

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

大西朔永¹ 児嶋祥成² 椎名広光³ 保森智彦⁴

¹ 岡山理科大学大学院 総合情報研究科 ² 岡山理科大学大学院 理工学研究科

³ 岡山理科大学 情報理工学部 ⁴ 岡山理科大学 教育学部

{i22ed08bf,r24smd8qv}@ous.jp {shiina,yasumori}@ous.ac.jp

概要

本研究は、教員支援を目的に、授業発話分析に基づくアドバイス生成と LLM-as-a-judge による自動評価手法を提案した。教員や児童を模倣した 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を行った。さらに、LLM による自動評価を用いて、生成されたアドバイスの質を客観的に評価する手法を提案した。実験では、影響推定を統合したアドバイス生成が、LLM による自動評価で高い評価を得る傾向を示した。提案手法の有効性が示唆されたが、自動評価の改善と信頼性向上が今後の課題である。

1 はじめに

近年、Transformer[1] を用いた大規模言語モデル (Large Language Model, LLM) の急速な発展により、自然言語生成の性能は飛躍的に向上している。そして、OpenAI の Generative Pre-trained Transformer (GPT) [2, 3, 4] を代表に、高コストではあるが高性能な LLM を API によって利用可能になっている。

一方、教育現場においては、学習者と教員を支援する機械的な方法への需要が高まっている。2017 年に示された小学校指導要領 [5] では「主体的・対話的で深い学び」を実現する授業が目指されており、小学校の授業では様々な工夫が試みられている。そのため、文部科学省中央教育審議会 [6] は継続的に学ぶ教員を求めており、その一環として省察活動 [7] が注目されている。省察活動とは、教員が教育実践を振り返り、改善策を検討する行為である。しかし、日本の小中学校教員は、研修に費やす時間が 48 ヶ国中最も少ないという調査結果 [8] がある。多忙な教員にとって、時間をかけずに効果的な省察活動ができる機械的な分析手法の開発は喫緊の

課題である。教員支援の研究としては、「主体的・対話的で深い学び」の観点から、省察活動のデジタル化に着目した研究 [9]、コンピュータシステムを用いた発話分析 [10] が行われている。しかし、学習者支援の研究に比べ、教員を支援する研究は少ない。

本研究では、教員支援を目的として、授業における教員と児童の発話を対話として捉えた研究を行っている。教員と児童を模倣するニューラル対話モデルを構築し、モデル内の Attention 機構を用いて、教員発話が児童発話から受けた影響や教員発話が児童発話に与えた影響を定量的に推定し、その推定影響に基づいて LLM でアドバイスを生成する手法を提案している。Attention 機構を用いた教師発話の影響推定により、教師が一方的に話す状況を分析・抽出でき、その影響を LLM への入力に用いることで、アドバイス生成の改善を行っている。

また、LLM の発展により、LLM による評価が可能となったため、人手評価の代替として、LLM を用いた自動評価 (LLM-as-a-judge) [11] が提案されている。LLM による自動評価を教育分野に用いた研究として、課題評価ツールとして LLM による自動評価を用いた研究 [12] があり、大学の 1000 人以上が受講する授業において、ティーチングアシスタントとして LLM を活用し、学生の提出物を GPT-4 によって評価している。

従来の自動評価では評価困難なアドバイスという対象に対して、LLM による自動評価を用いることで、人的コストを掛けない評価が可能である。そこで、本研究では、アドバイス生成における影響推定の有効性を確認するために、LLM による自動評価において定量的にプロンプトの評価も行っている。LLM による自動評価では、影響推定に基づいて生成されたアドバイスは、影響推定に基づかない場合に比べて高い評価を得る傾向が示唆されている。

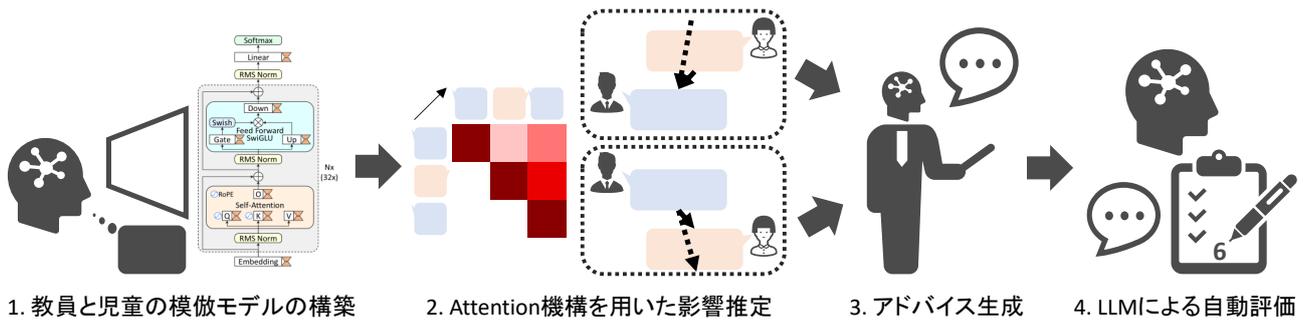


図 1 Attention 機構を用いた影響推定に基づくアドバイス生成と LLM による自動評価の概要

2 対話データ

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科目の 6 授業を用いている。小学校の授業（45 分）を録音し、教員と児童の発話を文字起こしすることで、テキストの対話データを構築している。対話データ全体の発話数は 1049 個であり、最小発話数が 88、最大発話数が 273 と、授業間で発話数に差が見られる。

また、LLM のファインチューニングに利用するために、授業全体の対話を最大発話数 6（最大 3-turn）の短い対話に分割したデータを構築している。

3 アドバイス生成

本研究では、LLM による教員発話の影響推定に基づいたアドバイス生成を行っている。アドバイス生成の手順としては、(1) 教員と児童を模倣するニューラル対話モデルを構築し、(2) 模倣モデル内の Attention 機構を用いて教員発話に対する影響を推定し、(3) 推定影響に基づいて LLM でアドバイスを生成している。本研究の概要を図 1 に示す。

3.1 教員と児童の模倣モデルの構築

本研究では、授業の対話データを用いて、LLM をファインチューニングすることで、教員と児童を模倣するモデルを構築している。小学校の授業に関する知識や対話方法を学習させるために、授業の対話データを用いて、QLoRA[13, 14] でベースモデルをファインチューニングし、教員や児童の応答を生成できるように訓練している。

LLM のベースモデルには Youri 7B Chat[15] を用いており、QLoRA のハイパーパラメータは、低ランク行列の r を 16、スケールリングの α を 16、LoRA の適用重みの種類を全てとしている。

3.2 模倣モデルの Attention Weight を用いた発話影響推定

自然言語処理の XAI（Explainable Artificial Intelligence）分野 [16, 17] では、説明可能な AI を目指し、AI の推論に対してその根拠を示す手法 [18] などが研究されている。授業の分析においても、教員や児童を模した AI モデルに対して、XAI の推論根拠を示す手法を用いて分析することで、発話の理由を示すとともに、発話に関して教員にフィードバックすることが可能になると考えられる。そこで、本研究では、模倣モデルの発話生成から得られる Attention Weight を用いて、授業発話の影響を推定している。

本研究で推定した影響の種類を以下に示す。

- (1) 教員発話が児童発話から受けた影響
- (2) 教員発話が児童発話に与えた影響

教員発話が児童発話から受けた影響の推定

Llama 2[19] のような Decoder のみを用いたモデルでは、トークンを生成する際にそれ以前のトークンの Attention Weight が計算される。そして、生成した応答のトークンに影響を与えたコンテキストのトークンがあり、影響度合いと Attention Weight の大きさは比例すると考えられる。そこで、Attention Weight を用いて、発話が受けた影響を推定している。

Attention Weight を用いた影響の推定手順を述べる。(1) コンテキストから応答発話を強制的に生成することで、Attention Weight を取得し、(2) Attention Weight から応答のトークン間の不要な重みを除去している。(3) 発話間の影響を推定するために、コンテキストと応答において、発話単位でトークンの重みを平均し、発話全体の重みに変換する。(5) 重みを 0-1 に正規化している。

教員発話が児童発話に与えた影響

本研究では、(1) 発話が受けた影響の推定結果か

ら、(2) 発話が与えた影響を推定している。授業対話中の発話数を N 、授業の対話は $D = u_1, u_2, \dots, u_N$ 、影響推定に用いる対話の発話数を 5 とする。発話を受けた影響では、最後の発話 u_i がそれまでの発話 $u_{i-5}, u_{i-4}, \dots, u_{i-1}$ から受けた影響を推定している。ここで、発話 u_i が発話 u_{i-1} から受けた影響は、発話 u_{i-1} が u_i に与えた影響であると考えられる。同様に発話を受けた影響の推定を全ての授業発話に対して行った結果、発話 $u_i, u_{i+1}, \dots, u_{i+4}$ が発話 u_{i-1} から受けた影響が推定されている。つまり、発話が与えた影響として、発話 u_{i-1} が発話 $u_i, u_{i+1}, \dots, u_{i+4}$ に与えた影響が得られる。

3.3 影響推定に基づくアドバイス生成

本研究では、LLM のプロンプトに影響推定結果を結合することで、影響推定に基づくアドバイスを生成している。そして、推定影響の活用によるアドバイスへの影響を確認するために、単純なアドバイス生成のプロンプト (1) から、現状分析として影響推定を活用したアドバイス生成のプロンプト (4) までの 4 種類のプロンプトでアドバイスを生成している。ただし、推定影響を用いない場合は対話の表を結合している。各プロンプトの概要を以下に示す。

- (1) アドバイス生成を指示
- (2) アドバイスの方向性とアドバイス生成を指示
- (3) アドバイスの方向性と推定影響に基づいたアドバイス生成を指示
- (4) 現在の推定影響と理想の発話影響の違いに基づいたアドバイス生成を指示. (4-1) 理想の発話影響を生成, (4-2) アドバイスの方向性 (現在と理想の違い) に基づいたアドバイス生成

本研究では、OpenAI の API を用いて、アドバイス生成を行っている。API では、Endpoint として `/v1/chat/completions`、Model には `gpt-4o-2024-08-06` を用い、生成するテキストのランダム性に影響する Temperature は 0.2 と低い値に設定している。

4 LLM による自動評価

LLM による自動評価を調査した研究 [11] では、評価結果の出力にはスコア、ランキング、選択の 3 種類があるとしている。スコアは、各サンプルに対して、連続値や離散値のスコアを出力し、定量的な比較を行うことが可能なため一般的に広く採用され

```
{アドバイス生成用のプロンプトと生成アドバイス}
## 指示
1. 評価基準に基づいて、教員の発話に対するアドバイスを評価してください。 ...
## 条件 ...
## 評価基準
- 指示に対する忠実性 ...
- 対話内容とアドバイスの関連性 ...
- アドバイスの適切さ ...
## 出力形式
### 評価理由
{評価理由}
### 評価
{評価値}
```

図 2 LLM による自動評価用プロンプトの概要

ている。ランキングは、各サンプルを比較し、相対的な順序を出力する。選択は、サンプルの中から最適な候補を 1 以上選択する手法である。教育分野において、課題評価ツールとして LLM による自動評価を用いた研究 [12] では、スコアを出力する手法を採用している。そこで、本研究では、アドバイスの評価手法として、スコアを出力する手法を用いた LLM による自動評価を行っている。

本研究では、影響推定を活用して生成したアドバイスに対して、LLM による自動評価を行っている。LLM で用いている評価用プロンプトは、(1) アドバイス生成用のプロンプトと生成されたアドバイス、(2) タスクに関する指示、(3) 指示の補足情報である条件、(4) 評価基準、(5) 出力形式で構成している。評価用プロンプトの概要を図 2 に示す。プロンプトにおいては、{アドバイス生成用のプロンプトと生成アドバイス} の部分を (1) アドバイス生成用のプロンプトと生成されたアドバイスに置換している。評価では、1-6 の 6 段階評価で、最高評価は 6 とし、3 種類の評価基準に基づく評価を指示している。

生成したアドバイス評価の実験設定を以下に述べる。評価には計 10 件の対話と推定影響を使用しており、(1) 教員発話が児童発話から受けた影響、(2) 教員発話が児童発話に与えた影響の 2 種類の影響からそれぞれ 5 件ずつを用いている。10 件の対話と推定影響に対して、4 種類のアドバイス用プロンプトでアドバイスを生成し、プロンプトの比較を行っている。LLM による自動評価では、計 40 件の評価用サンプルを評価用プロンプトで評価し、2 種類の影響と 4 種類のアドバイス用プロンプトごとに集計している。LLM には OpenAI の API を使用しており、その設定はアドバイス生成と同様である。

(1) 教員発話が児童発話から受けた影響に関して、生成されたアドバイスに対する LLM による自動評

表1 受けた影響の自動評価

プロンプト	平均	最小	最大	分散
プロンプト (1)	5.2	5	6	0.16
プロンプト (2)	5.4	5	6	0.24
プロンプト (3)	5.6	4	6	0.64
プロンプト (4)	5.6	5	6	0.24

表2 与えた影響の自動評価

プロンプト	平均	最小	最大	分散
プロンプト (1)	4.8	4	5	0.16
プロンプト (2)	4.4	4	5	0.24
プロンプト (3)	5.8	5	6	0.16
プロンプト (4)	5.8	5	6	0.16

価の結果を表1に示す。(2) 教員発話が児童発話に与えた影響に関して、生成されたアドバイスに対するLLMによる自動評価の結果を表2に示す。アドバイス生成用のプロンプト4種類それぞれの平均、最小、最大、分散を示している。推定影響を用いたプロンプト(3)と(4)によるアドバイスの平均評価が、推定影響を用いずにアドバイスを生成したプロンプト(1)と(2)に比べ、表1と表2の双方で上回っている。特に、表2の教員発話が与えた影響では、平均評価の差が1以上と大きくなっている。

LLMによる自動評価結果の最小値は4と高いが、同様に対話内容と無関係のアドバイスを評価した場合は低評価(2など)となった。つまり、本実験で生成されたアドバイスは評価基準から大きく逸脱していない。一方、高評価なアドバイス間の質的差異を明確にするには、プロンプトの改善が必要である。

また、LLMによる自動評価の例を図3に示す。図3では、(1)アドバイス生成用のプロンプト、(2)生成されたアドバイス、(3)生成された評価結果を示している。この例は、教員発話が児童発話に与えた影響を推定した結果を「発話影響ランク」として統合したプロンプト(3)によって、アドバイス生成を行っている。教員発話u1に対するアドバイスを生成しており、アドバイス内容は対話内容に関連した具体的なものとなっているため、LLMの評価である6は適切であると考えられる。

5 おわりに

本研究では、教員向けアドバイス生成及び、LLM-as-a-judgeを用いたアドバイスの評価を行った。小学校の授業における教員と児童の対話を分析し、その分析結果に基づいたアドバイス生成とLLMによるアドバイスの自動評価手法を提案した。具体的には、教員と児童の模倣モデルを構築し、その

```

## 指示
1. 児童の発話に影響をあまり与えられていない教員の発話 u1
   に対して、教員の発話が児童の発話に与えた影響が増加する
   方向に促すアドバイスをしてください。
## 条件 ...
- 発話 u1 が発話 u2-u6 に与えた影響度を 1-5 のランキング形
  式で評価した結果が発話影響ランクの表です。
## 発話影響ランク
Utterance | Speaker | Text | Impact Rank
---|---|---|---
u1 | Teacher | すごい。他にも決まり見つけてるんだけど、
   ちょっと1回先に行くね。ここでみんな問題解決してないよ
   ね? | 1
u2 | Student | 式。 | 4
u3 | Teacher | 最後、どこまで求めんといけんの? | 3
u4 | Student | 式。20。 | 5
u5 | Teacher | 式にして? | 2
u6 | Student | 周りの長さ。 | 1
## 出力形式 ...

### アドバイス
教員の発話が児童の発話に与える影響を増加させるためには、
児童に考えさせる質問を投げかけ、彼らの思考を深めるように
促すことが重要です。具体的には、...
1. **具体的な質問をする**：... 発話は、児童に対して具体的
   な考えを促すものではありません。例えば、「どんな決まりを
   見つけたのか、具体的に教えてくれる?」といった具体的な質
   問をすることで、児童が自分の考えを整理し、...
2. **児童の発言を引き出す**：... 「その決まりを使って、ど
   んな問題が解決できると思う?」といった形で、児童の発言を
   引き出すような質問をすることが効果的です。
3. **フィードバックを与える**：...

### 評価理由
- **指示に対する忠実性**：アドバイスは指示に従って、教員
  の発話 u1 に対するものとなっており、... 指示に忠実です。
- **対話内容とアドバイスの関連性**：アドバイスは、児童の
  発話を促すための具体的な質問を提案しており、対話の状
  況を踏まえたものとなっています。... 関連性があります。
- **アドバイスの適切さ**：アドバイスは具体的であり、教員
  が児童の発言を促進するための実践的な方法を提供していま
  す。具体的な質問やフィードバックの重要性を指摘しており、
  教員にとって有用なアドバイスです。
### 評価
6

```

図3 LLMによる自動評価の例(一部省略)

Attention機構を用いて教員発話に関する影響を定量的に推定した上で、その推定結果をLLMによるアドバイス生成に統合する手法を提案している。実験の結果、影響推定に基づいたアドバイスは、LLMによる自動評価において、推定影響を考慮しない場合に比べて高い評価を得る傾向が示唆された。これは、提案手法が教員の授業改善に資する質の高いアドバイス生成に寄与する可能性を示唆している。

今後の研究では、本研究において明らかとなった課題に取り組む必要がある。本研究で用いたデータセットは小規模であるため、アドバイス生成の精度向上に向けて、データの収集と構築が課題である。また、LLM-as-a-judgeによる自動評価の信頼性については、教育現場における実践的な効果を検証し、教員による主観的な評価との比較検討などが重要となる。さらに、LLMによる自動評価においては、ランキングによる相対評価の導入、教育学の知見を踏まえた評価用プロンプトの作成が必要である。

謝辞

本研究は JSPS 科研費 JP23K11378 の助成を受けたものです。

参考文献

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in neural information processing systems**, pp. 5998–6008, 2017.
- [2] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.
- [4] OpenAI et al. Gpt-4 technical report, 2023.
- [5] 文部科学省. 小学校学習指導要領 (平成 29 年告示) . https://www.mext.go.jp/content/1413522_001.pdf, 2016. 参照 2022-10-03.
- [6] 中央教育審議会. 教職生活の全体を通じた教員の資質能力の総合的な向上方策について (答申) , 2012.
- [7] 秋田喜代美. 変貌する教育学, 教師教育から教師の学習過程研究への転回—ミクロ教育実践研究への変貌—, pp. 45–75. 世織書房, 2009.
- [8] 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.
- [9] 保森智彦. 省察方法のデジタル化に関する一考察: 「主体的・対話的で深い学び」の観点から, pp. 3–11. 学習開発学研究, No. 14. 広島大学大学院人間社会科学部, 2022.
- [10] yuchen Wang, 大井翔, 松村耕平, 野間春生. 新任教員の授業力向上のための授業振り返りシステムに関する研究. 情報処理学会インタラクシオン, pp. 753–757, 2021.
- [11] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of llm-as-a-judge, 2025.
- [12] Cheng-Han Chiang, Wei-Chih Chen, Chun-Yi Kuan, Chienchou Yang, and Hung-yi Lee. Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 2489–2513, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
- [13] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In **International Conference on Learning Representations**, 2022.
- [14] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.
- [15] Tianyu Zhao and Kei Sawada. rinna/your-7b-chat.
- [16] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable AI for natural language processing. In **Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing**, pp. 447–459, Suzhou, China, December 2020. Association for Computational Linguistics.
- [17] Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural nlp: A survey. **ACM Comput. Surv.**, Vol. 55, No. 8, dec 2022.
- [18] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4443–4458, Online, July 2020. Association for Computational Linguistics.
- [19] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shriti Bhoosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kam-badur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

A アドバイス用プロンプト

教員発話が児童発話から受けた影響に対するアドバイス用プロンプトを図4と5に示す。図4はアドバイス用プロンプト(1)の例であり、推定影響を用いないため、{対話の表}部分に対話を結合している。図5はアドバイス用プロンプト(4)の例であり、推定影響を用いるため、{影響推定結果の表}部分を影響推定結果に置換している。

```
## 指示
1. 教員の発話 u6 に対して、アドバイスをしてください。
## 条件
- 以下の対話は小学校の授業で行われた教員と児童の対話の一部です。
## 対話
{対話の表}
## 出力形式
### アドバイス
{アドバイス}
```

図4 アドバイス用プロンプト(1)

```
## 指示
1. 発話 u6 が u1-u5 の発話から受けるべき理想的な発話影響ランクを教えてください。
2. 現在の発話影響ランクと理想的な発話影響ランクの違いに基づいて、u6 の発話はどのようにすべきかアドバイスをしてください。
## 条件
- 以下の対話は小学校の授業で行われた教員と児童の対話の一部です。
- 発話 u6 が発話 u1-u5 から受けた影響度を 1-5 のランキング形式で評価した結果が発話影響ランクの表です。
- 1 が最も影響度が高い発話、5 が最も影響度が低い発話としています。
## 発話影響ランク
{影響推定結果の表}
## 出力形式
### 理想的な発話影響ランク
{理想的な発話影響ランクの表}
### アドバイス
{アドバイス}
```

図5 アドバイス用プロンプト(4)

また、教員発話が児童発話に与えた影響に対するアドバイス用プロンプトを図6と7に示す。図6はアドバイス用プロンプト(1)、図7はアドバイス用プロンプト(2)の例である。

```
## 指示
1. 教員の発話 u1 に対して、アドバイスをしてください。
## 条件
- 以下の対話は小学校の授業で行われた教員と児童の対話の一部です。
## 対話
{対話の表}
## 出力形式
### アドバイス
{アドバイス}
```

図6 アドバイス用プロンプト(1)

```
## 指示
1. 児童の発話に影響をあまり与えられていない教員の発話 u1 に対して、教員の発話が児童の発話に与えた影響が増加する方向に促すアドバイスをしてください。
## 条件
- 以下の対話は小学校の授業で行われた教員と児童の対話の一部です。
## 対話
{対話の表}
## 出力形式
### アドバイス
{アドバイス}
```

図7 アドバイス用プロンプト(2)

B 評価用プロンプト

本研究で用いた LLM による自動評価の評価用プロンプトを図8に示す。{アドバイス生成用のプロンプトと生成アドバイス}の部分を(1)アドバイス生成用のプロンプトと生成されたアドバイスに置換している。

```
{アドバイス生成用のプロンプトと生成アドバイス}
## 指示
1. 評価基準に基づいて、教員の発話に対するアドバイスを評価してください。
2. 批判的な評価も可能な限り含むようにしてください。
3. 評価基準と出力形式を厳守してください。
## 条件
- この対話は小学校の授業で行われた教員と児童の対話の一部です。
- アドバイスに対する総合評価を 1-6 の 6 段階で評価してください。
- 最も高い評価は 6、最も低い評価は 1 とします。
- 評価理由を評価基準に基づいて教えてください。
## 評価基準
- 指示に対する忠実性
  - 指示の内容に合致しているか
  - アドバイスの対象発話以外に対するアドバイスを含んでいないか
- 対話内容とアドバイスの関連性
  - 対話の状況を踏まえたアドバイスになっていないか
- アドバイスの適切さ
  - 教員に有用なアドバイスでないか
  - アドバイスが具体的でないか
## 出力形式
### 評価理由
{評価理由}
### 評価
{評価値}
```

図8 LLM による自動評価における評価用プロンプト