# 対話データにおける個人の評価傾向の違いの分析 - 個人の評価傾向を反映した対話システム自動評価に向けて -

亀山 京右　駒谷 和範
大阪大学 産業科学研究所
keisuke-kameyama@ei.sanken.osaka-u.ac.jp　komatani@sanken.osaka-u.ac.jp

## 概要

良い対話システムの定義はシステムの用途や評価する個人によって様々である．本研究では，システム設計者など特定の個人の評価傾向を反映した対話システムの自動評価手法を目指す．事前調査として，対話評価における個人差を定量的に分析した．具体的には，対話評価用データセットに付与されている各評価者の評価値から評価傾向の違いを検証した．結果として，評価者間の相関は低く，重視する評価軸の相違が確認された．

## 1　はじめに

対話システムは使用環境や目的に応じてその振る舞いを調整すべきである．調整のためには，個々のシステムに合わせた評価をする必要がある．評価の質を確保するため人手による評価を行うことが多いが，時間的・金銭的コストから何度も行うことは難しい．そのため，FED [1] や ACUTE-EVAL [2] など人手で付与された評価値と相関の高い自動評価手法が多く提案されている．

近年，言語生成タスクにおいて LLM を用いた評価手法が多く用いられており，対話評価においても有効とされている [3]．プロンプトの調整のみで様々な評価軸を評価でき，新しく対話評価データを収集する必要がないため低コストである．LLM を用いた対話評価手法として GPT-Score [4] や G-EVAL [5] などが提案されており，人手評価と比較的相関が高いとされている．これらの手法では，評価対象とするシステムの情報は使用されておらず，システムに合わせた評価とはなっていない．

個々のシステムにはそれぞれの目的や用途があり，それに合わせた評価が必要である．設計者など特定の個人の評価傾向を反映することで個々の対話システムに合わせて評価できる．反映すべき評価傾向を明らかとするためには，人手評価の相違について分析する必要がある．

本研究では，対話データセットに含まれる人手評価の値から評価傾向の相違を明確にする．まず，同一対話データに付与された評価値の相違をスピアマンの順位相関を用いて分析する．個々の付与した評価値間の相関係数より，評価の個人差を定量的に示す．次に，評価者ごとの重視する評価軸の違いを総合評価と各評価軸の重回帰により分析する．回帰係数の比較より，個人の重視する評価軸を明らかにする．

本研究の分析より以下のことが明らかになった．

- 同一の対話データに対して付与される評価値は個人差があり，評価者間の相関は低い．
- 評価傾向の違いは，重視する評価軸の違いに現れる．

この結果より，対話評価における個人差が明確となった．また，開発者など特定の個人の評価傾向を反映するためには，重視する評価軸の傾向が有用であることが示された．

## 2　関連研究

対話システムは一つの入力に対して複数の発話が考えられるため，評価が難しい [6]．そのため，人手による評価が重要とされているが，付与する評価値に個人差があり，質の担保が難しい．Finch ら [7] は，対話システムと会話を行ったユーザの主観評価と第三者による評価では一致率が低いことを実験から明らかにした．これは，評価者が評価の際に利用する情報の量に差があるためとされている．これに対して本研究では，対話に参加していない第三者同士の評価値の差に着目する．評価者の立場が同じ場合の評価値を分析することで，対話評価における個人差がシステムとの対話の有無という立場の違い以外からも生じることを確認する．

対話システムを評価する際に，複数の側面から評価することが重要であるとされている [8]．個人の評価傾向の違いは重視する評価軸に現れる可能性がある．対話の総合的な評価値と他の評価軸との関係性に着目した研究として FineD-Eval [9] がある．この研究では FED 評価用データセット [1] に含まれる対話全体の印象評価である Overall と，Coherence，Diversity など 10 項目の評価軸の相互関係について分析を行っている．これにより，Coherence，Depth of topic，Likability の 3 つの評価軸が Overall 値の予測に有用であるとし，自動評価の手法を提案している．Overall 値の予測の際には， 各評価軸の予測値に対して平均を取り，それを Overall の評価値としている．そのため，各評価軸が Overall に対してどの程度重要なのかについては調査されていない．本研究では，Overall 値を予測する際に重視する評価軸が人によって異なる可能性を考え，重回帰により各評価軸の回帰係数を計算し，Overall 値の予測に対する重みを比較する．これにより，重視する評価軸から個々の評価傾向を明らかとする．

## 3　対象とするデータセット

本研究では，システムと人間の複数交換からなる対話に対する評価を分析対象とする．人間とシステムの対話を想定した評価用の対話データセットとして，FED 評価用データセット [1] や Persona-See データセット [10] などが存在する．個人差の分析にあたって，複数人による第三者評価が付与されている必要があるため，FED 評価用データセットを使用した．

**FED 評価用データセット [1]**　人間とシステムの対話を想定した 3 から 15 交換（平均 6 交換）の対話が収録されたデータセットである．各対話には 5 人の評価者（A, B, C, D, E）が Coherence や Diverse など 10 種の sub-metrics（0-2 の 3 段階）と対話全体の印象評価である Overall（0-4 の 5 段階）に付与した評価値が含まれている．

**評価値の分布**　各評価者の Overall に付与した値の分布を図 1 の左から 5 つに示す．各評価者の付与する評価値の度数に大きな差はなく，3 や 4 といった高めの値を付与していたという傾向が確認できる．評価者 E は特にその傾向が強く，0 や 1 といった低い評価値を付与する度数が小さい．

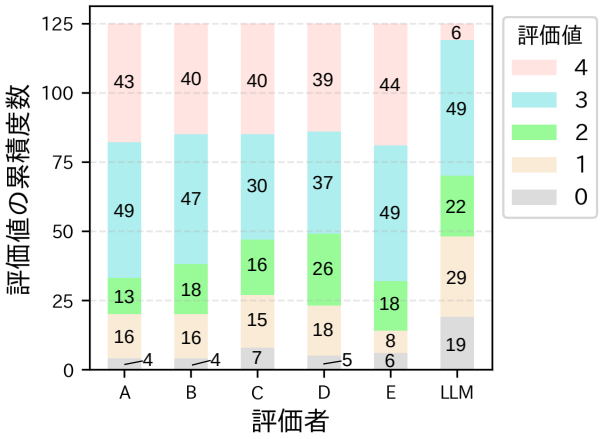


**図 1**　評価値の累積度数棒グラフ．数字は 125 対話中の各評価値の度数

**表 1**　評価者 A の付与した Overall 値とその順位

| 評価値 | 4 | 3 | 2 | 1 | 0 |
|---|---|---|---|---|---|
| 度数 | 43 | 49 | 13 | 16 | 4 |
| 順位 | 22 位 | 68 位 | 99 位 | 113.5 位 | 123.5 位 |

## 4　同一対話データに対する評価値の個人差分析

同一の対話データに付与された評価値の関係より，対話評価における個人差がどの程度存在するかを明確にする．分析としてデータセット内に含まれる 5 人の評価者の Overall 値の間でスピアマンの順位相関係数を算出した．計算にあたって，各評価者が 125 対話に対して付与した評価値を順位尺度に変換した．評価者 A に対する変換の例を表 1 に示す．変換の際には評価値 4 から順番に順位付けを行い，中央値を各評価値の順位とした．例えば，評価値 4 は 1 位から始まり，43 位までになるため，中央値である 22 位となる．評価値 3 は 44 位から始まり，98 位までなので中央値である 68 位となる．また，自動評価手法と個人の評価との間の関係性について明確にするため，LLM を用いた対話自動評価も行い，付与された値をもとに分析を行った．

### 4.1　LLM による評価値の予測

データセットに含まれる評価値に加えて，LLM による対話評価を行った．LLM には gpt-4o-2024-11-20 を用い，*temperature* = 0 とした．プロンプトは LLM-EVAL [11] で提案されているものをベースに作成した．実際に使用したプロンプトを付録 A に示す．

LLM を用いて付与した評価値の分布を図 1 の右

**表 2** Overall 値の各評価者 (A ~ E) と平均値 (Ave.)，LLM の予測の間のスピアマン順位相関係数.

| | A | B | C | D | E | Ave. | LLM |
|---|---|---|---|---|---|---|---|
| A | - | 0.21 | 0.48 | 0.33 | 0.32 | 0.67 | 0.53 |
| B | | - | 0.41 | 0.29 | 0.24 | 0.60 | 0.45 |
| C | | | - | 0.41 | 0.31 | 0.77 | 0.64 |
| D | | | | - | 0.38 | 0.72 | 0.53 |
| E | | | | | - | 0.60 | 0.50 |
| Ave. | | | | | | - | 0.78 |

端に示す．0 や 1 といった低い評価値が多く，4 の予測数が小さいことが確認できる．人手評価の分布と比較して，3 の値の度数が最も大きいことは共通しているが，それ以外の評価値の度数に相違が確認された．

### 4.2 評価者間の総合評価の順位相関

Overall 値について評価者間のスピアマンの順位相関係数を表 2 に示す．A, B, C, D, E は各評価者の付与した値の集合，Ave. は各対話データに対する 5 人の平均値の集合，LLM は gpt-4o-2024-11-20 の予測した値の集合を表す．最も相関の低かった組み合わせである（A, B）のヒートマップを付録 B に，最も相関が高い組み合わせである（Ave., LLM）の散布図を付録 C に示す．

**考察** 対話データセットに付与されている評価者（A~E）同士の相関係数は低く，無相関であることがわかる．LLM を用いて付与した評価値と各評価者間の相関係数は評価者間の相関係数と比べると比較的高く，5 人の平均値（Ave.）との相関係数が最も高い．

分析結果より，対話データに対して人手で付与された Overall 値同士の相関は低く，個人差が存在することが明らかとなった．また，LLM の予測する評価値は個々の評価者より平均値（Ave.）と高い相関があるため，5 人の平均値に近い評価傾向が確認された．

## 5 評価者ごとに重視する評価軸の傾向分析

個人の重視する評価軸の傾向を Overall 値と各評価軸の関係をもとに分析する．Overall の評価の際に，各評価者が異なる評価軸を重視して評価値を付与していると考えられるためである．これにより，評価傾向の違いを明らかにする．評価軸は

**表 3** Overall 値と各評価軸の間のスピアマン相関係数

| | Coherence | Depth | Likability |
|---|---|---|---|
| A | 0.52 | 0.57 | 0.69 |
| B | 0.56 | 0.71 | 0.74 |
| C | 0.68 | 0.72 | 0.75 |
| D | 0.69 | 0.66 | 0.70 |
| E | 0.66 | 0.62 | 0.64 |
| LLM | 0.72 | 0.72 | 0.71 |

FineD-Eval [9] で用いられている Coherence, Depth of topic，Likability の 3 つを使用した．

### 5.1 Overall と各評価軸の関係性

各評価者の Overall 値と Coherence, Depth of topic, Likability のスピアマンの順位相関係数を表 3 に示す．全ての評価軸との間で中程度以上の相関が確認できる．Likability は全ての評価者において高い相関が確認できる．LLM が予測した Overall 値は，同時に予測した全ての評価軸との間に高い相関が見られた．

評価者 A, B は他の評価者と比べて Coherence との間の相関係数が低く，Overall の評価の際に対話システムの一貫性を比較的重視していない傾向が確認できる．同様に Depth に関しても評価者 A, E は他の評価者と比べて相関が低いことが確認できる．このことより，Overall の評価値と Coherence や Depth の評価値の相関について個人差が生じることが確認された．

### 5.2 各評価軸の重視度分析

Overall 値が複数の評価軸の値をもとに付与され，線型結合によって表せるという仮定のもと，重回帰分析によって各評価軸の重みを計算する．得られた重みより，Overall の評価の際に重視する評価軸の相違を明らかにする．計算には以下の式 1 を用いた．

$$S_{p,Overall} = \alpha_{p,c} \cdot S_{p,c} + \alpha_{p,d} \cdot S_{p,d} + \alpha_{p,l} \cdot S_{p,l} \quad (1)$$

ここで $S_{p,Overall} \in \{1,2,3,4,5\}$，$S_{p,c}, S_{p,d}, S_{p,l} \in \{1,2,3\}$ はそれぞれデータセット内に含まれる各評価者の Overall，Coherence，Depth of topic，Likability の値である．3 章で述べたように最低値は 0 であり，分析に悪影響を及ぼす可能性があるため，回帰分析を行うにあたって全ての値に 1 を加算し，最低値を 1 とした．また，$p \in \{A, \ldots, E\}$ であり，各評価者を表す．上記のデータに加えて，4.1 節にて LLM を用いて付与した評価値も用いた．

**表 4**　回帰係数 $\alpha$ の値と MSE，決定係数 $R^2$
**太字**は回帰係数の値が最も大きいものを表す．

| $p$ | $\alpha_{p,c}$ | $\alpha_{p,d}$ | $\alpha_{p,l}$ | MSE | $R^2$ |
|---|---|---|---|---|---|
| A | 0.49 | 0.32 | **0.72** | 0.56 | 0.62 |
| B | 0.37 | 0.46 | **0.72** | 0.49 | 0.68 |
| C | 0.53 | 0.46 | **0.56** | 0.58 | 0.68 |
| D | 0.55 | 0.40 | **0.58** | 0.56 | 0.64 |
| E | **0.89** | 0.23 | 0.42 | 0.48 | 0.71 |
| LLM | **0.46** | 0.23 | 0.33 | 0.23 | 0.73 |

式 (1) によって計算された回帰係数とデータに対する平均二乗誤差（MSE），決定係数 $R^2$ を表 4 に示す．計算された回帰直線は $R^2$ の値が 0.6 以上であり，Overall 値に近い値を予測できるような回帰係数が算出されていることが確認できる．

**考察**　表 4 より，評価者 A, B は $\alpha_{p,l}$ の値が他の評価軸の回帰係数と比べて大きいため，システムに対する好感を表す Likability が Overall 値の評価に重要であることがわかる．評価者 E や LLM の付与した評価値では，$\alpha_{p,c}$ の値が最も大きく，Overall 評価の際にシステム発話の一貫性を表す Coherence の値が重要となっている．また，評価者 C や D においては，すべての評価軸に対する回帰係数の値が比較的高く，特定の評価軸を重視する傾向はないことがわかる．

結果より，評価傾向として，1. システムに対する好感という主観的印象を重視，2. システム発話の一貫性という客観的な印象を重視，3. 各評価軸をバランスよく重視，という 3 種類の評価傾向が確認された．これにより，評価傾向の個人差が重視する評価軸に現れることが確認された．

### 5.3　評価傾向に差が確認された対話例

評価傾向の個人差が確認された対話データの例を表 5 に示す．また，この対話例に関して各評価者が付与した値を表 6 に示す．U で始まるのはユーザ発話，S で始まるのはシステム発話を表す．

評価者 A, B, C は Overall 値として高い値を付与しているのに対して，D や E は 1 と低い値を付与している．また，LLM は，A~E の評価値の平均である 2.4 に近い値である 2 を付与しており，平均的な値になっていることがわかる．

評価者 C と D は各評価軸に対して同様の評価値を付与しているにも関わらず，Overall 値に付与した値が異なる．回帰分析の結果より，Overall 値の評価の際に評価者 C は各評価軸をバランス良く重視するため，Depth や Likability の値が高いことから Overall 値も高い値となっている．一方で，評価者 E は Coherence を最も重視するため，Coherence の値が 1 である部分が影響し，Overall 値が低くなっていると考えられる．

**表 5**　対話データと各評価者の評価値．U で始まるのはユーザ発話，S で始まるのはシステム発話を表す．

| | |
|---|---|
| U1 | Hi! |
| S1 | Hi! How are you today? |
| U2 | What's laser tag? |
| S2 | Like paintball, but with lasers! |
| U3 | lol good description |
| S3 | Do you know what paintball is? |
| U4 | yeah I played it before |
| S4 | Cool! What did you think? |
| U5 | It's somewhat exciting, but very tiring :) |
| S5 | That is very true. What is your favorite color? |
| U6 | I like red |

**表 6**　表 5 の対話例に対して各評価者（A~E）と LLM が付与した評価値

| | Coherence | Depth | Likability | Overall |
|---|---|---|---|---|
| A | 2 | 1 | 1 | 3 |
| B | 2 | 2 | 2 | 4 |
| C | 1 | 2 | 2 | 3 |
| D | 2 | 1 | 1 | 1 |
| E | 1 | 2 | 2 | 1 |
| LLM | 1 | 1 | 1 | 2 |

## 6　おわりに

本研究では，対話システムの設計者など特定の個人の評価を反映した対話評価手法の確立を目標とし，対話評価用データセットにみられる個人の評価傾向の違いについて分析した．評価者同士の相関は低く，対話データに対して付与する評価値の差が明らかとなった．また，LLM の予測は 5 人の評価の平均値と相関が高いが，付与される評価値の分布に違いが確認された．

本稿では，他のデータセットに対する分析は行っていないため，対話評価全体にこの分析が有効か検証できていない．そのため，本分析と同様に重視する評価軸の相違が個人の評価傾向を反映するのか引き続き調査し，特定の個人の評価を反映した対話評価の枠組みを検討していく．

## 謝辞

本研究の一部は科研費 (JP22H00536) の支援を受けた.

## 参考文献

[1] Shikib Mehri and Maxine Eskenazi. Unsupervised Evaluation of Interactive Dialog with DialoGPT. In **Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 225–235, July 2020.

[2] Margaret Li, Jason Weston, and Stephen Roller. ACUTE-EVAL: Improved Dialogue Evaluation with Optimized Questions and Multi-turn Comparisons. **arXiv CoRR**, Vol. abs/1909.03087, , 2019.

[3] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, NIPS '23, 2024.

[4] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as You Desire. In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 6556–6576, Mexico City, Mexico, June 2024.

[5] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and 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, December 2023.

[6] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 654–664, Vancouver, Canada, July 2017.

[7] Sarah E. Finch and Jinho D. Choi. Towards Unified Dialogue System Evaluation: A Comprehensive Analysis of Current Evaluation Protocols. In **Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 236–245, July 2020.

[8] Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. A survey of evaluation metrics used for nlg systems. **ACM Comput. Surv.**, Vol. 55, No. 2, January 2022.

[9] Chen Zhang, Luis Fernando D'Haro, Qiquan Zhang, Thomas Friedrichs, and Haizhou Li. FineD-Eval: Fine-grained Automatic Dialogue-Level Evaluation. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 3336–3355, December 2022.

[10] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. What makes a good conversation? How controllable attributes affect human judgments. 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. 1702–1723, Minneapolis, Minnesota, June 2019.

[11] Yen-Ting Lin and Yun-Nung Chen. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models. In **Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)**, pp. 47–58, July 2023.

## A　用いたプロンプト

LLM-EVAL [11] で提案されているプロンプトをもとに作成したプロンプトを以下に示す．各評価軸の説明は FED [1] にて人手評価の際に用いられた質問文と同様のものを使用した．

# Instruction

Evaluate the system with following multi-turn dialogue.

When evaluating system, first evaluate 'Coherence', 'Depth of topic', and 'Likability', and then evaluate 'Overall quality'.

# sub-metrics dimension

Coherence : Throughout the dialog, is the system coherent and maintain a good conversation flow? [0, 1, 2]

Depth of topic : Does the system discuss topics in depth? [0, 1, 2]

Likability : Does the system display a likable personality? [0, 1, 2]

# Score

0 : Bad.

1 : Neutral.

2 : Good.

# metrics

Overall quality : Overall impression of the dialogue? [0, 1, 2, 3, 4]

# Score

0 : Terrible.

1 : Bad.

2 : Neutral.

3 : Good.

4 : Excellent.

# Input

The dialogue is implemented by User and System.

—

システムとユーザの対話

—

## B　評価者 A と B の Overall 値のヒートマップ

4.2 節の分析において最も相関が低かった組み合わせである評価者 A と B の評価値のヒートマップを図 2 に示す．（A,B）の評価値が（1, 3）や（4, 1）といった評価値の組み合わせも存在し，無相関に近いことが読み取れる．

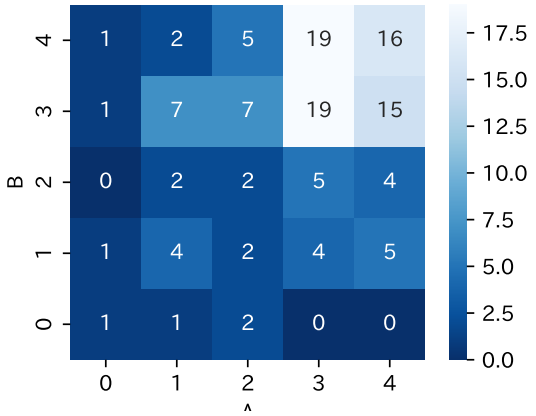


**図 2**　評価者 A と B の評価値のヒートマップ

## C　5 人の評価の平均（Ave.）と LLM の Overall 値の散布図

4.2 節の分析において最も相関が高かった組み合わせである 5 人の評価の平均値（Ave.）と LLM の予測の散布図を図 3 に示す．平均値（Ave.）は必ずしも整数値とはならないため，図 2 と異なり散布図によって表す．分析結果と同様に正の相関が確認できる．

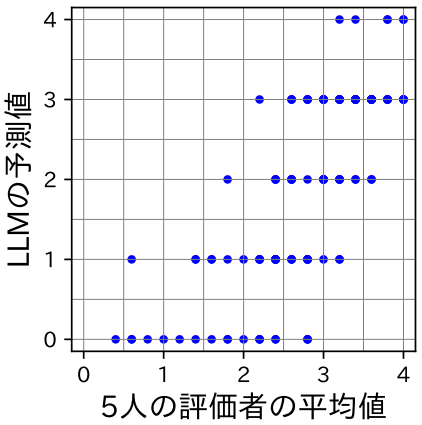


**図 3**　5 人の評価値の平均（Ave.）と LLM の予測値の散布図