MathMentorDB: A Massive Dataset of Authentic, Mathematics Tutoring Dialogues

Anonymous ACL submission

Abstract

Education researchers have traditionally relied on classroom data, such as video recordings and audio transcripts, to study teaching and learning. However, collecting and analyzing these data is often time- and resource-intensive, limiting the scope and generalizability of these findings. We address this gap by introducing *MathMentorDB*, a large-scale dataset of tutoring dialogues derived from an online Mathematics Discord server. Comprising 5.4 million messages across 200,332 conversations involving 43,249 unique users, *MathMentorDB* offers structured insights into authentic student-tutor interactions at an unprecedented scale, capturing both student learning processes and real-time tutoring strategies that take place in this widely available, publicly visible space. Integrating *MathMentorDB* with LLMs enables the identification of insights difficult to discern in smaller, more homogenous datasets. To demonstrate this, we explore question-asking patterns by mathematics students using pre-trained models to classify messages and categorize question types. The comprehensive capture of teaching expertise alongside student interactions positions *MathMentorDB* as a valuable resource for advancing NLP research and potential to develop better applications that enhance learning interactions. By offering a vast repository of authentic mathematics tutoring dialogues, *MathMentorDB* addresses limitations of traditional data collection methods and opens up new possibilities for large-scale education research and the development of NLP-assisted tutoring systems.

1 Introduction

The widespread use of smartphones and the internet among learners and educators has transformed the teaching and learning of mathematics (Engelbrecht et al., 2020). The information age has ushered in new modalities of learning where thousands of students can learn together in MOOCs, interactive learning applications, and online learning forums, all of which have become fundamental in reshaping education at all levels (Haleem et al., 2022). This shift necessitates research that goes beyond trying to understand what happens in traditional classroom settings, as students worldwide are turning to a diverse set of online resources for support with their coursework. This trend is not confined to a single educational platform; it is a widespread phenomenon seen across countless online learning environments. Van de Sande's (2011) study of an open, online calculus help forum provides one such example of an open-access, public platform where students from a wide range of backgrounds and abilities can ask questions and engage in subject-specific discussions.

When students transition from secondary to university mathematics, they find themselves at a challenging point where the familiar structures of academic support recede, forcing many of them to navigate their mathematics coursework more independently (Vollstedt et al., 2014). The study of this transition, often referred to as the *secondary-tertiary transition* (STT), has been studied extensively by scholars in mathematics education. Gueudet's (2008) review of the scholarship in this area highlights three main transitions involving individual, social, and institutional factors: the transition in ways of thinking, the transition to proof and the technical language of mathematics, and the institutional transition related to changes in the didactical contract. Furthermore, Bettinger and Long (2009) emphasize that many students entering university are academically under-prepared for such rigor, necessitating additional remedial support. Although students still engage with traditional forms of academic support, such as office hours, one-on-one tutoring, and discussion sections, many encounter barriers such as inflexible scheduling, financial constraints, and a hesitation to interact directly with authority figures, which can limit their

access to these resources (Pepin, 2014).

In recent years, the role technology plays in educational settings has become increasingly evident, requiring both teachers and students to have to adapt their methods for how they teach and learn (Escueta et al., 2017). As traditional classroom teaching approaches encounter their limitations in addressing diverse learning styles and adapting to a rapidly changing world, the need for innovative, technology-driven learning models has become more pronounced (Hiebert and Grouws, 2007). For example, research examining college students' attitudes and habits towards academic help seeking has indicated that students are using informal help-seeking options more than formal channels, as perceived threat is swaying students from approaching instructors for help (Qayyum, 2018). The adjustment towards the increased use of technology not only offers some new avenues for engagement and interaction for learners and educators that align with some of the demands of the 21st century, but also can help provide researchers with the means to study the ways in which students and teachers engage around content in these environments (Cohen et al., 2003).

This paper presents the construction and dissemination of the *MathMentorDB* dataset, a carefully curated, large-scale collection of over 5.4 million messages across 200,332 conversations involving 43,249 unique users. Derived from the Mathematics Discord Server (MDS, hereafter), a dynamic and collaborative space where students and peer mentors from across the world engage in mathematical discourse, the dataset captures a wide range of interactions, from homework help to open-ended discussions on specific mathematics topics. Our work offers the following contributions to the field: (1) The creation of a novel, large-scale dataset of authentic mathematics tutoring conversations, bridging the gap between educational research and NLP communities. (2) Development of an efficient methodology for extracting, cleaning, and disentangling conversational data from Discord servers for educational research purposes. (3) Application of large language models and text classification techniques to identify patterns in student question-asking behaviors at an unprecedented scale. (4) Insights into the dynamics of online mathematics tutoring and learning in a global, collaborative environment.

The creation of *MathMentorDB* involved a complex process of data extraction, cleaning, and conversation disentanglement from various channels devoted to providing individual help to students on homework problems. The resulting dataset provides a unique window into the real-time, dynamic process of learning and providing help with mathematical content at an unprecedented scale. By preserving the context and flow of these interactions, *MathMentorDB* offers a unique resource that combines the authenticity of real-world student-tutor interactions with the scale necessary for advanced machine learning applications. We argue the dataset's scale, diversity of topics, and authentic representation of student-tutor interactions from a global participant pool position it as an invaluable resource for a wide range of researchers working on problems related to teaching and learning, particularly in the intersection of NLP, education, and AI-assisted tutoring systems.

2 Related work

2.1 Mathematics Datasets in Education and NLP Research

Contemporary research and practice in mathematics education increasingly rely on classroom recordings, including video and audio transcripts. These data are fundamental for understanding important aspects of instruction (Major and Watson, 2018; Kim et al., 2023), and play a role in developing and training NLP models to train AI tutors (Malik et al., 2024). The construction and utilization of high-quality datasets in education research have become increasingly important, particularly in mathematics education. This importance is highlighted by the fact that the current iteration of baseline industry models (e.g., OpenAI's ChatGPT) are not yet capable of consistently delivering high quality calculations or proofs (Frieder et al., 2023).

Several notable projects exemplify advancements in this area. The "Million Tutor Moves Observatories Project" aims to crowd-source one million high-quality machine-readable student-tutor interactions to train intelligent tutoring systems tailored for diverse learning needs (Reich et al., 2023). This project is currently developing its dataset through an NSF mid-scale research infrastructure grant. In another study, Demszky et al. (2021) developed a framework for computationally measuring teachers' conversational uptake, releasing a dataset of 2,246 student-teacher interactions from US math classroom transcripts, annotated for uptake by domain experts. Additionally, the Talk-

moves dataset by Suresh et al. (2022) comprises 567 annotated transcripts from K-12 mathematics lessons, both in-person and online, providing a rich source for analyzing teacher and student discourse.

The main goal of this work is to identify **mathematics teaching** knowledge, addressing a challenge that extends beyond merely solving mathematics problems (see work by Ball et al. (2008) for more details). When developing student-facing applications, it is important to tap into the tacit knowledge that expert teachers and tutors possess when guiding students work through mathematics problems. This requires a deeper understanding of what constitutes "mathematical knowledge for teaching" as defined by experts in the field of mathematics education (Ball et al., 2005). Such specialized knowledge differs significantly from pure mathematical knowledge, encompassing pedagogical content knowledge, understanding of common misconceptions, and strategies for explaining complex concepts in accessible ways. By developing datasets that capture and better represent this specialized knowledge, we can create more effective models that not only solve mathematical problems but also emulate the nuanced approaches expert educators use in guiding students through their misconceptions and other various stages of learning.

While these real-world datasets provide valuable insights, the scale required for training fine-tuned LLMs often necessitates larger datasets. Recent work focusing on improving the mathematical reasoning of LLMs has led to the creation of massive, AI-generated mathematical problem-solution pairs. For example, NVIDIA's OpenMathInstruct-1 (Toshniwal et al., 2024) comprises 1.8 million such pairs, surpassing similar work by others such as $L\tilde{I}LA$ (Mishra et al., 2022), MetaMathQA (Yu et al., 2024), and MAmmoTH (Yue et al., 2024).

These synthetic datasets aim to address the current limitations of current open-source LLMs in mathematics problem-solving. As noted by Yu et al. (2024) in their work on MetaMathQA, "most existing open-source LLMs (e.g., LLaMA-2) are still far away from satisfactory for solving mathematical problems due to the complex reasoning procedures" (p. 1). Their approach of bootstrapping mathematical questions by rewriting them from multiple perspectives has shown promising results, with their MetaMath-70B model achieving accuracy comparable to GPT-3.5-Turbo on certain benchmarks. Specifically, their MetaMath-7B model achieved 66.5% accuracy on GSM8K and 19.8% on MATH, surpassing state-of-the-art models of the same size by significant margins. These datasets, generated and fine-tuned with both closed- and open-source pre-trained models, represent a significant step towards creating more capable and accessible AI systems for mathematics problem-solving. The development of these synthetic datasets offers the potential to enhance the mathematical reasoning capabilities of LLMs, potentially leading to more effective AI-assisted tutoring systems and educational tools.

However, while synthetic datasets offer the advantage of scale and can be tailored to cover a wide range of mathematical concepts, they may lack the nuanced, contextual richness found in authentic student-tutor interactions. The real-world complexity of mathematical discourse, including misconceptions, scaffolding, and the tacit knowledge of expert educators, is challenging to replicate synthetically. There is an additional step required to take in order to effectively transform a system that can solve mathematics problems into one that can effectively, and equitably teach that content to others. This is where our *MathMentorDB* dataset offers a unique contribution to the field. By providing a large-scale, machine-readable collection of authentic mathematics tutoring conversations, *MathMentorDB* bridges the gap between the authenticity of smaller, expert-annotated datasets and the scale of synthetic datasets. It captures the real-world complexity of mathematical discourse, including the nuanced approaches expert tutors use to guide students through misconceptions and learning stages. At the same time, its scale, comprising over 5.4 million messages across 200,332 conversations, provides the volume of data necessary for building research that can help train and fine-tune large language models working towards these goals.

2.2 Conversation Disentanglement

Most chat disentanglement models have been based on chats extracted from IRC channels (Kummerfeld et al., 2019; Shen et al., 2006), but until 2019 it was done with small datasets (less than 2500 messages) or non-open source datasets. Kummerfeld et al. (2019) manually annotated dataset of 77,563 messages for conversation disentanglement along with a model, which has served as a training set for many studies working to develop their own disentanglement models (see Chatterjee et al. (2020); Liu and Cohen (2021)). The researchers trained a model with some engineered features at

the message and pairwise message level. Some message level features included timestamps and answers to questions such as: how long ago this user last wrote, did the same user write right before or after, are they a bot, is this message targeted, and was the previous message targeted. Additionally, some pairwise message features included number of intervening messages or answers to questions like: is the user the same, did the user address another user, and do the two messages have the same target? Their results were state of the art at the time, establishing a benchmark for subsequent research in the field.

Recent advancements in conversation disentanglement have explored various techniques, including zero-shot learning (Chi and Rudnicky, 2023), unsupervised methods (Liu et al., 2021), and online disentanglement (Yu and Joty, 2020). These approaches have pushed the boundaries of what's possible in conversation analysis, particularly in unstructured chat environments. Our approach to conversation disentanglement with the *MathMentorDB* dataset offers several distinct advantages for educational research. By taking advantage of the structured nature of the Mathematics Discord Server, we achieve a nearly perfect degree of accuracy in delineating the conversations (via an inter-rater reliability check), minimizing the risk of erroneously splitting or combining conversations, a common challenge in traditional disentanglement methods (and one that was faced in the first couple of years of this project's life cycle). This method's scalability is particularly noteworthy, allowing for efficient processing of large volumes of data, which is crucial for the analysis of extensive educational datasets like *MathMentorDB*.

Another key strength of this approach lies in its simplicity, making it accessible to researchers across various disciplines, not just those with extensive machine learning expertise. This accessibility broadens the potential applications of the *MathMentorDB* dataset, encouraging interdisciplinary research in education, linguistics, and computer science. Moreover, while developed for a mathematics learning context, this approach demonstrates potential adaptability to other structured online learning environments, opening avenues for comparative studies across different educational domains. By blending the benefits of structured data collection with large-scale analysis capabilities, our work on *MathMentorDB* bridges a critical gap in educational research methodologies. It addresses the limitations of both traditional small-scale studies and the often context-poor large synthetic datasets. This approach not only facilitates the creation of a high-quality, large-scale dataset but also sets a precedent for how educational platforms can be designed to inherently support research endeavors.

3 Methods

In this section, we provide brief details about the Mathematics Discord community, data extraction methods, and insights into how and what was important to capture. Next, we describe information related to the second research question on our uses of two pretrained models: first, using a fine-tuned sequence classification model optimized for distinguishing between questions and statements, and second, hosted LLMs to classify the questions into different question types.

3.1 Building the Dataset

The dataset for this dissertation was sourced from the Mathematics Discord Community, an active online platform facilitated by Discord. The MDS, launched in January 2017, represents a unique convergence of individuals with a shared passion for mathematics. With over 51 million messages exchanged among more than 181,000 registered users (as of October 2024), the server provides a rich source of data for examining contemporary mathematical discourse and learning practices. This server was chosen due to its active engagement and substantial user base, which provide a diverse range of mathematical discussions and interactions. The broad yet focused nature of the community's rules has shaped the discourse, making it an ideal setting for this study. To extract the chat logs, we used the Discord Chat Exporter tool (Holub, 2023). This extraction process happened over multiple iterations, with a final download in January 2023. These messages were exported in JSON format, which offer a hierarchically-structured representation of the conversations, providing easy access to the metadata for each message. These metadata include message IDs, timestamps, message content, attachments, amongst other relevant information. The first author's university Institutional Review Board (IRB) determined this is not human subjects research, as the data is publicly available and granted an exemption from continuous review for the project, acknowledging its adherence to ethical research standards.

3.2 Characteristics about the Data

The dataset contains discrete communiation acts, or *messages*, which include student questions, peer responses, and mentor clarifications. These messages, often rich with mathematical notation and diagrams, form the atomic units of the dataset. When viewed collectively, these messages form *conversations*, encapsulated exchanges that begin with a student's introduction of a problem and conclude with its resolution.

Several useful features characterize the data that are important to consider when building the dataset and structuring metadata for future analysis. For one, the dataset encompasses a wide spectrum of mathematical domains, ranging from high school pre-calculus to graduate-level measure-theory and beyond. Messages are often multi-modal, as students often combine text with screenshots and diagrams. Additionally, the messages are timestamped, which allows for precise tracking of conversation flow and mapping learning trajectories over time. Importantly, there is automated assistance in the server, as there are a couple of bots that manage administrative tasks such as opening and closing threads, and providing activity statistics, based on text-based prompts provided by the users.

In late 2019, there was an administrative change to the server when the introduction of "help-channels" was introduced by the moderators. What this did was realign how the users in the server asked for help; that is, it directed question-asking to these channels that had specific rules and automated assistance from these bots we describe above, reserving topic and domain-specific channels for broader discussions on subject matter. Text-based commands like '.close' signify the end of a conversation, which triggers automated bot marks the conversation as complete, provide a grace period, and queue it among the rest of the available help channels. While the channels are content-agnostic (i.e., one could ask an abstract algebra question in the channel), the advantages of efficiently parsing chat logs into distinct dialogues outweigh any potential drawbacks. Prior to this implementation, much of the first author's doctoral research time was spent using conversation disentanglement methods as described in subsection 2.2. By implementing these norms, not only did it make it much more clear for the participants when it was their turn to ask and answer questions, but to outside researchers, it provided 1) a window to algorithmically disentangle these chat logs at scale and 2) learn from this implementation for future work on building chat-based learning communities. In subsection 4.1, we describe this algorithmic implementation in more detail.

3.3 Modeling Pipeline

This study uses the "Keyword Statement vs. Question Classifier for Neural Search" model developed by the Haystack team, which was designed to distinguish between real questions and keyword queries. This is useful as a preliminary step as in this dataset, a message might ask, 'Can someone help me understand how to find the zeros of this polynomial?' while another might simply say, 'Thanks, I got it.' Using this classifier provides a way to efficiently filter through the messages, distinguishing questions from other message types. The classification model takes in each message as an individual data point, and processes it through the classification function in sequence. Subsequently, the outputs are reincorporated into the dataset, now paired with labels that state whether the message is a 'question' or a 'statement'. With the messages categorized, the dataset can be filtered for the next step of analysis to look at the types of questions posed in the conversations. The results of the classification are shown later in Table 3; briefly, around 20% (approximately 1 million) of the messages were questions.

To address the second research question, which examines the types of questions present in the conversations in *MathMentorDB*, this part of the study builds on the foundational work of (Graesser and Person, 1994). Their seminal work on question asking during tutoring sessions provides a detailed framework for decomposing mathematics students' questions into detailed aspects of presupposition and focus, as elaborated in their prior work (Graesser et al., 1992). Their framework provides guidance on how to classify questions into 18 distinct types, capturing a range of both cognitive and linguistic diversity present while students ask questions about mathematics problems. The original taxonomy of 18 question types devised by Graesser et al. presents an extensive assortment, from ones that elicit short answers like verification (e.g., 'Is a fact true?'), disjunction (e.g., 'Is it X or Y?'), and quantitative questions (e.g., 'How many?'), to more elaborate ones that require open-ended longer answers, to those that are interpretational, and those that are procedural in nature. However, in examining the dialogues within *MathMentorDB*,

Category	Definition
Basic Inquiry	Seeks a simple, factual response or clarification of basic information.
Procedural Reasoning	Concerns the steps or procedures taken to achieve a certain goal or solve a problem.
Context Inquiry	Questions that require understanding and applying context or scenarios.
Exploratory Inquiry	Aims to explore concepts or definitions, often asking for elaboration or examples.
Assertive Communication	Involves making a statement or claim, often with confidence or certainty, sometimes to persuade.

Table 1: Question Type Coding Framework

I noticed a recurring theme: many questions often fit into several of these categories, blurring the lines of distinction set by the original framework. For example, there were often questions that would elicit an open-ended, yet procedural reply. To adapt Graesser et al.'s (1992) framework for this study, the goal here is to build a more streamlined set of clearly defined categories that would ensure each category was not only distinct but also fully representative of the types of questions posed within the questions posed in *MathMentorDB*.

The Question Type Coding Framework, as illustrated in Table 1, serves as the basis for the coding process. The categories are designed to capture the overarching nature of the questions asked in the conversations found in *MathMentorDB*, in some order of increasing abstraction. With the coding framework in place, the next phase of the study involves applying these categories to the dataset, both through manual annotation and the use of an LLM.

4 Findings

4.1 Contributions (1) and (2)

As cited above in subsection 3.2, the development of an algorithm to disentangle the dataset was pursued to address the complex challenge of parsing and organizing a substantial corpus of educational data, specifically housed within a number of data formats that can handle the hierarchical, nested structure of the conversations. The primary objective was to structure the data in a way such that it could be clear: (1) when the conversation started, (2) who started the conversation, (3) timestamps to identify when messages were sent and the duration of the message, (4) anonymize the usernames with pseudonyms, and (5) label each participant as a student or a helper.

Table 2: Conversation Disentanglement Algorithm

Algorithm 1: Conversation Disentanglement

Input: JSON files with messages
Output: Separate JSON files for each conversation

```
Initialize conversation_id = 1, conversations = {}, author_pseudonyms = {}
for each JSON file do
    Parse JSON file
    for each message in JSON file do
        if message timestamp ≥ cutoff date then
            Assign pseudonym to author
            Create conversation key
            if conversation key not in conversations then
                Initialize new conversation
            end if
            Process message content
            Add processed message to conversation
            if "Channel closed" message detected then
                Increment conversation_id
            end if
        end if
    end for
end for
for each conversation in conversations do
    Format conversation data
    Save conversation to separate JSON file
end for
```

Table 4.1 provides pseudocode for the conversation disentanglement algorithm. A key feature of this algorithm is its utilization of bot-generated 'Channel closed' messages, which serve as markers to delineate individual conversations. These messages are triggered when a user types '.close' or when the channel remains idle for a specified period. Upon encountering a 'Channel closed' message, the algorithm increments a conversation identifier (conversation_id), marking the start of a new conversation thread. This mechanism ensures each conversation is captured as a discrete entity, preserving the natural flow of dialogue. To address privacy concerns, the algorithm incorporates a pseudonymization function. While Discord usernames already provide some anonymity, this additional step ensures that users' personal information is not linked to their messages, maintaining both privacy and conversational integrity. This streamlined approach to conversation disentanglement, leveraging built-in Discord features and bot functionality, allows for efficient processing of large-scale chat data while preserving the context and structure of individual conversations.

4.2 Contributions (3) and (4)

In order to understand what types of questions are being asked in the *MathConverse* dataset, the first step is to distinguish between messages that pose questions and those that do not. This analysis serves a dual purpose: provides a quantitative measure of the proportion of questions versus statements within the dataset and refines the dataset for further in-depth examination of question-based messages.

Classification	Count	Percentage
Questions (True)	995,477	18.25%
Statements (False)	4,457,470	81.75%
Total	5,452,947	100%

Table 3: Classification Results: Questions vs Statements

Statement vs. Question Classifier Results As presented in Table 3, the application of the "Keyword Statement vs. Question Classifier" shows that approximately 20% of the messages, amounting to nearly 1 million, are classified as questions. This ratio is significant as it shows that within the conversations, there are many follow-up questions. The presence of such a large number of questions helps validate the notion that students feel empowered to ask questions in the MDS, aligning with the findings of (Graesser and Person, 1994) who found in their study that student questions were approximately 240 times as frequent in tutoring settings than in classroom settings.

Question classification on random sample We first took a random sample of 500 questions from the prior work done classifying questions from statements and used Label Studio, an open-source data labeling platform, to label the data. 494 out of 500 (98.8%) of the messages were classified correctly according to the graduate student annotator. We then used this time to hand-label 500 questions to compute reliability metrics with the LLM. Table 4 shows the final ground truth values after some calibration between two human coders.

As shown in the table, the most common questions were 'Basic Inquiry' ($\frac{199}{503} = 40.0\%$) and 'Procedural Reasoning' ($\frac{163}{503} = 32.4\%$). Basic Inquiry being the most common result aligns was an expected result, as this matched with the first authors' observations with their doctoral research observing the platform, as many of the student questions in the conversations started off with a question that can be answered quickly (e.g., 'Is $sin(x) + cos(x)$ defined for the entire domain?' or 'Does this reasoning look right?'), or finished off with some sort of clarifying question (e.g., 'Is that it?', 'What answer did you get?', or 'Would I use spherical coordinates here too?'), whether or not the question came from the student or the helper. The more surprising result was what the model was helpful for. Typically, qualitative coding is done by two or more researchers, as it is important to be able to check the work by establishing inter-rater reliability. After labeling the examples and running the messages through the model, the primary researcher assigned to the task assessed each predicted label against their own and found that *they needed to change many of labels in order to be consistent with the definitions they had in their framework*. In doing so, they went to another researcher, and went through the task of relabeling again. Specifically, the model did a much better job at identifying instances of procedural reasoning where the primary researcher was erroneously identifying the messages as exploratory inquiry. Table 4 shows hand-labeled counts, *after doing some calibration*, drawing another random sample, relabeling, and running the model again, which is typical of the cyclical process necessary in this type of work.

Category	Count	Percentage
Basic Inquiry	199	39.56%
Procedural Reasoning	163	32.41%
Contextual Inquiry	59	11.73%
Exploratory Inquiry	50	9.94%
Assertive Communication	32	6.36%
Total	503	100%

Table 4: Counts of Question Types ($n = 503$ questions)

Classification Results and Implications The evaluation metrics presented in Table 5 demonstrate the superior performance of GPT-3.5 Turbo compared to both the ZeroR baseline and the TF-IDF + Logistic Regression model. Notably, GPT-3.5 Turbo shows high precision and recall across all categories, including less common ones, indicating its ability to effectively distinguish between different question types.

This performance is particularly impressive given the nuanced nature of the question categories and the potential for overlap between them. For instance, the model achieves perfect precision (1.0) for Exploratory Inquiry, though with a lower recall (0.38), suggesting it accurately identifies this category when it does so, but may miss some instances.

Table 5: Category-wise Evaluation Metrics ($n = 503$)

	ZeroR		TF-IDF + Log. Reg.		GPT-3.5	
Category	P	R	P	R	P	R
Basic Inq.	.40	1.0	.76	.82	.95	.87
Proc. Reas.	0	0	.71	.68	.80	.93
Context Inq.	0	0	.65	.59	.81	.88
Explor. Inq.	0	0	.58	.52	1.0	.38
Assert. Comm.	0	0	.53	.47	.64	.71

* P = Precision, R = Recall

Question Type	Count	Percentage
Basic Inquiry	51372	42.71%
Procedural Reasoning	40637	33.80%
Contextual Inquiry	15742	13.09%
Exploratory Inquiry	10039	8.35%
Assertive Communication	2462	2.05%
Total	120,252	100%

Table 6: Model predicted question types on a large random sample ($n = 120,252$ questions)

Conversely, it shows balanced high performance for categories like Basic Inquiry and Procedural Reasoning, which are more prevalent in the dataset. To validate the model's performance on a larger scale, we applied it to a random sample of 120,252 questions from the *MathMentorDB* dataset. The results, presented in Table 6, show a distribution of question types that closely aligns with our manually labeled smaller dataset (Table 4). This consistency across different sample sizes provides strong evidence for the model's reliability and scalability.

The distribution of question types in the larger sample offers valuable insights into the nature of mathematical discourse in online tutoring environments. Basic Inquiry and Procedural Reasoning dominate the questions, accounting for 42.71% and 33.80% respectively, indicating a strong focus on fundamental understanding and problem-solving strategies. A significant portion of learners, represented by 13.09% of Contextual Inquiry questions, seek to apply mathematical concepts to real-world scenarios. The presence of Exploratory Inquiry questions (8.35%) reflects a subset of learners engaging in deeper, conceptual exploration. Notably, the low percentage of Assertive Communication (2.05%) aligns with the supportive, question-oriented nature of the tutoring environment.

5 Conclusion

In this paper, we introduced *MathMentorDB*, a large-scale dataset of tutoring dialogues derived from an online Mathematics Discord server. Our dataset, comprising 5.4 million messages across 200,332 conversations involving 43,249 unique users, provides a rich resource for understanding authentic student-tutor interactions at an unprecedented scale. We demonstrated the utility of *MathMentorDB* by exploring question-asking patterns using pre-trained models to classify messages and categorize question types.

The creation of *MathMentorDB* bridges a significant gap between education research and NLP communities by offering a large-scale, authentic dataset of mathematics tutoring conversations. Our methodology for extracting, cleaning, and disentangling conversational data from Discord servers ensures high-quality, structured data, making it a valuable resource for various research purposes. By applying large language models and text classification techniques, we identified patterns in student question-asking behaviors, providing insights into the dynamics of online mathematics tutoring and learning.

Our analysis revealed the prevalence of different question types, highlighting the focus on fundamental understanding and problem-solving strategies in online tutoring environments. The findings from *MathMentorDB* underscore the potential of large-scale, authentic datasets in advancing NLP research and developing better applications that enhance learning interactions. By offering a vast repository of authentic mathematics tutoring dialogues, *MathMentorDB* addresses the limitations of traditional data collection methods and opens up new possibilities for large-scale education research and the development of NLP-assisted tutoring systems.

Future work can build on this foundation by exploring additional aspects of student-tutor interactions, such as the effectiveness of different tutoring strategies and the impact of various types of feedback on student learning outcomes. Furthermore, integrating *MathMentorDB* with other educational datasets could provide a more comprehensive understanding of online learning environments, ultimately contributing to the development of more effective and equitable educational technologies.

5.1 References

References

Deborah Loewenberg Ball, Heather C Hill, and Hyman Bass. 2005. Knowing Mathematics for Teaching. *American Educator*, 29(1):14–17, 20–22, 43–46.

Deborah Loewenberg Ball, Geoffrey C Phelps, and M. H. Thames. 2008. Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, 59(5):389–407. Tex.ids: loewenbergball2008, loewenbergball2008a ISBN: 0022-4871 publisher: SAGE Publications Inc.

Eric P. Bettinger and Bridget Terry Long. 2009. Addressing the needs of underprepared students in higher education: does college remediation work? *Journal of Human Resources*, 44(3):736–771.

Preetha Chatterjee, Kostadin Damevski, Nicholas A. Kraft, and Lori Pollock. 2020. Software-related slack chats with disentangled conversations. In *Proceedings of the 17th International Conference on Mining Software Repositories*, pages 588–592, Seoul Republic of Korea. ACM.

Ta-Chung Chi and Alexander I. Rudnicky. 2023. Zero-Shot Dialogue Disentanglement by Self-Supervised Entangled Response Selection. *arXiv preprint*. ArXiv:2110.12646 [cs].

David K Cohen, Stephen W Raudenbush, and Deborah Loewenberg Ball. 2003. Resources, instruction, and research. *Educational evaluation and policy analysis*, 25(2):119–142. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2021. 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.

Johann Engelbrecht, Salvador Llinares, and Marcelo C. Borba. 2020. Transformation of the mathematics classroom with the internet. *ZDM - Mathematics Education*, 52(5):825–841. ISBN: 0123456789 publisher: Springer Berlin Heidelberg.

Maya Escueta, Vincent Quan, Andre Joshua Nickow, and Philip Oreopoulos. 2017. Education technology: an evidence-based review. Technical Report w23744, National Bureau of Economic Research, Cambridge, MA.

Simon Frieder, Luca Pinchetti, Alexis Chevalier, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, and Julius Berner. 2023. Mathematical Capabilities of ChatGPT. *arXiv preprint*. ArXiv:2301.13867 [cs].

Arthur C. Graesser, Natalie Person, and John Huber. 1992. Mechanisms that generate questions. In *Questions and information systems.*, pages 167–187. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US.

Arthur C Graesser and Natalie K Person. 1994. Question asking during tutoring. *American Educational Research Journal*, 31(1):104–137.

Ghislaine Gueudet. 2008. Investigating the secondary–tertiary transition. *Educational Studies in Mathematics*, 67(3):237–254.

Abid Haleem, Mohd Javaid, Mohd Asim Qadri, and Rajiv Suman. 2022. Understanding the role of digital technologies in education: A review. *Sustainable Operations and Computers*, 3:275–285.

James Hiebert and Douglas A Grouws. 2007. The effects of classroom mathematics teaching on students' learning. In F. Lester, editor, *Second Handbook of Research on Mathematics Teaching and Learning*, pages 371–404. Information Age, Charlotte, NC. 00000.

Oleksii Holub. 2023. Discord chat exporter.

Yongbeom Kim, Christoper Ivan Wijaya Ong, and Fun Man Fung. 2023. Supplementary discourse - forming an online learning community with asynchronous discussions. *Journal of Chemical Education*, 100(2):496–506.

Jonathan K. Kummerfeld, Sai R. Gouravajhala, Joseph J. Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros C Polymenakos, and Walter Lasecki. 2019. A large-scale corpus for conversation disentanglement. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3846–3856, Florence, Italy. Association for Computational Linguistics.

Hui Liu, Zhan Shi, and Xiaodan Zhu. 2021. Unsupervised Conversation Disentanglement through Co-Training. *arXiv preprint*. ArXiv:2109.03199 [cs].

Jing Liu and Julie Cohen. 2021. Measuring teaching practices at scale: a novel application of text-as-data methods. *Educational Evaluation and Policy Analysis*, 43(4):587–614.

Louis Major and Steven Watson. 2018. Using video to support in-service teacher professional development: the state of the field, limitations and possibilities. *Technology, Pedagogy and Education*, 27(1):49–68.

Ali Malik, Juliette Woodrow, and Chris Piech. 2024. Learners Teaching Novices: An Uplifting Alternative Assessment. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, pages 785–791, Portland OR USA. ACM.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. 2022. LILA: A Unified Benchmark for Mathematical Reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5807–5832, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Birgit Pepin. 2014. Student transition to university mathematics education: transformations of people, tools and practices. In Sebastian Rezat, Mathias Hattermann, and Andrea Peter-Koop, editors, *Transformation - A Fundamental Idea of Mathematics Education*, pages 65–83. Springer New York, New York, NY.

Adnan Qayyum. 2018. Student help-seeking attitudes and behaviors in a digital era. *International Journal of Educational Technology in Higher Education*, 15(1):17.

Justin Reich, Rachel Slama, Rene Kizilcec, David Joyner, and Ken Koedinger. 2023. Research infrastructure: million tutor moves observatories project (mt mop): incubating digital infrastructure for an equitable, intelligent, and resilient stem education system. NSF Project Proposal NSF Award # 2321499, MIT.

Dou Shen, Qiang Yang, Jian-Tao Sun, and Zheng Chen. 2006. Thread detection in dynamic text message streams. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, page 35, Seattle, Washington, USA. ACM Press.

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. *arXiv preprint*. ArXiv:2204.09652 [cs].

Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. 2024. OpenMathInstruct-1: A 1.8 Million Math Instruction Tuning Dataset. *arXiv preprint*. ArXiv:2402.10176 [cs].

Leanne Townsend and Claire Wallace. 2017. Chapter 8: The Ethics of Using Social Media Data in Research: A New Framework. In Kandy Woodfield, editor, *Advances in Research Ethics and Integrity*, volume 2, pages 189–207. Emerald Publishing Limited.

Carla Van De Sande. 2011. A description and characterization of student activity in an open, online, mathematics help forum. *Educational Studies in Mathematics*, 77(1):53–78.

Maike Vollstedt, Aiso Heinze, Kristin Gojdka, and Stefanie Rach. 2014. Framework for examining the transformation of mathematics and mathematics learning in the transition from school to university: an analysis of german textbooks from upper secondary school and the first semester. In Sebastian Rezat, Mathias Hattermann, and Andrea Peter-Koop, editors, *Transformation - A Fundamental Idea of Mathematics Education*, pages 29–50. Springer New York, New York, NY.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. MetaMath: Bootstrap Your Own Mathematical Questions for Large Language Models. *arXiv preprint*. ArXiv:2309.12284 [cs].

Tao Yu and Shafiq Joty. 2020. Online Conversation Disentanglement with Pointer Networks. *arXiv preprint*. ArXiv:2010.11080 [cs].

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mammoth: building math generalist models through hybrid instruction tuning.

6 Limitations and Ethics Statement

Limitations

While *MathMentorDB* offers a rich source of data for analyzing mathematics tutoring interactions, several limitations should be noted. The dataset is derived solely from Discord, which may not represent all online mathematics tutoring environments. Users of this platform may have specific characteristics that differ from the broader population of mathematics learners. Additionally, participants in the Mathematics Discord Server are self-selected, potentially leading to a sample that is not fully representative of all mathematics students or tutors. As an English-language platform, the dataset may not capture the full diversity of mathematical discourse across different languages and cultures. Furthermore, while the dataset captures text-based interactions, it may miss important non-verbal cues or contextual information that could be present in face-to-face tutoring sessions.

Ethics Statement

Our approach to addressing ethical concerns in this study encompasses several key aspects. We have taken extensive measures to safeguard the privacy of Discord users by pseudonymizing all usernames and not collecting any metadata that could be used to identify individual users. The dataset undergoes thorough processing to remove personally identifiable information and mask any named entities, ensuring user anonymity. While explicit consent was not obtained from individual users, we adhere to Discord's terms of service and widely accepted social media research ethics policies, which permit the use of publicly available data for research purposes. The dataset will only be shared with researchers who commit to ethical principles and agree to use the data solely for research purposes.

Our study obtained approval from our university Institutional Review Board (IRB), which deter-

mined that this research does not constitute human subjects research due to the public nature of the data. To further ensure ethical integrity, we are committed to preventing any potentially harmful uses of the dataset. Researchers accessing the data will be required to agree to terms that prohibit attempts to de-anonymize users or use the data for non-research purposes. We are transparent about our data collection and processing methods, allowing for scrutiny and replication of our approach. We acknowledge that this dataset may not represent all learners equally, particularly those without access to online resources or those from non-English speaking backgrounds.

Additionally, we adhere to the social media ethics framework developed by (Townsend and Wallace, 2017). This framework provides guidelines, rather than rules, recognizing that principles need to remain flexible to respond to the varied contexts in which social media data is found. It is designed to guide and support researchers' decisions, rather than providing definitive answers or a 'one size fits all' approach. We propose that researchers make reference to this framework when submitting applications for ethical approval for new projects in this domain. For further support with social media research ethics, we suggest researchers consult the 'New Social Media, New Social Science?' (NSMNSS) online network, as a place to pose questions to a community of social media researchers.

By adhering to these ethics, we aim to balance the potential benefits of this research with the protection of individual privacy and the responsible use of data in academic research.