ORACLE

AIエージェントの可観測性と評価

Oracle AI Agent Studioに組み込まれた、エージェントの品質とパフォーマンスを測定、評価、トレース、レポート、および監視するためのフレームワーク

目次

AIエージェントに堅牢な評価フレームワークが必要な理由 3
エージェントの品質：エンタープライズAIの次なる課題 5
AIエージェントの可観測性と評価フレームワークの概要 6
評価 7
LLM-as-a-judge 9
評価データセット 9
AIエージェントを評価するための実践的な5ステップのプロセス 10
トレース 11
レポートと可観測性 12
オラクルのAIエージェントの違い：可観測性と評価 13
オラクルが支援できること 14

AIエージェントに堅牢な評価フレームワークが必要な理由

価値の高いエンタープライズAIアプリケーションでは、エージェントのパフォーマンスを厳密に評価して改善するためのフレームワークが組み込まれていることが重要です。こうした仕組みの有無が、本番環境への導入の成功につながるか、あるいはプロトタイプ開発の無限ループに陥るかを分ける要因となり得ます。このガイドでは、可観測性と評価フレームワークが、正確で高いパフォーマンスなAIソリューションの提供を支援する、使いやすくコード不要のツールキットであるOracle AI Agent Studioに、自動化された測定、評価、トレース、レポート、および可観測性の機能をもたらす方法をご紹介します。また、組織で活用できる、エージェントを評価するための実践的な5ステップのプロセスもご確認いただけます。

エージェントの品質：エンタープライズAIの次なる課題

エンタープライズAIがプロトタイプから本番環境レベルのアシスタントや自律型エージェントに移行するにつれ、問いの焦点は「エージェントを構築できるか」ではなく、「そのエージェントが正確に、確実に、そしてコスト効率よく大規模に機能することを信頼できるか」へと移りつつあります。

AIエージェントはエンタープライズ・ワークフローに深く組み込まれるようになりつつあり、通常は人間が行うタスクを補強し、ビジネス・プロセスを完全に自動化することもあります。Oracle AI Agent Studioを使用すると、目標主導型で、目的の達成を支援するために計画、決定、および行動できるAIエージェントの作成が可能になります。これらのエージェントは動的で、自律型かつ適応型であり、高度な言語モデルと推論モデルを実装しています。しかし、その核心は非決定的であり、従来のソフトウェアとは異なり、これらのAIエージェントはあらかじめ決められた実行経路をたどるわけではありません。

AIエージェントのテストとソフトウェアのテストは別物

- エージェントのレスポンスは、モデルの更新やプロンプトの変更により異なることがあります。従来のソフトウェアでは扱うことのなかったハルシネーションが発生する可能性もあります。
- 「正確性」が二元的であることはほぼありません。回答は、意味的な整合性、関連性、完全性、明確性により判断する必要があります。
- エージェントには、マルチステップの実行や他のエージェントへの作業の委任を伴うことの多い複雑な実行パスが備わっており、自然言語の意図の解釈、ツール（内部および外部）のコール、ドキュメントのクエリなどを実行します。これらはすべて、トレースの必要な誤操作の原因になる可能性があります。
- 毒性、先入観、PII、迅速な注入、規制上の期待など、コンテンツの安全性に関する要件は時間とともに変化するため、継続的監視、監査、説明可能性が必要になります。

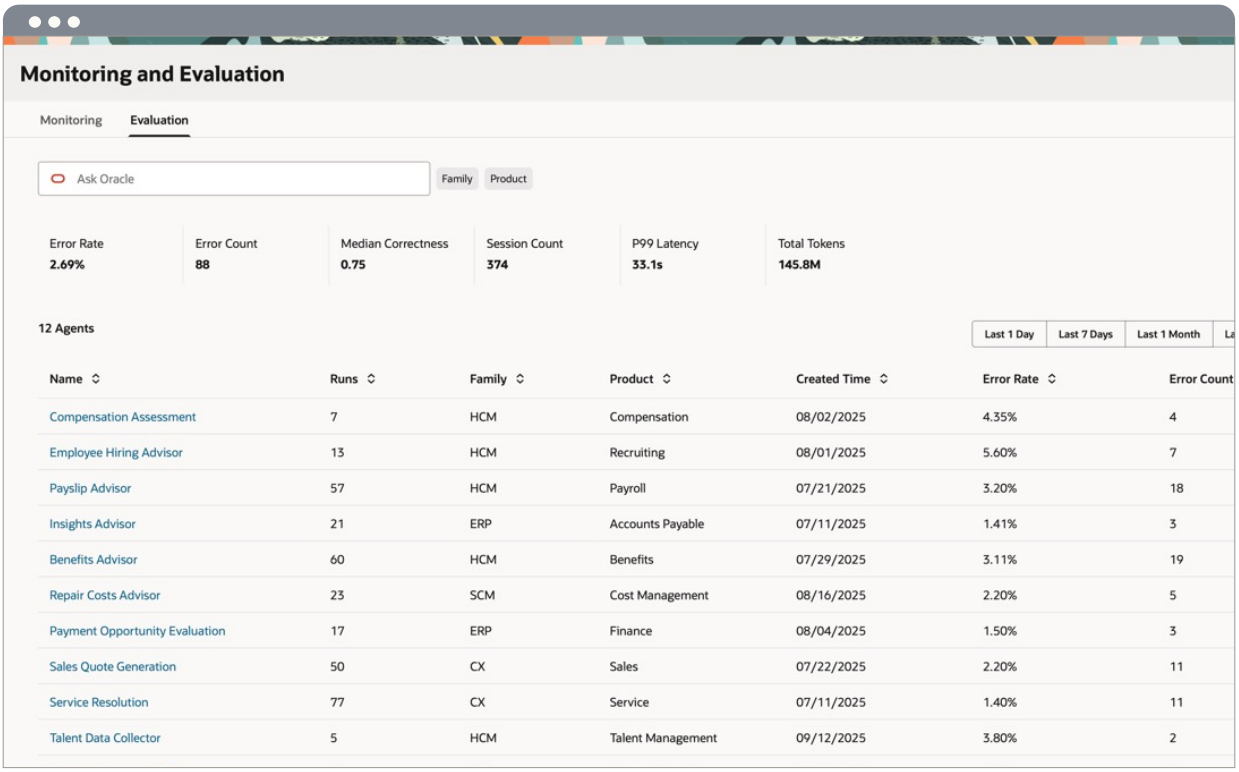
従来のQAテストの仕組みでは、このような分野では不十分です。AIエージェントのテストの難しさは、程度ではなく種類の問題です。AIエージェントの品質とパフォーマンスの評価と監視の問題は、決まったプロセスの検証から、適応型システムの安全で信頼できる正確な動作の検証へとシフトしています。このような現実に対処するために、Oracle AI Agent Studioには、チームがAIエージェントをエンドツーエンドで設計、テスト、導入、および継続的に改善できるように支援する統合的な組み込みのフレームワークを含む、可観測性と評価機能が備わっています。こうした機能により、設計時の評価、詳細なステップバイステップのエージェント実行トレース、および稼働時間の運用監視が単一のエクスペリエンスに統合されます。

> Oracle AI Agent Studioには、チームがAIエージェントをエンドツーエンドで設計、テスト、導入、および継続的に改善できるように支援する統合的な組み込みのフレームワークを含む、可観測性と評価機能が備わっています。

AIエージェントの可観測性と評価フレームワークの概要

Oracle AI Agent Studioの統合機能により、次のことが可能になります。

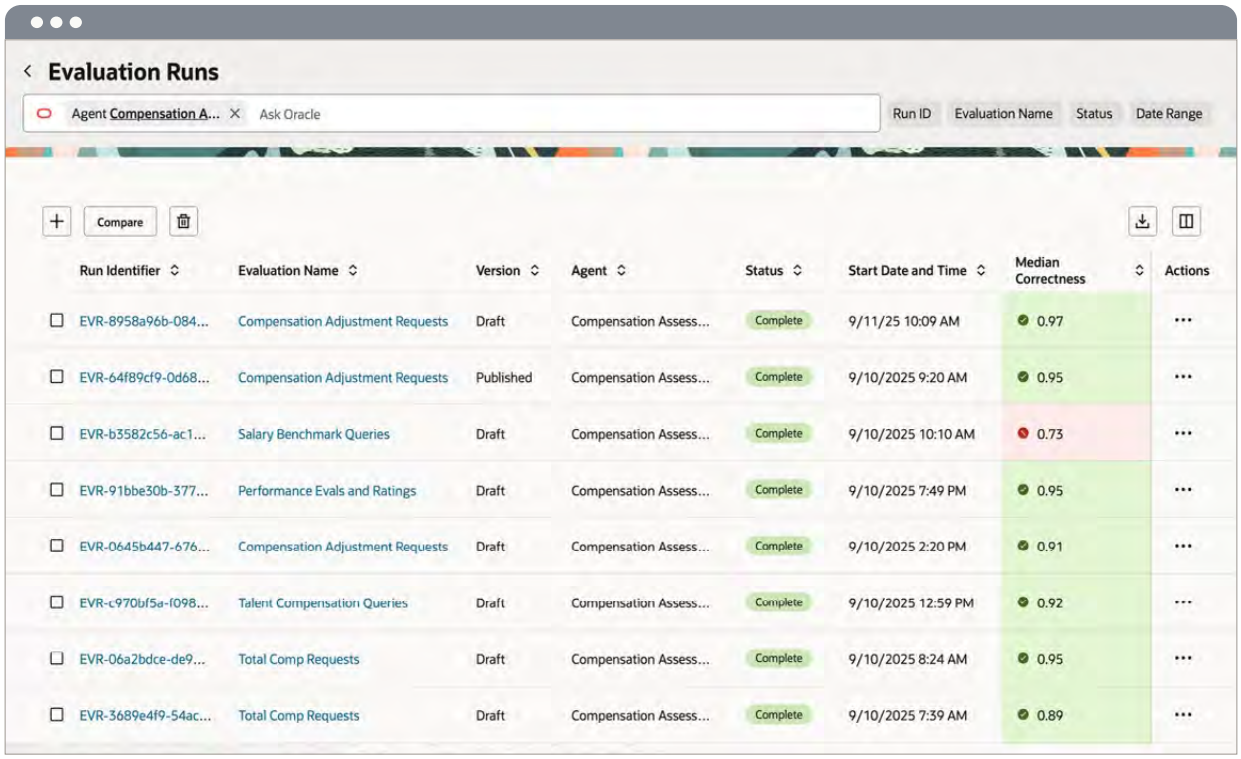
- **測定:**プロンプトとエージェントの品質、パフォーマンス、およびコスト・メトリックの一貫した取得
- **評価:**出力および設定可能なしきい値を使用した、設計時のテスト作成とバージョン間での自動実行
- **トレース:**根本原因分析を迅速に行うための、ツール・コール、LLMコール、レイテンシ、トークン、エラーを含むエージェント・セッションとターンのステップバイステップのイントロスペクション
- **レポート:**プロンプトとエージェントのドリルダウン・ダッシュボードと履歴ビュー、評価実行の比較、およびリーダーボードスタイルの要約
- **可観測性:** フィルタ、時間ウィンドウ、レイテンシ/エラーによる本番環境レベルの監視



評価

設計時の品質と安全、運用の実現

オラクルの評価機能は、導入前の厳しい反復可能なオフライン・テストをサポートします。チームは、入力と予想される回答/出力をペアにして評価セットを定義し、計算するメトリックを指定します。評価実行は、ドラフト・エージェントと公開されたエージェント、シードされたプロンプトと上書きされたプロンプトなど、特定のバージョンを対象に実行することができます。



主な機能:

評価データセットの管理

- エージェントのテスト・ケース（質問と期待される回答）の作成
- 参照評価セットの一括アップロード、編集、エージェントの進化に合わせた再実行
- データセットとターゲット・エージェントの関連付けと、メトリックの選択による合格/不合格のしきい値の計算と指定

LLM-as-a-judge (LaaJ)の正確性

- [ ] 必要に応じてオプションで人間が注釈を付ける専門の判定用大規模言語モデル(LLM)を使用する、自動化されたセマンティック・スコアリングと解答に対する正確性の説明

実行と結果

- [ ] 特定のバージョンを対象に実行し、正確性の中央値、レイテンシの中央値、トークンおよびエラーに関する実行レベルのサマリーの表示
- [ ] ドリルダウンによる個々のテスト・ケース、エージェントのレスポンス、および正確性、レイテンシ、トークン数、エラー・フラグを含むメトリックの調査
- [ ] しきい値駆動の合否指標による回帰の強調表示と、有用な場合には、実際の出力と期待される出力との差異の表示

A/B比較

- [ ] 同じ評価セットを2つ並べて実行して比較し、正確性、レイテンシ、トークン・コスト、エラー動作の変化の定量化

評価で利用可能なメトリック

- [ ] 品質：正確性(LaaJベースの評価)および合格数/率
- [ ] パフォーマンス：レイテンシ（中央値、P99）とエラー（APIコールのタイムアウト）
- [ ] コスト：プロンプト／入力トークン、出力トークン、トークンの総消費量、中央値、P99トークン数
- [ ] 使用状況（該当する場合）：ターン（エージェント）およびLLMコール

LLM-as-a-judge

セマンティック品質評価のスケーリング

従来のユニットテストは、オープンエンドの言語出力に苦労しています。これは、期待される回答に対するセマンティックな正確性を評価するために、独立した判定用LLMを使用する原則に基づいた方法であるLaaJによって対処されます。

- チームは評価セットで期待される回答/出力を定義します。
- 実行中、判定用LLMは各回答の意味的な整合性をスコアリングし、0～1の尺度で正確性のメトリックを提供するとともに、スコアリングの根拠を説明します。人間による評価者は、必要に応じて評価を調整し、また、監査など、適宜スコアや説明を追加することができます。
- 中央値のような集約された正確性は、実行する要約や比較において重要なシグナルとなります。
- LaaJの使用は、さまざまなメリットを数多くもたらします。LaaJは、包括的な人間による確認なしで、大規模なテスト・スイート全体に品質測定をスケールするため、人間の監視がより少ない言語で評価を実施できます。また、LaaJは完全一致のメトリックよりも、ユーザーが認識した品質をよりよく反映する繊細な評価を提供します。

テスト資産とメトリックの標準化と自動化されたバージョン対応の実行を可能にすることで、「品質」に関する主観的な議論の削減を支援し、評価を日常的なエンジニアリング業務にします。LaaJスコアリング、レイテンシ/コスト測定、実行比較を組み合わせることで、AIエージェント設計と反復的改善へのエビデンスに基づく経路を提供します。

評価データセット

必須要素

評価を確実に成功させるには、評価データセットがエージェントの機能をストレス・テストするように設計されていることが重要です。テスト・ケースは、ターゲット・オーディエンスからの幅広い入力と、言語、文法、冗長性の多様性を表すように選択する必要があります。また、エージェントのすべての機能要素および敵対的な攻撃、エッジ・ケース、障害ケースを完全にカバーする必要があります。可観測性と評価フレームワークの自動化と管理機能は、幅広い評価データセットを簡単に作成、実行、分析できるようにします。

AIエージェントを評価するための実践的な5ステップのプロセス

組織は、AIエージェントの評価を運用するために、次の5ステップのプロセスを導入することができます。

1 目標とメトリックの定義

- エージェントが解決を目指すビジネス上の問題の明確化
- 成功基準（解決率や顧客満足度など）の設定
- これら基準の測定可能なメトリックへの変換

2 代表的なデータセットの構築

- 実際のシナリオを反映したテスト・ケースの開発
- 「ハッピーパス」と曖昧なクエリ、敵対的なクエリ、スコープ外のクエリといった「アンハッピーパス」の両方のシナリオの組み込み

3 評価の実施

- データセットに対するエージェントの実行による、レスポンス、推論、ツールの使用状況の取得
- 自動評価手法と人間による評価手法の混合適用

4 結果の分析と解釈

- 許容可能なパフォーマンスのしきい値の決定
- 意図の解釈ミスや不正確なツールの使用といった障害パターンの特定
- 結果の集計と成功基準に対するベンチマーク

5 繰り返しと改善

- 評価を循環的なものとして扱うこと
- インサイトを使用したプロンプト、構造またはツール統合の改善
- 更新されたエージェントを再テストすることによる測定可能な改善の確認

トレース

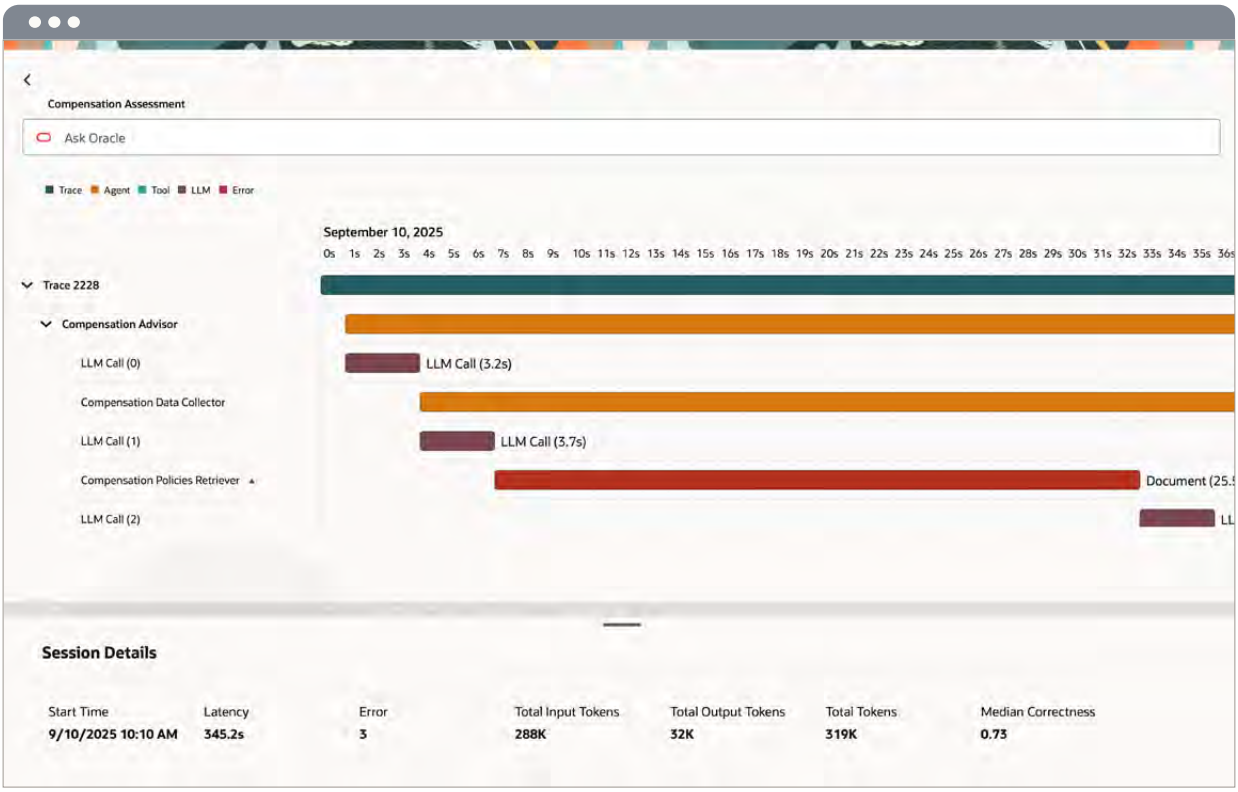
デバッグ対応と信頼を実現する深い透明性

本番環境で問題が発生した場合、またはチームがエージェントの設計を強化する場合、可観測性は集約よりも詳細である必要があります。トレースは、単一のプロンプトとマルチステップ・エージェントに対するステップレベルのイントロスペクションを提供し、開発者と運用担当者による実行フローの分析、ボトルネックの特定、適用の確認、エラーの診断を可能にします。

エージェント実行のトレース

- 全体の概要：開始時間、エンドツーエンドのレイテンシ、エラー、トークンの合計、および（評価のための）正確性。セッションの入力質問と最終レスポンスを表示します。
- インタラクティブなスケジュール：ステップ（従業員エージェント、LLMコール、ビジネス・オブジェクト、REST、ドキュメント/RAG、計算機、セッション、メール、ディープ・リンクなどのツール）の時系列的な可視化。各ステップには、タイプ、タイミング、期間、ステータスが表示され、ステップをクリックすると詳細が表示されます。
- エージェント・ステップ： プロンプト・テキスト、トピック/指示、レイテンシ、API障害の説明、トークン・メトリック、評価実行で表示される入力/出力。
- LLMコール：レイテンシ、エラーの詳細、トークン、評価実行時に表示される入出力。
- ツール：ツールごとの入出力と障害の説明、RAGの場合は取得したチャンクとソースの表示、REST/ビジネス・オブジェクトの場合は機能、パラメータ、ペイロード（評価実行）の表示。

トレースは「起きたこと」と「それが起きた理由」の間のギャップを埋めます。実行の詳細を示すことで、透明性を実現します。エージェントは、内部構造を示すことにより、自身を説明します。トレースはエージェントの推論とツールのオーケストレーションを可監査性にし、根本原因分析を加速します。



レポートと可観測性

本番環境での確実な運用

エージェントが導入されると、リーダーは、製品ファミリーおよび製品全体のSLA、予算、リスクを管理するための運用状況の把握が必要になります。モニタリング・ダッシュボードと履歴ビューは、利用状況、パフォーマンス、コスト、品質（評価）、およびプロンプトとエージェントのシグナルの一元化されたコマンドセンターを提供します。

メトリック・ダッシュボード

- デフォルトでは、すべての製品ファミリー/製品で集計し、ファミリー、製品、直近1日から直近3か月までの時間枠でフィルタします。
- リーダーボードとリスト：使用状況によってランク付けされたプロンプトとエージェント：フィルタリング、ソート、ページ区切りによる専用リストおよび実行履歴ビュー、トレースへのドリルダウンによる調査。

可観測性は、逸話を行動に変えます。チームは、ロールアップ、リーダーボード、フィルタ、およびドリルダウン履歴を使用して、コスト状況の継続的な可視化を維持しつつ、外れ値の検出、改善の追跡、および成果が得られる部分への最適化の取り組みの割り当てを行うことができます。

オラクルのAIエージェントの違い: 可観測性と評価

エージェントに特化した設計
可観測性と評価機能は、すべてを単一のLLMコールとして扱うのではなく、スーパーバイザー、従業員エージェント、プロンプト、ツール・コール、RAGといった、マルチステップのエージェントワークフローの複雑さに完全に対応します。

1つのフレームワークで完全なライフサイクル
評価データセットの管理や比較の実行から、本番用ダッシュボードやディープ・トレースまで、これらの機能は一貫したメトリックとインターフェースにより、設計時と実行時のニーズを統合します。

主要な設計要素としてのコスト
コンテンツ・ケア、プロンプト・インジェクション・フラグとトークン・エコノミクスは、後付けの機能ではなく、UI、メトリック、トレースにおいて、正確性やレイテンシと同等に位置づけられます。

エンタープライズクラスの可視性
製品ファミリ/製品ごとのフィルタリング、幅広い時間枠、リーダーボード、および実行履歴は、大規模な組織がポートフォリオ全体でSLAと予算を運用できるよう支援します。

オラクルが支援できること

AIエージェントには、エンジニアリング規律の新しい標準が必要です。Oracle AI Agent Studioの可観測性と評価フレームワークにより、チームはAIエージェントの設計、テスト、導入、および信頼性と効率の継続的な改善に必要な組み込み機能を取得できます。評価セットとLLM-as-a-judgeスコアリングは、本番前の品質を向上させます。トレースは、深い透明性と迅速な根本原因分析を実現します。また、ダッシュボードは、企業が必要とする運用の可観測性をもたらします。この統合アプローチにより、Oracle AI Agent Studioは、オラクルのスケールとオラクル・グレードのガバナンスで、精度、パフォーマンス、総所有コストに対する実際の期待に応えるAIエージェントを構築して立ち上げるための、堅牢で包括的な業界をリードするソリューションとなります。

デモのリクエスト

詳細

オラクルへのお問い合わせ

+050-3615-0035にお電話いただくか、oracle.com/jp/corporate/contactにアクセスしてください

国外の地域については、oracle.com/contactで最寄りのオフィスをお探しください。