

日本語小論文に対する LLM を用いた建設的フィードバックの生成と分析

中本 さや香[†] 嶋田 和孝^{††}

[†] 九州工業大学大学院 情報工学府 情報創成工学専攻 〒820-8502 福岡県飯塚市川津 680-4

^{††} 九州工業大学大学院 情報工学研究院 知能情報工学研究系 〒820-8502 福岡県飯塚市川津 680-4

E-mail: [†]nakamoto.sayaka478@mail.kyutech.jp, ^{††}shimada@ai.kyutech.ac.jp

あらまし 小論文教育は、学習者の論理的思考力や表現力を育成するために効果的である。効果的な小論文教育の形成には、学習者に対して即時かつ高品質なフィードバックを行うことが重要である。本研究では、高品質なフィードバックとして学習者の理解を深めることを目的とした「建設的フィードバック」に着目する。建設的フィードバックを行うためには、小論文をよく読み、学習者の理解度を考慮して適切な支援度でフィードバックを行う必要がある。このようなフィードバックを行うことは指導者の多大な労力を必要とするうえ、それを即時に行うことは難しい。そこで、本研究では、大規模言語モデル (LLM) を用いて建設的フィードバックを自動で生成することを目指す。近年の LLM は、非常に高度な言語生成能力を持ち、入力に指示を含めることでその指示に従った出力を行うことが可能である。本研究では、LLM の入力に建設的フィードバックの要素を知識として与え、これらの知識に従ってフィードバックを行うように指示する。実験の結果、小論文の内容のレベルが低いほど支援度の高いフィードバックが増えるなど、LLM 自身が学習者の能力に応じて支援度の調整を行える可能性が示された。今後の課題として、LLM の生成したフィードバックの教育効果の実証実験による効果性の検証が求められる。

キーワード フィードバック自動生成, 自動文書評価, 小論文, 大規模言語モデル

Generating and Analysing Constructive Feedback on Japanese Short Essays Using LLMs

Sayaka NAKAMOTO[†] and Kazutaka SHIMADA^{††}

[†] Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology 680-4 Kawazu, Iizuka, Fukuoka, 820-8502 Japan

^{††} Department of Artificial Intelligence, Kyushu Institute of Technology 680-4 Kawazu, Iizuka, Fukuoka, 820-8502 Japan

E-mail: [†]nakamoto.sayaka478@mail.kyutech.jp, ^{††}shimada@ai.kyutech.ac.jp

Abstract Essay education is an effective method for developing students' logical thinking and expressive skills. Enhancing the efficacy of this approach necessitates providing prompt and high-quality feedback. This study focuses on “constructive feedback” as a framework of high-quality feedback aimed at deepening students' understanding. To provide constructive feedback, instructors must thoroughly read essays to scale students' comprehension and then provide feedback tailored to their needs. However, constructing such feedback requires significant effort, complicating timely delivery. This study explores the ability of using the Large Language Models (LLMs) to generate constructive feedback automatically. LLMs have recently exhibited advanced language generation capabilities, enabling executing tasks by incorporating task-specific instructions into their input. In this research, we input elements of constructive feedback as knowledge and instructed the LLMs to generate feedback based on this knowledge. Our experiments indicated that LLMs can potentially adjust the level of assistance based on students' writing proficiency, as evidenced by an increase in supportive feedback for essays of lower content level. Future research should focus on validating the educational efficacy of feedback generated by LLMs through empirical experiments.

Key words Automated Feedback Generation, Automated Writing Evaluation, Short Essays, LLMs

1. はじめに

近年の学校教育では、知識や技能だけでなく、論理的思考力や表現力を育成することが求められている [1]。それらの能力の育成には、小論文指導が効果的であることが主張されている [2]。効果的な小論文指導を行うためには、即時的かつ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], [4]。学習効果を高める高品質なフィードバックとして、学習者の理解を深めることを目的とした「建設的フィードバック (constructive feedback)」という概念がある [5]。建設的なフィードバックを形成するために、Mahboob [6] は凝縮的かつ一貫性のある (cohesive and coherent) フィードバックを行うことを提案した。凝縮的かつ一貫性のあるフィードバックでは、フィードバックの焦点を絞り、適切な構成を守り、学習者の能力に応じて支援度を使い分けることが求められる。しかしこのようなフィードバックの作成を行うには、小論文をよく読み、問題点などを整理して適切に言語化するなど、そのプロセスに多大な労力を必要とする。そのため、学習者数が多い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、自然言語処理技術を用いた小論文のフィードバックの自動生成が注目されている。近年、非常に高度な言語生成能力を持つ大規模言語モデル (Large Language Models; LLMs) が開発されていることから、それらのモデルをフィードバック生成に活用する可能性が注目されている [7]~[9]。

本研究では、Mahboob [6] による小論文教育の知見を LLM に提供することで建設的なフィードバックを生成することを目指す。特に、LLM は学習者の能力に応じてフィードバックの支援度を使い分けることができるかなどの観点に着目して分析する。

2. 関連研究

本章では、建設的なフィードバックについて研究とフィードバック自動生成に関連する研究を紹介する。

2.1 建設的なフィードバック

フィードバックは小論文教育の効果を高めるために重要な要素である [6], [10]。しかし、フィードバックはマイナスの効果にもなり得ることがある [11]。例えば、指摘箇所が多すぎるフィードバックなどは却って学習者の気力を損ねてしまう場合がある [6]。そのため、フィードバック自体の質に注意を払うことは重要である。効果的なフィードバックとして建設的フィードバック (constructive feedback) という概念がある。建設的フィードバックは、学習者の理解や成長を意図したフィードバックである [5]。

Mahboob [6] は、建設的なフィードバックを行う方法の一つとして、*cohesive*(凝縮的)で *coherent*(一貫性のある)なフィードバックを提供することを提唱している。フィードバックにおける cohesion(凝縮性) は以下のように定義される。

- Purpose: フィードバックの目的が絞られている。
- Audience: フィードバックの対象者を認識している。
- Organization: 以下の3つの段階に分けたフィードバック構成となっている。

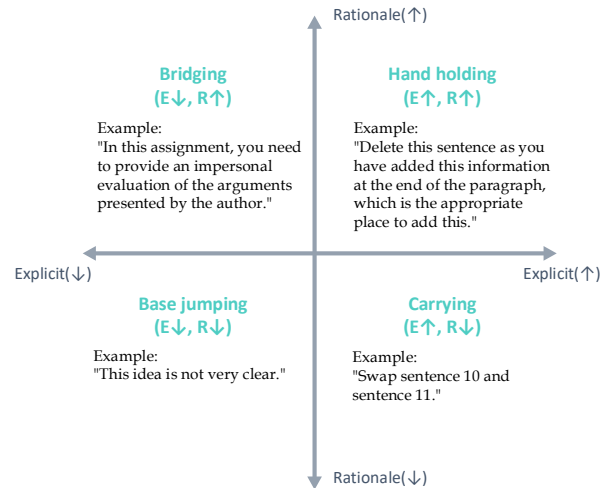


図1: フィードバックの類型と例。例は [6] より一部引用・改変。

- Feedback Preview (導入): 挨拶、フィードバックの概要
- Feedback: フィードバック本体
- Feedback Review (総括): 要点のまとめ、激励の言葉

また、coherent なフィードバックとは、フィードバックの各要素が学習者の問題理解を深めることを助けるフィードバックと定義される。Coherence(一貫性)を得るためには、生徒の理解度や問題点によって様々な feedback type を使い分けることが重要となる。ここで、Mahboob によって定義される feedback type は *explicitness* (E) と *rationale* (R) の2軸で分類される。*Explicitness* は明示性のことであり、*explicitness* が高いフィードバックは修正例を明示的に提案する。逆に *explicitness* が低いフィードバックは、生徒に対して修正例は与えず、生徒自身の修正を促す。*Rationale* は根拠・理由のことであり、*rationale* が高いフィードバックでは「なぜ問題なのか」「なぜ修正すべきなのか」の理由を詳しく説明する。*Rationale* が低いフィードバックでは問題点の説明をほとんど行わない。この2軸から図1のように類型化することができ、以下のように説明できる。

- Hand holding (E ↑, R ↑): 明示的に修正例を提供し、修正が必要な理由も提供するフィードバック。学生に対する支援レベルが一番高く、困難な問題や初学者に対して良いとされる。
- Carrying (E ↑, R ↓): 修正例のみを提供するフィードバック。スペルミスなどの低レベルなミスなどに適切とされる。
- Bridging (E ↓, R ↑): 修正例は与えず、修正が必要な理由のみを提供するフィードバック。学習者はフィードバックから問題点を理解し、自分自身で修正案を考えることが求められる。
- Base jumping (E ↓, R ↓): 修正例および修正が必要な理由をほとんど提供しないフィードバック。問題点に対するヒントのみを与えられ、学習者は自分自身で文章を見直して問題点を発見することが求められる。そのため、学習者にとって修正を行う難易度が高い。

2.2 フィードバック自動生成

フィードバック自動生成の先行事例・研究を紹介する。

2.2.1 従来のフィードバック自動生成

ETS (Educational Testing Service) は修正箇所をフィードバック

するウェブシステムとして Criterion© を開発した¹。Criterion© は文法、用法、文機構、スタイルの4つのカテゴリにおける間違い箇所を指摘して修正の提案を行う。間違いのある箇所を示して修正を行うこと (corrective feedback) が主な機能であるため、内容の適切性に対する指摘など、高レベルなフィードバックはされない。

2.2.2 大規模言語モデルの誕生とフィードバック自動生成

2022年に大規模言語モデル (Large Language Models; LLMs) である ChatGPT³ がリリースされ、その言語理解力や言語生成能力の高さに注目が集まった。ChatGPT などの LLM は入力にタスクの説明を含めることで追加学習を行わずにタスクを遂行できる。そこで、フィードバックの生成においても LLM が活用できることが期待されている。例えば Pankiewicz と Baker [12] は GPT-3.5 を用いてプログラミング課題に対するフィードバック生成し、フィードバックを受けた生徒はそうでない生徒よりも良い結果を得ることができることを確認した。Wan ら [13] は GPT-3.5 を用いて物理の短答記述式問題に対するフィードバック生成を行い、LLM を AI アシスタントとして使用する実現可能性を示した。

小論文に対するフィードバック自動生成では、Dai ら [7] が2つの GPT モデル (ChatGPT・GPT-4) を用いて読みやすさ・有効性・信頼性の観点から調査を行った。Dai らが GPT モデルに入力した指示は以下のようになっており、効果的なフィードバックを行うための枠組みが明記されていない。

Please give feedback on the following text in terms of a clear description of the goals of the project, appropriateness of the topic to data science, a clear description of the business benefits, novelty/creativity and overall clarity of the report. <INSERT THE TEXT OF A REPORT>

また、著者らは以前、GPT-4 を用いた日本語小論文に対するフィードバック自動生成を試みた [8]。この試みにおいても、効果的なフィードバックを行うための枠組みは明記していない。また、調査に使用したデータ数も29件と少なく、日本語小論文フィードバック自動生成における LLM の活用について十分な議論ができたとは言えない。

本研究では2.1節で示したような小論文教育の知見を LLM に提供することで、効果的なフィードバックを生成することを目指す。

3. データセット

本研究では、竹内ら [14] が公開⁴している小論文データセットを使用する。本データセットは、9つのテーマにわたって、1-3 件の小問を提示して小論文答案を収集している。解答者は大学生および大学院生である。小論文の指定文字数は各論題に

よって異なり、100文字以内のものから800字以内のものまで存在する。収集した答案数も各論題によって異なり、88件の論題から328件の論題まで存在する。

各論題には、理解力・妥当性・論理性・文書力の4つの観点についてルーブリックが存在する。それらのルーブリックに基づいて1点から5点の評点が各答案に付与される。

また、テーマによっては、小論文を書くにあたって講義資料や参考資料が与えられることがある。それらの資料もデータセットに含まれており⁵、本論文ではそれらを参考文献と呼ぶ。

4. 実験

本章では、LLM を用いたフィードバックの自動生成および分析を行う。本研究では特に、効果的な小論文教育を行うために有用なフィードバックを生成することを目指す。効果的なフィードバックとして、2.1節で示した「建設的なフィードバック」に着目する。建設的なフィードバックの要素を知識と小論文を与えられた時に LLM はどのようなフィードバックを生成するのか分析を行う。

4.1 LLM への入力方法と実験設定

LLM は追加学習を行わずにタスクの説明を入力に含めることでタスクを遂行できる。これを prompt-based learning と呼び、LLM への入力をプロンプトと呼ぶ。一般的に、プロンプトはタスクの説明をシステムプロンプトとユーザープロンプトに分けられる。本研究では、著者らによる日本語小論文の自動採点及びフィードバック生成の先行研究 [8] を基にプロンプトを作成する。

システムプロンプトでは [8] を基にタスクの説明を行い、それに追加してフィードバックに関する知識を埋め込む。フィードバックに関する知識として、2.1節で示した「建設的なフィードバック」の要素をプロンプト内で説明する。そして、それらの要素に従ってフィードバックを行うように指示する。また、分析を行うために以下の出力ルールを指示する。

- Feedback preview は<preview></preview>で囲むこと
- Feedback 本体は<feedback type={type}></feedback>で囲むこと。ここで {type} は a, b, c, d のいずれかであり、以下のように定義される。
 - a: Hand-holding feedback
 - b: Carrying feedback
 - c: Bridging feedback
 - d: Base-jumping feedback
- Feedback review は<review></review>で囲むこと

ここで、[8] のプロンプトは、自動採点タスクとフィードバック生成タスクの両方に同時に取り組む内容となっている。本研究ではフィードバック生成が主な目的である。しかし、小論文課題において何を書くことを求められているのか、与えられた答案はその目標とどのような差分があるのかを LLM 自身に考えさせるためにルーブリックを与え、採点も出力させることと

(注1) : <https://criterion.ets.org/Default.aspx?ReturnUrl=%2f> ただし、2025年8月31日にサービス提供終了予定。

(注2) : 文法ミスや単語の置き換えなど修正が容易なフィードバックを低レベル、内容や構成など修正が難しいものを高レベルとする。

(注3) : <https://chatgpt.com/>

(注4) : <https://www.gsk.or.jp/catalog/gsk2021-b/>

(注5) : 一部、参考文献なしの論題や、参考文献がデータセットに収録されていない論題がある。

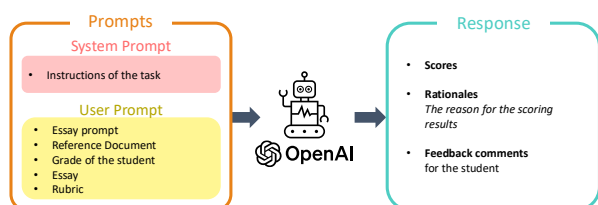


図 2: LLM を用いた自動採点のシステム図。

する。また、Naismith ら [15] に従い、なぜその点数をつけるにあたったかの理由を述べる根拠文の生成も行う。

ユーザープロンプトでは、フィードバック・採点の対象となるデータを与える。具体的には、論題、参考文書(参考文書が存在する場合のみ)、ルーブリック、学習者の学年、小論文本体を与える。また、本データセットではルーブリックに「備考」という、採点時に参考とするテキストが添付されている場合がある。その際は備考も入力に含める。

以上のプロンプトを用いてフィードバック作成と採点を行う様子を図 2 に示す。

LLM モデルは gpt-4o-mini-2024-07-18 を使用する。また、ランダム性を極力小さくするため、temperature は 0.0 に設定する。データセットはデータセット全件を用いる。採点タスクでは「理解力」の観点を採点するように指示をする⁶。

4.2 研究課題

フィードバックはいくつかの焦点に絞り適切な構成で書くこと、フィードバック種類は学習者のレベルを考慮した上で適切に選択することが求められる [6]。本研究では、LLM がどのようにフィードバックを構成し、どのようにフィードバック種類の使い分けを行うかを調査する。特に、フィードバック種類について、以下の研究課題 (RQ) に対して調査および分析を行う。

RQ 1. LLM は学年を考慮してフィードバックの種類を使い分けられるか。

学年情報を「大学生」と与える場合と「中学生」と与える場合でどのように変化が生じるか、という実験を行う。2.1 節で示したように、hand-holding フィードバックが最も手厚いタイプのフィードバックである。そのため、学年情報を「中学生」として与えた場合の答案には hand-holding が多くなると予想する。また、学生自身が修正方法を考える必要のある bridging や base-jumping は、学年情報が小さいほど少なくなると予想する。

RQ 2. 書き手の能力のレベルによってフィードバックの種類を使い分けられるか。

LLM が小論文から書き手の能力を推測し、フィードバックの種類を使い分けられるかどうかを分析する。書き手の能力は、小論文のクオリティに表れると考えられる。小論文のクオリティを小論文に対する採点結果とした場合、低点数の答案には hand-holding が多くなると予想する。

RQ 3. 参考文書がない場合 (LLM に十分な論題情報が渡されない場合)、LLM は具体的な修正の提案ができるのか。

(注6)：論題に沿っているかの内容を問う指標であり、内容に関するフィードバックを行うときの参考となると期待する。

表 1: 1 答案あたりのフィードバック文字数および個数の集計結果

学年 情報	文字数	フィード バック 個数	Hand- hol- ding 個数	Carr- ying 個数	Brid- ging 個数	Base- jum- ping 個数
大学生	161.67	1.740	0.665	0.006	1.068	0.000
中学生 に変更	181.45	1.743	0.630	0.009	1.105	0.000

表 2: 岡山大の 1 答案あたりのフィードバック文字数および個数の集計結果 (得点別)。ここで、得点は LLM の出力した採点結果とする。

集計対象	文字数	Hand-holding 個数	Carrying 個数	Bridging 個数
5 点	64.779	0.116	0.020	0.594
4 点	186.921	0.648	0.009	1.156
3 点	177.700	0.615	0.005	1.093
2 点	212.214	0.895	0.001	1.023
1 点	223.124	0.910	0.011	0.899

参考文書の有無によってフィードバックの傾向が異なるのかを調査する。特に、参考文書が与えられない場合、LLM は具体的に修正案を提案できない可能性がある。論題やルーブリック、小論文の情報のみでどのくらい修正案の具体性が変化するかに焦点を当てて分析を行う。

4.3 実験結果と考察

生成されたフィードバックについて、平均の文字数やフィードバック数の集計結果を表 1 に示す。

全体として、学習者にとって一番修正コストが高い base-jumping フィードバックは生成されなかった。また、修正例のみを提示する carrying フィードバックは 1 答案あたり 0.00–0.01 個程度となり、ほとんど含まれないことがわかった。Hand-holding フィードバックと bridging フィードバックは生成されやすく、特に修正例を具体的に提示しない bridging フィードバックは 1 答案あたり平均およそ 1 件が生成された。

RQ1. 学年とフィードバック種類の関係

表 1 より、学年情報を「大学生」とした場合と「中学生」とした場合ではフィードバックの種類による平均個数やフィードバック個数に大きい差は見られなかった。RQ1 では予想とは異なり、学年情報によってフィードバックの傾向はあまり変わらないと言える。

RQ2. 点数とフィードバック種類の関係

表 2 より、予想通り、点数が低いほど hand-holding フィードバックが出現しやすいことが示された。また、点数が高いほど bridging フィードバックが出現しやすいことがわかった。しかし、最大得点である 5 点が与えられた答案に対して bridging フィードバックは少ない。これは、満点の答案に対しては指摘すべき点が少ないため、フィードバック自体が少なくなったからである。

表 3: 参考文書の有無による, hand-holding と bridging のフィードバック個数の比較. 「人口減少」は参考資料がグラフであり, 参考文書としての情報はない. 「創造性」は参考文書なしで回答する形式であるため, 参考文書は存在しない. 「対照実験」では, 3 問中 2 問で参考文書を使用する.

	Hand-holding 個数		Bridging 個数	
	参考文書あり	参考文書なし	参考文書あり	参考文書なし
グローバル	0.808	0.504	1.048	1.129
自然科学	0.621	0.515	0.952	1.039
批判的思考	0.413	0.231	1.172	1.307
東アジア	0.772	0.444	1.078	1.167
人口減少	–	0.617	–	1.0341
文化的相対主義	0.678	0.505	1.176	1.277
創造性	–	0.933	–	1.162
民主主義	0.724	0.381	1.181	1.467
対照実験	0.679	0.533	1.029	1.016

満点の答案では, 他の得点帯と比較して carrying フィードバックが多い. Carrying フィードバックでは以下の例のように, 誤字や表現の改善を指摘することがほとんどである. 満点の答案に対しては, 採点には直接関係しないが修正するとさらに良くなる点をフィードバックとして提供したと考えられる.

「促進」→「促進」、「受け入れられた時の利益は大きくなるが、それと同時に、受け入れられなかった時の損害も大きくなるのが考えられる」→「受け入れられた場合の利益は大きい、受け入れられなかった場合の損失も大きい。」など、文法や表現の修正が必要です。

RQ3. 参考文書の有無によるフィードバックの変化

表 3 より, ほとんどの論題において参考文書が与えられない場合 hand-holding フィードバックの個数が少なくなり, bridging フィードバックが増えた. 表 4 に参考文書の有無によるフィードバック例を比較して示す. この例では, 参考文書を与えた場合は hand-holding フィードバックとして具体的な修正案(太字部分)を示すことができている. その一方, 参考文書を与えない場合はやや抽象的なフィードバックとなった.

点数とフィードバック個数の関係

表 2 より, 5 点(満点)の答案に対するフィードバックの平均件数は少ない. これは, 5 点の答案に対するフィードバックが 0 件となる場合があったからである. このような場合は以下の例のように, Feedback Preview や Review は存在し, できていることへの言及と少量のアドバイスが含まれている.

<preview>素晴らしい仕事です！批判的思考の3つの観点を明確に示し、それぞれの説明もわかりやすくなっています。今後の改善点としては、もう少し具体例を加えることで、さらに理解を深めることができるでしょう。</preview><review>全体的に非常に良い内容ですが、具体例を加えることで、より説得力のある文章になると思います。引き続き頑張ってください！</review>

表 4: 参考文書の有無による, フィードバック例(一部を抜粋)の比較.

フィードバック (参考文書あり)	<feedback type=c>韓国の財閥の問題については、財閥が中小企業の成長を妨げていることが述べられていますが、これがどのように他国との経済関係に影響を与えるのかを考える必要があります。特に、財閥の独占が国際的な競争にどのように影響するのかを考えてみてください。</feedback>
フィードバック (参考文書なし)	<feedback type=c>協調と対立の構図についての説明が不十分です。なぜこの構図が重要なのかを考え、各国の問題がどのように協調または対立に影響を与えるのかを明確にする必要があります。</feedback><feedback type=a>中国の問題についても、都市化の影響や農村人口の減少について具体的なデータや事例を挙げて説明すると良いでしょう。特に、これらの問題がどのように協調や対立に関連するのかを示すことが重要です。</feedback>

また, 1 点の答案に対する hand-holding フィードバックの個数は他の点数帯よりも多い. この原因として, 1 点の答案は修正すべき部分が多いことや, 学習者のレベルを考慮して高い支援レベルの hand-holding が多くなったと考えられる.

論題別のフィードバック分析

表 3 より, 「創造性」は他の論題に比べてフィードバックが多くなりやすい傾向がある⁷. 以下に示す通り, 「創造性」は他の論題と異なり, 学習者に参考文書を与えない形式である.

近年、「創造性を育む教育」の重要性が議論されている。「創造性を育む教育」の方法について、あなたの具体的なアイデアを600字以内で紹介してください。ただし、以下の4つの事項について必ず言及すること。1. 「あなたの考える創造性」とは何か。2. 「あなたの考える創造性」を育む教育の必要性。3. あなたの提唱する「創造性を育む教育」の方法を実現するときの障害や問題点。4. その障害や問題点を回避または解決するために必要なこと。

この形式では, 学習者は自分自身でアイデアを出し, それを条件に従って記述することが求められている. これは, 参考文書を読んで答えを探したり, 参考文書の言葉や表現を借りることができる形式に比べて難しい可能性がある. そのため, フィードバックを手厚くすることで学習者の理解をサポートしようとした可能性が考えられる. 表 5 に実際のフィードバック例を示す. Hand-holding フィードバック(feedback type = a)では, 太字部分のように学習者の記述を元に, LLM 自身が具体的な修正アイデアを提供していることがわかる.

Feedback Preview と Feedback Review

表 5 において Feedback Preview は<preview></preview>, Feedback Review は<review></review>で囲まれている. 2.1 節

(注7): ここで, 点数の分布が関係する可能性は考えられるが, この論題で LLM が出力した点数の平均は他の論題と大きく変わらなかった

表 5: 「創造性」に対するフィードバック例.

小論文	私に対して「創造性」とは今までなかったものを新に作りたす、先人が行った成果の基準のうに一層新しいものをつくる成果や能力である。(中略)この問題点を解決するために、まず、教師の人数を増やすことは第一歩である。多くの生徒と交流できる環境を作る。また、内気な生徒を発言を促進させ、すべての生徒を自分の意見を話せるように努力することが必要である。以上のことから創造性を育む教育を論じた。
フィードバック	<preview>このエッセイは「創造性を育む教育」についての具体的なアイデアをしっかりと述べており、特に創造性の定義やその必要性についての理解が見られます。しかし、いくつかの点で改善の余地があります。 </preview><feedback type=a>まず、創造性の定義については良いスタートを切っていますが、もう少し具体的に説明することが重要です。例えば、「創造性とは新しいアイデアを生み出す能力であり、既存の知識を基に新たな価値を創造すること」といった具体的な表現を使うと良いでしょう。 </feedback><feedback type=c>また、創造性を育む教育の必要性については触れていますが、具体的な理由や例を挙げることで、より説得力を持たせることができます。例えば、現代社会における創造性の重要性を具体的な事例を通じて説明することが効果的です。 </feedback><feedback type=c>さらに、提案した教育方法の実現における障害や問題点については、具体的な事例やデータを用いて説明することで、より明確にすることができます。例えば、教師の人数不足がどのように影響するかを具体的に示すと良いでしょう。 </feedback><feedback type=a>最後に、障害や問題点を解決するための具体的な方法についても、より詳細に述べるのが求められます。例えば、「教師の人数を増やすためには、教育予算の見直しが必要である」といった具体的な提案を加えると、説得力が増します。 </feedback><review>全体として、エッセイは良い方向に進んでいますが、具体性を増すことでさらに良くなるでしょう。引き続き、具体的な事例や例を挙げることを意識してみてください。頑張ってください！ </review>

で説明したようにそれぞれフィードバックの概要、まとめと激励の言葉を出力することができていることが確認できる。

5. おわりに

本研究では、LLM を用い、学習者にとって効果的な「建設的フィードバック」という概念に重点を置いて生成および分析を行った。建設的フィードバックを形成するために、Mahboob [6] が提案した要素を LLM に入力として与え、その要素に従ってフィードバックの生成をすることができた。

LLM は主に、「修正が必要な理由を説明し、修正案を提案する」学生への支援度が最大であるタイプの hand-holding 型フィードバックと「修正が必要な理由のみを説明し、具体的な修正は学生自身で考えさせる」タイプの bridging 型フィードバックを生成することがわかった。全体的には後者のタイプのフィードバックが多く生成される傾向があった。そして、点数が低い小論文に対しては前者のタイプが増えるなど、LLM が

学習者の能力に応じてフィードバックの種類を使い分けることができる可能性が示された。本研究では LLM 自身にフィードバック種類のタグ付けを出力させ、そのタグに基づいて分析を行なった。どの程度正確にフィードバック種類のタグ付けができているかは未調査であり、今後の課題となる。

また、フィードバックが実際にどのように学習者に影響を与えるか実証実験を行い学習者からの定性評価を得ることは、フィードバック自動生成の効果を確かめて改善点などの知見をさらに深めることができると考える。

文 献

- [1] 文部科学省. 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] Hillary G. Mullet, Andrew C. Butler, Berenice Verdin, Ricardo von Borries, and Elizabeth J. Marsh. Delaying feedback promotes transfer of knowledge despite student preferences to receive feedback immediately. *Journal of Applied Research in Memory and Cognition*, Vol. 3, No. 3, pp. 222–229, 2014. Cognition and Education.
- [4] Jessica Riddell. Performance, feedback, and revision: Metacognitive approaches to undergraduate. *Collected Essays on Learning and Teaching*, Vol. 8, pp. 79–96, 2015.
- [5] The Quality Assurance Agency for Higher Education. *Code of practice for the assurance of academic quality and standards in higher education Section6: Assessment of students*, 2006.
- [6] Ahmar Mahboob. Understanding and providing ‘cohesive’ and ‘coherent’ feedback on writing. *Writing & Pedagogy*, Vol. 7, No. 2-3, pp. 355–376, 2015.
- [7] Wei Dai, Yi-Shan Tsai, Jionghao Lin, Ahmad Aldino, Hua Jin, Tong-guang Li, Dragan Gašević, and Guanliang Chen. Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, Vol. 7, p. 100299, 2024.
- [8] Sayaka Nakamoto, Yoshi Okamoto, Takashi Nakakouchi, and Kazutaka Shimada. Towards human-level evaluation: Assessing the potential of gpt-4 in automated evaluation and feedback generation on japanese essays. In *2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 156–161, 2024.
- [9] Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and Alice Oh. Llm-as-a-tutor in efl writing education: Focusing on evaluation of student-llm interaction, 2024.
- [10] Jionghao Lin, Wei Dai, Lisa-Angelique Lim, Yi-Shan Tsai, Rafael Ferreira Mello, Hassan Khosravi, Dragan Gasevic, and Guanliang Chen. Learner-centred analytics of feedback content in higher education. In *LAK23: 13th International Learning Analytics and Knowledge Conference, LAK2023*, pp. 100–110, New York, NY, USA, 2023. Association for Computing Machinery.
- [11] John Hattie and Helen Timperley. The power of feedback. *Review of Educational Research*, Vol. 77, No. 1, pp. 81–112, 2007.
- [12] Maciej Pankiewicz and Ryan S. Baker. Large language models (gpt) for automating feedback on programming assignments, 2023.
- [13] Tong Wan and Zhongzhou Chen. Exploring generative ai assisted feedback writing for students’ written responses to a physics conceptual question with prompt engineering and few-shot learning. *Physical Review Physics Education Research*, Vol. 20, No. 1, p. 010152, 2024.
- [14] 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.
- [15] 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.