Field experiences and reflections on using LLMs to generate comprehensive lecture metadata

Sumit Asthana, Taimoor Arif, Dr. Kevyn Collins Thompson

University of Michigan

Towards equitable educational outcomes

Understanding students' conceptual knowledge gaps enables educators to personalize their instruction for equitable learning outcomes [1]. However, with increasing classroom sizes and more diverse populations of students in universities and MOOCs, addressing challenges faced by individual students is getting increasingly difficult for educators.

Challenges

Providing individual support to students requires understanding their background, and having an understanding of how to adjust curriculum to address student needs. In-class assessments used by instructors provide an aggregate understanding of the progress of students. While students can benefit from more personalized feedback, instructors do not have the time to address each student's individual needs.

Rich extraction of lecture representations

One pivotal aspect of leveraging LLMs in education is the extraction of rich metadata from lectures. By segmenting lectures into topically coherent "moments," we create structured natural language data that encapsulates key definitions, examples, and procedural knowledge. These "moments" serve as building blocks, allowing for a nuanced understanding of course content, laying the foundation for further exploration and application. From the moments we extract

- **Key concepts** discussed in the moment.
- **Key definitions** that the instructor provides in the moment. These are definitions of concepts that the instructor introduces in the lecture moment.
- **Key examples** that the instructor provides in the moment.
- **Procedural knowledge** captures knowledge related to creating the artifact associated with the concept, using the concept, or applying that concept.
- **Key questions** that test the relevant concepts discussed in the moment.

The moments are derived from high-quality transcripts of the lecture audio. Some challenges that we encountered with using only transcripts is that they are insufficient to capture references made to slides in the classroom. To improve the metadata, we plan to incorporate video data in the extraction process as well.

Lecture pipeline

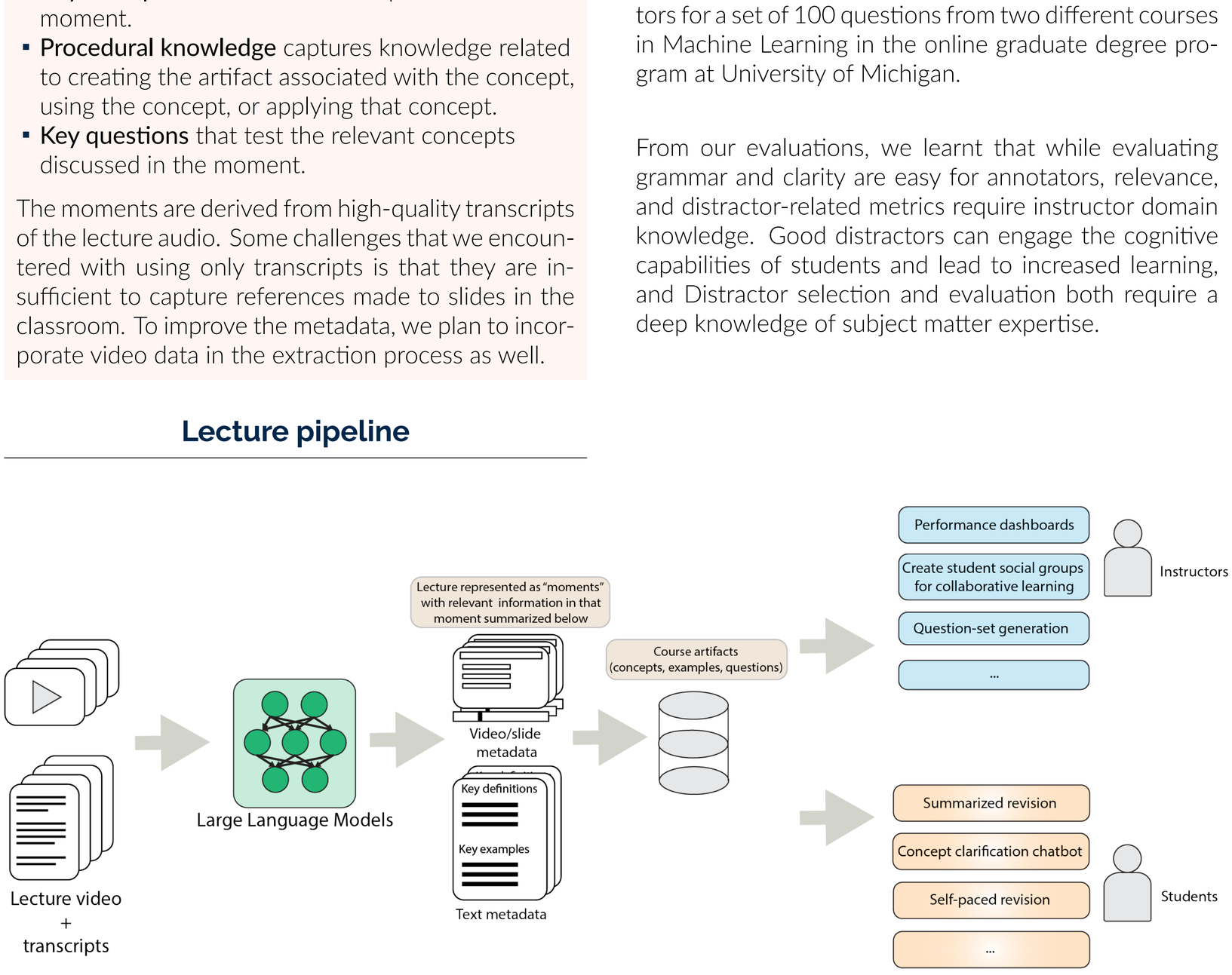

Figure 1. Lecture data extraction pipeline

Evaluating LLM capabilities

LLMs exhibit capabilities such as summarizing relevant lecture parts, ability to generate questions that test lecture concepts, ability to reason about student's mistakes, developing concept hierarchies.

To reliably integrate LLM capabilities in educational support tools, we need to evaluate their alignment with how instructors approach these educational activities. In this work, we provide a preliminary evaluation of their ability to generate questions to test student knowledge.

Question assessment statistics

Question assessment dimensions

Table 1. Accuracy and IRR Scores for question evaluation dimensions

Metric	Accuracy	κ score
Relevance	92%	0.89
Grammar	99%	0.99
Difficulty	-	0.60
Clarity	97%	0.94
Contextual Non-Specificity	90%	0.92
Question-Option Disjoint	-	0.90
Distractor Homogeneity	80%	0.77
Distractor Plausibility	65%	0.71

We evaluate the questions along several dimensions [3] such as plausibility, distractor quality, relevance. Two human raters, graduate students in data science, used an evaluation rubric to rate 100 questions drawn from two different courses. Table 1 summarizes the accuracy of GPT generated questions and IRR scores of the annotators for a set of 100 questions from two different courses in Machine Learning in the online graduate degree program at University of Michigan.

From our evaluations, we learnt that while evaluating grammar and clarity are easy for annotators, relevance, and distractor-related metrics require instructor domain knowledge. Good distractors can engage the cognitive capabilities of students and lead to increased learning, and Distractor selection and evaluation both require a deep knowledge of subject matter expertise.

Evaluating questions generated by LLMs

Comparison of manual and automated evaluations [3]

Table 2. Summary statistics for rule-based and LLM evaluation

Statistic	Rule based evaluation score	LLM-based evaluation score
Passes all metrics (%)	18	22
Passes at least half metrics (%)	100	91
Fails one or no metrics (%)	50	55
Fails two or fewer metrics (%)	79	69
Average IWF (failures) per MCQ	1.61	1.86

Future directions

LLMs open up exciting future directions for developing tools that provide students necessary support for their learning journeys [2].

- **Rich LLM-derived representations of learners and content**. Even limited additional metadata for educational content can provide significant new downstream prediction and modeling opportunities if chosen wisely.
- **Hybrid models for curriculum optimization**. In prototyping an adaptive study guide that used our generated questions, we became convinced of the need for new hybrid practice frameworks that combine existing scientifically validated statistical models of learning and memory with the representation and inference power of LLMs.
- **Education & theory of mind**. As we noted above, there is a need to assess the potential for even limited domain-specific theory-of-mind abilities in a LLM's educational interactions.
- **Impact of LLM recommendations on instructor decision-making**. Due to their potential for unrestricted and convincing text generation, LLMs can significantly impact human decision-making. Understanding whether LLM support augments instructor capabilities or restricts it is important to assess their usefulness.

References

[1] Lijia Chen, Pingping Chen, and Zhijian Lin. Artificial intelligence in education: A review. *Ieee Access*, 8:75264–75278, 2020.

[2] David R Krathwohl. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218, 2002.

[3] Steven Moore, Huy A. Nguyen, Tianying Chen, and John Stamper. Assessing the quality of multiple-choice questions using gpt-4 and rule-based methods. *arXiv preprint arXiv:2307.08161*, 2023.

Acknowledgements

This research was sponsored in part by a grant from the Michigan Institute for Data Science (MIDAS), with additional support from the University of Michigan School of Information.