Overview of ARQMath-3 (2022): Third CLEF Lab on Answer Retrieval for Questions on Math (Working Notes Version)

Behrooz Mansouri[1], Vít Novotný[3], Anurag Agarwal[1], Douglas W. Oard[2] and Richard Zanibbi[1]

[1]*Rochester Institute of Technology, NY, USA*
[2]*University of Maryland, College Park, USA*
[3]*Faculty of Informatics, Masaryk University, Czech Republic*

Abstract

This paper provides an overview of the third and final year of the Answer Retrieval for Questions on Math (ARQMath-3) lab, run as part of CLEF 2022. ARQMath has aimed to introduce test collections for math-aware information retrieval. ARQMath-3 has two main tasks, Answer Retrieval (Task 1) and Formula Search (Task 2), along with a new pilot task Open Domain Question Answering (Task 3). Nine teams participated in ARQMath-3, submitting 33 runs for Task 1, 19 runs for Task 2, and 13 runs for Task 3. Tasks, topics, evaluation protocols, and results for each task are presented in this lab overview.

Keywords

Community Question Answering, Open Domain Question Answering, Mathematical Information Retrieval, Math-aware Search, Math Formula Search

1. Introduction

Math information retrieval (Math IR) aims at facilitating the access, retrieval and discovery of math resources, and is needed in many scenarios [1]. For example, many traditional courses and Massive Open Online Courses (MOOCs) release their resources (books, lecture notes and exercises, etc.) as digital files in HTML or XML. However, due to the specific characteristics of math formulae, classic search engines do not work well for indexing and retrieving math.

Math-aware search systems can be beneficial for learning activities. Students can search for references to help solve problems, increase knowledge, reduce doubts, and clarify concepts. Instructors might also benefit from these systems by creating learning communities within a classroom. For example, a teacher can pool different digital resources to create the subject matter and then let students search through them for mathematical notation and terminology. Math-aware search engines can also help researchers identify potentially useful systems, fields, and collaborators. Good examples of this interdisciplinary approach benefiting physics include the AdS/CFT correspondence and holographic duality theories.

CLEF'22: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
bm3302@rit.edu (B. Mansouri); witiko@mail.muni.cz (V. Novotný); axasma@rit.edu (A. Agarwal); oard@umd.edu (D. W. Oard); rxzvcs@rit.edu (R. Zanibbi)

A key focus of mathematical searching is formulae. In contrast to simple words or other objects, a formula can have a well defined set of properties, relations, applications, and often also a 'result'. There are many (mathematically) equivalent formulae which are structurally quite different. For example, it is of fundamental importance to ask what information a user wants when searching for $x^2 + y^2 = 1$: is it the value of the variables x and y that satisfy this equation, all indexed objects that contain this formula, all indexed objects containing $a^2 + b^2 = 1$, or the geometric figure that is represented by this equation?

This third Answer Retrieval for Questions on Math (ARQMath-3) lab at the Conference and Labs of the Evaluation Forum (CLEF) completes our development of test collections for Math IR from content found on Math Stack Exchange,[1] a Community Question Answering (CQA) forum. This year, ARQMath continues its two main tasks: Answer Retrieval for Math Questions (Task 1) and Formula Search (Task 2). We also introduce a new pilot task, Open Domain Question Answering (Task 3).

Using the question posts from Math Stack Exchange, participating systems are given a question (in Tasks 1 and 3) or a formula from a question (in Task 2), and asked to return a ranked list of either potential answers to the question (Task 1) or potentially useful formulae (Task 2). For Task 3, given the same questions as Task 1, the participating systems also provide an answer, but are not limited to searching the ARQMath collection to find that answer. Relevance is determined by the expected utility of each returned item. These tasks allow participating teams to explore leveraging math notation together with text to improve the quality of retrieval results.

2. Related Work

Prior to ARQMath, three test collections were developed over a period of five years at the NII Testbeds and Community for Information Access Research (NTCIR) shared task evaluations. To the best of our knowledge, NTCIR-10 [2] was the first shared task on Math IR, considering three scenarios for searching:

- Formula Search: find similar formulae for the given formula query.
- Formula+Text Search: search the documents in the collection with a combination of keywords and formula queries.
- Open Information Retrieval: search the collection using text queries.

NTCIR-11 [3] considered the formula+text search task as the main task and introduced an additional Wikipedia open subtask, using the same set of topics with a different collection and different evaluation methods. Finally, in NTCIR-12 [4], the main task was formula+text search on two different collections. A second task was Wikipedia Formula Browsing (WFB), focusing on formula search. Formula similarity search (the *simto* task) was a third task, where the goal was to find formulae 'similar' (not identical) to the formula query.

An earlier effort to develop a test collection started with the Mathematical REtrieval Collection (MREC) [5], a set of 439,423 scientific documents that contained more than 158 million formulae. This was initially only a collection, with no shared relevance judgments (although

[1] https://math.stackexchange.com/

the effectiveness of individual systems was measured by manually assessing a set of topics). The Cambridge University MathIR Test Collection (CUMTC) [6] subsequently built on MREC, adding 160 test queries derived from 120 MathOverflow discussion threads (although not all queries contained math). CUMTC relevance judgments were constructed using citations to MREC documents cited in MathOverflow answers.

To the best of our knowledge, ARQMath's Task 1 is the first Math IR test collection to focus directly on answer retrieval. ARQMath's Task 2 (formula search) extends earlier work on formula search, with several improvements:

- **Scale.** ARQMath has an order of magnitude more assessed topics than prior formula search test collections. There are 22 topics in NTCIR-10, and 20 in NTCIR-12 WFB (+20 variants with wildcards).
- **Contextual Relevance.** In the NTCIR-12 WFB task [4], there was less attention to context. ARQMath Task 2, by contrast, has evolved as a contextualized formula search task, where relevance is defined both by the query and retrieved formulae and also the contexts in which those formulae appear.
- **Deduplication.** NTCIR collections measured effectiveness using formula instances. In ARQMath we clustered visually identical formulae to avoid rewarding retrieval of multiple instances of the same formula.
- **Balance.** ARQMath balances formula query complexity, whereas prior collections were less balanced (reannotation shows low complexity topics dominate NTCIR-10 and high complexity topics dominate NTCIR-12 WFB [7]).

In ARQMath-3, we introduced a new pilot task, Open Domain Question Answering. The most similar prior work is the SemEval 2019 [8] math question answering task, which used question sets from Math SAT practice exams in three categories: Closed Algebra, Open Algebra and Geometry. A majority of the Math SAT questions were multiple choice, with some having numeric answers.

While we have focused on search and question answering tasks in ARQMath, there are other math information processing tasks that can be considered for future work. For example, extracting definitions for identifiers, math word problem solving, and informal theorem proving are active areas of research: for a survey of recent work in these areas, see Meadows and Ferentes [9]. Summarization of mathematical texts, text/formula co-referencing, and the multimodal representation and linking of information in documents are some other examples.

3. The ARQMath Stack Exchange Collection

For ARQMath-3, we reused the collection[2] from ARQMath-1 and -2.[3] The collection was constructed using the March 1st, 2020 Math Stack Exchange snapshot from the Internet Archive.[4]

[2]By *collection* we mean the content to be searched. That content together with topics and relevance judgments is a *test collection*. There is only one ARQMath *collection*

[3]ARQMath-1 was built for CLEF 2020, ARQMath-2 was built for CLEF 2021. We refer to submitted runs or evaluation results by year, as AQRMath-2020 or ARQMath-2021. This distinction is important because ARQMath-2022 participants also submitted runs for both the ARQMath-1 and -2 test collections.

[4]https://archive.org/download/stackexchange

Questions and answers from 2010-2018 are included in the collection. The ARQMath test collection contains roughly 1 million questions and 28 million formulae. Formulae in the collection are annotated using XML elements with the class attribute `math-container`, and a unique integer identifier given in the `id` attribute. Formulae are also provided separately in three index files for different formula representations (LaTeX, Presentation MathML, and Content MathML), which we describe in more detail below.

During ARQMath-2021, participants identified three issues with the ARQMath collection that had not been noticed and corrected earlier. In 2022, we have made the following improvements to the collection:

1. **Formula Representations.** We found and corrected 65,681 formulae with incorrect Symbol Layout Tree (SLT) and Operator Tree (OPT) representations. This resulted from incorrect handling of errors generated by the LaTeXML tool that had been used for generating those representations.
2. **Clustering Visually Distinct Formulae.** Correcting SLT representations resulted in a need to adjust the clustering of formula instances. Each cluster of visually identical formulae was assigned a unique 'Visual ID'. Clustering had been performed using SLT where possible, and LaTeX otherwise. To correct the clustering, we split any cluster that now included formulae with different representations. In such cases, the partition with the largest number of instances retained its Visual ID; remaining formulae were assigned to another existing Visual ID (with the same SLT or LaTeX) or, if necessary, to a new Visual ID. To break ties, the partition with the largest cumulative ARQMath-2 relevance score retained its Visual ID or, failing that, choosing the partition with the lowest Formula ID. 29,750 new Visual IDs resulted.
3. **XML Errors.** In the XML files for posts and comments, the LaTeX for each formula is encoded as a XML element with the class attribute `math-container`. We found and corrected 108,242 formulae that had not been encoded in that way.
4. **Spurious Formula Identifiers.** The ARQMath collection includes an index file that includes Formula ID, Visual ID, Post ID, SLT, OPT, and LaTeX for each formula instance. However, there were also formulae in the index file that did not actually occur in any post or comment in the collection. This happened because formula extraction was initially done on the Post History file, which also contained some content that had later been removed. We added a new annotation to the formula index file to mark such cases.

The Math Stack Exchange collection was distributed to participants as XML files on Google Drive.[5] To facilitate local processing, the organizers provided python code on GitHub[6] for reading and iterating over the XML data, and for generating the HTML question threads. All of the code to generate the corrected ARQMath collection is available from that same GitHub repository.

[5] https://drive.google.com/drive/folders/1ZPKIWDnhMGRaPNVLi1reQxZWTfH2R4u3
[6] https://github.com/ARQMath/ARQMathCode

4. Task 1: Answer Retrieval

The goal of Task 1 is to find and rank relevant answers to math questions. Topics are constructed from questions posted to Math Stack Exchange in 2021, and the collection to search is only the answers to earlier questions (from 2010-2018) in the ARQMath collection. System results ('runs') are evaluated using measures that characterize the extent to which answers judged by relevance assessors as having higher relevance come before answers with lower relevance in the system results (e.g., using nDCG′). In this section, we describe the Task 1 search topics, participant runs, baselines, pooling, relevance assessment, and evaluation measures, and we briefly summarize the results.

4.1. Topics

ARQMath-3 Task 1 topics were selected from questions posted to Math Stack Exchange in 2021. There were two strict criteria for selecting candidate topics: (1) any candidate must have at least one formula in the title or the body of the question, (2) any candidate must have at least one known duplicate question (from 2010 to 2018) in the ARQMath collection. Duplicates have been annotated by Math Stack Exchange moderators as part of their ongoing work, and we chose to limit our candidates to topics for which a known duplicate question existed. We did this to avoid assessing topics with no relevant answers in the assessment pools or even the collection itself. In ARQMath-2 we had included 11 topics for which there were no known duplicates on an experimental basis. Of those 11, 9 had turned out to have no relevant answers found by any participating system or baseline.

We selected 139 candidate topics from among the 3313 questions that satisfied both of our strict criteria by applying additional soft criteria based on the number of terms and formulae in the title and body of the question, the question score that Math Stack Exchange users had assigned to the question, and the number of answers, comments, and views for the question. From those 139, we manually selected 100 topics in a way that balanced three desiderata: (1) A similar topic should not already be present in the ARQMath-1 or ARQMath-2 test collections, (2) we expected that our assessors would have (or be able to easily acquire) the expertise to judge relevance to the topic, and (3) the set of topics maximized diversity across four dimensions (question type, difficulty, dependence, and complexity).

In prior years, we had manually categorized topic type as *computation*, *concept* or *proof* and we did so again for ARQMath-3. A disproportionately large fraction of Math Stack Exchange questions ask for proofs, so we sought to stratify the ARQMath-3 topics in a way that was somewhat better balanced. Of the 100 ARQMath-3 topics, 49 are categorized as *proof*, 28 as *computation*, and 23 as *concept*. Question difficulty also benefited from restratification. Our insistence that topics have at least one duplicate question in the collection injects a bias in favor of easier questions, and such a bias is indeed evident in the ARQMath-1 and ARQMath-2 test collections. We made an effort to better balance (manually estimated) topic difficulty for the ARQMath-3 test collection, ultimately resulting in 24 topics categorized as hard, 55 as medium, and 21 as easy. We also paid attention to the (manually estimated) dependency of topics on text, formulae, or both, but we did not restratify on that factor. Of the 100 ARQMath-3 topics, 12 are categorized as dependent to text, 28 on formulae, and 60 on both. New this year, we

Task 1: Question Answering

```
<Topics>
  ...
  <Topic number="A.384">
    <Title>What does this bracket notation mean?</Title>
    <Question>
      I am currently taking MIT6.006 and I came across this problem on the
      problem set. Despite the fact I have learned Discrete Mathematics
      before, I have never seen such notation before, and I would like to
      know what it means and how it works, Thank you:
      <span class=``math-container'' id=``q_898''>
        $$f_3(n) = \binom n2$$
      </span>
    </Question>
    <Tags>discrete-mathematics, algorithms</Tags>
  </Topic>
  ...
</Topics>
```

Task 2: Formula Retrieval

```
<Topics>
  ...
  <Topic number="B.384">
    <Formula_Id>q_898</Formula_Id>
    <Latex>f_3(n) = \binom n2</Latex>
    <Title>What does this bracket notation mean?</Title>
    <Question>
      I am currently taking MIT6.006 and I came across this problem on the
      problem set. Despite the fact I have learned Discrete Mathematics
      before, I have never seen such notation before, and I would like to
      know what it means and how it works, Thank you:
      <span class=``math-container'' id=``q_898''>
        $$f_3(n) = \binom n2$$
      </span>
    </Question>
    <Tags>discrete-mathematics, algorithms</Tags>
  </Topic>
  ...
</Topics>
```

Figure 1: Example XML Topic Files. Formula queries in Task 2 are taken from questions for Task 1. Here, ARQMath-3 formula topic B.384 is a copy of ARQMath-3 question topic A.384 with two additional fields for the query formula (1) identifier and (2) LaTeX.

also paid attention to whether a topic actually asks several questions rather than just one. For these multi-part topics, our relevance criteria require that a highly relevant answer provide relevant information for all parts of the question. Among ARQMath-3 topics, 14 are categorized as multi-part questions.

The topics were published in the XML file format illustrated in Figure 1. Each topic has a unique Topic ID, a Title, a Question (which is the body of the question post), and Tags provided by the asker of the question on the Math Stack Exchange. Notably, links to duplicate or related questions are not included. To facilitate system development, we provided python code that participants could use to load the topics. As in the collection, the formulae in the topic file are placed in `<span>` XML elements, with each formula instance represented by a unique identifier and its LaTeX representation. Similar to the collection, there are three Tab Separated Value (TSV) files, for the LaTeX, OPT and SLT representations of the formulae, in the same format as the

Table 1
ARQMath-3: Submitted Runs. Baselines for Task 1 (5), Task 2 (1) and Task 3 (1) were generated by the organizers. Primary and alternate runs were pooled to different depths, as described in Section 4.4.

	Automatic		**Manual**	
	Primary	Alternate	Primary	Alternate
Task 1: Answer Retrieval				
Baselines	2	3		
Approach0			1	4
DPRL	1	4		
MathDowsers	1	2		
MIRMU	1	4		
MSM	1	4		
SCM	1	4		
TU_DBS	1	4		
Totals (38 runs)	*8*	*25*	*1*	*4*
Task 2: Formula Retrieval				
Baseline	1			
Approach0			1	4
DPRL	1	4		
MathDowsers	1	2		
JU_NITS	1	2		
XY_PHOC_DPRL	1	2		
Totals (20 runs)	*5*	*10*	*1*	*4*
Task 3: Open Domain QA				
Baseline	1			
Approach0			1	4
DPRL	1	3		
TU_DBS	1	3		
Totals (14 runs)	*3*	*6*	*1*	*4*

collection's TSV files. The Topic IDs in ARQMath-3 start from 301 and continue to 400. In ARQMath-1, Topic IDs were numbered from 1 to 200, and in ARQMath-2, from 201 to 300.

4.2. Participant Runs

ARQMath Participants submitted their runs on Google Drive. As in previous years, we expect all runs to be publicly available.[7] A total of 33 runs were received from 7 teams. Of these, 28 runs were declared to be automatic, with no human intervention at any stage of generating the ranked list for each query. The remaining 5 runs were declared to be manual, meaning that there was some type of human involvement in at least one stage of retrieving answers. Manual runs were invited in ARQMath to increase the quality and diversity of the pool of documents that are judged for relevance, but it is important to note that they might not be fairly compared to automatic runs. The teams and submissions are shown in Table 1. For the details of each run, please see the participant papers in the working notes.

[7] https://drive.google.com/drive/u/1/folders/1l1c2O06gfCk2jWOixgBXI9hAlATybxKv

4.3. Baseline Runs

For Task 1, five baseline systems were provided by the organizers.[8] This year, the organizers included a new baseline system using PyTerrier [10] for the TF-IDF model. The other baselines were also run for ARQMath 2020 and 2021. Here is a description of our baseline runs.

1. **TF-IDF.** We provided two TF-IDF baselines . The first uses Terrier [11] with default parameters and raw LaTeX strings, as in prior years of the lab. One problem with this baseline is that Terrier removes some LaTeX symbols during tokenization. The second uses PyTerrier [10], with symbols in LaTeX strings first mapped to English words to avoid tokenization problems.
2. **Tangent-S**. This baseline is an isolated formula search engine that uses both SLT and OPT representations [12]. The target formula was selected from the question title if at least one existed, otherwise from the question body. If there were multiple formulae in the field, a formula with the largest number of symbols (nodes) in its SLT representation was chosen; if more than one had the largest number of symbols, we chose randomly between them.
3. **TF-IDF + Tangent-S.** Averaging normalized similarity scores from the TF-IDF (only from PyTerrier) and Tangent-S baselines. The relevance scores from both systems were normalized in [0,1] using min-max normalization, and then combined using an unweighted average.
4. **Linked Math Stack Exchange Posts.** Using duplicate post links from 2021 in Math Stack Exchange, this oracle system returns a list of answers from posts in the ARQMath collection that had been given to questions marked in Math Stack Exchange as duplicates to ARQMath-3 topics. These answers are ranked by descending order of their vote scores. Note that the links to duplicate questions were not available to the participants.

4.4. Relevance Assessment

Relevance judgments for Tasks 1 and 3 were performed together, with the results for the two tasks intermixed in the judgment pools.

Pooling. For each topic, participants were asked to rank up to 1,000 answer posts. We created pools for relevance judgments by taking the top-k retrieved answer posts from every participating system or baseline in Tasks 1 or 3. For Task 1 primary runs, the top 45 answer posts were included; for alternate runs the top 20 were included. These pooling depths were chosen based on assessment capacity, with the goal of identifying as many relevant answer posts as possible. Two Task 1 baseline runs, PyTerrier TF-IDF+Tangent-S. and Linked Math Stack Exchange Posts, were pooled as primary runs (i.e, to depth 45); other baselines were pooled as alternate runs (i.e., to depth 20). All Task 3 run results (each of which is a single answer; see section 5.6) were also included in the pools. After merging these top-ranked results, duplicate posts were deleted and the resulting pools were sorted randomly for display to assessors. On average, the judgment pools for Tasks 1 and 3 contain 464 answer posts per topic.

[8]Source code and instructions for running the baselines are available from GitLab (Tangent-S: https://gitlab.com/dprl/tangent-s, PyTerrier: https://gitlab.com/dprl/pt-arqmath/) and GoogleDrive (Terrier: https://drive.google.com/drive/u/0/folders/1YQsFSNoPAFHefweaN01Sy2ryJjb7XnKF)

Table 2
Relevance Assessment Criteria for Tasks 1 and 2.

Score	Rating	Definition
Task 1: Answer Retrieval		
3	High	Sufficient to answer the complete question on its own
2	Medium	Provides some path towards the solution. This path might come from clarifying the question, or identifying steps towards a solution
1	Low	Provides information that could be useful for finding or interpreting an answer, or interpreting the question
0	Not Relevant	Provides no information pertinent to the question or its answers. A post that restates the question without providing any new information is considered non-relevant
Task 2: Formula Retrieval		
3	High	Just as good as finding an exact match to the query formula would be
2	Medium	Useful but not as good as the original formula would be
1	Low	There is some chance of finding something useful
0	Not Relevant	Not expected to be useful

Relevance definition. The relevance definitions were the same those defined for ARQMath-1 and -2. The assessors were asked to consider an expert (modeling a math professor) judging the relevance of each answer to the topics. This was intended to avoid the ambiguity that might result from guessing the level of math knowledge of the actual posters of the original Math Stack Exchange question. The definitions of the four levels of relevance are shown in Table 2. In judging relevance, ARQMath assessors were asked not to consider any link outside the ARQMath collection. For example, if there is a link to a Wikipedia page, which provides relevant information, the information in the Wikipedia page should not be considered to be a part of the answer.

4.5. Assessor Selection

Paid ARQMath-3 assessors were recruited over email at the Rochester Institute of Technology. 44 students expressed interest, 11 were invited to perform 3 sample assessment tasks, and 9 students specializing in mathematics or computer science were then selected, based on an evaluation of their judgments by an expert mathematician. Of those, 6 were assigned to Tasks 1 and 3; the others performed assessment for Task 2.

Assessment tool. As with ARQMath-1 and ARQMath-2, we used Turkle, a system similar to Amazon Mechanical Turk. As shown in Figure 2, there are two panes, one having the question topic (left pane) and the other having a candidate answer from the judgment pool (right panel). For each topic, the title and question body are provided for the assessors. To familiarize themselves with the topic question, assessors can click on the Thread link for the question, which shows the question and the answers given to it (i.e., answers posted in 2021, which were not available to task participants), along with other information such as tags and comments. Another Thread link is also available for the answer post being assessed. By clicking on that link, the assessor can see a copy of the original question thread on Math Stack Exchange in which the candidate answer was given, as recorded in the March 2020 snapshot used for the ARQMath test collection.

Note that these Thread links are provided to help the assessors gain just-in-time knowledge that they might need for unfamiliar concepts, but the content of the threads is neither a part of the topic nor of the answer being assessed, and thus it should have no effect on their judgement beyond serving as reference information.

In the right pane, below the candidate answer, assessors can indicate the relevance degree. In addition to four relevance degrees, there are two additional choices: 'System failure' to indicate system issues such as unintelligible rendering of formulae, and 'Do not know' which can be used if after possibly consulting external sources such as Wikipedia or viewing the Threads the assessor is simply not able to decide the relevance degree. We asked the assessors to leave a comment in the event of a 'System failure' or 'Do not know' selection.

Assessor Training. All training was done remotely, over Zoom, in four sessions, with some individual assessment practice between each Zoom session. As in ARQMath-1 and -2, in the first session the task and relevance criteria were explained. A few examples were then shown to the assessors and they were asked for their opinions on relevance, which were then discussed with an expert assessor (a math professor). Then, three rounds of training were conducted, with each round consisting of assessment of small judgment pools for four sample topics from ARQMath-2. For each topic, 5-6 answers with different ground truth relevance degrees (from the ARQMath-2 qrels) were chosen. After each round, we held a Zoom session to discuss their relevance judgements, with the specific goal of clarifying their understanding of the relevance criteria. The assessors discussed the reasoning for their choices, with organizers (always including the math professor) sharing their own judgments and their supporting reasoning. The primary

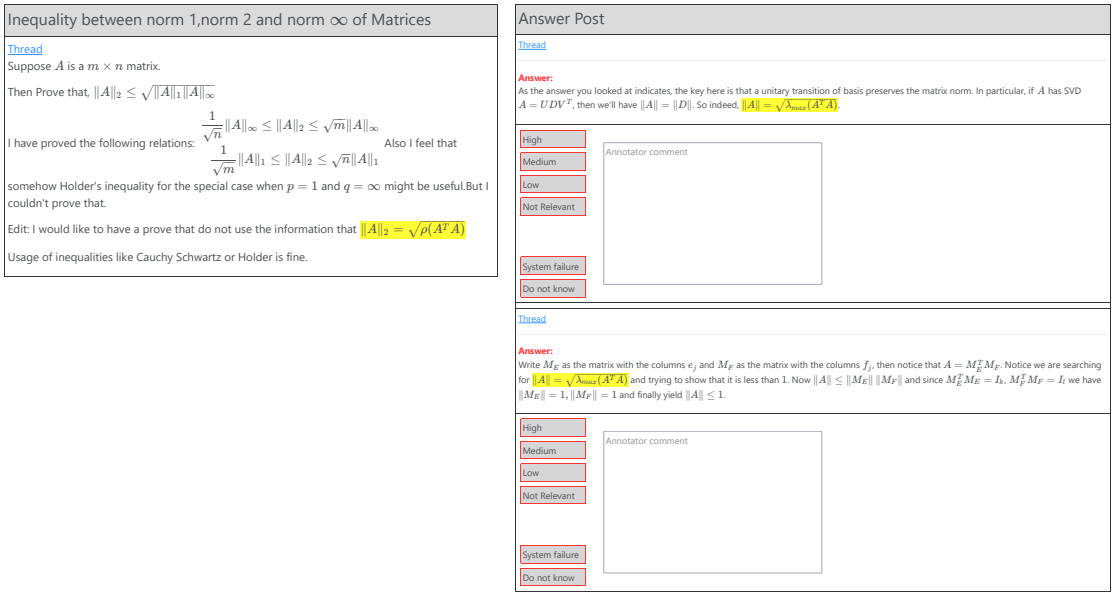

Figure 2: Turkle Assessment Interface. Shown are hits for Formula Retrieval (Task 2). In the left pane, the formula query is highlighted. In the right pane, two answer posts containing the same retrieved formula are shown. For Task 1, the same interface was used, but without formula highlighting, and presenting only one answer post at a time.

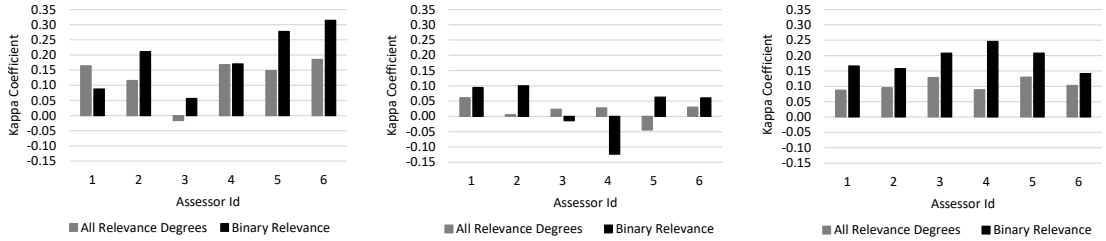

Figure 3: Inter-annotator agreement for 6 assessors during training sessions for Task 1 (mean Cohen's kappa), with four-way classification in gray, and two-way classification (H+M binarized) in black. Left-to-right: agreements for rounds 1, 2, and 3.

goal of training was to help assessors make self-consistent annotations, as topic interpretations will vary across individuals. Some of the topics involve issues that are not typically covered in regular undergraduate courses, and some such cases required the assessors to get a basic understanding of those issue before they could do the assessment. The assessors found the question Threads made available in the Turkle interface helpful in this regard (see Figure 2).

Figure 3 shows average Cohen's kappa coefficients for agreement between each assessor and all others during training. Collapsing relevance to binary by considering only high and medium as relevant (henceforth "H+M binarization") yielded better agreement among the assessors.[9] The agreement values in the second round are unusually low, but the third round agreement is in line with what we had seen at the end of training in prior years.

Assessment Results. Among 80 topics assessed, two (A.335 and A.367) had only one answer assessed as high or medium; these two topics were removed from the collection as score quantization for MAP′ can be quite substantial when only a single relevant document contributes to the computation. For the remaining 78 topics, an average of 446.8 answers were assessed, with an average assessment time of 44.1 seconds per answer post. The average number of answers labeled with any degree of relevance (high, medium, or low; henceforth "H+M+L binarization") over those 78 topics was 100.8 per question (twice as high as that seen in ARQMath-2), with the highest number being 295 (for topic A.317) and the lowest being 11 (for topic A.385).

Post Assessment. After assessments of 80 topics for Task 1 were done, each of the assessors for this task, assessed one topic assessed by another assessor.[10] With Cohen's kappa coefficient, a kappa of 0.24 was achieved on the four-way assessment task, and with H+M binarization, the average kappa value was 0.25.

4.6. Evaluation Measures

While this is the third year of the ARQMath lab, with several relatively mature systems participating, it is still possible that many relevant answers may remain undiscovered. To support fair comparisons with future systems that may find different documents, we have adopted evaluation

[9] H+M binarization corresponds to the definition of relevance usually used in the Text Retrieval Conference (TREC).

[10] One assessor (with id 8) was not able to continue assessment.

measures that ignore unjudged answers, rather than adopting the more traditional convention of treating unjudged answers as not relevant. Specifically, the primary evaluation measure for Task 1 is the nDCG′ (read as "nDCG-prime") introduced by Sakai and Kando [13]. nDCG′ is simply the nDCG@1000 that would be computed after removing unjudged documents from the ranked list. This measure has shown better discriminative power and somewhat better system ranking stability (with judgement ablation) compared to the bpref [14] measure that had been adopted for experiments using the NTCIR Math IR collections for similar reasons [12, 15]. Moreover, nDCG′ yields a single-valued measure with graded relevance, whereas bpref, Precision@k, and Mean Average Precision (MAP) all require binarized relevance judgments. As secondary measures, we compute Mean Average Precision (MAP@1000) with unjudged posts removed (MAP′) and Precision at 10 with unjudged posts removed (P′@10). For MAP′ and P′@10 we used H+M binarization. Note that the answers assessed as "System failure" or "Do not know" were not considered for evaluation, thus can be viewed as answers that are not assessed.

4.7. Results

Progress Testing. In addition to their submissions on the ARQMath-3 topics, we asked each participating team to also submit results from exactly the same systems on ARQMath-1 and ARQMath-2 topics for progress testing. Note, however, that ARQMath-3 systems could be trained on topics from ARQMath-1 and -2; Together, there were 158 topics (77 from ARQMath-1, 81 from ARQMath-2) that could be used for training. The progress test results thus need to be interpreted with this train-on-test potential in mind. Progress test results are provided in Table 3.

ARQMath-3 Results. Table 3 also shows results for ARQMath-3 Task 1. This table shows baselines first, followed by teams, and within teams their systems, ranked by nDCG′. As seen in the table, the manual primary run of the approach0 team achieved the best results, with 0.508 nDCG′. Among automatic runs, nDCG′, 0.504, was achieved by the MSM team. Note that the highest possible nDCG′ and MAP′ values are 1.0, but because fewer than 10 assessed relevant answers (with H+M binarization) were found in the pools for some topics, the highest possible P′@10 value in ARQMath-3 Task 1 is 0.95.

5. Task 2: Formula Search

The goal of the formula search task is to find a ranked list of formula instances from both questions and answers in the collection that are relevant to a formula query. The formula queries are selected from the questions in Task 1. One formula was selected from each Task 1 question topic to produce Task 2 topics. For cases in which suitable formulae were present in both the title and the body of the Task 1 question, we selected the Task 2 formula query from the title. For each query, a ranked list of 1,000 formulae instances were returned by their identifiers in the `<span>` XML elements and the accompanying TSV LaTeX formula index file, along with their associated post identifiers.

While in Task 1, the goal was to find relevant answers for the questions, in Task 2, the goal is to find relevant formulae that are associated with information that can help to satisfy an information need. The post in which a formula is found need not be relevant to the question post

in which the formula query originally appeared for a formula to be relevant to a formula query, but those post contexts inform the interpretation of each formula (e.g., by defining operations and identifying variable types). A second difference is that the retrieved formulae instances in Task 2 can be found in either question posts or answer posts, whereas in Task 1, only answer posts were retrieved.

Finally, in Task 2, we distinguish visually distinct formulae from instances of those formulae, and systems are evaluated by the ranking of the visually distinct formulae they return. The same formula can appear in different posts, and we call these individual occurrences *formula instances*. A *visually distinct formula* is a formula associated with a set of instances that are visually identical when viewed in isolation. For example, x^2 is a formula, $x \cdot x$ is a different (i.e., visually distinct) formula, and each time x^2 appears, it is an instance of the visually distinct formula x^2. Although systems in Task 2 rank formula instances in order to support the relevance judgment process, the evaluation measure for Task 2 is based on the ranking of visually distinct formulae. As shown by Mansouri et al. (2021) [7], using visually-distinct formulae for evaluation can result in a different preference order between systems than would evaluation on formula instances.

5.1. Topics

Each formula query was selected from a Task 1 topic. Similarly to Task 1, Task 2 topics were provided in XML in the format shown in Figure 1. Differences are:

1. **Topic Id.** Task 2 topic ids are in the form "B.x" where x is the topic number. There is a correspondence between topic id in tasks 1 and 2. For instance, topic id "B.384" indicates the formula is selected from topic "A.384" in Task 1, and both topics include the same question post (see Figure 1).
2. **Formula Id**. This added field specifies the unique identifier for the query formula instance. There may be other formulae in the Title or Body of the same question post, but the formula query is only the formula instance specified by this Formula_Id.
3. **LaTeX.** This added field is the LaTeX representation of the query formula instance, as found in the question post.

As the query formulae are selected from Task 1 questions, the same LaTeX, SLT and OPT TSV files that were provided for the Task 1 topics can be used when SLT or OPT representations for a query formula are needed.

Formulae for Task 2 were manually selected using a heuristic approach to stratified sampling over two criteria: complexity and elements. Formula complexity was labeled low, medium or high by the third author. For example, $[x, y] = x$ is low complexity, $\int \frac{1}{(x^2+1)^n} dx$ is medium complexity, and $\frac{\sqrt{1-p^2}}{2\pi(1-2p\sin(\varphi)\cos(\varphi))}$ is high complexity. These annotations, available in an auxiliary file, can be useful as a basis for fine-grained result analysis, since formula queries of differing complexity may result in different preference orders between systems [16]. For elements, our intuition was to make sure that we have formula queries that contain different elements and math phenomena such as integral, limit, and matrices.

5.2. Participant Runs

A total of 19 runs were received for Task 2 from a total of five teams, as shown in Table 1. Among the participating runs, 5 were annotated as manual and the others were automatic. Each run retrieved up to 1,000 formula instances for each formula query, ranked by relevance to that query. For each retrieved formula instance, participating teams provided the `formula_id` and the associated `post_id` for that formula. Please see the participant papers in the working notes for descriptions of the systems that generated these runs.

5.3. Baseline Run: Tangent-S

Tangent-S [12] is the baseline system for ARQMath-3 Task 2. That system accepts a formula query without using any associated text from its associated question post. Since a single formula is specified for each Task 2 query, the formula selection step in the Task 1 Tangent-S baseline is not needed for Task 2. Timing was similar to that of Tangent-S in ARQMath-1 and -2 (i.e., with an average retrieval time of around six seconds per query).

5.4. Assessment

Pooling. For each topic, participants were asked to rank up to 1,000 formula instances. However, the pooling was done using visually distinct formulae. The visual ids, which were provided beforehand for the participants, were used for clustering formula instances. Pooling was done by going down each ranked list until k visually distinct formulae were found. For primary runs (and the baseline system), the first 25 visually distinct formulae were pooled; for alternate runs, the first 15 visually distinct formulae were pooled.

The visual Ids used for clustering retrieval results were determined by the SLT representation when possible, and the LaTeX representation otherwise. When SLT was available, we used Tangent-S [12] to create a string representation using a depth-first traversal of the SLT, with each SLT node and edge generating a single item in the SLT string. Formula instances with identical SLT strings were then considered to be the same formula. For formula instances with no Tangent-S SLT string available, we removed the white space from their LaTeX strings and grouped formula instances with identical LaTeX strings. This process is simple and appears to be reasonably robust, but it is possible that some visually identical formula instances were not captured due to LaTeXML conversion failures, or where different LaTeX strings produce visually identical formulae (e.g., if subscripts and superscripts appear in a different order in LaTeX).

Task 2 assessment was done on formula instances. For each visually distinct formula at most five instances were selected for assessment. As in ARQMath-2 Task 2, formula instances to be assessed were chosen in a way that prefers highly-ranked instances and that prefers instances returned in multiple runs. This was done using a simple voting protocol, where each instance votes by the sum of its reciprocal ranks within each run, breaking ties randomly. For each query, on average there were 154.35 visually distinct formulae to be assessed, and only 6% of visually distinct formulae had more than 5 instances.

Relevance definition. To distinguish between different relevance degrees, we relied on the definitions in Table 2. The usefulness is defined as the likelihood of the candidate formula being

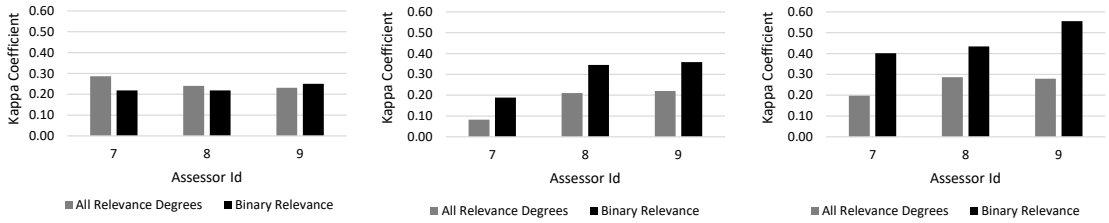

Figure 4: Annotator agreement for 3 assessors during training for Task 2 (mean Cohen's kappa). Four-way classification is shown in gray, and two-way (H+M binarized) classification in black. Left-to-right: agreements for rounds 1, 2, and 3.

associated with information (text) that can help a searcher to accomplish their task. In our case, the task is answering the question from which a query formula is taken.

To judge the relevance of a candidate formula instance, the assessor was given the candidate formula (highlighted) along with the (question or answer) post in which it had appeared. They were then asked to decide on relevance by considering the definitions provided. For each visually distinct formula, up to 5 instances were shown to assessors and they would assess the instances individually. For assessment, they could look at the formula's associated post in an effort to understand factors such as variable types, the interpretation of specific operators, and the area of mathematics it concerns. As in Task 1, assessors could also follow Thread links to increase their knowledge by examining the thread in which the query formula had appeared, or in which a candidate formula had appeared.

Assessment tool. As in Task 1, we used Turkle for the Task 2 assessment process, as illustrated in Figure 2. There are two panes, the left pane showing the formula query ($\|A\|_2 = \sqrt{\rho(A^T A)}$ in this case) highlighted in yellow inside its question post, and the right pane showing the (in this case, two) candidate formula instances of a single visually distinct formula. For each topic, the title and question body are provided for the assessors. Thread links can be used by the assessors just for learning more about mathematical concepts in the posts. For each formula instance, the assessment is done separately. As in Task 1, the assessors can choose between different relevance degrees, they can choose 'System failure' for issues with Turkle, or they can choose 'Do not know' if they are not able to decide on a relevance degree.

Assessor Training. Three paid undergraduate and graduate mathematics and computer science students from RIT were selected to perform relevance judgments. As in Task 1, all training sessions were done remotely, over Zoom.

There were four Task 2 training sessions. In the first meeting, the task and relevance criteria were explained to assessors and then a few examples were shown, followed by discussion about relevance level choices. In each subsequent training round, assessors were asked to first assess four ARQMath-2 Task 2 topics, each with 5-6 visually distinct formula candidates with a variety of relevance degrees. Organizers then met with the assessors to discuss their choices and clarify relevance criteria. Figure 4 shows the average agreement (kappa) of each assessor with the others during training. As can be seen, agreement had improved considerably by round three, reaching levels comparable to that seen in prior years of ARQMath.

Assessment Results. Among 76 assessed topics, all have at least two relevant visually

distinct formulae with H+M binarization, so all 76 topics were retained in the ARQMath-3 Task 2 test collection. An average of 152.3 visually distinct formulae were assessed per topic, with an average assessment time of 26.6 seconds per formula instance. The average number of visually distinct formulae with H+M+L binarization was 63.2 per query, with the highest number being 143 (topic B.305) and the lowest being 2 (topic B.333).

Post Assessment. After Task 2 assessments were done, each of the three assessors, assessed two topics, each assessed by the other two assessors. Using Cohen's kappa coefficient, a kappa of 0.44 was achieved on the four-way assessment task (higher than ARQMath-1 and -2), and with H+M binarization the average kappa value was 0.51.

5.5. Evaluation Measures

As in Task 1, the primary evaluation measure for Task 2 is nDCG′, with MAP′ and P′@10 also reported. Participants submitted ranked lists of formula instances used for pooling, but with evaluation measures computed over visually distinct formulae. The ARQMath-2 Task 2 evaluation script replaces each formula instance with its associated visually distinct formula, and then deduplicates from the top of the list downward, producing a ranked list of visually distinct formulae, from which our "prime" evaluation measures are then computed using `trec_eval`, after removing unjudged visually distinct formulae. For the visually distinct formulae with multiple instances, the maximum relevance score of any judged instance was used as the relevance visually distinct formula's relevance score. This reflects a goal of having at least one instance that provides useful information. Similar to Task 1, formulas assessed as "System failure" or "Do not know" were treated as not being assessed.

5.6. Results

Progress Testing. As with Task 1, we asked Task 2 teams to run their ARQMath-3 systems on ARQMath-1 and -2 Topics for progress testing (see Table 4). Some progress test results may represent a train-on-test condition: there were 70 topics from ARQMath-2 and 74 topics from ARQMath-1 available for training. Note also that while the relevance definition stayed the same for ARQMath-1, -2, and -3, the assessors were instructed differently in ARQMath-1 on how to handle the specific case in which two formulae were visually identical. In ARQMath-1 assessors were told such cases are always highly relevant, whereas ARQMath-2 and ARQMath-3 assessors were told that from context they might recognize cases in which a visually identical formula would be less relevant, or not relevant at all (e.g., where identical notation is used with very different meaning). Assessor instruction did not change between ARQMath-2 and -3.

ARQMath-3 Results. Table 4 also shows results for ARQMath-3 Task 2. In that table, the baseline is shown first, followed by teams and then their systems ranked by nDCG′ on ARQMath-3 Task 2 topics. As shown, the highest nDCG′ was achieved by the manual primary run from the approach0 team, with an nDCG′ value of 0.720. Among automatic runs, the highest nDCG′ value was the DPRL primary run, with an NDCG′ of 0.694. Note that 1.0 is a possible score for nDCG′ and MAP′, but that the highest possible P′@10 value is 0.93 because (with H+M binarization) 10 visually distinct formulae were not found in the pools for some topics.

6. Task 3: Open Domain Question Answering

The new pilot task developed for ARQMath-3 (Task 3) is Open Domain Question Answering. Unlike Task 1, system answers are not limited to content from any specific source. Rather, answers can be *extracted* from anywhere, automatically *generated*, or even written by a person. For example, suppose that we ask a Task 3 system the question "What does it mean for a matrix to be Hermitian?" An extractive system might first retrieve an article about Hermitian matrices from Wikipedia and then extract the following excerpt as the answer: "In mathematics, a Hermitian matrix (or self-adjoint matrix) is a complex square matrix that is equal to its own conjugate transpose." By contrast, a generative system such as GPT-3 can directly construct an answer such as: "A matrix is Hermitian if it is equal to its transpose conjugate." For a survey of open-domain question answering, see Zhu et al. [17]. In this section, we describe the Task 3 search topics, runs from participant and baseline systems, assessment and evaluation procedures, and results.

6.1. Topics and Participant Runs

The topics for Task 3 are the Task 1 topics, with the same content provided (title, question body, and tags). A total of 13 runs were received from 3 teams. Each run consists of a single result for each topic. 9 runs from the TU_DBS and DPRL teams were declared to be automatic and 5 runs from the approach0 team were declared as manual. The 4 automatic runs from the TU_DBS team used generative systems, whereas the remaining 9 runs from the DPRL and approach0 teams used extractive systems. The teams and their submissions are listed in Table 1.

6.2. Baseline Run: GPT-3

The ARQMath organizers provided one baseline run for this task using GPT-3. This baseline system uses the *text-davinci-002* model of GPT-3 [18] from OpenAI. First, the system prompts the model with the text `Q:` followed by the text and the LaTeX formulae of the question, two newline characters, and the text `A:` as follows:

```
Q: What does it mean for a matrix to be Hermitian?

A:
```

Then, GPT-3 completes the text and produces an answer of up to 570 tokens:

```
Q: What does it mean for a matrix to be Hermitian?

A: A matrix is Hermitian if it is equal to its transpose conjugate.
```

If the answer is longer than the maximum of 1,200 Unicode characters, the system retries until the model has produced a sufficiently short answer.

To provide control over how creative an answer is, GPT-3 resmooths the output layer L using the temperature τ as follows: $\text{softmax}(L/\tau)$ [19]. A temperature close to zero ensures deterministic outputs on repeated prompts, whereas higher temperatures allow the model's decoder to consider many different answers. Our system uses the default temperature $\tau = 0.7$.

6.3. Evaluation Measures

In this section, we first describe the measures we used to evaluate participating systems. Then, we describe additional evaluation measures that we have developed with the goal of providing a fair comparison between participating systems and future systems that return answers from outside Math Stack Exchange, or that are generated.

6.3.1. Manual Evaluation Measures

As described in Section 4.4, the assessors produced a relevance score between 0 and 3 for most answers from each participating system. The exceptions were 'System failure' and 'Do not know' assessments, which we interpreted as relevance score 0 ('Not relevant') in our evaluation of Task 3. To evaluate participating systems, we report the Average Relevance (AR) score and Precision@1 (P@1). AR is equivalent to the unnormalized Discounted Cumulative Gain at position 1 (DCG@1).[11] P@1 is computed using H+M binarization.

Task 1 systems approximate a restricted class of Task 3 systems. For this reason, we also report AR and P@1 for ARQMath-3 Task 1 systems in order to extend the range of system comparisons that can be made. To do this, we truncate the Task 1 result lists after the first result. Note, however, that Task 3 answers were limited to a maximum of 1,200 Unicode characters, whereas Task 1 systems had no such limitation. Approximately 15% of all answer posts in the collection are longer than 1,200 Unicode characters when represented as text and LaTeX. Therefore, the Task 3 measures that we report for Task 1 systems should be treated as somewhat optimistic estimates of what might have been achieved by an extractive system that was limited to the ARQMath collection.

6.3.2. Automatic Evaluation Measures

In Task 1, systems pick answers from a fixed collection of potential answers. When evaluated with measures that differentiate between relevant, non-relevant, and unjudged answers, reasonable comparisons can be made between participating systems that contributed to the judgement pools and future systems that did not. By contrast, the open-ended nature of Task 3 means that relevance judgements on results from participating systems can not be used in the same way to evaluate future systems that might (and hopefully will!) generate different answers.

The problem lies in the way AR and P@1 are defined; they rely on our ability to match new answers with judged answers. For future systems, however, the best we might reasonably hope for is similarity between the new answers and the judged answers. If we are to avoid the need to keep assessors around forever, we need automatic evaluation measures that can be used to compare participating Task 3 systems with future Task 3 systems. With that goal in mind, we also report Task 3 results using the following evaluation measures:

1. **Lexical Overlap (LO)** Following SQuAD and CoQA [20, Section 6.1], we represent answers as a bag of tokens, where tokens are produced by the MathBERTa[12] tokenizer.

[11] For ranked lists of depth 1 there is no discounting or accumulation, and in ARQMath the relevance value is used directly as the gain.

[12] https://huggingface.co/witiko/mathberta

For every topic, we compute the token F_1 score between the system's answer and each known relevant Task 3 answer (using H+M binarization). The score for a topic is the maximum across these F_1 scores. The final score is the average across all topics of those per-topic maximum F_1 scores.

2. **Contextual Similarity (CS)** Although lexical overlap can account for answers with high surface similarity, it cannot recognize answers that use different tokens with similar meaning. For context similarity, we use BERTScore [21] with the MathBERTa language model. As with our computation of lexical overlap, for BERTScore we also compute a token F_1 score, but instead of exact matches, we match tokens with the most similar contextual embeddings and interpret their similarity as fractional membership. For every topic, we compute F_1 score between the system's answer and each known relevant answer (with H+M binarization). The score for a topic is the maximum across these F_1 scores. The final score is the average across all topics of those per-topic maximum F_1 scores.

When computing the automatic measures for a participating system, we exclude relevant answers uniquely contributed to the pools by systems from the same team. This ablation avoids the perfect overlap scores that systems contributing to the pools would otherwise get from matching their own results.

6.4. Results

Task 3 runs were assessed together with Task 1 runs, using the same relevance definitions, although after that assessment was complete, we also did some additional annotation that was specific to Task 3. Here we present results for the baseline and submitted runs using manual and automatic measures, along with additional analysis that we performed using the additional annotation.

6.4.1. Manual Evaluation Measures

Table 5 shows ARQMath-3 results for Task 3 systems. This table shows baselines first, followed by teams ordered by their best Average Recall (AR), and within teams their runs are ordered by AR. As seen in the table, the automatic generative baseline run using GPT-3 achieved the best results, with 1.346 AR. Note that uniquely among ARQMath evaluation measures, AR is not bounded between 0 and 1; rather, it is bounded between 0 and 3.[13] Among manual extractive non-baseline runs, the highest AR was achieved by a run from the approach0 team, with 1.282 AR. Among automatic extractive non-baseline runs, the highest AR was achieved by a run from the DPRL team, with 0.462 AR. Among automatic generative non-baseline runs, the highest AR was achieved by the TU_DBS team, with 0.325 AR. No manual generative non-baseline runs were submitted to ARQMath-3 Task 3.

Table 7 shows ARQMath-3 Task 3 results for Task 1 systems. Similarly to Table 5, Table 7 shows baselines first, followed by teams ordered by their best AR, and within teams their runs are ordered by AR. As seen in the table, the *Linked MSE posts* baseline achieved the best

[13] Because some topics have no highly relevant answers, the actual maximum value of AR on the Task 3 topics is 2.346.

result, with 1.608 AR. Among non-baseline runs, the highest AR was achieved by a run from the approach0 team, with 1.377 AR. Among automatic runs, the highest AR was achieved by a run from the TU_DBS team, with 1.192 AR. Compared to ARQMath-3 Task 1 results in Table 3, the TU_DBS team's best run did relatively better, swapping order with the best runs from the MSM and MIRMU. Within teams, the fusion_alpha05 run from approach0, which achieved the highest nDCG′ on Task 1, did not do as well as that team's rerank_nostemmer system when both were scored using Task 3 measures. The RRF-AMR-SVM run from DPRL, which achieved the second highest nDCG′ score among DPRL runs on Task 1, received the lowest AR and P@1 among Task 1 systems. These differences result from the exclusive focus of Task 3 measures on the single highest-ranked result.

6.4.2. Automatic Evaluation Measures

At least one participating system produced a relevant answer (with H+M binarization) for 66 of the 78 Task 3 topics. However, automated evaluation can only be computed with ablation of each team's contributions if two or more of the three teams produced a relevant answer; there were only 35 such topics. We therefore expanded the set of references for automatic Task 3 measures to also include relevant answers (with H+M binarization) that were produced for ARQMath-3 topics by Task 1 systems, but only for relevant answers that were no longer than 1,200 Unicode characters.

As one measure of the suitability of our automatic evaluation measures for the evaluation of future systems, we report paired pointwise correlation measures between our automatic measures and manual measures, using Pearson's r to characterize the linear relationship between the measures, and Kendall's τ to characterize differences in how the evaluation measures rank systems.

Table 5 also shows results for automatic evaluation measures. The automatic generative baseline run using GPT-3, which achieved the best result using manual measures, scored below extractive runs from the approach0 and DPRL teams on both automatic measures. We theorize that this is because we used relevant answers produced by Task 1 systems in our automatic measures, which favors extractive systems over generative systems, because identical hits may be retrieved by extractive systems. Both automatic measures maintained the ordering of teams given by the manual measures.

Table 6 shows pointwise correlations between the manual and automatic measures. Both automatic measures show a strong linear relationship to the manual measures, with lexcial overlap (LO) and average relevance (AR) having Pearson's r of 0.837, and contextual similarity (CS) and AR having Pearson's r of 0.839. LO is better able to maintain the ordering of results given by the manual measures, having Kendall's τ with AR of 0.736, compared to CS, which has Kendall's τ with AR of 0.670. Furthermore, LO is also more easily interpretable than CS, because it only considers exact matches between tokens, and is independent of a specific BERT language model, which may have to be replaced in the future. This suggests that LO may be preferable as an automatic measure to evaluate future Math OpenQA systems.

6.4.3. Characterizing Answers

The answers for Task 3 were assessed together with the Task 1 results, using the same relevance definitions. We also provided a sample of Task 1 and Task 3 answers to assessors, and asked them to annotate:

1. Whether answers were machine-generated
2. Whether answers contained information unrelated to the topic question

In Tasks 1 and 3, answers are considered relevant if any part of the answer is relevant to the question. Annotating unrelated information allows us to determine whether extractive systems stuff answers with unrelated information, perhaps in the hope that some of it will be relevant, and whether generative systems generate off-topic content together with on-topic content. To support that analysis, assessors were asked to differentiate between undesirable answer stuffing and the possibly desirable inclusion of background information that is related to the question or relevant part(s) of the answer.

We report the answers to these questions using two measures:

1. **Machine-Generated (MG).** The fraction of answers assessed as machine-generated. Ideally this would always be zero, but in practice we are interested in whether it is larger for generative systems than for extractive systems.
2. **Unrelated Information (UI).** The fraction of answers assessed as containing information unrelated to the question. Again, ideally this would be zero.

We report these measures as averages over 73 of the 78 Task 3 topics because one assessor was unable to complete this post-evaluation assessment process.[14]

Table 5 includes results for these measures. The manual extractive run of approach0 produced the smallest fraction of answers annotated as machine-generated (11%). Among generative runs, the automatic baseline run using GPT-3 produced the fewest answers annotated as machine-generated (28.8%). With the exception of the automatic extractive SBERT-QQ-AMR run from DPRL, which had 34.2% of answers annotated as machine-generated, the generative runs are linearly separable from the extractive runs (by MG > 0.26). This suggests that even though people would perform worse than chance at identifying answers as machine generated for systems such as GPT-3, they would often be able to differentiate between extractive and generative systems after seeing many answers from a system.

We also see that UI has a strong inverse correlation with AR, with Pearson's r of -0.97 and Kendell's τ of -0.88. Moreover, we also see that 90.43% of answers that were annotated as containing information unrelated to the question had been assessed as not relevant (with H+M binarization), whereas only 79.03% of all answers were annotated as not relevant (with H+M binarization). That suggests that answer stuffing does not seem to have been a serious problem in our evaluation.

[14]The five topics for which results were not characterized in this way are A.301, A.314, A.322, A.324, and A.350.

7. Conclusion

Over the course of three years, ARQMath has created test collections for three tasks that together include relevance judgments for hundreds of topics for two of those tasks, and 78 topics for the third. Coming as it did at the dawn of the neural age in information retrieval, considerable innovation in methods has been evident throughout the three years of the lab. ARQMath has included substantial innovation in evaluation design as well, including better contextualized definitions for graded relevance, and piloting a new task on open domain question answering. Having achieved our twin goals of building a new test collection from Math Stack Exchange posts and bringing together a research community around that test collection, the time as now come to end this lab at CLEF. We expect, however, that both that collection and that community will continue to contribute to advancing the state of the art in Math IR for years to come.

7.0.1. Acknowledgements

We thank our student assessors from RIT: Duncan Brickner, Jill Conti, James Hanby, Gursimran Lnu, Megan Marra, Gregory Mockler, Tolu Olatunbosun, and Samson Zhang. This material is based upon work supported by the National Science Foundation (USA) under Grant No. IIS-1717997 and the Alfred P. Sloan Foundation under Grant No. G-2017-9827.

References

[1] B. Mansouri, R. Zanibbi, D. W. Oard, Characterizing Searches for Mathematical Concepts, in: Joint Conference on Digital Libraries (JCDL), 2019.

[2] A. Aizawa, M. Kohlhase, I. Ounis, NTCIR-10 Math Pilot Task Overview, in: Proceedings of the 10th NTCIR, 2013.

[3] A. Aizawa, M. Kohlhase, I. Ounis, NTCIR-11 Math-2 Task Overview, in: Proceedings of the 11th NTCIR, 2014.

[4] R. Zanibbi, A. Aizawa, M. Kohlhase, I. Ounis, G. Topic, K. Davila, NTCIR-12 MathIR Task Overview, in: Proceedings of the 12th NTCIR, 2016.

[5] M. Líška, P. Sojka, M. Růžička, P. Mravec, Web Interface and Collection for Mathematical Retrieval WebMIaS and MREC (2011).

[6] Y. Stathopoulos, S. Teufel, Retrieval of Research-level Mathematical Information Needs: A Test Collection and Technical Terminology Experiment, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2015.

[7] B. Mansouri, D. W. Oard, A. Agarwal, R. Zanibbi, Effects of context, complexity, and clustering on evaluation for math formula retrieval, arXiv preprint arXiv:2111.10504 (2021).

[8] M. Hopkins, R. Le Bras, C. Petrescu-Prahova, G. Stanovsky, H. Hajishirzi, R. Koncel-Kedziorski, SemEval-2019 Task 10: Math Question Answering, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019.

[9] J. Meadows, A. Freitas, A Survey in Mathematical Language Processing, arXiv preprint arXiv:2205.15231 (2022).

[10] C. Macdonald, N. Tonellotto, Declarative Experimentation in Information Retrieval using PyTerrier, in: Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval, 2020.
[11] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, D. Johnson, Terrier Information Retrieval Platform, in: European Conference on Information Retrieval, Springer, 2005.
[12] K. Davila, R. Zanibbi, Layout and Semantics: Combining Representations for Mathematical Formula Search, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017.
[13] T. Sakai, N. Kando, On Information Retrieval Metrics Designed for Evaluation with Incomplete Relevance Assessments, Information Retrieval (2008).
[14] C. Buckley, E. M. Voorhees, Retrieval Evaluation with Incomplete Information, in: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2004.
[15] B. Mansouri, S. Rohatgi, D. W. Oard, J. Wu, C. L. Giles, R. Zanibbi, Tangent-CFT: An Embedding Model for Mathematical Formulas, in: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR), 2019.
[16] B. Mansouri, R. Zanibbi, D. W. Oard, Learning to Rank for Mathematical Formula Retrieval, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021.
[17] F. Zhu, W. Lei, C. Wang, J. Zheng, S. Poria, T.-S. Chua, Retrieving and Reading: A Comprehensive Survey on Open-Domain Question Answering, arXiv preprint arXiv:2101.00774v3 (2021).
[18] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language Models are Few-Shot Learners, 2020.
[19] J. Ficler, Y. Goldberg, Controlling linguistic style aspects in neural language generation, in: Proceedings of the Workshop on Stylistic Variation, Association for Computational Linguistics, 2017.
[20] S. Reddy, D. Chen, C. D. Manning, CoQA: A Conversational Question Answering Challenge, Transactions of the Association for Computational Linguistics (2019).
[21] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, arXiv preprint arXiv:1904.09675 (2019).

Table 3
ARQMath 2022 Task 1 (CQA) results. **P**: primary run, **M**: manual run, **(✓)**: baseline pooled as a primary run. For MAP′ and P′@10, H+M binarization was used. (D)ata indicates use of (T)ext, (M)ath, (B)oth text and math, or link structure (*L).

		Type		ARQMath-1 77 Topics			ARQMath-2 71 topics			ARQMath-3 78 topics		
Run	D	P	M	nDCG′	MAP′	P′@10	nDCG′	MAP′	P′@10	nDCG′	MAP′	P′@10
Baselines												
TF-IDF(Terrier)	B			0.204	0.049	0.073	0.185	0.046	0.063	0.272	0.064	0.124
TF-IDF(PyTerrier) +Tangent-S	B	(✓)		0.249	0.059	0.081	0.158	0.035	0.072	0.229	0.045	0.097
TF-IDF(PyTerrier)	B			0.218	0.079	0.127	0.120	0.029	0.055	0.190	0.035	0.065
Tangent-S	M			0.158	0.033	0.051	0.111	0.027	0.052	0.159	0.039	0.086
Linked MSE posts	*L	(✓)		0.279	0.194	0.384	0.203	0.120	0.282	0.106	0.051	0.168
approach0												
fusion_alpha05	B	✓	✓	0.462	0.244	0.321	0.460	0.226	0.296	**0.508**	**0.216**	**0.345**
fusion_alpha03	B		✓	0.460	0.246	0.312	0.450	0.221	0.278	0.495	0.203	0.317
fusion_alpha02	B		✓	0.455	0.243	0.309	0.443	0.217	0.266	0.483	0.195	0.305
rerank_nostemer	B		✓	0.382	0.205	0.322	0.385	0.187	0.276	0.418	0.172	0.309
a0porter	B		✓	0.373	0.204	0.270	0.383	0.185	0.241	0.397	0.159	0.271
MSM												
Ensemble_RRF	B	✓		0.422	0.172	0.197	0.381	0.119	0.152	0.504	0.157	0.241
BM25_system	B			0.332	0.123	0.168	0.285	0.082	0.116	0.396	0.122	0.194
BM25_Tfldf _system	B			0.332	0.123	0.168	0.286	0.083	0.116	0.396	0.122	0.194
TF-IDF	B			0.238	0.074	0.117	0.169	0.040	0.076	0.280	0.064	0.081
CompuBERT22	B			0.115	0.038	0.099	0.098	0.030	0.090	0.130	0.025	0.059
MIRMU												
MiniLM+RoBERTa	B	✓		0.466	0.246	0.339	0.487	0.233	0.316	0.498	0.184	0.267
MiniLM +MathRoBERTa	B			0.466	0.246	0.339	0.484	0.227	0.310	0.496	0.181	0.273
MiniLM_tuned +MathRoBERTa	B			0.470	0.240	0.335	0.472	0.221	0.309	0.494	0.178	0.262
MiniLM_tuned +RoBERTa	B			0.466	0.246	0.339	0.487	0.233	0.316	0.472	0.165	0.244
MiniLM+RoBERTa	T			0.298	0.124	0.201	0.277	0.104	0.180	0.350	0.107	0.159
MathDowsers												
L8_a018	B	✓		0.511	0.261	0.307	0.510	0.223	0.265	0.474	0.164	0.247
L8_a014	B			0.513	0.257	0.313	0.504	0.220	0.265	0.468	0.155	0.237
L1on8_a030	B			0.482	0.241	0.281	0.507	0.224	0.282	0.467	0.159	0.236
TU_DBS												
math_10	B	✓		0.446	0.268	0.392	0.454	0.228	0.321	0.436	0.158	0.263
Khan_SE_10	B			0.437	0.254	0.357	0.437	0.214	0.309	0.426	0.154	0.236
base_10	B			0.438	0.252	0.369	0.434	0.209	0.299	0.423	0.154	0.228
roberta_10	B			0.438	0.254	0.372	0.446	0.224	0.309	0.413	0.150	0.226
math_10_add	B			0.421	0.264	0.405	0.566	0.445	0.589	0.379	0.149	0.278
DPRL												
SVM-Rank	B	✓		0.508	0.467	0.604	0.533	0.460	0.596	0.283	0.067	0.101
RRF-AMR-SVM	B			**0.587**	**0.519**	**0.625**	**0.582**	**0.490**	**0.618**	0.274	0.054	0.022
QQ-QA-RawText	B			0.511	0.467	0.604	0.532	0.460	0.597	0.245	0.054	0.099
QQ-QA-AMR	B			0.276	0.180	0.295	0.186	0.103	0.237	0.185	0.040	0.091
QQ-MathSE-AMR	B			0.231	0.114	0.218	0.187	0.069	0.138	0.178	0.039	0.081
SCM												
interpolated_text +positional_word 2vec_tangentl	B	✓		0.254	0.102	0.182	0.197	0.059	0.149	0.257	0.060	0.119
joint_word2vec	B			0.247	0.105	0.187	0.183	0.047	0.106	0.249	0.059	0.106
joint_tuned _roberta	B			0.248	0.104	0.187	0.184	0.047	0.109	0.249	0.059	0.105
joint_positional _word2vec	B			0.247	0.105	0.190	0.184	0.047	0.109	0.248	0.059	0.105
joint_roberta_base	T			0.135	0.048	0.101	0.099	0.023	0.060	0.188	0.040	0.077

Table 4
ARQMath 2022 Task 2 (Formula Retrieval) results. **P**: primary run, **M**: manual run, **(✓)**: baseline pooled as a primary run. MAP′ and P′@10 use H+M binarization. Baseline results in parentheses. DATA indicates sources used by systems: (M)ath, or (B)oth math and text.

		TYPE		ARQMATH-1 45 TOPICS			ARQMATH-2 58 TOPICS			ARQMATH-3 76 TOPICS		
RUN	DATA	P	M	NDCG′	MAP′	P′@10	NDCG′	MAP′	P′@10	NDCG′	MAP′	P′@10
Baselines												
Tangent-S	M	(✓)		0.691	0.446	0.453	0.492	0.272	0.419	0.540	0.336	0.511
approach0												
fusion_alph05	M	✓	✓	0.647	0.507	0.529	**0.652**	0.471	0.612	**0.720**	**0.568**	**0.688**
fusion_alph03	M		✓	0.644	0.513	0.520	0.649	0.470	0.603	**0.720**	0.565	0.665
fusion_alph02	M		✓	0.633	0.502	0.513	0.646	0.469	0.597	0.715	0.558	0.659
a0	M		✓	0.582	0.446	0.477	0.573	0.420	0.588	0.639	0.501	0.615
fusion02_ctx	B		✓	0.575	0.448	0.496	0.575	0.417	0.590	0.631	0.490	0.611
DPRL												
TangentCFT2ED	M	✓		0.648	0.480	0.502	0.569	0.368	0.541	0.694	0.480	0.611
TangentCFT2	M			0.607	0.438	0.482	0.552	0.350	0.510	0.641	0.419	0.534
T-CFT2TED+MathAMR	B			0.667	0.526	**0.569**	0.630	**0.483**	**0.662**	0.640	0.388	0.478
LTR	M			**0.733**	**0.532**	0.518	0.550	0.333	0.491	0.575	0.377	0.566
MathAMR	B			0.651	0.512	0.567	0.623	0.482	0.660	0.316	0.160	0.253
MathDowsers												
latex_L8_a040	M			0.657	0.460	0.516	0.624	0.412	0.524	0.640	0.451	0.549
latex_L8_a035	M			0.659	0.461	0.516	0.619	0.410	0.522	0.640	0.450	0.549
L8	M	✓		0.646	0.454	0.509	0.617	0.409	0.510	0.633	0.445	0.549
XYPhoc												
xy7o4	M			0.492	0.316	0.433	0.448	0.250	0.435	0.472	0.309	0.563
xy5	M			0.419	0.263	0.403	0.328	0.168	0.391	0.369	0.211	0.518
xy5IDF	M	✓		0.379	0.241	0.374	0.317	0.156	0.391	0.322	0.180	0.461
JU_NITS												
formulaL	M	✓		0.238	0.151	0.208	0.178	0.078	0.221	0.161	0.059	0.125
formulaO	M			0.007	0.001	0.009	0.182	0.101	0.367	0.016	0.008	0.001
formulaS	M			0.000	0.000	0.000	0.142	0.070	0.159	0.000	0.000	0.000

Table 5
ARQMath 2022 Task 3 (Open Domain QA) results for Task 3 systems. **P**: primary run, **M**: manual run, **G**: generative system, **(✓)**: baseline pooled as primary run. All runs use (B)oth math and text. P@1 uses H+M binarization. AR: Average Relevance. LO: Lexical Overlap metric. CS: Contextual Similarity metric. MG: Ratio of answers assessed as Machine-Generated. UI: Ratio of answers with Unrelated Information. Task 3 topics are the same as Task 1 topics except for MG and UI, where we only use a subset of 73 topics. Baseline results are in parentheses.

		Type			78 Topics				73 Topics	
Run	Data	P	M	G	AR	P@1	LO	CS	MG	UI
Baselines										
GPT-3	B	(✓)		✓	**(1.346)**	**(0.500)**	0.317	0.851	0.288	**(0.466)**
approach0										
run1	B		✓		**1.282**	**0.436**	0.509	**0.886**	**0.110**	**0.562**
run4	B		✓		1.231	0.397	**0.515**	**0.886**	0.123	0.616
run3	B		✓		1.179	0.372	0.467	0.879	0.247	0.658
run2	B		✓		1.115	0.321	0.427	0.868	0.164	0.616
run5	B	✓	✓		0.949	0.282	0.444	0.873	0.151	0.671
DPRL										
SBERT-SVMRank	B				0.462	0.154	0.330	0.846	0.205	0.767
BERT-SVMRank	B	✓			0.449	0.154	0.329	0.846	0.178	0.808
SBERT-QQ-AMR	B				0.423	0.128	0.325	0.852	0.342	0.877
BERT-QQ-AMR	B				0.385	0.103	0.323	0.851	0.260	0.863
TU_DBS										
amps3_se1_hints	B			✓	0.325	0.078	0.263	0.835	0.833	0.931
se3_len_pen_10	B			✓	0.244	0.064	0.248	0.806	0.877	0.890
amps3_se1_len_ pen_20_sample_hint	B			✓	0.231	0.051	0.254	0.813	0.959	0.932
shortest	B	✓		✓	0.205	0.026	0.239	0.820	0.849	0.918

Table 6
ARQMath 2022 Task 3 (Open Domain QA) correlations between automatic and manual evaluation measures from Table 5. P@1 uses H+M binarization. AR: Average Relevance. LO: Lexical Overlap metric. CS: Contextual Similarity metric. Task 3 topics are the same as Task 1 topics.

	AR	P@1	LO	CS
AR	1.000	0.989	0.837	0.839
P@1	0.989	1.000	0.787	0.802
LO	0.837	0.787	1.000	0.952
CS	0.839	0.802	0.952	1.000

(a) Pearson's r

	AR	P@1	LO	CS
AR	1.000	0.994	0.736	0.670
P@1	0.994	1.000	0.729	0.674
LO	0.736	0.729	1.000	0.805
CS	0.670	0.674	0.805	1.000

(b) Kendall's τ

Table 7
ARQMath 2022 Task 3 (Open Domain QA) results for Task 1 systems. **P**: primary run, **M**: manual run, **(✓)**: baseline pooled as a primary run. P@1 uses H+M binarization. AR: Average Relevance. (D)ata indicates use of (T)ext, (M)ath, (B)oth text and math, or link structure (*L). Baseline results are in parenthesis.

		Type		78 topics	
Run	Data	P	M	AR	P@1
Baselines					
Linked MSE posts	*L	(✓)		**(1.608)**	**(0.541)**
TF-IDF(Terrier)	B			0.590	0.154
TF-IDF(PyTerrier)+Tangent-S	B	(✓)		0.513	0.167
Tangent-S	M			0.410	0.128
TF-IDF(PyTerrier)	B			0.333	0.051
approach0					
rerank_nostemer	B		✓	**1.377**	**0.481**
fusion_alpha05	B	✓	✓	1.247	0.468
fusion_alpha03	B		✓	1.077	0.385
fusion_alpha02	B		✓	0.974	0.346
a0porter	B		✓	0.885	0.321
TU_DBS					
math_10	B	✓		1.192	0.372
math_10_add	B			1.128	0.321
Khan_SE_10	B			1.103	0.333
base_10	B			1.038	0.295
roberta_10	B			0.910	0.269
MIRMU					
MiniLM+RoBERTa	B	✓		1.143	0.377
MiniLM_tuned+RoBERTa	B			1.141	0.372
MiniLM+MathRoBERTa	B			1.013	0.338
MiniLM_tuned+MathRoBERTa	B			0.974	0.308
MiniLM+RoBERTa	T			0.679	0.205
MSM					
Ensemble_RRF	B	✓		1.026	0.295
BM25_system	B			0.718	0.218
BM25_TfIdf_system	B			0.705	0.218
TF-IDF	B			0.423	0.141
CompuBERT22	B			0.256	0.051
MathDowsers					
L8_a018	B	✓		1.038	0.333
L1on8_a030	B			0.936	0.308
L8_a014	B			0.910	0.282
DPRL					
QQ-QA-RawText	B			0.577	0.179
QQ-QA-AMR	B			0.526	0.179
SVM-Rank	B	✓		0.474	0.128
QQ-MathSE-AMR	B			0.423	0.128
RRF-AMR-SVM	B			0.064	0.013
SCM					
interpolated_text+positional_word2vec_tangentl	B	✓		0.551	0.179
joint_word2vec	B			0.551	0.154
joint_tuned_roberta	B			0.551	0.154
joint_positional_word2vec	B			0.551	0.154
joint_roberta_base	T			0.333	0.077