International Society of the Learning Sciences

Measuring Help-seeking in Online Course Discussion Forums with Privacy-preserving Large Language Models

Nigel Bosch, Destiny Williams-Dobosz, Michelle Perry
pnb@illinois.edu, destiny7@illinois.edu, mperry@illinois.edu
University of Illinois Urbana–Champaign

Abstract: Discussion forums in college courses play a vital role in enabling students to seek academic help from each other and from instructors, especially in online courses. However, it can also be difficult for instructors and researchers interested in help-seeking to identify when it occurs in large, active forums. This study examined large language models as measurement tools for automatically coding students' help-seeking forum posts, resulting in Spearman's *rho* up to .711 for the correlation between models' help-seeking ratings versus manually coding via an established schema. The large language model approach requires no problem-specific training data, enabling the help-seeking model to be constructed with minimal manual coding of forum posts compared to traditional machine learning approaches. Moreover, the models in this study are offline models, run on a typical laptop, which preserve students' privacy by eliminating the need to transmit forum data to a third-party service.

Introduction

Many college courses include an online discussion forum component, which can serve to foster a sense of community and support peer learning among students (Fiock, 2020), particularly in fully online courses where students would otherwise have limited opportunities to discuss the course with each other. One of the critical activities that occurs on course discussion forums is academic help-seeking—i.e., when students indicate to the course community that they would like assistance with something relevant to the course (Fong et al., 2023). Previous work has studied the benefits of help-seeking in course discussion forums via strategies like manually coding forum posts according to the type of help-seeking they exhibit (Jay et al., 2020), which is valuable but time consuming and limited in terms of application to practice (i.e., instructors manually coding their forums). Other constructs, such as emotions, have been studied with the help of artificial intelligence (AI) tools that automate the coding process (Crossley et al., 2017). In this paper, we explore AI-driven analysis of help-seeking as well: specifically, recent large language models (LLMs). If successful, LLMs could provide opportunities to study and adapt to help-seeking at a much larger scale than manual efforts in course forum contexts.

LLMs work by learning to predict what word (or part of a word) is most likely to come next in a piece of text, given the preceding context. They are then often further trained (i.e., *fine-tuned*) specifically to follow instructions or to answer questions. Doing so enables more straightforward interaction with users and has led to promising results on tasks like automatically grading students' short-answer responses (Kortemeyer, 2023). However, substantial issues can plague applying some common LLMs for such tasks, especially related to privacy concerns that are salient in education. Students, parents, instructors, and other stakeholders may object to transmitting student work (especially open-ended forum discussions that often include identifying and personal information) to online LLM services, even given assurances that the data will not be harvested. Moreover, regulations may restrict the transmission of such data (e.g., the Family Educational Rights and Privacy Act [FERPA] in the United States). Consequently, in this research we focus on freely available LLMs that can be applied without special computing resources, and, more importantly, that are offline—i.e., they do not require transmitting students' discussion forum text anywhere outside of the researcher or instructor's computer.

Given the potential opportunities to study help-seeking in course discussion forums via scalable automatic methods, we investigate the question *how well do current offline LLMs work for automatically detecting help-seeking?* We do so with minimal customization required for the help-seeking construct specifically, with the intention of informing future efforts to measure other constructs without requiring customized methods and models for each one. Simplicity is especially important given the time and monetary costs involved with developing bespoke machine learning models for tasks in educational data (Hollands & Bakir, 2015). The efforts required for detecting help-seeking using LLMs are relatively lower but still substantial, given that results must be measured against a validated coding schema, although this can be postponed until after initial results have face validity (as described next).

Method

In this study, we detected help-seeking events from students' discussion forum posts in an online, introductory-level natural science course at a large public university in the United States. Students in this course participated through a learning management system called LON-CAPA (Learning Online Network with a Computer-Assisted Personalization Approach), which delivers educational content, auto-graded exercises, and discussion forum functionality (Kortemeyer et al., 2003). Forum content included hierarchically threaded conversations between students and, in some cases, the instructor. For the purposes of this study, we consider only the top-level (i.e., initial, non-reply) forum posts from students. Forum posts constituted 5% of students' final grade in the class, which led to 860 posts from 82 students (and a further 70 posts from the instructor).

Many different LLMs, of varying complexity, have been trained on different text corpora; thus, we compared several to understand how much they might vary in their usefulness for measuring help-seeking. Here, we used 5 different fine-tuned offline LLMs to analyze 297 top-level student forum posts. These models were fine-tuned from two different "foundation" models, which are the LLMs trained to predict the next word in large text corpora before fine-tuning for instruction-following and question-answering. Specifically, we used models based on LLaMA 2 and Mistral (Jiang et al., 2023; Touvron et al., 2023), each of which was published with a fine-tuned version that was the one we used in this study. We also evaluated a publicly available alternative ("B" version) fine-tuned LLaMA 2 model and two ("B" and "C") alternatives based on Mistral. These alternatives were fine-tuned on additional text intended to improve their instruction-following properties.

We provided a prompt to each model that included a very short definition of help-seeking in the forum context (i.e., "*A request for help can consist of an explicit question or implicit indication that help is needed.*") and brief instructions to rate the forum post text that followed. These instructions were dramatically shortened and rewritten versus the instructions human raters followed, after we observed that the instructions for human raters were followed poorly by the LLMs. Specifically, LLMs tended to take an overly expansive interpretation of each instruction given, and to confuse long instructions with the forum post itself (despite appropriate delimiters). Hence, a less detailed prompt was more successful.

The prompt included instructions to "*Rate on a 0-9 scale if the following text contains a request for help, where 9 means that it definitely includes a request for help:*" We selected 0–9 because these LLMs generate numbers one digit at a time; if we had used a 1–10 scale, for example, the probability of a "1" and "10" rating would be more difficult to distinguish because a "10" rating would first require generating the "1". We also observed that the distribution of rating probabilities differed substantially, such that in some cases it was unimodal (i.e., one very likely rating) and in others multimodal (i.e., two or more similarly likely but dissimilar ratings). For example, if 3 is the most probable rating but only by a small margin over 7, 8, and 9, then perhaps 3 is an outlier and a higher rating is more appropriate. We rated each post nine times and computed the median as the final rating to avoid such outliers. LLMs were constrained to generate only values in the rating scale via the *Guidance* Python library (Lundberg, 2023). We also generated a brief explanation (approximately 100 words) from the LLMs for each rating. These are not a required part of the measurement process, and do not necessarily represent the actual reason why a rating was given. However, in developing the prompt for help-seeking detection, we found explanations useful for improving the prompt; there is at least some connection between the rating and the explanation, so clear mistakes in the explanation helped inform the prompt text (most of all, leading us to use a very short definition of help-seeking).

Finally, to determine how well the LLMs rated help-seeking, we compared LLM ratings to human expert ratings via an existing coding schema for help-seeking (Jay et al., 2020), which includes four levels: 1) no question or request for help; 2) question asked but no request for help; 3) implicit request for help; and 4) explicit request for help. We compared the models' ratings to human ratings via Spearman's *rho*, and compared models to each other with a *z*-test (Myers & Sirois, 2006). Note that while we tested a few prompt variations to produce reasonable-seeming explanations, we did so on only a few of the posts—moreover, we only compared LLM and human ratings once at the end of the process to avoid cherry-picking a prompt variation that works well on the measure of convergent validity. This also mirrors a typical use case where a researcher might adjust a prompt several times to achieve promising results and only spend time manually coding data if it seems like the LLM measure is promising.

We were particularly interested in LLMs that can run on typical consumer hardware. Hence, all experiments were conducted on a laptop with 16GB memory and an Intel Core i7-1165G7 processor (a four-core, 2.8GHz processor released in 2020). Mistral-based models have 7 billion parameters, while LLaMA-based models exist in several sizes (we used the 13 billion parameter size); both are too large fit in system memory, but can be *quantized* (i.e., reducing the precision of parameters while minimizing loss of generation quality) to fit. All of the code, prompts, links to quantized models, and instructions necessary for running our help-seeking experiments are available (https://github.com/pnb/llm-measurement).

Results and discussion

Help-seeking predictions from the original LLaMA 2 model correlated *rho* = .650 with human ratings, indicating substantial accuracy for the model. This and all correlations were significantly greater than 0; $p < .001$. We compared the other models to this model, which all had accuracy that was lower or statistically indistinguishable (Table 1). Only two models, one fine-tuned LLaMA 2 model and one fine-tuned Mistral-based model, had significantly lower correlations with the human ratings, with *rho* = .506 ($z = 2.653$, $p = .008$) and *rho* = .529 ($z = 2.269$, $p = .023$) respectively. Hence, all models produced ratings substantially better than chance, but choice of model did make a difference in some cases.

We also examined the pairwise correlations between the individual models to understand whether models were wrong in a similar way (which may suggest a common cause, such as a misleading element of the prompt) versus wrong in different ways (which might suggest errors were more due to the models). The highest correlation between all pairs of models was *rho* = .731, which was between the original LLaMA 2 model and the original Mistral model. The lowest was *rho* = .415, between the LLaMA 2 B version and the Mistral B version, and the mean of all correlations was *rho* = .566. Thus, models' ratings of help-seeking were only correlated with each other to approximately the same degree as those ratings correlated with human ratings. Hence, there may be opportunities for improvement in the LLMs themselves.

One consequence of having several measures that are not strongly related to each other is that, if averaged together, their errors cancel each other out to some extent (as opposed to correlated errors, which remain after averaging). Thus, we also analyzed a model consisting of the average rating of all five LLMs. The result (*rho* = .711) was not significantly better than the LLaMA 2 model ($z = 1.361$, $p = .173$), but certainly at least as accurate as the best model and suggestive of improvements that could be made with more (rather than only better) models.

Finally, we also analyzed explanations from incorrect predictions the LLaMA 2 model made as examples of limitations in the method that could inform future improvements. Specifically, we selected the forum posts with the highest LLM help-seeking rating given the lowest human rating, and the lowest LLM rating given the highest human rating. Two forum posts satisfied the first criteria; these posts represent "false positive" cases where the model identified help-seeking that was not there. In the first such post, the student outlined how to solve a problem, describing it as "tricky," which the LLM explanation identified as an indication that the student might need help (though they did not). In the second false positive case, a student again outlined how to solve a problem and ended with "what am I doing wrong?", which the LLM explanation identified as an indicator of help-seeking. This latter case may indicate an improvement that could be made to the help-seeking prompt, which does indicate that a question qualifies as help-seeking, whereas the coding schema from Authors (2020) differentiates between asking a question with vs. without recognizing the community (i.e., peers in the discussion forum) as an essential ingredient in help-seeking. In contrast, there was one forum post that had the lowest LLM help-seeking rating given the highest human rating; this "false negative" case involved the student recognizing the community by asking if anyone could verify the correctness of their understanding, which the LLM explanation incorrectly indicated was "asking for confirmation, rather than seeking assistance." Improvements to the help-seeking prompt may help in this case, but may also require newer, more capable LLMs to process prompts more accurately—in this study, when given the entire coding schema from Authors (2020), models consistently confused parts of the coding schema with the forum text despite delimiters (including delimiters officially supported by the models).

Table 1
Results of help-seeking detection by different LLMs. The z and p-values indicate difference in rho versus the first model.

Model	Spearman's *rho*	*z*	*p*-value
LLaMA 2	.650		
LLaMA 2 (fine-tuned version B)	.506	2.653	.008
Mistral	.644	0.128	.898
Mistral (fine-tuned version B)	.570	1.562	.118
Mistral (fine-tuned version C)	.529	2.269	.023

Conclusion

In this study, we sought to determine how well current offline LLMs work for detecting help-seeking in an online college course discussion forum. The results indicated that it was indeed feasible, with a medium-to-strong association between LLM ratings and human ratings of help-seeking. Moreover, it was possible to make these automatic ratings without transmitting students' forum posts to online LLM services, and using only typical laptop

hardware. Our approach also offered advantages over traditional machine learning approaches, which require large amounts of manually coded data to train a model and more to test it, whereas the approach in this study requires only data to test the model—and even then, only if the model appears worth testing. The advantages of this method thus open up new use cases for automatic analysis of course discussion forums with respect to help-seeking and perhaps other constructs for which effective prompts can be written. For example, automatic analysis of help-seeking in forums could provide statistical power to detect the size of effects that may be expected (Fong et al., 2023), power to analyze proportionally small groups in large courses, and opportunities for instructors of large courses to find forum posts where their input may be the most helpful.

There remains work to be done with this approach as well. LLMs are trained on vast amounts of text that represent the cultures, languages, and linguistic styles of some students much more than others; hence, research is needed to determine and counteract potential systematic biases in the help-seeking ratings generated by LLMs. Additionally, as the landscape of LLMs shifts rapidly, new offline LLMs will need to be compared to determine whether they are better able to follow prompts, and especially to implement an entire coding schema given as part of the prompt—something that larger models, such as GPT-4, can perhaps do more effectively. However, current results are already promising and applicable, yielding immediate opportunities for the study and improvement of online course discussion forums.

References

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods*, *49*(3), 803–821. https://doi.org/10.3758/s13428-016-0743-z

Fiock, H. (2020). Designing a community of inquiry in online courses. *The International Review of Research in Open and Distributed Learning*, *21*(1), 135–153. https://doi.org/10.19173/irrodl.v20i5.3985

Fong, C. J., Gonzales, C., Hill-Troglin Cox, C., & Shinn, H. B. (2023). Academic help-seeking and achievement of postsecondary students: A meta-analytic investigation. *Journal of Educational Psychology*, *115*(1), 1–21. https://doi.org/10.1037/edu0000725

Hollands, F., & Bakir, I. (2015). *Efficiency of automated detectors of learner engagement and affect compared with traditional observation methods* (pp. 1–37). Center for Benefit-Cost Studies of Education, Teachers College, Columbia University. https://repository.upenn.edu/cbcse/4

Jay, V., Henricks, G. M., Anderson, C. J., Angrave, L., Bosch, N., Williams-Dobosz, D., Shaik, N., Bhat, S., & Perry, M. (2020). Online discussion forum help-seeking behaviors of students underrepresented in STEM. *Proceedings of the 14th International Conference on Learning Sciences (ICLS 2020)*, 809–810.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). *Mistral 7B*. https://doi.org/10.48550/arXiv.2310.06825

Kortemeyer, G. (2023). *Performance of the pre-trained large language model GPT-4 on automated short answer grading* (arXiv:2309.09338). arXiv. https://doi.org/10.48550/arXiv.2309.09338

Kortemeyer, G., Albertelli, G., Bauer, W., Berryman, F., Bowers, J., Hall, M., Kashy, E., Kashy, D., Keefe, H., Behrouz, M.-B., Punch, W. F., Sakharuk, A., & Speier, C. (2003). The learning online network with computer-assisted personalized approach (LON-CAPA). *Computer Based Learning in Science (CBLIS 2003)*, 119–130.

Lundberg, S. M. (2023). *A guidance language for controlling large language models* [Python]. Microsoft. https://github.com/microsoft/guidance

Myers, L., & Sirois, M. J. (2006). Spearman correlation coefficients, differences between. In *Encyclopedia of Statistical Sciences*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0471667196.ess5050.pub2

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open foundation and fine-tuned chat models*. https://doi.org/10.48550/arXiv.2307.09288

Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education through Grant R305A180211 to the Board of Trustees of the University of Illinois. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education