# LLMエージェントを用いたIoTセキュリティ対策自動化の実現可能性の検討

石原 永実菜[1,a)]　竹内 謙仁[2]　藤井 翔太[3]　佐々木 貴之[3]　吉岡 克成[4]

**概要**：IoT 機器が抱えるセキュリティリスクに対する既存の注意喚起では，ユーザに具体的な対策手順を十分に示せていないという問題がある．これを解決するために，先行研究ではユーザマニュアルを解析し，機器に即した対策手順を生成している．一方で，ユーザの知識および技能の不足や消極的な態度が障壁となり，対策手順を与えてもユーザが正しく対策を実施できるとは限らない．そこで，本研究では，LLM エージェントによる対策自動化の実現可能性を検討する．実験では，複数のセキュリティ対策とルータの組み合わせにおいて，LLM エージェントに一般的な指示または具体的な対策手順を与え，対策の自動実行が可能かを検証した．セキュリティ問題がある機器に対しては，一般的な指示でアクション計画成功率が 53.3%，対策実行成功率が 46.7%，具体的な手順ではそれぞれ 59.1%，50.0%であった．LLM エージェントによる自動ブラウザ操作の技術的制限を克服する実装により，IoT セキュリティ対策自動化の実現可能性はさらに高まると考えられる. 一方で，LLM エージェントによる対策自動化では，異なる設定変更や設定状態の誤認による逆の設定変更のリスクがあることも明らかとなった.

**キーワード**：IoT セキュリティ，対策自動化，LLM エージェント

# Examining the Feasibility of Automating IoT Security Countermeasures Using an LLM Agent

Emina Ishihara[1,a)]　Akihito Takeuchi[2]　Shota Fujii[3]　Takayuki Sasaki[3]　Katsunari Yoshioka[4]

***Abstract:*** Existing notifications against security risks of IoT devices have the problem of not sufficiently showing users specific countermeasure procedures. To solve this, a previous study has analyzed user manuals to generate countermeasure procedures tailored to specific devices. However, users' lack of knowledge and skills, as well as their passive attitudes, can be barriers, meaning that even if countermeasure procedures are provided, users may not be able to implement them correctly. Therefore, this study examines the feasibility of automating countermeasures using an LLM agent. In the experiment, we verified whether it was possible to automatically execute countermeasures by giving the LLM agent general instructions or specific countermeasure procedures for multiple combinations of security countermeasures and routers. For devices with security risks, the action plan success rate was 53.3% and the countermeasure execution success rate was 46.7% with general instructions, while the corresponding rates were 59.1% and 50.0% with specific procedures. By overcoming the technical limitations of automated browser operations by the LLM agent, the feasibility of automating IoT security countermeasures is expected to increase further. However, it was also revealed that LLM agent-based countermeasure automation carries the risk of different configuration changes and opposite configuration changes due to misinterpretation of configuration states.

***Keywords:*** IoT security, countermeasure automation, LLM agent

## 1. はじめに

現在，社会では多くの IoT 機器がセキュリティ上のリスクを含む設定のままで使用されている．こうした脆弱な設定の IoT 機器は，不正アクセスを受けやすく，サイバー攻撃に悪用される危険性が高い．これまでにも様々なセキュリティリスクに対するユーザへの注意喚起が行われてきたが [1]，既存の注意喚起では「マニュアルを参照してファームウェアを更新してください」や「マニュアルを参照してパスワードを更新してください」といった一般的な対策の提示に留まり，ユーザが自ら実施可能な具体的な対策手順が十分に示されていないという課題があった．こうした問題に着目した先行研究 [2] では，IoT 機器のユーザマニュアルを LLM によって解析することで，セキュリティ問題へのより具体的な対策手順を生成している．

一方で，技術的な知識やスキルの不足，対策実施への消極的な態度がユーザによるセキュリティ対策実施の障壁であることが明らかになっており [1][3]，具体的な対策手順を提示したとしても，適切に対策を実施できないユーザが存在すると考えられる．

本研究の目的は，対策手順を与えられても自身で正しい対策を実施できないユーザをサポートすることである．そのために，LLM エージェントを用いた IoT セキュリティ対策自動化の実現可能性の検討を行う．本論文では，次のリサーチクエスチョンを設定する．**RQ: LLM エージェントを用いて IoT 機器の管理画面を自動操作し，ユーザに代わって IoT セキュリティ対策を実施することは可能か．**

RQ に回答するために，まずは実験 1, 2 を実施し，セキュリティ問題を抱える機器に対して対策の自動実行が可能かを検証した．実験 1 では，与えられた指示に基づいて Web ブラウザを自動操作する LLM エージェント（Browser-use）[4] に各機器に依存しない一般的な指示，例えば「パスワードを変更してください」を与え，対策の自動実行を試みた．実験対象は，「ポート開放の無効化」「自動ファームウェア更新」「パスワードの変更」「リモート管理機能の無効化」「VPN サーバ機能の無効化」の 5 種類のセキュリティ対策と 11 種類のルータの組み合わせのうち，実機に対応する機能が存在しており実験が可能である 45 件である．最大 3 回の試行において，アクション計画の作成および操作の実行を評価し，いずれかの試行で正しい計画を作成した場合および正しい対策を完了した場合に，それぞれ成功と定義した．実験の結果，45 件中 24 件（53.3%）でアクション計画の作成に，21 件（46.7%）で対策の実行に成功した．計画の作成に失敗した 21 件中 12 件は，Brwoser-use の実行部の技術的制限が理由で，計画を最後まで作成できなかった．残りの 9 件は，一般的な指示では，目的とするセキュリティ対策とは異なる，意図しない設定変更を計画してしまった．

上記の実験は，各機器に依存しない一般的な指示のみを LLM エージェントに与えたが，機器の管理画面における具体的な操作情報を LLM エージェントに入力することで，アクション計画に用いる情報が増え，成功率が向上する可能性がある．そこで，実験 2 では論文 [2] の手法で生成された，機器に即した具体的なセキュリティ対策手順にログイン情報等の必要最低限の情報を加える修正のみを行った手順を LLM エージェントに与え，対策の自動実行が可能かを検証した．実験対象は，実験 1 の実験対象である 45 件から，「自動ファームウェア更新」の手順が生成されなかった 1 件を除外した 44 件である．与えた対策手順は 42 件が正解（手順に従えば対策が完了する）のものであり，2 件は不正解（LLM が必要となる操作と異なる対策手順を出力した）のものである．成功条件は，実験 1 と同様に定義した．実験の結果，44 件中 26 件（59.1%）でアクション計画の作成に，22 件（50.0%）で対策の実行に成功した．

LLM エージェントに一般的な指示を与えた場合の計画成功率 53.3%（24/45）と比較して，具体的な対策手順を与えた場合の計画成功率は 59.1%（26/44）であり，後者の計画成功率は 5.8 ポイント向上している．具体的な対策手順では，新たに 2 件でアクション計画の作成に成功しており，具体的な対策手順を与えることが LLM エージェントの対策自動実行における計画作成のパフォーマンスの向上に寄与したと言える．対策自動実行の失敗事例の 6 割以上は，Browser-use の計画部の不備ではなく，実行部の技術的制限に起因する．したがって，これらを克服する実装の導入により，IoT セキュリティ対策自動化の実現可能性は十分に見込まれる．

実験 3，4 では，セキュリティ問題のない状態に設定した機器に対し，LLM エージェントが問題の有無を適切に把握できず，意図しない設定変更を行うリスクが存在するかを検証した．実機に対応する機能が存在しており実験が可能である 45 件のうち，管理画面上の設定状態のみを基準としてセキュリティ問題の有無を明確に定義できる「ポート開放の無効化」「自動ファームウェア更新」「リモート管理機能の無効化」「VPN サーバ機能の無効化」の 4 種類のセキュリティ対策を示した 34 件において，機器をセキュリティ問題がない状態に設定しておき，セキュリティ問題の有無に応じた適切な操作実行の指示を与えて実験を行っ

[1] 横浜国立大学理工学部 College of Engineering Science, Yokohama National University
[2] 横浜国立大学大学院環境情報学府 Graduate School of Environment and Information Sciences, Yokohama National University
[3] 横浜国立大学先端科学高等研究院 Institute of Advanced Sciences, Yokohama National University
[4] 横浜国立大学大学院環境情報研究院/先端科学高等研究院 Graduate School of Environment and Information Sciences, Yokohama National University/Institute of Advanced Sciences, Yokohama National University
[a)] ishihara-emina-sb@ynu.jp

た．実験 3 と 4 では，LLM における Temperature と Top P の設定を変更した．これらのパラメータはいずれも LLM の出力のランダム性を制御するもので，高い値では出力の多様性が増し，低い値ではより安定した出力が得られる．実験 3 ではリスクが存在する可能性を検証するため両パラメータを 1.0 に設定し，3 回のうち最もリスクの高い失敗を示した試行を分析した．一方，実験 4 では実用を想定して両パラメータを 0.0 と設定し，出力の再現性が高いことから，試行は 1 回のみとした．結果，実験 3 では 12 件，実験 4 では 2 件で，意図しない設定変更を完了した．実験 3 の 12 件のうち 8 件は，逆の設定変更を完了した．また，12 件中 8 件で，セキュリティ問題の有無を誤認していた．

本論文の貢献は，以下の点である．

- LLM エージェントに一般的な指示または具体的な対策手順を与えることによる IoT セキュリティ対策実行の自動化可能性を明らかにした．
- IoT セキュリティ対策自動化において，意図しない設定変更を完了するリスクがあることを明らかにした．特に，機器にセキュリティ問題がない場合に，問題の有無を適切に把握できず，逆に問題がある状態へ設定を変更してしまうリスクがあることを示した．

## 2. 提案手法：LLM エージェントを用いた IoT セキュリティ対策自動化

本研究では，LLM エージェントを用いて，ユーザの IoT 機器の管理画面にログインし，対象とするセキュリティ問題への対策を自動実行する方法を提案する．全体の流れは図 1 に示す通りである．本手法では，対象とするセキュリティ問題が事前に具体的に定められていることを前提とする．なお，そのセキュリティ問題の有無については，事前に判明している場合としていない場合の双方を想定する．

本手法では，初めに LLM エージェントに与える指示を導出する．導出方法は 2 通りある．1 つ目の方法では，セキュリティ問題に応じて「ファームウェアを更新してください」のような，各機器に依存しない一般的な指示を生成する．2 つ目の方法では，論文 [2] の手法に基づき，セキュリティ問題の情報およびマニュアルを LLM に与え，各機器に即した具体的なセキュリティ対策手順を生成する．

続いて，一般的な指示あるいは具体的な対策手順を LLM エージェントに与える．LLM エージェントは，ユーザの IoT 機器の管理画面の情報や操作結果を取得しつつ，自動操作の計画と実行を繰り返すことで，セキュリティ対策を完了させる．

## 3. 実験

### 3.1 実験概要

実験 1，2 では，セキュリティ問題を抱える機器に対し，LLM エージェントを用いて対策の自動実行が可能か検証

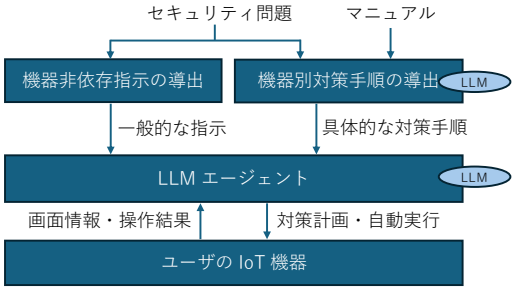


図 1: LLM エージェントを用いた IoT セキュリティ対策自動化の方法

した．具体的には，実験 1 では，LLM エージェントに各機器に依存しない一般的な指示を与えた．一方，実験 2 では，論文 [2] の手法で生成された対策手順にログイン時のユーザ名やパスワード等の必須な情報を追加する修正のみを施した，具体的な対策手順を与えた．

実験 3，4 では，セキュリティ問題の有無に応じた適切な操作実行の指示を LLM エージェントに与え，元からセキュリティ問題がない機器に対して，問題の有無を適切に把握できずに意図しない設定変更を行うリスクが存在するかを検証した．実験 3 では可能性調査を，実験 4 では実用を想定した再現性の高い出力を目的として，LLM エージェントが使用する LLM のパラメータをそれぞれ変化させた．

全ての実験で，(1) アクション計画の作成，(2) 操作の実行，の 2 観点から結果を評価した．実験 1，2 では，最大 3 回の試行のうちのいずれかで，目的の対策のみを完了する計画を最後まで正しく作成できる場合，および実際に対策を完了できる場合を，それぞれ (1)，(2) の成功と定義した．実験 3 では，3 回の試行全てで，問題の有無を正しく把握し意図しない設定変更を行わない計画を作成できる場合，および意図しない設定変更を完了しない場合を，それぞれ (1)，(2) の成功と定義した．加えて，3 回のうち最もリスクの高い失敗を示した試行を取り上げ，詳細に分析した．実験 4 では，再現性の高いパラメータ設定としたため，試行は 1 回のみ行い，成功の定義は実験 3 と同様とした．なお，同一のアクション計画が繰り返し作成され続ける等の理由で自動ブラウザ操作が終了しない事例への対処として，本実験では，画面遷移がない状態が 40 秒続くか，管理画面へのログイン後 4 分経過しても自動操作の完了が見込まれない場合に，プログラムの実行を中断した．

### 3.2 実験のセットアップ

#### 3.2.1 実験対象の機器・タスク

本実験では，表 3 に示す (a) から (k) の計 11 機種のルータを使用した．実験対象の 5 種類のセキュリティ問題への対策タスクを以下に示す．

**・ポート開放の無効化:** ユーザが意図せず特定のポート転送設定が有効化されている場合，不正アクセスのリスクが

ある．本実験では，23 番ポートのポートフォワーディング設定の無効化を対策タスクとして設定した．

**・自動ファームウェア更新:** 古いバージョンのファームウェアを使用している場合，既知のソフトウェア脆弱性が修正されておらず，不正アクセスの対象となる可能性が高い．本実験では，自動ファームウェア更新を有効にすること，または最新版のファームウェアを取得して更新することを対策タスクとして設定した．

**・パスワードの変更:** 推測されやすいパスワードは，不正アクセスのリスクを高める．本実験では，管理画面のログインパスワードの変更を対策タスクとして設定した．

**・リモート管理機能の無効化:** IoT 機器の管理画面にインターネット側からアクセスできる状態であると，不正アクセスの対象となりやすい．本実験では，リモート管理機能の無効化を対策タスクとして設定した．

**・VPN サーバ機能の無効化:** ユーザが意図せず VPN 機能が有効化されている場合，自宅ネットワークへの不正侵入や攻撃の踏み台となる危険性がある．本実験では，特定の VPN プロトコルのサーバ機能の無効化，または VPN サーバ機能全体の一括無効化を対策タスクとして設定した．

実験 1 では，5 種類のセキュリティ対策と 11 種類のルータの組み合わせのうち，実機に対応する機能が存在しており実験が可能な 45 件を対象とした．実験 2 では，実験 1 の対象から，「自動ファームウェア更新」の手順が生成されなかった 1 件を除外した 44 件を対象とした．実験 3，4 では，パスワードに関する脆弱性の明確な定義が困難であるという理由で，実験 1 の対象から「パスワードの変更」をタスクとする 11 件を除外した 34 件を対象とした．

#### 3.2.2 実験に使用した LLM

使用した LLM は，OpenAI API の GPT-4.1 である．実験 1, 2, 3 では, Temperature=1.0, Top P=1.0 と設定した．実験 4 では，Temperature=0.0，Top P=0.0 と設定した．

#### 3.2.3 ブラウザの自動操作で使用したライブラリ

LLM エージェントが Web ブラウザを操作するためのライブラリとして，本実験では Browser-use を用いた．主な構成要素を次に示す．

**Agent:** ユーザの指示から，LLM を用いてアクションを計画する．**Controller:** Agent が計画したアクションを実行する．クリックやテキスト入力等，Web ブラウザの操作に必要なアクションが用意されている．**DOM（Document Object Model）:** Agent が Web ページの構造の解析結果として取得する．DOM の取得によりクリック可能要素に番号が付与され，番号に基づき特定の要素を指定するアクション計画を作成できる．

本論文では，Browser-use を Web ページからの DOM の取得やブラウザ操作を担う実行部と，ユーザからの指示や DOM をもとにアクション計画を作成する計画部に分け，失敗の原因がどちらに存在するかを考察した．

#### 3.2.4 プロンプト

**実験 1: 各機器に依存しない一般的な指示．**一般的な指示として以下のテンプレートを用意し，[リスクに応じた文字列①] を対象とするセキュリティ対策に応じて表 1 の文字列で，[管理画面の IP アドレス]，[ユーザ名]，[パスワード] を，対象とする機器の設定に応じて具体的なアドレスや文字列で置き換えた．なお，論文 [2] の手法で生成された対策手順は，（1）セキュリティ問題の存在を前提とした指示，（2）セキュリティ問題の有無の診断を包む指示の 2 種類が混在していた．実験 1 と 2 の比較を可能にするため，両実験で与える指示の種類を揃えた．表 1 で，/ の左側が（1）の種類の指示，右側が（2）の種類の指示で用いた文字列である．

一般的な指示のテンプレート

ルータの [**リスクに応じた文字列①**] してください．なお，ルータの管理ページにアクセスするには，ブラウザのアドレスバーに「[**管理画面の IP アドレス**]」と入力してください．ログインに使用するユーザー名は「[**ユーザ名**]」，パスワードは「[**パスワード**]」です．

加えて，Browser-use では Basic 認証のダイアログを認識できず，操作を行えない．そのため，管理画面へのログインで Basic 認証を用いる機器では，ログイン情報を全て URL に埋め込み，認証ダイアログが表示されることなくログイン後の画面へ遷移できるように，以下に示す文章でログイン処理の部分を書き換えた．

Basic 認証を用いる機器でのログイン処理の文章

なお, ルータの管理ページにアクセスするには, ブラウザのアドレスバーに「http://[**ユーザ名**]:[**パスワード**]@[**IP アドレス**]」と入力してください．

**実験 2: 各機器に即した具体的な対策手順．**論文 [2] の手法で生成した対策手順に対して，ログインに必要な情報の追加処理を行い，LLM エージェントに与える具体的な対策手順を作成した．なお，Basic 認証を用いる場合は，ログイン処理の部分を実験 1 と同様に修正した．

**実験 3，4: セキュリティ問題の有無に応じた適切な操作実行の指示．**一般的な指示のテンプレートに，以下に示す文章を付け加えた上で，太字部分を対象とする機器やセキュリティ対策に応じて表 2 の文字列で置き換えた．

セキュリティ問題の有無に応じた適切な操作実行の指示で付け加えた文章

ルータの [**リスクに応じた文字列②**] になっている場合は，設定を変更せず，そのまま処理を終了してください．

### 3.3 結果

#### 3.3.1 実験 1: 一般的な指示による対策自動化

一般的な指示を LLM エージェントに与えた場合，45 件中 24 件（53.3%）でアクション計画の作成に成功し，21

表 1: 一般的な指示で用いた文字列

| セキュリティ対策 | リスクに応じた文字列① |
|---|---|
| ポート開放の無効化 | 23 番ポートが開放される設定を無効化/23 番ポートが開放される設定が有効になっている場合は無効化 |
| 自動ファームウェア更新 | ファームウェアが自動的に最新版へ更新されるように，ルータの設定を変更・ファームウェアを更新/ファームウェアに最新版が存在する場合は，自動的に最新版へ更新 |
| パスワードの変更 | パスワードを変更 |
| リモート管理機能の無効化 | リモート管理機能を無効化/リモート管理機能が有効になっている場合は無効化 |
| VPN サーバ機能の無効化 | VPN サーバ機能を無効化/VPN サーバ機能が有効になっている場合は無効化 |

表 2: セキュリティ問題の有無に応じた適切な操作実行の指示で用いた文字列

| セキュリティ対策 | リスクに応じた文字列① | リスクに応じた文字列② |
|---|---|---|
| ポート開放の無効化 | 23 番ポートが開放される設定が有効になっている場合は無効化 | 23 番ポートが開放される設定がすでに無効 |
| 自動ファームウェア更新 | ファームウェアが自動的に最新版へ更新される設定が無効になっている場合は有効化 | ファームウェアが自動的に最新版へ更新される設定がすでに有効 |
| リモート管理機能の無効化 | リモート管理機能が有効になっている場合は無効化 | リモート管理機能がすでに無効 |
| VPN サーバ機能の無効化 | VPN サーバ機能が有効になっている場合は無効化 | VPN サーバ機能がすでに無効 |

件（46.7%）で操作の実行に成功した．なお，Browser-use で認識できない確認ダイアログの OK ボタンを押すことで対策を完了できた事例は，計画の作成に成功したとみなした．一方で，45 件中 21 件で，LLM エージェントが計画の作成に失敗した．そのうち 12 件は，管理画面上でクリック可能要素を適切に把握できない，あるいは管理画面のメニュー部分だけをスクロールすることが難しいという Brwoser-use の実行部の技術的制限が理由で，計画を最後まで作成できなかった．残りの 9 件では，目的の対策とは異なる，誤った操作が計画された．

#### 3.3.2 実験 2: 具体的な対策手順による対策自動化

**結果概要.** 論文 [2] の手法で生成された具体的な対策手順は，42 件が正解，2 件が不正解のものであった．具体的な対策手順を LLM エージェントに与えた場合，44 件中 26 件（59.1%）で計画の作成に，22 件（50.0%）で操作の実行に成功した．一方で，44 件中 18 件で，LLM エージェントが計画の作成に失敗した．そのうち 11 件は，Brwoser-use の実行部の技術的制限が理由で，計画を最後まで作成できなかった．残りの 7 件では，誤った操作が計画された．

**一般的な指示と具体的な対策手順の比較.** 一般的な指示で計画に成功した組では，具体的な対策手順でも同様に計画に成功した．一方で，一般的な指示で誤った操作を計画した 9 件のうち，具体的な対策手順で計画に成功した組が，以下の 2 件確認された．

- （b）のリモート管理機能の無効化．一般的な指示では誤ったメニューに遷移した．具体的な対策手順では正しいメニュー遷移で，目的の対策実行を計画できた．
- （h）のパスワードの変更．一般的な指示では誤って WPA-PSK キーの変更を計画および実行した．具体的な対策手順ではステップにしたがって，管理画面のログインパスワードの変更を計画および実行できた．

#### 3.3.3 実験 3,4: セキュリティ問題の有無に応じた適切な操作実行

実験 3 では，セキュリティ問題がない機器に対して，意図しない設定変更を行わずに操作を終了できたのが 22 件（表 3 の●と ◑）であった．そのうち，問題がない状態を正しく把握し，意図しない設定変更を計画しなかった場合が 6 件（●），問題の有無を誤認，あるいは把握できなかったが意図しない設定変更を計画しなかった場合が 12 件，意図しない設定変更を計画したが，設定の変更には至らなかった場合が 4 件であった（両者の計 16 件を表中では ◑ で示す）．前者の 12 件のうち，3 件は問題の有無を誤認し，既に無効化されている設定を再度無効化する等，問題がない状態への設定変更を計画した．残りの 9 件は問題の有無を把握できなかった．また，後者の 4 件のうち，2 件は問題の有無を正しく把握し，2 件は問題の有無を把握できなかった．加えて，12 件（○）で意図しない設定変更を完了した．そのうち 8 件で逆の設定変更，4 件で異なる設定変更を完了した．また，12 件のうち，問題の有無を誤認したのが 8 件，把握できなかったのが 4 件であった．

実験 4 では，意図しない設定変更を行わずに操作を終了できたのが 32 件（●と ◑），意図しない設定変更を完了したのが 2 件（○）で，いずれも逆の設定変更であった．実験 3 と比較して，成功事例（●）が 6 件から 14 件に増加し，失敗事例（○）が 12 件から 2 件に減少した．また，問題の有無を誤認する事例が 11 件から 6 件に減少した．

## 4. 考察

### 4.1 失敗要因の分析

実験 1，2 で計画の作成や操作の実行に失敗した事例に

表 3: 実験 1，2，3，4 の結果

| | ポート開放の無効化 | 自動ファームウェア更新 | パスワードの変更 | リモート管理機能の無効化 | VPN サーバ機能の無効化 |
|---|---|---|---|---|---|
| (a) ASUS RT-AC66U B1 | ○ ○/◑ ◑ | ○ ○/ ◑ ◑ | ○ ○/ - - | ○ ○/◑ ◑ | ○ ○/◑ ◑ |
| (b) D-LINK DSR-250N | ○ ○/◑ ● | - - / - - | ○ ○/ - - | ○ ◐/◑ ◑ | ○ ○/◑ ◑ |
| (c) ELECOM WRC-300FEBKS | ○ ○/● ● | - - / - - | ● ●/ - - | - - / - - | - - / - - |
| (d) LINKSYS E1200 | ● ●/○ ◑ | - - / - - | ● ●/ - - | ● ●/◑ ◑ | - - / - - |
| (e) LOGITEC LAN-W301NR | ○ ○/◑ ◑ | ○ ○/◑ ◑ | ○ ○/ - - | - - / - - | - - / - - |
| (f) NETGEAR RAX50 | ● ●/○ ● | ● ●/◑ ◑ | ● ●/ - - | ● ●/○ ● | ● ●/○ ○ |
| (g) TP-Link Archer AXE5400 | ● ●/● ● | ○ ○/◑ ◑ | ● ●/ - - | ○ ○/○ ◑ | ○ ○/○ ◑ |
| (h) ASUS TUF Gaming AX4200 | ◐ ◐/◑ ● | ● ●/● ● | ○ ●/ - - | ◐ ◐/● ◑ | ○ ○/◑ ● |
| (i) NETGEAR DGN2200V4 | ● ●/○ ● | - - / - - | ◐ ◐/ - - | ● ●/○ ● | ● ●/○ ◑ |
| (j) Mercusys MR70X | ● ●/◑ ● | ○ - /○ ◑ | ● ●/ - - | ○ ○/○ ◑ | ○ ○/○ ○ |
| (k) BUFFALO WSR-1500AX2L | ● ●/● ● | ● ●/● ● | ● ●/ - - | ● ●/◑ ● | - - / - - |

表中の結果は，左から順に実験 1，2，3，4 に対応．円の左半分は「アクション計画の作成」，右半分は「操作の実行」に対応．

**実験 1，2:** セキュリティ問題のある機器の対策に，●成功 ◐計画のみ成功 ○失敗．

**実験 3，4:** セキュリティ問題の有無に応じた適切な操作に，●成功 (問題がない状態を正しく把握し，意図しない設定変更を行わない)，◑操作のみ成功 (問題がない状態の把握，または適切な操作計画の作成に失敗したが，意図しない設定変更を完了しない)，○失敗 (意図しない設定変更を完了する)

ついて，その原因を考察する．

#### 4.1.1 Browser-use の実行部の技術的制限

（a）と（e）の機器では，Browser-use が管理画面上でクリック可能要素を適切に把握できず，操作が行われなかった．また，（g）の自動ファームウェア更新とリモート管理機能の無効化は，メニュー部分とコンテンツ部分から構成される管理画面において，メニュー部分をスクロールできず，目的の設定画面に到達することができなかった．加えて，計画に成功したが実行に失敗した事例では，HTML でない確認ダイアログを認識することができない，または絵文字ボタンやトグルボタンのクリックを試みるが実行できない，のいずれかが理由で失敗した．

これらの失敗は本実験で用いた Browser-use の実行部の技術的制限に原因があるため，LLM エージェントによる WebUI 情報の取得およびアクション実行技術の高度化により，対策を自動実行できる可能性があると考えられる．

さらに（j）のリモート管理機能の無効化も，部分的にスクロールしてリモート管理設定の項目が画面上に見えるように操作することが困難であった結果，リモート管理設定項目の上に配置されていた別の設定項目を，リモート管理に関する項目であると誤認してしまった．本事例は，Browser-use の実行部の技術的制限に起因して LLM エージェントが計画作成を誤った事例であると言える．

#### 4.1.2 セキュリティ問題の有無の診断を包む指示

実験 1 および 2 で誤ったアクション計画が作成された事例のうち 4 件は，複数の VPN プロトコルに対する設定を確認した上で，有効になっているプロトコルを無効化するという診断がタスクに含まれるため失敗した．例えば，(b) の機器では複数の VPN プロトコルを一括して確認することができず，プロトコル毎に異なるメニュー遷移で画面を開き，設定を確認するような画面構成になっている．このような煩雑な画面構成の場合，LLM エージェントに自律的に意図した通りの操作を実施させるのは難しいと言える．また，複数の VPN プロトコルを一括して確認できるような画面構成においても，無効になっている VPN サーバ機能の有効化を計画する事例があったことから，LLM エージェントは現在の設定状態を正しく認識できない場合があることが明らかになった．

#### 4.1.3 具体的な対策手順の誤りと厳密性の欠如

（b）の機器のポート開放の無効化では，誤った対策手順を与えた結果，目的の対策を実施できる画面まで遷移することができなかった．このように，LLM を用いてマニュアルを解析した結果得られた対策手順が誤っている場合，LLM エージェントが正しく対策を実行できないことがある．一方で，（b）のパスワード変更では誤ったメニュー遷移を含む対策手順が与えられたにも関わらず，LLM エージェントは最終的に正しいメニュー遷移を実行できたことを確認した．この事例から，LLM エージェントは与えられた指示の意図を把握し，それに基づいて現状の画面から適切なメニュー遷移を自律的に判断できる場合もあると言える．

また，（c）の機器のポート開放の無効化では，一般的な指示と具体的な指示のいずれの場合も，ポートフォワーディング自体を無効化してしまった．具体的な指示でも正しい対策が実施できなかった一因として，先行研究 [2] の対策手順が，人が読むことを前提として作られていることが考えられる．本件では 23 番ポートのルールを選択した後に「選択して削除」ボタンを押すことが正しい操作となっているが，対策手順には「無効」を選択後「適用」ボタンを押す旨の記載がされており，厳密には正しくなかった．その結果，同じ画面上で「無効」という設定変更が可能なポートフォワーディング自体の無効化を計画したと考えられる．

このように，人間ならば多少ボタンの記載が異なっていても臨機応変に操作が可能であるのに対し，LLM エージェントを用いた自動ブラウザ操作ではそれが難しい場合がある．

### 4.2 一般的な指示および具体的な対策手順による対策自動化の比較

一般的な指示で計画に失敗し，具体的な対策手順で計画に成功した事例が 2 件存在しており，具体的な対策手順を与えた方が意図した対策の実行がより確実になると言える．

また，一般的な指示を与えた場合と比較して，具体的な対策手順を与えた場合には，管理画面へのログイン後に LLM エージェントが目的の設定変更を行うために様々なメニューを探索する様子が見られなかった．これは，具体的な対策手順では選択すべきメニューが逐次的に記述されているためと考えられる．目的の設定変更を行う画面に迷うことなくたどり着けることは，ユーザの機器での対策実行時間の短縮という利点がある．また，セキュリティ対策を実施するユーザの段階的サポートとして，まずは LLM エージェントが対策を自動実行する様子の録画を提供することも想定しているが，その場合，余計なメニュー遷移のない動画の方がユーザにとって分かりやすいと考えられる．さらに，自動実行によるサポートにおいても，LLM エージェントが余計なメニュー遷移を行わないことは，対策実行の様子を見ているユーザの不安感の軽減につながると考えられる．

以上の観点から，ユーザの機器に対応するマニュアルが取得できる場合には，具体的な対策手順を生成して LLM エージェントに与えることで，より正確で，ユーザにとって安心できる対策自動化の実現につながると考えられる．

### 4.3 IoT セキュリティ対策の自動化可能性

実験 1，2 で，Browser-use の実行部の技術的制限に起因して最後まで計画を作成できなかった事例を除いて LLM による計画作成能力を評価したところ，計画精度は一般的な指示で 72.7%（24/33），具体的な対策手順で 78.8%（26/33）であった．LLM は高精度で正しい対策の実施を計画できたと言える．

さらに，セキュリティ問題の有無の診断を包む指示を与えた事例を除外し，セキュリティ問題の存在を前提として対策の実行のみを指示した場合の計画精度を評価したところ，一般的な指示において，85.7%（24/28），具体的な対策手順で 92.9%（26/28）であった．ユーザの IoT 機器が抱えるセキュリティ問題が既知である条件下では，さらに高い精度で計画を作成できると言える．今後，LLM エージェントを用いた自動ブラウザ操作の実行部の技術的発展に伴い，一般的な指示と具体的な対策手順のいずれにおいても，IoT セキュリティ対策の自動化可能性は高まると考えられる．

### 4.4 セキュリティ対策自動化の課題

実験 2 の結果から，セキュリティ問題があることを前提として対策の実行のみを指示した具体的な対策手順を与えても，意図しない設定変更を完了してしまうことがあるという課題が明らかになった．セキュリティ対策では高い正確性が求められることから，実用化においては課題が残る．精度向上のためには，先行研究 [2] の対策手順作成において，LLM エージェントによる自動ブラウザ操作で与える指示として最適な対策手順を作成させるようなプロンプト設計が有効であると考えられる．また，機器が抱える具体的なセキュリティ問題が判明している条件下では，診断タスクを含めず，当該問題の改善に特化した指示を生成させるプロンプト設計が有効であると考えられる．さらに，対策手順にチェックリストを追加する等の工夫により各メニュー遷移や操作の実行が確実となり，LLM エージェントによる対策自動化の信頼性を高められる可能性があるため，今後検討が必要である．

実験 3，4 の結果から，機器にセキュリティ問題が存在しない場合であっても，LLM エージェントが設定状態を正確に認識できず，結果としてリスクのある設定変更をしてしまう可能性が確認された．ハルシネーション抑制のためのパラメータ調整により，このリスクを一定程度低減できたが，それでも設定状態を誤認する事例や意図しない設定変更を計画および完了してしまう事例が複数存在した．このことから，LLM エージェントを用いたセキュリティ問題の有無の診断，および問題の有無に応じた適切な操作の実行には，解決すべき課題が残されている．

実験 3，4 では，LLM エージェントを用いてセキュリティ問題の有無を診断したが，ユーザから管理画面へのログイン権限が与えられる場合には，他の診断手法の適用も考えられる．とりわけ，LLM エージェントのような確率的生成モデルではなく，ルールベースの決定的手法を用いることで，より高精度かつ安定的な診断ができる可能性がある．このような診断を前段とし，その結果に基づいて LLM エージェントが対策を実行する手順が，実用化の観点からは有望である．

## 5. 関連研究

**ユーザの IoT 機器に即した具体的な対策手順の生成．** IoT 機器のマニュアルを LLM によって解析することで，セキュリティ問題へのより具体的な対策手順を生成する手法が提案されている [2]．生成された対策手順は，具体性に欠ける，かつ情報が分散して記載されているというマニュアルの問題点を解消している．

**ユーザによるセキュリティ対策実施の難しさ．** ウイルス感染の疑いがある IoT 機器を前にしても，修復を試みなかったユーザーが全体の 74%にのぼるという実態が明らかになっている [3]．その理由として，「どこから始めたらよいか分からない」と答えたユーザが全体の 45%であった．IoT セキュリティ診断サービスである am I infected? [1] によるセキュリティ診断でセキュリティ問題が判明した後に

対策を試みたユーザのうち，17%が「セキュリティ問題を抱える機器の特定と操作が難しい」ことを理由に途中で対策を断念したことが明らかになった．また，セキュリティ問題があることが判明したにも関わらず対策を試みなかったユーザの 30%が「どう対策すればよいか分からなかった」と回答している．Chrome 拡張機能として実装したサイバーセキュリティ質問応答アシスタントを用いた調査 [5] では，行動変容に至らないユーザが一定数存在することも示されている．本研究の提案手法であるセキュリティ対策自動化によってユーザの負担が軽減され，技術不足や対策実施への消極的な態度が理由で自身で対策を実施できないユーザへの効果的なサポートが期待できる．

**LLM のセキュリティ関連タスクにおける性能．**LLM が一般的なセキュリティおよびプライバシー（S&P）に関する誤解を平均 21.3%の割合で誤って支持するなど，S&P 助言における信頼性の限界が示されている [6]．また，GPT-3.5 が有限状態機械の設計における 3 種類のセキュリティ規則違反の検出でそれぞれ 79.12%，82.30%，91.74%の精度を示し，専門ツールと比較しても一定の適用可能性があると確認されている [7]．また，温度パラメータの綿密な調整や詳細なプロンプト設計が一貫性と検出精度の向上に寄与する可能性が指摘されている．

本実験では自動ブラウザ操作の計画作成に LLM を用いた．実験 1，2 では，前述の 4.3 節で示した通り，LLM の計画精度が一般的な指示で 72.7%，具体的な対策手順で 78.8%となった．さらに，セキュリティ問題があることを前提として対策の実行のみを指示した場合の計画精度は一般的な指示で 85.7%，具体的な対策手順で 92.9%となった．これらの誤り率は，論文 [6] や [7] と同程度であることが確認された．本実験でも温度パラメータの調整やプロンプト設計を適切に行うことにより LLM の計画作成の精度向上が期待される．

**LLM エージェントの技術的脆弱性．**LLM エージェントは人間の指示の意図を十分に理解できず誤解を生じ，不適切または危険な行動を取る可能性がある．人の支援に特化したパーソナル LLM エージェントが結果の頻繁な検証を避けることでユーザへの中断を最小化するため，LLM ベースのチャットボットと比較して，誤った回答を生成した場合の影響が大きくなることが指摘されている [8][9]．本実験でも，LLM エージェントが指示の意図を正確に把握できず，意図しない設定変更を計画する事例が確認された．

## 6. 研究倫理

本研究は，LLM エージェントを用いて IoT 機器の管理画面を操作することによるセキュリティ対策の自動化可能性を検証するものである．実験で使用したルータの製品名および AI サービス名を論文中で明記しているが，ベンダーやサービス提供者に対する直接的な影響は想定されないため，研究倫理上の問題はないと判断した．

## 7. まとめ

本稿では，LLM エージェントを用いてユーザの IoT 機器の管理画面を自動操作してセキュリティ対策を実行する手法と，その可能性を評価するための実験を行った．実験では，一般的な指示，または各機器に即した具体的な対策手順を LLM エージェントに与えて，正しい対策の計画と実行ができるかを検証した．一般的な指示で，計画成功率は 53.3%，対策実行成功率は 46.7%であった．具体的な対策手順では計画成功率は 59.1%は，対策実行成功率は 50.0%であった．失敗事例の多くは Browser-use の実行部の技術的制限に起因しているため，これを克服する実装により，IoT セキュリティ対策の自動化可能性は高まると考えられる．一方で，具体的な対策手順を与えても意図しない設定変更を完了する事例や，セキュリティ問題がない状態を LLM エージェントが正しく認識できず，逆にリスクのある設定変更を完了する事例が存在することも明らかになり，対策自動化の実現の上では課題が残る．今後は，上記の技術的課題の解決と，LLM エージェントを用いた対策自動化に対するユーザ受容性について，ユーザスタディを通じて明らかにしていく．

**謝辞**　本研究の一部はＮＥＤＯ（国立研究開発法人新エネルギー・産業技術総合開発機構）の委託事業「経済安全保障重要技術育成プログラム／先進的サイバー防御機能・分析能力強化」（JPNP24003）によるものである．本研究の一部は JSPS 科研費 23K11099 の助成を受けて行われた．

### 参考文献

[1] Takayuki Sasaki et al.: Am I infected? Lessons from Operating a Large-Scale IoT Security Diagnostic Service, *USENIX Security '25* (2025).
[2] 竹内謙仁 ほか：LLM を用いたユーザマニュアル解析による機器に即した IoT セキュリティ対策手順の生成，CSS2024 論文集 (2024).
[3] Nissy Sombatruang et al.: Internet Service Providers' and Individuals' Attitudes, Barriers, and Incentives to Secure IoT, *USENIX Security '23* (2023).
[4] Müller, M. and Žunič, G.: Browser Use: Enable AI to control your browser (2024).
[5] Lea Duesterwald et al.: Can a Cybersecurity Question Answering Assistant Help Change User Behavior? An In Situ Study, *USEC '25* (2025).
[6] Yufan Chen et al.: Can Large Language Models Provide Security & Privacy Advice? Measuring the Ability of LLMs to Refute Misconceptions, *ACSAC '23* (2023).
[7] Dipayan Saha et al.: LLM for SoC Security: A Paradigm Shift, *IEEE Access*, Vol. 12, p. 155498–155521 (2024).
[8] Feng He et al.: The Emerged Security and Privacy of LLM Agent: A Survey with Case Studies (2024). Preprint available as arXiv:2407.19354.
[9] Yuanchun Li et al.: Personal LLM Agents: Insights and Survey about the Capability, Efficiency and Security (2024). Preprint available as arXiv:2401.05459.