提升生成式 AI 作文評價精確度之提示詞工程

陳相州

東吳大學日本語文學系副教授

摘要

本研究探討生成式 AI 與人類評價者的評價傾向的差異，並驗證了透過提示詞工程提升其精確度的成效。分析結果發現，評價差異的主要原因，在於生成式 AI 無法從提示詞中解讀人類評價者所具備的「內隱評價標準」。因此，在將此「內隱評價標準」明確化並具體地反映於提示詞後，生成式 AI 與人類評價者的評價一致性便顯著提升。本研究結論證實，分析生成式 AI 與人類評價者的評價差異並持續改善提示詞的提示詞工程，對於提升生成式 AI 評價的精確度極為重要。

關鍵詞：生成式 AI、作文評價、人類評價者、提示詞工程、精確度

受理日期：2025 年 08 月 25 日

通過日期：2025 年 11 月 07 日

DOI：10.29758/TWRYJYSB.202512_(45).0002

Prompt Engineering for Enhancing the Accuracy of Generative AI Essay Evaluation

Chen, Shiang-Jou
Associate Professor, Department of Japanese Language and Culture, Soochow University

Abstract

This study explores the differences in evaluation tendencies between generative AI and human raters, and assesses the effectiveness of prompt engineering in enhancing evaluation accuracy. The analysis revealed that the primary source of discrepancies lies in generative AI's inability to recognize and apply the "implicit evaluation standards" that human raters naturally employ. When these implicit standards were explicitly articulated and embedded in the prompts, the consistency between AI- and human-generated evaluations improved substantially. These findings demonstrate that identifying and addressing discrepancies between AI and human ratings, coupled with the continuous refinement of prompts, is essential for advancing the accuracy of AI-based evaluation.

Keywords: Generative AI, Essay Evaluation, Human Raters, Prompt Engineering, Accuracy

生成 AI を用いた作文評価の精度向上に向けた プロンプトエンジニアリング

陳相州
東呉大学日本語学科副教授

要旨

本研究は、生成 AI と人間評価者による作文評価の評価傾向の相違を分析し、プロンプトエンジニアリングによる精度向上を検証した。分析の結果、人間評価者が持つ文脈理解などの「暗黙的な評価基準」を、生成 AI がプロンプトから解釈できないことが評価が乖離する主因だと判明した。そこで、この「暗黙の基準」を言語化してプロンプトに具体的に反映させたところ、生成 AI と人間との評価一致度は著しく向上した。結論として、両者の評価差異を分析し、プロンプトを継続的に改善するプロンプトエンジニアリングが、生成 AI による評価の精度向上に極めて有効であることを実証した。

キーワード：生成 AI、作文評価、人間評価者、プロンプトエンジニアリング、精度

生成 AI を用いた作文評価の精度向上に向けた プロンプトエンジニアリング

陳相州
東呉大学日本語学科副教授

1. はじめに

生成 AI を活用した日本語学習者の作文評価は、近年その有効性が注目されている。例えば、生成 AI による自動採点は一定の精度と信頼性を持ち、人間の評価を補完する可能性が示されている（Mizumoto & Eguchi，2023）。一方で、石井（2025）が指摘したように、生成 AI を用いた自動採点を実施する際には、どのようなプロンプトを設定するかが極めて重要となる。生成 AI による日本語学習者の作文評価に関しては、簡潔な指示よりも詳細な指示を含んだプロンプトの方が良好な結果をもたらすことを示した研究も複数ある（Mizumoto & Eguchi，2023；李，2023；李ほか，2023；陳，2025a；陳，2025b など）。しかし、陳（2025b）の研究では、詳細な指示を含むプロンプトの方が評価の一致率は高くなるものの、それでも人間評価者との評価の一致性を示す kappa 係数が低い場合が多いことが報告されている。

本研究では、生成 AI と人間評価者の作文評価の評価傾向の相違を明らかにするとともに、評価精度を向上させるためのプロンプトエンジニアリング手法を検討することを目的とする。

2. 先行研究

生成 AI に与える指示は「プロンプト」と呼ばれ、これを作成する作業は「プロンプトエンジニアリング」と定義される（李 2025）。プロンプトエンジニアリングには、生成 AI の性能を最大限に引き出すための様々なテクニックが存在する。このプロンプトエンジニアリングは生成 AI の出力品質に大きく影響することが多くの研究で

指摘されている。本節では、特に作文評価の領域において、プロンプトエンジニアリングの重要性を論じた主な先行研究を概観する。

Mizumoto & Eguchi(2023)は、TOEFL11 に収録されている 1210 件の英語作文を ChatGPT-3.5 に評価させ、その英語能力を測定した。その結果、ChatGPT-3.5 による評価は、TOEFL11 が公式に提供する評価レベルと大差ないことが明らかになった。Mizumoto & Eguchi(2023)が使用したプロンプトを以下の表 1 に示す。このプロンプトでは、評価点の設定に加え、詳細な評価基準（IELTS rubric）が提示されていることがわかる。

表 1 Mizumoto & Eguchi(2023)が使用したプロンプト

I would like you to mark an essay written by English as a foreign language（EFL）learners. Each essay is assigned a rating 0 to 9, with 9 being the highest and 0 the lowest. You don’t have to explain why you assign that specific score. Just report a score only. The essay is scored based on the following rubric. [*IELTS rubric in a plain text format.*] ESSAY: [*Inserting each of the 12,100 essays using a for loop in Python code.*]

（Mizumoto & Eguchi2023：5）

また、Coyne et al.(2023)は、GPT-3.5 と GPT-4 の文法誤り訂正能力について調査した。この研究では、文全体の修正に焦点を当て、元の文の意味を保持したまま、いかに文法的な誤りを修正できるかが検証された。プロンプトエンジニアリングの観点から、ゼロショット(zero shot)やフューショット(few shot)設定を含む複数のプロンプトを設計し、モデルのパフォーマンスを最大化する最適な形式を模索した。その結果、プロンプトの内容が GPT の文法修正能力に大きく影響し、特に 2 つの修正例を含む「ツーショットプロンプト

(two shot prompt)」が最も効果的であったと報告している。研究で示された最適なプロンプトは以下の通りである。

表 2 Coyne et al.(2023)が示した作文修正に最適なプロンプト

Reply with a corrected version of the input sentence with all grammatical and spelling errors fixed. If there are no errors, reply with a copy of the original sentence. Input sentence: I think smoke should to be ban in all restarants. Corrected sentence: I think smoking should be banned at all restaurants. Input sentence: We discussed about the issu. Corrected sentence: We discussed the issue. Input sentence: {*x*} Corrected sentence:

（Coyne et al. 2023：5）

表 2 が示すように、Coyne et al.(2023)は、最適な修正結果を得るためには、明確な修正指示、2 つの修正例（two shot)、そして入力文（Input sentence:）と修正文（Corrected sentence:）のような明確なフィールド区分を提供することを推奨している。

日本語の作文評価に生成 AI を活用した研究としては、李(2023)、陳（2025a)、陳（2025b）が挙げられる。李(2023)は、日本語学習者が作成した 100 編の意見文を ChatGPT に評価させ、GPT モデルとプロンプトの違いが評価結果に与える影響を分析した。その結果、プロンプトに評価の観点を明記することで、評価の精度が向上することを示した。同様の観点から、陳（2025a）と陳（2025b）は、プロンプトの詳細度が生成 AI の評価に与える影響をさらに深く検証し

ている。陳（2025a）は、複数の生成 AI（ChatGPT、Gemini、Claude）の有償版と無償版を比較し、詳細な指示を含むプロンプトが簡潔な指示に比べて人間評価者との一致率を向上させることを示した。しかし、最も一致率が高かった GPT の有償版（ChatGPT o1 pro）でも約 61%に留まり、生成 AI のモデルや料金プランによって評価傾向に違いがあることも明らかにしている。さらに、陳（2025b）は、詳細なプロンプトが AI モデル内部の評価の一貫性を高める効果がある一方で、人間評価者との一致度を示す Kappa 係数は依然として「中程度の一致」には達しない場合が多いことを指摘し、生成 AI の評価が人間の評価を完全に代替するには課題が残ることを示唆した。これらの先行研究から、プロンプトエンジニアリングが生成 AI の出力の精度に大きな影響を与えることがわかる。

以上のように、生成 AI による作文評価の研究において、評価精度向上のためのプロンプトエンジニアリングが注目されている。プロンプトに詳細な評価基準や評価例を盛り込むことで、生成 AI が評価タスクをより正確に理解し、評価精度が向上することが報告されている。しかし、これまでの研究では生成 AI と人間評価者との評価結果の差異の原因に焦点を当てた研究が管見の限りないようで、また人間評価者の評価傾向に近づけるためのプロンプトエンジニアリングはどのようなものなのかは明らかにされていない。そこで、本研究では、これらの課題を踏まえ、人間評価者が持つ評価の観点や基準をより的確に反映させるプロンプトエンジニアリングを明らかにすることを目指す。

3. 研究方法

3.1 データ

本研究では、陳（2025a）と陳（2025b）と同様に、金澤（2014）の YNU 書き言葉コーパスに収録されている日本語学習者 57 名（中国語話者 29 名、韓国語話者 28 名）が執筆したタスク 8 の作文と、それらに対する人間評価者の評価結果を用いて分析を行う。タスク

8 の内容は以下のとおりである。

【タスク 8】友達と以下のケータイメールのやりとりをしました。先日あなたのクラブの先輩がちょっとした事件に遭ったという話を聞きました。(4 コマ漫画)。クラブの友達はその話を知りません。4 コマ漫画を見て、どんな事件だったか友達に詳細をメールで教えてあげてください。漫画の主人公は鈴木先輩です。

（金澤 2014：163）

3.2 評価対象

本研究では、陳（2025a）と陳（2025b）と同様に、「タスクの達成度」に関する評価結果を分析対象とし、「事件の詳細を正確に説明できているか」を評価の中心としている。YNU 書き言葉コーパスの評価は、3 名の人間評価者がそれぞれ評価を行った後、話し合いを通じて一致させた結果を最終評価としている。一方、本研究における生成 AI の評価方法では、各モデルが同じ作文を 3 回評価し、評価ごとにセッションをリセットして記憶をクリアする。その結果、最も多く出現した評価を最終評価とし、人間評価者による評価結果との比較を行う。

3.3 使用した生成 AI モデル

陳（2025a）の研究結果によれば、全体的に有償版の AI モデルは無償版よりも人間評価者との評価一致率が高いことが確認されている。そのため、本研究で利用する AI モデルは、有償版の ChatGPT o3 Pro、Gemini 2.5 Pro、Claude opus 4 と Grok 4[1]とし、これらはいずれも 2025 年 7 月末時点における最新のモデルである。

3.4 プロンプト設計

前節で論じたように、詳細な指示を含むプロンプトは、簡潔な指示のプロンプトに比べ、より良い評価結果が得られる。そのため、本研究では詳細な指示を含むプロンプトを採用する。プロンプトは、役割設定、課題内容、評価基準、評価方法、評価例とその評価理由で構成されている。プロンプトに提示した評価基準、評価方法、評価例とその評価理由はすべて金澤（2014）を参照している。なお、この詳細な指示を含むプロンプトの内容は、陳（2025a）と陳（2025b）で使用されたものと同様である。以下は実際に用いたプロンプトである。

表 3 本研究が使用したプロンプト

**あなたは経験豊富な日本語教師であり、以下の手順で学生の作文を評価してください。
1.「<作文>…</作文>」内の学生作文を最後まで読み、その内容を理解してください。
2.「課題内容」と「評価基準」に基づいて、学生作文を評価し、「○」「△」「×」のいずれかの評価を明記してください。
評価結果に続けて、その理由をわかりやすく述べてください。**
課題内容

[1] Claude opus 4 と Grok 4 のみ、ディープシンク（deep think）機能の有無を選択できるため、本研究ではこの機能を有効にした。ディープシンク機能を利用すると、生成 AI が評価過程でより時間をかけ、文脈を深く分析することが可能となると言われる。

「友達と以下のケータイメールのやりとりをしました。先日、あなたのクラブの先輩である鈴木先輩がちょっとした事件に遭ったという話を聞きました。(4 コマ漫画)。クラブの友達はその話を知りません。4 コマ漫画を見て、どんな事件だったか友達に詳細をメールで教えてあげてください。漫画の主人公は鈴木先輩です。」

※4 コマ漫画は添付ファイルをご参照ください。[2]

評価基準：タスクの達成

以下のポイントに基づいて作文を評価してください。

1. **事実の正確性と順序性**

- 4 コマ漫画の内容（新入社員歓迎会での出来事）が正しく時系列で説明されているか
- 「誰が・いつ・どこで・何をしたか」が明確か

2. **作文の構成**

- 冒頭文
- 4 コマの説明：①無理に飲まされる→②倒れる→③病院に運ばれる→④目を覚ます
- 感想

評価方法

- **○：十分に達成されている**
- 事実が正確で、時系列に沿って説明されている
- 作文の構成が適切である
- **△：部分的に達成されている**
- 一部の事実が不正確、または順序が不明確
- 作文の構成に一部不備がある
- **×：達成されていない**
- 事実の誤りが多い
- 作文の構成が不適切である

-以下の記述がある場合は×と評価する：

[2] 3.1 節に提示した金澤（2014：163）の図を添付する。

- 「無理に飲まされた」を「自分から積極的に飲んだ」と記載
- 「救急車で運ばれた」を「車で送ってもらった」と記載

評価例

評価例 1：

作文タスク《8》K039

先輩こないだ新入社員歓迎会に行って
お酒弱いのに上司がついでくれるのを
拒まず全部飲んじゃったんだって。
飲み屋でもうすでによっぱらってたのに
カラオケに行ってまたビールとか飲んだりして。
そこで倒れて求急車に運ばれたそうよ。
目が覚めたら病院だったという…
新入社員だからことわれなかったこともあったと思うけど、
上司もひどいよね…

スコア：○

理由：
4コマの流れが正しく描写できており、タスクが達成できている。

評価例 2：

作文タスク《8》K035

この前、鈴木さんの会社で新入社員歓迎会があったって、、
その時、鈴木さんが、まわりからすごく飲まされたって
まあ、新入社員って辛いよね。
それで歓迎会は2次会のカラオケまで
続けて、うたうとちゅうにかおが真っ白に
なって突然倒れたって。
それでおうきゅうしゃを呼んで、びょういんに
運ばれて、つぎの日に義●がもどって
きたらしいよ。急性アルコル症って

あぶないよな。お前も飲みでのりのり
キャラだから気をつけたほうがいいよ。
スコア：△
理由：
最も伝えるべき重たる事実の部分に誤りがあるためである。「おうきゅうしゃ（→救急車）を呼んで」、「つぎの日に義識（→意識）が戻ったらしいよ」のように、先輩が大変なことにあったという重要な事実を伝える部分の語彙に大きな誤りがあるため、内容が伝わりにくい。

評価例 3：
作文タスク《8》C025
先日、鈴木先輩は新入社員の歓迎会に
お酒をいっぱい飲んで、歌を歌う。
歌う時は、たぶん頭がふらふらして、突然に
倒れてと見たら、本当にびっくりした。
すぐ救急車を呼んで、病院に送った。
鈴木先輩は朝病院で目覚めて、何か発生
したら、自分が全然覚ていなかったと言った。
スコア：×
理由：
「お酒をいっぱい飲んで歌を歌う」という表現では、「上司に飲まされた」という事実が伝わらない。また、「本当にびっくりした」と書くことにより、まるで自分がその場を見たかのような描写になっている。

以下の学生の作文を評価してください。
〈作文〉

〈/作文〉

スコア：

理由：

4. 分析

4.1 人間評価者との一致率

生成 AI の評価結果と人間評価者の評価結果の一致度を検討するために、本研究では重み付き kappa 係数を使うことにした。重み付き kappa 係数を用いる理由としては、評価の○、△、×が順序のあるランクスコアであり、「全く異なる評価」と「一段階のみ異なる評価」を区別できるためである。重み付き kappa 係数を算出する前に、生成 AI が出力した評価結果（○・△・×）を数値化する。具体的には、金澤（2014）に従い、○＝7、△＝5、×＝1 と変換し、分析用のデータとした。

重み付き kappa 係数は 0 から 1 の範囲で示され、数値が高いほど生成 AI と人間評価者の評価一致度が高いことを表している。Landis & Koch（1977）の分類に従えば、kappa 係数が 0.61～0.80 の場合は「実質的な一致」、0.41～0.60 の場合は「中程度の一致」と解釈される。

以下の表 4 は、本研究において生成 AI モデルごとに算出された重み付き kappa 係数の結果を示している。

表 4 各生成 AI モデルの重み付き kappa 係数

生成 AI モデル	重み付き kappa 係数
ChatGPT o3 pro	0.594
Gemini 2.5 pro	0.617
Claude opus 4	0.633
Grok4	0.457

分析の結果、Gemini 2.5 pro と Claude opus 4 は「実質的な一致」を示し、概ね良好な性能を持つことが確認できた。一方、ChatGPT

o3 pro と Grok4 は「中程度の一致」にとどまっていることが明らかになった。

4.2 不一致事例の分析

この節では、生成 AI による評価コメントを詳細に検討する。本研究では、すべての生成 AI の評価結果は一致したものの、その評価が人間評価者とは異なった 11 件の作文を分析対象とし、評価が不一致であった要因を考察する。表 5 に、評価が一致しなかった 11 件の事例を示す。

表 5 人間評価者と生成 AI の評価が不一致であった事例

学習者	人間評価者	すべての生成 AI
C058	○	△
C005	△	○
C050	×	△
K036	○	△
K008	○	△
K003	×	△
K034	△	○
K028	×	△
K020	×	△
K033	×	△
K029	×	△

表 5 から、生成 AI が人間評価者より甘く評価した事例が 8 件、厳しく評価した事例が 3 件であったことがわかる。各事例における生成 AI の評価コメントを考察した結果、人間評価者と生成 AI との間には、以下の 3 つの評価傾向の差異が見られた。

① 人間評価者「○」に対し、すべての生成 AI「△」の事例

生成 AI は、「救急車で運ばれた」や「病院で目を覚ました」といった、課題の 4 コマ漫画における必須情報が部分的にでも欠落していると判断した場合、減点して「部分的に達成されている（△）」と評価する傾向が見られた。これに対し、人間評価者は、情報が簡潔であっても核心を押さえていれば、「十分に達成されている（〇）」と評価する傾向があった。

例えば、C058 の「病院まで運ばれて、翌日退院できたらしいけど」や、K008 の「先輩酔っぱらっちゃってカラオケで歌ってたらいきなり倒れちゃったそうだよ。そこで先輩は記憶とんじゃったけど起きたら病院だったって。」は、人間評価者には「〇」とされたが、生成 AI は情報の不足を指摘して「△」と評価した。

② 人間評価者「△」に対し、すべての生成 AI「〇」の事例

生成 AI は、「無理に飲まされた」、「救急車で搬送」といった必須事項が記述されていれば、誤字や語彙の誤用に対して比較的寛容であり、「十分に達成されている（〇）」と評価する傾向が見られた。一方、人間評価者は、語彙、誤字、文体の不適切さを重視し、これらが認められる場合は「部分的に達成されている（△）」と評価する傾向があった。

例えば、C005 の「その時、鈴木先輩がめちゃ飲ませられて、みんな上司だし、こどわれないじゃん。」という記述には、「ことわれない」を「こどわれない」とする誤字が含まれる。しかし、「飲ませられた」という必須情報が含まれるため、生成 AI は「〇」と評価した。同様に、K034 の「飲まれたらしい。でも上の人から進まれるから、がまんしてどんどん飲じゃったらしいよ。」という記述も、「勧められる」を「進まれる」、「飲んじゃった」を「飲じゃった」とする複数の誤字を含むが、必須情報が記述されていることから生成 AI は「〇」と評価した。

これらの事例から、言語的な正確性に対する評価基準の差異が、両者の評価を分ける一因であると考えられる。

③ 人間評価者「×」に対し、すべての生成 AI「△」の事例

生成 AI は、重大な事実誤認がなく、最低限の時系列が追えていれば「部分的に達成されている（△）」として一定の評価を与える傾向がある。一方で人間評価者は 4 コマすべてが適切に説明されていない、または重要な語句の誤表記が著しく多い場合では、「達成されていない（×）」と評価することがある。

例えば、C050 と K003 の作文は、いずれも 4 コマ目の結末である「病院で目を覚ます」という重要な情報が欠落していた。人間評価者はこれを物語の構成における重大な欠陥とみなし「×」と評価したが、生成 AI は先行する出来事の記述に一定の整合性があるとして「△」と評価した。K020 の「それで、どんどん飲ませて、結局はカラオケでうたってるとちゅうに倒れちゃったんだよ。おもしろいのは、先輩がその時からずっと寝て今日の朝 10 時におきたということがある」という記述も、病院に関する情報が欠落しているため、人間評価者は「×」としたが、生成 AI は大筋の流れが記述されていると判断し「△」と評価した。

また、K033「みんな鈴木せんぱいにお酒をさそって結局よっぱらっちゃったみたい！」や K029「鈴木先輩が新入社員歓迎会でよまらせてよっぱらになったよ。」では、「先輩が無理に飲まされた」という重要な背景が記述されていない。この点から人間評価者は「×」としたが、生成 AI は記述された情報のみで出来事の概要を理解できるとして「△」と評価した。

さらに、K028 の記述には「憶識がなかった」や「作日なにもなかったの姿で起きて」など、意味の理解を著しく妨げる表現が多数見られた。人間評価者は、物語が適切に伝達されていないとして「×」と評価した。しかし、生成 AI は「飲まされる」、「倒れる」、「病院」、「起きる」といった断片的なキーワードから最低限の時系列を把握できると判断し、「△」と評価した。

4.3 不一致事例の考察

以上の分析から、人間評価者と生成 AI の評価傾向には明確な差異が認められた。人間評価者は、必須情報が簡潔でも核心を捉えていれば肯定的に評価する一方、語彙や文法などの言語的な正確性を重視し、誤りには厳格な評価を下す傾向がある。また、物語の構成上、重要な情報が欠落している場合は、課題の達成度を「×」と厳しく判定した。

対照的に、生成 AI は情報の網羅性を重視し、必須情報が不足していると評価を下げる一方で、言語的な誤りには比較的寛容であった。さらに、断片的な情報からでも最低限の時系列が追えれば、人間評価者が「×」とするような作文に対しても「△」という一定の評価を与える傾向が見られた。

この評価傾向の差異は、人間評価者がプロンプトに明示されていない、文章の質や構成に関する暗黙的な評価基準を適用していることを示唆している。この暗黙知こそが、両者の評価を乖離させる要因であると考えられる。

したがって、生成 AI の評価を人間評価者の基準に近づけるためには、プロンプトにおける評価基準をより詳細かつ具体的に明示する必要がある。特に、「十分に達成されている（〇）」、「部分的に達成されている（△）」、「達成されていない（×）」の各基準について、許容される誤りの範囲や、必須情報の記述レベルなどを明文化することが不可欠である。以下に、これらの点を踏まえて修正したプロンプトを提示する。網掛け部分が修正した箇所である。

表 6 本研究が使用した修正後のプロンプト

**あなたは経験豊富な日本語教師であり、以下の手順で学生の作文を評価してください。

1.「<作文>…</作文>」内の学生作文を最後まで読み、その内容を理解してください。

2.「課題内容」と「評価基準」に基づいて、学生作文を評価し、「○」「△」「×」のいずれかの評価を明記してください。

評価結果に続けて、その理由をわかりやすく述べてください。**

課題内容

（課題内容は修正前と同様のため省略）

評価基準：タスクの達成

（評価基準は修正前と同様のため省略）

評価方法

- **○：十分に達成されている**
- 事実が正確で、時系列に沿って説明されている
- 作文の構成が適切である
- 簡潔な表現であっても、事件の核心（無理に飲まされた→倒れる→救急車で運ばれる→病院で目覚める）が明確に伝われば、十分に達成されていると見なす。
- 暗示的・推測的な記述（例：「起きたら病院だった」→救急車搬送を省略）を「満たしている」とみなす。
- **△：部分的に達成されている**
- 一部の事実が不正確、または順序が不明確
- 作文の構成に一部不備がある
- 事実関係の根幹をなす動詞の表現（受け身、使役など）の文法的な正確性に不備があるが、意図は推測できる。
- **×：達成されていない**
- 事実の誤りが多い
- 作文の構成が不適切である
- 表記誤りが重大で、日本語として意味が成立しない、または客観

的に内容を推測することが困難な場合

- 以下の重大な問題点を含む場合：

　　- 「無理に飲まされた」という事実を誤って表現している(例：「自分から積極的に飲んだ」「飲ませた」「飲まれた」と誤解している。または、強制されたのではなく、本人の意思や責任で飲んだかのような表現になっている。(例:「飲み過ぎた」「飲んじゃった」など)

　　- 「救急車で運ばれた」という事実を誤って表現している（例：「車で送ってもらった」「病院に運ばせた」など)。または自発的な行為や単純な移動と誤解させる表現(例:「病院に行った」など)。

　　- 4コマの核心的な出来事（①無理に飲まされる→②倒れる→③救急車で運ばれる→④病院で目覚める）のうち、いずれか一つでも明確に記述されていない、または客観的に推測できない場合は「×」とする。特に、目覚めた場所(病院)が明記されていない、または文脈から推測できない場合も「×」とする

（評価例は修正前と同様のため省略）

以下の学生の作文を評価してください。

<作文>

</作文>

スコア：

理由：

4.4 修正後のプロンプトの有効性検証

修正後のプロンプトを用いて再度評価実験を行い、人間評価者との一致度を算出した。図1は、プロンプト修正前後の重み付きkappa係数を比較したものである。

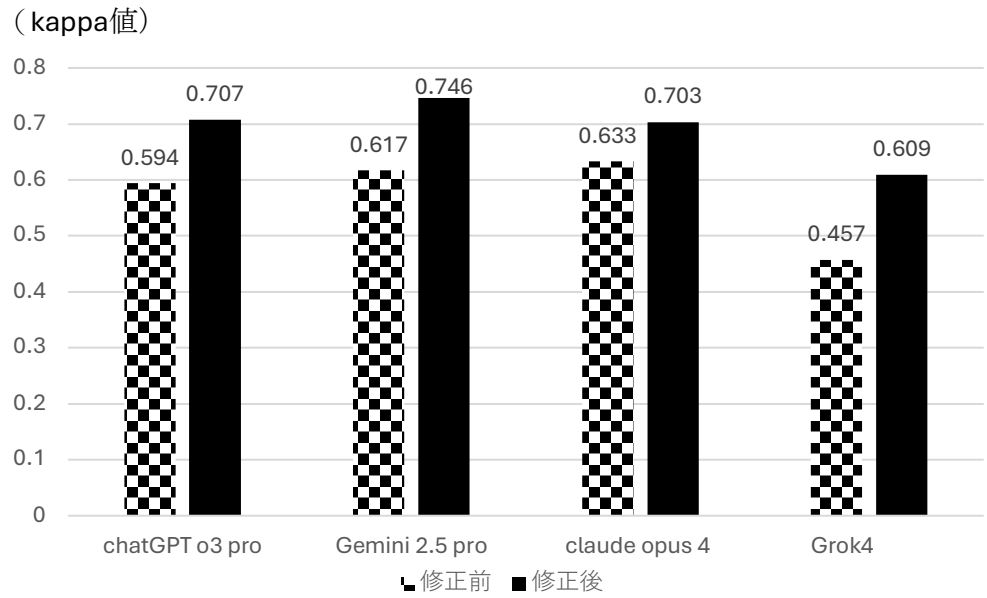

図 1　プロンプト修正前後の重み付き kappa 係数

図 1 が示すように、プロンプトを修正した結果、全ての AI モデルにおいて人間評価者との一致度が向上した。これは、評価基準を詳細に記述したプロンプトを用いることで、生成 AI が人間評価者の暗黙的な評価意図をより正確に解釈し、評価に反映できるようになったためと考察される。

特に、Gemini 2.5 Pro は kappa 係数が 0.617 から 0.746 へと大幅に向上しており、修正後のプロンプトが同モデルの評価精度に対して特に有効であったことを示唆している。Grok4 も大きく改善は見られた。

本研究ではすべての生成 AI モデルの評価結果は一致したものの、その評価が人間評価者とは異なった 11 件の作文を分析対象としたが、これは生成 AI 全体の評価傾向を捉えるための分析であった。生成 AI モデルごとの評価コメントを個別に分析し、それぞれの特性に合わせてプロンプトを最適化することで、さらなる精度向上が期待できるだろう。

以上の結果は、生成 AI を用いた作文評価において、評価基準を明確に示し、プロンプトを継続的に改善していくプロンプトエンジニ

アリングの重要性を示している。このプロンプトエンジニアリングを通じ、生成 AI は人間評価者の基準により近い、精度の高い評価を実現できると考えられる。

5. 考察

前節の不一致事例の分析から、生成 AI と人間評価者の評価傾向には明確な差異が認められた。人間評価者は、文脈から「病院まで運ばれて」や「起きたら病院だった」といった省略的な表現の含意を読み取るが、生成 AI はこのような情報を見逃す傾向があった。また、人間評価者は語彙や漢字の誤りといった言語的な正確性を厳しく評価するのに対し、生成 AI は必須情報が含まれていれば表記の誤りには比較的寛容であった。さらに、人間評価者は 4 コマすべてが適切に説明されていない場合は厳しい評価を下すのに対し、生成 AI は断片的な情報からでも最低限の時系列が追えると一定の評価を与える傾向があった。

このことから、両者の評価の乖離は、人間評価者が持つ暗黙的な評価基準を生成 AI が解釈できなかったことに起因すると考えられる。本研究では、生成 AI と人間評価者の評価傾向の差異を分析し、その結果に基づいてプロンプトを修正することで、両者の一致度が向上することを実証した。

以上の結果から、生成 AI を用いた作文評価の精度を向上させるためには、まず小規模な検証を通じて生成 AI と人間評価者との評価傾向の差異を分析し、そこから得られた知見を基に評価基準を明確化してプロンプトに反映させ、再度検証を行うという、一連のプロンプトエンジニアリングのプロセスが極めて重要である。このような継続的な調整を通じ、生成 AI は人間評価者の基準により近い、精度の高い評価を実現できると考えられる。

6. 日本語教育現場への示唆

Anthropic 社が提示したプロンプトエンジニアリングの概要によ

れば、一般的に効果的とされる手法として以下の九つが挙げられている[3]。具体的には、「プロンプトジェネレーター（prompt generator）の利用」、「明確で直接的な指示の提示」、「複数の例（multishot）の提示」、「思考の連鎖（Chain of Thought）の促進」、「XML タグを用いた構造の明確化」、「AI の役割設定」、「応答の一部を事前に提示すること」、「複雑なプロンプトの連鎖」、「長い文脈に関するヒントの活用」である。これまでの先行研究と本研究の結果を踏まえ、生成 AI による作文評価の精度向上に向け、以下のようなプロンプトエンジニアリングの示唆を提示する。

まず、プロンプトの基本として「AI の役割設定」をすること、「明確で直接的な指示の提示」をすること、さらに「複数の例の提示」をすることが必要である。本研究で用いたプロンプトの例として生成 AI に以下のような役割と指示を与えた。

「AI の役割設定」として「あなたは経験豊富な日本語教師である」と設定し、「明確かつ直接的な指示」として以下の指示を提示した。

・「<作文>…</作文>」内に記された学生作文を最後まで読み、その内容を十分に理解すること。

・「課題内容」と「評価基準」に基づいて作文を評価し、「○」「△」「×」のいずれかを明示すること。評価結果に続けて、その理由を分かりやすく記述すること。

・「課題内容」と「評価基準」の内容を詳述すること

また、「複数の例の提示」として実際に「○」「△」「×」の評価を行った例を挙げ、それぞれの理由を説明した。これらを前提として OpenAI 社[4]、Anthropic 社[5]、または Google 社[6]が提供する「プロンプ

---

[3] https://docs.anthropic.com/ja/docs/build-with-claude/prompt-engineering/overview （2025.08.08 閲覧）

[4] https://platform.openai.com/docs/guides/prompt-generation（2025.08.08 閲覧）

[5] https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/prompt-generator （2025.08.08 閲覧）

[6] https://cloud.google.com/vertex-ai/generative-ai/docs/learn/prompts/ai-powered-prompt-writing （2025.08.08 閲覧）

トジェネレーター」を利用し、「思考の連鎖の促進」と「XML タグを用いた構造の明確化」を効果的に取り入れる。さらに、有償版の生成 AI を使って少数の作文サンプルを評価させ、その結果とコメントを丁寧に検討しながら、人間評価者が持つ暗黙的な評価基準を明らかにし、プロンプトを継続的に改善・調整していくことが重要である。

このような反復的な改善を重ねるプロンプトエンジニアリングこそが、生成 AI を信頼性の高い評価支援ツールとして活用するための鍵となる。

7. おわりに

本研究では、プロンプトエンジニアリングを通じて生成 AI の評価精度を人間評価者の基準に近づける試みを行った。その結果、人間評価者の持つ暗黙的な評価基準を言語化し、プロンプトに具体的に反映させることで、両者の一致度が著しく向上することを実証した。このことは、生成 AI が単なる自動化ツールに留まらず、人間との相互作用を通じて性能を向上させられる可能性を示唆している。

今後の課題としては、生成 AI と人間との共生という視点から、より高度な連携モデルを模索することが挙げられる。その際に「ヒューマンインザループ（Human-in-the-Loop）機械学習」の考え方が参考になるのであろう。ヒューマンインザループは「AI を活用するアプリケーションにおいて人間と機械の知能を組み合わせる仕組み」（Robert 2021：4）を意味し、生成 AI の評価結果を人間が修正し、そのフィードバックを生成 AI が学習することで、継続的に評価精度を向上させていく仕組みである。このようなシステムを教育現場に導入することで、生成 AI は教師の評価負担を軽減する補助ツールとして機能し、教師はより創造的で個別化された指導に専念できるようになる。将来的には、このような人間と生成 AI の協働が、作文教育全体の質を向上させる上で不可欠な要素となるであろう。さらに、本研究は YNU コーパスの特定タスクを対象としたものであ

るが、今後は異なる課題やジャンルを対象とした検証を進める予定である。

<付記>

本研究は、113 年度国科会専題研究「初探 ChatGPT 運用在日語作文評価之可能性」（NSTC113-2410-H-031-028-）の助成を受けて実施されたものである。また、本研究は、2025 年 7 月 13 日に開催された「言語科学会第 26 回年次国際大会」において発表した内容に加筆・修正を加えたものである。

参考文献

石井雄隆（2025）「【コラム 5】AI と自動採点」李在鎬・青山玲二郎編『AI で言語教育は終わるのか？』くろしお出版、pp. 240−242.

金澤裕之（2014）『日本語教育のためのタスク別書き言葉コーパス』ひつじ書房

陳相州（2025a）「日本語作文評価における生成 AI の効果検証−プロンプト、AI プロバイダー、料金の影響を中心に」『台灣日語教育學報』44、pp.30−53.

陳相州（2025b）「大規模言語モデルを用いた日本語学習者作文の自動評価の可能性―プロンプト設計が評価の一貫性と人間評価者との一致度に及ぼす影響―」『台灣日本語文學報』57、pp.157−176.

李在鎬（2023）「ChatGPT による日本語作文の自動採点」『2023 年度日本語教育学会秋季大会予稿集』、pp.158−163.

李在鎬・加藤恵梨・堀恵子・村田裕美子・毛利貴美（2023）「ChatGPT の評価観点と人間の評価観点の比較−計量テキスト分析の手法を用いた分析−」『第二言語習得研究会（JASLA）第 34 回全国大会』、pp.37−42.

李在鎬（2025）「【コラム 2】プロンプトエンジニアリング」李在鎬・青山玲二郎編『AI で言語教育は終わるのか？』くろしお出版、pp. 229−231.

Landis, J. R., & Koch, G. G.（1977）. The measurement of observer agreement for categorical data. *Biometrics*, 33（1）, 159-174.

Mizumoto, A. & Eguchi, M.（2023）. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2（2）.

DOI：https：//doi. org/10. 1016/j. rmal. 2023. 100050.

Robert, M.（2021）. *Human-in-the-Loop Machine Learning：Active learning and annotation for human-centered AI*, Manning.