Expanding Spatial Regions and Incorporating IDF for PHOC-Based Math Formula Retrieval at ARQMath-3

Matt Langsenkamp, Behrooz Mansouri and Richard Zanibbi

Rochester Institute of Technology, 1 Lomb Memorial Drive, Rochester, NY, USA 14623

Abstract

For ARQMath 2022, the XY-PHOC Team has built upon their previous work from ARQMath 2021. We submitted multiple runs for the Formula Retrieval task (Task 2). Pyramidal Histogram of Character (PHOC) formula encodings capture the two-dimensional layout of symbols, using binary vectors to indicate the spatial regions where symbols appear. Our updated PHOC models add additional X/Y partitions, and introduce ellipsoidal regions to capture symmetric layouts (e.g., '$x + y$' and '$y + x$'). Our PHOC models were reimplemented in OpenSearch, which has decreased Mean Response Times (MRTs) by orders of magnitude. We also explored incorporating Inverse Document Frequency (IDF) weights in our PHOC similarity function. Despite their simplicity, our PHOC models were competitive at ARQMath 2022 in P'@10 measures. Our IDF models did not perform as expected, and we will be exploring ways to improve them in the future. We also introduce a new tool for visually comparing ARQMath Task 2 runs (`ARQMathCompare`). This along with Python libraries for PHOC embedding and retrieval (`AnyPHOC` and `PHOCindexing`) are publicly available at https://gitlab.com/dprl.

1. Introduction

The XY-PHOC team from The Document and Pattern Recognition Lab (DPRL) from the Rochester Institute of Technology (USA) participated in the ARQMath 2022 Formula Retrieval Task (Task 2). Task 2 deals with returning relevant formula based on a query formula taken from a Math Stack Exchange (MSE) post [1, 2]. In this paper, we continue our work on using spatial features for *isolated* formula retrieval, i.e., where both query and candidate formulas are considered without any surrounding context. Ad-hoc formula queries used to define unfamiliar notation is an example of where isolated formula retrieval may be useful. Browsing is another use case, where for example a user explores technical papers using a single formula query.

ARQMath Task 2 is actually a *contextualized* formula retrieval task, designed to test a systems' ability to retrieve relevant formula based on a given formula query and the question post where it appears. The relevance of a retrieved formula takes into account the question to be answered from which the formula is taken; for example, this is influenced by variable types and value ranges, and operator definitions that are not considered when retrieving formulas in isolation. Each Task 2 topic includes the topic number, the query formula identifier within the ARQMath

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

kulyabov-ds@rudn.ru (M. Langsenkamp); i.tiddi@vu.nl (B. Mansouri); Manfred.Jeusfeld@acm.org (R. Zanibbi)
https://yamadharma.github.io/ (M. Langsenkamp); https://kmitd.github.io/ilaria/ (B. Mansouri); http://conceptbase.sourceforge.net/mjf/ (R. Zanibbi)
0000-0002-0877-7063 (M. Langsenkamp); 0000-0001-7116-9338 (B. Mansouri); 0000-0002-9421-8566 (R. Zanibbi)

CEUR Workshop Proceedings (CEUR-WS.org)

collection, the query formula in LaTeX, and the full question or answer post that the query formula is taken from (with the formula included in-context). The collection to be searched for Task 2 includes both question and answer posts from Math Stack Exchange (MSE) [2]. The ARQMath collection includes posts published between 2010 and 2018. In ARQMath-1 topics are pulled from question posts in 2019, and in ARQMath-2 topics where pulled from question posts in 2020. Following this pattern, in ARQMath-3 topics are constructed from MSE question posts published in 2021. For evaluation, returned formulas are grouped by appearance using visual identifiers, and repeated visual identifiers are removed using a deduplication step before evaluation [2].

For ARQMath 2021 the XY-PHOC team submitted a run to Task 2. The model did not make use of the question post context in the Task 2 topics, instead using only the appearance of formulas in isolation. Formulas in LaTeX were rendered as Scalable Vector Graphics (SVG) images, which represent the identities and locations of symbols. Formula appearance was represented by binary vectors capturing the locations of symbols within hierarchically partitioned regions in both the X and Y directions (Pyramidal Histogram of Character (PHOC) vectors [3, 4]). Queries were disjunctive, with any document containing a query symbol being scored by the cosine similarity of the query PHOC vector with the candidate's PHOC vector. Despite its simplicity, the XY-PHOC model was competitive with other ARQMath 2021 Task 2 runs, scoring within 0.7% NDCG' 3.8% MAP' and 5.5% P'@10 of the best submission. Also, while many other systems fluctuated in effectiveness from ARQMath 2020 to ARQMath 2021, XY-PHOC had comparatively stable performance. One drawback was that the system was very slow, taking around ten minutes to run queries. With these promising results, we were interested in exploring ways to improve both the efficiency and effectiveness of PHOC models for ARQMath 2022.

For ARQMath 2022, we have used higher spatial resolutions (e.g., using 7 rather than 5 horizontal and vertical regions), and introduced ellipses for concentric regions that capture visual symmetry (e.g., to represent '$x + y$' and '$y + x$' similarly). We also introduced a variation of the model where PHOC matches are scaled by the Inverse Document Frequency (IDF) of symbols in formulas. IDF models did not improve effectiveness as we had hoped, and we will be looking to improve these models at a future time. Similar to ARQMath 2021, despite the absence of context features in our PHOC models, the P$'$@10 measures obtained for our non-IDF PHOC models in ARQMath 2022 Task 2 are again competitive, and now have an OpenSearch implementation that runs many times faster than the original implementation. The new implementation also supports a time/effectiveness tradeoff by constraining the number of query symbol matches required for a candidate to be scored. Using this matching constraint reduces retrieval time to less than 200ms on average for fully conjunct queries, where all query symbols must be present. This conjunct retrieval mode may be useful for autocompletion, where it is reasonable to expect all query symbols to be included in returned formulas.

To gain additional insight into the behavior of our models and Task 2 more generally, we created a tool to comparatively view qrel (relevance score) files and multiple Task 2 runs, showing returned formulas along with their associated relevance ratings. These visualizations can be found in the Appendix of the paper. The code for our models and visualization tool are available as Python libraries online.[1]

[1] Three libraries (`ARQMathCompare`, `AnyPHOC` and `PHOCindexing`) are available from https://gitlab.com/dprl

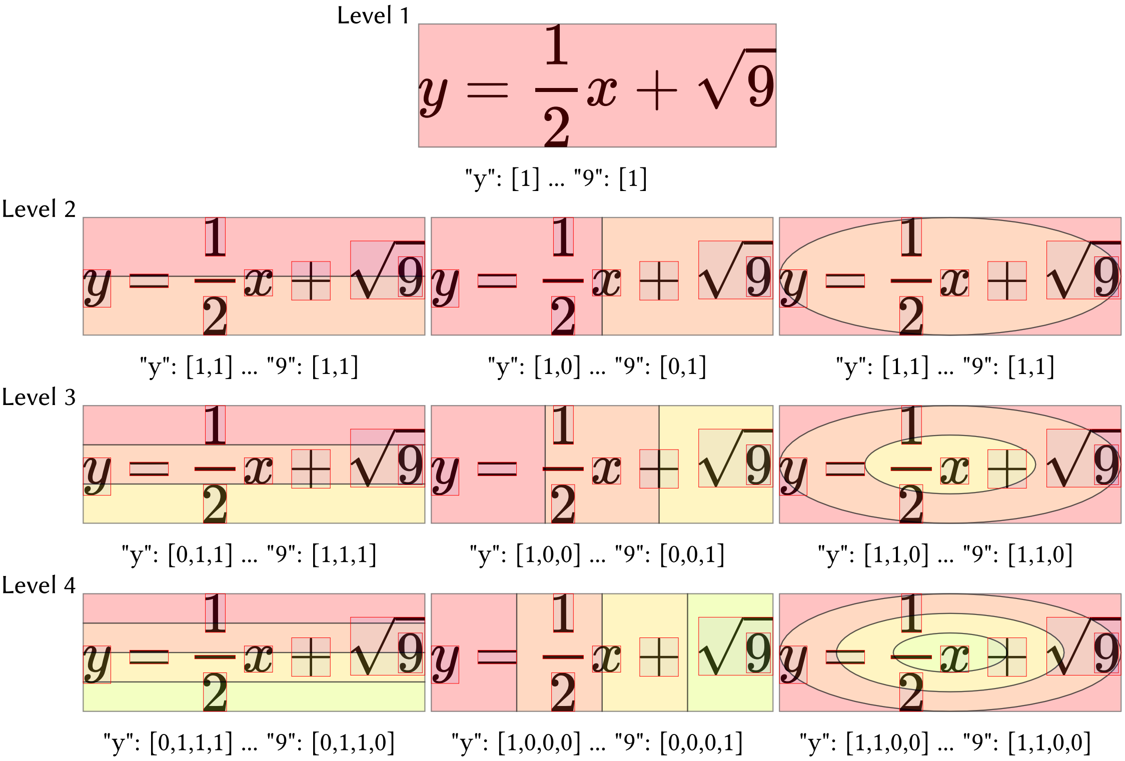

Level 1
"y": [1] ... "9": [1]
Level 2
"y": [1,1] ... "9": [1,1]
"y": [1,0] ... "9": [0,1]
"y": [1,1] ... "9": [1,1]
Level 3
"y": [0,1,1] ... "9": [1,1,1]
"y": [1,0,0] ... "9": [0,0,1]
"y": [1,1,0] ... "9": [1,1,0]
Level 4
"y": [0,1,1,1] ... "9": [0,1,1,0]
"y": [1,0,0,0] ... "9": [0,0,0,1]
"y": [1,1,0,0] ... "9": [1,1,0,0]

Figure 1: XYO4 PHOC Regions. X, Y, and elipsoid (O) regions are organized pyramidally, with equal sized partitions of 2 through 4 regions in Levels 2-4. Level 1 represents the entire expression. For space, the PHOC bit sub-vectors for symbols are shown for the leftmost and rightmost symbols (‘y’ and ‘9’).

2. Related Work

At ARQMath-2021, the XY-PHOC approach differed from other state of the art approaches: it did not represent formula structure using graphs or paths, but rather only the spatial positions of symbols [5]. This not only makes the model simpler to visualize and reason about, but also allows formulas to be represented in a forgiving way, as a slight positional shift of a symbol will not change its spatial region membership. The use of a binary vector also allow for smaller representations, as the 29 bit vectors used in the submission could easily fit into a 32 bit integer. An example of XY-PHOC regions can be seen in the left (Y) and middle (X) columns of Figure 1.

The XY-PHOC model is a variation of traditional vector space retrieval models (i.e., using ‘sparse vector’ representations). This is a well studied area for text, with many well-known variants (e.g., TF-IDF, BM25 [6, 7]). As is common for vector space models, XY-PHOC uses cosine similarity for ranking (inspired by its previous use with PHOC encodings by Sudholt et al. [8]). For XY-PHOC, a rank-equivalent version that takes advantage of the binary vectors used for PHOC was developed. Given a binary query PHOC vector **a**, and a binary candidate formula PHOC vector **b**, the binary vector cosine similarity (*bcos*) is defined as shown in Equation 1:

$$cos(\mathbf{a},\mathbf{b}) \overset{rank}{=} bcos(\mathbf{a},\mathbf{b}) = \frac{1}{\sqrt{|\mathbf{b}|_1}}|\mathbf{a} \wedge \mathbf{b}|_1 \quad (1)$$

$|\cdot|_1$ represents the Hamming weight (i.e., the number of '1' bits). $\wedge$ is logical AND, capturing where symbols are occupying the same region in $\mathbf{a}$ and $\mathbf{b}$. In essence, the vectors are scored by the number of shared PHOC symbol regions, scaled by the length of the candidate vector ($\sqrt{|b|}$). This scaling generally prefers candidates $\mathbf{b}$ with fewer additional symbols not present in $\mathbf{a}$. Note that the candidate scaling factor $\mathbf{b}$ can be computed at index time.

While surprisingly effective, the original XY-PHOC model used only axis-aligned horizontal and vertical regions, with the same number of regions in either direction. However, Avenoso noticed that most formulas tend to extend in the horizontal direction rather than in the vertical direction [5]. Another limitation is in the representation of *commutative* operators. A commutative operation is one where the order in which operands are applied does not impact the result, such as addition. We noted that some formulas containing commutative operations would produce different vectors in the XY-PHOC model. For example consider the following two equations:

$$y = \frac{1}{2}x + \sqrt{9} \quad \text{versus} \quad \sqrt{9} + x\frac{1}{2} = y \quad (2)$$

Despite having the same interpretation, these two equations will produce different XY-PHOC vectors. We wanted to find a way to capture a symbol's distance from the center of the equation, so that commutative operations that are presented differently could be better matched. This led us to using concentric ellipses to capture regions from the outside to the middle of an expression (see the right column of Figure 1).

3. PHOC Region Modifications and Generalization

The first modification we explored was adding more regions types. The motivation for this was to capture formulas with commutative operations, in which operands occur in a different order as described in the last section, but also to be able to work with the shape of the data we are trying to index, whether it be math, chemistry, or plots and figures. We switched our thinking from thinking in terms of X and Y to thinking in terms of regions with different shapes.

The first shape we added was an ellipse. An ellipse helps address the problem of commutative formulae being represented differently when using only horizontal (X) and vertical (Y) splits. Figure 2 illustrates this: despite being on opposite sides of the formula, the y and $\sqrt{}$ symbols produce identical PHOC vectors at all levels. From observing Figure 2 further, it can be seen that only the symbol 9 has a PHOC vector that differs at any level. In previous XY-PHOC work, symbol locations were represented by a horizontal line located at the vertical center of symbol bounding boxes. Since we are now introducing regions with more complex shapes that are not axis-aligned, we are using bounding boxes to represent symbol membership in PHOC regions.

PHOC Configuration Notation. To differentiate PHOC region configurations, we use a notation to represent what region shapes are being used, and the number of levels in each (e.g., 'xy5' or 'xy7o4'). 'x' denotes horizontally-split rectangles, 'y' denotes vertically-split rectangles, and 'o' denotes concentric ellipses. A number defines the number of levels to expand the

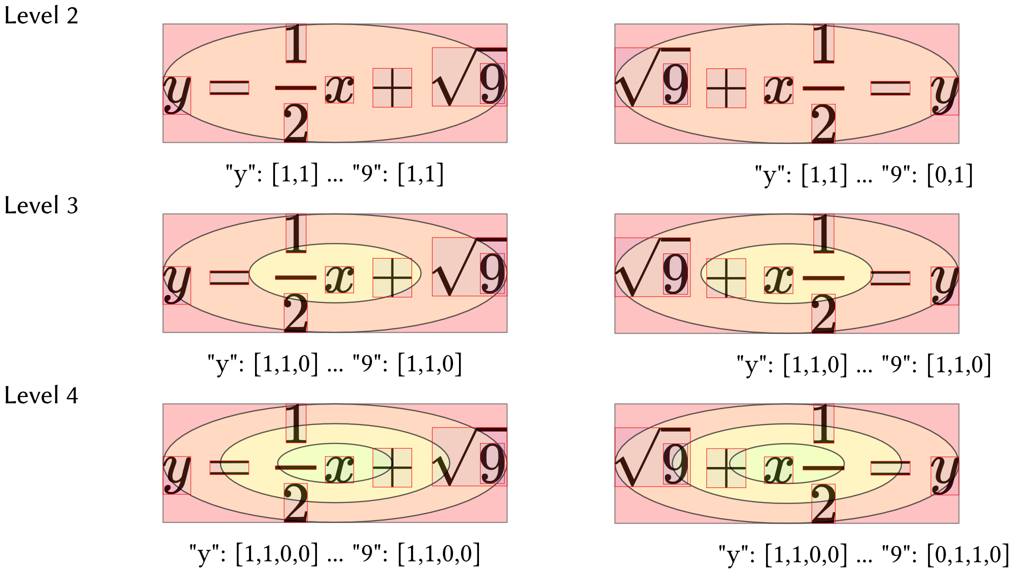

Level 2
"y": [1,1] ... "9": [1,1]
"y": [1,1] ... "9": [0,1]
Level 3
"y": [1,1,0] ... "9": [1,1,0]
"y": [1,1,0] ... "9": [1,1,0]
Level 4
"y": [1,1,0,0] ... "9": [1,1,0,0]
"y": [1,1,0,0] ... "9": [0,1,1,0]

Figure 2: Capturing Semantic Equivalence with Ellipsoidal Regions. Here a formula with commutative operations ('+' and '=') is flipped horizontally, producing nearly equivalent PHOC vectors.

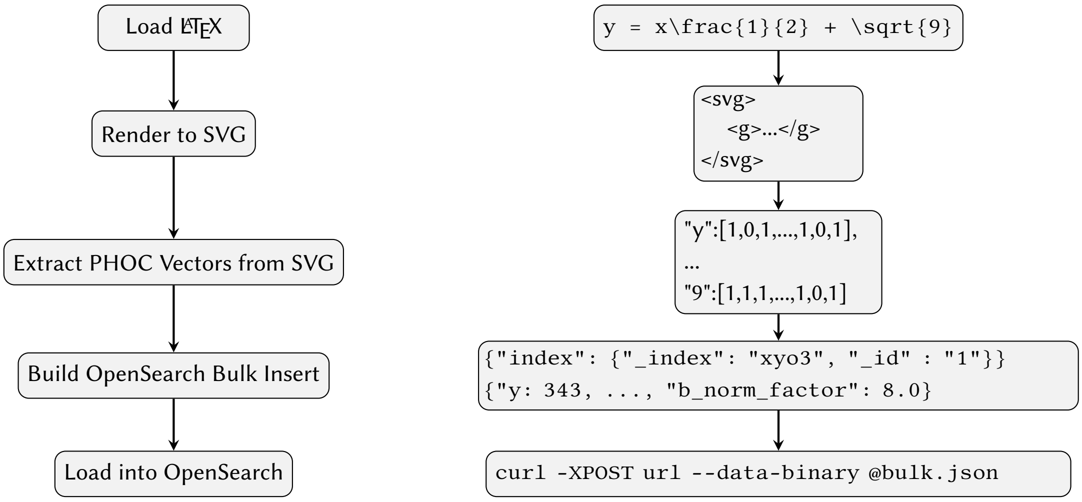
preceding list of region types to. For all region types, at each level $n > 1$, splits are made equally in the corresponding direction(s): n equal-size horizontally or vertically stacked rectangles for 'x' and 'y', and n ellipses that split the formula's width and height at equal intervals along lines running through the centroid of the formula. Figure 1 shows an 'xyo4' PHOC configuration. The top-level (Level 1) contains all symbols. The columns below Level 1 represent Levels 2 through 4 for the vertical rectangle ('y'), horizontal rectangle ('x'), and ellipsis ('o') region types. Shown below each level in Figure 1 are sub-vectors representing the portion of the PHOC associated individual symbols at that level.

4. Scoring with Symbol Inverse Document Frequency

We wanted to incorporate Inverse Document Frequency (IDF) into the model to increase the weight of rare symbols. IDF is implemented as described by Spärck Jones [6], shown in Equation 3. Given an index *I* which contains *N* formulas and has vocabulary *V*, token $k \in V$, and n_k, the number of formulas in which *k* occurs within index I, the Inverse Document Frequency of token k (idf_k) is:

$$idf_k = \log \frac{N}{n_k + 1} \tag{3}$$

Using this formula, we modify the $bcos$ function in Equation 1 as shown in Equation 4 to incorporate IDF weights. This produces a scoring function similar to TF-IDF. Given a query vector **a**, and a candidate formula vector **b** such that $\mathbf{a}_k$ and $\mathbf{b}_k$ denote the portion of each vector corresponding to token k:



```
y = x\frac{1}{2} + \sqrt{9}
```

```
<svg>
   <g>...</g>
</svg>
```

```
"y":[1,0,1,...,1,0,1],
...
"9":[1,1,1,...,1,0,1]
```

```
{"index": {"_index": "xyo3", "_id" : "1"}}
{"y: 343, ..., "b_norm_factor": 8.0}
```

```
curl -XPOST url --data-binary @bulk.json
```

Figure 3: Indexing LaTeX Formulas as PHOC Vectors. At left the indexing pipeline is shown: LaTeX is converted to SVG using MathJax, SVGs are converted to PHOC vectors for individual symbols, which are stored in JSON for ingestion by OpenSearch. At right, we show an example formula moving through the pipeline.

$$bcos_{idf}(\mathbf{a},\mathbf{b}) = \frac{1}{\sqrt{|\mathbf{b}|_1}} \sum_{k \in V} \left(|\mathbf{a}_k \wedge \mathbf{b}_k|_1 \times idf_k\right) \tag{4}$$

Note that we are using a *positional* term frequency for tokens which varies for matching regions in $\mathbf{a}$ and $\mathbf{b}$, but the IDF weight is fixed by token, and does not take positions into account.

5. Implementation: Indexing and Retrieval

We combined available software tools to create a full pipeline capable of running experiments from a single configuration file. These tools were MathJax (a Javascript Library used to render math in the web)[2], PySpark (a library for distributed computing)[3], and OpenSearch (a feature rich search engine that started as a fork of ElasticSearch[4])[5]. To make prototyping easier, we also wrote a library that allows experiments using different PHOC configurations to be defined declaratively, called AnyPHOC. AnyPHOC allows users to instantiate objects to represent different PHOC configurations, which can then be tested against a set of SVG images.

Indexing Pipeline. Figure 3 shows the process of a formula moving through the indexing pipeline. First '`y = x\frac{1}{2} + \sqrt{9}`' is loaded into a PySpark dataframe along with metadata such as the formula id and the associated post id. It is then sent to MathJax which produces an SVG that can be parsed into a set containing bounding boxes for each symbol in the SVG. The composed AnyPHOC object is then tested against the set of bounding boxes, creating

[2] https://www.mathjax.org/
[3] https://spark.apache.org/docs/latest/api/python/
[4] https://www.elastic.co/
[5] https://opensearch.org/

Table 1
PHOC Indices. The Index Type indicates whether individual formulas or visually unique formulas are indexed. Note that the last row provides statistics for the index used for XY-PHOC at ARQMath-2.

PHOC Configuration	Index Type	Formulas	Size on Disk (GB)
xy7o4	Individual	26,827,604	4.2
xy5	Individual	25,324,774	2.9
xy7o4	Visually Unique	8,268,110	2.7
xy5	Visually Unique	8,231,511	1.6
xy5 (XY-PHOC)	Visually Unique	9,326,795	2.3

a map from each symbol present in the SVG to a bitvector. These maps are then converted to the OpenSearch bulk index API file format and joined into a large file. The bitvectors are compressed into a 64-bit (long) or 32-bit (int) integer, depending on the size of the PHOC during the creation of the bulk index file. PySpark parallelizes this process in a map-reduce framework. Finally, a request is made to OpenSearch to index the bulk file, and queries can be made against the newly created index. Since we are using a novel representation for math formulas, we did not use any of the built in Open-Search scoring functions. We instead used the Painless scripting language[6] provided by both ElasticSearch and OpenSearch to implement our scoring function.

Retrieval. When performing retrieval for non-IDF scoring, we simply use the scoring function from the original XY-PHOC experiments (Equation 1). When using the modified IDF scoring (Equation 4), we were unable to get the inverse symbol frequencies from OpenSearch because of how PHOCs were encoded (as fields rather than documents). As a result, we needed to first cache the IDF map by getting the count of documents in the index and then getting the symbol frequency for each symbol that appears in the index. We could then build a map from each symbol to its IDF score. We send IDF values associated with the tokens in the query to OpenSearch when performing a search.

ARQMath 2022. A summary of the indices used for our experiments is shown in Table 1. When initially preparing for ARQMath-3, the full ARQMath collection was indexed. This included many formulas that were visually identical. Given that many formulas appear more than once, we refer to this dataset going forward as the ‘individual’ formula dataset. A total of 25,324,774 documents were included in the ‘xy5’ index, and 26,827,604 documents in the ‘xy7o4’ index. The difference in counts (more than one million for both the individual formula and visually unique formula indices) comes from errors in the SVG rendering and extraction process, which may impact results reported later in the paper.

Regarding the index sizes shown in Table 1, note that the default index compression algorithm used in OpenSearch is LZ4 [7]. For the individual formula index, the ‘xy5’ index is 30% smaller than the ‘xy7o4,’ and for the visually unique index, the ‘xy5’ index is 40% smaller than the ‘xy7o4’. This large difference is seen because ‘xy5’ is stored in an integer (32 bits), while ‘xy7o4’ is stored in a long integer (64 bits).

[6] https://www.elastic.co/guide/en/elasticsearch/painless/current/painless-guide.html
[7] https://github.com/lz4/lz4

Table 2
PHOC Encoding Comparison for ARQMath 2021 Formula Retrieval (Task 2), for Formulas in qrels. PHOCs are defined by a sequence of region types followed by number of PHOC levels (e.g., *xy7o5* indicates 7 regions for the x and y directions, and 5 ellipsoidal (o) regions)

Run	NDCG′	MAP′	P′@10	Vector Length (bits)
xy5	0.6135	0.3089	0.3879	29
xy7	0.6403	0.3369	0.4155	55
xy10	0.6305	0.3262	0.4241	109
x7y5	0.6352	0.331	0.4138	42
xyo5	0.6337	0.3323	0.4293	43
x7yo5	0.6343	0.3305	0.4172	56
xy7o5	**0.6415**	**0.3378**	**0.4328**	69

Using the individual formula indices led to degraded results, as many top results were identical. When de-duplication was performed, as is done for all Task 2 ARQMath submissions, we sometimes have few if any valid hits. To mitigate this, we switched to indexing only visually unique formulas (as done for XY-PHOC at ARQMath 2021), and this greatly improved results. This visually unique index contained 8,231,511 formulas that were successfully parsed, meaning that roughly 17 million formulas in the full ARQMath collection are duplicates.

Unfortunately, we did not make this switch until after the submission deadline for ARQMath 2022. In the next Section, we present a summary of both official runs using the individual formula index along with updated results using the visually unique index computed after submitting official runs.

6. Results

Here we report the results from preliminary experiments using Task 2 topics from ARQMath 2020 and 2021, along with results for official runs submitted to ARQMath 2022 and additional experiments performed after the ARQMath 2022 qrels became available.

Preliminary PHOC Configuration Experiment. For preliminary development, we generated a small dataset made up of only formulas with relevance scores included in the 2021 ARQMath Task 2 qrels. We took this approach to greatly reduce indexing time, and increase the number of PHOC configurations that we could quickly compare.

In our first experiment, we compared PHOC configurations with 5, 7, and 10 levels in the x and y directions, configurations with ellipse regions (with 5 levels to start), and configurations with more horizontal than vertical and ellipse regions (e.g., ‘x7y5’ and ‘x7yo5’). The largest configuration attempted was ‘xy10’, requiring 109 bits per vector. Avenoso showed previously that adding more levels in XY-PHOC can increase the robustness of the model[5], and so we wanted to see what happens when a large number of regions are employed.

Results for these experiments are shown in Table 2. Of the configurations tested, ‘xy7o5’ (69 bits long) had the most promising results in all metrics. The next most promising in terms of NDCG′ and MAP′ was ‘xy7’ while ‘xyo5’ was next most promising in terms of P′@10. As both of these are sub-vectors of ‘xy7o5’, we decided to use ‘xy7o5’ as our primary model for

Table 3
ARQMath-2020 and -2021 PHOC Results: Individual Formula Index vs. Visually Unique Index. Within each year, runs are sorted by nDCG$'$ for the Visually Unique index.

	ARQMath-2020					
	Individual Formulas			Visually Unique		
Run	NDCG$'$	MAP$'$	P$'$@10	NDCG$'$	MAP$'$	P$'$@10
xy7o4	**0.492**	**0.316**	**0.433**	**0.550**	**0.369**	**0.438**
xy5	0.419	0.263	0.403	0.547	0.319	0.438
xy5-IDF	0.379	0.241	0.374	0.492	0.317	0.404
xy7o4-IDF	—	—	—	0.489	0.318	0.411
	ARQMath-2021					
xy7o4	**0.448**	**0.250**	**0.435**	**0.505**	**0.284**	**0.428**
xy5	0.328	0.168	0.391	0.488	0.267	0.412
xy7o4-IDF	—	—	—	0.462	0.250	0.414
xy5-IDF	0.317	0.156	0.391	0.452	0.237	0.402

ARQMath-3. We then decided to reduce ‘xy7o5’ to ‘xy7o4’, which is exactly 64 bits long, so that it would fit into a long integer in our OpenSearch index.

Runs. We considered 4 PHOC configurations for runs this year:

1. xy7o4
2. xy5 (the configuration used for XY-PHOC at ARQMath-2)
3. xy5-IDF (adding IDF scaling)
4. xy7o4-IDF (adding IDF scaling)

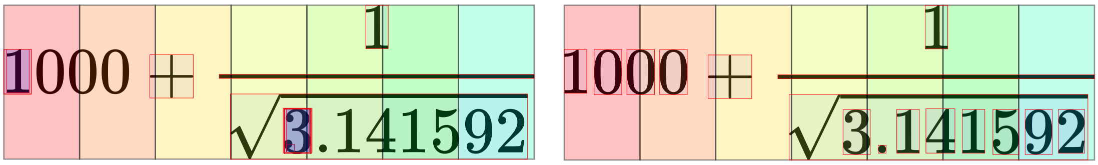
The first three configurations using ARQMath 2020 and 2021, and submitted for ARQMath-3; the fourth configuration (‘xy7o4-IDF’) was not submitted for ARQMath-3 due to a lack of time.

In the remainder of this Section, we report effectiveness measures for the ARQMath-1 and -2 collections (2020 and 2021), followed by a summary of results for ARQMath-3 (2022). We then provide information regarding the efficiency of our new PHOC implementation, which has reduced retrieval times dramatically when compared to the XY-PHOC system from ARQMath-2.

6.1. Effectiveness Measures for ARQMath-1 (2020) and ARQMath-2 (2021)

Table 3 provides effectiveness measures for our four configurations on both ARQMath 2020 and ARQMath 2021. Due to time constraints, we were unable to run ‘xy7o4-IDF‘ on the individual formula collections. The ‘Individual Formulas’ results correspond to our submitted runs, also shown later in Table 6. The P$'$@10 metrics are similar between the individual and visually distinct formula indices, as after the visually identical formula are removed, the remaining formulas at top ranks are often very similar. However the NDCG$'$ and MAP$'$ decrease when using the individual formula index, because of the fewer evaluated hits that remain after deduplication.

Within index types, the rankings of the PHOC models are stable across the 2020 and 2021 topics. There are some swaps in position for the IDF models, but the differences in effectiveness measures are quite small. Across all indices and collections, the P$'$@10 values fall within a



Old: Bounding Boxes Collapsed for Numbers **New:** Original Bounding Boxes

Figure 4: Effect of Collapsing Numbers on Y dimension PHOC Representation (Level 7 shown). Collapsing digits and decimal places (at left) allows digits and decimal points in a number to be matched at one position, which may be beneficial.

narrow range of 37.4-44%, and are within 0.2% for 'xy7o4' using the individual formula indices, and by just 1% for the visually unique formula indices. More variation is seen in the measures for complete rankings, with nDCG$'$ varying between 32-55%, and MAP$'$ varying between 16-37%.

Consistent with our preliminary experiment, for both ARQMath 2020 and 2021, and on both index types, the 'xy7o4' PHOC configuration performed better than 'xy5.' We suspect that the additional region type (ellipse) and higher spatial resolution in 'xy7o4' accounts for this difference, providing more spatial information and increasing precision as a result.

A surprising result was that for both 'xy5' and 'xy7o4', their accompanying IDF variation performed slightly worse. We believe that this may be due to a bug; but another possibility is that rarer tokens from queries are matched to rare tokens in a document in the lower levels, before the regions become smaller and more granular. Position is very important in the PHOC model, so if rare tokens exist in a query and a posting but are not close one another, the IDF weight applied at lower levels could provide an undesirable boost in scoring. It may also be helpful to weight matches in smaller regions for PHOC vectors themselves, as matching bits at higher levels indicate that symbols are spatially closer. Using IDF with PHOC, it may make more sense to use *individual* IDF weights defined for each region at each level, with higher weights for levels smaller spatial regions.

Differences Between XY-PHOC and 'xy5'. It is also worth noting that we did not match the ARQMath-2 XY-PHOC results using our 'xy5' PHOC model, which we had initially expected to be identical. After some investigation, we found that some queries retrieve identical formulas that are scored differently by XY-PHOC and the 'xy5' implementations. This led to identifying three differences in the models and their implementation: two are related to differences in how spatial information is represented, and the third to indexing differences.

First, while rebuilding the indexing module to incorporate the new AnyPHOC library and OpenSearch, what we thought was a bug in the SVG parser code was 'fixed,' but we now believe this may have been a feature in disguise (see Figure 4). Certain transformations were not being applied, causing symbol bounding boxes within tokens like numbers and function names to share the same starting position. This may actually be a feature, for example allowing the number π (3.141592. . .) to be matched using differing numbers of digits at one position. It would be interesting to see if instead of collapsing numbers as shown in Figure 4, each digit's bounding box was expanded to the bounding box for the complete number. For example in Figure 4 the digit '3' would use or add a bounding box encompassing all of '3.141592,' as would every subsequent digit within the number.

Table 4
Strong Retrieval Result for ‘xy7o4’ on ARQMath-2021 (Visually Unique Index)

	Q.202	
Rank	$[E : F] < \infty;$	Relevance
1	$[E : F] < \infty$	3
2	$[F : E] < \infty$	3
3	$[\overline{F} : F] < \infty$	3
4	$[E : F] = n < \infty$	3
5	$[E : F] = 2$	2

The second difference is that ‘xy5’ changes the representation of symbol locations to bounding boxes from the horizontal lines at the vertical center of symbols used by XY-PHOC. We intend to conduct future experiments to check the effect of how symbol locations are represented: we wish to go back to using just a line, as well as represent symbols at a single point by their bounding box centroids.

A third difference is the number of formulas indexed; as seen in Table 1, there are over a million more formulas that were indexed for the XY-PHOC model. Correcting a small number of frequent LaTeX parsing errors may close this gap.

Qualitative Results. We saw some confirmation of our hypothesis regarding ellipse regions being able to capture inverted formulas through spatial symmetry. As an example view Q.251 from the ARQMath 2021 topic file in Equation 5:

$$(k+1)^{\frac{1}{k+1}} < k^{\frac{1}{k}} \tag{5}$$

This query has multiple formula evaluated as highly relevant that are inverted. For example:

$$k^{\frac{1}{k}} > (k+1)^{\frac{1}{k+1}} \quad \text{(visual id: 1127482)} \tag{6}$$

$$n^{\frac{1}{n}} > (n+1)^{\frac{1}{n+1}}. \quad \text{(visual id: 9090717)} \tag{7}$$

‘xy7o4’ was able to retrieve Equation 6 at rank 25. Although the ellipses are having some effect, the larger number of horizontal and vertical regions may be limiting their influence.

Similar to the original XY-PHOC system, the new models perform well when relevant formulas have most of the symbols from the query, and symbol placements are not shifted far from their locations in the query. As an example, Table 4 shows the top 5 hits for topic Q.202 from the visually unique ‘xy7o4’ results. All returned hits have symbols from the query that are shifted only slightly. This feature is also the model’s weakness. Consider Table 5, which shows the top 5 hits for topic Q.229 using the same PHOC model. Many of the query symbols are present, and in some hits they are only shifted slightly or duplicated. However, a duplication can cause a formula to take on a very different meaning, as seen in the 2nd and 5th ranked hits in Table 5, where an additional integral is introduced.

Table 5
Weak Retrieval Result for ‘xy7o4’ on ARQMath-2021 (Visually Unique Index)

Rank	Q.229 $\int_0^1 \left\{\frac{1}{x}\right\}\left\{\frac{1}{1-x}\right\}\left\{1-\frac{1}{x}\right\}dx$	Relevance
1	$\int_0^1\{\frac{1}{x}\}dx$	1
2	$\int_0^1\int_0^1\{\frac{1}{x}\}\{\frac{1}{x\,y}\}dx\,dy$	1
3	$\int_0^1 \frac{1}{1-x}\,dx$	0
4	$\int_0^1 \left\{\frac{1}{x^{\frac{1}{6}}}\right\}\,dx$	1
5	$\int_0^1\int_0^1\left\{\frac{1}{x}\right\}\left\{\frac{1}{y}\right\}\frac{(1-x)(1-y)}{1-xy}dxdy$	1

Table 6
Submitted Runs for ARQMath-3 (results shown for 2020, 2021 and 2022 topics). The ARQMath-2 (2021) XY-PHOC run is provided for comparison. Systems ranked by nDCG′ for 2022 topics. *: manual run

		Effectiveness Metrics								
			2020			2021			2022	
Team	Run	NDCG′	MAP′	P′@10	NDCG′	MAP′	P′@10	NDCG′	MAP′	P′@10
approach0*	fusion_alph05	**0.647**	**0.507**	**0.529**	**0.652**	**0.471**	**0.612**	**0.720**	**0.568**	**0.688**
DPRL	TangentCFT2ED	0.648	0.480	0.502	0.569	0.368	0.541	0.694	0.480	0.611
MathDowsers	L8	0.646	0.454	0.509	0.617	0.409	0.510	0.633	0.445	0.549
Baseline	Tangent-S	0.691	0.446	0.453	0.492	0.272	0.419	0.540	0.336	0.511
XY-PHOC-DPRL	XY-PHOC	0.611	0.423	0.478	0.548	0.323	0.433	——	——	——
XY-PHOC-DPRL	xy7o4	0.492	0.316	0.433	0.448	0.250	0.435	0.472	0.309	0.563
XY-PHOC-DPRL	xy5-IDF	0.379	0.241	0.374	0.317	0.156	0.391	0.376	0.180	0.461
XY-PHOC-DPRL	xy5	0.419	0.263	0.403	0.328	0.168	0.391	0.369	0.211	0.518

6.2. ARQMath-3 (2022)

Results for our three submitted ARQMath-3 runs, along with the top runs from participating teams and the ARQMath-2 XY-PHOC run are shown in Table 6. These runs were submitted using the individual formula indices, which result in a reduction in retrieval effectiveness for nDCG′ and MAP′ as seen in Table 3 and discussed in the previous section. When we first realized this, we set the number of documents (formulas) retrieved to 30,000. Going higher often caused OpenSearch to raise an error and stop. After this large number of hits was returned, a de-duplication for visual formula identifiers was run, so that no more than 1000 hits were selected for each topic in the submitted runs.

For most other submissions NDCG′ is the highest scoring metric (including XY-PHOC, which indexed visually distinct formulas), but for most of our runs it is P′@10. While we could not match the XY-PHOC results using ‘xy5’, we were able to match P′@10 results for ARQMath 2021 using ‘xy7o4’. Also, for ARQMath-3 (2022), the P′@10 measures for ‘xy7o4’ are the second-highest for an automatic system after TangentCFT2ED, which is an n-gram embedding model applied to tree edges from Symbol Layout Trees (SLTs) and Operator Trees (OPTs), followed by reranking using a weighted tree edit distance.

Notes on Result Diversity in Formula Search. Related to indexing individual formulas,

Table 7
Example: A Candidate Grouping of Visually Similar but Non-Identical Formulae into an Information Nugget [10]. Subtle differences: lower and upper-case n, ellipsis dots that are centered or on the baseline.

Visual Id	Formula	LaTeX
469553	$A_1 \times \cdots \times A_n$	`A_{1} \times \cdots \times A_{n}`
22725	$A_1 \times ... \times A_n$	`A_1\times...\times A_n`
9014841	$A_1 \times ... \times A_N,$	`A_1 \times ... \times A_N,`

in a real-world setting you often cannot simply discard documents (formulas) because they are identical to other formulas in the collection, as a query may include text and/or additional formulas with matches that need to be merged. However, presenting a long list of identical results for a query is also unhelpful. To address this situation, we turn to research on search result diversity.

Clarke et al. devised an evaluation measure, where if a specific information nugget appears in the first $k - 1$ retrieved documents, then a repetition of that information nugget in the k^{th} returned document has less weight in the evaluation measure, as it will provide no additional benefit [9]. Information Nuggets can be described as atomic units of relevant information [10]. We could apply this measure as a way to select results shown to the user, to increase the diversity of results. Within the ARQMath Task 2 a nugget could be modeled by individual visual ids, or groupings of visual ids that are similar but contain different symbols, such as shown in Table 7. These formulae all look similar, and likely represent the same cartesian product in many contexts, but have distinct LaTeX strings and visual identifiers.

Another option would be to have results nested under visually distinct formulas within Search Engine Results Pages (SERP), for example as done previously for the original Tangent formula search engine (see Figure 8 of Zanibbi and Orakwue [11]). This would allow us to show only formulas associated with unique visual ids, so that diverse results can easily be scanned. To save space, clicking on a visually distinct formula could drop down a list of documents associated with the formula, rather than list them inline as done in the Tangent prototype. Building on the idea of grouping visual ids into nuggets, we could have each drop down be associated with a nugget, to allow for even more diverse results; again as an example, a similar thing was done for Tangent-3, where formulas that unify the same symbols in the query were grouped [12]. A tie-breaking algorithm could be applied to rank results within each of these drop down groupings.

Visual Comparison of qrels and Top-5 Results. We compared the top 5 results between our systems and the best runs submitted by each team using a tool we created, `ARQMathCompare`.[8] The tool takes Task 2 submitted runs as input, a qrel file, a topics file, and the ARQMath LaTeX formula index files, and then renders the results in a web page to make it is easy to visually scan and compare results. Approach0's L8 run and the DPRL's TangentCFT2ED run had the best P′@10 scores for primary runs for the ARQMath 2021 and 2022 topics, with Approach0's L8 being the best for the 2020 topics and DPRL's TangentCFT2ED being a close third, so we used these two systems in our visual comparisons.

[8] https://gitlab.com/dprl/arqmathcompare

Table 8

Efficiency and Effectiveness Trade-off for Minimum Symbol Match Threshold ('xy5' model, ARQMath-2 Test Topics, visually distinct formula index)

Match	MRT	Effectiveness Metrics		
Threshold	(secs)	NDCG′	MAP′	P′@10
0%	10.84	**0.4877**	**0.2665**	0.4121
25%	8.30	**0.4877**	**0.2665**	0.4121
50%	3.20	0.4831	**0.2665**	0.4121
75%	1.25	0.4770	0.2610	**0.4263**
85%	0.90	0.4107	0.2171	0.3877
95%	0.58	0.3552	0.1792	0.3404
100%	**0.18**	0.2572	0.1253	0.3024

First, in Appendix A we show an example of where our system outperformed top systems in Figure 5, for topic B.360. The non-relevant hits returned by TangentCFT2ED and L8 both contain symbols that are not present in the query, for example "log" (parsed as "l", "o", "g" by the XY-PHOC systems), "σ", "2", "(" and ")". Given that our systems only score symbols present in both the query and a candidate and penalize additional symbols in a candidate, it becomes less likely that formulas with additional symbols will have a high rank.

Next, Figure 6 shows results for topic B.363, where the inability to match symbols not in the query negatively affects PHOC results. From the top results from the L8 system and the qrel file, it can be seen that most relevant answers contain $\sup_{t\geq 0} M_t$. None of these symbols appear in the query; so in order for a candidate document with these symbols to score higher with our system, the candidate would need to contain most of the symbols in the query.

Even with these differences in systems, there are queries that all systems struggled with, such as for topic B.345 shown in Figure 7. All systems return mostly piece-wise function definitions rated as having low or non-relevance. None do a good job matching the functions with those of relevant documents. This shows that with some queries, additional context and/or mathematical knowledge will be needed to obtain strong retrieval results.

6.3. Exploring an Efficiency and Effectiveness Trade-Off

An important part of this work was accelerating the previous XY-PHOC model, which took 9.5 *minutes* on average per query [3]. We measure model efficiency using Mean Retrieval Time (MRT) in seconds. Retrieval run against the individual formula index is much slower, as more results are returned from the larger collection. All experiments run on the individual formula dataset were fully disjunctive, and had a Mean Response Time of ~110 seconds, which was much faster than XY-PHOC, but still much slower than we wished.

In this Section, we report additional experiments we carried out using our visually unique formula indices. After submitting our runs, we learned that OpenSearch supports adding a threshold percentage of query tokens (for PHOC, symbols) to be present in order for a document to be scored. A token match threshold of 0% yields a fully disjunctive search, while a threshold of 100% produces a fully conjunctive query. We ran multiple experiments in which we tested different percentages against the 'xy5' index built from visually unique formulas, and recorded both MRT and the ARQMath effectiveness metrics. The topic file used was that of Task 2 ARQMath 2021.

Results can be seen in Table 8. Note that setting the symbol matching threshold to 0% yields the same results seen in Table 6 for 'xy5': both use the same representation and disjunct retrieval, but use different indices (the submitted run uses an individual formula index, while Table 8 is for a visually distinct formula index). Table 8 shows that as the percentage of symbols that must match is increased, the MRT decreases as less candidates are considered for scoring. Surprisingly, effectiveness does not suffer until the threshold percentage reaches ~85%. Consider the rows for 0% and 75% as an example. Despite the MRT dropping from 10.84 to 1.25, an 88.46% decrease, metrics only degrade very slightly for NDCG′ and MAP′, while P′@10 actually improves. Going all the way down to the fully conjunctive (100%) retrieval, we see a more substantial decrease in effectiveness, but the MRT is only 0.18 seconds, which is a period of time that is close to what an end user may perceive as instantaneous [13], and with a P′@10 of 30%.

When actually deploying the models to a live application, the intended use case needs to be considered. Having a user wait 10+ seconds with the fully disjunctive query is likely not an acceptable user experience, however waiting 1.25 seconds for a query may be good enough. There are also use cases such as autocomplete, which was previously explored by Avenoso [5]. Autocomplete would likely benefit from a mostly conjunctive (95%) or a fully conjunctive search, as you would not want to have a recommended completion that does not contain most of the symbols already entered. While waiting 1.25 seconds for an ad-hoc query result may be acceptable, many users may hope to finish entering a query before 1.25 seconds pass. Fortunately, the 0.18 seconds obtained for the fully conjunct model should be fast enough to return autocompletion suggestions in real time.

We believe we can do better still in terms of reducing Mean Response Time by incorporating other strategies that have been used to accelerate text retrieval such as MaxScore [14] and Block-Max [15].

7. Conclusion

We have expanded the original bidirectional XY-PHOC model and implementation by adding more region shapes, and increasing retrieval speed substantially. We believe that there are additional opportunities to improve retrieval speed, for example using MaxScore [14] or Block-Max [15] strategies. The new implementation uses a modern search engine (OpenSearch), and the new AnyPHOC library will allow different spatial region configurations to be created more rapidly. Experiments we plan to undergo in the near term include using more ellipses than vertical or horizontal splits in order to place a stronger emphasis on capturing symmetric layouts, and making use of the line representation used in the original XY-PHOC experiments as it appears to be a more effective representation.

While we have been unable to quite match the results from XY-PHOC with our 'xy5' model, we feel that our results are promising. For the new implementation, the 'xy7o4' model performed better than the 'xy5' model we ran this year. Once we are able to replicate the results from last year, it is reasonable to expect that the 'xy7o4' model will perform even better. We also believe that weighting regions more at higher levels (i.e., smaller regions) and incorporating positional IDF could help us assign weight to rarer symbols and symbol locations more effectively. Lastly, we plan to incorporate result diversity techniques to formula retrieval; we are particularly interested in applying nugget-based analysis to group visual ids, so that end users are not flooded with overly similar formulas in search results.

Acknowledgments

We thank Robin Avenoso, whose original XY-PHOC code for ARQMath-2 we extended. This material is based upon work supported by the Alfred P. Sloan Foundation under Grant No. G-2017-9827, and the National Science Foundation (USA) under Grant No. IIS-1717997.

References

[1] B. Mansouri, A. Agarwal, D. W. Oard, R. Zanibbi, Advancing Math-Aware Search: The ARQMath-3 Lab at CLEF 2022, in: European Conference on Information Retrieval, Springer, 2022.

[2] B. Mansouri, V. Novotný, A. Agarwal, D. W. Oard, R. Zanibbi, Overview of ARQMath-3 (2022): Third CLEF lab on Answer Retrieval for Questions on Math (Working Notes Version), in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, 2022.

[3] R. Avenoso, B. Mansouri, R. Zanibbi, XY-PHOC Symbol Location Embeddings for Math Formula Retrieval and Autocompletion, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, volume 2936 of *CEUR Workshop Proceedings*, CEUR, Bucharest, Romania, 2021, pp. 25–35. URL: http://ceur-ws.org/Vol-2936/#paper-02, iSSN: 1613-0073.

[4] B. Mansouri, A. Agarwal, D. W. Oard, R. Zanibbi, Advancing Math-Aware Search: The ARQMath-2 Lab at CLEF 2021 | SpringerLink, 2021. URL: https://link.springer.com/chapter/10.1007/978-3-030-72240-1_74.

[5] R. Avenoso, Spatial vs. Graph-Based Formula Retrieval, Theses (2021). URL: https://scholarworks.rit.edu/theses/10784.

[6] K. Sparck Jones, A Statistical Interpretation Of Term Specificity And Its Application In Retrieval, Journal of Documentation 28 (1972) 11–21. URL: https://doi.org/10.1108/eb026526. doi:10.1108/eb026526, publisher: MCB UP Ltd.

[7] S. E. Robertson, S. Walker, Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval, in: B. W. Croft, C. J. van Rijsbergen (Eds.), SIGIR ’94, Springer, London, 1994, pp. 232–241. doi:10.1007/978-1-4471-2099-5_24.

[8] S. Sudholt, G. A. Fink, Evaluating word string embeddings and loss functions for cnn-based word spotting, in: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), volume 01, 2017, pp. 493–498. doi:10.1109/ICDAR.2017.87.

[9] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, I. MacKinnon, Novelty and diversity in information retrieval evaluation, in: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information

retrieval - SIGIR '08, ACM Press, Singapore, Singapore, 2008, p. 659. URL: http://portal.acm.org/citation.cfm?doid=1390334.1390446. doi:10.1145/1390334.1390446.

[10] V. Pavlu, S. Rajput, P. B. Golbus, J. A. Aslam, IR system evaluation using nugget-based test collections, in: Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12, ACM Press, Seattle, Washington, USA, 2012, p. 393. URL: http://dl.acm.org/citation.cfm?doid=2124295.2124343. doi:10.1145/2124295.2124343.

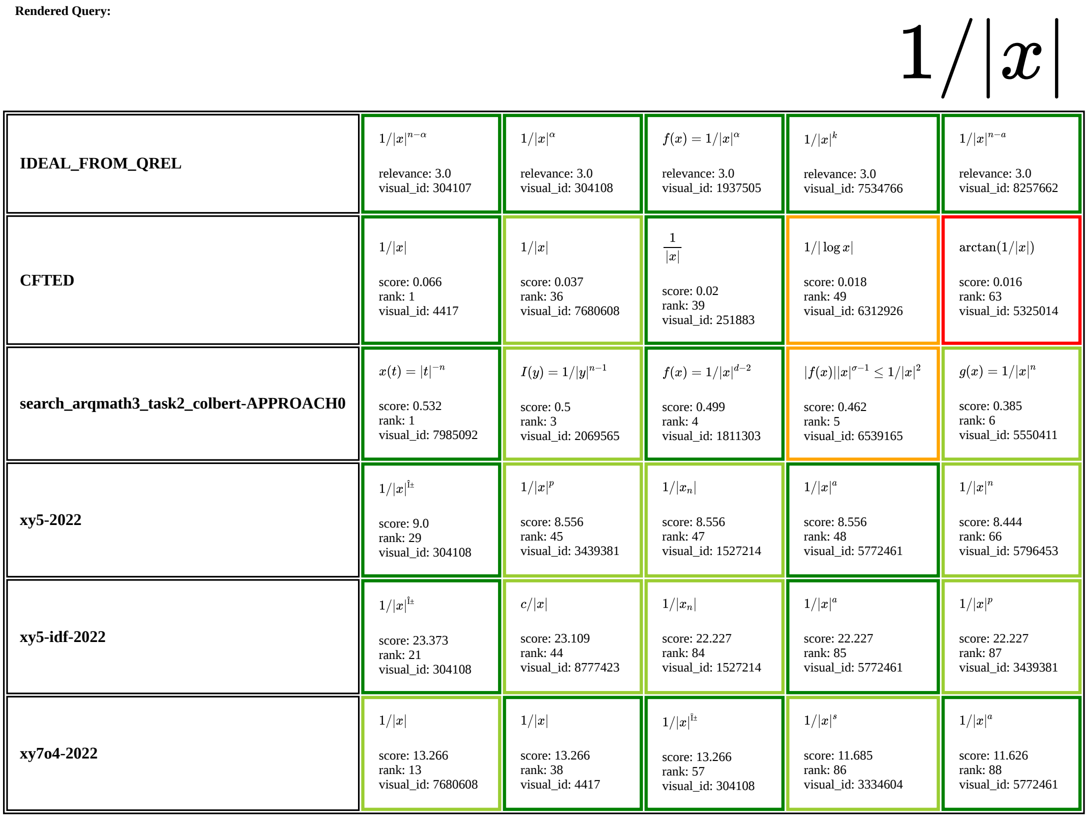
[11] R. Zanibbi, A. Orakwue, Math search for the masses: Multimodal search interfaces and appearance-based retrieval, in: M. Kerber, J. Carette, C. Kaliszyk, F. Rabe, V. Sorge (Eds.), Intelligent Computer Mathematics - International Conference, CICM 2015, Washington, DC, USA, July 13-17, 2015, Proceedings, volume 9150 of *Lecture Notes in Computer Science*, Springer, 2015, pp. 18–36. URL: https://doi.org/10.1007/978-3-319-20615-8_2. doi:10.1007/978-3-319-20615-8_2.

[12] K. Davila, Symbolic and Visual Retrieval of Mathematical Notation using Formula Graph Symbol Pair Matching and Structural Alignment, Ph.D. thesis, 2017.

[13] J. Serviere, D. Miceli, Y. Galifret, A psychophysical study of the visual perception of "instantaneous" and "durable", Vision Research 17 (1977) 57–63. URL: https://www.sciencedirect.com/science/article/pii/0042698977902012. doi:10.1016/0042-6989(77)90201-2.

[14] H. Turtle, J. Flood, Query evaluation: strategies and optimizations, Information Processing and Management: an International Journal 31 (1995) 831–850. URL: https://doi.org/10.1016/0306-4573(95)00020-H. doi:10.1016/0306-4573(95)00020-H.

[15] S. Ding, T. Suel, Faster top-k document retrieval using block-max indexes, in: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, SIGIR '11, Association for Computing Machinery, New York, NY, USA, 2011, pp. 993–1002. URL: https://doi.org/10.1145/2009916.2010048. doi:10.1145/2009916.2010048.



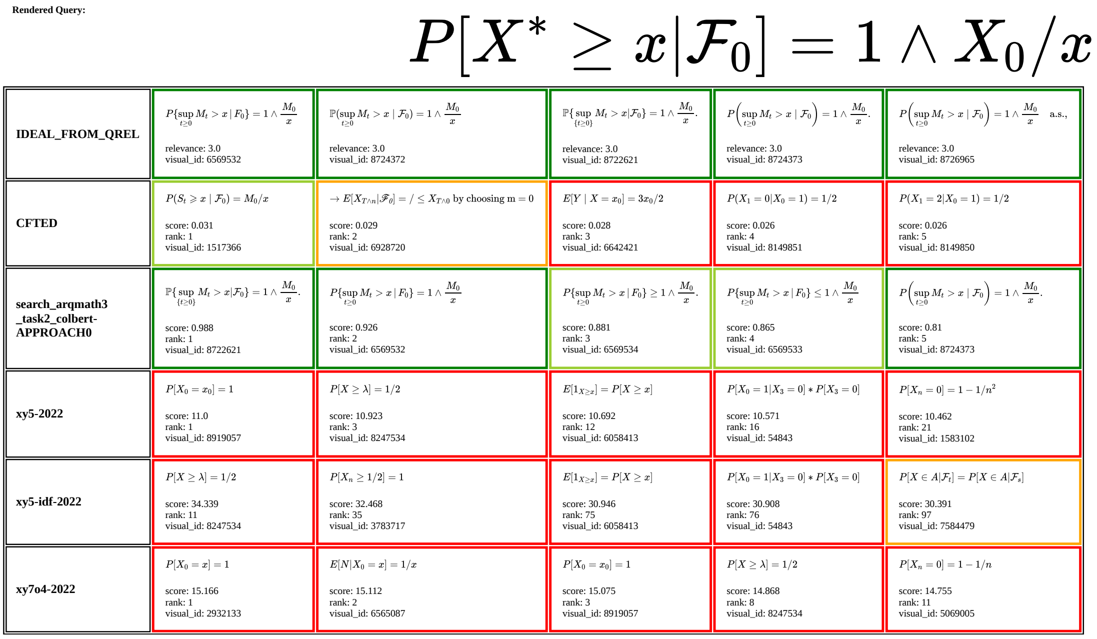
Rendered Query: $1/|x|$

IDEAL_FROM_QREL	$1/\lvert x\rvert^{n-\alpha}$ relevance: 3.0 visual_id: 304107	$1/\lvert x\rvert^{\alpha}$ relevance: 3.0 visual_id: 304108	$f(x) = 1/\lvert x\rvert^{\alpha}$ relevance: 3.0 visual_id: 1937505	$1/\lvert x\rvert^{k}$ relevance: 3.0 visual_id: 7534766	$1/\lvert x\rvert^{n-a}$ relevance: 3.0 visual_id: 8257662
CFTED	$1/\lvert x\rvert$ score: 0.066 rank: 1 visual_id: 4417	$1/\lvert x\rvert$ score: 0.037 rank: 36 visual_id: 7680608	$\frac{1}{\lvert x\rvert}$ score: 0.02 rank: 39 visual_id: 251883	$1/\lvert\log x\rvert$ score: 0.018 rank: 49 visual_id: 6312926	$\arctan(1/\lvert x\rvert)$ score: 0.016 rank: 63 visual_id: 5325014
search_arqmath3_task2_colbert-APPROACH0	$x(t) = \lvert t\rvert^{-n}$ score: 0.532 rank: 1 visual_id: 7985092	$I(y) = 1/\lvert y\rvert^{n-1}$ score: 0.5 rank: 3 visual_id: 2069565	$f(x) = 1/\lvert x\rvert^{d-2}$ score: 0.499 rank: 4 visual_id: 1811303	$\lvert f(x)\rvert\lvert x\rvert^{\sigma-1} \leq 1/\lvert x\rvert^{2}$ score: 0.462 rank: 5 visual_id: 6539165	$g(x) = 1/\lvert x\rvert^{n}$ score: 0.385 rank: 6 visual_id: 5550411
xy5-2022	$1/\lvert x\rvert^{\text{Î±}}$ score: 9.0 rank: 29 visual_id: 304108	$1/\lvert x\rvert^{p}$ score: 8.556 rank: 45 visual_id: 3439381	$1/\lvert x_n\rvert$ score: 8.556 rank: 47 visual_id: 1527214	$1/\lvert x\rvert^{a}$ score: 8.556 rank: 48 visual_id: 5772461	$1/\lvert x\rvert^{n}$ score: 8.444 rank: 66 visual_id: 5796453
xy5-idf-2022	$1/\lvert x\rvert^{\text{Î±}}$ score: 23.373 rank: 21 visual_id: 304108	$c/\lvert x\rvert$ score: 23.109 rank: 44 visual_id: 8777423	$1/\lvert x_n\rvert$ score: 22.227 rank: 84 visual_id: 1527214	$1/\lvert x\rvert^{a}$ score: 22.227 rank: 85 visual_id: 5772461	$1/\lvert x\rvert^{p}$ score: 22.227 rank: 87 visual_id: 3439381
xy7o4-2022	$1/\lvert x\rvert$ score: 13.266 rank: 13 visual_id: 7680608	$1/\lvert x\rvert$ score: 13.266 rank: 38 visual_id: 4417	$1/\lvert x\rvert^{\text{Î±}}$ score: 13.266 rank: 57 visual_id: 304108	$1/\lvert x\rvert^{s}$ score: 11.685 rank: 86 visual_id: 3334604	$1/\lvert x\rvert^{a}$ score: 11.626 rank: 88 visual_id: 5772461

Figure 5: Top-5 Hits for Task-2 Runs, Query B.360: Favorable Results For PHOC Runs. Here we see preferring symbols present in the query benefits the PHOC runs. Box colors represent high relevance (dark green), medium relevance (light green), low relevance (yellow), and non-relevance (red).

A. `ARQMathCompare` Formula Retrieval Result Visualizations

We include PDF screenshots of the HTML pages that we used to compare Task 2 runs for ARQMath-3 in Figures 5 to 7 below. A description may be found in Section 6.2.

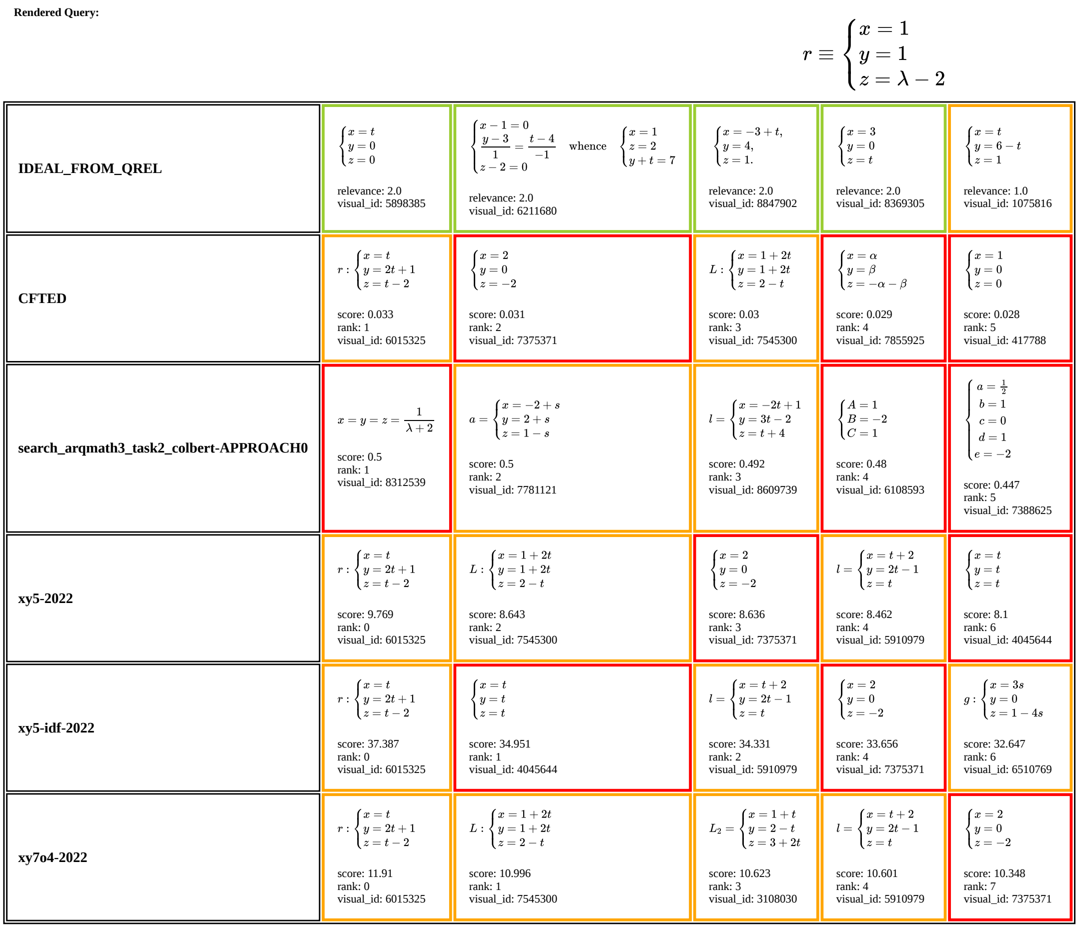


Rendered Query:

$$P[X^* \geq x \vert \mathcal{F}_0] = 1 \wedge X_0/x$$

IDEAL_FROM_QREL	$P\{\sup_{t\geq 0} M_t > x \mid F_0\} = 1 \wedge \frac{M_0}{x}$ relevance: 3.0 visual_id: 6569532	$\mathbb{P}(\sup_{t\geq 0} M_t > x \mid \mathcal{F}_0) = 1 \wedge \frac{M_0}{x}$ relevance: 3.0 visual_id: 8724372	$\mathbb{P}\{\sup_{\{t\geq 0\}} M_t > x \vert \mathcal{F}_0\} = 1 \wedge \frac{M_0}{x}.$ relevance: 3.0 visual_id: 8722621	$P\left(\sup_{t\geq 0} M_t > x \mid \mathcal{F}_0\right) = 1 \wedge \frac{M_0}{x}.$ relevance: 3.0 visual_id: 8724373	$P\left(\sup_{t\geq 0} M_t > x \mid \mathcal{F}_0\right) = 1 \wedge \frac{M_0}{x} \quad \text{a.s.},$ relevance: 3.0 visual_id: 8726965
CFTED	$P(S_t \geqslant x \mid \mathcal{F}_0) = M_0/x$ score: 0.031 rank: 1 visual_id: 1517366	$\rightarrow E[X_{T\wedge n} \vert \mathscr{F}_0] = / \leq X_{T\wedge 0}$ by choosing m $= 0$ score: 0.029 rank: 2 visual_id: 6928720	$E[Y \mid X = x_0] = 3x_0/2$ score: 0.028 rank: 3 visual_id: 6642421	$P(X_1 = 0 \vert X_0 = 1) = 1/2$ score: 0.026 rank: 4 visual_id: 8149851	$P(X_1 = 2 \vert X_0 = 1) = 1/2$ score: 0.026 rank: 5 visual_id: 8149850
search_arqmath3 _task2_colbert-APPROACH0	$\mathbb{P}\{\sup_{\{t\geq 0\}} M_t > x \vert \mathcal{F}_0\} = 1 \wedge \frac{M_0}{x}.$ score: 0.988 rank: 1 visual_id: 8722621	$P\{\sup_{t\geq 0} M_t > x \mid F_0\} = 1 \wedge \frac{M_0}{x}$ score: 0.926 rank: 2 visual_id: 6569532	$P\{\sup_{t\geq 0} M_t > x \mid F_0\} \geq 1 \wedge \frac{M_0}{x}.$ score: 0.881 rank: 3 visual_id: 6569534	$P\{\sup_{t\geq 0} M_t > x \mid F_0\} \leq 1 \wedge \frac{M_0}{x}$ score: 0.865 rank: 4 visual_id: 6569533	$P\left(\sup_{t\geq 0} M_t > x \mid \mathcal{F}_0\right) = 1 \wedge \frac{M_0}{x}.$ score: 0.81 rank: 5 visual_id: 8724373
xy5-2022	$P[X_0 = x_0] = 1$ score: 11.0 rank: 1 visual_id: 8919057	$P[X \geq \lambda] = 1/2$ score: 10.923 rank: 3 visual_id: 8247534	$E[1_{X\geq x}] = P[X \geq x]$ score: 10.692 rank: 12 visual_id: 6058413	$P[X_0 = 1 \vert X_3 = 0] * P[X_3 = 0]$ score: 10.571 rank: 16 visual_id: 54843	$P[X_n = 0] = 1 - 1/n^2$ score: 10.462 rank: 21 visual_id: 1583102
xy5-idf-2022	$P[X \geq \lambda] = 1/2$ score: 34.339 rank: 11 visual_id: 8247534	$P[X_n \geq 1/2] = 1$ score: 32.468 rank: 35 visual_id: 3783717	$E[1_{X\geq x}] = P[X \geq x]$ score: 30.946 rank: 75 visual_id: 6058413	$P[X_0 = 1 \vert X_3 = 0] * P[X_3 = 0]$ score: 30.908 rank: 76 visual_id: 54843	$P[X \in A \vert \mathcal{F}_t] = P[X \in A \vert \mathcal{F}_s]$ score: 30.391 rank: 97 visual_id: 7584479
xy7o4-2022	$P[X_0 = x] = 1$ score: 15.166 rank: 1 visual_id: 2932133	$E[N \vert X_0 = x] = 1/x$ score: 15.112 rank: 2 visual_id: 6565087	$P[X_0 = x_0] = 1$ score: 15.075 rank: 3 visual_id: 8919057	$P[X \geq \lambda] = 1/2$ score: 14.868 rank: 8 visual_id: 8247534	$P[X_n = 0] = 1 - 1/n$ score: 14.755 rank: 11 visual_id: 5069005

Figure 6: Top-5 Hits for Task-2 Runs, Query B.363: Weaker Results for PHOC Runs. Here the inability to match symbols not in the query negatively affecting PHOC run performance. Box colors represent high relevance (dark green), medium relevance (light green), low relevance (yellow), and non-relevance (red).



Rendered Query:

$$r \equiv \begin{cases} x = 1 \\ y = 1 \\ z = \lambda - 2 \end{cases}$$

IDEAL_FROM_QREL	🟩 $\begin{cases} x = t \\ y = 0 \\ z = 0 \end{cases}$ relevance: 2.0 visual_id: 5898385	🟩 $\begin{cases} x - 1 = 0 \\ \dfrac{y-3}{1} = \dfrac{t-4}{-1} \\ z - 2 = 0 \end{cases}$ whence $\begin{cases} x = 1 \\ z = 2 \\ y + t = 7 \end{cases}$ relevance: 2.0 visual_id: 6211680	🟩 $\begin{cases} x = -3 + t, \\ y = 4, \\ z = 1. \end{cases}$ relevance: 2.0 visual_id: 8847902	🟩 $\begin{cases} x = 3 \\ y = 0 \\ z = t \end{cases}$ relevance: 2.0 visual_id: 8369305	🟨 $\begin{cases} x = t \\ y = 6 - t \\ z = 1 \end{cases}$ relevance: 1.0 visual_id: 1075816
CFTED	🟨 $r : \begin{cases} x = t \\ y = 2t + 1 \\ z = t - 2 \end{cases}$ score: 0.033 rank: 1 visual_id: 6015325	🟥 $\begin{cases} x = 2 \\ y = 0 \\ z = -2 \end{cases}$ score: 0.031 rank: 2 visual_id: 7375371	🟨 $L : \begin{cases} x = 1 + 2t \\ y = 1 + 2t \\ z = 2 - t \end{cases}$ score: 0.03 rank: 3 visual_id: 7545300	🟥 $\begin{cases} x = \alpha \\ y = \beta \\ z = -\alpha - \beta \end{cases}$ score: 0.029 rank: 4 visual_id: 7855925	🟥 $\begin{cases} x = 1 \\ y = 0 \\ z = 0 \end{cases}$ score: 0.028 rank: 5 visual_id: 417788
search_arqmath3_task2_colbert-APPROACH0	🟥 $x = y = z = \dfrac{1}{\lambda + 2}$ score: 0.5 rank: 1 visual_id: 8312539	🟨 $a = \begin{cases} x = -2 + s \\ y = 2 + s \\ z = 1 - s \end{cases}$ score: 0.5 rank: 2 visual_id: 7781121	🟨 $l = \begin{cases} x = -2t + 1 \\ y = 3t - 2 \\ z = t + 4 \end{cases}$ score: 0.492 rank: 3 visual_id: 8609739	🟥 $\begin{cases} A = 1 \\ B = -2 \\ C = 1 \end{cases}$ score: 0.48 rank: 4 visual_id: 6108593	🟥 $\begin{cases} a = \frac{1}{2} \\ b = 1 \\ c = 0 \\ d = 1 \\ e = -2 \end{cases}$ score: 0.447 rank: 5 visual_id: 7388625
xy5-2022	🟨 $r : \begin{cases} x = t \\ y = 2t + 1 \\ z = t - 2 \end{cases}$ score: 9.769 rank: 0 visual_id: 6015325	🟨 $L : \begin{cases} x = 1 + 2t \\ y = 1 + 2t \\ z = 2 - t \end{cases}$ score: 8.643 rank: 2 visual_id: 7545300	🟥 $\begin{cases} x = 2 \\ y = 0 \\ z = -2 \end{cases}$ score: 8.636 rank: 3 visual_id: 7375371	🟨 $l = \begin{cases} x = t + 2 \\ y = 2t - 1 \\ z = t \end{cases}$ score: 8.462 rank: 4 visual_id: 5910979	🟥 $\begin{cases} x = t \\ y = t \\ z = t \end{cases}$ score: 8.1 rank: 6 visual_id: 4045644
xy5-idf-2022	🟨 $r : \begin{cases} x = t \\ y = 2t + 1 \\ z = t - 2 \end{cases}$ score: 37.387 rank: 0 visual_id: 6015325	🟥 $\begin{cases} x = t \\ y = t \\ z = t \end{cases}$ score: 34.951 rank: 1 visual_id: 4045644	🟨 $l = \begin{cases} x = t + 2 \\ y = 2t - 1 \\ z = t \end{cases}$ score: 34.331 rank: 2 visual_id: 5910979	🟥 $\begin{cases} x = 2 \\ y = 0 \\ z = -2 \end{cases}$ score: 33.656 rank: 4 visual_id: 7375371	🟨 $g : \begin{cases} x = 3s \\ y = 0 \\ z = 1 - 4s \end{cases}$ score: 32.647 rank: 6 visual_id: 6510769
xy7o4-2022	🟨 $r : \begin{cases} x = t \\ y = 2t + 1 \\ z = t - 2 \end{cases}$ score: 11.91 rank: 0 visual_id: 6015325	🟨 $L : \begin{cases} x = 1 + 2t \\ y = 1 + 2t \\ z = 2 - t \end{cases}$ score: 10.996 rank: 1 visual_id: 7545300	🟨 $L_2 = \begin{cases} x = 1 + t \\ y = 2 - t \\ z = 3 + 2t \end{cases}$ score: 10.623 rank: 3 visual_id: 3108030	🟨 $l = \begin{cases} x = t + 2 \\ y = 2t - 1 \\ z = t \end{cases}$ score: 10.601 rank: 4 visual_id: 5910979	🟥 $\begin{cases} x = 2 \\ y = 0 \\ z = -2 \end{cases}$ score: 10.348 rank: 7 visual_id: 7375371

Figure 7: Top-5 Hits for Task-2 Runs, Query B.345: Weak Results for All Runs Shown. Box colors represent high relevance (dark green), medium relevance (light green), low relevance (yellow), and non-relevance (red).