LLM を利用した Zero Shot 評判分析の性能調査

佐藤 匠真 [1]　新納 浩幸 [2]
[1] 茨城大学大学院　理工学研究科 [2] 茨城大学大学院理工学研究科情報科学領域
23nm724y@vc.ibaraki.ac.jp　hiroyuki.shinnou.0828@vc.ibaraki.ac.jp

概要

本論文では，LLM を利用した Zero Shot の評判分析における性能調査について報告する．実験では，評判分析のデータとして Webis-CLS-10 データセットを使用し，LLM として GPT-3.5-turbo，GPT-4o，Llama 3.1-Swallow-8B を用いた．結果として，極性判定では，LLM を利用した場合，Zero Shot でも高い精度が得られることが確認された．一方で，livedoor 記事のカテゴリ分類を行う場合には，Zero Shot の精度が低いという結果となった．また，Zero Shot で日本語の Amazon レビューの評判分析を実施する際，LLM にその判定根拠を提示させると精度が低下することが確認された．

1　はじめに

近年，大規模言語モデル（LLM）は，自然言語処理の分野で急速に発展しており，さまざまなタスクにおいて優れた性能を発揮している．特に，GPT-3.5[1] や GPT-4 [2] などの大規模なトランスフォーマーモデルは，従来の教師あり学習の枠を超え，事前学習された知識を活用することで，特定のタスクに対してデータセットを準備せずとも高い精度を達成できる「Zero Shot 学習」が可能となっている．このアプローチは，ラベル付けや学習データの収集が困難な場合においても有効であり，効率的な問題解決手段として注目を集めている．

本研究では，LLM を利用した Zero Shot の評判分析における性能調査について報告する．具体的には，Amazon レビューの評判分析をタスクとして選定し，日本語のレビューに対する評判分析を行う実験を実施した．実験では，GPT-3.5-turbo，GPT-4o，Swallow といった複数の LLM を使用し，レビュー文書のラベルを「ポジティブ」または「ネガティブ」といった 2 値分類および 4 値分類で評価した．また，極性判定以外の文書分類タスクについても性能調査を行った．具体的には，Livedoor ニュース記事の分類をタスクとして実験を行い，評判分析で使用したモデルと同じ LLM を利用した．

実験の結果，極性判定の 2 値分類および 4 値分類において，Zero Shot 評判分析が教師あり学習と同等以上の性能を示すことが確認された．一方で，文書分類タスクでは，Zero Shot 学習の精度が教師あり学習を下回る結果となった．また，LLM を利用した Zero Shot で日本語の Amazon レビューの評判分析を実施する際，LLM にその判定根拠を提示させると精度が低下することが確認された．

2　関連研究

2.1　LLama 3.1 Swallow

Llama 3.1 Swallow[3] は，Meta 社の Llama 3.1 を基盤に，日本語能力を強化した大規模言語モデルである．このモデルは，英語の性能を維持しつつ，日本語の理解・生成能力を向上させることを目的としている．モデルサイズは 8B と 70B が提供されている．Llama 3.1 Swallow は，日本語の理解・生成タスクにおいて高い性能を示している．特に，指示チューニングを施したモデル（Instruct モデル）は，日本語のマルチターン対話能力が向上し，日本語 MT-Bench の平均スコアで 13B 以下の日本語 LLM の中ではトップクラスの成績を収めている．

2.2　Chain of Thought

Wei ら [4] は，言語モデルが複雑な推論タスクを解く際の新しいアプローチとして，Chain of Thought (CoT) Prompting を提案している．この手法は，単に結論を出力するのではなく，問題解決のプロセスを逐次的に記述する形式を言語モデルに促すものである．特に，数学問題や常識推論タスクなど，複数ステップの推論が必要な場面において，CoT が性能を大幅に向上させることを実証した．

また，GPT-3[1] のような大規模言語モデルに CoT プロンプトを適用した際，従来のプロンプト方式と

比較して顕著な精度向上が観察された．また，CoT の効果はモデルサイズに強く依存しており，100B パラメータ以上の大規模モデルで特に効果的であることが示された．

さらに，CoT プロンプトは特定のデータセットやタスクに依存することなく，汎用的に使用可能である点も強調されている．この手法は，推論能力を引き出すだけでなく，プロセスの透明性や説明可能性を向上させる点で注目される．一方で，モデルサイズが小さい場合や，タスクが単純すぎる場合には効果が限定的であるという課題も指摘されている．

また，Kojima ら [5] は Chain of Thought (CoT) を Zero Shot で適用する手法を提案し，大規模言語モデルが複雑な推論タスクを効率的に解けることを示した．この研究では，CoT を活用して問題解決のプロセスを逐次的に展開することで，言語モデルの推論能力を向上させている．

3 LLM に与えるプロンプト

OpenAI が公開している GPT の API を利用する場合や Swallow を利用する場合，2 つのプロンプトを渡すことで生成を行う．以下にその 2 つを示す．

- system: チャットボットの動作や指針を設定するメッセージ．モデルに指示を与える役割．
- user: ユーザーからの入力を表すメッセージ．ユーザーが提供する情報．

実験で使用するプロンプトはタスクに応じて変更しているが，基本的にはタスクを説明する文，ラベルの説明，および判断理由の出力の有無を含めたものとしている．また，LLM が明確な回答を出さない場合があるため，必ず回答するように指示を加えた．

判断理由の出力については，LLM に理由を説明させることで性能が向上する「Chain of Thought」アプローチが日本語の Zero Shot 学習における評判分析において効果を持つかを検証する目的で含めている．一方，user では評判分析を行う文書のみを提示した．

4 実験

LLM に対して学習データを与えずに対話形式でタスクを解くことで，LLM の Zero Shot 学習における精度やタスクの問題について調査を行う．以下の 4 つの条件で調査を実施した．

- 理由出力なし，300 文書の Amazon レビューの評判分析 (2 値分類) (実験 1)
- 理由出力あり，300 文書の Amazon レビューの評判分析 (2 値分類)(実験 2)
- 理由出力なし，自作の 100 文書の Amazon レビューの評判分析 (2 値分類)(実験 3)
- 理由出力なし，300 文書の Amazon レビューの評判分析 (4 値分類)(実験 4)

また，考察においてタスクの違いによる性能の差を分析するために，以下の条件でも追加の実験を行った．

- 理由出力なし，900 文書の livedoor ニュース記事のカテゴリ分類 (実験 5)

4.1 大規模言語モデル

実験には，LLM として GPT-3.5-turbo，GPT-4o，Llama 3.1-Swallow-8B を用いた．GPT-3.5-turbo と GPT-4o はそれぞれ OpenAI の API でモデル名をそれぞれ’gpt-3.5-turbo’，’gpt-4o’ で利用できるモデルを使用した．Llama 3.1-Swallow-8B は Hugging Face 社の Transformers ライブラリから，モデル名’tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.1’ で利用できるモデルを使用した．

また，実験の比較対象として，rinna 社の日本語単言語モデルを使用した．このモデルは，Hugging Face 社の Transformers ライブラリから，モデル名’rinna/japanese-roberta-base’ として利用可能なものを使用した

4.2 実験用データセット

実験には Webis-CLS-10 データセット[1)] を用いた．ラベルはレビューの星の数であり，1 から 5 までの 5 段階評価である．ただし，ラベルが 3 のデータは存在しない．本実験では，ラベルが 1，2 のデータをネガティブ，4，5 のデータをポジティブとして評判分析（2 値分類）を行った．このデータセットには，日本語，英語，ドイツ語，フランス語それぞれに books，dvd，music の 3 つの領域がある．本実験では，すべて日本語の music の領域のデータを用いた．

また，自作のテストデータとして，2024 年 7 月以降の Amazon レビューを使用した．

さらに，livedoor ニュース記事分類の実験では，

1） https://webis.de/data/webis-cls-10.html

NHN Japan 株式会社が運営する「livedoor ニュース」の記事部分のみを使用した．livedoor ニュースのカテゴリは，トピックニュース，Sports Watch，IT ライフハック，家電チャンネル，MOVIE ENTER，独女通信，エスマックス，livedoor HOMME，Peachy の 9 つである．

4.3 実験結果

実験 1 では，日本語の Amazon レビュー 300 文書を用いたテストデータに対して，理由を出力させながら GPT-3.5-turbo，GPT-4o，Swallow を使用してタスクを解いた．ポジティブおよびネガティブ以外の出力があった場合は，不正解として集計した．また，比較対象として日本語の RoBERTa を 1600 文書で学習した場合の正解率も示した．それぞれの正解率を表 1 に示す．日本語の RoBERTa の正解率は 5 回の実験結果の平均値を使用した．また，監視する指標を loss と Accuracy の 2 パターンで評価を行った．

表 1 Amazon レビュー 300 文書の理由出力ありの正解率

モデル	正解率
GPT-3.5-turbo	0.8567
GPT-4o	0.9433
Swallow	0.9
RoBERTa(loss)	0.8933
RoBERTa(Accuracy)	0.8879

実験 2 では，日本語の Amazon レビュー 300 文書を用いたテストデータに対して，理由を出力させずに GPT-3.5-turbo，GPT-4o，Swallow を使用してタスクを解いた．また，比較対象として日本語の RoBERTa を 2000 文書で学習した場合の正解率も示した．それぞれの正解率を表 2 に示す．

表 2 Amazon レビュー 300 文書の理由出力なしの正解率

モデル	正解率
GPT-3.5-turbo	0.90
GPT-4o	0.9567
Swallow	0.9333
RoBERTa(loss)	0.8933
RoBERTa(Accuracy)	0.8879

実験の結果，GPT-3.5-turbo において理由出力ありの場合は教師あり学習よりも精度が劣る結果となった．一方で，理由出力ありの場合の GPT-4o と Swallow，理由出力なしの場合の GPT-3.5-turbo および GPT-4o，Swallow は，Zero-Shot 学習においても教師あり学習を上回る精度を達成した．

また，理由出力を求めた場合，理由出力を求めない場合と比較して精度が低下することが確認された．LLM に理由出力を求める場合，モデルは予測精度に加えて理由生成の品質も同時に最適化する必要があり，この二重の負担が Zero Shot 学習におけるモデルの能力を分散させる原因となった．その結果，日本語の Amazon レビューの評判分析において適切な予測を行うことが難しくなったと考えられる．また，Amazon レビューの評判分析は言語推論が重要ではないタスクであるため，数学的推論のように CoT（Chain of Thought）が有効に機能しなかった可能性がある．

そのため，これ以降の実験ではすべて理由出力を求めずに実施することとした．

Webis-CLS-10 データセットは 2010 年以前のデータであり，LLM の事前学習に含まれている可能性があるため，高い精度を示す結果となった可能性がある．そのため，2024 年 7 月以降の Amazon レビューを使用して，日本語 100 文書のテストデータを作成した．実験 3 では，最新のデータでも精度が高いのかを確認するため，作成した自前のデータセットで評判分析を行った．その結果を表 3 に示す．

表 3 自作の Amazon レビュー 100 文書の理由出力なしの正解率

モデル	正解率
GPT-3.5-turbo	0.93
GPT-4o	0.97
Swallow	0.93

実験 3 の結果，最新のデータでも精度は変わらないことが確認できた．

Zero Shot 学習でもポジネガ判定では高い性能を示したが，より細かいクラス分けを行う 4 値分類でも同様に高い精度を維持できるかを検証する．実験 4 では，実験 1 と 2 で使用したテストデータの星の数を予測する 4 値分類として評判分析を行った．

実験 1 と 2 同様に，日本語の Amazon レビュー 300 文書を用いたテストデータで，理由出力を求めずに GPT-3.5-turbo，GPT-4o，Swallow を使用して解いた．また，比較対象として，同様に日本語の RoBERTa を 2000 文書で学習した場合の正解率を示す．それぞれの正解率を表 4 に示す．

実験 4 の結果，4 値分類では精度が全体的に低下するものの，LLM の Zero Shot 学習でも RoBERTa の教師あり学習と同等の精度を示すことがわかった．

表 4　Amazon レビュー 300 文書の 4 値分類の理由出力なしの正解率

モデル	正解率
GPT-3.5-turbo	0.60
GPT-4o	0.63
Swallow	0.5933
RoBERTa(loss)	0.47
RoBERTa(Accuracy)	0.588

4.4　LLM が出力した判断理由について

LLM が評判分析を間違えた原因を調査するために，実験 2 で評判分析を行う際に出力した判断理由を分析する．

4.4.1　判断不能による間違い

GPT-3.5-turbo が間違えた 43 文書の中で 18 文書は判断不能により間違えてしまっている．判断不能の場合の理由出力の例は以下の通りである．

- 与えられた文書からは具体的な製品や作品の情報が得られないため，内容から評価を判断することが困難です．

このような間違いのすべては，正解ラベルがポジティブでありながら，誤ってネガティブと判断してしまっている．このような判断の原因としては，Zero Shot 学習におけるタスクやデータセットに関する知識不足や，レビュー文書の短さなどが考えられる．

4.4.2　GPT-4o の間違い

GPT-4o では不正解数が 17 文書となった．GPT-4o が間違えた文書は，反語表現や皮肉の表現，または人目で見てもわかりにくいものが多かった．判断不能による不正解はなくなった．

5　考察

極性判定のようにラベル間の境界がある程度わかりやすいタスクではなく，文書分類のようにラベル間の境界が曖昧なタスクでは，LLM の Zero Shot 学習における精度を検証するため，livedoor ニュース記事の文書分類を実験 5 として行った．

実験 5 では，livedoor のニュース記事 900 文書を GPT-3.5-turbo，GPT-4o，Swallow で文書分類した．また，比較対象として日本語の RoBERTa を 540 文書で学習した場合の正解率も示す．それぞれの正解率を表 5 に示す．

表 5　livedoor ニュース記事 900 文書の文書分類の正解率

モデル	正解率
GPT-3.5-turbo	0.5389
GPT-4o	0.7089
Swallow	0.57
RoBERTa(loss)	0.6287
RoBERTa(Accuracy)	0.8529

実験 5 の結果，livedoor ニュース記事の文書分類タスクでは，Zero Shot 学習の LLM よりも教師あり学習の RoBERTa の方が精度が高くなった．この結果から，Zero Shot 学習における LLM は，ラベル間の境界が曖昧なタスクでは RoBERTa の教師あり学習よりも精度が下がることがわかった．

6　おわりに

本研究では，LLM を学習データなしで利用し，Amazon レビューの評判分析および livedoor ニュース記事の文書分類タスクを解かせた．これにより，Zero-Shot 学習の精度比較，出力理由の分析，ならびにタスクの違いによる精度の比較を行った．その結果，GPT-4o や Swallow は Zero Shot 学習でも極性判定において高い精度を示した一方で，文書分類タスクでは精度が低下する傾向が見られた．また，LLM を用いた Zero Shot の日本語の Amazon レビューの評判分析において，CoT（Chain of Thought）を導入すると，精度が低下することが確認された．

今後の課題として，日本語の Zero Shot 学習における LLM の精度を向上させるプロンプトの開発や，LLM が Zero-Shot 学習で得意とするタスクの特定について検討する．

謝辞

本研究は国立国語研究所の共同研究プロジェクト「テキスト読み上げのための読みの曖昧性の分類と読み推定タスクのデータセットの構築」及び JSPS 科研費 23K11212 の助成を受けています.

参考文献

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[2] OpenAI. Gpt-4 technical report. https://openai.com/research/gpt-4, 2023. 2024-12 閲覧.

[3] 産業技術総合研究所岡崎研究室. Llama 3.1 swallow. https://swallow-llm.github.io/llama3.1-swallow.ja.html, 2024. 2024-12 閲覧.

[4] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[5] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023.