英日翻訳の話者個性保持に関する調査

傅 星儿[1]　永田 昌明[2]　Chenhui Chu[1]
[1] 京都大学大学院 情報学研究科　[2]NTT コミュニケーション科学基礎研究所
xinger@nlp.ist.i.kyoto-u.ac.jp　masaaki.nagata@ntt.com
chu@i.kyoto-u.ac.jp

概要

大規模言語モデル（LLM）の発展により機械翻訳の流暢性は飛躍的に向上したが、話者の性格特性が翻訳を通じてどの程度保持されるかは十分に検証されていない。先行研究では翻訳が性格予測精度に悪影響を与えることが示されているが、これらは主に欧州言語間を対象としており、役割語や敬語体系を持つ英日翻訳における検証は不十分である。実験の結果、mmBERT では英日間の比較的に低い予測誤差を達成し、これは予測モデル自体の誤差と同程度であり、翻訳による性格情報の大幅な損失がないことが確認された。また、Qwen3 では英日間で Pearson 相関 0.56 を達成し、特に外向性や開放性において高い一貫性を示した。さらに、日本語翻訳からの予測が英語原文よりも Ground Truth に近いという結果が得られ、役割語や語尾表現が性格シグナルを増幅している可能性が示唆された。

1　はじめに

近年、大規模言語モデル（LLM）の発展により、機械翻訳（MT）の性能は飛躍的に向上している。BLEU[1] や COMET[2] といった自動評価指標において高い翻訳精度が達成され、継続事前学習（CPT）や教師ありファインチューニング（SFT）、DPO や GRPO などの強化学習 [3, 4] の適用により、流暢性や意味的忠実度はさらに向上を続けている。

しかし、これらの指標は翻訳における意味的な忠実性を測定するものであり、原文に含まれる話者の性格特性や固有の言語スタイルの保持については十分に評価し得ない。映画字幕、アニメ、漫画、ゲームのローカライズといった、登場人物の個性表現が不可欠な翻訳タスクにおいて、こうした限界は実用上の大きな懸念点となっている。

機械翻訳における著者属性の保持に関する研究は限られている。Mirkin ら [5] は、翻訳（原言語から目的言語、およびその逆）が性格特性や社会人口学的特徴の予測精度に対して有害な影響を与えることを実証的に示した 。彼らの実験では、Big Five（外向性、情緒安定性、協調性、誠実性、開放性）の全次元が調査対象となっており 、機械翻訳モデルはユーザーに対して一般的であるため、原文に含まれる個別の言語的シグナルが保持されない可能性を指摘している 。Rabinovich ら [6] はこの知見を拡張し、特に英語等の記述言語では形態論的に明示されにくい性別シグナルの減衰を定量的に分析した。しかし、これらの研究は統計的機械翻訳時代のものであり、LLM ベースの翻訳における著者属性保持は未検証である。また、これらの先行研究は主に欧州言語間を対象としており、英日翻訳への適用は検討されていない。言語使用は性格特性や社会人口学的特徴に影響されるが 、英語で顕著な性格指標が他言語へ自然に翻訳されるとは限らない 。特に日本語には、話者の属性を示す「役割語」[7] や、英語には存在しな

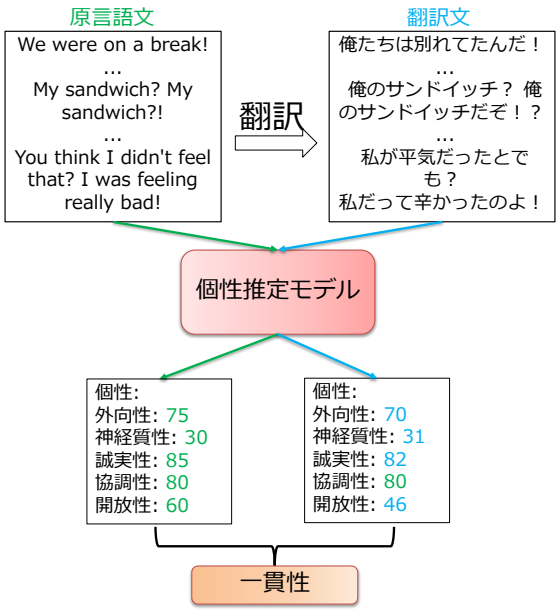

図 1　本研究の評価フレームワーク

い対人関係を規定する「敬語体系」といった独自の表出メカニズムが存在する。このような言語間の情報の非対称性は、翻訳プロセスにおいて性格シグナルの変容や消失を招く潜在的な要因となり得る。

本研究では、アメリカのドラマである「Friends」を人手で英語から日本語へ翻訳したデータ [8] を用いて、性格予測の一貫性の評価を試みる。図 1 に本研究で提案する評価フレームワークを示す。具体的には、(1) エンコーダベースの多言語モデル（mmBERT [9]）をファインチューニングした性格予測モデル、および (2) LLM（Qwen3 [10]）によるプロンプトベースの手法を用いて、翻訳前後のテキストから Big Five 性格特性を予測し、その一貫性を測定する。

本研究の貢献は以下の通りである。

- 人手による英日翻訳における性格予測の一貫性に関する、LLM を用いた初の体系的評価を行う。
- エンコーダベースモデル（ファインチューニング）と LLM（ゼロショット）の性格予測の有効性を比較検証する。

2 関連研究

2.1 多言語エンコーダモデルと性格予測

多言語事前学習モデルは、言語横断的なタスクにおいて広く利用されている。Marone ら [9] は mmBERT を提案し、ModernBERT アーキテクチャを基盤に 1800 以上の言語・3T トークンで学習を行った。mmBERT は段階的な言語追加スケジュールと逆マスク率スケジュールを導入し、XLM-R[11] を大幅に上回る性能を達成している。この言語横断的な転移学習能力は、ドメインや言語ペアを超えた性格表現の普遍的な抽出を可能にするため、本研究における英日間の性格予測一貫性評価において重要な基盤となる。

これらのエンコーダモデルに回帰ヘッドを付加することで、テキストからの性格予測が可能となる。Fu ら [12] は、PANDORA データセット [13] で Luke モデル [14] をファインチューニングし、Big Five および MBTI 特性を予測する手法 StyEmp を提案した。しかし、予測精度は限定的であり、Big Five の平均 Pearson 相関は 0.133、MBTI では 0.086 にとどまり、テキストからの性格予測の困難さを示している。このように、テキストからの性格予測は依然として困難なタスクであり、単一のモデルに依存した評価では信頼性に限界がある。本研究では、エンコーダベースモデルと LLM の双方を用いた比較評価により、この困難なタスクにおける評価の頑健性を検証する。

2.2 LLM の発展と機械翻訳への応用

デコーダベースの LLM は急速に発展している。Wei ら [15] が提案した Chain-of-Thought（CoT）推論や、OpenAI o1、DeepSeek-R1 等の推論モデル（Reasoning Model）は、回答前に長い思考過程を生成することで複雑なタスクの性能を向上させている。これらの高い推論能力を持つモデルは、文脈から微細な性格シグナルを抽出する能力に長けていることが期待される。

一方、LLM は特定のペルソナを演じるロールプレイ能力を持つことが知られており、プロンプトによって Big Five に基づく性格特性を制御した応答生成が研究されている [16]。しかし、翻訳タスクにおいて原文の性格特性が訳文にどの程度保持されるかは明らかでない。本研究では、原文と訳文に性格予測手法を適用し、英日翻訳における性格の一貫性を定量化する。

3 提案手法

3.1 タスク定義

本研究では、テキストから話者の性格特性を定量的に推定する Big Five 性格特性予測タスクを対象とする。Big Five モデルは、心理学において最も堅牢な性格モデルの一つであり、神経症傾向（N）、外向性（E）、開放性（O）、協調性（A）、誠実性（C）の 5 つの独立した次元で構成される。本研究では、入力テキストから各次元のスコア（0–99 の連続値）を推定する多変数回帰問題としてタスクを定式化する。

3.2 性格推定モデル

3.2.1 mmBERT-base（エンコーダベースモデル）

多言語事前学習エンコーダとして mmBERT-base[1] [9] を採用する。本モデルは、隠れ層次元 768、12 レイヤーから構成され、1,800 以上の言語をカバーしている。性格予測のため、[CLS] トークンの最終層出力に対してドロップアウト（$p = 0.1$）を

1） https://huggingface.co/jhu-clsp/mmBERT-base

適用後、Big Five の各次元に対応する独立した回帰ヘッド（768→1 の全結合層）へ入力する構造を採用した。訓練時は正解ラベルの Z-score 正規化を行い、推論時に元のスケール（0–99）へと逆変換を行う。損失関数には、Big Five の各次元の平均二乗誤差（MSE）の平均値を用いる：

$$\mathcal{L} = \frac{1}{5} \sum_{i \in \{N,E,O,A,C\}} (\hat{y}_i - y_i)^2 \tag{1}$$

mmBERT では、同一話者の全発話を [SEP] トークンで連結した長文テキストを入力とする。テキストがモデルの最大長（8192 トークン）を超える場合、重複を持たせたチャンク分割を行い、各チャンクの予測値を平均化する「話者レベル・チャンク分割推論」方式を採用した。

3.2.2 Qwen3-30B（LLM ベースモデル）

最新の多言語 LLM である Qwen3-30B-A3B-Thinking-2507[2] をゼロショットで性格特性の推定を行う。LLM は mmBERT と比較して十分なコンテキスト長 (32K) を持つため、チャンク分割なしで全発話を一括入力できる。また、本モデルは推論（thinking）機能を備えており、対話内容から話者の性格を推論する能力を検証する。推論効率向上のため vLLM フレームワークを用い、Guided Decoding 機能により出力を特定の JSON 形式に制約する。評価方式として、同一話者の全発話を連結して一括処理する「話者レベル推論」を検証する。

3.3 評価指標

翻訳一貫性指標（主要評価） 本研究の主目的は、翻訳プロセスにおける性格情報の保持度合いを定量化することである。翻訳一貫性の評価では、英語原文と日本語翻訳から得られた性格予測値の差異を、ベクトル間平均絶対誤差（MAE）、線形相関（Pearson r）、順位相関（Spearman ρ）によって測定する。MAE は予測値の絶対的なズレを、Pearson r は数値的な比例関係を、Spearman ρ はサンプル間の相対的順序の保持度を評価する。高い一貫性は、翻訳モデルが性格特性に関連する言語的特徴を目標言語の適切な表現へと変換・保持できていることを示唆する。

予測精度指標（補助評価） 性格予測モデルの基礎的な能力を検証するため、Ground Truth が存在するデータに対して翻訳一貫性と同様に MAE、Pearson r、Spearman ρ を評価し、補助的に報告する。

4 実験設定

4.1 データセット

訓練および予測精度評価データ 英日間の性格予測の一貫性を評価するには、両言語に対応した予測モデルが必要となる。そのため、性格予測モデルの訓練には、英語と日本語の 2 つのデータセットを統合して使用した。訓練データは、英語データとして PANDORA [13]（Reddit 由来、102,523 件）を、日本語データとして RealPersonaChat（RPC）[17]（8,544 件、233 話者）を使用した[3]。開発データは PANDORA 12,803 件、RPC 1,265 件を使用し、早期終了の判定に用いた。RPC の 1–7 尺度の自己評価スコアは、以下の線形変換により 0–99 尺度に正規化した：

$$y_{\text{norm}} = \frac{y_{\text{orig}} - 1}{6} \times 99 \tag{2}$$

全てのデータは話者単位で分割し、テストセットへの話者情報のリークを厳密に排除した。

予測精度の評価には、PANDORA テストセット（12,803 件）および RPC テストセット（2,304 件、47 話者）を使用した。PANDORA テストセットでの評価は、提案モデルが性格特性を抽出する基礎能力を検証するためのベンチマークとして位置づける。一方、RPC テストセットでの評価は、訓練データが限られる日本語におけるモデルの汎化性能を検証する目的で実施する。

翻訳一貫性評価データ 英日対訳コーパス MELD-ST [8] を使用する。MELD-ST は、アメリカのドラマ「Friends」に基づく感情ラベル付き音声翻訳データセットであり、以下 Friends データセットと呼ぶ。本研究では、性格情報の蓄積を確保するため、テストデータから 16 発話以上を持つ話者 34 名を選定した。このうち主要な主人公 6 名については、クラウドソーシング型性格評価データベース Personality Database[4] から性格特性の Ground Truth を取得した。各キャラクターの Big Five スコアは、ユーザー投票（0%, 25%, 50%, 75%, 100%の 5 段階）の投票数による加重平均として算出されている。一貫性評価では、同一発話の英語原文と日本語翻訳に

2） https://huggingface.co/Qwen/Qwen3-30B-A3B-Thinking-2507

3） 各サンプルは同一ユーザー/話者の複数投稿・発話を [SEP] で連結したテキスト。

4） https://www.personality-database.com

表 1　Friends における翻訳一貫性

モデル	Pearson	Spearman	MAE
mmBERT	0.217	0.204	17.6
Qwen3	0.559	0.550	29.0

表 2　データセット・モデル別の性格予測精度

データセット	モデル	Pearson	Spearman	MAE
PANDORA (英語)	mmBERT	0.433	0.415	23.9
PANDORA (英語)	Qwen3	0.124	0.117	26.9
RPC (日本語)	mmBERT	−0.059	−0.065	13.5
RPC (日本語)	Qwen3	0.076	0.072	14.9

対してそれぞれ性格予測を行い、両者の予測値の一貫性を測定する。

実験で使用したハイパーパラメータを付録 A に示す。

5　結果

5.1　翻訳一貫性

表 1 に、モデルごとの全体的な翻訳一貫性を示す。相関係数（Pearson および Spearman）に着目すると、Qwen3 は mmBERT と比較して大幅に高い値を示している。一方で、MAE に関しては Qwen3 の方が大きな値（29.0）となっている。これは、相関は高いものの、Qwen3 が予測するスコアの分散が mmBERT よりも広くなりやすいため、絶対値としての誤差が拡大した可能性がある。Big Five 次元別の結果と分析を付録 B.1 に示す。

5.2　性格予測精度

表 2 に、各データセットにおける性格予測精度を示す。PANDORA テストセット（英語）において、mmBERT は平均 Pearson 相関 0.433 を達成した。これは先行研究である Fu ら [12] の報告値（Big Five 平均: 0.133）を大幅に上回る結果であり、ファインチューニングによる性格予測能力の向上を示している。Qwen3 は同データセットで 0.124 にとどまり、先行研究と同程度の精度であった。一方、RPC テストセット（日本語）では、mmBERT・Qwen3 ともに平均 Pearson 相関が 0 付近となり、有意な予測性能を示さなかった。この結果の解釈については考察で議論する。Big Five 次元別の結果を付録 B.2 に示す。

5.3　Ground Truth との比較

表 3 に、Friends に登場する 6 名の主人公に対する Ground Truth との比較を示す。サンプル数が少ないため（n=6）、相関係数ではなく MAE で評価した。

表 3　主人公 6 名の Ground Truth との比較（MAE）

キャラクター	GT-EN	GT-JA	EN-JA
Rachel	26.0	22.4	17.8
Monica	30.4	24.4	15.8
Phoebe	25.3	26.9	22.1
Ross	35.1	19.9	19.7
Chandler	20.8	7.7	20.6
Joey	33.4	21.0	18.8
平均	28.5	20.4	19.1

なお、Ground Truth はクラウドソーシングによる投票の加重平均であり、参考値として位置づける。

興味深いことに、日本語翻訳からの予測（GT-JA: MAE=20.4）が英語原文からの予測（GT-EN: MAE=28.5）よりも Ground Truth に近い値を示した。ただし、Ground Truth 自体がキャラクターのステレオタイプに基づく可能性があるため、この結果の解釈には注意が必要である。より重要な知見は、英日間の予測一貫性（EN-JA MAE=19.1）が高いことであり、これは翻訳プロセスにおける性格情報の保持を示唆している。

5.4　mmBERT と Qwen3 の比較

mmBERT は英日間で MAE=17.6 を達成し、多言語エンコーダが言語を越えて共通の性格空間を学習している可能性を示唆する。一方、Qwen3 の話者レベル推論は次元別 Pearson 相関（平均 0.559）において mmBERT（平均 0.217）を上回り、LLM が高い推論能力に基づき、特定の次元における話者間の相対的な順位関係を捉えることに長けていることを示した。両アプローチの相補的な活用が、より頑健な評価に寄与する可能性がある。

6　おわりに

本研究では、英日翻訳における性格特性の保持能力を定量的に評価した。実験の結果、英日間で一定の予測一貫性が確認された（mmBERT: MAE=17.6、Qwen3: Pearson r=0.56）。mmBERT は絶対的な予測値の近さにおいて、Qwen3 は話者間の相対的な順位関係において優れた一貫性を示した。また、日本語翻訳からの予測が英語原文よりも Ground Truth に近いという結果が得られ、役割語による性格シグナルの増幅効果が示唆された。今後は日本語性格予測の精度向上に向け、より高品質なデータセットの構築が課題である。

謝辞

本研究は NTT コミュニケーション科学基礎研究所および JSPS 科研費 JP23K28144 の助成を受けたものです.

参考文献

[1] 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. Association for Computational Linguistics, 2002.

[2] Ricardo Rei, Jos'e G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and Andr'e F. T. Martins. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In **Proceedings of the Seventh Conference on Machine Translation**, pp. 578–585, 2022.

[3] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In **Advances in Neural Information Processing Systems**, Vol. 36, 2023.

[4] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. **arXiv preprint arXiv:2402.03300**, 2024.

[5] Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. Motivating personality-aware machine translation. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 1102–1108, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[6] Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. Personalized machine translation: Preserving original author traits. In **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers**, pp. 1074–1084, Valencia, Spain, April 2017. Association for Computational Linguistics.

[7] 金水敏. 仮想日本語 役割語の謎. 岩波書店, 2003. (Kinsui, Satoshi. Virtual Japanese: The Mystery of Role Language. Iwanami Shoten.).

[8] Sirou Chen, Sakiko Yahata, Shuichiro Shimizu, Zhengdong Yang, Yihang Li, Chenhui Chu, and Sadao Kurohashi. MELD-ST: An emotion-aware speech translation dataset. In **Findings of the Association for Computational Linguistics: ACL 2024**, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[9] Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. mmBERT: A modern multilingual encoder with annealed language learning. **arXiv preprint arXiv:2509.06888**, 2025.

[10] Qwen Team. Qwen3 technical report, 2025.

[11] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics.

[12] Yahui Fu, Simone Filice, and Ida Mele. StyEmp: Stylizing empathetic response generation with multi-grained prefix encoder and personality reinforcement. In **Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue**. Association for Computational Linguistics, 2024.

[13] Matej Gjurković and Jan Snajder. PANDORA talks: Personality and demographics on Reddit. In **Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media**, pp. 138–152, Online, June 2021. Association for Computational Linguistics.

[14] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6442–6454, Online, November 2020. Association for Computational Linguistics.

[15] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In **Advances in Neural Information Processing Systems**, Vol. 35, 2022.

[16] Tulika Saha, Sriparna Saha Reddy, and Pushpak Bhattacharyya Prakash. Stylistic response generation by controlling personality traits and intent. In **Proceedings of the 4th Workshop on NLP for Conversational AI**. Association for Computational Linguistics, 2022.

[17] 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.

A　モデル訓練ハイパーパラメータ

mmBERT の訓練は AdamW 最適化アルゴリズムを用い、学習率 2e-5、最大 5 エポック（早期終了適用）で実施した。入力の最大系列長は 256 トークンとした。Qwen3-30B の推論では、再現性を担保するため temperature=0.0（貪欲デコード）を採用し、bfloat16 精度で実行した。

B　Big Five 次元別結果

B.1　Friends における次元別翻訳一貫性

次に、次元ごとの詳細な分析結果を表 4 に示す。mmBERT の場合、誠実性（C）では一定の一貫性（Pearson=0.481）を維持できているものの、開放性（O）では負の相関（−0.045）、神経症傾向（N）でもほぼ無相関（0.035）となっている。これに対し、Qwen3 はすべての次元において mmBERT を上回る精度を示した。

表 4　Friends における次元別翻訳一貫性

次元	mmBERT		Qwen3	
	Pearson	Spearman	Pearson	Spearman
Neuroticism (N)	0.035	−0.024	0.239	0.249
Extraversion (E)	0.379	0.413	0.729	0.716
Openness (O)	−0.045	0.034	0.765	0.636
Agreeableness (A)	0.236	0.144	0.402	0.428
Conscientiousness (C)	0.481	0.453	0.662	0.719

B.2　次元別性格予測精度

表 5　PANDORA における次元別予測精度

次元	mmBERT		Qwen3	
	Pearson	Spearman	Pearson	Spearman
Neuroticism (N)	0.428	0.410	0.110	0.119
Extraversion (E)	0.431	0.396	0.199	0.178
Openness (O)	0.426	0.410	0.120	0.105
Agreeableness (A)	0.499	0.500	0.110	0.113
Conscientiousness (C)	0.384	0.358	0.082	0.070

表 6　RPC における次元別予測精度

次元	mmBERT		Qwen3	
	Pearson	Spearman	Pearson	Spearman
Neuroticism (N)	−0.124	−0.134	0.131	0.118
Extraversion (E)	−0.076	−0.101	−0.079	−0.103
Openness (O)	0.172	0.150	0.314	0.350
Agreeableness (A)	−0.144	−0.136	−0.213	−0.225
Conscientiousness (C)	−0.126	−0.104	0.227	0.218

表 5、表 6 にそれぞれ PANDORA、RPC テストセットでの Big Five 次元別の精度を示す。ただし、Qwen3 は開放性（O: 0.314）と誠実性（C: 0.227）で比較的高い相関を示しており、次元によっては一定の予測能力を持つ可能性が示唆された。

C　本研究の限界

本研究には主に二つの限界が存在する。

第一に、評価に用いた Friends キャラクターの Ground Truth（GT）の性質である。 本研究で使用した GT は、専門家による心理学的診断ではなく、Personality Database のユーザー投票に基づいた「社会的合意」としてのスコアである。 投票者はキャラクターのステレオタイプに基づいて判断している可能性があり、本研究の結果は、厳密な心理学的特性というよりも、「人間が抱く性格イメージ」と「翻訳後の言語表出」の一致度を示したものとして解釈すべきである。

第二に、日本語対話データ（RPC）における予測精度の低さと、実データを用いた学習の困難さである。 実験の結果、Friends データセットでは日本語翻訳からの予測が英語原文よりも GT に近い値を示し、モデルの日本語処理能力自体には一定の有効性が確認された。 しかし、RPC テストセットにおいては、mmBERT・Qwen3 ともに有意な予測性能を示さなかった。 この原因として、RPC の訓練データ量が PANDORA の約 8%と限定的であることに加え、日本語対話特有の短文・主語省略といった構造的特徴が、性格シグナルの抽出を困難にしていると考えられる。 さらに、RPC の性格スコアは自己申告に基づいているため、日本語話者特有の自己提示傾向（建前）が、テキスト表現と実際の性格スコアとの間に乖離を生んでいる可能性も否定できない。

以上のことから、日本語の実対話データにおける学習の困難さは本研究の枠組みでは十分に解決されていないと言える。 今後の課題として、テキストと性格ラベルの乖離を埋めるため、役割語や敬語といった日本語特有の表出メカニズムを陽に組み込んだモデルアーキテクチャの検討が必要である。