

Hot Fixing in the Wild

Carol Hanna, Karine Even-Mendoza, W.B. Langdon, Mar
Zamorano López, Justyna Petke, Federica Sarro

EASE 2026

背景：重要なバグ

バグ：

- プログラムに含まれる誤りや欠陥

すべてのバグが同じ重要度というわけではない

- 例：停電やセキュリティリークのような障害を引き起こす故障は重要度が高い

本番環境で重要度の高いバグが発生した場合迅速に修復する必要がある

背景：通常のバグ修復とHot Fixの違い

Hot Fix：

- 本番環境のソフトウェアシステムで発生した重要な問題に対する迅速で無計画な修復

修復の目的：

- 通常のバグの修復：永続的な解決策を作成
- Hot Fix：症状の緩和を行う

背景：LLMAgentのHot Fix

LLMAgentがソフトウェア開発に積極的に参加

- プルリクエスト(PR)を通じて、プログラムの修復にも関与

課題：Hot Fixは重要な問題であるにも関わらず、LLMAgentがどのように取り組むかは明らかになってない

研究概要

研究目的

- LLMAgentと人間はHot Fixをどのように行うか調査

アプローチ

- Hot Fixが行われているPRの収集
- Hot Fix中の開発者の行動と日常的なバグ修正行動を比較
- 人間主導のHot FixとLLMAgentが作成したHot Fixを比較

Hot Fixが行われているPRの収集

仮定：

- 重要度の高いIssueに含まれているPRはすべてHot Fixが行われている
- 重要度の高いIssueとそうでないIssueを分類する必要がある

AI-Devデータセットから収集

- AIが作成したPR 932,791件と人間が作成したPR 6,618件から成るデータセット
- 少し工夫をするとPRに関連したIssueも収集可能

Hot FixすべきIssueの検出

LLM+時系列データでHot FixすべきIssueを特定

LLMはLlama 3.2、Phi-4-mini、Qwen 2.5 の3つを使用

時系列データは2つ使用

- PRが作られてから閉じられるまでの時間差
 - 素早く解決されたPRを特定(12～24時間以内)
- PRが作られた時間と対応するIssueが作られた時間の時間差
 - 迅速に対応されたIssueを特定(12～24時間以内)

Hot FixすべきIssue収集の課題

Hot Fixすべき重要な問題を捉えることが難しい

- イシュートラッカーのラベルはプロジェクトによって用語が異なる
- ラベルがあってもそれが使えるとは限らない

例：

- HPC?の性能バグは特定のプロジェクトでは重要な問題
- 確かに重要ではあるが迅速に修正すべき問題ではない

緊急性を捉えるにはラベルだけではなく文脈を捉える必要がある

分類結果：Issueの分類

表：LLM+時系列データによる重要なIssueの分類

	Llama 3.2 (計5,695)	Phi-4-mini (計4,955)	Qwen 2.5 (計4,736)
重要でないIssue	4,078 (71.6%)	4,425 (89.3%)	4,555 (96.2%)
重要なIssue	1,617 (28.4%)	530 (10.7%)	181 (3.8%)

Llama 3.2：最も多くのIssueを重要と判定(積極的)

Qwen 2.5: 最も多くのIssueを重要でないと判定(保守的)

分類結果：抽出できたPR

表：Issueから抽出したPRの総数

- 特にフィルタリングは行っていない
- Bot(Agent)PRかそうでないかは、各データに元からラベル付けされている

		Llama 3.2	Phi-4-mini	Qwen
重要でないIssue	Bot PR	397	425	433
	Human PR	4,258	4,392	4,443
	合計	4,655	4,817	4,876
重要なIssue	Bot PR	41	13	6
	Human PR	229	99	31
	合計	270	112	37

収集したPRを元にホットフィックスの分析を行う

ホットフィックスの分析

定量指標:

- Contributor数
- レビューアー数
- コミット数
- 変更ファイル数
- 行の追加/削除数
- テストファイルの変更有無

定性指標:

- PRテキストの語彙特性
- ワードクラウド（上位40語）

分析結果：Hot Fix vs 非Hot Fix

Contributorの数

PR種別	統計	qwen2.5-3b	llama3.2	phi4mini
Hot Fixでない	中央値	1	1	1
	平均	1.3	1.3	1.3
Hot Fix	中央値	1	1	1
	平均	1.2	1.3	1.2

レビュアーの数

PR種別	統計	qwen2.5-3b	llama3.2	phi4mini
Hot Fixでない	中央値	2	2	2
	平均	2.3	2.3	2.3
Hot Fix	中央値	2	2	2
	平均	1.8	2.0	2.2

分析結果：Hot Fix vs 非Hot Fix

Contributor数もレビュアー数もHot Fixの時はそうでない場合よりも平均は小さい→Hot Fixは少人数のグループで行われる

PR種別	統計	qwen2.5-3b	llama3.2	phi4mini
Hot Fixでない	中央値	1	1	1
	平均	1.3	1.3	1.3
Hot Fix	中央値	1	1	1
	平均	1.2	1.3	1.2

レビュアーの数

PR種別	統計	qwen2.5-3b	llama3.2	phi4mini
Hot Fixでない	中央値	2	2	2
	平均	2.3	2.3	2.3
Hot Fix	中央値	2	2	2
	平均	1.8	2.0	2.2

分析結果：Hot Fix vs 非Hot Fix

コミット数

PR種別	統計	qwen2.5-3b	llama3.2	phi4mini
Hot Fixでない	中央値	3	3	3
	平均	4.9	4.9	4.9
Hot Fix	中央値	2	3	2
	平均	2.7	3.9	3.2

変更ファイル数

PR種別	統計	qwen2.5-3b	llama3.2	phi4mini
Hot Fixでない	中央値	4	4	4
	平均	27.7	28.3	27.9
Hot Fix	中央値	2	3	2
	平均	3.9	16.9	10.4

分析結果：Hot Fix vs 非Hot Fix

ホットフィックスPRはそうでないPRと比べてコミット数が少ない

PR種別	統計	qwen2.5-3b	llama3.2	phi4mini
Hot Fixでない	中央値	3	3	3
	平均	4.9	4.9	4.9
Hot Fix	中央値	2	3	2
	平均	2.7	3.9	3.2

変更ファイル数

PR種別	統計	qwen2.5-3b	llama3.2	phi4mini
Hot Fixでない	中央値	4	4	4
	平均	27.7	28.3	27.9
Hot Fix	中央値	2	3	2
	平均	3.9	16.9	10.4

分析結果：Hot Fix vs 非Hot Fix

ホットフィックスPRはそうでないPRと比べて変更ファイル数が大幅に少ない

PR種別	統計	qwen2.5-3b	llama3.2	phi4mini
Hot Fixでない	中央値	3	3	3
	平均	4.9	4.9	4.9
Hot Fix	中央値	2	3	2
	平均	2.7	3.9	3.2

変更ファイル数

PR種別	統計	qwen2.5-3b	llama3.2	phi4mini
Hot Fixでない	中央値	4	4	4
	平均	27.7	28.3	27.9
Hot Fix	中央値	2	3	2
	平均	3.9	16.9	10.4

分析結果：Hot Fix vs 非Hot Fix

コミット数

PR種別	統計	qwen2.5-3b	llama3.2	phi4mini
Hot Fixでない	中央値	3	3	3
	平均	4.9	4.9	4.9
Hot Fix PRはそうでないPRと比べて • コミット数が少ない • 変更ファイル数が大幅に少ない ➤ Hot Fixは、単一の重要な問題を解決することを目的とした、スコープの絞られた修正となっている				
PR種別	統計	qwen2.5-3b	llama3.2	phi4mini
Hot Fixでない	中央値	4	4	4
	平均	27.7	28.3	27.9
Hot Fix	中央値	2	3	2
	平均	3.9	16.9	10.4

分析結果：Hot Fix vs 非Hot Fix

テストファイル変更率

PR種別	qwen2.5-3b	llama3.2	phi4mini
Hot Fixでない	54.42%	53.57%	53.57%
Hot Fix	29.73%	50.14%	34.82%

Hot Fixしている時は、そうでない場合と比べてテストファイルを変更することが**少ない**

- チームが迅速な展開を優先してテストのベストプラクティスを意図的に省略している

分析結果：人間のHot Fix vs AgentのHot Fix

表：Contributor数

PR種別	作成者	qwen2.5-3b	llama3.2	phi4mini
Hot Fix	Human	1.2	1.2	1.1
	Bot	1.5	1.4	1.3

Hot Fix においてHuman PRはBot PRよりContributorの数が**多い**

- Hot Fixにおいて， Bot(Agent)が作成したPRは追加の議論や調整を必要とすることが多い

分析結果：人間のHot Fix vs AgentのHot Fix

行追加数

PR種別	作成者	qwen2.5-3b	llama3.2	phi4mini
Hot Fix	Human	24.4	66.0	18.4
	Bot	33.9	110.2	34.9

行削除数

PR種別	作成者	qwen2.5-3b	llama3.2	phi4mini
Hot Fix	Human	10.3	72.8	1.6
	Bot	3.6	91.2	0.5

人間よりBotの方が行追加数は少ない

人間よりBotの方が行削除数は多い

分析結果：人間のHot Fix vs AgentのHot Fix

テストファイル変更率

PR種別	作成者	qwen2.5-3b	llama3.2	phi4mini
Hot Total	Human	29.03%	47.50%	31.15%
	Bot	33.33%	72.97%	61.54%

人間よりBotの方がテストファイルを変更している場合が多い

考察：Hot Fixと通常のバグ修正の違い

Hot Fixは複数の側面で通常のバグ修正とは異なる構造

- 変更ファイル数が少ない
- 行追加・削除が少ない

コラボレーション：

- ホットフィックスは通常のバグ修正より少数の開発者によって取り組まれる

テストの取り組み

- テストコードの記述を省略しがち

考察：人間とBotの修復戦略の対比

指標	人間	Bot
変更ファイル数	多い	少ない
コード削除量	多い	少ない
コミット数	多い	少ない

変更の範囲：

- 人間：複雑・相互依存的な範囲の問題に広く対処
- Bot：局所的な範囲の修正に集中

論文が述べていた考察の結論：

人間は複雑さと相互依存性を管理する一方で、Botは局所的な修正に焦点を当てていることが示唆される。

結論

- | ホットフィックスの初の大規模実証研究かつ，人間vsエージェントの初の直接比較
- | ホットフィックスは通常のメンテナンスとは著しく異なる構造パターンを持つ
- | 本研究は野生のホットフィックス理解のベースラインであり，将来のホットフィックス自動化エージェント設計に示唆を与える

所感

人間は複雑さを管理している←本当？

削除数が多いのは人間が変更案考えるの面倒で、問題になった部分消してるだけなんじゃ・・・

重要な(criticalな)Issue、とは何なのか

- 開発者が重要と感じるIssueの包括的な分類をするだけでもこの論文が取り組む問題は解決しやすくなりそう

