Is GPT-3 a Good Course Forum Assistant?

Hannah Levin
Department of Computer Science
Stanford University
levinh@stanford.edu

Anjali Ragupathi
Department of Symbolic Systems
Stanford University
anjalirg@stanford.edu

Abstract

Massively Open Online Courses (MOOCs) substitute in-classroom discussion with course forums, where teaching staff responds to student queries. However, at a large scale, it becomes unfeasible to respond to every student that makes a request - in the process, some requests might be lost, leading to lower student uptake and increased frustration as their confusion is left unaddressed. To leverage recent developments with large language models, we explore how effective GPT-3 is in handling confusion on course forums. Our work analyzes human teachers' responses across 6 different Stanford MOOC courses from 2014 and compares them with those generated by GPT-3; we compare the two along the dimensions of semantic similarity preservation of the question and conversational uptake measured by Pointwise Jenson-Shannon Divergence (PJSD). Additionally, we conduct a field study to evaluate if students show a preference for human responses over GPT-3 responses. The results of our study demonstrate that there is an overwhelming preference by students for GPT-3 responses over human responses for posts classified as moderately confusing. Furthermore, GPT-3 outperforms teaching staff in conversational uptake, an essential indicator for student success.

Key Words: conversational uptake, natural language processing, online learning

1 Introduction

With distance learning and digital education gaining popularity since 2012 (Shah, 2020), enrollment in popular university courses has become more accessible to learners from around the world. With this proliferation of knowledge comes the disadvantage of impersonal interaction: in smaller, in-person classrooms, teaching staff can devote more attention to tracking the progress of individual students, addressing confusion more promptly, and tailoring their interactions to students' learning styles. With Massively Online Open Courses (MOOCs), however, classroom discussion is replaced by course forums, where it is easy for student inquiries to be left unanswered due to the sheer volume of requests. Geller et al. (2021) emphasizes that certain types of confusion left unaddressed could easily turn into frustration; by extension, this could lead to students losing interest in the course material or developing less faith in their ability to complete the course.

Recent work on large language models (e.g.(Brown et al., 2020)) has opened up the possibility of applying them to open-domain question answering (Yue et al., 2022; Pereira et al., 2022). Despite immense progress in this field, there has not been much significant work in applying this general principle of using large language models for question-answering to the educational field. This might be because education - which relies heavily on knowledge transfer - needs to be less susceptible to being misinformed using unverified data, which is a major problem point in using large language models. Hallucinations of information cannot be easily verified in the academic context, and most work in fact-checking has so far been restricted to the political domain. Moreover, pre-trained models like GPT-3 and T5 are black boxes with no way of informing their users of their limitations. Because of these ethical considerations, it becomes important to identify the strengths and weaknesses of such a model before releasing it into the field for public use.

Our study hypothesizes that large language models such as GPT-3 can leverage these capabilities to act as a teaching assistant on large-scale online educational course forums. This is distinct from the approach used by Tack and Piech (2022), which tests the capabilities of AI as a teacher by evaluating if AI can speak like a teacher, understand and help students. To answer our hypothesis, we feed questions expressing confusion about some course material to GPT-3 and compare its response to the actual first reply of the teaching staff to the same question. We measure the relevance of the answer and the conversational uptake involved in the exchange between the teaching staff and the student using cosine similarity and point-wise Jensen-Shannon Divergence respectively. Additionally, we conduct a field study asking current students to rate their preferred response (human or GPT-3) in terms of delivery in a blind trial.

Through these experiments, we observe if the performance of GPT-3 is dependent on the course - for instance, does it exhibit good conversational uptake consistently across all surveyed courses, or does it do better on a specific domain (such as Math or Writing)? Our results demonstrate that the average cosine similarity and conversational uptake of GPT-3 is comparable to or better than that of human teaching staff, indicating the potential of using GPT-3 as a teaching assistant. Additionally, human evaluation backs these findings by showing that students consider GPT-3 responses to be equivalent or better than human responses in terms of the way that the response is delivered.

2 Related Work

Research at the intersection of artificial intelligence and education has explored ways to use digital data (e.g. forum posts, lecture recordings, and course material) to improve the overall learning experience for students on online platforms. One direction of exploration focuses on identifying sources of confusion as the student learns new material and tracing how this knowledge is acquired and retained over the course of multiple lecture sessions. Another facet explored in these studies highlights how the teaching experience contributes to the acquisition and retention of student knowledge, including ways in which data from teacher-student discourse can be used to improve the quality of instruction.

For instance, Geller et al. (2021) propose a method to automatically detect confusion on course forums by training a machine learning model with word embedding vectors representing confusion-related terms to discriminate between confusing and non-confusing posts. The paper emphasizes the importance of teachers being able to quickly and accurately recognize not just a student's confusion, but also its source, in order to prevent confusion morphing into frustration. Additionally, since confusion can stem from numerous issues, including unfamiliarity with course material, teaching style that is incompatible with the student's learning style, or even logistical questions regarding the course format, the authors employ a labeling-tree method to distinguish between these different categories. This approach proves to be a helpful pedagogical tool in education to identify any misunderstandings in pre-lecture readings and properly clear up any doubt post-lecture by addressing the root of the misunderstanding. Further work by Wang et al. (2017) introduces a visual data mining technique to trace students' confusion as they work through a coding problem. This enables educators to pre-emptively identify when students get stuck on a problem and to provide timely assistance for improved instruction quality.

In analyzing discourse between educators and learners, Demszky et al. (2021b) propose a method to measure uptake through collaborative teacher-student exchanges using annotated math class transcripts. Here, "uptake" is defined as repeating what was said in a different way that consequently improves student performance. This study estimates uptake by using point-wise Jensen-Shannon divergence (PJSD) to measure the similarity between the distributions of the student's original inquiry and the teacher's reply; we employ the same approach in comparing uptake scores for GPT-3 and human teaching staff. The TalkMoves dataset (Suresh et al., 2022) and the NCTE Transcripts dataset (Demszky and Hill, 2022) on these ideas by constructing corpora of teacher-student exchanges containing phrases, words, and utterances that facilitate collaboration, encourage discussion, and promote reasoning. Overall, these studies empirically demonstrate a positive correlation between uptake and student success in a course, with the idea that validating the importance of a student's finding would initiate a problem-resolution process; this is further corroborated by Tack and Piech (2022) which uses an automated tool which gives teachers feedback about their uptake of student responses. Motivated by these foundations in learning science, our work aims to explore these principles when applied to large language models.

Cannon (2022) and Renduchintala et al. (2017) approach the problem of detecting confusion through the lens of knowledge tracing. (Cannon, 2022) uses an automatic Zoom recording analysis tool to create a profile of the presenter's teaching style based on how many times the presentation slides were changed, the variation in student sentiment over the course of the lecture, and the content of the slides - both stylistic and technical. On the other hand, Renduchintala et al. (2017) present a method of tracing a student's acquisition and retention of knowledge in a language-learning task. Current knowledge is modeled using an update rule that depends on knowledge the student acquired in the past. The student's learning experience is based on an update vector symbolizing the gradient of the log probability of the correct answer, which the student attempts to maximize. The authors compare how different approaches to correcting a student when they make a mistake on a question can alter the modeled knowledge state. To summarize, both studies take an indirect route to understanding where a student begins to get confused, rather than using direct indicators present in course forum posts or lecture transcripts. While these methods have great merit in the extension of our project to larger use-cases, we choose not to use them within the scope of our current task.

3 Approach

3.1 Topic Modeling

Since the Stanford MOOCPosts Dataset (Agrawal et al., 2015) contains forum posts from 11 courses and comprised technical, administrative, and content-focused questions, we attempt to ensure that our analyses are not confounded by irrelevant posts. For this reason, we eliminate the 5 courses with the lowest number of posts while preserving 6 courses spanning Education, Humanities and Sciences, and Medicine. Further, since we want to identify specific topics that pose problems for students in each course, we use topic modeling to map the distribution of topics throughout the quarter in which that course is taught.

Specifically, we use BERTopic (Grootendorst, 2022) to model the distribution of topics across the documents (i.e. posts). We restrict the posts included for this analysis to be questions having a confusion rating of 4.5 and above, to hone in on topics that students find especially confusing. To remove the impact of stopwords and irrelevant parts of speech (such as prepositions, adverbs, and conjunctions), we apply a custom pre-processing pipeline function that performs normalization, contraction replacement, part-of-speech tagging (keeping nouns, verbs, and adjectives), lemmatization, and stopword removal. The pre-processed questions are fed into the topic model and transformed to predict their topics. A topic distribution (see Figure 1 and Figure 2) is plotted over time for each course to highlight the most frequently talked about topics throughout the quarter, and the topic information is used to filter out the administrative or technical support questions from our final dataset.

In Education and the Humanities/Social Sciences (e.g., Fig. 2), the frequency of topics over time peaks at the beginning of the course and declines progressively towards the end, exhibiting a decline in discussion as the course ends. However, in Medicine (Fig. 1), the frequency of topics referenced across the course does not show a clear increase or decrease. Spikes in the use of specific keywords are observed throughout the time period, indicating sustained discussion and reuse of previous concepts in new material. This shows that different subjects propagate equally different discussion and learning styles.

3.2 GPT-3 Generation

Using the post IDs from every post or comment in the dataset, we map each student's post and its course name to a tuple containing the thread data such as each comments' text, timestamp, and confusion rating. Due to monetary constraints for calls to the GPT-3 API, we used stratification to sample a subset of our dataset to preserve the distribution of courses when randomly selecting posted student questions to generate GPT-3 responses to.

Using the post IDs from every post or comment in the dataset, we map each student's post and its course name

3.3 Questionnaire

Using random oversampling, we account for class imbalance in the dataset to select 12 questions equally representing Education, Statistics, Environmental Physiology, Science Writing, and Economics courses. Given a posted student question and two teacher responses, we prompt survey participants to mark the "better" response (defined here as one that is coherent, relevant, and supportive to the student's learning experience), and not evaluate subject matter/ factual correctness of either option. (e.g. Figure 3)

The original text contained special formatting such as '<REDACT>' and 'X777', presumably to replace usernames and other private information in the dataset. Thus, we use GPT-3 to remove these

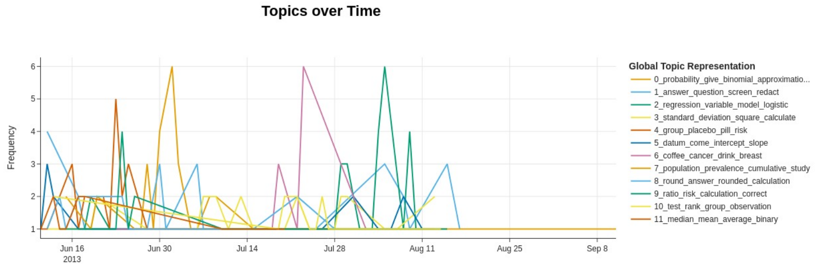


Figure 1: Topics over time in Statistics in Medicine course

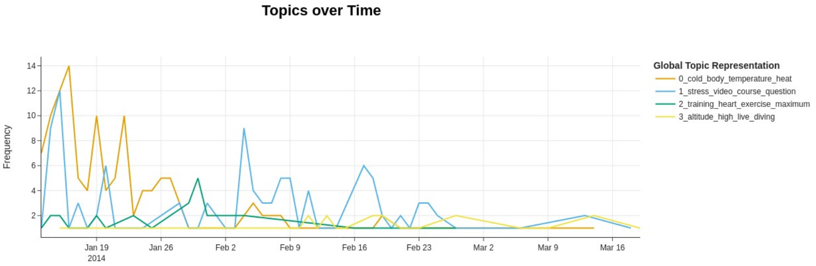


Figure 2: Topics over time in Environmental Physiology course

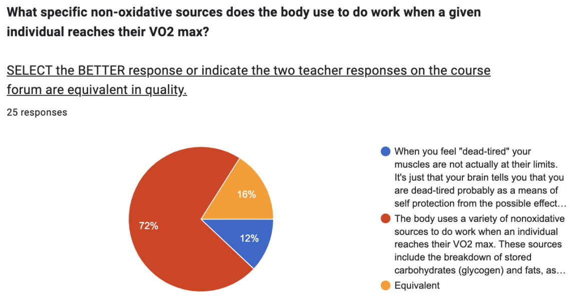
redactions and turn student questions into coherent sentences for the questionnaire, taking care to keep the cleaned texts as close as possible to the original wording of the student's question.

Initially, we wanted to send the questionnaire to alumni of the MOOC 2014 courses but this was not feasible as professors that responded to our requests did not have an easy way to reach MOOC alumni or their equivalent Stanford class was not in session at the time. Thus, survey participants were friends/associates not directly involved with the study. The participants were not informed that there were GPT-3 generated responses on the questionnaire; however, it is possible that some of participants could have guessed some of the teacher responses were GPT-3 generated, as seen in the comment section of our survey. Our original intention in getting alumni of the course to fill out the survey was to get data from people who have sufficient knowledge of the material to be able to make meaningful assessments of the teacher responses. To account for this, we prompted questionnaire participants to provide their level of involvement with the material. 80 percent were familiar with the topics in some capacity (including vague familiarity, high school/ college-level course/ academic specialization/ professional experience).

3.4 Uptake

Conversational uptake measures the overlap between a speaker's enquiry and the listener's response to it, as well as their mutual dependency. We follow Demszky et al. (2021a)'s approach of using point-wise Jensen-Shannon divergence (PJSD) to estimate this uptake score. We first remove punctuation and stopwords using NLTK, and stem the words using NLTK's SnowballStemmer. After tokenizing each student and teacher's texts, we create word embeddings using GloVe vectors.

Specifically following the interactions of the original student who posted a given thread, we only study the uptake of that student and the teacher (using the proxy of the most prominent poster for that course). Furthermore, we only consider threads of three or more turns to weed out questions that were never responded to and single question-answer threads of length two with no follow-up by student and teacher. This is because we want to study the back-and-forth interaction of students and teachers to get an average conversational uptake score. For example an interaction thread of (S,T,S,T) where S refers to the student and T refers to the teacher/teaching staff, we find each pair - the first (S,T) and the second (S,T) - in the thread and calculate the PJSD value for each pair. We take the average of the PJSD values of the pairs in the thread to calculate the conversational uptake score



What specific non-oxidative sources does the body use to do work when a given individual reaches their VO2 max?

SELECT the BETTER response or indicate the two teacher responses on the course forum are equivalent in quality.

25 responses

- ● When you feel "dead-tired" your muscles are not actually at their limits. It's just that your brain tells you that you are dead-tired probably as a means of self protection from the possible effect…
- ● The body uses a variety of nonoxidative sources to do work when an individual reaches their VO2 max. These sources include the breakdown of stored carbohydrates (glycogen) and fats, as…
- ● Equivalent

Figure 3: An example of the response distribution for a question on survey

for that interaction. Lower PJSD scores indicate a lower divergence between the distributions student and teacher text in the interaction, indicating a better uptake score.

4 Experimental Methods

4.1 Data

For the forum post component of our project, we use the 2014 Stanford MOOCPosts Dataset (Agrawal et al., 2015) which contains 29,604 discussion posts from eleven publicly available Stanford Online classes. Each post is categorized as either an opinion, question, or answer (or any combination of the three), annotated additionally with information about its sentiment (positive, neutral, negative), urgency (on a scale of 1 to 7) and confusion (on a scale of 1 to 7).

The dataset was completely anonymized and deidentified by the authors themselves, and annotated independently by nine paid coders with some subject knowledge (but not necessarily at the expert level). Gold standard scores for the annotated features were computed based on inter-coder agreeability.

4.2 Evaluation metrics

We use cosine similarity to measure the relevance of the teacher's response to the question asked through the semantic similarity preservation. We do this to compare the relevance of responses between teacher and GPT-3 to students' questions. Another metric we used was conversational uptake, measured through Pointwise Jensen-Shannon Divergence. PJSD shows the dependence of the teacher's response on the student's post as a measure of conversational uptake. While cosine similarity is a metric commonly used for semantic similarity, conversational uptake and educational course forums are perhaps more complex and better represented by the PJSD metric for each thread.

4.3 Experimental details

For the initial topic analysis, we initialize the BERTopic model with pre-trained universal sentence embeddings (all-mpnet-base-v2). We use the GPT-3 Davinci model with a frequency penalty of 0.6 and temperature of 0.3 with a prompt to answer and explain each question from the stratified sample in at least 4 to 6 sentences on a course forum, as a teacher. We choose a low temperature to create more deterministic and consistent responses, since on a course forum, answers to the same question should not vary. The moderate frequency penalty value in the configuration is intended to prevent the model response from being too repetitive in its explanation. A presence penalty of 0 is chosen because we want the model to exhibit signs of uptake; thus, we do not penalize it for using words present in the prompt. To measure the cosine similarity of the question with the human and GPT-3 responses, we convert the texts into sentence embeddings using a SentenceTransformer fine-tuned on the semantic textual similarity task (all-MiniLM-L12-v1). We then use a cosine similarity implementation from scikit-learn to measure the relative similarities of the responses to the original query, as well as to each other.

5 Results and Analysis

From the results in Table 2, we see that cosine similarity (between human and GPT-3 responses) is relatively higher for moderately confused posts, in particular for the Environmental Physiology and Summer 2014 MedStats courses (Table 2). This shows that the response deliveries of both GPT-3 and the teaching staff are approximately equivalent for those courses. Survey responses also indicate an overall preference towards GPT-3 answers in terms of response style and level of explanation.

To better understand how teaching staff capacity

Course Name	Human PJSD	GPT-3 PJSD
How_to_Learn_Math	0.123918	0.063128
StatLearning	0.121856	0.057794
Environmental_Physiology	0.089016	0.065785
Statistics_in_Medicine	0.117552	0.072898
MedStats	0.119954	0.082759
ScienceWriting	0.115420	0.052287
Average:	0.114619	0.065775

Table 1: Comparing Human versus GPT-3 with Pointwise Jensen-Shannon Divergence (PJSD)

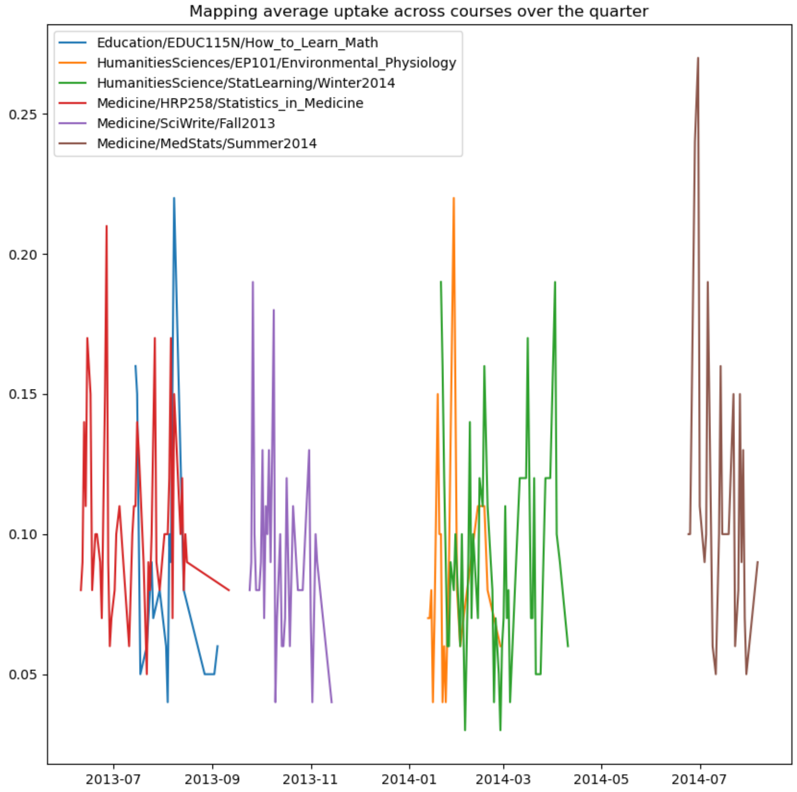


Figure 4: Average uptake per course over time without GPT-3 (note: smaller PJSD value means better uptake score)

(such as around mid-course for midterms or last week of the course for finals) and conversational uptake in course forum threads might relate, we graph average PJSD values over time for each course (see Figure 4). Smaller PJSD values mean that there is a smaller divergence from student's post and thus a larger uptake score.

Across courses, conversational uptake of between student and teacher responses on the course forums is not consistent throughout the span of the class. Our hypothesis was that GPT-3 would have more consistent uptake scores across the span of the class. To our surprise, GPT-3's PJSD values to these questions were also not consistent over the span of the class. However, GPT-3 outperforms teachers in average uptake for each course, indicating its potential in delivering more widely acceptable responses to student queries.(see Figure 5)

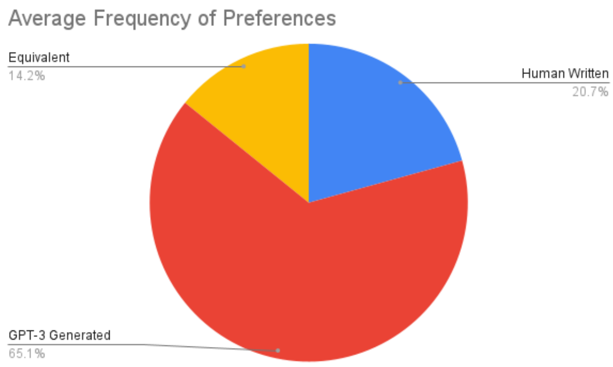


Figure 5: Average distribution of preferred responses on survey

6 Conclusion and Future Work

Our results demonstrate that GPT-3 outperforms humans on conversational uptake across courses in the dataset. However, GPT-3 generated responses are similarly inconsistent in their uptake scores over the period of the course as teachers' responses are. GPT-3 performs stronger on moderately confused classified questions and less so on low and high confusion, which means that while it can offload most of the moderately confused questions from the teaching staff to itself, the relatively critical posts can be handled by subject experts to preserve teaching quality and maintain some level of correctness for high-priority responses.

As a future extension to this project, we propose the following model for a collaboration between GPT-3 and teachers on MOOC forums:

- Automatically classify level of confusion (ranked 1-7) in posts
- GPT-3 generates responses to moderately confused (ranked 4-6) posts while teaching staff handles high-confusion / urgent posts.
- Teaching staff periodically checks samples of GPT-3 generated posts and flags any incorrect statement while providing corrections (i.e., annotation) to further tune model.
- Measure multi-exchange conversational uptake between student and teacher in threads (Demszky et al., 2021a) - improving upon what is covered at a single-exchange sentence level in Table 2

Limitations

Our study does not account for the factual correctness of the responses presented by GPT-3, which poses a problem when applied to the real-world assumption that teaching staff use verified information in their explanations. Additionally, GPT-3 does not generate significantly better responses to high-confusion, high-urgency, or low-urgency posts. This could likely be because such posts are outliers in the dataset, and we would need more examples of critical posts to accurately evaluate model performance on them. Student feedback has also revealed that GPT-3 responses are often too wordy, even when more concise alternatives exist. From a financial standpoint, large language models like GPT-3 are too expensive to use with massive datasets, and thus, our work is limited in the amount of data that could be analyzed. However, from a commercial standpoint, it is possible to use such models in various applications: currently, a startup called You.com is using GPT-3 requests in their search engine while providing the search for free to users.

Our project does not take into consideration the context that a teacher might have when interacting with students on forums such as current lectures, previously taught topics in the course, prerequisites, and course syllabus. It would be interesting to see how GPT-3 responses would change when trained on the context of the course and student's academic background when answering questions. For example, it may be helpful to train the model on lecture transcripts, previous discussion posts, and

Course	Human/GPT-3	Question/Human	Question/GPT-3	Count
How to Learn Math (EDUC)	0.414	0.407	0.716	21
Stat Learning (H&S)	0.416	0.389	0.729	153
Environmental Physiology (H&S)	0.504	0.492	0.787	54
Stats in Medicine 2013 (MED)	0.437	0.423	0.738	154
MedStats 2014 (MED)	0.456	0.435	0.732	67
Science Writing (MED)	0.419	0.404	0.723	69

Table 2: Cosine similarities per-course

the course syllabus. We would be curious to see if GPT-3 can provide tailored responses with references to course material such as how the teaching staff does on course forums.

Though GPT-3 cannot guarantee factual correctness, human error is also likely when at a large-scale. frequency could study cases where AI fails to generate factually correct or consistent answers across large-scale forums when compared to humans.

Ethics Statement

Our metrics for evaluating the "better" response between GPT-3 and the teacher should include factual correctness. This is the primary limitation of this project and is an important ethical consideration, especially if applied to real courses where misinformation can impact student understanding and grades. Even if redactions are made for incorrect GPT-3 issued responses, we can not guarantee that we can reach the majority of users who already viewed the incorrect post. Additionally, paying students may not approve of GPT-3-automated course forums as they are paying for an education from experts in the field, not from machines. Thus, it is vital to maintain a balance between the roles of humans and GPT-3 on MOOC forums.

GPT-3 does not have that same context that teaching staff have when interacting with students. For example, a teacher who has taught a class for ten years is likely more familiar with better ways to explain concepts than GPT-3. However, given the MOOCs nature, it is not feasible to scale that quality to all students.

In collecting information about student preference between GPT-3 and human responses, we distributed the questionnaire to a sampling of students in our acquaintance circles. This study was conducted voluntarily with no monetary compensation; perhaps in involving financial rewards, we could have incentivized more people to take the survey. We took the consent of all participants to collect their email information and anonymized the results in presenting them through this study.

7 Authorship Statement

Hannah Levin was responsible for programming and developing the mappings of student questions to threads in the dataset so that data analysis could be more easily viewable by the whole thread at a time. She also developed the GPT-3 API calls for response generation and storing the human response as comparison in a dataframe, and implemented Demtzky's PJSD approach to measuring conversational uptake for both teacher responses and GPT-3 generated responses.

Anjali Ragupathi was responsible for developing the preprocessing pipeline for topic modelling, conducting tests with the BERTopic model using different parameters, analyzing the results, and using them to filter out administrative and technical questions. She also evaluated the performance of GPT-3 in comparison to human teaching staff with respect to cosine similarity, plotted the graphs for average uptake per course, and helped design the survey questionnaire - in particular, the questions asking participants to reason about why they preferred one response over the other, and the ones collecting demographic information about participants' familiarity with the survey topics.

Acknowledgements

Thank you to Dr. Andreas Paepcke for his support and mentoring of this project. We also thank the participants of the survey for providing us with information to put our findings into context.

References

Akshay Agrawal, Jagadish Venkatraman, Shane Leonard, and Andreas Paepcke. 2015. Youedu: Addressing confusion in mooc discussion forums by recommending instructional video clips. In *Educational Data Mining*.

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jack Arlo II Cannon. 2022. An automated zoom class session analysis tool to improve education.

Dorottya Demszky and Heather Hill. 2022. The ncte transcripts: A dataset of elementary math classroom transcripts.

Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2021a. Measuring conversational uptake: A case study on student-teacher interactions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1638–1653, Online. Association for Computational Linguistics.

Dorottya Demszky, Devyani Sharma, Jonathan Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021b. Learning to recognize dialect features. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2315–2338, Online. Association for Computational Linguistics.

Shay A. Geller, Kobi Gal, Avi Segal, Kamali Sripathi, Hyunsoo G. Kim, Marc T. Facciotti, Michele Igo, Nicholas Hoernle, and David Karger. 2021. New methods for confusion detection in course forums: Student, teacher, and machine. *IEEE Transactions on Learning Technologies*, 14(5):665–679.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Jayr Pereira, Robson Fidalgo, Roberto Lotufo, and Rodrigo Nogueira. 2022. Visconde: Multi-document qa with gpt-3 and neural reranking.

Adithya Renduchintala, Philipp Koehn, and Jason Eisner. 2017. Knowledge tracing in sequential learning of inflected vocabulary. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 238–247, Vancouver, Canada. Association for Computational Linguistics.

Dhawal Shah. 2020. Capturing the hype: Year of the mooc timeline explained. `https://www.classcentral.com/report/mooc-hype-year-1/`.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022. The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4654–4662, Marseille, France. European Language Resources Association.

Anaïs Tack and Chris Piech. 2022. The ai teacher test: Measuring the pedagogical ability of blender and gpt-3 in educational dialogues.

Yiting Wang, Walker M. White, and Erik Andersen. 2017. Pathviewer: Visualizing pathways through student data. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*.

Xiang Yue, Xiaoman Pan, Wenlin Yao, Dian Yu, Dong Yu, and Jianshu Chen. 2022. C-MORE: Pretraining to answer open-domain questions by consulting millions of references. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 371–377, Dublin, Ireland. Association for Computational Linguistics.

A Example Appendix

This is a section in the appendix.