LLM を用いた ADV ゲームの自律プレイとゲームレビューの生成

末吉竜介[1]　星健太郎[2]　ソンチェヨン[3]　天野翔太[4]

概要：近年、AI がゲーム画面を認識し、マウスやキーボードで操作する技術が発展している。しかし、ゲームを進めることと、その内容を理解してレビューを書くことは異なる。本研究では、AI が ADV ゲームを自律的にプレイし、レビューを生成するシステム「ADV-Agent」を開発した。画面の読み取りには視覚言語モデル（VLM）を用い、会話や選択肢を物語の分岐構造「Story Graph」に記録する。Windows 用ゲーム『ハーバーランドでつかまえて』[1]では、1,384 ステップにわたり自律操作し、出現した 11 か所の選択肢をすべて認識・判断して BAD END「帰京」に到達した。続いて、時系列のプレイ記録（A）と Story Graph（B）から各 3 本のレビューを生成した。AI による事実照合では、正確率が A の 88.6% に対し B は 82.8%、重要な出来事を取り上げた割合が A の 34.8% に対し B は 23.2% であった。一方、7 名による予備的なアンケートでは、5 名が B 方式から選んだレビューを好み、1 名は A 方式を好み、1 名は判断できなかった。4 項目の平均評価は A の 2.54 点に対し B は 3.18 点（4 点満点）であった。これらは、記録方法によってレビューの事実の正確さと読者評価が異なる可能性を示唆する。ただし、実験は 1 作品・1 経路に限られ、生成数や回答者数も少ないため、優劣の一般化はできない。

Autonomous Gameplay and Game Review Generation for Adventure Games Using Large Language Models and Structured Story Graph Memory

Ryosuke Sueyoshi[1]　Kentaro Hoshi[2]　Chaeyoung Sung[3]　Shota Amano[4]

Abstract: Recent advances enable AI agents to perceive game screens and operate games via mouse and keyboard. However, completing a game does not necessarily imply understanding its story well enough to write a review. This paper presents ADV-Agent, a system that autonomously plays adventure games and generates reviews using a vision-language model (VLM) and a Story Graph that records dialogue and choices as branching structures. In the Windows game **Harborland de Tsukamaete**, the agent performed 1,384 steps, recognized and handled all 11 observed choice points, and reached a BAD END. We generated three reviews each from chronological play logs (A) and Story Graph records (B). AI-based fact checking showed factual accuracy of 88.6% for A versus 82.8% for B, and key-event coverage of 34.8% versus 23.2%. In a preliminary survey of seven participants, five preferred the selected review from B, one preferred A, and one could not decide. Mean ratings across four criteria were 2.54 for A and 3.18 for B (out of 4). These results suggest that recording methods may affect factual accuracy and reader evaluations differently. However, the findings cannot be generalized due to the single game and route, and the small numbers of reviews and participants.

1. はじめに（Introduction）

近年、AI がゲームを自動で操作する技術が発展している。従来、ゲームを自動でプレイする目的としては、攻略や動作テストなどが考えられてきた。しかし、ゲームの魅力は「最後まで進められるか」だけで決まるものではない。プレイヤーは登場人物の言動や物語の展開を振り返り、面白かった点や気になった点を含めて作品を評価する。では、**AI もゲームを自分でプレイし、その経験をもとにレビューを書くことができるのだろうか。**

[1]日本工学院、東京テクニカルカレッジ
Nihon Kogakuin College, Tokyo Technical College, Japan
[2]早稲田大学 グローバル・エデュケーション・センター
Global Education Center, Waseda University, Japan
[3]立教大学大学院人工知能科学研究科
Graduate School of Artificial Intelligence and Science, Rikkyo University, Japan
[4]関東学院大学 理工学部 情報ネット・メディアコース
Networking and Multimedia Course, College of Science and Engineering, Kanto Gakuin University, Japan
f23k6142@kanto-gakuin.ac.jp

© 2026 Information Processing Society of Japan

この問いを考えるためには、複数の技術的課題がある。まず、AIが実際のゲーム画面を読み取り、適切な操作を繰り返して結末まで進める必要がある。さらに、長時間のプレイで得られた会話や選択肢、出来事の前後関係を記録し、レビューの根拠として利用しなければならない。特にアドベンチャーゲーム（以下、ADVゲーム）では、登場人物の発言やプレイヤーの選択が物語を形作るため、物語の展開や分岐を整理することが重要となる。そして、これらの記録からレビューを生成し、その正確さや読者にとっての有用性を評価する必要がある。

画面を見て文章や物体を読み取るAIの一種に、視覚言語モデル（Vision-Language Model: VLM）がある。近年は、このVLMを使って画面上のボタンなどを見つけ、マウスやキーボードを操作する「GUIエージェント」の研究が進められている。Tanらの Cradle[2]は画面を使ったソフトウェア操作を扱い、Ahnらの FlashAdventure[3]はアドベンチャーゲームを物語の結末まで進める能力を評価している。一方、ゲームを操作できたとしても、画面ごとの文字列をただ並べただけでは、誰が何を話したのか、どの選択によって次の場面に進んだのかを把握しにくい。

本研究が目指すのは、AIがゲームを自律的にプレイし、その過程で得られた物語情報やプレイ経験を整理することで、人間にとって有用で魅力的なゲームレビューを生成する技術の確立である。

本稿では、その第一段階として、ADVゲームを対象に、画面認識・自律操作・物語記録・レビュー生成を統合した試作システム「ADV-Agent」を構築する。また、ゲーム中の会話と選択結果を、シーン同士のつながりとして記録する「Story Graph」を導入する。Story Graphとは、例えば「この場面でAを選んだ後、次の場面へ進んだ」という関係を、会話の内容とともに保存した記録である。

さらに、プレイ履歴を時系列に並べた記録とStory Graphの2方式からレビューを生成し、事実の正確さと人間による主観的評価の両面から予備的に比較する。本稿は、特定の記録方式の一般的な優位性を実証するものではなく、統合システムの実現可能性を確認するとともに、今後の研究に向けた課題を明らかにすることを目的とする。

以上を踏まえ、本稿では次の三つの研究課題（Research Question: RQ）を設定する。

- **RQ1**：軽量なVLMを用いた画面認識とGUI操作によって、実際のADVゲームの選択肢を認識し、一つの結末まで自律的に進めるか。
- **RQ2**：同じプレイ記録を「出来事を順番に並べた記録」と「Story Graph」の二つの形式でAIに与えた場合、生成レビューの正確さや内容にどのような違いが生じるか。
- **RQ3**：生成されたレビューを人間はどのように評価するか。また、異なる観点を持つ4人の架空のレビュアーによる評価は、ゲームの理解や作品選びに役立つか。

本稿では、2章で関連研究を整理し、3章で提案システムの仕組みを説明する。4章で実験方法、5章で結果を示し、6章で考察と限界、7章で結論を述べる。

2. 関連研究（Related Work）

2.1 画面を使った自律操作とゲームプレイ

Cradle[2]は、スクリーンショットから現在の画面を把握し、マウスやキーボードを使ってソフトウェアを操作する研究である。FlashAdventure[3]は、複数のアドベンチャーゲームを対象に、AIが一場面だけでなく物語の結末まで到達できるかを評価した。長いプレイの間に、過去の出来事を記憶して適切な操作を続けることが課題となる。これらの研究と本研究は、AIが実際のゲーム画面を通じて操作する点で共通している。

一方、TextWorld[4]や Jericho[5]は、文章によって状況を提示し、文章で操作するゲームを研究対象とする。本研究では、ゲームが表示する実際の画面をAIが読み取り、マウスやキーボードで操作する。ゲーム内部のプログラムや、未探索の分岐に関する情報を読み取って攻略する方式は採用しない。

また、本研究の主な目的は、ゲームを最短時間で攻略することではない。自律プレイで経験した内容を記録し、それをゲームレビューに利用できるかを調べる点に特徴がある。

2.2 物語の記録とレビュー生成

ADVゲームを長時間プレイすると、会話や選択の記録は大量になる。すべての文章を単純に並べる方法では、場面の区切りや登場人物の関係が分かりにくくなる可能性がある。そこで本研究では、選択肢が現れた位置で場面を区切り、会話の内容と場面同士の関係を記録する。

ただし、情報を整理すれば必ずレビューが良くなるとは限らない。例えば、情報が増えすぎるとAIが重要な出来事を見落とす可能性もある。そのため、通常の時系列ログとStory Graphの両方からレビューを作成し、実際のゲーム内容に合っているか、人間にとって読みやすいかを比較する。

3. 提案手法（Proposed Method）

3.1 ADV-Agentの基本構成

ADV-Agentは、ゲーム画面を見て、次に行う操作を判断し、その結果を記録するシステムである。処理は次の六つから構成される。(1)ゲーム画面を画像として取得する、(2)VLMで画面の文章と状態を読み取る、(3)マウスやキーボードを操作する、(4)操作が成功したか確認する、(5)

© 2026 Information Processing Society of Japan

物語の内容を記録する、(6)その記録からレビューを生成する（図 1）。

まず画面が文章の表示中なのか、物語上の選択肢を表示しているのか、システムメニューなのかを区別する。文章の表示中であれば先へ進め、選択肢が出ていれば候補を確認して選ぶ。ゲーム内のプログラムを変更するのではなく、人間のプレイヤーと同様に画面を通じて操作する。

画面の読み取りとプレイ中の判断には、ローカル PC で動作する Agents-A1-4B（Q4_K_M、画像処理用 mmproj）[6] を使用した。推論環境は KoboldCpp 1.117.1[7]である。レビュー生成には、保存した会話や選択結果を文章として扱う Qwen3-4B-Instruct-2507 を用いた。Agents-A1-4B は画面画像を含む入力から状態と文字列を認識し、次の操作を判断する役割、Qwen3 は記録済みの文章をもとにレビューを書く役割を担う。このように、画像を扱う処理と文章を生成する処理で必要な機能が異なるため、モデルを分けた。ただし、両モデルの性能比較を行ったわけではない。

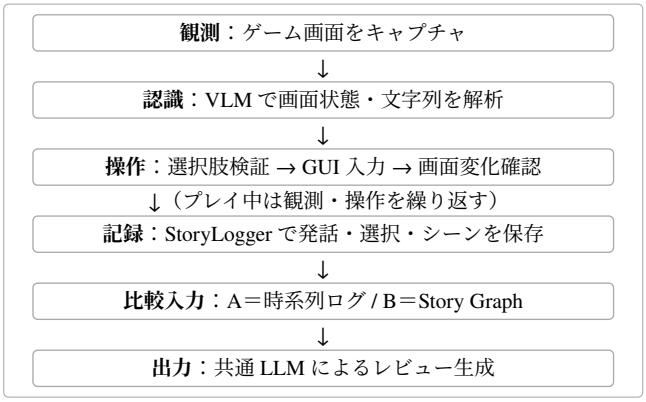


図 1　ADV-Agent の自律プレイ・物語記録とレビュー生成の流れ

3.2 選択肢の検出と安全な操作

ADV ゲームでは、本文と選択肢が似た文字で表示されることがある。今回の対象作品では、通常の文章は白く表示される一方、選択肢はマウスポインターを重ねていないときに灰色で表示される。この違いを利用し、VLM の判断だけでなく、文字色や位置も確認する。

さらに、選択肢の上へマウスポインターを移動すると表示が変化する「ホバー」の反応を調べ、クリックする位置を確かめる。例えば、VLM が通常の会話と判断しても、灰色の文字が並んでいれば選択肢の可能性を調べ直す。実際にクリックした後も、本文が切り替わるなどの画面変化を確認できるまでは、選択が成功したと記録しない。

複数回確認しても結果が一致しない場合は、操作を推測して続けるのではなく、安全のため実行を停止する。ゲームのウィンドウが前面にない場合や、ゲーム自体が終了した場合も同様である。これにより、誤ったクリックを正しい選択として記録してしまうことを防ぐ。

3.3 Story Graph による物語の記録

ADV-Agent は、画面から読み取った会話を Story Graph として保存する。ここでいう「シーン」は、選択肢によって区切られた物語のまとまりである。各シーンには、その場面で表示された会話、登場人物、選択肢、次に進んだシーンを記録する。これにより、例えば「主人公がある発言をした後、どちらの選択肢を選び、次にどの場面へ進んだか」をたどれるようにする。

処理の役割も分けた。VLM は画面上の文字を読み取ることに集中し、誰の発言かを分ける処理や、同じ文章の重複除去、シーン同士の関連付けは Python のプログラムで行う。後者は、同じ入力に同じ規則を適用する「決定論的な処理」である。AI に多くの判断を同時に任せず、規則で処理できる部分をプログラム側に分担させた。

保存形式には JSON を使用し、1 回のプレイ経路を `story_tree.json` に記録する。ただし、このグラフはゲーム内に存在するすべての分岐を示すものではない。**実際に画面で確認し、進んだことを確かめた場面だけを記録したものである。**

3.4 セーブ＆ロードを用いた分岐探索

長時間の自律プレイでは、途中で実行を停止したり、異なる選択肢を試したりする必要がある。そのため本システムでは、一定時間ごと、および選択肢を選ぶ直前にゲームのセーブを行い、そのときの画面と Story Graph の記録を対応付けて保存する。

さらに、未探索の選択肢を順番に試す深さ優先探索（Depth-First Search: DFS）を追加した。DFS は、ある分岐を先まで進み、行き止まりや結末に達した後、未探索の選択肢が残る直近の分岐点へ戻る方法である。本システムでは、ゲームのロード機能を使って分岐点に戻る。これにより、毎回ゲームの最初から同じ文章を読み直す必要がなくなる。このように以前の分岐点へ戻って別の選択を試す操作を、バックトラッキングと呼ぶ。

複数回のプレイで得た情報は、元の記録を上書きせず、どの試行で観測したかを区別して統合する。統合結果からは Mermaid 形式の分岐図も生成できる。ただし、本稿執筆時点で確定した結末は 1 種類であり、複数の結末に到達できることや、すべての分岐を探索できることまでは確認していない。

3.5 二つのレビュー生成方式とクロスレビュー

プレイ記録の整理方法がレビューに与える影響を調べるため、二つの方式を比較した。**方式 A** は、画面に表示された会話や選択結果を、プレイした順番に並べた「時系列ログ」を入力する方法である。**方式 B** は、同じ会話や選択結果に、シーンの区切りや場面同士のつながりを加えた「Story Graph」を入力する方法である。レビューを作る AI モデルと基本的な指示は両方式で共通とし、入力する記録の形式

© 2026 Information Processing Society of Japan

を変えた。時系列ログは、分岐構造を付加しない比較基準として採用した。

また、一つの長文レビューとは別に、複数の視点から短評と点数を提示する「クロスレビュー」も実装した。着想の一つは、ゲーム雑誌『週刊ファミ通』の「新作ゲーム クロスレビュー」である。同企画では 4 人のレビュアーが各 10 点満点で評価する[8]。本研究では 4 人の架空のレビュアーとして、「王道・シナリオ評論家」「辛口・ゲームシステム評論家」「人情・キャラクター愛好家」「ライト・カジュアル評論家」を設定し、各 10 点満点の採点と短評を生成させた。ただし、実際の編集者がプレイする同企画とは異なり、AI が観測した同じ 1 経路の記録を入力とする。各レビュアーには他者の採点を示さず、異なる観点から独立して評価するよう指示した。

4. 実験方法（Experiments）

4.1 対象作品と実行環境

対象作品には、JAM による Windows 用 ADV ゲーム『ハーバーランドでつかまえて』[1]を用いた。ゲームは検証用に複製し、NScripter 3.06 で動作させた。同梱の旧実行ファイルは NScripter 1.22 であり、以前の試行では物語の終盤でゲームのプロセスが終了する事象があった。この事象の原因が旧版にあるとは確定できていないが、終幕の判定やセーブ・ロードの検証を安定して行うため、開発元が公開する 3.06[9]に置き換えた。置換後はタイトルからの進行、保存・復元および終幕までの動作を実機で確認し、元のゲームデータは保全した。実験用 PC は Windows 11、Intel Core i7-9700K、NVIDIA GeForce RTX 2070 SUPER（8 GB）、メモリ 64 GB である。画面解析用の VLM は、KoboldCpp 1.117.1 を利用して PC 内で実行した。

ゲーム本体や以前の実験記録を誤って変更しないよう、操作は検証用の複製に限定した。ファイルの同一性はハッシュ値を用いて確認した。ハッシュ値とは、ファイルの内容から計算される値であり、操作前後で一致すれば内容が変わっていないことを確かめられる。

4.2 自律プレイの評価方法

ゲームをタイトル画面から開始し、AI が一つの結末まで進めるかを調べた。確認したのは、選択肢を見落とさずに検出したか、選んだ後に実際の画面が変わったか、途中で停止しても再開できるか、結末に到達したと判定できるか、そして会話が Story Graph に記録されたかである。

選択肢の件数は、保存された実際の画面を確認して数えた。したがって、ここで報告する検出件数は**今回通過した経路に現れた選択肢**についての結果であり、訪れていない場面の検出性能を示すものではない。

4.3 レビュー A/B の生成条件

完走した経路の記録は、12 シーン、会話 1,340 行であった。レビュー生成前に、重複表示や文章が途中まで表示された記録を、両方式で共通の規則により除去した。最終的に 1,112 行を入力に使用した。画面と照合して確認できた選択肢の文字の誤りは修正したが、実際に選んだ選択肢や進んだ経路は変更していない。

方式 A の入力は 18,148 トークン、方式 B は 22,802 トークンであった。「トークン」は AI が文章を処理するときの単位であり、文字数と完全には一致しない。方式 B ではシーンや分岐の情報を付け加えたため、A より 4,654 トークン多くなった。このため、両方式は物語の元情報を共通にしているが、入力の長さは同じではない。

レビュー生成には Qwen3-4B-Instruct-2507（UD-Q4_K_XL）[10]を使用した。画像を読む必要がないレビュー生成では文章処理用のモデルを採用し、方式 A と B には同一のモデルを用いた。モデル間の性能差は本研究の評価対象ではない。実行環境は KoboldCpp 1.117.1、入力上限に関わるコンテキスト長は 32,768、生成設定は temperature 0.7、top_p 0.8、top_k 20、max_tokens 1,200 とした。これらの設定は両方式で共通とし、乱数シード 20261025、20261026、20261027 をそれぞれ指定して各方式 3 本、合計 6 本のレビューを生成した。ただし、同じシードで生成結果が完全再現されることは検証していない。生成時の目標は本文 450〜550 字で、全 6 本が超過した。別に定めた採用条件は本文 400〜650 字などであり、A3（630 字）のみ条件を満たした。B の 3 本はいずれも条件を満たさなかったため、条件未達のまま字数が最も少ない B1（776 字）をアンケート用に選んだ。AI による事実照合では、選定に伴う偏りを抑えるため両方式の全 6 本を集計した。

4.4 レビューの事実確認と評価指標

レビューが読みやすくても、実際のゲームにない出来事を書いていては問題がある。そのため、レビューの文章を元のプレイ記録と照合し、内容の正確さを調べた。まず、主人公の転勤、少女との出会い、登場人物の関係の変化、帰京など、物語の重要な出来事を 23 件選び、比較前に一覧を固定した。

事実正確率は、レビュー中の事実に関する記述のうち、正しかった割合である。計算では、判断がつかない記述を除き、「正しい記述数÷（正しい記述数＋誤った記述数）」とした。**重要イベント網羅率**は、23 件の重要な出来事のうち、レビューに正しく取り上げられた割合である。さらに、人物や出来事の取り違え、具体的な出来事を根拠とした評価文の割合、根拠のある改善提案の件数も数えた。

この照合には、Claude Code を介して Claude Opus 5.5 を使用した。各判定について、根拠となるプレイ記録の行やシーンを保存した。ただし、同じ AI は実験資料の整備にも

© 2026 Information Processing Society of Japan

使用しており、独立した第三者による評価ではない。また、すべての判定を人間が確認したわけではないため、本稿の数値は AI による事実照合の結果として扱う。

4.5 人間による予備的なアンケート

次に、生成されたレビューを実際の読者がどう評価するかを調べた。時系列ログ方式で生成したレビュー A3 を「X」、Story Graph 方式で生成したレビュー B1 を「Y」とし、生成方法を明かさずに提示した。回答者には、文章の分かりやすさ、内容の具体性、作品の特徴や魅力の伝わりやすさ、ゲームを遊ぶかどうかを決める際の参考度を、それぞれ「1＝全くそう思わない」「2＝あまりそう思わない」「3＝ややそう思う」「4＝強くそう思う」の 4 段階で評価してもらった。また、X と Y のどちらを好むかを尋ねた。

さらに 4 人格クロスレビューを提示し、視点の違いが感じられるか、作品を理解しやすいか、ゲーム選びの参考になるか、今後も利用したいかを尋ねた。

本稿では、2026 年 10 月に得られた、研究利用に同意した 7 名の回答を分析した。普段のゲームプレイ頻度は「ほぼ毎日」が 4 名、「週に数回」が 1 名、「月に数回」が 2 名であり、対象作品をプレイしたことがある回答者は 3 名であった。回答者を無作為に選んだわけではなく、レビューの提示順も X、Y で固定した。また、人間評価に使用した文章は各方式 1 本ずつである。本稿では選択式の回答のみを集計し、自由記述は分析対象から除いた。

5. 実験結果（Results）

5.1 一つのエンディングまでの自律プレイ

『ハーバーランドでつかまえて』を最初から自律プレイした結果、**1,384 ステップを実行し、約 2 時間 55 分の経過時間で BAD END「帰京」に到達した。途中の約 3 分の停止・再開を除く実行時間は約 2 時間 52 分であり、VLM の推論待ちやセーブ操作を含む**。この間に実際に現れた 11 か所の選択肢をすべて認識し、判断した。保存された物語の記録は 12 シーン、会話 1,340 行であった。ゲーム内に BAD END の表示が出たことに加え、タイトル画面へ戻ったことも確認し、一つの経路が終わったと判定した。本稿でいう「完走」は、一つの物語経路が結末に到達したことを意味し、ゲーム内の全エンディングを見たことではない。

また、選択肢直前のセーブから異なる選択肢を選び直し、別経路で同じ BAD END「帰京」に再到達した試行も 2 件あった。再開からの所要時間はそれぞれ 23 分 50 秒、55 分 15 秒で、ロード、推論待ち、セーブ操作を含む。ただし、再開地点が異なるため初回の所要時間と直接比較できず、探索の高速化を示す数値ではない。新たな種類のエンディングへの到達も未確認である。

5.2 二つのレビュー生成方式の比較

表 1　AI によるレビューの事実照合（各方式 3 本、1 作品・1 経路）

評価項目	A：時系列ログ	B：Story Graph
レビュー本文の平均字数	787 字	789 字
事実正確率（判断不能を除外）	31/35＝88.6%	24/29＝82.8%
正誤を判断できなかった事実主張	0 件	4 件
人物・選択・展開の取り違え	4 件	5 件
重要イベント網羅率（各レビュー平均）	34.8%	23.2%
具体的な根拠を示した評価文の割合（平均）	46.7%	38.2%
根拠のある改善提案	5 文	1 文

表 1 に示すように、今回の AI による照合では、**時系列ログ方式 A の方が、事実の正確さと重要な出来事の取り上げ方の両方で良好な結果**となった。Story Graph 方式 B について、構造化した記録がレビューをより正確にするという効果は確認できなかった。ただし、B の 3 本の中には、照合対象となった事実記述の正確率が 100% のレビューもあり、生成ごとのばらつきが見られた。

両方式に共通して、物語の冒頭と結末は比較的取り上げられる一方、中盤以降の重要な出来事が抜け落ちる例があった。具体的には、A3・B1・B2 の 3 本で、主人公が「帰京」を自ら選んだかのような記述があった。実際の帰京は仕事上の失敗と辞令によるもので、プレイヤーが選んだ選択肢ではない。また B1 には、主人公が神戸に満足せず「帰国する」という記述も見られたが、実際には神戸から東京への帰京であり、主人公の心情も正しく捉えていない。文章が自然でも、物語の事実や結末の意味を取り違える問題が残った。

5.3 人間によるレビュー評価

表 2　アンケートにおけるレビュー評価（4 段階、回答者 7 名、各項目の平均）

評価項目	X＝A3（時系列ログ）	Y＝B1（Story Graph）
文章の分かりやすさ	2.71	3.29
内容の具体性	2.57	3.29
作品の特徴・魅力の伝わりやすさ	2.43	3.00
ゲーム選びの参考度	2.43	3.14
4 項目の平均	**2.54**	**3.18**

アンケートでは、**7 名中 5 名が「どちらかといえばレビュー Y」を、1 名が「どちらかといえばレビュー X」を選び、1 名は「判断できない」と回答した**。表 2 の 4 項目の平均評価は、いずれも Y の方が高かった。回答者ごとの 4 項目平均でも、6 名は Y の方が高く、1 名は X の方が高かった。なお、Y の項目平均が高かった 6 名のうち 1 名は、好みの設問では「判断できない」を選んだ。各項目の個別回答は X が 1〜4 点、Y が 2〜4 点に分布しており、回答者間にはばらつきがある。

© 2026 Information Processing Society of Japan

アンケートに用いた2本について、AIによる事実正確率はXが87.5%（7/8）、Yが72.7%（8/11）であった。一方、好みの設問では7名中5名がYを選んだ。

ただし、回答は7名分であり、比べたのは各方式から選んだ1本ずつである。XとYでは文章の長さや生成された回も異なるため、評価差がStory Graphという記録形式そのものによって生じたとは断定できない。

5.4 4人格クロスレビューの結果

4人の架空のレビュアーによるクロスレビューでは、白石カナメ（王道・シナリオ評論家）が8点、黒鉄ジロウ（辛口・ゲームシステム評論家）が8点、春野ひなた（人情・キャラクター愛好家）が9点、小町ルル（ライト・カジュアル評論家）が8点で、合計33/40点であった。短評はそれぞれ69〜86字であり、短い文章と点数で作品を評価する形式を生成できた。

ただし、4人とも関西弁に触れるなど、異なる担当を指定したにもかかわらず注目点が重なる例があった。また、キャラクター担当の短評には、地の文で説明された内容を登場人物の会話のように表現する誤りもあった。元の文章がログに存在していても、それを正しく解釈できているとは限らない。

表3　クロスレビューに対するアンケート結果（4段階）

評価項目	平均	有効回答数
4人の評価視点の違いを感じる	2.83	6
複数の短評によってゲームを理解しやすい	3.14	7
ゲーム選びの参考になる	2.86	7
今後も利用したい	3.00	7

表3にはクロスレビューに関する4項目の評価を示す。別に、ゲーム選びの参考になる形式を尋ねたところ、7名中4名がクロスレビューを、2名がまとまった長文レビューを、1名が「どちらも同程度」を選んだ。評価視点の違いを尋ねた項目は1名が無回答のため、有効回答数は6名である。また、クロスレビューがゲーム選びの参考になるかについては回答が1〜4点に広がっており、平均値だけで全員が高く評価したとは解釈できない。

6. 考察と研究の限界（Discussion and Limitations）

6.1 Story Graphはレビューを良くしたのか

今回の結果は、当初期待していたほど単純ではなかった。Story Graph方式では会話や選択肢のつながりを明示したにもかかわらず、AIによる事実照合では時系列ログ方式の方が良い結果となった。一方、人間による予備評価では、Story Graph方式から選んだレビューの方が好まれた。この違いを考えるうえで、いくつかの可能性がある。

第一に、**AIに渡す情報量の違い**である。Story Graph方式の入力は22,802トークンであり、時系列ログ方式の18,148トークンより約26%多い。Story Graph方式では、会話だけでなく、シーン番号や分岐のつながりなどの情報も読む必要がある。その結果、軽量な4Bモデルが物語の重要な部分を捉えにくくなった可能性がある。

第二に、**一つの経路の記録だけでは分岐条件を確定できないこと**である。ADVゲームの選択肢には、直後の展開を変えるものだけでなく、いったん同じ場面に合流した後で影響が現れるものや、会話の違いにとどまるものもある。Story Graphは今回選んだ選択肢と、その後に観測した場面を記録しているが、別の選択をした場合の展開までは示していない。Bのレビューには「選択の積み重ねが結末に直接影響する」など、今回の1経路からは確かめられない説明があった。ただし、こうした推測をStory Graph形式そのものが引き起こしたかは、本実験からは判断できない。

第三に、**今回は一つの経路しかレビューに利用していない**ことである。Story Graphの利点として期待されるのは、例えば、同じ人物が別の選択によってどのように振る舞うかを比較できる点である。しかし、今回のレビュー入力には、確定した一つの結末へ至る経路しか含まれていない。このため、複数の分岐を比較するという使い方は検証できなかった。

以上の三点はいずれも、今回の実験で原因を特定できたわけではない。今後は入力の長さをそろえる、複数の経路を用意するなど、条件を分けた実験が必要である。

一方、7名の予備評価では、AIによる事実照合で正確率が低かったYを5名が好み、1名はXを好み、1名は判断できなかった。**今回の結果は、ゲームの内容に忠実であることと、読者に好まれることが必ずしも一致しない可能性を示唆する**。ただし、各方式1本ずつの比較であり、両者が一般に独立した評価軸であると断定できるわけではない。自動ゲームレビューの評価では、原記録への忠実さと読みやすさの両方を調べる必要があると考えられる。

6.2 複数のレビュアーに分ける意味

4人格クロスレビューは、短評と点数から作品の特徴を概観できる形式として、回答者の一部から支持された。4人全員が関西弁の描写に触れていたが、**同じ点への言及は、必ずしも役割分担の失敗を意味しない**。複数の視点から共通の特徴が挙がることで、その重要性が伝わりやすくなる可能性もある。一方、担当を変えても実際の着眼点が十分に異なるかは別に検証する必要があり、今回は観点の多様性を定量的に測定していない。

また、引用した発言が実際にゲーム中にあっても、発言の意味や状況を取り違える場合があった。そのため今後は、どの場面を根拠として採点したのかを確認する仕組みや、担

© 2026 Information Processing Society of Japan

当ごとの評価内容がどれだけ異なるのかを測る仕組みが必要と考える。

6.3 ほかの ADV ゲームでも使えるか

今回のシステムは、画面に登場人物名とセリフが表示され、選択肢によって物語が分岐する ADV ゲームを主な対象としている。しかし、ADV ゲームにはほかにもさまざまな仕組みがある。

例えば『STEINS;GATE』では、メール本文中の単語を選んで返信したり、電話に応答したりする操作が物語の分岐に関わる[11]。これは、画面に独立して並ぶ選択肢を検出するだけでは対応できない例である。また、『街 ～運命の交差点～』では、8 人の主人公を切り替えるザッピングによって各人物の物語を進め、一人の行動が別の人物の展開に影響する[12]。『428 ～封鎖された渋谷で～』でも、5 人の主人公の選択と運命が相互に関係する[13]。これらの作品では、一人の経路だけでなく複数人物の進行状況を追跡しなければ、物語を最後まで進められない場合がある。

現在の Story Graph は、主に「どの選択肢を選び、どの場面へ進んだか」を記録する。したがって、こうした作品に対応するには、メール内の語句選択などの特殊な操作と、主人公ごとの進行状況・イベント発生条件を記録する仕組みが必要になる。また、画面全体に文章が表示されるノベル形式（NVL 型）や、発言者名が表示されない作品についても、認識方法を改めて検証する必要がある。

DFS によるセーブ＆ロードを使った分岐探索は、実機で別の選択肢へ進めることまで確認した。しかし、本稿時点では新しい結末への到達は確認できていない。また、把握できた選択肢をすべて調べても、隠れた条件による分岐まで存在しないことを証明できるわけではない。

6.4 今回の評価で分からないこと

本研究の実験対象は 1 作品であり、レビューに使用したプレイ経路は 1 本である。各方式のレビュー生成数は 3 本、アンケート回答者は 7 名である。したがって、別のゲームや別の回答者でも同じ結果になるかは分からない。アンケートは無作為抽出ではなく、文章の提示順も固定した。また、人間に見せたレビューは各方式 1 本であり、その長さも異なっていた。これらの条件が評価結果に影響した可能性がある。

加えて、事実正確率と重要イベント網羅率は AI による照合に基づく。23 件の重要イベントの選定も AI 支援で行っており、すべてを独立した人間が正解と確認したわけではない。したがって、今回の値を絶対的な正確さとみなすことはできない。

また、**出現した 11 か所の選択肢をすべて認識できたことと、Story Graph の全記録が正確であることは別の問題**である。発言者の識別、セリフの書き起こし、シーンの区切り、分岐先の対応などを正確に測るには、人間が作成した正解データと比較し、適合率（Precision）・再現率（Recall）・F1 値などを計算する必要がある。これらは今後の課題とする。

7. おわりに（Conclusion）

本研究では、AI が ADV ゲームの画面を読み取り、マウスやキーボードを操作しながら物語を記録する「ADV-Agent」を開発した。実際のゲームで 1,384 ステップの自律プレイを行い、出現した 11 か所の選択肢を認識して、一つの BAD END まで到達できた。また、このプレイ記録を時系列ログと Story Graph の二つの形式に整理し、それぞれからゲームレビューを生成した。

その結果、AI による事実照合では時系列ログ方式が良好であった一方、7 名による予備的な人間評価では、5 名が Story Graph 方式から選んだレビューを好み、1 名は時系列ログ方式のレビューを好み、1 名は判断できなかった。これは、**AI によるレビューには、物語の事実を正しく伝えることと、読者が読みやすく魅力的だと感じることの両方が求められる**ことを示す初期的な結果と考えられる。ただし、今回の結果だけで Story Graph 方式の一般的な優位性を示すことはできない。

本研究は、AI がゲームを最後まで進めるだけでなく、プレイで経験した物語を記録し、レビューとして表現するための試作基盤を示した。今後は複数のエンディングや別作品への適用、人間が確認した正解データによる記録精度の評価、より多くの回答者によるレビュー比較を進める。

謝辞（Acknowledgements）

本研究の実験およびアンケートにご協力いただいた皆様に感謝する。また、本稿に対して有益なご指摘とご助言をいただいた査読者の皆様に深く感謝する

なお、本研究の実装・分析および本稿の執筆・推敲にあたり、OpenAI 社の ChatGPT および Codex、Anthropic 社の Claude Code、Google 社の Gemini を利用した。

補足資料（Supplementary Materials）

本研究に関連する実験記録、生成レビューおよび評価結果などの補足資料は、Zenodo にて公開している。

公開先：［https://doi.org/10.5281/zenodo.23692972］

参考文献

[1] JAM,「ハーバーランドでつかまえて」. [オンライン]. 入手先: https://vector.co.jp/soft/win95/game/se168288.html

[2] W. Tan, W. Zhang, X. Xu, H. Xia, Z. Ding, B. Li, B. Zhou, J. Yue, J. Jiang, Y. Li, R. An, M. Qin, C. Zong, L. Zheng, Y. Wu, X. Chai, Y. Bi, T. Xie, P. Gu, X. Li, C. Zhang, L. Tian, C. Wang, X. Wang, B. F. Karlsson, B. An, S. Yan, と Z. Lu,「Cradle: Empowering Foundation Agents towards General Computer Control」, *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, vol. 267. PMLR, 2025, ページ 58658–58725. [オンライン]. 入手先: https://proceedings.mlr.press/v267/tan25h.html

© 2026 Information Processing Society of Japan

[3] J. Ahn, J. Kim, H. Yun, J. Son, D. Park, J. Cho, と G. Kim, 「FlashAdventure: A Benchmark for GUI Agents Solving Full Story Arcs in Diverse Adventure Games」, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China: Association for Computational Linguistics, 11 月 2025, ページ 23354–23384. doi: 10.18653/v1/2025.emnlp-main.1192.

[4] M.-A. Côté, Á. Kádár, X. Yuan, B. Kybartas, T. Barnes, E. Fine, J. Moore, M. Hausknecht, L. El Asri, M. Adada, W. Tay, と A. Trischler, 「TextWorld: A Learning Environment for Text-based Games」, *Computer Games Workshop at ICML/IJCAI 2018*, 2018. [オンライン]. 入手先: https://www.microsoft.com/en-us/research/publication/textworld-a-learning-environment-for-text-based-games/

[5] M. Hausknecht, P. Ammanabrolu, M.-A. Côté, と X. Yuan, 「Interactive Fiction Games: A Colossal Adventure」, *Proceedings of the AAAI Conference on Artificial Intelligence*, 巻 34, 号 5, ページ 7903–7910, 2020, doi: 10.1609/aaai.v34i05.6297.

[6] InternScience, 「Agents-A1-4B」. [オンライン]. 入手先: https://huggingface.co/InternScience/Agents-A1-4B

[7] LostRuins, 「KoboldCpp, version 1.117.1」. [オンライン]. 入手先: https://github.com/LostRuins/koboldcpp/releases/tag/v1.117.1

[8] ファミ通.com, 「百花繚乱、104 本の佳作の詰め合わせ！ 2019 年下半期のクロスレビュー殿堂入りソフトカタログ!!」. [オンライン]. 入手先: https://www.famitsu.com/news/202002/05191835.html

[9] N. Takahashi, 「NScripter最終版（公式リポジトリ）」. [オンライン]. 入手先: https://github.com/NaokiTakahashi/NScripter

[10] Qwen Team, 「Qwen3-4B-Instruct-2507」. [オンライン]. 入手先: https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507

[11] 5pb. and Nitroplus, 「STEINS;GATE：フォーントリガーシステム」. [オンライン]. 入手先: https://steinsgate.jp/sgflash.html

[12] K. Fukuyama, 「『街 ～運命の交差点～』×『428 ～封鎖された渋谷で～』×『WILL -素晴らしき世界-』開発者たちが語り合う群像ゲームの作り方（前編）」. [オンライン]. 入手先: https://automaton-media.com/articles/interviewsjp/20180529-68712/

[13] Spike Chunsoft, 「428 封鎖された渋谷で ：作品紹介」. [オンライン]. 入手先: https://www.spike-chunsoft.co.jp/pages/428/

© 2026 Information Processing Society of Japan