研究発表 / RESEARCH PRESENTATION

中高英語記述問題における 生成 AI 採点と手動採点の比較

— 採点作業の削減と評価の安定化に向けた実証研究 —

ジョシュクン エリフ
東京大学大学院

ジョシュクン セリム
株式会社 Nuginy

景山 開陽
株式会社 Nuginy

EMPIRICAL STUDY

手書き答案 7,721 件を対象とした 大規模実証分析

教育工学研究発表会 / 2026

AGENDA · 目次

本日の構成

発表 20 分 / 全 20 枚

01	研究背景	教員の業務負担と評価の揺らぎ
02	研究方法	7,721 件の手書き答案を OCR → AI 採点
03	結果	完全一致率・±1 点以内一致率・MAE
04	考察	人間採点者の揺らぎとハイブリッド運用
05	結論	実用化に向けた到達点と今後の課題

研究背景

近年、日本の学校教育現場では**働き方改革が急務**となっており、教員の業務負担軽減が大きな課題である。

中でも英語教育における**記述式問題（自由英作文・和文英訳）の採点**は、正答の多様性や文法・語彙の細かな評価が必要なため、多大な時間と精神的労力を要する。

Ifenthaler (2022) — 採点業務は教員の主要な負担要因として継続的に指摘されている。

> 記述式採点には、客観式採点にはない**判断の蓄積**が問題ごとに求められる。1 問あたり数十秒の判断が、数千件規模では 数十時間の業務 に直結する。

採点における二つの課題

課題 1

採点コスト

記述式問題は文法・語彙・語順・意味内容を複数観点から評価する必要があり、教員の時間的・精神的負担が大きい。

課題 2

評価の揺らぎ

採点者の主観による判定のばらつきを排除し、公平性をいかに担保するかが長年の懸念事項であった。

大規模言語モデル（LLM）の発展により、これら二つの課題を **同時に解決し得る可能性** が現実味を帯びてきた。

— Atkinson & Palma (2025)

関連する先行研究

Mizumoto & Eguchi · 2023

GPT-3 による 英作文採点

TOEFL11 コーパスを対象に、人間評価との **高い相関**を確認。

Atkinson & Palma · 2025

LLM ベースの ハイブリッド AES

大規模言語モデルと既存手法を組み合わせた自動エッセイ採点を提案。

三田・霜田 · 2025

初級学習者向け 評価システム

生成 AI を用いた英文の質の評価について試行的に分析。

既存研究の対象 ―多くがデジタル入力された大学生以上の長文エッセイ を対象。 中高生の手書き答案を OCR で扱う検証は十分に行われていない。

本研究の独自性

01

中高生の 短〜中程度記述

スペルミス・文法誤りが混在する学習初期段階の答案を対象。大学生以上のエッセイとは異なる難しさ。

02

手書き答案 × OCR の実地条件

紙の答案を AI-OCR でテキスト化する現場相当のパイプラインで、運用想定の精度を検証。

03

7,721 件の 大規模実証

14 設問にわたる答案を一括採点し、統計的に安定した比較を実施。

研究目的

中高生の英語手書き答案 **7,721 件** を対象に、 **生成 AI 採点** と **手動採点** の結果を直接比較し、 両者の差が ±1 点以内 に収まる割合を算出することで、 生成 AI 採点が**実用段階**にあるかを明らかにする。

採点作業の削減 × 評価の安定化 — 両立可能性の検証

研究対象データ

7,721 件

中学校・高等学校の英語手書き答案

A 社提供 / 実試験・教材演習で作成された紙答案

14 設問 **配点 4〜9 点** **100 点満点試験の一部**

問題形式	
和文英訳	提示された日本語を英語に書き換える形式
短文記述	英語による短い記述回答（数語〜一文）
条件付き英作文	指定条件を満たした英文を作成する形式

いずれも複数の正答表現が存在し、**文法・語彙・語順・意味内容** を複合的に評価する必要のある問題群である。

02 · METHOD

データ処理と比較フロー

STEP 01

手書き答案 画像取得

紙の答案を画像データとして取得

→

STEP 02

AI-OCR による 文字認識

光学文字認識で手書き文字をデジタル化

→

STEP 03

テキスト データ化

共通フォーマットへ整形・正規化

PIPELINE A

生成 AI 採点

大規模言語モデル（LLM）が採点基準に基づき得点と**信頼度スコア** を出力。

PIPELINE B

手動採点（Ground Truth）

教材制作・試験採点の経験を有する採点者による既存採点結果を正解データとして使用。

↓ 全 7,721 件で **直接比較** ↓

中高英語記述問題における生成 AI 採点と手動採点の比較

評価指標

Metric 01

完全一致率

生成 AI 採点と手動採点の得点が **完全に一致した割合** 。 厳密一致精度を測定。

Metric 02 · 主要指標

±1 点以内一致率

得点差が ±1 点以内 に収まった割合。 実運用で許容できる**誤差域**を評価。

Metric 03

平均絶対誤差 (MAE)

得点差の絶対値の平均。 全体誤差量を 1 指標で把握。

※ 生成 AI が出力する **信頼度スコア** は、ハイブリッド運用設計の根拠としても用いる（後述）。

結果の概要

83.64 %

全体の ±1 点以内一致率

N = 7,721 / 全 14 設問の合計

90 % +

一部問題群で達成した一致率

~1.0 点

平均絶対誤差 (MAE) の水準

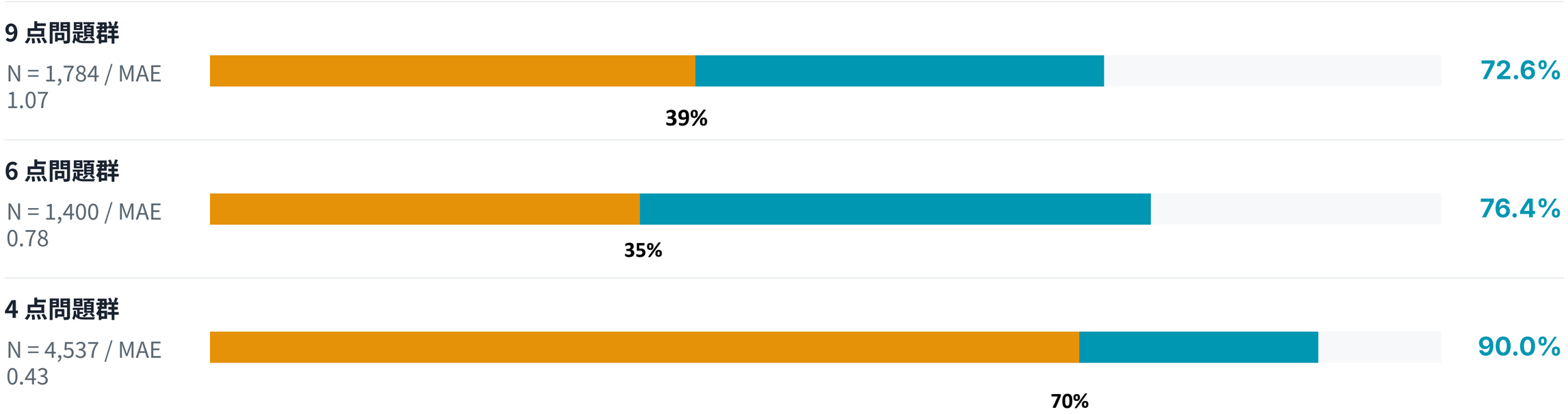
56.96 %

全体の完全一致率

設問別の一致率と MAE

表 1 — 14 設問 / N = 7,721

問題	総解答数	完全一致率 (%)	±1 点以内 (%)	MAE
9 点問題 - (1)	977	38.79	72.15	1.10
9 点問題 - (2)	807	40.15	72.99	1.04
6 点問題 - (1)	700	29.90	74.60	1.07
6 点問題 - (2)	700	40.00	78.14	0.49
4 点問題 - (1)	347	68.59	85.01	0.25
4 点問題 - (2)	337	**80.42**	**95.25**	0.65
4 点問題 - (3)	474	66.67	82.91	0.25
4 点問題 - (4)	477	**81.97**	**94.34**	0.43
4 点問題 - (5)	500	67.80	91.60	0.43
4 点問題 - (6)	500	70.80	90.00	0.32
4 点問題 - (7)	434	75.81	92.86	0.51
4 点問題 - (8)	490	65.51	87.35	0.42
4 点問題 - (9)	500	66.40	92.80	0.47
4 点問題 - (10)	478	65.70	90.59	0.97
合計 / 平均	**7,721**	**56.96**	**83.64**	**—**



配点が上がるほど絶対差が拡大する傾向は、**人間採点者間にも共通して観察される現象**であり、部分点幅の広さに起因する。

人間採点者間の揺らぎとの比較

生成 AI 採点の誤差は、**手動採点者間のばらつきと同程度、あるいはそれ以下**の水準に収まった。

含意 01

記述式問題の採点では、**人間同士でも完全一致しない** ため、AI が同程度に揺らぐことは現場運用上致命的ではない。

含意 02

AI が **人間の評価揺らぎの範囲内** に収まる以上、教育現場での実用的な評価手段となり得る。

考察

実用精度への接近 — 中高英語記述問題において、生成 AI 採点は実用的な精度水準に達しつつある。

揺らぎの吸収 — AI の誤差は人間採点者間のばらつき範囲内に収まる。

OCR を介した実地条件 — デジタル入力ではなく、手書き → OCR → AI 採点という現場相当の経路でも精度を維持。

先行研究との対比 Kim (2026) は、AI が一貫して人間より高得点を付与し、得点ばらつきが小さい傾向を報告。 本研究では AI のばらつきの小ささは確認されたが、**手動採点との誤差は限定的** であった。先行研究の課題に対し改善可能性を示す結果と言える。

ハイブリッド運用への展望

生成 AI が出力する**信頼度スコア** を活用し、高信頼の答案は AI が自動採点、低信頼の答案のみ人間が確認する運用が可能である。

INPUT

手書き答案

OCR でテキスト化された答案

→

AI 採点 + 信頼度

自動振り分け

信頼度スコアによって出力経路を決定

→

OUTPUT

得点確定

高信頼: AI 自動 / 低信頼: 人間確認

採点公平性

信頼度を介した一貫した基準により、採点者ごとの揺らぎが抑制される。

負担軽減

人間が確認するのは低信頼分のみで、採点工数を大幅に削減できる可能性。

中高英語記述問題における生成 AI 採点と手動採点の比較

研究の課題と今後の展望

課題 01

現場での 実証的導入

実教育現場での運用テスト、教員側の受容性、運用負荷の長期評価。

課題 02

OCR 精度の 採点への影響

OCR の誤認が AI 採点誤差に与える定量的影響、および OCR 改善の効果を測定する必要がある。

課題 03

他教科への 適用可能性

本研究は英語の記述問題に限定。国語・社会など他教科への一般化検証が必要。

結論

7,721 件

中高英語手書き答案

大規模実証データセット

83 % +

±1 点以内一致率（全体）

一部問題群では 90% 超

≦

人間採点者間の揺らぎ

AI 誤差はそれと同等以下

生成 AI 採点は、中高英語記述問題において**実用化可能な段階に近づいている**。 信頼度スコアを介したハイブリッド運用により、採点コストの削減と評価の安定化 を同時に実現できる可能性が示された。

REFERENCES

参考文献

ATKINSON & PALMA · 2025

Atkinson, J. and Palma, D. (2025) *An LLM-Based Hybrid Approach for Enhanced Automated Essay Scoring.* Scientific Reports, 15: 14551.
doi.org/10.1038/s41598-025-87862-3

IFENTHALER · 2022

Ifenthaler, D. (2022) *Automated Essay Scoring Systems.*
In O. Zawacki-Richter and I. Jung (Eds.), Handbook of Open, Distance and Digital Education. Springer, Singapore, pp. 1–15.
doi.org/10.1007/978-981-19-0351-9_59-1

KIM · 2026

Kim, S. (2026) *A Comparative Study of AI-Based and Human Scoring for Descriptive Writing Assessment.* Journal of Educational Measurement and Evaluation, 12(1): 45–62.

MIZUMOTO & EGUCHI · 2023

Mizumoto, A. and Eguchi, M. (2023) *Exploring the Potential of Using an AI Language Model for Automated Essay Scoring.*
Research Methods in Applied Linguistics, 2(2): 100050.
doi.org/10.1016/j.rmal.2023.100050

三田・霜田 · 2025

三田 薫, 霜田 敦子 (2025) *英語初級学習者のパラグラフ・ライティング自動評価システム開発の試み Part 3 ― 生成 AI を用いた英文の質を評価するシステムの分析.*
実践女子大学短期大学部紀要, 46: 71–95.
doi.org/10.34388/0002000234

ACKNOWLEDGEMENT

ご清聴
ありがとうございました

QUESTIONS & DISCUSSION