高等学校情報科のテスト問題を対象とした生成AIによる問題レベル自動判定の試み

Automating Rubric-based Level Classification of Japanese High School Informatics Test Items Using Generative AI

松島 拓路*1, *2　久保田 真一郎*1
Takumi MATSUSHIMA*1, *2　Shin-Ichiro KUBOTA*1

熊本大学*1　崇城大学*2
Kumamoto University*1　Sojo University*2

＜あらまし＞　高等学校情報科のテスト問題の分野や難易度（問題レベル）を，先行研究で開発されたルーブリックをもとに，生成AIで自動判定する仕組みについて検討した結果を報告する．大学入学共通テストの問題を対象とした実験の結果，プロンプトエンジニアリング技術（Chain-of-Thought，Few-shot，Self-Consistency）を組み合わせた手法（提案手法）では，単純な指示（Zero-shot）と比べて判定の正確性・安定性が向上することを確認した．
＜キーワード＞　高等学校情報科，生成AI，ルーブリック，プロンプトエンジニアリング

1. はじめに

筆者らは高等学校情報科における教師（特に新任教師や教職課程の学生）の指導力向上を目的としてインストラクショナルデザイン理論を基盤とした作問・評価力向上システムの開発を目指している(松島･久保田 2025)．このシステムの中核として，教師が作成したテスト問題をシステムが自動で分析・フィードバックする機能があるが，本稿では，入力されたテスト問題の分野や難易度（以下，問題レベル)を生成AIで自動判定する仕組みについて検討した結果を報告する．

2. 関連研究と本研究の位置付け

高等学校情報科の評価手法については，平成28〜30年度の文部科学省委託事業（以下，先行研究）において包括的な検討がなされている．この先行研究では，AI/ビッグデータ技術を用いて試験問題の特性を評価する試みも行われた．具体的には，受験者の正誤分布に基づき，クラスタリングや因子分析によって問題を分類する手法が検討されたが，人手による問題の特性分類との明確な関連性は見出せず,自動評価の難しさが報告されている(文部科学省 2019)．

一方で,近年の生成AIの技術発展は目覚ましく,その高度な文脈理解能力や学習能力は,これまで困難とされてきた自然言語で記述された問題の特性判定に応用できる可能性がある．そこで本研究は，高等学校情報科のテスト問題を対象に,生成AIとプロンプトエンジニアリング技術を組み合わせて，先行研究では達成が困難であった問題レベルの自動判定に再挑戦するものと位置付ける．

3. 問題レベル判定の基準

問題レベルを判定する基準として，先行研究で開発されたルーブリックを用いる．これは，高等学校の共通教科情報で育成すべき能力を12の分野で体系化したもので,各分野はさらに能力の到達度を示す 6 段階のレベル（1-1〜４）で定義されている．例として，アルゴリズムとプログラミング分野のルーブリックを抜粋して**表１**に示す．本研究では，この「分野」と「レベル」を統合したID（例: K2-1）を自動判定の対象とする．なお，結果出力の際は判断の理由とともに出力させる．

4. 提案手法

以下に示すプロンプトエンジニアリング技術を組み合わせたプロンプトによる判定手法を提案する（以下，提案手法)．

- Chain-of-Thought: AI に段階的な思考を促すワークフローを指示する手法.
- Few-shot Prompting: いくつかの具体例を与えることでタスクに対する精度を向上させる手法.
- Self-Consistency: 同じ質問に対して複数の回答を生成し，それらの中から最も一貫性のある（多数派の）答えを最終的な回答として採用する手法.

表１ アルゴリズムとプログラミング分野のルーブリック（ID は筆者が付与）

ID	能力の到達度（Can-do 形式）
K1-1	与えられたアルゴリズム・プログラムの記述を認識できる
K1-2	与えらえたアルゴリズム・プログラムの動作をトレースできる
K2-1	与えられたアルゴリズム・プログラムの動作を説明できる
K2-2	与えられたアルゴリズム・プログラムを、指示された動作になるように修正できる（デバッグを含む）
K3	与えられた目的に応じた機能を満たすプログラムを設計・作成できる
K4	与えられた目的に応じた機能・要求をよりよく満たすようにプログラムを評価・改善できる

5. 評価実験と考察

5.1. 実験概要

提案手法が，単純な指示のみを与える「Zero-shot」と比較して，問題レベルを正確に判定できるかを検証する．対象として，令和７年度大学入学共通テスト情報Ｉ第１問問３(チェックディジットの問題)を用いた．この問題は，題材としてはネットワーク分野の問題であるが，与えられた生成方法（アルゴリズム)に従って計算(トレース)したり，アルゴリズムを説明したりできるかという能力を問うものであり，空欄【キ】がルーブリックの K1-2 に，空欄【ク】が K2-1 に該当する．したがって，正解は″K-1 と K2-1″の両方を判定することとした.各手法で判定を 30 回ずつ実行し，その結果を比較した．なお，検証に利用したモデルは OpenAI 社の gpt-4.1-2025-04-14，temperature は 0.0 である．

5.2. 結果と考察

30 回の実行結果を**表２**に示す．どちらも分野を誤って判定することはなかったが，提案手法が 30 回すべてで″K1-2 と K2-1″を正確に判定したのに対し，Zero-shot の正解率は 23.3%にとどまった．

表２　各手法における判定回数

判定 ID	Zero-shot	提案手法
K1-2 と K2-1	7 (23.3%)	30(100.0%)
K1-2 のみ	23(76.7%)	0 (0.0%)

提案手法と Zero-shot で判定精度に大きな差が生じた理由は,プロンプトが生成 AI の思考プロセスをいかに設計･誘導したかにある．Zero-shot で″K1-2 のみ″と判定したケースでは，23 回中 21 回（91.3%）が，空欄【ク】について,『具体的な入力例でトレースし，結果を導く力を問う問題である』としてアルゴリズムの動作を説明するという高次の思考ではなく，トレース能力の一部と解釈してしまっていた．一方で，提案手法ではプロンプト内で,複数のルーブリック ID を許容し各問を個別に評価するよう明示的に指示している．この「思考の足場」が，生成 AI に各問を独立してルーブリックと照合させ，それぞれのレベルを正確に判定させたと考えられる.つまり，AI に思考の手順を設計して与えることで，単に問題を認識するだけでなく，ルールに基づき的確に分類するという高度なタスクを実行させることができたといえる．

6. まとめと今後の展望

本稿では,生成 AI を用いて高等学校情報科のテスト問題の分野や難易度を自動判定する手法を提案した．今後の展望として，今回は１問のみしか判定に用いていないため，まずは評価対象の問題数を増やし，他分野においても同様に判定が可能か,定量的に検証する．また,複数の異なる生成 AI モデルで実験を行い，本手法の汎用性やモデル間の性能差を明らかにしたい．これらの検証を経て，開発中の作問・評価力向上システムへ実装し，教育現場における実用性を高めていく所存である．

付記

センター試験および共通テスト等の試験問題の研究利用については，2025 年６月 18 日付で大学入試センターから許諾を受けている．

参考文献

松島拓路，久保田真一郎（2025）インストラクショナルデザインを基盤とした AI 支援による高等学校情報科教師の作問・評価力向上システムの設計.日本情報科教育学会第 18 回全国大会講演論文集(WEB): 3-A-1，44-45.

文部科学省（2019）情報学的アプローチによる「情報科」大学入学者選抜における評価手法の研究開発　最終成果報告書. https://www.mext.go.jp/content/1412881_3_1_1.pdf(参照日 2025.06.18)

独立行政法人大学入試センター（2025）令和７年度大学入学共通テスト「情報Ｉ」