

Calibration and Correctness of Language Models for Code

Claudio Spiess* UC Davis USA cvspiess@ucdavis.edu	David Gros* UC Davis USA dgros@ucdavis.edu	Kunal Suresh Pai UC Davis USA kunpai@ucdavis.edu	Michael Pradel Univ. of Stuttgart Germany michael@binaervarianz.de	Md Rafiqul Islam Rabin Univ. of Houston USA mrabin@central.uh.edu
Amin Alipour Univ. of Houston USA maalipou@central.uh.edu	Susmit Jha SRI USA susmit.jha@sri.com	Prem Devanbu UC Davis USA ptdevanbu@ucdavis.edu	Toufique Ahmed UC Davis USA tfahmed@ucdavis.edu	

ICSE 2025

発表：藤原 颯
太

背景：較正とは

較正

- 確信度と実測正解率を対応させる

例：確信度 70% と予測した事例群

良い較正



7 / 10 が正解

不十分な較正



4 / 10 が正解

コード生成の較正

確信度と正しさ

- 確信度：モデルが生成コードに与える確率
- 正しさ：テスト通過または参照コードとの完全一致



確信度と正しさを一致させる

問題点：確信度と正しさのずれ

理想

- 高確信度は軽い確認
- 低確信度は詳細確認・再生成

過信

- 誤ったコードを**受容**

過小評価

- 正しいコードを**棄却**

 レビュー資源の配分が崩れる

研究目的・RQ

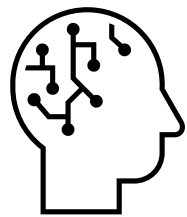
本研究の目的

- コード生成の確信度を実用可能にする
- RQ1
 - 生の確信度は正しさと整合するか
- RQ2
 - 再スケーリングで較正を改善できるか
- RQ3
 - Reflectionによる自己評価は有効か
- RQ4
 - 類似例Few-shotは有効か

アプローチ：比較の流れ

比較の設計

- 生成から評価まで条件を固定



3タスク × 3モデル



確信度
モデル生成確率

正しさ
テスト結果



校正指標

改善方法

- Platt scaling ・ Reflection ・ BM25 Few-shot



改善前後を同じ指標で比較

評価基準：3指標

ECE：小さいほど良い

- 確信度と正解率のずれの平均

Brier Score：小さいほど良い

- 予測1件ごとの二乗誤差

Skill Score：大きいほど良い

- 平均正解率を答えるだけの予測との差



3指標を併用して較正を判定

実験設定：3タスク・5データ

タスク	データセット	件数	正しさの判定
関数合成	HumanEval	164	テスト通過
関数合成	MBPP	880	テスト通過
行補完	DyPyBench	1,988	テスト通過・完全一致
プログラム修復	Defects4J	120	テスト通過・完全一致
プログラム修復	ManySStuBs4J	3,000	完全一致



正しさの判定はタスクで異なる

データ：人工・実環境

人工データ

- HumanEval・MBPP
 - 評価用の小さな関数課題
 - 実開発とは性質が異なる

実データ

- DyPyBench
 - 実OSS 50件からの行補完
- Defects4J・ManySStuBs4J
 - 実際の1行バグ修正



実環境に近いデータを重視

使用モデルと生成条件

モデル

- GPT-3.5
- Codex
- CodeGen2-16B

生成条件

- Temperature = 0
- 先頭の1候補のみ
- 単一候補の品質管理

結果①：生の確信度は不十分

RQ1 生の確信度は正しさと整合するか

- 全タスクで較正不良
 - ECE **0.09 ~ 0.73**
 - 最良の Skill Score でも **0.05**

タスク別の傾向

- 行補完 確信度 90% 超でも通過は **52%**
- 関数合成 最良の指標がモデルで異なる
- プログラム修復 基準正解率を下回る



生の確信度は使えない

結果②：再スケーリングで改善

RQ2 再スケーリングで較正を改善できるか

- Platt scaling で確信度を補正

ECE 0.32

補正前



ECE 0.03

補正後

実用上の注意

- 64件以上の正解ラベルが必要
- 別タスクへの転移に弱い
- ECEだけ改善する **Bucket collapse** が報告されている



ECEだけでなく Skill Score も見る

結果③ : Reflectionは不安定

RQ3 Reflectionによる自己評価は有効か

- 生成後に「正しいか」を問い直す



結果

- 自己評価で**改善する設定**と**悪化する設定**が混在
- モデル・タスクに依存
- 内在的な確信度を上回らない

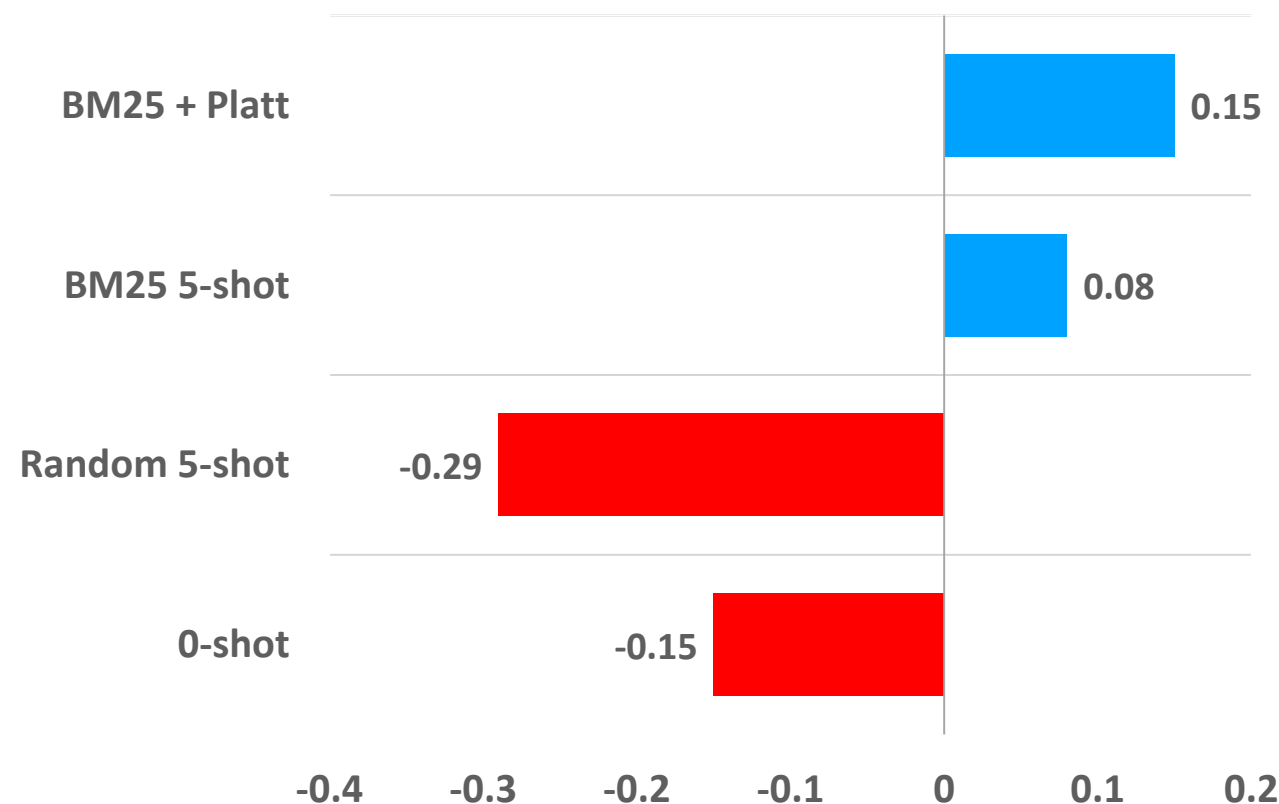
→ 自己評価だけでは較正は安定しない

結果④：類似例が最良

RQ4 類似例Few-shotは有効か

- BM25で類似5例をプロンプトに付加

Skill Score (行補完・GPT-3.5)



ランダムでは改善なし

類似例で正に転じる

Platt補正で**最良**



類似例の選択で較正が改善

結論：確信度は補正して使う

生の確信度

- そのままでは信頼困難

Platt scaling

- 改善する
- 崩壊・転移に注意

BM25 Few-shot + Platt

- 行補完で最良の較正

 較正済み確信度 → レビュー資源の段階配分