情報コミュニケーション学会第 22 回全国大会
CIS2025（2025.3.1～3.2）　発表番号：1B-3

探究のルーブリック評価における AI 活用の有効性の検証

永谷 研一*
Kenichi Nagaya.

*(株)ネットマン
Netman Corporation

あらまし：高校の探究の授業において生徒は課題研究の成果を論文やポスターにまとめて発表しているケースは多い。教員はルーブリック表に基づき評価を実施する場合もあるが、通常科目を受け持ちながら探究の授業も担当する場合、工数が問題となっている。そこで今回ルーブリック評価において AI を使った評価を行いその有効性について検証した。

キーワード：GPT-4o、ルーブリック評価、スーパーサイエンスハイスクール（以下 SSH）

1　はじめに

探究の授業において、生徒は課題研究の成果を論文やポスターにまとめて発表を行なっている場合が多い。その場合、教員はルーブリック表に基づき評価を行い生徒指導を行なっている。ルーブリック評価は教員間のブレを減らし評価の標準化を進められるとともに、生徒自身の成長においても有効な手段である。一方で課題として教員の評価工数の負担が大きい点がある。学生の論述を解読して適切な評価を行う必要があるため、レポート評価には時間が必要となる。教員の働き方改革が叫ばれる状況において、教員の負担を軽減するとともに生徒にとっての学びの質も落とさない評価方法を開発することは早急の課題となっている。今回、今まで用紙で行っていたルーブリック評価を、Chat GPT-4o（有料版）を使って行った。生徒の論文とルーブリックをスクリプトとして読ませることによって AI を活用した評価を行いその有効性を確認した。

2　ルーブリック表のスクリプト化

本研究を行うにあたり、SSH に指定されている東京都立富士高等学校・附属中学校に協力を頂いた。当校が利用しているルーブリック表をそのまま Chat GPT に読ませるよりスクリプト化することでより正確に理解させることができると考えた。図１が用紙のルーブリック表である。図２が ChatGPT に理解させやすいスクリプトである。評価項目（評価番号）は１から８までであり、それぞれに評価基準が ABC の三段階であったがそのすべてにおいてスクリプト化を行なった。

対象：論文
①研究課題　②著者名等　③研究概要（アブストラクト）　④背景　⑤目的　⑥仮説　⑦方法
⑧結果　⑨考察　⑩結論　今後の課題　⑪参考文献

ルーブリック（評価基準表）

項目番号			評価の観点	評価の対象	高いレベル	実力がある	一部に課題あり	Ⅰ 自己	Ⅱ 他者	Ⅲ 自己	Ⅳ 教員
					A	B	C				
1	挑戦力	価値創造	主体的に学習に取り組む態度	全て	研究課題、著者名等、研究概要（アブストラクト）、背景、目的、仮説、方法、結果、考察、結論、今後の課題、参考文献を記述しており、下記の分量でまとめている。（A４で４ページ目の半分以上かつ、４ページを超えない）	研究課題、著者名等、研究概要（アブストラクト）、背景、目的、仮説、方法、結果、考察、結論、今後の課題、参考文献を記述しており、下記の分量でまとめている。（A４で４ページ以上かつ、４ページを超えない）	研究課題、著者名等、研究概要（アブストラクト）、背景、目的、仮説、方法、結果、考察、結論、今後の課題、参考文献のいずれかを記述していない。また、下記の分量でまとめている。（A４で４ページ以上かつ、４ページを超えない）				
2	理数的解決力	結論伝達	思考・判断・表現	③	研究概要（アブストラクト）を各論の主題文と整合するように記述しており、研究の過程が分かるようになっている。背景、目的、仮説、方法、結果、考察、結論、今後の課題を全て記述している。	研究概要（アブストラクト）を各論の主題文と整合するように記述しており、研究の過程が分かるようになっている。背景、目的、仮説、方法、結果、考察、結論、今後の課題のいずれか一つを記述していない。	研究概要（アブストラクト）を記述しているが、各論の主題文と整合していない部分があり、研究の過程が分かりづらくなっている。背景、目的、仮説、方法、結果、考察、結論、今後の課題のいずれか一つを記述していない。				
3	理数的発見力	調査比較	思考・判断・表現	④⑤⑪	調査していた複数の先行研究についてまとめており、背景にその出典も明記している。また、先行研究で明らかになったことから、まだ明らかでない課題を明記し、研究の目的を設定している。	調査していた先行研究についてまとめており、背景にその出典も明記している。また、先行研究で明らかになったことから、まだ明らかでない課題を明記し、研究の目的を設定している。	調査していた先行研究についてまとめており、背景にその出典も明記している。また、研究の目的を設定している。				

図１：ルーブリック（評価基準表）用紙（一部）

```
添付ファイルを下記ルーブリックに沿って、結果と理由を教えて
# 対象
論文
# 評価項目
1. 研究課題　2. 著者名等　3. 研究概要（アブストラクト）
4. 背景　5. 目的　6. 仮説　7. 方法
8. 結果　9. 考察　10. 結論・今後の課題　11. 参考文献
# ルーブリック（評価基準表）
## 評価番号1
### 育成したい資質・能力
挑戦力
### グランドデザインの観点
価値創造
### 評価の観点
主体的に学習に取り組む態度
### 評価の対象
全て
### 評価基準
#### A（高いレベル）
研究課題、著者名等、研究概要（アブストラクト）、背景、目的、仮説、方法、結果、考察、結論、今後の課題、参考文献を記述しており、下記の分量でまとめている。
（A4で4ページ目の半分以上かつ、4ページを超えない）
#### B（実力がある）
研究課題、著者名等、研究概要（アブストラクト）、背景、目的、仮説、方法、結果、考察、結論、今後の課題、参考文献を記述しており、下記の分量でまとめている。
（A4で4ページ以上かつ、4ページを超えない）
#### C（一部に課題あり）
研究課題、著者名等、研究概要（アブストラクト）、背景、目的、仮説、方法、結果、考察、結論、今後の課題、参考文献のいずれかを記述していない。また、下記の分量でまとめている。
（A4で4ページ以上かつ、4ページを超えない）
---
## 評価番号2
### 育成したい資質・能力
理数的解決力
### グランドデザインの観点
```

図２：スクリプト化されたルーブリック（一部）

3 実験と考察

3.1 ４つの論文を６回ずつ評価

実験対象の論文を抽出するにあたり、探究テーマである「自然」、「社会」、「人文」、「数理」の各分野から１つずつ計４つの論文をランダムに選んで利用した。生成 AI はしばしば不正確な情報を提示するハルシネーションが報告されている[1]ため、それぞれの評価は６回ずつ行った。論文の PDF を添付したあと１回目のプロンプトでは、”添付ファイルを下記ルーブリックに沿って、結果と理由を教えて”と命令した。2～6 回目のプロンプトでは、”もう１度評価して、結果と理由を教えて”と命令した。最後に” ６回分の評価を表にして。縦に評価番号、横に回数で出力して”と命令した。結果を図３に示す。それぞれの最右列は、この実験の前に用紙で評価していた教員評価を追記した。

「自然」

番号	1	2	3	4	5	6	教員
1	B	A	B	A	A	A	A
2	B	A	A	A	A	A	B
3	B	B	B	B	B	B	A
4	B	A	B	A	A	A	A
5	A	A	A	A	A	A	B
6	C	B	B	B	B	B	A
7	C	B	C	B	B	C	B
8	C	B	B	B	B	B	B

「社会」

番号	1	2	3	4	5	6	教員
1	A	A	A	A	A	A	B
2	B	B	A	B	B	A	A
3	A	A	A	A	A	A	A
4	A	A	A	A	A	A	B
5	B	B	B	B	B	B	A
6	B	B	B	B	B	B	D
7	C	C	C	C	C	C	C
8	B	B	A	B	B	B	C

「人文」

番号	1	2	3	4	5	6	教員
1	B	B	A	B	B	A	C
2	A	A	A	A	A	A	A
3	B	B	B	B	A	B	C
4	A	A	A	A	A	A	C
5	A	A	A	A	A	A	C
6	B	B	B	B	B	B	C
7	C	B	C	C	C	C	D
8	B	B	B	C	B	B	C

「数理」

番号	1	2	3	4	5	6	教員
1	A	A	A	A	A	A	A
2	A	A	A	A	A	A	A
3	B	B	B	B	B	B	A
4	A	A	A	A	A	A	B
5	B	B	B	A	A	A	B
6	B	B	B	B	B	B	A
7	B	B	B	B	B	B	C
8	B	B	B	B	B	B	A

図３：AI によるルーブリック評価

また、AI が出力した評価理由を図４に示す。こちらは、「自然」の１回目の評価の出力内容である。

3.2 考察

懸案であったハルシネーションは大きくは見られなかった。AI は教員とほぼ同等の評価を行っている様子がわかる。また教員評価と異なる箇所もあった。たとえば「自然」の評価項目番号７：分析解析（図３太枠部分）は、AI 評価は３回 B、３回 C であった。教員評価は B である。AI の１回目の C 評価の理由を見ると、「図表を用いて実験結果を明確に記載していますが、仮説検定や回帰分析などの高度な分析は行われておらず、結果の解釈が簡易的に留まっています。」とある。ルーブリック表では高度な分析の評価基準 A であり、仮説検定をしてあれば B、そうでなければ C としている。論文を再確認すると仮説検証をしているとはいえず C 評価が妥当と思われる。AI が B 評価と C 評価を同回数つけたのも頷ける。一方教員が B をつけた理由は他の側面を見ている可能性があるのでここでの考察は行わない。

評価番号1:B
理由:研究課題、著者名、研究概要（アブストラクト）、背景、目的、仮説、方法、結果、考察、結論、今後の課題、参考文献が全て記載されており、分量もA4で4ページ以内に収まっている。しかし、記述が明確ではない部分が一部あり、Aには及ばない。

評価番号2:B
理由:研究概要（アブストラクト）は明確に記載され、各論の主題文と整合しています。ただし、研究過程で「周波数が高くなるにつれて発芽率が高くなる」という仮説については、すべての周波数での検証が十分ではないため、目的達成に課題が残る。

評価番号3:B
理由:背景では複数の先行研究に言及し、それらを基に研究課題を設定しています。また、参考文献も記載されています。しかし、背景情報において具体的な数値や引用の記載が不足しているため、Aには届かない。

評価番号4:B
理由:研究課題と仮説が新規性を持ち、観察や調査を通じて検証が可能なものであると考えられます。ただし、実験条件の制約があったため、仮説検証の深度が浅い点が課題です。

評価番号5:A
理由:仮説として「音の周波数が高くなるにつれて発芽率が高くなる」とし、独立変数（周波数）と従属変数（発芽率）が明確です。

評価番号6-1:C
理由:実験計画において、複数回の観察や実験を行う計画が立てられておらず、1回限りのデータに基づいています。図や表の利用は適切でしたが、科学的な精度を向上させるには更なる工夫が必要です。

評価番号7-1:C
理由:図表を用いて実験結果を明確に記載していますが、仮説検定や回帰分析などの高度な分析は行われておらず、結果の解釈が簡易的に留まっています。

評価番号8:C
理由:結果の数値を基にした考察が行われていますが、仮説の完全な検証には至っていません。また、今後の課題として具体的な提案が挙げられていますが、結果の解釈と結論の明確性がやや不足しています。

総合評価と改善提案
- **総合評価:** B
- **改善点:**
  1. 実験の信頼性を高めるため、複数回の実験を実施すること。
  2. 仮説検証のために統計的手法（仮説検定や回帰分析など）を導入すること。
  3. 背景情報に具体的な引用や数値データを加えることで、信憑性を高めること。

これにより、次回の研究がさらに高評価を得られるようになると考えられます。

図４：「自然」論文の AI 評価１回目（理由）

4 おわりに

探究論文のルーブリック評価に AI を活用することはとても有効であることがわかった。教員にとって工数削減のメリットは大きい。一方で AI は完全な教員の代替えというより１人の評価者としての位置付けとして有効であろう[2]。今後も教員の工数が削減されるツールの開発と研究をさらに続けていく。

参考文献

[1] 門谷宙ほか,大規模言語モデルは自身の Hallucination を検知できるか？,言語処理学会 第 30 回年次大会 発表論文集（2024 年 3 月）

[2] 石田 亨,LLM は学生評価のパートナー, 国立情報学研究所(NII) 第 77 回 教育機関 DX シンポ https://edx.nii.ac.jp/lecture/20240516-02,（2024 年 2 月 3 日アクセス