

# 日本語小論文に対する採点およびフィードバックの生成

中本 さや香<sup>1</sup> 嶋田 和孝<sup>1</sup> 岡本 芳明<sup>2</sup> 中河内 孝<sup>2</sup><sup>1</sup>九州工業大学大学院 <sup>2</sup>湘南ゼミナール

nakamoto.sayaka478@mail.kyutech.jp shimada@ai.kyutech.ac.jp

{okamoto\_1990,nakakouchi\_1219}@shozemi.com

## 概要

自動文書評価 (Automated Writing Evaluation) は教育工学分野で広く研究されているテーマのひとつである。本研究では生成 AI である GPT-4 を用い、児童が書いた日本語小論文の自動採点を行う。採点基準として複数の観点を設定し、各観点について点数、その点数を与える根拠、児童へのフィードバックを生成する。GPT-4 による採点および人手による採点を行い、その比較を行った。また、GPT-4 によって生成された根拠文やフィードバックに対し、定性評価や考察を行った。その結果、GPT-4 による高品質な採点結果およびフィードバックの生成が確認され、小論文教育への活用可能性が示された。

## 1 はじめに

近年の学校教育では Society 5.0 時代を見据え、知識や技能だけでなく、論理的思考力や表現力の育成が求められる [1]。それらの能力の育成には、小学生からの小論文指導が効果的であることが主張されている [2]。小論文を書く力を向上させるには、頻繁に高品質なフィードバックを受けて見直すことが有効である [3]。しかしフィードバックの作成には労力がかかるため学校や塾などの教育機関では十分な演習環境を用意することが困難である。そこで、近年登場している高性能な生成言語モデルを利用した小論文の自動評価が注目されている。

小論文の自動採点 (Automated Essay Scoring: AES) システムは、1966 年に Page ら [4] が開発した重回帰モデルを用いたシステム Project Essay Grade: PEG を筆頭に、言語処理技術の向上とともに様々なシステムが提案されてきた [5] [6] [7]。1990 年台から 2000 年台初頭にかけて、世界的なテスト機関である Educational Testing Service: ETS によって、e-rater [8] [9] が実用化されている。E-rater は自然言

語処理技術によって抽出した文法や用法、語彙などの多様な言語特徴量を基に予測する。E-rater を参考にして石岡らによって作成された Jess [10] は日本語の小論文に対する最初のシステムである。修辞・論理構成・内容などの観点から抽出した言語特徴量を利用して小論文の評価を行う。2010 年代頃よりニューラルネットワークを利用した自然言語処理が注目されはじめた。AES の分野においてもニューラルベースのシステム [11] [12] [13] や言語特徴量とニューラルネットワークを組み合わせたハイブリッド手法 [14] [15] が研究されてきた。

2023 年に、OpenAI による GPT-4 [16] などの最先端で高品質な生成系言語モデルが開発された。教育工学分野では、生成系言語モデルの教育応用が注目され、様々な研究がされている [17] [18] [19]。Mizumoto ら [20] は ChatGPT を用いた AES 応用について可能性を調査し、その有用性を示した。Naismith ら [21] は GPT-4 を用いて小論文の談話一貫性について採点および採点の根拠を生成し、人間評価者と採点結果および根拠文の比較を行った。

このような生成系言語モデルを用いた先行研究は主に英語の小論文を対象としており、日本語の小論文への応用はまだ研究されていない。本研究では GPT-4 を用いて日本語小論文の採点及びフィードバック生成を行い、それらの質を評価する。

## 2 実験設定

**小論文データ** 本研究では湘南ゼミナールが保持する 29 件の小論文を使用した。テーマは「外国人に紹介したい日本のいいところ」である。これらの解答者は全て小学 5 年生で、350–400 字程度の文字数が指定されている。表 1 は本研究のために国語教師が作成したループリック (評価指標) で、小論文の課題条件に沿った指標が設定されている。

**生成系言語モデルの設定** 使用した言語モデルは

表 1: ルーブリック.

Criterion 1 外国人に伝える日本のいいところ書かれている	
3	適切な記載があり, わかりやすい. (端的で一般的な言葉でまとまっている)
2	記載はあるが, 内容が適切でない (例: 日本特有のことでない), もしくはわかりにくい (例: 具体的すぎる・長すぎる.)
1	記載なし
Criterion 2 自分の経験が書かれている	
3	適切な内容 (主張と繋がっている) で, わかりやすい
2	記載はあるが, 内容が適切でない (例: 主張したい内容と繋がっていない), もしくはわかりにくい
1	記載なし
Criterion 3 日本の良さを残すため, もしくはもっとよくしていくために必要なことが書かれている	
3	適切な内容 (主張と繋がっている) で, わかりやすい
2	記載はあるが, 内容が適切でない (例: 主張したい内容と繋がっていない), もしくはわかりにくい
1	記載なし

“gpt-4-0613”<sup>1)</sup> (以降 GPT-4 と呼ぶ) である. GPT-4 には出力を制御するためのパラメータが存在する. 本研究では出力のランダム性を最小限にとどめるために “temperature” を 0.0 に設定し, 残りのパラメータは既定値のままとした.

### 3 手法

**プロンプト設計** GPT-4 は入力されたプロンプト (命令) を元に出力を行う. 一般的にプロンプトはシステムプロンプトとユーザープロンプトとに分けられる. 本研究で設計したプロンプトの概要を以下に示す. プロンプト全文は付録 A に表す.

- システムプロンプト: 期待される GPT-4 の振る舞い
  - 日本語小論文の専門家であるという役割
  - 採点タスクの説明.
- ユーザープロンプト: 小論文課題, ルーブリック (評価指標), 評価対象の小論文

1) <https://openai.com/research/gpt-4>

表 2: 各ペア間の RMSE 計算結果. 表中の K, N, S は人間採点者のイニシャルである. 小さい RMSE の値はペア間の誤差が小さいことを表す.

Pair	Criterion 1	Criterion 2	Criterion 3	Ave.
GPT-4, A	0.186	0.491	0.670	0.449
GPT-4, B	0.871	0.871	1.160	0.967
GPT-4, C	0.871	0.643	0.643	0.719
A, B	0.851	0.719	1.083	0.884
A, C	0.851	0.719	0.670	0.747
B, C	0.643	0.910	0.851	0.801
Ave.	0.712	0.7255	0.846	–

各評価指標に対する, 期待される GPT-4 の出力内容は以下の通りである.

- 点数 (1 ~ 3 点).
- その点数をつける根拠となる小論文内の文章.
- その点数をつける根拠の説明.
- 生徒に対するフィードバックコメント.

**人手採点** GPT-4 の採点結果と比較するために, 本研究では 3 名の国語教師が人手採点を行う. 本論文では以降 3 名を評価者 A, B, C と呼ぶことで区別する. 評価者 A と B は 22 年間, そして評価者 C は 3 年間の国語教師としての経歴を持つ.

## 4 結果と考察

**採点** GPT-4 および 3 名の人間採点者の採点結果から, 全ての組み合わせにおいて平均二乗誤差 (Root Mean Square Error: RMSE) を算出した結果を表 2 に示す.

RMSE は以下の式で計算する.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (s_{1i} - s_{2i})^2} \quad (1)$$

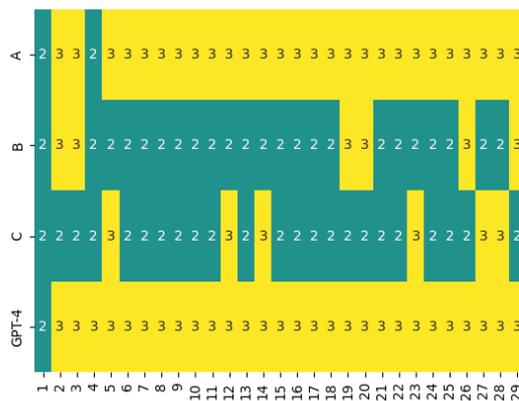
ここで  $n$  は小論文の総数,  $s_{1i}$  と  $s_{2i}$  は比較対象の 2 つの採点結果を示す. RMSE の値が小さいほど, ペア間の採点誤差が小さいといえる. GPT-4 と評価者 A のペアが平均的に一番小さい RMSE 値を得た. 一方, 評価者 B は他の評価者と離れた採点結果になる傾向にあり, 特に GPT-4 とは特に大きな差が生じた. また, 評価指標間でも誤差の程度に差が存在し, 特に Criterion 3 で大きな誤差を持つペアが多かった. 表 3 は, ペア間で発生した誤差の程度ごとに件数をまとめた結果を示す. Criterion 1 では全てのペアにおいて誤差が 2 点生じている採点は存在しなかった. Criterion 2 においては, 全ての人間評

表 3: ペア間の採点誤差の分布. 表中の数値は小論文の件数を示す.  $\pm 0$  はそのペア間の採点が一致していることを示す.

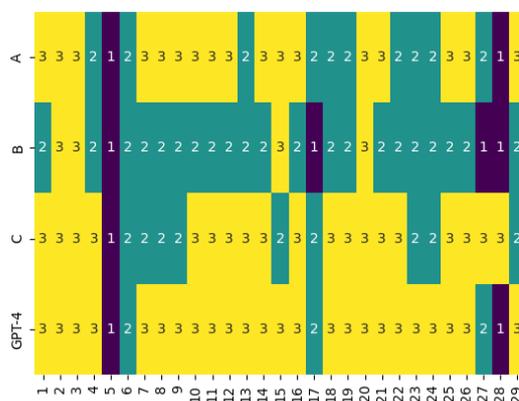
Pair	Criterion 1			Criterion 2			Criterion 3		
	$\pm 0$	$\pm 1$	$\pm 2$	$\pm 0$	$\pm 1$	$\pm 2$	$\pm 0$	$\pm 1$	$\pm 2$
GPT-4, K	28	1	0	22	7	0	19	9	1
GPT-4, N	7	22	0	7	22	0	11	11	7
GPT-4, S	7	22	0	7	22	0	20	8	1
K, N	8	21	0	14	15	0	10	14	5
K, S	8	21	0	17	11	1	19	9	1
N, S	17	12	0	11	16	2	17	9	3

評価者それぞれと GPT-4 の間では 1 点以内の誤差に収まっていることが確認できた. Criterion 3 では, GPT-4 と人間, および人間 2 名の間で 2 点の誤差を持つ答案が多く発生した. 図 1 は各評価者の採点結果の一覧を評価指標別で示したものである. ここで横軸は各小論文の ID を示し, 縦軸は評価者を示す. これらの結果より, 人間採点者間でも結果に大きくばらつきがある答案が存在することが確認できた. 評価指標ごとにばらつきの程度が異なることから, 人間評価者の中でもルーブリックの解釈や適用尺に差があり, 採点基準にも個人差が現れた可能性が考えられる. 3 名の人間評価者間では採点結果が一致しているが GPT-4 が異なる採点結果を与えている答案も確認できた. 例えば小論文 ID 10 の Criterion 3 における採点は, 人間評価者が全員 1 点を与えている一方で GPT-4 は 3 点を与えた. この小論文では「日本の科学の発展力」を日本の良さとして挙げ, 「日本の科学を勉強し, それをウズベキスタンの科学の発展力をあげるのに生かしていきたい」という結論を述べている. ウズベキスタンの発展力を上げることは日本の科学の発展力を残すもしくはもっとよくしていくために必要なことではない. このように GPT-4 が間違っただ判断を行う場合があるため, 実用化の際は人間が確認を行う段階を挟むなどの工夫が重要となる.

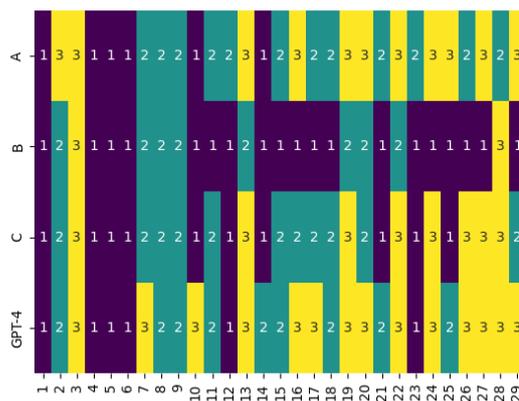
本研究で用いた GPT-4 は, 出力が毎回全く同じであるとは限らない. しかし, 小論文の採点において評価の基準や判断が一貫していることは重要である. そこで GPT-4 による採点の一貫性を確認するために, 同じ設定の GPT-4 を用いて追加で再度 29 件の小論文の採点を行った. 2 回目の採点結果との一致率は約 91% となり, 十分に一貫した採点を行える可能性が示された.



(a) Criterion 1: 外国人に伝える日本のいいところが書かれている.



(b) Criterion 2: 自分の経験が書かれている.



(c) Criterion 3: 日本の良さを残すため, もしくはもっとよくしていくために必要なことが書かれている.

図 1: 各評価者の採点結果の一覧. 横軸は各小論文の ID を示す. ヒートマップ内の数値は採点結果の点数を示す.

表 4: GPT-4 が生成した根拠文に対する説得力の定性評価。1: “全く同意できない”, 2: “あまり同意できない”, 3: “やや同意できる”, 4: “同意できる”。表中の数値は小論文の件数を表す。

	Criterion 1				Criterion 2				Criterion 3			
	1	2	3	4	1	2	3	4	1	2	3	4
C	0	3	1	25	0	2	1	26	1	2	0	26
著者	0	2	4	23	1	5	4	19	0	3	5	21

また、時間コストに対する有効性を確認するために、採点時間を比較した。3名の人間評価者が1答案あたりに要した採点時間はそれぞれ1分37秒、1分06秒、1分37秒であった。一方、GPT-4は採点、根拠文、フィードバックの生成を含めて1答案あたり48秒で出力し、十分に速いことを確認できた<sup>2)</sup>。

**採点根拠文の生成** GPT-4によって出力された採点の根拠文に対し、人間評価者2名(評価者Cと第一著者)が定性評価を行った。各根拠文に対し、小論文本文を踏まえた上で説得力があるかどうかを4段階で評価した。表4に示すように、ほとんどの根拠文が高評価を得た。付録Aに示す出力例は、高評価を得た根拠文である。

生成された根拠文に対して全く同意できない、と評価された答案について考察する。対象の答案は日本の良さとして「昔ながらの食べ物」を主張している。「最近、街にイタリアンやフレンチ、中華などがふえ、家の近くにあったそば屋もつぶれてしまった。もう少しでそう業100周年だった。そして、私も大好きな店で週に1回ほどは、いつも食べに行っていた。」という小論文内の文章に対して、GPT-4は「自身の経験として、地元のそば屋がつぶれてしまったことや、そのそば屋が好きだったことを述べているため、評価は3点です。」と根拠を示し、3点であると採点した<sup>3)</sup>。「そば屋がつぶれたことや好きだったこと」は「昔ながらの食べ物が日本のいいところである」という主張を支える経験ではないため不適切である。また、「日本の良さを残すために必要なことが書かれていますが、その内容が主張と直接繋がっているとは言えません。」など、減点された箇所や理由が明確に言及されていない根拠文が確認された。

このように一部未熟な結果もあるものの、概ね同

2) 実際には実行時間帯のアクセスの混み具合などで出力に要する時間は変動する。再度採点を行った際は43秒であった。

3) Criterion 2に対する採点。人間評価者は全員2点を与えた。

意できる根拠文が生成できた。根拠文によって採点者が注目すべき箇所を提案するなどの補助的な利用により、採点や自動採点の確認を行う際に効率化を図ることができるのではないかと考える。

**フィードバックの生成** GPT-4を用いた採点と根拠の生成に加え、生徒に対するフィードバックも出力した。いくつかの例を付録Bに示す。GPT-4が生成したフィードバックは小論文の内容と評価基準を踏まえたアドバイスを中心とした前向きなコメントになる傾向にあった。評価指標に基づいて最小の評価である1点が付与された答案には、小論文中で欠けている要素について指摘して具体的な改善案や解説を含むフィードバックが多く見られた。2点が付与された答案には、その小論文の良いところを褒めた上で改善に向けたアドバイスを行うフィードバックが生成された。そして、最大評価点である3点が付与された答案には、その評価項目を達成していることを褒めるフィードバックが生成され、時にはさらなるアドバイスを与えるものもみられた。

## 5 おわりに

本研究では、大規模言語生成モデルであるGPT-4を用い、ループリックに基づいた日本語小論文の自動採点および自動フィードバック生成について調査を行った。採点は人間評価者間の誤差と人間とGPT-4間の誤差が同程度以下に抑えることができていたことを確認した。また、採点結果に対して、人間が納得できるような根拠文を生成することが可能であることが示された。生徒に対するフィードバックの生成では、生徒の小論文能力を引き上げるための前向きで建設的なアドバイスがされることを確認した。人間より速く安定して、同等の質で採点が可能であるという結果より、練習問題や小テストなどの利用場面では自動採点システム単体への利用が可能であると考えられる。一方で、間違いが発生することは完璧には防げないため、高い精度が求められる定期テストや受験への応用については人間が内部に介在するようなシステムの構築が求められる。

我々の分析結果はGPT-4による自動評価の高い可能性を示すものであったが、本研究では、小学生の書いた29件の小論文という限定的なデータが対象であった。この結論を確固たるものにするためにはより幅広い属性の書き手による小論文を用いた分析が有益である。今後は教員や生徒が使用できる形のシステムを構築し、実証実験を実施していきたい。

## 謝辞

本研究を遂行するにあたり、データセットをご提供いただいた湘南ゼミナール御中、および人手評価にご協力いただいた中河内先生、近藤先生、佐々木先生にこの場を借りて深く感謝を申し上げます。

## 参考文献

- [1] 令和4年度文部科学白書, 2023.
- [2] 篠原京子. 小学校国語科「書くこと」における論理的思考力の育成. 東京未来大学研究紀要, Vol. 15, pp. 59–67, 2021.
- [3] Jessica Riddell. Performance, feedback, and revision: Metacognitive approaches to undergraduate. **Collected Essays on Learning and Teaching**, Vol. 8, pp. 79–96, 2015.
- [4] Ellis B. Page. The imminence of... grading essays by computer. **The Phi Delta Kappan**, Vol. 47, No. 5, pp. 238–243, 1966.
- [5] Ellis Batten Page. Project essay grade: Peg. In Mark D. Shermis and Jill Burstein, editors, **Automated essay scoring: A cross-disciplinary perspective**, pp. 43–54. Lawrence Erlbaum Associates Publishers, 2003.
- [6] Scott Elliot. Intellimetric: from here to validity. In Mark D. Shermis and Jill Burstein, editors, **Automated essay scoring: A cross-disciplinary perspective**, pp. 67–81. Lawrence Erlbaum Associates Publishers, 2003.
- [7] Giang Thi Linh Hoang and Antony John Kunnan. Automated essay evaluation for english language learners: A case study of my access. **Language Assessment Quarterly**, Vol. 13, No. 4, pp. 359–276, 2016.
- [8] Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, Martin Chodorow, Lisa Braden-Harder, and Mary Dee Harris. Automated scoring using a hybrid feature identification technique. In **36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1**, pp. 206–210, 1998.
- [9] Yigal Attali and Jill Burstein. Automated essay scoring with e-rater v.2. **The Journal of Technology, Learning and Assessment**, Vol. 4, No. 3, 2006.
- [10] Tsunenori Ishioka and Masayuki Kameda. Automated japanese essay scoring system based on articles written by experts. In **Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics**, pp. 233–240, 2006.
- [11] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In **Proceedings of the 2016 conference on empirical methods in natural language processing**, pp. 1882–1891, 2016.
- [12] Mitsuo Kiyono and Koichi Takeuchi. Automatic scoring of Japanese essay using neural networks. **Forum on Information Technology**, Vol. 18, No. 2, pp. 239–240, 2019. (In Japanese).
- [13] Misato Yamaura, Itsuki Fukuda, and Masaki Uto. Neural automated essay scoring considering logical structure. In Ning Wang, Genaro Rebolledo-Mendez, Noboru Matsuda, Olga C. Santos, and Vania Dimitrova, editors, **Artificial Intelligence in Education**, pp. 267–278, Cham, 2023. Springer Nature Switzerland.
- [14] Masaki Uto, Yikuan Xie, and Maomi Ueno. Neural automated essay scoring incorporating handcrafted features. In **Proceedings of the 28th international conference on computational linguistics**, pp. 6077–6088, 2020.
- [15] Sayaka Nakamoto and Kazutaka Shimada. Automated scoring of logical consistency of japanese essays. In Ning Wang, Genaro Rebolledo-Mendez, Vania Dimitrova, Noboru Matsuda, and Olga C. Santos, editors, **Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky**, pp. 652–658, Cham, 2023. Springer Nature Switzerland.
- [16] OpenAI. Gpt-4 technical report, 2023.
- [17] Ray Huang, Han-Shin Chen, Yuen-Hsien Tseng, and Liang-Yi Li. Gptutor: A chatgpt-powered programming. In **Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky: 24th International Conference, AIED 2023, Tokyo, Japan, July 3–7, 2023, Proceedings**, p. 321. Springer Nature, 2023.
- [18] Conrad W Safranek, Anne Elizabeth Sidamon-Eristoff, Aidan Gilson, and David Chartash. The role of large language models in medical education: applications and implications, 2023.
- [19] Stavros Athanassopoulos, Polyxeni Manoli, Maria Gouvi, Konstantinos Lavidas, and Vassilis Komis. The use of chatgpt as a learning tool to improve foreign language writing in a multilingual and multicultural classroom. **Advances in Mobile Learning Educational Research**, Vol. 3, No. 2, pp. 818–824, Aug. 2023.
- [20] Atsushi Mizumoto and Masaki Eguchi. Exploring the potential of using an ai language model for automated essay scoring. **Research Methods in Applied Linguistics**, Vol. 2, No. 2, p. 100050, 2023.
- [21] Ben Naismith, Phoebe Mulcaire, and Jill Burstein. Automated evaluation of written discourse coherence using GPT-4. In **Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)**, pp. 394–403. Association for Computational Linguistics, jul 2023.

## A 本研究で用いたプロンプト

### システムプロンプト

As a specialist in Japanese short essays (小論文), your task is to evaluate an essay in relation to the provided prompt (問題).

You will be given a prompt (小論文課題), an essay (小論文), and a rubric. The essay will be presented in individual sentences, each with a unique ID. The rubric will contain several evaluation axes, each with a scoring range of 1 to 3.

Your task is to score the essay according to the rubric, identify the key sentence IDs that form the basis for each score, provide a reason for the score in Japanese, and give formative feedback to the student for improvement. Please ensure to select only the most important sentences that serve as strong evidence for your scoring decisions.

## Input format (JSON)

```
{
  "prompt": <essay prompt>,"essay": [{"id":
  <integer>,"sentence":<sentence>}, {"id":
  <integer>,"sentence":<sentence>}, ...],
  "rubric": [{"axis-id": <axis-id>, "axis":
  <evaluation axis>, "scores":{"1":<goal>, "2":
  <goal>, "3":<goal>}}, ...]
}
```

## Output format (list of JSON)

```
[{"axis-id":<axis-id>,"score": <integer>,
  "key-sentences":<list of key sentence ids that
  provide the basis for the score>,"reason":<the
  reason of the score in Japanese>, "feedback":
  <formative feedback to the student to improve>,
  ...]
```

ユーザープロンプト (ここで<essay>は実際的小論文原文)

```
"prompt":
### テーマ
外国人に紹介したい日本の良さ
### 課題条件
- 「日本の良いところ」を自分の経験を踏まえて具体的に説明すること
- 「日本の良いところ」をそのまま残すために必要なこと もしくは もっと良くしていくために必要なことを、自分の経験を踏まえて書くこと
```

```
"essay":<essay>,
"rubric":
[
{"axis-id":0,"axis": "外国人に伝える日本の良いところが書かれている", "scores":{"1": "書かれていない", "2": "書いてあるが内容が適切でない (日本特有のことでない), もしくはわかりにくい (具体的すぎる・長すぎる)", "3": "適切な内容で、わかりやすい。端的で一般的な言葉でまとまっている"}},
{"axis-id":1,"axis": "自分の経験が書かれている", "scores":{"1": "書かれていない", "2": "書いてあるが内容が適切でない (主張したい内容と繋がっていない)", "3": "適切な内容 (主張と繋がっている) で、わかりやすい"}},
{"axis-id":2,"axis": "日本の良さを残すため、もしくはもっと良くしていくために必要なことが書かれている", "scores":{"1": "書かれていない", "2": "書いてあるが内容が適切でない (主張したい内容と繋がっていない), もしくはわかりにくい", "3": "適切な内容 (主張と繋がっている) で、わかりやすい"}}]
]
```

```
る"}},
{"axis-id":1,"axis": "自分の経験が書かれている", "scores":{"1": "書かれていない", "2": "書いてあるが内容が適切でない (主張したい内容と繋がっていない)", "3": "適切な内容 (主張と繋がっている) で、わかりやすい"}},
{"axis-id":2,"axis": "日本の良さを残すため、もしくはもっと良くしていくために必要なことが書かれている", "scores":{"1": "書かれていない", "2": "書いてあるが内容が適切でない (主張したい内容と繋がっていない), もしくはわかりにくい", "3": "適切な内容 (主張と繋がっている) で、わかりやすい"}}]
]
```

出力例 (「日本の文化や祭りの多さ」を紹介する小論文に対する出力)

```
[
{"axis-id": 0, "score": 3, "key-sentences": [0, 1], "reason": "日本の良さとして、文化や祭りの多さを具体的に述べており、外国人に伝える内容として適切でわかりやすい。", "feedback": "日本の良さを具体的に述べることでできています。このまま続けてください。"},
{"axis-id": 1, "score": 3, "key-sentences": [2, 3, 4, 5, 6, 7, 8, 9], "reason": "自身の経験として、台湾から来た天光生との交流を通じて日本の文化や祭りの多さを再認識し、それを外国人に伝えたいと思うようになったと述べている。内容が適切でわかりやすい。", "feedback": "自分の経験を具体的に述べることでできています。このまま続けてください。"},
{"axis-id": 2, "score": 2, "key-sentences": [10, 11], "reason": "日本の良さを残すために必要なこととして、一人一人が力を合わせて日本の文化を増やしていくことを述べているが、具体的な方法や行動が示されていないため、内容が適切でない。", "feedback": "日本の良さを残すために必要なことを述べる際には、具体的な方法や行動を示すとより良いエッセイになります。"}
]
```

## B 生成されたフィードバックの例

### 3点を与えた小論文へのフィードバック (Criterion 3)

日本の良さを守るための具体的な行動を提案することは非常に良いです。さらに、それらの行動がどのように自身の主張につながっているのかを明確にすることで、より説得力のあるエッセイになります。

### 2点を与えた小論文へのフィードバック (Criterion 3)

日本の良さを保つための意志を示すことができますが、具体的な行動や提案が不足しています。例えば、ゴミ拾いを続ける、リサイクルを推進する、環境教育を受けるなど、具体的な行動を提案すると、より説得力が増します。

### 1点を与えた小論文へのフィードバック (Criterion 2)

日本の四季の良さを述べる際に、自分の経験を具体的に書くと、より説得力が増します。例えば、自分が四季の中で特に好きな季節や、その季節に経験した思い出などを書くといいでしょう。