2025

2024 Vol.1
2024 Vol.2

White Paper Project

CONTENTS

2024 Vol.1

セールストークを対象とするエンゲージメントを考慮した目標指向対話データセット

邊土名 朝飛 Asahi Hentona
CyberAgent AI Lab
ML Engineer
hentona_asahi@cyberagent.co.jp

馬場 惇 Jun Baba
CyberAgent AI Lab
Research Scientist
baba_jun@cyberagent.co.jp

赤間 怜奈 Reina Akama
東北大学
Researcher
akama@tohoku.ac.jp

keywords: 対話コーパス、エンゲージメント、セールストーク、クラウドソーシング

Summary

本研究では、対話継続意欲、情報提供意欲、目標受容意欲の 3 種類のユーザーエンゲージメントを考慮しつつ対話目標を達成する対話タスクとして Engagement-motivated Goal-Oriented Dialogue (EGOD) を提案する。さらに、EGOD システムを実現に向けて、商品販売を対象としたセールス EGOD データセットを構築し、データセットの分析を行った。収集した 98 対話のうち、購買意欲が向上した対話は半数以上の 58 件を占めており、高品質なセールストークが収集できたと考えられる。データセットの分析の結果、対話目標である購買意欲向上のためには、エンゲージメントを向上させる発話を行うよりも、エンゲージメントを低下させる発話を避けることが重要であることが示唆された。

1. はじめに

エンゲージメントは、対話システムとの会話を続けるユーザの意欲を示す指標であり、主にオープンドメイン対話システムの評価において重要な指標である [Yu 16]。一方で、対話目標を持つ対話システムにおいては、迅速に目標を達成することが要求されているため、エンゲージメントはあまり考慮されていない [Siro 22]。しかしながら、実際に運用されている目標指向の対話システムにおいては、ユーザが対話途中で離脱するケースは多く、対話目標を達成するまでユーザとの対話を継続することが求められる。また、オープンドメイン対話と異なり対話目標が存在するため、対話を継続させると同時にユーザに対話目標を受け入れるよう促す必要がある。以上のことから、目標指向の対話システムにおいては、これらの観点を包括したエンゲージメントを考慮する必要があると考えられる。

本研究では、対話継続意欲、情報提供意欲、目標受容意欲の 3 種類のユーザーエンゲージメントを考慮しつつ対話目標を達成する対話タスクとして Engagement-motivated Goal-Oriented Dialogue (EGOD) を提案する。ここで、対話継続意欲とはユーザがシステムとの対話を続ける意欲、情報提供意欲とはユーザが自身の好みなどの情報をシステムに提供したいという意欲、目標受容意欲とはユーザがシステムの対話目標を受け入れたいと感じる意欲を指す。さらに本研究では、EGOD システムの実現に向けて、商品販売を対象としたセールス ETGD データセットを構築した。セールス EGOD タスクは、システムがユーザの購買意欲を向上させることを目標とした EGOD タスクである。

構築したセールス EGOD データセットの特徴として、以下の 3 点が挙げられる。第一に、ユーザの自然なエンゲージメントを可能な限り正確に計測するため、実験参加者であるユーザ役に対して、任意のタイミングで対話を離脱することを許可している点である。第二に、エンゲージメントの評価をユーザ役自身に行ってもらい、さらに対話レベルと発話レベルの両方で評価データを収集した点である。対話者であるユーザ役自身から、対話レベルだけではなく発話レベルの評価データも収集することで、実際のユーザに近い詳細な情報を獲得できると考えられる。第三に、対話システムのふりをしたセールス経験者 (セールス役) がユーザ役の対話相手となる Wizard of Oz (WOZ) 法 [Fraser 91] の設定で対話データを収集した点である。先行研究 [Hiraoka 16, Tiwari 23] では、主

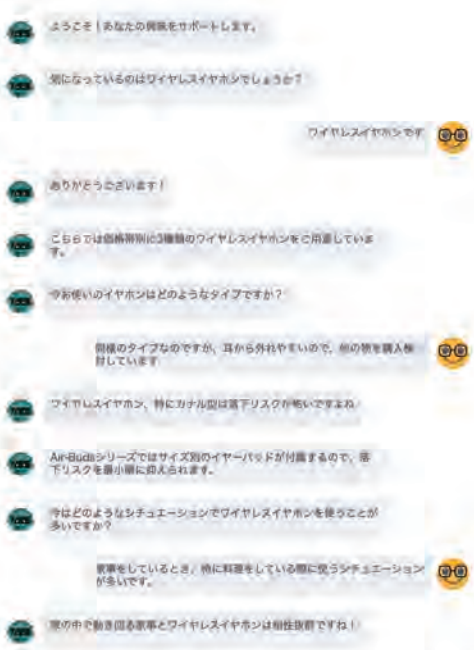


図 1: セールス対話の例（一部抜粋）

に人間同士の対話データが収集されているが、対話相手が人間かシステムかによって対話の特徴が変化することが示唆されている [Serban 18]。WOZ 法で収集した本研究の対話データセットは、実用的な対話システムの実現に寄与すると考えられる。

本稿では、セールス EGOD データセットの構築手順を説明し、データセットの特徴や分析結果について報告する。データセットの分析の結果、今回の対話目標である購買意欲向上のためには、エンゲージメントを向上させる発話を行うよりも、エンゲージメントを低下させる発話を避けることが重要であることが示唆された。

2. 関 連 研 究

セールストークのデータセット構築に取り組んだ研究として、Chiu らの SalesBot が挙げられる [Chiu 22]。Chiu らは、オープンドメインの対話から暗黙的なユーザの意図を検出し、タスク指向対話にスムーズに移行する対話データを自動的に作成するフレームワークを提案している。しかし、この研究ではユーザのエンゲージメントは考慮されておらず、オープンドメイン対話とタスク指向対話がスムーズに移行できているかを対話レベルで第三者評価するにとどまっている。

セールス経験者を対話実験に参加させ、人間同士のセールストーク対話を収集した研究もいくつか存在する。Hiraoka らは、カメラ販売を対象とした説得対話コーパスを構築した [Hiraoka 16]。この研究では、カメラの購入を検討している客と、特定のカメラを購入させることを意図した販売員との間で行われる説得対話を収集し、説得の成功と被説得者の満足度に影響を与える要因について分析を行っている。Tiwari らは、モバイル端末購入を対象とした大規模なパーソナライズド説得対話コーパスを構築した [Tiwari 23]。この研究では、客が要求する仕様を満たす商品がない状況で、類似する製品を購入するよう説得する説得対話タスクに取り組んでいる。しかしながら、これらの研究は人間同士の対話を収集したものである。人間同士の対話と人間対システムの対話では得られる対話データの特徴が異なる [Serban 18] ため、実際に対話システムを運用する環境に適用することは難しい。また、これらの実験設定では、購入を希望する商品もしくは推薦する商品が事前に設定されており、購入意欲があまり高くないユーザと対話して意欲を高めるといった状況は想定されていない。

これらの研究に対して、本研究では、セールス対話システムの実際の運用状況に近い設定で対話データを収集し、高品質なセールス EGOD データセットを構築することを試みる。

3. セールストーク対話の収集

本研究では、3 種類の架空のワイヤレスイヤホンの情報を掲載した Web ページに訪れたユーザと、その Web ページ上に設置されたセールス対話システムがテキストチャットを行うシナリオで対話データを収集した。収集した対話データの例を図 1 に示す。対話設定として WOZ 法 [Fraser 91] を採用し、ユーザ役のクラウドワーカーと、セールス対話システムのふりをしたセールス経験者 (セールス役) との間で対話を実施した。ユーザ役は、セールス対話システムを相手にしていると思いながら対話を行うため、実際のシステム運用状況に近い人間対システムの対話データが収集できると考えられる。実験終了後、ユーザ役の実験参加者に対してデブリーフィングを行い、対話相手がシステムではなく人間であることを通知した。対話時間は最長 30 分とし、ユーザ役は任意のタイミングで対話を終了してもよいことを事前に説明した。対話収集の実施にあたり、実験参加者に対する報酬として 2023 年 11 月 2 日時点の東京都の最低賃金を上回る時給 1500 円を稼働時間に応じて支払った。本研究のデータ収集手続きは、株式会社サイバーエージェント研究倫理審査委員会の審査を受け、承認を得た *1。

3·1 セ ー ル ス 役

セールス役は、日本語による円滑なテキスト対話が可能な営業経験者を対象として募集を行い、合計 5 名を採用した。実験に際しては、セールス役の対話品質向上のため、対話収集実験の前に実験ガイダンスとトークスクリプト作成を実施し、その上で対話練習を 2 回実施した。実験ガイダンスでは、対話相手となるユーザ役には「人

*1 承認番号：CAE-2023-06

間ではなくシステムと対話する実験である」と通知していることを説明した。さらに、購買意欲とエンゲージメント（対話継続意欲、情報提供意欲、目標受容意欲）の観点からユーザに対話を評価されることを事前に説明し、その上で購買意欲とエンゲージメントを可能な限り高める対話を行うよう指示した。

3·2 ユーザ役

ユーザ役は、日本語で円滑にテキスト対話ができる者を対象として募集を行い、合計 98 名が対話実験に参加した。WOZ 法を用いて対話データを収集する関係上、ユーザ役には「セールス対話システムとテキストチャットを行う実験である」と説明した。なお、実験終了後にユーザ役の実験参加者に対してデブリーフィングを行い、対話相手がシステムではなく人間であることを通知した。また、ユーザ役の自然な対話継続意欲を観測するために、任意のタイミングで対話から離脱してもよいことを実験前に説明した。

3·3 商品

サンプル商品として 3 種類の架空のワイヤレスイヤホンの商品情報を作成した。ワイヤレスイヤホンは商品の特性上、ある程度の専門性が要求される。専門知識が必要な商品の場合、ユーザは説得、すなわちセールストークの影響を受けやすくなるため、セールストークを対象とする本研究に適した商品と考えられる [Wilson 93]。さらに、ワイヤレスイヤホンは購買層が広く、年齢、性別、業種等による影響が少ない商品であり、かつユーザのコンプレックスに関わる商品ではないため、対話中に実験参加者のプライベートな内容が含まれる危険性は低い。商品価格については、低価格過ぎる商品の場合、ユーザのエンゲージメントを高めなくても商品が購入される可能性がある一方で、高価格過ぎる商品の場合、ユーザのエンゲージメントを高めても購買意欲が向上しない可能性が考えられる。そのため、高価格もしくは低価格すぎない価格帯と思われる 11,000 円、22,000 円、33,000 円を 3 種類のワイヤレスイヤホンそれぞれに設定した。

3·4 実験環境

実験で使用したチャットツールのインターフェースを図 2 に示す。テキストチャットツールは、Slurk[Götze 22] をベースに構築したシステムを、クラウド (Google Cloud Platform) 上にデプロイしたものを使用した。実験参加者がチャットツールの URL にアクセスすると、実験参加者同士のマッチングが行われ、ユーザ役とセールス役のペアが揃った時点でチャットルームが自動的に作成されて対話実験が開始される。本チャットツールには、基本的なテキストチャット機能の他、商品情報の表示および共有、対話終了などの機能が実装されている。

図 2: チャットツールのインターフェース

4. アノテーション

4·1 発話レベルのアノテーション

ユーザ役は、対話終了後にセールス役の各発話に対してエンゲージメントのアノテーションを行うよう指示した。各アノテーション項目とユーザ役に提示した質問文を表 1 に示す。各エンゲージメントは、それぞれ 3 段階（Positive: ポジティブな影響を与えた、Neutral: 全く影響がなかった、Negative: ネガティブな影響を与えた）で評価してもらった。

セールス役は、対話終了後にユーザ役のエンゲージメントを高める意図の発話を自分自身でアノテーションするよう指示した。このアノテーションにより、セールス役がどの程度エンゲージメントを向上させる発話を行ったのか、またそれらの発話がユーザ役の評価にどの程度影響を与えたかを計測することが可能になる。

表 1: セールス役の発話に対するアノテーション項目および各項目の質問文

項目	質問文
対話継続意欲	この発話は少しでも会話を続けたい気持ちに影響を与えましたか？
情報提供意欲	この発話は少しでも情報を提供したい気持ちに影響を与えましたか？
目標受容意欲	この発話は少しでも商品を買いたい気持ちに影響を与えましたか？

4·2 対話レベルのアノテーション

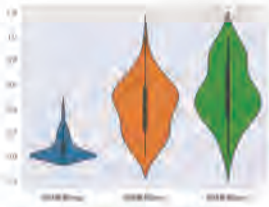
対話目標であるユーザ役の購買意欲は、対話実験の前後に実施した購買意欲アンケートにより測定した。これにより、対話実験前後の購買意欲の変化を計測することで、対話目標達成の度合いを評価することができる。

5. データセット分析

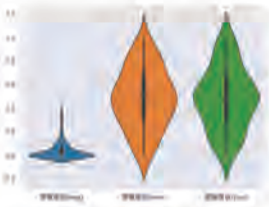
表 2 に収集した対話データセットの統計量を示す *2。収集した 98 対話のうち、購買意欲が向上した対話は半数以上の 58 件を占めており、また平均対話時間は 18.7 分とすぐに対話離脱するユーザ役が少なかったことから、ある程度高品質なセールストークが収集できたと考えられる。次に、1 対話中に含まれる各種エンゲージメントラベルの割合を図 3 に示す。図 3 からは、Negative ラベルがほとんど付与されず、付与された大半のラベルが Neutral、Positive であることが分かる。また、各エンゲージメントの Positive ラベル、Neutral ラベルの割合から、対話継続意欲、情報提供意欲、目標受容意欲の順にエンゲージメントを向上させやすいと考えられる。

表 2: セールス EGOD データセットの統計情報。平均は 1 対話あたりの値。

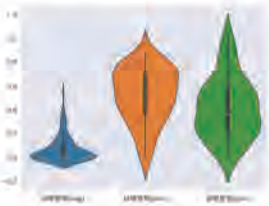
	合計	平均	最大	最小
対話数	98	-	-	-
発話数	3090	31.5	92	11
発話数 (ユーザ)	1063	10.8	41	3
発話数 (セールス)	2027	20.7	52	7
トークン数	49610	506.2	1405	153
語彙数	17687	-	-	-
対話時間 (分)	1828.1	18.7	36.5	5.9
目標達成対話数	58	-	-	-



(a) 対話継続意欲



(b) 情報提供意欲



(c) 目標受容意欲

図 3: 1 対話中の各エンゲージメントラベルの割合

次に、ユーザ役アンケートから得られた事前・事後の購買意欲と、対話中のエンゲージメント評価のデータから、各種エンゲージメントが対話目標である購買意欲にどの程度影響を及ぼしているかを分析した。まず、各対話における各種エンゲージメント評価ラベルの割合と購買意欲の相関を分析した。図 4 に相関ヒートマップを示す。エンゲージメント間の相関を見ると、Positive ラベルは Negative ラベルよりも Neutral ラベルに対して強い負の相関があることが分かる。一方、Neutral ラベルは、Negative ラベルとは弱い正の相関、またはほとんど相関がないことが分かる。このことから、ユーザの評価行動は、Positive/Negative 評価と Neutral/Negative 評価に二分されていると考えられる。続いて、各種エンゲージメントと購買意欲の相関を分析した。エンゲージメントを向上させた Positive な発話の割合は、購買意欲の増加とほぼ相関が見られなかった。一方で、エンゲージメントを低下させた Negative な発話と購買意欲は、やや負の相関が見られた。これらの結果から、対話目標である購買意欲増加を達成するためには、エンゲージメントを向上させる発話を行うことではなく、エンゲージメントを低下させる発話を避けることが重要であることが示唆される。

次に、エンゲージメントラベルを変更した場合の購買意欲の予測性能を調査した。実験では、対話中に含まれる各種エンゲージメントラベルの割合を特徴量として、購買意欲が向上するかどうかを予測する二値分類モデル (ロジスティック回帰モデル) を構築した。モデルの評価指標として、Accuracy、Precision、Recall、F1-score を採用し、5-fold cross validation の各評価指標の平均値を算出した。評価結果を表 3 に示す。表 3 に記載の全てのモデルの Accuracy は、全て購買意欲が向上したと予測した場合のチャンスレート (Accuracy 58%) よりも高い。また、全てのエンゲージメントを特徴量として用いる (All) よりも、1 種類もしくは 2 種類のエンゲージメントを特徴量として用いる場合の方が、予測性能が向上していることが分かる。F 値に注目すると、性能上位 3 件のモデルには全て情報提供意欲が含まれている。このことから、ユーザの発話から情報提供意欲を予測し対話戦略に活用することで、購買意欲を向上させる対話を行うことが可能であることが示唆される。

表 3: ロジスティック回帰モデルを用いた購買意欲増減予測結果

	Accuracy	F 値	Precision	Recall
目標受容	67.2	76.6	67.0	89.7
対話継続+目標受容	66.1	76.2	66.7	89.5
対話継続+情報提供	67.1	76.6	67.4	89.5
対話継続	66.0	75.8	66.9	87.9
情報提供+目標受容	70.2	78.2	69.7	89.5
情報提供	69.2	78.2	68.5	91.4
All(Positive のみ)	62.9	74.2	63.6	89.8
All(Negative のみ)	69.1	75.2	73.4	77.6
All	69.1	77.9	68.2	91.2

*2 トークン分割には日本語形態素解析機 MeCab を用いた

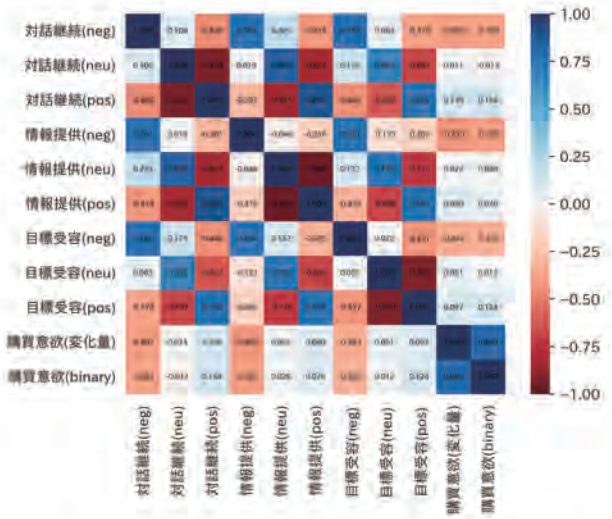


図 4: エンゲージメントおよび購買意欲間の相関

6. おわりに

本研究では、対話継続意欲、情報提供意欲、目標受容意欲の 3 種類のユーザーエンゲージメントを考慮しつつ対話目標を達成する対話タスクとして Engagement-motivated Goal-Oriented Dialogue (EGOD) を提案し、商品販売を対象としたセールス ETGD データセットを構築した。データセットの分析の結果、収集した 98 対話中 58 対話で購買意欲が向上していたこと、対話平均時間が 18.7 分ですぐに対話から離脱したユーザ役が少なかったことから、高品質なセールストークの対話データが収集できたと考えられる。また、対話目標の達成率向上のためには、エンゲージメントを向上させる発話を行うよりも、エンゲージメントを低下させる発話を避けることが重要であることが示唆された。今後の展望として、収集したデータセットを用いて大規模言語モデルを学習させ、購買意欲を向上させるセールストークをどの程度実行できるのかを検証することが挙げられる。

◇参考文献◇

[Chiu 22] Chiu, S., Li, M., Lin, Y.-T., and Chen, Y.-N.: SalesBot: Transitioning from Chit-Chat to Task-Oriented Dialogues, in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 6143–6158 (2022)

[Fraser 91] Fraser, N. M. and Gilbert, G. N.: Simulating speech systems, *Computer Speech & Language*, Vol. 5, No. 1, pp. 81–99 (1991)

[Götze 22] Götze, J., Paetzel-Prüsmann, M., Liermann, W., Diekmann, T., and Schlangen, D.: The slurk Interaction Server Framework: Better Data for Better Dialog Models, in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 4069–4078 (2022)

[Hiraoka 16] Hiraoka, T., Neubig, G., Sakti, S., Toda, T., and Nakamura, S.: Construction and Analysis of a Persuasive Dialogue Corpus, in Rudnicky, A., Raux, A., Lane, I., and Misu, T. eds., *Situated Dialog in Speech-Based Human-Computer Interaction*, pp. 125–138 (2016)

[Serban 18] Serban, I. V., Lowe, R., Henderson, P., Charlin, L., and Pineau, J.: A Survey of Available Corpora For Building Data-Driven Dialogue Systems, *Dialogue & Discourse*, Vol. 9, No. 1, pp. 1–49 (2018)

[Siro 22] Siro, C., Aliannejadi, M., and Rijke, de M.: Understanding User Satisfaction with Task-oriented Dialogue Systems, in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, p. 2018–2023 (2022)

[Tiwari 23] Tiwari, A., Khandwe, A., Saha, S., Ramnani, R., Maitra, A., and Sengupta, S.: Towards personalized persuasive dialogue generation for adversarial task oriented dialogue setting, *Expert Systems with Applications*, Vol. 213, p. 118775 (2023)

[Wilson 93] Wilson, E. J. and Sherrell, D. L.: Source effects in communication and persuasion research: A meta-analysis of effect size, *Journal of the Academy of Marketing Science*, Vol. 21, No. 2, p. 101–112 (1993)

[Yu 16] Yu, Z., Nicolich-Henkin, L., Black, A. W., and Rudnicky, A.: A Wizard-of-Oz Study on A Non-Task-Oriented Dialog Systems That Reacts to User Engagement, in *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 55–63 (2016)

著者紹介

邊土名 朝飛

2021 年長岡技術科学大学大学院工学研究科修士課程修了後、サイバーエージェント入社。AI Lab および AI Shift にて対話システムの研究開発に従事。

Source-Free Domain Adaptation with Class Distribution Shift via Generic Features

Leveraging features free of the source bias for robust nearest neighbors

テヘーロ デ パブロス アントニオ
Antonio Tejero-de-Pablos
CyberAgent - AI 事業本部 - AI Lab
Research Scientist
antonio_tejero@cyberagent.co.jp, https://cyberagent.ai/ailab/people/a_tejero/

富樫 陸
Riku Togashi
CyberAgent - AI 事業本部 - AI Lab
Research Scientist
togashi_riku@cyberagent.co.jp, https://cyberagent.ai/ailab/people/rtogashi/

大谷 まゆ
Mayu Otani
CyberAgent - AI 事業本部 - AI Lab
Research Scientist
otani_mayu@cyberagent.co.jp, https://cyberagent.ai/ailab/people/mayu-otani/

佐藤 真
Shin'ichi Satoh
CyberAgent - AI 事業本部 - AI Lab - 産学連携
Professor
satoh@nii.ac.jp, https://cyberagent.ai/ailab/academic-relations/sato_s/

keywords: Domain adaptation, source-free, class distribution shift, robust nearest neighbors, generic features

Summary

Source-free domain adaptation (SFDA) retrains a model fit on data from a source domain (e.g. drawings) to classify data from a target domain (e.g. photos) employing only the target samples. In addition to the domain shift, in a realistic scenario, the number of samples per class on source and target would also differ (i.e. class distribution shift, or CDS). By studying the SFDA pipeline, whose core is nearest neighbors (NN)-based pseudolabeling, we identify for the first time its sensibility to CDS, and propose a method to obtain robust NN. Since the class distribution of the target samples cannot be estimated via the source model, we leverage an external generic model that provides a "second-opinion" for calculating NN. Our method outperforms previous works in several datasets and tasks.

1. Introduction

After a deep neural network model is deployed, it normally finds data whose distribution is slightly shifted from that of the training data. This "domain shift" (also referred as *covariate shift*) worsens the performance of the model and limits its practical use. The research field of domain adaptation approaches this problem in order to make deep models more robust against label-less unseen data. Several domain adaptation scenarios have been proposed over the years, being unsupervised domain adaptation (UDA) the most basic [Cao 18, Saito 18, Chen 19]. Then, source-free domain adaptation (SFDA) methods were proposed [Yang 21, Dong 21, Litrico 23] to adapt the source model without accessing the source data.

Since source data is not accessible, real scenarios unknowingly present a class distribution shift (CDS) between both domains. That is, the ratio of samples for each class is significantly different in the source and target domains. If not dealt with properly, the model becomes sensitive to such imbalance, and its predictions become biased to the majority class in the source domain. This is a very challenging scenario, as both *covariate* and *class distribution* shifts need to be tackled without source data nor labels.

Most SFDA methods consist of a pseudolabeling pipeline, in which the target data is fed to the source model and assigned the most plausible labels. Pseudolabeling can be performed at either the feature level (i.e. observing the nearest neighbors of a given sample) or at the logits level (i.e. observing the predicted probabilities). Previous work approached the SFDA under CDS scenario as a problem of noise in the logits output by the source model [Li 21]. In order to avoid pseudolabels to be biased towards majority classes in the source data, they opt to modify the class distribution of the source data to be uniform. Although accessing the source data this way is against the

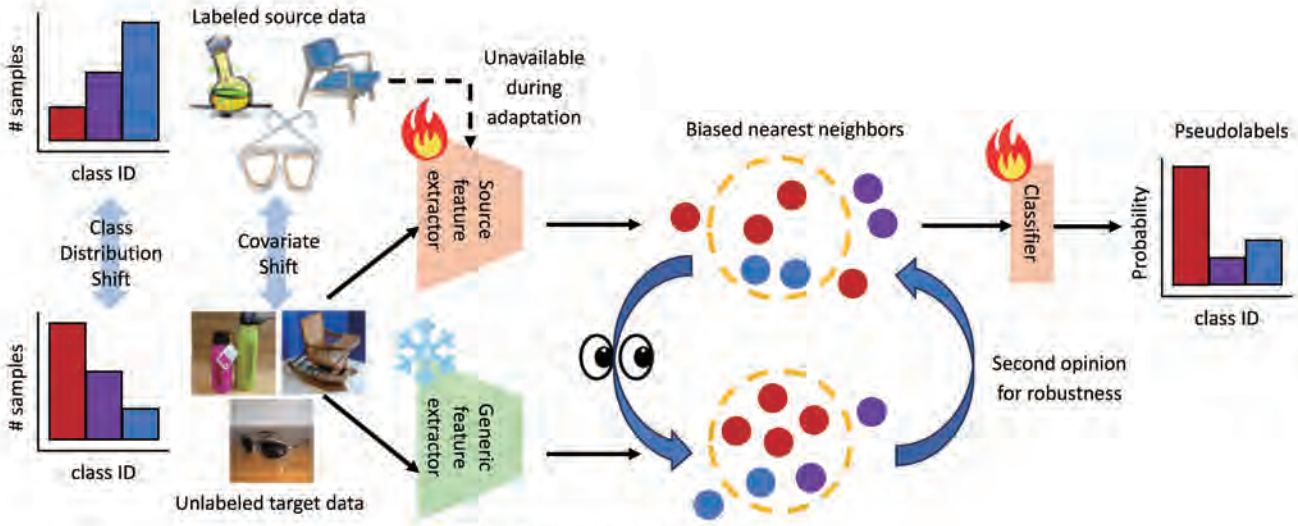


図 1 Overview of the proposed method

strict SFDA scenario, they had to control the distribution of the source data since the CDS cannot be estimated by the source model alone. However, we hypothesize that this problem can still be solved without accessing the source data. First, instead of addressing pseudolabeling at the logits level, we believe that a more fundamental problem lies in the previous step, the nearest neighbors calculation. Then, even though source and target data distributions are unknown during adaptation, a model unrelated to the source data can provide a "second opinion" on the target data without the bias. Distilling generic knowledge from powerful feature extractors was previously proposed in SFDA scenarios [Zara 23, Zhang 23], but neither in the CDS setting nor at the nearest neighbors level.

Figure 図 1 depicts the main idea of our proposal for SFDA under CDS. Unlike the previous work, we approach the majority/minority class bias at the nearest neighbors level by introducing an additional model to the SFDA pipeline. We opt for leveraging well-known pretrained models (i.e. ResNet, VisionTransformer, SwinTransformer) as they are widely available and have proved the generalizability of their features [Chen 21]. We regulate the nearest neighbors extracted in the source model's feature space by comparing them to the nearest neighbors in the generic feature space. We consider the neighbors are "robust" if they appear in both the source and generic models. In the teacher-student knowledge distillation fashion [Hinton 15], the generic model remains frozen as the source model is adapted by taking generic features as a reference. In summary, our contributions are:

- We study the problem of SFDA under CDS, and identify a weakness in the nearest neighbors (NN) calculation, which was unnoticed by previous works.
- We propose a method for reducing bias in the NN by introducing generic features of an auxiliary model.
- Our straightforward approach outperforms previous methods for both SFDA and TTA settings in several benchmarks under CDS, and for the first time in SFDA, without manipulating the training of the source model.

2. Methodology

Let's consider a source domain $\{x_s \in X_s, y_s \in Y_s\}$ and a target domain $\{x_t \in X_t, y_t \in Y_t\}$ with the same label space and a covariate shift (Fig. 図 1). The source model consisting of a pretrained feature extractor f_s and classifier h_s is trained on the source dataset in an unknown manner, and only the source model is provided, along with the unlabeled target samples x_t. For pseudolabeling, the target features are extracted $Z_s = f_s(X_s)$ and the k nearest neighbors $N = \{z_s^1, ..., z_s^k\}$ are calculated.

2·1 Effect of CDS on the nearest neighbors

While nearest neighbors (NN) for pseudolabeling is effective against SFDA's covariate shift [Yang 21, Dong 21, Litrico 23], the effect of simultaneously dealing with class distribution shift has not been studied yet. Previous works [Weiss 01, Shi 20] proved that the NN algorithm is significantly impacted by imbalanced class distributions, due to its sensitivity to the local data structure. In the labeled single-domain scenario the majority/minority bias can be mitigated via undersampling/oversampling techniques, or a weighted classification algorithm [Shi 20]. However, in the SFDA-CDS scenario, when a certain sample is misclassified as a majority class, it can be because of the bias of the model trained in the source data or because the bias

within the target data itself.

Providing a full theoretical proof of the impact of CDS on SFDA is complex and depends on specific assumptions about the data distribution. Therefore, we provide an empirical proof to illustrate this phenomenon. We trained a baseline classifier (i.e. ImageNet-pretrained ResNet101 [He 16]) in the source domain of the SFDA dataset VisDA-C [Peng 17] and applied the basic pseudolabeling pipeline to adapt the source model to the target domain. This pseudolabeling is used in both traditional and state-of-the-art SFDA methods (e.g. guided pseudolabels [Litrico 23]). Although we accessed the ground truth labels $y_t \in \{$plane, ..., truck$\}$ for observation, they are unavailable during the actual adaptation. Figure 図 2 (a) displays a histogram of the classes of the $k = 10$ neighbors N used to calculate the pseudolabel of an input target sample with label $y_t^i =$*train*. The majority of the neighbors belong to the correct class, and consequently $y_t'^i =$ *train* is predicted. The accuracy of the adapted model is 90.0%.

Next, we use the CDS version of the dataset, VisDA-C RSUT [Tan 20] (Fig. 図 3). For the same target sample, Fig. 図 2 (b) shows that the NN in the source model contain more samples from unrelated classes, including those with majority representation in the source domain (i.e. *bicycle*). As a result, the accuracy of the adapted model decreases to 83.59%. This bias noise is inherent to the CDS setting, but unfeasible to estimate without either source or target labels, as the source model predictions are influenced by both. As a reasonable way to consider an unbiased estimation of the class distribution of the target domain, we propose relying on an external reference model.

2·2 Generic features for robust nearest neighbors

We propose leveraging an additional feature extractor g free of the majority/minority bias of the source data. As an initial experiment, we use the same backbone as the source model before seeing the source data, i.e. ResNet101 pretrained on ImageNet-1K, which provides a set of features $g(x_t^i) = z_g^i \in Z_g$. As a multipurpose public dataset, ImageNet is less biased and more *generic* at least than the source data. Thus, while the feature space Z_g lacks bias in favor of generalizability, the feature space Z_s of the source model suffers from source bias but possesses domain knowledge (e.g. the label space). We hypothesize that combining both can lead to more "robust" nearest neighbors.

We propose replacing the set of source neighbors N with robust neighbors R that are present in both Z_s and Z_g. This requires looking further than the original $k = 10$ samples, so we introduce an additional hyperparameter $K = 100$. Therefore, $R = \{z_s^1, ..., z_s^K\} \cap \{z_g^1, ..., z_g^K\} = \{z_r^1, ..., z_r^o\}$. To validate our hypothesis, we consider a conservative setting prioritizing the source neighbors, so that:

$$\text{If } o == k, N \leftarrow R \tag{1}$$

$$\text{If } o > k, N \leftarrow \{z_r^1, ..., z_r^k\} \tag{2}$$

$$\text{If } o < k, N \leftarrow \{z_s^1, ..., z_s^{k-o}\} \cup \{z_r^1, ..., z_r^o\} \tag{3}$$

Whereas the feature bank F needs to be updated on each training iteration [Litrico 23], the bank of the generic features G only needs to be created once at the beginning.

Figure 図 2 (c) and (d) show the K NN in the source and generic feature spaces respectively, and (e) shows the set R of robust NN, which contains less bias noise. The final accuracy after adaptation is 86.6%, which means that pseudolabeling by employing NN of the correct class leads to higher classification accuracy.

§ 1 Source, generic and robust neighbors

Figure 図 4 compares the percentage of correct neighbors, i.e. those belonging to the same class as y_t^i, as the source model is adapted. While the generic neighbors do not improve during adaptation, the source neighbors improve getting closer to the performance of the generic neighbors. This results in robust neighbors that surpass the performance of both the generic model and the original guided pseudolabels in [Litrico 23].

3. Experimental results

3·1 Experimental settings

Metrics. As in the related work [Li 21], we calculate the class-wise mean accuracy, as it indicates if all classes (majorities and minorities) are properly classified.

Datasets. We evaluate our method in three standard SFDA datasets. **VisDA-C** [Peng 17] contains twelve object classes in two domains: *real* and *synthetic*. The source domain consists of computer generated *synthetic* images, while the target domain are *real* world photos. Although there are two domains, only the `synthetic` → `real` setting is considered. **Office-Home** [Venkateswara 17] contains sixty-five object classes in four domains, from which three are used as a benchmark in the relevant works [Li 21, Park 23]: *clipart*, *product* and *real*, which results in six different adaptation combinations. **DomainNet** [Peng 19] contains forty object classes in four domains: *clipart*, *painting*, *real* and *sketch*. The subset of DomainNet created by [Tan 20], also called DomainNet mini by [Li 21], contains four domains (*clipart*, *painting*, *real*, *sketch*). This dataset provides a train, validation and test splits for each domain. Following the related work, we use the RSUT

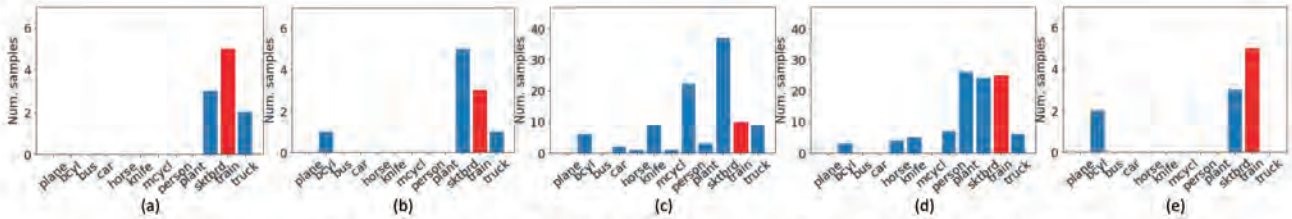


図 2 Class histogram of the nearest neighbors (NN) in SFDA pseudolabeling given a target sample of the class *train* (in red) on the VisDA-C dataset. (a) Source NN on VisDA-C ($k = 10$), (b) Source NN on VisDA-C RSUT ($k = 10$), (c) Source NN on VisDA-C RSUT ($K = 100$), (d) Generic NN on VisDA-C RSUT ($K = 100$), (e) Robust NN on VisDA-C RSUT ($k = 10$).

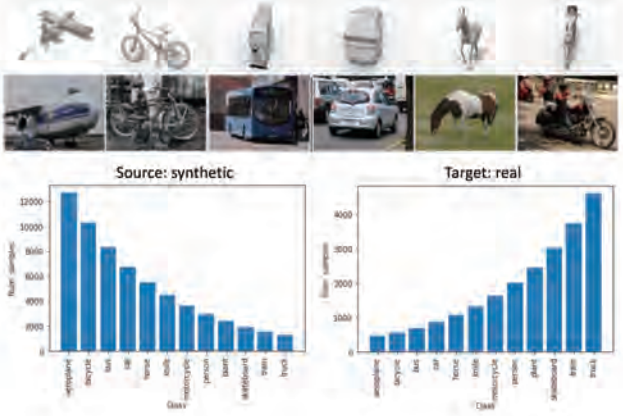


図 3 Class distribution in the VisDA-C RSUT dataset.

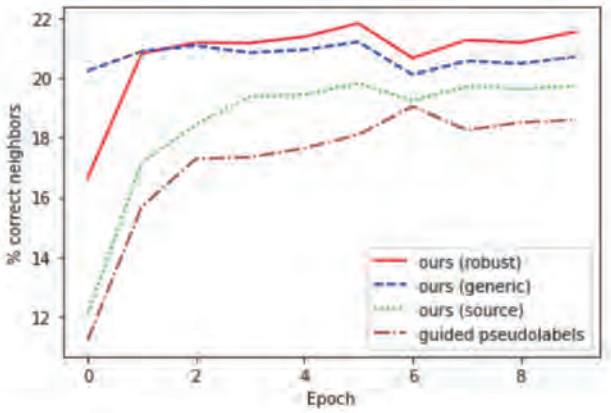


図 4 Percentage of correct NN used for pseudolabeling during the adaptation of the source model on VisDA RSUT.

(Reversely-unbalanced Source and Unbalanced Target) version of VisDA-C and Office-Home (see Fig. 図 3), created by [Tan 20] in order to apply CDS. It follows a Pareto distribution, which represents the worst case scenario for a class distribution shift between the source and the target domains. The source model is trained on the RS split of the source domain (20% of the data for validation [Litrico 23]) and adapted on the UT split of the target domain. The adaptation accuracy is calculated on a 20% of the UT split of the target domain. On the other hand, the domains in DomainNet are naturally class distribution shifted, so no class resampling is applied, but a subset of the dataset is used instead [Tan 20]. As a result, each dataset represents a different type of CDS.

Source/Adapted Model. For evaluation, we introduce our robust nearest neighbors into the SFDA state-of-the-art, guided pseudolabels (**PL guided**) [Litrico 23]. This method calculates pseudolabels for the target data via nearest neighbors, weights them according to the uncertainty of their predictions, and refines the feature space via contrastive learning. Additionally, as a reference, we provide the results on the pseudolabeling baseline (**PL base**), i.e. no weighting nor contrastive learning. Following the standard setting of the benchmark in [Litrico 23], the architecture used for the source models is: ResNet101 for VisDA-C and ResNet50 for Office-Home and DomainNet, pretrained in ImageNet-1K. The source model training and adaptation regimes are those of the original setting [Litrico 23]; stochastic gradient descent is run during 100 epochs and batch size 64 in VisDA-C (RSUT), and 200 epoch and batch size 128 in Office-Home (RSUT) and DomainNet (subset). Likewise, the source data is learned with the standard cross-entropy loss and label-smoothing.

Generic models. We employ three different backbones in our experiments. We use the same pretrained **ResNet** [He 16] backbones as the source/adapted model for each dataset. In addition, we employ the **Vision Transformer** (ViT-B/32) [Dosovitskiy 20] and the **Swin Transformer** (Swin-B/4) [Liu 21]. ResNet-101 is used in VisDA-C, and ResNet-50 is used in Office-Home and DomainNet. We use the pretrained models available in `torchvision`, which contain the training weights of ImageNet-1K. For Vision Transformer, we employ the image encoder provided by `huggingface transformers`. Specifically, we set a ViT-B/32 architecture pretrained on the model weights of `openai/clip-vit-base-patch32`. Finally, Swin Transformer uses also the base model available in `huggingface`, which is pretrained on ImageNet-21K at resolution 224×224 (`microsoft/swin-base-patch4-window7-224-in22k`).

表 1 Class-wise average accuracy of SFDA on VisDA-C RSUT, Office-Home RSUT, and DomainNet.

Method	VisDA-C	Office-Home	DomainNet
Source model	43.55	52.81	62.52
PADA [Cao 18]	42.06	42.66	64.48
MCD [Saito 18]	58.45	45.94	65.42
BSP [Chen 19]	46.15	40.97	74.09
COAL [Tan 20]	60.05	58.40	75.89
MDD (Implicit) [Jiang 20]	72.03	61.67	77.33
SHOT [Liang 20]	61.59	62.84	78.14
ISFDA [Li 21]	76.69	65.36	79.58
PL base	81.01	56.82	79.48
+ Ours (ResNet)	83.85	55.47	70.28
+ Ours (ViT-B)	83.88	58.71	82.51
+ Ours (Swin-B)	86.64	64.64	78.95
PL guided [Litrico 23]	83.59	61.05	80.12
+ Ours (ResNet)	86.6	59.67	72.74
+ Ours (ViT-B)	86.72	62.31	**83.9**
+ Ours (Swin-B)	**88.84**	**69.04**	81.4

Robust nearest neighbors are calculated with hyperparameters $k = 10$ and $K = 100$.

3·2 Source-free domain adaptation under CDS

Table 表 1 shows the classification accuracy after adaptation on the target domain (averaged for all domain combinations). The results are divided in seven blocks: (1) the source model without adaptation, (2) UDA methods, (3) UDA under CDS methods, (4) SFDA methods, (5) SFDA under CDS method, (6) SFDA via basic pseudolabeling with our proposal, and (7) SFDA via guided pseudolabeling with our proposal. The best performance is **bolded** and the second-best is underlined.

Similar patterns can be observed for all three datasets. Our method provides the best results when leveraging a strong generic feature extractor, outperforming the previous work in SFDA under CDS [Li 21] without needing to impose a uniform distribution on the source data. The reason is that, although approaching bias reduction at the logits level can correct the source model's predictions partially, the performance improvement is limited compared to reducing bias at the nearest neighbors level. As a result, by empirically exploring the essence of the problem, we successfully provided a solution to the most challenging SFDA-CDS setting for the first time, via a simple method. In particular, Swin-B provides the best results in two of the three datasets. Unlike ViT-B, Swin-B model extracts features at different local and global levels, which makes it more robust against covariate shifts [Zhang 23]. Regarding ResNet, while it is outperformed by the other generic models, it can provide comparable performance when the target domain are real images. In particular, given the similarity between the target domain in VisDA-C and ImageNet, ResNet's robust neighbors are as effective as the stronger generic models. Moreover, with our method, the simpler base pseudolabeling can surpass the original performance of the more complex guided pseudolabels method [Litrico 23] (e.g. `PL base + Swin-B` vs. `PL guided` in VISDA-C and Office-Home).

表 2 Class-wise average accuracy of TTA on VisDA-C RSUT, Office-Home RSUT, and DomainNet.

Method	VisDA-C	Office-Home	DomainNet
Source model	51.45	49.39	64.26
ONDA [Mancini 18]	50.68	49.02	67.89
LAME [Boudiaf 22]	50.72	47.46	62.20
CoTTA [Wang 22]	49.88	48.92	67.45
NOTE [Gong 22]	49.37	48.58	69.01
TENT [Wang 21]	48.68	51.15	70.34
+ Shift adapter [Park 23]	**72.97**	52.78	71.63
Pseudolabel	47.12	52.34	67.06
+ Ours (ResNet)	50.07	52.83	63.01
+ Ours (ViT-B)	49.60	53.95	**73.23**
+ Ours (Swin-B)	52.49	**60.16**	70.59

3·3 Test-time adaptation under CDS

The nature of our method also allows it to improve the adaptation without retraining the source model, which suits the time-test adaptation (TTA) setting. Thus, we run our method in inference mode, i.e. relying only on the predicted pseudolabels of the robust nearest neighbors. Note that, since no learning is involved, the accuracy results are the same for both the base and guided pseudolabels.

Table 表 2 shows the classification accuracy on the target domain (averaged for all domain combinations). The results are divided in four blocks: (1) the source model without adaptation, (2) TTA methods with partial support to CDS, (3) TTA method with full support to CDS, and (4) TTA via pseudolabeling with our proposal. Our method outperforms all TTA methods with the single exception of the state of the art [Park 23] on VisDA-C. However, unlike [Park 23], our method does not require optimizing an adapter module for CDS.

4. Discussion and conclusions

This paper studied the effect of class distribution shift (CDS) in the task of source free domain adaptation (SFDA). Instead of proposing additional modules and objective functions to improve the SFDA's pseudolabeling process, we study the weakness of the nearest neighbors algorithm used in many previous works. We proved that, by adding robustness to the nearest neighbors via an external feature extractor, the accuracy of the subsequent adaptation improves, outperforming previous methods in both SFDA and test-time adaptation (TTA) tasks under CDS.

We employ pretrained models publicly available with fixed parameters, so applying our method incurs no extra training cost. In general, the stronger the architecture (i.e. more parameters, more sophisticated), the higher the accuracy. However, there may be small differences depending on the domain. As a general result, our method performs better on the *real-world* target domain, since the training data of the generic models is mostly based on photos and real images.

◇参 考 文 献◇

[Boudiaf 22] Boudiaf, M., Mueller, R., Ben Ayed, I., and Bertinetto, L.: Parameter-free online test-time adaptation, in *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 8344–8353 (2022)

[Cao 18] Cao, Z., Ma, L., Long, M., and Wang, J.: Partial adversarial domain adaptation, in *Proc. European conference on computer vision*, pp. 135–150 (2018)

[Chen 19] Chen, X., Wang, S., Long, M., and Wang, J.: Transferability vs. Discriminability: Batch spectral penalization for adversarial domain adaptation, in *Proc. International Conference on Machine Learning*, pp. 1081–1090 (2019)

[Chen 21] Chen, W., Yu, Z., De Mello, S., Liu, S., Alvarez, J. M., Wang, Z., and Anandkumar, A.: Contrastive syn-to-real generalization, in *Proc. International Conference on Learning Representations*, pp. 1–12 (2021)

[Dong 21] Dong, J., Fang, Z., Liu, A., Sun, G., and Liu, T.: Confident anchor-induced multi-source free domain adaptation, *Advances in Neural Information Processing Systems*, Vol. 34, pp. 2848–2860 (2021)

[Dosovitskiy 20] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929* (2020)

[Gong 22] Gong, T., Jeong, J., Kim, T., Kim, Y., Shin, J., and Lee, S.-J.: NOTE: Robust continual test-time adaptation against temporal correlation, *Proc. Advances in Neural Information Processing Systems*, Vol. 35, pp. 27253–27266 (2022)

[He 16] He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition, in *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)

[Hinton 15] Hinton, G., Vinyals, O., and Dean, J.: Distilling the knowledge in a neural network, *arXiv preprint arXiv:1503.02531* (2015)

[Jiang 20] Jiang, X., Lao, Q., Matwin, S., and Havaei, M.: Implicit class-conditioned domain alignment for unsupervised domain adaptation, in *Proc. International Conference on Machine Learning*, pp. 4816–4827 (2020)

[Li 21] Li, X., Li, J., Zhu, L., Wang, G., and Huang, Z.: Imbalanced source-free domain adaptation, in *Proc. ACM International Conference on Multimedia*, pp. 3330–3339 (2021)

[Liang 20] Liang, J., Hu, D., and Feng, J.: Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation, in *Proc. International Conference on Machine Learning*, pp. 6028–6039 (2020)

[Litrico 23] Litrico, M., Del Bue, A., and Morerio, P.: Guiding Pseudo-Labels With Uncertainty Estimation for Source-Free Unsupervised Domain Adaptation, in *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 7640–7650 (2023)

[Liu 21] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B.: Swin Transformer: Hierarchical vision transformer using shifted windows, in *Proc. International Conference on Computer Vision*, pp. 10012–10022 (2021)

[Mancini 18] Mancini, M., Karaoguz, H., Ricci, E., Jensfelt, P., and Caputo, B.: Kitting in the wild through online domain adaptation, in *Proc. International Conference on Intelligent Robots and Systems*, pp. 1103–1109 (2018)

[Park 23] Park, S., Yang, S., Choo, J., and Yun, S.: Label shift adapter for test-time adaptation under covariate and label shifts, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16421–16431 (2023)

[Peng 17] Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., and Saenko, K.: VisDA: The visual domain adaptation challenge, *arXiv preprint arXiv:1710.06924* (2017)

[Peng 19] Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B.: Moment matching for multi-source domain adaptation, in *Proc. International Conference on Computer Vision*, pp. 1406–1415 (2019)

[Saito 18] Saito, K., Watanabe, K., Ushiku, Y., and Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation, in *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 3723–3732 (2018)

[Shi 20] Shi, Z.: Improving k-nearest neighbors algorithm for imbalanced data classification, *IOP Conference Series: Materials Science and Engineering*, Vol. 719, No. 1, p. 012072 (2020)

[Tan 20] Tan, S., Peng, X., and Saenko, K.: Class-imbalanced domain adaptation: An empirical odyssey, in *Proc. European Conference on Computer Vision Workshops*, pp. 585–602 (2020)

[Venkateswara 17] Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S.: Deep hashing network for unsupervised domain adaptation, in *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 5018–5027 (2017)

[Wang 21] Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T.: TENT: Fully Test-time Adaptation by Entropy Minimization, in *Proc. International Conference on Learning Representations*, pp. 1–15 (2021)

[Wang 22] Wang, Q., Fink, O., Van Gool, L., and Dai, D.: Continual test-time domain adaptation, in *Proc. Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211 (2022)

[Weiss 01] Weiss, G. M. and Provost, F.: The effect of class distribution on classifier learning: an empirical study, Technical report, Rutgers University (2001)

[Yang 21] Yang, S., Weijer, van de J., Herranz, L., Jui, S., et al.: Exploiting the intrinsic neighborhood structure for source-free domain adaptation, *Advances in Neural Information Processing Systems*, Vol. 34, pp. 29393–29405 (2021)

[Zara 23] Zara, G., Conti, A., Roy, S., Lathuilière, S., Rota, P., and Ricci, E.: The Unreasonable Effectiveness of Large Language-Vision Models for Source-free Video Domain Adaptation, in *Proc. International Conference on Computer Vision*, pp. 10307–10317 (2023)

[Zhang 23] Zhang, W., Shen, L., and Foo, C.-S.: Rethinking the Role of Pre-Trained Networks in Source-Free Domain Adaptation, in *Proc. International Conference on Computer Vision*, pp. 18841–18851 (2023)

著 者 紹 介

Antonio Tejero-de-Pablos

Dr. Tejero-de-Pablos received his PhD from the Nara Institute of Science and Technology in 2017. Then, he worked as a researcher at the University of Tokyo, and at RIKEN AIP. From 2021, he works as a research scientist at CyberAgent AI Lab, where he conducts research on Computer Vision and Machine Learning.

Kコア分解に基づくライブ配信プラットフォームのソーシャルネットワーク分析

武内 慎
Makoto Takeuchi
メディア統括本部 Data Science Center
Data Mining Engineer
takeuchi_makoto@cyberagent.co.jp

佐野 幸恵
Yukie Sano
筑波大学 システム情報系
sano@sk.tsukuba.ac.jp

keywords: ライブ配信，CGM，生態系，K コア分解，ソーシャルネットワーク

Summary

Consumer-Generated Media(以下，CGM) プラットフォームでは，コンテンツ生産者とコンテンツ消費者がソーシャルネットワークを形成し，互いに交流する．生産者は消費者にコンテンツを提供し，消費者は消費を通じて生産者にコンテンツ生成の動機を与える．これは，生産者と消費者が相互に依存し合う共生関係と言える．本研究では，CGM プラットフォームの 1 例としてライブ配信プラットフォームに着目し，ライブ配信者と視聴者の相利共生ネットワークのレジリエンスを分析する．ブログなどの CGM は，消費者にとってはオンデマンド形式でコンテンツを消費するメディアであるのに対して，ライブ配信の CGM は，ライブ配信者のコンテンツ生成と視聴者の消費が同時に発生する．これにより，ライブ配信者は消費者から直接的にフィードバックを受けるため，より強い相利共生関係が期待される．ライブ配信プラットフォームでは，ライブ配信者の離脱が視聴者の離脱を引き起こし，それがさらにライブ配信者の離脱を引き起こすような，離脱の連鎖が生じる可能性がある．したがって，ユーザーの共生ネットワークのレジリエンスは，ライブ配信プラットフォームの生態系の維持において重要な研究テーマである．我々は K コア分解を用いてネットワークのレジリエンスを調査し，ユーザー離脱のリスクを評価した．

1. はじめに

オンラインメディアサービスの中で, ユーザーが自主的にコンテンツを提供し，それを別のユーザーが消費する Consumer-Generated Media(以下，CGM) は, ビジネスと学術分野の両方で注目され, 計算社会科学やマーケティングサイエンスの分野における重要な研究対象となっている [Crowston 18] [Ogushi 21] [Lipka 10]．例えば，CGM のコンテンツ生産者に焦点を当てた，CGM を生成する動機に関する研究 [Crowston 18] や，コンテンツと生産者の評価に関する研究 [Ogushi 21] [Lipka 10] などが存在する．一方で，コンテンツ消費者がコンテンツ生産者に与える影響についての研究はあまり存在しない．しかしながら，多くの場合，CGM のコンテンツ生産者は，生産したコンテンツが消費されないと，コンテンツを生成するモチベーションを維持することができずに生成をやめてしまうと考えられる．そのため，CGM プラットフォームにおけるユーザーの生態系は，植物と花粉媒介者のように相互に依存する共生関係で維持されていると解釈できる．ゆえに，生産者と消費者がそれぞれ双方に与える影響を把握することは重要である．

本研究では，CGM プラットフォームの 1 例として，ライブ配信プラットフォームのライブ配信者と視聴者の相利共生ネットワークの頑健性や回復力 (レジリエンス) を分析する．ブログなどの CGM は，消費者にとってはオンデマンド形式でコンテンツを消費するメディアであるのに対して，本研究で着目するライブ配信の CGM は，ライブ配信者によるコンテンツ生成と視聴者の消費が同時に発生する．これにより，ライブ配信者は視聴者から直接的にフィードバックを受けるため，より強い相利共生関係が期待される．

この相利共生関係で構築されたユーザーネットワークのレジリエンスはライブ配信プラットフォームのユーザー生態系の維持にとって重要である．なぜなら，魅力的なライブ配信者の離脱は，視聴者の離脱を引き起こし，それがさらなるライブ配信者の離脱に繋がるという離脱の連鎖を引き起こすリスクがあるからである．生物生態系においては，K コア分解を使ってネットワークのレジリエンスを調べた先行研究が存在する [Morone 19] [Burleson-Lesser 20]．本研究においても K コア分解を用いて，ネットワークのレジリエンスを調べた．

2. データ

本研究では，日本の音楽ストリーミングサービス AWA[*1] が提供する音声ライブ配信環境「ラウンジ」のユーザーログを用いて，配信者と視聴者のソーシャルネットワークを構築した．ネットワークのノードは，配信者と視聴者を表し，エッジは，一定の期間内にある配信者の配信コンテンツを視聴者がある閾値の回数以上視聴したという関係性を表す．本研究において，ソーシャルネットワークは 1 週間の期間の視聴ログから構成した．エッジを張る視聴回数の閾値は，サービスの都合により非公開である．共生関係の生態系を表す既存研究では，ネットワークを植物と花粉媒介者のように 2 部グラフで表現することが多い．一方で，ラウンジの配信者は，他の配信者のラウンジに視聴者として参加することができるため，配信者でありかつ視聴者でもあるノードも存在する．よって，この配信者と視聴者のソーシャルネットワークは，厳密には 2 部グラフにならない可能性があることに注意が必要である．

得られたソーシャルネットワークの次数分布を図 1 に示す．累積次数分布は両軸対数プロットでほぼ直線で減衰しており，一般的に示されているソーシャルネットワークと同様に，ロングテールな分布になっていることがわかる．

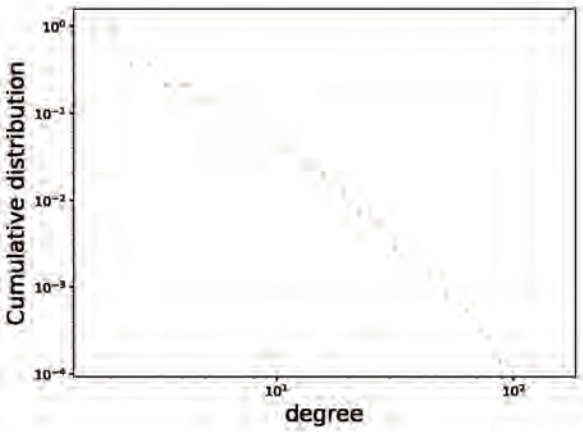


図 1　ライブ配信プラットフォームの配信者と視聴者のソーシャルネットワークにおける次数の累積分布 (両軸対数プロット)．次数分布はほぼ直線で，ロングテールな分布となっている．

3. 手法

まず，構築した配信者と視聴者のソーシャルネットワークに対して，K コア分解 [Seidman 83] [Carmi 07] を行う．ここで K コア分解とは，ネットワーク内の密に接続された部分グラフを特定するための分析手法である．K コア分解では，対象のネットワーク，つまりグラフを G とし，すべてのノードが少なくとも k の次数を持つ G の最大部分グラフを k-core と定義する．そして，k-shell [Carmi 07] と呼ばれる，k-core に属するが $(k+1)$-core には属さないノードの集合を計算し，各ノードの k-shell 数 k_{s} を求める．

次に，K コア分解を用いたネットワークのレジリエンスを調べる手法として，既存研究で提案された k_{risk} [García-Algarra 17] の指標を各ノードに対して計算する．k_{risk} は，式 (1) で定義される，あるノード m の削除に対するネットワークの脆弱性を定量化する指標であり，生物生態系ネットワークにおける種の絶滅の影響を測定する目的で導入された．

$$k_{\mathrm{risk}}(m) = \sum_{j,\, k_{\mathrm{s}}(j)<k_{\mathrm{s}}(m)} a_{mj}(k_{\mathrm{s}}(m) - k_{\mathrm{s}}(j)) + \epsilon k_{\mathrm{s}}(m) \tag{1}$$

ここで，m, j は G のノードを表し，a_{mj} は隣接行列の要素，k_{s} は k-shell 数である．右辺の第一項は対象のノード m の k-shell 数よりも小さい k-shell 数を持つ隣接ノードに対して和を取る．第二項は対象ノードが属する k-shell 数の違いを表現するための項であり，ϵ は第一項の和よりも第二項が常に小さくなるように調整するパラメータであり，本研究では先行研究 [García-Algarra 17] と同様に，0.01 とした．$k_{\mathrm{risk}}(m)$ は 0 以上の値をとり，値が大きいほど対象とするネットワーク G に対してノード m の除去が大きな影響を及ぼすことになる．

4. 結果

まず，k コア分解で各ノードの k-shell の値 k_{s} を求めた．図 2 は，k_{s} の値毎に階層を分け，配信者と視聴者のソーシャルネットワークをプロットしたものである．図の上部ノードほど k_{s} の値が高く，ネットワーク内で密に結合した中心部を構成するノードを表す．

図 3 に，k_{s} ごとの配信者と視聴者の割合を示す．$k_{\mathrm{s}} = 7$ を除き，k_{s} が増えるに従って，配信者ノードの割合が増加し，ネットワークの密な部分には配信者がより多く存在していることがわかる．その一方で，一部の視聴者ノードも高い k_{s} を持つが，これは配信者と視聴者の非対称性を考慮すると特筆すべき点である．つまり，配信者は一度の配信で複数の視聴者と接続される可能性があるのに対して，視聴者は基本的に配信への参加行動によって配信者と 1 人づつ接続されるため，各ノードから見た接続コストが，配信者と視聴者で異なるという背景がある．高い k_{s} を持つ視聴者ノードの存在は，自明ではなく，ネットワークのレジリエンスの観点でも興味深い結果である．

続いて，k_{s} を用いて，式 1 より，各ノードの k_{risk} を計算する．その結果得られた k_{risk} の累積分布を図 4 に示す．こちらの分布も，次数分布と同様にロングテールな分布となっている．この k_{risk} が高いノードほど，削除

*1　https://awa.fm/

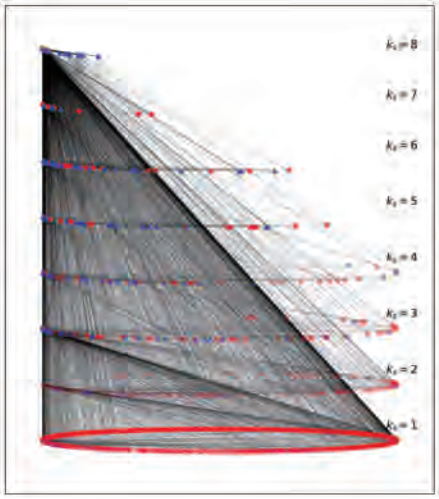


図 2　ライブ配信プラットフォームの配信者と視聴者のソーシャルネットワーク．ノードの青い三角形は配信者を，赤い丸は視聴者を表す．

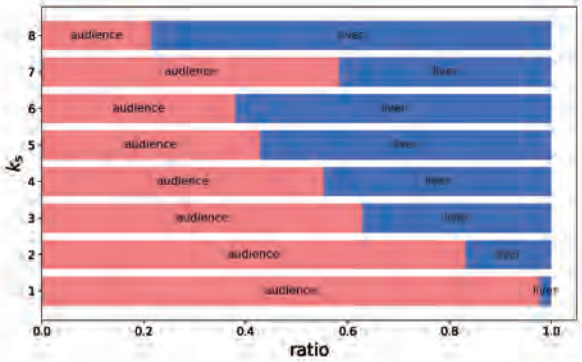


図 3　k-shell 数 k_s 毎の配信者と視聴者の割合．

した際にネットワークのレジリエンスに大きな影響を及ぼす．

k_{risk} は，その定義から，次数と正の相関を持つことが想定されるが，García-Algarra らによれば，次数が高いノードが常に k_{risk} も高い値を持つとは限らない [García-Algarra 17]．本研究で用いたネットワークにおける次数と k_{risk} の関係性を調べるため，図 5 に，次数と k_{risk} の散布図を示す．図から，同じ次数であっても k_{risk} の値には幅があることがわかる．この図 5 における縦方向の幅は，k_{risk} が持つ，ネットワーク上の 1 次の繋がりをみるだけでは捉えることができないレジリエンスに関する各ノードの特徴を示している．

ここで，本ネットワークの次数は，配信者ノードにとっては一定の回数閾値を超えて配信に参加してくれるファン視聴者の数であり，視聴者ノードにとっては一定の回数閾値を超えて配信に参加するような好みの配信者の数

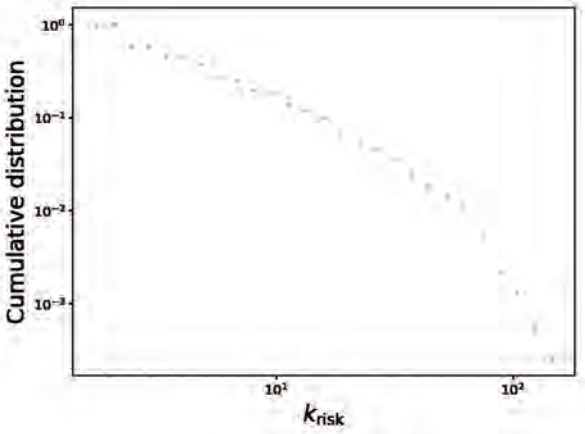


図 4　k_{risk} の累積分布 (両軸対数プロット)．ロングテールな分布となっていることがわかる．

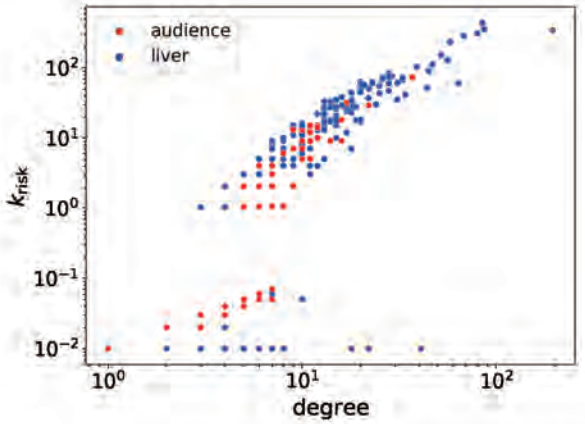


図 5　次数と k_{risk} の二次元平面上にノードをプロットした散布図 (両軸対数プロット)．赤色は視聴者を，青色は配信者をそれぞれ表す．

と解釈できる．そのような隣接ノードの数が少ないノードは，そのプラットフォームから離脱しやすいという仮定は自然である．この仮定が正しい場合，図 5 の 2 次元平面において，$k_{\mathrm{risk}}(m)$ が大きくかつ次数が小さいほど，より離脱しやすくかつ離脱したときの影響が大きい．逆に $k_{\mathrm{risk}}(m)$ が小さくかつ次数が大きいほど，より離脱しにくくかつ離脱したときの影響が少ないということになる．このように k_{risk} と他の特徴を組み合わせたユーザーの離脱リスク評価は，ユーザー生態系としてのソーシャルメディアプラットフォームを維持するために有用であると考えられる．

5. ま　と　め

本研究では，音声配信プラットフォームのユーザーログを用いて，配信者と視聴者のソーシャルネットワークを構築した．このネットワークは，プラットフォームにおけるユーザー同士の相利共生関係の生態系として解釈できる．我々は，生物生態系における生物種の相利共生

関係に使われる手法を用いて，このネットワークのレジリエンスを調べた．

まず，K コア分解を行い，ネットワークが密に接続している箇所を抽出し，各ノードの k-shell 数 k_s を計算した．k_s ごとの配信者ノードと視聴者ノードの割合は，k_s が高いほど配信者ノードの割合が高くなる傾向があり，配信者と視聴者の非対称性を示している．そして，k_s を用いて，各ノードを削除した際のネットワークへの影響を表す指標として k_{risk} を計算した．この k_{risk} の値は，次数と正の相関関係があるが，同じ次数においてもある程度の幅を持ち，ネットワーク全体のレジリエンスに関する各ノードの特徴を表現する．k_{risk} と次数などの特徴量を組み合わせて，各ノードの離脱リスクを評価できる可能性があり，プラットフォームにおける配信者と視聴者の相利共生関係の生態系を維持するための活用が期待される．

本研究の将来課題として，実際の音声配信プラットフォーム利用の時系列データから得られる，ユーザーの離脱フラグを活用し，ネットワークからの離脱予測因子を特定すること，また，その離脱予測因子と k_{risk} を組み合わせたリスク評価方法の確立とその検証が考えられる．また，今回確認された，音声配信プラットフォームの相利共生生態系における配信者と視聴者の非対称性が，ネットワークのレジリエンスに与える影響も解明すべき課題である．

◇ 参 考 文 献 ◇

[Burleson-Lesser 20] Burleson-Lesser, K., Morone, F., Tomassone, M. S., and Makse, H. A.: K-core robustness in ecological and financial networks, *Scientific Reports* (2020)

[Carmi 07] Carmi, S., Havlin, S., Kirkpatrick, S., Shavitt, Y., and Shir, E.: A model of Internet topology using k-shell decomposition, *Proceedings of the National Academy of Sciences*, Vol. 104, No. 27, pp. 11150–11154 (2007)

[Crowston 18] Crowston, K. and Fagnot, I.: Stages of motivation for contributing user-generated content: A theory and empirical test, *International Journal of Human-Computer Studies*, Vol. 109, pp. 89–101 (2018)

[García-Algarra 17] García-Algarra, J., Pastor, J. M., Iriondo, J. M., and Galeano, J.: Ranking of critical species to preserve the functionality of mutualistic networks using the k-core decomposition, *PeerJ*, Vol. 5, p. e3321 (2017)

[Lipka 10] Lipka, N. and Stein, B.: Identifying Featured Articles in Wikipedia: Writing Style Matters, in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, p. 1147–1148 (2010)

[Morone 19] Morone, F., Del Ferraro, G., and Makse, H. A.: The k-core as a predictor of structural collapse in mutualistic ecosystems, *Nature Physics* (2019)

[Ogushi 21] Ogushi, F., Kertész, J., Kaski, K., and Shimada, T.: Ecology of the digital world of Wikipedia, *Scientific Reports* (2021)

[Seidman 83] Seidman, S. B.: Network structure and minimum degree, *Social Networks*, Vol. 5, No. 3, pp. 269–287 (1983)

著 者 紹 介

武内慎

2015 年 株式会社サイバーエージェントに中途入社．2024 年 筑波大学大学院 システム情報工学研究群 博士後期課程修了 (社会工学)．メディアサービスのデータ分析に従事．

佐野幸恵

筑波大学 システム情報系 社会工学域 准教授．博士 (理学)．社会経済物理学やネットワーク科学に関する研究に従事．

2024 Vol.2

AdTEC: 検索連動型広告における テキスト品質評価のための統合ベンチマーク

AdTEC: A Unified Benchmark for Evaluating Text Quality in Search Engine Advertising

張 培楠 Peinan Zhang
AI Lab
Research Scientist
zhang_peinan@cyberagent.co.jp

坂井 優介 Yusuke Sakai
奈良先端技術大学院大学
Assistant Professor
sakai.yusuke.sr9@is.naist.jp

三田 雅人 Masato Mita
AI Lab
Research Scientist
mita_masato@cyberagent.co.jp

大内 啓樹 Hiroki Ouchi
AI Lab, 奈良先端技術大学院大学
Research Scientist, Associate Professor
ouchi_hiroki_xa@cyberagent.co.jp

渡辺 太郎 Taro Watanabe
奈良先端技術大学院大学
Professor
taro@is.naist.jp

keywords: テキスト品質評価、ベンチマーク、インターネット広告

Summary

自然言語生成技術によって生成された流暢な広告文が増加する中、これらのクリエイティブの品質を実際の環境で検証することが強く求められている。我々は、実際の広告運用の観点から広告文を多面的に評価する初の公開ベンチマークである AdTEC を提案する。我々の貢献は以下の通りである。(i) 広告文の品質を評価するための 5 つのタスクを定義し、実務経験に基づいた日本語データセットを構築した。(ii) 既存の事前学習済み言語モデル（PLM）と人間の評価者のデータセットにおける性能を検証した。(iii) ベンチマークの特性を分析し、課題を提供した。結果として、PLM がいくつかのタスクで既に実用レベルに達している一方で、人間が特定の領域で依然として優れていることを示しており、その分野には大きな改善の余地があることを示唆している。

1. はじめに

インターネット広告、特に検索連動型広告は、ベンダーが製品を宣伝するための主要な手段であり、市場規模は今後数年間で数十億ドル増加すると推定されている [Murakami 23]。広告運用（AdOps）の需要が高まる中で、商材情報から広告文を作成する（図 1のステップ 2）など、事前学習済み言語モデル（PLM）による自然言語生成（NLG）の顕著な成功が実応用に拍車をかけている [Dong 21, Vaswani 17, Brown 20, Touvron 23a]。このように、近年では広告は NLP の社会実装の主要な事例となっている [Hughes 19, Kamigaito 21, Golobokov 22]。本論文では、広告文の品質を評価することに焦点を当てる。このプロセスを広告文評価と呼び、図 1のステップ 3 で示されている。広告文の品質評価は重要である。なぜなら、低品質の広告文は流暢さの欠如、不適切な訴求、誤解を招く表現を通じて広告主にとって不利益となり得るからである。検索連動型広告のようなデータ量が大きいドメインで各テキストの品質を人間に確認させることは高コストであり、スケーラビリティに乏しいため、広告文の自動品質推定器の開発が強く求められている。品質には、適切な言葉遣い、効果的な訴求、広告文と商品情報の一貫性、高い配信パフォーマンスなど、複数の側面がある。これらの側面は自動品質評価器に含まれるべきであるが、そのようなベンチマークは存在しない。したがって、ボトルネックは、多数のクリエイティブを自動的に生成する能力があるにもかかわらず、広告文の質を検証することにあり、これが配信ボリュームのスケーラ

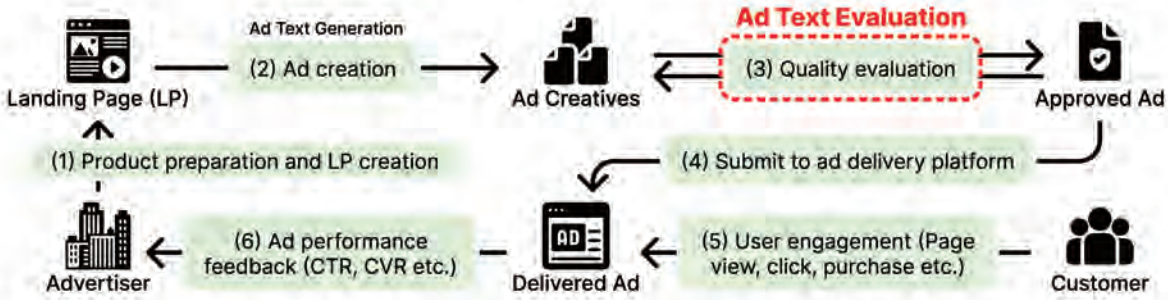


図 1: 広告運用の一般的なワークフローは、2 章で説明されている。(1) 広告主が製品を宣伝するためのランディングページ（LP）を作成する。(2) LP に記載された商材情報と顧客情報に基づいて、制作者がテキストとグラフィックを制作する。(3) 制作物は、流暢さ、魅力、規制、合法性、その他の要因に基づいて評価される。(4) 制作物が品質評価を通過すると、配信基盤に提出される。(5) 顧客は、表示された制作物に対して閲覧、クリック、購入などのアクションを取る。(6) 顧客のエンゲージメントに基づいて、制作物のパフォーマンスが広告主に報告され、LP と制作物の品質を向上させるためにステップ 1 に戻る。

ビリティを妨げている。したがって、我々は広告文の質を評価するためのベンチマークを構築することを目指す。

広告文評価ベンチマークの構築における主な課題は、タスクの明確な定義が存在しないことである [Murakami 22, Mita 24]。広告運用におけるドメイン知識の欠如は、高品質な広告文の基準の理解とタスクの正確な定義を複雑にしている。しかし、広告運用のワークフローは複雑であり、様々なプラットフォームやフォーマットに依存し、複数の方法論と指標を含んでいる。加えて、オンライン広告を大規模に運用する専門知識を持つ企業はごく少数である。法的および契約上の義務により、広告運用のワークフローとデータは主に社内で管理されており、公開されているデータセットが不足している。この不足により、多様な方法論を体系的に再現し、学術的に検証することが困難である。その結果、広告分野での研究はあまり活発ではなく、問題を見落としたり、最先端技術の応用と開発が遅れる可能性がある。これらの課題に対処するために、一般化された広告運用ワークフローに基づいてタスクを定義し統一する初の公開ベンチマークである **AdTEC** を提案する。我々の主な貢献は以下の通りである。

- 現実の広告運用ワークフローを整理し、このワークフローに密接に関連する 5 つのタスクを慎重に設計して広告文の品質を評価した。この取り組みでは、広告代理店の実務経験に基づき、日本の広告運用で使用されたデータをケーススタディとして用いた。このデータセットは広告文評価のための初の公開データセットである。
- 提案したベンチマークを用いて、BERT や RoBERTa、大規模言語モデル（LLMs）などの既存の PLM と人間の評価者も含めた性能を検証した。
- 実験を通じてデータセットの特性を分析し、潜在的な問題を特定した。この結果、我々のベンチマークが挑戦的であることを示し、改善の余地や今後の研究の可能性を明らかにした。

図 2: 検索連動型広告の概要

2. 検索連動型広告と運用の流れ

検索連動型広告は、ユーザーが検索エンジンに入力するキーワードに関連する広告のタイトルと説明文を表示し、検索結果の一部として表示される。図図 2に示すように、ユーザーが広告の URL をクリックすると、ランディングページ（LP）と呼ばれるウェブページに誘導される。LP には広告主の製品に関連するテキストや画像が含まれており、ユーザーが購入などのアクションを取ることができる。

このような広告を運用するには、配信基盤、特性、形式の多様性があるため、高度な専門知識が必要である。洞察を得るために、広告運用に精通した 2 種類の専門家、広告代理店の運用部門を監督する者と現場で直接運用に携わる者にインタビューを行った。これらのインタビューを通じて、広告運用の複雑さを探り、ワークフローを一般化し、図図 1に示すように 6 つのステップとして整理した。本研究では、主に品質評価のステップ 3 に焦点を当てる。

3. AdTEC の構築

我々の目標は、実際の広告運用のシナリオに即した品質評価の側面を捉え、現実世界のアプリケーションやさまざまな研究目的に価値を提供するよう慎重に設計されたベンチマークを開発することである。

広告文	LP テキスト	広告容認性/ 広告一貫性
マンション販売 / 無料査定実施中	XX であなたの一戸建て住宅を査定！	acceptable / inconsistent
エンジニアのキャリア / エンジニアの仕事	XX で仕事を探そう！	unacceptable / consistent

表 1: 広告容認性タスクと広告一貫性タスクの例。なお LP テキストは広告一貫性タスクでのみ使用される。

3.1 タスク設計

ワークフローは §2 章で示した手順に従い、インタビューで同じ専門家にステップ 3 の図図 1内の重要な部分を特定するよう依頼した。彼らの洞察に基づいて、広告文を直接または間接的に評価するタスクを定義した。 直接評価タスクは、二者択一の合格/不合格の結果やテキスト品質を定量化する数値スコアなど、厳格な基準でテキストを評価するために使用される。これらのタスクは、最低限の配信基準が満たされていることを確認するためのチェックリストとして機能する。間接評価タスクは、テキストを見直したり洗練したりする際に人間の評価者を支援するため、または下流のタスクに接続するための橋渡しとして機能する。上記の原則に基づき、3 つの直接評価タスク（広告容認性、広告一貫性、広告効果予測）と 2 つの間接評価タスク（訴求表現認識と広告類似性）の合計 5 つのタスクを設計した。

§1 広告容認性

多くの広告配信基盤では広告文の長さに制限を課しているため、限られたスペース内で可読性を高め、読者を惹きつけるための軽微な文法エラーは許容される。しかし、過度な圧縮は読者を誤解させる可能性があり、そのような低品質な広告は広告主への悪影響を避けるために配信前に検出されるべきである。これを評価するために、広告文全体の品質の容認度を予測するタスク「広告容認性」を定義し、二値ラベルで「acceptable (容認)」/「unacceptable (非容認)」を用いる。非容認な広告の現象には、専門家のフィードバックに基づく「記号の崩壊」、「不自然な繰り返し」、「意味不明」が含まれる。これは、CoLA [Warstadt 19] のような文法的正確性をチェックする一般的な言語容認性とは異なる。広告容認性の例は表 1に示されている。広告文「**エンジニアのキャリア / エンジニアの仕事**」は、意味が重複しているため「非容認」である。

§2 広告一貫性

広告文とランディングページ（LP）の内容における一貫性を確認することは重要である。広告文に記載された機能や価格が、対応する LP で言及されていない場合、不当景品類及び不当表示防止法に抵触し、広告主に損害を与える可能性がある。しかし、これらの不一致を検出するのは困難である。なぜなら、いくつかの事実表現が LP には現れないからである。例えば、広告文ではよく使われる「公式」という用語は、LP の内容にはほとんど現れない。これを評価するために、LP の内容と広告文の一貫性を二値ラベル「consistent (一貫)」/「inconsistent (非一貫)」で予測する広告一貫性タスクを定義した。広告一貫性の例を表 1に示す。一行目は、LP が「一戸建て」を指しているのに対し、広告文が「マンション」を言及しているため、不一致ラベルが付けられている。

§3 広告効果予測

広告の品質を測定する最も直接的な方法は、それらを実際に配信して最終顧客に評価させることである。しかし、すべての広告を修正なしで配信することは低品質の広告は広告主に悪影響を及ぼす可能性があるため現実的ではない。したがって、先行研究では、過去の配信履歴に基づいて、クリック率（CTR）などの顧客行動をシミュレートすることにより、広告文の品質を測定する方法が使用されてきた。これらの方法は現在、多くの組織で一般的な手段となっている。これらの研究 [Gharibshah 20, Niu 20, Yang 22] に触発され、広告文、キーワード、および業種から品質スコアを [0, 100] の範囲で推定する 広告効果予測 タスクを採用した。スコアは過去の配信履歴に基づいて顧客行動をシミュレートし、契約上の理由から元のラベル分布を維持するために非線形的に変換される。

§4 訴求表現認識

広告において最も重要な要素の一つは 訴求表現である。広告の本質は、広告主と顧客をつなげることであり、訴求表現はその橋渡しの役割を果たす。例えば、低価格を強調する広告は、価格に敏感な顧客からの共感を得るかもしれないが、高性能を重視する広告は同じ顧客からの共感は得られにくい。このように、広告における訴求表現を認識し、適切な訴求表現を使用することは、CTR 予測などの下流タスクを向上させることができる重要なファクターである [Murakami 22]。訴求表現認識タスクでは、先行研究 [Murakami 22] に従い、与えられた広告文におけるすべての関連する訴求表現ラベルを予測する。図 図 3 は、広告文と対応する訴求表現の例を示している。

§5 広告類似性

同じ広告を繰り返し顧客に表示すると、飽きに起因する広告効果の低下である「広告の疲弊」現象 [Abrams 07] が生じる。このため、同じ広告を長期間表示することを避け、定期的に異なる広告に置き換えることが重要である。しかし、古い広告から新しい広告への移行は慎重に管理する必要がある。なぜなら、製品とその魅力を維持しつつ、文言や表現を更新しなければならず、以前の広告に惹かれていた顧客の興味を失うリスクがあるからである。したがって、この状況に焦点を当てた類似性の測

Field	Example value
Title 1	[No.1] Card loan comparison site
Title 2	A must-see for those who in a hurry!
Title 3	Instant Loan Secure Card Loan
Desc. 1	The best place to get a card loan without telling anyone. You only need a driver's license to apply
Desc. 2	Convenient to use ATMs at convenience stores. Convenient and quick loans are available if you apply before 10:00 p.m.
Keyword	card loan
Industry	finance
Score	82.3

表 2: 広告効果予測 タスクの例。Desc. は説明文を表す。Score はラベルであり、その他は入力である。

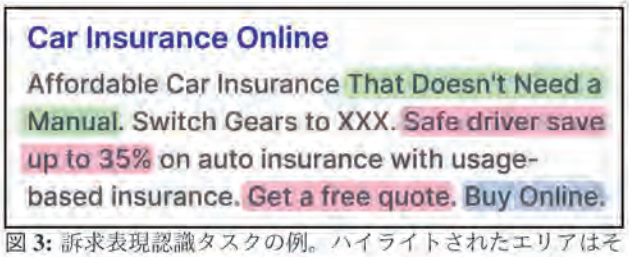


図 3: 訴求表現認識タスクの例。ハイライトされたエリアはそれぞれ特徴・メリット、特典、利便性の訴求表現を表している。

	Sentences	Score
S1	Suppon Black Vinegar with Luxury Ceramide	5.00
S2	Suppon Black Vinegar and Luxury Ceramide	
S1	Find a gift that fits your budget	2.33
S2	Save up to 40% on discounted products	

表 3: 広告類似性タスクの例。S1 と S2 はそれぞれペアになった文 1 と文 2 を表す。

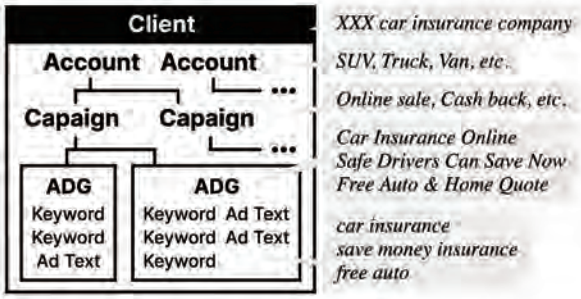


図 4: 広告配信のアカウント構造は階層的である。クライアントは単一の企業を表し、アカウントは通常、そのクライアントが提供する商業製品を包含する。キャンペーンはこれらの商業製品を宣伝するために作成され、広告グループはキーワードと広告文を整理するために使用される。階層の上位レベルでは、広告の数が多く、バリエーションも大きい。反対に、下位レベルでは広告の数が少なく、より似通っている傾向がある。

定が重要であり、定量化されたスコアに基づいて広告を置き換えるかどうかを判断することが重要である。専門家へのインタビューを通じて、広告の類似性が一般的な類似性とは異なり、商材や訴求表現の違いによって特徴付けられることが分かった。

上述の動機に基づき、我々は広告類似性タスクを定義し、広告文のペアに対する類似度スコアを [1, 5] のスケールで予測する。この値が低いほどペアの類似性は低く、逆に高いほど類似性は高い。例を表 3に示す。最初のペアは高い類似性を示しており、「すっぽん黒酢」という商材と「高級感」という 訴求表現が同一であるが、逆に、2 つ目の例のペアでは訴求表現 が予算と割引で異なっており、比較的低い類似性を示している。

3·2 データセット構築

§1 データ収集

広告容認性および広告一貫性のタスクでは、実際の広告運用ワークフローの制作フェーズから、人間のクリエイターと NLG モデルの出力の両方を含むデータを収集した。広告効果予測および 広告類似性タスクでは、2021 年から 2022 年に配信された日本のスポンサードサーチ広告を使用した。訴求表現認識タスクでは、[Murakami 22] のデータを使用した。

§2 データの前処理

広告効果予測 タスクでは、広告主にとってセンシティブな広告配信実績を含むため、クライアントと交渉して公開が承認されたデータのサブセットを使用する。また、CTR の生データに非線形変換を適用し、分布を保持するために [0,100] の範囲にスケーリングした。さらに、製品名や会社名のような固有名詞をマスクすることで広告主の特定を防ぎ、データ公開時の潜在的な悪影響を避けるようにした。

広告類似性 タスクでは、大半の広告文ペアが類似していないため、ランダムサンプリングによる文ペアの作成は非効率的である。したがって、広告配信時に設定されたアカウント構造を利用した（図 4参照）。予備実験の結果から、同じ広告グループからテキストをサンプリングすることで擬似的な類似ペアを作成した。

ラベル分布を均衡させるために、異なるクライアントに属する 2 つの広告文を確保することで擬似的な非類似ペアも作成した。その結果、擬似的な類似ペアと非類似ペアを 9:1 の比率でサンプリングした。

§3 アノテーション

すべてのアノテーターは日本語を母語とする話者であり、広告運用における豊富な経験を持つ専門家である。

Task	Setup: Input → Label	Metrics
広告容認性	分類: 広告文 → acceptable/unacceptable	Accuracy/F1-score
広告一貫性	分類: (広告文, LP 情報) → consistent/inconsistent	Accuracy/F1-score
広告効果予測	回帰: (広告文, キーワード, 業種) → [0,100]	Pearson/Spearman corr.
訴求表現認識	分類: 広告文 → マルチラベル	F1-micro/-macro
広告類似性	回帰: 広告文ペア → [1,5]	Pearson/Spearman corr.

表 4: ADTEC のタスク概要。 Corr. は相関係数の略である。

タスク	訓練	開発	テスト
広告容認性	13,265	970	980
広告一貫性	10,635	945	970
広告効果予測	125,087	965	965
訴求表現認識	1,856	465	410
広告類似性	4,980	623	629

表 5: 各データセットにおける各タスクのインスタンス数。

すべてのタスクにおけるアノテーションのワークフローは、広告効果予測 と訴求表現認識 を除いて以下の通りである：(1) まず、重複する項目を削除し、日本語以外の言語のデータをフィルタリングした。(2) 小規模なサンプルデータセットでのパイロットアノテーションを通じて、満足のいく合意レベルに達するまでアノテーションガイドラインを反復的に改訂した。(3) ガイドラインに従い、ステップ 2 で使用したサンプルデータを 3 人のアノテーターに再度アノテーションさせた。その後、アノテーターの結果を我々自身の結果と比較し、齟齬を解決するまでガイドラインをさらに改良した。このサイクルは少なくとも 2 回繰り返された。(4) ステップ 1 から 3 を完了した後、最終ガイドラインを用いてフルテストセットで本番アノテーションを実施した。

§4 データ分割

重複排除にもかかわらず、類似の広告表現は依然として簡単に見つけることができ、これは単純なランダム分割でデータリークにつながる可能性がある。さらに、データが業界で使用されると仮定した場合、特定の広告表現に過剰適合せずに効果的に一般化することが重要である。したがって、広告容認性および 広告一貫性に対しては、図 図 4 の広告階層構造を考慮してデータを分割し、訓練、開発、テスト間で クライアント が重複しないようにした。広告効果予測では、同じく広告階層構造のキャンペーンを重複しないように分割した。訴求表現認識に対しては、[Murakami 22] の同じ分割を使用した。広告類似性に対しては、ラベル分布の一貫性を維持するためにデータをランダムに分割した。表表 5は、我々のデータセットの統計を示している。

4. 実　　験

4·1 実 験 設 定

表 4は、各タスクの概要とタスクのパフォーマンスを測定するために使用された指標を示している。我々のタスクは、二値分類、マルチラベル分類、回帰の 3 つの設定に分類される。二値分類では、Accuracy と F1 スコアを使用し、マルチラベル分類では、[Murakami 22] に従い、マクロおよびマイクロの両方で F1 スコアを測定する。回帰では、ピアソンおよびスピアマンの相関係数を使用する。

我々は、PLM を含む 2 種類の評価器を使用した。これには、エンコーダモデルによる微調整設定と、LLM による文脈内学習設定、そして人手評価が含まれる。

§1 エンコーダモデルによる微調整

我々はベースラインとしてパブリックに利用可能なエンコーダーモデルを利用した。具体的には、Tohoku BERT、Waseda RoBERTa、XLM-RoBERTa [Conneau 19] であり、これらは日本語 NLP タスクで一般的に使用されている。これらのモデルは、トークナイザおよび事前学習データセットで異なる。前述のすべてのモデルは LARGE サイズである。

§2 LLM による文脈内学習

オープンな LLM のベースラインとして CALM2_{7b} と ELYZA_{7b} を採用した。CALM2_{7b} と ELYZA_{7b} は Llama 2 [Touvron 23b] アーキテクチャに基づいているが、訓練方法とデータが異なる。CALM2_{7b} はゼロからトレーニングされたのに対し、ELYZA_{7b} はオリジナルの Llama 2 から継続的にトレーニングされた。[*1]

§3 人手評価

広告運用に従事していない 3 人の人間評価者を募り、広告効果予測を除くすべてのタスクを評価した。3·2 節で説明した手順と同じ手順で指示を行った。本番評価実施前に、各タスクのトレーニングセットからランダムに抽出した 100 インスタンスについて、2 回のパイロット評価を実施した。最終評価では、広告容認性と 広告一貫性

*1 LLM のファインチューニングとゼロショット／少数ショット学習のギャップをさらに評価するために、LLM をファインチューニングした。これは我々の研究の範囲を超え、計算資源によって制約される実験であるため、ゼロショット／少数ショット設定で最も性能の良いオープンソースモデルのみを選択した。

評価器	広告容認性 Accuracy/F1-score	広告一貫性 Accuracy/F1-score	広告効果予測 Pearson/Spearman	訴求表現認識 F1-micro/-macro	広告類似性 Pearson/Spearman
エンコーダモデルによる微調整					
Tohoku BERT	0.711/0.688	**0.767**/0.552	**0.480**/**0.497**	0.774/**0.694**	0.773/0.807
Waseda BERT	0.598/0.637	0.755/0.474	0.445/0.457	0.663/0.517	0.740/0.800
XLM-RoBERTa	0.705/0.690	0.758/0.519	0.453/0.457	**0.778**/0.677	**0.878**/**0.878**
LLM による文脈内学習					
$CALM2_{7b}$	0.520/0.115	0.381/0.472	0.006/0.013	0.154/0.042	0.036/0.036
$ELYZA_{7b}$	0.352/0.520	0.628/0.771	0.003/0.046	0.196/0.044	0.015/-0.004
GPT-4	0.325/0.433	0.583/0.612	0.028/0.073	0.417/0.113	0.776/0.811
Human	**0.732**/**0.790**	0.703/**0.807**	—	0.564/0.538	0.699/0.765

表 6: テストセットにおける PLM と人間評価者のパフォーマンス。各設定での最良の結果を下線で示し、すべての手法での最良の結果を太字で示す。

タスクに対してはインスタンスごとに多数決を行い、他のタスクについては評価者の平均スコアを報告した。

4·2 結果と議論

表 6 は結果の概要を示している。

§1 エンコーダモデルによる微調整

XLM-RoBERTa と東北 BERT は、2 つ以上のタスクで最高または競争力のあるスコアを達成した。さらに、LARGE モデルはほとんどのタスクで BASE モデルを上回り、パラメータサイズの増加が広告表現の理解において重要な役割を果たしていることを示唆している。

§2 LLM による文脈内学習

GPT-4 は、すべてのタスクにおいて高いパフォーマンスを達成したが、$ELYZA_{7b}$ は広告一貫性で LLM の中で最も優れたパフォーマンスを示した。広告類似性タスクでは、オープンな LLM と OpenAI の LLM の間に大きな差が観察された。$CALM2_{7b}$ と $ELYZA_{7b}$ は 0 に近いスコアを記録し、相関がないことを示しているが、OpenAI のモデル、特に GPT-4 は 0.776/0.811 という競争力のあるスコアを達成した。したがって、この研究で使用されたオープンな LLM は、意味的な類似性や数値的な回答を扱うのに苦労するが、GPT-4 はかなり優れたパフォーマンスを発揮することがわかった。広告効果予測タスクでは、すべての LLM が 0 に近い相関のない回答を生成し、広告のパフォーマンスを正確に予測することがオープンな LLM にとって依然として課題であることが示された。

§3 エンコーダモデルによる微調整 vs. LLM による文脈内学習

全体として、微調整されたモデルは LLMs を上回った。差は大きく、0.2 から 0.6 の範囲であり、特に広告容認性、広告効果予測、および訴求表現認識のタスクで顕著であった。これはタスクの特性に起因するものである。広告効果予測は、[0,100] の範囲で数値を予測することを含み、訴求表現認識は 20 以上のラベルからすべての適切なラベルを選択することを要求するため、データの特徴と出力の多様性が少数ショットだけでは対処できないことを示唆している。広告容認性と広告一貫性はどちらも二値分類タスクであるが、微調整されたモデルと LLMs のパフォーマンスの差は、広告容認性タスクでより顕著であり、0.2 ポイントもの差があるのに対し、広告一貫性タスクでは比較的小さな差が観察された。最も差が小さかったタスクは広告類似性であり、差はわずか 0.06 から 0.10 であった。

§4 事前学習モデル vs. 人手評価

人間の評価者は、広告容認性および 広告一貫性において事前学習モデルを上回った。両方のタスクにおいて、モデルは高い精度と低い F1 スコアを持つ傾向がある一方で、人間は逆の傾向を示す。特にラベルの分布が不均衡である 広告一貫性タスクでは、最良のモデルでさえ F1 スコアが 0.55 に過ぎないが、人間のパフォーマンスは 0.8 に達する。

これは、人間が不均衡なデータにおいて、Precision と Recall の両方でより良い予測を行うことができることを示唆している。一方で、微調整されたモデルは、訴求表現認識 および 広告類似性 タスクにおいて人間を上回る。訴求表現認識 タスクでは、人間の評価者は LLMs と同様に出力ラベルの多様性に苦労していた。しかし、人間の結果では F1-micro と F1-macro のスコアの差は 0.03 ポイントと比較的小さい。対照的に、微調整された PLMs では、この差は 0.08 から 0.20 ポイントの範囲である。これは、人間の評価者が一般化する強い能力を持ち、ラベルが稀にしか現れない場合でも高いパフォーマンスを維持できることを示している。 広告類似性 タスクでは、微調整されたモデルに加えて、GPT-4 が人間の評価者を上回る。これは、LLM が人間を上回る唯一のタスクであり、GPT-4 が意味的類似性と数的理解において人間と高いレベルで一致していることを示唆している。

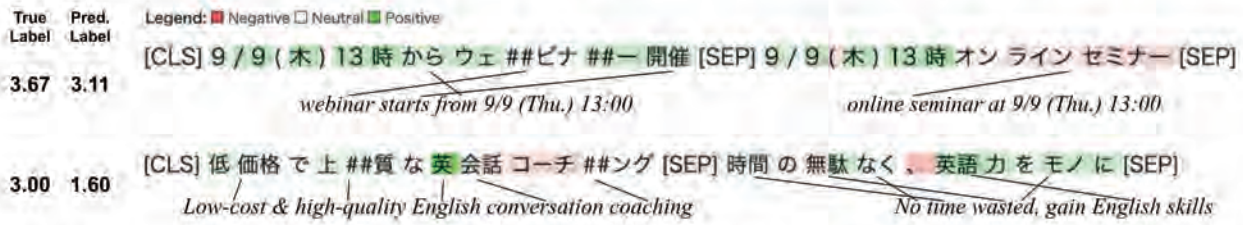


図 5: 東北 BERT モデルの出力における統合勾配可視化の例は、広告類似性 タスクにおいて真のラベルと予測ラベルの間のギャップが小さい（上）場合と大きい（下）場合の注意の違いを示している。赤は予測における負の影響を示し、緑は正の影響を示している。

Error	GT	H	M	広告容認性	広告一貫性
C-IC	T	T	F	30.1%	17.2%
	F	F	T	0.6%	5.5%
IC-C	T	F	T	0.6%	1.9%
	F	T	F	18.0%	10.9%
IC-IC	T	F	F	4.9%	6.2%
	F	T	T	3.9%	6.6%

表 7: Tohoku BERT（BERT）と XLM-RoBERTa（XLM-R）のタイプ別エラー率を広告容認性 タスクと広告一貫性 タスクで示す。グラウンドトゥルース（GT）、人間（H）、モデル（M）のラベルは、T（acceptable または consistent）と F（unacceptable または inconsistent）で表される。

§5 広告容認性 と広告一貫性 におけるエラー分析

我々は、モデルがまだ人間の評価者を上回っていない広告容認性と広告一貫性の 2 つのタスクについて詳細な分析を行った。我々は 3 種類のエラーを分析した: 人間が不正解でモデルが正解（IC-C）、人間が正解でモデルが不正解（C-IC）、そして両方が不正解（IC-IC）であり、これらは表 7に示されている。両方のタスクにおいて、モデルは IC-C および C-IC エラーに対して、True ラベルよりも False ラベルを予測することが多かった。対照的に、人間は正しい場合も誤った場合もより多くの True ラベルを提供した。これは、人間がモデルと比較して比較的高い寛容さを示すことを示唆している。モデルは、広告容認性 および 広告一貫性 タスクにおいて決定を下す際に過度に慎重になる傾向がある。

§6 広告類似性タスクにおけるモデルの挙動に関する事例分析

統合勾配法 [Sundararajan 17] を用いて、Tohoku BERT の注意機構を可視化し、モデルの動作をより深く理解することを目指した。この理解は、モデルの性能を向上させる可能性がある。図 5は、広告類似性タスクにおける予測と実際の値の間の誤差が小さい例と大きい例を示している。最初の例では、モデルが「オンラインセミナー」に十分な注意を払うことができず、これは「ウェビナー」と類似した意味を持つ。しかし、「*9/9（木）13:00*」という日付表現に高い注意を払うことで、正解に近いスコアを予測することができる。しかし、2 番目の例では、「英会話コーチング」に十分な注意を払わず、これは「英語スキルの習得」と類似した意味を持つため、不正確な非類似性の予測をしてしまう。このようなケースが他にも多く見られたため、モデルは特に日付、時間、数値などの表面的な情報を優先する傾向があると考えられる。これは、表面的な情報が異なる場合でも意味的に類似したケースを正しく識別することを妨げる可能性があるし、その逆もまた然りである。

モデルの欠点が観察されるもう一つの例は、推論においてである。例えば、「半額」や「2 つ買うと 1 つ無料」といったフレーズを、モデルが同じ意味として誤って解釈することがあるが、人間は容易に理解する。また、「24 時間営業」と「毎週水曜日休み」や「駅近くの物件」と「駅から徒歩 20 分」といったフレーズのペアも、モデルが理解する際に混乱することが多い。これらのようなパラフレーズは広告作成の過程でよく使用されるが、自然なデータセットで自動的に捉えることは難しい。そのため、これらの現象に焦点を当てた専門的なデータセットが、モデルが正確に捉えるために必要である。

5. おわりに

私たちは、実際の広告運用ワークフローに基づいて、広告文の品質を検証するために 5 つのタスクを定義し、NLP コミュニティのために構築された日本の広告データの大規模で多用途かつ包括的なベンチマークである AdTEC を構築した。私たちは AdTEC 上で事前学習モデルと人間の評価者の両方を用いて評価を行い、その特性を探求し、実践的なワークフローの応用や将来の改善と研究の可能性についての洞察を提供した。私たちの発見は、実際のデータから直接サンプリングすることがモデルにとって一般的に有益であり、自然言語推論と意味理解に焦点を当てたタスクの重要性を強調していることを示唆している。本論文で定義したタスクとデータセットの組み合わせが、広告文評価の研究を進展させ、広告と NLP の分野を橋渡しし、新たな発見と応用への道を開くことを望んでいる。

◇参 考 文 献◇

[Abrams 07] Abrams, Z. and Vee, E.: Personalized Ad Delivery When Ads Fatigue: An Approximation Algorithm, in *Proceedings of Workshop on Internet and Network Economics*, pp. 535–540 (2007)

[Brown 20] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D.: Language Models are Few-Shot Learners (2020)

[Conneau 19] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V.: Unsupervised Cross-lingual Representation Learning at Scale, *arXiv preprint arXiv:1911.02116* (2019)

[Dong 21] Dong, C., Li, Y., Gong, H., Chen, M. X., Li, J., Shen, Y., and Yang, M.: A Survey of Natural Language Generation, *ACM Computing Surveys*, Vol. 55, pp. 1 – 38 (2021)

[Gharibshah 20] Gharibshah, Z., Zhu, X., Hainline, A., and Conway, M.: Deep Learning for User Interest and Response Prediction in Online Display Advertising, *Data Science and Engineering*, Vol. 5, No. 1, pp. 12–26 (2020)

[Golobokov 22] Golobokov, K., Chai, J., Dong, V. Y., Gu, M., Chi, B., Cao, J., Yan, Y., and Liu, Y.: DeepGen: Diverse Search Ad Generation and Real-Time Customization, in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 191–199 (2022)

[Hughes 19] Hughes, J. W., Chang, K., and Zhang, R.: Generating Better Search Engine Text Advertisements with Deep Reinforcement Learning, in Teredesai, A., Kumar, V., Li, Y., Rosales, R., Terzi, E., and Karypis, G. eds., *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2019)*, pp. 2269–2277 (2019)

[Kamigaito 21] Kamigaito, H., Zhang, P., Takamura, H., and Okumura, M.: An Empirical Study of Generating Texts for Search Engine Advertising, in Kim, Y., Li, Y., and Rambow, O. eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers (NAACL-HLT 2021)*, pp. 255–262 (2021)

[Mita 24] Mita, M., Murakami, S., Kato, A., and Zhang, P.: Striking Gold in Advertising: Standardization and Exploration of Ad Text Generation (2024)

[Murakami 22] Murakami, S., Zhang, P., Hoshino, S., Kamigaito, H., Takamura, H., and Okumura, M.: Aspect-based Analysis of Advertising Appeals for Search Engine Advertising, in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track (NAACL-HLT 2022)*, pp. 69–78 (2022)

[Murakami 23] Murakami, S., Hoshino, S., and Zhang, P.: Natural Language Generation for Advertising: A Survey (2023)

[Niu 20] Niu, T. and Hou, Y.: Density Matrix Based Convolutional Neural Network for Click-Through Rate Prediction, in *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pp. 46–50 (2020)

[Sundararajan 17] Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic Attribution for Deep Networks, in Precup, D. and Teh, Y. W. eds., *Proceedings of the 34th International Conference on Machine Learning, (ICML 2017)*, Vol. 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328 (2017)

[Touvron 23a] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G.: LLaMA: Open and Efficient Foundation Language Models (2023)

[Touvron 23b] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T.: Llama 2: Open Foundation and Fine-Tuned Chat Models (2023)

[Vaswani 17] Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.: Attention is All you Need, in *Neural Information Processing Systems* (2017)

[Warstadt 19] Warstadt, A. and Bowman, S. R.: Grammatical Analysis of Pretrained Sentence Encoders with Acceptability Judgments, *arXiv preprint arXiv:1901.03438* (2019)

[Yang 22] Yang, Y. and Zhai, P.: Click-through Rate Prediction in Online Advertising: A Literature Review, *Information Processing and Management: an International Journal*, Vol. 59, No. 2 (2022)

著 者 紹 介

張　培楠

2018 年に AI Lab に中途入社し，リサーチサイエンティストとして広告文の自動生成や効果予測など，自然言語処理技術の広告分野適用についての研究開発に従事．

AdParaphrase: 魅力的な広告表現の分析を目的とした広告文言い換えデータセット

村上 聡一朗 Soichiro Murakami
株式会社サイバーエージェント AI 事業本部 AI Lab
Research scientist
murakami_soichiro@cyberagent.co.jp

張 培楠 Peinan Zhang
株式会社サイバーエージェント AI 事業本部 AI Lab
Research scientist
zhang_peinan@cyberagent.co.jp

上垣外 英剛 Hidetaka Kamigaito
奈良先端科学技術大学院大学
kamigaito.h@is.naist.jp

高村 大也 Hiroya Takamura
東京科学大学
takamura@pi.titech.ac.jp

奥村 学 Manabu Okumura
東京科学大学
oku@pi.titech.ac.jp

keywords: 自然言語処理，大規模言語モデル，広告文生成，言い換え，選好チューニング

Summary

広告の成功には人々を惹きつける効果的な言葉選びが欠かせない．本研究は広告文の言語表現に焦点を当て，どのような言語的特徴を持つ広告文が好まれるか明らかにすることを目的とし，選好評価データ付きの広告文言い換えデータセット AdParaphrase を提案する．AdParaphrase は広告文の言い換えペアから構成され，選好評価データを含む．これにより人々が魅力的に感じる広告表現の分析が可能となる．実験では広告文の言語的特徴量と選好評価データの関係を分析し，魅力的な広告文の特徴を明らかにした．またこれらの知見や提案データセットを活用し，魅力的な広告文を生成する手法を探求した．

1. はじめに

広告の目的は人々の注意を引き付け，クリックや購入等の行動を促すことである．そのためには人々の興味関心を引く内容を書くことが重要だが，それだけではない．その内容をどう表現して伝えるか，すなわち，広告の言語表現も広告の成功には欠かせない．本研究では広告文を魅力的にする言語表現に焦点を当て，どのような表現を持つ広告文が好まれ魅力的と感じるか明らかにすることを目的とする．

しかし言語表現の魅力度を分析する上で 2 つの課題に直面する．1 つ目は広告文の内容と表現の切り分けの難しさである．例えば，広告文「アディダス 50%OFF」と「ナイキ 半額」の比較で後者が好まれた場合，ブランド名（ナイキ）という内容が要因か「50%OFF」や「半額」等の表現が要因かを特定することは容易ではない．2 つの広告文の表現の差異に着目するためには内容を統一した上で比較する必要がある．2 つ目は広告文の魅力度を分析するためのオープンデータセットが不足している点である．一般に広告文の評価ではクリック率等の実績値や選好評価が用いられる．しかし広告文の選好評価データはこれまで公開されておらず，広告魅力度に寄与する要因を分析する上で障壁となっている [*1][Murakami 23].

表 1　AdParaphrase の例．() 内は各文を選好した人数．

初回購入で最大 50%割引 (0) ↔ **初回購入最大 50%オフ** (9)
【公式】マイナビバイト (8) ↔ マイナビバイト公式 (0)
業界一の安さ (3) ↔ **業界トップクラスの低価格** (7)

そこで本研究では，これら 2 つの課題を解決するために選好評価データ付きの広告文言い換えデータセット AdParaphrase を提案する [*2]．表 1 に例を示す．AdParaphrase は表現が異なるが同じ内容の広告文ペア，すなわち，言い

*1　特にクリック率などの実績値は多くの企業にとって秘匿情報に当たり，それが障壁の要因の一つとなっている．
*2　構築したデータセットは公開予定である．

換えペアから構成され，各ペアに対して 10 名の選好評価データが付与されている．これにより人々が魅力的に感じる広告表現の分析が可能となる．実験では AdParaphrase を用いて各文に含まれる言語的特徴量と選好評価データの関係を分析し，魅力的と感じる広告文の特徴を明らかにした（3·1 節）．さらに分析で明らかになった知見や選好評価データを活用し，魅力的な広告文を生成する手法を探求した （3·2 節）．その結果，分析で得られた知見が広告文の魅力度改善に寄与することを示した．

2. 言い換えデータセットの構築

選好に影響を与える広告表現を分析に向けて，内容は同一だが表現が異なる広告文ペアからなる言い換えデータセット AdParaphrase を構築する．本データセットにより，文ペアに対するペアワイズ比較を実施することで広告表現の違いに焦点を当てた選好評価データの収集や分析が可能となる．本章では始めに AdParaphrase の設計方針（2·1 節）を説明し，データセット構築手順である言い換え候補の収集 (2·2 節)，言い換え判定 (2·3 節)，選好評価データの収集 (2·4 節) を解説する．本研究では言い換え元の広告文として，広告文データセット CAMERA [Mita 24] に含まれる 16,365 件の広告文を用いる．

2·1 設 計 方 針

データセット構築では 3 つの設計方針を定めた: (1) 分析やモデル学習に活用できるデータセット規模; (2) 多様な言い換え現象を含む; (3) 一般に公開可能かつ研究開発目的で自由に使えるライセンス．

(1) の理由として，データ分析の信頼性を高めるためには十分なサンプル数が必要な点や分析結果を活用したモデル学習（例えば，広告文生成モデルの学習）ではある程度のデータセット規模が求められる点が挙げられる．(2) は広範な言語的特徴や表現を分析対象に含めることを念頭に置いたものであり，例えば「語順の変更」といった簡易な言い換え現象や偏った表現がデータセットを支配することを防ぐことを目的としている．(3) はこれまで一般利用可能なデータセットの不足等により留まっていた魅力的な広告表現の分析や生成に関する研究開発を推進することを目的としている．

2·2 言い換え候補の収集

言い換え候補の収集は，(1) 広告ライターによる言い換え例の作成，(2) 大規模言語モデル（LLM）やクラウドワーカーによる言い換え候補の作成の 2 段階で実施した．言い換え作成は理想的には広告ライターに依頼することが考えられるがリソースの限界により大規模には難しい．そこで広告ライターには広告文の魅力的な言い換え例の作成を依頼し，それらを参考事例として LLM やクラウドワーカーにより大量の言い換え候補を作成する方法を採用した．これにより広告ライターの知見を活用しながら大量の言い換え候補を作成・収集する．

i. 広告ライターによる言い換え例の作成

言い換え例の作成を広告制作の経験が豊富な広告ライター 2 名に依頼した．広告ライターには，原文をより魅力的な表現に言い換えること，全角 15 文字以内の文長制約 *3 を遵守することを指示した．100 件程度の言い換え例を作成するように依頼し，133 件の言い換え例が得られた．言い換え元の文は，CAMERA の開発セットから抽出した．

ii. LLM による言い換え候補の作成

LLM では文脈内学習 [Brown 20] により言い換え候補を作成する．具体的には広告ライターが作成した言い換え例を Few-shot 事例として与え，CAMERA に含まれる全ての広告文に対する言い換え文を生成した．また，設計方針（2·1 節）の 1 つである言い換えの多様性を考慮するために，プロンプトには言い換え生成におけるスタイル条件を加えた．具体的には「ひらがなを多く使ってください」や「よりシンプルな構文にしてください」などの条件を全 40 種類を定義し，各文の生成時に無作為選択した 1 つを用いた．この指示により LLM がスタイル条件を踏まえた多様な言い換え文を生成することを期待する．付録 A に LLM に与えたプロンプト，40 種類のスタイル条件を示す．なお，言い換え生成に使用する LLM は学習データやモデルサイズが異なる複数のモデルを使用した．これは [Shaib 24] らの LLM 生成文の構文は事前学習データに由来するといった報告を考慮し，単一のモデルだけでなく事前学習データが異なる複数のモデルで生成することでより多様な表現をもった言い換え文が生成されることを狙ったためである．言い換え生成に使用した LLM は設計方針 (3) を念頭に各モデルのライセンスを考慮したうえで 4 つのモデルを選定した．表 2 に各 LLM により生成した言い換え候補の数量を示す．

iii. クラウドワーカーによる言い換え候補の作成

クラウドワーカーによる作成ではクラウドソーシングを活用し，LLM と同じ指示文や言い換え例をガイドラインとして提示した．ワーカーの多くは広告制作の経験が乏しいと予想されるため，広告制作の知見をガイドラインに加えた．例えば，「行動を促す言葉を使う」や「重要な情報を先頭に書く」である．付録 B にクラウドワーカーに提示したガイドラインを示す．言い換え元には CAMERA から無作為抽出した 5,000 件を使用した．表 2 にクラウドワーカーが作成した言い換え候補数を示す．

2·3 言 い 換 え 判 定

言い換え候補 70,460 件が言い換えかを判定するアノテーションを実施した．まず作業効率化とデータ品質の

*3 一般的にウェブ広告では入稿する広告文に文長制約が設けられている．本研究では Google 広告等の検索連動型広告における見出し文の文長制約である全角 15 文字を採用した．

表 2　言い換えペアの収集結果

モデル	生成結果	フィルタ後	言い換え	言い換え通過率 (%)	言い換えを 8 名以上が魅力的と判定 (%)
CALM2-7B [Ishigami 23]	16,365	2,107	1,173	7.2	22.9
CALM3-22B [Ishigami 24]	16,365	6,287	4,551	27.8	21.4
Swallow-8B [Okazaki 24b, Fujii 24]	16,365	4,942	3,623	22.1	20.9
Swallow-70B [Okazaki 24a, Fujii 24]	16,365	5,226	4,174	25.5	19.5
Crowdworker	5,000	3,775	2,939	**58.8**	**25.8**
合計	70,460	22,337	16,460	23.4	21.7

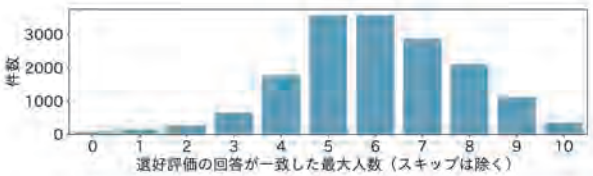


図 1　言い換えペアに対する選好評価の分布

向上のためルールによるフィルタリングを実施した．これには (1) 明らかに言い換えではない事例（例: 日付や金額が異なる）と (2) 文長制約を超過する事例を予め除外する目的がある．ウェブ広告では文長制約を超える文は入稿できないため，文長制約を満たす広告文を作成することが求められる．そのため実用上の観点から (2) を分析対象外にすることが妥当と判断し，除外した．

言い換え判定はクラウドソーシングにより各文ペアに対して 5 名で実施した．評価者には文ペアを提示し，言い換えとして成立するか二値で判定するよう依頼した．各事例の最終的なラベルは多数決で決定した．表 2 にルールによるフィルタリングと言い換え判定の結果を示す．16,460 件が言い換えペアとして判定された．Inter-Annotator Agreement（IAA）は 0.442（Fleiss' Kappa [Fleiss 71]）と中程度の一致である．

表 2 に各モデルの言い換え通過率を示す．言い換え通過率は生成文のうちフィルタリングおよび人手による言い換え判定を通過した割合を表す．クラウドワーカーが作成した言い換え文の通過率が最も高く，LLM はパラメータサイズが大きい CALM3-22B および Swallow-70B が高い傾向であることが分かった．また表 2 によると各モデルの生成文の多くはルールによるフィルタリングを通過できていない．そこでフィルタリングで言い換え候補が除外された要因を調査したところ，文長制約の超過が最も多い要因であることが判明した．この結果からより多くの言い換え候補を収集するためにはモデルの文長制御性能の改善が必要であると示唆された．

2·4　選好評価データの収集

言い換えペア 16,460 件に対して選好評価を実施した．評価はクラウドソーシングにより各文ペアに対して 10 名で実施した．評価者には各文ペアを提示し，より魅力的と感じる広告文を選択するよう依頼した．2 つの広告文の魅力度が同じ場合は「スキップ」を選択するように指示した．また選好評価の主観的な性質に配慮し，Wang らのガイドライン [Wang 21] を参考に評価観点を複数例示した．例えば「クリックしたいか」「理解しやすいか」等の観点がある．さらに，提示する広告文ペアの位置バイアスを軽減するために各広告文の表示位置はランダムに並び替えた．

図 1 に評価結果のヒストグラムを示す．X 軸は評価者 10 名中で 2 つの広告文に対する選好評価の回答が一致した最大人数を示す（スキップは除く）．例えば，6 は 10 名中 6 名が同じ広告文を選好し，0 は全員がスキップしたことを表す．評価結果のヒストグラムによると選好評価の回答が一致した最大人数は 5 名から 6 名が多いことが確認できる．この結果から広告文の言い換えペアに対する選好評価は全体的には回答が一致しづらいことが分かった．評価データ全体の IAA は 0.167（Fleiss'Kappa）であり，わずかな一致に留まっている [Landis 77]．一方，評価データ全体の約 22%を占める 3,570 件では 10 名中 8 名以上の選好が一致した．これらの IAA は 0.480（Fleiss' Kappa）であり，中程度の一致である．この結果は広告文ペアの表現の差異が評価者の選好に影響を与えた可能性があることを示唆している．

表 2 に各モデルが生成した言い換え文のうち，評価者 10 名中 8 名以上が魅力的と判定した割合を示す．例えば CALM2-7B では，言い換え文 1,173 件のうち約 22.9% の事例が評価者 8 名以上に魅力的と判定されたことを表す．各モデルを比較をすると，クラウドワーカーが作成した言い換え文が魅力的と判定される割合が最も高いことが分かった（25.8%）．LLM については CALM2-7B, CALM3-22B がやや高い傾向であることが分かった．またモデルのパラメータサイズに基づく比較では，CALM2-7B/CALM3-22B および Swallow-8B/Swallow-70B 間のそれぞれで大きな差は見受けられなかった．さらにクラウドワーカーと LLM を比較すると，前述の通りクラウドワーカーの言い換え文が魅力的と判定される割合は高かったものの，大きな差ではないことが確認できる．このことから LLM による言い換え候補の作成は人間にはやや劣るものの有用な手段であることが示唆される．

表 3　χ 二乗検定の結果

	特徴量		df	N	χ^2	p 値
基本特徴量	文長	文字	1	2,925	721.25	< 0.01
		単語	1	2,725	678.43	< 0.01
語彙的特徴量	内容語	名詞	1	1,406	326.61	< 0.01
		動詞	1	535	6.94	< 0.01
		形容詞	1	99	0.88	0.35
		副詞	1	127	0.68	0.41
	語彙選択	単語頻度	1	2,657	70.54	< 0.01
		一般名詞	1	1,397	288.12	< 0.01
		固有名詞	1	152	7.58	< 0.01
	文字種	ひらがな	1	2,047	23.24	< 0.01
		カタカナ	1	601	42.57	< 0.01
		漢字	1	1,503	257.72	< 0.01
		記号	1	2,332	795.93	< 0.01
統語的特徴量	係り受け	深さ (最大)	1	1,914	16.93	< 0.01
		長さ (最大)	1	2,349	1.89	0.17
	その他	Perplxity	1	3,570	223.26	< 0.01
文体的特徴量	感情	Joy	1	693	70.18	< 0.01
		Anticipation	1	683	89.29	< 0.01
	その他	具体性	1	186	116.44	< 0.01
		括弧	1	1,667	1372.57	< 0.01

3. 実　　験

選好評価（2·4 節）により 3,570 件の事例で 10 名中 8 名以上の選好が一致することが分かった．そこで実験ではこれらの事例に焦点を当て，2 文のどのような表現の差異が選好に影響を与えたか分析する（3·1 節）．さらに分析で明らかになった知見を踏まえ，魅力的な広告文の生成手法を探求する（3·2 節）．

3·1　選好に影響を与える広告表現の分析

実験の目的は言い換えペアの選好評価に影響を与えた要因を明らかにすることである．特に言い換えペアの表現の差異に着目し，どのような言語的特徴を持つ広告文が好まれる傾向があるか分析する．

§1　特徴量

広告の役割は人々の注意を引き付け，商品やサービスに興味を向けることである．そのため広告文の視認性や可読性，情報量は魅力度改善において重要な観点である [Murakami 23, Wang 12]．実験では広告文の表現やスタイルに関する様々な特徴量を定義し，選好評価との関係を分析した．特徴量は基本特徴量，語彙的特徴量，統語的特徴量，文体的特徴量に大別される．

i.　基本特徴量

可読性や情報量に関連する基本的な特徴量として文長（文字数，単語数）を使用した．

表 4　選好評価と文字数のクロス集計表

		文字数が多い広告文	
		広告文 1	広告文 2
選好された広告文	広告文 1	1,308	549
	広告文 2	200	868

ii.　語彙的特徴量

語彙的特徴量は内容語の数, 語彙選択, 文字種である．内容語を多く含む文は情報量が高く魅力的と仮説を立てた．語彙選択についてはより一般的な単語を含む広告文の方が好まれると仮説を立て，BCCWJ [Maekawa 10] に基づく平均単語頻度を算出した．加えて一般名詞と固有名詞の数も算出した．また各文に含まれる文字種の数を算出した．

iii.　統語的特徴量

統語的特徴量は文全体やその一部の構造に関する特徴量である．具体的には係り受け木の深さや依存リンクの長さ，Perplexity を含む．

iv.　文体的特徴量

文体的特徴量は広告文の言い回しやスタイルに関する特徴量である．本研究ではよりポジティブまたは具体的な広告文は好まれると仮定し，各文の感情と具体性に関するラベルを導入した．これらのラベルは感情と具体性を判別する独自の分類器を構築し判定した．分類器の詳細は付録 C を参照されたい．さらに，広告文で広く用いられる括弧記号（【】,「」）の有無も特徴量として用いる．これらの括弧記号は「【公式サイト】ABC 保険」のように視認性向上や重要情報の強調に使用される．

§2　分析方法

各特徴量と選好評価の関係を分析するために，独立性の χ 二乗検定を使用した．本手法は，2 つのカテゴリ変数の独立性を検証するものであり，本研究では (1) 多数の評価者に好まれた広告文と (2) 各特徴量の大小関係の関連性を検証する．例えば Perplexity の場合，好まれた広告文とスコアの関係を分析する．また分析では広告文ペアの表現の差異に着目するために各特徴量のスコアが異なる文ペアを分析対象とした．よって特徴量ごとに事例数が異なる．例えば文字数が異なる文ペアは 2,925 件である．

§3　実験結果

表 3 に χ 二乗検定の結果を示す．分析の結果，複数の特徴量が選好評価と有意な関係を持つことが明らかとなった（$p < 0.01$）．例えば，文字数や名詞の数，係り受けの深さ，Perplexity，括弧の有無などの特徴量が選好評価と有意な関係を持つことが分かった．一方で形容詞や副詞の数，係り受け木の依存リンクの長さについては有意な関係は見られなかった．

さらに各特徴量と選好評価のクロス集計表に基づき，各特徴量の大小関係と選好評価の関係を分析した．表 4

に広告文ペアに対する選好評価と文字数のクロス集計表を示す．ここで広告文 1 は言い換え元の広告文，広告文 2 は LLM またはクラウドワーカーにより作成された言い換え広告文（2·2 節）である．例えば表 4 では広告文ペアのうち広告文 1 の文字数が多く，選好評価で好まれた事例が 1,308 件存在したことを表す．以上の分析を各特徴量に対して実施し，例えば次のような特徴を持つ広告文が好まれる傾向があることが分かった: 文字数が多い，名詞の数が多い，係り受けの深さが小さい，Perplexity が低い（流暢性が高い），括弧記号を含む．以上の結果から，魅力的な広告文を作成するためにはこれらの特徴量を考慮することが重要であると示唆される．

3·2 広告文自動生成

各特徴量と選好評価の関係分析（3·1 節）により明らかになった知見を踏まえ，魅力的な広告文を生成する手法を探求する．本実験では与えられた広告文を情報の追加削除なしで魅力的な表現に言い換える広告文生成タスクに焦点を当てる [Youngmann 20, Mishra 20].

§ 1 実験設定

i. 生成手法

LLM を用いた生成手法を探求する．分析で得られた知見や評価者の選好を LLM に導入する方法は複数考えられるが，広告文生成においてどの手法が有用かは明らかではない．そこで実験では文脈内学習，指示チューニング [Wei 22]，Direct preference optimization（DPO）[Rafailov 23] を例として，各学習手法に基づく生成手法を比較する．文脈内学習では入力文を魅力的な表現に言い換えるように基本的な指示のみを与える zeroshot，分析で明らかになった知見を加えた zeroshot-findings, 言い換え例を複数与えた fewshot-findings の 3 パターンのプロンプトを検証した．付録 D に fewshot-findings のプロンプトを示す．fewshot-findings に与えた言い換え例は学習データから 20 件サンプリングした．このうち，少数の評価者が選好した広告文を入力，多数が選好した広告文を出力とした．また，魅力的な広告文を作成するための知見として文字数の多さ，流暢性の高さ（Perplexity の低さ），括弧記号の利用の 3 点をプロンプトに記載した．指示チューニングでは選好評価で少数の評価者に選好された広告文を入力，多数に選好された広告文を出力として追加学習した．また指示チューニング済みのモデルに対して DPO で選好チューニングを実施した．指示チューニングおよび DPO では QLoRA [Dettmers 23] を用い，1 epoch 学習した．実装は本コード *4 を使用した．モデルは CALM3-22B [Ishigami 24], Swallow70B [Fujii 24], GPT-4o [OpenAI 24b] を用いた．また zeroshot と同じ指示でクラウドワーカーが作成した言い換えも評価する．

*4 https://github.com/ghmagazine/llm-book

ii. 選好データセット

実験では AdParaphrase を選好チューニング向けに再構成したデータを使用する．具体的には，まず言い換え原文 x と 2 つのモデルが生成した言い換え文 y_1, y_2 の三つ組（x, y_1, y_2）を作成する．その後 y_1 と y_2 に対して新たに 10 名の選好評価データを収集した．これにより原文 x と選好評価付きの生成文 y_1^{pref}, y_2^{pref} からなる三つ組データ（$x, y_1^{pref}, y_2^{pref}$），合計 8,721 件を構築した．なお，学習, 開発, 評価用に 9:0.5:0.5 の比率で分割した．

iii. 評価方法

生成文を 4 つの観点で評価した．具体的には生成文が (1) 言い換えか，(2) 魅力的か，(3) 魅力的かつ文長制約を満たすか，(4) 予測クリック率（pCTR）の観点で評価した．(1) と (2) は 2·3 節及び 2·4 節と同手順で人手評価した．(1) は過半数が言い換えと判定した生成文の割合を算出する．(2) は過半数が言い換えと判定した生成文のうち，魅力的と判定された生成文の割合を報告する．例えば評価データ 200 件のうち 150 件が言い換えと判定された場合，150 件を分母としてそのうち過半数が魅力的と判定した割合を算出する．これは魅力的な広告文の生成能力に焦点を当てた評価を実施するためである．(3) は過半数が言い換えと判定した事例のうち，魅力的と判定かつ文長制約の全角 15 文字を満たす生成文の割合を算出する．ウェブ広告では魅力的な広告文であっても文長制約を超えた場合は入稿できないため，実用上の観点から魅力的かつ文長制約を満たす割合を評価する．(4) は自社で開発した広告文のクリック率予測モデルで入力文と生成文の pCTR を取得し，入力文に対する生成文の pCTR の改善率の平均を算出する．そのため 100%を超える場合は，生成文の pCTR が入力文よりも高いことを表す．なお，人手評価には評価データからサンプリングした 200 件を使用した．

§ 2 実験結果

表 5 に実験結果を示す．各評価観点について考察する．生成文が言い換えと判定された割合は指示チューニングにより追加学習した手法が高い傾向があった．文脈内学習では fewshot-findings が zeroshot に比べて改善する傾向を確認した．生成文が魅力的と判定された割合については DPO で選好チューニングしたモデルが高い傾向にあることを確認した．文脈内学習では zeroshot に対して知見を加えた zeroshot-findings が生成文の魅力度が高い傾向だった．これにより言語的特徴量と選好評価の関係性分析（3·1 節）で明らかになった知見が生成文の魅力度改善に寄与することが示唆された．一方で生成文が文長制約を満たしかつ魅力的と判定された割合については zeroshot-findings が最も高い傾向だった．DPO では文長制約を超えた文を生成する傾向があり，生成文の情報量が増えたことで魅力的と判定される割合が高くなったと考えられる．また pCTR ではほとんどの手法で改善は見られなかった．改善が確認できたのは GPT-4o-zeroshot-

表 5　広告文生成実験の評価結果（%）

モデル	言い換え	魅力度	魅力度&文長	pCTR
CALM3-22B				
zeroshot	74.0	23.0	12.8	97.2
zeroshot-findings	74.0	42.6	23.0	97.7
fewshot-findings	85.0	38.8	**31.2**	99.7
instruct-zeroshot	**90.5**	31.5	29.3	99.9
dpo-zeroshot	70.5	**84.4**	8.5	94.7
Swallow70B				
zeroshot	90.5	15.5	8.3	97.9
zeroshot-findings	80.0	44.4	17.5	99.1
fewshot-findings	86.5	40.5	**26.0**	99.5
instruct-zeroshot	**94.0**	18.6	17.6	99.7
dpo-zeroshot	62.5	**71.2**	8.0	95.7
GPT-4o				
zeroshot	86.0	12.8	12.8	99.1
zeroshot-findings	**95.5**	**39.3**	**34.6**	**100.7**
fewshot-findings	92.5	37.8	33.5	**100.3**
Crowdworker	89.1	23.9	22.3	99.3

findings，GPT-4o-fewshot-findings の 2 つであるが，その改善率は限られている．今後の課題として，選好評価だけではなくクリック率の改善にも寄与する言語的特徴量の考慮などが考えられる．

表 6 に生成文における各言語的特徴量を示す．表 6 には文脈内学習で知見としてプロンプトに記載した Perplexity（流暢性），文字数の多さ，括弧の有無の 3 点を示す．この結果から魅力度評価（表 5）で魅力度が高いと判定されたモデルは各特徴量のスコアが優れている傾向があることを確認できる．特に DPO では生成文の Perplexity の低さや文字数の多さが特徴的であり，これらの要因により生成文が魅力的と判定されやすかったことが推察できる．

表 7 に Swallow-70B [Fujii 24, Okazaki 24a] およびクラウドワーカーによる広告文の生成例を示す．これらの生成文のうち，評価者の過半数が魅力的と判定した文は Swallow70B-dpo-zeroshot だった．それ以外の生成文では言い換え元の入力文が魅力的と判定された．Swallow70B-dpo-zeroshot では他の生成文よりも文長が長く，他の文とは情報量が異なることから多くの評価者に好まれたと考えられる．

4. 分　　析

実験では選好評価に影響を与える言語的特徴を明らかにし，魅力的な広告文を生成する手法を探求した．広告の目的は人々の注意を引き付け，クリック等の行動を促すことである．そのため魅力的な広告表現へ言い換えることが実際のクリック等の行動にどのように影響を与えるかを明らかにすることは重要である．そこでまず本章では次の 2 つの分析を通してこの課題に取り組む．1 つ目の分析では広告文の言い換えペアに対する選好評価データと各広告文の pCTR の関係の分析する（4·1 節）．2 つ目の分析では 3·2 節の広告文生成手法で生成された広告文を実際に配信するオンライン評価を通して，魅力的な広告表現へ言い換えることがクリック等の行動に影響を与えるかを分析する（4·2 節）．

加えて本章では既存の自動評価指標と人手評価の関係を分析する（4·3 節）．本研究のデータセット構築や広告文生成実験における言い換え判定および魅力度評価は全て人手評価に依存している．しかしながら今後これらの研究をより大規模かつ効率的に推進するためには人手評価では限界があることから自動評価の導入が欠かせない．そのため本章では 3 つ目の分析として，既存の自動評価指標が人手による言い換え判定や魅力度評価の代替手段として有効であるかどうかを検討する．具体的には，広告文生成実験で得られた人手評価結果と既存の自動評価指標との関係を分析し，その有用性を評価する．

表 6　生成文の言語的特徴量

モデル	Perplexity↓	文字数↑	括弧（%）↑
CALM3-22B			
zeroshot	155.6	27.5	5.0
zeroshot-findings	157.6	30.7	64.5
fewshot-findings	146.7	27.0	**69.0**
instruct-zeroshot	168.5	24.1	48.5
dpo-zeroshot	**92.2**	**42.3**	37.0
Swallow70B			
zeroshot	158.3	27.7	13.0
zeroshot-findings	129.3	32.9	**89.0**
fewshot-findings	116.8	29.7	63.5
instruct-zeroshot	170.7	23.7	39.0
dpo-zeroshot	**70.5**	**42.4**	42.0
GPT-4o			
zeroshot	236.9	21.5	34.5
zeroshot-findings	228.9	25.0	**100.0**
fewshot-findings	**183.8**	**25.7**	73.0
Crowdworker	264.3	23.8	45.8
入力文	169.7	23.6	39.5

表 7　広告文生成実験の生成例．

モデル	生成例
Swallow-70B	
zeroshot	2022 年版おすすめクレカ 5 選
zeroshot-findings	【2022 年版】おすすめクレカ 5 選を紹介
fewshot-findings	【2022 年版】おすすめクレカ 5 選を紹介
instruct-zeroshot	【2022 年版】おすすめクレカ 5 社
dpo-zeroshot	【2022 年版】おすすめクレジットカード 5 選比較
Crowdworker	2022 年人気のクレカ TOP5
入力文	【2022 年版】おすすめクレカ 5 選

4·1　選好評価と pCTR の関係

本分析では言い換えペア (x_1, x_2) の pCTR と選好評価の大小関係を比較し，優劣が一致するかを確認する．例えば，

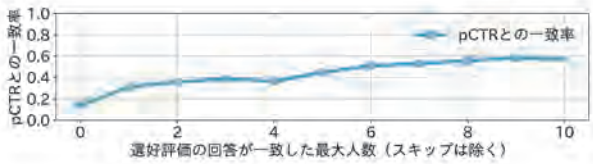


図 2　pCTR と選好評価の一致率

広告文ペア (x_1, x_2) の pCTR の大小関係が $\text{pCTR}(x_1) > \text{pCTR}(x_2)$ の時，各広告文の選好評価が $\text{HumanPref}(x_1) > \text{HumanPref}(x_2)$ であるかを確認する．ここで $\text{pCTR}(x)$, $\text{HumanPref}(x)$ はそれぞれ広告文 x の pCTR，広告文 x を選好した評価者数を表す．各広告文の pCTR は自社で独自に開発した広告文のクリック率予測モデルで取得する．

図 2 に AdParaphrase における選好評価データと pCTR の関係を示す．ここで x 軸は選好評価の回答が一致した最大人数であり，y 軸は pCTR と選好評価の大小関係が一致した割合を表す．例えば x 軸が 10 の時，評価者 10 名の選好が一致した事例のうち，その広告文の pCTR が他方よりも高い事例の割合を表す．また x 軸が 0 は全員がスキップを選択した場合であり，y 軸は $\text{pCTR}(x_1) = \text{pCTR}(x_2)$ である割合を示す．図 2 から選好評価の回答が一致する人数が多い場合，選好評価と pCTR の大小関係が一致しやすい傾向であることを確認した．例えば選好評価で評価者 10 名の選好が一致する場合，評価者が選好した広告文の pCTR の方が他方の広告文よりも高い割合は 56.9% だった．一方で選好評価の回答が一致する人数が少ない場合（多くの評価者がスキップを選択する場合），pCTR と選好評価の大小関係は一致しづらいことを確認した．ここで図 2 におけるピアソンの相関係数は 0.946 であり，強い相関関係が確認された．これらの結果から言い換えペアに対する選好評価データと pCTR には関係があることが分かった．したがって広告文をより多くの評価者にとって魅力的な表現に改善することはクリックなどの行動にポジティブな影響を与える可能性が推察される．

4·2　オンライン評価

オンライン評価では 3·2 節の文脈内学習に基づく fewshot-findings[*5] で生成した広告文を配信し，魅力的な表現へ言い換えることがクリック等の行動に影響を与えるかを分析する．具体的には既存の配信中の広告グループ（既存広告）をベースラインとして，既存広告を fewshot-findings により言い換えた広告グループ（生成広告）と比較する A/B テストを実施した．本検証では Google 広告 [*6] を評価プラットフォームとして利用し，2 週間の配信を実施する．また言い換え手法は既存広告における見出し文（15 文）に適用した．オンライン評価の評価対象として，事前に了承が得られた 2 社の広告グループ，合計 3 つの広告グループを利用した．広告グループの業種ドメインは人材系および教育系である．

*5　本検証ではモデルとして GPT-4 を利用した．
*6　https://ads.google.com/

表 8　オンライン評価の結果

広告グループ	閲覧数↑	クリック数↑	コスト↑
人材系 広告グループ A	1.628	1.233	1.531
教育系 広告グループ B	0.912	0.843	0.740
教育系 広告グループ C	0.627	0.365	0.401

表 9　システムレベルのメタ評価

	言い換え判定			魅力度評価		
評価指標	r	ρ	τ	r	ρ	τ
BLEU-4	**0.948**	0.950	0.831	-0.707	-0.484	-0.410
ROUGE-1	0.279	0.277	0.199	0.138	0.204	0.065
ROUGE-2	0.162	0.275	0.167	0.061	-0.113	-0.110
ROUGE-L	0.239	0.306	0.260	0.197	0.159	0.051
BERTScore	0.927	0.934	0.805	-0.769	-0.511	-0.385
GPT-4o	**0.948**	**0.965**	**0.895**	**0.886**	**0.758**	**0.615**

表 8 にオンライン評価の結果を示す．ここで既存広告と生成広告の閲覧数，クリック数，コストを比較する．表 8 では既存広告に対して生成広告が改善した割合を表しており，1 より高い場合は生成広告の各指標が改善されたことを表す．表 8 から生成広告における各指標の改善は広告グループによって異なることが分かった．例えば人材系の広告グループ A では全ての指標で既存広告よりも改善しているが，教育系の広告グループ B および C では改善が見られなかった．この結果から既存広告を魅力的な広告表現に言い換えることがクリック等の行動に影響を与えるかは業種や広告グループに依存することが推察できる．しかし一般的に広告配信によるオンライン評価は配信時期や業種・サービスのトレンドなど様々な外的要因の影響を受ける．そのため今後の方向性としてより大規模なオンライン評価の実施などが考えられる．

4·3　自動評価指標と人手評価の関係

本節では本研究で実施した人手評価（言い換え判定，魅力度評価）の代替手段として既存の自動評価指標が有効であるかどうかを検討する．具体的には，広告文生成実験（3·2 節）における人手評価結果と既存の自動評価指標とのシステム単位の相関関係を分析し，その有用性を評価する．相関係数には Pearson の積率相関係数（p），Spearman の順位相関係数（ρ），Kendall の順位相関係数（τ）を用いる．自動評価指標として BLEU-4 [Papineni 02]，ROUGE-1，ROUGE-2，ROUGE-L [Lin 04]，BERTScore [Zhang 20]，LLM（GPT-4o）による評価を採用した．ここで BLEU や ROUGE は参照文との n-gram の一致に基づく評価指標であり，BERTScore は参照文との意味的な類似度に基づく埋め込みベースの評価指標である．いずれも言い換え生成タスクの自動評価指標として広く用いられていることから採用した [Shen 22, Zhou 21].

ROUGE や BERTScore は F1 スコアを報告する．また近年の LLM によるテキスト自動評価の研究 [Liu 23, Gu 25] に着想を得て，LLM による評価も検証した．具体的にはモデルに GPT-4o を用い，言い換え判定と魅力度評価の人手評価ガイドラインをそれぞれプロンプトとして与えて生成文を評価した．よって BLEU，ROUGE，BERTScore は参照あり評価であり，LLM による評価は参照なし評価である．なお参照あり評価では，3·2 節の人手で作成した言い換え文を参照文として用いた．

表 9 に自動評価結果と人手評価結果の相関関数を示す．また各モデルの自動評価スコアは付録 E に示す．まず言い換え判定については言い換え生成タスクの自動評価指標として広く用いられる BLEU や BERTScore に加えて GPT-4o による評価も人手評価と強い正の相関があることが確認できた．また魅力度評価について GPT-4o による評価が人手評価と強い正の相関であったが，BLEU や BERTScore については負の相関が確認された．以上の結果から言い換え判定には BLEU, BERTScore, GPT-4o による評価が有用であり，魅力度評価については GPT-4o による評価が有用であることが示唆される．

5. 関 連 研 究

5·1 魅力的な広告文の分析

広告の成功のためには広告の魅力度や広告効果に影響を与える要因を明らかにすることは重要である．これまで広告の魅力度に影響を与える要因を分析する様々な研究が行われてきた．例えば，広告の訴求軸 [Murakami 22]，説得戦略 [Yuan 23]，感情表現 [Youngmann 20] など様々な観点における分析が取り組まれている．本研究では広告の言語表現，特に言語的特徴量に焦点を当てた分析に取り組んでいる．

広告効果や魅力度に影響を与える分析では実際の人々の行動を反映したクリック率や閲覧数といった実績値（ログデータ）に基づいた分析が一般的である．しかしこれら実績値は多くの企業にとって秘匿情報でありこれまで公開されておらず，学術研究のためのデータセット不足の問題が指摘されていた [Murakami 23]．この問題に対して本研究ではクリック率や閲覧数などのログデータの代替として広告文に対する選好評価データを収集し，選好評価データ付きの広告文データセットを研究用途として一般公開する．

本研究に関連した研究として，[Pryzant 18] らの広告文のライティングスタイル（文体）が広告効果に与える影響を調査した取り組みが挙げられる．この研究では文体と広告効果の因果関係を調査するために交絡因子の影響を抑える分析手法を提案した．具体的には広告文に含まれるブランド名が広告効果に影響を与える問題を指摘し，この影響を抑えた分析に取り組んでいる．しかしながらブランド名以外の交絡因子（例えば，価格やキャンペーン情報などの要因）は考慮されていない．これに対して本研究では意味的内容が同等である言い換えペアを用いることで交絡因子の影響を抑え，広告文の文体や言語表現に着目した分析が可能である．

5·2 広告文自動生成

インターネット広告の需要の増加により，大量に広告の運用や制作を担う広告代理店（広告制作者）では，これまでの手作業による制作・改善フローに限界が近づいている．そのため近年では広告文自動生成の研究に注目が集まっている [Murakami 23]．広告の目的は人々の注意を引き付けて商品やサービスに興味を持ってもらい，クリックや購入といった行動を促すことである．よって広告の成功のためには魅力的な広告文の制作が重要である．

広告文の自動生成手法は長年研究されており，テンプレートに基づく手法 [Bartz 08, Thomaidou 13] やニューラルネットワークに基づく手法 [Hughes 19, Kamigaito 21] など様々な手法が提案されている．本研究では広告文の言語表現やスタイルの魅力度に焦点を当てている点がこれまでの研究と大きく異なる．先行研究ではクリック率や選好評価により生成された広告文の魅力度を評価しているが，広告文の内容や表現といった要因がどのように魅力度に影響したか詳細な分析は進められていない．一方で本研究では広告文の言い換えペアを用いることで広告の言語表現に焦点を当てた分析や魅力度改善のための生成手法の比較検討が可能である．

6. お わ り に

本研究では魅力的な広告表現の分析を目的とした広告文言い換えデータセット AdParaphrase を提案した．広告文の言語的特徴と選好評価データの関係の分析を通して魅力的な広告文が持つ複数の特徴量を明らかにした．また魅力的な広告文を生成するための手法を探求した．

本研究の今後の方向性としていくつか考えられる．まず選好評価に影響を与える言語的特徴量の分析では，評価者の属性情報を考慮した分析やその他の言語表現やスタイルに関する要因を含めた分析などが挙げられる．広告文に対する選好評価は評価者の属性，例えば性別や年齢によって傾向が異なることが予想される．そのため評価者の属性情報などを追加的に収集し，評価者の属性と選好評価の関係を調査することが方向性として考えられる．これにより特定の属性を対象としたターゲティング広告の改善などの応用が期待できる．加えて，広告文の自動生成手法の改善も方向性の一つとして挙げられる．例えば DPO に基づく手法における文長制約の違反事例が多い問題に対して，文長制約の導入や強化学習による文長報酬モデル [Kamigaito 21] の導入などが考えられる．また文脈内学習に基づく手法では few-shot 事例の選定方法や様々なプロンプトの検証などが考えられる．

◇参 考 文 献◇

[Bartz 08] Bartz, K., Barr, C., and Aijaz, A.: Natural language generation for sponsored-search advertisements, in *Proceedings of the 9th ACM Conference on Electronic Commerce*, pp. 1–9 (2008)

[Brown 20] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D.: Language Models are Few-Shot Learners, in Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. eds., *Advances in Neural Information Processing Systems 33*, Vol. 33, pp. 1877–1901 (2020)

[Dettmers 23] Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L.: QLORA: efficient finetuning of quantized LLMs, in *Advances in Neural Information Processing Systems 36* (2023)

[Fleiss 71] Fleiss, J., et al.: Measuring nominal scale agreement among many raters, *Psychological Bulletin*, Vol. 76, No. 5, pp. 378–382 (1971)

[Fujii 24] Fujii, K., Nakamura, T., Loem, M., Iida, H., Ohi, M., Hattori, K., Shota, H., Mizuki, S., Yokota, R., and Okazaki, N.: Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities, in *First Conference on Language Modeling* (2024)

[Gu 25] Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, Y., and Guo, J.: A Survey on LLM-as-a-Judge (2025)

[Hughes 19] Hughes, J. W., Chang, K.-h., and Zhang, R.: Generating Better Search Engine Text Advertisements with Deep Reinforcement Learning, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2269–2277 (2019)

[Ishigami 23] Ishigami, R.: cyberagent/calm2-7b-chat (2023), Hugging Face

[Ishigami 24] Ishigami, R.: cyberagent/calm3-22b-chat (2024), Hugging Face

[Kajiwara 21] Kajiwara, T., Chu, C., Takemura, N., Nakashima, Y., and Nagahara, H.: WRIME: A New Dataset for Emotional Intensity Estimation with Subjective and Objective Annotations, in Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2095–2104 (2021)

[Kamigaito 21] Kamigaito, H., Zhang, P., Takamura, H., and Okumura, M.: An Empirical Study of Generating Texts for Search Engine Advertising, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pp. 255–262 (2021)

[Landis 77] Landis, J. R. and Koch, G. G.: The Measurement of Observer Agreement for Categorical Data, *Biometrics*, Vol. 33, No. 1, pp. 159–174 (1977)

[Lin 04] Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries, in *Proceedings of the ACL Workshop: Text Summarization Branches Out*, pp. 74–81 (2004)

[Liu 23] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C.: G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment, in Bouamor, H., Pino, J., and Bali, K. eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522 (2023)

[Maekawa 10] Maekawa, K., Yamazaki, M., Maruyama, T., Yamaguchi, M., Ogura, H., Kashino, W., Ogiso, T., Koiso, H., and Den, Y.: Design, Compilation, and Preliminary Analyses of Balanced Corpus of Contemporary Written Japanese, in Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D. eds., *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pp. 1483–1486 (2010)

[Mishra 20] Mishra, S., Verma, M., Zhou, Y., Thadani, K., and Wang, W.: Learning to Create Better Ads: Generation and Ranking Approaches for Ad Creative Refinement, in *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pp. 2653–2660 (2020)

[Mita 24] Mita, M., Murakami, S., Kato, A., and Zhang, P.: Striking Gold in Advertising: Standardization and Exploration of Ad Text Generation (2024)

[Murakami 22] Murakami, S., Zhang, P., Hoshino, S., Kamigaito, H., Takamura, H., and Okumura, M.: Aspect-based Analysis of Advertising Appeals for Search Engine Advertising, in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pp. 69–78 (2022)

[Murakami 23] Murakami, S., Hoshino, S., and Zhang, P.: Natural Language Generation for Advertising: A Survey (2023)

[Okazaki 24a] Okazaki, N., Mizuki, S., Ma, Y., Maeda, K., Hattori, K., Ohi, M., Shiotani, T., Saito, K., Yokota, R., Fujii, K., Nakamura, T., Okamoto, T., Shigeki, I., and Takamura, H.: tokyotech-llm/Llama-3.1-Swallow-70B-Instruct-v0.1 (2024), Hugging Face

[Okazaki 24b] Okazaki, N., Mizuki, S., Ma, Y., Maeda, K., Hattori, K., Ohi, M., Shiotani, T., Saito, K., Yokota, R., Fujii, K., Nakamura, T., Okamoto, T., Shigeki, I., and Takamura, H.: tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.1 (2024), Hugging Face

[OpenAI 24a] OpenAI, : GPT-4 Technical Report (2024)

[OpenAI 24b] OpenAI, : Hello GPT-4o (2024), Accessed: 2025-01-03

[Papineni 02] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J.: BLEU: A Method for Automatic Evaluation of Machine Translation, in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318 (2002)

[Pryzant 18] Pryzant, R., Basu, S., and Sone, K.: Interpretable Neural Architectures for Attributing an Ad's Performance to its Writing Style, in Linzen, T., Chrupała, G., and Alishahi, A. eds., *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 125–135 (2018)

[Rafailov 23] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C.: Direct preference optimization: your language model is secretly a reward model, in *Advances in Neural Information Processing Systems 36* (2023)

[Shaib 24] Shaib, C., Elazar, Y., Li, J. J., and Wallace, B. C.: Detection and Measurement of Syntactic Templates in Generated Text, pp. 6416–6431 (2024)

[Shen 22] Shen, L., Liu, L., Jiang, H., and Shi, S.: On the Evaluation Metrics for Paraphrase Generation, in Goldberg, Y., Kozareva, Z., and Zhang, Y. eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3178–3190 (2022)

[Thomaidou 13] Thomaidou, S., Lourentzou, I., Katsivelis-Perakis, P., and Vazirgiannis, M.: Automated Snippet Generation for Online Advertising, in *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, pp. 1841–1844 (2013)

[Wang 12] Wang, H.-C. and Pomplun, M.: The attraction of visual attention to texts in real-world scenes, *Journal of Vision*, Vol. 12, No. 6, pp. 26–26 (2012)

[Wang 21] Wang, X., Gu, X., Cao, J., Zhao, Z., Yan, Y., Middha, B., and Xie, X.: Reinforcing Pretrained Models for Generating Attractive Text Advertisements, in *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 3697–3707 (2021)

[Wei 22] Wei, J., Bosma, M. P., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V.: Finetuned Language Models are Zero-Shot Learners, in *The Tenth International Conference on Learning Representations* (2022)

[Yamada 20] Yamada, I., Asai, A., Shindo, H., Takeda, H., and Matsumoto, Y.: LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention, in Webber, B., Cohn, T., He, Y., and Liu, Y. eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6442–6454 (2020)

[Youngmann 20] Youngmann, B., Yom-Tov, E., Gilad-Bachrach, R.,

and Karmon, D.: The Automated Copywriter: Algorithmic Rephrasing of Health-Related Advertisements to Improve Their Performance, in *Proceedings of The Web Conference 2020*, pp. 1366–1377 (2020)

[Yuan 23] Yuan, Y., Xu, F., Cao, H., Zhang, G., Hui, P., Li, Y., and Jin, D.: Persuade to Click: Context-Aware Persuasion Model for Online Textual Advertisement, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 35, No. 2, pp. 1938–1951 (2023)

[Zhang 20] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y.: BERTScore: Evaluating Text Generation with BERT, in *The Eighth International Conference on Learning Representations* (2020)

[Zhou 21] Zhou, J. and Bhat, S.: Paraphrase Generation: A Survey of the State of the Art, in Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5075–5086, Association for Computational Linguistics (2021)

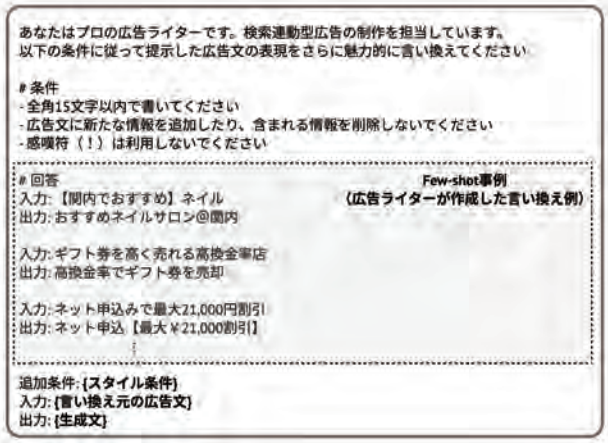


図 **A.1** LLM による言い換え候補の作成のプロンプト

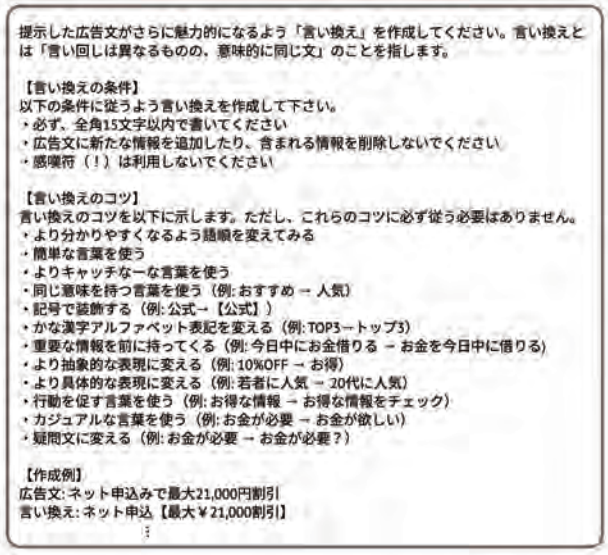


図 **B.2** クラウドワーカーに提示した言い換え作成のガイドライン

◇ 付　　録 ◇

A. LLM による言い換え候補の作成

図 A.1 に LLM による言い換え候補の作成（2·2 節）で使用したプロンプト，表 A.1 にスタイル条件 40 種類を示す．

B. 言い換え作成のガイドライン

言い換え候補の作成（2·2 節）でクラウドワーカーに提示したガイドラインを図 B.2 に示す．

C. 言語的特徴量の詳細

広告文の感情及び具体性ラベルを判定する独自の分類器を構築した．各分類器の詳細を以下に示す．

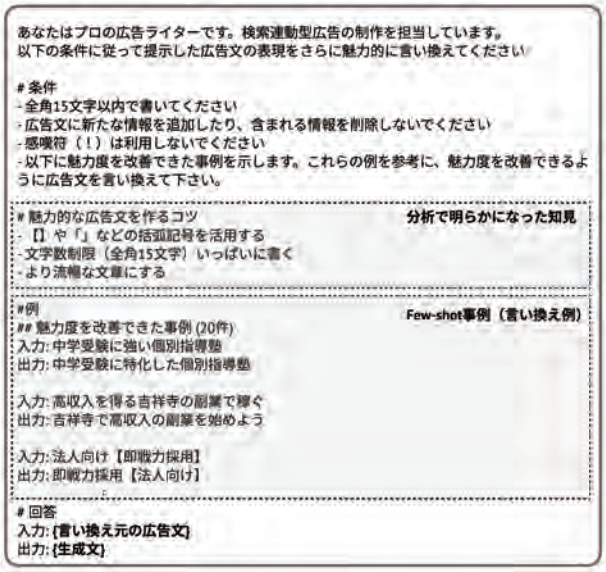
i. 感　情

各広告文の感情ラベルを判定するために WRIME [Kajiwara 21] で学習された LUKE [*7] [Yamada 20] を用いた．本モデルは 8 つの感情ラベル（joy, sadness, anticipation, surprise, anger, fear, disgust, trust）の中から最も相応しいラベルを予測する 8 クラス分類器である．本研究ではポジティブな表現を持つ広告文は好まれやすいと仮説を

*7 https://huggingface.co/Mizuiro-sakura/luke-japanese-large-sentiment-analysis-wrime

表 A.1　LLM による言い換え候補作成で使用したスタイル条件

番号	条件	番号	条件
(1)	ひらがなを多く使ってください	(21)	内容語を使ってください
(2)	カタカナを多く使ってください	(22)	一般的な言葉を使ってください
(3)	漢字を多く使ってください	(23)	専門用語を使ってください
(4)	ニュース記事の見出しのように書いてください	(24)	肯定的な言葉を使ってください
(5)	より具体的な表現を使ってください	(25)	否定的な言葉を使ってください
(6)	より抽象的な表現を使ってください	(26)	中立的な言葉を使ってください
(7)	第一人称代名詞を使ってください	(27)	堅い言葉を使ってください
(8)	第二人称代名詞を使ってください	(28)	カジュアルな言葉を使ってください
(9)	第三人称代名詞を使ってください	(29)	重要な情報を文章の左側に配置してください
(10)	より興奮を感じさせる表現を使ってください	(30)	重要な情報を文章の右側に配置してください
(11)	より喜びを感じさせる表現を使ってください	(31)	より複雑な構文にしてください
(12)	より安心感を感じさせる表現を使ってください	(32)	よりシンプルな構文にしてください
(13)	より緊急感を感じさせる表現を使ってください	(33)	疑問文にしてください
(14)	より行動を促す表現を使ってください	(34)	簡単な言葉を使ってください
(15)	【】や「」等の括弧を使ってください	(35)	難しい言葉を使ってください
(16)	数字を使ってください	(36)	ユーザーが得られるベネフィットを強調してください
(17)	動詞を使ってください	(37)	問題解決の方法を示してください
(18)	形容詞を使ってください	(38)	キャッチコピーを入れてください
(19)	名詞を使ってください	(39)	見やすい表現にしてください
(20)	副詞を使ってください	(40)	読みやすい表現にしてください


あなたはプロの広告ライターです。検索連動型広告の制作を担当しています。
以下の条件に従って提示した広告文の表現をさらに魅力的に言い換えてください

条件
- 全角15文字以内で書いてください
- 広告文に新たな情報を追加したり、含まれる情報を削除しないでください
- 感嘆符（！）は利用しないでください
- 以下に魅力度を改善できた事例を示します。これらの例を参考に、魅力度を改善できるように広告文を言い換えて下さい。

魅力的な広告文を作るコツ　　分析で明らかになった知見
- 【】や「」などの括弧記号を活用する
- 文字数制限（全角15文字）いっぱいに書く
- より流暢な文章にする

#例　　Few-shot事例（言い換え例）
魅力度を改善できた事例 (20件)
入力: 中学受験に強い個別指導塾
出力: 中学受験に特化した個別指導塾

入力: 高収入を得る吉祥寺の副業で稼ぐ
出力: 吉祥寺で高収入の副業を始めよう

入力: 法人向け【即戦力採用】
出力: 即戦力採用【法人向け】

回答
入力: {言い換え元の広告文}
出力: {生成文}

図 D.3　広告文生成実験で使用したプロンプト

立て，joy 及び anticipation と判定された広告文を分析に使用した．WRIME の評価データに対する正解率は 68.6%である．

ii.　具体性

本研究ではより具体的に書かれた広告文は好まれやすいと仮説を立て，広告文の具体性を判定する分類器を GPT-4 [OpenAI 24a] により構築した．本モデルは与えられた 2 つの広告文（言い換えペア）のうち具体性の高い広告文（広告 1 または広告 2）を選択する．具体性が同じ場合は「equal」を選択する．したがって，3 クラス分類器である．予測結果の中からランダムに選択した 100 件を人手評価したところ，正解率は 88.0%であった．

D.　広告文生成実験のプロンプト例

図 D.3 に広告文生成実験（3·2 節）で使用したプロンプト（fewshot-findings）を示す．

E.　広告文生成実験の自動評価指標

表 E.2 に広告文生成実験（3·2 節）の自動評価結果を示す．

著　者　紹　介

村上　聡一朗

2022 年東京工業大学工学院情報通信系情報通信コース博士後期課程修了，博士（工学）．2021 年株式会社サイバーエージェントに入社，AI Lab NLP チームにて自然言語処理，特に広告文の自動生成や広告効果の分析に関する研究に従事．言語処理学会，ACL 各会員．

表 **E.2**　広告文生成実験の自動評価結果

モデル	**BLEU-4**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BERTScore**	**GPT-4o**	
						言い換え	魅力度
CALM3-22B							
zeroshot	27.4	29.3	9.8	29.2	86.5	75.0	33.0
zeroshot-findings	30.2	30.8	10.5	31.0	86.8	66.5	49.0
fewshot-findings	40.0	32.0	10.5	31.3	89.5	79.0	25.5
instruct-zeroshot	46.8	32.4	11.0	32.3	90.3	89.5	11.0
dpo-zeroshot	15.9	29.8	10.5	29.7	81.5	59.5	80.0
Swallow-70B							
zeroshot	41.8	31.0	11.5	31.2	89.2	90.0	11.5
zeroshot-findings	37.8	31.5	10.3	31.0	87.8	78.0	32.5
fewshot-findings	44.5	32.5	10.5	31.6	89.2	83.0	18.0
instruct-zeroshot	50.5	29.4	10.7	29.0	90.9	90.5	6.5
dpo-zeroshot	20.1	30.0	10.5	29.7	82.8	61.5	65.5
GPT-4o							
zeroshot	37.7	27.1	9.2	26.9	88.1	90.0	3.0
zeroshot-findings	48.0	31.2	9.5	31.0	90.7	92.0	5.5
fewshot-findings	49.5	32.3	10.9	31.5	91.0	90.5	7.5

訓練不要な条件付きテキスト埋め込み

山田 康輔
Kosuke Yamada
株式会社サイバーエージェント AI Lab NLP チーム
Research Scientist
yamada_kosuke@cyberagent.co.jp

張 培楠
Zhang Peinan
株式会社サイバーエージェント AI Lab NLP チーム
Research Scientist
zhang_peinan@cyberagent.co.jp

keywords: テキスト埋め込み、意味的類似度算出、テキストクラスタリング

Summary

条件付きテキスト埋め込みは、特定の側面に焦点を当てたテキストの埋め込み表現であり、与えられた条件に基づくテキスト同士の類似度の算出を可能にする。従来手法は、大規模な訓練データによる指示学習や意味的テキスト類似度算出タスクによる微調整が求められ、開発コストが高い。そこで本研究では、生成型 LLM をテキストエンコーダとして条件付き一語制約プロンプトを用いる、訓練不要で高品質な条件付きテキスト埋め込み PonTE を提案する。条件付き意味的類似度テキスト類似度とテキストクラスタリングによる二つの実験を通じて、提案手法は追加の訓練なしで従来手法以上の性能を達成することを示す。

1. はじめに

テキスト間の類似度は類似文検索や文書クラスタリングなどにおいて重要な役割を果たす NLP タスクの一つであり [Agirre 12, Marelli 14, Cer 17]、効率的かつ類似度の算出を実現するために、テキストの埋め込み表現が一般的に使用される。ただし、従来のテキスト埋め込み手法 [Conneau 17, Cer 18, Reimers 19, Gao 21] は、一つのテキストに対して一つの汎用的な埋め込み表現を生成するものが主要であるが、テキストには多様な側面があることから、想定する類似度の算出が困難な場合がある。たとえば、表 1にあるようなレビューテキストの場合、T_1 と T_2 は類似したカテゴリの商品について言及しているものの、その感情極性は異なる。その一方で、T_1 と T_3 は異なるカテゴリの商品について言及しているが、どちらも肯定的な評価をしている。これらの事例では、着目する側面という条件を与えることなく、類似度の高低を判断することは難しい。

このような背景から、特定の側面に焦点を当ててテキストを埋め込む「条件付きテキスト埋め込み手法」が提案されている [Deshpande 23, Yoo 24]。しかし、これらの従来手法は、埋め込みモデルを訓練するための特定の条件に関するテキスト同士の類似度をアノテーションしたデータセットを必要とし、現在、英語の画像キャプションデータに付与されたデータしか存在しないため、分野や言語を超えた NLP タスクへの適用は容易ではない。また、テキスト埋め込み用に指示学習された手法も、タスクごとに指示文を与えるため条件付きテキスト埋め込みとして活用できるが、大規模な訓練データを整備し、長時間に渡って訓練する必要があり、開発コストが高い [Su 23, Li 23, Wang 24b]。

T_1: This camera is one of my favorites.
T_2: This smartphone cannot capture high-quality images.
T_3: Best fish I have ever had.

表 1 レビューテキストの例

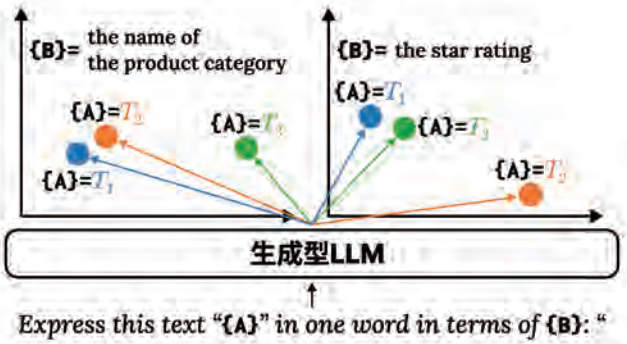


図 1 PonTE による条件付きテキスト埋め込みの可視化例。T_1、T_2、T_3 は表 1の事例に対応する。

そこで本研究では、訓練不要で様々な分野や言語に適用可能な条件付きテキスト埋め込み手法 **PonTE** (**P**rompt-based **Con**ditional **T**ext **E**mbedding) を提案する。概要を図 1に示す。Jiang ら [Jiang 23b] が提案した一語制約プロンプトを拡張した条件付き一語制約プロンプトを用い、与えられた条件を満たすようなテキスト埋め込みを、

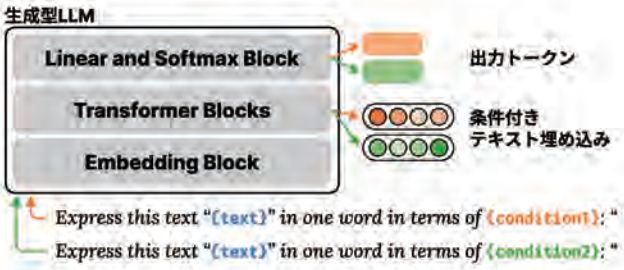


図 2　PonTE の概要

Mistral [Jiang 23a] や Llama-3 [Touvron 23] などの生成型 LLM から生成する。図 1は、表 1のテキストに対して PonTE を適用したときの埋め込み空間を可視化した結果を示し、条件に応じて事例間の距離が変わっていることを示している。

2. PonTE

従来の条件付きテキスト埋め込み手法は、特定の条件に関してテキスト間の類似度がアノテーションされたデータを用いて微調整する必要があり、分野や言語に依存しない汎用的な手法は探索されていない。また、テキスト埋め込み用に指示学習された手法も開発コストが高いため実現は容易ではない。そこで、開発コストの低い汎用的な手法を実現するため、プロンプトベースの条件付きテキスト埋め込み手法 PonTE を提案する。

PonTE の手法の概要を図 2に示す。PonTE はプロンプトを生成型 LLM に入力し、プロンプトの最終トークンに位置する Transformer ブロックの中間表現をテキスト埋め込みに用いる。PonTE では、PromptEOL [Jiang 23b] で使用される一語制約プロンプトを拡張した「条件付き一語制約プロンプト」を使用する。条件付き一語制約プロンプトには、情報を圧縮するための一語制約に加えて、圧縮する方向性を指定するための条件制約を導入する。たとえば、`{text}`を埋め込むテキスト本文、`{condition}`を条件として、条件付き一語制約プロンプトは「*Express this text* "`{text}`" *in one word in terms of* `{condition}`: "」になる。

また、プロンプトを生成型 LLM に入力し、Linear and Softmax ブロックの出力を通じて生成される一語を PonTE の埋め込みの解釈に利用する。この一語は「"」が生成されるまでの出力トークンを連結されたものを指し、予測結果の理解やプロンプトの決定に有用である。

3. 実　験: C-STS

PonTE が、条件に沿ったテキスト埋め込み表現が生成できているかを確認するために条件付き意味的テキスト類似度 (Conditional Semantic Textual Similarity; C-STS) の実験を行う。

手法	r_s	r_p
sup-SimCSE$_{\text{large}}$	3.4	4.1
GTE$_{\text{Qwen2-7B-Inst}}$	33.5	33.9
E5$_{\text{Mistral-7B-Inst}}$	**34.8**	**34.6**
unsup-SimCSE$_{\text{large}}$	2.3	1.7
PonTE$_{\text{Mistral-7B}}$	21.6	21.0
PonTE$_{\text{Mistral-7B-Inst}}$	30.6	28.9
PonTE$_{\text{Llama-3-8B}}$	21.7	19.7
PonTE$_{\text{Llama-3-8B-Inst}}$	**37.1**	**33.6**
PonTE$_{\text{Llama-3-70B}}$	11.3	10.9
PonTE$_{\text{Llama-3-70B-Inst}}$	35.1	31.0

表 2　C-STS の実験結果。上段が教師あり学習の手法、下段が教師なし学習の手法である。

3·1 実 験 設 定

データセットは、Deshpande ら [Deshpande 23] によって作成された C-STS データセットを用いた*1。各レコードは、二つのテキスト、条件、類似度ラベルから構成されている。予測時に使用した類似度は二つのテキストにおける埋め込み表現のコサイン類似度で、類似度ラベルと予測類似度から算出されるスピアマン順位相関係数 (r_s) とピアソン積率相関係数 (r_p) を指標として評価した。

PonTE でテキストエンコーダに用いる LLM として、Mistral 7B のベースモデルと指示学習モデル、Llama-3 8B のベースモデルと指示学習モデルを用いた*2。プロンプトテンプレートは複数の候補の中から各モデルで最もスピアマン順位相関係数の高いものを使用した*3。PonTE との比較手法として、追加の訓練を必要とする、指示学習された Qwen2 7B [Yang 24] と Mistral 7B を元に、それぞれテキスト埋め込み用にも指示学習された GTE [Li 23] と E5 [Wang 24a, Wang 24b] を用いた。また、SimCSE [Gao 21] の教師あり学習をしたモデルと教師なし学習をしたモデルを用いて、埋め込み対象のテキストと条件のテキストを連結する形式でモデルに入力して埋め込み表現を取得した。

3·2 実 験 結 果

実験結果を表 2に示す。PonTE はエンコーダに用いた LLM に関わらず高い性能を示した。PonTE$_{\text{Llama-3-8B-Inst}}$ が特に高いスコアを達成しており、スピアマン順位相関係数では追加訓練を必要とする手法を上回るなど、多大な開発コストをかけることなく性能の高い条件付きテキスト埋め込み手法を実現できることが示唆される結果となった。

同じ Mistral 7B をエンコーダに用いた E5$_{\text{Mistral-7B-Inst}}$ と PonTE$_{\text{Mistral-7B-Inst}}$ を比較すると、E5 が PonTE を上

*1　データセットの統計は付録 Dに示す
*2　使用したモデルは付録 Aに示す
*3　プロンプトテンプレートの候補は付録 Bに示す

	テキスト 1	テキスト 2	条件	ラベル	予測	生成語 1	生成語 2
(a)	T_{a1}: A group of elephants of different sizes walking together on dirt with a rock formation and trees in the background.	T_{a2}: One elephant is squirting water out of its mouth and the other is putting water into its mouth.	C_{a1}: the physical actions	1.0	1.30	Walking	water-squirting
			C_{a2}: the animal	5.0	4.77	Elephants	Elephant
(b)	T_{b1}: A man in a shirt and tie with his hands in his pockets leaning against a wall.	T_{b2}: The man is wearing a dress coat, suit and tie, but not dress pants.	C_{b1}: the attire of the person	2.0	4.52	Formal	Formal
			C_{b2}: the gender of the person	5.0	4.64	Male	Male

表 3 $\text{PonTE}_{\text{Llama-3-8B-Inst}}$ の出力例。「予測」は予測類似度を 0.5 から 5.5 で Min-Max 正規化を適用した値を示し、「生成語 1」と「生成語 2」は「テキスト 1」と「テキスト 2」のときの生成された一語を示す。

回っており、テキスト埋め込み用の指示学習が効果的であることがわかる。ただし、$\text{E5}_{\text{Mistral-7B-Inst}}$ は、15 万のユニークな指示文を用いて LLM で生成された 50 万の事例と、人手で作成された質問応答や検索のデータセットから収集した 180 万の事例を用いて訓練しており、その開発は容易ではない。

PonTE において、LLM のベースモデルと指示学習されたモデルによる手法を比較すると、後者が一貫して高いスコアを示した。これは、指示学習によってプロンプトの指示に従う能力が向上したことによるものだと考えられる。また、$\text{PonTE}_{\text{Llama-3-8B-Inst}}$ と $\text{PonTE}_{\text{Llama-3-70B-Inst}}$ では、$\text{PonTE}_{\text{Llama-3-70B-Inst}}$ の方が一般的に性能が高いとされているにも関わらず、$\text{PonTE}_{\text{Llama-3-8B-Inst}}$ の方がスコアが高かった。Jiang ら [Jiang 23b] の様々なパラメータ数のエンコーダを元にした PromptEOL を用いて STS の実験を行った結果においても、スケールの増大は性能に直接寄与しない傾向があり、実験結果はこの傾向と一致する。

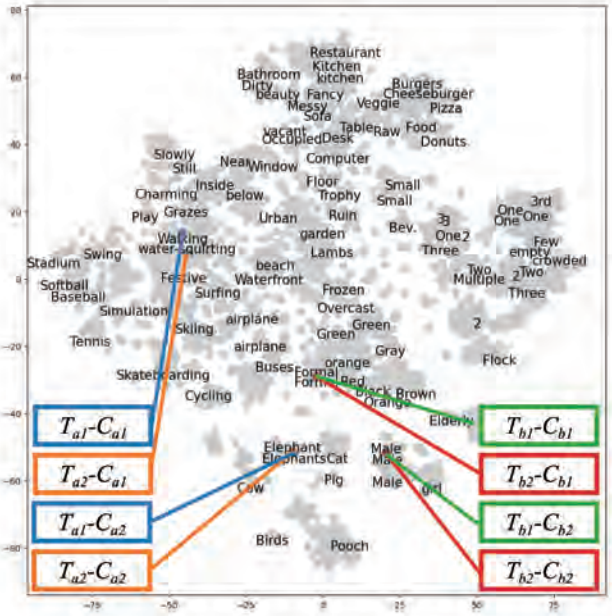


図 3 C-STS データセットにおける PonTE による埋め込みの二次元マッピング

3·3 分　　析

PonTE による条件付きテキスト埋め込みの挙動を明らかにするため、プロンプトを入力して生成された一語と埋め込みの二次元への射影結果を分析した。表 3に $\text{PonTE}_{\text{Llama-3-8B-Inst}}$ の出力例を示す。表には、二つのテキストに対して、二つの条件における類似度ラベルと予測類似度、生成された一語を示す。表 3 (a) より、PonTE は条件ごとに類似度ラベルと近い予測類似度を示していることが確認できる。生成された一語に関してもテキスト中の条件に関連した語を生成しており、さらに、生成された語の意味が似ているものは類似度が高く、そうではないものは低くなっている。この結果から、生成された語と予測類似度は関連が深いと考えられる。表 3 (b) では、類似度ラベルが低いにも関わらず、二つのテキストで同じ一語が生成されており、予測類似度が高くなっていることが確認された。これは、テキストを一語に要約するのが難しい事例では、的確な一語を生成できず、それに伴いその中間表現である条件付きテキスト埋め込みも適切に作用していないことが原因であると考えられる。

図 3に $\text{PonTE}_{\text{Llama-3-8B-Inst}}$ によるテキスト埋め込みを t-SNE によって二次元に射影した結果を示す。図中の点は、条件に基づきテキスト埋め込みを二次元に射影したベクトルを示し、生成された一語を付与している。表 3の二つの事例を色で強調している。図 3から、同じテキストであっても条件が異なれば、離れた位置にあることが確認でき、テキストの表層情報より条件に基づきテキストを埋め込んでいることがわかる。また、生成された一語を見ると、数量に関連する点は右側、生物は下側に集まっているなど、類似した意味の語は近い位置にあることが確認できる。人間の持つ語の意味の近さを表現したテキスト埋め込み空間が得られており、PonTE の有用性が確認できる。

4. 実験: テキストクラスタリング

条件付きテキスト埋め込みのもう一つの有力な応用先のテキストクラスタリングで実験を行った。

4·1 実験設定

条件付きテキスト埋め込みを用いることで、様々な側面を条件として埋め込み表現を生成することが可能である。その柔軟性を評価するため、以下の三つのデータセットを用いて複数の側面からテキストクラスタリングを行った。商品レビューのデータセットである Amazon reviews corpus はレビューに商品カテゴリ (Amazon-C) とレーティング (Amazon-R) のラベル、科学系の質問応答データセットである ScienceQA (SciQA) は質問文にトピックのラベル、Tweet emotion intensity dataset (Tweet Emotion) は X (旧 Twitter) の投稿に感情のラベルが付与されており、それぞれラベルごとにテキストクラスタリングを行う。クラスタリングには K-means を用い、クラスタ数にはデータセットごとにラベルの種類数を与えている。シード値を変えて五回実施し、評価指標を V-measure として各評価値の平均をスコアとする。

PonTE では、Mistral 7B と Llama-3 8B のそれぞれベースモデルと指示学習されたモデルの計四つのモデルを用いた。プロンプトテンプレートは、C-STS でスピアマン順位相関係数が最も高いものを用い、プロンプトテンプレートに挿入する条件は検証セットで V-measure の最も高いものを用いた*4。比較手法として、C-STS の実験と同じ GTE と E5、条件を与えないテキスト埋め込みとして SimCSE と PromptEOL を導入した。

4·2 実験結果

実験結果を表 4に示す。PonTE はすべてのデータセットで他の教師なし手法の性能を超え、C-STS の実験と同様に $\text{PonTE}_{\text{Llama-3-8B-Inst}}$ が全体的に高いスコアを示した。PonTE は、トピックのような全体的な意味でクラスタリングするときであっても、明示的に側面を指定することで性能が改善することを示唆している。

また、PonTE は教師あり手法と比較しても競争力のある高い性能を示している。特に、$\text{PonTE}_{\text{Mistral-7B-Inst}}$ や $\text{PonTE}_{\text{Llama-3-8B-Inst}}$ は、教師あり SimCSE よりすべてのデータセットで上回るスコアを示した。また、GTE や E5 はクラスタリング対象と類似したデータセットで訓練しているにも関わらず、PonTE はいくつかのデータセットで GTE や E5 を上回るスコアとなった。PonTE は、訓練せずとも有用な条件付きテキスト埋め込みを生成できるため、GTE や E5 がサポートしていない分野や言語であっても容易に応用できる可能性が見込まれる。

*4 条件の候補は付録 Cに示す

手法	Amazon -C	Amazon -R	SciQA	Tweet Emotion
$\text{sup-SimCSE}_{\text{large}}$	19.5	22.4	65.5	29.4
$\text{GTE}_{\text{Qwen2-7B-Inst}}$	**38.3**	36.8	73.9	36.8
$\text{E5}_{\text{Mistral-7B-Inst}}$	37.4	**37.6**	**74.0**	**41.3**
$\text{unsup-SimCSE}_{\text{large}}$	16.7	4.2	63.8	23.4
$\text{PromptEOL}_{\text{Mistral-7B}}$	8.6	27.2	66.0	6.5
$\text{PromptEOL}_{\text{Mistral-7B-Inst}}$	6.1	27.4	59.4	22.7
$\text{PromptEOL}_{\text{Llama-3-8B}}$	9.9	20.4	66.7	9.5
$\text{PromptEOL}_{\text{Llama-3-8B-Inst}}$	9.4	30.8	65.1	31.7
$\text{PonTE}_{\text{Mistral-7B}}$	27.7	27.7	**74.5**	18.1
$\text{PonTE}_{\text{Mistral-7B-Inst}}$	25.3	31.7	68.0	43.8
$\text{PonTE}_{\text{Llama-3-8B}}$	**30.9**	23.8	74.1	24.0
$\text{PonTE}_{\text{Llama-3-8B-Inst}}$	30.5	**34.1**	73.0	**45.9**

表 4 テキストクラスタリングの実験結果。上段が教師あり学習の手法、下段が教師なし学習の手法である。

5. まとめ

本研究では、訓練不要な条件付きテキスト埋め込み手法 PonTE を提案した。PonTE は、強力な LLM によるテキストエンコーダと条件付き一語制約プロンプトを用いることで、高品質な条件付きテキスト埋め込みを生成できることを実験的に示した。C-STS とテキストクラスタリングの実験では、PonTE は既存の教師なし手法の性能を超え、教師あり手法の性能と同等レベルの性能を達成することを示した。今後の展望として、PonTE の他の応用先や、PonTE が他の分野や言語に適用可能であるか検証したい。

◇参考文献◇

[Agirre 12] Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre, A.: SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity, in *Proceedings of *SEM-SemEval 2012*, pp. 385–393 (2012)

[Cer 17] Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L.: SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation, in *Proceedings of SemEval 2017*, pp. 1–14 (2017)

[Cer 18] Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., and Kurzweil, R.: Universal Sentence Encoder for English, in *Proceedings of EMNLP 2018*, pp. 169–174 (2018)

[Conneau 17] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A.: Supervised Learning of Universal Sentence Representations from Natural Language Inference Data, in *Proceedings of EMNLP 2017*, pp. 670–680 (2017)

[Deshpande 23] Deshpande, A., Jimenez, C., Chen, H., Murahari, V., Graf, V., Rajpurohit, T., Kalyan, A., Chen, D., and Narasimhan, K.: C-STS: Conditional Semantic Textual Similarity, in *Proceedings of EMNLP 2023*, pp. 5669–5690 (2023)

[Gao 21] Gao, T., Yao, X., and Chen, D.: SimCSE: Simple Contrastive Learning of Sentence Embeddings, in *Proceedings of EMNLP 2021*, pp. 6894–6910 (2021)

[Jiang 23a] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, de las D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P.,

Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E.: Mistral 7B, arXiv preprint: 2310.06825 (2023)

[Jiang 23b] Jiang, T., Huang, S., Luan, Z., Wang, D., and Zhuang, F.: Scaling Sentence Embeddings with Large Language Models, arXiv preprint: 2307.16645 (2023)

[Li 23] Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., and Zhang, M.: Towards General Text Embeddings with Multi-stage Contrastive Learning, arXiv preprint: 2308.03281 (2023)

[Marelli 14] Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R.: A SICK cure for the evaluation of compositional distributional semantic models, in *Proceedings of LREC 2014*, pp. 216–223 (2014)

[Reimers 19] Reimers, N. and Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in *Proceedings of EMNLP-IJCNLP 2019*, pp. 3982–3992 (2019)

[Su 23] Su, H., Shi, W., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., Yih, W.-t., Smith, N. A., Zettlemoyer, L., and Yu, T.: One Embedder, Any Task: Instruction-Finetuned Text Embeddings, in *Findings of ACL 2023*, pp. 1102–1121 (2023)

[Touvron 23] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G.: LLaMA: Open and Efficient Foundation Language Models, arXiv preprint: 2302.13971 (2023)

[Wang 24a] Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F.: Text Embeddings by Weakly-Supervised Contrastive Pre-training, arXiv preprint: 2212.03533 (2024)

[Wang 24b] Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F.: Improving Text Embeddings with Large Language Models, in *Proceedings of ACL 2024*, pp. 11897–11916 (2024)

[Yang 24] Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Yang, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Liu, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Guo, Z., and Fan, Z.: Qwen2 Technical Report, arXiv preprint: 2407.10671 (2024)

[Yoo 24] Yoo, Y., Cha, J., Kim, C., and Kim, T.: Hyper-CL: Conditioning Sentence Representations with Hypernetworks, in *Proceedings of ACL 2024*, pp. 700–711 (2024)

◇ 付　　録 ◇

A. PonTE で用いたモデルのリンク

PonTE は六つの LLM を用いて実験を行った。具体的には、Mistral 7B [Jiang 23a] のベースモデル*5と指示学習モデル*6、Llama-3 8B [Touvron 23] のベースモデル*7と指示学習モデル*8、Llama-3 70B のベースモデル*9と指示学習モデル*10を用いた。

B. プロンプトテンプレートの影響

表 B.1に $\text{PonTE}_{\text{Llama-3-8B-Inst}}$ の検証セットにおけるプロンプトテンプレートごとの C-STS の実験結果を示す。(1)-(6) は PromptEOL で使用されたものを発展したプロンプト、(7)-(12) は指示形式にしたプロンプトである。表 B.1から「in one word」がスコアに大きく影響していることがわかる。また、(7)-(12) は (1)-(6) より全体的にスコアが高く、指示形式にすることで、プロンプトの指示をより反映し、性能が高くなっていると考えられる。「in terms of」と「with respect to」による差は比較的小さかった。

*5 mistralai/Mistral-7B-v0.3
*6 mistralai/Mistral-7B-Instruct-v0.3
*7 meta-llama/Meta-Llama-3-8B
*8 meta-llama/Meta-Llama-3-8B-Instruct
*9 meta-llama/Meta-Llama-3-70B
*10 meta-llama/Meta-Llama-3-70B-Instruct

プロンプトテンプレート	r_s	r_p
{p} = This text: “{text}” means		
(1) {p} **in terms of {condition}**: “	18.8	17.1
(2) {p} **with respect to {condition}**: “	18.0	17.0
(3) {p} in one word **in terms of {condition}**: “	28.2	25.2
(4) {p} in one word with respect to {condition}: “	25.1	21.7
(5) {p} **in terms of {condition}** in one word: “	28.1	24.7
(6) {p} with respect to {condition} in one word: “	25.4	22.3
{p} = Express this text “{text}”		
(7) {p} **in terms of {condition}**: “	19.8	18.2
(8) {p} with respect to {condition}: “	19.1	17.6
(9) {p} in one word **in terms of {condition}**: “	**37.3**	**34.8**
(10) {p} in one word with respect to {condition}: “	36.4	33.9
(11) {p} **in terms of {condition}** in one word: “	33.1	30.4
(12) {p} with respect to {condition} in one word: “	30.4	27.9

表 B.1 プロンプトテンプレートごとの C-STS の実験結果

C. 条件の影響

表 C.2に各データセットの検証セットにおける条件ごとのテキストクラスタリングの実験結果を示す。Amazon-C や SciQA では、「name」や「product」、「question」を加えた条件のスコアが高く、Amazon-R でも「star」や「rating」だけよりも「star rating」を入れたもののスコアが高くなっている。これは条件を具体的にすることで LLM がプロンプトの意図をより反映した出力ができ、スコアが高くなったと考えられる。

	条件	V-measure
Amazon -C	(a) the category	22.8
	(b) the product category	26.5
	(c) the category name	21.6
	(d) the product category name	29.4
	(e) the name of the category	22.0
	(f) the name of the product category	**30.5**
Amazon -R	(g) the rating	30.4
	(h) the star	26.2
	(i) the star rating	**34.1**
	(j) the five-level rating	21.3
	(k) the five-level star rating	29.3
	(l) the emotion	33.4
SciQA	(m) the category	70.3
	(n) the question category	74.2
	(o) the name of the category	74.1
	(p) the name of the question category	**75.4**
Tweet Emotion	(q) the emotion	**45.9**
	(r) the feeling	44.7
	(s) the sentiment	43.6

表 C.2 条件ごとのテキストクラスタリングの実験結果

D. データセットの統計

C-STS の実験では、Deshpande ら [Deshpande 23] による C-STS データセット*11を用いた。テキストクラスタリングの実験では、Amazon reviews corpus*12、 ScienceQA*13、Tweet emotion intensity dataset*14の三つのデータセットを用いた。データセットの統計を表 D.3に示す。

データセット	ラベル数	検証セット	テストセット
C-STS	-	2,840	4,732
Amazon-C	31	5,000	5,000
Amazon-R	5	5,000	5,000
SciQA	25	4,241	4,241
Tweet Emotion	4	374	1,421

表 D.3 実験で使用したデータセットの統計

著 者 紹 介

山田 康輔

AI Lab リサーチサイエンティスト・名古屋大学大学院情報学研究科協力研究員。2024 年 3 月名古屋大学情報学研究科博士後期課程を修了し、博士（情報学）を取得。2024 年 4 月に AI Lab に新卒入社し、自然言語処理に関する研究開発に従事。「大規模言語モデル入門」「大規模言語モデル入門 II～生成型 LLM の実装と評価」を共著者として執筆。

張 培楠

2018 年に AI Lab に中途入社し、リサーチサイエンティストとして広告文の自動生成や効果予測など、自然言語処理技術の広告分野適用についての研究開発に従事。

*11 princeton-nlp/c-sts
*12 mexwell/amazon-reviews-multi
*13 derek-thomas/ScienceQA
*14 cardiffnlp/tweet_eval

JHARS: RAG設定における日本語Hallucination評価ベンチマークの構築と分析

亀井 遼平* Ryohei Kamei
東北大学 Researcher
ryohei.kamei.s4@dc.tohoku.ac.jp

坂田 将樹* Masaki Sakata
東北大学 Researcher
sakata.masaki.s5@dc.tohoku.ac.jp

邊土名 朝飛 Asahi Hentona
AI Lab, 株式会社 AI Shift Researcher
hentona_asahi@cyberagent.co.jp

栗原 健太郎 Kentaro Kurihara
株式会社 AI Shift ML Engineer
kurihara_kentaro@cyberagent.co.jp

乾 健太郎 Kentaro Inui
MBZUAI, 東北大学, 理化学研究所 Professor
kentaro.inui@mbzuai.ac.ae

keywords: Hallucination, RAG, LLM, データセット

Summary

大規模言語モデルの hallucination（与えられた情報源に存在しない内容を生成する現象）は，実応用上での重要な課題となっている．本研究では，日本語における hallucination 評価のための包括的なベンチマーク **JHARS**（**J**apanese **H**allucination **A**ssessment in **R**AG **S**ettings）を構築し，最新の GPT-4o を含む 3 つのモデルを対象に分析を行った．その結果，hallucination 発生率は低い一方，事実確認が必要な重大な hallucination が検出された．また，自動検出における高い適合率と再現率の両立は困難であるものの，重大な hallucination に関しては高い再現率で検出可能であることが示された．これは，LLM 自身による出力の検証が，ユーザへの事実確認支援として機能する可能性を示唆している．

1. はじめに

大規模言語モデル（LLM）は幅広い分野での活用が期待され，様々な業務プロセスの自動化や効率化に向けた検討が進められている．しかし，実用環境での LLM の適用における重要な課題として，hallucination の問題が指摘されている．一般に hallucination とは，与えられた情報源に存在しない内容を LLM が生成する現象を指す [Ji 23][Huang 24]．特に，検索拡張生成（Retrieval-Augmented Generation: RAG）の設定において，この問題は実用上の重要な課題となっている．図 1 に示すように，LLM の生成する応答は一見適切に見えるが，参考文献に記載のない情報が含まれることがある．それらの情報に対する事実確認や修正作業は，実務者のコスト増大を招く．

この問題に対処するための第一歩として，hallucination を定量的に評価するための包括的なベンチマークの整備が求められている．これまでにも，hallucination 評価のためのベンチマーク作成は行われてきた [Li 23][Liang 24][Niu 24][Mishra 24][Ji 24]．しかし，既存のベンチマークの多くは英語や中国語を対象としており，日本語における hallucination の評価データセットは十分に整備されていない．そのため，日本語に対応した LLM における hallucination の発生頻度や，その検出の困難さについては十分な知見が得られていない．この状況は，説明責任が求められる実務タスクへの LLM の導入を妨げる要因となっている．

そこで本研究では，RAG 設定における日本語 LLM の応答 450 件にアノテーションを行い，日本語 hallucination 評価ベンチマーク **JHARS** (**J**apanese **H**allucination **A**ssessment in **R**AG **S**ettings) を構築した．このベンチ

∗* Equal contribution

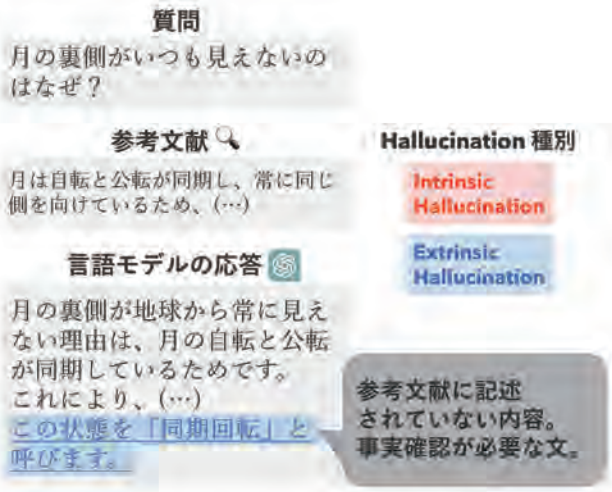


図 1　収集された hallucination の例.

マークを用いた分析の結果，LLM の応答文に対する hallucination 発生率は比較的低いものの，事実確認が必要な重大な hallucination が検出された．また，LLM の応答文に含まれる hallucination の自動検出において，高い適合率と再現率の両立は困難であることが明らかになった．一方で，事実確認が必要な重大な hallucination については高い再現率で検出可能であることが示された．この結果は，LLM 自身による生成結果の検証が，事実確認を要する箇所の特定と警告に有効である可能性を示唆している．

2. JHARS データセット

本研究の目的は，RAG 設定において日本語 LLM が生成する hallucination の種類と程度を調査することである．そのため，先行研究 [Niu 24][Mishra 24] に倣って RAG 設定での日本語 hallucination 評価ベンチマーク JHARS を作成した．

i. hallucination の定義

一般に，自然言語生成タスクにおける hallucination とは，生成された内容が，入力として与えられたテキストや参考文献に対して意味をなさない，または忠実でないことを指す [Huang 24]．この hallucination は Intrinsic Hallucination と Extrinsic Hallucination の 2 つの主要なタイプに分類することができる [Ji 23][Huang 24][Huang 23][Li 22]．Intrinsic Hallucination とは，LLM の出力が入力されたテキストや参考文献と矛盾していることを指す．Extrinsic Hallucination とは，入力されたテキストや参考文献から事実正誤性を検証できない LLM の出力を指す *1．本研究でもこちらの定義を採用してアノテーションを実施した *2．

*1　Intrinsic Hallucination と Extrinsic Hallucination の具体例については表 4 を参照されたい．

*2　なお本研究では常識的に正しい箇所は hallucination に含まないとした．(例：「岸田総理」＝「岸田文雄総理」であることは常識的に正しい．)

表 1　アノテーション時のアノテータのラベル一致率

	文数	割合
3 人一致	2130	90.37%
2 人一致	218	9.25%
一致なし	9	0.38%

2·1　データセットの構築

i.　応答生成のベースデータセット

我々は Wikipedia の記事検索を含む日本語の質問応答データセットである wikipedia-human-retrieval-ja[Baobab 24] を用いて応答の生成を行った．これは応答に必要な参考文献が与えられており，長文形式での応答が必要な点で本研究に適している．

ii.　LLM を用いた応答生成

hallucination 評価のためのデータセットの開発が活発になりつつある一方，それらの多くの場合で，特定の種類の hallucination を人為的に生成し，収集するためのテクニックが採用されている [Li 23][Liu 22][Longpre 21]．具体的には，hallucination が起こるようなプロンプトで指示したり，単に出力に矛盾を挿入したりするといったテクニックである．これは hallucination が発生する数を増やすのに有効であるものの，自然に発生した hallucination の分布と大きく異なることがある点が指摘されている [Niu 24]．我々は自然に発生する hallucination を評価するため，hallucination を人為的に発生させるテクニックは使用せず，むしろ参考文献以外の情報は使用しないように指示を与えた．

また，LLM の応答に hallucination が含まれる要因として，参考文献の品質と LLM の性能という 2 つの可能性が考えられる．本研究では，LLM の性能に起因する hallucination を評価するため，以下の 2 点を実施した．(i) 質問応答に必要な参考文献のみを使用した．(ii) 応答を生成する前に，GPT-4o[OpenAI 24a] を用いて参考文献が十分な情報を含んでいるかのチェックを行った．以上の 2 点により，RAG の設定において応答生成のための参考文献が正しく検索され，無関係なテキストは入っていないという前提で，LLM の応答に hallucination のラベルを付与する．

応答生成モデルは，GPT-4o，GPT-4o-mini*3[OpenAI 24b] に加え，Llama 3.1 の日本語機能を強化したものである Llama-3.1-Swallow-8B-Instruct-v0.1[Fujii 24][Okazaki 24] を使用した．GPT-4o，GPT-4o-mini は比較的最近にリリースされ広く社会で用いられているモデルである．Llama-3.1-Swallow-8B-Instruct-v0.1 はパラメータ数が小さい日本語 LLM がどの程度 hallucination を含む応答を生成するかを調査するために採用した．これらの 3 つのモデルを用いて，各モデル 150 件の質問文に対して応答（計 450 応答）を生成させた．

*3　GPT モデルは 2024 年 10 月時点でのモデルを使用．

図 2　データセット構築時に使用したフローチャート．

表 2　文単位の hallucination の統計情報．None は hallucination 無しを示す．括弧内は割合を表す．

モデル名	None	Intrinsic	Extrinsic	null (3 人のラベルが全て違う)	合計
GPT-4o-mini	959 (97.0%)	1 (0.1%)	24 (2.4%)	5 (0.5%)	989
GPT-4o	866 (97.9%)	0 (0.0%)	17 (1.9%)	2 (0.2%)	885
Llama-3.1-Swallow-8B-Instruct-v0.1	474 (98.3%)	0 (0.0%)	6 (1.2%)	2 (0.4%)	482

iii.　アノテーション

本研究のアノテーションのフローチャートを図 2 に示した．本研究では 1 文単位でアノテーションを付与した．hallucination に関するアノテーションを付与している先行研究 [Mishra 24][Dziri 22] に倣い，“はい/いいえ” で答えられる質問に複数回答えることで自動的にラベルが決まるようにした．

我々は，アノテータ間の合意を取ることを目的として，各文に 3 人のアノテータを割り当てた．なお，全体のアノテーションを実行する前に，アノテータに対し，ガイドラインを読んだ上での 10 件のテストアノテーションを実施した．その後，不明点や質問等に回答するための説明会を実施した．本研究におけるアノテータ間のラベルの一致率を表 1 に示した．表 1 より，全員不一致の割合は 0.38% と低く，妥当性の高いデータセットであると考えられる．

2·2　データセットの分析

定量的な分析として，本データセットのラベルの統計情報を表 2 に示した．Intrinsic Hallucination と Extrinsic Hallucination の発生件数は英語 hallucination 評価ベンチマークを構築している先行研究 [Niu 24][Mishra 24] と比較して少なかった．Intrinsic Hallucination の割合は特に少なく，GPT-4o-mini における 1 件のみであった．これらの要因として，RAG による応答生成のための参考文献が正しく検索された前提であるということと，LLM の性能が以前より向上していることが考えられる．また，応答単位の hallucination の発生件数を表 3 に示した．表 2，表 3 より，応答単位，文単位のいずれにおいても GPT-4o-mini の hallucination の割合が 3 つのモデルで最も高く，Llama-3.1-Swallow-8B-Instruct-v0.1 の割合が最も低かった．この原因として，Llama-3.1-Swallow-8B-Instruct-v0.1 の出力の文長が他のモデルに比べて短いことが考えられる．

表 3　応答単位の hallucination の発生件数（割合）

モデル名	件数 (割合)
gpt-4o-mini	16 (10.7%)
gpt-4o	11 (7.3%)
Llama-3.1-Swallow-8B-Instruct-v0.1	5 (3.3%)

定性的な分析として，hallucination の発生事例を 表 4 に示した．Intrinsic Hallucination の事例では，“14,534 人” の箇所は誤りであり，正しくは “約 1.6 万人” である．Extrinsic Hallucination の事例では，“この状態を「同期回転」と呼びます．” という箇所は参考文献には書いておらず，事実確認が必要である．Mishra らによると，Extrinsic Hallucination は，事実確認が必要となるものと，主観的な感想（例：“おもしろい”）に大別できる．本データセットにおいて，上記の違いについて著者らの目視で確認したところ，全応答中，平均約 1.6%が事実確認が必要となる事例となっていた．RAG 設定において参考文献が正しく検索されている前提を踏まえると，上記の hallucination 事例は文書検索処理の問題ではなく，LLM の生成特性に起因すると考えられる．このような応答はユーザに誤った知識を与える可能性があり，今後無くしていくべきである．

3.　実験：hallucination の自動検出

§2·2 節の分析により，最近の日本語 LLM であっても hallucination が一定数発生することが明らかになった．この結果は，LLM の実用において hallucination の検出が重要な課題となることを意味している．そこで本節では，「現状の日本語 LLM が起こす hallucination は自動検出可能か」という問いを検証する．実験の結果，日本語 LLM の生成結果に含まれる hallucination の自動検出において，高い適合率と再現率の両立は困難であることが明らかに

表 4　Intrinsic Hallucination（赤色塗りつぶし）と Extrinsic Hallucination（青色塗りつぶし）の事例.(...) は省略を表す.

応答モデル名：GPT-4o-mini
質問文：四万十町は何県ですか
参考文献：高知県西南部に位置し、高岡郡に属する約人口 1.6 万人（2021 年 8 月 31 日現在 [1]）の町。四万十川の中流に位置し、東は太平洋（土佐湾）に面し、(...)
応答文：四万十町は高知県に位置する町です。(...) 町の面積は 642.30 平方キロメートルで、2021 年 8 月 31 日現在の総人口は 14,534 人です。また、(...)

応答モデル名：GPT-4o
質問文：月の裏側がいつも見えないのはなぜ？
参考文献：月は自転と公転が同期し、常に同じ側を向けているため、(...)
応答文：月の裏側が地球から常に見えない理由は、月の自転と公転が同期しているためです。これにより、月は常に同じ側を地球に向けています。この状態を「同期回転」と呼びます。しかし (...)

表 5　hallucination 検出の F1 スコア（%）. Intrinsic Hallucination の事例は 1 件のみであることに注意.

	Generator: GPT-4o		Generator: GPT-4o-mini		Generator: Llama-3.1-Swallow-8B-Instruct-v0.1	
Detector	Intrinsic	Extrinsic	Intrinsic	Extrinsic	Intrinsic	Extrinsic
GPT-4o (zero shot)	-	**14.46**	0	15.75	-	10.81
GPT-4o w/ 5-shot	-	14.43	0	**17.07**	-	**14.63**
GPT-4o w/ 10-shot	-	10.08	0	16	-	10.53
GPT-4o w/ 30-shot	-	9.76	0	13.33	-	8.51

なった．一方，事実確認が必要な重大な hallucination については高い再現率での検出が可能であった．これは，LLM の生成結果を LLM 自身でチェックし，hallucination 発生の可能性がある箇所を警告するアプローチが実用的である可能性を示唆している．

3·1 実 験 設 定

検出モデルとして，HELM [Liang 23] 及び Nejumi LLM リーダーボード [Kamata 24] で高い性能を示している GPT-4o を採用した．検出モデルは，入力された質問と参考文献に基づき，各応答文に対して “hallucination なし”，“Intrinsic Hallucination”，“Extrinsic Hallucination”，の 3 ラベルのいずれかを付与する．JHARS の応答数は 450 件と限られているため追加学習は適切でないと判断し，文脈内学習を採用した．プロンプトには，Ji ら [Ji 24] の英語プロンプトを日本語に翻訳したものを使用した．評価には，JHARS を 8:2 の比率で分割し，8 割を評価データ，残り 2 割を少数事例学習用の例示データとして使用した．評価指標には，各クラスの適合率と再現率から算出した F1 スコアを採用し，hallucination の検出性能を測定した．

3·2 実 験 結 果

GPT-4o を用いた hallucination 検出の F1 スコアを表 5 に示した *4．hallucination 検出において，再現率は高い値を示したものの適合率が低く，結果として F1 スコアは 15%程度に留まった．これは，偽陽性が多く発生する傾向にあることを意味しており，高い適合率と再現率を両立した hallucination 検出は困難であることを示している．また，少数事例学習における事例数の増加が必ずしも F1 スコアの向上につながらないことも判明した．特に 30 事例を使用した場合において，最も低い F1 スコアを記録した．これについては，入力長の増加が LLM の性能低下を招くことが指摘されており [Levy 24]，本実験においても同様の現象が発生したと考えられる．

一方，§2·2 節で言及した事実確認が必要な Extrinsic Hallucination についての検出性能を分析したところ，7 件の事例全てを正しく検出できていることを確認した．この結果は，LLM 自身による生成結果のチェックが，hallucination 発生の可能性がある箇所をユーザに警告し，事実確認を促す有効な手段となり得ることを示唆している．

4. お わ り に

本研究では，日本語における hallucination 評価のための包括的なベンチマーク JHARS を構築し，3 つの LLM の応答を分析した．今後，本研究で用いたフローを活用してデータの拡充を目指す．加えて，本研究では文単位でアノテーションを付与したが，スパン単位でアノテーションを付与し，よりきめ細かい hallucination 評価データセットの構築を目指す．

*4　適合率と再現率の結果は Appendix 表 C.2 に記載.

◇参 考 文 献◇

[Baobab 24] Baobab, I.: wikipedia-human-retrieval-ja (2024), https://huggingface.co/datasets/baobab-trees/wikipedia-human-retrieval-ja [Accessed：10/2024]

[Dziri 22] Dziri, N., Kamalloo, E., Milton, S., Zaiane, O., Yu, M., Ponti, E. M., and Reddy, S.: FaithDial: A Faithful Benchmark for Information-Seeking Dialogue, *Transactions of the Association for Computational Linguistics*, Vol. 10, pp. 1473–1490 (2022)

[Fujii 24] Fujii, K., Nakamura, T., Loem, M., Iida, H., Ohi, M., Hattori, K., Shota, H., Mizuki, S., Yokota, R., and Okazaki, N.: Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities, in *Proceedings of the First Conference on Language Modeling*, COLM, p. (to appear) (2024)

[Huang 23] Huang, Y., Feng, X., Feng, X., and Qin, B.: The Factual Inconsistency Problem in Abstractive Text Summarization: A Survey (2023)

[Huang 24] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T.: A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, *ACM Transactions on Information Systems* (2024)

[Ji 23] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P.: Survey of Hallucination in Natural Language Generation, *ACM Comput. Surv.*, Vol. 55, No. 12 (2023)

[Ji 24] Ji, Z., Gu, Y., Zhang, W., Lyu, C., Lin, D., and Chen, K.: ANAH: Analytical Annotation of Hallucinations in Large Language Models, in Ku, L.-W., Martins, A., and Srikumar, V. eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics (2024)

[Kamata 24] Kamata, K., Ibi, T., Yamamoto, Y., Kurosawa, K., Kanazawa, R., and Shibata, A.: Nejumi LLM リーダーボード 3 (2024)

[Levy 24] Levy, M., Jacoby, A., and Goldberg, Y.: Same Task, More Tokens: the Impact of Input Length on the Reasoning Performance of Large Language Models, in Ku, L.-W., Martins, A., and Srikumar, V. eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15339–15353, Association for Computational Linguistics (2024)

[Li 22] Li, W., Wu, W., Chen, M., Liu, J., Xiao, X., and Wu, H.: Faithfulness in Natural Language Generation: A Systematic Survey of Analysis, Evaluation and Optimization Methods (2022)

[Li 23] Li, J., Cheng, X., Zhao, X., Nie, J.-Y., and Wen, J.-R.: HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models, in Bouamor, H., Pino, J., and Bali, K. eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, Association for Computational Linguistics (2023)

[Liang 23] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C. A., Manning, C. D., Re, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., WANG, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N. S., Khattab, O., Henderson, P., Huang, Q., Chi, R. A., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y.: Holistic Evaluation of Language Models, *Transactions on Machine Learning Research* (2023), Featured Certification, Expert Certification

[Liang 24] Liang, X., Song, S., Niu, S., Li, Z., Xiong, F., Tang, B., Wang, Y., He, D., Peng, C., Wang, Z., and Deng, H.: UHGEval: Benchmarking the Hallucination of Chinese Large Language Models via Unconstrained Generation, in Ku, L.-W., Martins, A., and Srikumar, V. eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5266–5293, Association for Computational Linguistics (2024)

[Liu 22] Liu, T., Zhang, Y., Brockett, C., Mao, Y., Sui, Z., Chen, W., and Dolan, B.: A Token-level Reference-free Hallucination Detection Benchmark for Free-form Text Generation, in Muresan, S., Nakov, P., and Villavicencio, A. eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6723–6737, Association for Computational Linguistics (2022)

[Longpre 21] Longpre, S., Perisetla, K., Chen, A., Ramesh, N., DuBois, C., and Singh, S.: Entity-Based Knowledge Conflicts in Question Answering, in Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, Association for Computational Linguistics (2021)

[Mishra 24] Mishra, A., Asai, A., Balachandran, V., Wang, Y., Neubig, G., Tsvetkov, Y., and Hajishirzi, H.: Fine-grained Hallucination Detection and Editing for Language Models, in *First Conference on Language Modeling* (2024)

[Niu 24] Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., and Zhang, T.: RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models, in Ku, L.-W., Martins, A., and Srikumar, V. eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, Association for Computational Linguistics (2024)

[Okazaki 24] Okazaki, N., Hattori, K., Shota, H., Iida, H., Ohi, M., Fujii, K., Nakamura, T., Loem, M., Yokota, R., and Mizuki, S.: Building a Large Japanese Web Corpus for Large Language Models, in *Proceedings of the First Conference on Language Modeling*, COLM, p. (to appear) (2024)

[OpenAI 24a] OpenAI, : GPT-4o (2024), https://platform.openai.com/docs/modelsgpt-4o [Accessed：10/2024]

[OpenAI 24b] OpenAI, : GPT-4o mini (2024), https://platform.openai.com/docs/modelsgpt-4o-mini [Accessed：10/2024]

◇ 付　　録 ◇

A. データセット作成の参考情報

アノテーション時のインターフェースを図 A.1 に示した．なお，アノテーションで使用したガイドラインは以下のリンクから確認できる．
https://docs.google.com/document/d/1T1iRGCdm384DTh7Hs11T4YRCpFaTDdK3-e$_g LSiiV_c/edit?usp = sharing$

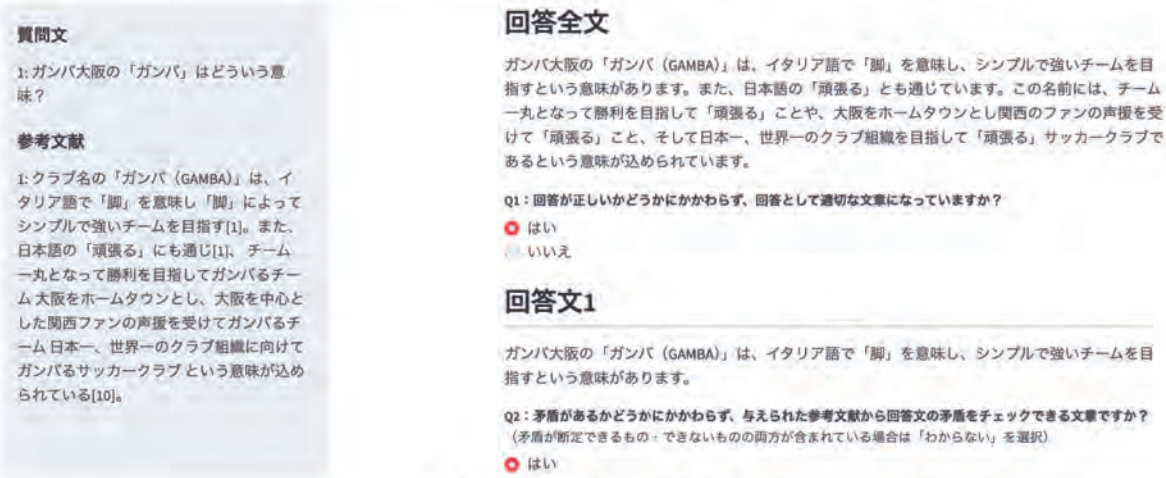


図 **A.1**　アノテーション時のインターフェース．

B. Extrinsic Hallucination の事例

本文に記載しなかった Extrinsic Hallucination の事例をいくつか表 B.1 に示した．

表 **B.1**　Extrinsic Hallucination（青色塗りつぶし）の事例．(...) は省略を表す．

応答モデル名：GPT-4o 質問文：ハイダイナミックレンジとはどのような機能ですか？ 参考文献：ハイダイナミックレンジ (英語: High dynamic range, HDR) は、通常よりも高いダイナミックレンジのことである。ハイダイナミックレンジは様々な信号、例えば、ビデオ、もしくは画像、(...) 応答文：ハイダイナミックレンジ（HDR）は、通常よりも高いダイナミックレンジを持つ技術や手法を指します。これは主に（...） このように、HDR は信号の特性を向上させ、より広範な明暗の情報を捉えることができる機能です。
応答モデル名：Llama-3.1-Swallow-8B-Instruct-v0.1 質問文：タヌキとアライグマを見分ける方法はありますか? 参考文献：灰褐色の体毛をもち、眼のまわりから頬にかけて黒い斑紋が（...） 応答文：タヌキとアライグマを見分ける方法はいくつかあります。(...) 一方で、ニホンアナグマはタヌキよりも小さく、体の毛並みが粗く、尾の毛が短いという特徴があります。

C. 実験結果の詳細

GPT-4o を用いた Extrinsic hallucination 検出の適合率と再現率を表 C.2 に示した．Intrinsic hallucination は生成モデルが GPT-4o-mini の場合に 1 件のみ，かつ検出できていなかったため省略している．

表 **C.2**　Extrinsic hallucination の検出の適合率と再現率（%）．

	Generator: GPT-4o		Generator: GPT-4o-mini		Generator: Llama-3.1-Swallow-8B-Instruct-v0.1	
Detector	適合率	再現率	適合率	再現率	適合率	再現率
GPT-4o (zero shot)	5.26	57.14	**11.01**	75.00	**6.25**	**100**
GPT-4o w/ 5-shot	**9.33**	**100**	9.70	81.25	5.41	**100**
GPT-4o w/ 10-shot	7.00	**100**	9.40	**87.50**	5.88	**100**
GPT-4o w/ 30-shot	5.83	**100**	8.86	**87.50**	**6.25**	**100**

著 者 紹 介

亀井　遼平

東北大学博士前期課程在学中. AI Shift との共同研究に従事. 本論文主著.

坂田　将樹

東北大学博士後期課程在学中. AI Shift との共同研究に従事. 本論文主著.

邊土名　朝飛

2021 年長岡技術科学大学大学院工学研究科修士課程終了後, サイバーエージェントに入社. AI Lab および AI Shift にて対話システムの研究開発に従事.

栗原　健太郎

2023 年早稲田大学大学院基幹理工学研究科修士課程終了後, サイバーエージェントに入社. 言語処理学会より 2023 年度最優秀論文賞を受賞. 現在は AI Shift にてプロダクト開発に従事.

乾　健太郎

Mohamed bin Zayed University of Artificial Intelligence (UAE) 客員教授，東北大学言語 AI 研究センター教授．1995 年東京工業大学大学院情報理工学研究科博士課程修了．同大学助手，九州工業大学助教授，奈良先端科学技術大学院大学助教授を経て，2010 年より東北大学教授．2016 年より理化学研究所 AIP センター自然言語理解チームリーダー，2023 年より MBZUAI 客員教授兼任.

リアルタイム性と柔軟性を兼ね備えた音声対話システムのための軽量かつ高速な処理手法の検討

大竹 真太
Shinta Otake
株式会社 AI Shift
ML Engineer
otake_shinta@cyberagent.co.jp

keywords: Spoken Dialogue System, Real-time Processing, Large Language Model

Summary

Voicebots currently used in call centers mainly adopt a system-driven approach, progressing conversations along predefined scenarios, which makes it difficult to respond flexibly like human operators. In recent years, spoken dialogue systems that use Large Language Models (LLM) to provide flexible responses have attracted attention. However, LLM has high computational costs and are not suitable for real-time responses. In this study, I propose a method for a spoken dialogue system targeting reservation tasks that performs high-speed intent understanding and end-of-conversation detection through streaming processing without using LLM. I also examine a method that incorporates the flexible responses of LLM into real-time processing. Furthermore, in this presentation, I will provide a demo of the implemented proposed method, consider the challenges of spoken dialogue systems based on the feedback obtained, and discuss future research directions.

1. はじめに

現在, 多くの企業が顧客サービスの向上と業務効率化を目的として, コールセンターに音声対話システム, いわゆるボイスボットの導入を進めている. これらのボイスボットは, システム主導で事前に定義されたシナリオに沿って対話を進める方式が主流となっている. この方式は, あらかじめ設定された質問と回答のパターンに基づいており, 特定の業務や問い合わせに迅速に対応できる一方で, ユーザからの予期せぬ発話や多様な要求に対しては柔軟な対応が難しいという課題がある.

この課題を解決するため，大規模言語モデル（Large Language Model; LLM）を活用した音声対話システムの開発・研究が盛んになっている [金子 23], [千葉 23]. LLM は, 高度な自然言語処理能力を有しており, シナリオベースの対話システム [熊谷 21], [熊谷 22] では困難であった柔軟かつ自然な対話が可能になる. しかし, LLM を音声対話システムに適用すると, 応答速度の遅延やコスト増大といった問題が生じる. また，同時に多数のユーザからの問い合わせに対応する場面では, リアルタイム性とスケーラビリティの観点からすべての対話シーンに LLM を直接適用することは現実的でない場合がある.

本稿では，音声対話システムの各モジュールの要素技術を組み替えてユーザ評価実験を実施し，リアルタイム性と柔軟性を兼ね備えた対話システムの要素技術の特定を試みる．特に，高度な言語理解が必要な処理については，LLM を部分的に活用し，リアルタイム性と柔軟性のトレードオフを評価する．実験では，飲食店の予約タスクを対象としたデモシステムを開発し，実際に被験者とシステムが対話するユーザ評価実験を実施する．

本稿の構成は以下のとおりである. 第 2 章では, 提案方式およびデモシステムの実装についての詳細を説明し, 第 3 章では, 実装したデモシステムの評価実験とその結果を示す. 第 4 章で本稿のまとめを述べる.

2. 提案手法

2·1 システム構成

本稿で提案する音声対話システムは, 大きく以下のモジュールから構成される. 音声をテキスト化する音声認識モジュール, テキストを入力としてユーザの意図を理解し, システムの現在の状態に応じて柔軟かつ適切な応答を返す対話モジュール, およびシステムの応答文を音声に変換する音声合成モジュールである.

音声認識モジュールとしては Google Speech-to-Text API[*1]（以下, Google STT）を, 音声合成モジュールとしては Azure Text-to-Speech API[*2]を利用する.

対話モジュールをさらに細分化すると, ユーザの発話終了を判定する発話終了検知機能, ユーザの意図を理解する意図理解機能, 現在の状態を管理する対話管理機能, お

∗1 https://cloud.google.com/speech-to-text?hl=ja
∗2 https://learn.microsoft.com/ja-jp/azure/ai-services/speech-service/index-text-to-speech

表 1　対話モデル一覧

モデル	発話終了検知	意図理解	割り込み発話機能	定型応答フィラー	逐次確認発話
1	stability	LLM	なし	あり	なし
2	stability	spaCy	なし	なし	あり
3	音量ベース VAD	LLM	あり	あり	なし
4	音量ベース VAD	spaCy	あり	なし	あり

よび状態に応じて適切な応答を返す言語生成機能が含まれる. また, 対話管理手法として, スロットフィリングを採用し, 埋まったスロットに基づいて対話を進行させる.

2·2　発話終了検知

汎用的な発話終了検知のために, 近年では深層学習ベース手法の研究が盛んになっている [佐藤 24], [Inoue 24] が, 本稿では, 軽量かつベースラインとなるような発話終了検知アルゴリズムを検証するために, 次の 2 種類のアルゴリズムを実装した. 1 つ目は, Google STT の `stability` を利用する方法である. `stability` は Google STT のストリーミングレスポンスに含まれるフィールドで, 音声認識結果の安定性を表す値である. 1.0 に近いほど安定していて 0.0 に近いほど不安定であることを表す. この確率が設定した閾値を超えたときに発話終了を検知する. 2 つ目は, 音量ベースの音声区間検出（Voice Activity Detection; VAD）を用いる方法である. 逐次送信される音声チャンクを処理して音量を取得し, 設定した閾値以上であれば有声音チャンクと判定し, 無声音チャンクが一定期間続いた場合に発話終了を検知する.

2·3　意図理解

ユーザの発話をテキストとして取得した後, 意図理解アルゴリズムを用いてスロットフィリングを行う. 本稿では, 意図理解アルゴリズムとして 2 つの実装を用意した. 1 つ目は, LLM を活用する方法である. LLM によるスロットフィリングはプロンプトを与えるだけで汎用的に機能する一方, レスポンスが返ってくるまでに一定の待ち時間が発生し, ユーザにストレスを与える可能性がある. これを防ぐため, ユーザの発話を受け取ったことを示すフィラーで間を埋める. なお, 本稿では LLM として Azure OpenAI Service[*3]の GPT-4o モデルを用いた.

2 つ目は, 自然言語処理ライブラリの spaCy[*4]（内部のモデルとして ja_ginza[*5]を採用）を活用した手法である. 本手法では, まずテキストの前処理として, ja_ginza で定義されている Entity 形式に適合させるため, 曖昧な日付表現や時間表現をルールベースで標準化する. その後, spaCy を用いて Entity を検知し, スロットフィリングを実行する. このアプローチは, LLM と比較すると汎用性では制限があるものの, 処理の効率性とコスト効率性の両面で優位であると考えられる. これにより, システムはユーザの発話からリアルタイムでスロット情報を抽出し, 即座に確認発話を生成することが可能となる. このような逐次確認発話は, システムの誤認識の検出と訂正のために重要なインタラクションである [平沢 99].

また, ユーザの曖昧な発話に対応するため, スロットが埋まらなかった場合には,「ご確認いたします」と応答し, LLM への問い合わせを行うようにする. LLM のハルシネーションを極力減らすため, あらかじめ用意した FAQ リストの中からユーザの発話に関連する質問に対する回答を返すようにする. 関連する質問がリストに含まれない場合は, 空の文字列を返すようにプロンプトを設定し, 空文字が返された際には, fallback の定型文言をシステムに発話させる.

2·4　スロット状態とテンプレートに基づく応答生成

本稿では, 言語モデルの生成過程におけるハルシネーションを防止するため, 応答生成においては定型的なテンプレートを用いた制御手法を採用する. まず, 現在のスロット状態 V において, 未充足のスロットの集合 $\mathcal{U}$ を以下のように定義する.

$$\mathcal{U} = \{ s \in \mathcal{S} \mid V(s) = \text{未充足} \} \tag{1}$$

ここで, $\mathcal{S}$ は全てのスロットの集合, $V(s)$ はスロット s の現在の状態を示す. 次に, 未充足スロット集合 $\mathcal{U}$ から次に問い合わせるべきスロット s_{next} を選択する. スロット選択関数は以下のように定義される.

$$s_{\text{next}} = \text{Select}(\mathcal{U}) \tag{2}$$

この関数はシナリオで設計された優先度に基づいてスロットを選択する. 選択したスロット s_{next} に対して質問テンプレート $Q(s_{\text{next}})$ を用いて質問文を作成し, その質問に対するユーザーの回答に応じて応答テンプレート $R(s_i, v_i)$ を用いて応答文を作成する.

3.　評価実験

3·1　実験設定

§1　実験概要

提案システムのインタラクションに関する主観評価を行うために評価実験を実施した. 実験では, 仮想の飲食店

*3　https://learn.microsoft.com/ja-jp/azure/ai-services/openai/overview
*4　https://spacy.io/
*5　https://megagonlabs.github.io/ginza/

表 2　飲食店予約の対話例

システム	お電話ありがとうございます. 新規のご予約を承ります.
ユーザ	はい, 新規予約をお願いします.
システム	ご希望の日付をお伺いしてもよろしいでしょうか.
ユーザ	来週の土曜日に予約したいです.
システム	ご希望の時間をお伺いしてもよろしいでしょうか.
ユーザ	19 時でお願いします.
システム	ご来店人数をお伺いしてもよろしいでしょうか.
ユーザ	5 人で.
システム	ご来店される代表者のお名前をお伺いしてもよろしいでしょうか.
ユーザ	田中です.
システム	10 月 26 日の 19 時, 5 名様で田中様, ご予約を確定させてもよろしいでしょうか. はい, またはいいえでお答えください
ユーザ	はい, 間違いありません.
システム	予約が完了しました. 当日はお気を付けてお越しください.

の新規予約受付タスクを設定し, 被験者に実際にシステムと対話を行ってもらった. 対話例としては表 2 のような対話になり, システム側の予約完了メッセージをもって対話終了とする.

被験者は社内のエンジニア 8 名で構成され, 各被験者は 4 種類の対話モデルを用いて対話を行った. 対話モデルの体験順序によるバイアスを軽減するため, 全ての順列 $4! = 24$ 通りから, 被験者ごとに異なる順序でモデルを体験するように割り当てを行った.

実験手順は次の通りである. まず, 被験者は指定された電話番号に架電し, 対話モデルの番号（1〜4）を選択する. 次に選択した対話モデルと対話し予約を完了させる. 対話が終了したら, 主観評価を評価シートに記入する. この一連の手順を, 割り当てられた対話モデルに対して繰り返し行った. また, バイアス軽減のために, 1 セット（4 回の架電）終了後は一定の時間をおいて次のセットの実験を実施した.

§2　4 種類の対話モデル

発話終了検知アルゴリズムと意図理解アルゴリズムの組み合わせに応じて 4 種類のモデルを作成した. それぞれのモデルの内容を表 1 に示す. 意図理解アルゴリズムに LLM を用いたモデルでは, LLM の処理遅延に対応するため,「はい」,「承知しました」などの定型応答をフィラーとして使用している. 一方, 音量ベースの VAD による発話終了検知を実装したモデルでは, 一定時間の有声区間を検知した際に割り込み発話を許可する機能を実装している. なお, LLM を用いた FAQ 応答機能は全ての対話モデルに共通して実装されており, LLM から空文字が返却された場合には fallback メッセージを出力する仕様となっている. また, 対話管理機能および応答生成機能も各モデルで共通であり, 現在のスロット状態に応じてあらかじめ作成した応答テンプレートを返す仕様となっている． 実験で用いたテンプレートを表 3 に示す. モデル 2 およびモデル 4 では更新されたスロットに応じて対話の各ターンで「{ 日付 } ですね.」や「{ 日付 } の { 時間 } ですね.」などの逐次確認発話を行う.

表 3　応答テンプレート一覧

シーン	応答テンプレート
初期発話	お電話ありがとうございます. 新規のご予約を承ります.
日付聴取	ご希望の日付をお伺いしてもよろしいでしょうか.
時間聴取	ご希望の時間をお伺いしてもよろしいでしょうか.
人数聴取	ご来店人数をお伺いしてもよろしいでしょうか.
名前聴取	ご来店される代表者のお名前をお伺いしてもよろしいでしょうか.
最終確認	{ 日付 } の { 時間 } に { 人数 } 名で { 名前 } さまのご予約を承りました. ご予約を確定させてもよろしいでしょうか. はい, またはいいえでお答えください.
予約完了	予約が完了しました. 当日はお気を付けてお越しください.
fallback	申し訳ございません. うまく聞き取れませんでした.

§3　評価項目

実験で用いる評価項目を表 4 に示す．これらの評価項目は，主に音声対話システムのユーザーとシステム間のインタラクションにおける自然さと, システムの応答文生成における自然さを評価することを目的として，[Mehri 20], [井上 19] を参考に設定した．　Q1〜4 に関しては,「全くそうは思わない」,「そうは思わない」,「ややそうは思わない」,「どちらとも言えない」,「ややそう思う」,「そう思う」,「とてもそう思う」の 7 段階のリッカート尺度

表 4　評価項目

Q1	応答は自然でしたか.
Q2	音声対話システムはあなたの質問や要望に柔軟に対応していましたか.
Q3	応答速度にストレスを感じましたか.
Q4	音声対話システムとの対話中に離脱したいと思いましたか.
Q5	どのタイミングで一番離脱したいと思ったか.
Q6	どこで一番ストレスを感じたか.

で評価してもらった．Q5〜6 は自由記述で回答してもらい，主観的な感想や具体的なストレス要因を明らかにし，システムの問題点や改善点を特定することを目的とした．

3·2 実 験 結 果

§1 定量分析

Q1〜4 の評価結果をすべての実験に渡って集計し，箱ひげ図としてまとめたものを図図 1 に示す．図 1 において，赤色がモデル 1，青色がモデル 2，緑色がモデル 3，橙色がモデル 4 の評価結果を表している．

応答の自然さ（Q1）については，全モデルにおいて中央値が 5.0 前後と比較的高い評価を得ているが，モデル 3，4 に関しては評価のばらつきが大きくなっている．これは音量ベースの VAD がうまく機能する場合もあれば音響環境やユーザーの音声の大きさによってうまく機能しないことがあり，不自然な発話終了検知が行われたためだと考えられる．

システムの柔軟性（Q2）に関しては，モデル 2 とモデル 4 が中央値 5.0 と同等の評価を示したものの，全モデルにおいて評価のばらつきが大きく，ユーザーによって評価が分かれる結果となった．モデル 2，4 の評価が高くなる要因は逐次確認発話があることで応答に柔軟性が生まれ，また，意図理解に誤りがある場合に訂正しやすくなり，インタラクションの面においてもより柔軟になるからだと考えられる．

応答速度に関するストレス（Q3）については，全モデルにおいて中央値が 2.0 と低くなっている．値が低いほど望ましい評価項目であり，全てのモデルが応答速度の面で優れたパフォーマンスを示していることを表している．これは発話終了検知アルゴリズムがユーザのストレスにならない程度の速度で概ね機能していることを示唆している．ただし，外れ値もいくつか含まれ，発話終了検知アルゴリズムの汎用的な性能に課題があることがわかる．

対話離脱意向（Q4）に関しては，全モデルで評価のばらつきが顕著に表れた．ユーザーごと，対話ごとに離脱の要因となる対話モデルの振る舞いが異なり，さまざまな要因が紐付いて離脱要因となっていることが示唆される．

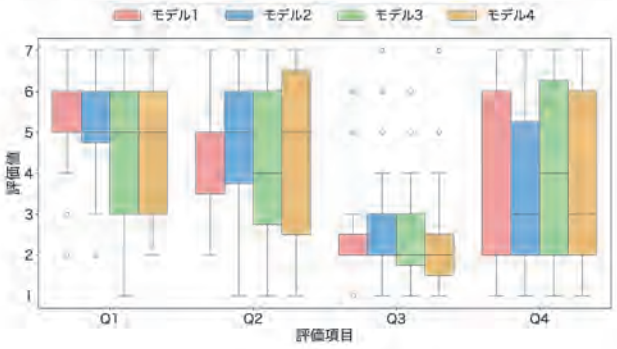


図 1 数値回答の評価結果

§2 定性分析

離脱タイミング（Q5）およびストレスを感じたタイミング（Q6）に関する自由記述回答と対話ログデータを定性的に分析した結果から考察をまとめる．

発話終了検知アルゴリズムに関して，stability ベースおよび音量ベースの VAD の両実装において，ユーザの離脱意向とストレス要因が確認された．stability ベースの手法では，言い淀みを含む長文発話における早期終了検知エラーが顕著であった．音量ベースの VAD においては，短文発話に対する検知遅延がユーザの不安感を惹起する一方，過度に早期の終了検知により発話が適切に処理されないケースも報告された．これらは音響環境や発話特性に依存するエラーとして特徴付けられる．ただし，音量ベース VAD を実装したモデル 3，4 では対話の自然性に対する肯定的評価も得られ，適切に機能する発話終了検知がユーザ体験の向上に寄与する可能性が示唆された．

意図理解アルゴリズムの評価において，LLM と spaCy の両実装におけるスロットフィリングのエラーが，ユーザエクスペリエンスの低下と離脱意向の主要因となることが明らかになった．LLM ベースのアプローチでは，時刻を回答する発話を解釈する際に，日付スロットを今日の日付で上書きしてしまう課題が観察された．特に，逐次確認機能の未実装により，予約の最終確認段階で初めてシステムの時間認識誤りが顕在化し，これがユーザ満足度の低下および潜在的な離脱要因となった．一方，spaCy ベースのアプローチにおいては，事前定義されていない時間表現形式（例：「19 時半」）に対する解釈失敗がスロットフィリングエラーを誘発し，同様にユーザの離脱行動とストレス増加を招いていた．これらの知見は，LLM ベースの手法における制御性の課題と，ルールベースの手法における汎用性の制約という，各アプローチに内在する技術的課題を示唆している．

定型的応答フィラーについては，ユーザ評価が顕著に一致した否定的な反応が観察された．具体的には，応答の不自然さや機械的な印象，「はい」や「承知しました」などの肯定的な応答に対して fallback 処理が実行される論理的矛盾などが指摘され，これらがユーザのストレス要因として作用することが明らかになった．一方，逐次確認発話機能については，対話の柔軟性向上に寄与する有効な要素として評価された．特に，ユーザがスロット情報を訂正する機会を自然な形で提供できる点が有用であることが示唆された．また，割り込み発話機能はシステムの確認発話に対する柔軟な訂正をユーザに許可する役割として上手く機能していた．一方で，割り込み発話機能がないモデルでは，システムの発話中に訂正ができないことがストレス要因となった．逐次確認発話と割り込み発話機能が適切に動作することで，システムが誤認識した際の訂正機会を柔軟に提供できると考えられる．

スロットが埋まらなかった時の挙動について，システムが「ご確認いたします」と応答した後に FAQ 検索が失敗

し，「聞き取れませんでした」というメッセージを出す流れに対する違和感が指摘された．この問題の要因として，FAQ の粒度が過度に詳細であることによるマッチング失敗と，エラー種別に応じた適切な応答メッセージの設定できていなかったことが考えられる．エラー処理の改善策として，音声認識失敗時は「聞き取れません」，該当する回答が存在しない場合は「その情報は持ち合わせていません」など，状況に応じた応答の使い分けが必要である．また，実験を通じてスロットが埋まらないケースが高頻度で発生することが判明したため，計算コスト削減の観点からも，LLM への問い合わせ前に表層的な判定を行う必要がある．さらに，本実験では FAQ リストに番号を振ってプロンプトに記載していたため，音声認識によって番号のみが認識された場合に，的外れな回答をしてしまう問題も発生した．

全てのモデルに共通する主要な離脱・ストレス要因として，対話シナリオの設計に関する問題が最も顕著であることが明らかになった．具体的には，応答テンプレートやフィラー文言の設計，初期発話から予約完了までの対話遷移設定，そして LLM を用いた FAQ 応答の設計などが課題として挙げられた．また，人名に対する音声認識誤りも重要な要因として挙げられた．より自然な対話システムの実現には，効率的な対話シナリオの設計手法の確立と，音声認識モデルの固有名詞認識性能の向上が重要であることが示唆された．

4. 結　　論

本稿では，飲食店予約を対象とした音声対話システムにおいて，意図理解や発話終了検知などの基本機能を軽量な手法で実装しつつ，複雑な言語理解が必要な場面では選択的に LLM を活用するハイブリッドアプローチを提案した．提案手法の有効性を検証するためデモを実装し，ユーザ評価実験を実施した．実験結果から，基本的な意図理解においては軽量な手法で十分な性能が得られることが確認された．一方で，より自然な対話の実現には，汎用的な発話終了検知アルゴリズムの開発，対話シナリオの精緻化，音声認識精度の向上が重要な課題として明らかになった．今後の展望として，発話終了検知の精度向上や対話シナリオの拡充を進めるとともに，多様なタスクや状況に適用可能な，LLM と軽量な手法を組み合わせたハイブリッドアプローチの最適な設計方式を確立していく．

謝　　辞

本稿の執筆にあたり，多大なるご協力を賜りました株式会社 AI Shift AI チームの皆様に，心より感謝申し上げます．

◇ 参　考　文　献 ◇

[Inoue 24] Inoue, K., Jiang, B., Ekstedt, E., Kawahara, T., and Skantze, G.: Multilingual Turn-taking Prediction Using Voice Activity Projection (2024)

[Mehri 20] Mehri, S. and Eskenazi, M.: Unsupervised Evaluation of Interactive Dialog with DialoGPT, in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 225–235 (2020)

[井上 19] 井上 昂治，ララ ディペッシュ，山本 賢太，中村 静，高梨 克也，河原 達也：自律型アンドロイド ERICA による傾聴対話システムの評価，人工知能学会研究会資料 言語・音声理解と対話処理研究会，Vol. 87，p. 04 (2019)

[金子 23] 金子 拓正，稲葉 通将：LLM に基づく音声対話システムのための非言語情報を活用したユーザ心情の考慮とリアルタイム性の向上，人工知能学会研究会資料 言語・音声理解と対話処理研究会，Vol. 99，pp. 131–133 (2023)

[熊谷 21] 熊谷 和実，徳永 清輝，三宅 徳久，田村 和弘，平良 弘之，大武 美保子：ベッドサイド見守り声掛けロボットのシナリオベース対話への応答と対話継続時間，人工知能学会全国大会論文集，Vol. JSAI2021，pp. 4E4OS11d02–4E4OS11d02 (2021)

[熊谷 22] 熊谷 和実，徳永 清輝，三宅 徳久，田村 和弘，水内 郁夫，大武 美保子：高齢者見守り声掛けロボットのシナリオベース対話中のユーザ発話終了予測に基づく話者交替タイミングの決定，人工知能学会全国大会論文集，Vol. JSAI2022，pp. 2N6OS7b01–2N6OS7b01 (2022)

[佐藤 24] 佐藤友紀，千葉祐弥，東中竜一郎：複数の日本語データセットによる音声活動予測モデルの学習とその評価，人工知能学会 言語・音声理解と対話処理研究会（第 100 回），pp. 192–197 (2024)

[千葉 23] 千葉祐弥，光田航，李晃伸，東中竜一郎：Remdis: リアルタイムマルチモーダル対話システム構築ツールキット，人工知能学会 言語・音声理解と対話処理研究会（第 99 回），pp. 25–30 (2023)

[平沢 99] 平沢 純一，宮崎 昇，中野 幹生，相川 清明：音声対話システムの誤解に対するユーザ応答の分析，音声言語情報処理，pp. 29–27 (1999)

著　者　紹　介

大竹　真太

2024 年 東京工業大学 修士課程終了

多面的なユーザ意欲を考慮したセールス対話データセットおよび対話システムの構築と評価

邊土名 朝飛 Asahi Hentona
CyberAgent AI Lab
Research Scientist
hentona_asahi@cyberagent.co.jp

馬場 淳 Jun Baba
CyberAgent AI Lab
Research Scientist
baba_jun@cyberagent.co.jp

佐藤 志貴 Shiki Sato
CyberAgent AI Lab
Research Scientist
sato_shiki@cyberagent.co.jp

赤間 怜奈 Reina Akama
東北大学
Assistant Professor
akama@tohoku.ac.jp

keywords: 対話データセット, 対話システム, セールストーク, クラウドソーシング

Summary

購買意欲を向上させるセールス対話システムを実現するためには多面的なユーザの意欲を考慮したデータセットが必要だが、既存データセットにはシステムの想定運用環境で収集された信頼性の高いユーザの意欲に関するデータが含まれていない。本研究では、想定運用環境に基づいた対話データ収集環境を開発し、3 種類のユーザ意欲データを含む日本語セールス対話データセットを構築した。ユーザ評価実験では、発話レベルでユーザの意欲を考慮し、さらにデータセット分析で得られたセールス対話戦略の知見を組み込むことが対話システムによる対話成功率の向上につながることが示唆された。

1. はじめに

購買意欲を向上させるセールス対話システムを実現するためには、(1) 対話継続意欲、(2) 情報提供意欲、(3) 目標受容意欲の少なくとも 3 種類のユーザ意欲を考慮しつつ対話を進行することが重要だと考えられる。そのようなシステムの開発には、生態学的妥当性 (ecological validity)[Brunswik 40, Brunswik 52] の高いセールス対話データセットの整備が必要不可欠である。生態学的妥当性とは、現実世界へのユーザへの実験結果の適用可能性を意味し、主に心理学や Human-Computer-Interaction の分野において用いられる概念である。しかし、既存のセールス対話データセット [Hiraoka 16, Tiwari 23] は、人間同士の対話設定であったり、販売対象となる商品が事前に固定されているなど、実際のセールス対話システムの想定運用環境とは乖離した設定でデータが収集されているため生態学的妥当性が高いとはいえない。

本研究では、実用的なセールストーク対話システムの実現を目指し、ユーザの多面的な意欲を考慮した日本語セールス対話データセットを構築する。データセットの生態学的妥当性を高めるため、実際のセールス対話システムの想定運用環境を可能な限り再現した実験設定下で、自然なユーザの対話および意欲データを収集することを試みる。具体的には、ユーザの自然なエンゲージメントを可能な限り正確に計測するため、実験参加者であるユーザ役に対して、任意のタイミングで対話を離脱することを許可し、対話システムのふりをしたセールス経験者（セールス役）がユーザ役の対話相手となる Wizard-of-Oz (WOZ) 法 [Kelley 84] の設定で対話データを収集する。さらに、ユーザの意欲の評価をユーザ役自身に行ってもらい、かつ対話単位のみならず発話単位での評価も収集する。本データセットにより、ユーザの購買意欲を高めるセールス対話戦略のきめ細やかな分析や、ユーザの反応に応じて柔軟に対話戦略を切り替える効果的なセールス対話システムの開発につながることが期待される。

本稿では、構築したデータセットを概説したうえで、本データセット上でユーザの購買意欲を向上させるセールス対話戦略を分析した結果を報告する。さらに、分析により得られた対話戦略を組み込んだ大規模言語モデルベースのセールス対話システムが、ユーザ評価実験にお

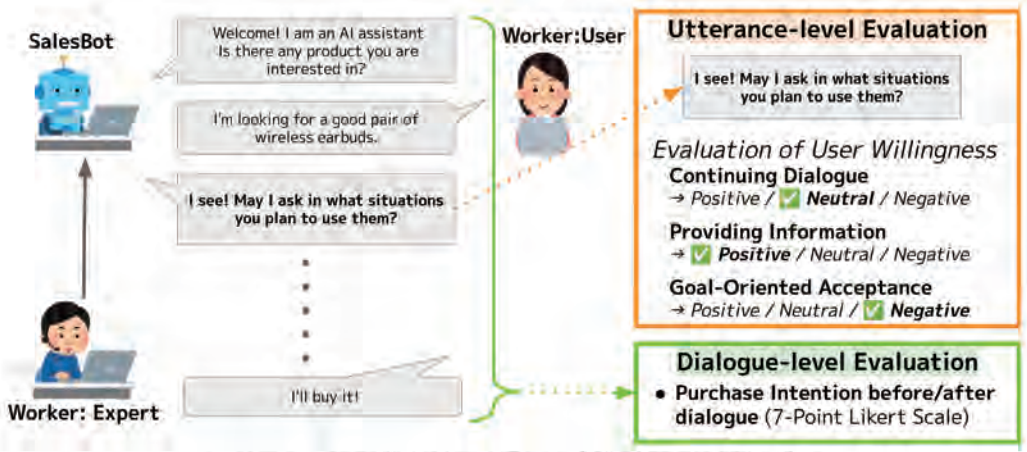


図 1: セールス対話データセット構築プロセスの概要図

表 1: セールス対話データセットの統計量。トークナイザーには MeCab[Kudo 04]+UniDic[Den 08] を使用した。

	Total	Mean	Max	Min
# dialogues	109	-	-	-
# success dialogues	63	-	-	-
# utterances	3289	30.2	92	11
— User	1144	10.5	41	3
— Sales	2145	19.7	52	7
# tokens	54301	498.2	1406	153

いて高い評価を得ることを通して，本データセットの有用性を示す。

2. データセット概要

2·1 データセット構築

図図 1 に、セールス対話データセットの構築プロセスの概要を示す。本研究では、我々の過去の収集 [邊土名 24] で用いた設定に従い、3 種類の架空のワイヤレスイヤホンの情報を掲載した Web ページに訪れたユーザと、その Web ページ上に設置されたセールス対話システムがテキストチャットを行うシナリオで対話データを収集した。構築したデータセットの統計情報を表表 1 に示す。

2·2 ユーザ意欲

本研究では、対話レベルのユーザ意欲データとして、対話実験の前後に 7 段階リッカート尺度評価の購買意欲データを収集した。また、発話レベルのユーザ意欲データとして (1) 対話継続意欲 (Continuing dialogue; CD)、(2) 情報提供意欲 (Providing information; PI)、(3) 目標受容意欲 (Goal-oriented acceptance; GA) の 3 種類の意欲データを収集した。対話継続意欲は、ユーザがシステムとの対話を続けたいと感じる意欲であり、主にオープンドメイン対話の分野で扱われてきた [Yu 16, Zhang 18, See 19, Ghazarian 20]。情報提供意欲は、ユーザが自分のニーズや要望をシステムに伝える意欲を指す。この意欲を高めることで、システムはユーザの嗜好に合わせたセールス対話が可能となり、説得力とユーザ満足度が向上することが期待できる [Wang 19, Sun 22, Berkovsky 12]。目標受容意欲は、ユーザがシステムの対話目標、すなわちセールス対話の最終目標である商品購入 [Jung 22] を受け入れたいと感じる意欲を指す。発話レベルのユーザ意欲は、それぞれ 3 段階 (Positive、Neutral、Negative) で評価してもらった。

3. データセット分析

上述の手順で構築した日本語セールス対話データセットを用いて、ユーザの購買意欲向上に寄与するセールス対話戦略について分析する。

3·1 対話レベル分析

各対話に占める各ユーザ意欲評価ラベルの割合と、ユーザの購買意欲の向上度合い（対話前後の 7 段階リッカート尺度評価の変化量）との間での相関分析結果を図 図 2 に示す。相関係数を見ると、Positive および Neutral 評価は購買意欲の向上とほぼ相関がなく、反対に Negative 評価との間には負の相関が示された。この結果から、ユーザの購入意欲を向上させるためには、各種意欲を高めるような発話を試みるよりも、対話を通じて意欲を低下させる発話を避けることが重要であることが示唆された。

3·2 発話レベル分析

対話ターンごとの意欲の推移：

図 3a、3b は、それぞれ成功対話（ユーザの購入意向が向上した対話）と失敗対話（ユーザの購入意向が向上しなかった対話）の平均意欲スコアの推移を示した図である。ここで、平均意欲スコアは、各発話に付与された評価ラベルについて Positive: +1、Neutral: 0、Negative: −1 のスコアを付与し、意欲スコアの移動平均を計算することで求めた。成功対話の平均意欲スコアの推移を見

表 2: ユーザ評価実験結果

モデル	学習データ	ユーザ意欲の考慮	対話戦略の考慮	対話成功率	平均ターン数
GPT-3.5	成功対話 (63 対話)	-	-	0.23	9.35
GPT-3.5W	全対話 (109 対話)	✓	-	0.33	9.23
GPT-3.5WD	全対話 (109 対話)	✓	✓	0.44	9.08
GPT-4o (reference)	-	-	-	0.58	5.81

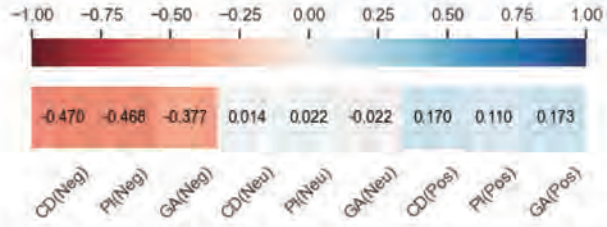
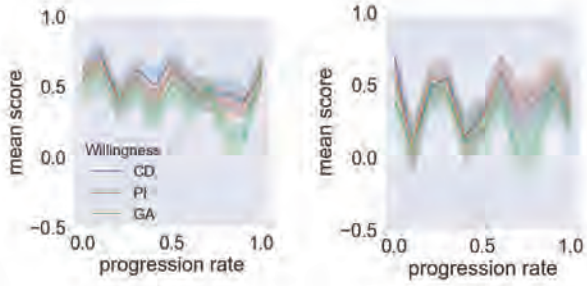
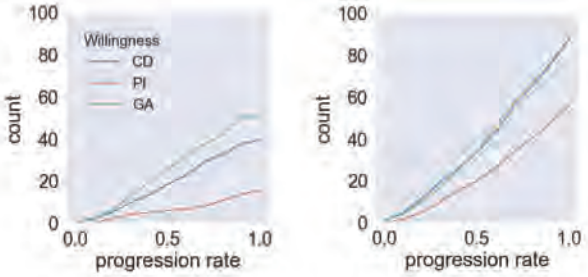


図 2: 各対話中の意欲評価ラベルの割合と購買意欲向上度合い間の相関分析結果



(a) 成功対話　(b) 失敗対話

図 3: 対話中のユーザ意欲スコアの推移



(a) 成功対話　(b) 失敗対話

図 4: 各意欲ラベルの累計 Negative 評価数の推移

ると、対話開始直後と終了直前に各種意欲スコアが上昇していることがわかる。特に、対話終盤にかけて目標受容意欲スコアが大幅に増加している。一方、失敗対話では、対話開始直後および終了直前に意欲スコアが減少していることがわかる。特に、対話開始直後に意欲スコアが大きく減少している。

図 4a、 4b は、それぞれ成功対話、失敗対話における累計 Negative ラベルの推移を表している。成功対話と失敗対話を比較すると、全ての種類の意欲について成功対話の Negative 評価数が低く抑えられている。特に、ユーザ意欲の中でも情報提供意欲の Negative 評価数は低く抑えられており、対話中盤においてこの傾向が顕著であることがわかる。

効果的なセールス対話戦略：

以上の分析から、ユーザの購買意欲を高めるためには、対話序盤、中盤、終盤ごとに以下の対話戦略が有効であると考えられる。(1) 対話序盤：対話継続意欲に Positive な印象を与え、早期の対話離脱を防ぐ対話の実施、(2) 対話中盤：情報提供意欲に Negative な印象を与えないユーザヒアリングの実施、(3) 対話終盤：目標受容意欲に Positive な影響を与える対話（例えば商品推薦）の実施。

4. ユーザ評価実験

セールス対話データセットの有用性を検証するために、データセットを用いてセールス対話システムを構築し、ユーザ評価実験を行った。

4·1 実 験 設 定

クラウドソーシングを通じて 48 名の作業者を募集し、複数のシステムと対話および評価を行うよう指示した。各システムとの対話方法および評価内容については、データセット構築時の設定と同様である。なお、各システムとの対話順序によってモデルの評価に順序バイアスが生じる可能性があるため、作業者間でシステムと対話する順序をランダムに割り振りバイアスの軽減を図った。

4·2 モ デ ル

以下の 3 種類のシステムを構築した。各モデルは、OpenAI の GPT-3.5 (`gpt-3.5-turbo-0125`) および Fine-tuning API[*1] を使用して開発した。

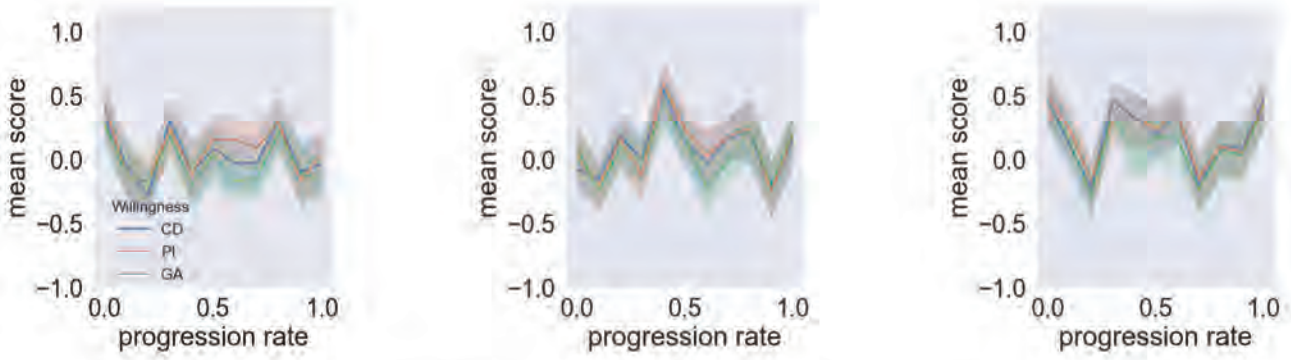
GPT-3.5 (baseline):

発話レベルのユーザ意欲を考慮せず、構築したデータセット中の成功対話 (計 63 件) のみを学習データとして使用して Fine-tuning したモデルである。

GPT-3.5 with willingness (`GPT-3.5W`):

baseline とは異なり、ユーザの発話レベルの意欲ラベ

*1 https://platform.openai.com.



(a) GPT-3.5　(b) GPT-3.5W　(c) GPT-3.5WD

図 5: ユーザ評価実験における各モデルのユーザ意欲スコアの推移

ルを考慮して Fine-tuning したモデルである。Fine-tuning のアプローチとして、Attribute-conditioned Supervised Fine-Tuning [Dong 23] を採用した。具体的には、各ユーザ発話の末尾に、次に発話されるシステム発話の各意欲ラベルの属性名、属性値のペアを CONTINUING_DIALOGUE:1 のように付与して学習することで、属性値に応じたシステム発話が出力されることを試みる *2。生成時の各意欲ラベルの属性値は全て 1(Positive) に設定した。

GPT-3.5 with willingness + dialogue strategy (GPT-3.5WD):

上記の GPT-3.5W をベースに、§3·2 節のセールス対話戦略を組み込んだモデルである。具体的には、1-3 ターン目は対話継続意欲、4-6 ターン目は情報提供意欲、7 ターン目以降は目標受容意欲がそれぞれ Positive になるよう属性値を指定した *3。

4·3 実験結果・議論

ユーザ評価実験の結果を表 表 2 に示す。対話成功率を見ると、発話レベルのユーザ意欲を考慮した GPT-3.5W、GPT-3.5WD が baseline モデルよりも高い。また、§3·2 節のセールス対話戦略を反映した GPT-3.5WD が最も対話成功率が高くなっている。

図 図 5 に、モデルごとの平均意欲スコアの推移を示す。最も低い成功率を獲得した baseline モデルは、対話開始時点の意欲スコアは GPT-3.5WD と並んでいるものの、その後スコアは減少し、対話終盤も向上することなく終了している。2 番目に対話成功率が高かった GPT-3.5W は、対話開始時点の意欲スコアは GPT-3.5、GPT-3.5WD と比較して低いものの、対話終盤には各種意欲スコアが向上している。最も対話成功率が高かった GPT-3.5WD は、対話序盤の意欲スコアは減少傾向を示しているものの開始時点のスコアは GPT-3.5W よりも高くなっている。対話中盤においては情報提供意欲のスコアが Positive 傾向で維持されており、対話終盤には各種意欲スコアが大幅に向上していることがわかる。

これらの結果から、発話レベルのユーザの意欲を考慮することで、ユーザの購買意欲向上につながるセールス対話を実現できることが示唆された。また、GPT-3.5WD の各種意欲スコアの推移から、§3·2 節のセールス対話戦略がシステムにある程度反映されており、その結果最も高い対話成功率を示したと考えられる。

5. お わ り に

本研究では、生態学的妥当性に基づいた対話データ収集環境を開発し、3 種類のユーザ意欲データを含む日本語セールス対話データセットを構築した。構築したデータセットの分析により、購買意欲向上につながると考えられるセールス対話戦略の知見が得られた。また、ユーザ評価実験では、発話レベルでユーザの意欲を考慮し、さらにデータセット分析で得られたセールス対話戦略の知見を組み込むことが対話成功率向上につながることが示された。今後は、ユーザの応答に応じて動的にセールス対話戦略を切り替えることが可能なシステムを開発し、その有効性を実証実験によって検証する予定である。

謝　　辞

本研究の一部は、JSPS 科研費 JP22K17943 の支援を受けたものである。

◇ 参 考 文 献 ◇

[Berkovsky 12] Berkovsky, S., Freyne, J., and Oinas-Kukkonen, H.: Influencing Individually: Fusing Personalization and Persuasion, *ACM Trans. Interact. Intell. Syst.*, Vol. 2, No. 2 (2012)

[Brunswik 40] Brunswik, E.: Thing constancy as measured by correlation coefficients., *Psychological Review*, Vol. 47, No. 1, p. 69 (1940)

[Brunswik 52] Brunswik, E.: *The Conceptual Framework of Psychology*, International encyclopedia of unified science, University of Chicago Press (1952)

*2 属性値は Positive=1、Neutral=0、Negative=-1 と設定した。
*3 その他の意欲ラベルの属性値は 0(Neutral) とした。

[Den 08] Den, Y., Nakamura, J., Ogiso, T., and Ogura, H.: A Proper Approach to Japanese Morphological Analysis: Dictionary, Model, and Evaluation, in Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D. eds., *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco (2008), European Language Resources Association (ELRA)

[Dong 23] Dong, Y., Wang, Z., Sreedhar, M., Wu, X., and Kuchaiev, O.: SteerLM: Attribute Conditioned SFT as an (User-Steerable) Alternative to RLHF, in Bouamor, H., Pino, J., and Bali, K. eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11275–11288, Singapore (2023), Association for Computational Linguistics

[Ghazarian 20] Ghazarian, S., Weischedel, R., Galstyan, A., and Peng, N.: Predictive Engagement: An Efficient Metric for Automatic Evaluation of Open-Domain Dialogue Systems, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 05, pp. 7789–7796 (2020)

[Hiraoka 16] Hiraoka, T., Neubig, G., Sakti, S., Toda, T., and Nakamura, S.: Construction and analysis of a persuasive dialogue corpus, *Situated Dialog in Speech-Based Human-Computer Interaction*, pp. 125–138 (2016)

[Jung 22] Jung, Y.: *Sales Talk*, pp. 87–114, Springer Nature Singapore, Singapore (2022)

[Kelley 84] Kelley, J. F.: An iterative design methodology for user-friendly natural language office information applications, *ACM Trans. Inf. Syst.*, Vol. 2, No. 1, p. 26–41 (1984)

[Kudo 04] Kudo, T., Yamamoto, K., and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, in Lin, D. and Wu, D. eds., *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 230–237, Barcelona, Spain (2004), Association for Computational Linguistics

[See 19] See, A., Roller, S., Kiela, D., and Weston, J.: What makes a good conversation? How controllable attributes affect human judgments, in Burstein, J., Doran, C., and Solorio, T. eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1702–1723, Minneapolis, Minnesota (2019), Association for Computational Linguistics

[Sun 22] Sun, S., Zhang, J., Zhu, Y., Jiang, M., and Chen, S.: Exploring users' willingness to disclose personal information in online healthcare communities: The role of satisfaction, *Technological Forecasting and Social Change*, Vol. 178, p. 121596 (2022)

[Tiwari 23] Tiwari, A., Khandwe, A., Saha, S., Ramnani, R., Maitra, A., and Sengupta, S.: Towards personalized persuasive dialogue generation for adversarial task oriented dialogue setting, *Expert Systems with Applications*, Vol. 213, p. 118775 (2023)

[Wang 19] Wang, X., Shi, W., Kim, R., Oh, Y., Yang, S., Zhang, J., and Yu, Z.: Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good, in Korhonen, A., Traum, D., and Màrquez, L. eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5635–5649, Florence, Italy (2019), Association for Computational Linguistics

[Yu 16] Yu, Z., Nicolich-Henkin, L., Black, A. W., and Rudnicky, A.: A Wizard-of-Oz Study on A Non-Task-Oriented Dialog Systems That Reacts to User Engagement, in Fernandez, R., Minker, W., Carenini, G., Higashinaka, R., Artstein, R., and Gainer, A. eds., *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 55–63, Los Angeles (2016), Association for Computational Linguistics

[Zhang 18] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J.: Personalizing Dialogue Agents: I have a dog, do you have pets too?, in Gurevych, I. and Miyao, Y. eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, Melbourne, Australia (2018), Association for Computational Linguistics

[邊土名 24] 邊土名 朝飛, 馬場 惇, 赤間 怜奈：セールストークを対象とするエンゲージメントを考慮した目標指向対話データセット, 人工知能学会全国大会論文集, Vol. JSAI2024, pp. 3Xin240–3Xin240 (2024)

著 者 紹 介

邊土名 朝飛

2021 年長岡技術科学大学大学院工学研究科修士課程修了後、サイバーエージェント入社。AI Lab および株式会社 AI Shift にて対話システムの研究開発に従事。

公平なマッチング相互推薦

ユーザーに被推薦機会の不公平感を抱かせない相互推薦システム

冨田 燿志
Yoji Tomita

AI Lab
Research Scientist
tomita_yoji@cyberagent.co.jp

keywords: 推薦システム, 相互推薦, マッチングマーケット, 公平性, 公平分割理論

Summary

オンラインデーティングや求職サービスなどのマッチングプラットフォームにおいて，ユーザーに他のユーザーを推薦する相互推薦システムは重要な役割を果たしている．相互推薦システムにおいては一部の人気ユーザーに被推薦機会が集中してしまう混雑の問題が発生してしまいやすく，そのような状況ではユーザー間に不公平感を抱かせてしまう課題が指摘される．本稿ではマッチングプラットフォームにおける不公平性を，公平分割理論における無羨望性の概念を用いて定義し，公平な相互推薦手法としてナッシュ厚生交互最適化法を提案する．また，人工データを用いたシミュレーション実験によって，ナッシュ厚生交互最適化法は比較手法と比べても比較的期待マッチ数を多くし，一方でほぼ羨望を発生させない非常に公平な推薦を達成することを示した．

1. はじめに

オンラインデーティングや求職サービスなどのマッチングプラットフォームにおいて，ユーザーに他のユーザーを推薦する相互推薦システムは，プラットフォームの成功に対し非常に重要な役割を果たしている．相互推薦システムにおいて発生しやすい課題として，一部の人気ユーザーばかりを集中して推薦してしまう混雑の問題は研究においても実務においても認識されており，近年研究が進んでいる [Tomita 23].

また，これに関連する課題として，ユーザー間の被推薦機会についての不平等性・不公平性の問題がある．多くのマッチングプラットフォームにおいて，あるユーザーがどれだけのマッチを得られるかどうかは，そのユーザーが相手側ユーザーに対してどれだけ推薦されるかに大きく依存する．したがって，被推薦機会が偏ってしまい，それによって得られるマッチ数にも差が付いてしまう場合，ユーザー間に不公平感を抱かせてしまう可能性がある．同じサブスクリプションプランに加入しているにもかかわらず被推薦機会の偏りによりマッチ数に差がついてしまう状況はユーザーにとってサービスに対する不満となりうるため，ユーザー体験を高めるためにはそのような被推薦機会の不公平は生じないようにすることがサービスにとって望ましいと考えられる．しかし，推薦システムにおける公平性は近年盛んに研究が進められているものの，マッチングプラットフォームの相互推薦システムの問題において公平性に着目する研究は非常に限られている．

本稿では，マッチングプラットフォームにおける相互推薦システムにおいて，被推薦機会の公平性の問題を考える．まず，マッチングプラットフォームの推薦の問題を表現する基本モデルを 2 章で示した後，3 章において，公平分割理論の分野で知られる「無羨望性」の概念を用いて被推薦機会の公平性を定義する．次に，4 章において，プラットフォーム全体の期待マッチ数として定義される社会厚生を最大化することを企図する社会厚生交互最適化法を導入する．5 章では，経済学・公平分割理論の分野で知られるナッシュ厚生をマッチングプラットフォームのモデルにおいて定義し，それを最適化するナッシュ厚生交互最適化法を公平相互推薦手法として提案する．最後に，6 章において人工データによるシミュレーション実験の結果を示す．実験において，社会厚生交互最適化法は期待マッチ数を高くするが，多くの羨望を発生させる不公平な推薦となるため期待マッチ数と公平性にはトレードオフがあることが示された．また，ナッシュ厚生交互最適化法は社会厚生交互最適化法には劣るものの他の比較手法と比べると同等以上の期待マッチ数を達成する一方で，公平性については羨望をほぼ発生させない非常に公平な推薦となることを示した [*1].

2. 準備：基本モデル

オンラインデーティングサービスを模した，二つのグループに分かれたユーザー間のマッチングプラットフォームを考える．ここでは二グループはそれぞれ左側ユーザー・

*1 本稿は，本稿著者が筆頭著者として発表した論文 [Tomita 24] を日本語で紹介するものである．紙幅の関係で詳細な説明を省略した箇所もあるため，詳細は元論文を参照してほしい．

右側ユーザーと呼ぶ．左側ユーザーの集合を $\mathcal{L}$ ($|\mathcal{L}| = L < \infty$)，右側ユーザーの集合を $\mathcal{R}$ ($|\mathcal{R}| = R < \infty$) とする．

プラットフォームは以下の流れでマッチングを行う．

(1) 各左側ユーザー $\ell \in \mathcal{L}$ は，K 人の右側ユーザーのランク付き推薦リストを受け取り，そのうち気に入ったユーザーに “Like” をする．

(2) 同様に，各右側ユーザー $r \in \mathcal{R}$ は，K 人の左側ユーザーのランク付き推薦リストを受け取り，そのうち気に入ったユーザーに “Like” をする．

(3) 互いに “Like” をし合ったユーザーは “マッチ成立” となる．

ここでプラットフォームが考えたい問題は，ユーザー間に不公平感を抱かせないようにしつつマッチ数を最大化する推薦の仕方である．

2·1 推薦ポリシー

プラットフォームは各ユーザーにパーソナライズされた推薦リストを確率的に選ぶことができ，その確率的推薦ポリシーを $(A,B) = ((A_\ell)_{\ell\in\mathcal{L}}, (B_r)_{r\in\mathcal{R}})$ と表す．ここで，$A_\ell \in \mathbb{R}_{\geq 0}^{R\times K}$ は左側ユーザー $\ell \in \mathcal{L}$ に送る確率的推薦を表す $R \times K$ の行列で，$A_\ell(r,k) \geq 0$ は ℓ への推薦リスト内で右側ユーザー r が第 k 番目の位置で推薦される確率を表し，確率のため制約として

$$\sum_{r\in\mathcal{R}} A_\ell(r,k) = 1 \ \ \forall \ell \in \mathcal{L}, k \leq K \tag{1}$$

$$\sum_{k=1}^{K} A_\ell(r,k) \leq 1 \ \ \forall \ell \in \mathcal{L}, r \in \mathcal{R} \tag{2}$$

を満たす．同様に，$B_r \in \mathbb{R}_{\geq 0}^{L\times K}$ は右側ユーザー $r \in \mathcal{R}$ に送る確率的推薦を表す $L \times K$ の行列で，$B_r(\ell,k) \geq 0$ は r への推薦リスト内で左側ユーザー ℓ が k 番目の位置で推薦される確率を表し，

$$\sum_{\ell\in\mathcal{R}} B_r(\ell,k) = 1 \ \ \forall r \in \mathcal{R}, k \leq K \tag{3}$$

$$\sum_{k=1}^{K} B_\ell(\ell,k) \leq 1 \ \ \forall r \in \mathcal{R}, \ell \in \mathcal{L} \tag{4}$$

を満たす．以下簡単のため，確率的推薦ポリシー (A,B) がプラットフォームによって決定された後，実際にユーザーに送られる推薦リストを確率分布から引く操作は各ユーザーについて独立に行われると仮定する．

2·2 ユーザー行動モデル

左側ユーザー $\ell \in \mathcal{L}$ は，プラットフォームから送られた K 人の右側ユーザーの推薦リストを確認し，“Like” を送る相手を決定する．このとき．ユーザーの行動は Position-based Model（PBM）に基づいて決定されるものとする．つまり，左側ユーザー ℓ が，自身が受け取った推薦リスト内で k 番目に位置付けられた右側ユーザー r に “like” を送る確率は，

$$\Pr(\ell \text{ likes } r \mid r \text{ is on } k\text{'th position}) = e(k)\cdot p_{\ell,r}$$

とする．ただし，$e(\cdot)$ は検査関数（exanmination function）で，k についての減少関数（たとえば，$e(k) = 1/k$ や $1/\log_2(k+1)$），$p_{\ell,r}$ は左側ユーザー ℓ の右側ユーザー r に対する選好度合いを表すスコアとする．

同様に，右側ユーザー r が推薦リスト内 k 番目の左側ユーザー ℓ に “like” を送る確率は，

$$\Pr(r \text{ likes } \ell \mid \ell \text{ is on } k\text{'th position}) = e(k)\cdot q_{r,\ell}$$

で，$q_{r,\ell}$ は右側ユーザー r の ℓ に対する選好度合いを表す．ここで両側の選好スコア，$p_{\ell,r}$ や $q_{r,\ell}$ は ALS などなんらかの通常の推薦システムの手法によってプラットフォームによって正確に計算されているものとする．

2·3 社会厚生

上記の PBM によるユーザー行動を仮定すると，確率的推薦ポリシー (A,B) を所与としたとき，左側ユーザー ℓ が右側ユーザー r に “like” する確率は，

$$\Pr(\ell \text{ likes } r \mid A_\ell) = \sum_{k=1}^{K} A_\ell(r,k)\cdot e(k)\cdot p_{\ell,r},$$

右側ユーザー r が ℓ に “like” する確率は，

$$\Pr(r \text{ likes } \ell \mid B_r) = \sum_{m=1}^{K} B_r(\ell,m)\cdot e(m)\cdot q_{r,\ell}$$

なので，ユーザーペア (ℓ,r) がマッチする確率 $\mathbb{P}_{\ell,r}$ は，

$$\mathbb{P}_{\ell,r} = \sum_{k=1}^{K} A_\ell(r,k)\cdot e(k)\cdot p_{\ell,r} \sum_{m=1}^{K} B_r(\ell,m)\cdot e(m)\cdot q_{r,\ell}$$

となる．

このプラットフォームの目的関数として社会厚生関数 SW をマッチ数として定義すると，

$$SW(A,B) = \sum_{\ell\in\mathcal{L}}\sum_{r\in\mathcal{R}} \mathbb{P}_{\ell,r} \tag{5}$$

とかける．

3. 公平性の公理：無羨望性

左側ユーザー $\ell \in \mathcal{L}$ の厚生を得られるマッチ数とすると，期待厚生は

$$U_\ell(A_\ell, B_\cdot(\ell,\cdot)) = \sum_{r\in\mathcal{R}} \mathbb{P}_{\ell,r}.$$

となる．ここで，U_ℓ は ℓ が受け取る確率的推薦 (A_ℓ) と，ℓ が右側ユーザーにどのように推薦されるか $(B_\cdot(\ell,\cdot))$ によって決まる．

ここで考えたい不公平性は，以下のようなものである．

Algorithm 1 社会厚生交互最適化

Require: 選好スコア $(p_{i,j})_{i,j}$, 検査関数 $e(\cdot)$, 繰り返し回数 T, 学習率 $(\eta_t)_{t\in[T]}$.

A, B を適当な値で初期化.
for $t = 1, 2, \ldots, T$ **do**
  $A^* = \arg\max_{A'} SW(A', B)$ s.t. A' は制約 1-2 を満たす.
  $A \leftarrow (1-\eta_t)A + \eta_t A^*$.
  $B^* = \arg\max_{B'} SW(A, B')$ s.t. B' は制約 3-4 を満たす.
  $B \leftarrow (1-\eta_t)B + \eta_t B^*$
end for
return A, B

定義 1. 推薦ポリシー (A, B) において左側ユーザー $\ell \in \mathcal{L}$ が被推薦機会について別の左側ユーザー ℓ' に**羨望**を持つとは,

$$U_\ell(A_\ell, B_\cdot(\ell, \cdot)) < U_\ell(A_\ell, B_\cdot(\ell', \cdot))$$

が成り立つことをいう．また，(A, B) が左側ユーザーについて**無羨望性**を満たすとは，他の左側ユーザーに羨望を持つ左側ユーザーが存在しないことをいう.

左側ユーザー ℓ が仮に ℓ' のように右側ユーザーへ推薦された場合より多くの期待マッチ数が得られる場合，ℓ は「ℓ' は自分より多く良い位置で推薦されていて不公平だ」と感じるだろう．そのような不公平が生じない状態が無羨望性である.

同様に，右側ユーザーの厚生は得られる期待マッチ数

$$V_r(B_r, A_\cdot(r, \cdot)) = \sum_{\ell \in \mathcal{L}} \mathbb{P}_{\ell, r}$$

とする.

定義 2. 推薦ポリシー (A, B) において右側ユーザー $r \in \mathcal{R}$ が被推薦機会について別の左側ユーザー r' に**羨望**を持つとは,

$$V_r(B_r, A_\cdot(r, \cdot)) < V_r(B_r, A_\cdot(r', \cdot))$$

が成り立つことをいう．また，(A, B) が右側ユーザーについて**無羨望性**を満たすとは，他の右側ユーザーに羨望を持つ右側ユーザーが存在しないことをいう.

定義 3. 推薦ポリシーが (A, B) が**無羨望性**を満たすとは，左側ユーザーについての無羨望性と右側ユーザーについての無羨望性の両方を満たすことをいう.

4. 社会厚生交互最適化

公平な相互推薦手法を考える前に，社会厚生（総期待マッチ数）最大化を目指す手法を考える.

社会厚生関数 (5) は $A_\ell(r, k), B_r(\ell, k)$ の凹とは限らない二次の関数であり，直接この制約付き最適化問題

$$\max_{A,B} SW(A, B)$$
$$\text{s.t. } A_\ell, B_r \text{ は制約 (1)-(4) を満たす.}$$

を解くことは計算的に困難である.

ここで，社会厚生関数 (5) は，B について固定すると A についての線形関数であり，また A について固定すると B についての線形関数である．制約 (1)-(4) もそれぞれ A, B について線形であるため，一方を固定すればもう一方については最適解を線形計画法によって解くことができる．このアルゴリズムは Algorithm 1 としてまとめた.

この社会厚生交互最適化の収束解は大域最適解となる保証はないが，後に示すように他手法と比べて社会厚生を大きくすることが実験によって示されている.

4·1 社会厚生最適解の不公平性

しかし，社会厚生を最大化する解は全体の期待マッチ数は最大化するが，前章で定義した無羨望性の意味では公平でない．以下の例を見てみよう.

例 1. 左側ユーザーは $1, 2$ の 2 人，右側ユーザーは 1 の 1 人とする．左側ユーザーの選好は等しく $p_{1,1} = p_{2,1} = 1$ とし，一方右側ユーザーは左側ユーザー 1 をわずかに好む $q_{1,1} = 1$, $q_{1,2} = 1 - \epsilon$ $(\epsilon > 0)$ とする．また $K = 1$ で，$e(1) = 1$ とする．このとき，マーケットが操作できるのは右側ユーザーに左側ユーザー 1 を推薦する確率 $r \in [0, 1]$ である.

r を固定したときの期待マッチ数は，$r * 1 + (1-r) * (1-\epsilon)$ であるため，厚生を最大化する r は $r^* = 1$ である．しかし，このとき左側ユーザー 2 は推薦される確率が 0 であるため期待利得は 0 である．一方，1 のように確率 $r^* = 1$ で推薦されると期待利得は $1 - \epsilon$ となるため，左側ユーザー 2 は 1 に対して羨望を持つ.

この状況では，$r = 1/2$ とする，つまり左側ユーザー 1 を確率 $1/2$，2 を確率 $1/2$ で推薦する状況が唯一の無羨望推薦である.

5. 公平推薦：ナッシュ厚生交互最適化

前章では，全体のマッチ数を最大化する社会厚生最大化推薦を考え，それがユーザー間の被推薦機会の不公平性を引き起こすことを示した．次に，公平な推薦を実現

Algorithm 2 ナッシュ厚生交互最適化

Require: 選好スコア $(p_{i,j})_{i,j}$, 検査関数 $e(\cdot)$, 繰り返し回数 T, 学習率 $(\eta_t)_{t\in[T]}$.

1: A, B を適当な値で初期化
2: **for** $t = 1, 2, \ldots, T$ **do**
3: $\quad B^* = \arg\max_{B'} \sum_{r,\ell,k} \frac{\partial NW_L(A,B)}{\partial B_r(\ell,k)} B'_r(\ell,k)$ s.t. B' は制約 1-2 を満たす.
4: $\quad B \leftarrow (1-\eta_t)B + \eta_t B^*$.
5: $\quad A^* = \arg\max_{A'} \sum_{\ell,r,k} \frac{\partial NW_R(A,B)}{\partial A_\ell(r,k)} A'_\ell(r,k)$ s.t. A' は制約 3-4 を満たす.
6: $\quad A \leftarrow (1-\eta_t)A + \eta_t A^*$
7: **end for**
8: **return** A, B

する手法を考える．そのために，経済学・公平分割理論において知られているナッシュ厚生関数を定義しよう．

5·1 ナッシュ厚生関数とナッシュ厚生最適推薦

左側ユーザーについてのナッシュ厚生（Nash Welfare）は，各左側ユーザーの期待厚生を掛け合わせたもの，つまり

$$NW_L(A,B) = \prod_{\ell\in\mathcal{L}} U_\ell(A_\ell, B_{\cdot}(\ell,\cdot))$$

とする．また，右側ユーザーについてのナッシュ厚生も同様に，

$$NW_R(A,B) = \prod_{r\in\mathcal{R}} V_r(B_r, A_{\cdot}(r,\cdot))$$

とする．

ナッシュ厚生最適推薦を以下で定義する．

定義 4.
- 左側ユーザーへの推薦 A に対して右側ユーザーへの推薦 B^* が左側ナッシュ厚生最適であるとは，$B^* \in \arg\max_B NW_L(A,B)$ を満たすことをいう．
- 右側ユーザーへの推薦 B に対して左側ユーザーへの推薦 A^* が右側ナッシュ厚生最適であるとは，$A^* \in \arg\max_A NW_R(A,B)$ を満たすことをいう．
- (A^*, B^*) が両側ナッシュ厚生最適であるとは，互いが他方に対してナッシュ厚生最適であることをいう．

ナッシュ厚生最適な推薦は近似的に無羨望性を満たすことが知られている（詳細は [Tomita 24] を参照）．

5·2 ナッシュ厚生交互最適化

左側ユーザーのナッシュ厚生について log をとると，$\log NW_L(A,B) = \sum_{\ell\in\mathcal{L}} \log U_\ell(A_\ell, B_{\cdot}(\ell,\cdot))$ は A を固定した場合 B についての凹関数となる．一方，右側ユーザーのナッシュ厚生についても log をとると，$\log NW_R(A,B) = \sum_{r\in\mathcal{R}} \log V_r(B_r, A_{\cdot}(r,\cdot))$ は B を固定した場合に A についての凹関数となる．したがって，社会厚生交互最適化と同様に，ナッシュ厚生もそれぞれ Frank-Wolfe 法による制約付き凸最適化を交互に行うことによって，ナッシュ厚生最適な推薦を求めることができる（Algorithm 2）．

6. 実　　験

本章では，人工データによるシミュレーション実験の結果を示す．

6·1 実 験 設 定

左側ユーザーの数は $n = 50$ または 75，右側ユーザーの数は $m = 50$ とし，検査関数は $e(k) = 1/k$（“inv”）または $e(k) = 1/\log_2(1+k)$（“log”）とし，推薦される最大数 K は左側ユーザーに対しては $K = m$，右側ユーザーに対しては $K = n$ とする．

選好スコアについては，左側ユーザー $i \in \{1,2,\ldots,m\}$ と右側ユーザー $j \in \{1,2,\ldots,n\}$ に対し，左側ユーザー i が右側ユーザー j をどれだけ好むかを表す $p_{i,j}$ は $p_{i,j} = (1-\lambda)\widetilde{p}_{i,j} + \lambda\frac{j-1}{m-1}$，逆の右側ユーザー j が左側ユーザー i をどれだけ好むかを表す $q_{j,i}$ は $q_{j,i} = (1-\lambda)\widetilde{q}_{j,i} + \lambda\frac{i-1}{n-1}$ とする．ここで，$\widetilde{p}_{i,j}, \widetilde{q}_{j,i} \in U[0,1]$, i.i.d. であり，$\lambda \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ とした．添え字 i, j が大きくなるほど多くの相手ユーザーに好まれる “人気ユーザー” となり，λ が大きければ大きいほど人気ユーザーに人気が集中し混雑が発生しやすくなる状況となり，一方 λ が小さければ小さいほど人気が分散する状況となる．

各シナリオに対して選好のランダム項 $\widetilde{p}_{i,j}, \widetilde{q}_{j,i}$ を発生させ，各推薦手法により推薦を行った時の期待マッチ数（社会厚生）と，左側ユーザー・右側ユーザー間の羨望の数を計測する実験を 10 回ずつ行った．

6·2 比 較 手 法

この実験において比較する推薦手法は以下の通りである．

- Naive : 各ユーザーに対し最も好む相手ユーザーを順に推薦する．つまり，左側ユーザー i に対しては右側ユーザーを $p_{i,j}$ の降順に，右側ユーザー j に対しては左側ユーザーを $q_{j,i}$ の降順に推薦する．

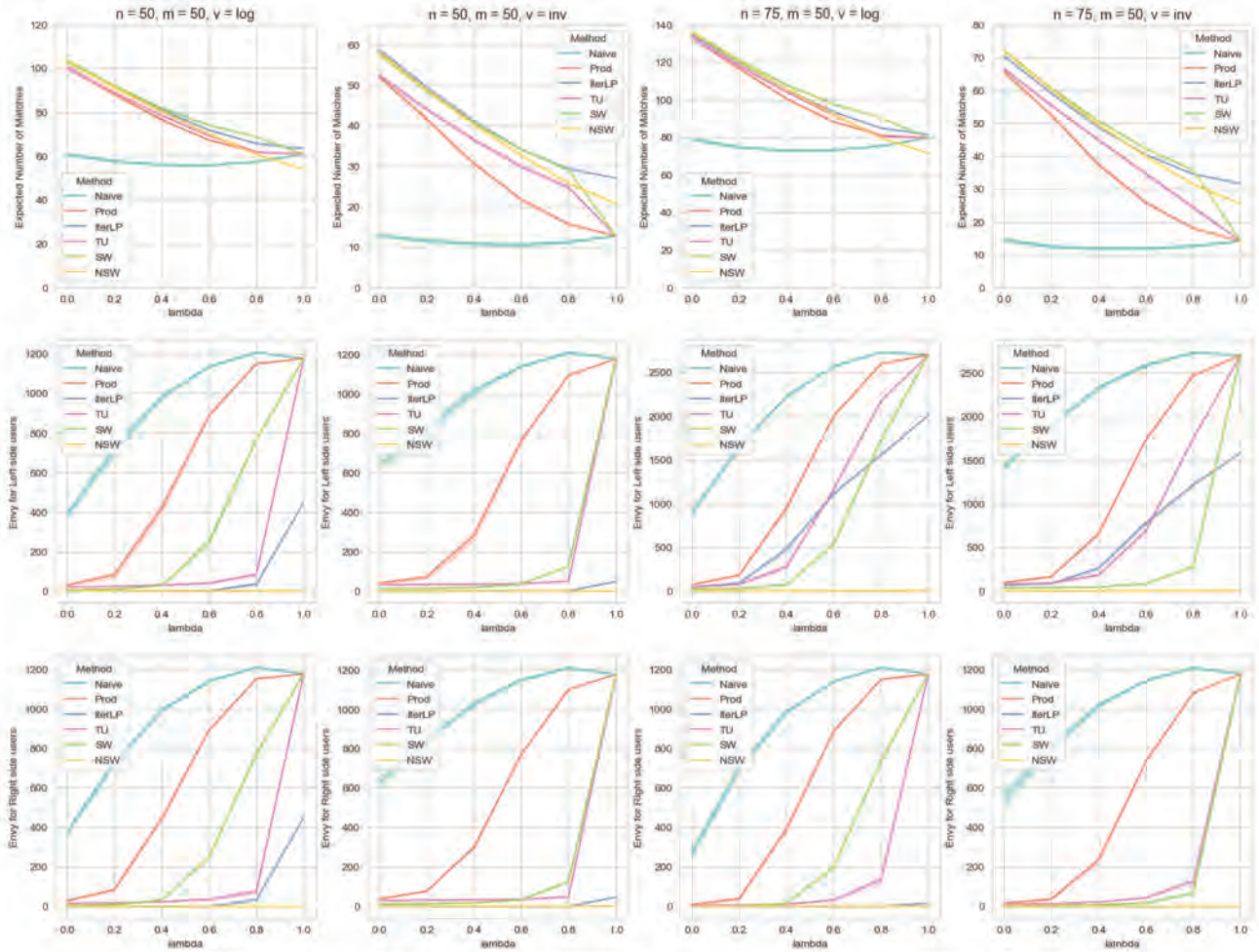


図 1　実験結果．各シナリオについて 10 回ずつ実験を行い，期待マッチ数（上段），左側ユーザーの羨望の数（中段），右側ユーザーの羨望の数（下段）の平均と 95%信頼区間を示した．

- Prod：各ユーザーに対して選好スコアの掛け合わせた値の順に推薦する．つまり，左側ユーザー i に対しては右側ユーザーを $p_{i,j} \cdot q_{j,i}$ の降順に，右側ユーザー j に対しては左側ユーザーを $q_{j,i} \cdot p_{i,j}$ の降順に推薦する．
- IterLP：1 番目に推薦する相手から順に線形計画法によって計算された最適な推薦相手を順に推薦していく手法．詳細は [Tomita 24] を参照．
- TU：TU マッチングモデルによって計算された均衡マッチングの値によって推薦する手法 [Tomita 23]．
- SW：社会厚生交互最適化法（Algorithm 1）．
- NSW：ナッシュ厚生交互最適化法（Algorithm 2）．

6·3 結　　果

各シナリオに対して 10 回ずつ実験を行い，それらの平均を実験結果として 図 1 として示した．

まず，期待マッチ数については Naive 法を除くすべての手法で λ が大きくなるほど小さくなっている．これは，一部の人気ユーザーに人気が集中する状況になるほど全体のマッチ数を高めることが困難になることを示唆する．また，$\lambda < 1.0$ の多くのケースで，他の手法と比べ SW（社会厚生交互最適化法）が期待マッチ数を高くしている．社会厚生交互最適化法の収束解は社会厚生関数の大域最適解であることを保証するものではないものの，多くのケースで社会厚生関数を比較的大きくしていることが示された．一方，$\lambda = 1.0$ の極端に人気が集中するケースにおいては，SW は局所最適解に陥ってしまうことも示唆している．NSW（ナッシュ厚生交互最適化法）は，SW ほどではないものの多くのケースにおいて比較的高い期待マッチ数を得ていることが分かる．

次に，不公平性についての指標となる羨望の数については，すべての推薦手法において傾向として λ が大きくなり人気が一部のユーザーに集中するほど羨望の数が多くなり，より不公平な推薦となる．また，左側ユーザーと右側ユーザーの数が異なる状況では，数の多い左側ユーザーの方でより多くの羨望が発生し不公平となりやすいことがわかる．社会厚生を高くする SW 法においても，$\lambda = 0.8$ 程度と十分人気が集中する状況になると多くの羨望が発生してしまっていることがわかる．しかし，NSW（ナッシュ厚生交互最適化法）においてはすべてのケースで羨望の数はほぼ 0 に近く，非常に公平な推薦が達成されている．これにより，NSW 法は期待マッチ数は比較的

多く，公平性については非常に公平な推薦となっていることが示された．

7. 結　　論

本稿では，マッチングプラットフォームにおける相互推薦システムの，被推薦機会の公平性について検討した．公平分割理論における無羨望性の概念を用いて被推薦機会の公平性を定義し，ナッシュ厚生を交互に最適化する公平相互推薦手法を提案した．実験により，期待マッチ数と公平性にはトレードオフがあるが，ナッシュ厚生交互最適化法は他の手法と比べて比較的多くの期待マッチ数を実現する一方で，羨望がほぼ0となる非常に公平な相互推薦手法であることを示した．

本稿で示したナッシュ厚生交互最適化法については，実務上応用するにはいくかの注意が必要である．まずナッシュ厚生交互最適化法は，本稿で示したマッチングプラットフォームモデルに依存したものとなっている．しかし，このモデルは現実のマッチングプラットフォームの問題を簡単化したものであるため，実際のプラットフォームにおいては異なる点がある可能性が高い．そのような少し相違点のある現実のプラットフォームにおいても本稿の手法が有効かどうかは，オフライン実験等を行い十分検討する必要がある．また，本稿の手法は $2RLK$ 個の変数を最適化する手法となっているため，ユーザー数の多い大規模なプラットフォームにおいてはこの手法をそのまま実行することは計算的に困難である．そのようなプラットフォームにおいても一部のユーザーに対してこの手法を適用することは可能であるが，そのように応用することが有効であるかは十分検討する必要がある．また，より多くのユーザー数においても適用可能となるスケーラブルな公平相互推薦手法の開発は，今後の研究課題の一つとなるだろう．

◇参　考　文　献◇

[Tomita 23] Tomita, Y., Togashi, R., Hashizume, Y., and Ohsaka, N.: Fast and examination-agnostic reciprocal recommendation in matching markets, in *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 12–23 (2023)

[Tomita 24] Tomita, Y. and Yokoyama, T.: Fair Reciprocal Recommendation in Matching Markets, in *Proceedings of the 18th ACM Conference on Recommender Systems*, pp. 209–218 (2024)

———著　者　紹　介———

冨田　燿志

2019 年東京大学大学院経済学研究科修士課程修了。2020 年サイバーエージェント新卒入社。AI Lab 経済学社会実装チームでリサーチサイエンティストとしてマッチング理論・マーケットデザイン・相互推薦システムの研究と社会実装プロジェクトを行う。

GPTはデザインの原則に注目したグラフィックデザインの評価はできるのか？

原口 大地
Daichi Haraguchi

AI Lab
リサーチサイエンティスト
daichi_haraguchi_xa@cyberagent.co.jp

keywords: LLM-as-a-judge, GPT, グラフィックデザイン, 評価

Summary

大規模マルチモーダルモデル（LMMs）の発展に伴い，さまざまな研究分野でLMMsを用いたタスクの評価が行われている．グラフィックデザイン生成研究においても，LMMsが生成されたグラフィックデザインの品質を適切に評価できると仮定し，LMMsを用いて評価する試みが普及しつつある．しかし，その妥当性の検証は十分には行われておらず，その評価が信頼できるかどうかは不明である．グラフィックデザインの品質を評価する一つの方法は，グラフィックデザインが基本的なデザイン原則に従っているかどうかを評価することである．本稿では，実験を通して収集した人手によるデザインの品質に関するアノテーションを用いて，デザイン原則に基づくヒューリスティック評価とGPT評価の挙動を比較する．実験では，GPTが細かいディテールを区別することはできないものの，人手によるアノテーションと一定の相関を持ち，デザイン原則に基づくヒューリスティック評価よりもGPTを活用した評価の方が人手によるアノテーションと相関の高い傾向を示すことを明らかにする．

1. はじめに

基盤モデルは大規模なコーパスで学習され，様々なタスクにおいて顕著な汎化能力を示している．グラフィックデザイン関連では生成タスクにおいて顕著な研究の進展がみられている [Chen 24, Cheng 24, Jia 23, Inoue 24]．生成などの特定のグラフィックデザインタスクにおいて成功を収めている一方で，GPT-4oのような基盤モデルがグラフィックデザインの品質を信頼性をもって評価できるかどうかは依然として明らかではない．それにも関わらず，近年のグラフィックデザイン生成の研究では，大規模マルチモーダルモデル（LMMs），特にGPT-4V [Achiam 23] を用いて品質を直接評価する試みが行われている．

適切にグラフィックデザインを評価するアプローチとして，人間による主観評価が活用されることもある．しかし，その評価には莫大な時間と人的コストがかかるため大規模な評価を行うのは容易ではない．グラフィックデザインの自動生成の初期の試みの一つとして，ヒューリスティック評価を用いた最適化が導入された事例がある [Peter 14]．これは，一般に高品質なグラフィックデザインは，整列や反復といったデザイナーの共通のデザイン原則 [Graham 02, Williams 14] に従う傾向があることを活用したものである．それらの原則を定式化し，最適化するアプローチをグラフィックデザイン生成に導入している．

本研究でも，高品質なグラフィックデザインの評価の基準としてデザインの原則を活用する．理由は以下の2点である．第一に，ある程度ルールの決まっている原則を利用することで，人間でも評価がしやすい点である．第二に，定式化されていることからGPT評価の性能を客観的に検証することが可能であるためである．

本研究では，図1に示すように，グラフィックデザインの評価におけるGPTの挙動を定量的に調査する．GPT評価の性能を調査するために，グラフィックデザインのヒューリスティック評価をベースラインとして採用する．整列，重なり，空白の3つの代表的なデザイン原則にわたって，グラフィックデザインを評価することにより，これらのアプローチを比較する．オンラインサービスから収集したグラフィックザインおよびそれらに摂動（ノイズ）を加えたグラフィックデザインを評価に利用する．被験者にこれらのデザインをデザイン原則に基づいて評価してもらう．この人手による評価アノテーションを用いて，ヒューリスティック評価とGPT評価のどちらが人間の評価とより一致するかを確認する．また，ヒューリスティック評価とGPT評価の定性的評価についても議論する．以下に本研究の貢献をまとめる．

- グラフィックデザインにおけるGPT評価が人間の評価に対してどの程度の相関があるのかを定量的に分析した．
- デザイン品質を評価するために，異なる品質のグラフィックデザインの人手アノテーション付き評価データセットを構築した．

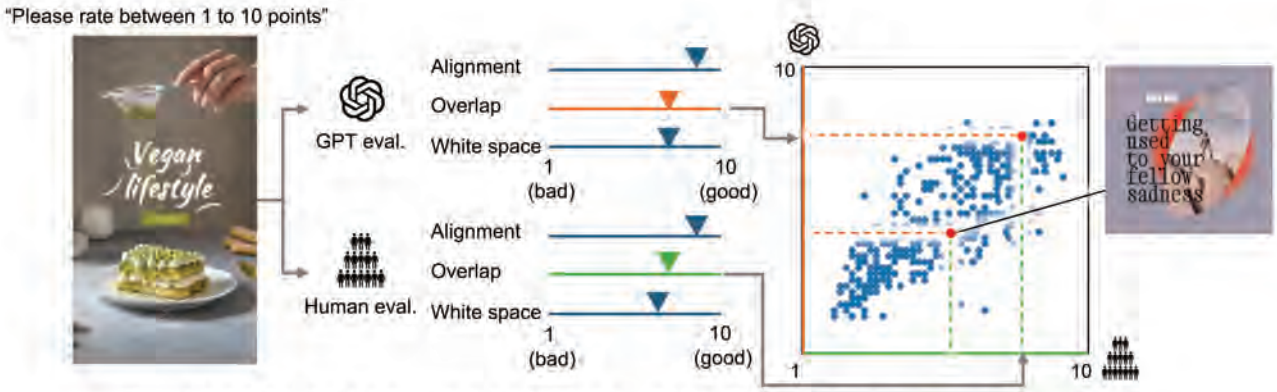


図 1　本研究の全体像．グラフィックデザインにおける「整列」,「重なり」，および「余白」の 3 つのデザイン原則に基づき GPT 評価の能力を調査する．

図 2　3 つのデザイン原則に反する例

- 人手のアノテーションと GPT 評価はヒューリスティック評価よりも高い相関が見られ，特定の条件下で GPT がグラフィックデザインの評価に利用可能であることが明らかになった．

2. グラフィックデザインの原則

グラフィックデザインの原則は，美的な作品を制作するためのデザイナーの共通の原則である [Graham 02, Williams 14]．本研究では，グラフィックデザイン生成のいくつかの研究で使用されている代表的なデザイン原則である整列，重なり，および余白を採用する [Peter 14, Kong 22]．整列と重なりは，レイアウト生成の評価において一般的に使用される [Li 20]．余白もまた，グラフィックデザイン生成アプローチにおいて重要な要素である [Kong 22]．本研究では，文献 [Peter 14] に基づく 3 つのデザイン原則に従う．以下に，それぞれの原則を簡単に説明する．

i. 整　列

要素の配置が共通の行または列に沿って並ぶように配置されることで，秩序と構造を表現する．

(1) 水平方向および垂直方向の整列が考慮される．

(2) 一見して整列しているがわずかにずれている要素は，視覚的に不快であるため，ペナルティを課される．

(3) 大きな整列グループ（すなわち，互いに離れた位置にある整列された要素）は，要素間の統一感が増すため，好まれる．

ii. 重なり

不適切な重なりは可読性を低下させる．

(1) テキスト間の重なり，テキストとグラフィック要素の重なり，グラフィック要素間の重なりの 3 種類の重なりが考慮される．

(2) テキストと背景色の色のコントラストが不十分で読みづらいテキストはペナルティを課される．

(3) 要素が境界を超えて広がっているグラフィックデザインもペナルティを課される．

iii. 余　白

余白は，デザインにおいて適切なスペースを確保することで，可読性を向上させる．

(1) デザイン要素（例：グラフィックやテキスト）で覆われていない大きな割合の余白が好まれる．

(2) ただし，画像上の空白の領域が大きすぎるグラフィックデザインは望ましくない．

(3) 各要素間の距離が大きい方が好まれる．

(4) 各テキスト要素の均一な垂直間隔が好まれる．

(5) 各要素の周りの余白（すなわち，画像の端の余白）が広い方が好まれる．

これらのデザイン原則に反する視覚的に魅力のない例を 図 2 に示す．

3. 評　価　方　法

本研究では，ヒューリスティック評価と GPT ベース評価を比較する．両アプローチにおいて，グラフィックデザインを入力し，その入力のスコアを取得する．両指標の下限と上限を設定し，値が高いほど品質が良いことを示す．

3·1 ヒューリスティック評価指標

ヒューリスティック評価指標として，O'Donovan ら [Peter 14] が上述のデザインの基本原則を定式化したものを採用する．これらはグラフィックデザインの最適化のために設計されたものである．この指標は，テキストやそ

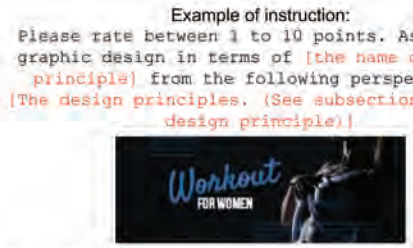


図 3　GPT によるグラフィックデザインの評価方法．赤文字で記載の箇所はデザインの原則に関するプレースホルダーである．

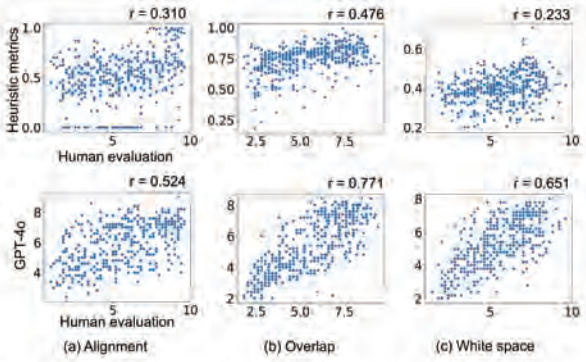


図 4　人手によるアノテーションとヒューリスティック評価または GPT 評価の相関．r はピアソン相関係数である．

図 5　整列に関するスコアの例．

の他のグラフィック要素のサイズと座標などの情報から直接品質評価をするものである．本実験では指標の範囲は 0 から 1 に正規化し，使用する．

3·2　GPT 評価指標

GPT-4o [*1] にプロンプトを与え，整列，重なり，および余白に関するスコアを求める．プロンプトは，[Peter 14] に記載されたデザイン原則と数式の説明に基づいて作成した．元のデザインをレンダリングおよびラスタライズし，入力プロンプト用の画像を作成する．図 3 に例を示すように，GPT-4o は特定の原則に基づいて入力に対して 1 から 10 のスコアを付与する．

4.　データセット

VistaCreate [*2] からグラフィックデザインテンプレートを収集した．VistaCreate には多数のバナーやポスターデザインがホストされている．テンプレートには，デザイン内のグラフィックおよびテキスト要素の座標，サイズ情報などが含まれている．VistaCreate からランダムに 100 のテンプレートをサンプリングし，座標およびサイズのパラメータに摂動を加え，美的に劣るサンプルを作成した．実験では，x 座標とフォントサイズの 2 種類の摂動を適用した．テキスト位置の x 座標の摂動データにより整列を評価し，フォントサイズの摂動データにより重なりと余白を評価する．指標のセンシティビティを調査するために，摂動の範囲を小，中，大の 3 つに設定した．100 のオリジナルサンプルと 600 の摂動サンプルを組み合わせて，合計 700 のサンプルからなるデータセットを構築した．各デザインについて，整列，重なり，および余白に関する人間のスコアのために，クラウドソーシングを通じて各グラフィックデザインごとに 5 人からアノテーションを収集した．参加者には，1 から 10 の範囲でスコアを付けるよう依頼した．実験では，5 人のアノテーションの平均スコアを使用した．

5.　実　　験

5·1　定　量　評　価

i.　設　定

人手の評価との公平な比較と詳細な分析のため，GPT 評価を 5 回実施し，平均スコアを使用する．実験では多様性を考慮し，GPT のランダム性を制御する温度パラメータは 1 に設定した．なお，温度パラメータは最小値 0，最大値 2 であり，温度が高いほど多様な出力を示す．

ii.　人手による評価との相関

人手のアノテーションと自動評価スコアの相関を分析するため，評価スコアの散布図を 2 種類準備した．ヒューリスティック評価と人間の評価を比較するものと，各デザイン原則について GPT 評価と人間の評価を比較するものである（図 4）．興味深いことに，GPT 評価はヒューリスティック評価よりも人間の評価との相関が良好であった．これは，ヒューリスティック評価がグラフィックデザインの修正前後を比較してデザインを最適化するために設計されているため，絶対的な評価に向かないことが要因にあると考えられる．異なるグラフィックデザインとそのスコアの比較例を 図 5 に示す．ヒューリスティック評価では，デザインのスコアの関係は (a)>(b)>(c) である．し

*1　https://openai.com/index/hello-gpt-4o/
*2　https://create.vista.com

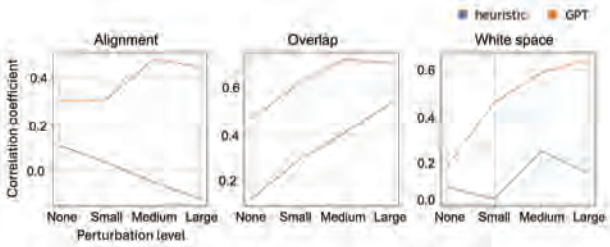


図 6　人間の評価と各手法のスコア間のピアソン相関係数.

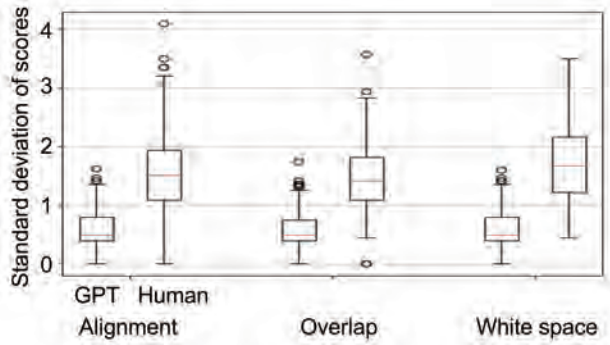


図 7　スコアの標準偏差の箱ひげ図.

かし，GPT 評価と人間の評価による順序は (a)>(c)≥(b) である．ヒューリスティック評価では，(c) のグラフィックデザインが人間のデザイナーによって作成されたにもかかわらず，スコアが非常に低くなっている．一方，GPT 評価と人間の評価はそれほど大きな違いがない．この結果は，GPT 評価がヒューリスティック評価と比較し，グラフィックデザインの絶対評価のような実用的な評価に活用可能であることを示唆している．

iii.　センシティビティ

図 6 に示すように，各摂動レベルに対する相関係数を計算することで，GPT 評価が多様な品質のデザインに対して有効かどうかを調査した．GPT 評価は，全てのデザイン原則および摂動レベルにおいて，ヒューリスティック評価よりも人手によるアノテーションとの相関が強い．また，摂動レベルが上がるにつれて，GPT 評価の相関も増加する傾向がある．これは，摂動が顕著な場合，例えば，見た目が明らかに悪いグラフィックデザインの場合，安定した評価を人間よりも容易かつ効率的に実施できることを示唆している．

iv.　信頼性

GPT を複数回実行すると異なるスコアが得られる可能性があるため，GPT 評価がどれほど安定しているかを調査した．図 7 に GPT 評価と人手によるアノテーションの標準偏差を示す．GPT 評価の標準偏差は，人間の評価スコアの標準偏差よりも低い．この結果は，各原則に対して 1 回の実行で得られる GPT 評価が実用的かつコスト効果の高い評価指標であることを示唆している．

図 8　GPT 評価による正しい余白評価のサンプル.

図 9　ヒューリスティック評価によって正しく評価されたサンプルの例.

5·2　定性的分析

GPT 評価が成功し，ヒューリスティック評価が失敗する典型的なケースとその逆のケースを示す．図 8 に示すように，背景にオブジェクトが含まれているものの，ヒューリスティック評価は背景を一貫して余白と見なすため，GPT 評価の方が正しく評価できている．背景に埋め込まれたオブジェクトを含むグラフィックデザインの評価は，今回使用したヒューリスティック評価では困難である．また，図 9 にヒューリスティック評価のみが成功するケースも示す．ヒューリスティック評価はベクトルベースであり，座標などの値からデザインの微細な違いを直接捉えることができる．しかし，GPT 評価はそのような微細な違いを正確に検出するのが難しい．

6.　まとめと今後の課題

本研究では，デザイン原則に焦点を当てたグラフィックデザインの評価において，GPT 評価の有効性を調査した．この目的を達成するために，大規模な人間のアノテーションを収集し，アノテーションと各評価指標の相関を分析した．実験の結果，GPT 評価が人間のアノテーションと一定の相関を示すことが明らかになった．今後の課題として，複数のデザイン原則を統合的に評価し，グラフィックデザインの全体的な良さを評価することを検討する．また，フォントの選択など，デザイン原則を超えたグラフィックデザインの評価も計画している．加えて，人間とあまり相関が見られなかったようなデザインの細かいところも正しく評価できるように，プロンプトの改良や LMMs を活用した評価方法を検討していく．

謝　　辞

本研究は九州大学内田研究室との共同研究を通して得られた成果である．ご指導いただいた内田誠一教授および議論していただいた九州大学修士2年生の三谷勇人氏に感謝申し上げる．また，本研究に関してご指導・ご協力くださったAI Labの井上直人氏，下田和氏，山口光太氏にも感謝を申し上げる．

◇参　考　文　献◇

[Achiam 23] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report, *arXiv preprint arXiv:2303.08774* (2023)

[Chen 24] Chen, J., Huang, Y., Lv, T., Cui, L., Chen, Q., and Wei, F.: TextDiffuser-2: Unleashing the Power of Language Models for Text Rendering, in *ECCV* (2024)

[Cheng 24] Cheng, Y., Zhang, Z., Yang, M., Hui, N., Li, C., Wu, X., and Shao, J.: Graphic Design with Large Multimodal Model, *arXiv preprint arXiv:2404.14368* (2024)

[Graham 02] Graham, L.: *Basics of Design: Layout and Typography for Beginners*, Delmar Cengage Learning (2002)

[Inoue 24] Inoue, N., Masui, K., Shimoda, W., and Yamaguchi, K.: OpenCOLE: Towards Reproducible Automatic Graphic Design Generation, in *CVPRW* (2024)

[Jia 23] Jia, P., Li, C., Liu, Z., Shen, Y., Chen, X., Yuan, Y., Zheng, Y., Chen, D., Li, J., Xie, X., et al.: COLE: A Hierarchical Generation Framework for Graphic Design, *arXiv preprint arXiv:2311.16974* (2023)

[Kong 22] Kong, W., Jiang, Z., Sun, S., Guo, Z., Cui, W., Liu, T., Lou, J., and Zhang, D.: Aesthetics++: Refining graphic designs by exploring design principles and human preference, *IEEE TVCG*, Vol. 29, No. 6 (2022)

[Li 20] Li, J., Yang, J., Zhang, J., Liu, C., Wang, C., and Xu, T.: Attribute-conditioned layout gan for automatic graphic design, *IEEE TVCG*, Vol. 27, No. 10 (2020)

[Peter 14] Peter O'Donovan, Agarwala, A., Hertzmann, A.: Learning layouts for single-pagegraphic designs, *IEEE TVCG*, Vol. 20, No. 8 (2014)

[Williams 14] Williams, R.: *Non-Designer's Design Book*, Peachpit Press (2014)

——著　者　紹　介——

原口　大地

株式会社サイバーエージェントのリサーチサイエンティスト．2024年3月に九州大学大学院システム情報科学府博士後期課程を修了し，博士(情報科学)を取得．同年4月に株式会社サイバーエージェント中途入社．専門とする研究領域はビジュアルデザインインフォマティクスである．

編集後記

White Paper Project（以下、WPP）は、サイバーエージェントの各組織で行われている AI/Data系の研究開発を論文化するプロジェクトです。2017年より年2回の頻度で発行を続けてきました。
WPPは、以下の目的のもとに実施されています。

AI/Data系の研究開発に関する社内認知の向上
社内に散在するAI/Data系の技術資産の集約
秘匿情報を含む研究成果の適切なアウトプット
技術共有による車輪の再発明の防止
長期的な研究開発における中間的な情報共有
今回は、社内限定で集められた論文集の中から社外公開可能なものを選定し、社外公開版WPPを発行する運びとなりました。

本プロジェクトを通じて、サイバーエージェントの多様な事業ドメインにおける研究開発の取り組みを知っていただく機会となれば幸いです。

White Paper Project 運営

White Paper Project

著者

赤間 怜奈
Antonio Tejero-de-Pablos
乾 健太郎
大内 啓樹
大竹 真太
大谷 まゆ
奥村 学
上垣外 英剛
亀井 遼平
栗原 健太郎
坂井 優介
坂田 将樹
佐野 幸恵
佐藤 志貴
佐藤 真
高村 大也
武内 慎
張 培楠
富樫 陸
冨田 燿志
馬場 惇
原口 大地
邊土名 朝飛
三田 雅人
村上 聡一朗
山田 康輔
渡辺 太郎

※著者の所属は発行当時のものです

運営

井上 翔太
下田 和
鈴木 智之
田中 宏樹
友松 祐太
原口 大地
松月 大輔

デザイン

後谷 莉子（Design Factory）


CyberAgent