

研究ノート

生成 AI を活用したレポート自動採点における プロンプト構築の研究

薦 田 勇 智

要旨

本研究では、生成 AI の自然言語処理能力を教育分野で活用する可能性を探るため、大学レポートの採点に適したプロンプト設計に焦点を当てた。生成 AI は 2018 年頃から急速に注目され、多くの分野で応用が進められている。特に ChatGPT のようなモデルは、対話生成や文章作成などの分野で高い能力を示しているものの、生成される内容の精度や適切性には課題が残る。そのため、生成 AI の利用効果を最大化するには、適切なプロンプト設計や評価基準の策定が不可欠である。本研究では、プロンプト設計の重要性を踏まえ、教育分野における応用可能性を検討した。

1. はじめに

2018 年頃から、ChatGPT を始めとした生成 AI が注目され、その流行は大きな話題を呼んだ。それ以降、生成 AI 技術は急速に進歩し、現在では多くの分野で活用されている。例えば、医療や教育、ビジネスにおけるデータ分析やクリエイティブなコンテンツの生成といった多岐にわたる用途が挙げられる。

しかし、その一方で、生成 AI が扱う自然言語の精度や適切性にはまだ課題が残されており、さらなる研究が必要である。特に、自然言語処理の応用範囲が拡大する中で、生成された内容の評価基準をどのように設計するかが重要な課題となっている。

また、生成 AI を効果的に活用するためには「プロンプト」の設計が鍵を握っており、プロンプト技師という専門職が登場するほど、その重要性が認識されている。プロンプトは、生成 AI が適切かつ有用な応答を出力するための指示文であり、その内容や構造が生成結果に大きな影響を及ぼ

す。

本研究では、生成 AI の自然言語処理能力を評価する一環として、本学のレポートを採点するためのプロンプト構築に焦点を当てる。この取り組みにより、教育分野における生成 AI の可能性をさらに引き出し、効率的かつ公平な評価手法の実現を目指す。

2. レポートの評価基準

レポートを採点する際に、基準になるものが必要である。コーパスというソフトウェアを利用し、レポートの自動採点を行おうとした山本 *et. al* (2017) では、「自動採点モデルの概要採点システムの中核として、5 種の評価観点 (Contents : 課題の理解度と解答内容の妥当性, Structure : 論理的な展開, Evidence : 資料と根拠の妥当性, Style : 文章作法の遵守と適切な推敲, Skill : 読みやすさ・表現の巧みさ) を持つレポート採点用ルーブリックを提案した」とある。また、一般教養科目で学生の文章力を向上させる取り組みを実践し

た和田(2016)では、「チェックリストを設け、それをレポートの評価基準にする。明確な目標を示すことが、学生に成長につながる。また教育の質を保証することができる。」とあった。これらのことから、5種の評価観点を基にチェックリストを作成し、ChatGPTにそのチェックリストに沿って採点をさせる。また、このチェックリストの妥当性を判断するために、筆者が直接採点したものと比較し、どれだけ人間の評価に近くなっているかを判断する。

3. プロンプトの作成

ここからは、実際にプロンプトの作成を行う。まずは5種の評価観点について、改めて確認する。1つ目は「Contents: 課題の理解度と解答内容の妥当性」について。これは、山本 *et. al* (2016)によると、「解答内容が、課題とは無関係である(評価D, 0-1点)。課題を理解し解答しているが、誤りがある(評価C, 2-3点)。課題を理解し解答しているが、記述が不足している(評価B, 4-5点)。課題を理解し的確な解答であるが、改善の余地がある。的確な解答である(評価A, 6-7点)。関連用語を適正に用いている。改善の必要はない(評価A+, 8-9点)。」これらをまとめると、課題の内容を正確に理解し、的確に回答できているかを採点している。特に、改善の余地があるかどうかで判断しているため、どのラインで改善させるかを決める必要があると考える。2つ目は「Structure: 論理的な展開」について。これは、山本 *et. al* (2016)によると、「記述内容にまとまりがない(評価D, 0-1点)。理論の展開に矛盾がある(評価C, 2-3点)。順序立てて理論を展開しているが、改善すべき点が複数ある(評価B, 4-5点)。順序立てて理論を展開しているが、説得力がない(評価A, 6-7点)。順序立てて理論を展開している。意見・主張があり、説得力がある(評価A+, 8-9点)。」これらをまとめると、文章を順序だてて組み立てているかどうか、文章に意見・主張・説得力があるかどうかを採点

している。3つ目は「Evidence: 資料と根拠の妥当性」について。これは、山本 *et. al* (2016)によると、「資料を全く参照していない。根拠を示していない(評価D, 0-1点)。資料を参照していないが、根拠を示そうとしている(評価C, 2-3点)。参照しようとしている資料は相応しくない、または信頼性がない(評価B, 4-5点)。信頼でき、関連性のある資料を参照しているが、引用・参照の仕方に誤りがある(評価A, 6-7点)。当該の学問分野に相応しく信頼でき関連性のある資料を、うまく使いこなしている(評価A+, 8-9点)。」これらをまとめると、関連資料を的確に引用し、その資料の信頼性を確保しているかどうかを採点している。4つ目は「Style: 文章作法の遵守と適切な推敲」について。これは、山本 *et. al* (2016)によると、「複数にわたってルールを守っていない。文章が全く推敲されていない(評価D, 0-1点)。ルールを守っていない、誤字・脱字、文体の誤りなどが複数ある(評価C, 2-3点)。大よそのルールを守っているが、訂正すべき点が複数ある(評価B, 4-5点)。訂正すべき点はないが、改善の余地がある(評価A, 6-7点)。よく推敲してある。誤りが全くない(評価A+, 8-9点)。」これらをまとめると、レポートを書く際のルールを守っているかどうかを採点している。5つ目は「Skill: 読みやすさ・表現の巧みさ」について。「文章が読み辛い。明らかに文章スキルがない(評価D, 0-1点)。文章が長すぎるなど、複数の改善すべき点がある(評価C, 2-3点)。文章が概ねまとまっているが、改善すべき点がある(評価B, 4-5点)。文章が読みやすい。語彙が豊富である(評価A, 6-7点)。読み手に明確に意味を伝えることができ、読みやすい。語彙が豊富である(評価A+, 8-9点)。」これらをまとめると、表現力、可読性があるかどうかを採点している。

以上のことをチェックリストに反映させていく。すると、図表1のようになった。

レポート自動採点チェックリスト	
評価観点	配点
課題の内容を正確に理解し、的確に回答できているか	20
文章を順序だてて組み立てているかどうか、文章に意見・主張・説得力があるかどうか	20
関連資料を的確に引用し、その資料の信頼性を確保しているかどうか	20
レポートを書く際のルールを守っているかどうか	20
表現力、可読性があるかどうか	20
合計	100

図表1 レポート自動採点チェックリスト

4. まとめ

前章では山本 *et. al* (2016) で示されていた基準を基に作成をした。また、配点は計算が楽になるよう合計が100点になるように考えた。次回の研究では、実際に採点を行い、人間が採点したものと比較を行っていく。現状考えられる改善点としては、基の基準（ループリック）のようにさらに細かいレベルを作成する必要があるのではないかと考えている。他には、配点はすべて同じ数字で良いのか（比重を変える必要はないのか）、評価観点の文言はこのままで大丈夫なのかを考察していきたい。

5. 参考文献

- 1) 山本恵, 梅村信夫, 河野浩之, (2017), コーパスを利用したレポート自動採点品質の向上, 第16回情報科学技術フォーラム, 第4分冊, 279-280,
- 2) 和田正法, (2016), レポート採点基準の開発と大学の一般教養科目で学生の文章力を向上させる取り組み, 三重大学教養教育機構研究紀要, 第1号, 37-48,
- 3) 武田俊之, (2023), 大学は生成系 AI の影響をいかに認識しているか?, 日本教育工学会研究報告集, 88-94,

Research on prompt construction in automatic report grading using generative AI

KOMODA Yuchi

Abstract

To explore the potential of using the natural language processing capabilities of generative AI in the field of education, this study focuses on designing prompts suitable for grading college reports. Generative AI has rapidly gained attention since around 2018 and is being applied in many fields. In particular, models such as ChatGPT have shown high performance in areas such as dialogue generation and writing, but the accuracy and appropriateness of the generated content remains a challenge. Therefore, in order to maximize the effectiveness of the use of generated AI, it is essential to develop appropriate prompt design and evaluation criteria. Based on the importance of prompt design, this study examined the potential applications in the field of education.