E-013

大規模言語モデルは年齢を超えて文体を生成できるか
Can a Large Language Model Generate Writing Styles across Ages

井下 敬翔† ‡
Keito Inoshita

1. はじめに

自然言語処理（NLP）における深層学習に基づいたモデル構築では,Web ページ,SNS,学術論文などから得られる膨大なデータをもとに訓練され,汎用的かつ高精度な予測を実現してきた.しかし,これらの学習データの多くは成人向けテキストに偏っており,文体の多様性には依然として課題が残る.特に,児童や高齢者の言語使用に関しては,データ収集に伴う倫理的・制度的制約が大きく,十分な学習資源が存在しない.このような偏りは,モデルの汎用性や公平性を制限する可能性があり,より多様な言語スタイルを使用したモデル構築が求められている.

このような課題に対し,LLM の柔軟な言語生成能力を活用することで,年齢ごとの文体に対応したテキストを人工的に生成し,新たな学習資源として活用する可能性が考えられる [1].とりわけ日本語は,ひらがな・カタカナ・漢字という 3 種類の文字体系が発達段階に応じて使用比率を変化させるという特徴を有しており,文体の発達的変化をモデル化する上で極めて興味深い対象となる.しかし,LLM が日本語生成においてどれほど年齢ごとの文体を再現できるかは明らかになっていない.

そこで本研究では,LLM によって生成された日本語テキストが,実際に年齢相応の発達段階を模倣できているのか,またそれがどれほどもっともらしいのかを検証することを目的とする.具体的には,複数の質問に対する回答を,LLM に 8 段階の年齢スタイルで生成してもらい,その結果を日本語難易度推定器 (jReadability) [2] および,LLM-as-a-Judge による自動的なマッチング評価を実施し,生成文がどの程度もっともらしい文体を再現できているかを評価する.これにより,LLM を活用した年齢別データ拡張の新たな可能性を提示することを目指す.

2. 年齢別スタイル生成と評価フレームワーク

2.1 質問項目の設計

年齢ごとの語彙選択,構文の複雑さ,思考様式の違いを適切に引き出すため,本研究では 10 個の自由記述型質問を設計した.これらの質問は,対象となる年齢層における内面の表現や言語発達の特性が自然に表出されるよう,テーマの選定において幅広い抽象度と感情的深度を意識して構成されている.具体的には,「春と聞いて思い浮かべるもの」といった日常的かつ具体的な問いから,「ひとりぼっちを感じるとき」などの感情表現や自我形成に関わる問い,さらには「大切にしていること」といった抽象的・社会的な主題に至るまで,段階的な認知的要求を含むよう設計している.これにより,発達段階ごとの自然な語り方や価値観の違いを浮かび上がらせることを目的としている.

† 関西大学 商学研究科
Kansai University, Faculty of Business and Commerce
‡ 滋賀大学 データサイエンス・AI 研究推進センター
Shiga University, Data Science and AI Innovation Research

2.2 LLM による年齢別スタイル生成

生成対象は,次の 8 段階の年齢カテゴリに設定した：小学校低学年（6–8 歳）,小学校高学年（9–12 歳）,中学生（13–15 歳）,高校生（16–18 歳）,若年成人（19–29 歳）,中堅社会人（30–44 歳）,壮年層（45–64 歳）,高齢者（65 歳以上）.各質問に対し,OpenAI 社の GPT-4o モデルの API を用い,各カテゴリに 1 文ずつ生成させた.使用したプロンプトは,年齢層に応じた語彙,構文,文字種（ひらがな・カタカナ・漢字）の比率,一人称の語り口,感情表現などに言及し,自然な年齢別スタイル模倣が行われるよう設計されている.生成には temperature=0.7 を指定し,創造性を保持しつつ,過度な逸脱を抑制した.質問 10 件×年齢 8 カテゴリ＝計 80 文を生成した.

2.3 日本語難易度推定と LLM による検証

- jReadability による日本語難易度の推定

機械的なスタイル変化の客観評価には,jReadability を用いた.jReadability は,文数・語数・語彙階層・文字種分布・語源比率など 30 種類以上の統計的特徴を基に,重回帰モデルで読みやすさをスコア化（リーダビリティ・スコア）する.数値が大きいほど平易,小さいほど難解とされる.これにより,年齢カテゴリに応じた文体的特徴を定量的に評価した.

- LLM-as-a-Judge による主観的一致率・妥当性の検証

GPT-4o より高い知能を持つ OpenAI 社の GPT-4.1 モデルに対し,生成文と 8 つの年齢スタイルラベルを提示し,「この文はどの年齢スタイルに最も近いか」を重複なしに選択させるマッチング形式の評価と,その妥当性（その文がその年齢にふさわしいと思うか）を 3 段階（はい／どちらともいえない／いいえ）で尋ねた. temperature=0.7 とし,プロンプトは人間の段階的な認知プロセスを再現するように設計した.これを 3 回繰り返すことで,評価のばらつきを考慮した LLM の年齢別スタイル生成の精度と信頼性を評価する.

3. 実験

3.1 jReadability による文体的特徴の評価

各年齢カテゴリで生成された回答文に対し,jReadability を用いて 30 種以上の文体的特徴量を算出した.図 1 は,主要指標である総文字数,ひらがな数,漢字数,助詞数のカテゴリ別平均を示す.まず総文字数は,年齢の上昇に伴って増加し,中堅社会人で最大となった後,高齢者層でやや減少した.この傾向は, 語彙や構文に制限のある若年層から,中年層で構成力が最も高まり,高齢層では表現が簡素化する傾向を反映している.ひらがなは,総文字数に対する比率で見ると,幼年層ほど高く,高齢層で低下する.これは,漢字習得が未熟な段階ではひらがなが多く用いられる特徴をよく再現している.漢字の使用数は加齢とともに増加し,成人期に最大となるが,高齢層ではやや減少に転じており,助詞数は年齢に応じて緩やかに増加し,発達に伴う構文の複雑化や平易化を反映していると考えられる.

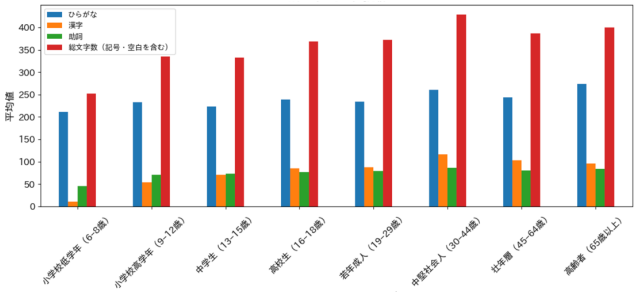

図 1　主要指標におけるカテゴリ別平均の分布

図 2 は,各特徴量を総文字数で割った比率を用い,8 カテゴリにおける文体的傾向をヒートマップとして示したものである.例外として,リーダビリティ・スコアと総文数は元の値を正規化した結果を用いている.特徴的なのは,語彙階層,品詞構成の使用比率に明確な発達段階の差が表れている点である.まず語彙階層では,初級前半語彙が小学生層に多く,中級後半や上級前半語彙は成人層で顕著となった.これは,語彙の発達段階を LLM が再現していることを示している.品詞構成では,助動詞や動詞の比率が年齢とともに上昇し,文構造の複雑化を反映している.特に高齢層で助動詞が顕著に増加しており,丁寧さや文末表現の多様性が再現されていると考えられる.以上より,LLM は語彙・文法・文字構成の複合的な変化を捉えた文体生成が可能であることが確認された.

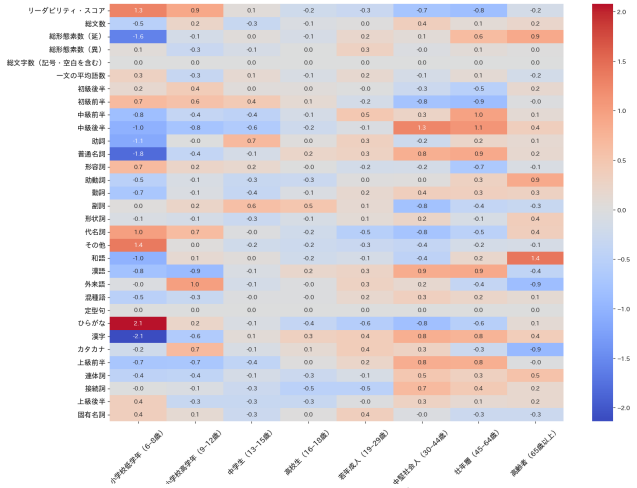

図 2　文体的特徴量のカテゴリ別の傾向

3.2 LLM による生成文の自動評価

生成文の文体がどの年齢カテゴリに最も近いと感じられるかを評価するため,GPT-4.1 に対して 8 カテゴリ×10 問の合計 80 文を 3 度提示し,質問ごとに各文に最も適したカテゴリを重複なく一度ずつ割り当てるマッチング評価を実施した.その結果,どの試行でも表 1 に示すような高い一致率を示し,多くの文においてカテゴリを識別できるような文体を再現していることが確認された.また,妥当性において,全てが「はい」との回答であり,妥当性に関しても高水準であった.

表 2　LLM-as-a-judge における試行ごとの一致率

試行番号	試行 1	試行 2	試行 3
一致率	95.0	92.5	95.0

図 3 に示す混同行列を見ると,中学生と高校生の誤分類が多く,両カテゴリ間での混同が顕著であった.これは,語彙や構文の成熟度に関して明確な境界が捉えにくい年代であるためと考えられる.これらの結果から,LLM は全体的に発達段階を反映した文体生成が可能である一方,中高生における微細なスタイル差の再現にはさらなる改善の余地があることが示唆される.

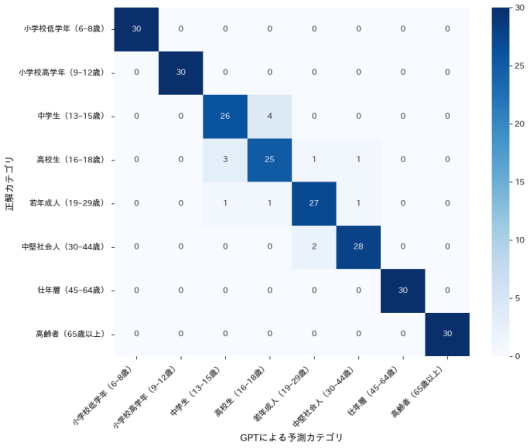

図 3　LLM-as-a-Judge における誤分類の分布

4. 議論

本研究は,LLM による年齢別スタイル生成が一定の精度で達成可能であることを示したが,より人間らしい文体再現のためにはプロンプト設計の改良が必要である.特に児童層では,その年齢で学習する常用漢字や語彙を辞書として参照することで,より自然な文体生成が可能になると考えられる.一方で,本研究ではうまく識別できたが,中堅社会人と壮年層の区別については,実際の人間による文でも区別が困難であることから,両層の文体的差異をより精緻に定義・抽出する必要がある.

また,本評価にはいくつかの限界が存在する.第一に,正確なスタイルの定義が曖昧であり,文字種や語彙階層といった表層的特徴に依存せざるを得ない点である.第二に,人間評価は評価者の主観に基づくものであり,一般性や再現性には限界がある.今後は,より客観的かつ多様な視点を取り入れた評価指標の導入が求められる.

5. 結論

本研究は,LLM が日本語における年齢別文体を一定程度再現可能であることを示した.語彙や構文の傾向が発達段階に沿って現れており,年齢スタイル模倣の有効性が確認された.今後は,プロンプト設計の工夫により,さらなる精度向上が期待される.

謝辞

本研究は,JST 次世代研究者挑戦的研究プログラム JPMJSP2150 の支援を受けたものである.

参考文献

[1] N. Bui et al., “Mixture-of-Personas Language Models for Population Simulation”, arXiv, (2025).
[2] S. Ishikawa et al., “Japanese Students’ L1 Story Writing Corpus (JASWRIC): A New Dataset for Analysis of L1/L2 Japanese”, Proceedings of Language Resources Workshop, Issue. 1, pp. 393-416, (2023).