

ChatGPT を用いた日本語対話応答の多面的自動評価

寺内祐希¹ 加藤恒夫¹ 田村晃裕¹ 池田和史²

¹同志社大学大学院 ²KDDI 総合研究所

¹{ctwh0179@mail4, tsukato@mail, aktamura@mail}.doshisha.ac.jp

²kz-ikeda@kddi.com

概要

人間と雑談を行う非タスク指向型対話システムの性能を簡便に測ることのできる自動評価指標が望まれている。対話の応答は多様であり、評価は様々な観点から多面的に行われることが望ましい。そこで本研究では、現在最も性能が優れた大規模言語モデル(LLM)と呼ばれる ChatGPT を用いた、日本語対話応答の多面的な評価を行う。対話の品質を左右すると考える 6 つの評価項目における日本語対話応答の主観評価データを収集し、ChatGPT による自動評価との比較を行う。その結果、ChatGPT と人間の主観評価の相関は、人間同士の相関に近いことを確認した。また、評価対象の文を英語に翻訳することにより、主観評価との相関が強まることを確認した。

1 はじめに

非タスク指向型対話システムの性能評価においては、人手による主観評価が一般的である[1]。しかし人手による評価は時間とコストがかかるため、特に開発段階においては、簡便にシステムの性能を測ることが出来る自動評価指標が望まれている。

対話の自動評価指標には、参照型の指標と非参照型の指標がある。参照型の指標としては BLEU[2]や FBD[3]等があるが、対話の応答は多様であり、単一の正解文との比較では適切に評価することが難しい。非参照型の指標としては FED[4]や GPTScore[5]など LLM を用いた手法が近年主流であり、これらは人手による主観評価との相関が比較的高いことが示されている。その中でも、人間のフィードバックにより学習した報酬モデルを用いた強化学習を行った LLM である ChatGPT によって対話の評価を行う研究が特に活発に行われており、従来手法より人手との相関が高いことが報告されている[6,7,8]。しかし、これらは英語のデータセットにより検証されている。日本語対話データでは FED の検証が行われている[9,10]が、ChatGPT を用いた日本語対話応答の評価は

著者の知る限り現在行われていない。そこで本研究では、ChatGPT により日本語対話応答の評価を行い、日本語対話応答の自動評価における ChatGPT の活用の有効性を検証する。

対話の応答は多様であり、評価の際には全体的な品質だけでなく、複数の側面を多面的に評価することが望ましいとされる[11]。一方で、評価項目が統一化されていないという問題点がある[12]。対話には様々な評価項目が存在するが、統一された定義や名称がなく、似た意味の評価項目でも研究によって定義や名称が微妙に異なる。また、人手による主観評価データを収集する際、コストや評価者の負担を考慮して特定の評価項目に絞ることが多い。これらの要因から、対話の評価項目は研究により様々で統一されていない。この問題点に留意して、本研究では先行研究をベースに対話の品質を測るうえで重要な 6 つの評価項目を選定し、主観評価データの収集を行う。

主観評価データを収集した対話応答について ChatGPT により同じ評価項目の自動評価を行った結果、自動評価と主観評価の相関は、人間同士の相関に近いことを確認した。また、ChatGPT では学習データの性質上、要約などのタスクにおいて、英語での精度が他言語に比べて高いことが報告されている[15]。本研究では対話の自動評価に関して日本語と英語間の性能差を比較するために、評価対象の文を英語に翻訳して検証した。その結果、ほとんどの評価項目で主観評価との相関が向上した。

2 関連研究

FED[4]は DialoGPT[13]を用いて、対話履歴が与えられた際に評価用に指定したフォローアップ(FU)発話で強制デコーディングを行い、その FU 発話の対数尤度を測ることで評価を行う指標である。FED では計 18 個の細かい評価項目が用いられ、収集した主観評価データと中程度の相関があることが報告されている。

表1 TopicalChat データの評価項目と定義

評価項目	定義
Understandable (0,1)	理解可能であるか
Natural (1,2,3)	自然な文であるか
Maintains Context (1,2,3)	対話文脈の続きとして有効な文であるか
Interesting (1,2,3)	興味深いか
Uses Knowledge (0,1)	対話文脈で提示された知識を利用した応答であるか
Overall Quality (1,2,3,4,5)	上記5項目を踏まえた総合的な品質

G-EVAL[6]ではChatGPTを用いて対話の自動評価を行っている。評価を行う際にChatGPTに与えるプロンプトとして、評価タスクの定義と、各評価項目の評価基準に加え、Chain of Thought(CoT)という中間命令を与える。G-EVALはCoTを用いて評価手順をLLMに詳細化させることで、人手による評価と高い相関を実現した。

複数のモデルを組み合わせて対話の評価を行うUSR[14]では、人間同士の会話を収集したTopicalChatコーパスを対話文脈として、複数のシステムによって生成された応答文に対して人手による主観評価が行われている。TopicalChatデータにおける評価項目とその定義を表1に示す。TopicalChatデータは最新の対話評価に関する研究においてよく用いられており[6,7,8]、対話における重要な品質を測定していると考えられる。

ChatGPTにおける英語と他言語での性能差の検証はいくつか行われている。Laiら[15]は、要約など7つの自然言語処理タスクにおいて、日本語を含む37の言語でChatGPTを評価しており、ほとんどのタスクにおいて英語が他言語に比べて高い性能を示すことを報告している。尚、この研究では対話の自動評価に関する検証は行われていない。

3 主観評価データの収集

3.1 クラウドソーシングによる主観評価

ChatGPTを用いて日本語対話応答の多面的な評価を行うために、クラウドソーシングによって評価対象となる主観評価データを収集する。

評価対象となる応答文を生成するために、NTT社の

表2 主観評価データの評価項目と定義

評価項目	定義
自然さ(1,2,3)	日本語として自然な文であるか
文脈的整合性 (0,1)	応答文が対話文脈に対して噛み合っているか
興味深さ (1,2,3)	応答文が興味深い文であるか
話題の関連性 (1,2,3)	応答文がそれまでの対話文脈の話題に沿った応答であるか
首尾一貫性 (0,1)	応答文が会話を通して一貫した内容で矛盾していないか
総合的な品質 (1,2,3,4,5)	上記5項目を踏まえた応答文に対する全体的な印象

Transformer[16]をベースとした対話モデル[17]と、ELYZA社の日本語LLM¹⁾、そしてrinna社の日本語対話GPT²⁾の3種のモデルを用いる。入力文には、NTT社が公開している日本語版PersonaChatデータから抽出した4ターンの対話文脈を与え、それに続く応答文をモデルごとに1文ずつ、計3種類生成する。評価者は分類の定義や方法を示した作業手順書を読んだ後、割り当てられた応答文を評価する。本研究では750種の対話文脈に対して、3つのモデルで生成した応答文、合計2,250種の応答文の評価を行う。1ファイルあたりの対話文脈が30種、応答文が90種となるように均等に分割し、25種類のファイルを作成した。1つのファイルにつき5人の評価者が評価を行う。評価者による偏りを防ぐために、1人あたりが評価できるファイル数の上限を3個とした。

評価項目については、2節で示したTopicalChatデータをベースに、対話の品質にとって重要であると考えられる6項目を評価項目に選定し、主観評価を行う。選定した評価項目とそれらの定義を表2に示す。

TopicalChatデータの評価項目より、「Natural」を「自然さ」、「Interesting」を「興味深さ」、「Overall Quality」を「総合的な品質」として評価項目に選定した。「Understandable」については近年の対話システムではほとんどの応答文が理解可能であることからG-EVAL[6]において評価項目から除外されており、本論文でも除外した。「Uses Knowledge」については、対話文脈で何らかの知識が与えられるTopicalChatの性質に依存した評価項目であり、本研究では除外した。残る「Maintains Context」について

1) [elyza/ELYZA-japanese-llama-2-7b-instruct · Hugging Face](https://huggingface.co/elyza/ELYZA-japanese-llama-2-7b-instruct)

2) [rinna/japanese-gpt-neox-3.6b-instruction-ppo · Hugging Face](https://huggingface.co/rinna/japanese-gpt-neox-3.6b-instruction-ppo)

表 3 各評価項目間の主観評価値の相関係数（スピアマン相関）

	自然さ	文脈的整合性	興味深さ	話題の関連性	首尾一貫性	総合的な品質
自然さ	—————	0.71	0.36	0.63	0.70	0.77
文脈的整合性	0.71	—————	0.25	0.65	0.82	0.87
興味深さ	0.36	0.25	—————	0.57	0.29	0.47
話題の関連性	0.63	0.65	0.57	—————	0.63	0.76
首尾一貫性	0.70	0.82	0.29	0.63	—————	0.81
総合的な品質	0.77	0.87	0.47	0.76	0.81	—————

は、対話文脈の続きとして有効であるかという項目であるが、他の項目に比べて、話題に沿っているのか、それまでの対話の経緯を無視していないかといった幅広い観点を含んでおり、FED においては「Coherent」と「Relevant」といった別の項目として扱われているため、「文脈的整合性」と「話題の関連性」の2つに分割した。

加えて、FED の評価項目に含まれる「Consistent」（発話に矛盾がないか）については対話における矛盾を検出する研究もいくつか行われている[18,19]ことから、発話が一貫して矛盾のないことは対話において重要だと考えられるため、「首尾一貫性」として評価項目に選定した。評価のレベル数は TopicalChat データに準拠したうえで、新たに追加した項目は各レベルの定義が明瞭になるように設定した結果、「文脈的整合性」、「首尾一貫性」を2段階、「自然さ」、「興味深さ」、「話題の関連性」は3段階、「総合的な品質」は5段階とした。どの項目も評価値が大きくなるほど良い。（付録 A を参照）

3.2 主観評価データの分析

本節では収集した主観評価データの分析を行う。表 3 に各評価項目間のスピアマンの相関係数を示す。尚、相関係数の算出には 5 人の評価者の評価値の平均値を用いている。TopicalChat データの評価項目を2つに分割した「文脈的整合性」と「話題の関連性」の相関係数は 0.65 であり、他の項目間の相関係数と比較して高すぎる値では無く、分割は妥当であるといえる。また、「興味深さ」は他の項目との相関が突出して低い。これは他の項目が対話の継続に大きく関わる項目であるのに対し、「興味深さ」が低い場合、即ち「そうですね」といった汎用的な応答でも対話は継続する機会が多いためであると考えられる。「総合的な品質」との相関は「文脈的整合性」が最も高く、評価者はこの項目を応答文の品質を判断する際に重視しているといえる。表 4 に各評価項目の評価値の平均、標準偏差、評価者間の相

表 4 各評価項目の平均値、標準偏差、評価者間の相関係数(r_s :スピアマン相関, r_p :ピアソン相関)

評価項目	平均	標準偏差	r_s	r_p
自然(1,2,3)	2.39	0.48	0.44	0.46
文脈(0,1)	0.60	0.36	0.58	0.59
興味(1,2,3)	2.14	0.50	0.45	0.49
話題(1,2,3)	2.35	0.54	0.52	0.54
首尾(0,1)	0.68	0.32	0.48	0.49
総合(1,2,3,4,5)	3.04	1.02	0.63	0.63

関係数を示す。ここで相関係数は、1 人の評価値と他 4 人の評価値の平均との相関係数を求め、それを 5 人の評価者全員について行いその平均値を用いている。各項目の相関係数は TopicalChat データと比較すると少し低い値であるが、どの指標についても 0.4 を上回る相関がある。TopicalChat データでは比較的少量のデータ（360 文）を対話研究者が評価しているのに対して、本研究ではクラウドソーシングにより 2,250 文の評価を行ったことを考慮すると、ある程度評価者間の評価が一致しているといえる。

4 ChatGPT による自動評価

ChatGPT を用いて対話応答文における先述の 6 つの項目の評価を行う。プロンプトにおいて、各評価項目の定義やレベルごとの基準についてはクラウドソーシングで用いた評価作業手順書の内容（付録 B を参照）を用いる。また、「総合的な品質」以外の 5 項目については、各レベルに 1 文ずつ、評価作業手順書と同様の例文を与える。さらに各評価項目の評価値のみを出力することを促す文と、フォーマットの指定を加えることで、評価値の集計の簡略化を図る。評価値は主観評価と同様に各評価項目で定められた範囲の整数値を、ChatGPT に直接出力させる。

日本語と英語での自動評価の性能差を検証するため、評価対象の文およびプロンプトを英語に翻訳した場合でも評価を行う。

表 5 実験結果 (左: スピアマン相関/右: ピアソン相関)

	自然	文脈	興味	話題	首尾	総合
BERTScore(日)	0.02/0.02	0.04/0.02	0.22/0.25	0.24/0.24	0.02/0.00	0.11/0.09
BERTScore(英)	0.13/0.14	0.15/0.14	0.26/0.28	0.32/0.32	0.10/0.08	0.21/0.20
ChatGPT(日)	0.34/0.35	0.33/0.34	0.46/0.49	0.39/0.41	0.38/0.40	0.38/0.39
ChatGPT(英)	0.41/0.42	0.39/0.41	0.48/0.49	0.36/0.42	0.25/0.27	0.45/0.46
BERTScore×ChatGPT(日)	0.26/0.26	0.16/0.20	0.39/0.47	0.39/0.44	0.24/0.27	0.34/0.34
BERTScore×ChatGPT(英))	0.38/ 0.43	0.31/0.40	0.47/ 0.52	0.44/0.46	0.18/0.24	0.45/0.46

さらに、本論文ではベースラインとして2つの文のトークン間のBERT[20]による埋め込み表現におけるコサイン類似度からスコアを算出する手法であるBERTScore[21]を用いるが、トークン間のコサイン類似度を用いるという性質上、BERTScoreは「話題の関連性」を測るのに比較的適していると考えられる。そこで、ChatGPTに生成させた評価値にBERTScoreによる評価値を加算することで人手との相関の向上を図る。具体的な加算方法は以下の通りである。

$$score = c_i + w(b_i - \bar{b}) \quad (1)$$

ここで c_i はChatGPTによる評価値、 b_i はBERTScoreによるスコア、 \bar{b} は応答文に対するBERTScoreの平均値、 w はBERTScoreの重みを決定するハイパーパラメータを表す。

5 実験

5.1 実験設定

本実験では、主観評価データの評価値とChatGPTによる評価値の相関係数を測ることで、日本語対話応答におけるLLMによる自動評価の有効性を検証する。ChatGPTのモデルにはgpt-3.5-turboを用いる。パラメータはデフォルトの値に従うが、スコアの揺らぎを減らすため、応答の多様さを制御するパラメータであるtemperatureの値を最も低い0に設定する。BERTScoreは4ターンの対話文脈の最後の発話と、それに続く応答文の2文間で算出する。また、式(1)の重み w の値は10とする。応答文やプロンプトの英訳には、翻訳ツールであるDeepL³⁾を用いる。

5.2 実験結果

各手法の評価値と主観評価値との相関係数を表5に示す。ChatGPTによる評価は、表4で示した評価

者間の相関に近く、BERTScoreによる評価を大きく上回っており、日本語の対話評価においてもChatGPTを用いることの実効性が確認できる。

また、評価対象の文やプロンプトを英語に翻訳することで、ほとんどの項目で人手との相関が上がっていることから、対話の自動評価においても日本語と英語の間でChatGPTにおける性能差があることがわかる。BERTScoreにおいても応答文を英語に翻訳することで人手との相関が上がっている理由としては、英語は日本語と比べてトークン数が少なく、英訳することで無駄なトークンが削減されて1トークンあたりの重要度が増すことや、埋め込み表現の算出に用いられるBERTの学習データが影響していると考えられる。また、BERTScoreとChatGPTを組み合わせることで仮説通り「話題の関連性」の相関が上がっており、各項目に適したモデルを組み合わせることで、単一のモデルによる評価よりも人手との相関が高まる可能性があることが考えられる。

6 おわりに

本論文では日本語対話の自動評価におけるChatGPTの有効性を検証するため、対話の重要な品質であると考えられる6つの評価項目に関する主観評価データを収集し、ChatGPTによる評価を行った。その結果、ChatGPTによる評価と主観評価データとの相関が人間同士の相関に近いことを確認し、日本語においてもChatGPTを対話評価に用いる有効性を確認した。一方で、応答文を英語に翻訳することで相関が高くなることから、対話評価におけるChatGPTの日本語と英語での性能差についても確認した。今回はChatGPTを対象に、主観評価で用いた評価作業手順書に合わせたプロンプトによる評価を行ったが、今後は別のLLMを用いた評価や、異なるプロンプトでも評価を行い、主観評価との相関をより向上させたい。

3) <https://www.deepl.com/translator>

参考文献

- [1] Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, Qun Liu. Achieving reliable human assessment of open-domain dialogue systems. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, pp.6416-6437, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [2] Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 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, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.
- [3] Jiannan Xiang, YahuiLiu, Deng Cai, Huayang Li, Defu Lian, Lemao Liu. Assessing dialogue systems with distribution distances. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 2192-2198, Online, August 2021. Association for Computational linguistics.
- [4] Shikib Mehri, and Maxine Eskenazi. Unsupervised Evaluation of Interactive Dialog with DialoGPT. In Proceedings of the SIGdial 2020 Conference, pp. 225-235, 1st virtual meeting, 01-03 July 2020. Association for Computational Linguistics.
- [5] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, Pengfei Liu. Gptscore: Evaluate as you desire. arXiv preprint arXiv:2302.04166, 2023.
- [6] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, Chenguang Zhu. G-Eval:NLG Evaluation using Gpt-4 with Better Human Alignment. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 2511-2522, Singapore, December 2023.
- [7] Yen-Ting Lin, Yun-Nung Chen. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-domain Conversations with Large Language Models. arXiv preprint arXiv:2305.13711, 2023.
- [8] Peiyuan Gong, Jiabin Mao. CoAScore: Chain of Aspects Prompting for NLG Evaluation. arXiv preprint arXiv:2312.10335, 2023.
- [9] 川本稔己, 岡野裕紀, 山崎天, 佐藤敏紀, 船越孝太郎, 奥村学. フォローアップ発話を用いた日本語対話の自動評価. 言語処理学会 第 29 回年次大会 発表論文集 (2023 年 3 月), p.1659-1663.
- [10] 佐藤志貴, 岸波洋介, 杉山弘晃, 赤間怜奈, 徳久良子, 鈴木潤. 多様な話者との自動対話に基づく雑談システムの自動評価. 言語処理学会 第 28 回年次大会 発表論文集(2022 年 3 月), p.1099-1104.
- [11] Ananya B Sai, Akash Kumar Mohankumar, Mitesh M Khapra. A survey of evaluation metrics used for NLG systems. ACM Computing Surveys (CSUR) 55, 2 (2022), 1-39.
- [12] Shikib Mehri, Jinho Choi, Luis Fernando D'sHaro, Jan Deriu, Maxine Eskenazi, Milica Gasic, Kallirroi Georgila, Dilek Hakkani-Tur, Zekang Li, Verena Rieser, Samira Shaikh, David Traum, Yi-Ting Yeh, Zhou Yu, Yizhe Zhang, Chen Zhang. Report from the NSF Future Directions Workshop on Automatic Evaluation of Dialog: Research Directions and Challenges. arXiv:2203.10012, 2022.
- [13] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DIALOGPT: Large-scale generative pre-training for conversational response generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 270-278, Online, July 2020. Association for Computational Linguistics.
- [14] Shikib Mehri, and Maxine Eskenazi. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,
- [15] Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, Thien Huu Nguyen. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language models in Multilingual Learning. Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 13171-13189, December 2023. Association for Computational Linguistics.
- [16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, Vol. 30, 2017.
- [17] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima and Toyomi Meguro. Empirical Analysis of Training Strategies of Transformer-based Japanese Chat-chat Systems. arXiv preprint arXiv:2109.05217, 2021.
- [18] Sarik Ghazarian, Nuan Wen, Aram Galstyan, Nanyun Peng. DEAM: Dialogue Coherence Evaluation using AMR-based Semantic Manipulations. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, pp.771-785, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [19] Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, Jie Zhou. Addressing Inquiries about History: An Efficient and Practical Framework for Evaluating Open-domain Chatbot Consistency. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp.1057-1067, August 2021. Association for Computational Linguistics.
- [20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In 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. 4171-4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [21] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi. BERTScore: Evaluating Text Generation with BERT., in ICLR2020, OpenReview. net, 2020

A 各評価項目のレベルごとの定義

<p>①自然さ</p> <ul style="list-style-type: none"> ・1 (不自然) : 応答文が日本語として不自然 ・2 (やや不自然) : 応答文が日本語として違和感があるものの、完全に不自然ではない ・3 (自然) : 応答文が日本語として自然 <p>②文脈的整合性</p> <ul style="list-style-type: none"> ・0 (噛み合っていない) : 応答文が対話文脈と噛み合っていない ・1 (噛み合っている) : 応答文が対話文脈と噛み合っている <p>③興味深さ</p> <ul style="list-style-type: none"> ・1 (興味深くない) : 応答文に具体性がなく、会話が発展しないような応答 ・2 (やや興味深い) : 応答文にある程度具体性があり、会話が続きそうな応答 ・3 (興味深い) : 応答文に具体性があり、会話を発展させるような応答 <p>④話題の関連性</p> <ul style="list-style-type: none"> ・1 (関連していない) : 応答文がそれまでの話題と全く異なる話題に関する応答 ・2 (どちらでもない) : 応答文が特定の話題を含まず、どちらともいえない ・3 (関連している) : 応答文がそれまでの話題に沿っている <p>⑤首尾一貫性</p> <ul style="list-style-type: none"> ・0 (一貫していない) : 応答文が過去の発話と矛盾している ・1 (一貫している) : 応答文が一貫していて、過去の発話と矛盾がない

図 1:各評価項目のレベルの定義 (総合的な品質を除く)

B ChatGPT のプロンプト (抜粋)

<p>①自然さ (1~3)</p> <p>この項目では応答文が日本語として自然な文であるかどうかを 1~3 の整数値で評価して下さい、</p> <ul style="list-style-type: none"> ・1 (不自然) : 応答文が日本語として不自然である際は、スコアを 1 として下さい。 <p>例: B さん「おなかがすいてきましたね。なにか食べたいものはありますか？」</p> <p>→A さん「私は私はお寿司が食べたいですよ？」</p> <ul style="list-style-type: none"> ・2 (やや不自然) : 日本語として違和感があるものの、完全に不自然ではない際はスコアを 2 として下さい。 <p>例: B さん「おなかがすいてきましたね。なにか食べたいものはありますか？」</p> <p>→A さん「私は食べたいですよお寿司が」</p> <ul style="list-style-type: none"> ・3 (自然) : 応答文が日本語として自然である際は、スコアを 3 として下さい。 <p>例: B さん「おなかがすいてきましたね。なにか食べたいものはありますか？」</p> <p>→A さん「私はお寿司が食べたいです。」</p> <p>②文脈的整合性(0~1)</p> <p>~~~~~中略~~~~~</p> <p>それでは評価していただく対話を示します。</p> <p>A さん「〜」 B さん「〜」 A さん「〜」 B さん「〜」</p> <p>A さんの応答 (評価対象) 「〜」</p> <p>評価結果は必ず以下の形式で出力してください。</p> <p>自然さ: 「」 文脈的整合性: 「」 興味深さ: 「」</p> <p>話題の関連性: 「」 首尾一貫性: 「」 総合的な品質: 「」</p>

図 2 : ChatGPT のプロンプト (一部)

C 評価文の例

<p>例①</p> <p>対話文脈</p> <p>A さん: 「こんにちは！お元気でしたか？」</p> <p>B さん: 「ああ、久しぶりです。こないだ成人しましたよ。2000 年ミレニアム生まれなもんで。」</p> <p>A さん: 「おめでとうございます！2000 年と言うことは平成生まれですよ。私と一緒にだ！」</p> <p>B さん: 「そうですね。令和になっちゃったものなあ。急に老け込んだ感じがしますよ。運動とかかかれていますか？」</p> <p>評価対象の文</p> <p>A さんの応答文①: 「はい、週に数回ジムに行って、ランニングや水泳をしています。」</p> <p>A さんの応答文②: 「はい、運動はしていますよ。最近は何日ジョギングをしています。この前の土日は 30 キロのウルトラマラソンに出場してきました。」</p> <p>A さんの応答文③: 「いや、特にしてないな。あなたは何かしてるんですか？」</p> <p>例②</p> <p>対話文脈</p> <p>A さん: 「こんにちは。最近の調子はいかがですか？」</p> <p>B さん: 「最近体調を崩しまして、ただいま休職中なんですよ。。」</p> <p>A さん: 「それは、大変ですね。どんなお仕事をなさっていたんですか？」</p> <p>B さん: 「北陸に住んでいるので、そこの地方銀行で働いていたのですが、地方はいろいろ厳しくて。。」</p> <p>評価対象の文</p> <p>A さんの応答文①: 「こんにちは。そうだったんですね、お大事に。わたしは派遣社員として働いていますよ。」</p> <p>A さんの応答文②: 「なるほど。では、復職したくなったら、いつでも戻ってきてくださいね。良い一日を!」</p> <p>A さんの応答文③: 「それは大変ですね。地方銀行というところは、このところ経営が厳しいと聞きますよ。」</p>

図 3 : 評価文の例