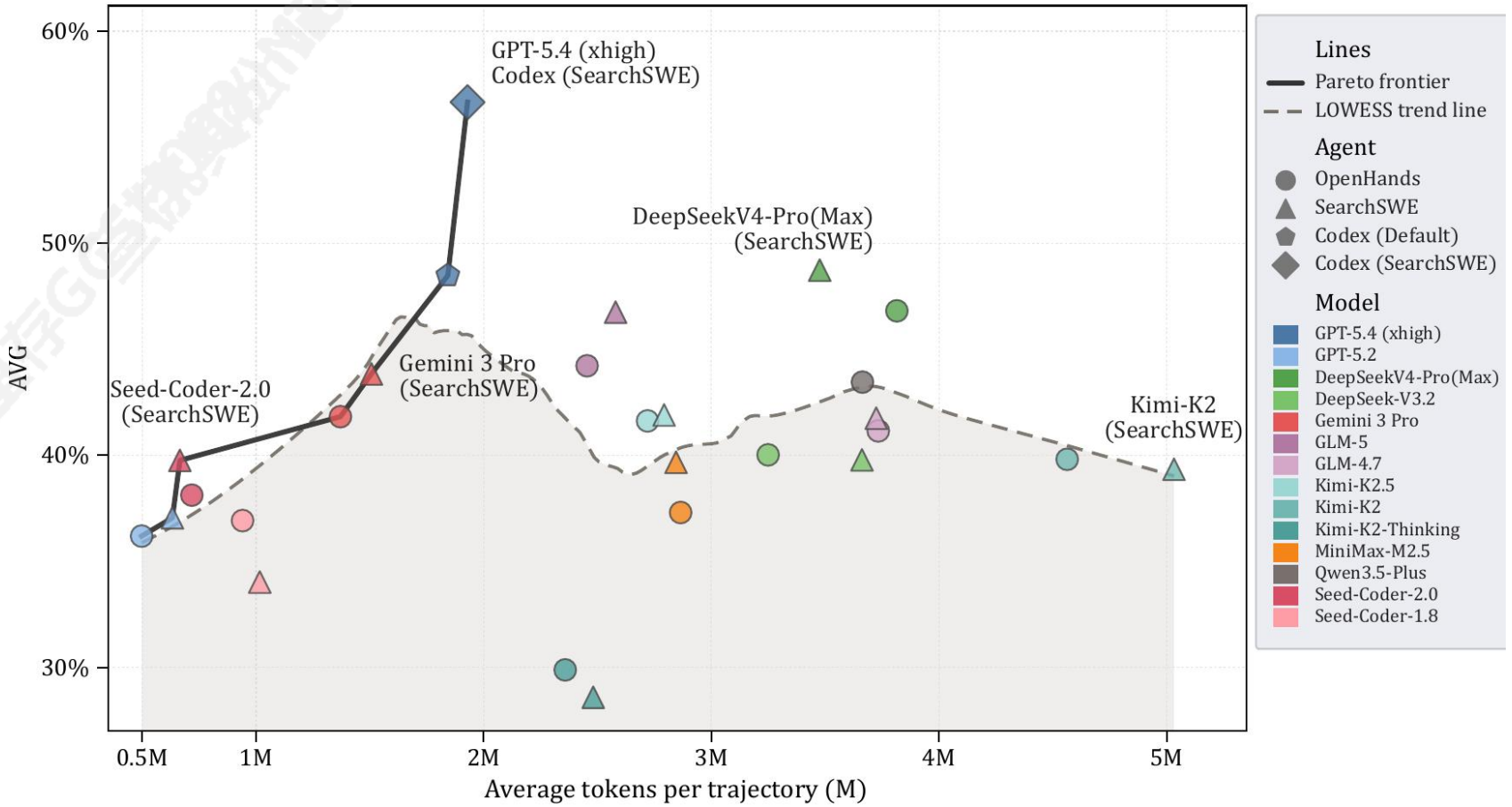


Figure 7b: Token budget vs performance

Finding 7: 失败在证据转代码之间

AI Box

- 三道关： granularity / version grounding / semantic filtering

Failure Mode	Instance	Search Behavior	Local Failure	Takeaway
Evidence availability	unidata_siphon_pr234	Search returns high-level documentation instead of source-level backend logic	Agent implements a brittle interpretation of an ambiguous API description	Search must recover evidence at the right technical granularity
Version grounding	behave_behave-django_pr162	Agent searches around an unsupported newer Django assumption instead of local constraints	Agent applies a method pattern incompatible with the repository's legacy test lifecycle	External evidence must be filtered through local dependency versions
Evidence filtering	abravalheri_validate-pyproject_pr105	Keyword-matched results drift into unrelated domains due to overloaded terminology	Agent falls back to a generic plugin-registration pattern that creates test side effects	Agents need semantic source discrimination, not only retrieval

Table 3. Summary of search-augmented coding failure modes analyzed in the case studies.

Outlines

AI Box

- Why Beyond SWE-Bench?
- BeyondSWE: Seeing the Boundary
- Data Scaling: Learning from Verifiable Tasks
- Towards Long-Horizon Code Agents

Code Agent数据规模化方法

AI Box

- 以往的数据集的规模化往往受限于仓库数目或者合成数据。

Dataset	Exec. instances	Primary Source	Repo.	Traj.
R2E-Gym (Jain et al., 2025)	4.6k	Synthetic	10	3.3k
SWE-Gym (Pan et al., 2024)	2.4k	Real	11	491
SWE-smith (Yang et al., 2025a)	50k	Synthetic	128	5k
SWE-Mirror (Wang et al., 2025a)	60k	Synthetic	40	12k
SWE-rebench (Badertdinov et al., 2025)	7.5k	Real	3.5k	N/A
Scale-SWE	**100k**	Real	**5.2k**	**71k**

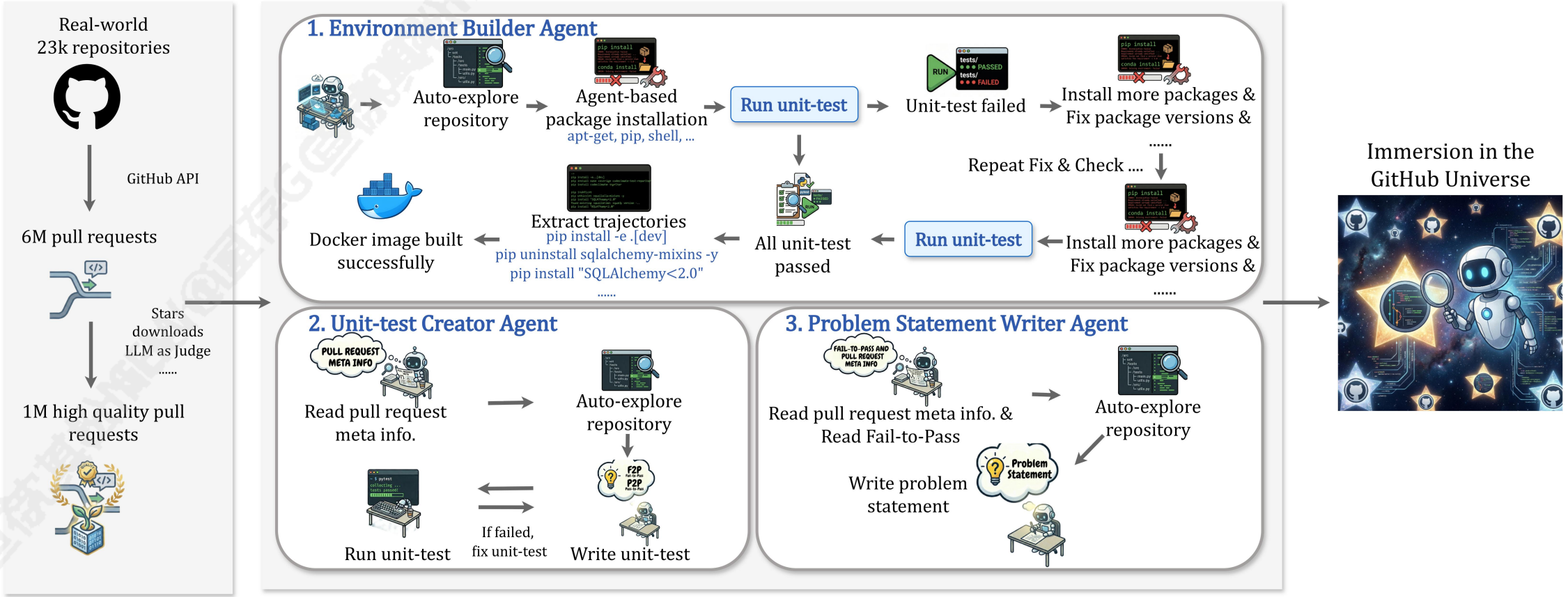
- 为什么以往方法无法规模化高质量数据呢？
  - 多样的GitHub代码仓库环境没有统一的方法来配置。
  - 众多高质量的pull request并没有作者自带的单元测试来作为Fail-to-Pass。
  - pull request本质是提出者解决完问题之后提出的，因此pull request的描述往往会泄露解决bug的方案和路径，另外有大量pull request不会些描述。

Immersion in the GitHub Universe: Scaling Coding Agents to Mastery, Arxiv, 2026

Scale-SWE：Code Agent数据规模化方法

AI Box

- 针对数据规模化痛点，提出沙盒中实时与环境交互的多agent系统
  - 主要包括：环境构建Agent，单元测试创建Agent，问题描述创建Agent。

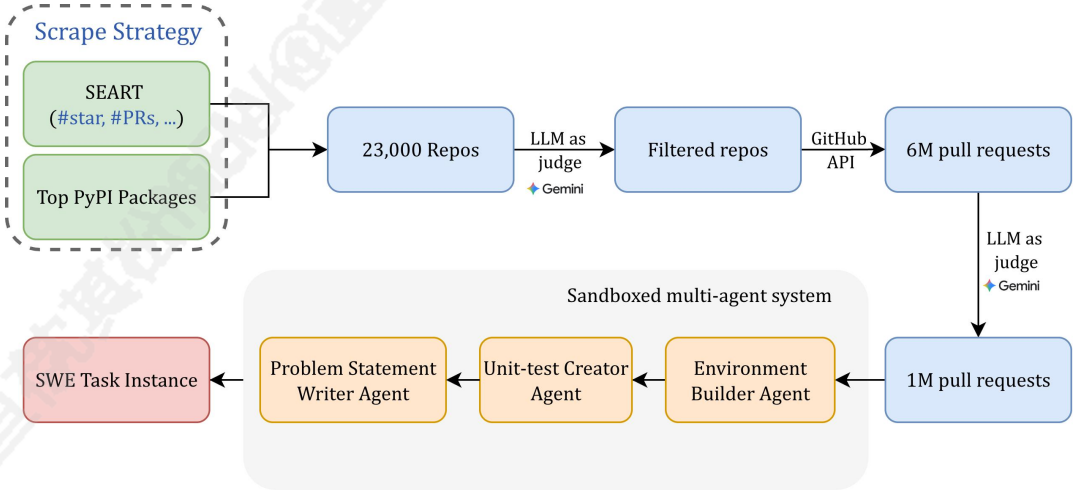


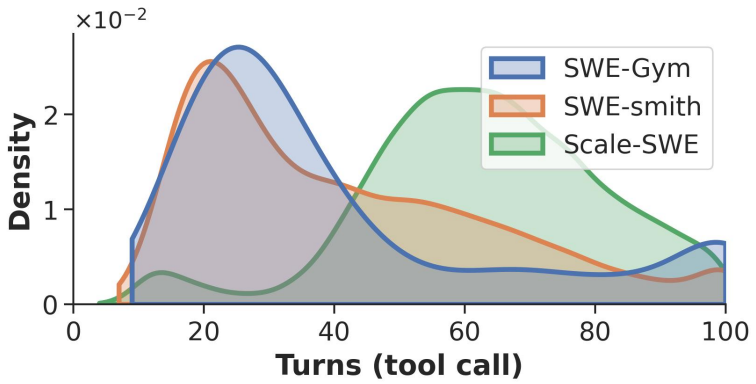
Immersion in the GitHub Universe: Scaling Coding Agents to Mastery, Arxiv, 2026

Scale-SWE：Code Agent数据规模化方法

AI Box

- 具体地
  - 1. 从GitHub 6百万条pull request与2万3千个仓库出发，开始重重过滤
  - 2. 经过环境构建Agent，Agent与环境实时交互构建环境，根据反馈调整环境
  - 3. 经过单元测试创建Agent，给没有单元测试的pull request编写单元测试
  - 4. 经过问题描述创建Agent，创建不泄露解决方案的问题描述





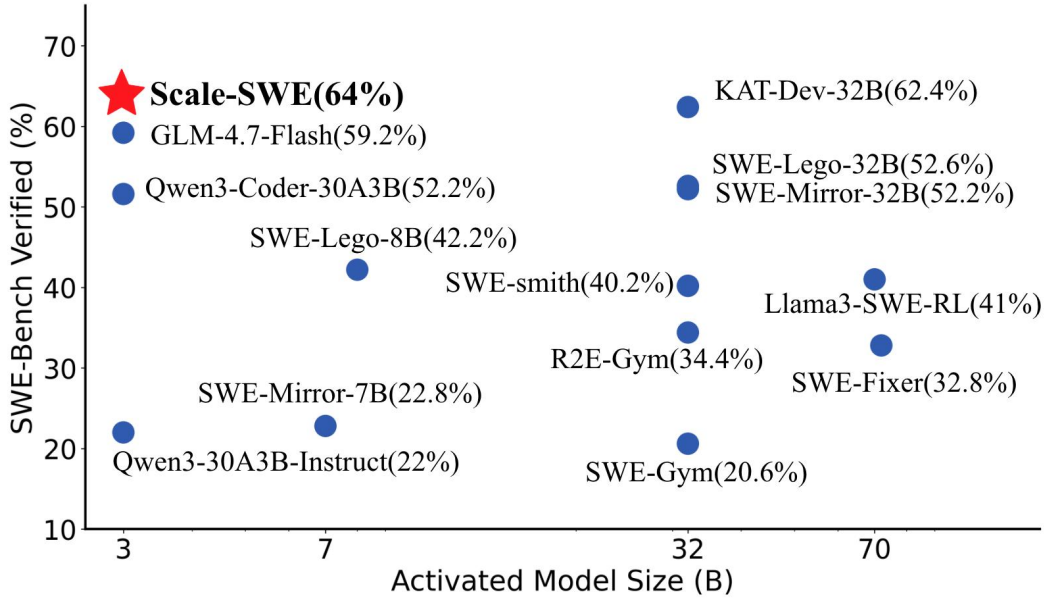
构造不泄露解决方案的问题描述，能够极大增加题目质量

Immersion in the GitHub Universe: Scaling Coding Agents to Mastery, Arxiv, 2026

Scale-SWE：Code Agent数据规模化方法

AI Box

- 效果
  - 构建了10万条SWE数据（开源了2万条数据）蒸馏DeepSeek v3.2得到7万条轨迹
  - 基于Qwen3-30BA3B-Instruct进行训练，在同等规模大小下模型效果超过GLM-4.7-Flash、Qwen3-Coder
  - 与更大模型相比也超过代码领域模型KAT-Dev-32B，以及SWE-RL



Immersion in the GitHub Universe: Scaling Coding Agents to Mastery, Arxiv, 2026

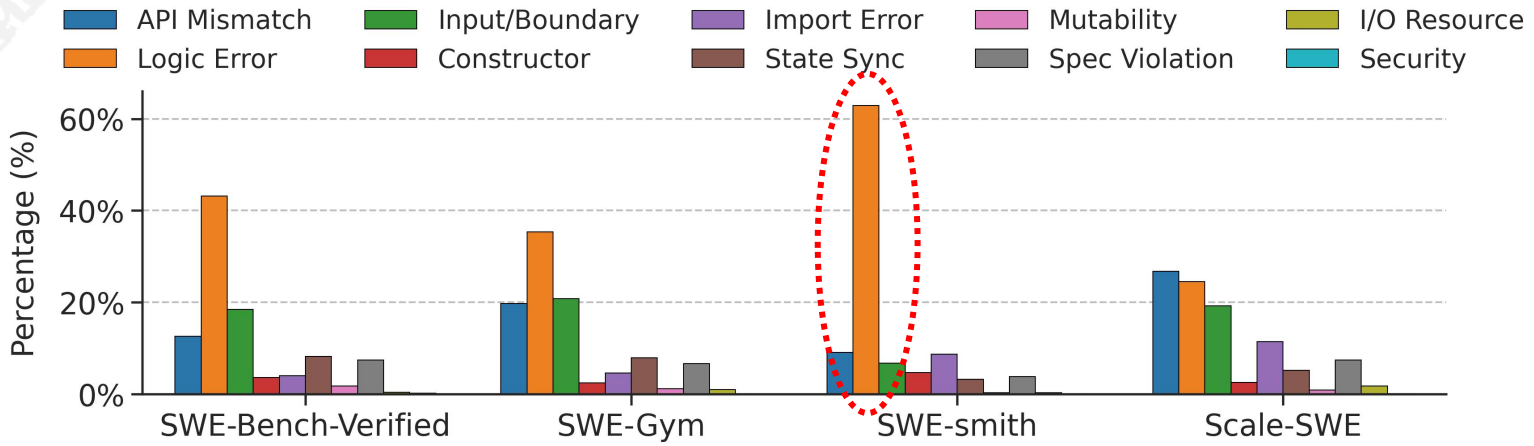
Scale-SWE：Code Agent数据规模化方法

AI Box

- 即使以往合成数据的规模更大，效果仍然不如以往小规模真实数据

Dataset Name	SWE-bench Verified
SWE-Gym	54.8
SWE-smith	54.6
Scale-SWE	**64.0**

- 合成数据往往由于合成方法与仓库候选数目的限制导致数据不够多样



Immersion in the GitHub Universe: Scaling Coding Agents to Mastery, Arxiv, 2026

Outlines

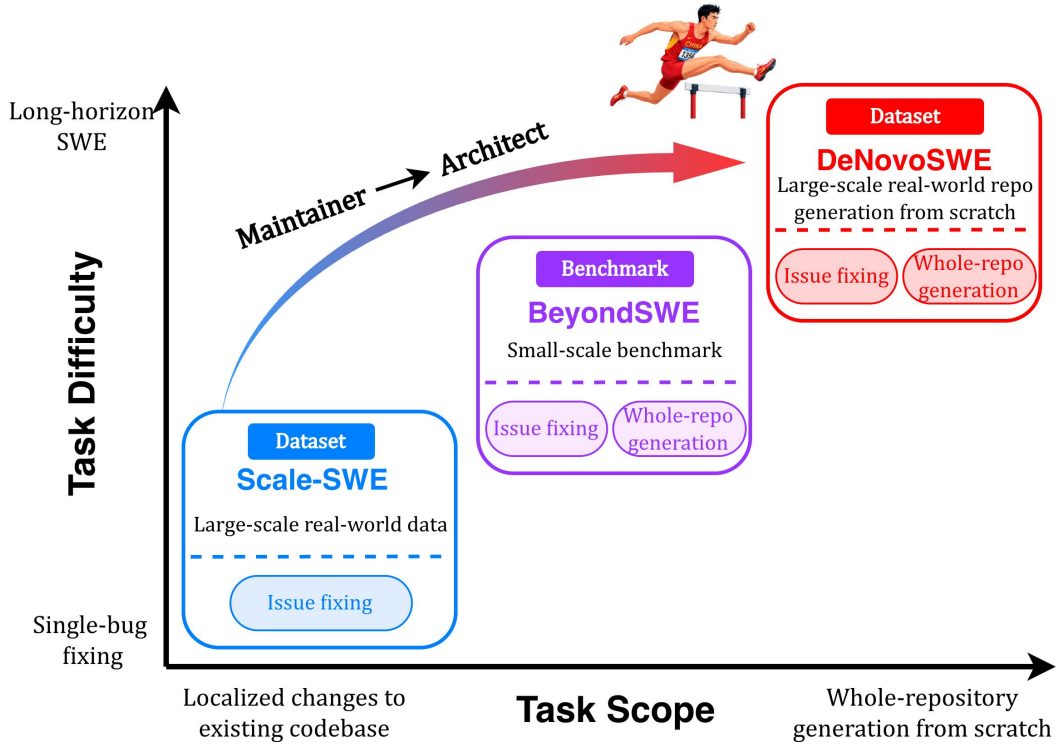
AI Box

- Why Beyond SWE-Bench?
- BeyondSWE: Seeing the Boundary
- Data Scaling: Learning from Verifiable Tasks
- Towards Long-Horizon Code Agents

DeNovoSWE：长程Code Agent数据规模化方法

AI Box

- 理想的Code Agent是什么样的？
  - handle超长工作时间、超长程context
  - 可以从零实现一个仓库来完成用户任务

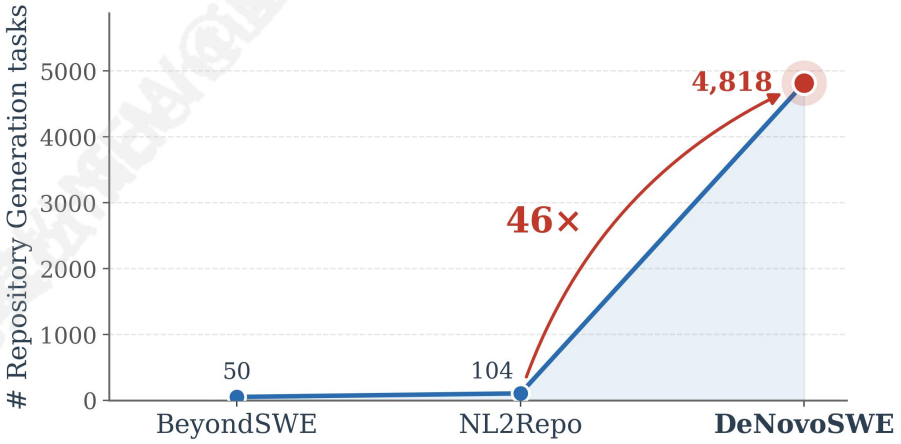


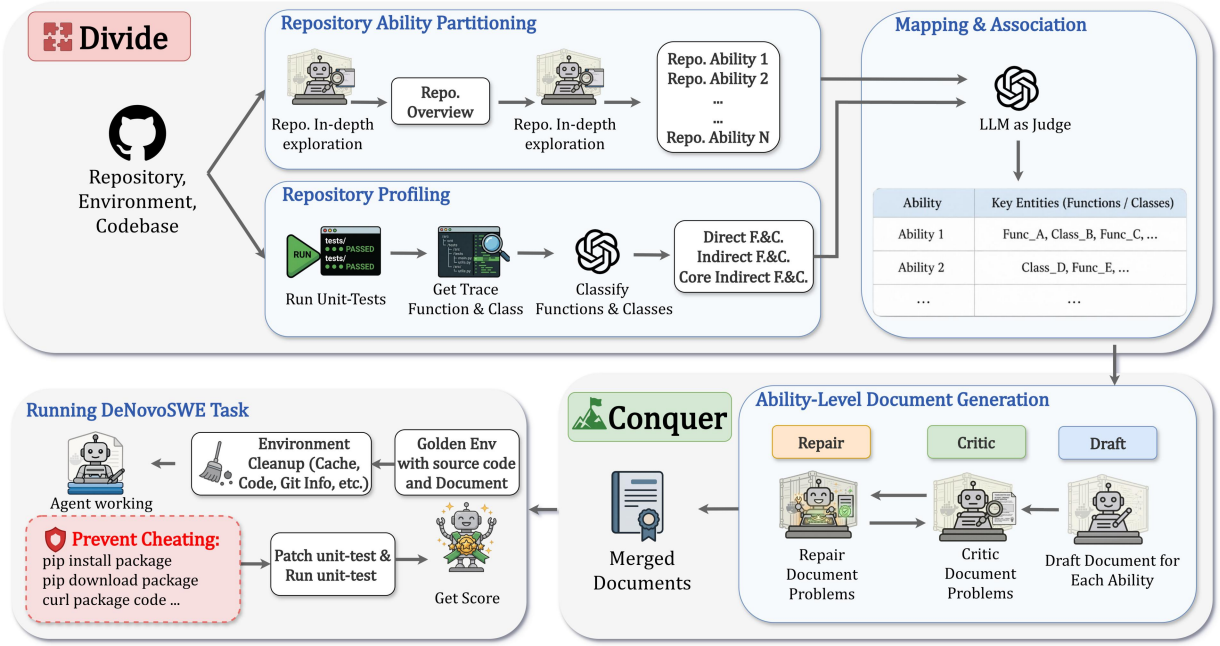
DeNovoSWE: Scaling Long-Horizon Environments for Generating Entire Repositories from Scratch, Arxiv, 2026

DeNovoSWE：长程Code Agent数据规模化方法

AI Box

- 目前的仓库从头生成的数据极其稀少
  - 整个仓库高质量文档生成十分具有挑战性
  - 高质量的文档往往同时也需要和测试样例一致
  - 为了克服以上困难采用Divide & Conquer和Critic & Repair机制



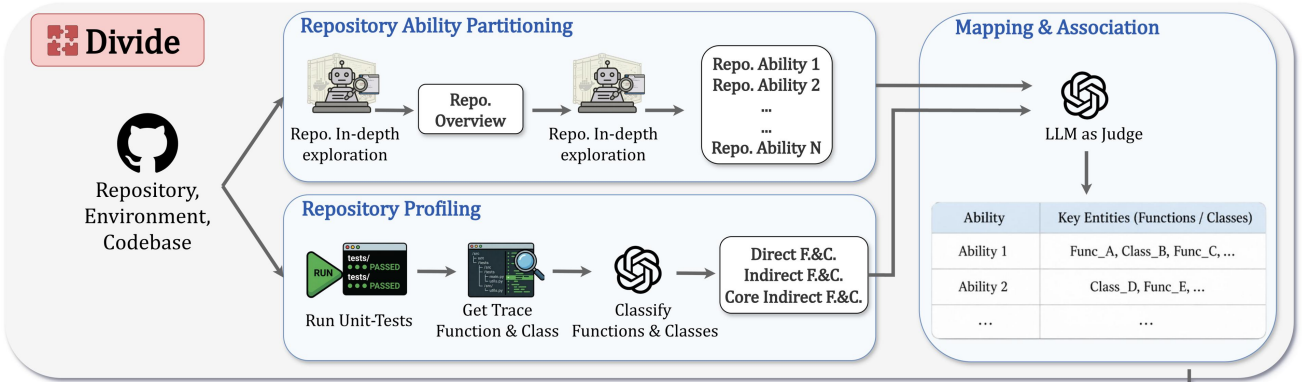


DeNovoSWE: Scaling Long-Horizon Environments for Generating Entire Repositories from Scratch, Arxiv, 2026

DeNovoSWE：长程Code Agent数据规模化方法

AI Box

- Divide & Conquer
  - 完成整个仓库的overview描述作为文档开头
  - 把整个仓库分解为多个ability，比如：数据预处理、参数loading.....
  - 根据unit-test的轨迹，提取出所有仓库原子化元素：函数 & 类
  - 把所有的函数和类归类到不同的ability中
  - 后续以ability粒度来生成文档

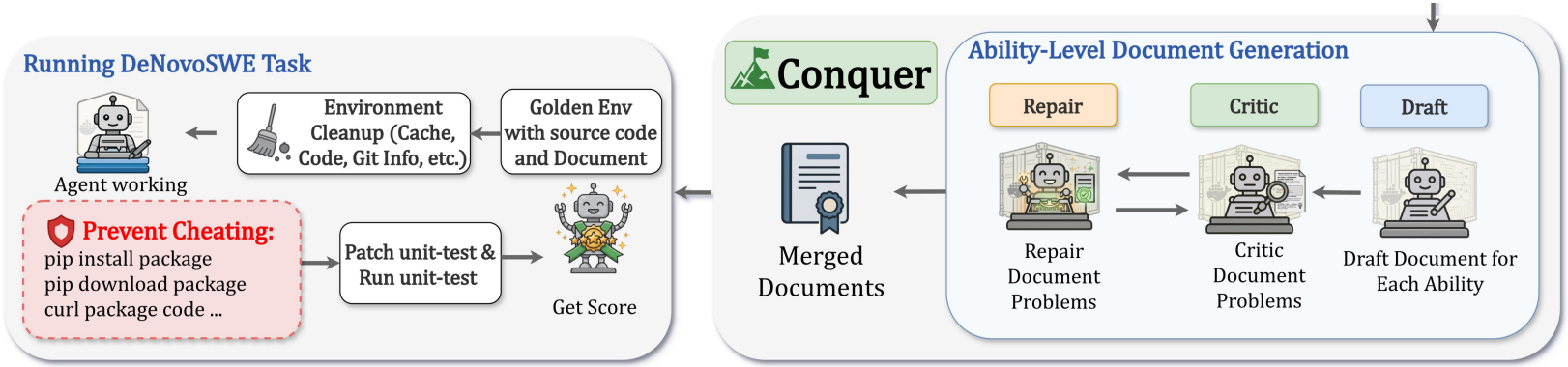


DeNovoSWE: Scaling Long-Horizon Environments for Generating Entire Repositories from Scratch, Arxiv, 2026

DeNovoSWE：长程Code Agent数据规模化方法

AI Box

- Critic & Repair
  - 生成所有ability的draft
  - critic：unit-test直接调用的函数/类、文档rubric等信息让其检查文档存在的问题
  - repair：根据critic生成的问题进行文档修改
  - 重新交给critic持续多轮文档润色
  - 最终融合overview以及所有的ability粒度的文档

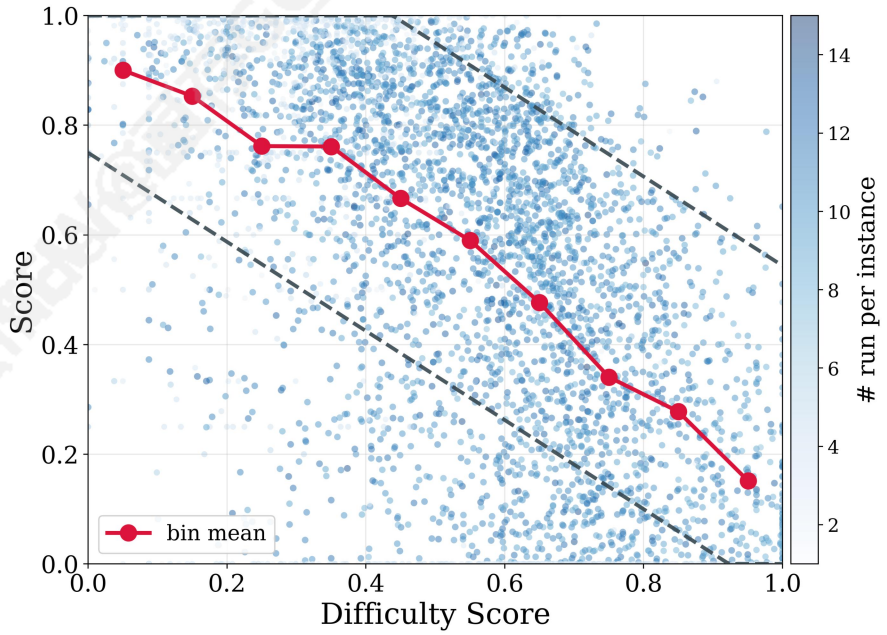


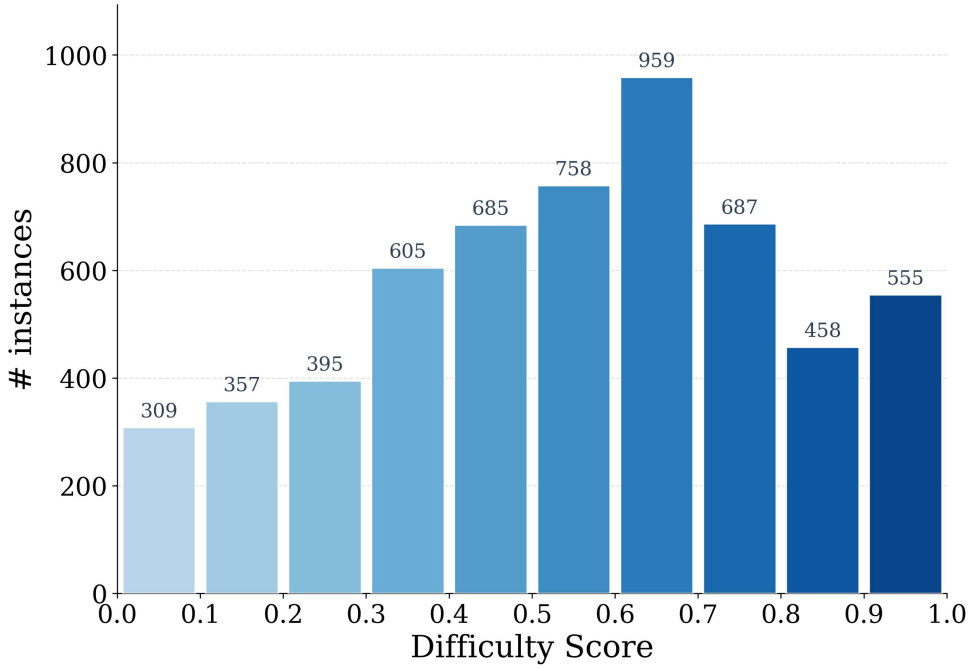
DeNovoSWE: Scaling Long-Horizon Environments for Generating Entire Repositories from Scratch, Arxiv, 2026

DeNovoSWE：长程Code Agent数据规模化方法

AI Box

- 如何选择轨迹进行训练？
  - 与普通简单的issue resolving任务不同，仓库从头生成任务比较难得到完美解决问题的轨迹。
  - 如果只取高分轨迹，将会产生很大的数据bias，往往轨迹都对应于一些简单仓库





DeNovoSWE: Scaling Long-Horizon Environments for Generating Entire Repositories from Scratch, Arxiv, 2026

DeNovoSWE：长程Code Agent数据规模化方法

AI Box

- 如何选择轨迹进行训练？
  - 于是采用基于难度的分数阈值过滤，首先对题目难度进行打分

$$d_i = w_s \cdot d_i^{\text{struct}} + w_g \cdot \tilde{\ell}_i^{(g)} + w_q \cdot \tilde{\ell}_i^{(q)}$$

$$\text{s.t.} \quad w_s, w_g, w_q \geq 0, \quad w_s + w_g + w_q = 1$$

  - 其次对不同难度的题目用不同的分数阈值过滤

Score Threshold per Difficulty Range					Doc2Repo	NL2Repo
$[0.0, 0.2)$	$[0.2, 0.4)$	$[0.4, 0.6)$	$[0.6, 0.8)$	$[0.8, 1.0]$		
Difficulty Score Independent Filtering						
0.95	0.95	0.95	0.95	0.95	0.481	0.250
0.80	0.80	0.80	0.80	0.80	0.485	0.254
0.60	0.60	0.60	0.60	0.60	0.488	0.264
Difficulty Score Dependent Filtering						
0.90	0.85	0.80	0.70	0.60	**0.500**	**0.271**

DeNovoSWE: Scaling Long-Horizon Environments for Generating Entire Repositories from Scratch, Arxiv, 2026

DeNovoSWE：长程Code Agent数据规模化方法

AI Box

➢ 最终带来了显著的效果提升

Models	Doc2Repo	NL2Repo
Proprietary Models		
GPT-5.4(CodeX) (OpenAI, 2026a)	0.617	-
GPT-5.4 (OpenAI, 2026a)	0.563	-
DeepSeek-V4-Pro (DeepSeek-AI, 2026)	0.566	-
GLM-5 (GLM-5-Team et al., 2026)	0.568	-
Seed-Coder-2.0 (Seed et al., 2025)	0.568	-
Qwen3.5-Plus (Qwen Team, 2026)	0.524	-
Gemini3-Pro (Google, 2025)	0.520	-
Qwen3-30B-A3B		
Qwen3-30B-A3B-Instruct (Team, 2025)	0.058	0.043
Scale-SWE-Agent (Zhao et al., 2026)	0.292	0.183
DeNovoSWE-Agent-30A3B	0.472	0.230
Qwen3.5-35B-A3B		
Qwen3.5-35B-A3B (Qwen Team, 2026)	0.438	0.235
DeNovoSWE-Agent-35A3B	**0.500**	**0.271**

DeNovoSWE: Scaling Long-Horizon Environments for Generating Entire Repositories from Scratch, Arxiv, 2026

感谢聆听

BeyondSWE

ScaleSWE

20k数据（docker）
DeepSeek V4蒸馏轨迹

DeNoveSWE

3.7k数据（docker）
DeepSeek V4蒸馏轨迹

AweAgent

轻松上手
接入内部infra