Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025 ASEE
Paper ID #48147

Democratizing the Analysis of Unprompted Student Questions Using Open-Source Large Language Models

Brendan Lobo, University of Toronto

An MASc candidate in the Integrative Biology and Microengineered Technologies Laboratory at the University of Toronto.

Sinisa Colic, University of Toronto

Sinisa Colic is an Assistant Professor, Teaching Stream with the Department of Mechanical and Industrial Engineering. He completed his PhD at the University of Toronto in the area of personalized treatment options for epilepsy using advanced signal processing techniques and machine learning. Sinisa currently teaches several courses at University of Toronto covering a broad range of topics in mechatronics, data science and machine learning / deep learning.

Chirag Variawa, University of Toronto

Prof. Chirag Variawa is the Director, First Year Curriculum, and Associate Professor, Teaching-stream, at the Faculty of Applied Science and Engineering, University of Toronto. He received his Ph.D. from the Department of Mechanical and Industrial Engineering, and his B.A.Sc. in Materials Science Engineering, both from the University of Toronto. His multidisciplinary teaching and research bring together Engineering Education and Industrial Engineering to identify and mitigate learning barriers for diverse student populations.

Democratizing the Analysis of Unprompted Student Questions Using Open-Source Large Language Models

Analyzing student questions can help instructors make informed pedagogical improvements by providing a better understanding of student thinking. In past literature, the analysis of student questions (SQs) has primarily been conducted using taxonomic categorization [1]. These taxonomies focus on various aspects of learning. For instance, utilizing taxonomies based on Bloom's taxonomy [2] can reveal what cognitive levels students are utilizing or struggling with. On the other hand, the taxonomy proposed by Scardamalia and Bereiter in 1992 [3] can be used to determine how familiar students are with a certain topic. Consequently, the application of differing taxonomies allows for a fuller understanding regarding the performance and struggles of the student body. Modern university courses have access to a wealth of student questions in the form of online course discussion boards. These platforms (e.g. Piazza, Blackboard Discussions, Canvas Discussions, etc.) allow students to pose class-wide questions that can be answered by instructors and/or their peers. However, while these discussion boards provide student questions, the questions are not guided or structured for a given taxonomy. In order to glean any insight from them using taxonomic categorization, instructors would have to undertake the tedious but non-trivial task of recategorizing the entire corpus of discussion board entries for every taxonomy they wish to use. This effectively gatekeeps instructors from leveraging these discussion boards in this manner if they do not have surplus time and human labour required for manual question categorization. This problem is magnified in first-year engineering courses which tend to have the largest enrollment sizes with students who are the most unfamiliar with the content. A potential approach to overcome this hurdle of laborious cognition involves leveraging large language models (LLMs) that have been pre-trained on expansive datasets. These models have the potential to categorize questions into a wide variety of taxonomies without rigorous fine-tuning due to their advanced language comprehension. This not only makes LLMs a versatile solution but also an accessible solution as instructors, who may not be experts in natural language processing (NLP), can still utilize these models with the help of simple tutorials or guided code snippets.

Due to the current gap in methodologies using pre-trained LLMs "as-is" to taxonomically categorize student questions, this exploratory paper aims to propose a structured and accessible procedure to fulfill this task. The LLMs utilized will be limited to open-source offline models that do not require the sharing of sensitive data in order to maximize the number of instructors that would be permitted to use this procedure. Pre-trained, but not finetuned, models are used as they allow for an approach that does not require large amounts of labelled data or high levels of programming expertise. This preliminary paper will focus on the implementation of an engineering-focused question taxonomy based on that proposed by Goldberg et al. in 2021 [4] to find areas of struggle for the student body in APS106H1, an introductory Python course for first-year engineering students at the University of Toronto. The scope of this pilot will be limited to using LLMs to categorize student questions from a course discussion board with this taxonomy.

Manual Taxonomic Categorization

The course discussion board for APS106H1 in the winter semester of 2017 contained a total of 785 entries. The data is downloaded from Piazza's servers as a CSV file and is anonymized. The script used to anonymize the entries also rectifies HTML artifacts using the corresponding ANSI

text found in Table 1. Since the exported data does not tag individual entries as questions, a researcher who is a subject matter expert on the course content sorts the entries as either containing a question or not containing a question. In total, 286 entries were identified as containing questions with the rest being non-questions. The researcher then categorizes this subset according to an altered taxonomy, found in Table 2, based on that outlined by Goldberg et al in 2021 [4]. The original taxonomy is altered for two reasons. This first is to isolate student questions from questions instructors sometimes embed with their responses (e.g. "… Think about this: do you want the evaluation to occur first or the execution? …"). This is done by injecting a short phrase shown italicized in the second column of Table 2. This kind of alteration should be done to all taxonomies that do not account for instructor questions in the data if the instructor questions are not removed from the data prior to categorization. The second alteration is to mitigate false positives by creating a "catch-all" category. This should be done for all taxonomies that do not claim to account for all possible questions. While the catch-all category can usually be added as an additional category, it has been merged with the "Question is unspecific" category in this taxonomy, indicated by the underlined text in the second column of Table 2. This is because Category 6 from the original taxonomy is an artifact of the prior study's instruction for students to ask questions that will fit into the five other categories [4]. The number of questions identified for each of the six altered categories can be found in the fourth column of Table 2.

Automated Taxonomic Categorization

The LLMs used in question categorization are all pre-trained downloadable models from HugingFace's model repository [5]. The process of taxonomic categorization with LLMs mimics the manual process. As such, the first step is to use LLMs to identify entries containing questions. Zero-shot classification, an NLP task where a language model classifies a passage with no context outside the classification labels [6], is an efficient solution for this task. Using this approach, a model asked to classify a discussion board entry as either a "question" or a "non-question". An entry is only labelled as a question if the model has confidence of 95% or more in classifying it as a question. Otherwise, the entry is labelled as a nonquestion. The models considered for this task are chosen due to their popularity within the HuggingFace platform and their compatibility with the HuggingFace "zero-shot-classification" pipeline. The models used were selected for diversity in architecture as opposed to diversity in training protocols or model size. The chosen models include: MoritzLaurer/deberta-v3-large-zeroshot-v2.0 (DeBERTa V3) [7], [8], MoritzLaurer/bge-m3-zeroshot-v2.0 (BGE-M3) [8], [9], joeddav/xlm-roberta-large-xnli (XLM-RoBERTa) [10], [11], cross-encoder/nli-MiniLM2-L6-H768 (MiniLMv2) [12], [13], and facebook/bart-large-mnli (BART) [14], [15]. The F1 scores (the harmonic mean between recall and precision) [16] and accuracy scores of these models against the manually sorted data are presented in Fig. 1a.

The next step is to categorize the identified questions according to the revised taxonomy from Table 2. To do this, a conversational prompt approach is used where a LLM is informed of its role as a data analyst, is provided with the course description, is provided with the taxonomy for categorization, and is instructed to return the corresponding category number for a provided question. The pre-trained LLMs used for this step are selected from popular text generation models available on HuggingFace. The selection is narrowed by enforcing a maximum model size of 12 GB due to limitations in available memory.

Table 1: Unicode replacements for HTML artifacts

HTML Artifact	â€™	Â	Ã±	&amp;	&gt;	&lt;	€“
ANSI Replacement	’	*whitespace character*	n	&	>	<	-

Table 2: Alterations done to the taxonomy proposed by Goldberg et al. along with the corresponding course statistics

Original Taxonomy	Altered Taxonomy	Cat. Num.	Ct.
Question is about a definition	Question is *asked by a student and* is about a definition	1	18
Question is about how to do something	Question is *asked by a student and* is about how to do something	2	12
Question is about how to do something if conditions of the problem changed	Question is *asked by a student and* is about how to do something if the conditions of the problem changed	3	9
Question is about understanding how or why something happens	Question is *asked by a student and* is about understanding how or why something happens	4	63
Question extends knowledge to a new circumstance beyond that of the problems solved in class	Question *is asked by a student and* extends knowledge to a new circumstance beyond that of the problems solved in class	5	2
Question is unspecific	Question is unspecific or does not match any other category	6	182

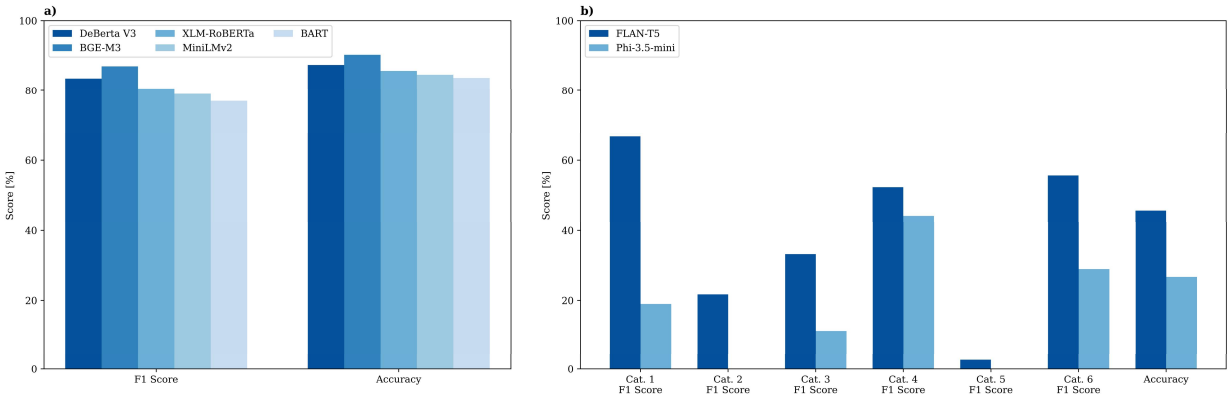

Figure 1. LLM performance comparisons for various tasks in the taxonomic question categorization procedure. a) a comparison between zero-shot classification models used to identify entries that contain questions from those that don't. b) a comparison between conversational models used to identify and taxonomically categorize student questions.

After taking into account the aforementioned privacy considerations, the models evaluated include: google/flan-t5-xl (FLAN-T5) [17], [18] and microsoft/Phi-3.5-mini-instruct (Phi-3.5-

mini) [19], [20]. In order to avoid compounding errors, the manually labelled questions are fed into these models as opposed to those identified in the zero-shot classification task. The performance of the categorization models is shown in Table 4 using per-category F1 scores and overall accuracy when compared against the manually labelled data.

Insights From Using the Taxonomy

When analyzing the categorized questions, several trends become apparent. The first is that most questions (excluding the catch-all Category 6) pertain to asking how something works or why a certain result is achieved. Taking a closer look at questions in Category 4 reveals that students often ask why code snippets from the textbook examples behave in the way that they do. Instructors could use this knowledge to augment the code explanations provided in the textbook or inform their choice when choosing reference texts in the future. The next largest category concerns definitions. Considering that most of these questions are asked in the first two months of the semester, the instructors could reference these questions to improve their introductory lessons which are usually catered towards students who do not have prior programming experience.

Viability of Using Large Language Models for Question Categorization

The LLMs used to identify discussion board entries containing questions performed well achieving F1 scores above 0.8 and accuracy scores above 80%. Specifically, the models based on DeBERTa V3 and BGE-M3 performed exceptionally well considering that they were not fine-tuned for this task. The discrepancies present are likely due to rhetorical questions used by instructors as teaching tools within their responses to student questions. However, the LLMs used to categorize the questions are less impressive. The likely reason for this is that the models used are relatively small compared to the current state-of-the-art generative models. While FLAN-T5 and Phi-3.5-mini models have about 3 billion and 3.8 billion parameters respectively, models such as Meta Llama 3 or DeepSeek-R1can have over an order of magnitude more parameters. However, since the price for the components required to operate these models with hardware acceleration is well beyond the budget of an average consumer, these LLMs were not considered. Having said that, Goldberg et al. observed that the agreement between faculty raters when labelling questions was only about 65%. While the accuracy scores in Table 4 show that agreement between the LLMs and the manual labelling is well below this level, the questions analyzed are unprompted and unguided unlike those that Goldberg et al. studied.

Future Work

The immediate next steps for this work entail improvements to the current methods employed. On the validation side, this involves using multiple human raters to mitigate potential bias in the ground truth. Concerning the usage of LLMs, the tasks of identifying questions, specifically isolating student questions, and taxonomically categorizing student questions can be merged in different ways to identify the most accurate and robust solution. Further work would involve evaluating additional NLP techniques (e.g. few-shot classification) for the various tasks. Building this foundation in LLM-powered taxonomical question categorization would also pave the way to further process the discussion board questions into digestible data for instructors (e.g. FAQ lists, timelines correlating question type and frequency to course events [21], etc.).

References

[1] C. Chin and J. Osborne, "Students' questions: a potential resource for teaching and learning science," *Stud Sci Educ*, vol. 44, no. 1, pp. 1–39, 2008, doi: 10.1080/03057260701828101.

[2] B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, and D. B. Krathwohl, *TAXONOMY OF EDUCATIONAL OBJECTIVES; The Classification of Educational Goals*. 1956.

[3] M. Scardamalia and C. Bereiter, "Text-Based and Knowledge Based Questioning by Children," *Cogn Instr*, vol. 9, no. 3, pp. 177–199, Sep. 1992, doi: 10.1207/S1532690XCI0903_1.

[4] S. R. Goldberg, C. Venters, and A. Masnick, "Refining a Taxonomy for Categorizing the Quality of Engineering Student Questions," *ASEE Annual Conference and Exposition, Conference Proceedings*, Jul. 2021, doi: 10.18260/1-2--37649.

[5] "Hugging Face – The AI community building the future." Accessed: Jan. 14, 2025. [Online]. Available: https://huggingface.co/

[6] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot Learning with Semantic Output Codes," *Adv Neural Inf Process Syst*, vol. 22, 2009.

[7] M. Laurer, "MoritzLaurer/deberta-v3-large-zeroshot-v2.0," Hugging Face. Accessed: Feb. 20, 2025. [Online]. Available: https://huggingface.co/MoritzLaurer/deberta-v3-large-zeroshot-v2.0

[8] M. Laurer, W. van Atteveldt, A. Casas, and K. Welbers, "Building Efficient Universal Classifiers with Natural Language Inference," Dec. 2023, Accessed: Feb. 20, 2025. [Online]. Available: https://arxiv.org/abs/2312.17543v2

[9] M. Laurer, "MoritzLaurer/bge-m3-zeroshot-v2.0," Hugging Face. Accessed: Feb. 20, 2025. [Online]. Available: https://huggingface.co/MoritzLaurer/bge-m3-zeroshot-v2.0

[10] J. Davison, "joeddav/xlm-roberta-large-xnli," Hugging Face. Accessed: Feb. 19, 2025. [Online]. Available: https://huggingface.co/joeddav/xlm-roberta-large-xnli

[11] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Nov. 2019, doi: 10.18653/v1/2020.acl-main.747.

[12] N. Reimers, "cross-encoder/nli-MiniLM2-L6-H768," Hugging Face. Accessed: Jan. 14, 2025. [Online]. Available: https://huggingface.co/cross-encoder/nli-MiniLM2-L6-H768

[13] W. Wang, H. Bao, S. Huang, L. Dong, and F. Wei, "MiniLMv2: Multi-Head Self-Attention Relation Distillation for Compressing Pretrained Transformers," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2140–2151, Dec. 2020, doi: 10.18653/v1/2021.findings-acl.188.

[14] J. Chaumond, “facebook/bart-large-mnli,” Hugging Face. Accessed: Jan. 14, 2025. [Online]. Available: https://huggingface.co/facebook/bart-large-mnli

[15] M. Lewis *et al.*, “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Oct. 2019, doi: 10.18653/v1/2020.acl-main.703.

[16] D. M. W. Powers, “EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION,” *Journal of Machine Learning Technologies* , vol. 2, no. 1, pp. 37–63, Feb. 2011, Accessed: Feb. 19, 2025. [Online]. Available: http://www.bioinfo.in/contents.php?id=51

[17] A. Zucker, “google/flan-t5-xl,” Hugging Face. Accessed: Feb. 20, 2025. [Online]. Available: https://huggingface.co/google/flan-t5-xl

[18] H. W. Chung *et al.*, “Scaling Instruction-Finetuned Language Models,” Oct. 2022, Accessed: Feb. 20, 2025. [Online]. Available: https://arxiv.org/abs/2210.11416v5

[19] A. Garg, “microsoft/Phi-3.5-mini-instruct,” Hugging Face. Accessed: Feb. 20, 2025. [Online]. Available: https://huggingface.co/microsoft/Phi-3.5-mini-instruct

[20] M. Abdin *et al.*, “Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone,” Apr. 2024, Accessed: Feb. 20, 2025. [Online]. Available: https://arxiv.org/abs/2404.14219v4

[21] B. Lobo, J. Chen, S. Colic, and C. Variawa, “Excuse me, but can you repeat the question? Identifying correlations between thousands of engineering student questions and their impacts on the content and delivery of course materials,” *Proceedings of the Canadian Engineering Education Association (CEEA)*, Mar. 2024, doi: 10.24908/pceea.2023.17070.