モジュラー型大規模言語モデル活用自動作文評価システム

代表研究者　Qiao WANG　早稲田大学　理工学部　助教
共同研究者　John Maurice Gayed　早稲田大学　Global Education Center 助教

1 はじめに

本研究では、自動作文評価システムを提案する。本システムは、論証型エッセイに対して、精度の高い信頼性のあるスコアと、効果的かつ包括的なフィードバックを提供する。本システムは、スコアリング、表層的フィードバック、深層的フィードバックの 3 つのモジュールに分割されたモジュラー設計を採用している。システムの構築に際しては、LLaMA-3.3-70B-Instruct、GPT-4o、Claude 3.7 などの複数の大規模言語モデル（LLM）を、直接プロンプトによるアプローチおよび教師ありファインチューニング(SFT)を試した。使用したデータは、Educational Testing Service (ETS) が提供する、ベンチマークスコア付きの TOEFL Independent Writing 480 本である。評価の結果、システムはスコアリングにおいて、0-5 スケールでのベンチマークスコアに対し、二乗加重カッパ係数（QWK）0.84、平方平均二乗誤差（RMSE）0.44 という最先端の性能を達成した。また、システム生成フィードバックは人間評価者からも高く評価され、表層的フィードバックで 96.14%、深層的マクロフィードバックで 93.03%、ミクロフィードバックで 94.69%の肯定的評価を得た。さらに、インタラクティブなユーザーインターフェースも開発されていた。

2 先行研究

2-1 初期の自動作文評価

論証型ライティングは、思考力を育成するという観点から、中等教育および高等教育のカリキュラムにおいて重視されている（Graff, 2003; Kuhn, 2005）。しかし、たくさんの文章を評価することは、人間の評価者にとって極めて困難な作業である。評価には多大な時間がかかり、たとえ訓練された評価者であっても、判断には主観的バイアスや不一致が見られることがある（Shermis and Burstein, 2003）。こうした背景から、自動作文評価（Automated Writing Evaluation: AWE）システムは、大規模に学生のエッセイをスコア化し、フィードバックを提供する有望な解決策として注目されている。AWE は、「文章をコンピュータプログラムによって評価・採点するプロセス」と定義され、スコアリングとフィードバックという二重のタスクを包含する（Shermis & Wilson, 2024, p.1）。

この分野は自然言語処理（NLP）および教育研究の中でも長い歴史を持ち、その起源は Page（1967）による先駆的研究にまで遡る。彼は初めてコンピュータによるエッセイ採点の可能性を提案し、Project Essay Grade（PEG）を開発した（Page, 2003）。初期の AWE システムは、主として採点に焦点を置いており、文法や語彙の複雑性といった特定のテキスト特徴を対象とするもの、あるいは潜在的意味解析（LSA）などの意味的類似度に基づくものが存在した。例えば、ETS の E-rater（Attali & Burstein, 2006）は、文法エラー、語彙の洗練度、構成指標など、手作業で設計された言語的特徴のスイートを利用して作文を評価する。一方、Intelligent Essay Assessor（IEA）は、LSA を活用して、エッセイと高得点回答との意味的類似性を測定する。

近年の AWE システムは、エッセイの分散表現を学習し、同等の品質を持つ文章が類似のベクトル空間にマッピングされるような、深層学習手法を採用している（Li & Ng, 2024）。例えば、Taghipour and Ng（2016）は、CNN で n-gram レベルの局所依存性を抽出し、LSTM で長距離のグローバル依存性を捉えるという二段階モデルを用いて、全体的なエッセイスコアを算出した。これらのシステムの一部は、採点において高い効果を

示した（例：E-rater：Burstein et al., 2004）が、内容理解、特に論理性や主張といった高次思考の捉え方に限界があるという共通の課題を抱えている。したがって、初期の AWE システムが提供するフィードバックが存在したとしても、それは形式的で表層的（例：文法・表記ミスの指摘）にとどまり、主張の妥当性や一貫性に関する洞察は欠如していた（Li & Ng, 2024）。

2-2 大規模言語モデル（LLM）を活用したアプローチ

近年の LLM の急速な発展により、AWE 分野においても新たな研究と応用が促進されている。LLM は高い読解力と自然言語生成能力を持ち、従来型 AWE システムのようなテンプレート的フィードバックを超える、詳細かつ人間らしい批評を生成することができる。

LLM を AWE に応用する研究は、直接プロンプトによるアプローチとファインチューニングによるアプローチの双方が存在する。例えば、Mansour et al.（2024）は、ChatGPT および LLaMA-2 を使用し、プロンプト工学によるエッセイスコアリングを評価した。その結果、両モデルは比較的良好なパフォーマンスを示し、ChatGPT がやや優れていたと報告している。

また、Stahl et al.（2024）は、Chain-of-Thought プロンプト（Wei et al., 2022）に基づくゼロショットおよび少数ショットのスコア・フィードバック同時生成を試み、スコアとフィードバックを分けて生成するよりも、同時生成の方が全体的な品質が向上すると報告した。

ファインチューニングに関する研究では、LLM をそのまま使用するよりも、やや旧式あるいは小規模なモデルをファインチューニングする方がスコアリングにおいて有効であるとされている（Li & Ng, 2024）。Wang and Gayed（2024）は、GPT-3.5 モデルを TOEFL の論証型エッセイコーパスでファインチューニングし、GPT-4 などの最新モデルと比較したところ、ファインチューニング済みモデルの方がスコアの精度と再現性の両面で優れていたと報告している。

Cai et al.（2025）は、Rank-Then-Score（RTS）という 2 段階構成のフレームワークを提案し、まずファインチューニングモデルでエッセイをランキングし、その順位を基にスコアリングを行う方式で、特に中国語データセットにおいて高い効果を示した。また、Chu et al.（2024）は、プロンプト工学による LLM とファインチューニングモデルを組み合わせた、Rationale-based Multi-Trait Scoring（RMTS）モデルを提案し、ルーブリックに即した多面的で精緻な説明を生成することで、採点の信頼性を向上させた。

2-3 ギャップ

Li and Ng（2024）は、AWE システムにおいて、全体的または特性別のスコア、文章フィードバック、改訂作文という複層的な構造が必要であると提案している。この提案は、第二言語ライティング研究におけるフィードバックの重要性に呼応しており、文法・語彙・表記などの表層的誤りに対する訂正と、構成・論理性・主張の明確さといった深層的な観点に基づくフィードバックの双方が必要とされている（Bitchener and Knoch, 2008）。一方で、学習者に対する認知的負荷の懸念も指摘されている（Truscott, 1996）が、学習者が自身にとって必要な側面に選択的に注意を払うことで、包括的フィードバックの有用性が担保されると考えられる。しかし、既存の AWE システムの多くは、表層的なフィードバック（例：文法エラー訂正システム）または深層的フィードバック（例：「主張が不明瞭で例示が不足している」など）を個別に提供するにとどまり、かつそれらのフィードバックは本文とは別のパラグラフとして提示されることが多く、利用者が自分の記述と照合しにくい（Stahl et al., 2024）。

本研究では、論証型エッセイに対応した LLM ベースの AWE システムを構築し、スコアと、表層・深層の包括的フィードバックをインタラクティブな UI を通じて提供する。本システムは、Li and Ng（2024）の提案

した上位 2 層（スコアおよびフィードバック）を基に、スコアリング、表層的フィードバック、深層的フィードバックの 3 つの独立モジュールを実装した。

3 方法

3.1 データセットとサブセット

本システムの開発には、ベンチマークスコアおよび包括的フィードバックを伴う論証型エッセイのデータセットが必要であった。そのため、本研究では、Educational Testing Service（ETS）が提供する、TOEFL 独立ライティング課題の受験者エッセイ 480 本からなる専有データセットを取得した。TOEFL の独立ライティング課題とは、たとえば「Do you agree or disagree with the following statement…」という形式の問いに対し、受験者が自らの立場を論じる典型的な論証型タスクである。

このデータセットでは、エッセイは 2 つの異なる課題に均等に分配され、2 名の ETS 評価者が 0〜5 の整数スコアを特定のルーブリックに基づいて付与している。2 人の評価者のスコアの差が 1 以下であれば、その平均値が最終スコアとなり、それ以上の乖離がある場合は第 3 の評価者が介入して最終スコアを決定した（Blanchard et al., 2013）。なお、スコア 0 のエッセイは除外されているため、本研究の対象スコアは 1〜5（0.5 刻み）である。

スコアリングモジュールでは、スコアのバランスを考慮し、2 つの課題から均等にサンプリングした 120 本のエッセイをファインチューニング用のサブセットとした。スコア 1 のエッセイが少なかったため、他のスコアのエッセイを補ってバランスを調整した。残りの 360 本はテスト用サブセットとして使用した。

フィードバックモジュールに関しては、当該データセットにはフィードバックが含まれていなかったため、独自にフィードバック用データを作成した。表層的フィードバック（文法・表記の訂正）については、先行研究により LLM への直接プロンプトが有効であることが示されているため（Davis et al., 2024）、教師あり学習および対応するデータセットの作成は不要と判断した。

一方、深層的フィードバック（構成や主張の論理性など高次的観点）については、スコア別に均等にサンプリングした 90 本のエッセイを選定し、大学でアカデミックライティングを教える経験豊富な教員 8 名を招聘して、MS Word のコメント機能を用いてフィードバックを記述してもらった（図 1 参照）。事前に、各教員にはスコア 1〜5 のベンチマーク付き模範エッセイ 9 本を提示し、ルーブリックの解釈について訓練を行った。90 本のエッセイはすでに表層訂正済みで、かつスコアを削除してあり、教員の注意が深層的特徴のみに向けられるよう配慮した。提出されたコメントは、別の独立した教員により再確認され、正確性と表層誤りの除去が担保された。

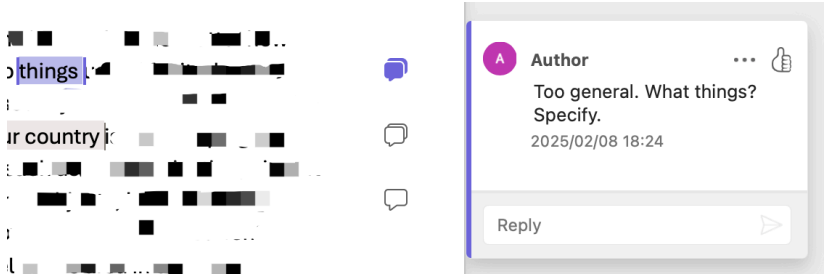

図 1. MS Word のコメント機能を用いてフィードバック

3.2 LLM の選定

スコアリングモジュールでは、汎用性を確保するため、オープンソースモデルと商用モデルの両方を評価した。オープンソースモデルとしては、命令追従タスクで最高性能を誇る LLaMA-3.3-70B-Instruct を選定した。商用モデルとしては、Wang and Gayed (2024) の示唆に従い、ファインチューニングが可能な GPT シリーズの最新モデルである GPT-4o を使用した。ファインチューニング時のプロンプトは付録 C に記載されている。

表層的フィードバックモジュールは直接プロンプトのみに依存しているため、まず GEC（文法訂正）タスクにおいて LLaMA-3.3-70B-Instruct、GPT-4o、Claude 3.7 Sonnet の 3 モデルを用いた予備実験を 1 本のエッセイで実施した。その結果、GPT-4o が最も優れたパフォーマンスを示し、次点が LLaMA、Claude は過剰訂正（例：語の意味は同じでも文体レジスタが高い同義語への置換）傾向を示したため、最終的に GPT-4o を採用した。プロンプトには、文法・表記の訂正のみを指示し、語彙や文体の変更は控えるよう明示した。

深層フィードバックモジュールでは、前述の 90 本の注釈付きデータにより、LLaMA および GPT-4o の教師ありファインチューニングを行った。LLaMA ではエントロピーベースの損失関数を用いて 8 エポックで学習した。GPT-4o に関しては商用モデルであるため、ファインチューニングの詳細パラメータは不明である。

直接プロンプトによる深層フィードバック生成では、Chu et al. (2024) の多面評価の枠組みに着想を得て、教員によるコメントに基づくテーマ別分析を実施し、複数のフィードバック観点（マクロ／ミクロ）を抽出した（表 1 参照）。この 2 分類に従い、深層フィードバックをマクロパイプラインとミクロパイプラインの 2 系統に分けた。

表 1：深層レベル分析のためのフィードバック分類

カテゴリー	サブカテゴリー
マクロフィードバック	適切な段落構成；テキストの長さの妥当性；構成（導入部・本文・結論の各セクション）
ミクロフィードバック	文脈依存の文法的問題；表現の明確さ；アイデア間の一貫性；論証の質と論理的な流れ；形式性とアカデミック・レジスター（学術的文体）

モデル選定のため、Claude 3.7、GPT-4o、DeepSeek-R1、LLaMA-3.3-70B-Instruct の 4 モデルを、1 本のテスト用エッセイで予備比較した結果、Claude 3.7 が最も満足度の高い出力を示したため、本パイプラインの主モデルとして採用した。その後、各サブカテゴリに焦点を当てた長めの few-shot プロンプトを設計し、Claude 3.7 がそれに基づいて詳細なフィードバックおよび修正文案を生成するよう調整した。

3.3 検証方法

3.3.1 スコアの検証

スコアリングモジュールでは、ファインチューニング済みの LLaMA-3.3-70B-Instruct および GPT-4o モデルを使用し、360 本のテスト用エッセイを評価した。ファインチューニング時と同じプロンプトを用いて推論を実施した。

スコアの精度と信頼性の評価には、以下の 3 つの指標を用いた：

- 平方平均二乗誤差(RMSE):システムスコアとベンチマークスコアの絶対的な差異を測る精度指標。値が小さいほど性能が高いとされる（Chai & Draxler, 2014）。
- 二乗加重カッパ係数（QWK）：スコアの順序性を考慮した一致度指標。人間評価との一致度を定量的に測定する（Li & Ng, 2024）。
- 一致率（Percentage Agreement）：QWK を補完する指標であり、完全一致（スコアが同一）および隣接一致（スコア差が 0.5）の比率をそれぞれ算出した。

3.3.2 フィードバックの検証

表層的フィードバックについては、原文と訂正済みエッセイを ERRANT v3.0.0（Bryant et al., 2017）で整合させ、全ての編集操作を自動でタグ付け・分類した。2 名の専門評価者が以下の 2 軸で各編集操作を評価した：

- 必要性（Necessity）：修正対象が実際に誤りとして訂正が必要であったか（LLM の過剰訂正傾向を考慮）。
- 有効性（Effectiveness）：提案された修正が実際にその誤りを適切に修正しているか。

2 名の評価者は全編集について協議し、不要または無効と判断されたケースにはコメントを添えた。

深層的フィードバックについては、マクロパイプラインでは、各エッセイの段落ごとに Claude 3.7 がフィードバックを生成するようプロンプトで指示し、各段落コメントが有効であったかどうかを評価者が判断した。ミクロパイプラインでは、表層的フィードバックと同様に「必要性」と「有効性」の 2 軸で評価した。2 名の評価者による評価結果をもとに、Gwet の AC1 係数（Gwet, 2008）を用いて信頼性を算出し、意見が分かれた場合は第 3 の評価者が最終判断を下した。

4 結果

4.1 スコアリングモジュールの結果

2 つのファインチューニング済みモデル（GPT-4o、LLaMA）と、Wang and Gayed（2024）によるベースラインモデルのスコア評価結果を表 2 に示す。使用された指標は RMSE、QWK、一致率（完全一致・隣接一致・合計）の３つである。結果から、我々のモデルはいずれも SOTA（最先端）のスコアリング性能を示し、特に GPT-4o が全指標で最良の結果を示した。両モデルの QWK スコアは 0.8 を上回り、「ほぼ完全一致」の閾値（Sim & Wright, 2005）を超えた。なお、ETS の E-rater が TOEFL で採用するための QWK 基準は 0.7 とされており（Williamson et al., 2012）、本モデルはそれを大きく上回る。RMSE に関しては、GPT-4o で 0.44、LLaMA で 0.53 と良好であり、ETS が人間評価者間に許容している 1 点の差（最終スコアの平均が 0.5 ずれる可能性）を下回る結果であった。

表 2：ファインチューニング済みモデルのテストサブセットにおける性能指標

モデル	RMSE	QWK	一致率（絶対）	一致率（隣接）	一致率（合計）
GPT-4o	0.44	0.84	0.45	0.48	0.93
LLaMA	0.53	0.81	0.41	0.45	0.86

ベースライン	0.57	0.78	0.33	0.52	0.85
(Wang & Gayed, 2024)					

4.2 表層的フィードバックモジュール

テストサブセットに含まれる 40 本のエッセイに対して表層的訂正を行った結果、ERRANT は合計 2049 件の編集操作（edit operations）をタグ付けした。人間による評価の結果、これらのうち 1985 件（96.88%）が「修正が必要である」と判定され、さらにそのうち 1970 件（96.14%）が「有効である」と評価された。不要と判定された編集には、以下の 2 つの傾向が確認された：

1. 英米英語の違いに起因する訂正（N=7）
   例として、「favorite」が「favourite」に変更されたり、「...store.”」が「...store”.」に変更されたりするなど、GPT-4o がイギリス英語の表記を好む傾向が見られた。
2. カンマの追加（N=15）
   特に、名詞のリストにおいて最後の項目の前にオックスフォードカンマを追加するケース（N=10、例：「A, B and C」→「A, B, and C」）があったが、評価者からは不要と判断された。

4.3 深層的フィードバックモジュール

4.3.1 マクロフィードバックパイプライン

テストサブセットの 40 本のエッセイには、合計 201 段落が含まれており、それに対応して Claude 3.7 は 201 件のマクロコメントを生成した。2 名の評価者による有効性判定のクロス表は以下の通りである：

表 3. 2 名の評価者によるマクロコメント評価のクロス表

評価者 A／評価者 B	有効	無効
有効	179	9
無効	13	0

このデータに基づき、Gwet の AC1 係数（Gwet, 2008）を算出したところ、以下のような結果が得られた：AC1 = 0.89（SE = 0.03、95%信頼区間［0.82, 0.93］、p < .001）。これは統計的に有意かつ非常に高い一致度を示しており、「ほぼ完全一致」とみなせる（Landis & Koch, 1977 の基準による）。両評価者の不一致に対して第 3 評価者が介入し、最終的に 187 件（93.03%）が有効、27 件（6.97%）が無効と確定した。無効と判断されたコメントには、Claude 3.7 が予期しない入力に対応できなかった例が多かった。具体例としては：

- 受験者がエッセイに不要なタイトルをつけており、Claude 3.7 がそれを第 1 段落と誤認（N=1）
- 各文を段落として誤って改行しているケースにおいて、毎段落に「段落が短すぎる」とコメント（N=8）
- エッセイが時間内に完了していないケースに対して、結論の書き方を長文で指導（N=2）
- 引用資料の使用が禁止されているテストにもかかわらず、「出典を使え」という誤コメント（N=1）

4.3.2 ミクロフィードバックパイプライン

同じ 40 本のエッセイに対して、Claude 3.7 は合計 630 件のミクロフィードバックコメントを生成した。必要性の判断における評価者間一致のクロス表は以下の通り：

表 4.2 名の評価者によるミクロフィードバックの必要性に関するクロス表

評価者 A／評価者 B	必要	不要
必要	579	28
不要	22	1

有効性の判断（必要だと両者が合意したコメントのみ）のクロス表は以下の通り：

表 5.2 名の評価者によるミクロフィードバックの有効性判断に関するクロス表

評価者 A／評価者 B	有効	無効
有効	577	1
無効	1	0

いずれの次元においても、Gwet の AC1 係数に基づく評価は以下のような結果であった：

- 必要性：AC1 = 0.91（SE = 0.01、95% CI [0.89, 0.94]、p < .001）
- 有効性：AC1 = 1.00（SE = 0.00、95% CI [0.99, 1.00]、p < .001）

第 3 評価者による最終調整の結果、以下のように確定した：

- 必要とされたコメント：600 件（95.24%）
- 不要とされたコメント：30 件（4.76%）
- 必要かつ有効なコメント：596 件（94.69%）

評価者のコメントによれば、いくつかのミクロフィードバックはマクロフィードバックと重複しており、Claude 3.7 が段落全体について繰り返し言及する傾向もあった（N=18）。たとえば、導入段落の冒頭文が不十分な場合、Claude は「背景説明が不十分」というコメントに加え、「良い導入段落の書き方」全体についてコメントすることがあった。

4.4 最終システム構成と UI 設計

これらの評価結果に基づき、最終的なシステム構成を決定した：

- スコアリングモジュール：ファインチューニング済み GPT-4o
- 表層フィードバックモジュール：GPT-4o（直接プロンプト）
- 深層フィードバックモジュール：Claude 3.7（マクロ／ミクロ両パイプライン）

図 2 に本システム全体のワークフローを示す。

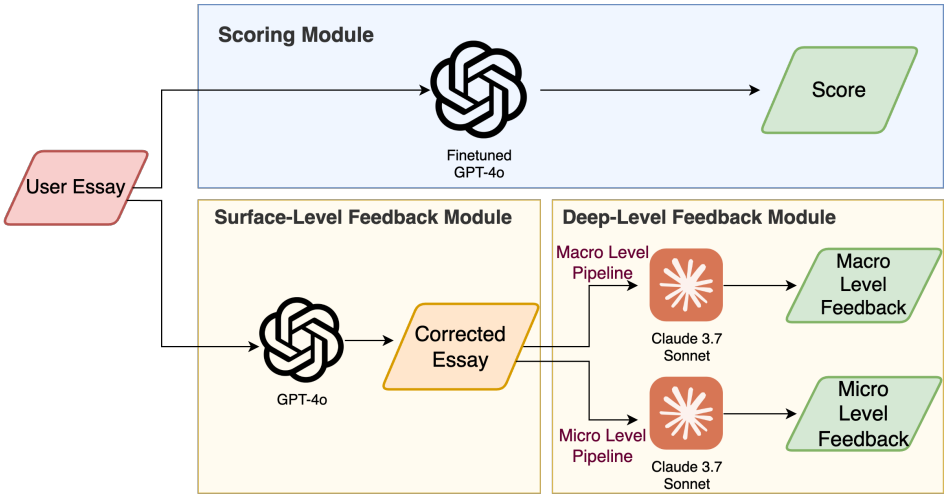

図 2. システム全体のワークフロー

また、人間のフィードバックの仕組みを模倣するため、次のようなインタラクティブ UI を開発した：

- 入力ページ：エッセイ課題と本文を入力（Word アップロードまたはテキスト貼り付け）
- 出力ページ：スコア、表層的・深層的フィードバックの表示タブ

- 表層的フィードバック表示：①元の文の誤り部分のハイライト、②訂正文との比較、③訂正文の提示（付録 1 の図 3）
- 深層的フィードバック表示：マクロコメントは段落ごとに左側パネルに配置、ミクロフィードバックは該当文に色分けでハイライトし、マウスホバーで右パネルにコメント表示（付録 1 の図 4）

5 議論

本研究で構築したモジュール型のシステムは、ETS 基準に準拠した正確かつ信頼性の高いスコアと、人間教員によって検証された高品質なフィードバックを提供することに成功した。この成果は、Stahl et al.（2024）や Chu et al.（2024）の研究結果を支持するものであり、ライティング評価の各要素に特化したモデルを用いるアプローチが、単一モデルによる一括処理よりも優れた評価をもたらすことを示している。

表層的フィードバックについての人間評価からは、文法誤り訂正（GEC）タスクは、複雑なファインチューニングを行わずとも、設計の工夫されたプロンプトだけで十分に対応可能であることが示された。この結果は、Zeng et al.（2024）の研究とも一致しており、適切なプロンプト設計によって、LLM が特化モデルと同等の GEC 性能を発揮できることが分かる。特に、本研究のプロンプト設計により、従来 LLM ベースの GEC において課題とされていた過剰訂正を最小限に抑えることができた（Katinskaia & Yangarber, 2024；Davis et al., 2024）。実際、修正箇所のうち 96.14%が「必要かつ有効」と評価されており、システムが誤りを検出した場合、それはほぼ確実に妥当なものであることを示している。深層的フィードバックの人間検証により、構造化された few-shot プロンプトによって、高次的思考に関する効果的なフィードバックが生成可能であることが示唆された。さらに、教師ありファインチューニングよりも有効な場合がある可能性も見出された。

ここで特筆すべきは、深層フィードバックに対する SFT の限界である。GPT-4o のファインチューニングモデルは、ほぼすべての語・句・文にフィードバックを生成し、その多くが不要であった。これはおそらく、出力が非決定的（non-deterministic）であるタスクに対し、ファインチューニングが適していないことに起因していると考えられる。数値的・カテゴリ的出力とは異なり、「どのコメントが必要・効果的か」についての明確な基準は存在しないため、今後は人間の好みに基づく強化学習（RLHF）を導入する必要があるだろう。

一方、LLaMA モデルは出力フォーマットに不備があったものの、生成されたフィードバックの内容自体は、ファインチューニング用データとよく一致しており、Claude 3.7 には及ばないまでも、一定の改善が確認された。Yao et al.（2025）の研究でも、LLaMA における出力フォーマットエラーが指摘されており、将来的にはフォーマットチェッカー等の補助ツールを併用することで、オープンソースモデルの実用性がさらに高まると考えられる。

本システムのインタラクティブ UI には、教育現場における実践的意義が大きい。特に、論証型ライティングに取り組む学生にとって、従来の授業では時間のかかるフィードバックを即時に得られることは大きな利点である。また、教員にとっても、表層的な訂正作業をシステムに任せることで、構成や論理性といった高次思考の指導に集中できるという利点がある。

さらに、大学の大規模授業などで見られる教員ごとの採点基準のばらつきを緩和し、公平性を担保する採点リファレンスツールとしても有効である。特に、複数教員が関与する協調授業などでは、基準の一貫性を保つ手段として活用できる可能性がある。

6 本研究の限界

本研究にはいくつかの限界が存在する。第一に、本システムの開発には、過去にわずか一件の研究でしか使用されていない専有の TOEFL エッセイデータセットを用いた。このため、本研究で報告されたスコアリング性能を他の研究と直接比較することは困難であり、結果の一般化可能性には一定の制約がある。

第二に、本研究におけるフィードバックの妥当性検証は、精度（precision）に焦点を当てており、再現率（recall）を考慮していない。すなわち、システムが誤って修正した箇所や不適切なコメントを検出する評価は行っているものの、本来修正すべきであった誤りを見逃していなかったか、または必要であったはずのフィードバックを生成していなかったかについては検証していない。より包括的な評価には、すべての誤りとフィードバック機会を網羅した「ゴールドスタンダード」を確立し、それとの照合によってシステムの網羅性を評価する必要がある。しかしながら、そのために必要となる人的資源と時間的コストは本研究の範囲を超えていた。

第三に、本研究ではシステムの有効性を人間評価者によって検証したが、実際に学習者がこのシステムを使用した場合に、ライティング能力がどの程度向上するかといった教育的効果の測定は行っていない。したがって、今後の研究では、従来のフィードバック方式と比較しつつ、システムを用いた指導が学習者のライティング能力に及ぼす中長期的な影響を測定する実験的研究が求められる。

加えて、システム内部では、限定的ながら LLM による「幻覚（hallucination）」が観察された。これは、実際には存在しない誤りを検出したかのように表示し、その訂正案が元の文と同一であるようなケースである。こうした誤検出を防ぐためには、今後のシステム設計において、フィードバックを検証する別の LLM など、二重化された品質管理機構の導入が検討されるべきである。

本システムは論証型ライティング、かつ課題文に基づく作文に特化して開発されたものであり、情報源を引用・参照することを求められるソースベース型ライティングや、他ジャンルの文章にはそのまま適用できない。したがって、将来的には対応するライティングジャンルを拡張し、タスクごとの汎化性を検証する必要がある。

また、現時点で提供されるフィードバックはすべて英語で提示されるため、英語初級者にとっては理解と活用が困難となる可能性がある。多言語対応機能、とりわけ学習者の母語への翻訳表示などを組み込むことで、フィードバックの可読性と実用性を高める余地がある。

最後に、システムは GPT-4o や Claude 3.7 といった商用 LLM API に依存して構築されている。このため、利用にあたってはコストやレート制限、長期的な持続可能性といったスケーラビリティの課題が存在する。加えて、商用モデルの仕様変更がシステムの性能に予期せぬ影響を与える可能性も否定できない。また、開発および運用時には匿名化されたテキストのみを送信しているものの、教育現場において学生の作文データを第三者 API に通す場合にはデータプライバシーとガバナンスの観点から慎重な検討が求められる。

7 結論

本研究では、論証型エッセイに特化したモジュラー型の大規模言語モデルベース自動作文評価システムを開発した。スコアリング、表層的フィードバック、深層的フィードバックという 3 つのモジュールにタスクを分離することで、TOEFL ライティングの専有データセット上において最先端のスコアリング性能を達成した。

さらに、システムによって生成されたフィードバックは、人間教員の評価により高い有効性と実用性が確認された。本研究の成果は、スコアと包括的フィードバックの両方を提供するモジュール型 AWE システムの設計に関する将来の研究に対して、有効な基盤を提供するものである。

本システムは、誰でも無料で使用可能なインタラクティブ UI としても公開されており、教育現場への即時的な導入が可能である。

【参考文献】

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v.2. Journal of Technology, Learning, and Assessment, 4(3).

John Bitchener and Ute Knoch. 2008. The value of written corrective feedback for migrant and international students. Language Teaching Research, 12(3):409431.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. ETS Research Report Series, 2013:i–15.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2004. Automated essay evaluation: The criterion online writing service. Ai magazine, 25(3):27–27.

Yida Cai, Kun Liang, Sanwoo Lee, Qinghan Wang, and Yunfang Wu. 2025. Rank-then-score: Enhancing large language models for automated essay scoring. arXiv preprint arXiv:2504.05736.

T. Chai and R. R. Draxler. 2014. Root mean square error (rmse) or mean absolute error (mae)? – arguments against avoiding rmse in the literature. Geoscientific Model Development, 7(3):1247–1250.

SeongYeub Chu, JongWoo Kim, Bryan Wong, and MunYong Yi. 2024. Rationale behind essay scores: Enhancing s-llm's multi-trait essay scoring with rationale generated by llms. arXiv preprint arXiv:2410.14202.

Christopher Davis, Andrew Caines, Øistein E. Andersen, Shiva Taslimipoor, Helen Yannakoudakis, Zheng Yuan, Christopher Bryant, Marek Rei, and Paula Buttery. 2024. Prompting open-source and commercial language models for grammatical error correction of English learner text. In Findings of the Association for Computational Linguistics: ACL 2024, pages 11952–11967, Bangkok, Thailand. Association for Computational Linguistics.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attentionbased recurrent convolutional neural network for automatic essay scoring. In Proceedings of the 21st Conference on Computational Natural

Language Learning (CoNLL 2017), pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

Gerald Graff. 2003. Clueless in Academe: How Schooling Obscures the Life of the Mind. Yale University Press.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2019. Minimally-augmented grammatical error correction. In Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), pages 357–363, Hong Kong, China. Association for Computational Linguistics.

Kilem L Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology, 61(1):29–48.

Kenji Imamura, Kuniko Saito, Kugatsu Sadamitsu, and Hitoshi Nishikawa. 2012. Grammar error correction using pseudo-error sentences and domain adaptation. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 388–392, Jeju Island, Korea. Association for Computational Linguistics.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. TDNN: A two-stage deep neural network for promptindependent automated essay scoring. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1088–1097, Melbourne, Australia. Association for Computational Linguistics.

Anisia Katinskaia and Roman Yangarber. 2024. GPT3.5 for grammatical error correction. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 7831–7843, Torino, Italia. ELRA and ICCL.

Deanna Kuhn. 2005. Education for Thinking. Harvard University Press.

Thomas K. Landauer, Darrell Laham, and Peter W. Foltz. 2003. Automated scoring and annotation of essays with the intelligent essay assessor. In Automated Essay Scoring: A Cross-disciplinary Perspective, pages 87–112. Lawrence Erlbaum Associates.

J. Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. Biometrics, 33(1):159–174.

Jianwei Li and Jiahui Wu. 2023. Automated essay scoring incorporating multi-level semantic features. In International Conference on Artificial Intelligence in Education, pages 206–211. Springer.

Shengjie Li and Vincent Ng. 2024. Automated essay scoring: A reflection on the state of the art. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1787617888, Miami, Florida, USA. Association for Computational Linguistics.

Sha Liu and Antony Kunnan. 2015. Investigating the application of automated writing evaluation to chinese undergraduate english majors: A case study of writetolearn. Assessing Writing, 24:1–15.

Watheq Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can large language models automatically score proficiency of written essays? arXiv preprint arXiv:2403.06149.

Ellis B Page. 1967. Grading essays by computer: Progress report. In Proceedings of the invitational Conference on Testing Problems.

Ellis Batten Page. 2003. Project Essay Grade: PEG. Lawrence Erlbaum Associates Publishers.

Jim Ranalli and Junko Yamashita. 2021. L2 student engagement with automated feedback on writing. System, 99:102512.

Alla Rozovskaya and Dan Roth. 2019. Grammar error correction in morphologically rich languages: The case of Russian. Transactions of the Association for Computational Linguistics, 7:1–17.

Mark D. Shermis and Jill C. Burstein. 2003. Automated Essay Scoring: A Cross-disciplinary Perspective. Lawrence Erlbaum Associates.

Mark D Shermis and Joshua Wilson. 2024. Introduction to automated essay evaluation. In The Routledge international handbook of automated essay evaluation, pages 3–22. Routledge.

Julius Sim and Chris C Wright. 2005. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. Physical Therapy, 85(3):257268.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 234–245.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

John Truscott. 1996. The case against grammar correction in l2 writing classes. Language Learning, 46(2):327–369.

Ivo Verhoeven, Pushkar Mishra, Rahel Beloch, Helen Yannakoudakis, and Ekaterina Shutova. 2024. A (more) realistic evaluation setup for generalisation of community models on malicious content detection. In Findings of the Association for Computational Linguistics: NAACL 2024, pages 437–463, Mexico City, Mexico. Association for Computational Linguistics.

Q. Wang and M. Gayed. 2024. Effectiveness of large language models in automated evaluation of argumentative essays. Computer Assisted Language Learning, 37(1):1–25.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:2482424837.

David M Williamson, Xiaoming Xi, and F Jay Breyer. 2012. A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practice, 31(1):2–13.

Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. 2013. A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Medical Research Methodology, 13(1):61.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Jiashu Yao, Heyan Huang, Zeming Liu, Haoyu Wen, Wei Su, Boao Qian, and Yuhang Guo. 2025. Reff: Reinforcing format faithfulness in language models across varied tasks. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 25660–25668.

Min Zeng, Jiexin Kuang, Mengyang Qiu, Jayoung Song, and Jungyeul Park. 2024. Evaluating prompting strategies for grammatical error correction based on language proficiency. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 6426–6430, Torino, Italia. ELRA and ICCL.

付録 1

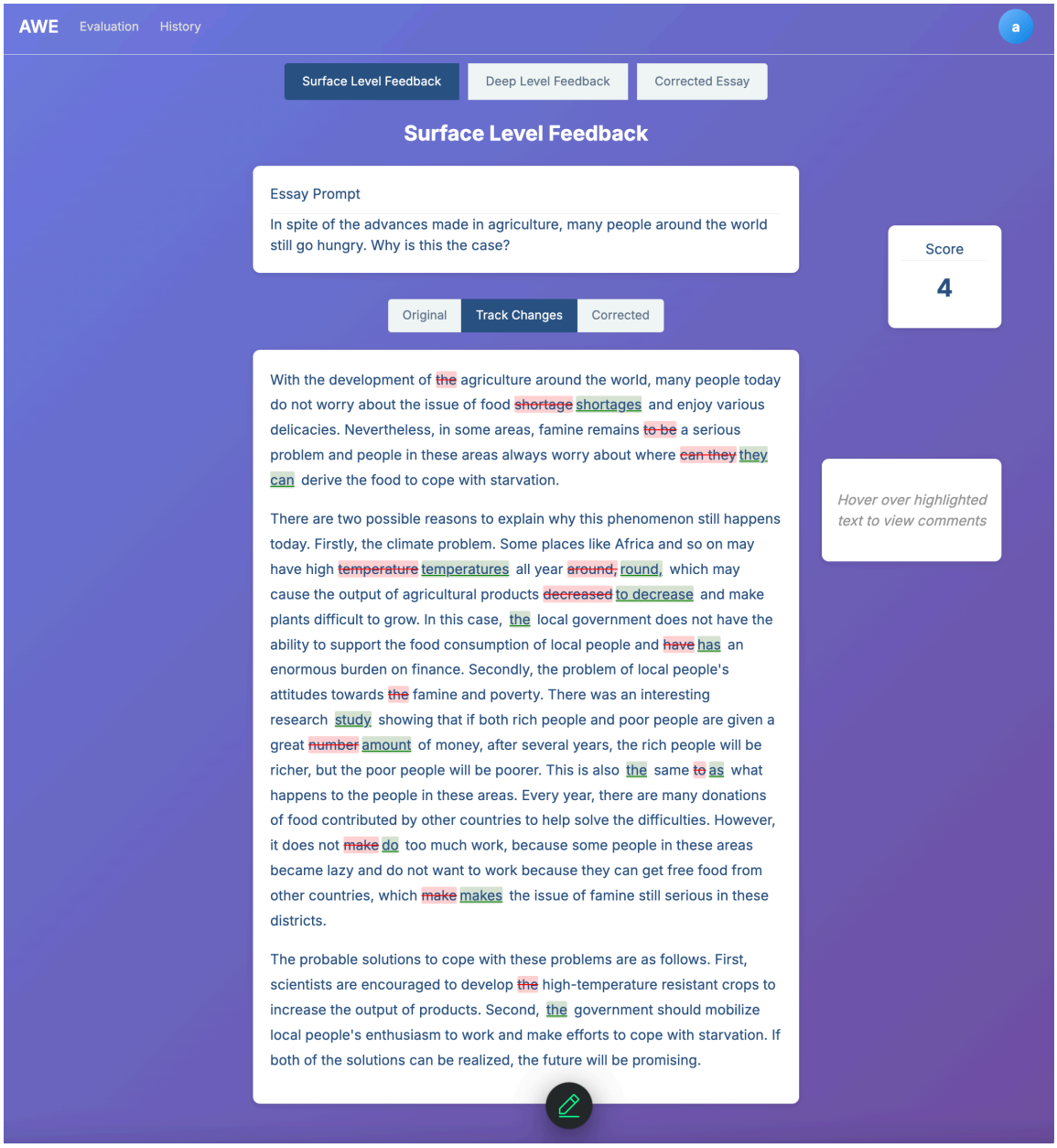

図 3.　スクリーンショット：表層的フィードバックページ（変更履歴表示）

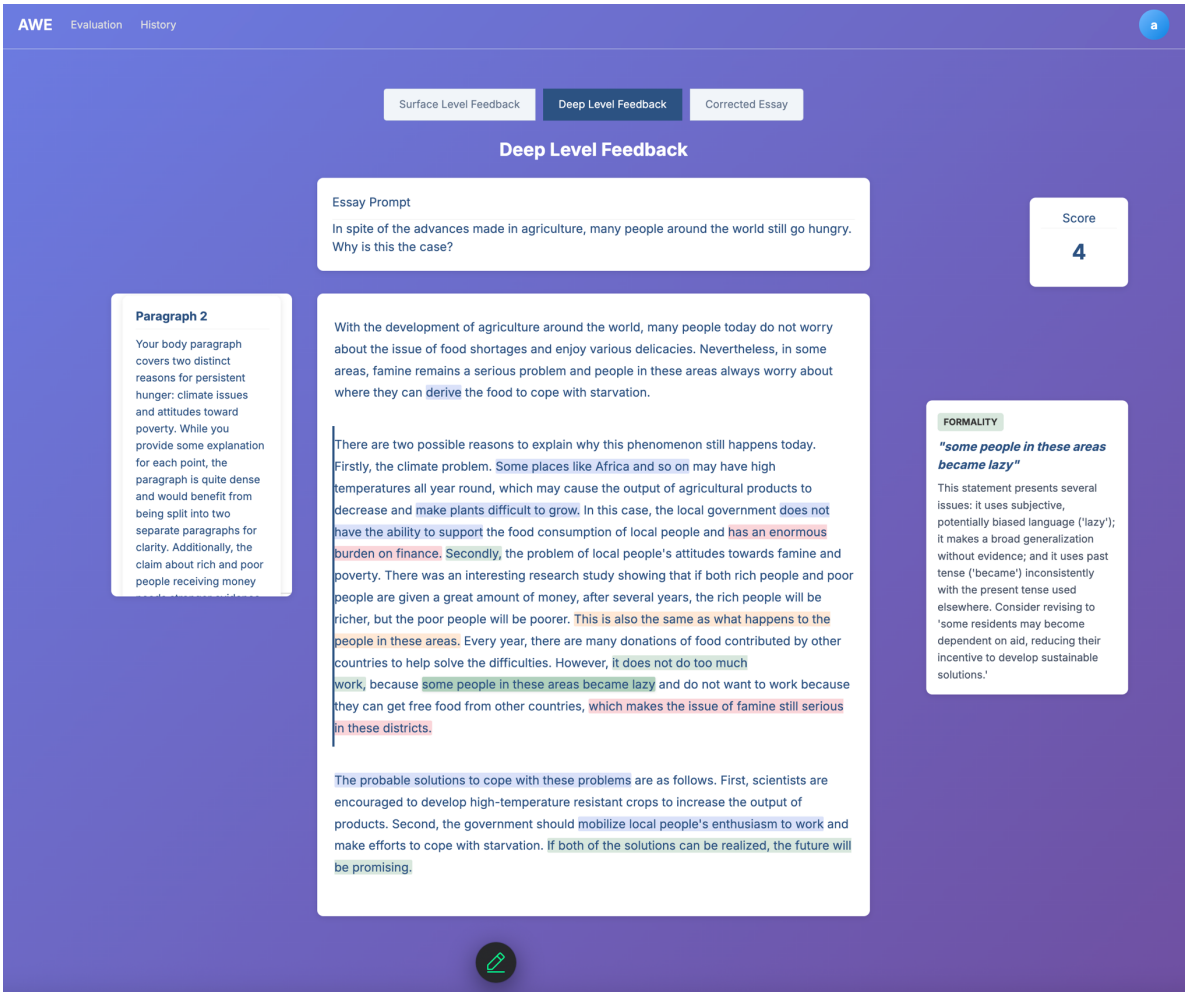

図 4. スクリーンショット：深層的フィードバックページ

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
WrAFT ウェブサイト	awe.judywang.jp	2025 年 5 月
WrAFT: A Modular Large Language Model-Powered Automated Writing Evaluation System for Argumentative Essays	Emperical Methods in Natural Language Processing (査読中)	2025 年 5 月
Effectiveness of large language models in automated evaluation of argumentative essays	Computer Assisted Language Learning	2024 年 5 月