探討生成式 AI 應用在日語作文評價的效果：以提示詞設計、AI 供應商和付費方案為中心

陳相州

東吳大學日本語文學系副教授

摘要

本研究探討了生成式 AI 在日語作文評價中的應用效果，重點考察了提示詞設計、AI 供應商和付費方案三個因素對評價結果的影響。研究使用了「YNU 書面語料庫」中 57 名學習者完成的作文任務，並透過兩種不同的提示詞（簡潔指示和詳細指示）對 ChatGPT、Gemini 和 Claude 的付費版和免費版進行測試。研究發現：(1)相較於簡潔指示的提示詞，詳細指示的提示詞與人類教師的一致率更高；(2)付費版 GPT o1 pro mode 的一致率最高(61.40%)，其次是付費版 Gemini 2.0 experimental advanced、Claude 3.5 sonnet 以及免費版 Claude 3 haiku(49.12%)；(3)各 AI 供應商在評價傾向上存在差異，如 GPT o1 pro mode 在評價上位群學習者時表現較好，而 Claude 3 haiku 則整體評價偏寬鬆；(4)在上位群學習者作文的評價上，AI 與人類教師的一致率較高，而隨著學習者作文能力降低，AI 與人類教師的一致率也下降。

關鍵詞：生成式 AI、日語作文評價、提示詞設計、AI 供應商、付費方案

受理日期：2025 年 02 月 26 日

通過日期：2025 年 06 月 06 日

DOI：10.29758/TWRYJYSB.202506_(44).0002

Verification of Generative AI Effectiveness in Japanese Composition Evaluation: Focusing on the Impact of Prompts, AI Providers, and Pricing Plans

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

Abstract

This study evaluates generative AI in Japanese composition, analyzing prompt design, AI providers, and payment plans. Using 57 learner compositions from the "YNU Written Corpus," both paid and free versions of ChatGPT, Gemini, and Claude were tested with concise and detailed prompts. Key findings include: (1) detailed prompts improved agreement with human evaluators; (2) GPT o1 pro mode (paid) had the highest agreement (61.40%), followed by Gemini 2.0 experimental advanced and Claude 3.5 sonnet, with Claude 3 haiku (free) at 49.12%; (3) AI providers had distinct evaluation tendencies. For instance, GPT o1 pro mode performed better in evaluating high-level learners, whereas Claude 3 haiku tended to give more lenient evaluations overall; (4) agreement was higher for advanced learners and declined with lower proficiency.

Keywords: Generative AI, Japanese composition evaluation, prompt design, AI provider, payment plan

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―

陳相州
東呉大学日本語学科副教授

要旨

本研究では、日本語作文評価における生成 AI の効果を検証し、プロンプト設計、AI プロバイダー、料金プランという三つの要因が評価結果にどのように影響するかを明らかにした。「YNU 書き言葉コーパス」の 57 名の学習者による作文を対象とし、二種類のプロンプト（簡潔な指示と詳細な指示）を用い、ChatGPT、Gemini、Claude の有償版と無償版をテストした。その結果、(1)詳細な指示を含むプロンプト B が簡潔な指示のプロンプト A よりも人間評価者との一致率が高いこと、(2)有償版の GPT o1 pro mode が最も高い一致率(61.40%)を示し、次いで有償版の Gemini 2.0 experimental advanced と Claude 3.5 sonnet、および無償版の Claude 3 haiku(49.12%)が続くこと、(3)各 AI プロバイダーによって評価傾向に特性があり、例えば GPT o1 pro mode は上位群学習者の評価に強く、Claude 3 haiku は全体的に甘い評価に偏る傾向があること、(4)上位群学習者の作文評価における一致率が高く、中位群学習者と下位群学習者になるにつれて一致率が低下する傾向があることが明らかになった。

キーワード：生成 AI、日本語作文評価、プロンプト設計、AI プロバイダー、料金プラン

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―

陳相州
東呉大学日本語学科副教授

1. はじめに

近年、AI 技術の急速な発展に伴い、日本語教育現場において AI の効果的な活用方法が模索されている。特に作文指導では、教師が学習者全員の作文を添削し、評価する必要があり、この作業は教師にとって体力的にも時間的にも大きな負担となっている。そのため、AI 技術を活用して作文評価を効率化することは、作文授業を担当する教師にとって非常に重要な課題となっている。

このような状況の中、ChatGPT、Claude、Gemini などの生成 AI（Generative AI）の登場により、作文指導に新たな可能性が開かれている。具体的には、教師が AI を活用して学習者の作文を初期段階で処理することが可能となり、AI が学習者の日本語の誤りを添削し、簡単な評価を提供する。これにより、教師はそのフィードバックを基に内容を補足して強化することができ、指導効率の向上が期待される。

一方で、生成 AI の使用においては、プロンプトの設計、生成 AI のプロバイダー、料金プラン（有償と無償）の違いが評価の結果に影響を与える可能性がある。教師の立場からすると、どのようなプロンプトを使用するか、更にどの AI プロバイダーや料金プランを選ぶべきか判断が難しい場合もあるであろう。そこで本研究では、AI を活用した日本語作文指導において、「プロンプトの設計」、「AI プロバイダーの選択」、「料金プランの選択」が学習者の作文評価にどのような影響を与えるのかを明らかにする。

2. 先行研究

AI モデルを活用した作文評価は近年、積極的に研究されている。英語学習者を対象とした研究では、Mizumoto & Eguchi(2023) が ChatGPT-3.5 を利用し、TOEFL11 コーパスに基づく英作文を評価し、その結果が既存の分級レベルと概ね一致していることを示した。一方で、日本語学習者の作文評価に関する研究としては、李(2023) や李・加藤・堀・村田・毛利(2023) などが挙げられる。以下では、これらの研究について詳しく説明する。

李(2023)の研究は、ChatGPT のモデル比較（ChatGPT-3.5 と ChatGPT-4）とプロンプト（プロンプト A とプロンプト B）の差異に焦点を当てている。具体的には、プロンプト A では内容評価に重点を置き、プロンプト B では文法や語彙の誤りに基づく減点方式を採用している。この研究の結果、次の 3 点が明らかになった。1) ChatGPT-4 を用いた日本語作文の自動採点は、客観試験のスコアと中程度の相関が認められること、2) プロンプトに言語的要件を明記することで採点精度が向上すること、3) 学習者の習熟度が上がるにつれて採点のずれが比較的小さくなることが判明した。

一方、李・加藤・堀・村田・毛利(2023)の研究は、日本語教師による評価と ChatGPT による評価を比較したものである。李・加藤・堀・村田・毛利(2023)では、李(2023)で使用された同一の 100 篇の意見文から、評価が最も高い 5 篇と最も低い 5 篇を選び出し、それらの作文に対する 4 名の日本語教師と ChatGPT による評価を比較した。ChatGPT はモデル（ChatGPT-3.5 と ChatGPT-4）とプロンプト（簡潔な指示と明確な指示）の違いによって作成した 4 条件でコメントを生成した。この研究の結果、次の 4 点が明らかになった。1) 日本語教師と ChatGPT の双方が、文章構成、主張の有無、読み手への配慮、具体例の有無といった基本的な評価基準を共有していること、2) 日本語教師は意見文の評価において、抽象的で広範な観点（積分型の視点）を用いる傾向があること、3) ChatGPT は文章内の具体的な問題点を指摘する形で評価を行う（微分型の視点）こと、4) プ

ロンプト設定の影響よりもGPTモデルの性能が大きく、意見文のフィードバックにはChatGPT-4が最適であることが示された。

これまでの研究では、ChatGPTのモデルやプロンプトの相違が作文評価に及ぼす影響について検討が行われてきたが、ChatGPT以外のAIプロバイダーの違いや、有償版と無償版の違いが日本語作文評価に与える影響については十分に解明されていない。そのため、本研究では、これらの未解明の課題に焦点を当て、プロンプト、AIプロバイダーや料金プランといった異なった条件が評価結果にどのような影響を及ぼすのかを体系的に明らかにすることを目指す。

3. 研究目的と研究意義

本研究では、プロンプト、AIプロバイダー、料金プランといった要因の組み合わせは、学習者の作文評価の結果にどのような影響を及ぼすか、またどの条件で評価結果がもっとも人間による評価に近づくのかを検証することを目的とする。本研究では、プロンプトやAIプロバイダー、料金プランの違いによる最適なアプローチを提案し、作文指導の現場における応用価値を明確に示すことが重要であると考えられる。

4. 研究方法

4.1 使用資料と分析対象

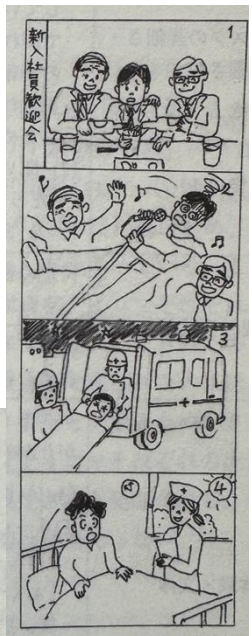
本研究は、主に大学生レベルの日本語学習者が作成した作文を収録した「YNU書き言葉コーパス」（金澤2014）を用いて実証的な分析を行った。このコーパスを選定した理由は、以下の3点にある。(1)学習者による日本語作文が網羅的に含まれていること、(2)作文ごとにタスク達成度が明確に示されていること、(3)作文を評価するための客観的な評価の基準が整備されていることである。なお、コーパス作成当時の参加者は、金澤（2014）の説明によると、授業の受講に支障のない「上級レベル」の学習者とされている。

「YNU書き言葉コーパス」に収録されている作文には、「タスク

の達成」、「タスクの詳細さ・正確さ」、「読み手配慮」、「体裁・文体」の 4 つの評価項目が設定されている。本研究では、この中から「タスクの達成」のみを分析対象とした。さらに、作文タスクは全 12 タスクあるが、各タスクにおける学習者の達成度の分布状況を考慮し、「タスク 8」を選定した。タスク 8 では、4 コマ漫画の内容をメールで友人に伝えるという状況設定がなされており、「事件の詳細を正確に説明できているか」が評価の要点となっている。以下に「タスク 8」の具体的な内容を示す。

【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。



イラストは『実用英語技能検定準 1 級平成元年度第 2 回二次試験』より使用。

（金澤 2014：163）

「YNU 書き言葉コーパス」のレベル分けは、学習期間や日本語能力試験の結果に基づくものではなく、各タスクの評価を総合的に考

慮して上位群・中位群・下位群のグループに振り分けられている（金澤 2014）。すなわち、「YNU 書き言葉コーパス」のレベル分けは学習者の作文能力に基づいて行われている。本研究では、以下で説明するプロンプト B の評価例として提示した 3 名の学習者を除き、上位群 19 名、中位群 19 名、下位群 19 名、合計 57 名の学習者の「タスク 8」の作文を分析に用いた。

4.2 プロンプトの設定

本研究では、以下のような 2 種類のプロンプトを用意し、調査に用いた。詳細は以下の通りである。

プロンプト A

あなたは経験豊富な日本語教師であり、学生の作文を評価しています。以下に示す作文は、あなたが指導している学生が書いたものです。

この作文のタスクは「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言う話を聞きました。（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。」です。

4 コマ漫画は添付ファイルをご参照ください[1]。評価の項目は「タスクの達成」です。結果は○△×で評価してください。

プロンプト B

**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] 4.1 節に提示した金澤（2014：163）の図を添付する。

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

「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

重要：以下の記述がある場合は×と評価します：

- 「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

評価方法

- **○：十分に達成されている**

[2] 4.1 節に提示した金澤（2014：163）の図を添付する。

- 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

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：

以上のプロンプトの内容から分かるように、プロンプトAは基本的なタスク指示と評価観点（「タスクの達成」）だけを簡潔に示す。学習者の作文を「○・△・×のいずれかで評価してください」という指示付きの形式である。一方、プロンプトBは、(1)課題内容の具体的な説明、(2)作文評価基準の明示、(3)評価例の提示という三つの段階で構成し、生成AIが参照できる情報をより丁寧に提示した。例えば、「誰が・いつ・どこで・何をしたか」という時系列や重要人物の記述が正確かどうか、文章構成や誤りの程度に着目するような基準を明確にしている。また、「×と判定する条件」なども具体的に示している。なお、プロンプトBの作文評価基準と方法、および○（十分達成）、△（部分的達成）、×（未達成）の例と理由は、すべて金澤（2014）を参照している。

4.3 生成AIのプロバイダー、料金プランとモデル

本研究に利用した生成AIは2024年12月下旬の時点において写真を添付できる最新バージョンで、生成AIのプロバイダー、料金プランとモデルを以下の表にまとめた。

表1　生成AIのプロバイダー、料金プランとモデル

プロバイダー	料金プラン	モデル
ChatGPT	有料	GPT o1 pro mode

	無料	GPT-4o
Gemini	有料	Gemini 2.0 experimental advanced
	無料	Gemini 1.5 flash
Claude	有料	Claude 3.5 sonnet
	無料	Claude 3 haiku

5. 評価の実施

まず、先述の「YNU 書き言葉コーパス」における 57 名の学習者のタスク 8 の作文を対象とし、各生成 AI モデルに対し、同一作文を 3 回ずつ評価させた。評価するたびに、前回の評価履歴が残らないように対話メモリをリセットしている。次に、モデルによる三回の評価の中で最も多く出現した評価を最終評価の結果とした。最後に、この結果を金澤（2014）の人間評価者の評価結果と比較した。

6. 分析

6.1 人間評価者の評価結果

金澤(2014)の人間評価者による「タスク 8」の評価結果は、以下の通りであった。

表 2　金澤(2014)の各レベルの学習者の評価結果

	上位群学習者		中位群学習者		下位群学習者	
評価結果	評価数	割合	評価数	割合	評価数	割合
○	17	89.5%	6	31.6%	4	21.1%
△	2	10.5%	7	36.8%	0	0.0%
×	0	0.0%	6	31.6%	15	78.9%

上位群学習者では約 9 割が「○」であり、中位群学習者では評価

がほぼ均等に分布、下位群学習者では約 8 割が「×」となっている。そのことから、人間評価者の評価結果は学習者の作文能力レベルを適切に反映していると考えられる。

6.2 生成 AI の評価結果

6.2.1 全体的な評価の一致率

以下の表 3 では、全 57 作文を対象とした場合における、生成 AI と人間評価者との評価一致率を示す。表 3 に記載されているとおり、プロンプト B を使用し、有償版の GPT o1 pro mode を利用した場合が 61.40%と最も高い一致率を示した。次いで、プロンプト B を使用した有償版の Gemini 2.0 experimental advanced、Claude 3.5 sonnet、および無償版の Claude 3 haiku が 49.12%の一致率で並んでいる。一方で、プロンプト A を使用した場合、全体的に一致率が低い傾向が見られ、最も低かったのはプロンプト A を使用した無償版の Gemini 1.5 flash で 17.54%となった。

表 3　生成 AI と人間評価者との評価一致率

プロンプト	料金プラン	AI モデル	作文数	一致数	一致率
B	有料	GPT o1 pro mode	57	35	61.4%
B	有料	Gemini 2.0 experimental advanced	57	28	49.1%
B	有料	Claude 3.5 sonnet	57	28	49.1%
B	無料	Claude 3 haiku	57	28	49.1%
A	無料	Claude 3 haiku	57	27	47.4%
A	有料	GPT o1 pro mode	57	25	43.9%
A	有料	Claude 3.5 sonnet	57	25	43.9%
B	無料	GPT-4o	57	23	40.4%
A	有料	Gemini 2.0 experimental advanced	57	22	38.6%
A	無料	GPT-4o	57	21	36.8%

B	無料	Gemini 1.5 flash	57	19	33.3%
A	無料	Gemini 1.5 flash	57	10	17.5%

以上の結果から、ChatGPT、Gemini、Claude の有償版（GPT o1 pro mode、Gemini 2.0 experimental advanced、Claude 3.5 sonnet）がプロンプト B を使用した場合において、その作文評価結果と人間評価者の評価結果の一致率が高いことがわかった。特に注目すべきは、Claude の無料バージョン（Claude 3 haiku）がプロンプト B を使用した場合にも高い一致率を示すという優れた結果を見せた。すなわち、人間評価者との一致率を重視する場合、有償版の GPT o1 pro mode が最適な選択肢となり、次いで同じく有償版の Gemini 2.0 experimental advanced と Claude 3.5 Sonnet が優れた選択肢となる。一方、予算が制約されている場合には、無償版の Claude 3 Haiku も非常に有用な代替案として考えられる。

以下では、人間評価者との評価一致率が比較的高かった条件、すなわちプロンプト B を用いた有償版 GPT や Gemini、Claude（GPT o1 pro mode、Gemini 2.0 experimental advanced、Claude 3.5 sonnet）、およびプロンプト B を用いた無償版 Claude（Claude 3 haiku）を中心に、学習者の作文能力別（上位群・中位群・下位群）にそれぞれの一致率を詳しく分析しつつ、各生成 AI を実際に利用する際の注意点を検討する。

6.2.2 GPT o1 pro mode の場合

表 4 は、プロンプト B と GPT o1 pro mode を組み合わせた評価結果を示している。

表 4　プロンプト B を用いた GPT o1 pro mode の評価結果

対象	評価	作文数	一致数	一致率
上位群学習者	○	17	15	88.2%

	△	2	1	50.0%
	×	0	0	0.0%
	合計	19	16	84.2%
中位群学習者	○	6	6	100.0%
	△	7	0	0.0%
	×	6	5	83.3%
	合計	19	11	57.9%
下位群学習者	○	4	3	75.0%
	△	0	0	0.0%
	×	15	5	33.3%
	合計	19	8	42.1%

表 4 の結果から、GPT o1 pro mode はプロンプト B を用いた場合、上位群学習者の作文評価において人間評価者との一致率が 84.2%と非常に高い値を示した。特に「○」評価の一致率が 88.2%と高く、上位群学習者の優れた作文を適切に評価できていることがわかる。中位群学習者に対しては全体で 57.9%の一致率であり、特に「○」評価では 100.0%、「×」評価では 83.3%と高い一致率を示した。しかし「△」評価については一致率が 0.0%と極めて低く、評価が「○」または「×」のいずれかに明確に分かれる傾向が顕著に見られる。

下位群学習者においては、「○」評価の一致率が 75.0%と高いものの、「×」評価の一致率は 33.3%と低く、全体的な一致率は 42.1%に留まる。

以上のことから次のようなことが言える。GPT o1 pro mode は、プロンプト B を用いることで、特に上位群と中位群の学習者に対して高い評価一致率を示すことが明らかになった。しかし、中位群の「△」評価の一致率が低いこと、下位群全体の評価一致率が相対的

に低いことから、GPT o1 pro mode は学習者の作文能力レベルによって評価の一致率に差が生じる可能性があることが示唆される。特に、下位群の評価においては、人間評価者とのずれが生じやすく、「×」評価の作文を高く評価してしまうことに注意が必要であろう。

6.2.3 Gemini 2.0 experimental advanced の場合

表 5 は、プロンプト B と Gemini 2.0 experimental advanced を組み合わせた評価結果を示している。

表 5　プロンプト B を用いた Gemini 2.0 experimental advanced の評価結果

対象	評価	作文数	一致数	一致率
上位群学習者	○	17	7	41.2%
	△	2	2	100.0%
	×	0	0	0.0%
	合計	19	9	47.4%
中位群学習者	○	6	1	16.7%
	△	7	4	57.1%
	×	6	4	66.7%
	合計	19	9	47.4%
下位群学習者	○	4	2	50.0%
	△	0	0	0.0%
	×	15	8	53.3%
	合計	19	10	52.6%

表 5 によると、Gemini 2.0 experimental advanced はプロンプト B を用いた場合、学習者の作文能力のレベル間で比較的に均等な一致率を示しており、上位群学習者で 47.4%、中位群学習者で 47.4%、

下位群学習者で 52.6%となっている。

上位群学習者に対しては「△」評価で 100.0%の一致率を達成した一方、「○」評価では 41.2%と低い値となった。つまり、上位群の作文に対して実際は「○」とすべき作文をやや過小評価する傾向がある。

中位群学習者に対しては「△」評価で 57.1%、「×」評価で 66.7%の一致率を示したが、「○」評価では 16.7%と非常に低い値となった。つまり、上位群学習者の場合と同様、人間評価者が評価した「○」の作文をやや厳しめに評価している可能性が考えられる。下位群学習者に対しては「○」評価で 50.0%、「×」評価で 53.3%の一致率を示したが、誤りの多い作文をある程度「×」と判定する傾向が見られる。

これらの結果から、Gemini 2.0 experimental advanced は、GPT o1 pro mode と比較すると、学習者の作文能力のレベルによる評価一致率の変動幅が小さいことが特徴であることがわかる。また、「○」評価において人間評価者よりも厳しい評価傾向があると考えられる。全体としては、下位群においても比較的安定した一致率が見られる。

6.2.4 Claude 3.5 sonnet の場合

表 6 は、プロンプト B と Claude 3.5 sonnet を組み合わせた評価結果を示している。

表 6　プロンプト B を用いた Claude 3.5 sonnet の評価結果

対象	評価	作文数	一致数	一致率
上位群学習者	○	17	9	52.9%
	△	2	2	100.0%
	×	0	0	0.0%
	合計	19	11	57.9%
中位群学習者	○	6	1	16.7%
	△	7	3	42.9%

	×	6	3	50.0%
	合計	19	7	36.8%
下位群学習者	○	4	3	75.0%
	△	0	0	0.0%
	×	15	7	46.7%
	合計	19	10	52.6%

表 6 の結果から、Claude 3.5 sonnet はプロンプト B を用いた場合、上位群学習者で 57.9%、中位群学習者で 36.8%、下位群学習者で 52.6%の一致率を示した。上位群学習者に対しては「△」評価で 100.0%と高い一致率を示し、「○」評価でも 52.9%の一致率となった。中位群学習者に対しては「○」評価で 16.7%と低く、「△」評価で 42.9%、「×」評価で 50.0%の一致率にとどまり、中位群学習者の作文の微妙な完成度を捉えきれない様子がうかがえる。下位群学習者に対しては「○」評価で 75.0%と高い一致率を示した一方、「×」評価では 46.7%の結果となった。下位群の「○」評価の一致率が高いことから、下位群学習者の優れた側面を適切に評価できる可能性がある。

全体的には、中位群学習者の一致率は他の AI モデルと比較して低い傾向にあることから、中位群学習者の評価に課題があるといえる。

6.2.5 Claude 3 haiku の場合

Claude 3 haiku は、上記で示した有償版の AI モデルとは異なり、唯一の無償版 AI モデルである。表 7 は、プロンプト B と Claude 3 haiku を組み合わせた評価結果を示している。

表 7　プロンプト B を用いた Claude 3 haiku の評価結果

対象	評価	作文数	一致数	一致率
上位群学習者	○	17	17	100.0%

	△	2	0	0.0%
	×	0	0	0.0%
	合計	19	17	89.5%
中位群学習者	○	6	6	100.0%
	△	7	1	14.3%
	×	6	0	0.0%
	合計	19	7	36.8%
下位群学習者	○	4	4	100.0%
	△	0	0	0.0%
	×	15	0	0.0%
	合計	19	4	21.1%

Claude 3 haiku の全体一致率は 49.12%であったが、詳細をレベル別に見ると非常に特徴的な傾向が現れている。Claude 3 haiku はプロンプト B を用いた場合、上位群学習者では「○」評価の作文に対する一致率が 100.0%と高く、合計でも 89.5%と優れた結果となった。一方、中位群学習者では「○」評価の作文に対しては 100.0%の一致率を示したが、「△」と「×」がほぼ正しく判定されておらず、合計の一致率は 36.8%に留まる。下位群学習者では「○」評価の作文に対しては 100.0%の一致率を示したものの、「×」評価の作文はすべて外れており、合計の一致率はわずか 21.1%と低い数値に留まっている。

以上の結果から、Claude 3 haiku は無償版にもかかわらず上位群学習者の作文評価においては非常に高い評価の一致率を示す一方、「△」と「×」の評価をほとんど行わない傾向にあり、全体として非常に甘い AI 評価基準を持つことがわかった。このことから、Claude 3 haiku は上位群学習者の優れた作文を評価する場合には有

用であるが、中位群と下位群学習者の作文に対する評価には注意が必要であるといえる。

7. 考察

本研究の分析結果から、日本語作文評価における生成 AI の効果に関して、以下のようなことが明らかになった。

まず、プロンプトの設計が評価の一致率に大きく影響することが明らかとなった。全体的には、プロンプト B（詳細な指示）がプロンプト A（簡潔な指示）よりも人間評価者との一致率が高かった。これは、李（2023）の「プロンプトにおいて言語的要件を明記することで採点精度があがる（158 頁)」という知見と一致しており、生成 AI に対して明確な作文評価基準や参考例を提示することの重要性を示している。

次に、料金プランについては、有償版が無償版よりも全体的に高い評価の一致率を示す傾向が見られた。特に GPT o1 pro mode は有償版で 61.4%という最も高い一致率を示した。しかし、興味深いことに Claude 3 haiku の無償版もプロンプト B を用いた場合には 49.1%と比較的高い一致率を示した。この結果は、予算制約がある教育現場においても、適切なプロンプト設計により無償版の生成 AI でも一定の成果が期待できることを示唆している。

また、プロバイダーごとの生成 AI の特性の違いも明確になった。GPT o1 pro mode は上位群学習者の評価に優れ、Gemini 2.0 experimental advanced は学習者の作文能力のレベルによる評価一致率の変動が小さく、「○」評価においては人間評価者よりも厳しい評価傾向がある。Claude 3.5 sonnet は中位群学習者の評価に課題がある一方で、Claude 3 haiku は全体的に甘めの評価傾向があり、「△」と「×」の評価をほとんど付けない傾向にある。

更に、各 AI モデルの評価結果から見ると、上位群学習者の作文評価に対する一致率が高く、中位群と下位群になるにつれて一致率が低下する傾向がある。これは、李（2023）の「習熟度が上がるにつ

れて採点の誤差が比較的小さくなる（158 頁）」という知見と一致している。

また、Claude 3 haiku のような無償版でありながら、上位群学習者の「〇」評価において 100.0%の一致率を示しているため、AI モデルの特性を理解し、適切に使い分けることの重要性が明らかとなった。例えば、上位群学習者の優れた作文を評価する際には、Claude 3 haiku が十分な精度を持つ可能性がある。

これらの結果から、日本語作文評価における生成 AI 活用の可能性と限界が明らかになった。生成 AI の全体的な評価一致率は GPT o1 pro mode でも 61.40%にとどまるため、生成 AI は人間の評価者を完全に代替するものではない。しかし、プロンプト B のように作文評価基準を詳細に提示し、各 AI プロバイダーの得意・不得意を見極めて適切に活用することで、教師の作文評価を効果的に支援できる可能性がある。教師は各 AI プロバイダーの特性を理解した上で、初期スクリーニングや評価の補助ツールとして生成 AI を活用することで、作文指導の効率化が期待できる。

8. まとめと今後の課題

本研究では、日本語作文評価における生成 AI の効果を、プロンプト、AI プロバイダー、料金プランという三つの要因に焦点を当てて検証した。本研究の主な研究成果は以下にまとめる。

- プロンプト B（詳細な指示）が、プロンプト A（簡潔な指示）よりも人間評価者との一致率が高い。
- 有償版の GPT o1 pro mode が最も高い一致率を示し、次いで有償版の Gemini 2.0 experimental advanced と Claude 3.5 sonnet、および無償版の Claude 3 haiku が続いた。
- 各 AI プロバイダーによって評価傾向に特性があり、GPT o1 pro mode は上位群学習者の評価に強く、Gemini 2.0 experimental advanced は「〇」評価の作文を厳しく評価する傾向があり、Claude

3.5 sonnet は中位群学習者の評価に課題がある一方、Claude 3 haiku は全体的に甘い評価に偏る傾向がある。

・全体として上位群学習者の作文評価における一致率が高く、中位群と下位群になるにつれて一致率が低下する傾向がある。

これらの結果は、教師が生成 AI を日本語作文の評価に活用する上で、以下のような示唆を与える。まず、生成 AI を作文評価に活用する際には、作文評価基準や具体的な評価例を明確に示した詳細なプロンプトを設計することが重要である。次に、有償版 AI を利用することでより人間評価者に近い評価を得られるが、無償版 AI でも使い方やプロンプトの工夫次第では十分に活用可能な場合がある。さらに、学習者の作文能力のレベルによって判断のずれが生じやすく、特に中位群と下位群では微妙な誤りや構成不足を正確に見極めるのが難しいため、最終的な判断には人間の教師によるチェックが必須となる。

日本語教育における生成 AI の活用はまだ始まったばかりであり、今後も技術の進化とともに新たな可能性が広がっていくことが期待される。本研究がその一助となれば幸いである。今後は、生成 AI と人間評価者の評価内容の違いを詳しく分析するとともに、生成 AI を活用した作文評価が学習者にどのような影響を与え、自己学習にどのように結びつくのかといった学習効果を検証していく予定である。

<付記>

本研究は 113 年度国科会専題研究「初探 ChatGPT 運用在日語作文評價之可能性」（NSTC113-2410-H-031-028-）の研究成果の一部である。

参考文献

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

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.

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.