生成 AI 組み込み型 SaaS におけるテスト評価の実践

株式会社ブレインパッド　山崎 清仁
株式会社ベリサーブ　松木 晋祐

生成 AI を組み込んだ SaaS が普及するにつれ、品質保証の鍵が「仕様適合性」から「生成品質」へと移りつつある。株式会社ブレインパッドが提供する生成 AI 組み込み型 SaaS であるレコメンド検索サービス「Rtoaster GenAI」を題材にして、株式会社ベリサーブの生成 AI アプリケーション評価検証の知見を生かし、品質基準の策定から評価データ設計、運用時のモニタリングに至るまでの一連のプロセスを行った。本稿では、責任範囲の明確化から評価フレームワークの実装までを事例として紹介する。

背景

従来の SaaS では、仕様通りに動くかを確認することが主にシステムテストレベルにおける関心ごとの中心であった。一方、生成 AI を組み込んだ機能では、ハルシネーション、バイアス、毒性、プロンプトインジェクション、規制対応など、従来の SaaS では想定されなかった品質リスクが顕在化する。これらは「仕様通り動くか」では捕捉できず、「生成された内容が妥当か」という評価軸が新たに必要となる。
産業技術総合研究所が公開する「機械学習品質マネジメントガイドライン」は、その第 4 版において機械学習タスクの品質体系を整理しており、2025 年 5 月には「生成 AI 品質マネジメントガイドライン第 1 版」が公開された。これらのガイドラインは品質保証の土台を提供するものの、具体的な評価指標や合格基準はプロダクトごとに構築する必要がある。同ガイドライン（および関連する品質要件定義）では、生成AIシステムにおける複雑なサプライチェーンを踏まえたステークホルダーの責任分界や、生成AI特有のリスクに対する抽象的な「品質要件」が網羅的に示されている。しかし、これらのガイドラインは品質保証の土台を提供するものの、実際のビジネス要件や対象システムの特性と照らし合わせて、どのように具体的な評価指標や合格基準へとテーラリング（適合化）するかはプロダクトごとに構築する必要がある。本稿は、産総研のガイドラインが求める要件を、実際のSaaS開発プロセスにおいていかにテーラリングし、運用可能な評価フレームワークへと落とし込んだかを示す実装解の一つである。

品質保証の責任範囲の整理

生成 AI 組み込み型 SaaS の品質を論じる上で、まず整理すべきは「誰が何の品質に責任を持つか」である。生成 AI の挙動は LLM 提供者、SaaS 事業者、利用ユーザーの三者の関与によって決まるため、責任範囲を明確に分解しなければ、品質改善の打ち手が定まらない。まずは、責任範囲を以下のように整理した。

- LLM 提供者の責務： 基盤モデル自体の性能（学習データの素性、モデル更新の方針等）

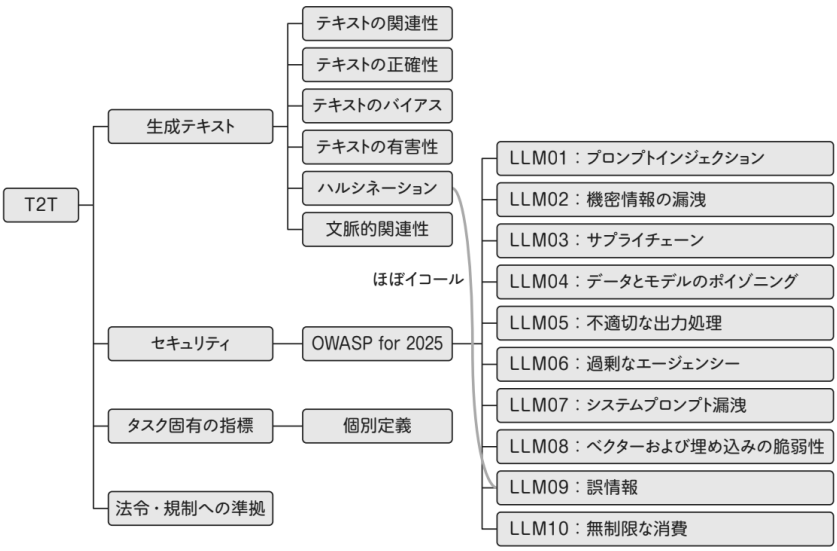
- SaaS 事業者の責務： プロンプト設計、RAG の品質、ガードレール設計、入力データの評価・品質、UI/UX 上での出力の扱い
- 利用ユーザー： 利用するときの文脈、入力内容、出力の最終的な解釈

利用ユーザーから見ると、応答品質が LLM 提供者の責務によるものか SaaS 事業者の責務によるものか、責任主体の区別はつきにくい。そのため、SaaS 事業者は「自社が責務を負う範囲についてどのような評価を行っているか」を説明可能な状態にしておく必要がある。もちろん、SaaS 事業者は、サービス提供者という立場は免れないため、実質的には利用ユーザーから見えるのと同じように LLM 提供者の責務も負うこととなるが、この透明性が生成 AI 組み込み型 SaaS への信頼を構築する出発点となる。

評価指標の体系化

評価指標を考える際に、従来のような機械学習タスクで用いられてきた「正解率」のような単一指標では不十分である。生成 AI では「正解が一意に定義できない」ケースが多く、妥当性や安全性といった複数の評価軸を組み合わせる必要がある。
品質基準策定には、テキスト入力・テキスト出力を行う生成 AI アプリケーション全般に適用可能な観点を構造化した「T2Tテスト観点基盤モデル」（下図、T2Tは Text to Text の略）を起点とした。このモデルは、産総研のガイドラインが定義する網羅的な品質要件（機能正確性、安全性、公平性など）を包含しつつ、それらを実際のテスト設計の現場で測定可能な「評価指標」へとテーラリングしたものである。T2Tテスト観点基盤モデルはベリサーブの松木が提案している概念で、テスト要求分析・基本設計の段階で検討すべき〝どんなテストを行うか〟を考えるための基礎となる観点を体系化したモデルである。これはあくまで「基盤」であり、個別のアプリケーションに対してはカスタマイズして用いることを前提としている。

【図】T2Tテスト観点基盤モデル



出典：生成AIアプリケーションの評価入門 第2章 図2，技術評論社，2026

T2Tテスト観点基盤モデルは、生成 AI アプリケーションの評価観点を以下の四つの大分類に整理する。

- 生成テキストの品質： 「テキストの関連性」「テキストの正確性」「テキストのバイアス」「テキストの有害性」「ハルシネーション対策」「文脈的関連性」の６つの観点に細分化される。テキストの関連性は利用ユーザーの入力に対して適切かつ妥当な応答になっているかを測り、正確性は出力情報が事実とどの程度合致するかを評価する。文脈的関連性は複数ターン対話における履歴の反映度を、ハルシネーション対策は学習データに存在しない情報を事実のように出力する現象への検知・抑制を扱う。
- セキュリティ： 従来の Web アプリケーション脆弱性に加えて、プロンプトインジェクションや機密情報漏洩など LLM 特有の攻撃手法・脆弱性への対応を求める。OWASP Top 10 for LLM Applications 2025 を参考に、規格の改訂に追随することが推奨される。
- タスク固有の指標： アプリケーション目的に応じた評価指標を採用する。要約タスクでは ROUGE スコアによる網羅性・簡潔性、翻訳タスクでは BLEU スコアによる正確性、対話型アシスタントではやり取りの円滑さや感情トーン、クリエイティブ生成では表現力・独創性といった評価メトリクスが用いられる。
- 法令・規制への準拠： 個人情報保護(GDPR 等)、著作権・知的財産権、医療・金融などの業種別規制、ヘイトスピーチ等のコンテンツ規制への適合を確認する。グローバル展開では地域ごとの規制差を踏まえた監査証跡・ドキュメント整備が重要となる。

品質基準策定のプロセス

品質基準策定にあたっては、次の 5 ステップからなるプロセスで行う。

1. 指標選定： ユースケースから逆算して、必要な品質要求を抽出する。検索系では関連性とハルシネーション対策、対話系では文脈的関連性、すべての系で安全性を必須項目として組み込む。
2. 重み付け： 事業特性に応じた優先順位付けを行う。Rtoaster GenAI のような検索系サービスでは「関連性」を重視するが、「安全性」を無条件に妥協しない。
3. 閾値決定： 業務要件に基づいた合格ラインを策定する。閾値は初期段階ではベースライン測定からスタートし、運用を通じて見直す。
4. 評価実施： 評価データを用いて指標ごとのスコアを測定する。
5. 運用モニタリング： リリース後も同一指標を継続的に測定し、改善サイクルを回す。

T2Tテスト観点基盤モデルを用いて、指標を選定するときのポイントは次の三点である。これらは、本事例の品質基準策定全体を貫く原則でもある。

- 基盤モデルから個別アプリケーションへの具体化という階層的アプローチで設計する。
- サービスの特性に応じて評価軸の優先度や採否を判断するというコンテキスト依存性を持つ。
- LLM 固有の脆弱性や規制は急速に進化するため、評価基準そのものを継続的に更新する必要がある。

指標の選定では、例えば、マルチターン対話を伴う機能であれば「文脈的関連性」を上位に置くが、単発の問い合わせ応答であれば優先度を下げたり、EC 検索のように検索性が中心となるサービスでは「関連性」と「ハルシネーション対策」を重視したりして、T2Tテスト観点基盤モデルをそのサービスに適合するように評価軸に強弱をつける。
一方、医療・金融・教育のようにサービスが利用されるドメインにも配慮が必要である。ドメイン固有のリスクが大きい場合は、汎用観点に加えて業界固有の観点を加味するとよい。汎用指標だけでは業界固有の要件やリスクを拾いきれないためである。

評価データ設計が品質を決める

本事例の評価実装は、ベリサーブが提供する「QA4AIエージェント」サービスを軸に進めた。同サービスでは、テキスト生成型 AI の汎用テスト観点策定、プロダクト特性に応じたテーラリング、評価指標の設計指針、評価メトリクスの自動計測ツールを提供するもので、これらをもとにして Rtoaster GenAI の評価フレームワークを構築した。

オープンソースの評価フレームワーク「DeepEval」は、T2Tテスト観点基盤モデルの「生成テキスト」および「セキュリティ」に対応した評価機能を既に提供しており、実装の足掛かりとして活用することができる。DeepEvalは、GEValという評価ステップ、評価メトリクスの計測をQAエンジニア自身で構築できる機構を持っており、これは「個別のタスク」の評価に役立てることができる。また、DeepEval は、Pytest ライクな記述で評価ロジックを組み立てられるため、既存のテストパイプラインへの統合が容易である。評価メトリクスごとに閾値をCI に組み込めば、リリース前の品質ゲートとしても機能する。評価結果は日本語で自動的にレポーティングされるよう仕組みを整えることで、品質管理業務の負荷軽減と改善サイクルの高速化に寄与した。

ここで重要となるのが、評価そのものを担う LLM-as-a-Judge の運用方針である。被評価対象である Rtoaster GenAI に組み込まれた生成 AI は、基盤モデルの更新やプロンプト・RAG コーパスの調整に伴って、応答内容が次第に変化する。これに対応するために、評価する側の LLM を固定化することで、被評価モデルが変更されたときに「どこをチューニングすべきか」を素早く特定できる。評価メトリクスが動くと変化の切り分けができなくなるため、このような設計は、運用品質を維持するうえで効果的な手立てとなる。

従来のテスト手法と同様ではあるが、LLMの評価において特に重要な点はテストケースと評価データの品質が、そのままプロダクト品質に直結するということである。どれほど優れた評価指標を選定しても、評価データが代表性を欠いていれば、評価結果は実運用での挙動と乖離する。
そこで評価データ設計では以下の三点を意識した。

- 代表性のある評価データセット： 実利用での問い合わせ分布を反映したサンプル設計
- エッジケース： 境界条件や想定外の入力を意図的に含める
- 失敗パターン： 実運用で起きやすい誤回答や不適切回答を再現するケース

近年は評価データそのものを生成 AI で自動生成するアプローチも普及しているが、AI が作ったデータの偏りや誤りは、そのまま評価の歪みになる。生成効率は向上するものの、人による妥当性検証は不可欠である。

本事例では、全数検査は現実的でないため、サンプルの代表性、重要ユースケース、高リスクケースを重点的に検証するサンプリング設計を採った。実プロダクトでは、利用ログの分布を参照しつつ、カテゴリ別の代表ケースと高リスクケースを織り込み、評価指標の信頼性と評価コストのバランスから 500 件規模の評価データセットを構築した。評価者間のばらつきを抑えるため、サービス評価ガイドラインを社内で明文化し、複数評価者で同一サンプルを評価して整合性を確認した。

運用時のモニタリング

安全基準を高く設定するほど、サービスの回答は消極的になる。「お答えできません」という回答が頻発すれば、利用ユーザー体験は損なわれる。安全性と UX のバランスを取るために「どの失敗を許容できるか」を運用ルールとして明文化した。
具体的には以下のような運用設計を採用した。

- サービス運営上認められていない行為を示唆する回答は、フィクションを装った問い合わせであっても抑制する。
- 「答えられない」だけの回答ではなく、答えられない理由を利用ユーザーに説明する。
- 安全性を理由に回答を抑制した場合、その判断ログを蓄積して後で分析する。

この運用設計のもと運用のなかで評価を行うが、その閾値の合理化は段階的に進めた。初期段階ではベースライン測定からスタートし、運用を進めながら利用ユーザー満足度や問い合わせ完了率との相関を監視する。最終的には A/B テストやサンプルレビューを通じて、体験スコアと評価スコアが整合するラインを探索した。

生成 AI 組み込み型 SaaS の品質は、リリース時点で固定されるものではない。LLM 提供者がモデルを更新したり、参照ナレッジが更新されたり、季節要因で利用ユーザーの問い合わせ傾向が変化したりすることで、品質スコアは時間とともに変動する。
このため、リリース後も開発時と同一の評価指標を継続的に測定する。さらに、以下のメタ情報を評価結果と紐付けて記録する。

- モデルバージョンの更新履歴
- システムプロンプトの変更履歴
- 参照ナレッジ（RAG コーパス）の更新履歴
- 季節要因や外部イベントの記録

これらと評価スコアを多層的に分析することで、品質劣化の原因を切り分けられる。「先週からハルシネーションスコアが低下した」という事象が観測されたとき、モデル更新が原因か、ナレ

ッジ更新が原因か、それとも利用ユーザー入力の傾向変化が原因かを特定できなければ、適切な対応ができなくなる。

SaaS として得られた効果

サービス評価ガイドラインを策定し、その評価を運用することにより次のような効果が得られた。

- **客観的な品質説明**： 品質をスコア化することで判断基準が明確になり、SaaS を導入するユーザーに対して、品質の良し悪しを説明しやすくなった。
- **属人化の解消**： 開発・運用の現場で属人化しがちな品質判断業務を、評価フローと自動レポーティングによって標準化できた。
- **改善サイクルの高速化**： 評価結果を日本語で即時にレポーティングする仕組みにより、品質管理業務の負荷が軽減され、改善のサイクルを回しやすくなった。
- **テストの網羅性・客観性の向上**： 従来は経験に依存していたテスト観点を体系化したことで、網羅性と客観性が同時に高まり、社内外に対して評価指標に基づく説明が可能になった。

レコメンドや対話型検索のような自由形式・非構造化データを扱うアプリケーションでは、直交表やデシジョンテーブルといった従来のテスト技法が適用しにくく、テスターの経験依存が大きくなりがちであった。本事例で構築した評価プロセスは、こうした領域における QA の体系化に対する一つの実装解として位置づけられる。

両社では、本事例で構築した品質基準を Rtoaster GenAI という単一サービスの運用に閉じず、他社が提供する生成 AI サービスにも適用範囲を広げていくことを想定している。生成 AI プロダクトが業界全体として健全に発展するためには、品質基準の体系化とその社会的共有が不可欠であるという認識のもと、ベースとなる基準の継続的なアップデートを進めていく。さらに、AI 駆動開発のように生成 AI が開発工程の上流に入り込む流れも踏まえ、開発と QA の双方向で生成 AI に向き合う方法論の整備を進めていく。

事例のまとめ

本事例では、生成 AI 組み込み型 SaaS の品質保証を実プロダクトに適用する過程で整理した「責任範囲の明確化」「T2Tテスト観点基盤モデルを起点とした評価指標の体系化」「5 ステップの品質基準策定プロセス」「評価データ設計」「運用時のモニタリング」について紹介した。要点を改めて整理すると、

- **責任範囲の明確化**： LLM 提供者と SaaS 事業者の責務を分解し、自社の責務範囲についての評価を説明可能にする。
- **基盤モデルの活用**： T2Tテスト観点基盤モデルのような構造化された評価観点を起点に、プロダクト特性で選別する。

- **データ駆動設計**： 評価データの品質管理がそのままプロダクト品質に直結することを認識し、データ設計に投資する。
- **運用ループへの組み込み**： リリース前の評価ルールをリリース後も継続し、メタ情報と紐付けて継続的改善を回す。
- **透明性の確保**： 評価基準・手法・結果をステークホルダーに説明できる体制を構築する。

生成 AI をビジネスに取り入れる企業は急速に増えている一方、導入時にリスクを十分に算出できないまま運用を始め、利用ユーザーからの指摘で初めて課題に気付くケースは依然として多い。ハルシネーションのリスクは、現時点の技術ではゼロにすることはできない。このため、リスクが残存することを前提に、サービスの設計とサポート体制を構築する姿勢が求められる。産総研が公開する「生成 AI 品質マネジメントガイドライン」および品質要件定義は、業界の重要な土台である。しかし、ガイドラインが示す要件を実社会のプロダクトへ適用するには、事業のドメイン知識に基づく「テーラリング」が欠かせない。本事例では、ガイドラインの「サプライチェーンの責任構造」を事業者とLLM提供者の責任分解へ、「抽象的な品質特性」をT2Tモデルによる具体的な指標へ、そして「継続的なリスク管理要件」をメタ情報と紐づいた運用モニタリングへと、それぞれテーラリングして実装した。本取り組みが、同様の課題に取り組む事業者の参考になれば幸いである。

参考文献：

- 産業技術総合研究所「機械学習品質マネジメントガイドライン 第 4 版」(2023 年 12 月)
- 産業技術総合研究所「生成 AI 品質マネジメントガイドライン 第 1 版」(2025 年 5 月)
- DeepEval（オープンソース LLM 評価ライブラリ）
- 松木 晋祐「生成 AI アプリケーションのテスト観点基盤モデル：T2T(Text to Text)」note, 2025 年 2 月 16 日 (https://note.com/_snsk/n/n1450f1018962)
- ブレインパッド・ベリサーブ共同プレスリリース「生成 AI プロダクトの品質基準を策定、新製品『Rtoaster GenAI』へ搭載」(2025 年 7 月 16 日) (https://www.brainpad.co.jp/news/2025/07/16/23671)
- ベリサーブ事例インタビュー「株式会社ブレインパッド様 – Rtoaster GenAI の品質保証への取り組み」(2026 年 3 月 17 日) (https://www.veriserve.co.jp/asset/achievement/interview12.html)
- ブレインパッド DOORS「ビジネスを取り巻く AI・DX の現状と未来 ～生成 AI 組み込み型 SaaS のテスト評価」(2026 年 3 月 27 日) (https://www.brainpad.co.jp/doors/contents/02_ai_dx_in_business_10/)