K-014

適応学習における生徒作問課題の生成 AI による自動評価

Automatic Evaluation of Student-Created Questions in Adaptive Learning Using Generative AI

加藤 空‡　　小林 学‡
Sora Kato　Manabu Kobayashi

1. はじめに

近年，オンライン学習環境の急速な発展に伴い，学習者一人ひとりの理解度に応じて学習内容を動的に調整する適応学習（adaptive learning）の重要性が高まっている．特に ICT の活用が進む現代の教育実践においては，学習者の反応や成果をリアルタイムに評価し，その結果に基づいて最適な学習支援を提供することが求められている．

このような背景のもと，生成 AI（generative AI）技術の進展は，学習評価の自動化に新たな可能性をもたらしている．たとえば，記述式問題に対する答案の自動採点や，個々の解答に応じた即時かつ個別化されたフィードバックの生成が可能となりつつある．これにより，これまで人手に大きく依存していた評価プロセスの効率化や質の向上が期待されている．

しかしながら，日本の教育現場においては，児童・生徒・学生の理解度を的確に把握し，それに基づいて個別に評価を行うことは，教員にとって非常に大きな負担となっている．特に，記述式の学習評価は採点やフィードバック作成に多大な時間と労力を要するため，実際の教育実践において十分に活用されていないのが現状である．このような課題を踏まえると，生成AIを活用した学習評価の支援技術は，教育現場の負担軽減と質的改善の両面において，極めて有用なアプローチとなり得る[1].

本研究では，生徒に対して学習内容に関連する問題作成（作問）課題を提示し，その生徒によって作成された問題の内容を生成AIが分析することによって，生徒自身の学習内容に対する理解度を評価する手法の可能性を検討する．従来の評価方法は，主に学習者が与えられた問題に対して正答を導き出す能力に基づいて行われることが一般的であるが，本研究ではその枠を超え，学習者自身が問題を構成する能力を評価の指標とする点において新規性を有する．

本研究の目的は，生成AIを活用することにより，生徒が解答する側に回る従来型の評価手法では捉えきれない，より高次の認知的能力や概念的理解の程度を可視化・定量化する評価モデルを構築する点にある．最終的には本手法により，生徒が単に知識を再現する能力だけでなく，学習内容を再構成し，他者に伝達可能な形式で表現する力，すなわち深い学びに至っているかどうかを自動的に評価，フィードバックできる可能性を示すことを目的とする．

2. 方法

本研究では生徒に対して学習内容に関連する問題作成（作問）課題を提示し，その生徒によって作成された問題を生成AIが受け取る．次に生徒が作成した問題文を評価対象とし，その内容を生成AIにより自動分析させる手法を提案する．最終的に，得られた評価は教師および生徒にフィードバックを行う．このシステムの概要を図 1 に示す．この際，評価基準としては日本の学習指導要領における学習到達目標に加え，教育評価理論として広く用いられているブルームの改訂版タキソノミー（Revised Bloom’s Taxonomy）を導入する．これにより，作成された問題が，記憶・理解・応用・分析・評価・創造といった認知プロセスのいずれの水準に該当するかを客観的に判定し，生徒の理解の深さを多角的に捉えることを目指す．

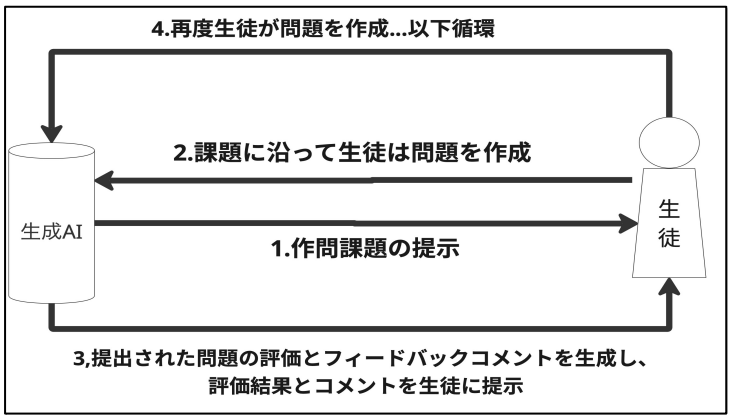

図 1　生成 AI による生徒の理解度評価システム

本研究では，生徒によって作成された問題文を自動的に評価する手法として，生成 AI を活用した分析モデルを構築・検証した．生成 AI には OpenAI が提供する ChatGPT（モデル：GPT-o4-mini）を採用し，同社の提供する API を通じてアクセス・利用を行った．API 実行環境としては Google Colaboratory を使用し，Python を用いて実装した．

評価基準としては，日本の学習指導要領（平成31年度版）[2]，[3]を参照し，10 の観点をそれぞれ 1-10 のスコアとして評価した．主たる評価範囲は中学校 3 年生の数学における二次方程式とした．しかし，プロンプトのトークン制限を考慮すると学習指導要領の原文をそのままプロンプトに含めることは困難であった．そこで，Google 社の大規模言語モデルアシスタントである NotebookLM を用いて，学習指導要領の各記述を要約・抽出し，評価に必要な観点情報を整理したうえで，生成AIへの入力形式を最適化した．また，学習内容の理解度を多層的に評価するため，ブルームの改訂版タキソノミー（Revised Bloom’s Taxonomy）を評価基準の一部として導入した．具体的には，生徒が作成した問題文が，認知プロセスのどの段階（記憶・理解・応用・分析・評価・創造）に該当するかをAIに判定させることで，理解の深さを分類的に評価した．スコアは記憶が 10，理解が 20，応用が 30，分析が 40，評価が 50，創造が 60 とした．最後に問題文の破綻や誤りがある場合最大で 160 点

‡早稲田大学　Waseda University

の減点とした．本研究における AI モデルの実行および検証に使用した物理的環境は，Apple 社製 M4 チップ搭載 MacBook Air（macOS 15.3.2）であり，ブラウザとしては Safari を使用した．これらの環境において，Google Colaboratory 上での AI モデル実行と NotebookLM による資料要約・整理を適切に行うことが可能であった．

3. 実験と結果

ここでは実験を行うにあたり，本来生徒に課題として提示し，作成を行わせる問題文をあらかじめ用意する．本研究では生成 AI に評価を行わせる対象の問題文として，1) 著者自身が作成した問題，および 2) 生成 AI によって作成された問題の 2 種類を準備した．各問題は，大問および小問の構成を有し，単一問または 10 問から成る形式とした．また一部の問題については，意図的に誤りを含むよう設計した．

プロンプト（作成した問題含む．文字数の関係上プロンプトと結果の一部抜粋とする）

{“～中略～学習指導要領を参考にしつつ、二次方程式の問題を 1 問作成して下さい．"}
作成した問題　{ある高さ２０ｍの台の上から，小球を鉛直上方に初速２０ｍ／ｓで投げ上げた．地上からの高さ h（ｍ）は，時間 t（秒）後に $h(t)=-5t^2+20t+20$ で与えられる．問１ $h(t)=0$ となる時刻 t を求め，それぞれが何を意味するか説明せよ．～以下略～}

生成 AI の評価（文字数の関係上結果の一部抜粋とする）

～中略～
8. 問題解決能力： 9
（複数の解法手段を使い分けて答えを導く）
9. 数学的な推論： 9
（平方完成や不等式の利用など推論過程が明確）
10. 表現力： 8
（解答上，式と文章の融合を問うが，さらに簡潔化の余地あり）
２．ブルームの改訂版タキソノミーによる評価
レベル：応用（Application） → 30 点
〈指摘とアドバイス〉
– 頂点を求める平方完成と微分的（変化率の変化）な見方の両面から扱うよう問いを追加。
〈生徒の理解不足と学習ポイント〉
- 二次関数のグラフと現象の対応付け（特に負の解の扱い）

評価の結果，問題構成に論理的破綻が見られないものは高得点を得る傾向があり，逆に構成上の不整合や欠陥が認められた問題は低得点となった．いずれの場合においても，生成 AI による客観的なフィードバックが付与されたことを確認している．さらに AI による分析において，誤りを含む問題が誤って正当と評価されるケースは一度も観察されなかった．加えて，構成に破綻のある問題に対しては，暗記偏重の出題内容や記述上の不備（例：表記ミス）など，誤りの要因に関する推論も行われていた．なお AI によるスコアリングには 10～20 点程度のばらつきが確認されており，例えばスコア 160 点を目指してプロンプトを設計した問題文であっても，実際の評価結果が 160 点ちょうどとなることはなかった．さらに問題数を 10 問とした場合には，ブルームの改訂版タキソノミーに基づき，「理解」から「創造」までの認知レベルにおいて偏りが生じないよう，生成 AI から具体的なアドバイスが提示された．

次に生成 AI が作成した問題と著者が作成した問題との比較した結果について述べると，生成 AI による評価において両者に顕著な差異は見られなかった．すなわち，生成 AI による問題作成は，一定の条件下において人間と同等の評価結果が得られることが示唆された．次に複数問題を一気に提示する場合において問題数を増加させた場合，生成 AI は問題群を大問構成として適切に捉え，それに応じた助言を提供した．これにより，AI は単なる作成した問題の採点機能にとどまらず，問題構成の分析や出題意図の解釈に基づく批判的かつ的確なコメントを付与する能力を有することが確認された．

生成 AI による作問に対する評価およびフィードバックコメントの妥当性について，いくつかの実験を通して著者で内容を個別に確認したところ，かなり複雑な評価ではあるが十分な妥当性を有しているものと判断した．より詳細な結果については[4]を参照されたい．

4. まとめと今後の課題

本研究では，生徒に作問課題を提示し，その結果として得られた問題を生成 AI によって分析・評価する手法を検討した．その結果，生成 AI は生徒の作成した問題を的確に採点し，具体的かつ建設的なフィードバックを提供できることが示された．これにより，本手法は学習者自身の理解度を客観的に可視化する自己評価ツールとしての有用性を有することが明らかになった．加えて，生成 AI を活用した出題支援は，教員の問題作成業務の効率化や，学習者一人ひとりに応じた個別最適化学習の推進にも寄与し得ることが示唆された．

本研究にはいくつかの限界が存在する．第一に，対象とした学習内容が中学校 3 年生の数学「二次方程式」に限定されており，他教科・他学年における一般化可能性には慎重な検討が求められる．第二に，問題作成を学習者自身が担う形式であったため，個々の作問スキルや認知負荷の違いが結果に影響を与える可能性がある．第三に，用いた生成 AI のモデルは特定のバージョンに依存しており，他のモデルとの間で採点精度や助言の質に差異が生じる可能性がある．また評価対象となった問題数が限定的であるため，統計的な客観性には一定の制約がある．これらは今後の課題である．以上の成果は，教育現場における AI 活用の新たな可能性を拓くものであり，今後は他教科・他学年への適用検証や，実運用に向けたインターフェース設計・運用ガイドラインの整備が求められる．

謝辞

本研究は JSPS 科研費 JP24H00370，JP23K04293 の助成を受けたものです．

参考文献

[1] Mi, Zejia, Li, Kangkang, “A Comparative Analysis of Different Large Language Models in Evaluating Student-Generated Questions”, 2024 13th International Conference on Educational and Information Technology (ICEIT), IEEE, pp. 24–29, 2024
[2] 文部科学省，中学校学習指導要領（平成 29 年告示），2017
[3] 文部科学省，中学校学習指導要領（平成 29 年告示）解説：数学編，2017
[4] https://github.com/hakubishin014/Automatic-Evaluation-of-Student-Created-Questions