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

Sumit Asthana
Department of Computer Science
University of Michigan,
Ann Arbor, MI
asumit@umich.edu

Taimoor Arif
School of Information
University of Michigan,
Ann Arbor, MI
taimoora@umich.edu

Kevyn Collins Thompson
School of Information
University of Michigan,
Ann Arbor, MI
kevynct@umich.edu

Abstract

We describe an ongoing initiative to incorporate generative AI for online higher education classes at a large public U.S. university. Our specific online-only class setting poses special challenges: the technical backgrounds of incoming learners tend to vary widely across domains, making the potential for personalized adaptation especially compelling; the majority of instruction is via pre-recorded video, with some live office hours support and forum discussions, making it critical to promote additional effective engagement and self-assessment. Toward these goals, we describe what we have learned and been thinking about in our early explorations, with an initial framework starting from using generative AI to create questions and other rich metadata from lecture video that can support instructor- and student-facing affordances for learning and discovery.

1 Introduction

Understanding students' conceptual knowledge gaps enables educators to personalize their instruction for equitable learning outcomes [7]. However, with increasing classroom sizes and more diverse populations of students in universities and MOOCs [37], addressing challenges faced by individual students is getting increasingly difficult to handle at scale for educators [22].

While AI research has developed numerous tools to support instructors such as automatically generating practice question sets [34], personalized question set recommendations to students [29], visualization of student performance, and generating lecture summaries [14], each such tool requires careful engineering efforts and domain knowledge to collect and store data about courses as a basis for developing student- or instructor-facing applications [5, 49].

Generative AI systems such as Large Language Models (LLMs) allow easy curation of data from existing course materials (text, audio and video) through simple natural language prompts [21]. However, while applications of LLMs are rapidly growing due to their ease of on-demand personalized content generation, we lack an understanding of how their use impacts various stages of current educational scenarios [49]. Mapping out where LLMs can provide value in existing research in education from Computer Science, Learning Analytics, and Data science can help us identify concrete directions to integrate their capabilities for improving educational outcomes [9].

In this paper we describe ongoing field experiences towards incorporating generative AI for online instruction settings in higher education at a large public U.S. university. Our specific online-only learning setting poses special challenges: the technical backgrounds of incoming learners tend to vary widely across domains, making the potential for personalized adaptation especially compelling; the majority of instruction is via pre-recorded video, with some live office hours support and forum discussions, making it critical to promote effective engagement and self-assessment. Toward these

NeurIPS'23 Workshop on Generative AI for Education (GAIED).

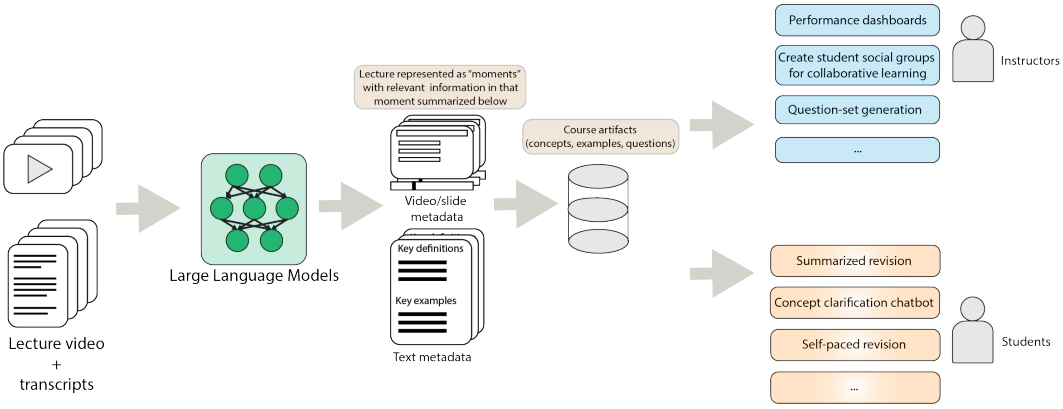

Figure 1: An overview of our data extracting from lectures using LLMs to create rich representations of courses. We plan to use these representations to support several end use cases for both instructors and enabling self-learning for students.

goals, we describe what we have learned and been thinking about in our early explorations and initial framework for using generative AI to curate rich metadata from lecture video that can support instructor- and student-facing affordances for learning and discovery.

We describe how we curated structured natural language data using LLMs for online degree courses, how we evaluated the quality of the data, and the potential directions to integrate the data for building applications that support student learning outcomes. Based on our work with this data curated from lectures (transcripts and videos) using LLMs we highlight immediate applications (e.g., developing question sets) [34], and also highlight what we believe are important new research directions, such as experiments to evaluate LLM theory of mind [1] capabilities in assessing student's concept representations.

2 Holistic view of course data to support educational objectives

Any typical learning system [30] requires data about the concepts taught in the course. In our system, for each lecture's transcript in the course, we first segment it into segments using GPT-4. We used simple but effective prompts such as "Split the given lecture into 4-5 segments. Each segment should be topically coherent and the main topics across segments should be different", along with the lecture transcript as input. We use insights from NLP research that segment boundaries split passages into units that are topically coherent [17]. We call these segments "moments" of a lecture. For each such "moment" we extract its summary, key definitions, key examples, and types of procedural vs factual knowledge described in that moment. We used the prompt "List 5-7 key definitions, key examples, procedural and factual knowledge described in the moment" and provided the associated moment text. Next, for each moment, we also extract a set of questions representing that moment. Thus, for each lecture, we have a set of key definitions, key examples, summary, and questions by the topics covered in the lecture. Figure 1 provides a high level overview of this data extraction and some potential applications it can enable.

For the moment, the lecture content being analyzed comes from high-quality transcripts of the lecture audio, but we intend to extend the system to also process video images in preparation for exploiting multi-modal LLMs. Our system also uses GPT-4 to extract metadata from other course resources, such as learning objectives from the course syllabus, so that question generation can be applied to ask about any relevant learning resources in the system. To date we have generated detailed meta-data for 11 courses in our curriculum, containing thousands of topical questions and metadata about key specific learning moments in a class video.

3 Generating rich course representations to support educational objectives

Extensive research in learning analytics [43] has developed and investigated tools to support instructors in improving educational outcomes. Examples of such tools are test-directed learning [40] or a

tool to provide natural language feedback to student essays [46]. However, an important challenge in using analytics derived from course activities is interpreting them to derive actionable insights or give actionable feedback to students [19]. In the sections below we describe several areas where, based on our field experiences so far, LLMs display increasingly impressive ability to generate output in natural language descriptions, often with clear explanations, which can enhance communication of insights for instructors and feedback for students.

3.1 Concept-level abstraction of lectures

Knowledge tracing [15] is one of the most widely used methods to track student's progress on concepts covered in the course. Knowledge tracing models students' progress on concepts as as a Markov decision process with transitions on correct and incorrect outcomes on the concept. Extracting relevant concepts from courses is the most challenging aspect, and data mining approaches rely on extracting important terms from textbooks [6] or online lectures and apply dimensionality-reduction techniques such as matrix factorization [44]. However, it is difficult to control the granularity of concepts without direct supervision, and extraction does yield high quality results on lecture dialogues [42].

LLMs can identify and represent important concepts in natural language even from challenging sources like dialogues [2]. Instructors can control the granularity of concepts by ensuring alignment with an existing taxonomy or course syllabus. In our own data extraction of concepts from lecture transcripts using GPT-4, using simple prompts like "Extract the key concepts from the lecture that are present in the pre-defined course syllabus" has yielded reasonable success. The meta-data generated is highly relevant to the lecture material. For example, for a lecture covering Random Forests, the list of key concepts given was: Ensembles, Bagging, Boosting, Random forest Decision trees, Overfitting, Supervised Learning, and Regression. Most of these concepts are highly relevant to Random forests and are used in teaching it as well. For another lecture on KNN, the topics mapped were: KNN, Supervised Learning, Classification, Feature Engineering, and Majority vote - all of which are directly related to KNN. This shows that our method of mapping topics onto lectures is yielding fairly accurate results.

3.2 Metadata generation from lectures

Other than questions and topic mappings, we also generate additional metadata from our lectures – Lecture Summary, Procedural knowledge (the "How to" descriptions), Key Definitions, and Key Examples. Following is a description of the metadata around the concept "Ensemble model".

- **Procedural Knowledge**: This metadata captures knowledge related to creating the artifact associated with the concept, using the concept, or applying that concept. For example, following is the procedural knowledge associated with how to create an ensemble model: "An ensemble model is created by combining multiple individual learning models to produce an aggregate model that is more powerful than any of its individual learning models alone. This is effective because different learning models, although each of them might perform well individually, they'll tend to make different kinds of mistakes on a data set. Typically this happens because each individual model might overfit to a different part of the data. By combining different individual models into an ensemble, we can average out their individual mistakes to reduce the risk of overfitting while maintaining strong prediction performance".
- **Key Definitions**. Key definitions capture the salient description of the concept as it would occur in an encyclopedia or a course material. Following are example definitions for the concept "Ensemble" and "Overfitting".
  - Ensembles: A method in machine learning that involves creating learning models by combining multiple individual learning models to produce an aggregate model that is more powerful than any of its individual learning models alone.
  - Overfitting: A modeling error in machine learning that occurs when a function is too closely fit to a limited set of data points.
- **Key Examples**. Key examples describe the instantiation of the concepts, so that students can concretely understand the concept. For example GPT provides the following description for "Random Forests", which it provides as an example of the ensemble idea applied to decision trees – "They are widely used in practice and achieve very good results on a wide variety of problems. Random forests can be used as classifiers via the scikit learn random

forest classifier class or for regression using the random forest regressor class both in the sklearn ensemble module. The use of random forests helps to overcome the disadvantage of using a single decision tree, which is prone to overfitting the training data".

The purpose of this metadata is to supplement the questions as a study guide which the students can use to review the lecture in a way that will help them in attempting the questions in our database.

3.3 Synthesizing questions to test concepts

Generating questions that tests student knowledge on concepts is a time-intensive process that requires instructor domain knowledge and training. Methods to support instructors by automatically generating questions include using templates or ML models trained on existing questions. However, many approaches still generate questions starting with "what", fill in the blanks and controlling difficulty has been challenging [29]. LLMs can ease generation of questions to test concepts but we need robust evaluations of the quality of such questions [10] along dimensions like clarity, informativeness, and distractor quality.

As part of our field study preparation we systematically assessed the quality of AI-generated questions (using GPT-4) along quality dimensions based on recent question generation research (e.g. [33]). Two human raters, graduate students in data science, used an evaluation rubric to rate 100 questions drawn from two different courses. The raters assigned a binary label to the questions on the dimensions – relevance (92%), grammar (99%), clarity (97%), answerability (97%), distractor plausibility (65%), distractor homogeneity (80%), Difficulty, Contextual Specificity (10%) and Question-option disjoint. The percentages indicate the percentage of questions that the raters rated as 1 for the respective dimension. For difficulty of questions, the raters assigned the levels "Beginner", "Intermediate" and "Advanced". For the question-option disjoint metric, a lower score translates to a higher question quality since a question tagged "1" would mean there is an inconsistency between the question statements and the options. Similarily, a "1" for contextual specificity indicates that the question requires the context of the lecture to answer correctly (e.g., reference to a lecture figure). Thus, low percentage of contextually specific questions is better.

We plan to vet all questions by human experts before deployment in any intelligent educational tool, but for our initial offline experiments, the relatively high scores of the generated questions on the dimensions suggests that most candidates could be retained, resulting in approximately 500 useful generated questions per 4-week course, tagged with one or more key concepts/skills, that covered all key moments of all lectures in multiple ways. We highlight that the metrics pertaining to the distractors (distractor plausibility and distractor homogeneity) exhibited lower scores than other metrics. To this end, we are working on a method to replace the distractors that are not up to the required standard, which will greatly improve the overall quality of the questions. We also used Cohen's Kappa metric [32] to calculate the inter-rater agreement between the human annotators. Table 1 summarizes the IRR agreement scores. We performed the evaluations for the questions generated for Machine Learning/Data Science courses – 1) Supervised Learning and 2) Unsupervised Learning. We found that the questions are better on some dimensions than others - i.e. some metrics show much higher IRR agreement scores than others. The Grammatical correctness, Clarity, and Answerability metrics have very high scores across majority of the questions. We plan on expanding our method to other domains and do cross-domain evaluation. The IRR is lower for difficulty because it can be a subjective metric for annotators, and we can not have objective criteria for annotators.

Table 1: IRR Scores for question evaluation dimensions

Metric	κ score
Relevance	0.89
Grammar	0.99
Difficulty	0.60
Clarity	0.94
Contextual Specificity	0.92
Question-Option Disjoint	0.90
Distractor Homogeneity	0.77
Distractor Plausibility	0.65

Other than human evaluation, we also used automated evaluation methods defined by Moore et al [33] using a set of 19 item writing flaws (IWFs) from their study. These flaws are different errors that can be committed while writing multiple choice questions – We replicated both their rule-based and LLM-based methods for the complete set of 1850 questions for both the aforementioned courses. Table 2 summarizes the results for both automated methods.

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

For both methods, the majority of the questions pass a significant number of IWFs, with the average failure being less than 2 IWFs for both methods. Other than that, almost all the questions pass at least half of the metrics for both evaluation methods. While more work is needed to further decrease the rate of IWFs, by making a few adjustments noted earlier, our results show that a significant fraction of questions (about 20%) pass all IWF tests. Thus, based on the large scale of the initial question candidate generation, our system produces many thousands of high-quality questions that will be useful for deployment in live instructional settings, in concert with human verification.

3.4 Evaluating LLMs capabilities to generate natural language insights for courses

Some recent Theory of Mind experiments with LLMs suggest that LLMs may show at least limited capabilities to reason about mental representations of characters in stories [27], although there have been conflicting conclusions regarding those abilities [39]. While there is need for robust experiments to establish Theory of mind capabilities of LLMs in social scenarios [38], evaluating whether LLMs can reason about student's conceptual knowledge gaps given their mistakes can help instructors as well. For example, if instructors provide LLMs a list of questions and a student's answers, can a LLM accurately reason about the concepts that students are unaware of, diagnose why a student may have made an incorrect inference, or develop strategies and activities to remediate those specific errors in a given instructional setting? More research on methods, metrics, and datasets for assessing and optimizing the use of educational theory-of-mind abilities of LLMs appears needed: the recent study by Wang et al. [45] on how LLMs can assist with remediation of student math errors is a notable example in this direction.

One example of an important consideration in this type of evaluation is to what extent LLMs representation of student knowledge gaps allows instructors to take informed decisions about their course structure. We intend to address this with the help of the thousands of high-quality generated questions contained in our database: combined with empirical evidence and guidance from instructors, we are implementing a multi-course experiment within the same STEM program that will gather data for developing and validating specific forms of educational theory-of-mind measures around key concept dependencies, misconceptions, and learner inference errors.

3.5 Evaluating LLMs capability to provide feedback on areas for improvement

In our online degree program, platform support for learner help-seeking is critical, and a key part of seeking help is knowing what one doesn't know. Thus, our field prototype uses the extensive question set in our course metadata database to enable a variety of adaptive self-assessments. First, given the wide variance in student backgrounds, response data from an initial entrance evaluation is extremely valuable as a baseline against which to track student background expertise. Second, self-assessment within specific courses can provide ongoing measures of progress for both learners and instructors.

In addition to self-assessment and reflection on one's own progress, providing students relevant feedback has been shown to be critical to improve learning outcomes [16]. However, instructors

are not able to provide personalized student feedback due to challenges of scale [11]. Automated methods for student feedback rely on feedback templates informed by instructor's domain knowledge or recent neural methods to automatically provide natural language feedback on code [24]. However, these approaches require careful data training or domain knowledge making their wide applicability difficult.

LLMs have been shown to generate flexible natural language feedback given a question and student's answer choice [25]. However, to help instructors provide feedback to students at scale, we need to understand the quality of feedback that LLMs can provide and what are the gaps in different levels of LLM generated feedback according to Bloom's taxonomy of educational learning [28]. We also need studies to understand bias in LLM feedback [3]. A recent study on writing feedback with LLMs demonstrates that writing produced by users taking the aid of writing assistants with specific viewpoints has a higher tendency to reflect those viewpoints [20].

3.6 Aligning course activities with objectives

In this respect, our extensive database of AI-generated metadata across lectures and courses in our degree program is already proving to be of great benefit for doing fine-grained curriculum mapping. Both instructors and administrators now are able to analyze a complete picture of exactly which concepts are being taught in each class and how, which important concepts or methods are not currently covered, how the class materials correspond to overall learning objectives, and how well a class curriculum is aligned with important learning objectives in upstream and downstream courses within the program as a whole. Given that online courses tend to be much shorter and faster-paced than their residential counterparts (e.g. 4 weeks online vs 10-13 weeks on campus), instruction is more like co-teaching, so it is especially critical to have reliable tools to make sure content stays coordinated across inter-dependent course series as well as the entire program.

4 Challenges around data collection to support educational objectives

4.1 Privacy considerations

Personalizing education requires data on student's performance in courses raising privacy considerations of student data [30]. However, LLMs add another dimension to privacy challenges of personalization [12] due to their potential to infer students' knowledge states from their data. Most state-of-the-art LLMs are not accessible outside of the APIs hosted by large organizations. Querying the LLMs using student data implies agreeing to the terms and conditions of use, which may not be agreeable to every organization. Student data is protected by strong regulations [13] and in our own experiments, we are mindful of the need to use student data with LLMs in privacy and regulation compliant ways. For example, we strictly use University storage for student related data, and ensure that we de-anonymize any student metadata before querying the GPT api. Our current experiments include coarse grained description of student performance in terms of concepts that they get right or wrong in tests. However, future algorithms to support learning that may require more student data such as their prior background and demographics need to ensure that several student attributes even if anonymized do not violate privacy when combined together [35].

4.2 Ethical considerations

Use of LLMs in education also raises ethical considerations as LLMs have been known to exhibit biases and active research is underway to address such issues [3]. LLMs are still in their nascent stage and despite their ease of text generation through prompting, lay users need training on how to use them [47]. Their effective use in education requires careful training for instructors and students alike [4]. Experimental uses in educational scenarios need appropriate safeguards so that no student's learning experience is affected due to problems with generative AI that we are still uncovering.

5 Summary of proposed future research directions

Based on our experience so far with exploring how to integrate LLMs at multiple interaction points of an online academic degree platform, we advocate for further research in the following areas. This

list is not exhaustive, but represents those areas that we believe are important high priority directions in service of our goal of providing both students and instructors with effective, robust affordances for understanding and supporting learning, especially at scale.

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. For example, robust reading difficulty metadata generated from a basic language-modeling approach is one simple example from older research [26] that directly enabled better personalization of results, and characterization of user and site expertise. In general, given interaction traces between learners and content, improvements in content representation can lead directly to more expressive learner models. The ability of generative AI to detect and summarize deeper semantic events, such as key educational moments based on what was said and shown in class, remains a fruitful area to explore. Beyond fixed, pre-computed metadata, opportunities also exist for hybrid approaches where existing pre-computed metadata can be transformed and adapted for specific learner scenarios. For example, metadata describing an important learning moment for a concept in a video could be 'translated' in the context of a team's project report to explain how that new concept relates to the team's choice of methods.

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. For example, we are currently investigating combining the recent DAS3H [8] practice model, which provides success/recall probabilities for specific skills/concepts based on past question practice history, with LLM-based functions that can add additional components to the utility objective based on natural language descriptions of important learner-specific attributes such as their future goals.

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. New LLM evaluation instruments would assess the accuracy, scope, and robustness of a LLM's ability to diagnose and explain student misconceptions, knowledge gaps, and incorrect inferences. This would be related to existing measures such as Mathematical Knowledge for Teaching (MKT) [18], but in a framework that could be extended to arbitrary domains and wider diagnostic categories. Beyond evaluation, developing more accurate educational theory-of-mind abilities for interaction would likely require new rich data representations of content and learners, AI-expert collaboration, and enhanced latent representations in learner models informed by the significant existing line of educational research on MKT and general help seeking (e.g. [23]).

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 [3]. We have limited understanding of the extent of bias in LLM recommendations [48]. Instructors using LLM support to draw insights or create questions can over-rely on LLM generations substantially changing the outcome of their work, depending on the model's data representation [36]. We can draw inspiration from the rich literature on Human-AI interaction [31], and Cognitive science [41] to study the interaction aspects of LLMs and how they can best fit with the decision-making of instructors to augment their capabilities, instead of overriding it.

6 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.

References

[1] Ian A. Apperly. What is "theory of mind"? concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology*, 65(5):825–839, 2012. PMID: 22533318.

[2] Sumit Asthana, Sagih Hilleli, Pengcheng He, and Aaron Halfaker. Summaries, highlights, and action items: Design, implementation and evaluation of an llm-powered meeting recap system. *arXiv preprint arXiv:2307.15793*, 2023.

[3] Perpetual Baffour, Tor Saxberg, and Scott Crossley. Analyzing bias in large language model solutions for assisted writing feedback tools: Lessons from the feedback prize competition series. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 242–246, Toronto, Canada, July 2023. Association for Computational Linguistics.

[4] Simon Buckingham Shum, Lisa-Angelique Lim, David Boud, Margaret Bearman, and Phillip Dawson. A comparative analysis of the skilled use of automated feedback tools through the lens of teacher feedback literacy. *International Journal of Educational Technology in Higher Education*, 20(1):40, 2023.

[5] Rodrigo Campos, Rodrigo Pereira dos Santos, and Jonice Oliveira. A recommendation system based on knowledge gap identification in moocs ecosystems. In *Proceedings of the XVI Brazilian Symposium on Information Systems*, SBSI '20, New York, NY, USA, 2020. Association for Computing Machinery.

[6] Hung Chau, Igor Labutov, Khushboo Thaker, Daqing He, and Peter Brusilovsky. Automatic concept extraction for domain and student modeling in adaptive textbooks. *International Journal of Artificial Intelligence in Education*, 31:820–846, 2021.

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

[8] Benoît Choffin, Fabrice Popineau, Yolaine Bourda, and Jill-Jênn Vie. Das3h: modeling student learning and forgetting for optimally scheduling distributed practice of skills. *arXiv preprint arXiv:1905.06873*, 2019.

[9] Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International Journal of Information Management*, 71:102642, 2023.

[10] Sabina Elkins, Ekaterina Kochmar, Iulian Serban, and Jackie CK Cheung. How useful are educational questions generated by large language models? In *International Conference on Artificial Intelligence in Education*, pages 536–542. Springer, 2023.

[11] Peter Ferguson. Student perceptions of quality feedback in teacher education. *Assessment & evaluation in higher education*, 36(1):51–62, 2011.

[12] Sameera Ghayyur, Jay Averitt, Eric Lin, Eric Wallace, Apoorvaa Deshpande, and Hunter Luthi. Panel: Privacy challenges and opportunities in {LLM-Based} chatbot applications. 2023.

[13] Ann Gilley and Jerry W Gilley. Ferpa: What do faculty know? what can universities do? *College and University*, 82(1):17, 2006.

[14] Hannah Gonzalez, Jiening Li, Helen Jin, Jiaxuan Ren, Hongyu Zhang, Ayotomiwa Akinyele, Adrian Wang, Eleni Miltsakaki, Ryan Baker, and Chris Callison-Burch. Automatically generated summaries of video lectures may enhance students' learning experience. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 382–393, Toronto, Canada, July 2023. Association for Computational Linguistics.

[15] José González-Brenes, Yun Huang, and Peter Brusilovsky. General features in knowledge tracing to model multiple subskills, temporal item response theory, and expert knowledge. In *The 7th international conference on educational data mining*, pages 84–91. University of Pittsburgh, 2014.

[16] John Hattie and Helen Timperley. The power of feedback. *Review of educational research*, 77(1):81–112, 2007.

[17] Marti A. Hearst. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

[18] H.C. Hill, S.G. Schilling, and D.L. Ball. Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal*, 105:11–30, 2004.

[19] Monika Hooda, Chhavi Rana, Omdev Dahiya, Ali Rizwan, Md Shamim Hossain, et al. Artificial intelligence for assessment and feedback to enhance student success in higher education. *Mathematical Problems in Engineering*, 2022, 2022.

[20] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2023.

[21] Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. Promptmaker: Prompt-based prototyping with large language models. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–8, 2022.

[22] Srećko Joksimović, Oleksandra Poquet, Vitomir Kovanović, Nia Dowell, Caitlin Mills, Dragan Gašević, Shane Dawson, Arthur C Graesser, and Christopher Brooks. How do we model learning at scale? a systematic review of research on moocs. *Review of Educational Research*, 88(1):43–86, 2018.

[23] S. A. Karabenick and J. R. Knapp. Relationship of academic help seeking to the use of learning strategies and other instrumental achievement behavior in college students. *Journal of Educational Psychology*, 83(2):221–230, 1991.

[24] Hieke Keuning, Johan Jeuring, and Bastiaan Heeren. A systematic literature review of automated feedback generation for programming exercises. *ACM Trans. Comput. Educ.*, 19(1), sep 2018.

[25] Natalie Kiesler, Dominic Lohr, and Hieke Keuning. Exploring the potential of large language models to generate formative programming feedback. *arXiv preprint arXiv:2309.00029*, 2023.

[26] Jin Young Kim, Kevyn Collins-Thompson, Paul N Bennett, and Susan T Dumais. Characterizing web content, user interests, and search behavior by reading level and topic. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 213–222, 2012.

[27] Michal Kosinski. Theory of mind might have spontaneously emerged in large language models. *arXiv 2302.02083*, 2023.

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

[29] Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204, 2020.

[30] Danny Yen-Ting Liu, Kathryn Bartimote-Aufflick, Abelardo Pardo, and Adam J Bridgeman. Data-driven personalization of student learning support in higher education. *Learning analytics: Fundaments, applications, and trends: A view of the current state of the art to enhance e-learning*, pages 143–169, 2017.

[31] Mansoureh Maadi, Hadi Akbarzadeh Khorshidi, and Uwe Aickelin. A review on human–ai interaction in machine learning and insights for medical applications. *International journal of environmental research and public health*, 18(4):2121, 2021.

[32] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[33] 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.

[34] Lidiya Murakhovs' ka, Chien-Sheng Wu, Philippe Laban, Tong Niu, Wenhao Liu, and Caiming Xiong. Mixqg: Neural question generation with mixed answer types. *arXiv preprint arXiv:2110.08175*, 2021.

[35] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125, 2008.

[36] Samir Passi and Mihaela Vorvoreanu. Overreliance on ai literature review. *Microsoft Research*, 2022.

[37] Sandra Sanchez-Gordon and Sergio Luján-Mora. Design, implementation and evaluation of moocs to improve inclusion of diverse learners. In *Accessibility and diversity in education: Breakthroughs in research and practice*, pages 52–79. IGI Global, 2020.

[38] Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[39] Natalie Shapira, Mosh Levy, Hossein Seyed Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv 2305.14763*, 2023.

[40] Dirk T Tempelaar, André Heck, Hans Cuypers, Henk van der Kooij, and Evert van de Vrie. Formative assessment and learning analytics. In *Proceedings of the third international conference on learning analytics and knowledge*, pages 205–209, 2013.

[41] Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. Do large language models know what humans know? *Cognitive Science*, 47(7):e13309, 2023.

[42] Don Tuggener, Margot Mieskes, Jan Milan Deriu, and Mark Cieliebak. Are we summarizing the right way?: a survey of dialogue summarization data sets. In *Conference on Empirical Methods in Natural Language Processing (EMNLP), Punta Cana, Dominican Republic (online), 7-11 November 2021*, pages 107–118. Association for Computational Linguistics, 2021.

[43] Olga Viberg, Mathias Hatakka, Olof Bälter, and Anna Mavroudi. The current landscape of learning analytics in higher education. *Computers in human behavior*, 89:98–110, 2018.

[44] Chunpai Wang, Shaghayegh Sahebi, Siqian Zhao, Peter Brusilovsky, and Laura O. Moraes. Knowledge tracing for complex problem solving: Granular rank-based tensor factorization. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '21, page 179–188, New York, NY, USA, 2021. Association for Computing Machinery.

[45] Rose E. Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. Step-by-step remediation of students' mathematical mistakes. *arXiv preprint arXiv:2310.10648*, October 2023.

[46] Denise Whitelock, Alison Twiner, John TE Richardson, Debora Field, and Stephen Pulman. Openessayist: A supply and demand learning analytics tool for drafting academic essays. In *Proceedings of the fifth international conference on learning analytics and knowledge*, pages 208–212, 2015.

[47] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2023.

[48] Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. *arXiv preprint arXiv:2305.07609*, 2023.

[49] Olga Zlatkin-Troitschanskaia, Hans Anand Pant, and Hamish Coates. Assessing student learning outcomes in higher education: Challenges and international perspectives, 2016.