システム発話を起点とした雑談会話における パーソナリティを考慮した話題推薦の検討

藤本裕之 [1]　島田陽介 [1]　大野実 [1]
[1] セコム株式会社　IS 研究所
{hiroyu-fujimoto,yusuk-shimada, m-ono}@secom.co.jp

概要

ユーザのパーソナリティ情報を用いて、興味に合わせた会話を提供することが求められている。本研究では、特にシステム発話を起点とした雑談会話を想定し、ユーザのパーソナリティ情報を利用してユーザの興味のありそうな話題かどうかを推定することができるかを検討した。追加学習した BERT がベースラインを上回る性能だった一方、GPT-4o はベースラインを下回った。パーソナリティ情報から興味の有無を一定の性能で判定できることを確認した。

1　はじめに

ユーザとシステムが長く雑談会話をする上では、信頼や親近感を感じてもらうことが重要である [1]。人は嗜好が一致する対話相手に親近感を抱きやすいと言われている [2]。そのような背景からユーザのパーソナリティ情報を用いて、会話をパーソナライズする研究が進められている。

高齢者への声かけによって社会的なつながりを促進できるという報告がある [3]。ユーザにとって興味のある話題であれば親近感を感じてもらいやすく、声かけの効果が高まると考えられる。ユーザに対して「最近どんなことがありましたか？」などのようにユーザに対して話題提供を促す場合もある [4]。システム側からユーザの興味のある話題を提供できると、嗜好の一致や親近感を抱かせやすいと考えられる。

本研究では、システム発話を起点とした雑談会話を想定し、ユーザのパーソナリティ情報を利用してユーザの興味のありそうな話題かどうかを推定する。

2　関連研究

雑談会話中に、ユーザ発話文や音声や画像などのマルチモーダル情報を用いて、感情極性分析やシステム発話に対するユーザの嗜好や興味有無を推定する研究 [5, 6, 7, 8] や、パーソナリティを考慮して会話継続の可否を推定する研究がいくつかある [9, 10]。これらは雑談会話中の推定を想定しており、ある話題に対するユーザの反応を使って話題への興味有無を推定する。

性別や年代ごとに話題推薦する研究もあるが、個人の嗜好を考慮できない [11]。また、ユーザの過去の会話履歴を基に話題を推薦する研究 [12, 13] もあるが、実用的な精度でない。

大規模言語モデルを評価者として様々なタスクの評価をさせる LLM-as-a-Judge の研究も多く報告されている [14, 15]。また推薦タスクに利用する研究も報告されている [16, 17]。しかし、雑談会話の話題推薦での有効性は明らかでない。

本研究では、雑談会話中の発話は利用せず、性別・年代を含むより多くのパーソナリティ情報から興味有無を判定することが実用的な精度で可能なのかを検証する。

3　提案手法

本研究では、話題推薦用データセットの作成方法と話題推薦手法を提案する。話題推薦手法としては、追加学習を必要としない OpenAI の gpt-4o(2024-05-13) [1)] を利用した OpenAI モデルと、追加学習を必要とする BERT モデルを提案する。

1）https://learn.microsoft.com/en-us/azure/ai-services/openai/

3.1 話題推薦用データセットの作成

RealPersonaChat[18, 19] は実在する 233 話者同士による約 14,000 件の雑談会話データセットで、全話者の統計情報, ペルソナ文, 性格特性などが付与されている。また各会話終了時に、それぞれの話者がお互いに興味度を含む 6 つの観点で会話を評価した人手評価値も付与されている。 RealPersonaChat に対し、OpenAI の gpt-4o(2024-05-13) [1] を用いて、各会話の話題ラベルを推定した。

本研究では、会話に付与された 5 段階の興味度を、その話題に対する興味度として扱うことで、話題と興味度を対応付けた。話題ラベルとその話者のパーソナリティ情報を入力し、その話者が実際に付与した人手評価値である興味度を予測できるかというタスクとする。予測した興味度は推薦スコアとする。

話題ラベルの推定は、各会話ごとに OpenAI の gpt-4o(2024-05-13) [1] に会話全文を入力し、自然文で話題を出力することで取得した。

3.1.1 パーソナリティ情報

統計情報 性別、年齢、学歴、職業、居住地

ペルソナ文 話者についての自己紹介文。本研究では、会話中に現れたペルソナも gpt-4o(2024-05-13) [1] に会話全文を入力し、自然文で話者ごとに出力することで取得した。RealPersonaChat でのペルソナ文が話者ごとに 10 件だったが、それに加えて話者ごとに 5 件から 1000 件超のペルソナ文を追加した。

性格特性 BigFive 特性（開放性、誠実性、外向性、協調性、神経症傾向ごとに 1 から 7 の範囲のスコアが付与された心理尺度）[18, 19] のみ。

3.2 話題推薦手法

提案手法の概要を図 1 に示す。入力に関しては 3.2.1 節、モデルに関しては 3.2.2, 3.2.3 節で述べる。

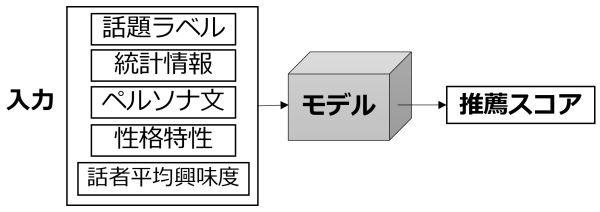

図 1 提案手法の概要

3.2.1 パーソナリティ情報の入力形式

話題ラベルと、推薦したい話者の話者統計情報、ペルソナ文、性格特性を結合して入力とする。ペルソナ文は OpenAI の text-embedding-3-large [1] を用いて、埋め込みベクトル化し、同様に埋め込んだ話題ラベルとの類似度が高い上位 10 件を各入力に用いるペルソナ文として用いる。なお、入力する該当会話から推定されたペルソナ文は除外する。

また、興味度分布には個人差があることが想定される。そこで話者平均興味度もその個人差を考慮するためにパーソナリティ情報と併せて結合・入力する。これによって、興味度の個人差を考慮した出力が期待できる。話者平均興味度は入力する該当会話以外を基に算出する。

3.2.2 OpenAI モデル

gpt-4o(2024-05-13) [1] に話題ラベルとその話者に関するパーソナリティ情報を入力し、1 から 5 の範囲の推薦スコアを出力することで、ある話者に対してある話題が推薦できるかを推定する。なお、OpenAI モデルの場合は、入力文にはタスク指示も加え、判断根拠も生成させた後に推薦スコアを生成させた。

3.2.3 BERT モデル

比較のために、東北大学が公開している bert-large-japanese-v2 [2] を追加学習する。話題とパーソナリティ情報を、異なる segment として埋め込んで、推薦スコアを出力する回帰モデルとして学習した。分類 Head は全結合層であり、[CLS] トークン特徴を入力する。学習の詳細については 4.2 節で述べる。なお、BERT モデルの場合は入力文に [CLS] トークンなどの特殊トークンも加わる。

3.2.4 ベースライン

興味度分布には個人差があり、かつ分散が小さいことがわかっている。そのため話者ごとの平均興味度を算出しておき、話題ラベルによらず、話題の興味度の推定値とする方法も有効であると考えられる。したがって、これをベースラインとした。なお、ベースラインはモデルを介さず、単に話者平均興味度を推薦スコアとして出力する。

2） https://github.com/cl-tohoku/bert-japanese

4　実験

4.1　評価データセット

RealPersonaChat の会話数は 13,581 件で、各話者ごとに話題ラベル、パーソナリティ情報、興味度のペアがあるため、その 2 倍の 27,162 件が利用できる。

本研究では、興味度の個人差を排除するため、話者・興味度ごとにサンプリングした評価データセットを利用する。話者ごとに興味度の 1 から 5 の評価レベルごとに 1 件ずつサンプリングする。もし該当する評価レベルが 0 件ならサンプリングしない。各評価レベルごとに同数でサンプリングすると、特定話者の占める割合が大きくなってしまうためである。話題推薦のデータセットとして扱うため、母集団の割合に則る必要はない。

その結果、評価データセットは計 890 件となった。それ以外の 26272 件を学習用データセットとして利用する。

4.2　BERT の追加学習

4.1 節で述べた学習データセットを追加学習に利用する。訓練データは 95%、検証データは 5%の割合とし、損失関数は平均二乗誤差 (MSE) とした。分類 Head は 1 層の全結合層とした。学習率は 1e-5、Epoch 数は 10 で Early Stopping を導入した。

4.3　評価指標

モデル出力の推薦スコアと正解の興味度の相関をみるために、ピアソンの積率相関係数 r とスピアマン順位相関係数 ρ を用いる。また二値分類としての性能もみるために、AUROC も用いる。

5　結果

評価データセットに対する性能を表 1 に示す。なお、ベースラインとして話者ごとの興味度平均をそのまま出力とした場合の性能も比較する。

表 1　話題推薦の性能

	評価指標		
	r	ρ	AUROC
OpenAI モデル	0.2867	0.3892	0.7091
BERT モデル	**0.4809**	**0.5047**	**0.7649**
ベースライン	0.3808	0.4492	0.7329

6　考察

6.1　話者平均興味度をベースラインとする妥当性

RealPersonaChat での興味度分布には個人差がある。図 2 にある 4 話者の興味度の分布を示す。5 が一番多く、1 が一番少ない話者もいれば、1 か 5 しかない話者もいる。話者平均興味度はこのような興味度分布の代表値であり、個人差を端的に表現できる。

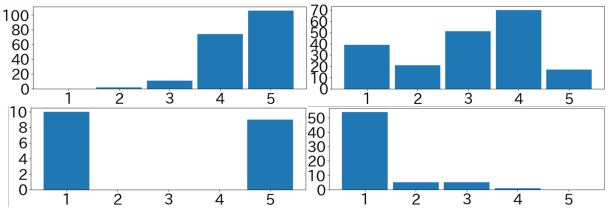

図 2　ある 4 話者の興味度分布

図 3 に話者ごとの興味度の平均と標準偏差のヒストグラムを示す。これより多くの話者は平均的に高い興味度を小さい分散で付与していることがわかる。

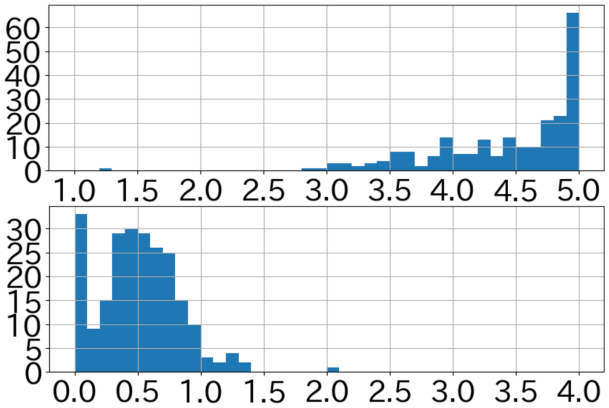

図 3　話者ごとの興味度の平均 (上) と標準偏差 (下) のヒストグラム

このことから、話者平均興味度が話題の興味度と近い値を取る場合が多いことがわかる。そのため、ベースラインの性能が高かったという側面もあると考えられる。

6.2　話題推薦の難しさ

パーソナリティ情報（主にペルソナ文）が話題に対してポジティブな情報だが、興味度は低いケース、あるいはパーソナリティ情報が話題に対してネガティブな情報だが、興味度は高いケースが散見された。例えば、以下のようなものがあった。

表 2　OpenAI モデルが有効だったケース

	興味度	スコア		
話題ラベル		OpenAI モデル	BERT モデル	ベースライン
ペット（特に猫）についての会話	5.0	4.5	4.8	2.8
体重増加と運動不足	2.0	3.0	3.8	4.3

1. ペルソナ文よりスイーツ好きな話者がスイーツの話題にて低い興味度 (1) を付与。
2. ペルソナ文より掃除は嫌いだが掃除機は好き。な話者が掃除の話題に高い興味度 (5) を付与。
3. ペルソナ文より食中毒を恐れている話者が食中毒の話題に高い興味度 (5) を付与。

1 に関しては、途中で栗のスイーツの話になり、対話相手との意見が分かれた。自分の嗜好に関わる話題だからこそこだわりがあり、関心が持てない場合があると考えられる。

2 に関しては、掃除自体は嫌いだが、掃除の話は嫌いではないように見受けられた。また 2 の話者は BigFive 特性も誠実性、外向性、協調性が高く、100 件以上の会話に全て興味度 4 以上を付与している。どんな話題でも興味を持てる人物であると考えられる。

3 に関しては、食中毒に恐れているがゆえに関心が高いことが見受けられた。

したがって、パーソナリティ情報に話題に関連するポジティブ/ネガティブな情報があったとしても興味度が高い/低いという単純な関係にはならない場合があり、話題の性質や文脈等を考慮して推論する能力が必要であることがわかる。

6.3　パーソナリティ情報が有効な例

一方で、パーソナリティ情報を考慮した推論ができたケースもあった。

ベースラインである話者平均興味度と正解である興味度の差が大きく、OpenAI モデルと興味度の差が少ないデータを表 2 に示す。OpenAI モデルには推論根拠も生成させており、それを基に考察した。

「ペット（特に猫）についての会話」に興味度 5 を付与した話者は、ペット飼育不可の物件に居住しているが、猫が一番好きな動物であり、飼育経験もある。また、この話者は興味度分布が 1 か 5 の両極端となっているため、話者平均興味度も 2.8 となっている（図 2 左下参照）。一方で OpenAI モデルはペット飼育不可というネガティブな情報を認識しつつ猫への関心の高さを考慮し、高い興味度を推定した。

「体重増加と運動不足」に興味度 2 を付与した話者は、運動はあまりしない、ダイエットについて諦めている、身体的な問題を抱えているという。また BigFive 特性の神経症傾向も高い。この話者は高い興味度を付与する場合が多く話者平均興味度も高い（図 2 左上参照）が、この話題には低い興味度を付与した。OpenAI モデルはこの話者の個人性を総合的に評価し、低い興味度を推定した。

関連するパーソナリティ情報を適切に選択し組み合わせることで、一部のケースにおいて有効な結果が得られることが確認できた。

6.4　BERT モデルで注目されたパーソナリティ情報

BERT モデル推論時にどういったパーソナリティ情報が注目されているかについて、Attention Weight を分析し考察した。

話者平均興味度が最も注目されており、単体で注目される場合が多かった。一方で話題・ペルソナ文・性格特性は同時に注目されることが多かった。入力に応じてこの 2 つの使い分けをしており、この判断が困難であったため性能が伸び悩んだと考えられる。

詳細は付録 A に記載した。

7　おわりに

本研究では、システム発話起点の雑談会話を想定し、パーソナリティ情報から興味有無を判定する手法を提案し、追加学習した BERT モデルで一定の性能で判定できることを確認した。

一方で話題推薦は同じ話題でも話者の個人性によって推薦可否が異なり、関連するパーソナリティ情報を適切に選択し組み合わせて推論することが必要であることを確認した。

今後は、話題推薦におけるパーソナリティ情報を考慮した推論をより高度化することが求められる。

参考文献

[1] Timothy W Bickmore and Rosalind W Picard. Establishing and maintaining long-term human-computer relationships. **ACM Transactions on Computer-Human Interaction (TOCHI)**, Vol. 12, No. 2, pp. 293–327, 2005.

[2] 東中竜一郎, 堂坂浩二, 磯崎秀樹. 対話システムにおける共感と自己開示の効果. 言語処理学会第 15 回年次大会発表論文集, pp. 446–449, 2009.

[3] Kazuko Obayashi and Shigeru Masuyama. Pilot and feasibility study on elderly support services using communicative robots and monitoring sensors integrated with cloud robotics. **Clinical Therapeutics**, Vol. 42, No. 2, pp. 364–371.e4, 2020.

[4] 小林峻也, 萩原将文. ユーザの嗜好や人間関係を考慮する非タスク指向型対話システム. 人工知能学会論文誌, Vol. 31, No. 1, pp. DSF–A_1, 2016.

[5] 西本遥人, 駒谷和範. 対話におけるマルチモーダル情報を用いたユーザの興味の有無の推定. 人工知能学会全国大会論文集 第 32 回 (2018), pp. 3C2OS14b04–3C2OS14b04. 一般社団法人 人工知能学会, 2018.

[6] 松本紗規子, 荒木雅弘. 雑談対話におけるマルチモーダル情報を統合した興味判定手法. 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 84 回 (2018/11), p. 23. 一般社団法人 人工知能学会, 2018.

[7] 稲葉通将, 高橋健一. ニューラルネットワークを用いた雑談対話からのユーザの興味推定. 人工知能学会論文誌, Vol. 34, No. 2, pp. E–I94_1, 2019.

[8] 目良和也, 青山正人, 黒澤義明, 竹澤寿幸. 発話内容と口調の関係に基づく発話者の嗜好情報推定. 知能と情報, Vol. 31, No. 5, pp. 816–825, 2019.

[9] 佐藤明智, 南泰浩, 金子俊太, 谷口伊織, 郭恩孚. 話題継続とペルソナを考慮した雑談対話システムの構築. 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 96 回 (2022/12), p. 47. 一般社団法人 人工知能学会, 2022.

[10] 蔦侑磨, 吉永直樹, 佐藤翔悦, 豊田正史. パーソナリティを考慮した雑談対話の会話継続可能性評価. 言語処理学会第 28 回年次大会発表論文集, pp. 583–587, 2022.

[11] 南光, 芋野美紗子, 土屋誠司, 渡部広一. 性別・年代別の嗜好情報を基にした話題語提供システム. In **IEICE Conferences Archives**. The Institute of Electronics, Information and Communication Engineers, 2012.

[12] 蔵内雄貴, 倉島健, 岩田具治, 星出高秀, 藤村考. Twitter の会話ログを利用した複数ユーザに対する話題推薦. In **DEIM Forum**, pp. A1–4, 2012.

[13] 横山慎, 馬強ほか. 話題の新鮮度を考慮したマイクロブログ推薦手法の提案. 2017 年度 情報処理学会関西支部 支部大会 講演論文集, Vol. 2017, , 2017.

[14] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of llm-as-a-judge, 2024.

[15] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. A survey on llm-as-a-judge, 2024.

[16] Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. Recommendation as instruction following: A large language model empowered recommendation approach, 2023.

[17] Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi R. Fung, Hou Pong Chan, Kevin Small, ChengXiang Zhai, and Heng Ji. Persona-db: Efficient large language model personalization for response prediction with collaborative data refinement, 2024.

[18] Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities. In **Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation**, pp. 852–861, 2023.

[19] 山下紗苗, 井上昂治, 郭傲, 望月翔太, 河原達也, 東中竜一郎. RealPersonaChat: 話者本人のペルソナと性格特性を含んだ雑談対話コーパス. 言語処理学会第 30 回年次大会発表論文集, pp. 2738–2743, 2024.

[20] Xiaobing Sun and Wei Lu. Understanding attention for text classification. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 3418–3428, Online, July 2020. Association for Computational Linguistics.

[21] 佐藤拓真, 窪田愛, 峯島宏次. BERT はどのように逆接の談話関係を判定しているか— attention と品詞を手がかりとして—. 言語処理学会第 30 回年次大会発表論文集, pp. 757–762, 2024.

A　BERT 推論時の Attention Weight 分析

BERT モデルの Attention を分析することで、推論過程や特徴重要度を説明しようとする研究が進められている [20, 21]。しかし、Attention Weight はトークン間の重要度に関する相対的な尺度でしかないと言われている [20]。また特に自然言語では長さやコンテキストの位置が変わるため、入力トークンに依存すると言われている [20]。

本考察では、BERT モデルにおいて [CLS] トークン特徴に対する各パーソナリティ情報に該当するトークン領域の Attention Weight 平均を可視化することでどのパーソナリティ情報が相対的に注目されたかを確認する。

AttentionWeight 平均の算出方法は次の通りである。興味度を出力する直前の最終層の全 Head の Attention Weight を平均し、Attention Weight(Head 平均) を取得する。 その後、Attention Weight(Head 平均) 内で [CLS] トークン特徴に対する重み、かつ各パーソナリティ情報での全該当トークンに対応する重みを平均して、Attention Weight 平均を取得した。これによって長さやコンテキストの位置を固定化し入力トークン依存を解消する。

図 4 に最終層での各パーソナリティ情報ごとの Attention Weight 平均を示す。評価データセット全件で算出し、各 ID ごとに Max-min 正規化した。

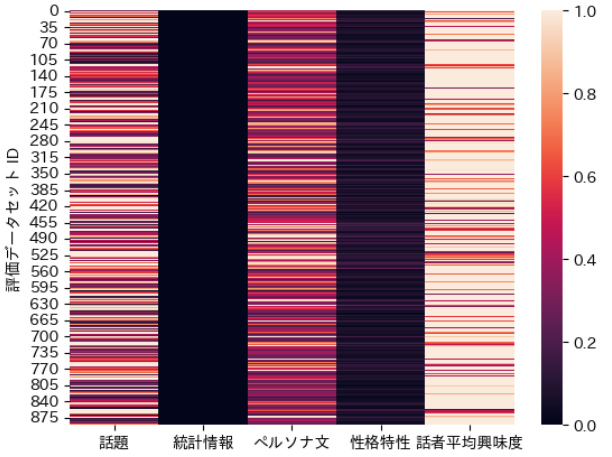

図 4　最終層での各パーソナリティ情報ごとの Attention Weight 平均

統計情報はほとんど注目されておらず、話者平均興味度話題、ペルソナ文、性格特性が主に注目されていることがわかる。また ID によって注目度が高いパーソナリティ情報は異なる。また ID によって注目度が高いパーソナリティ情報は異なる。

次に各パーソナリティ情報に対応する Attention Weight 平均同士の相関関係を図 5 に示す。ピアソン積率相関係数 r を用いた。

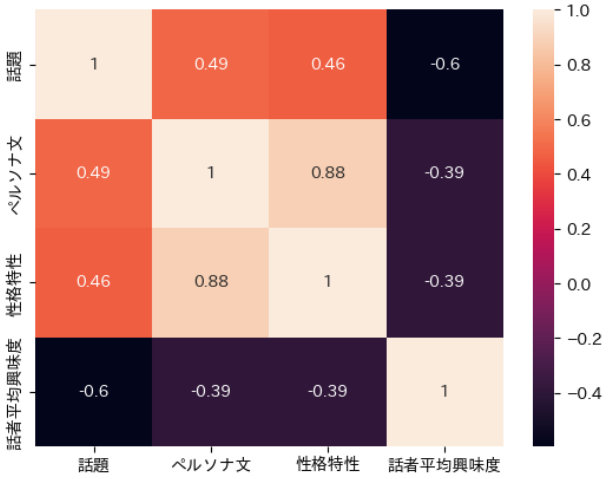

図 5　各パーソナリティ情報の Attention Weight 平均の相関関係

話者平均興味度はそれ以外のパーソナリティ情報と負の相関となっている。また、話題・ペルソナ文・性格特性に関しては互いに正の相関がある。特にペルソナ文と性格特性は強い正の相関である。このことから、話者平均興味度は単体で注目されることが多い一方、話題・ペルソナ文・性格特性は同時に注目されることが多いと考えられる。

BERT モデルでは、6.3 節で述べたようなパーソナリティ情報が必要なケースで話題・ペルソナ文・性格特性を活用できたと考えられる。一方で、どのケースで話者平均興味度のみを使うのか、他のパーソナリティ情報を使うかの判断が困難だったため、ベースラインに比べ、性能が伸び悩んだとも考えられる。

OpenAI モデルはこの推論が十分でなかったため、ベースラインを下回り、一方で BERT モデルは話題推薦における複雑な推論過程を学習によってわずかに獲得したことでベースラインを上回ったと考える。このような複雑な推論過程を解明することができれば、OpenAI モデルでもその推論過程をプロンプトで教示することで同程度の性能となる可能性がある。