WebArena Mod：WebArena ベンチマークの修正

張皓辰　榎本昌文　竹岡邦紘　小原涼馬　小山田昌史
NEC データサイエンスラボラトリー
{haochen-zhang,masafumi-enomoto,k_takeoka,ryoma-obara,oyamada}@nec.com

概要

WebArena は Web 操作エージェントの標準的なベンチマークとして広く利用されている．しかし，実行環境と評価系には実験の安定性や評価の妥当性を損なう課題が残されている．本稿では，WebArena を網羅的に調査し，全タスクの約 40%に影響する問題群を発見し，その要因を特定した：(1) インフラの動作不良，(2) 評価ロジックの不備，(3) タスク定義の曖昧さや矛盾，である．これらに包括的に対処する修正パッチ群「WebArena Mod」を提案し，75 件のタスクサブセットを用いた評価実験により，偽陰性率が 21.3%から 1.3%へと **20 ポイント低減する**ことを確認した．修正パッチ群およびプログラムは GitHub[1)] にて公開し，エージェントの能力を正確に測定するための信頼できる実験基盤を提供する．

1　はじめに

自律型 Web 操作エージェントは，大規模言語モデル（LLM）や視覚言語モデル（VLM）の実世界応用における重要な研究領域である．その能力を正確に評価するベンチマークの整備は，手法間の公正な比較と該当分野の発展に不可欠である．WebArena[1] は，EC・フォーラム・ソフトウェア開発・コンテンツ管理・地図という実サービスを模した Web サイト群と 812 件の人手検証済みタスクを提供し，自律型 Web エージェント評価の事実上の標準として広く採用されている [2, 3, 4, 5].

しかし，我々が WebArena の評価プロセスを網羅的に調査した結果，全タスクの約 **40%**において評価の妥当性が損なわれている可能性が明らかとなった．詳細な要因分析により，以下の 3 つの課題が特定された．

実行環境の不安定性　ベンチマークが提供する環境構築方法には，リソース管理と実験サイト設定の両面に不備がある．具体的には，メモリ不足による

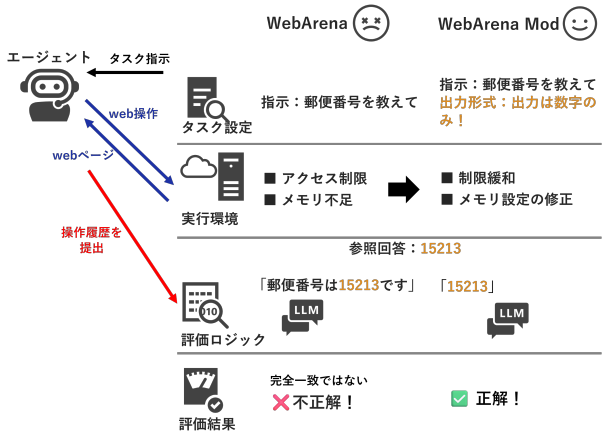

図 1　Overview of WebArena Mod

システムの再起動や，過度なアクセス制限による正常な操作の拒絶が発生している．これらはエージェントの能力とは無関係なタスク失敗の要因となっている．

評価ロジックの限界　評価プログラムの実装にはいくつかの不備がある．LLM ジャッジの判定基準が厳格すぎること，複数正答の処理の不備，URL 正規化の未実装などにより，意味的に正しい回答であっても形式的な差異のみで不当なペナルティが生じている．

曖昧・不正確なテストデータ　タスク設定には，指示と評価基準の矛盾，曖昧な記述，不正確な参照回答，フォーマット要件の未記載といった問題が含まれており，正しい回答が不当に棄却される原因となっている．さらに，一部のタスクは原理的に実行不可能である．

本稿では，これらの問題に包括的に対処する修正パッチ群 **WebArena Mod** を提案する (図 1)．WebArena が推奨するエージェントフレームワークである BrowserGym[6]，および WebArena の評価機能をパッケージ化した `libwebarena`[7] と互換性を維持しつつ，既存のコードを変更せずに適用できる．各課題に対する具体的な対処は以下の通りである：

- **環境の安定化**（3.1 節）：実験サイトに対してリ

1）https://github.com/nec-research-labs/WebArena-Mod

ソース設定の修正とアクセス制限の緩和により，環境起因のタスク失敗を防止する．

- **評価ロジックの改善**（3.2.1 節）：判定処理と正規化処理を改善し，意味的に正しい回答の誤判定を防止する．
- **テストデータの修正**（3.2.2 節）：指示と評価の不整合の解消，曖昧な記述の明確化，実行不可能なタスクの特定を行う．

提案手法の有効性を検証するため，修正対象となる 325 件のタスクから無作為に抽出した 75 件のサブセットを用いて評価実験を実施した．その結果，偽陰性率が **21.3%**から **1.3%**へと大幅に低減し，評価の信頼性が向上することを確認した．

本稿の貢献は以下の 3 点である：

1. WebArena ベンチマークを網羅的に調査し，評価の信頼性を損なう問題群を発見し，その要因を明らかにした．
2. 上記の問題に対処する修正パッチ群 WebArena Mod を開発し，オープンソースとして公開した．
3. 人手評価との比較実験により，提案手法が評価精度を向上させることを実験的に検証した．

2 背景

本節では，WebArena の構成を概説し（2.1 節），先行研究や実務で報告されている課題を整理する（2.2 節）．

2.1 WebArena の概要

WebArena[1] は，自律型 Web エージェント評価のための現実的かつ再現可能なベンチマークである．従来のベンチマークにおける合成タスク [8, 9]，制限された行動空間 [10]，単一の参照軌跡との一致に依拠した硬直的な評価手法 [11] といった課題に対処するため，完全に機能するセルフホスト型 Web サイトと機能的正確性に基づく評価を提供する．

環境構成　WebArena のデプロイメントは 2 層構成である．公式には AWS 上で AMI を利用する構築方法が提供されているが，コストや環境の柔軟性から，ローカルマシン上で Docker コンテナを用いる方法が広く採用されている．WebArena 公式が推奨する BrowserGym[6] も，この方法を提供する非公式の `webarena-setup` リポジトリ [12] を参照している．

ローカル層には，5 つの Web サイトを Docker コンテナで実行する：EC サイトとその管理画面，フォーラム（Postmill），ソフトウェア開発 (GitLab)，Wikipedia（Kiwix で提供），地図（OpenStreetMap）フロントエンドである．

クラウド層は，地図機能のためのバックエンドサービスをホストする．AWS EC2 インスタンス上で，タイルサーバー（地図画像レンダリング），Nominatim（ジオコーディング），OSRM（ルーティングエンジン）をホストする．

評価パイプライン　ベンチマークは 812 件のタスクを提供し，各タスクには自然言語による意図記述，開始 URL，参照回答が含まれる．評価方式には，文字列の完全一致・部分一致，URL 比較，HTML コンテンツ検証，LLM ジャッジによる意味的判定がある．エージェントフレームワークである BrowserGym[6] は，WebArena の評価プログラムとタスクデータを Python パッケージ化した `libwebarena`[7] を通じてこの評価機能を利用している．

2.2 報告されている課題

WebArena は広く採用されている一方で，その運用と評価に関する課題が複数報告されている．環境面では，地図バックエンドの障害やフォーラムのレート制限，セルフホスト環境の構築困難が指摘されており，一部の研究では地図関連タスクを評価から除外している [13, 14, 15, 16, 17]．評価面では，厳格な文字列照合による形式的差異へのペナルティ，曖昧なタスク定義，LLM ジャッジの判定誤りなどが報告されている [18, 19, 13, 14]．

これらの課題に対し，評価系を再設計する WebArena Verified[20] が並行して開発されている．しかし，既存の報告は個別の問題指摘にとどまり，問題の体系的な分類と包括的な対処は行われていなかった．本稿では，環境・評価ロジック・テストデータの 3 側面から WebArena の問題を体系的に分類し，既存フレームワークとの互換性を維持しつつ包括的に対処する修正パッチ群を提案する．

3 修正内容

本節では，WebArena Mod の修正内容を説明する．3.1 節で環境の安定化，3.2 節で評価とテストデータの修正について述べる．

3.1 環境の安定化

3.1.1 OpenStreetMap バックエンド

地図機能を担うクラウド層（2.1 節）について，WebArena が当初提供していた公開サーバーは頻繁に停止している．これに対し WebArena チームは AWS 上でのセルフホスト方法を提供しているが，このセルフホスト環境もメモリ不足や設定不備により不安定である．

公式のセルフホスト用スクリプトは `t3a.xlarge` インスタンス（16 GB RAM）を想定しているが，OSRM が各 3.5–4.5 GB × 3 コンテナ，Nominatim が最大 6 GB，タイルサーバーがピーク時 4 GB を必要とし，メモリ枯渇が発生する．また，タイルサーバーの `-shm-size` 未指定により間欠的な 404 エラーも発生する．

これらの問題により，高負荷時にコンテナが再起動し経路検索が応答しなくなる，地図画像が描画されないといった障害が発生し，エージェントの能力とは無関係にタスクが失敗する．これに対し，適切なメモリ制限とコンテナ設定を施した `t3a.2xlarge` 向け修正版スクリプトを提供する．

3.1.2 ローカル WebArena スタック

ローカル層の構築には，BrowserGym が参照する `webarena-setup` リポジトリ [12] のスクリプトが広く利用されている．しかし，(1) フォーラムのレート制限およびローカル URL ブロッキングにより正当な操作が 429/500 エラーで失敗する，(2) PHP ベースサービスのデフォルト設定が並行リクエストに不十分でタイムアウトが発生する，(3) パッチスクリプトがデータベース初期化前にコンテナを変更し SQL エラーを引き起こす，といった問題があった．

その結果，正常な操作が拒否されたり，並行実行時にサービスが応答しなくなったりといった障害が発生する．これに対し，レート制限の緩和とローカル URL ブロッキングの無効化，最適化された PHP OPcache 設定と明示的な共有メモリ割り当て，データベースサービスの準備完了チェックを実装した．

3.2 評価とテストデータの修正

3.2.1 評価ロジック

`libwebarena` の評価コードには以下の問題があった：(1) LLM ジャッジが「意味的等価性」のみを判定基準としており，正しい情報を含むが追加の文脈を伴う回答が不正解となることがある（付録 A 参照），(2) 文字列・URL 評価における複数正答の処理に不備があり，有効な回答が棄却される，(3) 参照回答の URL（`www.reddit.com` 等）がローカル環境の URL（`localhost:9999` 等）と一致せず，正しい回答が棄却される．

これらに対し，2 段階 LLM ジャッジ（意味的等価性チェック＋ファジー包含チェック）を実装し，いずれかで合格すれば正解とする．また，複数正答を正しく処理するよう修正と，参照回答の URL をローカル環境のホスト名に正規化する処理を実装した．

3.2.2 タスクデータ

812 件のタスク設定を検証し，以下の問題を特定した：(1) 評価が厳密な文字列照合を用いるにもかかわらず，タスク指示に出力形式の要件が記載されていない（例：参照回答「15213」に対し「郵便番号は 15213 です」は棄却される），(2) 指示と評価基準の矛盾（例：「価格を上げる」指示で値下げをチェック），(3) 参照回答がベンチマーク環境の実データと不一致，，(4) 曖昧な指示例：「Nic が 4 月にしたコミット数は？」に対し，Nic を含むユーザ名が複数存在），(5) Chromium の特定クッキー存在時にフォーラムログインが失敗する認証バグ，(6) 実行不可能なタスク．

これらに対し，フォーマット要件の明示，指示と参照の整合，参照回答の更新，曖昧な指示の明確化，認証順序の修正を行い，実行不可能なタスク（ID: 2, 5, 184, 425；詳細は付録 B 参照）を特定した．すべての変更はタスク設定ファイルを置換で修正する単一パッチスクリプトとして提供する．

4 実験

WebArena Mod による評価精度の改善を検証するため，比較実験を実施した．

表 1　評価パイプラインの判定精度の比較（$n = 75$）．エージェント出力に対する人手評価を正解とし，WebArena Mod 適用前後の評価パイプラインによる判定を分類した（TP: 正しく成功と判定，FN: 成功を失敗と誤判定，TN: 正しく失敗と判定，FP: 失敗を成功と誤判定）．

構成	TP	FN	TN	FP
適用前	23	16	36	0
適用後	40	1	34	0

4.1　実験設定

タスクサンプリング　WebArena Mod の修正対象となるタスク群から 75 件を無作為に抽出した．

エージェント設定　GPT-5.2[2] を基盤とした AgentLab[6, 21] の GenericAgent[3] を使用し，`use_screenshot` を有効化した．

比較条件　同一のエージェントを以下の 2 条件で実行した：

- **オリジナル**：公式スクリプトで構築した環境，`libwebarena` の評価プログラムおよびタスクデータ
- **WebArena Mod**：修正スクリプトで構築した環境，パッチ適用済み評価プログラムおよびタスクデータ

評価指標　各エージェント出力に対し，2 名のアノテータががタスク指示と行動軌跡に基づき成否を独立に判定し，不一致があった場合は議論を経て最終判定を決定した．評価パイプラインの判定を人手評価と比較し，偽陽性（FP），偽陰性（FN），真陽性（TP），真陰性（TN）を算出した．

実装　BrowserGym（コミット `254938b`）および AgentLab（コミット `519abed`）を使用した．

4.2　結果

表 1 に評価精度の比較結果を示す．

WebArena Mod により，偽陰性率は 21.3%（16/75）から 1.3%（1/75）へと **20 ポイント改善**した．一方，偽陽性は両条件で 0 件であり，不正解を誤って正解と判定するケースは増加しなかった．なお，2 条件間で実際の成功数に差異があるが（エージェントの非決定性による），本実験の主眼は評価パイプラインの判定精度である．偽陰性の大幅な削減は，WebArena Mod が評価の信頼性を向上させることを示している．

4.3　実験分析

オリジナル環境で発生した 16 件の偽陰性のうち，WebArena Mod により 15 件が解消された．内訳は，タスクデータの修正（§ 3.2.2）によるものが 9 件，評価ロジックの修正（§ 3.2.1）によるものが 6 件であった．タスクデータでは参照回答の誤りや出力形式要件の欠如が主な原因であり，評価ロジックでは URL 正規化と 2 段階 LLM ジャッジが効果を示した．

なお，本実験は逐次実行であり負荷が低いため，環境の安定化（§ 3.1）に関する問題は顕在化しなかった．並行実行を行う大規模実験では，環境修正の効果も期待される．

以下に，解消された偽陰性の具体例を示す．

タスク 181（出力形式要件の欠如）　「issue がクローズ済みか確認」に対し，エージェントは「Its status badge shows Open, so it is not closed.」と正しく回答した．しかし参照回答「No」との文字列照合により不正解と判定された．指示へのフォーマット要件の明記により修正後の環境においては正しく評価された．

タスク 294（ローカル URL 正規化の不備）　「SSH で clone するコマンドを表示」に対し，エージェントは正しいコマンドを出力した．しかし参照回答の外部ホスト名とローカル環境のホスト名が一致せず不正解と判定された．URL 正規化により修正後の環境においては正しく評価された．

5　おわりに

本稿で提案した WebArena Mod は，WebArena ベンチマークの評価信頼性を損なう問題群を網羅的に調査し，全タスクの約 40%に影響する課題を特定した．これらに対処する修正パッチ群を開発し，環境の安定化，評価ロジックの改善，テストデータの修正を行った．75 件のタスクサブセットを用いた評価実験により，偽陰性率が 21.3%から 1.3%へと 20 ポイント低減することを確認した．

WebArena Mod は既存の BrowserGym および `libwebarena` と互換性を維持しており，GitHub にて公開している．Web エージェント研究における信頼性の高い評価基盤として活用されることを期待する．

2）　gpt-5.2-2025-12-11

3）　https://huggingface.co/spaces/ServiceNow/browsergym-leaderboard/blob/main/results/GenericAgent-GPT-5/README.md

参考文献

[1] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In **ICLR**, 2024.

[2] Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. VisualWebArena: Evaluating multimodal agents on realistic visual web tasks. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 881–905, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[3] Jiayuan Wang, Zhiruo anf Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. **arXiv preprint arXiv:2409.07429**, 2024.

[4] Jing Yu Koh, Stephen McAleer, Daniel Fried, and Ruslan Salakhutdinov. Tree search for language model agents. **arXiv preprint arXiv:2407.01476**, 2024.

[5] Ido Levy, Ben Wiesel, Sami Marreed, Alon Oved, Avi Yaeli, and Segev Shlomov. St-webagentbench: A benchmark for evaluating safety and trustworthiness in web agents. **arXiv preprint arXiv:2410.06703**, 2024.

[6] Gasse Chezelles, et al. The browsergym ecosystem for web agent research. **Transactions on Machine Learning Research**, 2025. Expert Certification.

[7] Maxime Gasse. libwebarena: A library to adapt webarena to browsergym. https://pypi.org/project/libwebarena/, 2024. Python Package Index.

[8] Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, and Percy Liang. World of bits: An open-domain platform for web-based agents. In **ICML**, 2017.

[9] Evan Zheran Liu, Kelvin Guu, Panupong Pasupat, Tianlin Shi, and Percy Liang. Reinforcement learning on web interfaces using workflow-guided exploration. In **ICLR**, 2018.

[10] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In **NeurIPS**, 2022.

[11] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. In **NeurIPS**, 2023.

[12] Maxime Gasse. webarena-setup: Docker scripts for webarena deployment. https://github.com/gasse/webarena-setup.

[13] Atsuyuki Miyai, Zaiying Zhao, Kazuki Egashira, Atsuki Sato, Tatsumi Sunada, Shota Onohara, Hiromasa Yamanishi, Mashiro Toyooka, Kunato Nishina, Ryoma Maeda, Kiyoharu Aizawa, and Toshihiko Yamasaki. Webchorearena: Evaluating web browsing agents on realistic tedious web tasks. **arXiv preprint arXiv:2506.01952**, 2025.

[14] Jace AI. Awa 1.5 achieves breakthrough performance on webarena benchmark. https://jace.ai/blog/awa-1-5-achieves-breakthrough-performance-on-web-arena-benchmark, 2024.

[15] Su Kara, Fazle Faisal, and Suman Nath. Warex: Web agent reliability evaluation on existing benchmarks. **arXiv preprint arXiv:2510.03285**, 2025.

[16] WebArena Community. Issue #66: Environment is very slow due to playwright and inspect.stack(). GitHub Issue, 2023.

[17] WebArena Community. Issue #168: Environment issue (gitlab 502/500 errors). GitHub Issue, 2024.

[18] Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun, Andy Zhang, Shu Liu, Sasha Cui, Sayash Kapoor, Shayne Longpre, Kevin Meng, Rebecca Weiss, Fazl Barez, Rahul Gupta, Jwala Dhamala, Jacob Merizian, Mario Giulianelli, Harry Coppock, Cozmin Ududec, Antony Kellermann, Jasjeet Sekhon, Jacob Steinhardt, Sarah Schwettmann, Matei Zaharia, Ion Stoica, Percy Liang, and Daniel Kang. Establishing best practices for building rigorous agentic benchmarks. **arXiv preprint arXiv:2507.02825**, 2025.

[19] Invariant Labs. What we've learned from analyzing hundreds of ai web agent traces. https://invariantlabs.ai/blog/what-we-learned-from-analyzing-web-agents, 2024.

[20] Amine El hattami, Megh Thakkar, Nicolas Chapados, and Christopher Pal. Webarena verified: Reliable evaluation for web agents. In **Workshop on Scaling Environments for Agents**, 2025.

[21] Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. WorkArena: How capable are web agents at solving common knowledge work tasks? In Kolter Salakhutdinov, et al., editors, **Proceedings of the 41st International Conference on Machine Learning**, Vol. 235 of **Proceedings of Machine Learning Research**, pp. 11642–11662. PMLR, 21–27 Jul 2024.

A　LLM ジャッジの詳細

WebArena は，LLM ジャッジによる意味的判定において「意味的等価性（semantically equivalent）」のみを判定基準としている．このため，正しい情報を含むが追加の文脈を伴う回答が不正解と判定される場合がある．

誤判定の例　以下に具体例を示す．

- **タスク**：“Tell me the full address of all international airports that are within a driving distance of 5 km to Carnegie Mellon University”
- **参照回答**：“There is no airport within 5 km of Carnegie Mellon University”
- **エージェント出力**：“There is no international airport in that range. The closest international airport is 25 km away.”
- **判定結果**：“Partially correct. The reference says ‘no airport’ while the student says ‘no international airport’ — a narrower claim. Also ‘in that range’ is vaguer than ‘within 5 km of Carnegie Mellon University.’”

この出力は質問に対する正しい回答を含むが，表現の差異により「部分的に正解」と判定され，最終的に不正解として扱われる．

修正内容　WebArena Mod では，2 段階の LLM ジャッジを実装した．第 1 段階では従来通り意味的等価性を判定し，第 2 段階では回答が参照回答の重要情報を「含むか」を判定する．いずれかの段階で正解と判定されれば，最終判定を正解とする．

B　実行不可能なタスク

本節では，原理的に実行不可能と判断した 4 件のタスクの詳細を示す．

曖昧な概念定義（タスク 2, 5）　両タスクは「最も売れた製品タイプ」を問うが，ベンチマーク環境のショッピング管理画面において「製品タイプ（product type）」の定義が明示されていない．カテゴリ，属性セット，商品分類など複数の解釈が可能であり，エージェントの回答が参照回答と一致することを期待するのは不合理である．

- **タスク 2**：“What is the top-1 best-selling product type in Quarter 1 2022”（参照回答：“Yoga ball”）
- **タスク 5**：“What is the top-1 best-selling product type in Jan 2023”（参照回答：“Yoga ball”）

評価基準とタスク範囲の不整合（タスク 184）　在庫が 0 の商品名を列挙するタスクであるが，該当商品は数百件存在する．一方，評価は単一の参照回答（“Sinbad Fitness Tank”）との厳密な文字列照合を用いており，他の正しい回答はすべて棄却される．

データ範囲の制約（タスク 425）　“Find the page of the longest bridge in the Western hemisphere on the map” というタスクであるが，参照回答（“Mackinac Bridge”）の所在地域はベンチマークの地図データダンプに含まれていない．したがって，該当ページへのアクセスは原理的に不可能である．

これらのタスクは，環境の大幅な変更またはタスク設計の見直しなしには達成できないため，評価から除外することを推奨する．