LLM を用いたブラウザ操作エージェントによる社内システムの UI の自動評価手法の検討

安達太祐 [1]　中西惇 [2]　淡島英輝 [2,3]　村井貴彦 [1]　西郷彰 [1]
[1] 全日本空輸株式会社　[2] ちゅらデータ株式会社　[3]DATUM STUDIO 株式会社
d.adachi@ana.co.jp　m.nakanishi@churadata.okinawa
h.awashima@churadata.okinawa　t.murai@ana.co.jp　saigo@ana.co.jp

概要

社内の専門的な業務システムの UI 品質は業務効率に直結する重要な要素だが，大規模なユーザビリティテストが困難である．本研究では，ANA グループのシステム刷新に際し，LLM を用いたエージェントによる自動評価手法の導入可能性を検証した結果を報告する．実験の結果，エージェントによる指摘は実務担当者の評価と高い整合性を示し，開発側が見落としがちな現場のつまずきを客観的に可視化できることが確認された．これにより，本手法が評価サイクルの高速化に寄与し，実務への導入が極めて有効であることが示唆された．

1　はじめに

航空運送事業をはじめとして多角的に事業を展開する ANA グループでは，現場オペレーションのデジタル化による生産性向上を強力に推進している．グループ内には 200 を超える基幹システムに加え，多種多様な業務支援ツールが存在しており，これらツールの UI の最適化は，社員の業務効率化のみならず，最終的なお客様へのサービス品質の向上に直結する極めて重要な要素である．しかし，膨大な数のシステムに対して一貫して高品質な UI/UX を実現することは容易ではない．

第一の課題は評価コストと精度である．従来の UI 評価はユーザビリティテストやヒューリスティック評価に依存してきたが，これらの人手を前提とする手法は，頻繁なシステム更新に対して継続的に実施することはコスト面で現実的ではない．また，評価者のスキルによって品質にばらつきが生じやすく，レビューの質の高度化や標準化が課題となる．

第二の課題は社内システム特有の制約である．専門性の高いツールほど機能の拡充に伴って UI が複雑化し，IT スキルが必ずしも高くない現場利用者にとって「操作のつまずき」や「利用放棄」を招くリスクを孕んでいる．また，マニュアル展開の都合から UI の並行運用を維持しづらく，A/B テストなどによる定量的実績の収集が難しいという制約が存在する．

本研究では，これらの課題を解決するため，LLM を用いたブラウザ操作エージェントによる UI 自動評価手法を提案し，その検証を行った．本手法は以下の役割を持つ 2 種類のエージェントを用いる点に特徴がある．

- **利用者役エージェント**：定性評価を担当する．IT スキルが低い社員を模倣し，作業遂行中に生じる困惑点や気づきにくい UI の問題を自然言語で報告する．
- **専門家役エージェント**：定量評価を担当する．Nielsen のユーザビリティ原則に基づき UI の構造的問題を体系的に点検し，スコアとして出力する．

両者を組み合わせることで，利用者が遭遇しやすい具体的な問題と，専門家視点で本来改善すべき構造的な問題の双方を自動的に検出し，UI 改善に有用なフィードバックを得ることができた．本稿では，この手法を社内の膨大なアンケートやサービス改善に関する報告書の分析を効率化する AI ツール「ANAlyzer-Buddy」の刷新プロセスと比較した結果を報告し，AI が生成する指摘が要件定義を行う ANA の IT 担当者の評価，指摘といかに整合するかを明らかにする．

2　関連研究

2.1　Web UI の評価方法

UI の評価方法は現在までに様々なものが考案されている．主なものとしてユーザビリティテストとヒューリスティック評価がある．

ユーザビリティテストは実際の利用者にシステムを操作してもらい，使用感や問題点を直接収集する手法である．確実ではあるが，画面改修のたびに協力者を募集する必要があり継続的な実施は難しい．

1990 年代に Nielsen が提唱したヒューリスティック評価 [1] は，経験則から定めた原則に基づいて専門家が UI を点検する．

Nielsen の 10 原則

1. 進行状況の表示によって，システムが今どのような状態にあるのかがユーザーに伝えられているか．
2. 用語やアイコンが現実世界でよく用いられているものに一致しているか．
3. ユーザーが誤って行った操作を取り消したり，やり直したりできるか．
4. ボタンやリストなど繰り返し使用されるデザインパーツや文言表記がシステム内で一貫しているか．
5. アカウントの削除など重大なミスを引き起こす可能性のある操作には，ポップアップなどでユーザーに確認を求めているか．
6. 情報をシンプルにした上でシステムが次のアクションや選択肢を提示することで，ユーザーが操作に迷わないようにしているか．
7. 初心者にはメニューから操作を選択できるようにし，上級者にはキーボードショートカットを提供するなど，ユーザーのスキルレベルに応じた操作方法を提供しているか．
8. 必要ない情報はなるべく配置せず，最小限で美しいデザインになっているか．
9. エラーが発生した場合，ユーザーにわかりやすいメッセージを表示し，エラーの原因や解決方法を示しているか．
10. ユーザーがヘルプやマニュアルに容易にアクセスできるようになっており，それらは具体的かつ簡潔で，検索しやすい形式で提供されているか．

実利用者を集める必要がないという点で費用対効果に優れ，システムによらない UI の構造的問題の把握に有効である．しかし，専門家による評価は工数を要し，頻繁な画面改修に対して繰り返し実施することが現実的ではないという問題は同様である．

このように，ユーザビリティテストとヒューリスティック評価はそれぞれ異なる観点から有用な洞察を与えるものの，人手を前提とする以上，その自動化が本質的に困難であった．

2.2　ブラウザ操作エージェント

2024 年以降， LLM が外部ツールを呼び出してタスクを遂行するエージェント技術は急速に発展している．特にブラウザ上で DOM 要素の探索，入力，スクロール，遷移などを行う能力は， Web 操作の自動化に広く応用されており， ANA でも導入に向けて様々な調査を行っている．

ユーザビリティテストとヒューリスティック評価は人手を必要とする点で自動化が困難であったが，ブラウザ操作エージェントを用いれば自動化ができる可能性がある．実際， Lu らは多数の擬似利用者を自動生成してユーザビリティテストを行う手法を提案している [2]．ただし，先行研究では評価のために多数の擬似利用者を生成する構成をとっており， UI 改善サイクルの中で短時間かつ反復利用することは困難である．また，評価はユーザビリティテストの設計そのものに焦点を当てていた．本研究の独自性は，先行研究とは異なり UI そのものの評価に焦点を当て，利用者視点と専門家視点を分離した二重構造のエージェント設計により，実務的な UI 改善サイクルに短時間で組み込める実用的手法を提示した点にある．

3　提案手法

本研究では，ユーザビリティテストを行う利用者役エージェントと，ヒューリスティック評価を行う専門家役エージェントを組み合わせ， Web UI の操作性を自動評価する枠組みを導入する．

プロンプトには評価対象とする操作手順を明示的に与える．そのため，システムに存在していても手順に含まれない機能は評価されない．

3.1　利用者役エージェント

利用者役エージェントは IT スキルが低い利用者を模倣するよう設計した．利用中に遭遇する疑問

点，認知負荷の高い表示，見つけにくいボタンなどを随時記述し，利用者が遭遇する潜在的なつまずきを抽出することを目的とする．

評価は 5 回実行し，得られた評価文を要約・統合して最終的な指摘として提示する．これは Nielsen が示した「5 ユーザーで主要な問題の大半を検出できる」という知見による [3]．同一モデル・同一プロンプトであるが， LLM の非決定性を利用して 5 回試行することで，操作の迷いや実行操作の差異を擬似的に抽出した．

3.2 専門家役エージェント

専門家役エージェントのプロンプトには Nielsen の 10 原則を与え，原則に基づいた評価を行う専門家を模倣するよう設計した．評価のばらつきを抑え，短時間での反復実行を可能にするため，各原則を「満たす (1) ／満たさない (0)」の二値スコアで評価した．ヒューリスティック評価を厳密に再現することを目的とするものではなく， UI 改修前後の変化を検知するための簡略化された指標として設計した．

評価は 13 回実行し，平均スコアを出力する．各評価をベルヌーイ試行と仮定すると，その平均は中心極限定理より近似的に正規分布に従う．

UI 改修前後の平均スコアの差は t 検定を用いて比較し，改修後のスコアが有意に低下した場合にはユーザビリティの悪化を検知する．本研究では UI 改修のタイミングごとに短時間で評価を完了できることを重視しており，必要な評価回数を事前に設計可能であることが重要である．t 検定では，検出したい差分と標準偏差の関係から必要な評価回数を比較的容易に定めることができる．評価回数は，有意水準 5 % の左片側検定で改修前後での各原則の平均点の差が標準偏差以上のとき検出力が 80 % となるように設定した [4].

4 実験設定

ANA が開発した ANAlyzer-Buddy は，度重なる機能追加によって煩雑になっていた UI の大幅刷新を本研究と同時期に実施している．そこで， UI 刷新前後に対してエージェントの出力結果を比較する実験を行った．

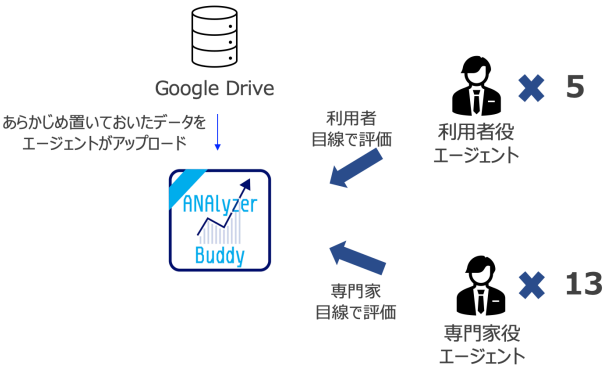

図 1 エージェントによる UI 評価

4.1 対象システム

ANAlyzer-Buddy は，社内アンケートや各種レポートなどの自然言語で記述された社内文書に対して LLM を用いてタグ付けを行うシステムである．複数ステップからなる操作手順を持ち，少人数の利用者が業務に合わせて利用することを想定している．

4.2 使用モデルとライブラリ

エージェントを動作させるのは Gemini 2.5 Pro , 出力の要約・統合には Gemini 2.5 Flash を用いた．エージェントライブラリには Browser Use を用いた．

4.3 タスク設定

プロンプトには評価対象とする操作手順を与えることが可能である．今回は最もオーソドックスなタスクを設定した．

1. 分析ファイルを新規作成する
2. Google Drive に置かれているタグ付け対象のデータをシステムにアップロードする
3. アップロードしたデータに対してタグ付けを行う
4. 作成した分析ファイルを削除する

4.4 評価指標

評価は以下の観点で行った．

- 利用者役エージェントが出力した評価文が 実運用を把握している IT 担当者の評価と整合しているか．
- 専門家役エージェントの平均スコアについて UI 刷新前後で t 検定を行い，ユーザビリティの悪化が検知されないか．

Nielsen の原則 7,10 においてはプロンプトには含め

たものの，アプリとしての実装がないため評価対象外とする．

4.5 結果と考察

表 1 改修前の UI への指摘例

問題点	指摘回数
自動タグ付けやサンプルサイズ適用など，各種処理の開始，進行中，完了のフィードバック（ローディング表示，成功メッセージなど）が不足しており，ユーザーが処理状況を把握できず不安になる．	4/5
自動タグ付けのフローが直感的でなく，特に「タグ候補生成」と「自動タグ付け」の連携や，次のステップへの誘導が不明確．また，プロンプト編集画面のボタン文言「適用」も分かりにくい．	3/5

表 2 改修後の UI への指摘例

問題点	指摘回数
タグ付けや分析ファイルのプレビュー画面の読み込みが非常に長く，表示されない，または先に進めない．	2/5
ファイル処理が長時間完了せず，進捗状況も表示されない．	1/5

表 3 専門家役エージェントによる評価の平均点

原則	改修前	改修後
1	0.08	0.00
2	0.85	0.92
3	0.08	0.08
4	0.69	0.92
5	0.38	0.77
6	0.62	0.77
7	0.00	0.00
8	0.92	1.00
9	0.00	0.00
10	0.00	0.00

実験の結果，改修後の UI では利用者役エージェントによるネガティブフィードバックが減少し，専門家役エージェントのスコアも改善する傾向を確認した．観測された平均値はいずれも左片側検定の帰無仮説の棄却域に入らず，悪化は検知されなかった．

定性的な側面では，表 1 および表 2 の比較から明らかなように，一貫性のある文言や step-by-step の操作誘導といった，刷新時に意図された改善点がエージェントの評価に明確に反映されていた．

本手法における専門家役エージェントは，各評価指標に対して相互の影響を受けずに独立して評価できていると判断できる．改修において重点的に改修した「UI の一貫性 (原則 4)」や「ミスの防止 (原則 5)」のスコアが大きく改善した一方で，改修対象としなかった「進行状況表示 (原則 1)」「ユーザーの制御 (原則 3)」「エラー対応 (原則 9)」については，改修前後で一貫して低スコアを維持した．これは，LLM が各原則を独立した評価軸として正しく機能させ，客観的に UI の実装状態を評価できていることを示唆している．

一方で，行動ログから判明した「タグ未選択でのボタン押下」といった人間固有のミスは検知されなかった．これは，エージェントが目標に対して合理的な操作を選択する性質を持つため，「アプリの利用目的である分類のタグ設定を行わない」という非合理的な行動を想定できなかったものと推察される．本手法は UI の構造的欠陥の特定には極めて有効だが，こうした人間の認知的な隙に起因する問題を網羅するには，ログ分析等との併用が不可欠である．

5 まとめ

本研究では， LLM を用いたブラウザ操作エージェントによる業務システム UI の自動評価手法を提案し，その実効性を検証した．利用者役・専門家役という役割の異なる二種類のエージェントを組み合わせることで，利用者が直面するつまずきと，UI 設計における構造的なヒューリスティック違反の双方を検出できることを示した．ANA で実際に運用している ANAlyzer-Buddy での比較実験により，提案手法による指摘はドメイン知識を持つ社員による評価と整合する傾向が確認された．今回の検証では大規模な UI 改修を対象としたため， Nielsen の 10 原則の項目別のスコア変化と，具体的な改修内容との間の詳細な因果関係の特定には課題が残るものの，全体的なユーザビリティの客観的な評価に成功した．今後は評価の粒度をより微細化し，特定の UI 要素の変更が各評価指標に与える影響を検証することで， LLM による評価の信頼性と妥当性をさらに高めていく必要がある．また，エージェントによる合理的な操作では検知困難な，人間特有の認知的隙に起因するミスについても，意図的なペルソナ設定によってエージェントの合理性を排除できるかの実験や，操作ログ分析等と統合したハイブリッドな評価基盤を構築することで補完していく予定である．

謝辞

本研究の遂行にあたり，リードエンジニアの金城氏には，本研究に必要な各種調整において重要なご助力をいただいた．さらに，検証環境の構築にご尽力くださったMarcellin氏をはじめ，エンジニアチームの皆様には技術的なアドバイスを頂戴した．ここに深く感謝の意を表する．

参考文献

[1] Jakob Nielsen. 10 usability heuristics for user interface design. https://www.nngroup.com/articles/ten-usability-heuristics/, 1994. Accessed: 2025-12-09.

[2] Yuxuan Lu, Bingsheng Yao, Hansu Gu, Jing Huang, Jessie Wang, Yang Li, Jiri Gesi, Qi He, Toby Jia-Jun Li, and Dakuo Wang. Uxagent: A system for simulating usability testing of web design with llm agents. **arXiv:2504.09407**, 2025.

[3] Jakob Nielsen. Why you only need to test with 5 users. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/, 2000. Accessed: 2025-12-09.

[4] 永田靖. サンプルサイズの決め方. 朝倉書店, 2003.