

# ChatGPTによるルーブリック生成及びレポート自動採点の検討 —アントレプレナーシップ教育におけるルーブリックの活用—

溝上 広樹\* 中山 泰宗\*

## A Study on Rubric Generation and Automated Scoring Using ChatGPT

—The Use of Rubrics in Entrepreneurship Education—

by

Hiroki MIZOKAMI\* and Yasumune NAKAYAMA\*

### 要 旨

本研究は、ChatGPTによるルーブリック生成およびルーブリックを利用した評価について、その利用可能性を検討することを目的として、信頼性および妥当性の検証を行った。ChatGPTにより、修正前・修正後の2つのルーブリックを得た。アントレプレナーシップ基礎講座を受講する大学生のレポートについて教師とChatGPTによる採点を行った。教師による採点で上位グループ・中位グループ・下位グループそれぞれ6つを得た。各グループのサンプルレポート1つずつを利用して、ChatGPTによる評価観点別の採点を実施した。次に、サンプルレポートと教員によるその採点結果を利用した学習有り・無しの条件で各グループ5つのレポートを採点した。その結果、修正後ルーブリックを用いたChatGPTによる採点結果では高い一貫性がみられた。また、上位レポートでは、高い信頼性が得られた。評価観点について、構成や表現に関する1観点以外では妥当な結果が得られた。教師とChatGPTによる採点の比較では、上位グループでは同等の結果となったが、中位グループと下位グループでは教員よりも高い得点分布となった。ChatGPTによって十分利用可能なルーブリックは得られたが、自動採点への利用は難しいと判断した。

**Key Words :** ChatGPT、ルーブリック、レポート、自動採点、アントレプレナーシップ

### 1. はじめに

現在、世界的に生成AIの利活用が進んでいる<sup>1)</sup>。特にテキスト生成AIは大規模言語モデル(LLM)の開発と応用により、ユーザーが容易に生成AIにアクセスできる状況にある。OpenAIのChatGPTは特に利用が進むテキスト生成AIで、その週間アクティブユーザーは2024年8月時点で2億人を超過していると報道されている<sup>2)</sup>。

教育分野への応用も検討されており、ルーブ

リックと関連付けた研究もされている。各インプットに対して、柔軟なアウトプットが得られることから、ルーブリックと生成AIを組み合わせることで、個別化教育等における教員の負荷低減の可能性があると考えられる。

実際に国外の研究においては、HornによるChatGPTを利用したルーブリック生成方法や<sup>3)</sup>、YavuzらによってLLMを活用したルーブリックによるエッセイの評価の信頼性と妥当性の検討<sup>4)</sup>等が複数報告されている。

一方で、国内の研究において、ChatGPTを利用したルーブリック自動採点に関する研究は、MizumotoとEguchiによる英語学習者が書いた

\* 崇城大学総合教育センター准教授

エッセイの評価において正確性と信頼性の報告<sup>5)</sup>、笹原・高橋による児童の自由記述の3段階評価とフィードバックに関する試みや<sup>6)</sup>、脇谷によるルーブリックでの採点利用についての基礎検討等が報告されているが<sup>7)</sup>、限定的である。

本研究は、ChatGPTによるルーブリック生成やChatGPTによるルーブリックを利用した評価について、繰り返しによる信頼性の検討や教師採点との比較による妥当性の検証を行い、その利用可能性を検討することを目的とした。

対象として、アントレプレナーシップ科目のレポート課題を利用した。アントレプレナーシップは「起業家精神」と訳され、日本の大学等でもアントレプレナーシップ教育による若手起業家育成が行われてきた<sup>8)</sup>。そして、近年は「急速な社会環境の変化を受容し、新たな価値を生み出していく精神を備えた人材の創出」といったアントレプレナーシップ教育の目的と人材育成の親和性の高さから<sup>9)</sup>、キャリア教育プログラムとしての実施も増えている<sup>8)</sup>。崇城大学においても、アントレプレナーシップを「起業に限定せず新しいことに挑戦するマインド」と定義し、学生の将来の可能性を広げる教育が展開されている。

ここでは、学生が何を知っているかだけでなく、何ができるかという側面を捉えるためにも、レポート課題等のパフォーマンス課題及び評価が重要となる。しかし、レポート課題の採点において客観性と信頼性を確保するためには、教員の認知的リソースと膨大な時間的リソースが必要であり、困難さを伴うことがある<sup>10)</sup>。

このような課題を解消するために、ルーブリックを利用した採点が利用される。ルーブリックは、複数の評価観点及び尺度、それらに対応した評価基準から構成されており、正しく設計されたルーブリックを利用することで、採点の客観性や信頼性、効率性を高めることができる<sup>11)</sup>。一方で、評価基準が数段階あるため、作成に時間がかかる欠点がある。栗田は、作成の効率性を高めるために生成AIの利用を提案している<sup>11)</sup>。

## 2. 材料と方法

### 材料

ChatGPTはChatGPT-4oを利用した。利用期間は、2024年9月13日～20日。個人の有料アカウントでブラウザからアクセスして利用した。

表1 レポート用ルーブリック生成のためのプロンプト

- ①大学の\*初年次教育を担当する教師になってください。
  - ②レポート採点用のルーブリックを作成してください。
  - ③ルーブリックは次の課題に対するものです：  
「資質を統合したあなたの人間性」、「あなたが好きなこと・興味があることに関する仕事」、「大学での学び」  
について (i) これら3つを満たす仕事、(ii) その仕事の具体的な仕事内容、(iii) あなたの興味関心や学びとその仕事内容の関連性、(iv) あなたの人間性がどのようにその仕事内容と結びつくかを記載してください。(i)～(iv)を必ず記載するようにしてください。
  - ④ルーブリックは、評価観点、尺度と得点、評価基準の3つの部分を含めてください。
  - ⑤ルーブリックは、次の到達度目標に沿った内容にしてください：自分自身の現在や将来について考えることができる。
  - ⑥ルーブリックには、次の尺度と得点を使用してください：  
・素晴らしい！(4点)  
・もう少し！(2点)  
・残念(0点)
  - ⑦ルーブリックの評価観点には、次の2つの観点も加えてください。  
・文章構成  
・日本語表現
  - ⑧ルーブリックは、表で作成してください。表の最初の行には、尺度と点数を記載します。表の一番左の列には評価観点を記載します。評価観点と基準に沿って評価基準を作成してください。
  - ⑨評価基準の記述は、学習者及び教員にとって適切かつ明確な表現にしてください。
- \* 下線部を中心に変更することで、様々な場面で利用可能なルーブリックが生成できる。

### 方法

#### (1) ChatGPTによるルーブリックの生成

レポート採点用ルーブリックを作成するため、ChatGPTを利用し、HornのChatGPTによるルーブリック生成方法を参考に表1のプロンプト(指示)を入力した<sup>3)</sup>。表1-①では、役割を指定している。表1-②は、生成するルーブリックの種類を指示している。表1-③では、具体的な課題を指示している。表1-④は、ルーブリックの基本的な構造であり、この指示を加えないと点数が表示されない可能性がある。表1-⑤は、ディ

プロマポリシーに紐づく科目の到達度目標である。本課題と紐づく到達度目標に絞って指示した。表 1-⑥尺度は、レポート課題については、まずは 3 段階で指示した。表 1-⑦では、レポートの内容に関する部分だけでなく、基本的な構成や表現についても評価するため追記した。表 1-⑦は、補足の部分であり、場合によっては、評価基準に必要な要素を指示することも想定される（例：⑦次の要素を評価基準に含めること：参考文献、文章構成、日本語表現）。表 1-⑧については、ルーブリックの基本的な構造であり、特に指示をしなくても多くの場合正しく生成され则认为られる。表 1-⑨については、表現の明確性を指示し、学習者の段階に応じて生成される評価基準の表現を調整している。

なお、表 1 のプロンプトによって、生成される修正前 3 段階のレポート採点用ルーブリック（以降、修正前ルーブリック）を元に、採点者によって解釈の幅が小さくなるように改善することと、ChatGPT によるより細かな採点の可能性を調査するために、次のプロンプトを指示し、修正後 5 段階のレポート採点用ルーブリック（以降、修正後ルーブリック）を得た。①ルーブリックには、次の尺度と得点を使用してください：たいへん素晴らしい！（4 点）素晴らしい（3 点）もう少し！（2 点）頑張れ（1 点）残念（0 点）、②採点のブレが少なくなるよう、具体的な数値で示せる部分は数値を入れてください。

## （2）データ収集と採点

初年次教育としてアントレプレナーシップ基礎講座を受講する大学生のレポートを利用した。

採点は、アントレプレナーシップ教育担当教員 2 名と、ChatGPT による自動採点を実施した。採点の際には、ID を振ったレポートの PDF ファイルを利用し、レポートの質に関する事前情報は分からない状態で採点した。また、教員は、別々に採点し、採点基準に関する事前・事後の調整や意見交換は行わなかった。

修正後ルーブリックを利用して教員による採点を実施し、上位グループ（ $n = 6$ ）、中位グループ（ $n = 6$ ）、下位グループ（ $n = 6$ ）を得た。各グループから 1 つのサンプルレポートを選び、修正前・修正後 2 種類のルーブリックを使用した。ChatGPT を利用して、後述する不連続な 5 回の評価により、ChatGPT 採点の信頼性を検討した。

ファイン・チューニングのため、Yavuz らが報告しているプロンプトを参考に表 2 のプロンプトを使って指示をした<sup>4)</sup>。なお、ワークシートの一部に今回のターゲットである自由記述のレポートを含む PDF ファイルを読み込むため、Yavuz のプロンプトの要素にはない採点部分を具体的に指示する④のプロンプトを追加した。なお、6 番目の評価観点に抜けることがあったため、[ルーブリックの日本語表現の観点が抜けています。日本語の観点に従って、採点を追加してください。] の追加のプロンプトを入力した。また、点数が出力されない場合や点数配点に不備がある際には、[ルーブリックに厳密に従って、観点別の得点と総点を出してください。][ルーブリックの配点に従って採点してください。] 等の追加のプロンプトを入力した。

なお、連続した採点を行おうとすると、出力される内容が減少したり、連続して満点が出力されたりする等の不具合が生じたため、毎回 [新しいチャット] を開き、表 2 のプロンプト及びルーブリックのファイル、レポートファイル 1 つを入力した。なお、ChatGPT の一時チャットモードを利用し、採点が終わる度にページの再読み込みをした。一時チャットモードは、入力した内容が履歴に残ったり、モデルの学習に使用されたりすることが無い機能である。

## （3）データ分析

ChatGPT によって生成されたスコアの信頼性を検討するため、評価者が対象に対して行った評価の一致度や信頼性を検討するために用いられる級内相関係数（ICC）を求めた。また、分散分析によって上位・中位・下位の各サンプルレポート採点 5 回分のデータを調査し、統計的に有意な差がある採点ができているのか確認した。さらに、観点別と総合の平均得点（ $M$ ）及び標準偏差（ $SD$ ）を算出した。この際、修正前・修正後の両ルーブリックを利用した。

次に、サンプルレポートを除く、上位グループ（ $n = 5$ ）、中位グループ（ $n = 5$ ）、下位グループ（ $n = 5$ ）のレポートについて、修正後ルーブリックを利用して教員 2 人による採点を実施した。教員採点の結果についても、一貫性を調査するため ICC を求めた。さらに、同じ修正後ルーブリックを利用して、ChatGPT による 2 回の採点を実施した。

表2 採点時のプロンプト

- ①大学の初年次教育を担当する教師になってください。
  - ②学生にレポート課題を与え、ルーブリックに基づいて学生の作文を採点します。
  - ③まずは、添付するルーブリックを読んでください。
  - ④その後、添付する学生のレポートの間7について、添付するルーブリックに従い採点してください。観点別の得点と総点を出してください。
  - ⑤評定の根拠を裏付けるため、学生のレポートから根拠を示してください。
- [ルーブリックの Excel ファイルを添付]  
[採点するレポートの PDF ファイルを添付]

表3 サンプル学習を伴う採点時のプロンプト

- ①大学の初年次教育を担当する教師になってください。
  - ②学生にレポート課題を与え、ルーブリックに基づいて学生のレポートを採点します。
  - ③まずは、添付するルーブリックを読んでください。
  - ④次に、サンプルレポート3つとその採点結果を読んでください。
  - ⑤その後、添付する学生のレポートの間7について、添付するルーブリックに従い採点してください。観点別の得点と総点を出してください。なお、サンプルレポートとその採点結果を参考にしてください。
  - ⑥評定の根拠を裏付けるため、学生のレポートから根拠を示してください。
- [ルーブリックの Excel ファイル]  
[サンプルレポートの PDF ファイルを添付]  
[採点するレポートの PDF ファイルを添付]

また、ChatGPT 採点精度を上げるため、サンプルレポートとその採点結果を学習する表3のプロンプトを使用した。上位グループ ( $n = 5$ )、中位グループ ( $n = 5$ )、下位グループ ( $n = 5$ ) のレポートについて ChatGPT による修正後ルーブリックを利用した2回の採点を行った。

#### (4) 倫理的配慮

本調査は、著者が所属する機関における倫理委員会の承認を得て実施した。対象者には成績確定後に、研究目的、方法、参加は自由意志で拒否による不利益はないこと個人情報の保護について説明を行った。研究には同意を得た対象者のレポートのみを匿名化して使用した。

### 3. 結果

#### (1) ChatGPT によるルーブリックの生成

レポート用ルーブリック生成のためのプロンプト (表1) によって、ChatGPT による修正前ルーブリックを得た (表4)。条件に合った観点別の分析的ルーブリックになっているものの、「一部具体的」等採点者によって解釈の幅が生じる表現が散見された。追加のプロンプトによって、表5の ChatGPT による修正後ルーブリックを得た。尺度が3段階から5段階とより細くなり、評価基準に具体的な数値が記載された。

#### (2) ChatGPT 採点の信頼性の検討

教員採点を基準に選んだ次の3つのサンプルレポートを利用した：上位レポート ( $M = 23.5$ )、中位レポート ( $M = 18.0$ )、下位レポート ( $M = 5.5$ )。級内相関係数の分析では、修正前ルーブリックにおける ChatGPT の ICC スコアは5回の測定で 0.63、各レポートを水準としたとき自由度 (水準間の自由度、水準内の自由度) 及び観測された分散比 (F 比)、有意水準は  $F(2,12) = 10.75$ 、 $p < 0.01$  となった。F 比が十分大きければ、グループ間の平均の差は偶然生じ得ないことを示し、F 比から求められる  $p$  値も小さくなる。なお、自由度は F 比が大きいのかを確認する際に必要な値であり記載している。また、ICC スコアは 1.0 に近いほど一貫性が高いことを意味する。ここでの分散比は、0.63 と 0.6 ~ 0.8 のしっかりと一貫性の範囲にはあるものの、高い一貫性があると判断する目安である 0.7 よりも小さい値となった。修正後ルーブリックにおける ChatGPT の ICC スコアは5回の測定では 0.82、となり高い一貫性を示した。また、 $F(2,12) = 32.07$ 、 $p < 0.0001$  となり、3つのグループに統計的有意差があることを示した。

次に各サンプルレポートの5回の測定データを比較する (図1)。まず上位レポートの ChatGPT による得点を見ると、修正前 ( $M = 23.6$ ,  $SD = 0.8$ ) と修正後 ( $M = 23.8$ ,  $SD = 0.4$ ) とほぼ一致していた。いずれの基準においてもほとんどの得点が24点であり、1度だけ22点か23点になっていた。上位レポートの修正後ルーブリックでの ChatGPT 採点は、特に高い信頼性が示された。

次に中位レポートの ChatGPT による得点を見ると、修正前の平均点 ( $M = 21.6$ ) と修正後の



表 4 ChatGPT が生成した 3 段階のレポート採点用修正前ルーブリック

評価観点 / 尺度	素晴らしい！ (4 点)	もう少し！ (2 点)	残念 (0 点)
(i) 3つの条件を満たす仕事の記載	3つの条件（資質、人間性、興味）が明確かつ具体的に結びついており、理解しやすい形で説明されている	3つの条件が一部曖昧または不十分だが、ある程度の結びつきが確認できる	3つの条件が曖昧で結びつきが弱い、または記載が不足している
(ii) 仕事の具体的な仕事内容	仕事内容が具体的かつ詳細に記載され、明確なイメージが持てる	仕事内容が一部具体的だが、全体像が不明確	仕事内容が抽象的で具体性に欠け、理解が難しい
(iii) 興味関心や学びと仕事内容の関連性	興味関心や大学での学びが仕事内容に密接に関連し、論理的に説明されている	興味関心や学びが仕事内容にある程度関連しているが、関連性がやや不明瞭	興味関心や学びと仕事内容の関連性がほとんど見られない、または説明が不足している
(iv) 自分の人間性と仕事内容の結びつき	自分の人間性が仕事内容と深く結びついており、自己理解が明確に表現されている	人間性と仕事内容の結びつきが一部見られるが、深く掘り下げられていない	自分の人間性が仕事内容とほとんど結びついていない、または記載が不十分
文章構成	構成が論理的で一貫性があり、スムーズに読み進められる	文章構成に一部不明瞭な部分があり、読み進める際に多少の混乱がある	文章構成が不十分で、内容を理解するのが困難
日本語表現	表現が明確かつ適切で、誤字脱字がほとんどなく、読解に支障がない	表現に一部曖昧さや誤字が見られるが、理解は可能	誤字脱字や不適切な表現が多く、内容の理解が困難

表 5 ChatGPT が生成した 5 段階のレポート採点用修正後ルーブリック

評価観点 / 尺度	大変素晴らしい！ (4 点)	素晴らしい (3 点)	もう少し！ (2 点)	頑張れ (1 点)	残念 (0 点)
(i) 3つの条件を満たす仕事の記載	3つの条件（資質、人間性、興味）が具体的に結びつき、100%以上の達成度で示されている	3つの条件が明確に結びついているが、一部抽象的(達成度 80-99%)	3つの条件が関連しているが、説明が不十分 (達成度 60-79%)	3つの条件が曖昧で関連性が弱い (達成度 40-59%)	条件が満たされていない、または説明が不足している (達成度 40%未満)
(ii) 仕事の具体的な仕事内容	仕事内容が具体的で、実際の業務内容を理解できるレベル (具体例 3つ以上)	仕事内容が具体的だが、やや抽象的な部分がある (具体例 2つ)	仕事内容が大まかで、全体像が不明瞭 (具体例 1つ)	仕事内容が曖昧で具体性が欠けている (具体例なし)	仕事内容の記載がない、またはほとんど理解できない
(iii) 興味関心や学びと仕事内容の関連性	興味や学びが仕事内容と非常に密接に関連している。論理的な説明 (具体例 2つ以上)	関連性があり、論理的な説明ができている (具体例 1つ)	関連性があるが、説明が不十分	関連性が曖昧で理解が困難	関連性がない、または説明がない
(iv) 自分の人間性と仕事内容の結びつき	自分の人間性が仕事内容と深く関連し、明確に表現されている (具体例 2つ以上)	人間性と仕事内容が明確に結びついている (具体例 1つ)	人間性と仕事内容が一部関連しているが、説明が不十分	結びつきが曖昧、または一部にのみ見られる	結びつきが全く見られない、または記載がない
文章構成	構成が非常に論理的で一貫性があり、段落が3つ以上に分かれており、情報が順序立てられている	段落が2つ以上に分かれていて、概ね論理的な流れがある	段落が1つか不十分で、情報が順序立てられていないが、最低限の構成がある	段落がなく、全体の流れが不明確である	段落構成が全くなく、論理的な流れがない
日本語表現	誤字脱字がなく、文法的にも完全に正しい (誤字脱字 0 個)	誤字脱字が少なく、文法的にも概ね正しい (誤字脱字 1-2 個)	誤字脱字がいくつか見られ、文法的な誤りも多少ある (誤字脱字 3-4 個)	誤字脱字が多く、文法的な誤りが多数 (誤字脱字 5 個以上)	誤字脱字が非常に多く、文法的な誤りが目立ち、内容がほとんど理解できない

平均点 ( $M = 21.4$ ) は、近いが、標準偏差は修正前 ( $SD = 4.8$ ) では、修正後 ( $SD = 2.2$ ) よりも幅が大きいことが確認できる。修正前のルーブリックによる採点では、12点から24点と幅があり、上位レポートと下位レポートの得点とも一致しており、中位レポートとして一貫した採点ができなかった。修正後ルーブリックでの採点では、18点から24点の間にあり、修正後ルーブリックの結果よりも信頼性が改善した。

最後に、下位レポートの ChatGPT による得点を見ると、標準偏差においては、修正前ルーブリックでの採点 ( $SD = 2.33$ ) と修正後ルーブリックでの採点 ( $SD = 2.32$ ) ではいずれも同等であった。得点では、修正前 ( $M = 14.4$ ) よりも修正後 ( $M = 15.2$ ) の方が平均点はやや高くなった。下位レポートでは、いずれのルーブリックでも採点の信頼性が概ね確認できた。

### (3) 評価観点別の ChatGPT 採点結果の検討

レポート課題のルーブリックの評価観点は次の5つのから構成されている：①人生の目標の表明、②目標設定と行動、③資質と仕事の結びつき、④文章構成、⑤日本語表現。①～③は、自己の振り返りに関する領域であり、④～⑤は構成や表現に関する領域である。

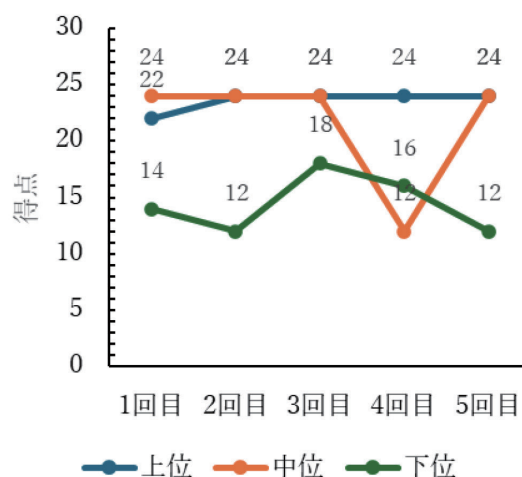
ルーブリックの評価観点別の ChatGPT 採点の信頼性を検討するため、サンプル学習前後の、上位・中位・下位の各5つのレポートの各観点別の平均値 ( $M$ ) と標準偏差 ( $SD$ ) を算出した (表6)。

上位レポートでは、観点別の得点平均値は、サンプル学習前 ( $M = 3.7$ ,  $SD = 0.5$ ) とサンプル学習後 ( $M = 3.6$ ,  $SD = 0.6$ ) では、平均点において0.1ポイント差があるのみで同等の結果となった。観点別に標準偏差を確認すると、サンプル学習前は0.3～0.5であったが、サンプル学習後は日本語表現で0.7と相対的にやや大きな幅が見られた。

中位レポートでは、観点別の得点平均値はサンプル学習前 ( $M = 3.3$ ,  $SD = 0.6$ ) とサンプル学習後 ( $M = 3.0$ ,  $SD = 0.6$ ) では、平均点において0.3ポイント差があった。観点別に標準偏差を確認すると、学習前は0.5～0.7、学習後は0.4～0.7とその幅は同程度であった。

下位レポートでは、観点別の得点平均値はサンプル学習前 ( $M = 2.8$ ,  $SD = 0.6$ ) とサンプル学習後 ( $M = 2.1$ ,  $SD = 0.6$ ) では、平均点におい

(修正前レポートでの採点結果)



(修正後レポートでの採点結果)

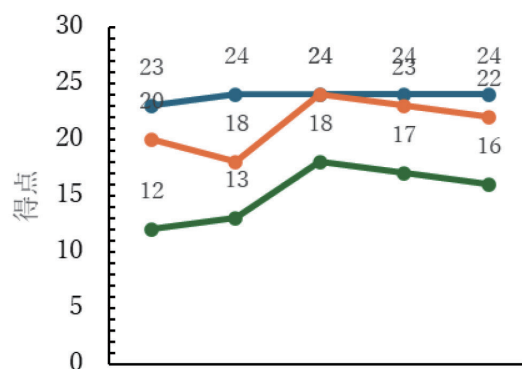


図1 ChatGPT が学生レポートに付けた得点

て0.7の差があり、学習の影響を最も受ける結果となった。観点別に確認すると、サンプル学習前は自己の振り返りに関する領域1つの観点 (iv) で標準偏差が0.6と相対的に大きな値になった。また、文章構成の観点でも標準偏差が0.7と大きな値になった。サンプル学習後の値では、自己の振り返りに関する領域について、さらに1つの観点 (i) で標準偏差が0.7になった。一方で、サンプル学習後の標準偏差では、文章構成の値は0.3と相対的に小さくなった。

サンプル学習後は、観点別の平均点においても、日本語表現の観点を除き、上位・中位・下位の順で得点が高くなった。

### (4) 教師採点と ChatGPT 採点

3グループ各5つの合計15レポートについて、修正後ルーブリックを利用した教員2人による採点結果の ICC およびグループ間の F 値を求めた。その結果、ICC スコアは0.96、グループ間

表 6 ChatGPT による修正後ルーブリックを用いたサンプル学習前後の観点別採点結果

	上位レポート				中位レポート				下位レポート			
	学習前		学習後		学習前		学習後		学習前		学習後	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
(i) 3つの条件を満たす仕事の記載	3.6	0.5	3.6	0.5	3.6	0.5	3.2	0.4	3.1	0.3	2.4	0.7
(ii) 仕事の具体的な仕事内容	3.4	0.5	4	0	2.6	0.7	2.9	0.7	2.3	0.5	1.8	0.4
(iii) 興味関心や学びと仕事内容の関連性	3.9	0.3	3.6	0.5	3.3	0.5	2.9	0.7	2.7	0.5	1.6	0.5
(iv) 自分の人間性と仕事内容の結びつき	3.9	0.3	3.7	0.5	3.3	0.6	3.1	0.7	2.7	0.6	2.2	0.6
文章構成	3.5	0.5	3.6	0.5	3.1	0.5	2.8	0.6	2.5	0.7	2.1	0.3
日本語表現	3.6	0.5	3.1	0.7	3.6	0.5	3.2	0.4	3.2	0.4	2.7	0.5
平均得点	3.7	0.5	3.6	0.6	3.3	0.6	3	0.6	2.8	0.6	2.1	0.6

の  $F(2,24) = 165.35$ 、 $p < 0.0001$  となり非常に高い一貫性が確認できた。サンプルレポート学習前後における ChatGPT による各グループ 5 個のレポート 2 回の採点結果を図 2 に示す。

上位グループのレポートにおいて、いずれの評価においても、ChatGPT の平均スコア（学習前： $M = 21.9$ 、学習後： $M = 21.6$ ）は、教師のスコア（ $M = 22.4$ ）よりもわずかに低くなった。

中位グループのレポートにおいては、いずれのルーブリックにおいても、ChatGPT の平均スコア（学習前： $M = 19.3$ 、学習後： $M = 18.1$ ）は、教師のスコア（ $M = 16.1$ ）よりも高くなったが、学習後は教員評価と同等の低い評価も見られた。

下位グループのレポートにおいては、いずれのルーブリックにおいても、ChatGPT の平均スコア（学習前： $M = 16.5$ 、学習後： $M = 12.8$ ）は、教師のスコア（ $M = 7.5$ ）よりも明らかに高くなったが、学習後は顕著に低くなった。

標準偏差については、上位グループでは、教員採点（ $SD = 2.0$ ）と比較して、学習前（ $SD = 1.6$ ）、学習後（ $SD = 2.0$ ）ともに同等もしくはそれ以下の値になった。中位グループでは、教員採点（ $SD = 1.6$ ）と比較すると、学習前（ $SD = 2.0$ ）学習後（ $SD = 2.5$ ）ともに値が大きくなった。下位グループでは、学習前（ $SD = 2.0$ ）は幅が大きいものの、学習後（ $SD = 1.6$ ）は教員採点（ $SD = 1.6$ ）と同等になった。

#### 4. 考察

ChatGPT によって、レポート採点に使用可能なルーブリックを生成することができた。特に修正後のルーブリックにおいては、数値が示され、より詳細に尺度を分けたことで、一貫性が保ちやすくなったと考えられる。実際に、教員 2 名の ICC スコアは 0.96 と非常に高く、客観性が担保されていることを示唆している。レポートの目的と内容を精査した上で、ルーブリックを生成することで比較的質の高いルーブリックを得ることができると考えられる。レポートの採点用のルーブリックにおいては、学生との活動意図の共有のしやすさ採点の効率性の観点から尺度や観点多すぎない方が良いと言える。しかし、今後 LLM 採点の精度が向上した際には、尺度数や観点を増やしたり、詳細な評価基準を設定したりしても、採点の効率性は担保される可能性がある。その際の運用方法として、学生との活動意図の共有の際には、各観点の最高基準のみを示す採点指針ルーブリックを示し、LLM 採点の際により複数の尺度や評価基準からなるルーブリックを読み込ませる方法も考えられるかもしれない。

ChatGPT による採点においては、特に修正後のルーブリックを使用することで、ICC スコア 0.82 の高い一貫性を確認することができた。上位レポートについては、5 回の採点においてルーブリックの高い信頼性が認められた。修正後ルーブ

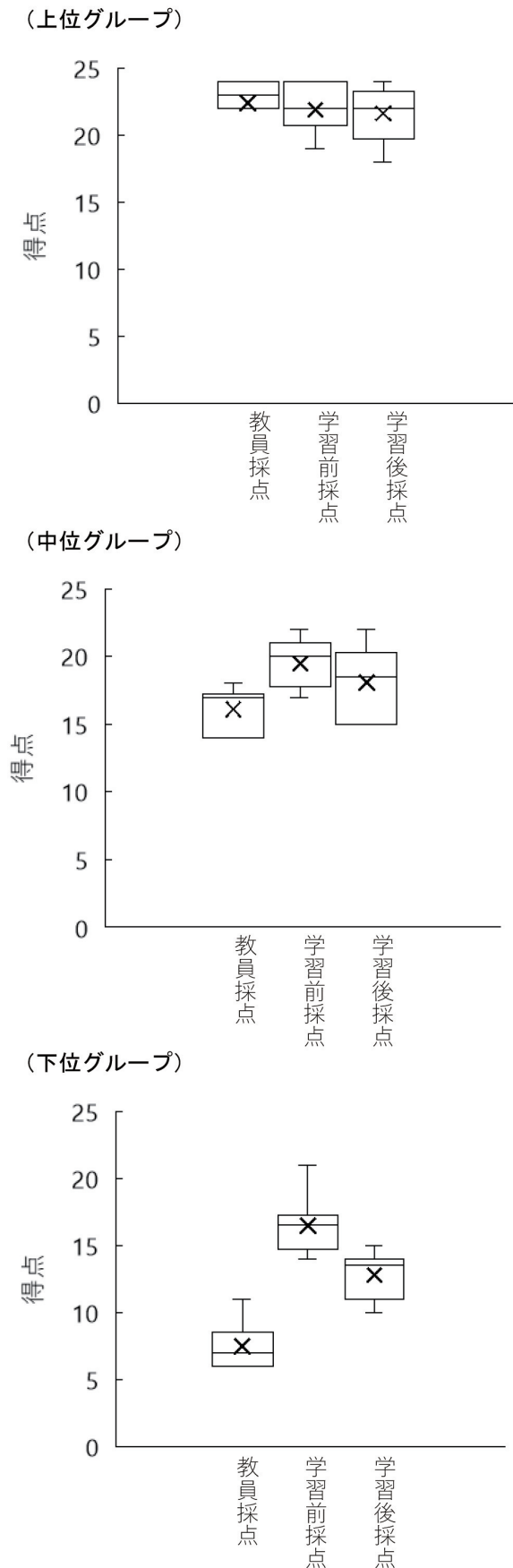


図2 教師及び ChatGPT 学習前後の採点結果

各グループ5本の2回採点の結果。箱ひげ図は上から値の最大値、75%、中央値、25%、最小値、×は平均値を示す。

リックを利用することで、中位レポートの信頼性は高まったが、上位レポートの採点結果には及ばず、下位レポートの採点結果も同等となった。上位レポートと下位レポートの区別は十分できているものの、中位レポートの得点は幅が大きかったことから、ループリックの採点基準をさらに修正する等し、信頼性を向上させる必要がある。

評価観点別の ChatGPT 採点では、全般的にはサンプル学習後の方が平均点は低くなる傾向にあった。しかし、自己の振り返りに関する4評価観点では、(ii)のみは学習後に平均点が高くなった。ループリックには「具体的」「抽象的」といった表現があり、教員は経験則で理解できても、ChatGPT にとっては幅がある解釈ができる表現になっていることも影響しているかもしれない。また、日本語表現の観点では、上位レポートと中位レポートでは平均点において、中位レポートの方が0.1点高くなっていた。上位レポートは長文になる傾向があり、その分ミスが起こりやすくなったり、中位レポートでは比較的短い文章のためミスが起こりにくかったりすることが要因の一つになっているかもしれない。この点については、別途文字数による減点項目を加える等の検討も必要だと考えられる。

教師採点と ChatGPT による採点では、上位グループの得点は22点前後に分布しており、類似したものとなった。中位グループと下位グループでは、いずれも教員採点の結果よりも ChatGPT による採点の方が高くなる傾向にあった。しかし、サンプルレポート学習後の得点では、教員の採点の分布に近づいていた。特に下位グループでは平均値が3.7点低くなり比較的大きな変化が見られた。しかしながら、ChatGPT による中位グループと下位グループの採点結果は、教員採点の結果とはそれぞれ平均値で2.0点、5.3点と差があり、教員採点の結果に十分に近づけることはできなかった。

英語のエッセイの LLM 採点の結果では、今回の結果とは異なり、下位グループの評価で人間と生成 AI の採点結果で一致が見られ、上位グループと中位グループでは採点の幅がより大きくなることが報告されている<sup>4), 5)</sup>。米国における英語教育においては、エッセイの採点において厳格な文章形式が重要視されると言われている<sup>12)</sup>。また、そのためのチェックリストなどもオンライン上で多数見受けられる (<https://www.google.com/> で



"essay checklist"で検索、access on 2024/09/24)。そのため、GPTが減点対象も含めて十分に学習している可能性が挙げられる。

なお、教師採点とChatGPT採点の差を改善する方法としては、ルーブリックをより洗練させることが考えられる。今回は研究の性質上実施していないが、通常ならばサンプルレポートを採点した後に、教員同士が差異を解消するモデレーションというプロセスを入れることが考えられる<sup>11)</sup>。また、授業者の考えや経験則をルーブリックに反映するためには、追加の指示や、その後の授業者による微調整も必要である。この際、ルーブリックを総合的に評価するメタ・ルーブリック<sup>13)</sup>、既存のルーブリック作成の際のチェックリスト<sup>6)</sup>を参考にプロンプトを指示することで、より正確なルーブリックが生成されることが期待できる。今回最初の修正時に利用したプロンプト(①ルーブリックには、次の尺度と得点を使用してください、②採点のブレが少なくなるよう、具体的な数値で示せる部分は数値を入れてください)は、チェックリストの「評価尺度の表現が教育的観点から適切な表現になっているか」「評価尺度ごとに基準間で明確にかき分けられているか」、メタ・ルーブリックの「達成段階の数は評価対象者の年齢や価値の内容に対応している」、「各評価基準は明確であり、類似の評価基準がない」と関連している。

ChatGPTによる採点では、採点結果が数値で出力されなかったり、ルーブリックとは異なる新たに生成された観点や配点で結果が出力されたりする場合があった。その際は、出力を中断して、再度入力をし直す必要があった。また、1回のプロンプトで連続した採点を行おうとすると、明らかに出力される情報量が減少したり、満点が連続して出力されたりする等の不具合が生じた。そのため、今回の研究では毎回ページを更新し、ファイル添付を伴うプロンプト入力を行ったため、効率性の面で課題が残った。この点については、ChatGPTに搭載されている「マイGPT」という特定用途に特化したチャットボットを作成する機能を使用することで、入力の手間を軽減することができるかもしれない。ただし、連続した入力の際には、採点精度の変化等に注視する必要がある。

なお、有料版のChatGPTにおいても、連続した入力を行うと使用上限に達して、数時間程度使

用できなくなることもあった。

以上の結果より、ChatGPTによるルーブリック生成は、現時点でも十分に利用可能なレベルに達していると言える。一方で少なくとも今回のような課題においては、実際の場面でChatGPTによる自動採点の結果をそのまま利用できるレベルには達していないと考えられた。しかしながら、現時点においても、ある評価基準でレポートを分類する場面では、一定の信頼性と妥当性が発揮できる可能性はある。

## 6. おわりに

アントレプレナーシップ教育に限らず、教育場では、何ができるかも重視され、パフォーマンス課題の使用頻度が上がることが予想される。このような中で、ルーブリックは必須のツールになると考えられる。ルーブリックは、特に初めて作成する場合には多大な時間と労力が必要である<sup>13)</sup>。しかし、本研究により、ChatGPTを活用することで、ルーブリック作成の負担が軽減できることが示唆された。一方で、学修目標や教員の意図を反映させるためには、生成されたものをそのまま受け入れるのではなく、追加のプロンプトを指示し改善をしたり、暗黙知となっている経験則を表出したりすることも重要である。

2023年に閣議決定された教育振興基本計画の5つの基本的な方針のひとつに「教育デジタルトランスフォーメーション(DX)の推進」が掲げられている<sup>14)</sup>。その中で、生成AIについては、教育現場での利用により効果をもたらす可能性と生じうるリスクを踏まえて対応することが必要であるとされている。この点を踏まえても、生成AIの教育活動における利用については継続した調査研究が欠かせないと言える。現段階では、ChatGPT採点の信頼性を考えると、そのまま自動採点に利用できるものではない。しかし、幅広いユーザーが自然言語でアクセスできる生成AIが登場後、そのインパクトは大きく、目まぐるしく発展している。そして、今後より改良されたLLMが登場してくることが予想される。

現在課題となっている、教育現場の負担軽減や、個別最適な学び推進をサポートするためのツールとして、ChatGPTを始めとするLLM環境の推移をモニターし、その都度検討と提案を行っていくことが重要である。今後もLLM採点の信頼性

を上げるための工夫について検討を行うと共に、これから出現してくる環境に合わせた調査研究を継続していきたい。

## 謝辞

本研究は崇城大学令和6年度若手重点研究の助成を受けたものです。御助言いただいた川副智行教授、御協力いただいた学生の皆さんに感謝申し上げます。

## 参考文献

- 1) Gozalo-Brizuela, R., & Garrido-Merchán, E. C. : A survey of Generative AI Applications. arXiv preprint, arXiv:2306.02781, 2024, <https://arxiv.org/abs/2306.02781>, (2024/09/24) .
- 2) Reuters : OpenAI says ChatGPT's weekly users have grown to 200 million. Reuters, <https://www.reuters.com/technology/artificial-intelligence/openai-says-chatgpts-weekly-users-have-grown-200-million-2024-08-29/>, (2024/09/24) .
- 3) Horn, E. : How to Create a Rubric with ChatGPT. TCEA, <https://blog.tcea.org/how-to-create-a-rubric-with-chatgpt/>, (2024/09/11) .
- 4) Yavuz, F., Çelik, Ö., Çelik, G., Y. : Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. British Journal of Educational Technology, 00: 1-17, 2024, <https://doi.org/10.1111/bjet.13494>, (2024/09/11) .
- 5) Mizumoto, A., Eguchi, M. : Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics, 2 (2) : 100050, 2023, <https://www.sciencedirect.com/science/article/pii/S2772766123000101>, (2024/09/11) .
- 6) 笠原秀浩・高橋純：文章生成 AI を活用した児童の自由記述からの指導・助言生成の試み，日本教育工学会研究報告集，135-140, 2023.
- 7) 脇谷伸：生成 AI によるレポート評価支援システム設計に関する一検討．第 68 回システム制御情報学会研究発表講演会，24-26, 2024.
- 8) 文部科学省：アントレプレナーシップ教育の現状について．資料 1 科学技術・学術審議会産業連携・地域振興部会（第 2 回），[https://www.mext.go.jp/content/20210728-mxt\\_sanchi01-000017123\\_1.pdf](https://www.mext.go.jp/content/20210728-mxt_sanchi01-000017123_1.pdf) (2024/09/19 閲覧)
- 9) 牧野恵美：特集・アントレプレナーシップ教育プログラム海外における起業家教育の先行研究レビュー．研究 技術 計画，33 (2) : 92-100, 2018.
- 10) Meyer, J., Jansen, T., Fleckenstein, J., Keller, S., & Köller, O. : Machine learning im Bildungskontext: Evidenz für die Genauigkeit der automatisierten Beurteilung von Essays im Fach Englisch. Zeitschrift Für Pädagogische Psychologie, 37 (3) : 203-214, 2023.
- 11) 栗田佳代子・中村長史（編）・日本教育研究イノベーションセンター（協力）：インタラクティブ・ティーチング 実践編 3 学びを促す評価ールーブリックの作法と事例ー，河合出版，2024.
- 12) 渡辺雅子：納得の構造：日米初等教育に見る思考表現のスタイル，東洋館出版，2006.
- 13) ダネル・スティーブンス，アントニア・レビ，佐藤浩章（監修，翻訳），井上敏憲（翻訳），俣野秀典（翻訳）：大学教員のためのルーブリック評価入門，玉川大学出版，2014.
- 14) 日本政府：教育振興基本計画 令和 5 年 6 月 16 日閣議決定．文部科学省，[https://www.mext.go.jp/content/20230615-mxt\\_soseisk02-100000597\\_01.pdf](https://www.mext.go.jp/content/20230615-mxt_soseisk02-100000597_01.pdf), (2024/09/19) .