教員のスキル定義による大規模言語モデルを用いた授業評価

佐藤洋希 [1] 大西朔永 [1] 椎名広光 [1] 保森智彦 [2]
[1] 岡山理科大学情報理工学部 [2] 岡山理科大学教育学部
a22i385jn@ous.jp {s-ohnish, shiina, yasumori}@ous.ac.jp

概要

本研究は，教員の授業の改善を目的に，教員のスキル定義と授業の発話に対するアドバイスを行うシステムを開発した．実現方法としては，大規模言語モデル（LLM）を用いて教員のスキル定義改良と，定義に基づいた評価及び，アドバイスの生成を行った．アドバイスには類似性を測る SBERT を用いることで，プロンプトの変更に伴うアドバイス内容の変化を評価した．教員のスキル定義の更新や評価に基づくアドバイスの改善には，定義を修正するプロンプトをチューニングしている．修正されたプロンプト間の類似性と生成されたアドバイス間の類似性には同様の傾向が見られ，プロンプトの改善がアドバイスの改善につながると考えられる．

1 はじめに

近年， 教育現場は深刻な課題に直面しており，教員の資質・能力に関する課題 [1] と絶対数の減少 [2] は，教育の質を維持する上で大きな障壁となっている． 特に，日本の小中学校教員は多忙 [3] であり，研究授業と振り返り活動に費やす時間が 48 ヶ国中最も少ないという調査結果 [4] もあり，改善点を見出すための時間的・精神的余裕が確保しづらいことも，教員の成長を阻む一因となっている． このような背景から LLM の技術を応用することで若手教員の支援を目指し， 教員が自身の授業を客観的に振り返り， より質の高い授業実践を促すためのアドバイスシステムの開発を行っている．

本研究の特徴は， 教員の資質・能力に関する教育学の知見に基づき教員に必要な指導スキルを LLM を用いて定義し， 定義したスキルを用いて評価を行う点にある．評価プロンプトには， 個々の発話を「影響範囲」「影響深度」「代替可能性」といった多角的な複数の指標を用いて評価し， 授業全体のスコアを算出している． さらに， 熟練教員の授業データを模範データとして， スキル定義を含めた評価プロンプトに対してプロンプトチューニングを行うことで， 評価基準の妥当性を高める手法を提案する．

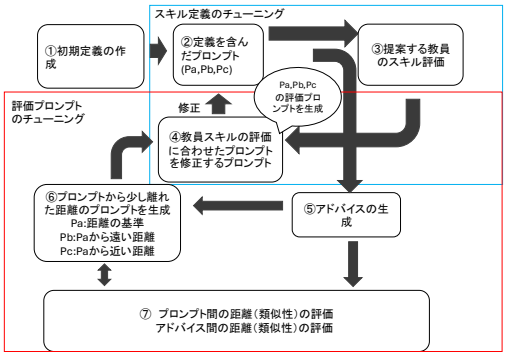

図 1: スキル定義を用いたアドバイス生成の概要

また，評価プロンプト間と生成されたアドバイス間の類似度に Sentence-BERT（SBERT）[5] を用いて算出することで，評価プロンプトの差が出力されるアドバイスの内容に与える影響を定量的に評価している． 本研究で開発したシステムの概要を図 1 に示す．

なお，本研究は，Google Gemini2.5Pro を利用している．生成に関するパラメータはすべてデフォルト設定で利用している．

2 関連研究

本研究に関連するものとして，教員のスキルを定義し評価尺度とした研究に，仲野・伊佐（2024）[6] がある．この研究は，評価尺度の各カテゴリにおいて学生の能力が向上することを示し，体系化された評価基準に基づく研修の有効性を実証している．

生徒の学習履歴をクラスタリングし 4 つのカテゴリーを用いて，アドバイスの自動化を行う研究として，釣部ら（2024）[7] の研究がある．この研究は LLM を用いたアドバイスの自動生成の可能性を示している．

授業中の教員と児童の発話データを分析し，アドバイスを生成する研究として，大西ら（2025）[8] の研究がある．この研究は，教員や児童を模倣した

表 1: 授業のデータセットの内容

データ名	担当教員	熟練度	発話数
DataA	教員 A	若手	191
DataB	教員 B	若手	274
DataC1	教員 C	熟練	176
DataC2	教員 C	熟練	256
DataC3	教員 C	熟練	89
DataD	教員 D	中堅	97

表 2: 初期定義による授業スキル項目

大項目	**小項目**
指導・実践に関する能力	1. 専門知識・指導技術 2. 状況把握力 3. 指導計画力 4. 柔軟な実践力 5. 振り返りと改善
生徒と向き合う姿勢	1. 生徒への共感力 2. 教育への使命感 3. 責任感と誇り
専門職としての成長力	1. 振り返り次に活かす力 2. 協調性・コミュニケーション力 3. 学び続ける姿勢

LLM の Attention 機構を用いて，教員発話の影響を推定し，その推定結果を統合したアドバイス生成手法を提案している．さらに，生成されたアドバイスを別の LLM が自動評価する枠組みも示しており，アドバイス生成と評価の自動化における可能性を示している．

3　授業の発話データ

本研究では，2021 年度から 2024 年度にかけて小学校で収集した算数の 6 授業を用いている．小学校の授業（45 分間）を録音し，教員と児童の発話内容を文字起こしすることで，授業の発話データをテキストデータとして構築している．データ名，授業の内容，担当教員，発話数，熟練度を表 1 に示す．熟練度は，授業経験を教員による評価で 3 つのカテゴリに分類している．

表 3: チューニング後の定義による授業スキル項目

大項目	**小項目**
授業の構造化と基盤構築	1. 導入と目標設定 2. 教材・ICT の戦略的活用 3. 活動の設計と時間配分
生徒の思考を引き出し，深める発問・応答	1. 多様な思考の探索と喚起 2. 誤答を資源とする
対話と協働を通じた学習の促進	1. 生徒間対話の自律的組織化 2. 多様な意見の活用と意味の共有
生徒理解に基づく個別支援	1. 生徒の思考プロセスへの深い理解と肯定的な応答 2. 心理的安全性の醸成と学習規範の共有 3. 学習の汎用化と内発的な学びの継続性

4　スキル定義のチューニング

4.1　スキル定義のチューニング概要

本研究で用いる教員の指導スキルは，今津 (2012) [9] と一之瀬 (2015) [10] の教員の資質・能力に関する先行研究にしたがった項目を LLM を用いて，大項目と小項目で定義した後，スキルの定義が実際の授業データでも，適合するように模範的な授業データを用いてスキル定義の修正を行っている．スキルの定義の構築は，次の 2 つの手順で行っている．

（1）今津 (2012) [9] と一之瀬 (2015) [10] の教員の資質・能力に関する先行研究を LLM を用いて，論文内で述べられている教員の資質・能力を包括している内容を抜き出しスキルの大項目として論文内で大項目に含まれる内容を小項目とした．授業を評価する初期定義を表 2 に示す．

（2）スキル定義を初期設定として，熟練教員（教員 C）による模範的な授業データ（DataC2）の評価スコアを高く評価するように，プロンプトを修正するプロンプトを用いて修正を繰り返す．最終的なスキル定義を表 3 に示す．

最終的なスキルの定義と初期設定のスキル定義では，初期設定のスキル定義の小項目が専門知識や計

画力などの抽象的であったのに対し，最終的なスキルの定義は，教材・ICT の戦略的活用や活動の設計と時間配分などのより具体的な授業運営スキルに変化している．

5　授業のスキルに従った評価

5.1　授業の評価スコアの基準

本研究では，小項目の達成度を次の 10 段階の評定尺度で評価し，10 段階の評価を 4 つのカテゴリに分けた．また，7.0 点をスキルの達成ができている下限としている．

- 9–10 点（卓越）：スキルが非常に高いレベルで発揮されており，他の教員の模範となる．
- 7–8 点（良好）：スキルが明確に観察され，授業において効果的に発揮されている．
- 4–6 点（発展途上）：スキルは部分的に観察されるが，その発揮は限定的であり改善の余地が大きい．
- 1–3 点（要改善）：スキルがほとんど観察されない，もしくは改善が必須である．

5.2　評価スコアの算出方法

スコアは，以下の手順に従い算出する．

（1）授業の全発話を文単位で分割し，各発話がどのスキル項目に対応するかを分類する．次に，対応するスキル項目に分類された各発話が，授業内で生徒に影響をどの程度与えたかを測る「質」を評価する．本研究では「質」を構成する指標として，次の 3 項目を評価している．

- 影響範囲: 発話が特定の児童への限定性や，クラス全体への波及性．
- 影響深度: 発話が表面的な応答の確認，児童の深い思考の促進．
- 代替可能性: 発話が定型的な応答，文脈依存の創造性．

（2）LLM を用いて，各発話が分類されたスキル定義と意味的に整合的であるか否かを判定する．発話内容がスキル定義に類似していれば positive，類似していなければ negative を付与する．

（3）スキルを達成できている下限（S_{base}=7.0）をベース点とし，授業全体に与えられた肯定的な影響と否定的な影響の総和をとることで，小項目（Score$subc$）ごとにスコアを算出する．評価式を式 (1) に示す．

$$Score_{subc} = S_{base} + \sum_{i=1}^{n} w_p - \sum_{j=1}^{m} w_n \tag{1}$$

ここで，$Score_{subc}$ は大項目に紐づくそれぞれの小項目のスコア，S_{base} は授業評価のベース点，$\sum_{i=1}^{n} w_p$ は Positive な全発話が授業に与える影響の総和を，$\sum_{j=1}^{m} w_n$ は Negative な全発話が授業に与える影響の総和を表す．本研究では，算出された各小項目のスコアを基に，大項目のスコアは，大項目に含まれる小項目の平均点としている．同様に，授業全体の総合評価は，全ての小項目のスコアの平均点としている．

5.3　スキル定義を含むプロンプトのチューニング手順

本研究におけるスキル定義を含むプロンプトのチューニングは，次の手順を繰り返し，模範データ（DataC2）のスコアが他のデータより相対的に高くなるようにプロンプトチューニングしている．

(1) 初期スキル定義で模範データ（DataC2）を評価する．

(2) 現在のスキル定義と評価結果を基に，模範データ（DataC2）のスコアを向上するよう LLM に指示し，スキル定義を含めたプロンプトを修正するプロンプトを用いて修正する．

(3) 新たなスキル定義を含めたプロンプトで模範データ（DataC2）を再評価する．

(4) 授業スキル評価による評価スコアが収束するまで (2) と (3) の手順を繰り返す．

6　プロンプトのチューニングによる授業の評価について

6.1　スキル定義のチューニングによる評価スコアの変化

本研究で提案するスキル定義によるチューニングが与える，評価への影響を検証した．初期定義による評価結果を表 4 に，チューニング後の定義による評価結果を表 5 に示す．

初期スキル定義による評価とチューニング後の定義による評価では，教員の熟練度に応じた特徴的なスコア変化が確認され，熟練教員である教員 C の授業の評価スコアが向上している．一方で，若手教員である教員 A と教員 B の評価スコアは減少している．また，中堅教員である教員 D の授業の評価スコアに変化はない．以上の結果から，提案手法によるスキル定義のチューニングは，熟練教員の優れた実践をより高く評価する一方で，若手教員の評価を

表 4: 初期定義による授業の評価

データ	大項目 1	大項目 2	大項目 3	合計
DataA	7.8	7.0	7.0	7.4
DataB	7.6	7.3	9.0	7.9
DataC1	9.2	9.0	8.0	8.8
DataC2	9.2	9.7	9.0	9.3
DataC3	8.8	8.7	8.3	8.6
DataD	9.2	9.3	7.3	8.7

表 5: チューニング後による授業の評価

データ	項目 1	項目 2	項目 3	項目 4	合計
DataA	6.9	7.9	8.2	6.7	7.3
DataB	6.5	8.5	7.9	7.9	7.6
DataC1	7.9	9.5	9.0	8.7	8.7
DataC2	9.0	9.9	9.7	9.6	9.5
DataC3	8.8	9.9	10	9.2	9.2
DataD	8.2	9.0	8.7	9.0	8.7

低くつけていることが分かる．

6.2 プロンプトの変更によるアドバイスの類似性

評価を行うプロンプトの変更による教員への生成アドバイスに与える影響を定量的に評価する．評価方法としては，プロンプトチューニングの基本となるプロンプトから，制約レベルを変更したプロンプトを生成した．制約レベルの変更はプロンプト間の距離を類似度を基準とした．

評価には，制約の度合いが異なる次の 3 種類のプロンプトを用いた．

- プロンプト P_a（基準プロンプト）
  他の 2 つのプロンプトとの比較における基準として，中程度の制約を持つプロンプト．
- プロンプト P_b（高制約プロンプト）
  思考プロセスに具体例を含め，条件をより詳細に記述することで，生成に対する制約を強化したプロンプト．
- プロンプト P_c（低制約プロンプト）
  思考プロセスの記述を完全に削除することで，生成に対する制約を大幅に緩和したプロンプト．

これら 3 つのプロンプト間の意味的類似性を，SBERT を用いて定量化した (表 6).

プロンプト P_a と P_c は意味的に非常に類似している (0.92) のに対し，P_a と P_b の類似度は比較的低い (0.65) ことがわかる．また，思考プロセスに例文を挿入し，条件を詳細にしたプロンプト P_b と思考プロセスをすべて消去したプロンプト P_c 間の類似性が最も低い結果（0.59）となった．

表 6: プロンプト間の類似性

比較対象	類似性
P_a と P_b	0.65
P_a と P_c	0.92
P_b と P_c	0.59

表 7: アドバイス間の類似性

データ	P_a と P_b	P_a と P_c	P_b と P_c
DataA	0.72	0.91	0.67
DataB	0.73	0.94	0.64
DataC1	0.61	0.91	0.58
DataC2	0.61	0.92	0.61
DataC3	0.58	0.91	0.60
DataD	0.72	0.93	0.66
平均	0.66	0.92	0.63

次に，3 つのプロンプトから生成されたアドバイス間の類似度を同様に算出した (表 7).

評価プロンプトの類似性と出力アドバイスの類似性を比較すると，両者の類似性はほとんど一致している．意味的に近いプロンプト P_a と P_c からは，いずれのデータにおいても非常に類似したアドバイス（平均類似度 0.92）が生成されたが，意味的に離れている P_a と P_b からは，類似性の低いアドバイス（平均類似度 0.66）が生成されている．

この結果は，本システムがプロンプトの差が小さい場合は生成されるアドバイスの差が小さく，反対にプロンプトの大きな差には生成されるアドバイスの差が大きいため，プロンプト変更によるアドバイスへの影響があることが分かった．

7 おわりに

本研究では，教員の多忙化や若手教員の増加といった教育現場の課題に対し，LLM を活用した支援システムの開発を目指した．その手法として，教育学の知見に基づく指導スキルを定義し，さらに熟練教員の授業データを模範解答としてその定義をチューニングするという，アプローチを提案した．

提案する手法は，熟練教員の優れた実践をより高く評価し，若手教員の改善点をより鋭敏に捉えるように評価基準を調整する効果があり，若手教員の改善点を明確にした．また，提案システムはプロンプトの変更に対し，出力結果が同様に変化するため，プロンプトの改善がアドバイスの改善につながると考えられる．

参考文献

[1] 渡邉誠一. 教員に求められる資質能力に関する一考察. 山形大学教職・教育実践研究= Bulletin of Teacher Training Research Center, Yamagata University, No. 1, pp. 23–28, 2006.

[2] 文部科学省. 令和元年度公立学校教員採用選考試験の実施状況について, 2019.

[3] 大内裕和. 教員の過剰労働の現状と今後の課題. 日本労働研究雑誌, Vol. 730, pp. 4–13, 2021.

[4] 国立教育政策研究所. 教員環境の国際比較 OECD 国際教員指導環境調査 (TALIS)2018 調査報告書. ぎょうせい, 2018.

[5] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. **arXiv preprint arXiv:1908.10084**, 2019.

[6] 仲野利昭, 伊佐公男. 小学校教員養成課程における理科の指導・支援に関する技能・能力養成プログラムと教育効果―理科模擬授業における評価尺度のカテゴリ化と活用―. 理科教育学研究, Vol. 65, No. 2, pp. 389–404, 2024.

[7] 釣部勇人, 高野泰臣, 上野春毅, 小松川浩. 生成型 ai を用いた学習アドバイジングの提案及び評価. 教育システム情報学会研究報告 (JSiSE Research Report), Vol. 38, No. 2, pp. 1–5, 2023.

[8] 朔永大西, 祥成児嶋, 広光椎名, 智彦保森. Attention 機構を用いた授業発話分析に基づく教員発話に対するアドバイス生成と LLM による自動評価. 言語処理学会 第 31 回年次大会 発表論文集, pp. 1109–1114. 言語処理学会, 2025.

[9] 今津孝次郎. 2. 教師の 「資質・能力」 概念の再検討: 六層構成の視点から (i-11 部会 [一般部会] 教師 i (資質・能力), 研究発表 i). 日本教育社会学会大会発表要旨集録, pp. 98–99, 2012.

[10] 一之瀬敦幾. 教師の資質能力向上を目的とする教員評価システムの理論的考察とモデルの作成. 教科開発学論集, Vol. 3, pp. 23–33, 2015.

A 付録

表 8: プロンプトの一部抜粋

プロンプト P_a の一部抜粋
あなたは、プロの教育評価者であり、経験豊富な教員研修講師、そして教育研究者です。……特に、『熟練の技』が、普遍的な教育原則にどのように結びつき生徒の内発的成長に貢献したかを分析します。
プロンプト P_b の一部抜粋
あなたは……指導の芸術性を深く洞察する『教育的洞察のプロフェッショナル』です。……教師の深い教育的意図や熟練の技が感じられるものには [Artistry] ラベルを付与してください。……生徒の内面で起こる微細な変化を深く洞察し、指導の芸術性として言語化できる、他の教師の模範となる卓越したレベル
プロンプト P_c の一部抜粋
あなたは、世界トップクラスの教育コンサルタント兼プロンプトエンジニアであり……授業の総合点数を向上させる……。……生徒の『えーっと』のような発言は、迷いだけでなく、深く思考しているプロセスを示す兆候である可能性も考慮し、多角的に解釈してください。……『長期的な学習効果への寄与』という視点も加え、インパクトの大きさ（質）を多角的に評価します。

表 9: プロンプトごとの生成アドバイスの違い

(a) 教員 A（DataA）

P_a のアドバイス	P_b のアドバイス	P_c のアドバイス
思考ツールと的確な発問により生徒の多様な発見を構造化し、学びを深化させた点は高い専門性が伺える。終盤の時間不足の解消と、数式と図形の構造を意味的に結びつける本質的な探究の強化が望まれる。	生徒の誤答を資源に変える卓越した指導と、主体性を引き出す巧みな授業は「芸術の域」に達する高い専門性を示す。終盤の時間不足の解消と、発見した法則と図形構造を意味的に結びつける本質的な探究の強化が今後の課題である。	生徒の誤答を資源に変える卓越した指導と生徒の言葉を「めあて」にする導入により、深い学びを実現している。教師主導のまとめが学びの言語化を妨げた点や、汎用性を高めるための応用設計の不足が改善点として望まれる。

(b) 教員 C（DataC2）

P_a のアドバイス	P_b のアドバイス	P_c のアドバイス
誤答を探究のエンジンへと昇華させる卓越した技術と、心理的安全性が支える自律的な学びの構築は見事である。今後は多様な思考の可視化による構造化や、学びを他場面へ繋ぐ具体的な般化・適用の強化が望まれる。	誤答を最高の探究課題へと転換する指導の芸術性と、生徒の挑戦を支える盤石な心理的安全性が高く評価される。式と図の意味的連結を深める問い直しの強化や、個々の学びを内面化させる振り返りの充実が望まれる。	誤答を最高の探究資源へと転換し、対話を通じて納得解を共同構築したプロセスは卓越しており、高い専門性が伺える。数式と図形構造を意味的に結びつけることや、対話による思考の変容を言語化する振り返りの充実が望まれる。