日本語小論文におけるルーブリックに対応した評価部分の Zero-shot 予測

成岡 智也 [1]　竹内 孔一 [2]
[1, 2] 岡山大学大学院環境生命自然科学研究科
[1]pak13nxr@s.okayama-u.ac.jp　[2]takeuc-k@okayama-u.ac.jp

概要

本研究では，小論文自動採点を行うにあたり重要である，小論文内の得点に関係する評価部分の予測を行う．ルーブリックの評価部分に対応した日本語小論文の文書内構造の情報を正解データとし，複数のオープンウェイト LLM を用い Zero-shot で評価部分の有無を予測させ，比較を行った．その結果から，従来の小論文自動採点の課題である学習データが大量に必要であること，採点根拠が不明瞭であることを解決するアプローチに繋がる点，及び各オープンウェイト LLM の特性と評価部分の有無についての予測可能性を明らかにした．

1　はじめに

大量の小論文を人間が平等に採点することは困難であり，また採点にかかる労力も大きい．近年，この問題を解決するために，大規模言語モデルを用いた様々な小論文自動採点手法の研究が行われている．しかし，小論文自動採点には，更なる精度の向上や，学習データを大量に必要にしてしまう点など様々な課題が見られる．

そこで，本研究では学習データを大量に必要にしてしまう点に着目し，Zero-shot で小論文を採点するための要素の有無を予測する手法について考える．先行研究では，大規模言語モデルを用いた小論文の評価においてモデルのファインチューニングを行っているもの [1][2][3][4] が存在する．しかし，小論文自動採点の実利用を考えた際，採点対象の小論文の一部が既に採点済である事象は考えにくい．そのため，本研究では一貫して Zero-shot で実験を行う．また，小論文は基本的に採点基準 (ルーブリック) を基準に採点が行われる．ある小論文の得点を決定する際，その小論文内にルーブリックに記載されている評価部分がどの程度含まれるかによって判断される．先行研究では，小論文の構造やルーブリックに着目した研究 [2][3][5][6][7] があるが，ルーブリックの評価部分がアノテーションされたデータを用い，日本語小論文自動採点に Zero-shot でアプローチした例は我々の知る限り見当たらない．このような背景から，本論文ではルーブリックに基づいた評価部分のアノテーションがされた日本語小論文データを正解データとして用い，Zero-shot での小論文内にルーブリックの評価部分が存在するか否かの予測可能性を複数のオープンウェイト LLM で検証・比較し、その有効性を明らかにする．

2　関連研究

本研究と同様に，採点対象データの情報を得ることが実利用面で考えにくいことに着目したもので，Li らの研究 [8] が挙げられる．この研究は，採点対象と異なる設問のエッセイを学習データとして用いることで，様々なエッセイに共通する一般的な特徴を元に採点対象エッセイの評価を行う手法を提案している．実際に文法や論述方法など，異なるエッセイ間である程度評価軸が一致する採点基準についてこの手法は有効であるが，本研究で使用する小論文のような，設問ごとに評価される内容が大きく異なる場合には，この手法で得られた得点が採点基準通りに採点されている確証が得られないため，あまり適していない．

また，本研究で使用しているものと同じ日本語小論文データを用いたもので，水野らの研究 [9] が挙げられる．この研究は，BERT モデルでの小論文自動採点に対し，Attention，Masking Token，Sparse Autoencoder の 3 手法で採点根拠を同定し，採点の信頼性を担保しようとしたものである．しかし，得られた特徴量と得点の間の関連性は不明瞭であると結論付けられている他，本来結果に表れるべきであるルーブリックの評価部分について有意な結果は得ら

表 1　文書構造分析タグの種類

課題名	タグ
global_q1	光，影，格差縮小，格差拡大
global_q2	光，影，具体例
science_q1	実証性の説明，再現性の説明， 客観性の説明
science_q2	自然相手，持続役割，根拠，客観確保， 共通役割
easia_q1	相互依存 (データあり)， 協力・協業の実態，具体例
easia_q2	概略，脱する方法
easia_q3	日本，韓国，中国，協調と対立
criticize_q1	論理的・合理的思考の説明， 目標志向的思考の説明， 内省的・熟慮的思考の説明
criticize_q2	カラーテレビ
criticize_q3	方法

れていない．

3　ルーブリック評価部分がアノテーションされた日本語小論文データ

本研究で使用している小論文データは，言語資源協会から公開されている日本語小論文データ [10] である．この小論文データは，2016 年から 2018 年に開講された講義の受講者の回答データから作成されている．今回利用した小論文データは「グローバリゼーションの光と影」，「自然科学の構成と科学教育」，「東アジア経済の現状」，「批判的思考とエセ科学」の計 4 つのテーマであり，それぞれの講義テーマごとに 3 つの設問が用意されている．各設問ごとに 300 件前後の小論文がある．以降，各講義テーマは「グローバリゼーションの光と影」を global，「自然科学の構成と科学教育」を science，「東アジア経済の現状」を easia，「批判的思考とエセ科学」を criticize と表現する．加えて各設問を q1，q2，q3 と表現し，これらを組み合わせて「グローバリゼーションの光と影」の設問 1 を global_q1，または簡略化し g1 と表現する．先行研究 [11] にて，この各設問に対する各解答に，ルーブリックの評価部分をもとにしたタグを付与するアノテーションが行われている．以降，これらのタグを単に「タグ」と呼ぶ．本研究では，これらのタグのうち一部について取り扱う．本研究で使用するタグの種類は，表 1 の通りとなっている．

これらのタグに加え，一部のタグには「光?」のような形式で?が付随しているものも存在する．これは，その部分の内容が評価部分に相当するが，採点者の確信度が低い場合に付与されている．アノテーションされた小論文は，図 1 のような形式になっている．

図 1 は，設問 global_q1 の小論文の一部である．例えば，1 行目に着目すると，「グローバリゼーションは世界全体の所得格差を縮小する」の部分に「光」「格差縮小」タグが，「国内及び各国間での所得格差を拡大している」の部分に「影」「格差拡大」タグがアノテーションされていることがわかる．本研究では，全 12 設問のうち global_q3，science_q3 を除く計 10 設問の日本語小論文データ及びそれらに対応する講義資料データを使用する．講義資料データは，学生が小論文の答案を作成する際に事前知識として講義された内容をテキストに起こしたものである．

4　実験

本章では，複数の LLM を用いての評価実験及び結果の比較を行うことで，提案手法の有効性を確かめる．以下，各節で実験設定，実験結果，考察を述べる．

4.1　実験設定

以下，実験では Llama-3-ELYZA-JP-8B[1] と gpt-oss-20B[2] の 2 つのモデルを扱う．両モデルとも，十分な日本語能力を有する点，ダウンロード可能でありローカルで推論可能である点，モデルサイズが軽量である点を理由に実験対象とした．一点目は，日本語小論文を扱うため必要不可欠である．二点目は，実利用を考慮した際に，実際の試験答案のデータを外部サーバに渡す行為が実現しにくいこと，日本語小論文採点という特性上入出力トークン総量が膨大であり，効率化及び高精度化が API 利用料金に見合わない可能性があることが考えられるため必要である．本実験では，Llama-3-ELYZA-JP-8B 及び gpt-oss-20B を用い，Zero-shot で各小論文に対し設問に対応するタグが存在するか否かを予測させ，その結果について比較・評価を行う．評価指標には Accuracy，F-1score を用いる．それぞれ，以下の式で表す．

$$P(a,b) = \begin{cases} 1, & \text{if } a = b \\ 0, & \text{if } a \neq b \end{cases} \tag{1}$$

$$\text{Accuracy} = \frac{\sum_{l=1}^{n} P(A_l, B_l)}{n} \tag{2}$$

1）https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B
2）https://huggingface.co/openai/gpt-oss-20b

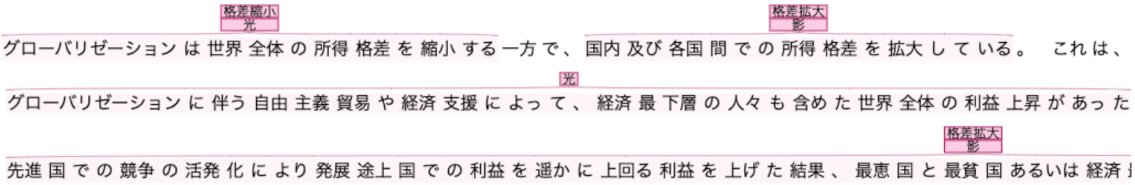

図 1　アノテーションされた小論文データ

$$\text{F-1score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

Accuracy は，式 (1)，(2) によって求められる．n は各設問における小論文の総数であり，A_l，B_l は l 番目の小論文に対してそれぞれタグ有無の正解，タグ有無の予測結果である．Accuracy は 0 から 1 の値を取り，1 に近いほど一致率が高い．F-1score は，それぞれ式 (3) によって求められる．また，タグが付与されていないものの予測に対しての F-1score を F1_0，タグが付与されているものの予測に対しての F-1score を F1_1 とする．推論過程については，図 2 のとおりである．

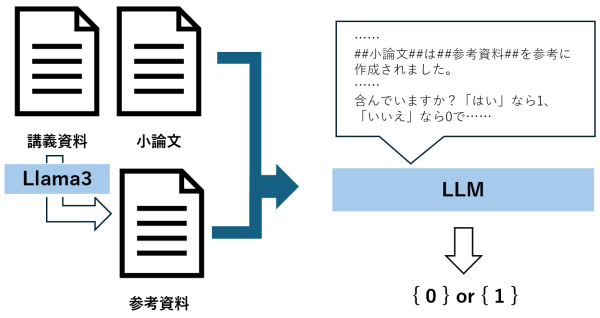

図 2　タグ有無の推論過程

事前準備として，講義資料の情報が必要なタグに関して，講義資料からタグに対応した参考資料を 1 タグにつき 1 つ作成する．参考資料作成は，Llama-3-ELYZA-JP-8B に資料内の該当部分を抽出させることで行った．タグ有無の予測を指示するプロンプトは，参考資料と小論文 1 件を提示し，小論文内に参考資料と同等の内容が含まれているかどうかを推論させるものである．なお，プロンプトに関して両モデルで全く同じものを使用しているが，gpt-oss-20B は事前学習の特性上ハーモニー記法に従う必要があるため，Llama-3-ELYZA-JP-8B のものに role の記述を加えたものとなっている．また回答を確実に二値化するため，参考資料と同等の内容を含んでいるならば 1，含んでいないならば 0 と回答するよう指示している．

4.2　実験結果

以降，前節の設定で行った実験の結果を示す．表 2 に，各タグの有無予測に関しての Accuracy，F1_1，F1_0 を示している．

Llama3-ELYZA-JP-8B でタグ有無予測を行った際の Accuracy の平均値は 0.7309，F1_1 の平均値は 0.7903，F1_0 の平均値は 0.1046 であった．対して，gpt-oss-20B でタグ有無予測を行った際の Accuracy の平均値は 0.4859，F1_1 の平均値は 0.4288，F1_0 の平均値は 0.4100 であった．全体的に，Accuracy と F1_1 に関しては Llama3 のほうが高く，F1_0 に関しては gpt-oss のほうが高い結果が得られた．なお，一部タグの F-1score について値が 0 となっているものがあるが，これはタグ有無の予測が全ての小論文において同じ結果となったことに起因している．本来は 0 除算となるため数値は得られないが，本実験では便宜上 0 として扱う．

4.3　考察

それぞれのタグについて Accuracy，F1_1 を比較すると Llama3 が高く、F1_0 については gpt-oss が高い傾向にある．これらの結果より，Llama3 は質問に対し肯定しやすく，gpt-oss は質問に対して否定しやすい傾向があることが考察される．

実際に，Llama3 はタグ該当部分を含むという予測が多く，gpt-oss はタグ該当部分を含まないという予測が多い結果が見られた．データセットの特性上，基本的にタグの種類を問わずタグ該当部分を含む小論文の方が含まない小論文より多い為，結果的に Accuracy は Llama3 の方が高く出やすくなっている．その結果 gpt-oss に精度面で優位であるような結果に見て取れるが，これはデータの偏りによる影響であると考えるのが妥当である．これらのことから，予測の信頼度を高める為両モデルを使用し，2 つの予測結果を考慮して最終的な予測結果を出力するというアプローチの有効性が新たに考えられる．先行研究で大規模言語モデルの回答バイアスについて述

表 2　実験結果

	Llama3			gpt-oss		
タグ	Acc.	F1_1	F1_0	Acc.	F1_1	F1_0
g1 光	**0.8384**	**0.9094**	0.2535	0.1311	0.0000	**0.2318**
g1 影	**0.8384**	**0.9094**	0.2535	0.1890	0.0362	**0.3000**
g1 格差縮小	0.4878	**0.6300**	0.1683	**0.6128**	0.2395	**0.7403**
g1 格差拡大	**0.8262**	**0.9048**	0.0000	0.2561	0.2278	**0.2824**
g2 光	**0.8872**	**0.9400**	0.0513	0.2409	0.2567	**0.2243**
g2 影	0.6341	**0.6648**	0.5973	**0.6921**	0.3034	**0.8023**
g2 具体例	0.6951	0.8201	0.0000	**0.7713**	**0.8577**	**0.4186**
s1 実証性の説明	**0.8960**	**0.9452**	0.0000	0.7676	0.8571	**0.3770**
s1 再現性の説明	0.8960	0.9452	0.0000	**0.9144**	**0.9544**	**0.3000**
s1 客観性の説明	**0.8960**	**0.9452**	0.0000	0.8287	0.8967	**0.5000**
s2 自然相手	**0.6850**	**0.8131**	0.0000	0.5046	0.5207	**0.4873**
s2 持続役割	**0.8043**	**0.8915**	0.0000	0.2813	0.2395	**0.3188**
s2 客観確保	0.4220	0.5935	0.0000	**0.7217**	**0.6128**	**0.7828**
s2 共通役割	**0.4557**	**0.6261**	0.0000	0.4220	0.4290	**0.4149**
e1 相互依存 (データあり)	**0.2897**	**0.4213**	0.0004	0.1379	0.1935	0.0741
e1 協力・協業の実態	**0.7759**	**0.8738**	0.0000	0.2241	0.0000	**0.3662**
e2 概略	**0.7690**	**0.8657**	0.1728	0.4759	0.4722	**0.4795**
e2 脱する方法	**0.7379**	**0.8492**	0.0000	0.2759	0.0367	**0.4199**
e3 日本	**0.7310**	**0.8361**	**0.2500**	0.1241	0.0000	0.2209
e3 中国	**0.8931**	**0.9427**	**0.2051**	0.0828	0.0000	0.1529
e3 韓国	**0.9207**	**0.9580**	**0.3030**	0.0586	0.0000	0.1107
e3 協調と対立	**0.6103**	**0.7570**	0.0174	0.4241	0.0973	**0.5772**
c1 論理的・合理的思考の説明	0.6483	0.7866	0.0000	**0.7690**	**0.8424**	**0.5677**
c1 目標志向的思考の説明	0.6586	0.7942	0.0000	**0.7069**	**0.8156**	**0.2857**
c1 内省的・熟慮的思考の説明	0.9345	0.9661	0.0000	**0.9828**	**0.9908**	**0.8718**
c2 カラーテレビ	0.5759	0.7309	0.0000	**0.6000**	**0.7422**	**0.1077**
c3 方法	**0.9276**	**0.9606**	0.5532	0.9241	0.9574	**0.6562**
Average	0.7309	0.7903	0.1046	0.4859	0.4288	0.4100

べられているもの [12] が存在するが，異なるバイアスを持つモデルを併用して予測を行う研究は我々の知る限り見当たらなかった為，新たな精度改善手段の 1 つとして考えられる．

また，s1，c1 の「〜の説明」タグについて，該当部分を含まないと予測する傾向に囚われず，gpt-oss は多くの小論文のタグ予測を成功させている．特に c1 のタグにおいては，全ての評価指標で gpt-oss の値が Llama3 を上回る結果が得られた．このことからも，全体的な予測精度は Llama3 より gpt-oss の方が高いと考えるのが妥当である．しかし，両モデル間でモデルサイズが異なる為，本来は 20B 相当の Llama3 モデルと比較した際に同等の結果が得られるかどうかで評価するべきである．そのため，本実験の結論としてはモデルサイズの違いによる影響も含めた上で，gpt-oss の方が優位であったと結論づける．

5　まとめ

本研究では，Llama3-ELYZA-JP-8B，gpt-oss-20B の 2 つの大規模言語モデルを用い，Zero-shot での小論文内にルーブリックの評価部分が存在するか否かの予測手法を提案し，評価実験によりその精度評価及び 2 モデル間の比較を行った．結果として，Accuracy と F-1score の数値では Llama3-ELYZA-JP-8B が上回るものが多かったが，その内訳を観察すると gpt-oss-20B のほうがより優れた予測結果を示していることが明らかになった．また，Llama3 に肯定性，gpt-oss に否定性が見られたことから，その特性を活かした予測手法についても一考の余地があることが示された．しかし，両モデルともほぼ完全に予測ができていないタグが一部存在する点や，「〜の説明の分かりやすさ」等の度合いを考慮する評価部分に関しても予測する必要がある点，タグ有無に加えてタグ該当部分の位置同定も行う必要がある点など，課題があることも確認された．今後は，本研究で予測精度が低かったものについての精度改善，除外したタグについての予測手法の提案及び Zero-shot での最終的な小論文の得点予測について取り組んでいく予定である．

謝辞

議論に参加して下さいました竹内研究室の諸氏に心より感謝致します.

本研究は JSPS 科研費 JP22K00530 の助成を受けたものです.

参考文献

[1] Xinlin Zhuang, Hongyi Wu, Xinshu Shen, Peimin Yu, Gaowei Yi, Xinhao Chen, Tu Hu, Yang Chen, Yupei Ren, Yadong Zhang, Youqi Song, Binxuan Liu, and Man Lan. TOREE: Evaluating topic relevance of student essays for Chinese primary and middle school education. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 5749–5765, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[2] SeongYeub Chu, Jong Woo Kim, Bryan Wong, and Mun Yong Yi. Rationale behind essay scores: Enhancing S-LLM's multi-trait essay scoring with rationale generated by LLMs. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, **Findings of the Association for Computational Linguistics: NAACL 2025**, pp. 5796–5814, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[3] 加藤嘉浩. 論理構造グラフを用いた自動採点モデル. 言語処理学会　第 30 回年次大会　発表論文集, E11-6, 2024.

[4] Yida Cai, Kun Liang, Sanwoo Lee, Qinghan Wang, and Yunfang Wu. Rank-then-score: Enhancing large language models for automated essay scoring, 2025.

[5] 石井雄隆, 舟山弘晃, 松林優一郎, 乾健太郎. 国語記述問題自動採点システムの開発と評価. 日本教育工学会研究報告集, Vol. 2024, No. 1, pp. 215–222, 2024.

[6] Yuning Ding, Marie Bexte, and Andrea Horbach. Score it all together: A multi-task learning study on automatic scoring of argumentative essays. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 13052–13063, Toronto, Canada, July 2023. Association for Computational Linguistics.

[7] 藤田晃輔, 山田寛章, 徳永健伸, 石井雄隆, 澤木泰代. 自動アノテーションを導入した g-eval による英文要約課題評価. 言語処理学会　第 31 回年次大会　発表論文集, P3-2, 2025.

[8] Yuan Chen and Xia Li. PLAES: Prompt-generalized and level-aware learning framework for cross-prompt automated essay scoring. 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. 12775–12786, Torino, Italia, May 2024. ELRA and ICCL.

[9] 水野友暉, 竹内孔一. 解釈可能性の高い自動採点モデルを用いた小論文採点支援システムの構築. 言語処理学会　第 31 回年次大会　発表論文集, P2-3, 2025.

[10] 竹内孔一, 大野雅幸, 泉仁宏太, 田口雅弘, 稲田佳彦, 飯塚誠也, 阿保達彦, 上田均. 研究利用可能な小論文データに基づく参照文書を利用した小論文採点手法の開発. 情報処理学会論文誌, Vol. 62, No. 9, pp. 1586–1604, 2021.

[11] 成岡智也, 竹内孔一. ルーブリックに基づいたタグを付与した日本語小論文データの構築と自動採点への効果. 言語処理学会　第 31 回年次大会　発表論文集, Q6-12, 2025.

[12] Daniel Braun. Acquiescence bias in large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Findings of the Association for Computational Linguistics: EMNLP 2025**, pp. 11341–11355, Suzhou, China, November 2025. Association for Computational Linguistics.

表 3　追加実験結果

タグ	タグ無し	タグあり	タグあり?
s1 実証性の説明	0.7843	0.8643	0.8720
s1 再現性の説明	0.7985	0.8898	0.9177
s1 客観性の説明	0.6524	0.8083	0.8123
c1 論理的・合理的思考の説明	0.9153	0.9264	0.9216
c1 目標志向的思考の説明	0.7846	0.8519	0.8655
c1 内省的・熟慮的思考の説明	0.8556	0.7988	0.8995

A　追加実験

Llama3-ELYZA-JP-8B の予測結果について，全ての小論文で該当タグ部分が存在すると予測してしまった結果，F1_0 の値が 0 となってしまったタグの例が多く見られた．この結果から，小論文内に該当タグ部分が存在すると予測されたものの間に差があるのかについて検証するため追加実験を行った．本追加実験では，Llama3-ELYZA-JP-8B が該当タグ部分が存在すると予測する確率について，実際に該当タグがアノテーションされている小論文と該当タグがアノテーションされていない小論文の間で比較する．前処理として小論文内の一部文字列 mask，プロンプトの明確化を行い，「〜の説明」タグについて，?が付随しているタグを含めた 3 種類間での比較を行った．前者に関して，s1 実証性の説明タグであれば小論文内の文字列「実証性」を mask し，その他についても同様の規則で mask する．後者に関して，プロンプト内の「はい」「いいえ」の部分を「含んでいる」「含んでいない」に変更した．比較する確率値 P については，式 (4) によって求められる．

$$P(t) = \frac{1}{n}\sum_{l=1}^{n} \mathrm{softmax}\left(\mathrm{logits}_l\right)_1 \tag{4}$$

この式は，l 番目の小論文内に対するタグ t に該当する部分の有無について Llama3 が回答に"1"のトークンを生成する確率を，実際にタグが付与されている (いない) 小論文 n 件で平均した値を算出している．実験結果は表 3 に示すとおりである．

追加実験の結果より，タグ該当部分を含む小論文と含まない小論文の確率平均値に大きな差があるものと，あまり差が無いものの両パターン存在することが分かる．また，タグ無しの列の全ての種類のタグについて確率の平均値が 0.5 を超えていることから，多くの小論文で「タグ該当部分を含んでいる」と予測していたことが改めて確認できる．タグ該当部分を含む小論文と含まない小論文の確率平均値の差が顕著な例として，s1 実証性の説明，s1 再現性の説明，s1 客観性の説明が挙げられる．これらすべては，Llama3 のタグ有無予測実験で全ての小論文に対してタグ該当部分が存在すると予測し，F1_0 の値が 0 となってしまったものである．しかし，追加実験ではタグ無しの確率平均値は 0.5 を超える値となってしまってはいるものの，実際にタグ該当部分が存在する小論文の確率平均値と比較すると大きな差があり、Zero-shot であってもタグ有無での差別化を十分に行うことができることが分かる．この結果から，トークン生成確率に明確な閾値を設定することが可能であると仮定すると，Llama3 であってもタグ有無予測を行うことが可能となる可能性があることが考察できる．しかし，c1 論理的・合理的思考の説明など，タグ無しとタグありの間で大きな差が見られないものに関しては閾値による識別も難しいため，予測精度改善には更なる工夫が必要であると考える．