Pull Requests as a Training Signal for Repo-Level Code Editing

Clean-PR: A Verified Mid-Training Corpus from 2M GitHub PRs

Qinglin Zhu, Tianyu Chen, Shuai Lu, Lei Ji, Runcong Zhao,
Murong Ma, Xiangxiang Dai, Yulan He, Lin Gui, Peng Cheng, Yeyun Gong

KCL · MSRA · CUHK · NUS · Alan Turing Institute

ICML 2026

Motivation: Where do SWE-bench gains come from?

SOTA SWE-bench systems = heavy scaffolding (agentic loops, retrieval, test-time scaling).
Hard to attribute gains to any single factor.

Our question
How much repo-level editing can be encoded directly into model weights — under a simple scaffold?

The data gap:

- SWE-bench-style: high fidelity, **small** (1k–50k)
- Pretraining corpora: **huge**, no *how-to-modify* signal

Dataset	#Lang	#Instances
SWE-rebench	1	21,324
SWE-smith	1	50,000
CodeReview	9	534k
commitbench	6	1.17M
The Stack	30	317M
Clean-PR (ours)	**12**	**2.0M**

Verified, multi-file, S/R edits.

Our Idea: Pull Requests as a Training Signal

Open-source PRs are a natural middle ground:

- Human intent (Issue + PR description)
- Real multi-file code changes
- Massive scale, many languages

But the raw stream is extremely noisy:

- 16.4M raw PRs $\rightarrow$ only **18.6% usable**
- Bots, docs-only changes, unmerged PRs, broken patches

Contribution

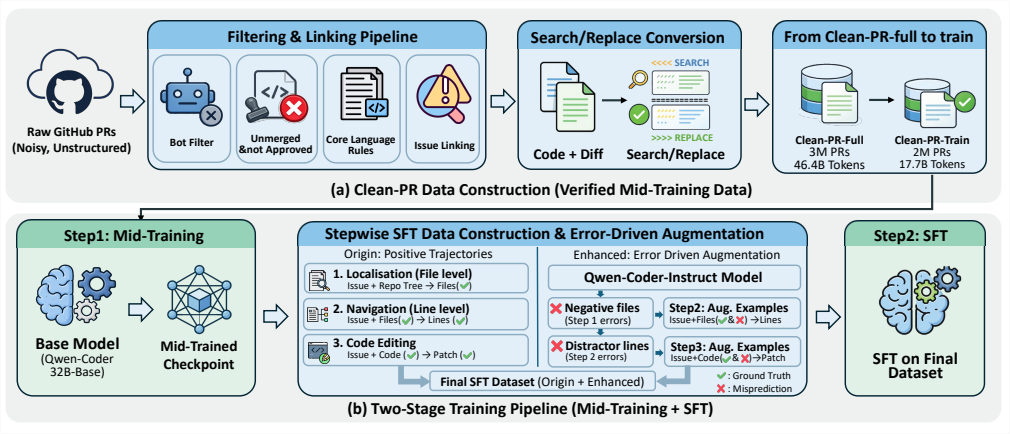
A rigorous pipeline that converts noisy PR diffs into **verified Search/Replace** examples — the largest such public corpus.

Noise in raw GitHub PRs

Noise category	Ratio (%)
Non-core source changes	38.1
Suspected bot activity	25.1
Unmerged / not approved	24.5
Empty base file or diff	13.7
Patch validation failure	4.1
Clean (kept)	**18.6**

16.4M raw $\rightarrow$ 3.05M verified $\rightarrow$ 2.0M train (17.7B tokens, 12 languages).

The Clean-PR Framework



(a) Data: filter → link Issues → reconstruct before/after → minimal-unique S/R → round-trip *verify*.

(b) Training: Mid-train → Agentless-aligned Stepwise SFT (Localise → Navigate → Edit) + **Error-Driven Augmentation**.

Main Results on SWE-bench (Qwen2.5-Coder-32B)

Base	Mid-Train	SFT	Valid	Line Acc.	Pass@1
	SWE-bench Lite (300)				
Qwen-Coder-32B-Instruct	None	✗	77.0	38.3	10.7
Qwen-Coder-32B-Base	None	✓	84.0	46.7	11.3
Qwen-Coder-32B-Base	StarCoder2-style (17.4B)	✓	89.7	47.0	15.7
Qwen-Coder-32B-Base	**Clean-PR-train (17.7B)**	✓	**96.3**	**55.7**	**24.3**
	SWE-bench Verified (500)				
Qwen-Coder-32B-Instruct	None	✗	77.6	42.3	18.3
Qwen-Coder-32B-Base	None	✓	81.8	46.6	17.6
Qwen-Coder-32B-Base	StarCoder2-style (17.4B)	✓	82.4	48.4	20.4
Qwen-Coder-32B-Base	**Clean-PR-train (17.7B)**	✓	**95.2**	**52.2**	**30.6**

+13.6% on Lite, +12.3% on Verified over Instruct, *without* agentic loops or test-time search.

Search/Replace lifts **valid-patch** 89.7%→96.3%; line accuracy 47%→56%.

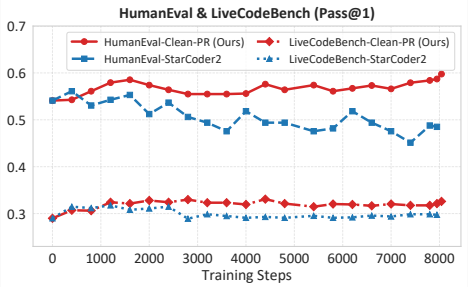
Beats Larger Open-Source Models with a Simple Scaffold

vs. open-source SOTA (Pass@1)

Method	Framework	Params	Lite	Verified
SWE-Gym	OpenHands	32B	15.3	20.6
Lingma-SWE	SWESynInfer	**72B**	22.0	30.2
SWE-Fixer	SWE-Fixer	**72B**	22.0	30.2
Clean-PR	Agentless	32B	**24.3**	**30.6**

A 32B model + lightweight Agentless matches 72B agentic systems.

Generalisation during mid-training



No catastrophic forgetting. HumanEval +5.7, LiveCodeBench +3.6 — verified S/R *sharpens* general coding; diff-style training degrades it.

Generalises Across Scale, Language, and Inference Paradigm

1. Scale — Qwen-Coder-7B

	Lite	Verified
Base + SFT	10.3	14.2
+ **Clean-PR**	**14.5**	**20.4**

2. Multilingual — Multi-SWE-bench Flash (7 langs)

	Valid	Pass@1
Base + SFT	73.0	7.0
+ StarCoder2-style	76.3	8.7
+ **Clean-PR**	**81.7**	**12.3**

3. Agentic transfer — OpenHands CodeActAgent

	Empty↓	Resolve↑
Lite		
SWE-Gym	–	15.3
Base + SFT	21.0	16.1
+ **Clean-PR** + SFT	**16.3**	**20.7**
Verified		
SWE-Gym	–	20.6
Base + SFT	16.6	19.5
+ **Clean-PR** + SFT	**14.3**	**24.7**

Clean-PR is a *prior*, not a scaffold-specific trick: gains transfer back into multi-turn agents.

Takeaways

- **Pull requests are a high-signal training source** — once you filter, link issues, and *verify* edits round-trip.
- **Verified Search/Replace** > raw unified diffs: higher valid-patch rate, better line localisation, *no* catastrophic forgetting.
- **Heavy scaffolding is not the only way to scale SWE-bench.** A 32B model + simple Agentless reaches **24.3% Lite / 30.6% Verified** — matching 72B agentic systems.
- **Released:** largest verified PR corpus to date — **2M PRs, 12 languages, 17.7B tokens**, plus the full pipeline.

Thank you for watching!