LLM の文体差理解向上のためのデータセットの開発

森貞堅湧[1] 田窪洋介[2, 3] 高品良[4] 狩野洋一[4] Jeongmyeong Lee[4]
[1]新居浜工業高等専門学校 専攻科 電子工学専攻
[2]新居浜工業高等専門学校 電気情報工学科
[3]高エネルギー加速器研究機構 素粒子原子核研究所
[4]株式会社 APTO
{e1212038，Y.Takubo }@niihama.kosen-ac.jp
{takashina，y.karino，j.lee}@apto.co.jp

概要

大規模言語モデル（LLM）は高い日本語生成性能を示す一方で，話し言葉や書き言葉といった文体差に対する理解には課題が残されている．本研究では，話し言葉および書き言葉の理解向上を目的とした日本語データセットを新たに作成し，LLM に対する学習を行った．さらに，学習前後のモデル出力を代表的な LLM の 1 つである GPT-5 を評価器として用いて比較し，文体別の理解性能の変化を定量的に分析した．その結果，作成したデータセットによる学習が話し言葉および書き言葉の理解向上に寄与することを確認した．

1 はじめに

近年，大規模言語モデル（LLM）は，日本語の質問応答や文章生成において高い性能を示しており，実社会での利用が急速に進んでいる．一方で，話し言葉と書き言葉のような文体差に着目すると，モデルの理解挙動が人間の期待と必ずしも一致しない場合がある．特に話し言葉においては，砕けた表現や会話的文脈を適切に捉えられない事例が見られ，文体差を考慮した理解性能の向上が課題となっている．

このような背景のもと，本研究では，話し言葉および書き言葉の理解向上を目的として，同一の質問に対して異なる文体の回答を対応付けた日本語データセットを新たに構築する．さらに，構築したデータセットを用いて複数の日本語対応 LLM に対して追加学習を行い，文体別の理解性能がどのように変化するかを分析する．

評価には，高性能な LLM を評価器として用い，学習前後のモデル出力を同一基準で比較することで，話し言葉および書き言葉の理解性能の変化を定量的に評価する．これにより，文体差を明示的に考慮したデータセット構築が，日本語 LLM の理解性能向上に有効であるかを検証する．

2 関連研究

2.1 文体差を扱う研究

話し言葉と書き言葉のような文体差は，自然言語処理において古くから指摘されてきた課題であり，特に日本語においては，口語的表現と書面語的表現の差が大きいことが知られている［1］．これまで，文体分類や文体制御に関する研究が行われており，文体情報を明示的に扱うことで，生成文の表現を制御する手法や，文体差を考慮した処理の有効性が報告されている［2］．一方で，これらの研究の多くは，生成される文の文体制御や分類精度に主眼を置いており，LLM が話し言葉や書き言葉といった文体をどの程度「理解」しているかを，同一の質問条件下で比較・分析した研究は限定的である．本研究は，この点に着目し，文体差を明示的に対応付けたデータセットを用いて，LLM の文体別理解性能を分析する点に特徴がある．

2.2 日本語 LLM 向けデータセット

日本語 LLM の性能向上を目的として，質問応答や対話を中心とした多様な日本語データセットがこれまでに構築されてきた［3,4］．これらのデータセットは，モデルの汎用的な生成能力を高める上で重要な役割を果たしている．

しかし，多くの既存データセットでは，話し言葉と書き言葉が混在している場合や，文体差が明示的に区別されていない場合が多い．同一の質問に対して，話し言葉と書き言葉という異なる文体の回答を

体系的に対応付けたデータセットは限定的であり，文体差と理解性能の関係を直接的に分析することは困難であった．本研究では，文体差を明確に区別した構成を持つ日本語データセットを構築することで，既存研究とは異なる観点から文体理解性能の分析を試みる．

2.3 自動評価および LLM を用いた評価

LLM の評価において，人手評価は信頼性が高い一方で，評価コストや一貫性の確保が課題とされてきた．このため，自動評価指標や分類器を用いた評価手法が広く検討されている［5］．近年では，高性能な LLM を評価器として用いるメタ評価の手法が提案されており，人手評価に近い判断を自動的に行える可能性が示されている［6］．

本研究においても，この流れに基づき，高性能な LLM を評価器として用いることで，話し言葉および書き言葉の理解性能を同一基準で評価する．これにより，文体差に着目した学習前後の性能変化を，大規模かつ一貫した条件下で比較することを可能としている．

3 データセットの構築

本章では，話し言葉および書き言葉の理解向上を目的として構築した，日本語データセットの作成手法について述べる．本研究では，最終的な学習データセットの設計に先立ち，文体データの構成や品質が理解性能に与える影響を確認するための予備的な検討を行い，その知見を踏まえて最終データセットを構成した．

3.1 質問文の作成

質問文の生成には openai/gpt-oss-20b を用いた．生成した質問文は，本研究を通して固定されており，学習データおよびテストデータの作成過程において変更していない．これにより，文体理解性能の変化が，質問内容ではなく回答文の文体表現や学習データ構成に起因するものであることを明確にしている．質問文生成に用いたシステムプロンプトの詳細は，付録 A に示す．

3.2 回答文の作成

質問文に対する回答文は，話し言葉および書き言葉で異なる文体条件を与えて生成した．話し言葉の回答生成には，砕けた表現や会話的文脈を促すシステムプロンプトを用い，自然な話し言葉表現が得られるよう設計した．一方，書き言葉の回答生成には，丁寧で論理的な文体（です・ます調）を指示するシステムプロンプトを用い，説明的で構造化された文章となるよう制御した．それぞれの生成に用いたプロンプトの詳細は，付録 B および付録 C に示す．

3.3 予備的検討とデータ構成の方針決定

最終的なデータセット設計に先立ち，複数のデータ構成を用いた予備的な学習および評価を行った．その結果，書き言葉については学習前の段階から高い理解性能が確認される一方で，話し言葉についてはデータ量や表現の自然さが理解性能に大きく影響することが示唆された．

この知見に基づき，本研究では，話し言葉の理解向上を主目的として，話し言葉データを重点的に拡充した学習データセットを構成する方針とした．一方で，書き言葉については，既存の高い理解性能を維持できるよう，文体の一貫性を保った形でデータを構成した．

3.4 最終学習データセット

以上の検討を踏まえ，本研究の最終実験では，話し言葉 2,000 件，書き言葉 1,000 件からなる学習データセットを構築した．各データは，同一の質問に対して話し言葉または書き言葉の回答が対応付けられており，文体差を明示的に扱える構成となっている．この最終データセットを用いて，各対象モデルに対する追加学習を実施した．

3.5 テストデータ

評価に用いるテストデータは，学習データとは独立に生成した質問文から構成される．質問数は 500 問とし，日常会話，教育，歴史，医学，時事，科学，経済，ライフスタイル，SF，スポーツの計 10 ジャンルを対象とした．これにより，特定分野に依存しない文体理解性能を評価可能な構成としている．テストデータに対する回答生成および評価方法については，次章で述べる．

4 学習および評価方法

本章では，前章で述べた最終学習データセットを用いた LLM の追加学習方法および，学習前後の理解性能を評価する手法について述べる．

4.1 学習対象モデル

本研究では，日本語対応の LLM として，tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5，Qwen/Qwen3-32B，abeja/ABEJA-Qwen2.5-32b-Japanese-v0.1 の 3 種類を対象とした．これらのモデルはいずれも日本語生成能力を有しており，話し言葉および書き言葉の理解性能を比較する対象として適していると判断した．

4.2 学習方法

前章で構築した最終学習データセットを用い，各対象モデルに対して追加学習を実施した．学習では，同一の質問文に対して話し言葉または書き言葉の回答が与えられる構成とし，文体差を明示的に含むデータに基づく学習を行った．本研究では，学習データの構成を固定し，学習前後のモデル出力を比較することで，文体理解性能の変化を分析する．

4.3 評価方法

学習前後のモデルの理解性能を評価するため，3.5 節で述べたテストデータを用いた．評価は，学習前モデルおよび追加学習後モデルに対して，同一のテストデータおよび評価基準を用いて実施した．評価には，高性能な LLM を評価器として用い，モデル出力が「話し言葉」「書き言葉」「どちらともいえない」のいずれに該当するかを自動的に判定した．評価に用いたシステムプロンプトおよび出力形式は固定し，モデル間および学習前後で一貫した条件下で評価を行った．評価に用いたプロンプトの詳細は，付録 D に示す．

4.4 評価指標

本研究では，生成された回答が，意図した文体として正しく判定された割合を理解性能の指標として用いた．具体的には，話し言葉の回答が「話し言葉」と判定された割合，および書き言葉の回答が「書き言葉」と判定された割合を算出し，学習前後で比較した．本指標により，文体差を考慮した学習が，理解性能に与える影響を定量的に評価する．

4.5 評価手法の妥当性

LLM を評価器として用いることで，人手評価と比較して一貫した基準で大量の評価を実施できる利点がある．一方で，評価結果が単一の評価器に依存する点は，本手法の制約として挙げられる．本研究では，学習前後の相対的な性能変化に着目することを目的としており，この目的に対して本評価手法は妥当であると判断した．なお，一部の条件において見られる性能の変動については，統計誤差の範囲内に収まっていることを事前に確認している．

5 実験結果

本章では，前章で述べた方法に基づき実施した追加学習および評価の結果について述べる．評価は，学習前モデルおよび最終学習後モデルに対して，同一のテストデータ（500 問）を用いて行った．

5.1 話し言葉における理解性能

表 1 に，話し言葉における理解性能を示す．表中の数値は，500 問のテストデータに対して，生成された回答が「話し言葉」と正しく判定された割合（%）である．

表 1　話し言葉における理解性能（500 問）

学習段階	Swallow	Qwen3	ABEJA-Qwen
学習前	87.6	99.8	88.6
学習後	97.8	99.0	97.8

表 1 より，Swallow および ABEJA-Qwen では，学習後に話し言葉として正しく判定される割合が大きく向上していることが分かる．一方，学習前の段階から高い性能を示していた Qwen3 については，学習後も同程度の性能が維持されている．これらの結果は，話し言葉データを重点的に含む学習データセットを用いた追加学習が，話し言葉理解性能の向上に有効であることを示唆している．

5.2 書き言葉における理解性能

表 2 に，書き言葉における理解性能を示す．表中の数値は，500 問のテストデータに対して，生成された回答が「書き言葉」と正しく判定された割合（%）である．

表 2　書き言葉における理解性能（500 問）

学習段階	Swallow	Qwen3	ABEJA-Qwen
学習前	100.0	100.0	100.0

学習後	99.8	100.0	100.0

表 2 より，書き言葉については，いずれのモデルにおいても学習前の段階から高い理解性能が確認され，追加学習後においても顕著な性能低下は見られない．一部の条件においてわずかな性能変動が確認されるものの，これらは統計誤差の範囲内に収まっていることを確認しており，実質的な性能劣化は生じていないと判断できる．

以上の結果から，本研究で構築した文体に特化したデータセットを用いた学習は，話し言葉の理解性能を向上させつつ，もともと高い性能を有していた書き言葉の理解性能を維持できることが示された．

6 考察

本章では，前章で示した実験結果に基づき，話し言葉および書き言葉の理解性能の変化について考察する．特に，文体差を明示的に考慮した学習データ構築が，理解性能に与えた影響について議論する．

6.1 話し言葉理解性能が向上した要因

実験結果より，話し言葉については，学習前と比較して学習後に理解性能が向上する傾向が確認された．特に，学習前の段階で話し言葉理解が十分でなかったモデルにおいて，顕著な改善が見られた．これは，話し言葉の表現や文体的特徴を重点的に含むデータを用いた学習により，話し言葉特有の表現を適切に捉える能力が向上したためであると考えられる．

また，同一の質問に対して話し言葉と書き言葉の回答を明確に区別した構成を採用したことにより，モデルが文体差を意識した形で学習できた点も，話し言葉理解性能の向上に寄与した可能性がある．

6.2 書き言葉理解性能が維持された理由

書き言葉については，学習前の段階からいずれのモデルも高い理解性能を示しており，追加学習後においても大きな性能低下は確認されなかった．これは，本研究で構築したデータセットが，話し言葉の理解向上を目的としつつも，書き言葉の文体的特徴や構造を損なわないよう設計されていたためであると考えられる．

一部の条件において，ごくわずかな性能変動が見られたが，これらは統計誤差の範囲内に収まっていることを確認しており，実質的な性能劣化は生じていないと判断できる．本研究では，各評価結果を 500 問に対する割合として算出しており，数問程度の差は評価誤差として生じ得ることを事前に確認している．このため，観測されたわずかな性能変動は，統計的なばらつきの範囲内であると判断できる．

6.3 データセット構築方法の有効性と制約

以上の結果から，文体差を明示的に区別したデータセットを構築し，話し言葉データを重点的に含めた学習を行う手法は，日本語 LLM の話し言葉理解性能向上に有効であることが示唆された．一方で，本研究では，評価に単一の LLM を評価器として用いており，評価結果が特定モデルの判断に依存する点は制約として挙げられる．

また，本研究の評価指標は，文体理解に焦点を当てたものであり，回答内容の正確性や詳細な意味理解については直接的には評価していない．今後は，人手評価との比較や，内容理解を含めた多角的な評価手法を検討する必要がある．

7 まとめ

本研究では，話し言葉および書き言葉の理解向上を目的として，同一の質問に対して異なる文体の回答を対応付けた日本語データセットを構築し，日本語対応 LLM に対する追加学習および評価を行った．特に，話し言葉理解に着目し，話し言葉データを重点的に含む学習データセットを設計した点に本研究の特徴がある．

実験の結果，提案データセットを用いた学習により，話し言葉における理解性能が向上する一方で，学習前から高い性能を示していた書き言葉の理解性能は低下することなく維持されることを確認した．一部の条件において見られた性能変動についても，統計誤差の範囲内に収まっており，実質的な性能劣化は生じていないと判断できる．

以上より，文体差を明示的に考慮したデータセット構築と学習手法は，日本語 LLM の話し言葉理解性能を改善しつつ，書き言葉理解性能を維持する上で有効であることが示唆された．今後の課題としては，人手評価との比較による評価妥当性の検証や，文体理解と内容理解の関係を考慮した，より多角的な評価手法の検討が挙げられる．

謝辞

本研究は JSPS 科研費(JP23K17512) ，2023 年度 IU-REAL 異分野融合・新分野 創出プログラム・スタートアップ (IU-REAL23p03)の助成を受けたものである.

参考文献

[1] 田中牧郎.
話し言葉と書き言葉の差異.
日本語学，23(4)，pp. 4‐13，明治書院，2004.
[2] Fujiwara, Y., Takamura, H., and Okumura, M.
Stylistic Variation in Japanese Texts and Its Control.
Proceedings of the Pacific Asia Conference on Language, Information and Computation (PACLIC),
pp. 1–10, 2013.
[3] Kobayashi, S., Inoue, N., and Yamada, I.
Japanese Dialogue Corpus for Conversational Agents.
Proceedings of the 12th Language Resources and Evaluation Conference (LREC),
pp. 682–689, 2020.
[4] Yamada, I., Asai, A., Shindo, H., Takeda, H., and Matsumoto, Y.
Large-scale Japanese Text Corpus for Language Models.
Proceedings of the 2020 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT),
pp. 4133–4142, 2020.
[5] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J.
BLEU: a Method for Automatic Evaluation of Machine Translation.
Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL),
pp. 311–318, 2002.
[6] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhang, J., Li, Z., Lin, Z., and Stoica, I.
Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.
arXiv preprint arXiv:2306.05685, 2023.

A 質問文生成プロンプト

本研究における質問文生成には，以下のシステムプロンプトを用いた．

【System Prompt】

あなたは「質問文」を生成するアシスタントです。

あなたの役割は、指定されたジャンルやテーマに基づき、AI が一般的な知識と論理的推論によって回答できる質問文を作成することです。

■ 言語に関する制約

・出力する質問文は、すべて自然な日本語で書いてください。

・日本語以外の言語での質問文は生成しないでください。

■ 生成方針

・教科書や一般書に掲載されるような、知識・概念・思考を問う質問を作成してください。

・特定の個人情報に依存する質問は作成しないでください。

・質問文のみを番号付きリストで出力してください。

■ 出力形式

・質問文は 1〜3 文とする。

B 話し言葉生成プロンプト

話し言葉の回答生成には，以下のシステムプロンプトを用いた．

【System Prompt】

あなたは友達のように自然な話し言葉で答えるアシスタントです。

砕けた日本語で、会話っぽく回答してください。

C 書き言葉生成プロンプト

書き言葉の回答生成には，以下のシステムプロンプトを用いた．

【System Prompt】

あなたは学術的かつ明確な書き言葉で回答するアシスタントです。

丁寧で論理的な文体（です・ます調）で、段落構成を意識して回答してください。

D GPT-5 による評価プロンプト

学習前後のモデル出力の評価には，GPT-5 を評価器として用い，以下のシステムプロンプトを固定して使用した．

【System Prompt】

あなたは文章スタイル判定器です。

次の文章が「話し言葉」「書き言葉」「どちらともいえない」のどれに該当するかを 1 つだけ答えてください。

【基準】

・砕けた表現や会話的な文脈が多い → 話し言葉

・丁寧で論理的，説明的な文体 → 書き言葉

・判定が困難な場合 → どちらともいえない

【出力形式】

話し言葉

書き言葉

どちらともいえない