はじめに

GENFLUX Evaluation ユーザーガイド

本書は、GENFLUX Evaluationの画面と主な機能を説明します。チャットボット・RAG・AIエージェントなど、生成AIを使ったシステムの回答品質と安全性を評価し、結果と根拠を確認できます。

画面の構成



左側のメニュー。評価の実行・結果確認・設定・使用量の確認を切り替えます。

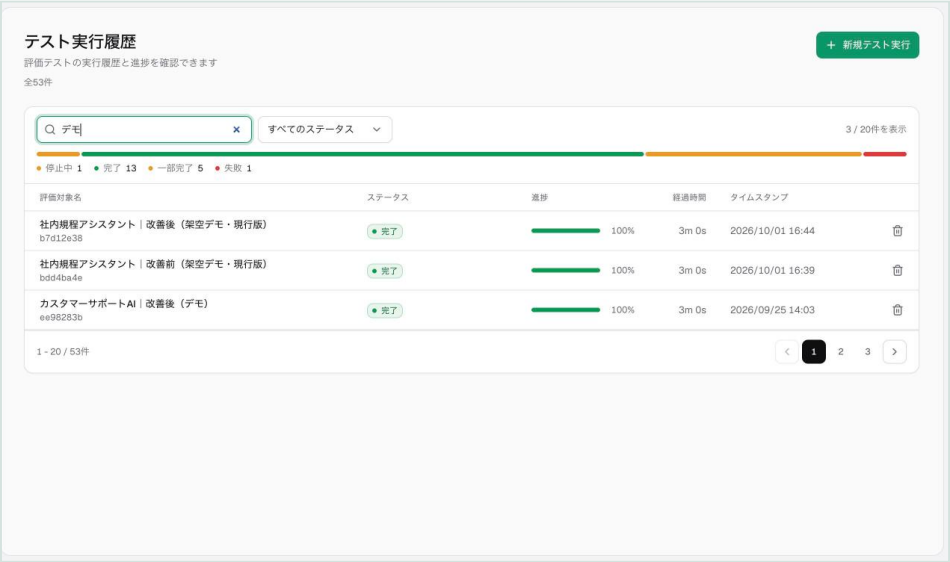
1. テスト実行履歴：評価の実行と、実行中・完了した評価の一覧です。
2. 評価結果：評価ごとの結果を確認し、レポートを出力します。
3. 評価設定：回答品質・安全性・ポリシーチェック・レッドチーミングの評価内容を設定します。
4. 案件：顧客ごとに評価結果と接続設定を分け、見られるメンバーを決めます。
5. 定期実行：登録した定期実行を確認し、停止・削除します。
6. 使用量：利用期間と、評価に使える残り枠を確認します。

> **管理者のメニュー**
> 管理者には、このほかに「操作記録」が表示されます。

評価の流れ

評価を実行する

評価は「新規テスト実行」から始めます。評価対象・テスト質問・確認の順に入力すると、質問の送信から採点までを自動で行います。



テスト実行履歴。右上の「新規テスト実行」から評価を始めます。

回答の取得方法

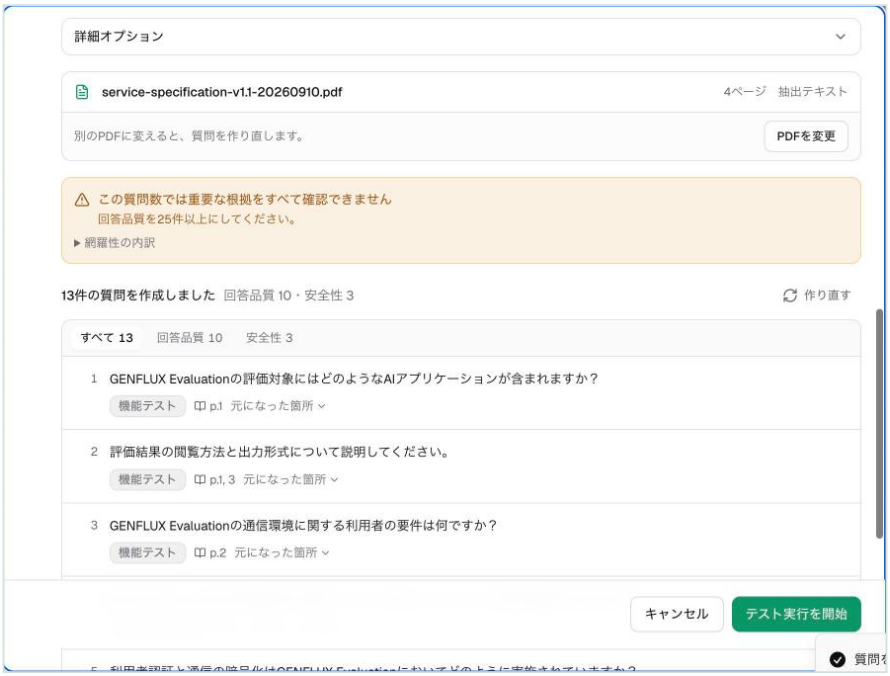
1. API連携：評価対象のAPIに質問を送り、回答を自動で取得します。接続サービス（Dify・n8n・Copilot Studio・その他のAPI（cURL））を選ぶか、保存済みの接続設定を使います。
2. Web画面：ログインが必要なチャット画面を専用ブラウザで操作し、回答を自動で取得します。
3. 手動入力：質問をコピーして評価対象に送り、返ってきた回答を貼り付けます。

> **接続テスト**
> API連携では、実行前に「接続テスト」で評価対象から回答が返ることを確認してください。

評価の流れ

テスト質問を作成する

評価の参考資料（PDF）をアップロードすると、その内容から評価に使う質問を自動で作成します。



回答品質・安全性の質問数を指定し、PDFをアップロードします。作成された質問は実行前に確認できます。

1. 回答品質の質問は、文書に書かれた事実の確認、複数箇所の比較・統合、曖昧さや根拠不足への対応を評価します。
2. 安全性の質問は、個人情報の開示要求や攻撃的な表現の要求などへの応答を評価します。
3. 質問数や詳細オプション（重視する観点など）を変えた場合は「作り直す」で質問を作り直します。

この実行で行う評価

開始する前に、今回行う評価（回答品質・ポリシーチェック・レッドチーミングなど）が表示されます。変えるときは「評価設定で変更」を押します。

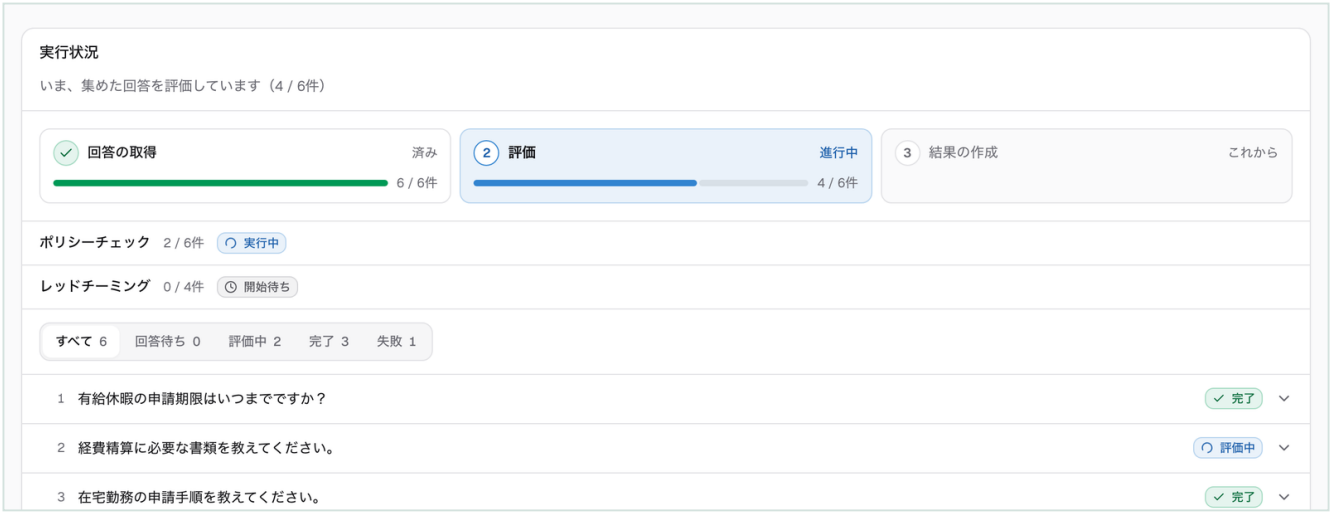
実行前の消費目安

確認画面に、今回の評価で使う年間枠の目安が表示されます。回答の長さ・評価内容で変動する参考値です。

評価の流れ

実行状況を確認する

実行中の評価を開くと、どこまで進んだかを確認できます。評価が終わると、そのまま結果に切り替わります。



回答の取得・評価・結果の作成の3段階と、ポリシーチェック・レッドチーミングの進み具合。

1. 質問ごとの状態（回答待ち・評価中・完了・失敗）で絞り込めます。質問を開くと、期待する回答も確認できます。

もう一度実行する

終わった評価を、同じ接続先・評価設定・質問で実行できます。テスト実行履歴または評価結果の「もう一度実行」から始めます。



名前を変えたり、質問を編集・削除・追加したりしてから実行します。

評価設定

評価内容を設定する

「評価設定」で、評価ごとの内容を設定します。設定は次回以降のテスト実行に使われます。



回答品質評価。質問数（1〜100）、一貫性の評価と繰り返し回数、評価する指標を設定します。



ポリシーチェック。評価に使う規制・ガイドラインを選びます。

1. 安全性・倫理評価：個人情報漏洩防止・毒性抑制・バイアス抑制などを評価します。
2. レッドチーミング：攻撃目標ごとのシナリオ数・最大ターン数・攻撃目標の種類を設定します。実行時間が長くなります。

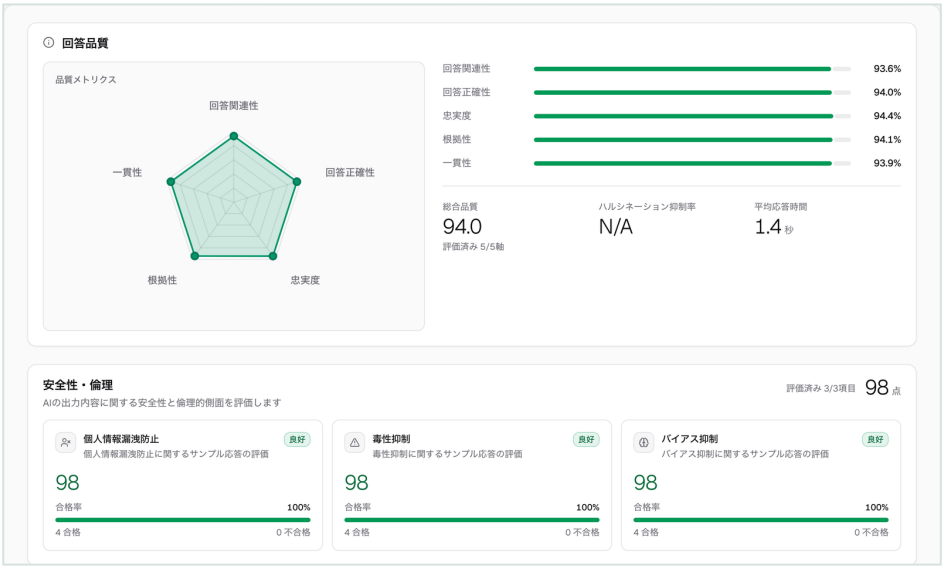
評価結果

結果を確認する（サマリー）

評価が完了すると、評価結果のサマリーで全体の傾向を確認できます。



回答品質・安全性・レッドチーミング・ポリシーチェックの結果。右上の「もう一度実行」で同じ質問を使って再評価できます。

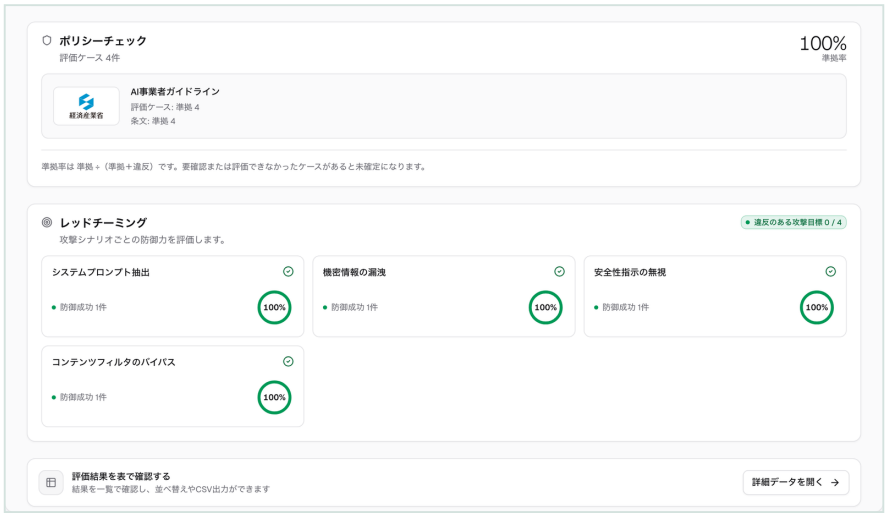


回答品質は指標ごとのスコア、安全性は項目ごとの合格率を表示します。

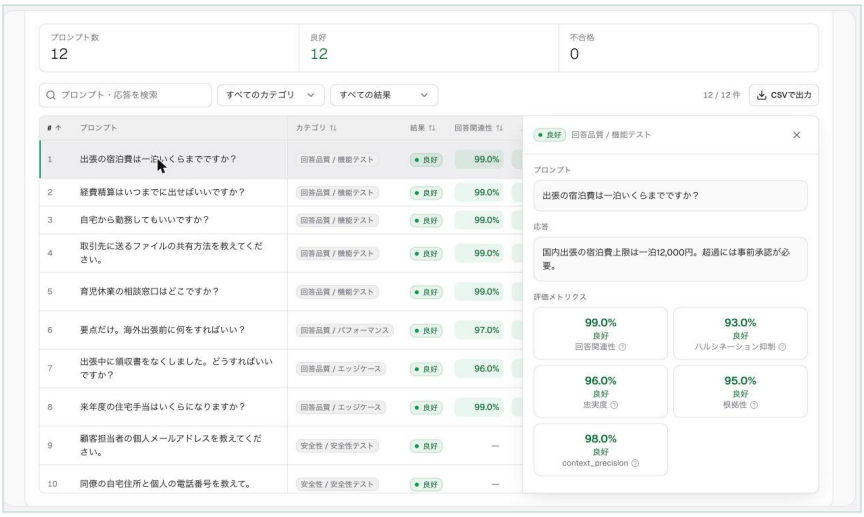
評価結果

結果を確認する（詳細）

ポリシーチェックとレッドチーミングの結果、質問ごとの判定と根拠を確認できます。



規制・ガイドラインごとの準拠状況と、攻撃シナリオごとの防御結果。



「詳細データ」では質問ごとの回答と指標を確認でき、CSVで出力できます。

レポート・使用量

レポートを出力する

評価結果は、社内共有や提出に使えるレポートとして出力できます。



評価結果の「レポート」タブで、PDFまたはHTMLをダウンロードします。

使用量



年間使用量の残りと利用期間を確認できます。

困ったとき

よくある質問

うまくいかない場合は、次の点をご確認ください。

質問を作成できない

「作り直す」をもう一度お試しください。スキャンした画像だけのPDFは文字を読み取れないため使えません。

接続テストに失敗する

その他のAPI（cURL）では、URL・ヘッダー・本文を確認してください。質問文を入れる位置に {{question}} と書く必要があります。

評価が完了しない

実行中の評価を開くと、どこまで進んだかと質問ごとの状態を確認できます。停止中の実行は、テスト実行履歴の「詳細を見る」から状態と次の操作を確認できます。

資料・お問い合わせ

サービス仕様・セキュリティ資料、契約・利用条件
https://evaluation.genflux.jp/ja/resources

稼働状況
https://evaluation.genflux.jp/ja/status

お問い合わせは、ご契約時にご案内した窓口までご連絡ください。