プロンプトの言語による数値時系列解釈能力の変化

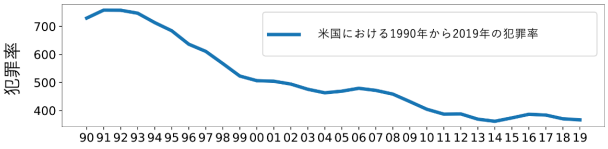
新井 深月 [1] 石垣 達也 [2] 宮尾 祐介 [2,3] 高村 大也 [2] 小林 一郎 [1,2]
[1] お茶の水女子大学 [2] 産業技術総合研究所 [3] 東京大学
g2120503@is.ocha.ac.jp ishigaki.tatsuya@aist.go.jp
yusuke@is.s.u-tokyo.ac.jp takamura.hiroya@aist.go.jp koba@is.ocha.ac.jp

概要

本研究では，大規模言語モデル（LLM）の数値時系列解釈能力を測る 16 の評価タスクにおいて，プロンプトの記述言語の影響を検証する．数値時系列は，マルチモーダル言語生成や時系列予測などの多くの問題で重要な入力情報となっている．従来，LLM の数値理解能力を評価するタスクは，非系列データの算術演算や数値を含む表の論理推論が中心であったが，数値時系列に特化した評価タスクの開発が進んでいる．そこで本研究では，1) イベント検出，2) 計算，3) 比較の 3 つのカテゴリの評価タスクに対し，プロンプトの言語（日本語と英語）の違いによる性能差について考察する．実験より，タスクによってプロンプトの言語選択が数値解釈能力に影響を与えることが明らかとなった．

1 はじめに

数値時系列（データ）は時系列予測 [1]，言語生成 [2, 3] など多くのタスクで重要な入力情報となる．数値の正確な解釈は下流タスクの性能向上に不可欠であり，例えば図 1 のように数値時系列から説明テキストを生成するタスク [4] を正確に解くためには，数値時系列から最小の値と対応する年を抽出する必要がある．しかし，LLM の数値時系列解釈能力は不明確であり，従来の評価は算術演算や表に対する論理推論が中心であったが，最近では数値時系列に特化した評価タスクの開発が進められている [5]．本研究では，環境測定データのような数値時系列における，LLM の数値解釈能力について評価を行う．さらにプロンプトの言語（日本語と英語）の違いが性能に与える影響にも焦点を当てる．具体的には，数値時系列解釈タスクにおける日本語と英語のプロンプトによるモデルの性能の違いを調査し，言語の特性が LLM の数値解釈能力にどのように影響するかを評価する．



数値時系列データからの言語生成タスクの例:
*この統計は1990年以降の米国における犯罪率の変化を示しており，2019年には100,000人あたり**366人と最小の犯罪率**を示した.*

図 1 数値時系列から生成される説明テキストの例

2 関連研究

LLM の数値解釈能力に関する分析は現在も進行中であり，特に数学的な推論能力や計算能力に注目した研究が多い．例えば，Ahn ら [6] は，数学問題に対する LLM の解決能力を調査し，Li ら [7] は，LLM が数学的知識を理解しているのか，ショートカットに頼っているのかを検証している．一方で，数値データが「時系列」として与えられた場合の LLM の解釈能力に関する研究は少なく，特に時系列データを用いたタスクの評価は十分でない．また，使用するプロンプトの言語（日本語と英語）による性能差についての研究も限られている．

本研究では，LLM の数値時系列解釈能力に対する影響を，モデルの学習背景から分析することができる．特に，プロンプトの言語による性能差が高い言語能力に起因するのか，あるいは個別タスクに対する継続的な事前学習の効果によるものなのかを明らかにすることができる．

3 評価タスク

本研究では，数値時系列データにおける解釈能力を評価するために，3 つのタスクカテゴリを設定した．具体的なタスクは以下の通りである．

3.1 イベント検出 (10 タスク)

数値時系列データ内で発生する特定の数値的イベントを検出する能力を評価する．まず，**最大値/最小**

値の検出では，データ全体における極値を特定し，全体のスケールで特徴点の識別能力を評価する．次に**部分最大値/部分最小値の検出**では，特定の区間内での極値を検出し，局所的な特徴点の識別能力を評価する．さらに，**最大値/最小値の時刻の特定**では，最大値や最小値が発生した時刻を特定し，時間的変動の把握能力を検証する．また，**ピーク点/ディップ点の検出**では，転換点を特定し，重要な変動を捉える能力を評価する．最後に，**超過点/未満点の検出**では，閾値を超えた点や下回る点を特定し，異常点の識別能力を評価する．

3.2 計算 (4 タスク)

数値時系列データにおける基本的な計算能力を評価する．具体的には，**平均値**や**累積和**といった基本的な統計量を計算する能力を検証する．さらに，**部分平均値/部分累積和**についても検証する．これらはそれぞれ全体や局所的なデータの傾向を理解するための基盤となるタスクである．

3.3 比較 (2 タスク)

異なる時刻間での値の比較能力を評価する．タスクとして，**差**と**大小比較**の 2 つを含む．差では異なる時刻間の変化を評価し，データの変動を適切に理解できるかを検証する．大小比較では，2 つの異なる時刻における値の大小関係を正確に解釈できるかを評価する．

4 実験

4.1 データセット

実験には Kawarada ら [8, 3] により前処理された Chart-to-text データセット [4] を使用する．犯罪率，死亡者数，国家債務などに関する 2,360 個の数値時系列が含まれている．

4.2 比較する LLM

使用する LLM は以下の通りである (括弧内に本稿での略称を記す)．：

–**API ベースの LLM**:
GPT-3.5-turbo(3.5) , GPT-4(4) , GPT-4o(4o) , GPT-4o-mini(mini)

–**オープンソース LLM**:
Llama-3.1-8B(L3-8B) , Llama-3.1-Swallow-8B(S3-8B), Gemma2-9B(G2-9B)

4.3 計算機環境

実験には，GPU メモリ 40GB の Nvidia A30 を 8 台使用した．推論は Python3 と Hugging Face の `transformers` ライブラリ[1)]で行った．

4.4 評価指標

タスクの特性に応じて，以下の評価指標を使用：

- **正解率**：単一の値が求められるタスクに対して，出力と正解が一致した割合を評価．計算タスクに関して，小数点以下については 5 %の誤差範囲を許容．
- **F1 スコア**：複数の値が正解値となるタスクに対して使用する．モデルが検出したイベントの正確性（適合率）と検出漏れ（再現率）をバランスよく評価するために F1 スコアを用いる．

4.5 出力形式エラーの軽減手法

予備実験で，出力形式の不一致（例えば，数式形式やプログラムコード形式での出力）が観測された．これを改善するために，以下の手法を採用する：

- 少数ショット学習
- 「数値のみで答えてください」という指示をプロンプトに加える

この手法により形式エラーが全体的に減少し，特に計算タスクでは顕著に効果が確認された [5].

また，評価時に出力の冒頭の数字を LLM の回答として採用しているため，この評価方法が結果に影響を与えると考えられる（表 1, 表 5）．形式エラー率が高いタスクは目視で出力を確認し，最も回答が出現する可能性の高い位置の数値を採用することにした．次節以降ではこの軽減手法と評価手法を導入した実験結果を報告する．

5 結果および考察

本研究では，LLM における日本語および英語プロンプトの性能差について評価を行う．

5.1 言語による比較

Swallow モデルでは，日本語の継続学習の効果が顕著に確認され，日本語プロンプトが高い精度を示した．しかし，英語プロンプトとの性能差が課題と

1） https://huggingface.co/docs/transformers/index

表 1　出力形式の例（冒頭と末尾の比較）

回答位置	出力例
冒頭	"**332.5** • Created September 15, 2015 • Files Included • Post your question"
末尾	"3569.97 + 569.02 = 4138.99 / 4138.99 + 552.23 = 4691.22 / 4691.22 + 561.67 = **5252**"

表 2　イベント検出タスクの正解率および F1 スコア.

タスク	言語	3.5	4	4o	mini	L2-7B	L3-8B	S3-8B	G2-9B
正解率									
最大値	日本語	0.96	**1.00**	**1.00**	0.98	0.64	0.83	0.86	0.95
	英語	0.98	**1.00**	**1.00**	0.98	0.65	0.83	0.25	0.74
最小値	日本語	0.96	**1.00**	**1.00**	0.97	0.63	0.76	0.82	0.90
	英語	0.93	**1.00**	**1.00**	0.98	0.63	0.82	0.34	0.64
部分最大値	日本語	0.12	0.87	**0.92**	0.48	0.22	0.33	0.32	0.29
	英語	0.36	0.63	**0.76**	0.75	0.33	0.35	0.10	0.45
部分最小値	日本語	0.05	0.75	**0.83**	0.23	0.21	0.19	0.18	0.16
	英語	0.11	**0.94**	0.90	0.82	0.50	0.59	0.10	0.15
最大値の時刻	日本語	0.21	0.73	**0.83**	0.47	0.06	0.08	0.04	0.23
	英語	0.15	0.74	**0.84**	0.42	0.12	0.10	0.03	0.18
最小値の時刻	日本語	0.16	**0.72**	0.68	0.29	0.06	0.15	0.05	0.17
	英語	0.13	0.68	**0.81**	0.35	0.11	0.18	0.05	0.19
F1 スコア									
ピーク点	日本語	0.25	0.50	**0.60**	0.30	0.09	0.20	0.19	0.20
	英語	0.28	0.50	**0.63**	0.25	0.09	0.11	0.05	0.23
ディップ点	日本語	0.22	0.40	**0.42**	0.27	0.07	0.15	0.16	0.13
	英語	0.23	0.51	**0.60**	0.10	0.09	0.12	0.03	0.15
超過点	日本語	0.21	0.23	**0.38**	0.29	0.08	0.16	0.26	0.11
	英語	0.21	0.38	**0.67**	0.39	0.12	0.18	0.05	0.20
未満点	日本語	0.26	0.30	**0.65**	0.46	0.08	0.42	0.54	0.25
	英語	0.28	0.28	**0.67**	0.43	0.18	0.30	0.21	0.35

して挙げられる．特に累積和においては日本語プロンプトに比べて英語プロンプトの性能が大幅に低下する（日本語プロンプトでは 0.50 であるが英語プロンプトでは 0.02 まで落ち込む）課題が見られた．

一方で，GPT 系モデルでは全体的に英語プロンプトの方が優勢である一方，計算タスクでは日本語プロンプトが高い精度を示した．

5.2　タスク別の性能差

イベント検出タスクにおいて，GPT および Llama モデルは英語プロンプトで高い精度を示し，Swallow モデルは日本語プロンプトの方が優位であった．計算タスクでは，GPT および Swallow モデルが日本語プロンプトで優れた性能を発揮したが，比較タスク全般では日本語プロンプトの方が高い精度を示す傾向が見られた．

これらの結果は，モデルの設計や学習データの違いがタスクごとの言語依存性に大きく影響を与えていることを示している．

5.3　プロンプト設計の改善方向性

一部のモデルでは数値解釈に関する誤答が見られた．特に GPT モデルでは英語プロンプトで年号の数値を計算対象に含める誤りが確認されている（表 5）．

このような誤答を防ぐためにはプロンプトの指示内容を明確化し，モデルがタスクの意図を正確に理解できるよう設計を改善する必要がある．特に，タ

表 3　計算タスクの正解率 (±5% の許容範囲)

タスク	言語	3.5	4	4o	mini	L2-7B	L3-8B	S3-8B	G2-9B
平均値	日本語	0.66	**0.83**	0.80	0.60	0.34	0.37	0.47	0.57
	英語	0.50	0.37	0.47	**0.53**	0.21	0.20	0.15	0.36
累積和	日本語	0.12	**0.57**	**0.57**	0.20	0.10	0.20	0.50	0.48
	英語	0.12	0.10	0.10	0.06	0.04	0.09	0.02	**0.17**
部分平均値	日本語	0.79	0.73	**0.98**	0.83	0.12	0.13	0.77	0.02
	英語	0.34	0.16	0.19	**0.70**	0.60	0.53	0.04	0.36
部分累積和	日本語	0.36	0.93	**0.97**	0.59	0.10	0.10	0.10	0.16
	英語	0.10	0.05	0.04	0.06	0.10	**0.80**	0.05	0.36

表 4　比較タスクの正解率.

タスク	言語	3.5	4	4o	mini	L2-7B	L3-8B	S3-8B	G2-9B
正解率									
大小比較	日本語	0.93	0.98	**0.99**	0.95	0.44	0.64	0.92	0.45
	英語	0.54	**0.99**	0.94	0.93	0.43	0.47	0.57	0.53
正解率 (±5% の許容範囲)									
差	日本語	0.26	**0.69**	0.67	0.58	0.17	0.51	0.52	0.30
	英語	0.17	**0.63**	0.40	0.47	0.18	0.50	0.08	0.41

スク指示に具体例や文脈を盛り込むことで，LLM の数値解釈能力を最大限に引き出すことが期待される．

6　おわりに

LLM の数値時系列データに対する解釈能力を評価するため，「イベント検出」「計算」「比較」の 3 つのカテゴリに分類される 16 の評価タスクを使用して，性能を比較した．

その結果，日本語の継続学習が施されたモデルでは日本語プロンプトでの精度が顕著に高く，各モデルが比較的得意とする言語で性能が向上する傾向が確認された．プロンプト言語やタスクの種類に応じて，LLM の性能に差異が見られた．このことから，LLM の数値解釈能力には言語依存性が存在することが示唆される．加えて，結果から各モデルが持つタスクに対する独自の特徴を把握することができる．

今後の研究では，学習背景が数値解釈能力にどのように起因しているのかを明らかにすることが重要であると考えられる．特に，LLM の学習過程におけるデータの性質やモデルのアーキテクチャが，数値解釈タスクに対する理解にどのように影響を与えるのかを深く掘り下げることが求められる．この探求によって，数値解釈能力の向上に向けた新たなアプローチが見つかるとともに，今後のモデル改善に向けた有益な知見を得ることが期待される．

表 5　誤答例（累積和計算タスク）

入力データ（年）	2020, 2021, 2022, 2023
入力データ（値）	14.8, 14.7, 14.5, 15.3
正解値	59.3
LLM 出力例	8089.3

謝辞

この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の助成事業 (日本語 NP20006) による支援および産総研政策予算プロジェクト「フィジカル領域の生成 AI 基盤モデルに関する研究開発」の結果得られたものである．

参考文献

[1] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In **The Twelfth International Conference on Learning Representations**, 2024.

[2] Soichiro Murakami, Akihiko Watanabe, Akira Miyazawa, Keiichi Goshima, Toshihiko Yanase, Hiroya Takamura, and Yusuke Miyao. Learning to generate market comments from stock prices. In Regina Barzilay and Min-Yen Kan, editors, **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1374–1384, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[3] Masayuki Kawarada, Tatsuya Ishigaki, and Hiroya Takamura. Prompting for numerical sequences: A case study on market comment generation. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 13190–13200, Torino, Italia, May 2024. ELRA and ICCL.

[4] Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4005–4023, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[5] 深月新井, 達也石垣, 祐介宮尾, 大也高村, 一郎小林. 大規模言語モデルの数値時系列解釈能力の検証. Technical Report 41, お茶の水女子大学／産業技術総合研究所, 産業技術総合研究所, 東京大学／産業技術総合研究所, 産業技術総合研究所, お茶の水女子大学／産業技術総合研究所, dec 2024.

[6] Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. In Neele Falk, Sara Papi, and Mike Zhang, editors, **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop**, pp. 225–237, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

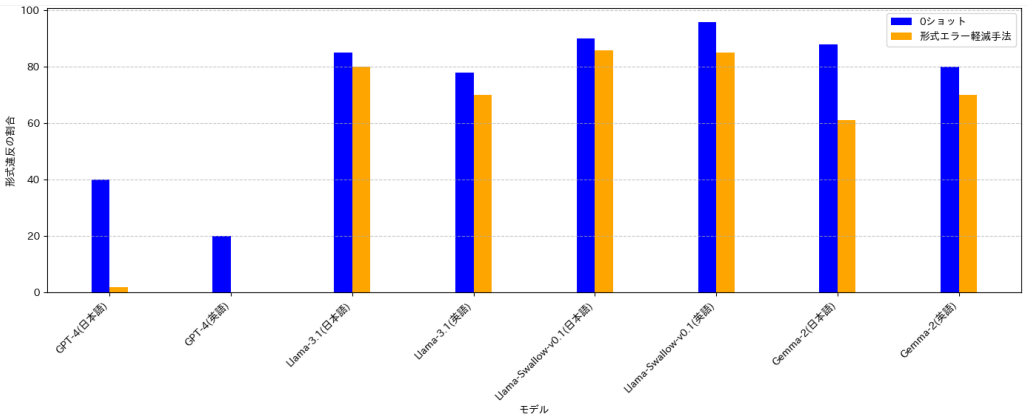
[7] Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. GSM-plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2961–2984, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[8] Masayuki Kawarada, Tatsuya Ishigaki, Goran Topić, and Hiroya Takamura. Demonstration selection strategies for numerical time series data-to-text. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 7378–7392, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

表 6　プロンプト例一覧

タスク	言語	プロンプト例
イベント検出（最大値）	日本語	次の時系列データから最大値を検出してください．数値のみで答えてください． 例 1: 年: 2001, 2002, 2003, 2004, 2005．値: 10, 20, 30, 40, 50．最大値: 50. ： 例 4: 年: {years_str}．値: {values_str}．最大値:
	英語	Which is the value that is the maximum value? Please respond with the maximum value only as a number. Example 1: Years: 2001, 2002, 2003, 2004, 2005. Values: 10, 20, 30, 40, 50. Maximum value: 50. : Example 4: Years: {years_str}. Values: {values_str}. Maximum value:
計算（平均値）	日本語	次の時系列データから平均値を計算してください．数値のみで答えてください． 例 1: 年: 2001, 2002, 2003, 2004, 2005．値: 10, 20, 30, 40, 50．平均値: 30. ： 例 4: 年: {years_str}．値: {values_str}．平均値:
	英語	Calculate the average value of the following time series data. Please respond with the average value only as a number. Example 1: Years: 2001, 2002, 2003, 2004, 2005. Values: 10, 20, 30, 40, 50. Average value: 30. : Example 4: Years: {years_str}. Values: {values_str}. Average value:
比較（差）	日本語	次の時系列データの 2001 年の値と 2003 年の値の差を計算してください．数値のみで答えてください． 例 1: 年: 2001, 2002, 2003, 2004, 2005．値: 10, 20, 30, 40, 50．差: 20. ： 例 4: 年: {years_str}．値: {values_str}．差:
	英語	Calculate the difference between the value for the year 2001 and the value for the year 2003. Please respond with the difference only as a number. Example 1: Years: 2001, 2002, 2003, 2004, 2005. Values: 10, 20, 30, 40, 50. Difference: 20. : Example 4: Years: {years_str}. Values: {values_str}. Difference:

図 2　累積和計算タスクにおける形式違反の割合



※本研究におけるタスクの中で累積和計算タスクが最も形式違反の割合が高かったため，その結果を図示した．