

# PAIChecker: Uncovering and Checking PR-Issue Misalignment in SWE-Bench-Like Benchmarks

---

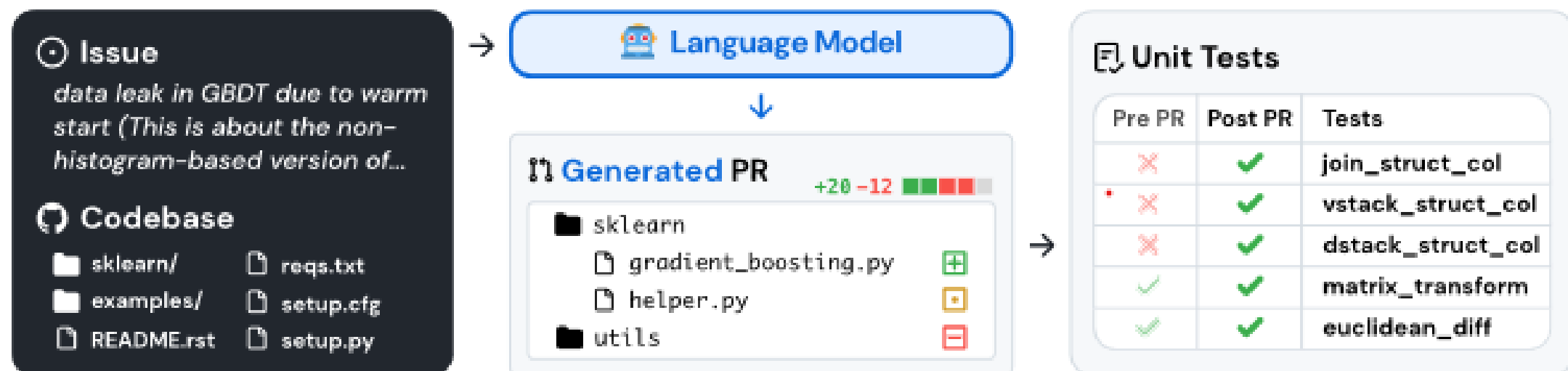
Manyi Wang, Junjielong Xu\*, Pinjia He\*  
香港中文大学 · 深圳

ASE2026

# 背景：SWE-Bench

## SWE-Bench:

- LLMの問題解決能力を評価するためのベンチマークの1つ
- 能力評価プロセス：
  - 与えたIssueとCodebaseに対するPull Request(PR)をLLMに作成させる
  - 作ったPRに含まれるコードがUnit Testに通過するかどうかを検証
  - UnitTestに通過するPRを作成できたIssueがどれだけあるかで能力を評価
- 模範解答として，元々Issueに関連付けられていたPRが用意されている



# 背景：SWE-Benchを構成するデータの収集法

## PR Title #124

Merged

Conversation

Commits

Files changed

username commented 3 days ago

...

このPRは以下の問題を修正します。

Fixes [#123](#)

Add a reaction

PRそのものと、PRのタイムラインで言及されているIssueを「当PRが唯一解決したIssue」と仮定して収集する

1 commit: バグを修正

commit message / 2 hours ago

reviewer が変更を承認しました

review approved / 1 hour ago

疑問：PRで言及されているIssueは必ずしも「PRが唯一解決したIssue」といえるのか？

背景：PRで言及されているIssueは必ずしも「当PRが唯一解決したIssue」といえるのか？

---

「当PRが唯一解決したIssue」といえるには以下2つの条件を満たす必要がある

- 関連元のIssueただ1つに対処
- 関連元のIssueをバグを混入させることなく解決

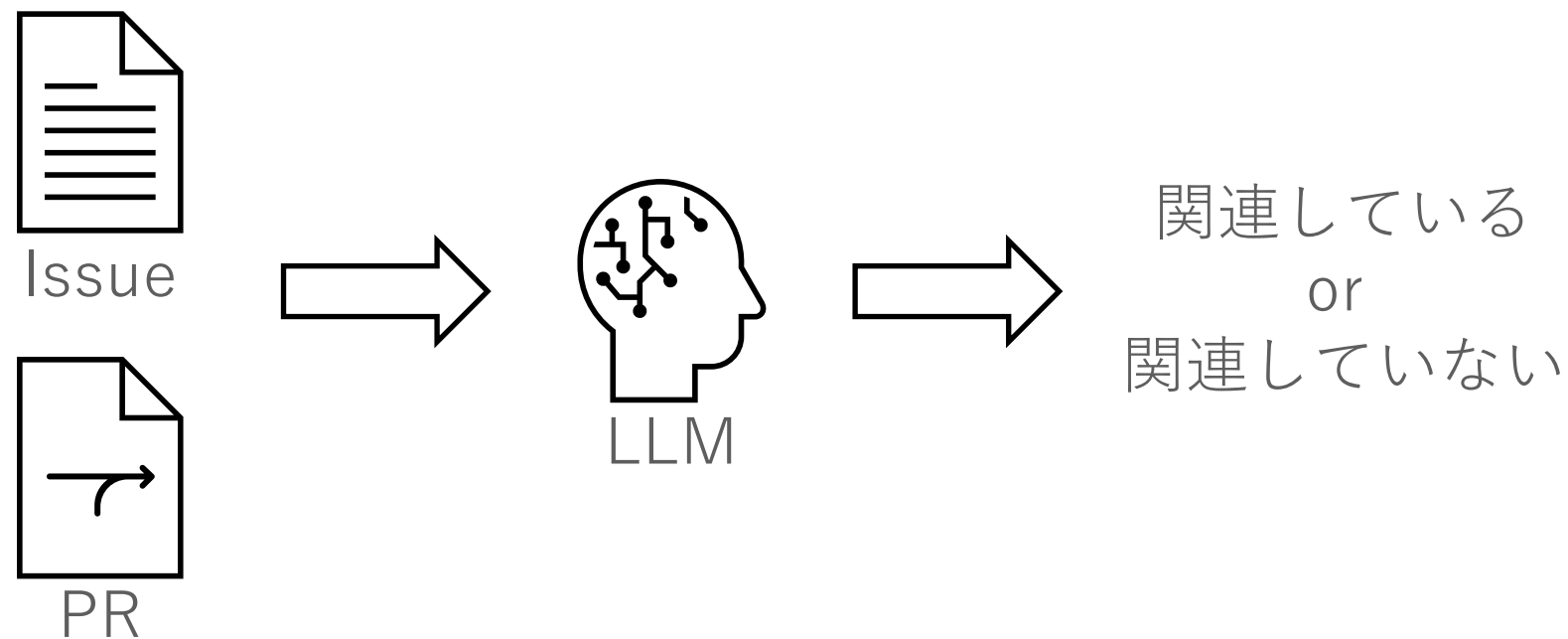
関連づいたPRが必ずしも上記2つの条件を満たしてゐるとは限らない

- ex)1つのPRで複数のIssueに対処
- ex)Issueは解決できたが、バグが混入してしまった

模範解答の条件を満たさないPRがベンチマークに混入により、ベンチマークが正しく機能しなくなってしまう→そのようなPRとIssueのずれ(PR-Issue misalignment)を検出できるようにしたい

# 考えるアプローチ

LLMに関連しているかどうか判定させる



## 問題点

- 推論プロセスが過度に一般化されている
  - 「関連づいていない」のには複数のパターンが考えられる
- 仮にパターンを定義しても、定義したパターンに当てはまるとは限らない

# 研究概要

---

## 目的:

- PRとIssueのずれの種類・蔓延度がベンチマークに与える影響の分析
- PRとIssueのずれを自動検出・分類する手法の提案

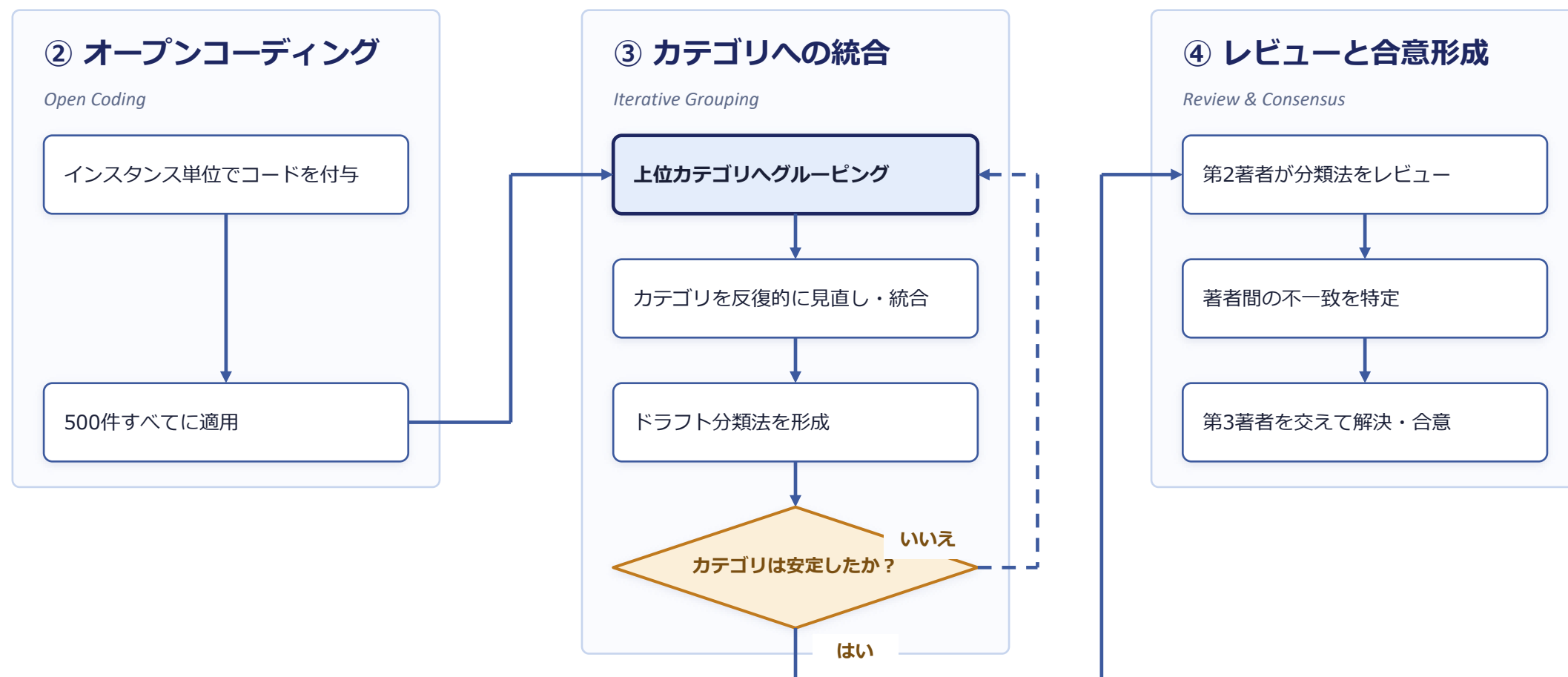
## リサーチクエスチョン

- RQ1: PRとIssueのずれにはどのような種類があり、どの程度蔓延しているか？
- RQ2: PRとIssueのずれはエージェント評価の信頼性にどう影響するか？
- RQ3（有効性）：PAIChecker はPRとIssueのずれの検出・分類にどの程度有効か？
- RQ4（アブレーション）：各コンポーネントは全体の検出性能にどう寄与しているか？
- RQ5（アブレーション）：3つのパイプラインフェーズ間でラベルはどのように、なぜ変化するのか？

# RQ1:実験設定

使用データセット：SWE-bench Verified(500件)

オープンコーディングで分類法を導出



コード：データに付与する記述子

- コードの例：「PR本文が problem\_statement がない別issueも Fixes として列挙している」

# RQ1:PRとIssueのずれの種類

表：PRとIssueのずれの分類

| パターン                         | 定義                                                   |
|------------------------------|------------------------------------------------------|
| SC(PR Scope Creep)           | PRがissueの要求を超える機能を実装している                             |
| DP(Defective PR)             | PR自体が欠陥を含む。新たなバグを導入している                              |
| FP(Follow-up PR)             | DPの逆。同一issueに対する先行PRを補完・修正するPR                       |
| IS(Incomplete Specification) | issueの記述だけでは、Issueの解決に必要な要素が揃わない                     |
| UL(Unspecified Literal)      | テストパッチが、PRで導入されissueには存在しない厳密なリテラル値（例外メッセージ等）を検証している |

# RQ1:PRとIssueのずれの蔓延具合

表：SWE-Bench verifiedの500件のデータに含まれるPRとIssueのずれ

- 1つのIssue-PRのペアには，複数種類のラベルを同時に付与できる

| パターン               | 件数       | 割合    |
|--------------------|----------|-------|
| SC                 | 22       | 4.4%  |
| DP                 | 30       | 6.0%  |
| IS                 | 18       | 3.6%  |
| FP                 | 3        | 0.6%  |
| UL                 | 1        | 0.2%  |
| ラベル合計              | 74       |       |
| misaligned インスタンス数 | 68 / 500 | 13.6% |

# RQ2:実験設定

---

## 使用データ：

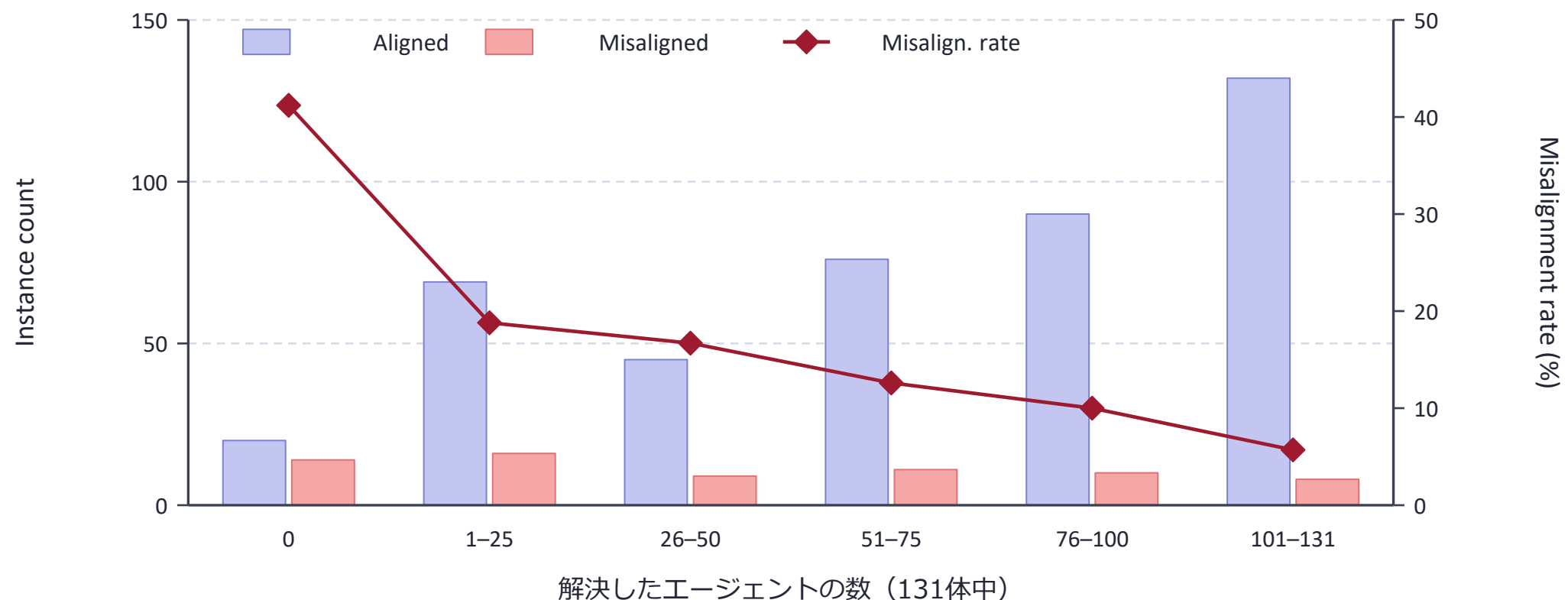
- SWE-bench Verified 500件
- 公式 experiments リポジトリの 131 エージェント分のインスタンス別解決データ

分析1：500件を「解決したエージェント数」でバケットに分け、各バケットのずれ率を算出

分析2：ずれのある68件を除外した 432件で pass rate を再計算し、131体を並べ替えて元の順位と比較

# RQ2: ずれが評価に及ぼす影響

図：PRとIssueのずれの割合と解決したエージェントの数

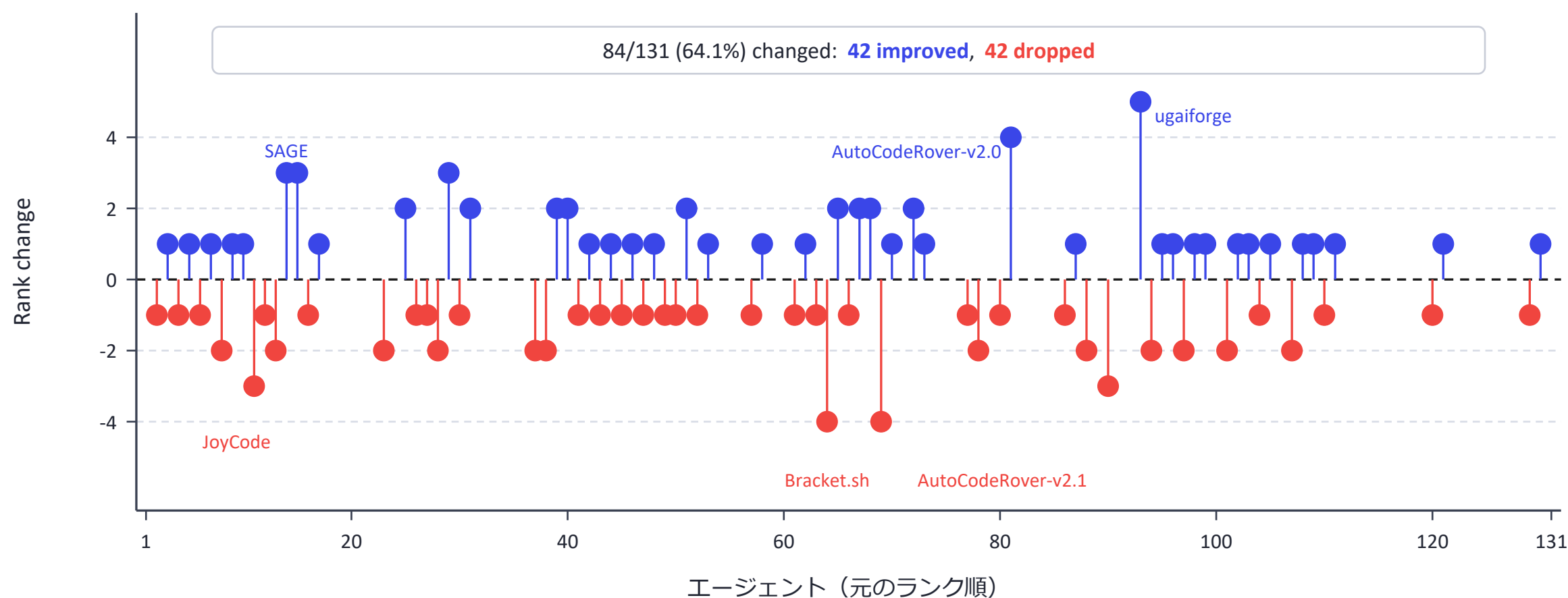


どのエージェントにも解けなかった34件のうち、14件 (41.2%) にPRとIssueのずれの存在を確認

解決したエージェント数が増えるにつれ、ずれがあるIssueの割合は単調に**減少**

# RQ2: ずれが評価に及ぼす影響

図： ミスアライメントが発生しているIssueを抜いたSWE-Bench-verifiedのランクの変化



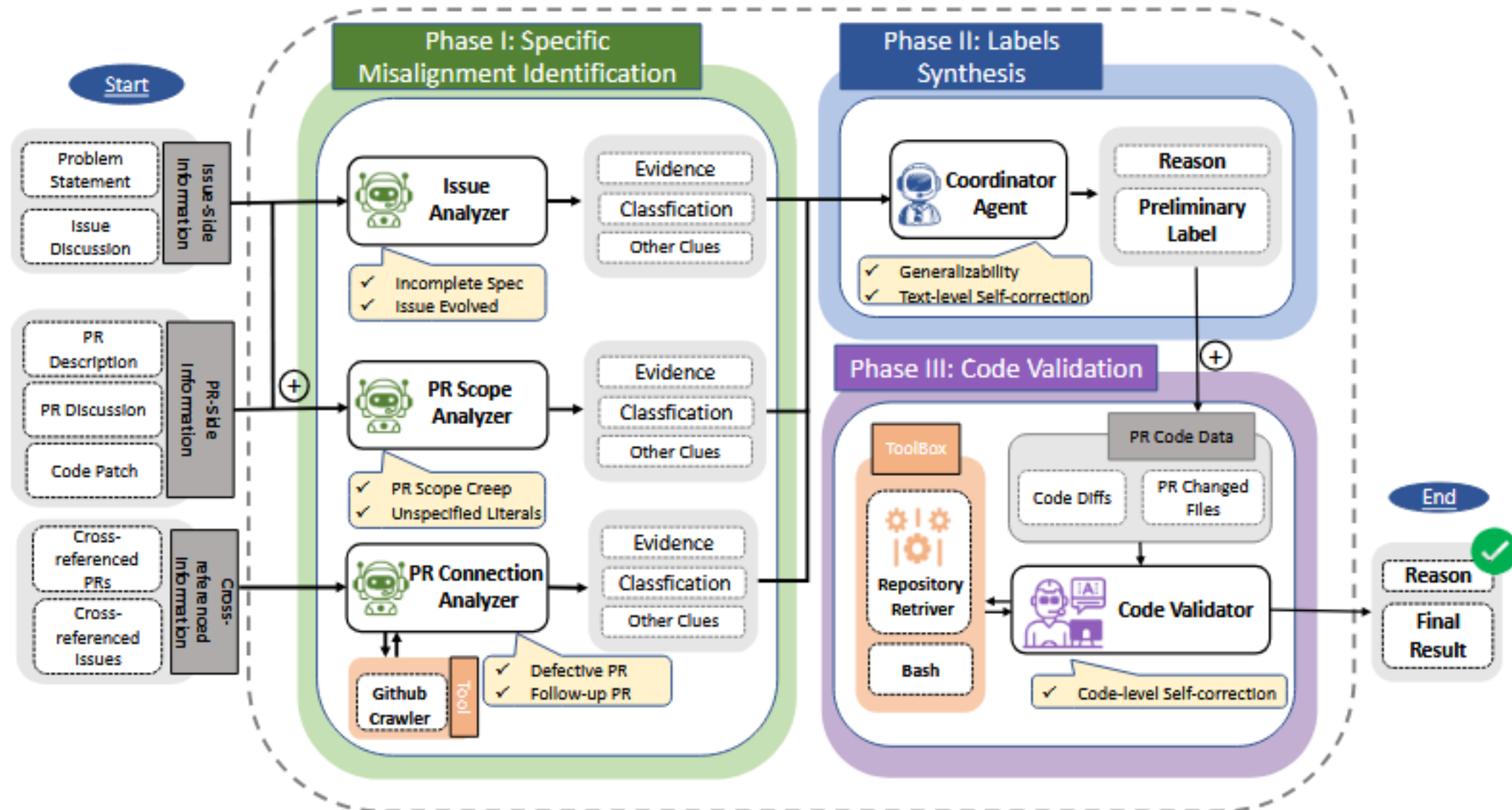
68件除外で 84/131体（64.1%）の順位が変動

トップ10で9体が入れ替わり

# 提案手法

## PAIChecker

- PRとIssueのずれを検出するためのマルチエージェントフレームワーク



# 提案手法の流れ

## Phase 1

- 3つのエージェントでズレの有無を確認
- 各エージェントは判定, 根拠, suspicious clues (担当外で気づいた異常) を出力

## Phase 2

- エージェントの clues を集約し、5パターンのどれでもないズレを指していれば Others
- 根拠が不十分・自己矛盾しているラベルを落とす

## Phase 3

- テキストで下した判断をリポジトリのソースファイル全体と突き合わせ、偽陽性を除去

# 実験設定

---

## 使用データセット:

- SWE-Gym, SWE-bench Multilingual

## 比較手法:

- プロンプトエンジニアリング
  - Zero-Shot, Few-Shot, Chain of Thought
- エージェント
  - Mini-SWE-Agent, OpenHands, ClaudeCode, Codex

## 評価タスク

- 二値分類（ズレの有無） : BA(Binary Accuracy)で評価
- マルチラベル分類（どのパターンか） : EM(Exact Match)で評価

## 使用LLM

- GPT-5.3-Codex, Claude-Sonnet-4.6, Gemini-3.1-Pro-Preview, Qwen-3.5-Plus

# RQ3

すべてのベースラインで最高の性能を発揮

| Backbone              |              | SWE-Gym |       | SWE-bench Multilingual |       |
|-----------------------|--------------|---------|-------|------------------------|-------|
|                       |              | BA      | EM    | BA                     | EM    |
| GPT-5.3<br>Codex      | 最良ベース<br>ライン | 75.72   | 62.02 | 88.00                  | 83.33 |
|                       | PAIChecker   | 88.11   | 79.78 | 91.00                  | 88.00 |
| Claude-<br>Sonnet 4.6 | 最良ベース<br>ライン | 82.36   | 65.88 | 86.00                  | 80.67 |
|                       | PAIChecker   | 87.49   | 76.62 | 91.67                  | 88.33 |
| Gemini-3.1<br>Pro     | 最良ベース<br>ライン | 83.59   | 69.40 | 86.67                  | 82.67 |
|                       | PAIChecker   | 92.12   | 84.66 | 91.33                  | 90.67 |
| Qwen-3.5<br>Plus      | 最良ベース<br>ライン | 74.77   | 62.02 | 83.00                  | 79.67 |
|                       | PAIChecker   | 80.56   | 70.59 | 85.33                  | 83.33 |

# RQ4, RQ5 : アブレーションスタディ

---

Phase I ~ III いずれも検出精度の向上に貢献

- Phase I 除去で最大の低下
- Phase II・III は低下幅は比較的小さい

3サブエージェントは単独・ペアどちらの除去でも性能低下

Phase間の移行において, 「正しいラベルが誤ったラベルになる」場合より「誤ったラベルが正しいラベル」になることの方が多い

# 考察

実世界での検証：実リポジトリ9つの open PR 200件に適用し、ずれ78件を検出。GitHub上でコメント報告

➤ 返信17件中16件（94%）が確認

- 返信の内容は、論文に記載されていない

3段階もLLMを通したら、高くて遅いのでは？

- 1件あたりの費用：\$0.08～\$0.23（10～35円くらい）
- 1件あたりの時間：30～100秒
- 使うトークン：4.2万～5.9万

ズレ検出をZero-shotとPAICheckerにやらせた場合の推定料金の比較

|            | SWE-Gym 2,438件 |
|------------|----------------|
| Zero-Shot  | \$7～\$118      |
| PAIChecker | \$195～\$561    |

# まとめ

---

- | SWE-bench Verified 全500インスタンスの手動分析により、5パターン11シナリオにわたり13.6%のPRとIssueのずれを発見
- | PRとIssueのずれの検出のためのツール PAIChecker を提案
- | SWE-Gym と SWE-bench Multilingual で評価し、それぞれ最高 92.12%、91.67% の精度を達成

# 所感

---

## 「PRとIssueがずれている」という大きく漠然とした問題

- この論文の定義以外にも「ずれ」と呼べるものはまだありそう

## UL(Unspecified Literal)の存在が意外

- 確かにIssue本文で求められてないリテラルを使うことで、テストケースを必要以上に厳しくしているケースは一定数見たことがある

| LLM    | Meth.       | SWE-Gym      |              |              |              | Multilingual |              |              |              |
|--------|-------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|        |             | BA           | BP           | BR           | BF1          | BA           | BP           | BR           | BF1          |
| GPT    | ZS          | 74.36        | 73.14        | 66.57        | 69.70        | 79.67        | 57.32        | 64.38        | 60.65        |
|        | FS          | 75.72        | 75.31        | 67.22        | 71.04        | 79.33        | 57.33        | 58.90        | 58.11        |
|        | CoT         | 75.72        | 73.33        | 71.02        | 72.15        | 83.33        | 68.25        | 58.90        | 63.24        |
|        | MSA         | 72.48        | 83.04        | 47.59        | 60.51        | 88.00        | 84.91        | 61.64        | 71.43        |
|        | OH          | 70.10        | 79.59        | 48.13        | 59.99        | 83.67        | 83.33        | 41.10        | 55.05        |
|        | CX          | 62.34        | 61.18        | 44.60        | 51.59        | 82.67        | 81.82        | 36.99        | 50.94        |
|        | <b>Ours</b> | <b>88.11</b> | <b>83.47</b> | <b>91.20</b> | <b>87.17</b> | <b>91.00</b> | <b>85.94</b> | <b>75.34</b> | <b>80.29</b> |
| Claude | ZS          | 67.02        | 65.23        | 54.72        | 59.52        | 56.33        | 33.71        | 82.19        | 47.81        |
|        | FS          | 69.03        | 68.24        | 56.30        | 61.69        | 70.33        | 43.94        | 79.45        | 56.59        |
|        | CoT         | 73.54        | 68.20        | 75.46        | 71.65        | 75.33        | 49.54        | 73.97        | 59.34        |
|        | MSA         | 82.36        | 81.13        | 78.43        | 79.76        | 86.00        | 70.13        | 73.97        | 72.00        |
|        | OH          | 77.66        | 76.82        | 73.24        | 74.99        | 75.67        | 50.00        | 68.49        | 57.80        |
|        | CC          | 78.09        | 75.30        | 76.20        | 75.75        | 72.00        | 43.82        | 53.42        | 48.15        |
|        | <b>Ours</b> | <b>87.49</b> | <b>82.76</b> | <b>90.65</b> | <b>86.52</b> | <b>91.67</b> | <b>77.91</b> | <b>91.78</b> | <b>84.28</b> |
| Gemini | ZS          | 74.36        | 67.93        | 79.81        | 73.39        | 63.67        | 37.32        | 72.60        | 49.30        |
|        | FS          | 81.05        | 79.37        | 77.31        | 78.33        | 77.67        | 53.19        | 68.49        | 59.88        |
|        | CoT         | 81.09        | 80.49        | 75.65        | 78.00        | 83.33        | 64.94        | 68.49        | 66.67        |
|        | MSA         | 83.59        | 83.60        | 78.33        | 80.88        | 59.33        | 26.67        | 38.36        | 31.46        |
|        | OH          | 79.65        | 82.02        | 69.26        | 75.10        | 86.67        | 78.95        | 61.64        | 69.23        |
|        | <b>Ours</b> | <b>92.12</b> | <b>91.19</b> | <b>91.02</b> | <b>91.10</b> | <b>91.33</b> | <b>88.52</b> | <b>73.97</b> | <b>80.60</b> |
| Qwen   | ZS          | 74.36        | 82.92        | 53.06        | 64.71        | 83.00        | 77.50        | 42.47        | 54.87        |
|        | FS          | 73.75        | 80.73        | 53.52        | 64.37        | 82.33        | 76.32        | 39.73        | 52.25        |
|        | CoT         | 72.76        | 84.90        | 46.85        | 60.38        | 81.67        | 76.47        | 35.62        | 48.60        |
|        | MSA         | 74.77        | 77.65        | 60.46        | 67.99        | 74.33        | 45.00        | 24.66        | 31.86        |
|        | OH          | 60.53        | 65.97        | 28.25        | 39.56        | 81.33        | <b>79.31</b> | 31.51        | 45.10        |
|        | <b>Ours</b> | <b>80.56</b> | <b>90.51</b> | <b>62.69</b> | <b>74.07</b> | <b>85.33</b> | <b>78.43</b> | <b>54.79</b> | <b>64.52</b> |

| Backbone               | Method           | EM    | Macro |       |       |       | Per-Label F1 |       |       |       |       |        |         |
|------------------------|------------------|-------|-------|-------|-------|-------|--------------|-------|-------|-------|-------|--------|---------|
|                        |                  |       | Acc.  | Prec. | Rec.  | F1    | SC           | FP    | DP    | IS    | UL    | Others | No Mis. |
| SWE-Gym                |                  |       |       |       |       |       |              |       |       |       |       |        |         |
| GPT-5.3<br>Codex       | Zero-Shot        | 58.20 | 88.40 | 49.36 | 41.26 | 41.01 | 64.70        | 63.60 | 37.30 | 38.80 | 4.90  | 0.00   | 77.80   |
|                        | Few-Shot         | 60.21 | 89.10 | 56.26 | 41.14 | 42.97 | 67.40        | 62.10 | 30.50 | 40.10 | 21.60 | 0.00   | 79.10   |
|                        | Chain-of-Thought | 60.34 | 89.20 | 53.59 | 49.37 | 49.81 | 67.10        | 70.90 | 48.70 | 44.80 | 38.80 | 0.00   | 78.40   |
|                        | Mini-SWE-Agent   | 62.02 | 89.40 | 62.34 | 39.63 | 45.07 | 62.10        | 59.50 | 55.40 | 42.00 | 17.70 | 0.00   | 78.80   |
|                        | OpenHands        | 61.59 | 90.00 | 74.80 | 39.27 | 46.61 | 54.16        | 54.01 | 11.85 | 46.96 | 59.22 | 7.40   | 79.37   |
|                        | Codex            | 54.36 | 89.40 | 71.28 | 34.87 | 43.32 | 51.67        | 46.02 | 17.39 | 46.80 | 52.39 | 6.90   | 75.62   |
|                        | PAIChecker       | 79.78 | 94.80 | 78.39 | 68.07 | 69.49 | 81.40        | 79.40 | 63.40 | 78.00 | 84.20 | 11.10  | 88.90   |
| Claude-Sonnet<br>4.6   | Zero-Shot        | 53.16 | 86.80 | 47.14 | 35.14 | 38.79 | 58.40        | 47.60 | 41.20 | 30.60 | 21.50 | 0.00   | 72.20   |
|                        | Few-Shot         | 55.58 | 87.70 | 49.77 | 39.70 | 43.29 | 63.10        | 61.20 | 43.50 | 29.10 | 24.70 | 7.40   | 74.00   |
|                        | Chain-of-Thought | 57.47 | 88.50 | 57.83 | 50.79 | 53.09 | 65.60        | 63.30 | 50.20 | 58.70 | 37.90 | 20.70  | 75.20   |
|                        | Mini-SWE-Agent   | 65.26 | 91.50 | 67.07 | 55.97 | 60.34 | 61.00        | 62.00 | 60.00 | 57.40 | 71.90 | 26.70  | 83.40   |
|                        | OpenHands        | 65.88 | 91.51 | 67.80 | 54.77 | 58.81 | 65.53        | 68.26 | 45.65 | 63.32 | 71.47 | 15.38  | 82.07   |
|                        | Claude Code      | 64.30 | 90.74 | 66.24 | 56.93 | 59.90 | 61.68        | 66.25 | 52.38 | 62.35 | 66.38 | 29.41  | 80.84   |
|                        | PAIChecker       | 76.62 | 94.20 | 71.64 | 74.39 | 72.79 | 79.60        | 77.10 | 72.30 | 76.10 | 79.70 | 36.40  | 88.30   |
| Gemini-3.1<br>Pro      | Zero-Shot        | 56.15 | 88.20 | 58.06 | 56.94 | 54.61 | 67.70        | 71.00 | 50.70 | 61.90 | 34.90 | 20.80  | 75.30   |
|                        | Few-Shot         | 66.28 | 91.20 | 64.70 | 59.79 | 59.31 | 66.80        | 63.50 | 53.90 | 61.10 | 65.40 | 21.30  | 83.20   |
|                        | Chain-of-Thought | 66.20 | 91.20 | 65.74 | 58.06 | 60.19 | 68.90        | 67.00 | 52.70 | 63.70 | 54.80 | 30.80  | 83.40   |
|                        | Mini-SWE-Agent   | 69.40 | 92.40 | 69.16 | 57.10 | 61.36 | 61.10        | 68.20 | 73.20 | 61.40 | 70.90 | 9.10   | 85.60   |
|                        | OpenHands        | 67.50 | 91.77 | 69.60 | 52.67 | 57.56 | 66.99        | 76.84 | 36.46 | 63.30 | 58.82 | 16.22  | 84.28   |
|                        | PAIChecker       | 84.66 | 96.20 | 79.60 | 80.46 | 79.21 | 85.10        | 80.00 | 77.60 | 82.80 | 89.60 | 46.50  | 92.90   |
| Qwen-3.5<br>Plus       | Zero-Shot        | 62.02 | 89.60 | 67.86 | 43.04 | 49.04 | 61.50        | 58.50 | 44.40 | 53.50 | 13.50 | 25.00  | 79.90   |
|                        | Few-Shot         | 61.81 | 89.80 | 74.11 | 41.34 | 49.83 | 58.20        | 52.90 | 40.40 | 50.70 | 38.80 | 21.30  | 79.20   |
|                        | Chain-of-Thought | 61.48 | 89.10 | 66.81 | 39.36 | 44.73 | 55.20        | 58.70 | 37.60 | 52.80 | 11.40 | 18.20  | 79.20   |
|                        | Mini-SWE-Agent   | 61.57 | 89.90 | 70.57 | 42.23 | 50.61 | 54.80        | 44.80 | 51.50 | 47.60 | 51.90 | 25.00  | 78.70   |
|                        | OpenHands        | 54.37 | 87.81 | 63.79 | 25.77 | 31.24 | 48.11        | 30.89 | 23.23 | 28.15 | 13.79 | 0.00   | 74.51   |
|                        | PAIChecker       | 70.59 | 92.05 | 75.90 | 52.24 | 58.98 | 74.20        | 68.50 | 52.20 | 54.00 | 52.40 | 27.30  | 84.40   |
| SWE-bench Multilingual |                  |       |       |       |       |       |              |       |       |       |       |        |         |
| GPT-5.3<br>Codex       | Zero-Shot        | 73.00 | 92.38 | 41.13 | 34.23 | 30.44 | 67.57        | -     | 15.38 | 18.18 | 25.64 | -      | 86.29   |
|                        | Few-Shot         | 75.00 | 92.86 | 41.19 | 35.77 | 31.14 | 72.00        | -     | 15.38 | 10.00 | 34.29 | -      | 86.28   |
|                        | Chain-of-Thought | 79.67 | 94.14 | 47.14 | 38.56 | 35.43 | 77.78        | -     | 8.00  | 28.57 | 44.44 | -      | 89.22   |
|                        | Mini-SWE-Agent   | 83.33 | 95.33 | 49.04 | 35.91 | 38.69 | 79.41        | -     | 27.59 | 28.57 | 42.86 | -      | 92.41   |
|                        | OpenHands        | 81.00 | 94.43 | 53.97 | 32.43 | 35.77 | 66.67        | -     | 8.00  | 19.05 | 66.67 | -      | 90.02   |
|                        | Codex            | 80.67 | 94.14 | 57.50 | 30.15 | 34.23 | 58.82        | -     | 8.00  | 33.33 | 50.00 | -      | 89.47   |
|                        | PAIChecker       | 88.00 | 96.33 | 82.22 | 70.91 | 71.85 | 82.19        | -     | 28.57 | 63.64 | 84.21 | -      | 94.37   |
| Claude-Sonnet<br>4.6   | Zero-Shot        | 45.33 | 83.76 | 32.56 | 34.49 | 29.48 | 58.14        | -     | 55.00 | 25.64 | 4.96  | -      | 62.64   |
|                        | Few-Shot         | 60.33 | 88.81 | 36.56 | 37.91 | 34.33 | 67.61        | -     | 60.00 | 24.24 | 10.81 | -      | 77.66   |
|                        | Chain-of-Thought | 67.33 | 90.33 | 36.48 | 35.66 | 31.63 | 71.88        | -     | 51.43 | 0.00  | 15.58 | -      | 82.49   |
|                        | Mini-SWE-Agent   | 80.67 | 92.29 | 47.60 | 51.28 | 48.03 | 70.97        | -     | 79.07 | 40.00 | 53.85 | -      | 92.31   |
|                        | OpenHands        | 70.67 | 91.67 | 39.93 | 42.80 | 37.68 | 63.64        | -     | 51.43 | 33.33 | 32.43 | -      | 82.90   |
|                        | Claude Code      | 68.67 | 90.95 | 37.62 | 35.30 | 33.52 | 61.54        | -     | 47.06 | 28.57 | 16.67 | -      | 80.82   |
|                        | PAIChecker       | 88.33 | 97.05 | 62.40 | 72.33 | 65.95 | 84.85        | -     | 87.50 | 63.16 | 81.82 | -      | 94.33   |
| Gemini-3.1<br>Pro      | Zero-Shot        | 55.33 | 80.10 | 45.56 | 52.64 | 42.64 | 68.85        | -     | 63.16 | 32.26 | 11.36 | -      | 72.82   |
|                        | Few-Shot         | 71.00 | 84.95 | 44.43 | 40.22 | 39.43 | 66.67        | -     | 57.14 | 33.33 | 33.33 | -      | 85.51   |
|                        | Chain-of-Thought | 78.00 | 89.10 | 46.66 | 41.62 | 42.26 | 77.19        | -     | 55.56 | 45.71 | 27.27 | -      | 90.09   |
|                        | Mini-SWE-Agent   | 57.00 | 64.14 | 59.67 | 30.02 | 38.95 | 59.57        | -     | 52.94 | 34.78 | 50.00 | -      | 75.38   |
|                        | OpenHands        | 82.67 | 95.14 | 53.45 | 41.39 | 42.23 | 78.12        | -     | 28.57 | 54.55 | 42.86 | -      | 91.49   |
|                        | PAIChecker       | 90.67 | 95.19 | 82.80 | 64.68 | 71.77 | 82.76        | -     | 81.82 | 81.82 | 94.12 | -      | 95.24   |
| Qwen-3.5<br>Plus       | Zero-Shot        | 79.33 | 94.00 | 38.84 | 26.25 | 28.57 | 58.62        | -     | 34.48 | 17.39 | 0.00  | -      | 89.53   |
|                        | Few-Shot         | 79.33 | 94.10 | 45.69 | 29.26 | 32.53 | 64.29        | -     | 28.57 | 19.05 | 26.67 | -      | 89.16   |
|                        | Chain-of-Thought | 78.67 | 93.10 | 43.63 | 24.56 | 26.74 | 52.17        | -     | 15.38 | 30.77 | 0.00  | -      | 88.84   |
|                        | Mini-SWE-Agent   | 73.00 | 77.67 | 56.29 | 22.61 | 28.64 | 37.21        | -     | 40.00 | 11.11 | 20.00 | -      | 75.97   |
|                        | OpenHands        | 79.67 | 93.48 | 40.26 | 23.88 | 26.69 | 58.82        | -     | 28.57 | 10.53 | 0.00  | -      | 88.93   |
|                        | PAIChecker       | 83.33 | 95.24 | 50.40 | 37.89 | 42.38 | 73.68        | -     | 63.16 | 31.58 | 37.50 | -      | 90.76   |