LLM を用いた主観的難易度に基づく問題推薦システムの提案と評価

Proposal and Evaluation of a Problem Recommendation System Based on Subjective Difficulty Using LLM

三浦 翔大[*1]，高木 正則[*1]
Shota Miura[*1], Takagi Masanori[*1]
[*1] 電気通信大学
[*1]The University of Electro-Communications
Email: s.miura@uec.ac.jp

あらまし：本研究では，学習者が求める適切な難易度の問題推薦を目的とし，大規模言語モデル (LLM) を活用して学習者の主観的フィードバック（自信度・挑戦度）に基づき問題推薦を行うシステムを提案する．提案システムは，学習者のプロファイルを生成・更新しながら「自信と挑戦度が釣り合う」問題を推薦する．評価実験の結果，LLM による自信度の予測は学習者プロファイルの活用により精度が向上し，ベースライン手法を上回ることが示された．
キーワード：問題推薦，適応型支援，生成 AI，数学科教育

1. はじめに

令和 3 年 1 月の中央教育審議会答申において「個別最適な学び」の実現が提言されて以来，ICT を活用した適応型学習支援の重要性が高まっている(1)．従来の適応型学習システム（Adaptive Instructional Systems: AIS）では，Bayesian Knowledge Tracing（BKT）(2)等を用いて学習者の知識状態を推定するとともに, 項目反応理論（Item Response Theory: IRT）(3)に基づき項目情報量を最大化する問題，すなわち正答確率が 0.5 付近の問題を提示する手法が広く用いられてきた．しかし，測定効率（項目情報量）を最大化するアプローチでは，学習者は常に正誤の境界線上にある問題への解答を強いられることになり，学習意欲を減退させる懸念がある．これは，既存手法が客観的な正誤情報のみを利用しており，学習者の「自信」や「挑戦度」といった主観的な心理的側面を十分に考慮できていないことに起因すると考えられる．Csikszentmihalyi のフロー理論(4)によれば，学習者の「スキル（自信）」と課題の「挑戦度」が高度に均衡した際に, 最も効果的な学習が生じる「フロー状態」に入るとされている．そこで，本研究では，学習者が求める心理的に適切な難易度の問題推薦を目的とし，大規模言語モデル（Large Language Models: LLM）を活用して，学習者の主観的フィードバックに基づいて「自信と挑戦度が釣り合う（Gap≈0）」問題を推薦するシステムを提案する．ここで，「Gap = 自信度 − 挑戦度」と定義する．

2. 関連研究

Labaj ら(5)は，学習者の主観的難易度を協調フィルタリングに活用する手法を提案したが，類似ユーザーのデータ蓄積が必要であり，推薦の説明性にも課題がある．Xu ら(6)は，LLM を推薦システムに応用するフレームワーク「ProLLM4Rec」を提案し，Role prompting や Chain-of Thought prompting（CoT）の有効性を示した．本研究では，類似ユーザーのデータ蓄積や正誤情報を必要とせず, LLM の推論能力により学習者個人の主観的フィードバックのみから学習者が求める適切な難易度の問題を推薦する点に特徴がある．本稿では，以下の 2 つのリサーチクエスチョンを設定して実施した評価実験の結果を報告する．

RQ1 : LLM は学習者の主観的難易度（自信度・挑戦度）をどの程度正確に予測できるか？

RQ2：プロンプトの工夫や学習者プロファイルの付与により主観的難易度（自信度・挑戦度）の予測精度は向上するか？

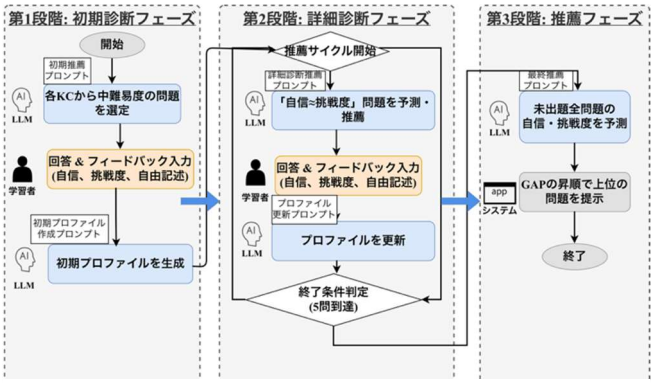

図 1 提案システムの診断フロー

3. システムの提案と開発

本システムの概要を図 1 に示す．システムは 3 段階で構成される．

第 1 段階（初期診断）では，LLM が対象とする学習単元（Knowledge Component．以下，KC） の中で，中程度の難易度の問題を 1 問選定し，その問題に対する学習者の主観的フィードバック（自信度・挑戦度・自由記述）を収集する．また，収集したフィードバックから学習者の初期プロファイルを生成する．

表 1 予測精度のベースラインと LLM との比較

手法	自信度		挑戦度	
	相関	MAE	相関	MAE
提案手法	0.600	1.09	0.371	1.23
問題平均	0.591	1.19	**0.632**	**0.83**
ユーザー平均	0.501	1.26	0.416	0.97
全体平均	0.000	1.65	0.000	1.18

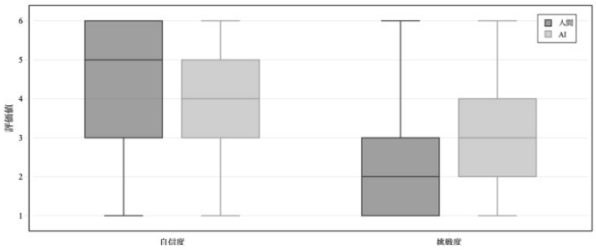

図 1　LLM と人間の自信・挑戦度の分布

第 2 段階（詳細診断）では，学習者の初期プロファイルと直近の解答履歴に基づき，Gap ≈ 0 となる問題を予測し，学習者に出題する．また，その問題に対する学習者の主観的フィードバックを収集し，学習者のプロファイルを更新する．これを 5 回繰り返す．

第 3 段階（推薦）では，更新されたプロファイルに基づき未出題の全問題に対する自信度・挑戦度を予測し，Gap の昇順で問題を推薦する．

プロンプト設計には，Xu ら[6] の知見に基づき，(1) Role prompting（適応学習の専門家ロール），(2) CoT prompting（段階的推論）を採用した．

4.　評価実験

4.1 実験概要

線形代数を履修済みの大学生・大学院生 10 名を対象に実験を実施した．被験者は初期診断，詳細診断フェーズ後に主観的フィードバックを入力し，その後 30 問に対して自信度・挑戦度を回答した．LLM には「gemini-3.0-pro-preview」を使用した．

4.2 RQ1：主観的難易度の予測精度

LLM の予測精度をベースライン手法（全体平均・問題平均・ユーザー平均）と比較した結果を表 1 に示す．ここでベースラインは，自信度と挑戦度を全データの平均（全体平均），同一問題への回答の平均（問題平均），同一学習者の回答の平均（ユーザー平均）でそれぞれ予測し，その差分として GAP を算出した．自信度予測では提案手法が全ベースラインを上回り最良の精度を示した．相関係数は 0.600，MAE は 1.09 であった．この成功要因として，学習者プロファイルから「自分がどの程度できると思っているか」という主観的自己認識を推定できた点が挙げられる．

一方，挑戦度予測では問題平均ベースラインが提案手法を上回った．相関係数は問題平均が 0.632 であったのに対し，LLM は 0.371 にとどまった．これは，学習者の過信傾向を LLM が適切にモデル化できなかったためと考えられる．図 1 に LLM と人間の自信度・挑戦度分布の比較を示す．

表 2 予測精度のベースラインと LLM との比較

指標	提案手法	ノーマル	p 値
自信度 MAE	1.09	1.38	<0.001*
自信度 相関	0.600	0.421	0.003*
挑戦度 MAE	1.23	1.30	0.727
挑戦度 相関	0.371	0.385	0.837
GAP MAE	2.11	2.39	0.004*

*: $p < 0.05$ で有意

4.3 RQ2：プロンプト設計の効果

学習者プロファイルを含む提案手法と，プロファイルを含まないノーマルプロンプトを比較した結果を表 2 に示す．学習者プロファイルを含めることで，自信度 MAE は 0.29 向上し（$p < 0.001$），GAP MAE も 0.28 向上した（$p = 0.004$）．これらの改善は統計的に有意であった．この結果は RQ1 の考察と整合的であり，自信度は学習者の主観的自己認識に依存するため，プロファイル情報を明示的に与えることで予測精度が向上したと考えられる．一方，挑戦度には効果が限定的であった．

5.　おわりに

本研究では，提案手法を用いた主観的難易度に基づく問題推薦システムを提案し，予測精度とプロンプト設計の効果を検証した．自信度予測では LLM が有効であり，学習者プロファイルを活用したプロンプト設計により予測精度が向上することが示された．今後は，挑戦度予測の改善に向けて，問題の特徴情報を付与したプロンプト設計を検討する．

参考文献

(1) 中央教育審議会.「令和の日本型学校教育」の構築を目指して（答申）．文部科学省，2021. https://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/toushin/1412985_00001.htm.

(2) Albert T. Corbett and John R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, Vol. 4, No. 4, pp. 253–278, 1994.

(3) Wim J. van der Linden and Ronald K. Hambleton, editors. Handbook of Modern Item Response Theory. Springer, New York, 1997.

(4) Mihaly Csikszentmihalyi. Flow: The Psychology of Optimal Experience. Harper & Row, New York, 1990.

(5) Martin Labaj and Mária Bieliková. Using perceived difficulty for personalized exercise recommendation. In Proceedings of the 2014 IEEE 14th International Conference on Advanced Learning Technologies (ICALT), pp. 21–23. IEEE, 2014.

(6) Lanling Xu, Jianxun Lian, Wayne Xin Zhao, Ming Gong, Linjun Shou, Daxin Jiang, and Ji-Rong Wen. Prompting large language models for recommender systems: A comprehensive framework and empirical analysis. arXiv preprint arXiv:2401.03787, 2024.