人間同士の対話を対象とする確信度を考慮した LLM 性格推定

望月 敦史　高橋 空大　土田 陸斗　宇津呂 武仁
筑波大学大学院 システム情報工学研究群

概要

本論文では，LLM を用いた性格推定における課題に着目し，一部の対話データには正確な推定に必要な十分な情報が含まれていないことを示す．自己申告による性格特性と第三者によって知覚された性格特性との間に乖離が存在することが報告されていることを踏まえ，本論文では，推定結果の信頼性を評価する指標として確信度推定を導入する．実験結果から，人手による確信度が高いほど，LLM による性格特性推定の精度が向上することが示された．さらに，CNN モデルによって学習された確信度指標を用いて LLM の推定結果を補正することで，外向性，神経症傾向，および全性格特性を通した全体性能において性能向上が観測された．

1　はじめに

個人の性格は，対象の行動や言動，本人でも気付かないような長期的なキャリア，対人関係の考え方に大きな影響を及ぼすため，パーソナリティに関する研究は分野を問わず長く行われてきている．一方近年の LLM の発展に伴い，LLM を用いて性格の推定を行う研究も行われており，CoT を用いるものや，複数 LLM との対話を用いたものなど多くの研究が存在している．一方 LLM による性格推定に用いられるコーパスや，心理学の分野における性格算出方法である質問表に関する研究において，本人により申告された性格と，他者から見た性格に隔たりがあることが報告されている．そのため本論文では性格推定において，対象となるデータ内に存在する情報のみでは，推定することが困難な対象が存在すると考え，アンサンブル学習に着想を得た，推定における確信度という指標を導入する．これにより性格推定に用いるのに十分な対象と，不十分な対象を分けることができ，結果として性格推定を実用的に用いることができるようになることが期待される．

本論文における貢献をまとめると以下の通りとなる

- 2 名の参加者による対話を参照し，複数の人手アノテータによって性格特性推定を行うことで，性格特性推定の難易度が対話ごとに異なることを示し，推定が容易な対象および困難な対象が存在することを明らかにした．
- 以下の 2 つの方法により，推定確信度を考慮した性格特性推定の概念を導入した．
  - (i) 人手によって性格特性が正しく推定された対話，および (ii) 人手による性格特性推定の一致が得られた対話に対して、LLM による性格特性推定を行った．(i) では十分に高い精度 (約 0.80 以上) を達成し，(ii) においても、ランダムベースラインより (有意ではないものの) 高い精度を達成した．
  - さらに，(iii) LLM 同士の性格特性推定の一致が得られた対話に対しても LLM による性格特性推定を行い，Big Five の 5 特性のうち 2 特性において，単一モデルの中で最も性能の高いベースラインと比較して、統計的に有意な性能向上を達成した．

2　関連研究

個人の性格をモデルを用いて表現しようとする試みが長年にわたり行われてきた．代表的なモデルとしては，Big Five [7, 13]，MBTI [14]，および Dark Triad [15] が挙げられが，本論文では，最も一般的に使用されている性格モデルである Big Five を用いる．Big Five における性格特性は，通常質問紙への回答によって測定される．代表的な質問紙としては，BFI [10]，BFI-2 [16]，IPIP-NEO-PI [6]，および TIPI [8] がある．これらのうち本論文では，質問項目数が最も少ない TIPI と，広く利用されている BFI-2 を用いる．なお，使用に際しては各原著者によって提供されている日本語版を使用する．自然言語処理の分野においても，個人の性格に関する研究は長年

にわたり行われており，様々な手法で個人の性格を推定し，その精度や傾向が調べられている．LLM を用いない研究としては，TF-IDF や特徴選択に基づく手法 [2] や，グラフニューラルネットワークを用いる手法 [5] が挙げられ，LLM を用いたものとしては，Facebook 投稿を推定対象とする研究 [18]，ビデオ面接を対象とする研究 [24]，対話データを用いる研究 [3]，および推定時に CoT を用いる手法 [22] などがある．また，ユーザと LLM との相互作用を通じて性格を推定する研究も行われており，対話ターン数に着目した研究 [23] や，ユーザ応答の特徴と LLM との相互作用を比較する研究 [4] がある．さらに特定の性格特性を再現するよう LLM にプロンプトを与える研究 [9] や，LLM 自体の性格を推定する研究 [12] も存在している．一方複数のモデルを用いることで性能向上を図る試みもなされている．さまざまなタスクに適用可能なフレームワークである Reconcile [17] や，性格推定において複数の SVM カーネルを用いる研究 [11]，SVM，BERT，および XGBoost を組み合わせる研究 [1] などが挙げられる．これらに対し本論文では，対象コーパスには，推定に十分な情報を含む対象と，そうでない対象の双方が含まれていると仮定し，これらを区別する手段として推定確信度を導入し，その有効性を評価する．

3 Big Five [7, 13]

Big Five は人の性格構造を 5 つの性格特性から構築する．以下が性格特性とその特徴である．

- Openness to experience (O)
  (経験への) 開放性を表す要素であり本論文では O と略す．想像力豊かであり，新しい経験や冒険を好む性格の場合，この値が高くなる．
- Conscientiousness (C)
  勤勉性を表す要素であり本論文では C と略す．注意深く意欲に富んでおり，規律を重視する性格の場合，この値が高くなる．
- Extraversion (E)
  外向性を表す要素であり本論文では E と略す．陽気で自己主張が激しく，人と関わることを好む性格の場合，この値が高くなる．
- Agreeableness (A)
  協調性を表す要素であり本論文では A と略す．利他的で親身であり，人を信じ謙虚な性格の場合，この値が高くなる．
- Neuroticism (N)
  神経症傾向を表す要素であり本論文では N と略す．心配性で悲観的であり，ストレスに弱く憤りやすい性格の場合，この値が高くなる．

4 対話コーパス [21, 20]

今回使用する対話コーパスは，日本語の RPC [21, 20] となる．このコーパスには 233 人の話者それぞれのペルソナと性格特性が付与された，13,583 対話が存在している．本論文においては，人手による性格推定において 100 人による 50 対話，LLM による性格推定において，全ての話者が登場するようにしつつ，話者ごとの登場回数ができるだけ均等になるように選択した 233 人による 1,000 対話を選択した．このコーパスにおける性格は，BFS [19] の質問表を用いて測定された，発話者個人の Big Five が付与されている．1 対話あたりの平均発話数は 30.08 対話，発話あたりの平均文字数は 13.16 文字となっている．

5 性格推定

5.1 人手による推定

4 節にて説明した，50 対話 100 名分の対象に対して，本論文では，3 人のアノテーターによる性格推定を行う．人手による性格推定において用いた質問表は，TIPI [8] となる．TIPI は少ない質問数で，Big Five を測定する手法であり，10 個の質問 $qx_i(x \in \{O, C, E, A, N\}, i = 1, 2)$ に対する 1~7 の 7 段階の離散値の回答 ax_i で Big Five の測定を行うことができる．その後，各対象ごとの各 Big Five に対して，2 値分類を行う．この際の分類基準は，各ラベルの個数が同数になる様に基準値を設定した．

5.2 LLM による推定

本論文で用いる LLM のモデルは gpt-4o-2024-08-06，claude-sonnet-4-5，gemini-2.5-flash，grok-4-1-fast-reasoning の 4 つであり，API を経由して用いた．再現性を確保するために，全てのモデルにおいて tempature を 0 とし，集計のために LLM の出力は json 形式となるように．各種設定，プロンプトを利用する．LLM による性格推定において用いた質問表は，BFI-2 [16] となる．BFI-2 は Big Five の 5 つの性格特性それぞれに対して，12 個の質問に対する 1~5 の 5 段階の離散値の回答を用いることで，Big Five の測定を行うことができる．結果として算出された各性格特性に対して，人手の際と同じように 2

表 1　LLM 性格推定評価結果 (人手による確信度ごと．**下線・太字**は最良の結果を示す)

(a) 正答数を確信度として用いた場合

人手確信度 (正答数)	O	C	E	A	N	全性格特性
0	0.455	0.235	0.284	0.154	0.09	0.232
1	0.323	0.346	0.384	0.485	0.314	0.371
2	0.757	0.548	0.576	0.235	0.636	0.589
3	**0.810**	**0.769**	**0.833**	**0.792**	**0.810**	**0.802**
合計	0.600	0.500	0.540	0.430	0.500	0.514

(b) 一致数を確信度として用いた場合

人手確信度 (多数派-少数派)	O	C	E	A	N	全性格特性
2-1	0.559	0.456	0.491	0.400	0.471	0.480
3-0	**0.688**	**0.558**	**0.610**	**0.460**	**0.563**	**0.566**
合計	0.60	0.50	0.54	0.43	0.50	0.514

表 2　人手による確信度推定 (正答数) 結果の分布

人手確信度 (正答数)	O	C	E	A	N	全性格特性
0	11	17	17	26	11	82
1	31	26	26	33	35	151
2	37	31	33	17	33	151
3	21	26	24	24	21	116
合計	100	100	100	100	100	500

値分類を行う．この際 LLM に対しては，対話履歴，タスクの説明，出力形式の指定を入力した．なお，推定の結果の表を表 4 (付録の節 A) に示すように，O，C，E の推定においては Claude が，A，N の推定においては GPT が最も良い精度となり，全性格特性に対しては Claude が最も良い精度となった．

6　確信度推定

本論文で用いたコーパス [21] や質問表に関する研究 [16] において，個人の申告する性格と，他者が読み取ることのできる性格の間に隔たりがあることが報告されている．コーパスに関する研究 [21] では，ピアソンの相関係数が最も良い性格特性においては 0.17，最も悪い場合においては-0.09，質問表に関する研究 [16] では，最も良い性格特性において相関係数が 0.49，最も悪い場合は 0.02 となっている．そのため推定タスクにおいて，全てのデータを対象にして推定を行った際に，推定対象に含まれる情報のみでは，推定が困難な対象が存在すると考えられる．そのためその対象が推定するにあたって十分な情報を有しているかどうかを判断するために，確信度という考えを導入する．これにより推定対象に対する結果が，確信に足るものかどうかを判定することができ，結果として確信度が高いと判断された対象に対する結果のみを用いることで，より精度が高く，実用的な推定結果が得られると考えられる．

6.1　人手による確信度推定

6.1.1　手法

本論文では，人手による確信度推定として 2 つの手法を提案する．第 1 の手法は，対象となる人物および性格特性ごとに，アノテータのうち正しく性格特性を推定できた人数に基づいて確信度を算出する方法であり，これを正答数と呼ぶ．第 2 の手法は，対象となる人物および性格特性ごとに，アノテータ同士の推定結果が一致した人数に基づいて確信度を算出する方法であり，これを一致数と呼ぶ．

6.1.2　結果

まず，表 2 は，人手確信度推定 (正答数) の結果を示している．表に示すように，全てのアノテータが正しい推定に失敗したケースは 82 件存在し，アノテータの過半数が誤った推定を行ったケースは 233 件存在する．その結果，人手による性格特性推定の全体精度は 0.534 となった．次に，人手確信度 (正答数) を適用した場合が表 1 (a) である．LLM による性格特性推定の全性格特性にわたる精度は 0.802 に達する．さらに各性格特性ごとに見ても，人手確信度推定 (正答数) が最も高い対象に対して，LLM は高い推定精度を達成している．次に，付録の A 節の

表 3　LLM による確信度を用いた場合の LLM 性格推定評価結果 (∗ は best single model との間で有意差有 ($p < 0.05$))

—		O	C	E	A	N	全性格特性
best single model		0.533	0.521	0.528	0.521	0.497	0.520
LLM 確信度 (一致数)	3-1	0.539	0.505	0.514	0.521	0.492	0.517
	4-0	0.506	0.525	0.522	0.514	0.473	0.509
LLM 確信度 (調整後一致数)	3-1	0.533	0.534	0.527	0.523	0.502	0.524
	4-0	0.517	0.555	0.576*	0.552	0.540*	0.549*

表 5 は人手確信度推定 (一致数) の結果を，表 2 (b) は人手確信度推定 (一致数) を用いた場合の結果を示しており，LLM による性格推定の全性格特性にわたる精度は 0.566 となる．同様に各性格特性ごとに見ても，人手確信度推定 (一致数) が高い対象ほど，LLM はより高い推定精度を示す傾向が確認される．

6.2　LLM による確信度推定

6.2.1　手法

本論文で用いる人手による確信度の推定方法は 2 つあり，1 つ目は先ほどの人手の例の二つ目と同じく，対象となる人物，性格特性の LLM による推定結果において，いくつの LLM の回答が一致したかどうかによる算出方法 (一致数) となる．2 つ目は，教師あり学習を用いた手法となる．本論文にて使用する質問表である BFI-2 は，Big Five の各特性 {O, C, E, A, N} それぞれに対応する 12 個の質問の計 60 個の質問から構成されている．結果として，1 つの LLM は 1 つの特性に対して，12 個の回答を返す形になり，対象となる話者のそれぞれの性格特性に対しては，4 モデルが 12 個の回答を返すため，48 個の回答が存在する．各モデルの質問に対する回答を合わせたものを説明変数，各モデルの回答が，正答しているかどうかを目的変数とした上で，CNN を使って確信度算出モデルを作成する．その結果，各モデルの回答が，正答しているか，誤答しているかの結果が得られる．この結果と各モデルの元々の回答を用い，各モデルの回答を，正答していると判断された場合そのままに，誤答していると判断された場合，2 値分類の他方の結果に調整する．

6.2.2　結果

LLM による確信度推定の対象数の表が表 6 (付録の節 A) であり，LLM 確信度を用いた LLM 性格推定の結果が表 3 である．精度の表においては，一致数 2，つまりモデルが 2 つずつ別の回答をしている場合の結果を求めることはできないため，省略している．1 つ目の確信度算出方法である，一致数を LLM 確信度として用いた場合の結果が，表の上部となる．LLM の推定性能に関しては，単一の最良モデルに比べて，LLM 確信度 (一致数) による有意差は見られない．次に 2 つ目の確信度算出方法である，LLM 性格推定の結果の正誤を，機械学習により求めた後に，LLM の回答を調整し，調整した結果の一致数を確信度とした，際の結果が表の下部となる．結果としては外向性，協調性，そして全性格特性の合計に関しては，単一の最良モデルと比べ，LLM 確信度 (調整後一致数) が 4 の場合において，カイ二乗検定を行った結果，有意差が見られた．

7　おわりに

本論文では性格推定タスクにおいて，対象となるデータ内に推定困難な対象が存在すると考え，人手 2 種，LLM2 種の推定確信度の導入を行った．結果として，人手による確信度推定においては，どちらにおいても確信度が高い対象において，LLM の推定精度も向上した．一方で複数の LLM による確信度推定においては，複数 LLM の回答の一致数を推定確信度とした場合においては，最も良い単一のモデルと比べて有意差は見られなかった．しかし，CNN モデルを用いて LLM の出力の正誤を推定し，その結果に基づいて推定結果を調整した上で，複数 LLM の調整後の回答の一致数を推定確信度とした場合においては，最も良い単一のモデルと比べ，外向性，協調性，そして全ての性格特性の合計において，有意差が見られた．これは各 LLM が特定の領域の得手不得手があり，複数 LLM の質問表に対する回答を用いて機械学習を行うことで，それぞれの欠点を補い合った結果によるものだと考えられる．今後はプロンプトの調整，性能の大きく異なる LLM の追加，及び各 LLM に複数プロンプトを与えることで，より実用的で頑健な推定性能の実現を目指す．

謝辞

本論文は，一部，科研費 25K03416 の支援を受けたものである．

参考文献

[1] Mohammad Hossein Amirhosseini and Hassan Kazemian. Machine learning approach to personality type prediction based on the myers–briggs type indicator®. **Multimodal Technologies and Interaction**, Vol. 4, No. 1, p. 9, 2020.

[2] Alessandro Bruno and Gurmeet Singh. Personality traits prediction from text via machine learning. In **Proc. IEEE AIC 2022**, pp. 588–594. IEEE.

[3] Erik Derner, Dalibor Kučera, Nuria Oliver, and Jan Zahálka. Can ChatGPT read who you are? **Computers in Human Behavior: Artificial Humans**, Vol. 2, No. 2, p. 100088, 2024.

[4] Jinyan Fan, Tianjun Sun, Jiayi Liu, Teng Zhao, Bo Zhang, Zheng Chen, Melissa Glorioso, and Elissa Hack. How well can an ai chatbot infer personality? examining psychometric properties of machine-inferred personality scores. **Journal of Applied Psychology**, Vol. 108, No. 8, p. 1277, 2023.

[5] Yahui Fu, Haiyue Song, Tianyu Zhao, and Tatsuya Kawahara. Enhancing personality recognition in dialogue by data augmentation and heterogeneous conversational graph networks. In **Proc. 14th IWSDS**, pp. 1–17, 2024.

[6] Lewis R. Goldberg. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. **Personality psychology in Europe**, Vol. 7, No. 1, pp. 7–28, 1999.

[7] Lewis R. Goldberg. An alternative 'description of personality': The big-five factor structure. In **Personality and personality disorders**, pp. 34–47. Routledge, 2013.

[8] Samuel D. Gosling, Peter J. Rentfrow, and William B. Swann Jr. A very brief measure of the big-five personality domains. **Journal of Research in personality**, Vol. 37, No. 6, pp. 504–528, 2003.

[9] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. PersonaLLM: Investigating the ability of large language models to express personality traits. In **NAACL 2024**, pp. 3605–3627, June.

[10] Oliver P. John, Eileen M. Donahue, and Robert L. Kentle. Big five inventory. **Journal of personality and social psychology**, 1991.

[11] Akshi Kumar, Rohit Beniwal, and Dipika Jain. Personality detection using kernel-based ensemble model for leveraging social psychology in online networks. **ACM Transactions on Asian and Low-Resource Language Information Processing**, Vol. 22, No. 5, pp. 1–20, 2023.

[12] Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beongwoo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. In **NAACL 2025**, pp. 8397–8437, April.

[13] Robert R. McCrae and Oliver P. John. An introduction to the five-factor model and its applications. **Journal of personality**, Vol. 60, No. 2, pp. 175–215, 1992.

[14] Isabel Briggs Myers and Peter B. Myers. **The myers-briggs type indicator**, Vol. 34. Consulting Psychologists Press Palo Alto, CA, 1962.

[15] Delroy L. Paulhus and Kevin M. Williams. The dark triad of personality: Narcissism, machiavellianism, and psychopathy. **Journal of research in personality**, Vol. 36, No. 6, pp. 556–563, 2002.

[16] Christopher J. Soto and Oliver P. John. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. **Journal of personality and social psychology**, Vol. 113, No. 1, p. 117, 2017.

[17] Lei Sun, Jinming Zhao, and Qin Jin. Revealing personality traits: A new benchmark dataset for explainable personality recognition on dialogues. In **EMNLP 2024**, pp. 19988–20002, November.

[18] Adithya V. Ganesan, Yash Kumar Lal, August Nilsson, and H. Andrew Schwartz. Systematic evaluation of GPT-3 for zero-shot personality estimation. In **Proc. 13th Workshop on WASSA**, pp. 390–400, July 2023.

[19] 和田さゆり. 性格特性用語を用いた big five 尺度の作成. 心理学研究, Vol. 67, No. 1, pp. 61–67, 1996.

[20] 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 回年次大会論文集, pp. 2738–2743, 2024.

[21] Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities. In **Proc. 37th PACLIC**, pp. 852–861, 2023.

[22] Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiaxiang Wu. PsyCoT: Psychological questionnaire as powerful chain-of-thought for personality detection. In **EMNLP 2023**, pp. 3305–3320, December.

[23] Baiqiao Zhang, Zhifeng Liao, Xiangxian Li, Chao Zhou, Juan Liu, Xiaojuan Ma, and Yulong Bian. Rethinking personality assessment from human-agent dialogues: Fewer rounds may be better than more. In **EMNLP 2025**, pp. 5357–5380, November.

[24] Tianyi Zhang, Antonis Koutsoumpis, Janneke K. Oostrom, Djurre Holtrop, Sina Ghassemi, and Reinout E. de Vries. Can large language models assess personality from asynchronous video interviews? a comprehensive evaluation of validity, reliability, fairness, and rating patterns. **IEEE Transactions on Affective Computing**, Vol. 15, No. 3, pp. 1769–1785, 2024.

A　付録

四つの LLM による性格推定の結果を表 4 に示す．O，C，E の推定においては Claude が，A，N の推定においては GPT が最も良い精度となり，全性格特性に対しては Claude が最も良い精度となった．人手確信度推定 (一致数) の結果を表 5 に示す．このうち，表 5(a) には対象数を，表 5(b) には対象に対する人手の推定結果を示す．全体として人手一致数が多い対象ほど，人手による推定性能が高くなる．LLM による 2 種類の確信度推定の結果の対象数を表 6 に示す．調整後には一致数 4 の対象数が減少する傾向が見られる．

表 4　LLM 性格推定評価結果 (確信度推定結果を用いない場合．**下線・太字**は各性格特性における最良の結果)

Model	O	C	E	A	N	全性格特性
GPT	0.513	0.502	0.500	**0.521**	**0.497**	0.506
Gemini	0.479	0.500	0.504	0.497	0.488	0.494
Claude	**0.533**	**0.521**	**0.528**	0.508	0.496	**0.517**
Grok	0.522	0.512	0.515	0.508	0.490	0.510
合計	0.512	0.509	0.512	0.509	0.493	0.507

表 5　人手による確信度推定 (一致数) を用いた場合の LLM 性格推定評価結果

(a) 人手による確信度推定 (一致数) 結果の分布

人手確信度 (多数派-少数派)	O	C	E	A	N	全性格特性
2-1	68	57	59	50	68	302
3-0	32	43	41	50	32	198
合計	100	100	100	100	100	500

(b) LLM 性格推定評価結果 (**下線・太字**は各性格特性における最良の結果)

人手確信度 (多数派-少数派)	O	C	E	A	N	全性格特性
2-1	0.544	0.544	0.559	0.340	0.485	0.500
3-0	**0.656**	**0.605**	**0.585**	**0.480**	**0.653**	**0.586**
合計	0.580	0.570	0.570	0.410	0.540	0.534

表 6　LLM による確信度推定 (一致数) 結果の分布

(a) CNN モデルによる調整前

LLM 確信度 (多数派-少数派)	O	C	E	A	N	全性格特性
2-2	776	773	652	1,014	855	4,070
3-1	2,047	2,028	1,872	2,003	1,964	9,914
4-0	1,177	1,199	1,476	983	1,181	6,016
合計	4,000	4,000	4,000	4,000	4,000	20,000

(b) CNN モデルによる調整後

LLM 確信度 (多数派-少数派)	O	C	E	A	N	全性格特性
2-2	1,269	1,205	1,226	1,349	1,302	6,351
3-1	2,004	1,950	1,986	1,984	2,040	9,964
4-0	727	845	788	667	658	3,685
合計	4,000	4,000	4,000	4,000	4,000	20,000