セキュリティ教育用ビデオゲームのテーマ抽出に向けたプロンプトエンジニアリング

新井 美音[1,2]　矢内 直人[3]　猪俣 敦夫[1]　花岡 悟一郎[2]

概要：ユーザの関心を得やすいセキュリティ教育手法としてビデオゲームが注目されており, これまで様々なビデオゲームが提案されてきた. しかしながら, これらのビデオゲームに関して, その設計に密接な関りがあるテーマの抽出手法は確立されていない. 本稿では, 大規模言語モデル (LLM) を用いてビデオゲームのテーマを効果的に抽出する方法を, LLM のプロンプトエンジニアリングの観点から議論する. LLM を用いてゲームの紹介ページから得られる情報からテーマを生成できるようなプロンプトを複数提案する. 提案手法の効果を明らかにするために Steam にて公開されているビデオゲームの紹介文を用いて複数の調査も行った. 本稿で明らかにした知見は二つである. 第一に LLM ごとに適切なテーマが抽出しやすいプロンプトが異なることを確認した. 第二に複数の LLM でテーマを抽出するのに適したゲームの紹介情報は希少と考えられる. 最後に, これらの知見を踏まえ, 今後推奨される研究も議論する.

キーワード：セキュリティ教育, ビデオゲーム, 大規模言語モデル, テーマ, 体系調査

Prompt engineering for extracting themes from video games for security education

Mine Arai[1,2]　Naoto Yanai[3]　Atsuo Inomata[1]　Goichiro Hanaoka[2]

Abstract: Video games have recently gained attention as a security education tool that is likely to attract user interest, and various video games have been proposed to date. However, there is no established method for extracting themes closely related to the design of these video games. In this paper, we propose a method for extracting themes from video games using large language models (LLMs) and discuss effective prompts for extraction. In general, keywords can be generated from information obtained from game introduction pages using LLMs, enabling the objective extraction of game themes. To clarify the effectiveness of the proposed method, we also conducted multiple surveys using the descriptions of video games published on Steam. Then, we found two insights: first, we identify that prompts suitable for extracting appropriate themes are different for each LLM; second, it is considered that the game introduction information rarely appears. We also discuss several recommendations for subsequent works.

Keywords: security education, video game, large language models, theme, systemization of knowledge

1. 序論

1.1 背景

情報セキュリティは今や全てのエンドユーザにおいて必要不可欠な技術となっている. しかし, 情報セキュリティは技術の変遷が早いことに起因して, 技術的な対策だけではユーザを保護することは難しい [16]. とくに非専門家であるエンドユーザを保護するためには, 情報セキュリティの教育が必須である [18].

一方, 高い教育効果を発揮するためには受講者の関心を上げることが望ましいが, 情報セキュリティはユーザに

[1] 大阪大学. Osaka University
[2] 産業技術総合研究所, AIST
[3] パナソニックホールディングス株式会社, Panasonic Holdings Corporation

とって二次的な産業であることから [16], 全ての受講者が情報セキュリティに興味を持っているとは言い難い. このため, 情報セキュリティへの関心を高める教材として, ビデオゲームが注目されてきた [15,16]. ビデオゲームは情報セキュリティ技術を仮想的な体験を通じて学習できる点で, 情報セキュリティの教材として優れている [8].

近年では大規模言語モデル (LLM) が様々な分析業務に利用されていること [3,11] に関連して, セキュリティ教育ビデオゲームに関しても LLM を用いた分析がされつつある. 一般にユーザ調査は分析を含めた作業が複雑であることに起因して高負荷なことに加え, 分析者の地域特性といった背景による影響を受けることに影響を受ける懸念がある. これに対し, LLM はユーザ調査の分析作業が人間よりも効果的に行えることが可能であり [7], また, 関連するデータセットも含めて LLM を公開することで結果の一般性と再現性の担保が期待できる. しかしながら, 著者の知る限り唯一の既存研究 [2] は LLM を平易に用いただけであり, プロンプトの影響は考慮できていなかった.

1.2 本稿の学術的問い

本稿の問いは次の通りである: **どのようなプロンプトがセキュリティ教育ビデオゲームのテーマ抽出に良いか？**

上述した問いは, LLM がビデオゲームの分析に有効か, プロンプトエンジニアリングの観点から明らかにする狙いがある. プロンプトの設計は一般に LLM による分析に重要であり, その内容に応じて結果に影響することが考えられる. 本稿ではこの分析への影響を, ビデオゲームのテーマを通じて確認する. ここでいうビデオゲームのテーマとは, ゲームの内容を補助的に表す情報 [9] を意味する. プロンプトの設計を通じて, より適切なテーマが抽出されるか確認する.

1.3 貢献

本稿では LLM によるビデオゲームのテーマ抽出に適したプロンプトを模索する. 具体的には, プロンプトエンジニアリングの観点として, LLM の位置づけを明確にするペルソナ, 思考の展開を示す Chain-of-Though, 例を示す One-Shot Prompting に着目してプロンプトを提案するとともに, その分析を行っている. 最大のゲーム市場である Steam[*1]から 20 本のゲームに着目し, そのコンテンツを紹介するウェブページの情報を用いて実験評価したところ, 以下の知見を得た.

- モデルの違いによりテーマの抽出に適したプロンプトが異なると考えられる.
- 複数の LLM でテーマを抽出するのに適したセキュリティ教育用ビデオゲーム自体の紹介文が希少である.

これらの知見を踏まえて, 後続の研究で推奨されるべき観点についても議論する.

2. 関連研究

本節では関連研究として, 情報セキュリティを扱うゲームに関する既存の体系調査, LLM とそのユーザ調査への応用について述べる.

2.1 セキュリティ教育ビデオゲームとその体系調査

セキュリティ教育用ビデオゲームは, ユーザが情報セキュリティを学ぶために設計されたゲームの総称である. セキュリティ教育用ビデオゲームは学術と産業の二種類に大別され, 学術のものは各技術に特化しており, 産業のものは一般的なセキュリティを扱う傾向がある [1]. セキュリティ教育用ビデオゲームに関する体系調査 [15,16,18] も様々に行われている. 上述した体系調査は LLM を用いていない. 前節でも述べた通り LLM を用いることでユーザ調査において様々な恩恵を得ることが可能となる.

本稿に最も近い既存研究は著者らの先行研究 [2] である. 先行研究において著者らは LLM を用いることで, ビデオゲームのテーマが分析できるか議論した. 前述したとおりテーマとはゲーム内容を補助的に表す情報であり, ユーザがビデオゲームを遊ぶ動機としてジャンルと同様に密接なかかわりがあることが知られている [14]. 例えば, 代表的なテーマであるファンタジーは学習内容に対する宣言的知識を向上させる [17]. 本稿では上述したような知見を後続研究で得られることを期待して, より効果的にテーマを抽出できるプロンプトを検討する点が異なっている.

2.2 大規模言語モデル (LLM) とユーザ調査への応用

LLM は与えられたトークンの頻度を学習することで, トークン列を確率的に生成する機械学習モデルである [12]. ここでトークンとはある言語における固有文字の集合を表す. すなわち, 入力として文字列を与えられ, 次のトークンとして最も確率の高い文字列を生成する. 人間が LLM へ与える動作指示はプロンプトと呼ばれ, 本質的には同じ意味となる指示においても, そのプロンプトに応じて結果が異なることがある. LLM は一般には数百億個のパラメータを包括し, 代表的なものには OpenAI 社の ChatGPT や Google の Gemini がある. LLM は近年では様々な分野で研究ツールとして利用されている [11,13].

ユーザ調査の支援に LLM が有効か明らかにする検討も行われており, 改善できる可能性が示されている [7]. 個々の分析手法に関しては, LLM は帰納的主題分析が可能であること [5], 質的飽和を LLM による主題分析の評価軸として利用できること [6], また, 演繹的手法を通じて内容分析を行う手法の提案 [4] が, 著者の知る限り示されている. 実際の調査分析としても, LLM はヘルスケアデータから有意

[*1] https://store.steampowered.com/

義なテーマを得ることが可能である [11]. 本稿ではセキュリティ教育ビデオゲームを分析する点が, 上記の既存研究とは異なっている.

3. 研究設計

1.2 節で述べた問いを明らかにするための研究設計を述べる. まず分析対象とするセキュリティ教育用ビデオゲームの収集について述べる. 次に, 本稿の主な手法となるプロンプトエンジニアリングについて述べ, それから専門家によるコーディング方法について述べる. 最後に, 問題設定について述べる.

3.1 セキュリティ教育ビデオゲームの収集方法

本稿ではゲーム販売プラットフォーム Steam[*2]から情報セキュリティを扱うゲームを選択し, それらのゲームに関する情報を Steam の API[*3] を介して収集する.

まず収集対象は, セキュリティ教育ビデオゲームの収集は著者の過去の体系調査 [2] に従う. 文献 [2] ではセキュリティ教育用ビデオゲームとしてゲーム販売プラットフォーム Steam[*4]から, 情報セキュリティを扱うゲームとして"cybersecutiy", "e-safety", "security"を検索ワードに, 収集対象として地理的制限を設けず, 英語と日本語のものを 125 本収集した. このうち, 専用の端末などビデオゲームではないもの, ゲームの紹介画面から内容が判断できないもの, 体験版か追加コンテンツに相当するもの, 明らかに情報セキュリティとは関係がないと判断できるものを除外したところ, 108 本が残った. この中から無作為に 20 本のゲームを選択し, Steam からその情報を収集した.

上述したとおり, これらのゲームの情報は Steam の API を介して収集している. このとき, 画像までは API では取得できないため, 画像ファイルは手作業で収集して, それらの画像を API から取得した情報と結合して pdf ファイルとして出力する. これらの pdf ファイルを収集した検体として LLM に与えることで, テーマを抽出する.

3.2 プロンプトエンジニアリング

高品質なプロンプトは AI に実行してほしい命令, AI がタスクを理解するための文脈, 処理させたい具体的な情報としての入力データ, 及び, 回答の形式となる出力形式の四つからなる [10].

本節では本稿の調査に用いた手法と実際に用いたプロンプトを記載する. 本稿ではまず基盤となるプロンプトを用意し, それに追記する形で後続のプロンプトを作成した. 基盤となるプロンプトは以下の通りである.

基盤となるプロンプト

以下のゲームに関連する情報から, 最も適切だと思われる Steam のタグを英語で 20 個推測してください.
【入力】ゲームタイトル：XXX
ゲーム紹介文：XXX
【出力】タグ：

ここで XXX には各セキュリティ教育用ビデオゲームのタイトルあるいは Steam の画面上に表示される紹介文が与えられる.

上述した通り, このプロンプトに逐次追記をしていく. 各プロンプトの追記内容を含めた全体像を表 1 に示す. ここで表中の○は該当の情報が含まれることを表す. プロンプト 1 は上述した基盤となるプロンプトそのものだが, プロンプト 2 ではペルソナとして「あなたはゲーム作成者であり, これから作成したゲームを Steam に公開します. 」という文節を追記することで, LLM にゲームの開発者であることを認識させる. 次にプロンプト 3 はゲームの紹介画像を添付し, プロンプト 4 では Chain-of-Thought として「ただし, ステップバイステップで考えてください. 」と文節を追加することで思考の過程を表示するようにした. プロンプト 5 では one-shot prompting として, Shadows of Doubt[*5]における入力と出力の例を文例として与える. ここで, Shadows of Doubt は著者のうち 2 名が紹介文およびタグを確認し, セキュリティ教育ゲームとして一般的な内容であると判断したため例として使用された. プロンプト 6 では Steam におけるタグの分類に従って, 代表的なタグを 35 個与えた. プロンプト 5 で与えた例はセキュリティ教育ゲームという前提があるのに対し, プロンプト 6 ではセキュリティ教育ゲームに限らない全てのゲームにおいて代表的なタグを与えることで, より多種多様なゲームを評価できることを期待した.

3.3 問題設定

本稿では 1.1 節に示した問いについて, Steam におけるタグに相当するものを LLM により作成できるか議論する. Steam においてタグはゲームの製作者およびユーザが自由に編集できるいくつかの単語からなり, ゲームの特徴を端的に表すとしてゲームの検索などに活用されている. 研究目的で作成されたセキュリティ教育ゲームには一般にタグは設定されていないが, タグを設定することで, ユーザが好きなゲームを探すこと, 研究者がゲームの分析を行うことなどに役立てることができ, 作成されたセキュリティ教育ゲームの幅広い活用が期待できる. Steam においてタグは複数人のユーザにより確認および編集されることが一般的であるため, 集合知である LLM を使用してゲームにタグ

[*2] https://store.steampowered.com/
[*3] https://steamcommunity.com/dev
[*4] https://store.steampowered.com/
[*5] https://store.steampowered.com/app/986130/Shadows_of_Doubt/

表 1: プロンプトの全体像

プロンプト	命令	ペルソナ	入力データ		思考	例	
			タイトルと紹介文	ゲーム紹介画像		ゲーム	タグ
1	○	×	○	×	×	×	×
2	○	○	○	×	×	×	×
3	○	○	○	○	×	×	×
4	○	○	○	○	○	×	×
5	○	○	○	○	○	○	×
6	○	○	○	○	○	×	○

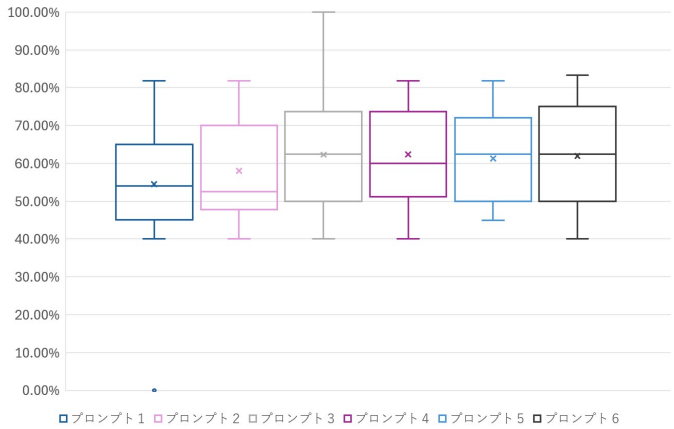

図 1: 各プロンプトにおける正解率 (Gemini).

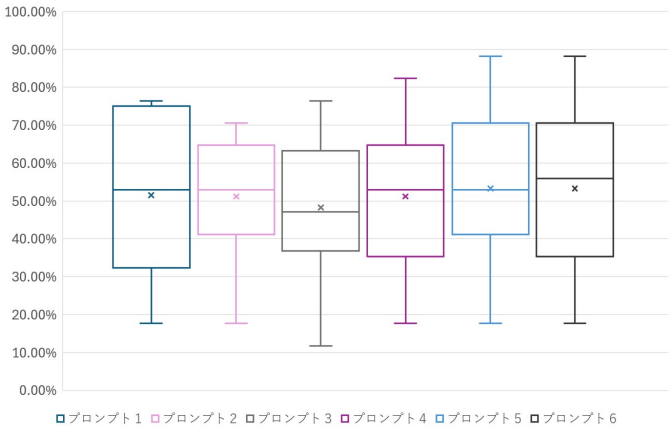

図 2: 各プロンプトにおける正解率 (GPT).

を設定することが期待できる. そこで, LLM が処理できる情報の組み合わせおよび適したプロンプトを模索し, 従来のタグと同等の意義を持つテーマ抽出を実現する.

詳細は次節以降で述べるが, いくつのテーマが LLM に生成されたかという観点と, どのようなテーマが生成されたかという観点の双方からそれぞれについて議論する.

4. 分析結果

本節では前節で調査した結果を述べる. まずプロンプトごとの結果を示したのち, 個々の結果に関する分析としてゲームごとの違いを示す. 最後に, テーマごとの違いを示す.

4.1 プロンプトごとの結果

プロンプトごとの結果を図 1 に示す. 図によると Gemini ではいずれも最も低い正解率が 40%以上はあり, 平均すると 60.0%の正解率になっている. 一方, ChatGPT は Gemini と比べて正解率が不安定になっている. 最も低い正解率は 20% を下回っており, 平均してみても 51.4% と Gemini との差が見られている. これらの差が出た理由について, 画像を追加したプロンプト 3 から 10 ポイント以上の差が見られている点で, Gemini の方が画像の処理に長けているためと考えられる.

4.2 ゲームごとの結果

前述の結果の詳細な分析として, 各セキュリティ教育用ビデオゲームごとの傾向を以下に述べる. まず表 2 に Gemini における各ゲームの正解率を, 表 3 に ChatGPT における各ゲームの正解率をそれぞれ示す. なお, いずれも正解率が最大のものを太字で示している. それぞれの表によると, まず Gemini では Cyber Puzzle HackRow が正解率の最低値が 80%を越えており, 高い精度でテーマが抽出できている. また, 正解率の最高値が 80%を越えてものは Code7: A Story-Driven Hacking Adventure, Cyber Puzzle HackRow, DAB'S NOT DEAD, Deus Ex: Breach, Hack Run Zero, HeistGeist とその数も多い. 一方, ChatGPT では前小節で述べた通り正解率が低いことに起因して, 80%もの数値を達成したゲームは Rendezvous だけだった.

上述した結果を鑑みるに, Gemini と ChatGPT では正解率が高いゲームが異なっていた. 詳細は今後の研究で継続して調査するが, これはモデルの特性に起因すると考えられる. また, 正解率がゲームごとに大きく異なっており, 複数の LLM でテーマを抽出するのに適したセキュリティ教育用ビデオゲーム自体の紹介文が希少と考えられる.

4.3 テーマごとの結果

本節では各セキュリティ教育用ビデオゲームから抽出されたテーマが, Steam におけるタグの区分[*6]において, いず

[*6] https://store.steampowered.com/games

表 2: 各プロンプトにおける正解率 (Gemini)

	プロンプト					
ゲーム名	1	3	3	4	5	6
Anonymous Hacker Simulator	52.94%	47.06%	47.06%	**58.82%**	52.94%	41.18%
ByteBurst: Hacking Simulator	40.00%	40.00%	**46.67%**	**46.67%**	**46.67%**	40.00%
Code7: A Story-Driven Hacking Adventure	**80.00%**	70.00%	70.00%	70.00%	**80.00%**	**80.00%**
Crash Override	45.00%	45.00%	40.00%	45.00%	45.00%	**60.00%**
CTRL Phreak	50.00%	50.00%	50.00%	**55.56%**	50.00%	50.00%
Cyber Puzzle HackRow	**81.82%**	**81.82%**	**81.82%**	**81.82%**	**81.82%**	**81.82%**
DAB'S NOT DEAD	65.00%	55.00%	**80.00%**	70.00%	65.00%	60.00%
DataJack 2020	50.00%	50.00%	**60.00%**	55.00%	50.00%	50.00%
Deus Ex: Breach	0.00%	81.82%	**100.00%**	81.82%	72.73%	72.73%
Dogs and Pigs	45.00%	40.00%	**50.00%**	**50.00%**	45.00%	45.00%
Gunpoint	55.00%	50.00%	**70.00%**	**70.00%**	65.00%	60.00%
Hack Run Zero	50.00%	50.00%	50.00%	66.67%	66.67%	**83.33%**
Hacker Evolution - 2019 HD remaster	**75.00%**	**75.00%**	**75.00%**	**75.00%**	**75.00%**	**75.00%**
HeistGeist	70.00%	75.00%	75.00%	**80.00%**	70.00%	75.00%
MINDHACK	55.00%	**70.00%**	65.00%	60.00%	60.00%	65.00%
nullptr	60.00%	60.00%	65.00%	60.00%	**70.00%**	65.00%
OFF GRID: Stealth Hacking	45.00%	**50.00%**	**50.00%**	40.00%	**50.00%**	**50.00%**
Project RyME	60.00%	**65.00%**	**65.00%**	60.00%	60.00%	**65.00%**
Rendezvous	65.00%	60.00%	60.00%	**75.00%**	**75.00%**	**75.00%**
Retro Hacker	**45.45%**	**45.45%**	**45.45%**	**45.45%**	**45.45%**	**45.45%**

表 3: 各プロンプトにおける正解率 (ChatGPT)

	プロンプト					
ゲーム名	1	3	3	4	5	6
Anonymous Hacker Simulator	29.41%	**52.94%**	41.18%	47.06%	**52.94%**	47.06%
ByteBurst: Hacking Simulator	29.41%	**41.18%**	29.41%	35.29%	**41.18%**	35.29%
Code7: A Story-Driven Hacking Adventure	**41.18%**	**41.18%**	35.29%	35.29%	17.65%	35.29%
Crash Override	**52.94%**	**52.94%**	47.06%	35.29%	**52.94%**	35.29%
CTRL Phreak	41.18%	41.18%	47.06%	52.94%	64.71%	**70.59%**
Cyber Puzzle HackRow	**47.06%**	**47.06%**	**47.06%**	**47.06%**	**47.06%**	**47.06%**
DAB'S NOT DEAD	**76.47%**	64.71%	47.06%	58.82%	64.71%	64.71%
DataJack 2020	52.94%	52.94%	52.94%	64.71%	**70.59%**	64.71%
Deus Ex: Breach	**52.94%**	41.18%	47.06%	47.06%	41.18%	47.06%
Dogs and Pigs	47.06%	47.06%	41.18%	**52.94%**	**52.94%**	**52.94%**
Gunpoint	**76.47%**	70.59%	70.59%	70.59%	70.59%	70.59%
Hack Run Zero	23.53%	**29.41%**	23.53%	23.53%	23.53%	23.53%
Hacker Evolution - 2019 HD remaster	**17.65%**	**17.65%**	11.76%	**17.65%**	**17.65%**	**17.65%**
HeistGeist	**76.47%**	70.59%	**76.47%**	70.59%	70.59%	70.59%
MINDHACK	**76.47%**	70.59%	58.82%	52.94%	52.94%	58.82%
nullptr	58.82%	64.71%	64.71%	64.71%	64.71%	**70.59%**
OFF GRID: Stealth Hacking	58.82%	58.82%	52.94%	64.71%	**70.59%**	58.82%
Project RyME	**70.59%**	64.71%	**70.59%**	**70.59%**	**70.59%**	**70.59%**
Rendezvous	76.47%	64.71%	70.59%	82.35%	**88.24%**	**88.24%**
Retro Hacker	23.53%	29.41%	29.41%	29.41%	29.41%	**35.29%**

れに該当するか議論する．これはある抽出されたテーマがあったとして, そのテーマがもともと抽出されやすいものであったのか, 明らかにする意図がある．このとき, 位置づけが近いテーマ群として各テーマを包括的に考えたときに, それぞれのテーマ群ごとに高い正解率を持つものがあるか確認する．この方法としては表 4 に示されるような Steam におけるタグの区分を利用している．なお, ここに含まれなかったタグを本稿ではその他と括る．Gemini の結果は図 3 に, ChatGPT の結果は図 4 に示される．

表 4: タグの区分と代表例

区分	例				
トップレベル のジャンル	アクション	カジュアル	アドベンチャー	シミュレーション	ストラテジー
ジャンル	パズル	アクション アドベンチャー	アーケード	シューティング	プラットフォーム
サブジャンル	探検	2D プラットフォーム	ローグライト	FPS	3D プラットフォーム
ビジュアルと視点	2D	3D	カラフル	かわいい	ドット絵
テーマと雰囲気	雰囲気	ファンタジー	リラックス	笑える	ホラー
特徴	物語性	コンバット	コントローラ	女性主人公	選択型進行
プレイヤー	シングルプレイヤー	マルチプレイヤー	協力プレイ	オンライン協力プレイ	ローカル マルチプレイヤー

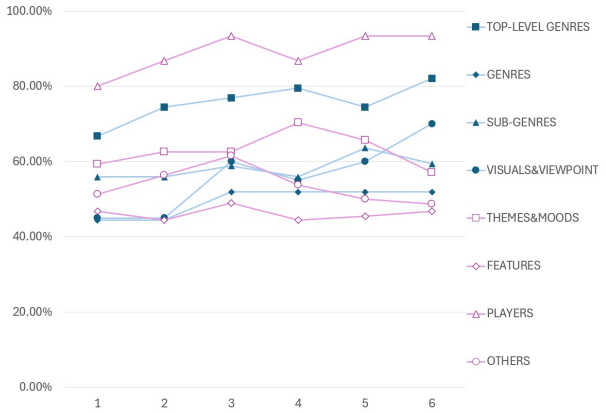

図 3: タグの属性ごとの正解率 (Gemini). 横軸はプロンプト番号を表す.

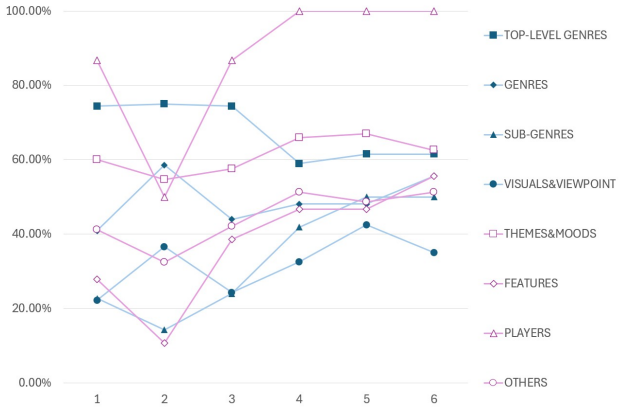

図 4: タグの属性ごとの正解率 (GPT). 横軸はプロンプト番号を表す.

4.3.1 Gemini での結果

まず Gemini ではどの区分においても 40%以上の正解率となった. 特にプレイヤー はすべてのプロンプトで 90% を超えており, 次にトップレベルのジャンルが高く, すべてのプロンプトで 60% を超えている. また各テーマ群において効果的なプロンプトとして, トップレベルのジャンルはプロンプト 1 からプロンプト 2 において正解率が最も向上したことから, ペルソナとして追加した文節がプロンプトとして効果的だったと考えられる. プレイヤーとその他においても同様にプロンプト 1 からプロンプト 2 で正解率が最も向上したが, 特徴では低下していた. 一方, ジャンル, ビジュアルと視点, 特徴, プレイヤー, その他 ではプロンプト 2 からプロンプト 3 が正解率の上昇が大きく, ゲーム紹介画像を入力したことが向上に寄与したとみなせる. プロンプト 3 で正解率が変化しなかったテーマと雰囲気 などもあるが, どの属性においても数値の低下はなかった.

興味深い結果として, この テーマと雰囲気 はプロンプト 3 からプロンプト 4 で正解率が大きく上昇しており, Chain-of-Though として追加した思考の提示が効果的だったと考えられる. このとき, プロンプト 3 からプロンプト 4 においては 8 個中 5 個の属性, すなわちサブジャンル, ビジュアルと視点, 特徴, プレイヤー, その他 において正解率が低下している. 類似した現象として, サブジャンル はプロンプト 4 からプロンプト 5 で正解率が最も向上しており, One-Shot Prompting として例を提示した効果があったものと考えられる. このとき, トップレベルのジャンル, テーマと雰囲気, その他では プロンプト 4 からプロンプト 5 で数値が低下している. また, ビジュアルと視点, プレイヤー ではプロンプト 4 からプロンプト 6 で高い正解率の向上が見られた. しかし, これらに関してはプロンプト 3 からプロンプト 4 にかけて低下した数値が回復しただけであり, Chain-of-Though による影響が例の提示によって打ち消されたとみなせる. なお, ほとんどのタグの区分において正解率の向上がみられたが, テーマと雰囲気, その他においては低下がみられた.

4.3.2 ChatGPT での結果

ChatGPT では 4.1 節と 4.2 節に共通して, 正解率の最低値が Gemini よりも低く, 20% 以下になるものもあった. 最も高いものはトップレベルのジャンル, テーマと雰囲気, プレイヤー などの区分であるが, これらですら 50% 程度となっている. また各テーマ群において効果的なプロンプトとして, ジャンル, ビジュアルと視点ではプロンプト 1 か

らプロンプト 2 が正解率の上昇が大きいことから, ペルソナとして追加した文節がプロンプトとして効果的だったと考えられる. しかし 8 種類のテーマ群のうち 5 種類, すなわちサブジャンル, テーマと雰囲気, 特徴, プレイヤー, その他がプロンプト 1 からプロンプト 2 で正解率が低下していた. 一方特徴, プレイヤー, その他ではプロンプト 2 からプロンプト 3 で最も正解率が向上しており, ゲーム紹介画像を入力したことが向上に寄与したとみなせる. 一方 3 種類のテーマ群, すなわちトップレベルのジャンル, ジャンル, ビジュアルと視点で正解率が低下した.

サブジャンル, テーマと雰囲気はプロンプト 3 からプロンプト 4 で正解率が最も向上しており, Chain-of-Though として追加した思考の提示が効果的だったと考えられる. このとき, トップレベルのジャンルではプロンプト 3 からプロンプト 4 で正解率が低下していた. またトップレベルのジャンルはプロンプト 4 からプロンプト 5, プロンプト 4 からプロンプト 6 で正解率が最も向上したことから, 例の提示が効果的だったと考えられる. プロンプト 4 からプロンプト 5 は 4 種類のテーマ群, すなわちトップレベルのジャンル, サブジャンル, ビジュアルと視点, テーマと雰囲気正解率が向上した一方, その他はプロンプト 4 からプロンプト 5 で正解率が低下した. またプロンプト 4 からプロンプト 6 は 5 種類のテーマ群, すなわちトップレベルのジャンル, ジャンル, サブジャンル, ビジュアルと視点, 特徴で正解率が向上した一方, テーマと雰囲気はプロンプト 4 からプロンプト 6 で正解率が低下した.

4.3.3 モデルごとの比較

上述した結果について, モデルごとにどのような違いがあるか比較して議論する. Gemini では, 多くのテーマ群に対しペルソナとゲーム紹介画像の入力が正解率の上昇に効果的だった. 具体的には, プロンプト 1 からプロンプト 2 ではすべてのテーマ群, プロンプト 2 からプロンプト 3 では 1 個を除く 7 種類のテーマ群において正解率が向上した. 一方, プロンプト 4 における Chain-of-Thought の追加では正解率の低下が目立った. とくに表 4 におけるタグの区分のうち, プロンプト 3 からプロンプト 4 では 5 種類の区分において低下がみられた.

ChatGPT では, 多くのテーマ群に対しプロンプト 4 における Chain-of-Thought とプロンプト 5 における One-Shot Prompting, および, プロンプト 6 における例の提示が正解率の上昇に効果的だった. 具体的には, プロンプト 3 からプロンプト 4, プロンプト 4 からプロンプト 5, プロンプト 4 からプロンプト 6 において 7 種類のテーマ群で正解率の向上があった. 一方, プロンプト 2 におけるペルソナの導入では正解率の低下が顕著だった. 具体的にはプロンプト 1 からプロンプト 2 で 5 種類のテーマ群において低下がみられている. とくにプロンプト 2 でのペルソナの導入は Gemini では正解率の向上に貢献しているが ChatGPT では低下を招いており, プロンプト 4 での Chain-of-Though は ChatGPT で正解率の向上に貢献しているが Gemini では低下を招いている. このことから, モデルの違いによりテーマの抽出に適したプロンプトが異なると考えられる.

4.4 制限事項

本稿における制限事項を以下に述べる. まず主な制限事項として, 本稿の分析ではテーマ抽出の成否と入力として与えられたゲーム紹介文の特性に関する対応を十分に評価できていない点が挙げられる. 具体的には, ゲーム紹介文中の語彙選択, 構文的明示性, 記述の具体性・冗長性といった, 一般的な文章に含まれるであろう要因が大規模言語モデルからの出力に影響すると考えられるが, それらの要因については考慮できていない. これらの観点も考慮した分析は今後の課題である.

5. 推奨される研究内容

本節では結果の考察として, 今後推奨される研究内容について述べる. まず本稿の結果から, LLM によってタグが抽出されやすいゲーム紹介文は, 人間にとっても特徴が把握しやすい紹介文である可能性が示唆された. 今後, タグが効果的に抽出された事例とそうでない事例を比較・分析することで, ユーザにとってより魅力的かつ理解しやすい紹介文の特性を明らかにできると考えられる. これにより, ゲーム紹介文の質を高め, ひいてはセキュリティ教育ゲームの受容や普及に資する知見を提供できるだろう.

さらに, 本研究はゲーム紹介文に関する研究の萌芽的試みとして位置づけられる. 従来, 数多くのゲームが制作されてきたにもかかわらず, それらをユーザに効果的にアピールする方法論は十分に検討されてこなかった. 本研究は, ユーザへの訴求力を高める紹介文の特性を明らかにするための第一歩であり, 今後の教育ゲームの設計および普及戦略における基盤的知見を提供するものである.

6. 結論

本稿では, 大規模言語モデル (LLM) を用いてビデオゲームのテーマを抽出する方法を提案するとともに, テーマ抽出に有効なプロンプトエンジニアリングの観点から検討した. とくに LLM にゲームの紹介ページに掲載される紹介文と紹介画像を入力することで, それらからテーマを生成することで, ゲームのテーマを効果的に抽出できるか明らかにした. 提案手法を用いて, Steam にて公開されているビデオゲームの紹介文を用いて複数の調査も行ったところ, 二つの知見を得ている. まず, LLM ごとに適切なテーマが抽出しやすいプロンプトが異なることを確認した. 次に, 複数の LLM でテーマを抽出することに適したゲームの紹介情報は希少と考えられる. 最後に, これらの知見を踏まえ, 今後推奨される研究も議論した. LLM によってタグが抽

出されやすいゲーム紹介文は, 人間にとっても特徴が把握しやすい紹介文である可能性が示唆されたが, 今後はその可能性が真であるか明らかにしていく.

謝辞 本稿は JSPS 科研費（課題番号 23H00479）および JST さきがけ事業 (課題番号 JPMJPR23P6) の支援を受けている.

実験データ 実験データは, 結果の再現性と更なる発展研究の促進のために, 読者の要求に応じて共有を検討する.

参考文献

[1] F. Alotaibi, S. Furnell, I. Stengel, and M. Papadaki. A review of using gaming technology for cyber-security awareness. *International Journal for Information Security Researc*, 6(2):660–666, 2016.

[2] M. Arai, N. Yanai, A. Inomata, and G. Hanaoka. An investigation method in video games for education in cybersecurity through large language models. In *Proc. of Cyberpsychology 2025*, pages 1–1. The British Cyberpsychology Society, 2021.

[3] L. Cheng, X. Li, and L. Bing. Is GPT-4 a good data analyst? In *Proc. of EMNLP 2023*, pages 9496–9514, Singapore, 2023. ACL.

[4] R. F. Chew, J. Bollenbacher, M. Wenger, J. Speer, and A. Kim. Llm-assisted content analysis: Using large language models to support deductive coding. *CoRR*, abs/2306.14924, 2023.

[5] S. De Paoli. Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review*, 42(4):997–1019, 2024.

[6] S. De Paoli and W. S. Mathis. Reflections on inductive thematic saturation as a potential metric for measuring the validity of an inductive thematic analysis with llms. *Quality & Quantity*, pages 1–27, 2024.

[7] M. S. Jalali and A. Akhavan. Integrating ai language models in qualitative research: Replicating interview data analysis with chatgpt. *System Dynamics Review*, 40:1–9, 2024.

[8] Z. Kilhoffer, Z. Zhou, F. Wang, F. Tamton, Y. Huang, P. Kim, T. Yeh, and Y. Wang. "how technical do you get? i'm an english teacher": Teaching and learning cybersecurity and ai ethics in high school. In *Proc. of IEEE S&P 2023*, pages 2032–2049. IEEE, 2023.

[9] D. King, P. Delfabbro, and M. Griffiths. Video game structural characteristics: A new psychological taxonomy. *International Journal of Mental Health and Addiction*, 8:90–106, 11 2009.

[10] G. Marvin, N. Hellen, D. Jjingo, and J. Nakatumba-Nabende. Prompt engineering in large language models. In *Proc. of ICDICI 2023*, AIS, pages 387–402. Springer, 2023.

[11] W. S. Mathis, S. Zhao, N. Pratt, J. Weleff, and S. D. Paoli. Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: How does it compare to traditional methods? *Computer Methods and Programs in Biomedicine*, 255:108356, 2024.

[12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[13] D. Saha, S. Tarek, K. Yahyaei, S. K. Saha, J. Zhou, M. Tehranipoor, and F. Farahmandi. Llm for soc security: A paradigm shift. *IEEE Access*, 12:155498–155521, 2024.

[14] M. Sjöblom, M. Törhönen, J. Hamari, and J. Macey. Content structure is king: An empirical study on gratifications, game genres and content type on twitch. *Computers in Human Behavior*, 73:161–171, 2017.

[15] S. Weitl-Harms, A. Spanier, J. Hastings, and M. Rokusek. A systematic mapping study on gamification applications for f cybersecurity education. *Journal of Cybersecurity Education, Research and Practice*, 2023(1), 2023.

[16] L. Zhang-Kennedy and S. Chiasson. A systematic review of multimedia tools for cybersecurity awareness and education. *ACM Computing Surveys*, 54(1), jan 2021.

[17] T. Zuo, M. V. Birk, E. D. Van der Spek, and J. Hu. The effect of fantasy on learning and recall of declarative knowledge in ar game-based learning. *Entertainment Computing*, 46:100563, 2023.

[18] 美. 新井, 直. 矢内, 敦. 猪俣, 悟. 花岡, A. Mine, Y. Naoto, I. Atsuo, and H. Goichiro. Sok：セキュリティ教育ビデオゲームにおけるゲームジャンルに関する体系的調査. **情報処理学会論文誌**, 65:1644–1681, 2024.