ISSN 2186 − 3989

ローカル大規模言語モデルを用いたレポート自動採点システムの開発と再現性の検証

田部田　晋、篠原　史成、斎藤　英明、丸山　洋三

Development and Reproducibility Verification of an Automated Report Grading System Using Local Large Language Models

Shin Tabeta, Fuminari Shinohara, Hideaki Saito and Yozo Maruyama

北 陸 大 学 紀 要
第59号(2025年9月)抜刷

北陸大学紀要
第 59 号（2025 年度）
pp.107 ～ 120
［原著論文］

ローカル大規模言語モデルを用いたレポート自動採点システムの開発と再現性の検証

田部田　晋*※、篠原　史成*、斎藤　英明*、丸山　洋三*

Development and Reproducibility Verification of an Automated Report Grading System Using Local Large Language Models

Shin Tabeta*, Fuminari Shinohara*, Hideaki Saito* and Yozo Maruyama*

Received July 25, 2025
Accepted August 5, 2025

抄録

大学教員の教育・管理業務の増大に伴う研究時間の減少は、日本の高等教育における喫緊の課題である。特にレポート採点業務は教員に大きな負担を強いており、その効率化が求められている。本研究は、この課題に対し、学生のデータプライバシーを保護しつつ業務負担を軽減するため、学内 PC 上で完結するローカル大規模言語モデル（LLM）を用いたレポート自動採点システムを開発し、その採点の再現性を定量的に検証することを目的とした。

システムは Ollama フレームワークを用いて構築し、日本語処理に優れた 3 つの LLM（Llama-3-ELYZA-JP-8B、Llama-3-ELYZA-JP-70B、DeepSeek-Coder-V2）を使用した。実際の授業で提出されたレポート 46 件に対し、教員が作成したルーブリックに基づき、各 LLM が 10 回ずつ繰り返し自動採点を行った。この 10 回の試行結果内の一致度（再現性）を、評価者間信頼性係数である Fleiss' κ 統計量および Gwet's AC1 統計量を用いて評価した。

結果、各モデルの 10 回の採点における平均一致率は 0.87 と高かった。しかし、Fleiss' κ 統計量の平均は 0.36 と低い値を示した。これはデータの偏りに起因する統計的特性によるものであり、偏りの影響を受けにくい Gwet's AC1 統計量の平均は 0.80 と高い値を示した。このことから、本システムは実質的に安定した再現性の高い評価を行うことが示された。特に、基準が明確な観点別評価では、AC1 統計量が 0.98 を超えるなど、極めて高い信頼性が確認された。

本研究により、開発したローカル LLM システムは、採点業務において再現性の高い支援ツールとして機能しうることが実証された。本システムは教員の定型的な評価作業を代替することで、教員がより本質的な教育活動に注力する時間を創出し、教育の質の向上に貢献することが期待される。

Key Words（キーワード）： 大規模言語モデル（LLM）、レポート評価、自動採点システム

* 北陸大学経済経営学部　Faculty of Economics and Management, Hokuriku University
※責任著者　田部田晋 Shin Tabeta　s-tabeta@hokuriku-u.ac.jp

はじめに

近年、日本の大学教員を取り巻く状況は大きな転換期を迎えている。文部科学省の調査では、大学教員の研究時間は減少傾向にあり、教育活動やその他の管理運営業務が増加していることが指摘されている（文部科学省、2022）。研究時間の減少は平成 14 年度から継続しており、教員が本来注力すべき研究活動に十分な時間を割けない状況が常態化している。この問題は日本に限定されるものではなく、18 ヶ国を対象とした国際比較調査においても、研究と教育の両立に困難を感じていると調査に参加した教員の約半数が回答しており、国際的な課題であることが示唆される（有本、2008）。

このような状況下で、大学教員が限られた時間の中で質の高い研究と教育を両立させるためには、業務の効率化が不可欠である。特に、レポートや記述式試験の採点、および個別のフィードバック作成は、教育の質に直結する重要な業務でありながら、教員に多大な時間的負担を強いている。積み上げ型の評価方式を採用している授業が多い北陸大学では、学生一人当たりのレポート提出量が膨大になる傾向があり、教員の採点業務負担は深刻な課題となっている。

自動採点システムの歴史は古く、1960 年代のマークシート方式から、プログラミング課題や小論文の自動評価へと進化してきた（Hollingsworth, 1960; Page, 1966）。従来のシステムは、テキストの表層的な類似性や事前に定義されたルールに基づき、学生の答案と模範解答を比較する手法が主流であった（Heilman and Madnani, 2013; Liu et al., 2019）。

しかし、近年の大規模言語モデル（Large Language Model: LLM）の発展は、自動採点システムに新たな発展をもたらした。BERT（Devlin et al., 2019）や GPT-4（Achiam et al., 2023）に代表される LLM は、文章の文脈や意味を深く理解する能力を有しており、プロンプトエンジニアリングを適用することで、追加の事前学習なしに高度な評価タスクを実行可能とする。Mizumoto and Eguchi (2024) は、LLM にルーブリックと学生の小論文を直接入力し、得点を予測させるゼロショット手法の有効性を報告している。

一方で、ライティング能力の向上には、教員からの質の高いフィードバックが不可欠である（野瀬ら、2022; 北澤ら、2010）。西口（2016）は、従来のフィードバックが誤字脱字といった形式面の指摘に偏りがちであったことを指摘し、今後は内容の妥当性や論理構成といった、より深い内容面に関する指導の重要性を強調している．個別のフィードバックは、学生の批判的思考態度を醸成し（神山 and 藤原、1991）、内発的学習動機を高める上で極めて重要である（Hattie and Timperley, 2007）.

これらのことから本研究は、LLM 技術を活用し、大学のレポート採点業務を支援する自動採点システムを開発することを目的とする．また、クラウドサービスを介さず、学内ネットワークで完結するローカル LLM を用いることで、学生の個人情報や成果物といった機微な情報を外部に送信することなく、セキュリティとプライバシーを確保可能なシステムとする.

本システムは、教員が設定したルーブリックに基づき、レポートの自動採点と個別フィードバックコメントの生成をおこなう．これにより、教員の採点業務負担を軽減し、迅速かつ個別化されたフィードバックを学生に提供することで、教育の質向上に貢献することを目指す．本稿では、開発したシステムの採点精度と再現性を、複数の LLM モデルを用いて定量的に評価し、その有効性と課題を考察する.

方法

システム構成

本研究で開発する自動採点システムは、セキュリティとデータプライバシーを最優先に考慮し、外部のクラウドサービスを利用せず、学内 PC 上のローカル環境で LLM を動作させる構成とした．これにより、学生のレポートデータが外部に送信されることや、LLM の学習データとして利用されるリスクを完全に排除する．

LLM の実行環境には、多様なモデルをローカルで容易に管理・実行できるフレームワークである Ollama を用いた．評価に使用する LLM には、日本語処理能力が高いとされる複数のモデル、Llama-3-ELYZA-JP-8B、Llama-3-ELYZA-JP-70B、DeepSeek-Coder-V2 を採用した．ELYZA モデルは、Meta 社の Llama-3 をベースに日本語の追加事前学習を施したモデルであり、国内の日本語性能ベンチマークで高いスコアを記録している．DeepSeek は、GPT-4o や GPT-o1 を上回る性能と低い計算コストを実現したモデルである．

システムの操作インターフェース（GUI）は、Python のフレームワークである Flet を用いて開発した（図 1、2）．GUI は「評価基準」「学生回答」「実行」「結果」のタブで構成され、教員が直感的に操作できる設計とした．LLM を搭載する PC は G-Tune (mouse)(CPU: 12th Gen Intel Core i9-12900KF、RAM: 32GB、GPU: NVIDIA GeForce RTX 3070、OS: Windows 11 Home）とした。

本システムは以下の 4 ステップで採点を実行する。まずは教員が採点基準となるルーブリック（観点別の評価基準、配点、減点項目）をテキスト形式でシステムに入力する。次に学生が提出したレポートのテキストデータをシステムに入力する。その後、LLM が入力されたルーブリックに基づき、レポートを評価する。このとき、LLM は詳細な評価指示を記述したシステムプロンプトを用いて、評価結果として観点ごとの点数と評価コメントを生成する。 評価結果は、GUI の結果タブもしくは、CSV ファイルとして出力される。

LLM の応答における再現性を確保しつつ、フィードバックコメントの多様性を許容するため、応答のランダム性を制御するパラメータである temperature は 0.2 に設定した。

LLM への指示（プロンプト）設計

LLM への評価指示は、評価の一貫性と再現性を高めるために役割、評価基準、手順、制約条件、出力形式を明確に定義したシステムプロンプトによっておこなう。観点別評価と減点項目評価では、プロンプトテンプレートを使用した（図 3、4)。レポートの入力は、Python にて算出した文字数とともに入力をおこなう（図 5)。

図 1 GUI 評価基準 （左）観点、(右）減点項目

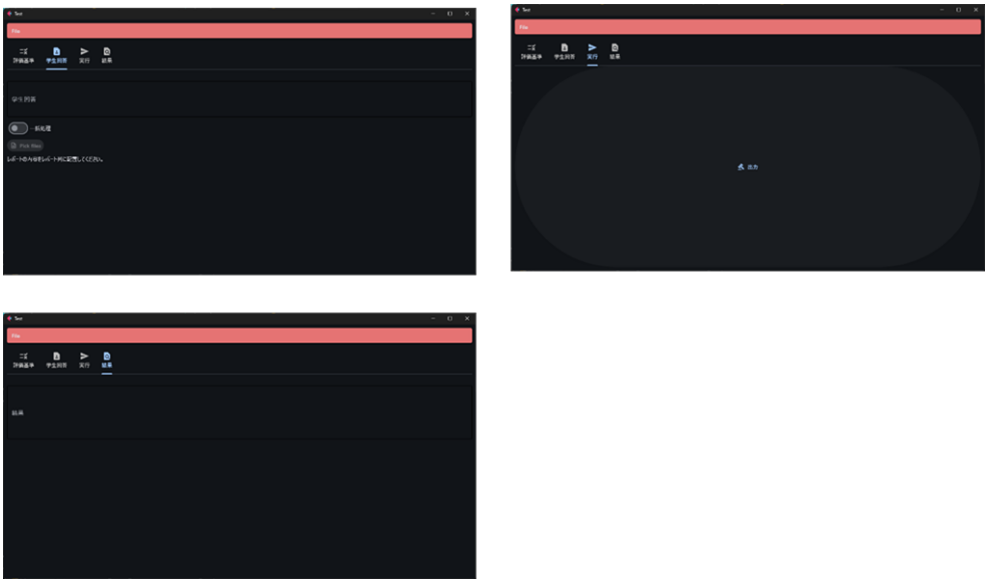

図 2 GUI（左上）学生回答、（右上）実行、（左下）結果

```
#指示：下記の評価観点でレポートの採点をおこなってください。
#役割：大学の教員

観点1: 観点1の説明
2点: 2点の説明
1点: 1点の説明
0点: 0点の説明

#手順：
観点の点数を算出し、評価コメントを作成してください。
観点の点数、評価コメントを表示してください。

#制約条件：
点数には小数点を含めないように採点してください。

#出力形式：
下記のように結果を出力してください。
##観点１：
点数：
評価コメント:
```

図 3　観点ごとの System Prompt（観点 1 の例）

#指示：下記の評価観点でレポートの採点をおこなってください。
#役割：大学の教員

減点項目：
〇〇
〇〇
（具体的な減点項目をリストアップ）

#手順：
減点項目から該当するものを見つけ出してリストアップしてください。このとき、項目は重複しないものとします。
該当項目を表示してください。該当しない項目は表示しない。

#出力形式：
下記のように結果を出力してください。
##減点項目：
該当項目:

図 4　減点項目の System Prompt

#レポートの文字数：〇〇
#レポートの本文:
〇〇〇〇〇〇〇〇〇〇〇〇〇〇〇〇〇〇〇〇〇〇〇〇〇
〇〇〇〇〇〇〇〇〇〇〇〇〇〇
（学生のレポート本文）

図 5　ユーザーへの Question（入力例）

表 1　ルーブリック

観点	2点（すばらしい）	1点（よい）	0点（もうちょっと）
観点1 協働とは何か	協働とは何か具体的に説明している。	協働とは何か説明しているが具体的でない。	協働とは何か説明していない。
観点2 協働に必要なこと	協働に必要なことを具体的に説明している。	協働に必要なことを説明しているが具体的でない。	協働に必要なことを説明していない。

表 2　減点項目

	減点項目
1	複数の段落で文書を作成していない（段落が分けられていない）
2	各段落の冒頭を全角 1 字分空けて書いていない※段落冒頭に字下げがない場合は、減点対象となる
3	段落と段落の間に空行（行頭に改行があり何の文字も含まない行）を作っている
4	文体が「ですます調」になっている（「である調」にしていない）※「ですます調」の文があまりにも多すぎる場合は、大幅に減点される
5	「話し言葉」で書いている（「書き言葉」で書いていない）
6	誤字・脱字が目立つ、「てにをは」の誤り（助詞の使い方の誤り）が目立つ
7	レポート答案に不適切な一人称を使っている
8	レポート答案に不適切な記号を使っている
9	不必要な「と思う」や「と考える」という表現がある
10	一文一義にしていない、1 つの文が長すぎるために内容が理解しづらい
11	主語と述語が整合していない（主述のねじれがある、主語が不明確、述語がない）
12	接続表現が文脈に合っていない

評価実験

開発したシステムの採点精度を検証するため、北陸大学経済経営学部の基礎ゼミナールで実際に提出されたレポート 46 件を対象に評価実験を行った。

レポートとルーブリックを本システムに入力し、3 種類の LLM（Llama-3-ELYZA-JP-8B、Llama-3-ELYZA-JP-70B、DeepSeek）に繰り返し 10 回評価させた。

各モデルにおける 10 回の評価の一致度の評価には、評価者間信頼性係数として Fleiss' κ (kappa)統計量（Fleiss, 1971）と Gwet's AC_1 統計量（Gwet, 2008）を用いた。κ 統計量は偶然の一致を除外して一致度を測る代表的な指標であるが、評価カテゴリの分布に偏りがある場合に値が著しく低くなる問題が知られている。AC_1 統計量は、この問題に対応するために提案された指標であり、より安定した評価が可能であるとされる（西浦、2010)。本研究では、両指標を併用することで評価結果の一致度を評価する。

結果

すべての LLM における一致率の平均は 0.87、κ 統計量の平均値は 0.36、AC_1 統計量は 0.80 と κ 統計量は低い値であったが高い一致率と AC1 統計量を示した。観点ごと、減点項目における一致率、評価値の平均値、κ 統計量、AC_1 統計量を示す（表 3、4、5、6、7)。

観点ごとの評価コメントの例を示す（表 8、9)。

観点別評価では、全モデルで高い再現性が示された。Llama-3-ELYZA-JP-8B は観点 1 において、10 回の評価が完全に一致し、κ 統計量 1.0、AC_1 統計量 1.0 であった。観点 2 においても各モデルは高い再現性を示した。減点項目の評価では、κ 統計量は低い値に留まったが、AC1 統計量は高い水準を維持した。

異なるモデル間の評価結果を比較したところ、評価傾向に顕著な違いが見られた(表 10、11)。観点 1 において、Llama-3-ELYZA-JP-70B と DeepSeek 間の一致度は AC_1 統計量が 0.83、κ 統計量が 0.77 と高く、両モデルが類似した評価を行ったことが示された。しかし、Llama-3-ELYZA-JP-8B と他の 2 モデルとの間の一致度は AC_1 統計量、κ 統計量ともに低い値であった。観点 2 においても、Llama-3-ELYZA-JP-70B と DeepSeek 間の一致度は AC_1 統計量が 0.98、κ 統計量が 0.82 と高い一致度を示した。一方で、Llama-3-ELYZA-JP-8B と他の 2 モデルとの間の一致度は AC_1 統計量、κ 統計量ともに低い値であった。これは、両者の間で評価が偶然以下の一致度でしかなく、むしろ逆の評価傾向が存在することを示唆している。

観点別評価とは対照的に、減点項目におけるモデル間の一致率は、いずれのモデルペアにおいても平均 0.79 と、比較的一貫していた（表 12、13、14)。

評価コメントの質的分析では、3 モデルとも指摘事項を適切に説明し、観点 1、2 の両方で改善案を提示していた（表 8、9)。

表 3　観点 1 における一致度

			Fleiss' kappa		AC1	
	一致率	平均値	kappa	P値	AC1統計量	P値
Llama-3-ELYZA-JP-8B	1.00	1.96	1.00	0.00	1.00	0.00
Llama-3-ELYZA-JP-70B	0.86	1.47	0.73	0.00	0.81	0.00
deepseek	0.87	1.45	0.76	0.00	0.83	0.00

表 4　観点 2 における一致度

			Fleiss' kappa		AC1	
	一致率	平均値	kappa	P値	AC1統計量	P値
Llama-3-ELYZA-JP-8B	0.98	1.93	0.82	0.00	0.98	0.00
Llama-3-ELYZA-JP-70B	0.98	0.94	0.80	0.00	0.97	0.00
deepseek	0.98	0.94	0.87	0.00	0.98	0.00

表 5 減点項目における一致度（Llama-3-ELYZA-JP-8B）

			Fleiss' kappa		AC_1	
	一致率	平均値	κ 統計量	P値	AC_1統計量	P値
1	1.00	1.00	1.00	0.00	1.00	0.00
2	1.00	1.00	1.00	0.00	1.00	0.00
3	0.99	1.00	0.00	0.16	0.99	0.00
4	1.00	1.00	1.00	0.00	1.00	0.00
5	0.98	0.99	-0.01	0.05	0.98	0.00
6	1.00	1.00	0.00	0.32	1.00	0.00
7	0.79	0.86	0.15	0.01	0.73	0.00
8	0.82	0.15	0.31	0.00	0.76	0.00
9	0.71	0.78	0.14	0.00	0.55	0.00
10	0.60	0.49	0.21	0.00	0.21	0.00
11	0.62	0.60	0.22	0.00	0.27	0.00
12	0.62	0.59	0.22	0.00	0.27	0.00

表 6 減点項目における一致度（Llama-3-ELYZA-JP-70B）

			Fleiss' kappa		AC_1	
	一致率	平均値	κ 統計量	P値	AC_1統計量	P値
1	0.98	0.99	0.05	0.16	0.98	0.00
2	1.00	1.00	1.00	0.00	1.00	0.00
3	0.84	0.88	0.21	0.05	0.79	0.00
4	0.98	0.99	0.05	0.16	0.98	0.00
5	1.00	1.00	1.00	0.00	1.00	0.00
6	0.94	0.96	0.16	0.00	0.93	0.00
7	0.60	0.59	0.18	0.00	0.23	0.00
8	0.70	0.23	0.17	0.00	0.54	0.00
9	0.93	0.96	0.15	0.00	0.92	0.00
10	0.81	0.88	0.14	0.00	0.76	0.00
11	0.93	0.96	0.15	0.00	0.93	0.00
12	0.71	0.80	0.11	0.00	0.58	0.00

表 7　減点項目における一致度（deepseek）

			Fleiss' kappa		AC_1	
	一致率	平均値	κ統計量	P値	AC_1統計量	P値
1	0.98	0.99	0.10	0.00	0.98	0.00
2	0.99	1.00	0.11	0.00	0.99	0.00
3	0.86	0.90	0.22	0.06	0.82	0.00
4	0.98	0.99	0.10	0.00	0.98	0.00
5	1.00	1.00	1.00	0.00	1.00	0.00
6	0.94	0.96	0.16	0.00	0.93	0.00
7	0.60	0.59	0.18	0.00	0.23	0.00
8	0.70	0.23	0.17	0.00	0.54	0.00
9	0.93	0.96	0.15	0.00	0.92	0.00
10	0.81	0.88	0.14	0.00	0.76	0.00
11	0.93	0.96	0.15	0.00	0.93	0.00
12	0.71	0.80	0.11	0.00	0.58	0.00

表 8 観点 1 の評価コメントの例

		観点1
	点数	評価コメント
Llama-3-ELYZA-JP-8B	0	協働とは何か具体的に説明していないため、点数は0点です。レポートでは「私は協働とは競争であると考える」という主張がなされていますが、協働の定義や概念についての説明が不足しています。協働を理解する上で、まずはその意味や特徴を正しく把握する必要があります。
Llama-3-ELYZA-JP-70B	1	レポートには「協働とは競争である」という主張が記述されていますが、その根拠となる具体的な事例や説明が不足しています。理想とするリーダー像については、競争を促進するような環境作りをするリーダーであると推測できますが、それ以上の具体的な描写がありません。 より具体的に理想とするリーダー像を記述することで、評価点数を2に引き上げることが可能です。例えば、 * 競争を促進するためにどのような施策を講じるのか * チームメンバーの意見を尊重しながらも、建設的な議論を導くための具体的な方法 * 失敗を恐れずに挑戦できる雰囲気作りについて などを記述することで、より説得力のあるレポートになります。
deepseek	0	レポート本文には、自分の理想とするリーダー像についての記述は見当たりません。

表 9　観点 2 の評価コメントの例

	観点2	
	点数	評価コメント
Llama-3-ELYZA-JP-8B	0	協働に必要なことについての説明が不足しています。具体的な例や方法は提示されていますが、協働に必要な要素や心構えなどを明確に説明することができていません。したがって、この観点では0点と評価します。
Llama-3-ELYZA-JP-70B	1	レポートでは、協働における競争の重要性について論じていますが、理想とするリーダー像になるための具体的な行動は記述されていません。 例えば、「より良い意見を考え、ほかの人の話をもとに改良し続けることが大事」と述べていますが、どのように改良していくのか、具体的な方法や努力について言及がありません。 理想とするリーダー像になるためには、 * **具体的な行動**: 例えば、「積極的に意見交換に参加する」「他者の意見を丁寧に聞き、理解しようと努める」「自分の意見を明確に伝え、根拠を提示する」など * **目標設定**: どのようなリーダーになりたいのか、そのために必要なスキルや経験は何かを具体的に示す といった要素が重要です。 レポートでは、これらの点についてより深く掘り下げることで、理想とするリーダー像への道筋が明確になるでしょう。
deepseek	1	レポートでは、協働における競争の重要性について論じていますが、理想とするリーダー像になるための具体的な行動は記述されていません。 例えば、「より良い意見を考え、ほかの人の話をもとに改良し続けることが大事」という点は、リーダーシップを発揮する上で重要な要素ですが、それをどのように実践していくのか具体的な方法論が示されていません。 理想とするリーダー像になるためには、 * どのような行動を心掛けるべきか * 具体的にどのようなスキルを磨く必要があるか * 周囲の人々とどのようにコミュニケーションを取るべきか など、具体的な行動指針が示されることが望ましいです。

表 10　観点 1 におけるモデル間の一致度

		Fleiss' kappa		AC1	
	一致率	κ統計量	P値	AC_1統計量	P値
Llama-3-ELYZA-JP-8B vs Llama-3-ELYZA-JP-70B	0.66	-0.05	0.16	0.59	0.00
Llama-3-ELYZA-JP-8B vs DeepSeek	0.65	-0.04	0.27	0.59	0.00
Llama-3-ELYZA-JP-70B vsDeepSeek	0.88	0.77	0.00	0.83	0.00

表 11　観点 2 におけるモデル間の一致度

		Fleiss' kappa		AC1	
	一致率	κ統計量	P値	AC_1統計量	P値
Llama-3-ELYZA-JP-8B vs Llama-3-ELYZA-JP-70B	0.03	-0.80	0.00	-0.33	0.00
Llama-3-ELYZA-JP-8B vs DeepSeek	0.03	-0.80	0.00	-0.33	0.00
Llama-3-ELYZA-JP-70B vsDeepSeek	0.98	0.82	0.00	0.98	0.00

表 12　減点項目における Llama-3-ELYZA-JP-8B と Llama-3-ELYZA-JP-70B の一致度

		Fleiss' kappa		AC1	
	一致率	κ統計量	P値	AC_1統計量	P値
1	0.99	0.00	0.05	0.99	0.00
2	1.00	1.00	0.00	1.00	0.00
3	0.88	-0.06	0.00	0.86	0.00
4	0.99	0.00	0.05	0.99	0.00
5	0.99	0.00	0.05	0.99	0.00
6	0.96	-0.02	0.00	0.96	0.00
7	0.58	-0.05	0.30	0.30	0.00
8	0.72	0.09	0.06	0.59	0.00
9	0.76	-0.06	0.15	0.69	0.00
10	0.50	-0.15	0.00	0.12	0.02
11	0.58	-0.21	0.00	0.36	0.00
12	0.57	-0.01	0.77	0.26	0.00

表 13　減点項目における Llama-3-ELYZA-JP-8B と DeepSeek の一致度

		Fleiss' kappa		AC1	
	一致率	κ統計量	P値	AC_1統計量	P値
1	0.99	0.00	0.05	0.99	0.00
2	1.00	0.00	0.16	1.00	0.00
3	0.89	-0.06	0.00	0.88	0.00
4	0.99	0.00	0.05	0.99	0.00
5	0.99	0.00	0.05	0.99	0.00
6	0.97	0.10	0.35	0.97	0.00
7	0.55	-0.08	0.07	0.23	0.00
8	0.72	0.10	0.04	0.60	0.00
9	0.76	-0.05	0.19	0.69	0.00
10	0.50	-0.17	0.00	0.13	0.02
11	0.58	-0.23	0.00	0.37	0.00
12	0.59	0.01	0.86	0.29	0.00

表 14　減点項目における Llama-3-ELYZA-JP-70B と DeepSeek の一致度

		Fleiss' kappa		AC1	
	一致率	κ統計量	P値	AC_1統計量	P値
1	0.99	-0.01	0.00	0.98	0.00
2	1.00	0.00	0.16	1.00	0.00
3	0.88	0.14	0.02	0.79	0.00
4	0.99	-0.01	0.00	0.98	0.00
5	0.99	1.00	0.00	1.00	0.00
6	0.96	0.08	0.28	0.93	0.00
7	0.58	0.13	0.01	0.16	0.00
8	0.72	0.19	0.00	0.56	0.00
9	0.76	0.12	0.14	0.92	0.00
10	0.50	0.24	0.00	0.81	0.00
11	0.58	0.09	0.27	0.93	0.00
12	0.57	0.17	0.00	0.62	0.00

考察

本研究で開発したローカル LLM 自動採点システムは、平均的に高い一致率を示した。κ統計量は低い値であったが、AC_1統計量は高い値であった。そのため、本システムが実質的に高い信頼性で採点を行っていることが示唆された。特に、明確な基準が示されたルーブリックの観点別評価（観点 1、2）では、κ統計量、AC_1統計量ともに極めて高い値を示し、LLM が安定的かつ再現性の高い評価をおこなうことができる能力を持つことが実証された。これは、同一のレポートに対して常に一貫した評価を下せる可能性を示しており、教員間の評価のばらつきを抑制する上でも有用である。

減点項目の評価においてκ統計量が低かった一因は、前述のデータの偏りに加え、評価の難しさにあると考えられる。減点項目には、「誤字・脱字」「不適切な表現」といった、文脈上の微妙なニュアンスの判断を要するものが含まれる。LLM がこれらの細かな点を100%正確に検出することは依然として課題であり、モデルによって検出能力にばらつきが見られた。そのため、LLM を使用することを考慮したルーブリック作成が必要であることが示唆された。

しかしながら、AC_1統計量では高い値が維持されたことから、実用上は多くの項目を適切に検出できていると解釈できる。現状のシステムでも、学生へのフィードバック作成の第一稿として十分に機能し、教員がゼロから全ての誤りを指摘する手間を大幅に削減できると考えられる。教員は LLM が検出した項目を確認・修正し、AI が見逃した点を補うという協調的な運用が現実的であろう。

モデル間の比較では、Llama-3-ELYZA-JP-70B と DeepSeek が類似した評価傾向を示したのに対し、Llama-3-ELYZA-JP-8B は大きく異なる評価を行った。特に観点 2 で見られた負の相関は、モデルの規模やアーキテクチャの違いが、プロンプトの解釈や評価の内部基準に根本的な差異を生じさせた可能性を示唆する。小規模な 8B モデルがプロンプトの指示をより厳格に、あるいは字義通りに解釈した結果、大規模モデルとは異なる評価軸で判断した可能性が考えられる。一方で、減点項目の評価ではモデル間の不一致が比較的小さかったことから、客観的でルールベースに近い評価ではモデル間の差は出にくいが、より解釈を要する評価ではモデルの特性が顕著に現れることが示唆される。この結果は、自動採点においてモデル選定が評価結果そのものを大きく左右する重要な要素であること

を示している。

評価コメントの質に関しては、全モデルが改善案を提示する能力を示した。しかし、西口（2016）が指摘するような、学生の思考プロセスに踏み込む深いフィードバックの生成は依然として課題であり、今後の研究では、生成されるコメントの質を教員のそれと比較・検証する必要がある。

本研究は、ローカル LLM を用いたレポート自動採点システムを開発し、その採点が高い再現性を持つことを実証した。本システムは、教員の採点業務負担を軽減し、教育の質向上に貢献する有効な支援ツールとなりうる。モデルによって評価傾向が大きく異なるという知見は、今後のシステム設計やモデル選定における重要な指針となる。今後は、フィードバックの質的向上や、多様な課題形式への対応を進め、より実用的な教育支援システムの実現を目指す。

謝辞

本研究は北陸大学特別研究助成（学内環境向上システムの開発、代表；田部田晋）を受けたものである。

参考文献

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & Zoph, B. (2023). Gpt-4 technical report. *arXiv*. https://arxiv.org/abs/2303.08774.

有本，章. (2008). 変動する大学教授職. *玉川大学出版部*.

DeepSeek-AI. (2024). DeepSeek-V2 technical report. *arXiv*. https://arxiv.org/abs/2405.04434.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.

Gwet, K. L. (2008). Computing inter‐rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1), 29-48. https://doi.org/10.1348/000711006X126600.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. https://doi.org/10.3102/003465430298487.

Heilman, M., & Madnani, N. (2013). ETS: Domain adaptation and stacking for short answer scoring. *In *Second Joint Conference on Lexical and Computational Semantics (SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 275–279. Association for Computational Linguistics.

Hollingsworth, J. (1960). Automatic graders for programming classes. *Communications of the ACM*, 3(10), 528–529.

神山，貴弥，& 藤原，武弘. (1991). 認知欲求尺度に関する基礎的研究. *社会心理学研究*, 6(3), 184-192.

北澤，武，永井，正洋，& 上野，淳. (2010). 大学情報教育のブレンディッドラーニング環境における e ラーニングシステムを用いたフィードバックの効果. *日本教育工学会論*

文誌, 34(1), 55-66.

Liu, X., Wang, S., Wang, P., & Wu, D. (2019, May). Automatic grading of programming assignments: An approach based on formal semantics. I*n 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET)*, 126–137. IEEE.

Messer, M., Brown, N. C., Kölling, M., & Shi, M. (2024). Automated grading and feedback tools for programming education: A systematic review. *ACM Transactions on Computing Education*, 24(1), 1–43.

Mizumoto, A., & Eguchi, M. (2023). Exploring the feasibility of ChatGPT, GPT-3.5, and GPT-4 for automated scoring of English essays. *Journal of Educational Technology & Society*, 26(4), 1-15.

株式会社ブレインアカデミー. (2022). 「大学教員の勤務実態に係る調査研究」調査報告書. https://www.mext.go.jp/content/20221124-mxt_daigakuc01-000026106_1.pdf.

西口, 啓太. (2016). アメリカ合衆国におけるアカデミックライティングとその評価―ライティングを通じた学生の情緒的発達と評価活動としてのフィードバックに着目して―. *初年次教育学会誌*, 8(1), 157-165.

西浦, 博. (2010). 観察者間の診断の一致性を評価する頑健な統計量 AC1 について. *日本放射線技術学会雑誌*, 66(11), 1485-1491.

野瀬, 由季子, 三井, 規裕, 福山, 佑樹, 西口, 啓太, & 時任, 隼平. (2022). 教員からのフィードバックを踏まえた文章執筆における初年次学生の修正の傾向―成績下位群の文章の変化に着目して―. *日本教育工学会研究報告集*, 2022(4), 320-324.

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5), 238–243.

Ureel II, L. C., & Wallace, C. (2019, February). Automated critique of early programming antipatterns. *In Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, 738–744. Association for Computing Machinery.