国語記述式答案に対する LLM を用いた OCR 誤り訂正と自動採点への影響

鈴木里菜 [1]　齊藤隆浩 [1]　臼井久生 [1]　尾崎太亮 [1]　グエントゥアンフーン [1]　古宮嘉那子 [1]
石岡恒憲 [2]　中川正樹 [1]
[1] 東京農工大学大学院　[2] 大学入試センター
{s248289x, s245145v, h-usui, hiroaki-ozaki}@st.go.tuat.ac.jp,
{fx7297@go, kkomiya@go, nakagawa@cc}.tuat.ac.jp,　tunenori@rd.dnc.ac.jp

概要

本研究では，国語科記述式問題の手書き答案を対象とし，OCR 誤り訂正モデルの性能と自動採点への影響を評価する．T5 及び GPT 系モデルに対して，対象文章全体を入力として訂正する 1 段階訂正と，誤りの可能性がある箇所に<error>タグを付与し訂正の手掛かりとして利用する 2 段階訂正を適用し，訂正精度と採点精度を比較した．既存研究では，T5 による 1 段階訂正と 2 段階訂正の双方が OCR 出力そのままと比較して文字起こし精度の向上を示しており，特に 2 段階訂正がより大きな改善を示すことが報告されていたが，自動採点への影響は検討されていなかった．本研究では，中学生 185 名の手書き答案を用いて各手法を評価した結果，2 段階訂正はいずれのモデルでも OCR 訂正精度を向上させた一方で，採点精度は 1 段階訂正と大きな差は見られなかった．このことから，訂正性能の比較的小さな改善は採点性能に直結しないことが分かった．

1　はじめに

国語科記述式問題の自動採点では，手書き答案の文字認識（OCR）が採点精度に直接関わる重要な処理となる．学習者の手書き答案は筆跡や文字形状にばらつきがあり，文字の潰れや濃淡の不均一が生じることもあるため，OCR が誤認識を起こす場合がある．OCR の誤認識が残った状態で自動採点を行うと，答案の内容が正しく評価されないケースが生じ，採点の正確性に影響を与える可能性がある．したがって，OCR 出力に対する誤り訂正は，自動採点を行う上で重要な前処理である．

これまで，OCR 誤り訂正の手法として，N-gram やルールベース手法，BERT[1] による文脈理解に基づく訂正などが提案されてきた．近年では，生成モデルである T5[2] を用いる手法も試みられ，一定の訂正性能が報告されている [3, 4]．また，全文を対象とする 1 段階訂正に加えて，誤りの可能性がある箇所を推定し，訂正モデルへの手掛かりとして利用する 2 段階訂正は，不必要な訂正を抑えつつ訂正精度を向上させることが示されている [5]．しかし，これらの研究は主に訂正精度の向上に焦点を当てており，訂正後の文章が自動採点モデルにどのような影響を与えるかについては十分に検討されていない．

本研究では，中学生 185 名の手書き答案を対象として，先行研究で用いられてきた T5 系の 1 段階訂正および 2 段階訂正の設定に加え，同様の枠組みを GPT 系モデルにも適用し，OCR 訂正精度と，訂正文を GPT ベースの自動採点モデルに入力した際の採点精度の変化を比較する．

2　関連研究

OCR 誤り訂正に関する研究はこれまで多く行われてきた．竹内ら [6] や Nagata ら [7] は，統計言語モデルや文字 N-gram に基づく訂正手法を提案している．Sakamoto ら [8] や Nguyen ら [9] は，編集距離や探索手法を活用した OCR エラー訂正に取り組み，誤り候補の探索と最適解の選択により訂正性能の向上を図った．

大規模言語モデルを活用した手法では，謝ら [10] が BERT による文脈理解を用いた OCR 誤り訂正を行い，文脈に基づく柔軟な書き換えの有効性を示している．また，中村ら [11] や藤武ら [12] は T5 を OCR や音声認識の誤り訂正に適用し，生成モデルによる訂正が高い性能を示すことを報告している．

また，誤り箇所を推定してから訂正を行う二段階型の手法も提案されている．Schaefer ら [13] は

Bi-LSTM による誤り箇所推定と機械翻訳モデルを組み合わせ，Nguyen ら [14] は BERT による推定と文字レベルの変換モデルを用いた訂正手法を示している．

記述式答案の自動採点では，GPT 系モデルの活用が広がりつつあり，少数例や zero-shot 設定でも一定の採点性能が報告されている [15, 16]．日本語短答式問題に対しては，BERT による専用モデルを構築した事例もあり，高精度な採点が可能であることが示されている [17]．また，Suzuki ら [18] は，T5 を用いた OCR 誤り訂正が自動採点精度に与える影響を検証し，データ拡張を用いた T5(DA) が，通常の T5 よりも訂正精度および採点精度の双方を向上させることを示している．

3　データ

本研究では，既存研究 [3, 4, 5, 18] と同様に「10 分間復習ドリル国語読解」に対する中学生 185 名の手書き答案を用いた．対象 25 問はいずれも 60 字以内の記述式短答問題である．答案画像はアイラボ株式会社の縦書き OCR システムによりテキスト化され，最大 5 候補の出力から計 6,656 件の OCR 結果を得た．本研究ではこれらを訂正対象とし，人手転記データを正解として利用した．各答案には教員による正誤ラベルが付与されている．

4　LLM を用いた OCR 誤り訂正

既存研究 [5] で用いられた T5 による 1 段階訂正および 2 段階訂正の設定を踏襲しつつ，本研究では同じ枠組みを GPT 系モデルにも適用した．[1)]

4.1　1 段階訂正

1 段階訂正では，OCR 結果の全文をそのまま入力として与え，LLM に訂正後のテキストを生成させる．T5 では，text-to-text 形式で，OCR 結果を入力とし訂正後の文を生成する．GPT 系モデルでも同じく OCR 結果を入力し，OCR 誤りを修正するよう指示したプロンプトを与えて生成を行い，その出力を訂正文とした．

4.2　2 段階訂正

2 段階訂正では，まず誤りの可能性がある部分を推定し，その情報を手掛かりとして訂正を行う．

1)　2 段階訂正モデルに関する詳細な検討については，現在ジャーナル論文として投稿中である．

(1) 誤り箇所推定　T5 ベースの 2 段階訂正では，RoBERTa を用いて OCR 結果をトークン単位で分類し，各トークンが誤りを含むかを判定した．この推定結果に基づき，誤りと判定された箇所に対して後処理として`<error>`タグを付与した．GPT 系モデルでは，OCR 結果を入力とし，誤りと推定した箇所に`<error>`タグを付与したテキストを生成させた．いずれの場合も，誤り箇所を`<error>`タグで明示したテキストを後段の訂正モデルへの手掛かりとして用いた．

(2) 誤り訂正　誤り訂正では，前段の誤り箇所推定で得られた`<error>`タグ付きの OCR 結果と，タグなしの OCR 結果を `[sep]` トークンで連結した形式を入力として用いた．T5 では，この入力をそのまま与え，訂正後のテキストを生成させた．GPT 系モデルでは，T5 に与えたものと同じ入力に加え，`<error>`タグで示された OCR 誤りを修正するよう指示したプロンプトを与え，生成結果を誤り訂正結果とした．

5　実験

本実験では，4 節で述べた OCR 誤り訂正手法について，訂正精度と，訂正後の文章を入力とした自動採点結果への影響を評価した．

5.1　実験設定

OCR 誤り訂正手法　OCR 誤り訂正では，次の 4 手法を比較対象とした．

- **T5-1step**：OCR 結果全体を入力とし，T5 により訂正後の文を直接生成する 1 段階訂正
- **T5-2step**：誤り箇所推定の結果を手掛かりとして訂正を行う，T5 ベースの 2 段階訂正
- **GPT-1step**：OCR 結果全体を入力とし，GPT 系モデルにより訂正後の文を生成する 1 段階訂正
- **GPT-2step**：誤り箇所推定の結果を手掛かりとして訂正を行う，GPT ベースの 2 段階訂正

T5-1step および T5-2step の実験設定は，既存研究 [3, 5] と同一の条件を用いた．以下では，GPT 系手法の実験設定について述べる．GPT 系手法では，モデルとして `gpt-4o-mini-2024-07-18` を使用し，いずれの手法においても OpenAI の API を用いて fine-tuning を行った．また，データ分割には，T5 ベースの手法と同一の 5 分割を用い，GPT 系モデルでは評価データを用いた途中評価が行えないため，学習データお

よびテストデータのみを用いて評価を行った．

GPT-1step　GPT-1step の学習では，入力として OCR 結果全体と，OCR 誤りを訂正するよう指示するプロンプトを与え，出力として人手による正確な文字起こしを用いた．使用したプロンプトの詳細は付録 A.1 に示す．

GPT-2step　GPT-2step における誤り箇所推定の学習では，OCR 結果と人手による正確な文字起こしとの差分を基に，人手の文字起こしと一致しない箇所を<error>タグで囲んだテキストを正解データとして作成した．学習データは，入力として OCR 結果と誤り箇所を推定するよう指示するプロンプトを与え，出力として上記の方法で<error>タグを付与したテキストを用いた．誤り訂正の学習では，正解の<error>タグ付きテキストと OCR 結果を [SEP] トークンで連結した入力に，OCR 誤りを訂正するよう指示するプロンプトを付与し，出力として人手による正確な文字起こしを与えた．使用したプロンプトの詳細は付録 A.2,A.3 に示す．

自動採点実験　訂正後テキストが自動採点に与える影響を評価するため，先行研究 [19] の設定に基づき，GPT 系モデルを用いた自動採点実験を行った．具体的には，訂正後の答案テキストを入力とし，GPT-4o-mini を用いた zero-shot 設定で正誤判定を行った．

5.2　評価指標

誤り箇所推定には，既存研究と同様に Accuracy，Precision，Recall，F1 を用い，トークン単位の誤りラベルに基づいて算出した．誤り訂正の評価には BLEU，CRR（Character Recognition Rate），WRR（Word Recognition Rate）を用い，訂正結果と正解文字列の一致度を文単位・文字単位・語単位で評価した．また，各答案について，モデルが出力した正誤判定結果を用い，訂正手法ごとの採点精度を比較した．使用したプロンプトは付録 A.4 に示す．

6　実験結果

本節では，5 節で述べた実験設定に基づき，OCR 誤り箇所推定，OCR 誤り訂正，及び訂正後文章に対する自動採点の結果を示す．

6.1　誤り箇所推定の結果

表 1 に，誤り箇所推定における評価結果を示す．RoBERTa を用いた T5-2step に加え，GPT-2step による誤り箇所推定についても，同一の評価指標を用いて比較した．

表 1　OCR 結果に対する誤り箇所推定性能の比較

	Accuracy	Precision	Recall	F1
RoBERTa	91.48	91.40	91.40	91.40
GPT-2step	**98.72**	**98.61**	**98.79**	**98.69**

RoBERTa による誤り箇所推定では，Accuracy が 91.48%，Precision が 91.40%，Recall が 91.40%，F1 スコアが 91.40%と，いずれの指標においても高い値を示した．一方，GPT-2step では，Accuracy97.82%，Precision98.61%，Recall98.79%，F1 スコア 98.69%と，RoBERTa を上回る結果が得られた．

6.2　誤り訂正の結果

表 2　T5 および GPT 系モデルによる OCR 誤り訂正前後の性能評価

	文字起こし評価指標			採点精度	
	BLEU	CRR	WRR	4o	4o-mini
Raw	48.56	74.68	51.18	50.66	52.19
T5-1step	52.70	67.91	59.73	59.84	58.09
T5-2step	54.11	69.73	62.91	57.58	54.59
GPT-1step	65.36	77.71	74.22	**69.61**	**66.55**
GPT-2step	**68.69**	**81.05**	**78.19**	68.73	65.67
Human Trans.[2)]	100.00	100.00	100.00	76.09	72.89

表 2 に OCR 誤り訂正の評価結果を示す．**Raw** は訂正前の OCR 結果，**T5-1step**，**T5-2step**，**GPT-1step**，**GPT-2step** は各手法による訂正後の結果を表す．太字は各指標の最高値である．

T5 系手法では，1step・2step のいずれも BLEU および WRR が訂正前より向上し，特に 2step では 1step を上回る結果が得られた．一方，CRR は訂正前より低下したが，文脈に基づく修正による影響と考えられる．GPT 系手法では，全体として T5 系より高い訂正精度が得られ，BLEU，CRR，及び WRR の全ての指標において 2step によりさらに向上した．以上より，T5 系および GPT 系のいずれにおいても，2 段階訂正は 1 段階訂正と比較して OCR 訂正精度を向上させる傾向が確認された．

6.3　自動採点結果

表 2 には，OCR 誤り訂正後の文章を入力とした自動採点の精度も併せて示している．採点の結果，い

2)　Human Trans. は，手書き答案画像を人手で文字起こしした結果を指す．

表 3　誤り訂正後における文字起こしおよび採点結果の改善・悪化・維持の割合

	文字起こし				採点			
Methods	**T5-1step**	**T5-2step**	**GPT-1step**	**GPT-2step**	**T5-1step**	**T5-2step**	**GPT-1step**	**GPT-2step**
改善	14.07%	7.07%	28.86%	**31.41**%	16.03%	12.68%	**24.78**%	23.76%
悪化	6.12%	3.21%	1.46%	**1.31**%	6.85%	5.76%	5.83%	**5.69**%
改善-悪化	7.94%	3.86%	27.40%	**30.10**%	9.18%	6.92%	**18.95**%	18.07%
正解維持	8.89%	11.81%	13.56%	13.70%	43.80%	44.90%	44.83%	44.97%
不正解維持	70.92%	77.92%	56.12%	53.57%	33.31%	36.66%	24.56%	25.58%
変化なし	79.81%	89.73%	69.68%	67.27%	77.11%	81.56%	69.39%	70.55%
Accuracy	22.96%	18.88%	42.42%	**45.11**%	59.84%	57.58%	**69.61**%	68.73%

ずれの訂正手法でも自動採点精度は OCR 出力そのままより向上した．一方，1 段階訂正と 2 段階訂正の差は小さく，訂正精度の向上が採点性能に直接結びつくとは限らないことが分かる．なお，1 段階訂正と 2 段階訂正の採点精度の差について，カイ二乗検定を用いて検定を行った結果，有意水準 0.05 において統計的に有意な差は確認されなかった．

7　考察

表 3 は，文字起こしおよび採点結果が，訂正前の OCR 結果 (Raw) と比較して，各訂正手法により改善・悪化・変化なしとなった割合を示している．まず文字起こしに着目すると，T5 系では，**T5-1step** より **T5-2step** の方が改善の割合は小さいものの，悪化が抑えられ，特に維持の割合が増えていることがわかる．これは，T5 系の 2 段階訂正が，積極的に大きな書き換えを行うというより，不必要な訂正を抑え，既存の内容を安定して保持する方向に働いていることを示唆する．一方，GPT 系では，**GPT-2step** が **GPT-1step** よりも改善率が高く，悪化率が低い結果となっており，誤り箇所情報を活用した段階的な処理が，GPT 系モデルではより効果的に働き，積極的な訂正と安定性の両立に寄与していることが示唆される．

一方で，採点結果に着目すると，1 段階訂正と 2 段階訂正の間で改善・悪化の割合に大きな差は見られない．GPT 系では，いずれの手法においても T5 系より高い採点精度が得られているものの，**GPT-1step** と **GPT-2step** の間では，改善率・悪化率ともに近い値を示している．これらの結果から，2 段階訂正は文字起こしの品質向上には有効である一方，採点結果全体に対しては一貫した改善効果をもたらすとは限らないことが分かる．

次に，**GPT-1step** と **GPT-2step** で採点結果が異なった答案を対象に，その特徴を定性的に分析する．**GPT-1step** では誤採点された一方で，**GPT-2step** では正しく採点できた例として，不要な書き換えを抑えつつ適切に訂正が行われたケースが確認された (付録 B 表 4 参照)．このような場合は，**GPT-2step** では OCR 誤りに対応した最小限の訂正が行われ，答案の意味内容が保持された結果，採点モデルが正しく判断できたと考えられる．一方で，**GPT-1step** では正しく採点できたものの，**GPT-2step** では誤採点となった例も確認された．これらの例には，元の OCR 出力の崩れが大きく訂正後も意味解釈が困難なものや，1step と 2step で訂正結果がほぼ同一であるにもかかわらず，採点結果にばらつきが生じたものが含まれていた (付録 B 表 5 参照)．

以上の考察から，2 段階訂正は誤り箇所情報を利用することで訂正対象を絞り込み，文字起こしの品質を安定的に向上させる効果を持つ一方，採点結果への影響は限定的であることが分かった．このことは，自動採点においては OCR 訂正の精度向上だけでなく，採点モデル自体の解釈能力や安定性が重要な要因となる可能性を示している．

8　おわりに

本研究では，国語科記述式問題の手書き答案を対象に，T5 および GPT 系モデルにおける 1 段階訂正と 2 段階訂正の比較を通して，OCR 誤り訂正性能と自動採点への影響を評価した．実験の結果，2 段階訂正はいずれのモデルでも OCR 訂正精度の向上に寄与した一方，採点精度に与える影響は限定的であり，訂正性能の比較的小さな改善は採点性能に直結しないことが分かった．

今後は，訂正結果と採点モデルの相互作用をより詳細に検討することで，自動採点における OCR 誤り訂正の位置づけをさらに明確にしていくことが課題となる．

謝辞

本研究は JSPS 科研費 JP22K12145，JP23K28201 JP24H00738 の助成を受けたものです．答案収集は，本学における人を対象とする研究に関する倫理審査委員会の承認を得て実施しました (No.230402-0411)．また，本研究において答案データを提供くださったワコム株式会社，および文字認識データを提供くださったアイラボ株式会社に深く感謝申し上げます．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. **NAACL-HLT2019**, p. 4171–4186, 2019.

[2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **J. Mach. Learn. Res.**, Vol. 21, No. 140, pp. 1–67, 2020.

[3] 鈴木里菜, 臼井久生, 尾崎太亮, Nguyen Tuan Hung, 古宮嘉那子, 石岡恒憲, 中川正樹. T5 を用いた日本語記述式答案の文字認識誤り訂正. 言語処理学会 第 30 回年次大会 発表論文集, pp. 1148–1153, 2024.

[4] Rina Suzuki, Hisao Usui, Hiroaki Ozaki, Hung Tuan Nguyen, Kanako Komiya, Tsunenori Ishioka, and Masaki Nakagawa. Error Correction of Japanese Character-Recognition in Answers to Writing-Type Questions Using T5. **Document Analysis Systems**, pp. 229–243, 2024.

[5] 鈴木里菜, 臼井久生, 尾崎太亮, Nguyen Tuan Hung, 古宮嘉那子, 石岡恒憲, 中川正樹. RoBERTa と T5 を用いた 2 段階モデルによる国語答案の文字認識誤り訂正. 言語処理学会 第 31 回年次大会 発表論文集, pp. 1051–1055, 2025.

[6] 竹内孔一, 松本裕治. 統計的言語モデルを用いた OCR 誤り訂正システムの構築. 情報処理学会論文誌, Vol. 40, No. 6, pp. 2679–2689, 1999.

[7] Masaaki Nagata. Japanese OCR error correction using character shape similarity and statistical language model. In **COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics**, 1998.

[8] 阪本浩太郎, 阿部川明優, 佐竹真樹, 岸川至白, 阪本エリーザ, 石下円香, 渋木英潔, 森辰則. 契約書 OCR の単語誤り訂正における漢字の偏旁冠脚を考慮した木編集距離の検討. **The Association for Natural Language Processing**, pp. 137–140, 2020.

[9] Quoc-Dung Nguyen, Nguyet-Minh Phan, Pavel Krömer, and Duc-Anh Le. An efficient unsupervised approach for OCR error correction of Vietnamese OCR text. **IEEE Access**, Vol. 11, pp. 58406–58421, 2023.

[10] 謝素春, 松本章代. 日本語 BERT モデルによる近代文の誤り訂正. 言語処理学会 第 29 回年次大会 発表論文 集, pp. 1616–1620, 2023.

[11] 中村朝陽, 李聖民, 田村鴻希, 吉永直樹. 前後の発話を文脈として考慮するニューラル音声認識誤り訂正. 情報処理学会, pp. 1–7, 2022.

[12] Nao Soma, Teruno Kajiura, Mai Takahashi, and Kimio Kuramitsu. Applying error correction models in code with additional pre-training to large language model (Daikibo gengo model heno tsuika jizengakusyu niyoru ayamariteisei model no code heno tekiyou). **DEIM Forum 2023**, pp. 1b–5–4, 2023.

[13] Robin Schaefer and Clemens Neudecker. A Two-step Approach for Automatic OCR Post-Correction. **Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage**, pp. 52–57, 2020.

[14] Thi Tuyet Hai Nguyen, Adam Jatowt, Nhu-Van Nguyen, Mickael Coustaty, and Antoine Douset. Neural machine translation with BERT for post-ocr error detection and correction. **Proceedings of the ACM/IEEE Joint Conference on Digital Libraries**, 2020.

[15] Lui Yoshida. The impact of example selection in few-shot prompting on automated essay scoring using GPT models. In **International Conference on Artificial Intelligence in Education**, p. 61–73, 2024.

[16] Shigeng Chen, Yunshi Lan, and Zheng Yuan. A multi-task automated assessment system for essay scoring. In **International Conference on Artificial Intelligence in Education**, p. 276–283, 2024.

[17] Haruki Oka, Hung Tuan Nguyen, Cuong Tuan Nguyen, Masaki Nakagawa, and Tsunenori Ishioka. Fully automated short answer scoring of the trial tests for common entrance examinations for Japanese university. In Maria Mercedes Rodrigo, Noburu Matsuda, Alexandra I. Cristea, and Vania Dimitrova, editors, **Artificial Intelligence in Education**, pp. 180–192, Cham, 2022. Springer International Publishing.

[18] Rina Suzuki, Takahiro Saito, Hisao Usui, Hiroaki Ozaki, Hung Tuan Nguyen, Kanako Komiya, Tsunenori Ishioka, and Masaki Nakagawa. Investigating the influence of automated transcription error correction on automated grading for handwritten Japanese answers. In Alexandra I. Cristea, Erin Walker, Yu Lu, Olga C. Santos, and Seiji Isotani, editors, **Artificial Intelligence in Education**, pp. 372–379, Cham, 2025. Springer Nature Switzerland.

[19] Takahiro Saito, Hung Tuan Nguyen, Kanako Komiya, Tsunenori Ishioka, and Masaki Nakagawa. Similarity-based scoring model for handwritten answers in Japanese workbooks, artificial intelligent in education. In **Artificial Intelligent in Education**, 2025.

A　使用したプロンプト

A.1　GPT-1step に用いたプロンプト

prompt = "次の文章は手書き文字の OCR 結果です。OCR による文字の誤りのみを訂正し、本来書かれていた通りの文字列に戻してください。"

A.2　GPT-2step における誤り箇所推定のプロンプト

prompt = "次の文章は OCR の出力結果です。文字認識が誤っている部分にのみ<error>タグ</error>を付けてください。
注意: ・誤りのある語句だけを<error>タグ</error>で囲むこと。・それ以外の補足説明、前置き、警告などは一切出力しないこと。・タグ付きの文章だけを返してください。"

A.3　GPT-2step における誤り訂正のプロンプト

prompt = "次の文章は手書き文字の OCR 結果です。<error>タグで囲まれた箇所に OCR による誤りがあります。<error>タグで囲まれた部分のみを訂正し、本来書かれていた通りの文字列に戻してください。入力は、<error>タグ付きの文と、タグなしの文が<sep>で連結された形式です。"

A.4　自動採点に用いたプロンプト

prompt = "You are a grader. Read the provided model answer and the student's answer. If you consider the student's answer correct, output "1"; if you consider it incorrect, output "0". The output must be a single character, either "1" or "0". Even if the judgment is difficult, choose and output the option that seems more likely."

B　実際の採点例

表 4　GPT-1step で誤採点された一方で GPT-2step で正しく採点できた例

項目	内容	採点
OCR 出力	がコごに父を科学者だと思っていて、科学、・者に憧れを抱に１：：にいたから	0
GPT-1step 訂正	がここに父を科学者だと思っていて、科学者に憧れを抱いていたから	0
GPT-2step 訂正	父を科学者だと思っていて、科学者に憧れを抱いていたから	1
正解（人手）	父を科学者だと思っていて、科学者に憧れを抱いていたから	1

表 5　GPT-1step で正しく採点できたものの GPT-2step では誤採点となった例

項目	内容	採点
OCR 出力	ピッケルにしがみっいた	0
GPT-1step 訂正	ピッケルにしがみついた	1
GPT-2step 訂正	ピッケルにしがみついた	0
正解（人手）	ピッケルにしがみついた	1