Clarifying Questions in Math Information Retrieval

Behrooz Mansouri
behrooz.mansouri@maine.edu
University of Southern Maine
Portland, Maine, USA

Zahra Jahedibashiz
zahra.jahedibashiz@maine.edu
University of Southern Maine
Portland, Maine, USA

ABSTRACT

One of the challenges of math information retrieval is the inherent ambiguity of mathematical notation. The use of various notations, symbols, and conventions can lead to ambiguities in math search queries, potentially causing confusion and errors. Therefore, asking clarifying questions in math search can help remove these ambiguities. Despite advances in incorporating clarifying questions for search, little is currently understood about the characteristics of these questions in math. This paper investigates math clarifying questions asked on the MathStackExchange community question answering platform, analyzing a total of 495,431 clarifying questions and their usefulness. The results of the analysis uncover specific patterns in useful clarifying questions that provide insight into the design considerations for future conversational math search systems. The formulae used in clarifying questions are closely related to those in the initial queries and are accompanied by common phrases, seeking for the missing information related to the formulae. Additionally, experiments utilizing clarifying questions for math search demonstrate the potential benefits of incorporating them alongside the original query.

CCS CONCEPTS

• **Information systems → Information retrieval query processing**.

KEYWORDS

Clarifying Questions, Math Information Retrieval, Conversational Information Seeking

ACM Reference Format:
Behrooz Mansouri and Zahra Jahedibashiz. 2023. Clarifying Questions in Math Information Retrieval. In *Proceedings of the 2023 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '23), July 23, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3578337.3605123

1 INTRODUCTION

In information retrieval systems, users typically articulate their information needs through queries that may exhibit ambiguity, incompleteness, or inaccuracies. This poses a challenge for search engines to accurately provide relevant results. However, the challenge is amplified in the domain of mathematical information retrieval, where the inherent ambiguity of mathematical expressions renders the task more intricate.

As a symbolic language, math is prone to multiple interpretations and can be difficult to parse and understand without context. For example, the equation "$x + y = z$" could represent any number of mathematical concepts depending on the context in which it is used. This inherent ambiguity can make it difficult for search algorithms to accurately retrieve relevant content, particularly when searching for specific formulas or concepts. Moreover, it is important that the users specify what domain they are looking for to get better search results. One can search for the Taylor series in the physics domain to understand asymptotic behavior, while another may search for the same phenomenon in electrical engineering to analyze circuits. There could also be constraints on the formulae that users need to specify before searching. For instance, in the ARQMath [15] formula search task (finding similar formulae for a query formula), two identical instances of the formulae $x^n + y^n + z^n$ were considered as dissimilar as one instance had the constraint of x, y, z to be any real numbers and the other instance considered these three variables to be only integers.

To deal with ambiguous queries, conversational search systems allow users to search for information through dialogue in order to better express their needs. The large-scale study by Zamani et al. [20] demonstrated significant improvements in the search system's ability to provide relevant results and clarify user information needs, indicating that asking clarifying questions is a crucial aspect of a successful search. However, it is also important for these systems to maintain a balance between asking too many clarifying questions and providing non-relevant results [1]. To gain a deeper understanding of clarifying questions, several studies have analyzed patterns in these types of questions on community question answering websites [2, 19], resulting in the development of new taxonomies for classifying these questions.

Despite the existence of some research on clarifying questions for search, to the best of our knowledge, there was no attempt on exploring these questions for math searches. In this paper, we will take the first step to analyze math clarifying questions. For our study, we rely on the community question answering website, MathStackExchange. We propose an approach to extract these clarifying questions and identify their usefulness with a new scoring scheme. With our approach, approximately 500K clarifying questions are extracted. We then study these clarifying questions to answer the following research questions:

- How often clarifying questions are asked on math community question answering websites? And how useful are they?
- What are the characteristics of useful clarifying questions?
- What are the common types of clarifying questions?
- How often clarifying questions are asked about formulae?

ICTIR '23, July 23, 2023, Taipei, Taiwan

ACM ISBN 979-8-4007-0073-6/23/07…$15.00
https://doi.org/10.1145/3578337.3605123

- Can clarifying questions help with search results?

Our study shows the ratio of responses to the math clarifying questions is almost 6 times higher than other topics in different community question answering websites. Clarifying questions regarding the formulae are also asked with common phrases such as "What is ...?" or "what does ... means?". Looking at the categories, clarifying questions related to incomplete information are asked more often than other categories. Also, clarifying questions regarding correcting the original question or confirming information tend to be more useful than suggestion clarifying questions. Our experiment on using clarifying questions for math searches shows that there is a potential in improving search results by considering clarifying questions and their corresponding response.

The remainder of this paper is organized as follows. We first review related work on math information retrieval and clarifying questions in Section 2. Then, Section 3 introduces our proposed approach for identifying and scoring math clarifying questions from MathStackExchange. In Section 4, we review the characteristics of math clarifying questions. Section 5 explores whether clarifying questions can help with math searches. Finally, in Section 6 we conclude the paper and provide a few remarks about the next steps.

2 RELATED WORK

In this study, we aim to bridge the gap between math information retrieval and clarifying questions for search. To achieve this, we review the existing work on both topics.

Math Information Retrieval. Math Information Retrieval (MIR) refers to information retrieval tasks where the user's information need pertains to mathematics. Previous research has shown that MIR poses unique challenges due to the specialized nature of mathematical queries [14]. For instance, math searches tend to have a higher proportion of question queries (20%) compared to general searches (2%). The importance of finding answers to math questions in MIR has led to the development of the recent ARQMath [10, 15, 22] test collection. The main goal of ARQMath is to identify relevant answers to math questions, which are sourced from MathStackExchange. One of the main challenges of this task is the need to handle mathematical formulae, which are often represented as trees using Symbol Layout or Operator trees [21], different from linear text and making it difficult to apply traditional IR techniques.

Several approaches have been proposed to address the challenges of MIR. These include traditional information retrieval as well as deep neural network models. Kane et al. [5] introduced the *msearch* engine, which combines search results from both text and math. For math, the system utilizes a modified version of Tangent-L [3] and for text, it employs the BM25 model [18]. To find answers to math questions, in addition to the candidate answer itself, the msearch engine also considers the question to which the answer belongs, as well as comments on both the question and answer, and question tags when performing text searches. Using deep neural networks, Mansouri et al. [11] used Sentence-BERT embeddings to compute similarities between pairs of questions and answers, and applied SVM-rank to learn a ranking model that combines these similarity scores with other features such as the number of comments on the question. Reusch et al. [17], in turn, employed a ColBERT model [6] to find relevant answers to math questions. The model was trained on 10 pairs of correct and 10 pairs of incorrect answers, with answers paired for each question.

Formula search is another important task in MIR, where the goal is to identify relevant formulae for a given formula query. The approaches for this task can be divided into two categories: isolated and contextual formula search models. Isolated formula search models compare candidate and query formulae in isolation (no context), and can use tree representations of formulae [8, 23] or embedding models [13, 16]. Contextual approaches, on the other hand, consider not only the formulae themselves, but also the context in which they appear. Some contextual formula search models perform searches for formulae and text separately and then combine the retrieval results [5, 24]. Other approaches, such as contextual formula search with MathAMR [12], represent both text and formulae in a unified tree-based representation and perform a search on this tree.

Context is crucial for formula search, as a formula query in isolation can be ambiguous. For example, a user may search for a formula with constraints on its variables, but an exact match of the formula may be irrelevant if the same constraints are not applied to the variables. This ambiguity can also arise in text searches, as users may have different intentions when searching for a term such as "triangle inequality," including seeking the definition, proof, equation, or application of the concept [14]. Given these considerations, we believe there is a need to use clarifying questions in MIR to address the issue of ambiguous math queries.

Clarifying Questions. Several studies have investigated the use of clarifying questions on community question answering websites. Braslavski et al. [2] conducted the first study on clarifying questions on Stack Exchange data and found that two categories, "more information" and "check," were dominant (60% of questions). In another work, Tavakoli et al. [19] studied clarification questions in asynchronous information-seeking conversations on three Stack Exchange websites, and introduced a new taxonomy of clarifying questions including ambiguity/incompleteness, confirmation, general, incorrectness, paraphrasing, and suggestion, with ambiguity/incompleteness being the most frequent type. To extract clarifying questions, the authors considered comments with question marks that were not asked by the user posting the question. Potential answers to clarifying questions were defined as comments posted after the clarifying question, mentioning the clarifying question asker. The study by Zamani et al. [20] found that clarifying questions provided both functional and emotional benefits to users, including helping them reach the right conclusion and feeling understood by the search system. The authors also introduced a taxonomy of clarification types, including disambiguation, preference, topic, and comparison.

There have been several attempts to generate clarifying questions, using either question selection or question generation methods. Aliannejadi et al. [1] focused on selecting clarifying questions from a set of human-generated questions for open-domain information seeking. In contrast, Zamani et al. [20] trained a neural sequence-to-sequence model to generate clarifying questions in response to open-domain search queries using weak supervision. To the best of our knowledge, there is no attempt on investigating clarifying questions for math search. In this work, we will take the first step, by characterizing the clarifying questions in math searches and suggesting design considerations for future work.

Table 1: Example of extracted clarifying questions using three approaches: CF, CBA, and CMA. The second column shows the response provided by the user who asked the question. The last row show mentions by '@' (with anonymized username).

Type	Clarifying Question	Response
CF	How is gin_{lex} defined?	Rough speaking, lex is a kind of order <, the initial ideal in ...
CBA	Have you tried Euler-Maclaurin formula?	No, I haven't. I supposed that it would not help because ...
CMA	@Asker Do you mean that $\lim(f(x) - g(x)) = 0$?	@CQAsker yes, exactly.

3 COLLECTION

MathStackExchange (MathSE) is a popular community question answering website for mathematics, hosting a wide range of questions spanning various levels of difficulty, from basic school homework questions to more complex and advanced topics. Researchers have previously utilized data from MathSE to address a variety of tasks in the field, such as formula search and answer retrieval [10, 15, 22] and formula markup revision [9]. For this study, we used the September 27, 2022 snapshot of MathSE from the Internet Archive.[1] This snapshot includes questions and their related answer posts (including the accepted answer), comments on each post, the creation time of each post and comment, and the information about the users. In total, there are approximately 1.5 million questions, of which approximately 1 million have at least one comment, with an average of 2.44 comments per question.

With MathSE as the source, we will use the following terminology in the rest of the paper:

(1) *asker*: the user who posts the question on MathSE
(2) *CQ asker*: the user who asks the clarifying question
(3) *question*: the original question post

The asker posts a question regarding a math information need. Then CQ askers can comment on that question and ask clarifying questions. We will next explain how we extract clarifying questions from these comments, along with their related responses.

Identifying Clarifying Questions and Related Answers. Clarifying questions are extracted only from the comments on questions. To identify clarifying questions, we first imposed the constraint that the comment must contain at least one sentence ending with a question mark, similar to the approach proposed by Tavakoli et al. [19]. To split the comments into individual sentences, we utilized the Natural Language Toolkit (NLTK) library.[2]

To further filter out non-clarifying questions, we introduced the following three categories of comments:

(1) **CF**: **C**omments that are the **f**irst to be posted on the question
(2) **CBA**: **C**omments written **b**efore a comment posted by **a**sker
(3) **CMA**: **C**omments that contain a **m**ention of the **a**sker

In this study, we focus specifically on clarifying questions that are directed towards the asker. Therefore, comments that mention other users are ignored. To extract candidate clarifying questions, we followed the priority order outlined; this means that if a clarifying question is identified using the first approach, it is not subsequently extracted using the other two approaches. To identify responses to the clarifying questions, we considered the comments written by the asker only. Potential responses to clarifying questions are comments that: (1) are posted right after the clarifying questions, or (2) have mention of the CQ asker.

Table 2: Clarifying question usefulness scoring scheme. Symbols: ✓ action taken, × not taken, ⋆ taken or not taken.

CQ replied by asker	×	✓	✓	✓	✓	✓
Post edited	⋆	✓	×	✓	×	✓
CQ asker posted an answer	⋆	×	✓	✓	✓	✓
CQ asker posted an accepted answer	⋆	×	×	×	✓	✓
Score	0	1	1	2	2	3

Table 1 illustrates examples of clarifying questions extracted by each approach, along with their responses. It is worth noting that it is possible for the answer to a clarifying question to be given by another user rather than the asker. In our study, we ignore these answers and only consider responses from the asker.

Clarifying Question Usefulness. We introduce a new measure for clarifying question usefulness that removes the effort for manual annotation. To measure the usefulness of a clarifying question, we first consider whether a conversation is formed between the CQ asker and the asker. This can be simply measured by receiving a response to the clarifying question by the asker. If a conversation is formed, we then check if the clarifying question has led to (1) post edit, (2) answer by the CQ asker, (3) accepted answer by the CQ asker (meaning the answer worked for their question). We propose a scoring scheme for usefulness, where clarifying questions with no conversation formed get a score of 0. If the conversation is formed, we then consider one score per each of the following activities happening after the conversation: 1) Post being edited, 2) CQ asker posts an answer, and 3) CQ asker posts an answer and that is the accepted answer.

To examine the relationship between a clarifying question and the CQ asker's subsequent answer, we analyzed the dates of both the answer, and the end of the conversation between the CQ asker and the asker (as indicated by the asker's response). We found that the average time between the end of the conversation and the CQ asker's answer was approximately 35 days. Based on this, we then examined the time of the post edit and the end of the conversation. If the edit occurred within 35 days of the end of the conversation, we considered it to be related to the usefulness of the clarifying question. Otherwise, we assigned a score of 0 for post editing, ensuring that we do not attribute post edits to the clarifying question unless there is a connection.

The final usefulness score is the sum of the scores obtained from the criteria described above. Table 2 illustrates the usefulness scores for different combinations of these criteria. For instance, if a

[1] https://archive.org/
[2] https://www.nltk.org/

Table 3: Examples of clarifying questions missed by our extraction pipeline. The 'Mention' category includes comments that mention the asker's username (anonymized), while the 'First' category includes the first comment in a thread.

Type	Comment
Mention	@Asker: it might help if you gave us the original function.
Mention	@Asker Please explain what you mean by the symbol $\approx$
First	Are you asking about $\int_0^1 dx\, x^3(x^4-1)^3$
First	I am not sure what g^* denotes here

clarifying question leads to a conversation between the asker and the CQ asker, and CQ asker's answer is the accepted answer, the clarifying question will have a score of 2 (sixth column). On the other hand, if the asker never responds to the clarifying question, it will have a score of 0 (first column).

Extraction Approach Analysis. To assess whether our approach is failing to extract a higher number of actual clarifying questions (i.e., low recall), we selected 500 comments [3] on questions that were not extracted as clarifying questions, with 250 being the first comment and 250 containing a mention of the asker (but not being the first comment). We then manually evaluated whether these comments qualified as actual clarifying questions according to a similar definition provided in [19]: being on the topic of the post, appearing clear, and not containing sarcastic/humorous questions. Annotation was done by two assessors, with given description of clarifying questions as, along with examples. With Cohen's kappa coefficient, a kappa of 0.86 was achieved, with disagreement being resolved after discussion.

Our analysis showed that among these 500 comments, 45 comments (9%) were assessed as actual clarifying questions, with 9 of them containing a mention of the asker. There were two main reasons for these extraction failures: (1) the comment corrected a wrong statement in the question, and (2) the comment requested missing information but did not end with a question mark. Table 3 provides examples of four comments (two from each category) that were actual clarifying questions but not extracted by our method. One possible solution to this issue might be considering comments containing key phrases such as "what do you mean by", "please explain", and "are you asking about" as clarifying questions.

To assess the precision of our extracted clarifying questions, we selected 500 randomly chosen samples from our collection, with 340 from the CF category, 155 from the CBA category, and 5 from the CMA category (the same ratio of extracted clarifying question by each category). We applied the same definition and assessment protocol as in the previous experiment to determine whether these samples were actual clarifying questions. The inter-annotator agreement was 0.81 (Cohen's kappa), with disagreement being resolved after discussion. Our analysis showed that only 3.6% of the analyzed extracted clarifying questions were not considered actual clarifying questions. Table 4 provides examples of wrongly extracted clarifying questions. The main patterns observed in these examples were sarcasm in the comments and unrelated questions, often regarding the MathSE website (e.g., "Why is this downvoted?").

Table 4: Examples of wrong extracted clarifying questions by our proposed approach. These are not genuine clarifying questions, but were mistakenly identified as such by our extraction pipeline.

People still buy books? Impressive!
why would you expect anything pretty to happen ?
What did you write on this question on the test?
this question has been on my mind too, what blogs do you keep up with?
Why is this downvoted?

Extracted Clarifying Questions. Table 5 summarizes the characteristics of the extracted clarifying questions and their responses. Using our proposed method, 495,431 clarifying questions were identified, taking account for 21% of all comments posted on math questions by users other than the asker. As can be seen, the majority of the clarifying questions are extracted using the first approach (CF). Of these clarifying questions, on average, 59.64% received a response from the asker. The average usefulness score was 0.44. 66% of the clarifying questions received a score of 0, while 26% received a score of 1, 5% received a score of 2, and a minority of 3% received a high score of 3.

We have built, to the best of our knowledge, the first test collection on math clarifying questions. This collection contains math clarifying questions with a usefulness score. This collection is publicly available,[4] and includes the original math questions (title, body, and tags), the clarifying questions, their associated usefulness scores, and responses to the clarifying question if available. Several tasks such as predicting the usefulness of clarifying questions, generating math clarifying questions, and deciding when to ask a clarifying question in a math search can use the proposed collection.

4 ANALYSIS OF CLARIFYING QUESTIONS

In this section, we explore the different characterizations of clarifying questions and their responses in math searches.

4.1 Responses to clarifying questions

The results presented in Table 5 indicate that a high proportion of clarifying questions in math receive a response from the asker, with approximately 60% of such questions being replied to. This is 6 times higher than the number reported by Tavakoli et al. [19], with a response ratio of only 10% on other Stack Exchange sites.

One possibility for this high response rate might be more active users on MathSE, with higher reputation scores. To explore the relationship between asker response and user reputation score, we compared the reputation scores of askers who responded to clarifying questions against those who did not. Our analysis revealed that the average reputation score for both groups was roughly 2.8K, indicating that asker reputation alone is not a strong predictor of

[3] similar sample size of related work

[4] https://github.com/AIIRLab/CQMath

Table 5: Number of extracted clarifying questions by each approach, the ratio of response by the asker, and the (average, standard deviation) usefulness score. The count and ratio of responses are presented as percentages.

Type	Total Count	Ratio (%)	Ratio of Response (%)	Usefulness Score (Avg., STD.)
CF	344,227	69.48	53.02	(0.41, 0.70)
CBA	145,388	29.35	77.00	(0.53, 0.75)
CMA	5,813	1.17	17.19	(0.10, 0.36)
All	495,431	100	59.64	(0.44, 0.71)

response behavior. We also calculated Kendall's Tau correlation between user reputation and response time, defined as the elapsed time between the posting of a clarifying question and its response. The resulting correlation was -0.03, suggesting that there is no clear relationship between response time and reputation.

Looking at clarifying questions with no response, one possibility is that the asker has already received an answer to their original question. To investigate this, we compared the timing of clarifying questions with the posting of accepted answers for the related questions. Our analysis showed that 9% of clarifying questions that received no response had an accepted answer, compared to 6% for clarifying questions that received a response. This suggests that receiving an answer may be a contributing factor to the lack of response to a clarifying question. Additionally, we looked at the timing of the first answer posted for a given question. Our analysis revealed that 15% of questions that received no response to their clarifying question had at least one answer posted before the clarifying question was asked, while this was the case for only 10% of questions that received a response to their clarifying question.

Overall, while we found that a higher proportion of clarifying questions on MathSE receive a response compared to other Stack Exchange sites, we were unable to identify a strong relationship between user reputation and asker response. Our analysis showed that the presence of an answer, either in the form of an accepted answer or an answer posted prior to the clarifying question, can be one reason why some clarifying questions go unanswered.

4.2 Useful clarifying questions

Our next analysis is on the usefulness of clarifying questions. As described in Section 3, a clarifying question is considered useful if it receives a response and leads to at least one of the following actions: a post edit, an answer by the CQ asker, or an accepted answer by the CQ asker. Our analysis revealed that approximately 34% of clarifying questions led to one of these actions. An example of a clarifying question with a usefulness score of 3 is shown in Table 6. In this case, the clarifying question led to edits to the question's title and body, and the answer provided by the CQ asker was selected as the accepted answer.

To further explore the contribution of clarifying questions to these actions, we analyzed the distribution of each action among clarifying questions that received a response from the asker, as shown in Figure 1. Our analysis showed that 49% of replied clarifying questions led to post edits, 16% led to an answer by the CQ asker, and 9% led to an accepted answer by the CQ asker. Notably, 44% of replied clarifying questions did not lead to any of these three useful actions, indicating that receiving a response does not always result in a useful outcome.

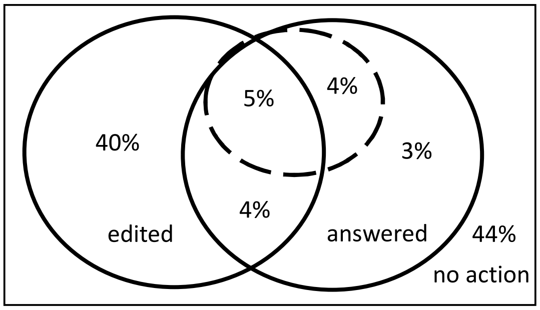

Figure 1: Distribution of each action among clarifying questions that received a response from the asker. 44% of clarifying questions with a response had no further action. The dotted circle shows accepted answers.

To investigate the usefulness of clarifying questions, we first examined the clarifying questions that resulted in accepted answers given by the CQ asker. It is possible that the asker selects the first given answer as the accepted answer, even if other answers are more comprehensive [4]. Overall, 96% of accepted answers to all math questions (with accepted answers) were the first answer given to the question. Similarly, the accepted answers given by CQ askers had a similar proportion of being the first answer. Among the answers given by CQ askers that received a response from the asker and were selected as the accepted answer, 98% were the first answer. However, among the answers provided by the CQ askers (that received a response), 58% of the answers were the first answer given to the question, but were not selected as the accepted answer. Of these questions, 72% did not have an accepted answer. Therefore, we can conclude that users value a quick response and tend to select the first answer as the accepted answer. In terms of response times, CQ askers with accepted answers provided their answers within an average of approximately 163 minutes of receiving a response from the asker, while those with no accepted answer provided their answers within a significantly higher average of approximately 172 minutes (Welch's t-test $p < 0.05$).

Next, we analyzed the types of edits made to posts on MathSE after a response was received to a clarifying question. We found that the most common type of edit was to the question body, accounting for 65% of all edits. Edits to the question title accounted for 15% of edits, while edits to the question tags accounted for 12%. Other types of edits, such as closing or deleting a post, accounted for the remaining 8% of edits. These results suggest that the majority of clarifying questions pertained to the body of the original question, and that after a clarifying question was asked, many of these posts were subsequently edited to address the additional information provided. Additionally, it is worth considering that the need for clarifying questions and subsequent edits may not always be due to ambiguity or incompleteness in the initial question. Other factors,

Table 6: Example of a clarifying question with usefulness score of 3 that has led to edits on question's title and body, and received an accepted answer from the CQ asker.

Action	Text	Date, Time
Original Title Posted	Area of revolution of a triangle.	Jan 16, 12:53
Original Body Posted	...So the triangle volume would be $\frac{1}{2}bh \cdot 2\pi r$...	Jan 16, 12:53
Clarifying Question Asked	Are you trying to find the volume of revolution, or the surface area?	Jan 16, 12:55
Response Posted	The volume of revolution.	Jan 16, 12:55
Title Edited	Volume of revolution of a triangle.	Jan 16, 12:56
Body Edited	...So the triangular doughnut volume would be $\frac{1}{2}bh \cdot 2\pi r = bh \cdot \pi r$...	Jan 16, 12:56
Question Answered	Your approach doesn't work: the outer parts of the triangle are being swept...	Jan 16, 13:16

Table 7: The top-5 most frequent 4-grams in clarifying questions among different usefulness categories.

	Usefulness			
Rank	3	2	1	0
1	what do you mean	what do you mean	what do you mean	what do you mean
2	what have you tried	what have you tried	what have you tried	what have you tried
3	I don't understand	do you know the	do you want to	do you know the
4	do you want to	do you want to	I don't understand	do you want to
5	do you know the	are you familiar with	do you know the	do you know about

such as the complexity of the topic being discussed or the level of expertise of the person asking the question, could also contribute to the need for clarification. For instance, for a question related to "abelian category", a clarifying question was posted to introduce related question tags to the asker. Therefore, while it may be reasonable to conclude that systems designed to improve the quality of information on online Q&A platforms should address the issue of ambiguous or incomplete initial questions, it would be important to consider a range of potential contributing factors such as incorrect information.

Next we explore the characteristics of useful clarifying questions. To do this, we examined the common n-grams (word sequences) present in each category of usefulness, to determine if there were any specific templates that useful clarifying questions tended to follow. We focused on the top frequent 4-grams (four-word sequences) in clarifying questions (similar to [7]). The results of this analysis are shown in Table 7, which lists the top-5 frequent 4-grams for each category of usefulness. We found that the common 4-grams among the different categories of usefulness were largely similar, with the top-2 being identical across all categories. However, there were some differences in the frequencies of these 4-grams between the categories. For example, the phrase "what do you mean" appeared nearly twice as often in clarifying questions with a score of 3 compared to those with a score of 0. Additionally, phrases such as "are you trying to" and "is your definition of" were more commonly used in clarifying questions, with a usefulness score of 3 compared to the other categories. Future conversational math search systems can utilize such information, following a similar rule-based approach suggested by Zamani et al. [20].

Krasakis et al. [7] studied another aspect of clarifying questions: the polarity of responses to these questions. In their study, the authors aimed to explore how clarifying questions and their answers could impact the ranking of documents on a search engine. They found that positive answers (those containing the word "yes") to clarifying questions, whether single-word or multi-word, could be useful for ranking documents. In contrast, single-word negative answers ("no") decreased the performance of ranking, while negative answers with multiple words significantly improved ranking. On MathSE, there was no response to clarifying questions that were a single word of "yes" or "no". However, about 3% of responses in each category of usefulness started with the word "no". In contrast, roughly 12% of responses with a usefulness score of 3 started with "yes", while this percentage dropped to around 7% for clarifying questions with a score of 0 that received a response. These responses often served to confirm or correct information, such as "Is a a constant?" with the response "yes, a is a constant", or "Do you mean $q^2p^3 < 0$? Square roots are always positive." with the response "Yes, sorry- will adjust." This indicates that confirmation and correction clarifying questions can be more useful, which we will explore in the next section.

Considering other aspects, our analysis of the length (number of terms) of clarifying questions and responses in different categories of usefulness found that the average number of terms used in both was around 21–24 words. We also analyzed the use of formulae in clarifying questions and found that clarifying questions with a usefulness score of 3 had an average of 1.5 formulae, while clarifying questions with a score of 0 had an average of 1.2 formulae. Finally, we found that useful clarifying questions were less likely to contain web links, with 2% of clarifying questions with a usefulness score of 3 containing links, compared to 5% of clarifying questions with a score of 0. Of the links that were used, 42% were to pages on MathSE, while 34% were to Wikipedia pages. These links were often used to refer askers to pages with solutions or additional information on concepts such as "Euler-Maclaurin" and "Pigeonhole principle".

4.3 Clarifying questions types

Our next analysis is on the types of clarifying questions. Based on the categories from previous works [19, 20] and our observation on the analyzed instances, we identified the following categories of clarifying questions:

1. **Incompleteness:** About an unclear concept in the question. (e.g., "What's Ω?")
2. **Confirmation:** To confirm the CQ asker understanding of the question. (e.g., "Just to confirm, your sequence is a subset of the natural numbers. Right?")
3. **Correction:** To correct wrong information in the original question. (e.g., "There are a couple of issues. The first equality is not correct. Can you see why?")
4. **Suggestion:** Suggesting a part of a solution or a reference to the asker. (e.g., "Have you tried using $\log(x) = \lim_{h \to 0} \frac{x^h - 1}{h}$?")
5. **Combination:** Combination of two or more of the categories above. (e.g., "Complex or real? If complex, you need $(x^* Ay)(y^* Ax)$ on the left." as Incompleteness and Suggestion.)
6. **Others:** If not fitting in any of the above categories. This is considered in the case that our previous categories are not comprehensive. (e.g., "It's theorem 12.36 in Rudin. Shall I reproduce the proof, or do you prefer to look it up yourself?")

In this study, we analyzed 500 randomly selected clarifying questions with an equal distribution of usefulness categories (125 per usefulness score). These clarifying questions were classified into the categories described above by reading both the clarifying questions and the original questions they pertained to. With the same assessment protocol as previous experiments, the inter-annotator agreement was 0.82 (Cohen's kappa). Table 8 shows the distribution of different categories of clarifying questions among different usefulness scores and among all clarifying questions.

We found that clarifying questions about incomplete information (Incompleteness) were the most common type of clarifying question on MathSE, which is consistent with the findings of Tavakoli et al. [19] on three other Stack Exchange websites: Quantitative Finance, English Language and Usage, and Science Fiction and Fantasy. However, only similar to Science Fiction and Fantasy, the second most common type of clarifying question on math was Suggestion. When we examined the distribution of types based on clarifying question usefulness, we found that the Incomplete category had a similar distribution across all usefulness scores. However, the Confirmation and Correction categories were more common in useful clarifying questions, while the Suggestion category was more frequent in clarifying questions with a usefulness score of 0.

Table 8: Distribution of different categories of clarifying questions (numbers in percentage).

	Usefulness Score				
Category	0	1	2	3	All
Incomplete	46.4	42.4	47.5	45.6	45.4
Confirmation	10.4	16.0	16.8	15.2	14.6
Correction	1.6	11.2	12.0	12.0	9.2
Suggestion	30.4	17.6	16.0	16.0	20.0
Combination	8.0	10.4	4.8	5.4	7.4
Others	3.2	2.4	3.2	4.8	3.4

Upon examining common n-grams, certain templates emerged for different categories of clarifying questions. For example, Confirmation and Correction clarifying questions often used templates such as "Do you mean", "Are you sure", and "Should that be". Suggestion clarifying questions commonly used templates like "Do you know" and "Have you tried". In the Incomplete category, a more diverse range of templates was observed, including "How did you", "What is the", "What do you", and "Do you want".

Another observation was that, on average, Correction and Confirmation clarifying questions contained 1.7 and 1.5 formulae, respectively, while Incomplete clarifying questions contained fewer formulae (an average of 1). Additionally, 72% of Correction clarifying questions and 67% of Confirmation clarifying questions contained at least one formula, suggesting that askers may introduce errors when writing formulae. For example, the clarifying question "Are you sure the RHS is not $w(|a|^2)^{1/2}$?" led to correction of the formula $|\omega(a)| \leq \omega(|a|^{\frac{1}{2}})^2$ to $|\omega(a)| \leq \omega(|a|^2)^{\frac{1}{2}}$. In another example, the clarifying question "Should it be $x^2 \tan x \mid_{x=\pi}$?" prompted the asker to respond "Yes, sorry for the formatting" and edit the original formula in the post, which was $x^(2)tanx|x = pi$. Clarifying questions may also be needed to provide missing information about certain symbols in formulae. For example, the clarifying question "Is $f : X \to Y$ continuous?" requests clarification on the symbol f. These examples demonstrate the importance of providing useful feedback on the formulae used in input queries and recommending corrections if necessary. Therefore, next, we analyze clarifying questions related to formulae.

4.4 Clarifying questions about formulae

Over 50% of clarifying questions contained at least one formula, with certain categories of clarifying questions using more formulae than others. In this section, we aim to explore the relationship between formulae in clarifying questions and the posts. To do so, we used the LaTeXML[5] tool to extract the MathML representations of the formulae from their LaTeX strings. Then, we employed the Tangent-CFT [13] pre-processing pipeline to generate the Symbol Layout Tree (SLT) and Operator Tree (OPT) representations of the formulae. The SLT representation is a tree in which nodes represent the elements of a formula, with labels indicating the type

[5] https://math.nist.gov/~BMiller/LaTeXML

(e.g., variables, and numbers) and value of each element. The edge labels in these trees show the spatial relationships between the elements of the formula, providing further detail on the structure of the formula. The OPT representation of a formula has the same node of as SLT, but the edge labels show the order in which the operands should be calculated.

Using these tools and representations, we first analyzed the extent to which the formulae in the clarifying questions were related to the variables used in the related post's formulae. For this, we used the SLT representation of formulae. In particular, if a formula in the clarifying question had only one variable node, and that node was also present in any of the SLT representations of the formulae in the post, we considered it to be a variable that the clarifying question was asking about. For example, in the clarifying question "What is α?", α is referring to the variable in the formula $\alpha(u, v) = \frac{\pi}{3}$ that has appeared in the post. Our analysis found that approximately 15% of clarifying questions contained a symbol used in the formulae in the body of the question, while 5% had symbols used in the question's title.

Additionally, common n-grams such as "what is your definition" and "how do you define" were frequently used in clarifying questions to disambiguate symbols. For example, in a question about calculating the series $\frac{x!}{1!} + \frac{(x+1)!}{2!} + \frac{(x+2)!}{3!} + \cdots + \frac{(x+n-1)!}{n!}$, clarifying question "Is x a real number? If yes, how do you define $x!$ for real numbers?" was asked. The average usefulness score for these clarifying questions was 0.48. In cases where the clarifying question was deemed useful, like the previous example, the asker received a response such as "no, x and n are natural numbers" and was provided with an answer to the question using the Pascal formula.

Our next analysis is on the similarity of formulae used in clarifying questions compared to the formulae in the original questions. To do this, we utilized the formulae embedding model from Tangent-CFT [13]. This model generates embeddings for formulae using different representations, including Symbol Layout Tree (SLT), Operator Tree (OPT), and SLT TYPE. SLT captures the appearance of the formulae, while OPT considers their semantics. For example, the formulae $a = b$ and $b = a$ are considered identical under the OPT representation, but not by SLT, as their semantics are unique, but their appearance is different. SLT TYPE represents formulae in a manner similar to SLT, but instead of individual elements (such as symbols), it considers the types of these elements. Using this representation, formulae such as $a + b$ and $x + y$ are considered identical. We applied these trained models on the ARQMath collections, to measure the similarity between formulae. Two formulae were considered identical if their cosine similarity was 1, almost identical if it was above 0.9 but not 1, and not similar if their cosine similarity was less than 0.5.

Table 9: Percentage of clarifying questions at least one formula similar to formulae in the question post. SLT uses the full representation of the formula based on its appearance, while SLT TYPE considers only element type of formula (unification). OPT captures semantics of formulae.

Representation	Identical	Almost Identical	Not Similar
		Title	
SLT	3.12	11.79	5.03
OPT	3.12	11.88	3.57
SLT TYPE	4.54	14.57	0.83
		Body	
SLT	16.45	26.20	5.11
OPT	16.81	26.33	3.48
SLT TYPE	22.22	24.99	0.56

Table 9 illustrates the percentage of clarifying questions that contain at least one formula that is (dis)similar to formulae in the question. We can see that the proportion of formulae that are identical or almost identical to those in the title or body of the question is higher than those that are not similar. This suggests that formulae clarifying questions are closely related to the formulae in the question post, particularly those in the body. When using the OPT representation, we observe that more formulae are detected as being similar. For example, in a question with the title "If $a^2 + b^2 = c^2$, then $m^2 + n^2 = c$, and vice versa. Why?", the clarifying question "If $c = m^2 + n^2$, surely you realize you can use this m, n and plug it into the formula $(m^2 - n^2)^2 + (2mn)^2 = (m^2 + n^2)^2$ to find the corresponding a, b?" includes the two formulae $m^2 + n^2 = c$ and $c = m^2 + n^2$, which are detected as identical using the OPT representation, but as almost similar using the SLT representation. When using the SLT TYPE representation, formulae tend to have higher similarity scores due to the generalization. For instance, the formulae $\lim_{x \to a} f(x)$ and $\lim_{x \to c} F(x)$ are considered identical using the SLT TYPE representation, but almost identical using the SLT.

Looking at the usefulness scores, the average score for clarifying questions with formulae was similar among the different categories of similarity (around 0.45) with the highest average being 0.49 for the clarifying questions with identical formulae to those used in the question body. Not similar formulae were mostly useful in the two cases: (1) being a part of bigger formulae in the question or (2) being formulae related to math concepts in the questions. For example, the clarifying question "Where do you want the half-circle? $\Re(z) > 0$? $\Im(z) > 0$? the opposite of these? Something else?" has a formula not used in the post, but is closely related to the question about the parametrization of a half circle.

Finally, we explore what clarifying questions are asked about the formulae. Table 10 shows the patterns commonly used with the formulae. As can be seen, the majority of clarifying questions are about incomplete data regarding the formulae or symbols used in them. This includes clarifying questions such as "What is ω?" to get information about the formula $\psi(\lambda) = \frac{e}{\pi\lambda} Im(e^{-\omega(\lambda-1)^{\frac{1}{4}}})$ used in the question. The next common category belongs to the Correction clarifying questions, where the CQ askers aim to provide the correct formulae that should have been used. For instance, after the clarifying question "Do you mean $\lim_{n \to \infty}$?", the original formula in the question, $\lim_{x \to \infty}$, was corrected. It can be useful to have a correction tool to check if the formulae used in the question are correct, and to have an autocorrection tool for math. Overall, it can be seen that clarifying questions with formulae are about disambiguating them and getting more information about symbols or the whole formula.

Table 10: Common patterns used in clarifying questions containing formula(e).

Rank	Pattern	Example	Rank	Pattern	Example
1	what is FORMULA?	What is ω?	6	FORMULA?	$\mathbb{Z}/3$?
2	do you mean FORMULA?	Do you mean $\lim_{n\to\infty}$?	7	how do you define FORMULA?	How do you define sin?
3	what are FORMULA and FORMULA?	What are n and p?	8	is FORMULA supposed to be...	Is $\frac{x(t)}{dt}$ supposed to be...
4	what does FORMULA mean?	What does C mean?	9	by FORMULA do you mean	By $\leq$ do you mean $\subset$?
5	what about FORMULA?	What about $\mathbb{Z}_4$?	10	is it FORMULA or FORMULA?	Is it u or f?

Table 11: NDCG′@5 values for re-ranking ARQMath topics. The values are average over the 39 topics that had at least one clarifying question with a response.

Data	Q	Q-RCQ	Q-CQ-RCQ
nDCG′@5	0.124	0.126	0.150

5 CLARIFYING QUESTIONS FOR SEARCH

This section presents our final analysis of whether clarifying questions and their responses can be useful for math searches. As shown in previous sections, knowing the answer to a certain clarifying question can lead to generating an answer to a math question that satisfies the user's information need. Without disambiguating, correcting certain errors, or confirming a setting, it is not always possible to post an answer. For instance, for the proof question on *Banach space* regarding the formula

$$\sup_{x\in X} |f(x)| \leq c \sup_{x\in E} |g(x)|$$

the asker did not use the condition $f|_E = g$ and a CQ asker had to confirm this before answering.

To study the effect of clarifying questions and their responses, we focus on the answer retrieval task of ARQMath: given a math question (Q), find the relevant answer (A) among different answer posts. Among the 300 topics provided in the ARQMath test collections (ARQMath-1 [10], ARQMath-2 [15], and ARQMath-3 [10]), 75 topics had a clarifying question and 39 of them received a response. For these 39 topics, we use the assessed answers from ARQMath and re-rank them. We consider nDCG′@5 as the evaluation measure. For ranking, we use the TF-IDF PyTerrier baseline system provided in ARQMath. In this system, symbols in LaTeX strings are mapped to English words to avoid tokenization problems. We consider three inputs as the query:

(1) **Q**: Only the math question on MathSE is the input.
(2) **Q-RCQ**: The math question concatenated with the response to the clarifying question is considered as the query.
(3) **Q-CQ-RCQ**: The math question, the clarifying question, and its response are all concatenated and considered as the query.

With these settings, Table 11 shows the nDCG′@5 for different input queries averaged across 39 topics in ARQMath that have clarifying questions with responses. Adding both the clarifying question and its response to the original question improved nDCG′@5. However, this improvement was not significant (p < 0.05, two-tailed paired t-test with Bonferroni correction). In cases where the clarifying questions were useful for search, they were mostly those that suggested an approach or a part of the solution to the question. For instance, for the question on uniformly continuous functions, the CQ asker asked a question about the range of the function and then hinted at a solution by checking if the function is differentiable and has a bounded derivative. The ndcg′@5 increases from 0.65 to 0.89 by including the clarifying questions and their related responses. Answers considering bounded interval were retrieved in the top-5 results when the clarifying question was included along with the original questions.

Cases in which adding the clarifying questions and responses to the question did not help the search results were mostly those in which the clarifying question was not useful or provided a hint to a wrong solution. For example, for the question with the title "Divisibility of n^3+6n^2-7n", the clarifying question "Have you tried factoring the polynomial?", was replied with "Factoring resulted in (n-1)n(n+7) from where i don't see how this would help?". Overall, our experiment showed that certain clarifying questions can be helpful for improving search results, depending on whether the clarifying question provides useful information relevant to the original question (usefulness score). We studied the use of clarifying questions as a form of query expansion, but another approach could be to reformulate the query, left for future work.

6 CONCLUSION

In this study, we explored the role of clarifying questions in math searches by analyzing a large dataset of such questions collected from MathStackExchange. We proposed a new approach to detect clarifying questions on MathStackExchange along with their usefulness scores, analyzing approximately 500K clarifying questions. Our analysis revealed that approximately 34% of these clarifying questions were found to be useful, often leading to a post edit or an answer from the user who asked the clarifying question. The majority of clarifying questions in math are related to incomplete information, and that correction and confirmation clarifying questions are more useful than suggestion-type questions. In addition, we observed that formulae in clarifying questions are often closely related to those in the original question. Clarifying questions containing formula(e) were aimed to get more information about its symbols or the whole formula. To further assess the impact of clarifying questions on math searches, we conducted an experiment on the ARQMath dataset. While the results showed that the inclusion of these elements did not significantly improve search results, we did find that clarifying questions that providing a hint towards a solution were more beneficial. For future work, we plan to explore generating clarification questions for math, determining the need for asking clarifying questions and classifying the usefulness of such questions.

REFERENCES

[1] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2019. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[2] Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. 2017. What Do You Mean Exactly? Analyzing Clarification Questions in CQA. In *Proceedings of the 2017 Conference on Human Information Interaction and Retrieval*.

[3] Dallas Fraser, Andrew Kane, and Frank Wm Tompa. 2018. Choosing Math Features for BM25 Ranking with Tangent-L. In *Proceedings of the ACM Symposium on Document Engineering 2018*.

[4] Agnieszka Geras, Grzegorz Siudem, and Marek Gagolewski. 2022. Time to vote: Temporal clustering of user activity on Stack Overflow. *Journal of the Association for Information Science and Technology* (2022).

[5] Andrew Kane, Yin Ki Ng, and Frank Tompa. 2022. Dowsing for Answers to Math Questions. Doing Better with Less. *Proceedings of the Working Notes of CLEF 2022*.

[6] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[7] Antonios Minas Krasakis, Mohammad Aliannejadi, Nikos Voskarides, and Evangelos Kanoulas. 2020. Analysing the Effect of Clarifying Questions on Document Ranking in Conversational Search. In *Proceedings of the 2020 acm sigir on international conference on theory of information retrieval*.

[8] Giovanni Yoko Kristianto, Goran Topic, and Akiko Aizawa. 2016. MCAT Math Retrieval System for NTCIR-12 MathIR Task. In *NTCIR*.

[9] Suyu Ma, Chunyang Chen, Hourieh Khalajzadeh, and John Grundy. 2021. Latexify Math: Mathematical Formula Markup Revision to Assist Collaborative Editing in Math Q&A Sites. *Proceedings of the ACM on Human-Computer Interaction*.

[10] Behrooz Mansouri, Vít Novotnỳ, Anurag Agarwal, Douglas W Oard, and Richard Zanibbi. 2022. Overview of ARQMath-3 (2022): Third CLEF Lab on Answer Retrieval for Questions on Math. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer.

[11] Behrooz Mansouri, Douglas W Oard, and Richard Zanibbi. 2021. DPRL Systems in the CLEF 2022 ARQMath Lab: Introducing MathAMR for Math-Aware Search. *Proceedings of the Working Notes of CLEF 2022*.

[12] Behrooz Mansouri, Douglas W Oard, and Richard Zanibbi. 2022. Contextualized Formula Search Using Math Abstract Meaning Representation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*.

[13] Behrooz Mansouri, Shaurya Rohatgi, Douglas W Oard, Jian Wu, C Lee Giles, and Richard Zanibbi. 2019. Tangent-CFT: An Embedding Model for Mathematical Formulas. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*.

[14] Behrooz Mansouri, Richard Zanibbi, and Douglas W Oard. 2019. Characterizing Searches for Mathematical Concepts. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE.

[15] Behrooz Mansouri, Richard Zanibbi, Douglas W Oard, and Anurag Agarwal. 2021. Overview of ARQMath-2 (2021): Second CLEF Lab on Answer Retrieval for Questions on Math. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer.

[16] Shuai Peng, Ke Yuan, Liangcai Gao, and Zhi Tang. 2021. MathBERT:: A Pre-trained Model for Mathematical Formula Understanding. *arXiv preprint arXiv:2105.00377* (2021).

[17] Anja Reusch, Maik Thiele, and Wolfgang Lehner. 2021. An ALBERT-based Similarity Measure for Mathematical Answer Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[18] Stephen Robertson, Hugo Zaragoza, et al. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*.

[19] Leila Tavakoli, Hamed Zamani, Falk Scholer, William Bruce Croft, and Mark Sanderson. 2022. Analyzing Clarification in Asynchronous Information-Seeking Conversations. *Journal of the Association for Information Science and Technology*.

[20] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating Clarifying Questions for Information Retrieval. In *Proceedings of the Web Conference 2020*.

[21] Richard Zanibbi and Dorothea Blostein. 2012. Recognition and Retrieval of Mathematical Expressions. *International Journal on Document Analysis and Recognition (IJDAR)*.

[22] Richard Zanibbi, Douglas W Oard, Anurag Agarwal, and Behrooz Mansouri. 2020. Overview of ARQMath 2020: CLEF Lab on Answer Retrieval for Questions on Math. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer.

[23] Wei Zhong, Shaurya Rohatgi, Jian Wu, C Lee Giles, and Richard Zanibbi. 2020. Accelerating Substructure Similarity Search for Formula Retrieval. In *European Conference on Information Retrieval*. Springer.

[24] Wei Zhong, Xinyu Zhang, Ji Xin, Richard Zanibbi, and Jimmy Lin. 2021. Approach Zero and Anserini at the CLEF-2021 ARQMath Track: Applying Substructure Search and BM25 on Operator Tree Path Tokens. In *CLEF (Working Notes)*.