LUNON: 相対尤度に基づくテキストの本人らしさの評価指標

多田龍之進 [1]　石井太河 [1]　宮尾祐介 [1,2]
[1] 東京大学　[2] 国立情報学研究所大規模言語モデル研究開発センター
tada31@g.ecc.u-tokyo.ac.jp, {taigarana,yusuke}@is.s.u-tokyo.ac.jp

概要

特定の個人の文体や考え方を本人らしく模倣したテキスト生成において、生成されたテキストの「本人らしさ」を定量的に評価することは重要な課題である。しかし従来の自動評価指標は、一般的な言語パターンと区別される「本人らしさ」の微妙な違いを十分に捉えられていない。本論文では、対象個人の文章で訓練した言語モデルでは高確率だが、一般的な文章で訓練したモデルでは低確率となるテキストほど本人らしい、という単純な仮説を立て、その検証のために、両モデル間の正規化された対数尤度差に基づく評価指標 LUNON を提案する。日本人アイドルのブログ記事を用いた実験では、LUNON のスコアは専門知識を有するファンによる本人らしさ評価と中程度の相関（Spearman の $\rho \approx 0.58$）を示し、この仮説を支持した。一方で、モデルの学習データと人間評価者の知識量の差に起因する乖離も観察された。

1　はじめに

大規模言語モデル（LLM）の発展により、特定の個人やキャラクターを模倣する AI システムの構築が現実的になってきた。ファインチューニングやプロンプトエンジニアリングなどの技術を用いることで、対象個人の文章を高い精度で模倣できるようになった [1]。最近の研究では、LLM が生成した応答が実際の人間の応答よりも本物らしいと判断される場合があることも示されている [1]。

しかし、**生成されたテキストが本人らしいかどうかを定量的に評価する**という根本的な課題が残されている。既存研究の多くは対話シナリオにおけるペルソナ一貫性の評価に焦点を当てているが [2, 3]、本研究では、単一のテキストが特定の人物らしさをどの程度示しているかを測定する。このような文レベル・発話レベルの評価は、特定の個人を模倣するように訓練された言語モデルの品質評価や、何がテキストを本人らしく代表的なものにするかを理解する上で不可欠である。

本研究では、**本人らしさ**を、対象個人によって生成される可能性が高い一方で、他者によって生成される可能性が低い言語パターンをテキストが示す度合いと定義する。本人らしいテキストは、単に対象個人が頻繁に用いる表現を含むだけでなく、言葉の流れや世界の見方といった、対象個人に固有の確率的生成過程を反映していると考えられる。確率論的観点から見ると、本人らしい発話とは、対象個人の文章で訓練した言語モデルの下では尤もらしいが、一般的な文章で訓練したモデルの下では相対的に尤もらしくないものと仮定できる。しかし、この根底にある確率的非対称性を明示的にモデル化した既存手法は存在しない。

既存の評価指標は、このような確率的観点から本人らしさを評価するものではない。単語の重なりに基づく従来の指標（例：BLEU [4]）や埋め込み類似度に基づく指標（例：BERTScore [5]）は、参照テキストとの類似性を測定するものであり、対象個人と他者を区別する確率的言語パターンを直接評価するものではない。PersonaCLR [6] のような埋め込みベースのアプローチも、対照学習を通じてペルソナ特性を埋め込み空間に投影しているが、本研究で提案する確率的非対称性の観点からの評価は行っていない。

この確率的非対称性の仮説を検証するため、本研究では **LUNON**（**L**og-likelihood-based **U**niqueness using **N**ormalized **O**ffset to a **N**on-personalized model）を提案する。これは、対象個人の文章で訓練した言語モデル（ペルソナ特化型モデル）と一般的な文章で訓練した言語モデル（一般モデル）との間の尤度の乖離を定量化する指標である（図 1）。本人らしいテキストは、ペルソナ特化型モデルによって高い確率が割り当てられる一方で、一般モデルによっては比較的低い確率が割り当てられると予想され、LUNON はその非対称性を直接測定する。

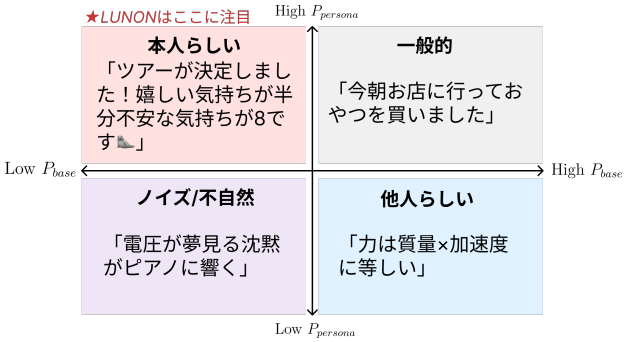

図 1　尤度に基づくペルソナ評価の概念図。LUNON は、一般言語モデルの下では比較的尤度が低く（低 P_{base}）、ペルソナ特化型モデルの下では比較的尤度が高い（高 P_{persona}）テキストに高いスコアを割り当てる。

本研究では、日本のアイドルグループメンバーの文章から構築されたデータセットで LUNON を検証する。アイドルは定期的にブログ投稿等で個性を表現する文章を生成し、熱心なファン層が各メンバーの言語パターンを熟知しているため、本人らしさ評価の検証に適している。

本研究の貢献は三つある：(1) ペルソナ特化型モデルと一般モデルとの間の尤度の非対称性に基づく本人らしさの仮説を提示し、それを検証する評価指標 LUNON を提案する。(2) 専門知識を有するファンによる人手評価を用いた実験により、提案手法の妥当性を示す。(3) 定性分析により、尤度ベースのアプローチが捉える本人らしさの側面と、その限界について考察する。

2　提案手法

本研究の中心仮説は、本人らしいテキストが特徴的な確率パターンを示すことである。すなわち、ペルソナ特化型モデルの下では尤度が高いが、一般モデルの下では尤度が低い。この仮説は、本人らしいテキストは対象個人にとって自然に生成することは比較的容易だが、他の書き手が似たような文章を生成するのは困難であるという観察から生じる。

仮説を検証するため、LUNON を以下のように定式化する。LUNON は、候補テキスト $x = (w_1, \ldots, w_{|x|})$ が与えられたペルソナに対してどの程度特徴的かを定量化する：

$$\text{LUNON}(x) = \frac{\log P_{\text{persona}}(x) - \log P_{\text{base}}(x)}{|x|}, \quad (1)$$

ここで、P_{persona} はペルソナ特化型モデルを示し、P_{base} は一般モデルを示す。異なる長さのテキスト間で公平な比較を保証するため、テキスト長 $|x|$ で正規化する。

標準の LUNON 指標は、真にペルソナ特化的なテキストではなく、ドメイン特化的なテキスト（例：アイドルらしい言語パターン）に高いスコアを割り当てる可能性がある。この制限に対処するため、一般モデルの代わりに、**類似した**ペルソナ（例：アイドルグループの他のメンバー）でファインチューニングされたモデルを P_{base} として使用する。この変種を **LUNON-Sim** と呼び、同じドメイン内のペルソナが実質的な話題の語彙、文体の慣習を共有する場合に、個人の本人らしさをより識別しやすい尺度となる。

3　実験

本実験では、提案手法 LUNON が本人らしさを適切に評価できるかを検証する。実験の流れは以下の通りである：(1) 日本のアイドルグループメンバーのブログデータを収集し、(2) 各ペルソナ特化型モデルを訓練し、(3) 5 種類の生成手法を用いて各ペルソナにつき 30 の評価用テキストを生成し、(4) 熱心なファンによる人手評価で本人らしさのスコアを取得し、(5) LUNON および既存の評価指標で自動評価を行い、(6) 人手評価と自動評価の相関を分析する。

3.1　データセット

仮説を検証するため、日本のアイドルグループメンバーのブログ投稿を用いた実験を実施した。5 人のメンバー（A–E と表記、以降それぞれを 1 つのペルソナとして扱う）の公開ブログ投稿を収集した。各ブログは複数年にわたり、全ブログテキストを約 140 文字のセグメントに分割し、メンバーあたり 400 から 3,000 のテキストセグメントを得た。データは訓練セットと検証セットにランダムに分割した（詳細な統計情報は付録 A を参照）。

3.2　モデル訓練

各ペルソナの訓練データを用いて、Llama-3.1-Swallow-8B [7] を継続事前学習により各ペルソナに特化させた言語モデルを構築した。これらのモデルは、LUNON スコアの計算における P_{persona} として使用される（詳細な実装設定は付録 B を参照）。

3.3　評価用テキスト生成

本人らしさの程度が異なるテキストを得るため、5 種類の生成手法を用いて多様な評価データセットを作成した。評価用プロンプトとして、GPT-4o

で 6 つの中立的なブログ開始文を生成した。各プロンプトに対して以下の 5 種類の手法でテキストを生成し、ペルソナあたり 30 のテキストを得た：(1) Llama-3.1-Swallow-8B-Instruct [7] を用いた対象ペルソナのファインチューニング、(2) GPT-4o[8] を用いた対象ペルソナのファインチューニング、(3) 他のペルソナのファインチューニング、(4) 50 例による few-shot 生成、(5) zero-shot 生成。

3.4 既存の評価指標

以下の評価指標と比較した：**MaxBLEU** [9] は、生成テキストと訓練コーパス内の任意の参照テキスト間の最大 BLEU スコアを測定する。**BERTScore** [5] は、事前訓練された BERT モデルからの文脈埋め込みを用いて意味的類似性を計算する。**uPPL** [10] は、ペルソナ特化型言語モデルの下での生成テキストの perplexity を測定する指標である。**Persona Speaker Probability（PSProb）** [11] は、多項分類器を用いて特定のペルソナによってテキストが生成される確率を測定する。**Persona Term Salience（PTSal）** [11] は、特定のペルソナに対する用語（n-gram）の重要性を測定することでペルソナ特性を定量化する。**PersonaCLR** [6] は、話者の同一性に基づく対照学習により、発話のペルソナ特性の強度を測定する。**GPT-4o** [8] は、1 から 5 のスケールで本人らしさを評価する自動評価器として採用した。

3.5 人手評価

X（旧 Twitter）[1] から 39 人の熱心なファンをボランティアベースで募集し、人手評価に参加してもらった。アノテーターは主に 20 歳から 50 歳の日本人男性である。これらのファンは、各自が好むメンバーを平均 3 年間応援しており、非常に熱心で知識豊富な評価者である。各ファンは、好むメンバーについて生成された 30 のテキストすべてを、本人らしさの 5 段階リッカート尺度で評価した。また、各評価者から、応援期間、ブログ閲覧頻度、エンゲージメントレベルなどの背景情報を収集した。

アノテーター間一致度は Krippendorff の $\alpha = 0.34$ であった。「本人らしさ」という主観的な評価基準である以上、ある程度の不一致は予想されるが、比較的低い一致度（$\alpha = 0.336$）は、評価者の背景がスコアの変動に影響を与えた可能性を示唆している。

1） https://x.com/

3.6 相関分析

自動指標とテキストあたりの平均人手評価スコアとの間の Spearman の ρ を報告する。

モデルの訓練および評価に関する詳細な実装設定については付録 B を参照されたい。

4 結果

表 1 の相関結果は、仮説をある程度支持している。LUNON は、人手による本人らしさ評価と平均 Spearman の $\rho = 0.58$ の相関を示した。この中程度の相関は、提案した対数尤度パターンが本人らしさの一側面を捉えていることを示している。

ペルソナごとの相関 結果から、異なるペルソナ間で大きな変動が見られた。LUNON は、ペルソナ A（$\rho = 0.83$）とペルソナ C（$\rho = 0.56$）で特に強い相関を示し、LUNON-Sim はペルソナ B（$\rho = 0.70$）と D（$\rho = 0.55$）で高い相関を示した。

ベースライン手法との比較 MaxBLEU のような従来の指標は、弱い相関や負の相関を示しており、語彙の重なりだけでは本人らしさを測る指標として不十分であることが示された。PTSal（平均 $\rho = 0.46$）と BERTScore（平均 $\rho = 0.44$）は中程度の相関を示した。PersonaCLR（平均 $\rho = 0.38$）と PSProb（平均 $\rho = 0.39$）は、ペルソナを意識した手法であるにもかかわらず、提案手法を一貫して下回った。

uPPL から得られる示唆 重要なことに、ペルソナ特化型モデルの確率のみを用いる uPPL 指標は、人手評価との相関が低い（平均 $\rho = -0.08$）。この結果は、ペルソナ特化型言語モデルの確率だけでは本人らしさを十分に捉えられないことを示唆しており、仮説の重要な側面を支持している。ペルソナ特化型モデルと一般モデルの確率間の乖離を計算する LUNON によって達成されたより高い相関は、提案した非対称確率パターンが本人らしいテキストに存在する可能性を示している。

5 考察

提案した仮説の妥当性を検証するため、生成テキストの定性分析を実施した。図 2 は人手評価と LUNON の両方で高スコアを得たケースを示す。特徴的な言い回しや絵文字の使用、口調、話の流れなど、複数の側面でペルソナらしさを捉えていた。

一方、図 3 に示すように、人手評価と LUNON の

表 1 人手評価スコアと自動評価スコアとの間の Spearman の順位相関係数

指標	A	B	C	D	E	平均
LUNON	**0.83**	0.61	**0.56**	0.48	0.43	**0.58**
LUNON-Sim	0.76	**0.70**	0.27	**0.55**	0.44	0.54
GPT-4o	0.78	0.61	0.46	0.38	**0.56**	0.56
PSProb	0.55	0.57	0.31	0.26	0.25	0.39
PTSal	0.57	0.52	0.39	0.41	0.39	0.46
PersonaCLR	0.37	0.45	0.50	0.34	0.25	0.38
BERTScore	0.41	0.59	0.50	0.30	0.38	0.44
uPPL	−0.35	0.08	−0.09	0.24	−0.27	−0.08
MaxBLEU	0.13	0.60	−0.12	0.53	−0.38	0.15

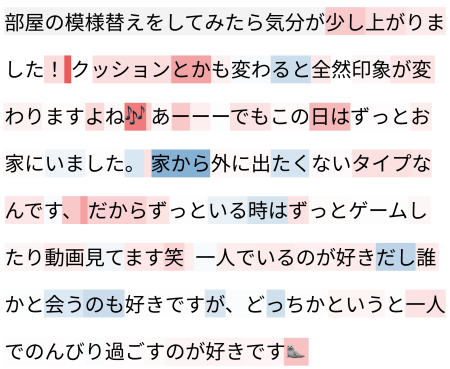

図 2 サンプル 23：人手評価（4.1）と LUNON（0.44）の高い一致。赤色はペルソナモデル優位、青色はベースラインモデル優位を示し、色の濃さは $|\Delta \log p|$ を表す。灰色はプロンプト。

間に乖離が見られるケースも観察された。LUNON が高スコアを付けたが人手評価が中程度だったケースでは、人間評価者は「このペルソナはこのような詳細に物事を話したりしない」といった、人物に対するより深い理解に基づいて評価していた。このような乖離は、モデルの学習データに含まれる情報と人間が持つ知識量の差に起因すると考えられる。人間評価者は数年間にわたる継続的なエンゲージメントを通じて、学習データには明示的に現れない性格特性や行動パターンを理解している可能性がある。この乖離は尤度ベースの指標の根本的限界を示唆する一方で、より多くの学習データを用いることで改善できる可能性もある。

6 おわりに

本研究では、本人らしいテキストは、ペルソナ特化型モデルの下では尤もらしいが、一般言語モデルの下では比較的尤もらしくないという特徴的な確率パターンを示すという仮説を調査した。LUNON の

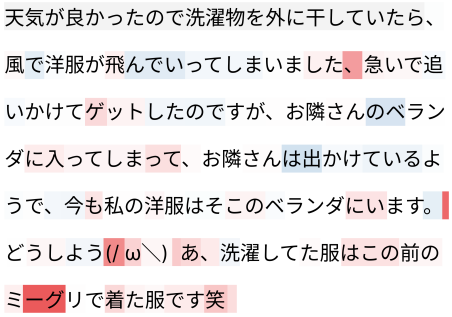

図 3 サンプル 25：アイドル特有の用語により LUNON は高スコア（0.36）だが、人間は中程度の評価（3.3）。色の配置は図 2 と同様。

開発と日本のアイドルブログ投稿を用いた実験を通じて、この仮説の妥当性を示す結果を得た。ペルソナ特化型モデルと一般モデル間の尤度乖離は、人間による本人らしさ判断と中程度の相関（Spearman の $\rho = 0.58$）を示した。しかし、分析により限界も明らかになり、人間評価者が学習データに含まれない性格特性や行動パターンに基づいて評価する場合、LUNON との乖離が生じることが示された。

今後の課題として、多様なドメインや他言語での LUNON の有効性を評価することで、提案手法の一般化可能性を検証する必要がある。

謝辞

本研究は、文部科学省補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の支援を受けたものです。

参考文献

[1] Steffen Herbold, Alexander Trautsch, Zlata Kikteva, and Annette Hautli-Janisz. Large language models can impersonate politicians and other public figures, 2024.

[2] Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11836–11850, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[3] Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Compu-*

tational Linguistics (Volume 1: Long Papers), pp. 1840–1873, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[5] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

[6] Michimasa Inaba. PersonaCLR: Evaluation model for persona characteristics via contrastive learning of linguistic style representation. In Tatsuya Kawahara, Vera Demberg, Stefan Ultes, Koji Inoue, Shikib Mehri, David Howcroft, and Kazunori Komatani, editors, *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 674–685, Kyoto, Japan, September 2024. Association for Computational Linguistics.

[7] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In *Proceedings of the First Conference on Language Modeling*, COLM, p. (to appear), University of Pennsylvania, USA, October 2024.

[8] OpenAI. Gpt-4o system card, 2024.

[9] Zhen Xu, Nan Jiang, Bingquan Liu, Wenge Rong, Bowen Wu, Baoxun Wang, Zhuoran Wang, and Xiaolong Wang. LSDSCC: a large scale domain-specific conversational corpus for response generation with diversity oriented evaluation metrics. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2070–2080, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[10] Bowen Wu, MengYuan Li, Zongsheng Wang, Yifu Chen, Derek F. Wong, Qihang Feng, Junhong Huang, and Baoxun Wang. Guiding variational response generator to exploit persona. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 53–65, Online, July 2020. Association for Computational Linguistics.

[11] Chiaki Miyazaki, Saya Kanno, Makoto Yoda, Junya Ono, and Hiromi Wakaki. Fundamental exploration of evaluation metrics for persona characteristics of text utterances. In Haizhou Li, Gina-Anne Levow, Zhou Yu, Chitralekha Gupta, Berrak Sisman, Siqi Cai, David Vandyke, Nina Dethlefs, Yan Wu, and Junyi Jessy Li, editors, *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 178–189, Singapore and Online, July 2021. Association for Computational Linguistics.

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[13] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

A　データセット詳細統計

表 2 に、各ペルソナのブログデータセットの詳細統計を示す。カウントはテキストセグメント数（ブログ投稿数ではない）を示す。総文字数は全テキストセグメントの合計文字数、平均/中央値/最小/最大文字数は個々のテキストセグメントあたりの文字数を表す。

表 2　ブログデータセット詳細統計

	A		B		C	
	訓練	検証	訓練	検証	訓練	検証
カウント	2454	306	572	71	332	41
総文字数	348,770	43,388	84,300	10,806	45,242	5,281
平均文字数	142.1	141.8	147.4	152.2	136.3	128.8
中央値文字数	149.0	146.0	156.5	161.0	146.0	134.0
最小文字数	52	52	50	48	51	54
最大文字数	366	219	214	200	206	173
標準偏差	36.0	35.0	34.1	32.8	32.8	28.7

	D		E		全体	
	訓練	検証	訓練	検証	訓練	検証
カウント	3032	379	1265	158	7,655	955
総文字数	431,039	53,691	177,647	21,903	1,086,998	135,069
平均文字数	142.2	141.7	140.4	138.6	142.0	141.5
中央値文字数	149.0	148.0	145.0	145.0	148.0	146.0
最小文字数	47	55	50	66	47	48
最大文字数	342	254	242	199	366	254
標準偏差	34.3	33.1	30.8	30.4	34.2	32.9

B　実装詳細

すべての実験は、48GB メモリを持つ単一の NVIDIA RTX A6000 GPU で実施した。Llama モデルでは、rank 16、alpha 32、学習率 5×10^{-5}、埋め込み学習率 3×10^{-5}、バッチサイズ 32、AdamW 8 ビットオプティマイザ [12] で LoRA [13] を採用した。8 エポック訓練し、検証損失が最も低いモデルチェックポイントを選択した。

テキスト生成には Llama-3.1-Swallow-8B-Instruct-v0.3 [7] を使用し、評価指標には Llama-3.1-Swallow-8B-v0.5 [7] を使用した。GPT-4o [8] ファインチューニングは、バッチサイズ 8、4 エポック、learning rate multiplier 2 を使用した。

LUNON スコアでは、プロンプトと生成部分に分割せず、約 140 文字のテキストセグメント全体で継続事前訓練を用いてペルソナ特化型モデルを訓練した。これらのモデルを P_{persona} として用い、元の Llama-3.1-Swallow-8B-v0.5 を P_{base} として用いた。

LUNON-Sim では、19 人のアイドルグループメンバーのブログデータ（約 1,500 セグメント）でモデルを訓練し、一般的なアイドルの文章特性を捉え、これを P_{base} として用いることで、グループパターンから個人の特性を分離した。