Few-Shot 学習を用いた未見の設問に対する手書き答案の採点

齊藤隆浩 [1]　Hung Tuan Nguyen[1]　古宮嘉那子 [1]　石岡恒憲 [1,2]　中川正樹 [1]
[1] 東京農工大学大学院　[2] 大学入試センター
s245145v@gmail.com, fx7297@go.tuat.ac.jp,
kkomiya@go.tuat.ac.jp, tunenori@rd.dnc.ac.jp, nakagawa@cc.tuat.ac.jp

概要

本稿では Few-Shot 学習を用い、学習時に未見の問題に対する日本語手書き答案の採点について検証した。手書き答案を人手及び自動のいずれかで書き起こしした解答データを、LLM に模範解答とのペアとして与えて Few-Shot 学習を行い、類似度ベースで採点を行わせた。Few-Shot 事例データには採点対象とは異なる設問に対する解答ペアを使用した。多くの場合で採点性能が Zero-Shot 学習時よりも向上し、Zero-Shot 学習と Fine-Tuning を行った場合の中間の性能を示すことが分かった。また、ノイズを含む自動書き起こし答案の採点には、Fine-Tuning を行った場合と比較して Few-Shot 学習の性能は不十分であることが分かった。さらに、条件を変えていくつかの実験を行い、複数の知見が得られた。

1　序論

近年、教育現場の業務負担軽減のため、自然言語処理技術を用いた採点プロセスの自動化が注目されている [1, 2, 3, 4, 5]。採点プロセスの完全自動化を実現する上での障壁はいくつかある。第一に、設問特化型 (Instance ベース) モデルを訓練して採点する場合、設問ごとに数千以上もの採点済み答案データを訓練データとして用意する必要がある。そのため、学校のクラスのように採点すべき設問数が多く、設問当たりの解答数が少ない環境には Instance ベースモデルの使用は向かない。この問題を解決するために、一つ以上のサンプル解答と比較することで、学習データが無いか少量でも採点可能な、類似度ベースの採点モデルが研究されている [6, 7, 8]。

第二に、教育現場では手書きで答案を作成することが一般的であり、純粋な自然言語処理モデルでは直接採点できない。このため、まず画像処理モデルを用いて手書き文字認識 (OCR) を行ってから採点モデルに入力する必要がある。この際、手書き答案の筆跡が乱雑である場合が多く、文字認識の精度によって採点精度が左右されるという課題がある。

以上の二つの問題を踏まえた研究として Saito ら [9] がある。この研究では、GPT-4o を Fine-Tuning することで、ノイズを含む手書き答案を高い精度で採点可能であることが示されている。一方で、GPT 系モデルの Fine-Tuning には高いコストがかかり、実用面で課題が残る。

本研究では、より低コストで利用可能な Few-Shot 学習を、未見の設問に対する手書き答案の類似度ベースの採点に適用し、Zero-Shot 学習や Fine-Tuning を行ったモデルとパフォーマンスを比較する。また、ノイズを含むデータを使用した場合の傾向について調査する。さらに、Few-Shot 事例 (以下、FS 事例と略記) に関する条件を変更した場合の性能の変化についても検証する。

2　関連研究

Bexte ら [6] は、類似度に基づく採点のための効率的なアーキテクチャを提案し、それが代表的なベンチマークデータセットである ASAP 上で BERT ベースの分類モデルと同等の性能を達成できることを示した。また彼らは類似度ベースの採点手法の利点を調査し、ラベル付き訓練データの必要性の低減や、設問間での汎化性能の向上などは確認されなかったものの、解釈可能なフィードバックの提供可能性に言及した [7]。さらに彼らは、類似度ベースの採点における信頼度の閾値が設問ごとに大きく異なることを示した [8]。

日本語の短答式記述問題の手書き答案の採点を行った研究として Oka ら [10] がある。彼らは手書き文字認識によるデータを用い、Instance-based モデルを訓練して採点を行った。すべての設問で高い性能を達成した一方で、約 6 万件のデータを用いても性能が収束しないことが明らかになった。Saito ら [9] は、類似度ベースの採点を手書き答案の採点に

適用した。GPT-4o を Fine-Tuning したモデルが未見の設問を高い性能で採点可能であることや、自動認識によるノイズを含む訓練データを使用することで、同様のノイズを含む答案を高い精度で採点可能であることなどが示された。

3 データ

中学生向けの国語ドリル 3 冊に含まれる 25 問の短答式記述問題およびその答案を使用する[1]。答案サンプルは各設問につき 55〜66 件、計 1,547 件あるが、実際に使用したのは白紙答案などを除外した 1,372 件である。各データは (模範解答, 生徒の解答, 正誤ラベル) の組である。生徒の解答は、手書き答案を人手で文字起こしした**人手**と、手書き文字認識システムを用いて文字起こしした**自動**の 2 種類のデータを使用する。前者はクリーンなテキストであり、後者は生徒の乱雑な筆跡に起因するノイズを多く含む。手書き文字認識システムにはアイラボ社から提供されているものを用いた [11]。正誤ラベルは教師によって採点され、値は正答/誤答[2]のバイナリである。正誤ラベルの分布は 1(正答) が 885 件 (64.5%)、0(誤答) が 487 件 (35.5%) である。

4 手法

本研究では、Fine-Tuning を行っていない GPT 系モデルを用いて、Few-Shot 学習による採点性能を評価する。モデルには、OpenAI から提供されている API を、Python の公式ライブラリを通じてアクセスして利用する。モデルの推論パラメータには `temperature=0.0`、`max_tokens=1` を設定し、その他のパラメータは全てデフォルトを使用する。

モデルに与えるプロンプトは 2 種類あり、まず system プロンプトでモデルに採点者の役割や期待される出力形式を説明する。次に、user プロンプトで FS 事例及び採点対象となる模範解答と生徒の解答を与える。ここで、採点者の役割は FS 事例を参考に、模範解答と生徒の解答を比較して採点を行うことであり、期待される出力形式は "1" (正答と判定) または "0" (誤答と判定) である。なお、user プロンプトは [12] を参考に英語で作成した。system プロンプト及び user プロンプトの例は付録 A に示した。

5 実験

デフォルトの実験設定では、FS 事例数は $n = 4$ とし、正答例と誤答例を等しい割合で無作為に選択する。また、FS 事例となる n 個の解答が属する設問は、採点対象とは異なる設問の中から、それぞれ重複しないように無作為に選択する。モデルには、GPT-4o 及び GPT-4o-mini [3] の 2 種類を用いる。FS 事例データ及び評価データには、**人手**と**自動**の 2 種類のデータをそれぞれ使用して比較する。ただし、FS 事例データに**自動**、評価データに**人手**を用いる組み合わせでは実験しない。これは、一般に人手書き起こしテキストの方が入手しづらく、実用に即さない組み合わせであると考えられるためである。

モデルの評価には、正解率 (Accuracy) を用いる。これは、全ての評価事例数に対して、モデルが正誤を正しく予測した数の割合である。

5.1 A: 学習方法の比較

5.1.1 実験設定

この実験では、全てデフォルトの設定で Few-Shot 学習を用いた採点を行い、Zero-Shot 学習及び Fine-Tuning を行う場合と比較する。

5.1.2 評価

結果を表 1 に示す。本研究の手法を `Few` としている。`Zero` 及び `FT` はベースライン (先行研究 [9] による[4]) であり、それぞれ Zero-Shot 学習、Fine-Tuning を行った場合を表している。`Few` では、訓練データの列に FS 事例データを記している。`Zero` では訓練データは用いていない。なお、設問単位で分割して実験を行ったマクロ平均の値を示す。

表 1 実験 A (学習方法の比較) 結果

モデル	訓練データ	評価データ	Accuracy		
			`Zero`	`Few`	`FT`
GPT-4o	**人手**	**人手**	78.8 %	80.7 %	**81.7 %**
	人手	**自動**	51.1 %	55.5 %	**76.0 %**
	自動			64.2 %	**80.8 %**
GPT-4o -mini	**人手**	**人手**	75.0 %	76.2 %	**78.9 %**
	人手	**自動**	52.2 %	47.4 %	**75.2 %**
	自動			53.3 %	**79.0 %**

1) 答案収集は東京農工大学において、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た (No.220707-04111)。

2) 本稿では、正答 (又は誤答) という語は、生徒が設問に正答 (又は誤答) することを指す。正解 (又は Accuracy) という語は、採点システムが正しく (=教師によって採点された正誤ラベルと等しく) 採点すること (又はその割合) を指す。

3) `gpt-4o-2024-08-06`, `gpt-4o-mini-2024-07-18`

4) 引用元と同じ手法を用いたが、表 1 の実験結果は推論パラメータのみ本研究と同一に変えて実験し直したものである。

人手データに対する採点では、Few は Zero と FT の中間の性能を示した。Few でモデルに与えている FS 事例の数 (4 件) は、Zero(0 件) と FT(数百件)[5] の中間にあたるため、予想通りの結果である。

自動データに対する採点では、Few は両モデルにおいて、**人手**データを FS 事例にした場合よりも**自動**データを FS 事例にした場合の方が性能が高かった。これは、与えられた FS 事例から、**自動**データがもつノイズに関する情報をモデルが得ているためと考えられ、FT と場合と同様の傾向である。一方で、**自動**データで Fine-Tuning した場合は、**人手**データに対する採点と遜色ない Accuracy が得られたのに対し、**自動**データを FS 事例にして Few-Shot 学習を利用した場合の Accuracy は最大でも 60 %台と著しく低かった。このことから、数件の FS 事例だけでは、**自動**データがもつノイズの傾向などをモデルが得るには不十分であることが分かった。

5.2 B: Few-Shot 事例数の変更

5.2.1 実験設定

この実験では、FS 事例数をデフォルトの $n = 4$ から、$n = 2, 8$ に変化させて性能の変化を観察する。

5.2.2 評価

表 2 実験 B (FS 事例数の変更) 結果

モデル	FS 事例データ	評価データ	Accuracy (n: FS 事例数)		
			$n = 2$	$n = 4$	$n = 8$
GPT-4o	**人手**	**人手**	80.0 %	80.7 %	**80.8 %**
	人手	**自動**	53.5 %	55.5 %	**56.4 %**
	自動		59.6 %	64.2 %	**69.8 %**
GPT-4o -mini	**人手**	**人手**	**76.5 %**	76.2 %	75.0 %
	人手	**自動**	**48.5 %**	47.4 %	46.9 %
	自動		52.1 %	53.3 %	**55.8 %**

6 つの条件のうち、4 つは FS 事例数の増加に従って Accuracy が上昇し、2 つは低下した。FS 事例数の増加に従って Accuracy が大きく増加したのは、両モデルで FS 事例データ、評価データとも**自動**を使用した場合である。この条件では、**自動**がもつノイズの情報を FS 事例でなるべく多く与えることが性能向上に寄与すると考えられる。一方、FS 事例数の増加に従って Accuracy が低下したのは、GPT-4o-mini で FS 事例データに**人手**を使用した 2 条件である。

5) Fine-Tuning では五分割交差検定を行っているため、訓練データは 1372 件のうち 5 分の 3 である。

このうち、評価データには**自動**を用いた場合では、FS 事例データが評価データと類似していないため、FS 事例を与えるほどモデルが混乱して性能が低下するのであろう。

5.3 C: Few-Shot 事例の正誤割合の変更

5.3.1 実験設定

この実験では、FS 事例数は $n = 8$ に固定したうえで、FS 事例中における正答例と誤答例の割合をデフォルトの (正 : 誤) = 4 : 4 から、(正 : 誤) = 1 : 7 または (正 : 誤) = 7 : 1 に変化させて性能の変化を観察する。

5.3.2 評価

この実験では、正答例の割合が大きい方が性能が向上すると予想していた。その根拠として、(1) ラベルの分布が正答に偏っている、(2) 正答例は複数あるであろう採点基準にすべて合致していると考えられるので、採点基準に関する情報をより多く含んでいると考えられる、の 2 点である。さらに**自動**データの採点では、(3) ノイズを含んでいるために誤答に見える解答のうち、どれが実際には正答であるかを判断するためには正答例が有効な情報になる、という理由も挙げられる。

表 3 実験 C (FS 事例の正誤割合の変更) 結果

モデル	FS 事例データ	評価データ	Accuracy (正 : 誤)		
			1:7	4:4	7:1
GPT-4o	**人手**	**人手**	**82.0 %**	80.8 %	80.0 %
	人手	**自動**	**58.4 %**	56.4 %	52.0 %
	自動		66.6 %	69.8 %	**72.1 %**
GPT-4o -mini	**人手**	**人手**	**77.2 %**	75.0 %	74.3 %
	人手	**自動**	**47.3 %**	46.9 %	45.7 %
	自動		54.7 %	55.8 %	**58.2 %**

実際の実験結果は表 3 のようになった。FS 事例データが**人手**の場合、予想に反して誤答例の割合が多い方が性能が高かった。特に、正答例の割合が多いほど、モデルが解答を正答と判定する割合は却って減少し、Recall が低下するという現象が観察された (付録 B を参照)。この現象を説明する仮説として、モデルは「FS 事例の分布に左右されずに出力分布を調整しようとする」性質を持っており、それが過剰に発現したのではないかと考えている。

一方で、FS 事例データ及び評価データに**自動**を用いた場合では、事前の予想通り、正答例が多いほ

ど性能が高くなった。これは、事前予想の根拠 (3) によるものと考えている。

5.4 D: Few-Shot 事例とする設問の変更

5.4.1 実験設定

この実験では、FS 事例とする設問を変えた場合に性能がどのようにばらつくかを観察する。評価データの各設問に対して、FS 事例として与える n 題の設問を無作為に選んで採点実験を行う。この試行を 20 回繰り返し、Accuracy のばらつきを観察する。モデルは GPT-4o-mini に固定する。FS 事例数 n は 4 または 8 とする。

5.4.2 評価

実験結果の概要を表 4 に示す。全ての設問・試行における Accuracy の平均、設問内標準偏差 (FS 事例の設問に対する Accuracy のばらつきの平均) 及び設問間標準偏差 (評価事例の設問に対する Accuracy のばらつき) を掲載する。

表 4 実験 D (FS 事例とする設問の変更) 結果

FS 事例データ	評価データ	事例数 n	Accuracy		
			平均	標準偏差	
				設問内	設問間
人手	**人手**	4	75.1 %	3.8 %	6.3 %
		8	74.0 %	3.3 %	7.1 %
人手	**自動**	4	47.7 %	2.4 %	11.3 %
		8	46.8 %	2.0 %	11.4 %
自動	**自動**	4	53.7 %	3.2 %	10.1 %
		8	55.9 %	3.1 %	10.4 %

FS 事例数 n がいずれの場合でも、設問内標準偏差よりも、設問間標準偏差のほうが大きかった。つまり、FS 事例と採点対象の設問の相性よりも、採点対象の設問自体の採点難易度の方が採点性能に大きく影響すると言える。これは、FS 事例としてより効果的な設問を選ぶだけでは、採点難易度の高い設問に対する採点性能を向上させるのに限界があることを示唆している。

さらに、FS 事例数 $n = 4$ と $n = 8$ の場合を比較すると、設問間標準偏差は FS 事例数 $n = 8$ のほうが大きかったが、設問内標準偏差は $n = 8$ のほうが小さかった。FS 事例数を増加させることで、FS 事例データに選ぶ設問による性能の上下を抑えて安定した採点が可能になると考えられる。

次に、採点するデータを比較すると、**自動**データを採点する場合は、**人手**データを採点する場合より設問間標準偏差が大きかった。このことから、**自動**データは設問による採点難易度のばらつきが大きいと考えられる。このばらつきが、手書き文字認識システムの設問ごとの認識精度のばらつきと、採点システム自体の特性のどちらに起因するものなのかは今後の検討を要する。

6 まとめ

本研究では Few-Shot 学習を用い、学習時に未見の問題に対する日本語手書き答案の採点について検証した。手書き答案を人手または自動で書き起こしした解答データを、LLM に模範解答とのペアとして与えて Few-Shot 学習を行い、採点を行わせた。FS 事例データには採点対象とは異なる設問に対する解答ペアを使用した。実験の結果、多くの場合で Zero-Shot 学習と Fine-Tuning を行った場合の中間の性能を示すことが分かった。また、ノイズを含む自動書き起こし答案の採点には、Fine-Tuning を行った場合と比較して Few-Shot 学習の性能は不十分であることが分かった。さらに、**人手**データの採点には FS 事例の数や正誤割合の変化に対して採点システムが頑健であることや、**自動**データの採点には FS 事例数および正答例の割合が多いほどよいことが分かった。加えて、FS 事例の設問による採点性能のばらつきは両データともに小さいものの、**自動**データの採点においては評価事例の設問による採点性能のばらつきが大きくなることが分かった。

本研究では、採点対象の設問に対する解答が訓練 (FS 事例) データとして入手しづらいことを前提として、FS 事例データに採点対象とは異なる設問に対する解答を用い、Few-Shot 学習を Zero-Shot 学習及び Fine-Tuning を行った場合と比較した。一方で、Few-Shot 学習では数個の FS 事例を用いて採点が可能であり、採点対象と同じ設問の解答データの入手が容易である。採点対象と同じ設問の解答を FS 事例とした場合における採点性能向上の有無や、利用可能な FS 事例数が限られている場合の性能のばらつきの検証を今後の研究課題としたい。また、実験 C (FS 事例の正誤割合の変更) で観察された、モデルの出力分布が FS 事例の分布と逆の傾向を示す現象が、GPT 系モデル固有の特性によるものなのかについても調査したい。

謝辞

本研究は、科研費 JP23H03511 と JP24H00738 の助成を受けたものです。また、本研究において答案データを提供いただいたワコム株式会社、および文字認識データを提供いただい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] Aoife Cahill, James H Fife, Brian Riordan, Avijit Vajpayee, and Dmytro Galochkin. Context-based automated scoring of complex mathematical responses. In **Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 186–192, 2020.

[2] Tasuku Sato, Hiroaki Funayama, Hanawa kazuaki, Yuya Asazuma, and Kentaro Inui. Explanation of the automatic grading results based on the reference sections. **28th The Association for Natural Language Processing**, pp. 459–464, 2022.

[3] Yuya Asazuma, Hiroaki Funayama, Yuichiro Matsubayasi, Tomoya mizumoto, and Kentaro Inui. Verification of consistency with scoring criteria for the descriptive answer grading model. **29th The Association for Natural Language Processing**, pp. 1868–1873, 2023.

[4] Lui Yoshida. The impact of example selection in few-shot prompting on automated essay scoring using gpt models. In Andrew M. Olney, Irene-Angelica Chounta, Zitao Liu, Olga C. Santos, and Ig Ibert Bittencourt, editors, **Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky**, pp. 61–73, Cham, 2024. Springer Nature Switzerland.

[5] Wenchao Li and Haitao Liu. Applying large language models for automated essay scoring for non-native Japanese. **Palgrave Communications**, Vol. 11, No. 1, pp. 1–15, December 2024.

[6] Marie Bexte, Andrea Horbach, and Torsten Zesch. Similarity-based content scoring - how to make S-BERT keep up with BERT. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch, editors, **Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)**, pp. 118–123, Seattle, Washington, July 2022. Association for Computational Linguistics.

[7] Marie Bexte, Andrea Horbach, and Torsten Zesch. Similarity-based content scoring - a more classroom-suitable alternative to instance-based scoring? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 1892–1903, Toronto, Canada, July 2023. Association for Computational Linguistics.

[8] Marie Bexte, Andrea Horbach, Lena Schützler, Oliver Christ, and Torsten Zesch. Scoring with confidence? – exploring high-confidence scoring for saving manual grading effort. In Ekaterina Kochmar, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan, editors, **Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)**, pp. 119–124, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[9] Takahiro Saito, Hung Tuan Nguyen, Kanako Komiya, Tsunenori Ishioka, and Masaki Nakagawa. Similarity-based scoring model for handwritten answers in japanese workbooks. In Alexandra I. Cristea, Erin Walker, Yu Lu, Olga C. Santos, and Seiji Isotani, editors, **Artificial Intelligence in Education**, pp. 470–477, Cham, 2025. Springer Nature Switzerland.

[10] 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.

[11] Hung Tuan Nguyen, Thanh-Nghia Truong, Nam Tuan Ly, Masaki Nakagawa, and Toshihiko Horie. Automatic scoring for handwritten answers from elementary to junior high school grades in japan. In **IEICE Tech. Rep.**, Japan, 2025. IEICE.

[12] Chenyan Zhao, Mariana Silva, and Seth Poulsen. Language models are few-shot graders. In Alexandra I. Cristea, Erin Walker, Yu Lu, Olga C. Santos, and Seiji Isotani, editors, **Artificial Intelligence in Education**, pp. 3–16, Cham, 2025. Springer Nature Switzerland.

付録

A　プロンプト

```
You are a grader. First, several model answers, student answers, and their
correctness are given as examples. Then, read the provided model answer and
the student's answer given last, which is the target to be graded.
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.
```

図 1　system プロンプト

```
Example Model Answer: 戦後三年しかたっていないので、すべてが乏しい状況。
Example Student Answer: 戦争が終ってまだ三年で、すべてのものが乏しかった時代。
Grade to this example student answer: 1

Example Model Answer: 完全に移動するための道具にすることで、手が自由に使えるようになった
Example Student Answer: 四本足での歩行には戻らずに、後ろ足だけを使い二本足で地上を歩行する
Grade to this example student answer: 0

Example Model Answer: 客観的な原理に基づいて制作された作品。
Example Student Answer: 客観的な原理に基づく秩序が美を生み出す
Grade to this example student answer: 0

Example Model Answer: いくら大声で叫んでみても、それがどうなるものでもないことに気づいたから。
Example Student Answer: 大声で叫んでも？？？でもないことに気づいたから。
Grade to this example student answer: 1

Example Model Answer: 東洋では、身体と精神を結びつける思想があるので、姿勢を整えて臨むことが非常に重要である。
Example Student Answer: 東洋では身体と精神を結びつける思想がある
Grade to this example student answer: 1

Model Answer: 斉藤多恵の姿を見られるかもしれないと期待する気持ち。
Student Answer: もしかしたら斉藤多恵が見えるのではないかと期待する気持ち。
How would you grade this? Please return '1' for correct and '0' for incorrect.
```

図 2　user プロンプト

B　実験 C (FS 事例の正誤割合の変更): 実験結果 (Presicion 及び Recall)

実験 C (Few-Shot 事例の正誤割合の変更) の実験結果 (Precision と Recall) を示す。Precision は、正答例の割合が多いほうが高い。Recall は、FS 事例データが**人手**の場合は誤答例が多いほうが高く、**自動**の場合は正答例が多いほうが高い。

表 5　実験 C (FS 事例の正誤割合の変更) 結果 (Precision)

モデル	FS 事例データ	評価データ	Precision (正 : 誤)		
			1:7	4:4	7:1
GPT-4o	**人手**	**人手**	84.1 %	85.5 %	**87.4 %**
	人手	**自動**	83.7 %	**88.7 %**	87.5 %
	自動		83.3 %	84.6 %	**86.2 %**
GPT-4o -mini	**人手**	**人手**	91.3 %	92.9 %	**93.2** %
	人手	**自動**	90.9 %	92.8 %	**93.8 %**
	自動		91.1 %	89.2 %	**92.8 %**

表 6　実験 C (FS 事例の正誤割合の変更) 結果 (Recall)

モデル	FS 事例データ	評価データ	Recall (正 : 誤)		
			1:7	4:4	7:1
GPT-4o	**人手**	**人手**	**89.4 %**	84.8 %	80.7 %
	人手	**自動**	**44.3 %**	37.4 %	30.0 %
	自動		60.7 %	65.4 %	**67.6 %**
GPT-4o -mini	**人手**	**人手**	**71.3 %**	66.4 %	64.9 %
	人手	**自動**	**20.3 %**	19.3 %	16.9 %
	自動		33.0 %	36.0 %	**38.3 %**