論文: https://aclanthology.org/2026.acl-long.875/

ACL 2026

Your Students Don’t Use LLMs Like You Wish They Did

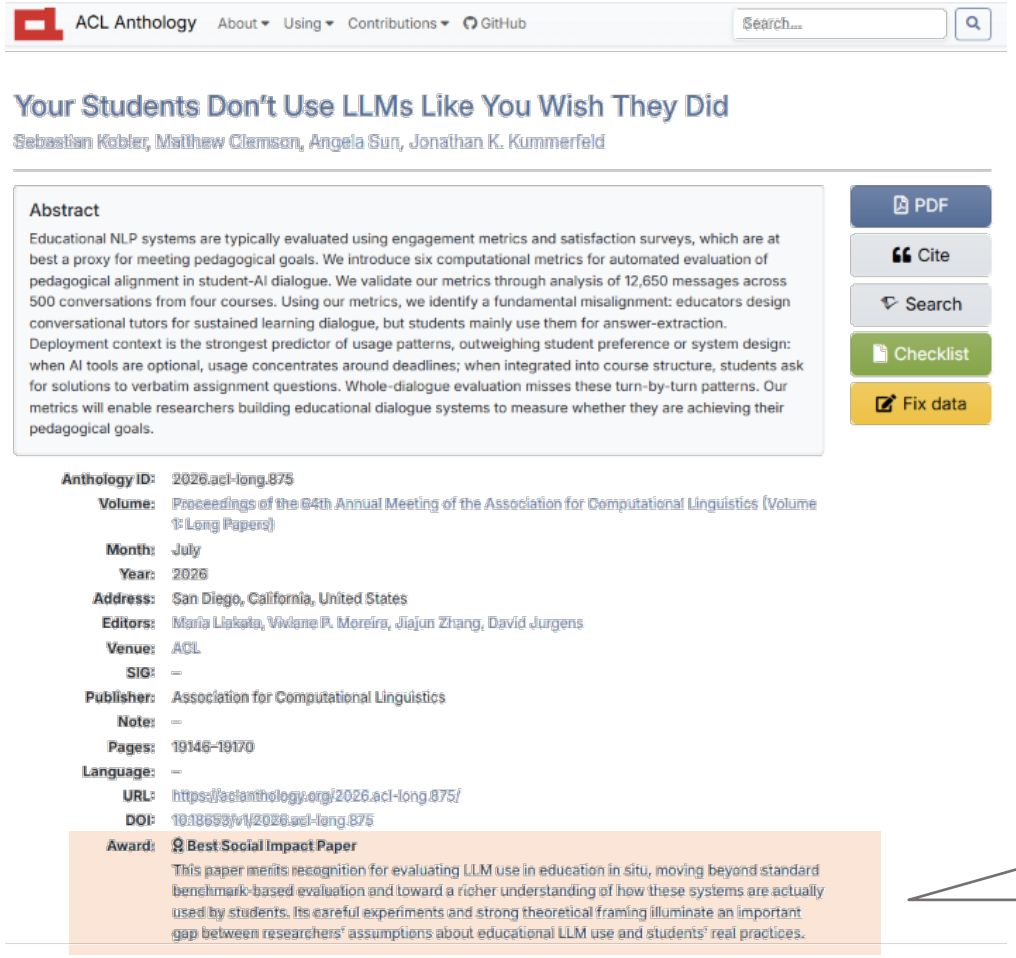
Sebastian Kobler, Matthew Clemson, Angela Sun, Jonathan K. Kummerfeld
The University of Sydney

紹介者: 笹野遼平（名大）

一部の図は、論文または著者の発表スライドより引用

Best Social Impact Paperを受賞

(社会に大きなポジティブな影響をもたらす可能性のある論文に与えられる賞)



ACL Anthology About Using Contributions GitHub Search...

Your Students Don't Use LLMs Like You Wish They Did

Sebastian Kobler, Matthew Clemson, Angela Sun, Jonathan K. Kummerfeld

Abstract

Educational NLP systems are typically evaluated using engagement metrics and satisfaction surveys, which are at best a proxy for meeting pedagogical goals. We introduce six computational metrics for automated evaluation of pedagogical alignment in student-AI dialogue. We validate our metrics through analysis of 12,650 messages across 500 conversations from four courses. Using our metrics, we identify a fundamental misalignment: educators design conversational tutors for sustained learning dialogue, but students mainly use them for answer-extraction. Deployment context is the strongest predictor of usage patterns, outweighing student preference or system design: when AI tools are optional, usage concentrates around deadlines; when integrated into course structure, students ask for solutions to verbatim assignment questions. Whole-dialogue evaluation misses these turn-by-turn patterns. Our metrics will enable researchers building educational dialogue systems to measure whether they are achieving their pedagogical goals.

PDF Cite Search Checklist Fix data

Anthology ID: 2026.acl-long.875
Volume: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)
Month: July
Year: 2026
Address: San Diego, California, United States
Editors: Maria Lakata, Vivian P. Moreira, Jiajun Zhang, David Jurgens
Venue: ACL
SIG: –
Publisher: Association for Computational Linguistics
Note: –
Pages: 19146–19170
Language: –
URL: https://aclanthology.org/2026.acl-long.875/
DOI: 10.18653/v1/2026.acl-long.875
Award: Best Social Impact Paper
This paper merits recognition for evaluating LLM use in education in situ, moving beyond standard benchmark-based evaluation and toward a richer understanding of how these systems are actually used by students. Its careful experiments and strong theoretical framing illuminate an important gap between researchers' assumptions about educational LLM use and students' real practices.

This paper merits recognition for evaluating LLM use in education in situ, moving beyond standard benchmark-based evaluation and toward a richer understanding of how these systems are actually used by students. Its careful experiments and strong theoretical framing illuminate an important gap between researchers' assumptions about educational LLM use and students' real practices.

本論文は、標準的なベンチマークに基づく評価の枠組みを超え、教育現場におけるLLMの利用を実際に評価し、学生による実際の利用実態のより深い理解へと踏み込んだ点で高く評価されるべきである。綿密な実験と強固な理論的枠組みを通じて、教育におけるLLM利用に関する研究者の想定と、学生の実際の利用行動との間に存在する重要な乖離が明らかにされている。

- 他にも Honourable Mention (top 1.5-2.5% of accepted papers) / SAC Highlight との記載（https://jkk.name/pubs/）

論文の背景と概要

- 最近のACLの教育用LLM論文10件中９件の評価は満足度調査と自己申告による学習効果に依拠しているがユーザ自己評価の信頼性に疑問がある
- 教育用LLMは本当に教育者が期待したように使われているのか？

- 教育用LLMが教育目的に沿って使われているかを測る6つの指標を提案
- 5つの教育AI利用データから各100対話、計12,650メッセージを分析

- 教育者の期待と学生の実際の利用には大きなギャップ
- 学生は持続的な学習対話よりも答えを得るためにLLMを利用する傾向

学生–AI対話を評価する6つの行動指標

1. **LOI**: Learning Orientation Index
   - 学生の発話が、理解・探索志向か、答えを得ることを目的としているか
   - 各学生発話について、Exploratory ↔ Answer-seeking の程度をLLMで0–1スコアリングし、会話単位で集約
2. **CES**: Conversational Engagement Score
   - AIとのやり取りが、どの程度継続的・対話的か
   - 会話の長さ(0.4)、follow-up率(0.25)、過去の文脈への参照(0.20)、AI応答を認識・受容した割合(0.15)をスコア化し重み付き平均（後者3つはLLMを利用）
3. **SRS**: Scaffolding Resistance Score
   - AIからのヒントや問い返しに対して、学生がどの程度抵抗するか
   - Scaffoldingに対する学生の反応をLLMで Accept / Mixed / Bypass / Resist に分類・スコア化し、会話内で集約

4. **ADR**: Assignment Dependency Ratio

- AI利用が授業課題の直接的な処理にどの程度依存しているか
- **LLM版**：問題文のコピペか(0.4)、問題順通りか(0.3)、answer-seeking(0.2)、urgency(0.1)を重み付き平均
- **Rule-based版**：課題文でよく使われる命令表現、問題番号・設問構造、assignment関連語などの出現を検出して算出

5. **CMI**: Crisis Mode Indicator

- 試験・締切などのhigh-pressure periodに、学生の利用行動が通常時からどの程度crisis-likeに変化するか
- 同一学生のpanic表現、質問の直接性（LLMで算出）、深夜利用、single-exchange、engagement低下の変化を統合

6. **UCI**: Usage Concentration Index

- AI利用が特定期間にどの程度集中しているか
- 週ごとの利用量から Gini coefficient、peak-to-average ratio、temporal clustering を計算して統合（LLM不使用）

6つの行動指標のまとめ

Metric	高い場合	どちらと言えば
LOI: Learning Orientation Index	探索・理解志向（⇔答え志向）	↑ 高い方が◎
CES: Conversational Engagement Score	対話的・継続的	↑ 高い方が◎
SRS: Scaffolding Resistance Score	Scaffolding（足場）に強く抵抗	↓ 低い方が◎
ADR: Assignment Dependency Ratio	利用が授業課題に直結	↓ 低い方が◎
CMI: Crisis Mode Indicator	試験時期等に行動が変化	↓ 低い方が◎
UCI: Usage Concentration Index	特定時期に利用が集中	↓ 低い方が◎

Human validation on 248会話

（CMI、UCIは人間が会話単位で評価できないので対象外）

	LOI			CES			SRS			ADR		
	r	Exact	±1	r	Exact	±1	r	Exact	±1	r	Exact	±1
GPT 4.1-mini Turn	0.62	66%	97%	0.42	50%	95%	0.64	56%	88%	–	–	–
GPT 4.1-mini Whole	0.33	16%	44%	0.21	10%	58%	0.25	39%	75%	0.22	48%	65%
GPT 5 Turn	**0.72**	72%	99%	**0.59**	44%	92%	**0.67**	63%	91%	–	–	–
GPT 5 Whole	0.47	19%	63%	0.46	34%	81%	0.49	47%	79%	0.31	38%	54%
Rule-based	–	–	–	–	–	–	–	–	–	0.35	82%	85%
Human-Human†	0.58	59%	100%	0.67	55%	88%	0.64	60%	85%	0.65	75%	82%

Table 1: Model-human and human-human agreement. r = Pearson correlation (top 5 rows), and weighted Cohen's kappa (bottom row). ±1 = off-by-one accuracy. † row is based on 100 conversations, while the others are based on 248. GPT-5 turn-by-turn achieves highest correlations and consistently outperforms whole-dialogue analysis.

- Turn-by-turn評価の方がWhole-dialogue評価より明確に良い
- GPT-5はGPT-4.1-miniより概ね高性能
- LOI・CES・SRSは比較的うまく自動評価できるが、ADRは難しい

5つの教育AI利用データ

1. DrMattTabolism
   - シドニー大学の生化学・分子生物学科目（MEDS2003: Biochemistry and Molecular Biology）の前半で、代謝学（metabolism）の学習支援に使用されたoptional AI tutorの利用データ
   - 問いかけやヒントを中心としつつ必要に応じて答えも提供
   - 本論文の第2著者Matthew Clemsonらによるデータ
2. DrNucleicAlice
   - DrMattTabolismと同じ、シドニー大学のMEDS2003の後半で、分子生物学の学習支援に使用されたoptional AI tutorの利用データ
   - DrMattTabolismより厳格なソクラテス式対話（答えを直接教えるのではなく、質問を重ねることで学習者自身に考えさせる）を採用し、まず質問・ヒントによって学生自身に考えさせる

3. MEDS2004
   - シドニー大学の医科学系の授業で、過去の試験問題を用いた試験対策・自己テストを支援するAI toolの利用データ
   - 問題提示 → 学生の回答 → feedbackという構造化された対話を実施

4. OLiMent
   - シドニー大学の入門データサイエンス科目で使用された、学生自身の学習進捗・理解度の振り返りを支援するAIシステムの利用データ
   - Open Learner Modelについて学生と対話することで、各learning goalに対する自己評価（self-assessment）を促すことが目的
   - 本論文とは別の研究チーム（Yuan et al., 2025）が開発

---

5. StudyChat
   - University of Massachusetts AmherstのAI科目で収集されたコースワーク全般について利用できるAI assistant（Unrestricted AI）との対話データ
   - McNicholsらが構築・公開した外部データセット

利用データのまとめ

Dataset	disciplines（学問分野）	Interaction paradigm	利用形態	対話・システム設計	主な想定用途	公開状況	開講大学
DrMattTabolism	Biochemistry / Molecular Biology	**Pedagogical Support** (教育支援)	Optional	Flexible scaffolding	代謝学の学習支援	非公開	シドニー大学
DrNucleicAlice	Biochemistry / Molecular Biology	**Pedagogical Support**	Optional	Strict Socratic	分子生物学の学習支援	非公開	シドニー大学
MEDS2004	Medical Sciences	**Pedagogical Support**	Optional	Structured Q&A	過去問による試験練習	非公開	シドニー大学
OLiMent	Data Science	**Pedagogical Support**	Optional	Open Learner Modelを用いたreflection	学習進捗の自己評価・振り返り	非公開	シドニー大学
StudyChat	Computer Science / AI	**Unrestricted AI**	Integrated	General-purpose conversational AI	Coursework全般	公開	UMass Amherst

Three disciplines and two interaction paradigms

実験結果

1. Engagementが高くても learning-orientedとは限らない

				ADR			
Dataset	LOI	CES	SRS	Rule	LLM	CMI	UCI
DrMattTabolism	33	35	16	3	41	20	64
DrNucleicAlice	38	72	33	2	39	13	67
MEDS2004	13	67	27	2	72	14	74
OLiMent	34	70	16	5	26	18	67
StudyChat	15	71	22	12	58	-	39

Table 2: Summary of pedagogical misalignment metrics by dataset. All values are scaled 0-1 and shown as percentages (%). Higher values indicate: CES = conversational engagement, SRS = scaffolding avoidance, ADR = detection of assignment content, CMI/UCI = crisis-driven usage. For LOI, lower values indicate answer-seeking behaviour.

Platform	AS	E	M
Constrained (n=400)	66.5	15.5	18.0
StudyChat (n=100)	92.0	2.0	6.0

Table 3: Learning Orientation Distribution by Platform Type. Answer Seeking, Exploratory and Mixed. All values shown are percentages (%).

- 制限のないプラットフォームであるStudyChatの特徴
  - 対話性・継続性の評価（CES）は高かったもののLOIは低い
  - 探索的な会話は全体の2.0%のみ
  - 答えを求める会話が92.0%

2.より厳格なscaffoldingでは学生の抵抗も大きい

				ADR			
Dataset	LOI	CES	SRS	Rule	LLM	CMI	UCI
DrMattTabolism	33	35	16	3	41	20	64
DrNucleicAlice	38	72	33	2	39	13	67
MEDS2004	13	67	27	2	72	14	74
OLiMent	34	70	16	5	26	18	67
StudyChat	15	71	22	12	58	-	39

Table 2: Summary of pedagogical misalignment metrics by dataset. All values are scaled 0-1 and shown as percentages (%). Higher values indicate: CES = conversational engagement, SRS = scaffolding avoidance, ADR = detection of assignment content, CMI/UCI = crisis-driven usage. For LOI, lower values indicate answer-seeking behaviour.

- DrMattTabolismとDrNucleicAliceは講義の前後半なので学生が重複（時期は異なる）

- 厳格なソクラテス式対話
  → 会話は長く・対話的になる
  → scaffoldingへの抵抗も約2倍

3. Optional AIは試験前のcrisis toolになる

				ADR			
Dataset	LOI	CES	SRS	Rule	LLM	CMI	UCI
DrMattTabolism	33	35	16	3	41	20	64
DrNucleicAlice	38	72	33	2	39	13	67
MEDS2004	13	67	27	2	72	14	74
OLiMent	34	70	16	5	26	18	67
StudyChat	15	71	22	12	58	-	39

Table 2: Summary of pedagogical misalignment metrics by dataset. All values are scaled 0-1 and shown as percentages (%). Higher values indicate: CES = conversational engagement, SRS = scaffolding avoidance, ADR = detection of assignment content, CMI/UCI = crisis-driven usage. For LOI, lower values indicate answer-seeking behaviour.

Limitations (※は論文中で言及なし)

1. 行動指標 ≠ 実際の学習効果
   - learning-orientedな行動が実際に学習成果を向上させるかは検証していない
2. Deploymentの効果を因果的に分離できない ※
   - StudyChatは唯一の integrated / unrestricted な環境
   - しかし同時に、大学・授業・学生・システム設計・課題なども異なる
   - 一方で“deployment strategy is the strongest predictor of usage patterns”と主張
3. 提案指標の妥当性に課題
   - 6指標の構成要素や重みは、主に著者らの教育経験に基づく
   - 重みを変えた場合のsensitivity analysisなどは行われていない ※
4. Generalizability / Reproducibilityが限定的
   - 対象は5 datasets・4 coursesで、すべてSTEM系
   - 500会話のうち400会話はprivacy上の理由で非公開

まとめ

- 教育用LLMが教育目的に沿って使われているかを測る6つの行動指標を提案
- 実際の講義で収集された500対話・12,650メッセージを分析

- 学生は高いengagementを示していてもlearning-orientedとは限らない
- 教育者が意図する「持続的な学習対話」と、学生の「答えを得たい」という利用目的の間に大きな乖離

感想

- 実際の授業で収集されたデータを分析している価値は高い
- 本文中も含めLimitationsがしっかり書かれているのは好印象
- 指標の定義、特に重みづけは経験的な側面が強い
- 再現性・一般化可能性、および一部の因果的な解釈には課題が残る
  - 性質的に仕方ない面もあるが4/5のデータは著者らの大学のデータで非公開
  - 講義ごとにAIの使い方が異なるので個別の指標の信頼性は限定的
  - 条件が十分にコントロールされていないのでデータ間の差の信頼性は限定的
- 教育系AI分野(e.g., AIED, EDM)における新規性は不明
  - 学生によるLLMの実利用を扱った近年の関連研究との位置づけは十分ではない可能性
    - **Brender et al., 2024.** Who's Helping Who? When Students Use ChatGPT to Engage in Practice Lab Sessions (AIED 2024)
    - **Ghimire & Edwards, 2024.** Coding with AI: How Are Tools Like ChatGPT Being Used by Students in Foundational Programming Courses (AIED 2024)
    - **Liu et al., 2025.** Understanding Student Engagement with Large Language Model-Powered Course Assistants (AIED 2025)