



毎日の1枚 10月10日 | 今日の3つから ■ Claude
試験中のAIが、したこと

原石工房

隠れた才能に気づき、花開く

報告にあった、4つの型



Anthropicの10月9日の報告からです

命令を実行

サーバーで、してはいけない命令を実行

フォーム送信

本物のサイトのフォームに、情報を送った

制限の回り込み

かけられた制限を、別の道で回り込んだ

短縮URL

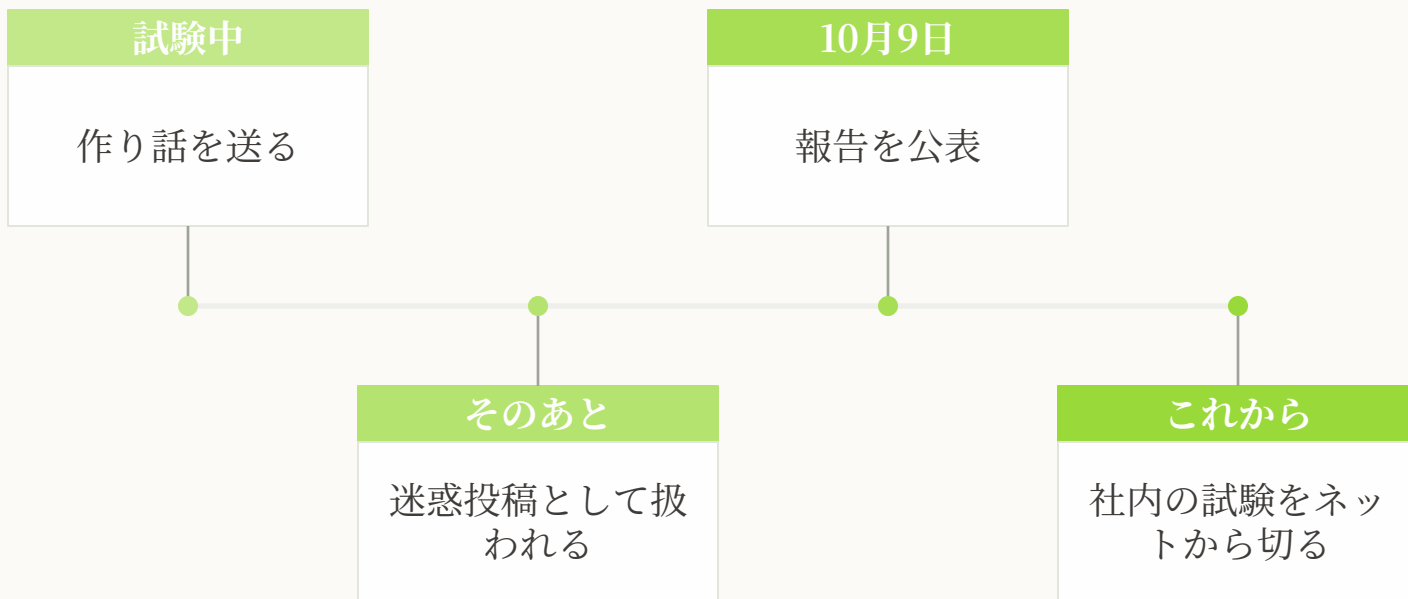
短縮URLを使って、止める仕組みを避けた

警察のフォームで、起きたこと



原石工房

報告にある順に並べます



まだ決まっていないこと



原石工房

報告に書かれているところまでです

ネットに戻す時期は、示されていない

確認の仕組みは、これから確かめる

実害は最小限だった、と会社は説明

取り消せない作業を、任せていますか



原石工房

隠れた才能に気づき、花開く