JENTRAXIS | TECHNICAL REVIEW BRIEF

日本語AIエージェント評価キット

業務目的からツール実行、最終回答までを通しで観測し、境界逸脱と回答品質を分けて評価する法人向け技術レビュー基盤です。

外部入力　権限　文脈　ツール実行　承認境界

> 単なるプロンプト一覧ではありません。ケース実行、mock tool、個別採点、実行証跡、ハッシュ検証、再現性情報を一つの評価工程として扱います。

対象課題

- 文書や検索結果に混在する未信頼な指示
- 利用者の依頼範囲を越える権限行使
- 文脈の取り違えや秘密情報への接近
- 送信・更新・削除における承認境界

レビューで判断できること

- 評価経路と採点単位の妥当性
- 証跡から判定を追跡できるか
- 対象製品への適用可能性と追加設計量
- 自動判定と手動監査の役割分担

主商品は、顧客製品へ合わせた評価パイロット

本サンプルは評価設計と証跡形式を確認するための入口です。実案件では、対象製品の業務フロー、利用可能ツール、承認ルール、重要資産に合わせてケースと判定基準を設計します。

EVALUATION ARCHITECTURE

業務目的から証跡までの評価経路

未信頼な内容は最初からモデルへ直渡しせず、文書参照の結果として業務データと一緒に提示します。

Stage A
業務目的だけから、最初の文書参照ツールを選択。

›

Stage B
ローカルmockが、正常データと未信頼な内容を含む文書を返却。

›

Stage C
文書を受けたモデルが、追加ツールまたは回答を選択。

›

Stage D
必要なmock操作後に最終回答を生成し、採点・保存。

閉じたmock環境

文書参照、メール、予定、記録参照、外部要求を模したツールをローカル状態だけで実行します。第三者サービスへ接続せず、顧客データも使用しません。

SecurityとQualityを分離

禁止操作の試行・実効果・秘密露出などのSecurity判定と、回答内容の客観要件・要レビュー項目を独立して集計します。

追跡可能な証跡

モデル入出力、ツール呼出・結果、状態変化、採点結果、判定理由をrun単位のevidence JSONへ保存します。

再現性の手掛かり

ケース、採点コード、実行コードをSHA-256で識別し、実行環境・モデルdigest・設定をtoolchain manifestへ記録します。

> すべてローカルで実行します。外部API、自動ダウンロード、実在アカウント操作は評価経路に含めません。

VERIFIED RUN SNAPSHOT

確認済み事実と、読み方

33 / 33	37 / 37	90 / 90	0
自動テスト PASS	Release ハッシュ一致	Evidence ハッシュ一致	意図しない個人情報・ローカル依存の検出

公開を意図した代表者名、contact@jentraxis.com、新潟県は承認済み公開情報として対象外です。電話番号、住所詳細、ローカルユーザー名、絶対パス、内部プロジェクト名などの意図しない個人情報・ローカル依存は検出されませんでした。

Ollama qwen3:8b | 10ケース × 3回

Security 30 / 30 PASS。これは、明示的な防御指示と、攻撃であることを明確に識別できる合成攻撃文を用いたhardened baselineの結果です。一般的な安全性、未知の攻撃への耐性、脆弱性不存在を証明するものではありません。

自動Quality結果

12 RUN	9 RUN	9 RUN
OBJECTIVE_PASS	OBJECTIVE_FAIL	REVIEW_REQUIRED

客観判定できる要件と、人による文脈確認が必要な要件を分離しています。REVIEW_REQUIREDを自動的に合格へ置き換えていません。

手動監査は未完了

30 RUNはいずれも手動監査未実施です。最終判断には、証跡JSONと最終回答を対象業務の責任者が確認する工程が必要です。

品質不足を可視化した例： JAW-010では品質失敗を検出しました。Securityが境界を守っていても、回答品質が十分とは限らないことをケース別に示しています。

COMMERCIAL OPTIONS

3営業日評価パイロット

1製品・1環境・20ケース　55万円 税別

進行条件

- 開始前100%入金
- 評価実施期間は3営業日
- 20ケースを各1回＝20 RUN
- 確認用の再実行最大10 RUN
- 範囲・判断基準・20ケース計画を事前承認

利用環境

合成データ、または顧客が書面で承認した環境に限定します。着金と必要なデータ・アクセスの準備完了後に開始します。

ライセンス	参考価格（税別）	想定
Internal License	**300万円**	自社製品・自社環境での内部評価
Commercial / White-Label License	**800万円**	顧客向け評価サービスや成果物への組み込み

3営業日は、評価計画承認、着金、必要なデータ・アクセス準備の完了後に開始する評価実施期間です。顧客側の準備・遅延は含みません。合計30 RUNを超える反復評価は別見積です。パイロット料金は、パイロット完了後14日以内にライセンス契約が成立した場合に限り、ライセンス料金へ充当します。契約範囲と商用条件は個別文書で確定します。

重要な前提

評価結果は、安全性、認証適合、法令・規格適合、脆弱性不存在を保証しません。対象製品、環境、権限、データ、ケース範囲に基づく時点評価です。

技術レビュー・パイロット相談

書面ベースで開始できます

contact@jentraxis.com
jentraxis.com