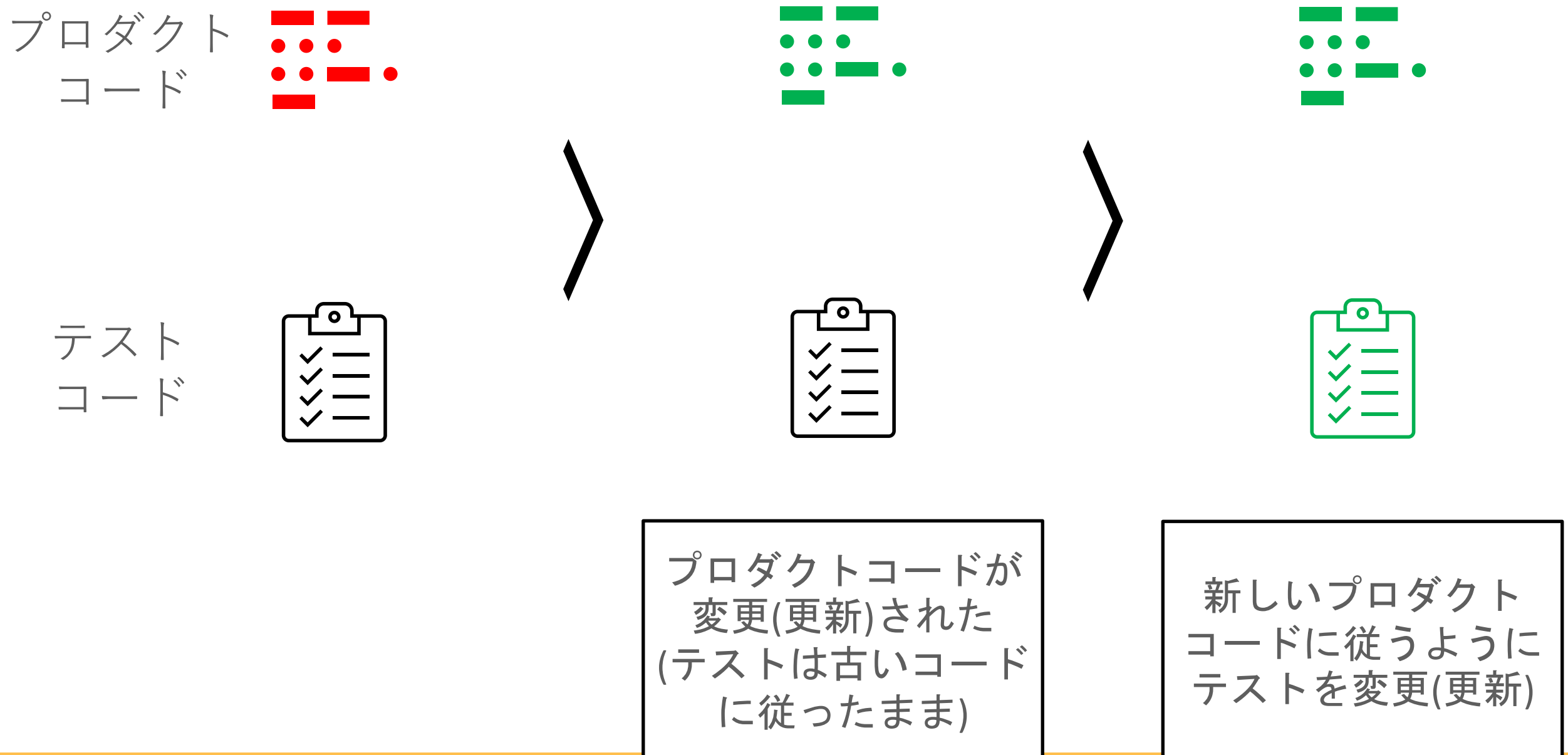


PT共進化PRとレビューコメント の関係性調査

テスト更新の流れ

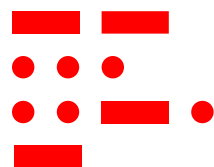
図：テストを更新するに至る流れ



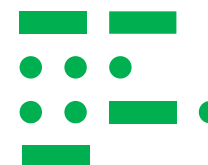
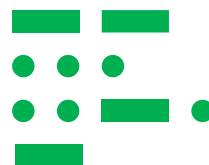
テスト更新の流れ

このプロセスが同時(同一PR, 同一commit, etc...)に行われた場合, プロダクトコードとテストが共進化してる(PT共進化)とみなす

プロダクト
コード



テスト
コード

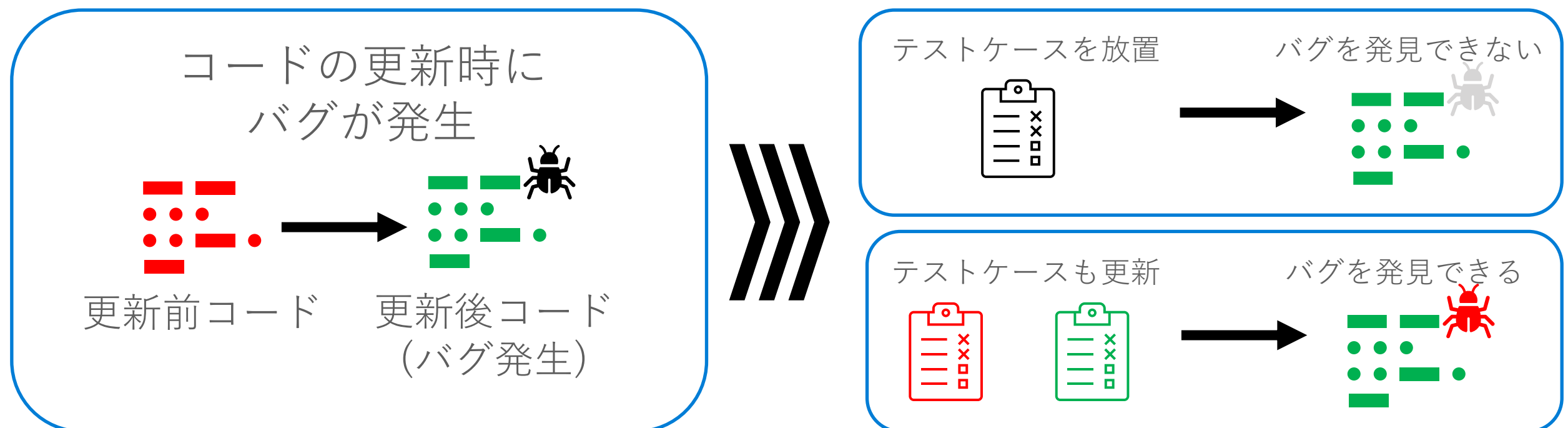


プロダクトコードが
変更(更新)された
(テストは古いコード
に従ったまま)

新しいプロダクト
コードに従うように
テストを変更(更新)

PT共進化の重要性

- 変更したプロダクトコードに対応するテストがあるのにPT共進化しない
- プロダクトコードの更新に伴い発生したバグの検出が遅れる[1].



[1] Chi, J., Wang, X., Huang, Y., Yu, L., Cui, D., Sun, J., & Sun, J. (2025). Reaccept: Automated co-evolution of production and test code based on dynamic validation and large language models. *Proceedings of the ACM on Software Engineering*, 2(ISSTA), 1234-1256.

背景

先行研究[2]

- Milanese et al. 2026: 人間とAI Agentが協働して作ったPRと人間が作ったPRとではPR単位でPT共進化が行われる頻度はどちらも8割程度

先行研究のギャップ

- PR単位での共進化の頻度は明らかになったが、レビュアーのコメントとPT共進化の関係は未分析
- AI Agent種別・PRの進化種別の共進化の違い・AIレビューツールの役割も未検証

本調査の問い

- PT共進化PRにおいて、AIと人間ではどのようなレビューを受けるのか？
- レビュー内容を規定する要因は何か？（作者種別 / エージェント / 進化型 / テスト品質）

[2]Human-Agent versus Human Pull Requests: A Testing-Focused Characterization and Comparison.
arXiv preprint arXiv:2601.21194.

データセット

先行研究[2]で作成されたデータセットを使用

- AI-Devデータセットv3がベース
 - 人間とAIが協働して作成したPR(HAPR) 932,791件と人間が作成したPR(HPR) 6,618件から成るデータセット
- AI-Devデータセットv3 [3]を以下の条件でフィルタリング
 - 対象言語 : Java、JavaScript、Python、TypeScript
 - 変更がないPRは除外
 - スター数500未満のリポジトリは除外
- フィルタリングで得られた6,582件のHAPRと3,122件のHPRから、HAPRとHPR各339件および294件の層別有意サンプルを抽出
 - 共進化してるかしてないかを学生2人で目視でラベル付け

[3] Li, H., Zhang, H., & Hassan, A. E. (2025). The rise of ai teammates in software engineering (se) 3.0: How autonomous coding agents are reshaping software engineering. *arXiv preprint arXiv:2507.15003*.

データ収集・タクソノミー構築

PRコメントを分析しやすくするために、PRコメントの分類体系の構築と分類を行った。

- 先行研究[4]で提案されているコメント分類手法を使用。

フェーズ1: GitHubレビューコメント収集

- PyGithubで339+294件のPRに含まれるPRコメントを取得
- 取得したPRコメント
 - 人間: 1,664件 / Bot: 1,367件 / 合計: 3,033件

フェーズ2: BERTopic + GPT-5 タクソノミー構築

- sentence-transformers/all-MiniLM-L6-v2 + BERTopicでPR単位でトピック抽出
- GPT-5で全トピックを統合 → 12カテゴリを確定
- GPT-5-nanoで全3,033件を自動アノテーション

[4] Haider, M. A., & Zimmermann, T. (2026). Understanding Dominant Themes in Reviewing Agentic AI-authored Code. *arXiv preprint arXiv:2601.19287*.

12カテゴリ

タグ	カテゴリ名
python_impl_review	Pythonの実行結果レビュー
bun_buildkite_tests	Bun Buildkitのテストレビュー
js_test_execution	JavaScriptのテスト実行
vercel_previews_security	Vercelのプレビューとセキュリティ
app_preview_links	アプリプレビューのデプロイURL
deps_and_release_mgmt	依存関係とリリースの管理
checks_and_e2e	E2Eテストのチェック
ci_bot_comment_noise	CI botによるコメントの自動管理
sonar_quality_reports	コード品質分析レポート
coverage_reports	テストカバレッジレポート
build_network_issues	ビルドおよびネットワークの問題
ai_review_summaries	AI PullRequestの要約

生成されたカテゴリが明らかにノイズの影響を受けてしまった

結果 ①:PRの内容とPT共進化の関係分析

COC vs TFC — 全カテゴリ有意差なし

- Stage 1: 全カテゴリ adj.p = 1.0(12カテゴリあるので1カテゴリごとに12回検定を実行)
- Stage 2: 全カテゴリ非有意
- → レビュアーはPRがCOCかTFCかを意識してレビュー内容を変えていない

HAPR-COC vs HPR-COC

- CI Bot Moderation: OR=2.01* / Build & Network: OR=8.86*
- Python Implementation Review: OR=0.54* (HPR優位)
- HAPR-TFC vs HPR-TFCの結果との比較が必要 (今後の課題)

OR: オッズ比 (>1 = HAPR優位、<1 = HPR優位)

adj.p: Holm-Bonferroni 補正済み p 値 | * p<0.05 ** p<0.01 *** p<0.001

Cliff's δ : 効果量 (negligible <0.15、small 0.15-0.33、medium 0.33-0.47、large ≥ 0.47)

Stage 1: χ^2 検定 (カテゴリ保有率の二値比較) | Stage 2: Mann-Whitney U 検定 (件数の量的比較)

COC:共進化, TFC:テストコードのみ変更

結果 ②:HAPRとHPRにおける進化型の割合

PR進化型の内訳（共進化 PR 506件）

- always_mixed : PRのはじめから共進化
- prod_to_coc : PRの最初はプロダクトコードのみ
- Test_to_coc : PRの最初はテストコードのみ

進化型	合計	HAPR	HAPR内%	HPR	HPR内%	HPR比率
always_mixed	301件	158件	87.3%	143件	62.4%	47.5%
prod_to_coc	109件	23件	12.7%	86件	37.6%	78.9%
test_to_coc	6件	2件	1.1%	4件	1.7%	66.7%
合計 (unknown 除く)	416件	183件	100%	233件	100%	—

結果 ③: 各カテゴリのコメントの保有率, 件数が各共進化に与える影響

比較	カテゴリ	prod_to_coc %	always_mixed %	OR / δ	adj.p	判定
保有率	Checks & E2E Tests	62.4%	36.2%	OR=2.92	0.000	***
	Dependency & Release Mgmt	28.7%	13.5%	OR=2.58	0.032	*
件数	Python Impl Review	多い	少ない	δ =+0.320	0.005	*
	AI Review Summaries	多い	少ない	δ =+0.283	0.015	*

OR: オッズ比 (>1 = prod_to_cocグループ優位、<1 = always_mixedグループ優位)
adj.p: Holm-Bonferroni 補正済み p 値 | * p<0.05 ** p<0.01 *** p<0.001
Cliff's δ : 効果量 (negligible <0.15、small 0.15-0.33、medium 0.33-0.47、large $\geq$ 0.47)
Stage 1: χ^2 検定 (カテゴリ保有率の二値比較) | Stage 2: Mann-Whitney U 検定 (件数の量的比較)
prod_to_coc :最初はプロダクトコードだけ変更だったが, 最終的にプロダクトコードとテストコードが同時に変更されたもの
always_mixed :最初からプロダクトコードとテストコードが同時に変更されたもの

考察

PRコメントの内容と共進化の有無は無関係

PRの作成者が人間かそうでないかとPRコメントの内容は関係あり

- しかし、これは先行研究[4]で判明している事実

今後の予定

タクソノミーの改善

- ai_review_summaries を ai_coding_activity と ai_code_review に分割して全分析を再実行
- GPT-5-nanoアノテーションの精度を人手検証で評価

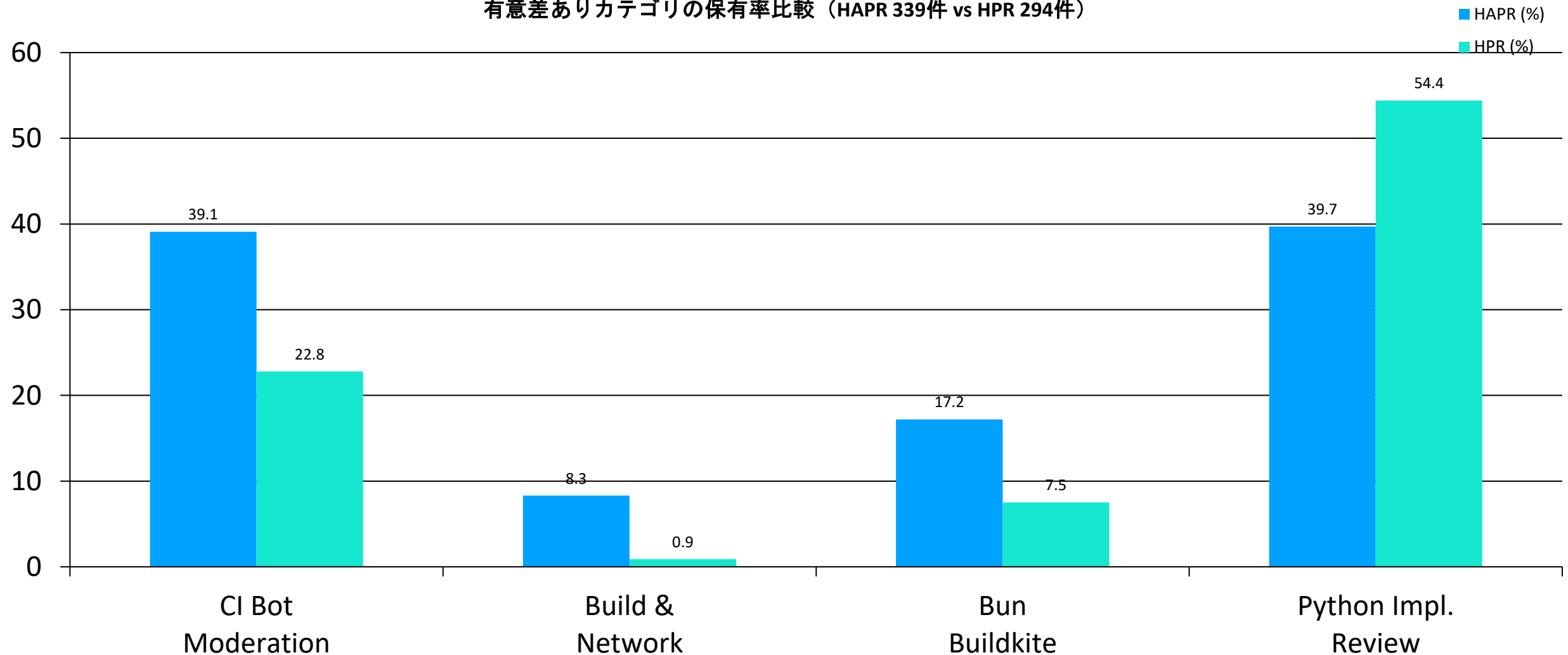
因果関係の解明

- prod_to_coc PRにおいて「レビューコメントがテスト追加を引き起こす」時系列検証
- test_requestコメント後にテストが実際に追加されるかをPR履歴で追跡

補足資料

補足: HAPR vs HPR — レビューカテゴリ保有率

有意差ありカテゴリの保有率比較 (HAPR 339件 vs HPR 294件)

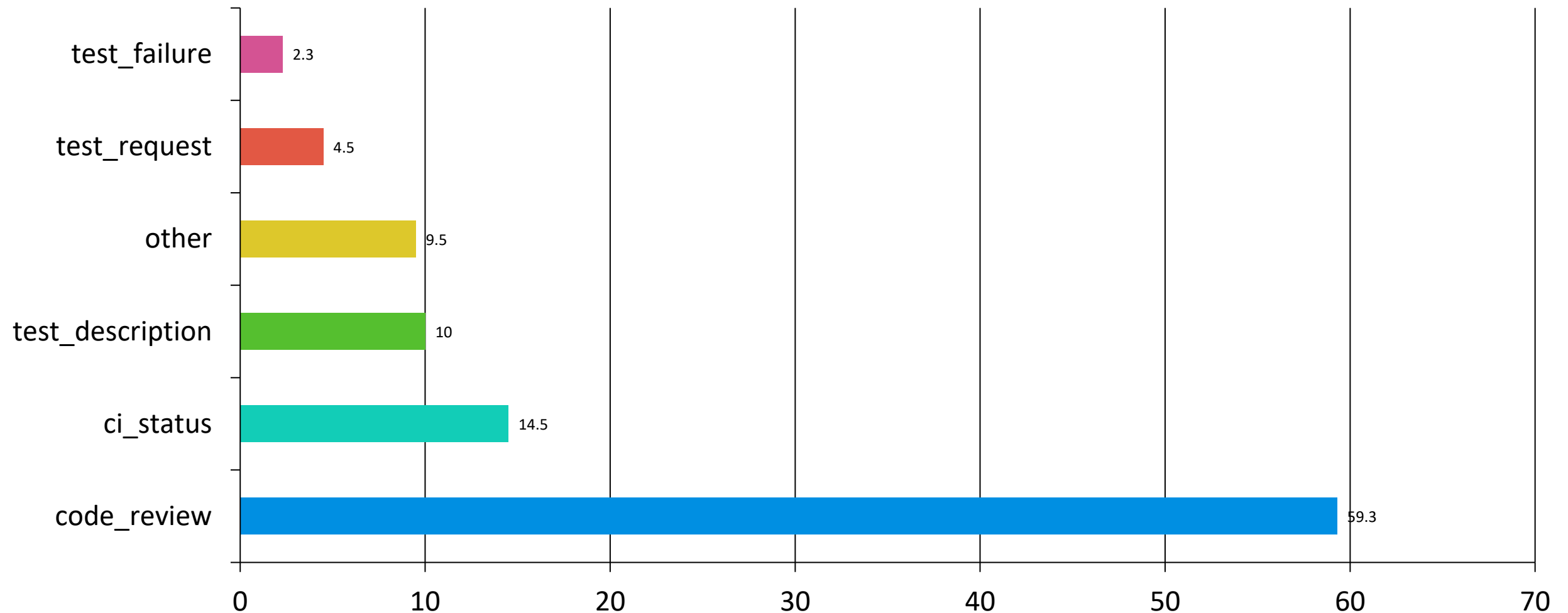


CI Bot: OR=2.17** | Build & Network: OR=10.17** | Bun Buildkite: OR=2.58* | Python Impl: OR=0.55* (HPR優位)
Stage 2: Python Impl ReviewのみHPRがコメント数多い (Cliff's δ =-0.221、small、 $p=0.041$)

OR: オッズ比 (>1 = 左グループ優位、<1 = 右グループ優位)
adj.p: Holm-Bonferroni 補正済み p 値 | * $p<0.05$ ** $p<0.01$ *** $p<0.001$
Cliff's δ : 効果量 (negligible <0.15、small 0.15-0.33、medium 0.33-0.47、large ≥ 0.47)
Stage 1: χ^2 検定 (カテゴリ保有率の二値比較) | Stage 2: Mann-Whitney U 検定 (件数の量的比較)

補足: AIレビューボットのコメント意図分布（フェーズ5）

221件のAIレビューボットコメントの意図分類（LLM分類）



仮説「AIレビューボットがprod_to_cocのテスト欠如を検出して指摘」→棄却
prod_to_coc vs always_mixed: test_request率2.4% vs 4.2% (OR=0.569、adj.p=1.000) 有意差なし

OR: オッズ比 (>1 = 左グループ優位、<1 = 右グループ優位)

adj.p: Holm-Bonferroni 補正済み p 値 | * p<0.05 ** p<0.01 *** p<0.001

Cliff's δ : 効果量 (negligible <0.15、small 0.15-0.33、medium 0.33-0.47、large ≥ 0.47)

Stage 1: χ^2 検定（カテゴリ保有率の二値比較） | Stage 2: Mann-Whitney U 検定（件数の量的比較）

補足: エージェント別レビュープロフィール

エージェント	特徴的なレビューカテゴリ（多い）	OR / δ	解釈
Devin	CI Bot Comment Moderation	OR=8.6	CIを壊す頻度が最高
	AI Review Summaries	OR=6.2	自動サマリーが支配的
	JS Test Execution（少）	OR=0.08	JS系レビューは逆に少
Copilot	Build & Network Issues	OR=66.4	ビルド失敗が突出
	Checks & E2E Tests	OR=3.2 / $\delta=+0.466$	E2Eに課題・実装議論も最活発
OpenAI_Codex	Bun Buildkite Tests	OR=14.6	Bunプロジェクト特化
	Python Impl Review（少）	OR=0.24	実装品質議論が少ない
Cursor	Coverage Reports	OR=6.1	カバレッジが注目される

OR: オッズ比（>1 = 左グループ優位、<1 = 右グループ優位）

Cliff's δ : 効果量（negligible <0.15、small 0.15-0.33、medium 0.33-0.47、large ≥ 0.47 ）

Stage 1: χ^2 検定（カテゴリ保有率の二値比較） | Stage 2: Mann-Whitney U 検定（件数の量的比較）

補足: 人間 vs AIレビューアーの test_request 率

分析対象	進化型	test_request率 (PR 単位)	n (PR数)	比較
人間レビューアー (1,205件)	prod_to_coc	18.2% (14/77 PR)	77 PR	
	always_mixed	9.6% (14/155 PR)	155 PR	prod vs always: OR=2.23、 p=0.054 (傾向)
AIレビューボット (221件)	prod_to_coc	2.4% (2件)	—	人間の7.6倍のギャップ
	always_mixed	4.2% (3件)	—	
テストスメル相関 (496 PR)	全体	有意な相関なし	496 PR	最大 r =0.102 (adj.p=0.274)

OR: オッズ比 (>1 = 左グループ優位、<1 = 右グループ優位)
adj.p: Holm-Bonferroni 補正済み p 値 | * p<0.05 ** p<0.01 *** p<0.001
Cliff's δ : 効果量 (negligible <0.15、small 0.15-0.33、medium 0.33-0.47、large ≥ 0.47)
Stage 1: χ^2 検定 (カテゴリ保有率の二値比較) | Stage 2: Mann-Whitney U 検定 (件数の量的比較)