大規模言語モデルの数値時系列解釈能力の検証

新井 深月　(指導教員：小林 一郎)

1　はじめに

数値時系列データは，時系列予測 [1] や言語生成 [2, 3] などの多くのタスクにおいて重要な情報源であり，その正確な解釈は下流タスクの性能向上に寄与する．本研究では，大規模言語モデル（LLM）の数値時系列解釈能力に着目し，その性能を「イベント検出」「計算」「比較」の 3 カテゴリに分類された計 16 の評価タスクを通じて評価を行う．また，プロンプト言語（日本語および英語）がモデル性能に与える影響についても検証する．

2　評価タスク

2.1　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．

2.2　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．

2.3　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．

3　実験

3.1　データセット

Kawarada ら [3] により前処理済みの Chart-to-text データセットに含まれる折れ線グラフを用いる．このデータセットは，犯罪率，死亡者数，国家債務などに関する 2,360 個の数値時系列から成る．

3.2　比較する LLM

- **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)

3.3　評価指標

タスクの特性に応じて以下の指標を用いる：

- **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 %の誤差範囲を許容する．
- **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．

3.4　形式エラー軽減手法

本研究では，自動評価を適切に行うために，出力を意図した形式に統一する必要がある．その対策として，以下の 2 つの手法を採用した．

- 少数ショット学習を用いる
- 「数値のみで答えてください」という指示を加える

4　結果および考察

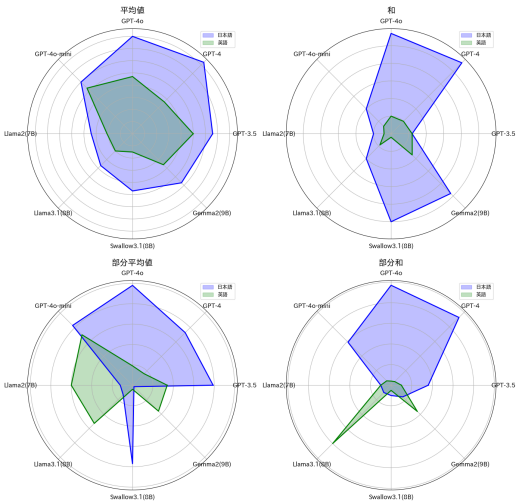

図 1: 計算タスクの正解率（青：日本語；緑：英語）

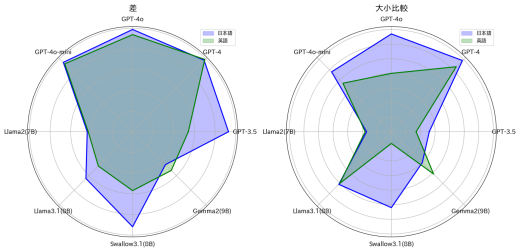

図 2: 比較タスクの正解率（青：日本語；緑：英語）

表 1: プロンプト例（太字部はタスクにより変更する）

タスク	言語	プロンプト例
イベント検出	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。**最大値**:
	英語	Which is the value that is **the maximum value**? **Maximum value**:
計算	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．**平均値**:
	英語	Calculate **the average value** of the following time series data. **Average value**:

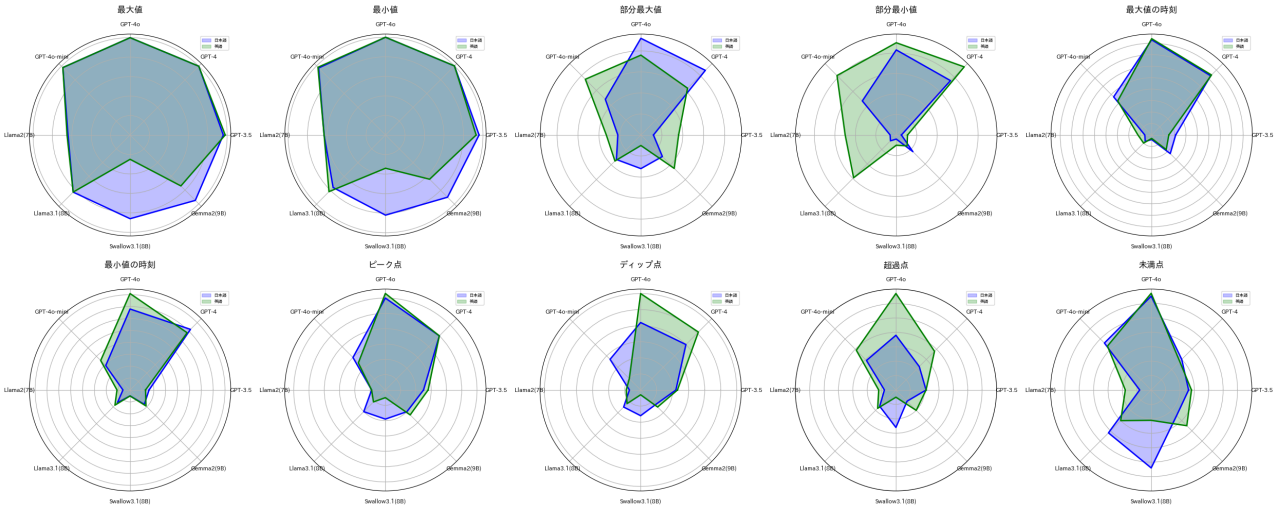

図 3: イベント検出タスクの正解率および F1 スコア（青：日本語；緑：英語）

4.1 タスク別の比較

最大値・最小値以外のタスクについては全体的に精度が良いとは言い難い結果となった．その中で GPT-4 および GPT-4o は全般的に高い精度を示した．一方で範囲を絞った数値検出タスクでは精度の低下が見られた．さらに，GPT 系と Llama，Gemma，Swallow の間では，複数点検出タスクや計算タスクにおいて大きな性能差が見られた．

4.2 プロンプト言語による比較

プロンプト言語による性能の違いも明確にみられた．GPT および Llama，Gemma は英語プロンプトで高い精度を示したのに対し，Swallow は日本語プロンプトの方が精度が高い傾向にあった．この結果からモデルが学習時に使用したデータの言語バイアスに起因すると考えられる．

4.3 考察

本研究の評価タスクにおける精度は，モデルの言語能力に依存する傾向があると考えられる．しかし言語間における性能差が単純な言語理解能力の違いによるものなのか，それとも数理解釈能力に起因するのかについては，さらなる詳細な分析が必要である．本研究ではこの点を十分に解明するには至らなかったため，今後の課題として検討を進める必要がある．

また，複雑なタスクや部分的な検出を行う際，対象とするデータを事前に切り出す適切な前処理が必要であることが示唆された．プロンプト手法やデータ形式の工夫が現存の LLM の数値解釈能力を最大限活用するために必要であると言える．

5 まとめ

本研究では，様々な評価タスクを設定し，LLM の数値解釈能力を評価した．結果として，GPT 系モデルは多様なタスクで一貫して高い精度を示し，特に英語プロンプトでの性能が優れていることが確認されるなどプロンプト言語がモデル性能に与える影響が示された．また，範囲を絞った数値検出タスクでは精度が低下する傾向が観察され，適切な前処理やモデル改良が必要であることが示唆された．

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.

A 付録

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
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
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

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（差は ±5% の許容範囲）

タスク	言語	3.5	4	4o	mini	L2-7B	L3-8B	S3-8B	G2-9B
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
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