大規模言語モデルが持つ選好情報と少数正解事例の統合による絶対評価較正を用いた複数観点同時小論文自動採点

柴田拓海 [1]　宮村祐一 [1]
[1] 有限責任監査法人トーマツ デロイトアナリティクス R&D
{takumi.shibata, yuichi.miyamura}@tohmatsu.co.jp

概要

近年，大規模言語モデル（LLM）を用いた小論文自動採点手法が多数提案されている．特に小論文答案間のペアワイズ比較に基づいて得点を算出する従来手法は高精度である一方，推定された得点が相対評価に留まり，ルーブリックに基づく絶対評価得点の算出が困難であるという課題がある．また，従来手法は単一観点の評価に限られるため，結果の解釈性にも課題が残る．そこで本研究では，LLM の選好データに基づく学習と，少数の正解事例を用いた学習とを統合し，絶対評価得点への較正と複数観点採点を同時に実現する手法を提案する．

1　はじめに

小論文自動採点（Automated Essay Scoring：AES）は，小論文の採点を機械学習を用いて自動化するタスクである．AES は採点にかかるコストの削減や公平性を担保する技術の一つとして近年注目されている [1].

AES 研究の多くは，教師あり問題固有型と問題横断型のタスクを対象としている [2]．教師あり問題固有型手法では，採点対象の小論文問題の採点済み答案を用いて深層学習モデル等の機械学習モデルを学習し，同じ小論文問題に対して採点を行う．一方で，問題横断型手法は，採点対象の問題に対しては十分な量の答案が得られていないもしくは存在しないが，それ以外の問題に対しては十分な量の採点済み小論文が得られている状況を仮定し，ドメイン適応やドメイン汎化等の技術を用いて採点対象の問題に対する採点モデルの構築を目指す．

一方で，近年の大規模言語モデル（Large Language Model：LLM）の言語理解能力や推論能力を活用して，教師データを用いずにゼロショットで小論文採点を行う手法がいくつか提案されている [3, 4, 5, 6]．例えば，柴田・宮村 [3] は，多数の小論文ペアの優劣を LLM に比較させ，そのペアワイズ選好データを Bradley–Terry モデルに基づく出力層を持つ深層ニューラルネットワークで統合し，各小論文の得点を算出する LLM Comparative Essay Scoring (LCES) を提案している．

しかし，この LLM を用いたペアワイズ比較に基づくアプローチは有望である一方，次のような課題が残る．(1) ペアワイズ選好データにより推定される得点は相対尺度に留まり，ルーブリックに基づく絶対的な得点スケールを推定することができない．(2) 単一の観点に関する採点しか行えず，評価観点ごとの得点（フィードバック）が得られないため，採点結果の解釈性が低い．(3) ペアワイズ比較時における LLM の位置バイアスに対応するために，順番を入れ替えて合計 2 回，同じ小論文ペアに対して問い合わせを行っており計算コストが高い．

そこで本研究では，これらの課題を解決するため，LCES を拡張した手法を提案する．提案手法は，複数観点採点モデルを共通潜在表現に基づくマルチタスク学習モデルとして定式化する．この枠組みの中で，LLM のペアワイズ選好データからの学習と少数の採点済みデータによる絶対評価較正とを，位置バイアスを抑制しつつ統合的に実行する．

提案手法は，複数観点小論文採点において一般的な ASAP，ASAP++データセットを用いて，得点予測精度の評価実験を行う．実験により，提案手法は，従来の手法と比較して高い得点予測精度を達成することを示す．

2　提案手法

本研究では，小論文 $\mathcal{D} = \{x_i \mid i \in \mathcal{I}\}$ に対して評価観点別の得点 $\{y_{ik} \mid i \in \mathcal{I}, k \in \mathcal{K}\}$ を予測することを目指す．ここで，x_i は $i \in \mathcal{I} = \{1, \ldots, N\}$ 番目の小論文を表し，y_{ik} は小論文 x_i における $k \in \mathcal{K} = \{1, \ldots, K\}$

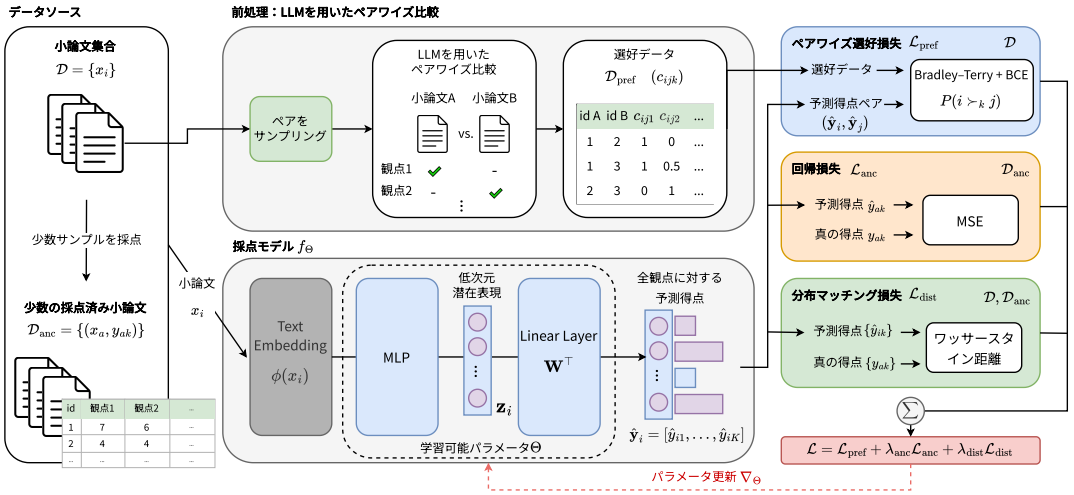

図 1　提案手法の概念図.

番目の評価観点の得点を表す．また N，K はそれぞれ小論文の数と評価観点の数を表す．さらに本研究では小論文集合 $\mathcal{D}$ のうち，少数の小論文 $\{x_a \mid a \in \mathcal{A} \subset \mathcal{I}\}$ について人手採点による正解得点データ $\{y_{ak} \mid a \in \mathcal{A}, k \in \mathcal{K}\}$ が得られている状況を仮定する．ここで $\mathcal{A}$ は $\mathcal{I}$ からサンプリングされた小論文のインデックス集合であり $|\mathcal{A}| \ll N$ を想定している．

これらの状況のもとで，本研究の目的は，小論文 $\mathcal{D}$ に対して，LLM のペアワイズ選好に基づく学習を行いつつ，ルーブリックで定義された絶対的な得点スケールに近づけるようにアンカーとなる小論文 $\mathcal{D}_{\text{anc}} = \{(x_a, y_{ak}) \mid a \in \mathcal{A}, k \in \mathcal{K}\}$ を用いて絶対評価較正を行うことで，高精度な得点予測を目指すことである．なお，実際にモデルが得点予測を行うのは $\{x_i \mid i \in \mathcal{I} \setminus \mathcal{A}\}$ であることに注意されたい．

次節から，具体的な複数観点採点モデルの定式化（2.1）およびその学習方法（2.2）について説明する．

2.1　複数観点採点モデル

本研究では，各小論文 x_i の評価観点別の得点 $\boldsymbol{y}_i = [y_{i1}, \ldots, y_{iK}]^\top$ を予測する採点モデル $f_{\boldsymbol{\Theta}}$ を低次元の共有潜在表現に基づいて定式化する．この定式化は，小論文試験では複数の評価観点の背後に，それらの得点を共通して説明する因子が存在するという考えに基づいている [7]．具体的には，評価観点別の得点 $\boldsymbol{y}_i$ を次式で算出する．

$$\boldsymbol{y}_i = f_{\boldsymbol{\Theta}}(x_i) = \boldsymbol{W}^\top \boldsymbol{z}_i(x_i) \tag{1}$$

ここで，$\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K] \in \mathbb{R}^{d \times K}$ は k 番目の評価観点の重みベクトル $\boldsymbol{w}_k \in \mathbb{R}^d$ からなる行列，$\boldsymbol{z}_i(x_i) \in \mathbb{R}^d$ は x_i の観点に共通する潜在表現を表すベクトルである．また，d は潜在表現ベクトルの次元を表し，$d < K$ を想定している．

また，式 (1) の潜在表現ベクトル $\boldsymbol{z}_i$ を得るため，本研究では，まず小論文 x_i を重み固定の文章埋め込みモデルに入力し，その出力をさらに多層パーセプトロン（Multi-Layer Perceptron：MLP）で変換する．具体的には，d_{in} 次元のベクトルを出力する文章埋め込みモデルを ϕ とすると，

$$\boldsymbol{z}_i(x_i) = \mathrm{MLP}(\phi(x_i)) = \boldsymbol{W}_2 \tanh(\boldsymbol{W}_1 \phi(x_i) + \boldsymbol{b}_1) + \boldsymbol{b}_2 \tag{2}$$

のように算出できる．ここで，$\boldsymbol{W}_1 \in \mathbb{R}^{d \times d_{in}}$，$\boldsymbol{W}_2 \in \mathbb{R}^{d \times d}$ は重み行列，$\boldsymbol{b}_1 \in \mathbb{R}^d$，$\boldsymbol{b}_2 \in \mathbb{R}^d$ はバイアスベクトル，tanh は双曲線正弦関数を表す．この採点モデル $f_{\boldsymbol{\Theta}}$ の学習可能パラメータ $\boldsymbol{\Theta} = \{\boldsymbol{W}, \boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{b}_1, \boldsymbol{b}_2\}$ は次節で説明する 3 つの目的関数を最適化することで学習される．

2.2　モデル学習

本節では，前節で定式化した採点モデル $f_{\boldsymbol{\Theta}}$ を学習するための目的関数を定義する．目的関数は，LLM が生成するペアワイズ選好に基づく損失関数 $\mathcal{L}_{\text{pref}}$ と，少数の採点済み小論文 $\mathcal{D}_{\text{anc}}$ を用いた絶対評価較正のための 2 つの損失関数，すなわち回帰損失 $\mathcal{L}_{\text{anc}}$ および分布マッチング損失 $\mathcal{L}_{\text{dist}}$ から構成される．これらを重み付き和で合成し，

$$\mathcal{L} = \mathcal{L}_{\text{pref}} + \lambda_{\text{anc}}\mathcal{L}_{\text{anc}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} \tag{3}$$

と定義する．ここで，λ_{anc}，λ_{dist} はそれぞれの損失項に対する重み係数である．採点モデル $f_{\boldsymbol{\Theta}}$ のパラメータ $\boldsymbol{\Theta}$ は，この損失関数 $\mathcal{L}$ を最小化することで学習される．以下，各損失項の詳細を述べる．

2.2.1 ペアワイズ選好損失

LLM のペアワイズ選好に基づく学習では，まず小論文集合 $\mathcal{D}$ から，LLM を用いたペアワイズ比較を行うペアの集合 $\mathcal{P}_{\text{sub}}$ を決定する．具体的には，$\mathcal{D}$ 内の小論文の任意の組み合わせ集合 $\mathcal{P} = \{(i,j) \mid i \neq j, x_i, x_j \in \mathcal{D}\}$ から M 個をランダムにサンプリングして，ペア集合 $\mathcal{P}_{\text{sub}}$ を生成する．

次に，上記のようにして作成したペア集合に対して LLM を用いて実際にペアワイズ比較を行うことで，モデル学習用のデータセット $\mathcal{D}_{\text{pref}} = \{(i,j,k,c_{ijk}) \mid (i,j) \in \mathcal{P}_{\text{sub}}, k \in \mathcal{K}\}$ を生成する．ここで，c_{ijk} はペア (i,j) の k 番目の評価観点に対する LLM の比較結果を表す．本研究では，$c_{ijk} = 1$ はペア (i,j) の k 番目の評価観点において小論文 x_i が x_j よりも優れていることを，$c_{ijk} = 0$ はその逆を，$c_{ijk} = 0.5$ は両者が同等であることを表す．

最後に学習のための損失関数を定義する．ペアワイズ選好学習では，各ペア (i,j) に対して，式 (1) における採点モデルを用いて各小論文の k 番目の評価観点の得点 $(\hat{y}_{ik}, \hat{y}_{jk})$ を算出する．これらの得点を所与としたとき，小論文 x_i が小論文 x_j よりも観点 k において高い得点を得る確率 $P(i \succ_k j \mid \hat{y}_{ik}, \hat{y}_{jk})$ を Bradley–Terry モデルでモデル化する．具体的には，

$$P(i \succ_k j) = \sigma(\hat{y}_{ik} - \hat{y}_{jk} + \delta_{\text{pos}}) = \sigma(\boldsymbol{w}_k^{\top}(\boldsymbol{z}_i - \boldsymbol{z}_j) + \delta_{\text{pos}})$$

で計算する．ここで，σ はシグモイド関数，δ_{pos} は LLM が持つ位置バイアスを補正する項を表し，学習可能なパラメータとして設定する．なお，この補正項は Davidson and Beaver による提示順序効果のモデル化に基づいている [8, 9]．この確率値 $P(i \succ_k j)$ を LLM が出力した選好情報と整合するように，以下のバイナリ交差エントロピー損失を用いて学習を行う．具体的には，

$$\mathcal{L}_{\text{pref}} = -\frac{1}{MK} \sum_{\mathcal{D}_{\text{pref}}} \Big[c_{ijk} \log P(i \succ_k j) + (1 - c_{ijk}) \log\big(1 - P(i \succ_k j)\big) \Big]$$

を損失関数とする．この学習は，LLM が生成したペアワイズ選好データを最もよく説明する得点 $\hat{y}_{ik}$ を学習していると解釈できる．

2.2.2 回帰損失

前述のペアワイズ選好学習では小論文間の得点の相対的な関係のみを学習するため，予測される得点がルーブリックに基づく得点段階と乖離する可能性がある．そこで本節では，少数の採点済み小論文 $\mathcal{D}_{\text{anc}}$ を用いた回帰学習により，予測スコアの絶対的な値を較正する．これにより，実際の採点スケールに整合した得点予測が可能になるとともに，LLM の選好データに含まれるノイズの補正も期待できる．

具体的には，採点モデルによる予測得点 $\hat{y}_{ak}$ と真の得点 y_{ak} との最小二乗誤差を最小化する．このとき，評価観点 k によって得点段階の多寡が異なる場合，そのままでは損失のスケールが観点ごとに不均一となる．そこで，ルーブリックで定義された観点 k の最大点 $r_{max}^{(k)}$ と最小点 $r_{min}^{(k)}$ の差 $R_k = r_{max}^{(k)} - r_{min}^{(k)}$ で予測得点と真の得点をそれぞれ正規化する．これにより，すべての評価観点を等しく重み付けた損失関数を次式で表せる．

$$\mathcal{L}_{\text{anc}} = \frac{1}{|\mathcal{A}| \cdot K} \sum_{a \in \mathcal{A}} \sum_{k=1}^{K} \left\| \frac{\hat{y}_{ak} - r_{min}^{(k)}}{R_k} - \frac{y_{ak} - r_{min}^{(k)}}{R_k} \right\|^2 \quad (4)$$

2.2.3 分布マッチング損失

回帰損失 $\mathcal{L}_{\text{anc}}$ はアンカー小論文上でのサンプル単位の誤差を抑える一方で，小論文集合 $\mathcal{D}$ 全体における予測得点の分布形状が調整されるとは限らない．そこで本研究では，観点 k ごとに予測得点 $\{\hat{y}_{ik}\}_{i \in \mathcal{I}}$ の分布とアンカー得点 $\{y_{ak}\}_{a \in \mathcal{A}}$ の分布を近づけるため，これらの分布間の 1-ワッサースタイン距離を最小化する．具体的には，それぞれの分布の各経験分布の一般化逆累積分布関数 $F^{-1}(p) = \inf\{x \in \mathbb{R} \mid F(x) \geq p\}$ を用いて，分布マッチング損失 L_{dist} を以下のように定義する．

$$\mathcal{L}_{\text{dist}} = \frac{1}{K} \sum_{k=1}^{K} \int_0^1 \left| F_{\hat{\mathcal{Y}}}^{-1}(p) - F_{\mathcal{Y}_{\text{anc}}}^{-1}(p) \right| dp \quad (5)$$

ここで，$F_{\hat{\mathcal{Y}}}^{-1}$，$F_{\mathcal{Y}_{\text{anc}}}^{-1}$ はそれぞれ，予測得点とアンカー得点の経験分布に対応する一般化逆累積分布関数を表す．

3 実験

本章では，提案手法の得点予測精度を評価する実験について述べる．

3.1 データセット

本実験では，得点予測精度を評価するためのデータセットとして，Automated Student Assessment Prize (ASAP)[1] および ASAP++ [10] を用いた．ASAP は Kaggle のコンペティションで初めて利用されて以降，多くの自動採点研究においてベンチマーク

1） https://www.kaggle.com/c/asap-aes

表 1　得点予測精度の評価結果．各小論文問題（P1〜P8）の列では，その問題に付属する評価観点に対する平均値を示している．また各手法のうち，平均値が最も高い手法を**太字**で示している．

Model	Method	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
Gemma 3n E2B	Zero-shot	0.272	0.494	0.404	0.262	0.313	0.271	0.190	0.059	0.283
	Few-shot (1/Score)	0.036	0.160	0.546	0.553	0.398	0.493	0.509	0.064	0.345
	Few-shot (2/Score)	0.061	0.112	0.553	0.345	0.430	0.392	0.483	0.093	0.309
	Few-shot (3/Score)	0.020	0.038	0.473	0.303	0.410	0.414	0.369	0.056	0.260
	SFT	0.520	0.606	0.484	0.621	0.286	0.651	0.190	0.022	0.423
	Proposed	**0.695**	**0.676**	**0.672**	**0.736**	**0.680**	**0.711**	**0.702**	**0.523**	**0.674**
GPT-5 mini	Zero-shot	0.226	0.500	0.464	0.564	0.505	0.392	0.200	0.379	0.404
	Few-shot (1/Score)	0.345	0.549	0.534	0.609	0.522	0.576	0.195	0.137	0.433
	Few-shot (2/Score)	0.382	0.587	0.525	0.649	0.534	0.609	0.281	0.162	0.466
	Few-shot (3/Score)	0.401	0.569	0.560	0.637	0.567	0.639	0.372	0.171	0.490
	Proposed	**0.641**	**0.682**	**0.700**	**0.748**	**0.707**	**0.722**	**0.576**	**0.652**	**0.678**
GPT-5	Zero-shot	0.280	0.528	0.510	0.601	0.588	0.488	0.235	0.523	0.469
	Few-shot (1/Score)	0.451	0.655	0.605	0.674	0.641	0.640	0.258	0.305	0.529
	Few-shot (2/Score)	0.519	0.649	0.656	0.714	0.675	0.668	0.355	0.418	0.582
	Few-shot (3/Score)	0.538	0.679	0.672	0.743	0.687	0.686	0.421	0.416	0.605
	Proposed	**0.676**	**0.699**	**0.687**	**0.750**	**0.703**	**0.723**	**0.631**	**0.675**	**0.693**

データセットとして広く用いられている．ASAP には，8 種類の小論文問題に対する受検者の答案 12,978 個と，各答案に対する総合得点が含まれている．また，ASAP++では，ASAP データセットと同じ答案に対して評価観点別の得点が付与されている．

3.2　ベースライン

提案手法と比較する手法として以下を採用した．

ゼロショット（Zero-shot）： LLM に小論文問題，ルーブリック，答案を入力して得点を出力させる手法．採点は観点ごとに独立に行った．

少数ショット（Few-shot）： この手法では，LLM にゼロショット採点で述べた 3 つの入力情報に加え，少数の正解事例を与えた上で採点対象の小論文の得点を出力させる．本実験では，各得点段階につき 1，2，3 個のサンプルを与える 3 パターンを用いた．より多くの正解事例を与えることも可能であるが，入力トークン数の増加に伴い多大なコストがかかるため，本実験では上記のパターンに限定した．

教師あり微調整（Supervised Fine-Tuning：SFT）： ゼロショット採点と同じプロンプトテンプレートを利用し，アンカー集合 $\mathcal{D}_{\text{anc}}$ に対して正解得点を出力するように LLM を微調整する手法である[2)]．本実験では，全観点を同時に学習させた．

3.3　実験設定

LLM． 本実験では，Gemma 3n E2B [11]，GPT-5 mini，GPT-5 の 3 つの LLM[3)] に対してベースラインおよび提案手法を適用し，得点予測精度を評価した．

ハイパーパラメータ． アンカー集合 $\mathcal{D}_{\text{anc}}$ の大きさは 100 とし，LLM による小論文のペア比較生成数 M は 5000 と設定した．小論文の埋め込みモデル ϕ としては，OpenAI の文章埋め込みモデルである *text-embedding-3-large* を使用した．λ_{anc}，λ_{dist} はそれぞれ 5.0，3.5 と設定した．モデルの最適化には，学習率を 0.001 に設定した Adam [12] を用い，50 エポック学習した．また，共有潜在表現の次元数 d は 3 とした．

評価指標． 本実験では，モデルの評価指標として二次重み付きカッパ係数（Quadratic Weighted Kappa：QWK）を採用した．QWK は二者の評価者（例：モデルの予測得点と真の得点）が与える得点の一致度を測る指標であり，多くの AES 研究で使用されている [13, 14, 15, 16]．

3.4　結果

実験結果を表 1 に示す．表より，実験で使用したすべての LLM およびすべての小論文問題において，提案手法がベースラインを上回る結果を得た．また，Gemma 3n E2B といった小規模モデルにおいても提案手法が高精度を達成していることが読み取れる．これらの結果から，提案手法が最も高い得点予測性能を持つことが確認できる．

4　おわりに

本研究では LLM が生成した選好情報に少数の採点済み小論文情報を統合することで絶対評価較正を行う複数観点採点手法を提案した．提案手法は従来手法と比べて高い得点予測精度を達成した．より詳細な提案手法の性質分析は今後の課題としたい．

2）　2025/12/15 現在，GPT-5，GPT-5 mini は SFT に未対応であるため，SFT は実施しなかった．

3）　それぞれ，gemma-3n-E2B-it，gpt-5-mini-2025-08-07，gpt-5-2025-08-07 を使用した．

謝辞

本研究は有限責任監査法人トーマツの研究環境のもとで実施した.

参考文献

[1] Masaki Uto. A review of deep-neural automated essay scoring models. **Behaviormetrika**, Vol. 48, No. 2, pp. 1–26, 2021.

[2] Takumi Shibata and Masaki Uto. Cross-prompt automated essay scoring via reinforcement learning-based data valuation. **IEEE Access**, Vol. 13, pp. 184792–184808, 2025.

[3] Takumi Shibata and Yuichi Miyamura. LCES: Zero-shot automated essay scoring via pairwise comparisons using large language models. In **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 29976–29989, 2025.

[4] Atsushi Mizumoto and Masaki Eguchi. Exploring the potential of using an AI language model for automated essay scoring. **Research Methods in Applied Linguistics**, Vol. 2, No. 2, p. 100050, 2023.

[5] Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. Unleashing large language models' proficiency in zero-shot essay scoring. In **Findings of the Association for Computational Linguistics, the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 181–198, 2024.

[6] Jinhee Jang, Ayoung Moon, Minkyoung Jung, YoungBin Kim, and Seung Jin Lee. LLM agents at the roundtable: A multi-perspective and dialectical reasoning framework for essay scoring. In **Findings of the Association for Computational Linguistics, the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 19674–19687, 2025.

[7] Takumi Shibata and Masaki Uto. Analytic automated essay scoring based on deep neural networks integrating multidimensional item response theory. In **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 2917–2926, 2022.

[8] Roger R. Davidson and Robert J. Beaver. On extending the bradley-terry model to incorporate within-pair order effects. **Biometrics**, Vol. 33, No. 4, pp. 693–702, 1977.

[9] Satoshi Usami. Analyzing paired-comparison data in the situation where judgment is affected by multiple factors. **The Japanese Journal of Psychology**, Vol. 79, No. 6, pp. 536–541, 2009.

[10] Sandeep Mathias and Pushpak Bhattacharyya. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation**, 2018.

[11] Gemma Team. Gemma 3n. 2025.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **Proceedings of the 3rd International Conference on Learning Representations**, 2015.

[13] Xia Li and Wenjing Pan. KAES: Multi-aspect shared knowledge finding and aligning for cross-prompt automated scoring of essay traits. In **Proceedings of the 39th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence**, Vol. 39, pp. 24476–24484, 2025.

[14] Heejin Do, Yunsu Kim, and Gary Geunbae Lee. Prompt- and trait relation-aware cross-prompt essay trait scoring. In **Proceedings of Findings of the Association for Computational Linguistics, the 61st Annual Meeting of the Association for Computational Linguistics**, pp. 1538–1551, 2023.

[15] Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In **Findings of the Association for Computational Linguistics, the 2020 Conference on Empirical Methods in Natural Language Processing**, pp. 1560–1569, 2020.

[16] Fei Dong, Yue Zhang, and Jie Yang. Attention-based recurrent convolutional neural network for automatic essay scoring. In **Proceedings of the 21st Conference on Computational Natural Language Learning**, pp. 153–162, 2017.

表 2　ASAP と ASAP++データセットの詳細.

問題	小論文数	平均単語数	評価観点	得点範囲 全体	得点範囲 観点別
1	1783	350	Overall, Content, Organization, Word Choice, Sentence Fluency, Conventions	2-12	1-6
2	1800	350	Overall, Content, Organization, Word Choice, Sentence Fluency, Conventions	1-6	1-6
3	1726	150	Overall, Content, Prompt Adherence, Language, Narrativity	0-3	0-3
4	1772	150	Overall, Content, Prompt Adherence, Language, Narrativity	0-3	0-3
5	1805	150	Overall, Content, Prompt Adherence, Language, Narrativity	0-4	0-4
6	1800	150	Overall, Content, Prompt Adherence, Language, Narrativity	0-4	0-4
7	1569	250	Overall, Content, Organization, Conventions, Style	0-30	0-6
8	723	650	Overall, Content, Organization, Word Choice, Sentence Fluency, Conventions, Voice	0-60	2-12

表 3　評価観点ごとの平均 QWK．また各手法のうち，平均値が最も高い手法を**太字**で示している．なお，Cont. は Content，Org. は Organization，WC は Word Choice，SF は Sentence Fluency，Conv. は Conventions，PA は Prompt Adherence，Lang. は Language，Narr. は Narrativity を表す．

Model	Method	Overall	Cont.	Org.	WC	SF	Conv.	PA	Lang.	Narr.	Style	Voice	Avg.
Gemma 3n E2B	Zero-shot	0.305	0.266	0.296	0.239	0.313	0.200	0.254	0.362	0.339	0.112	0.067	0.250
	Few-shot (1/Score)	0.378	0.354	0.256	0.067	0.080	0.151	0.453	0.490	0.504	0.414	-0.001	0.286
	Few-shot (2/Score)	0.385	0.338	0.214	0.058	0.090	0.121	0.364	0.366	0.403	0.499	0.099	0.267
	Few-shot (3/Score)	0.333	0.233	0.086	0.050	0.024	0.105	0.388	0.344	0.421	0.431	0.033	0.223
	SFT	0.379	0.456	0.272	0.405	0.389	0.327	0.515	0.504	0.556	0.228	0.016	0.368
	Proposed	**0.719**	**0.679**	**0.629**	**0.654**	**0.614**	**0.614**	**0.705**	**0.653**	**0.695**	**0.681**	**0.529**	**0.652**
GPT-5 mini	Zero-shot	0.460	0.400	0.395	0.285	0.335	0.327	0.420	0.468	0.521	0.089	0.282	0.362
	Few-shot (1/Score)	0.419	0.448	0.358	0.289	0.306	0.311	0.530	0.564	0.574	0.188	0.248	0.385
	Few-shot (2/Score)	0.457	0.454	0.393	0.296	0.371	0.363	0.560	0.572	0.611	0.466	0.194	0.431
	Few-shot (3/Score)	0.466	0.491	0.451	0.351	0.349	0.350	0.593	0.613	0.608	0.468	0.226	0.451
	Proposed	**0.687**	**0.681**	**0.624**	**0.671**	**0.669**	**0.654**	**0.726**	**0.682**	**0.719**	**0.539**	**0.649**	**0.664**
GPT-5	Zero-shot	0.539	0.491	0.462	0.440	0.416	0.293	0.492	0.500	0.546	0.083	0.555	0.438
	Few-shot (1/Score)	0.538	0.582	0.457	0.450	0.488	0.309	0.612	0.588	0.638	0.225	0.442	0.485
	Few-shot (2/Score)	0.595	0.635	0.471	0.583	0.499	0.391	0.673	0.618	0.694	0.292	0.513	0.542
	Few-shot (3/Score)	0.620	0.657	0.541	0.544	0.548	0.400	0.687	0.633	0.703	0.364	0.547	0.568
	Proposed	**0.711**	**0.696**	**0.658**	**0.698**	**0.694**	**0.673**	**0.724**	**0.669**	**0.706**	**0.609**	**0.677**	**0.683**

A　データセットの詳細

本研究で使用した ASAP，ASAP++データセットの詳細を表 2 に示す.

B　評価観点ごとの予測精度

第 3 章の表 1 では，各小論文問題を単位として評価観点ごとの QWK を平均化した．これに対し本節では，評価観点を軸とし，各小論文問題における QWK を平均化した結果を表 3 に示す．この表から，第 3 章での議論と同様に，提案手法がすべての評価観点で従来手法を上回っていることが確認できる.