Overview of the TREC 2024 NeuCLIR Track

Dawn Lawrie,[†] Sean MacAvaney,[‡] James Mayfield,[†]
Paul McNamee,[†] Douglas W. Oard,[o†] Luca Soldaini,[*] Eugene Yang[†]
[†]Johns Hopkins University Human Language Technology Center of Excellence,
[‡]University of Glasgow, [o]University of Maryland, [*]Allen Institute for AI
lawrie@jhu.edu,sean.macavaney@glasgow.ac.uk,mayfield@jhu.edu
mcnamee@jhu.edu,lucas@allenai.org,oard@umd.edu,eugene.yang@jhu.edu

ABSTRACT

The principal goal of the TREC Neural Cross-Language Information Retrieval (NeuCLIR) track is to study the effect of neural approaches on cross-language information access. The track has created test collections containing Chinese, Persian, and Russian news stories and Chinese academic abstracts. NeuCLIR includes four task types: Cross-Language Information Retrieval (CLIR) from news, Multilingual Information Retrieval (MLIR) from news, Report Generation from news, and CLIR from technical documents. A total of 274 runs were submitted by five participating teams (and as baselines by the track coordinators) for eight tasks across these four task types. Task descriptions and the available results are presented.

1 INTRODUCTION

This is the third and final year of the TREC Neural Cross-Language Information Retrieval (NeuCLIR) track.[1] The first year of the track included News CLIR tasks [10]. The second year of the task added News MLIR and Technical Documents CLIR tasks [11]. In this third and final year of NeuCLIR those three task types continued, and we added a new Report Generation pilot task. This overview describes NeuCLIR's each of these four task types.

There are three News CLIR tasks, each of which has topics in English and news documents in one other language (Chinese, Persian or Russian).[2] CLIR is the most mature of the track's task types, and the capabilities that CLIR provides are foundational to the other three task types. Current CLIR systems face two broad challenges that distinguish CLIR from monolingual retrieval: (1) there is less robust training data than is available for monolingual ranked retrieval tasks; and (2) there is misalignment of term representations for different languages in multilingual embeddings. The news test collections in Chinese, Persian and Russian are the same as in the TREC 2022 NeuCLIR track. In 2024, the CLIR tasks provide relevance judgments for 56 new topics in Chinese, 68 new topics in Persian, and 65 new topics in Russian. Three of the five participating teams submitted CLIR runs in 2024, and the track coordinators also created baseline runs.

There is one News MLIR task, in which the topics are in English and the documents to be searched are news stories from the union of the Chinese, Russian and Persian news collections. This task requires generating a single ranked list for each topic over the unified collection. The principal additional challenge in this task is that scores computed independently for documents in different languages might not be as easy to compare as are scores for documents in a single language. This is the second year of the News MLIR task. In 2024, the news MLIR task has relevance judgments for 51 topics, each of which is also present in the 2024 topic set for two or more of the news CLIR tasks. Four of the five participating teams submitted results for the news MLIR task in 2024, and the track coordinators also created baseline runs.

There is one Technical Documents CLIR task in which the topics are in English and the documents to be searched are Chinese academic abstracts. The principal additional challenge in this task is that some standard tools such as multilingual embeddings, pretrained language models such as BERT [7], or generative large language models may be less well suited to the highly-technical language of these abstracts than to more general news text. In 2024, the Technical Documents CLIR task has relevance judgments for 71 new topics. Three of the five participating teams submitted results for the Technical Documents CLIR task, and the track coordinators also created baseline runs.

Report Generation is a new task type this year. In the Report Generation task, systems receive a report request in English and are asked to respond with an English report that meets the requirements specified in the request, and in which substantive statements in the report are supported by a citation to one or more documents in one of the three news collections. There are three report generation tasks, one each for reports with citations to Chinese documents, Russian documents, and Persian documents. Evaluation used a recently-proposed evaluation framework for such reports [15]. Four of the five participating teams submitted results for the Report Generation task.

The remainder of this paper is organized as follows. We begin with a brief summary of the News CLIR and News MLIR tasks, emphasizing changes since TREC 2023. This is followed by results for those tasks. Next, we present a brief summary of the Technical Documents CLIR task and the results for that task. Additional details for the News CLIR tasks can be found in the TREC 2022 and TREC 2023 NeuCLIR track overview papers [10, 11]; additional details on the News MLIR and Technical Documents CLIR tasks can be found in the TREC 2023 track overview paper. Following that, we provide details for our new Report Generation task. This will be the final year of the TREC NeuCLIR track. Report generation will continue to be evaluated in a new TREC 2025 RAGTIME track, so we conclude with a brief overview of our plans for RAGTIME.

2 NEWS RETRIEVAL TASKS

In this section we describe the CLIR and MLIR tasks.

[1] https://neuclir.github.io
[2] The news test collections also provide topics in Chinese, Persian and Russian, which can be used to create monolingual baselines or for other purposes.

2.1 Task Definitions

We have two news retrieval task types, CLIR and MLIR. The News CLIR task includes two “settings” (i.e., sub-tasks): ad hoc CLIR and (for pool enrichment) monolingual retrieval. The two settings use the same document collections, topics, and relevance assessments. Monolingual runs use topics manually translated into the language of the documents; ad hoc runs use the original English topics and rank documents from the entire collection.

2.1.1 Ad Hoc CLIR. For the ad hoc CLIR setting, systems receive a document collection in Chinese, Persian, or Russian, and a set of topics written in English. For each topic, the system must return a ranked list of 1,000 documents drawn from the entire target language document collection, ordered by likelihood and degree of relevance to the topic. Runs that use a human in the loop for ad hoc retrieval (or had design decisions influenced by human review of the topics) are indicated as “manual” runs; all others are considered “automatic.”

2.1.2 Monolingual Retrieval. While monolingual retrieval is not a focus of the NeuCLIR track, monolingual runs can improve assessment pools and serve as points of reference for cross-language runs. The monolingual retrieval setting is identical to the ad-hoc setting, but it uses topic files that are human translations of the English topics into a target language in a way that would be expressed by native speakers of the language. This setting is suitable for teams looking to explore monolingual ranking in languages other than English. It also has a lower barrier to entry than do the cross-language tasks.

2.1.3 Multilingual Information Retrieval (MLIR) Task. NeuCLIR 2023 added a multilingual retrieval task. This task is identical to the CLIR task described in §2.1.1, except systems must search all three document collections and produce a single unified ranked list. In other words, systems should treat the three document collections (§2.2) across all three languages as a single corpus. Participants were informed that, since topics for this task are derived from those of the CLIR task, there is no guarantee that each topic will have relevant results in every language.

2.2 Documents

NeuCLIR 2024 continues to use the NeuCLIR 1 document set, which was also used for NeuCLIR 2022 and 2023. The collection consists of roughly two million Persian documents, three million Chinese documents, and almost five million Russian documents spanning the years 2016 to 2021. For more information about how to extract text from Common Crawl News documents and how the collection can be obtained, see the NeuCLIR 2022 Overview paper [10].

This year we discovered a problem with language identification that affected 1,157 CommonCrawl files of the 16,951 files in the time window of August 2016 to July 2021. None of the documents in the affected files is included in the collection, which means that after pre-filtering and de-duplication, 765,299 Chinese documents, 317,392 Persian documents, and 3,410,884 Russian documents were excluded from the collection. This has less impact on the Russian collection, which was down-sampled to five million documents. For the other two languages, the collections would likely have included about half of the missing documents.

2.3 Topics

NeuCLIR 2024 topic development was completed entirely by NIST assessors.

The topic development process was identical to that used in 2023. Two paired assessors with language skills in two different languages met virtually to brainstorm a topic together. Good topics were described as topics that “revolve around events, people, and places, and [are] significant enough to have coverage in more than one language.” After exploring the collection with monolingual searches in the two assessor languages, a description was written, followed by a first draft of the narrative, and finally the title. This year assessors used a neural retrieval engine in hopes that initial statistics would better capture the prevalence of a topic in the collection. As usual a single monolingual search was initiated in each language, and each assessor counted the number of relevant documents in the top twenty-five returned documents. They were instructed to revise the topic if the count in either language was less than one or greater than twenty. Once the topic was appropriately scoped, the narrative was revised and the topic was included in the topic set for 2024. Ninety-two topics were created.

Because it was the third year using this document set, some topics were removed from consideration because they were semantically too similar to topics that appeared in 2022 or 2023. In addition, some topics were released as development data for the report generation task. Since the report generation development data included associated documents, these topics were also not considered for inclusion in the topic set. Each language ended up with a different number of assessed topics because the languages differ in the time required to judge a topic.

2.4 Relevance Judgments

Once all submissions were made, by-language pools were created that integrated the top-ranked documents from both CLIR and MLIR runs as well as documents cited by reports for the associated report request if the topic had one. The top 100 documents for runs that teams prioritized as their top nine runs were included in the pools. Such runs have a checkmark in the JFD columns of Tables 8, 9, and 10. For other runs, a depth of fifty was used. The Coordinators created some runs based on last year’s system submissions as baselines. The Coordinators were considered to be a unique team and followed the same cutoffs as other teams. The same four-point scale as NeuCLIR22 [10] was used to judge relevance. The four-point scale was converted to a three-point scale for the qrels,[3] again as was done for NeuCLIR22 [10] and NeuCLIR23 [11].

After pooling, some topics were dropped from the CLIR and MLIR tasks according to the following rules:

- If more than 40% of the judged documents were judged to be somewhat or very valuable in a particular language, drop the topic from the CLIR task in that language and from the MLIR task.
- If the relevance judgments for a topic had fewer than two documents in the somewhat or very valuable categories in a language, drop that topic from the CLIR task in that language, but include it in the MLIR task.

[3] The mapping from four relevance grades was 3->3, 2->1, 1->0, and 0->0.

Table 1: Relevance judgment statistics for 2024 News topics.

Topics Developed	Chinese	Persian	Russian	MLIR
# Topics Retained	56	68	64	52
Avg. # Judgments / Topic	700	661	577	1999
Avg. Prevalence	0.022	0.036	0.040	0.022

- If a topic had fewer than two relevant documents across all languages, drop it from the MLIR task.

Of the 92 topics created, 56 were retained for Chinese, 68 for Persian, 64 for Russian, and 52 for the MLIR task. All MLIR topics are judged in all three languages, though not all topics have relevant documents in all languages. Table 1 describes features of the topics that were retained based on the criteria described in Section 2.3.

2.5 Additional Resources

To support the aims of the CLIR and MLIR tasks and to lower the barrier to entry for new participants, the track made available additional resources beyond the document collection and topics. These resources included machine translated versions of queries and document collections, translations of the widely used MS MARCO collection into the three NeuCLIR-1 document languages, and previously-used IR test sets in the three NeuCLIR languages.

Machine translated versions of the queries were created by the online *Google Translate* service.

English versions of the document collections were created by directional machine translation Transformer models using the Sockeye version 2 toolkit. As the document collection for the CLIR and MLIR tasks did not change from 2022, these are the same translations that were provided in the previous two years of the track and that are described in the TREC 2022 NeuCLIR Overview paper[10].

As many neural IR systems are trained using data derived from the MS MARCO dataset[2], translations of this English resource into different languages were provided. We provided a version on Hugging Face called *NeuMARCO*.[4] We also provided links to similar translations from the *mMARCO* project[3] on the NeuCLIR website.

The track website also collected a number of multilingual and bilingual resources in the languages of the track including HC4 – a CLIR collection built over three years of Common Crawl data in the same three languages [12], as well as two multilingual CLIR datasets based on Wikipedia, known as CLIRMatrix [19] and WikiCLIR [18].

Finally, the topics and relevance judgments from the previous iterations of NeuCLIR (2022 and 2023) were available to track participants, either from NIST, or in `ir_datasets`.[5] These datasets could be used for system tuning and validation.

2.6 Participation

The News MLIR task had four participating teams, three of which also submitted runs to News CLIR tasks:

- Johns Hopkins University HLTCOE [?]
- University of Amsterdam (MLIR only) [9]
- University of Southern California ISI [1]
- University of Waterloo [17]

In addition to track participants, the track coordinators contributed baseline runs to ensure representation of a wide variety of retrieval approaches in the judgment pools. Table 2 shows the number of runs submitted under each category.

2.7 Track Coordinator Baselines

The track coordinators also prepared several runs to include as baselines. The foci of these runs were monolingual dense retrieval and sparse retrieval. The run names are outlined in Table 3. Notice that the MLIR run conditions are a subset of the CLIR run conditions. This is because there are fewer options for sparse retrieval when the documents are in multiple languages, and the idea of monolingual retrieval is nonsensical in this scenario.

2.7.1 *Monolingual Dense Retrieval.*

For monolingual non-English retrieval, ColBERT models were trained using translate-distill [24]. While the student model was shown translated queries and documents from MS MARCO in the non-English setting (e.g., Chinese for Chinese retrieval), scores came from the teacher model applying scores to the English queries and documents. This way the scores were not influenced by any 'translationese' introduced by machine translation. These runs used the human-translated queries along with the native documents. Because there is no notion of monolingual retrieval in MLIR, this approach is not used in MLIR.

2.7.2 *Sparse Retrieval.*

The coordinators included two broad categories of sparse retrieval. Probabilistic Structured Queries relies on a probabilistic translation table to cross the language barrier. BM25 relies on translation to cross the language barrier. Machine translation of either the query or document side each resulted in a CLIR run. Monolingual runs rely on queries expressed in the document language. Since NeuCLIR produces queries first in English, these non-English queries are referred to as human translations.

Probabilistic Structured Queries (PSQ) [6] is a translation approach that probabilistically matches a token from one language to a distribution of tokens in another. This technique can be used to translate queries, documents, or both. Prior work [20] has concluded that mapping documents to the query language at indexing time achieves the best effectiveness while minimizing query latency. The resulting documents are bags of probabilistic tokens in the query language. They can be indexed as ordinary documents in a sparse retrieval model such as BM25 or HMM with real-valued weights.

Our submission uses PSQ [26] to translate the documents and uses a Hidden Markov Model (HMM) [21] for retrieval.

[4] https://huggingface.co/datasets/neuclir/neumarco
[5] https://ir-datasets.com

Table 2: Number of News CLIR and MLIR runs submitted by language and team. Incorrectly submitted runs by h2oloo are reassigned to the correct tasks in the table but not in the actual pools created for relevance judgments.

Team	Persian	Russian	Chinese	MLIR	Total
IRLabAmsterdam	0	0	0	3	3
ISI-SEARCHER	1	1	1	1	4
h2oloo*	11	11	11	5	38
hltcoe	15	15	15	12	57
Track Coordinator Baselines	21	21	21	8	71
Total	37	37	37	62	173

The Patapsco framework [5] supports CLIR lexical retrieval through Pyserini [14]. Patapsco ensures that language-specific processing is consistent for both queries and documents. The coordinators submitted BM25 monolingual runs that used human-translated queries to search documents in their native language (QHT), CLIR runs that used the track-provided machine query translations to search the native documents (QGT), and runs that used English queries to search the track-provided document translations (DT). All languages used spaCy [8] for tokenization. For Russian and English machine translation, spaCy also provided stemming, while Parsivar [16] was used for Persian stemming. (We did not stem Chinese.) We explored three query variants: title, description, and title+description. We also explored the addition of ten RM3 expansion terms.

2.8 Results

2.8.1 Effectiveness and Run Diversity. Submissions this year feature LLM reranking. Tables 8, 9, and 10 present the evaluation results of the News CLIR submissions. Based on their run ids, all top-scoring runs use a generative model as the final-stage reranker, which has been the trend over the past year in the retrieval research community. Interestingly, as summarized in Figure 1 the most effective system for the Russian task used only the original documents, unlike others that used both original and translated documents. Such systems show the potential of developing effective neural retrieval pipelines without using any machine translation during indexing and search. While this year marks the final year of NeuCLIR, we believe there is still great room for future improvement in CLIR tasks.

The News MLIR task received fewer submissions but still included a wide range of approaches. nDCG@20 results are summarized in Figure 2. Please refer to Table 11 for the full results. One noticeable difference from the CLIR tasks is the emphasis on fusion techniques. Top-scoring runs in the MLIR task use systems that fuse multiple runs, possibly including CLIR runs, before reranking. Such approaches exemplify the need for strong end-to-end MLIR first (or early) stage retrievers that provide strong coverage in all languages. Similar to the CLIR tasks, top-scoring runs all use both original and translated documents.

All participant submissions include neural models in their pipelines, a common theme in modern retrieval research. To ensure that the pools still include documents that have surface forms matching with the queries, the coordinators submitted a number of BM25 variations to enrich the pools. While most of these runs are less effective than participant submissions, they ensure the quality of our pools and of the resulting collection.

2.8.2 Reusability. We conducted leave-one-run-out and leave-one-team-out experiments to assess the reusability of the collection. Since the pools for the CLIR and MLIR tasks are constructed from both CLIR and MLIR runs, leave-one-team-out removes all runs from the team across both CLIR and MLIR tasks to simulate the absence of the team entirely. Figures 3 and 4 summarize the results of these experiments.

There are many reranking runs, so the differences between the top-ranked document sets across runs are small (though the ranking may be drastically different, resulting in different effectiveness); this is shown in Figures 9, 10, 11, and 12. nDCG@20 is stable when leaving runs out of pooling across both CLIR and MLIR tasks. However, the relevant documents brought in by each run are more different, resulting in larger differences in R@1000. When leaving a team out from pooling, the reduced pools provide similar metric values and thus stable system ordering.

It is rare to see leave-one-run-out lead to less stable pools than leave-one-team-out experiments. We suspect this phenomenon is due to the submission error of one of the participating teams, who submitted all their MLIR runs to the CLIR task and vice versa; this resulted in shallower pool depths for their runs. In this experiment, we reassigned each run to its correct task; however, the actual pools were already created using the errorful runs. Despite this incident, we believe the collection to be reusable.

3 TECHNICAL DOCUMENTS TASK

The Technical Documents Task is in its second year. It is a cross-language ad hoc retrieval task, with English queries and Chinese documents. The key distinguishing feature of this task is the technical nature of the documents. The document collection is the same as was used for the Technical Documents Task in 2023. While last year's pilot task used a small number of topics, this year's full task allows researchers to gauge the effectiveness of existing CLIR approaches on technical documents, and to identify along which dimensions those systems need improvement.

This task contains the same two settings as the newswire CLIR task, namely ad hoc CLIR, and monolingual retrieval.

3.1 Documents

The documents for this task were abstracts from the Chinese Scientific Literature (CSL) dataset [13]. The dataset contains 396,209

Table 3: Track Coordinator baseline runs for CLIR and MLIR tasks. Run names in italics are monolingual runs.

Run Name	Type	Model	Query	Description
News CLIR and Technical Documents CLIR Baseline Runs				
patapscoBM25htRM3desc	Sparse	BM25	D	Monolingual Patapsco with RM3
patapscoBM25htRM3td	Sparse	BM25	TD	Monolingual Patapsco with RM3
patapscoBM25htRM3t	Sparse	BM25	T	Monolingual Patapsco with RM3
patapscoBM25htnoRM3desc	Sparse	BM25	D	Monolingual Patapsco without RM3
patapscoBM25htnoRM3td	Sparse	BM25	TD	Monolingual Patapsco without RM3
patapscoBM25htnoRM3title	Sparse	BM25	T	Monolingual Patapsco without RM3
patapscoBM25qtRM3desc	Sparse	BM25	D	Patapsco Google query translation with RM3
patapscoBM25qtRM3td	Sparse	BM25	TD	Patapsco Google query translation with RM3
patapscoBM25qtRM3title	Sparse	BM25	T	Patapsco Google query translation with RM3
patapscoBM25qtnoRM3desc	Sparse	BM25	D	Patapsco Google query translation without RM3
patapscoBM25qtnoRM3td	Sparse	BM25	TD	Patapsco Google query translation without RM3
patapscoBM25qtnoRM3title	Sparse	BM25	T	Patapsco Google query translation without RM3
patapscoBM25dtRM3desc	Sparse	BM25	D	Patapsco indexing translated documents with RM3
patapscoBM25dtRM3td	Sparse	BM25	TD	Patapsco indexing translated documents with RM3
patapscoBM25dtRM3title	Sparse	BM25	T	Patapsco indexing translated documents with RM3
patapscoBM25dtnoRM3desc	Sparse	BM25	D	Patapsco indexing translated documents without RM3
patapscoBM25dtnoRM3td	Sparse	BM25	TD	Patapsco indexing translated documents without RM3
patapscoBM25dtnoRM3title	Sparse	BM25	T	Patapsco indexing translated documents without RM3
plaid_distill_mono_ht	Dense	PLAID	TD	TD PLAID with monolingual training
fast_psqtd	Sparse	PSQ	TD	PSQ-HMM
fast_psqtitle	Sparse	PSQ	T	PSQ-HMM
	News MLIR Baseline Runs			
patapscoBM25dtRM3desc	Sparse	BM25	D	Patapsco indexing translated documents with RM3
patapscoBM25dtRM3td	Sparse	BM25	TD	Patapsco indexing translated documents with RM3
patapscoBM25dtRM3title	Sparse	BM25	T	Patapsco indexing translated documents with RM3
patapscoBM25dtnoRM3desc	Sparse	BM25	D	Patapsco indexing translated documents without RM3
patapscoBM25dtnoRM3td	Sparse	BM25	TD	Patapsco indexing translated documents without RM3
patapscoBM25dtnoRM3title	Sparse	BM25	T	Patapsco indexing translated documents without RM3
fast_psqtd	Sparse	PSQ	TD	Combining CLIR PSQ-HMM scores using score fusion
fast_psqtitle	Sparse	PSQ	T	Combining CLIR PSQ-HMM scores using score fusion

journal abstracts from 1,980 academic Chinese journals spanning 67 general disciplines, where Engineering, Science, Agriculture, and Medicine dominate. This is the same document set as was used in NeuCLIR 2023 for the technical documents task [11].

3.2 Topics

Topic creation in 2024 was accomplished by fifteen graduate students and one postdoc in Biology, Computer Science, Earth Science, Economics, Engineering, Math, and Physics from The Johns Hopkins University and from the University of Maryland, College Park. Annotators were hired based on their Chinese language skills and their familiarity with scientific research. During an interview, students were asked to describe their research area in both Chinese and English. They were then asked to choose a research topic they were familiar with, enter an English, Chinese, or mixed language query on that topic into an interactive search system that returned documents from the CSL dataset, and read and briefly summarize the top returned documents to determine whether they were relevant to their search. The purpose of this part of the interview was to ensure that the collection contained documents related to their area of research and to determine whether they could assess documents accurately in a timely fashion. Of the fifteen students, eleven were Ph.D. students and the others were Masters students.

Once hired, each annotator participated in a three-hour online training session. During the training, the topic creation task was explained. Then each person worked independently to create their first topic. During that process, two of the Coordinators reviewed their ongoing work. This exercise was used to ensure that topics had a suitable level of specificity, and that the tool was being used properly to determine whether abstracts on the topic existed in the collection. After the training, assessors were asked to spend up to a total of ten hours creating five to eight topics. One assessor created five topics, six assessors created six topics apiece, seven assessors created seven topics, and two assessors created eight topics, yielding a total of 106 topics.

The English title, description, and narratives were reviewed by a coordinator to ensure that the topic was sufficiently descriptive. The topic was also checked for grammar and spelling. In some

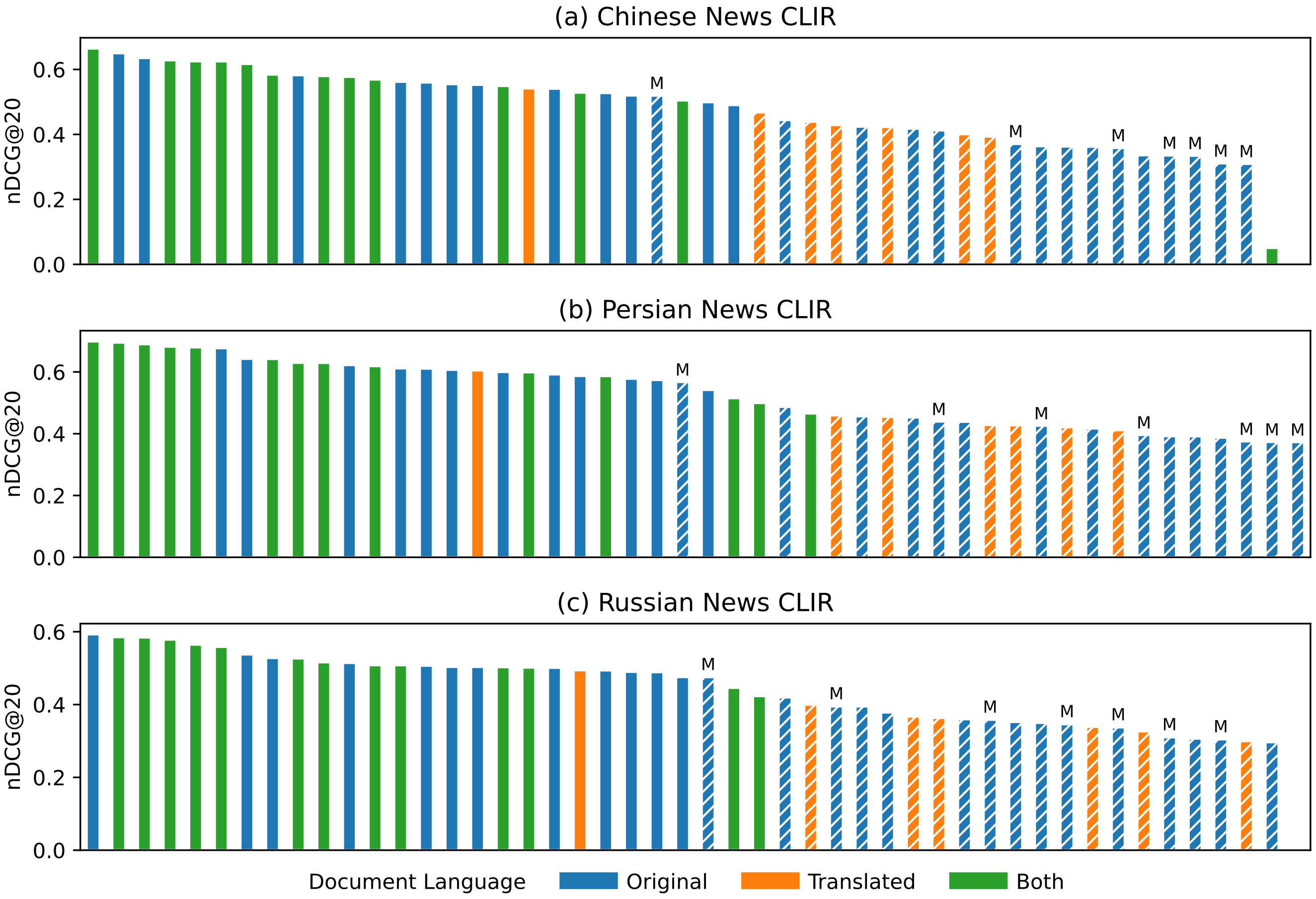


Figure 1: News CLIR nDCG@20. Coordinator runs are marked with slashes. Monolingual runs (i.e., using human-translated topics) are marked with "M" at the top of the bar.

cases, the assessors were asked to revise topics that appeared to be too vague or were not understandable. Narratives were also given special attention to ensure they were sufficiently detailed. After any revision, assessors checked the translation to ensure that it incorporated any changes. The translations did not undergo any external quality control. In the end 106 topics were distributed to participants.

3.3 Relevance Judgments

Assessors then participated in a second online training session, lasting two hours, that focused on relevance judgment. In addition, instructions written in English were provided for completing relevance judgments. Relevance for the Technical Documents pilot task differed somewhat from the usual TREC view of relevance. Assessors were asked to imagine that they were writing the background section or the related work section of a scientific paper on the topic they had created. They were asked to evaluate whether they would plan to read the paper being judged based on its abstract so as to possibly cite the paper in their related work section.

They answered two questions about each document:

Question 1. Does this document contain central information?

Yes There is information in the abstract related to their search topic.

No There is no information in the abstract related to their search topic.

Unable to judge The document was not viewable in the document viewer panel.

Question 2. How valuable is the most important information in this document?

Very Valuable One would definitely read the paper associated with this abstract when writing the related work section for this research topic.

Somewhat Valuable If one had enough time one would read the paper, because it might have something that could appear in the related work section, but confidence about that is low.

Not that Valuable One is unlikely to read the paper because one does not expect to find in it information that one would cite in the related work section.

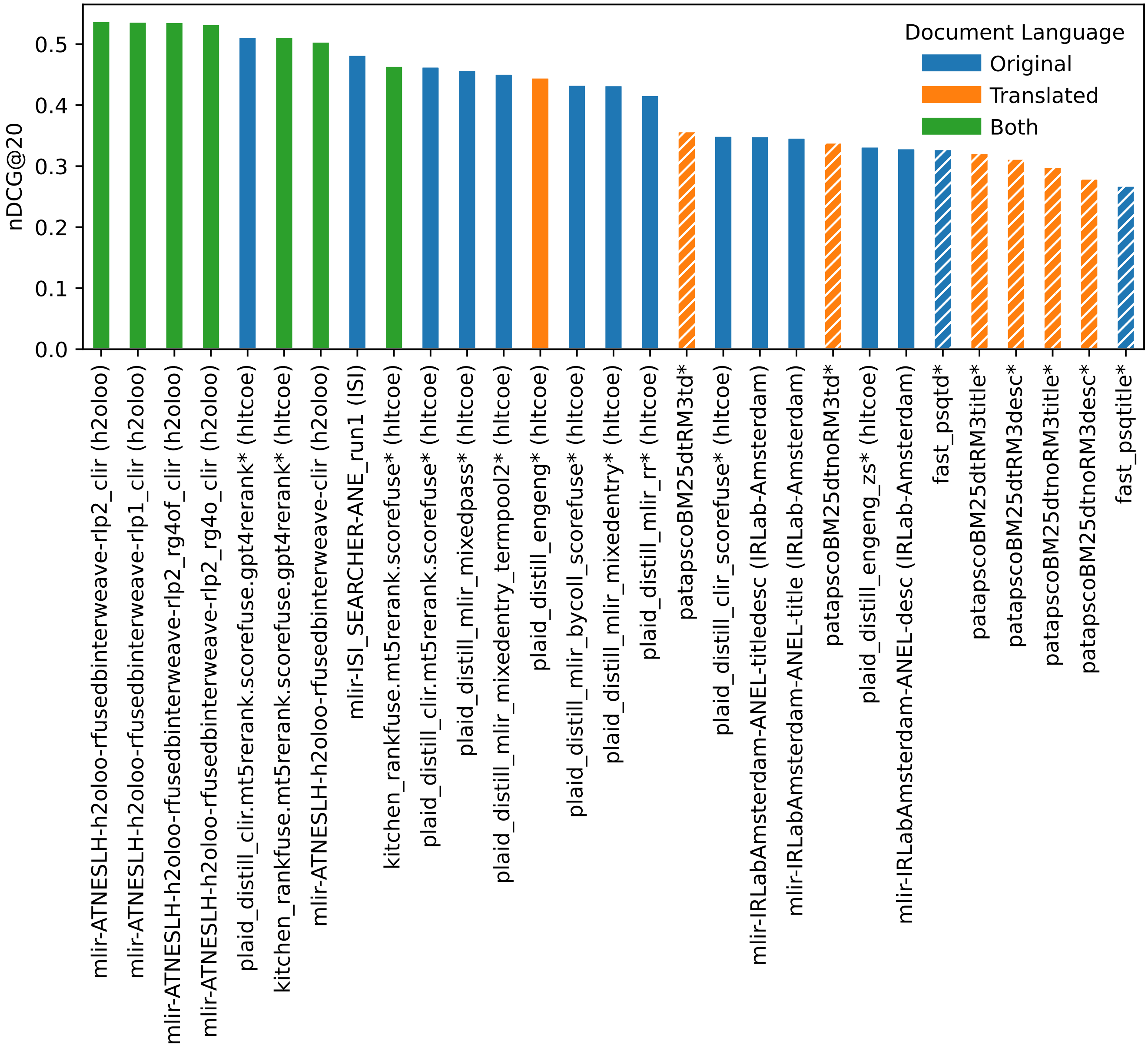


Figure 2: News MLIR nDCG@20. Coordinator runs are marked with slashes.

Table 4: Relevance judgments for the Technical Documents CLIR task.

	Biology	Computer Science	Earth Science	Economics	Engineering	Physics	Overall
# Topics Included	13	14	19	2	10	13	71
Avg. # Judgments / Topic	367.92	470.43	292.53	375	315.2	311.54	350.41
Avg. # Somewhat Valuable / Topic	8.69	17.36	10.58	28	12	7.08	11.62
Avg. # Very Valuable / Topic	4.46	9.07	9.11	27.5	9.8	6.46	8.46

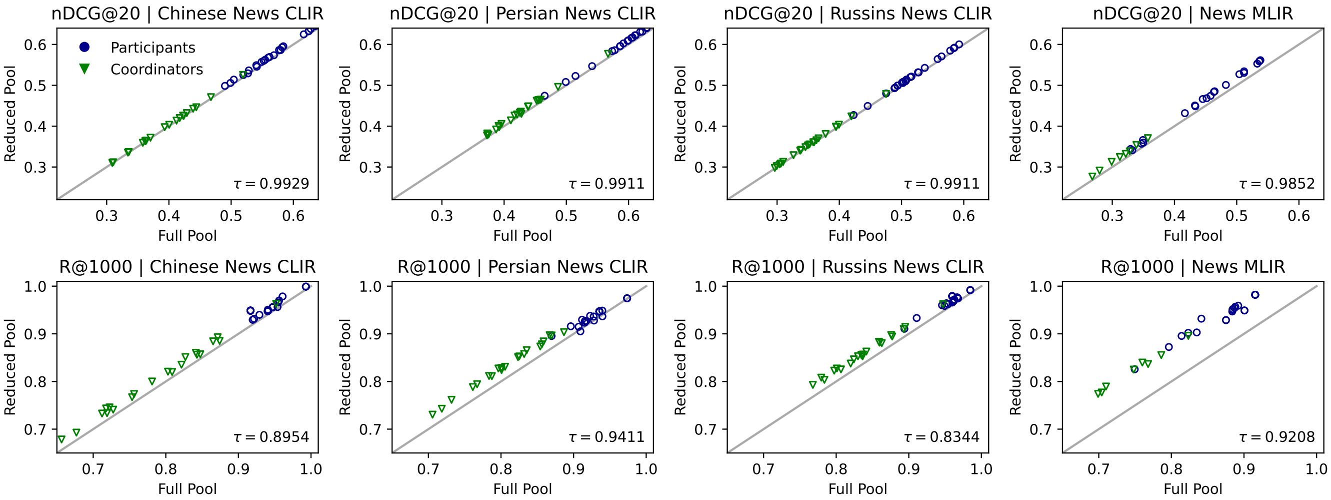


Figure 3: News CLIR and MLIR Leave-One-Run-Out Experiments.

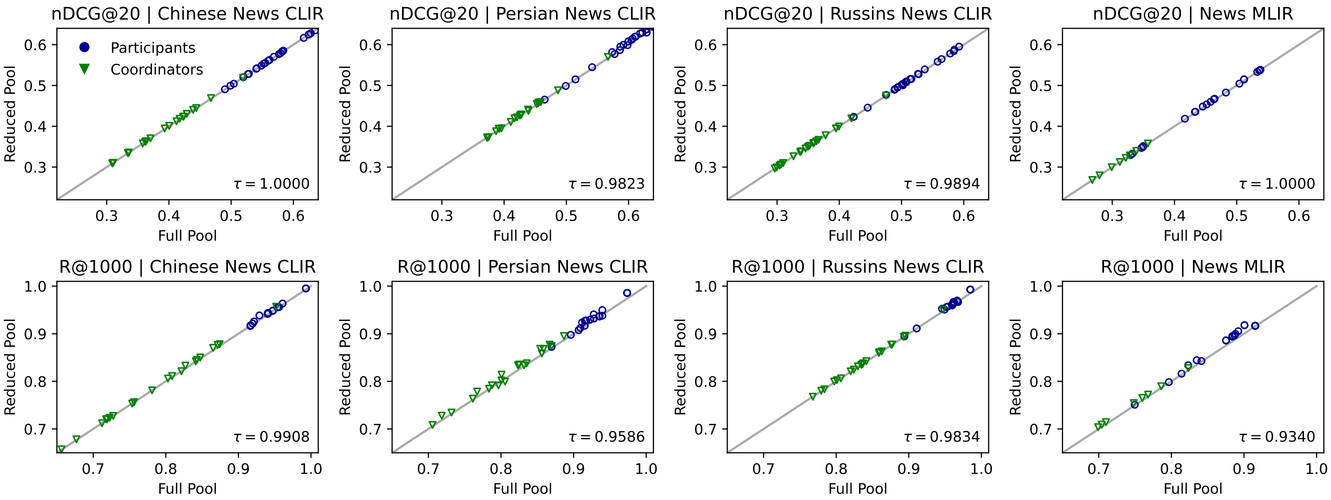


Figure 4: News CLIR and MLIR Leave-One-Team-Out Experiments.

After training, they were asked to judge part of one topic and then given an opportunity to ask questions. Most assessors finished the first topic the same day as the training.

Assessors were asked to spend at most ten hours judging, and their progress was tracked. Some assessors ran out of time, leading to seven unjudged topics. Topics were removed for having fewer than three relevant documents (affecting eight topics) or for judging more than 20% of the pool to be somewhat or very valuable (affecting sixteen topics). Four other topics were removed because the assessor experienced technical difficulties while judging the task. No assessor had both a topic removed for having too few relevant documents and a topic removed for having too many relevant documents.

Pools were created from the top thirty-five documents of each submitted run. In the end seventy-one topics were used to judge system performance. Table 4 contains information on the judgments, on the average number of very valuable per topic and the average number of somewhat valuable per topic in the judged pools. Documents that the assessor identified as relevant during topic development were also included in the pools.

3.4 Additional Resources

In addition to the document collection itself and the resources already described in Section 2.5, the track provided translations into Chinese of the topic fields, and translations into English of the document texts. These translations were obtained from the online *Google Translate* service in 2023.

3.5 Participation

The Chinese Technical Documents CLIR task had three participating teams that together submitted 28 runs:

- Johns Hopkins University HLTCOE [25]
- University of Southern California ISI [1]
- University of Waterloo [17]

3.6 Track Coordinator Baselines

The track coordinators also created several runs to include as baselines. The foci of these runs were monolingual dense retrieval and sparse retrieval. The run names are outlined in Table 3. See Section 2.7 for more information about the algorithms used in these runs.

3.7 Results

3.7.1 Effectiveness and Run Diversity. During this final year of NeuCLIR we primarily focused on the utility and the reusability of the resulting collection. Summarized in Figure 5, we observe a wide range in effectiveness with a clear separation between the neural and statistical methods of around 0.3 in nDCG@20. The coordinators submitted the relevant documents found during topic development as a manual run (`topic_dev`); among all the runs in Figure 5 this run ranked directly below the runs that use heavy rerankers at the end of their retrieval pipeline Detailed effectiveness measures are presented in Table 12.

Most systems used both translated and original document text to increase effectiveness, with the exception of the best system. Top-ranked runs all use a large generative model as a reranker (based on the run IDs and personal communications with the participants); this aligns with findings in recent monolingual neural retrieval literature.

Despite the number of participating teams being less than ideal, the submitted systems included a wide variety of models and approaches. The coordinators also contributed several statistical runs that are less popular nowadays and that are missing from the participant submissions. While these are significantly less effective than the neural runs, they provide enrichment to the pools and assurance of measurability on the lower effectiveness ranges. Captured in the overlapping graphs in Figure 13, the coordinator runs are less similar to the participants' runs (darker lower left corner) for both retrieved documents (Figure 13(a)) and retrieved relevant documents (Figure 13(b)). In particular, the topic development manual run provides a very different set of retrieved documents (the black strip in Figure 13(a)) than all other system submissions. Annotators tended to issue several queries. They tended to stick to one query language, only occasionally entering mixed-language searches.

3.7.2 Reusability. We directly test reusability by conducting leave-one-run-out (LORO) and leave-one-team-out (LOTO) experiments to verify the robustness of the set of relevant judgments produced from pooling.

nDCG@20 on both LORO and LOTO experiments indicates very stable pools with 0.99 and 0.97 Kendall's τ respectively. While leaving one team out from the pool leads to slight instability in the results, this is due to having only four (including the coordinators) teams in the pool. However, since teams all provide diverse sets of runs (especially for `h2oloo`), leaving one team out would remove a large and diverse set of runs.

With the large number of reranking runs, LORO on R@1000 is less meaningful, since runs from a single organization tend to share retrieval results. These trends are captured in the overlapping graphs in Figure 13. LOTO indicates a stronger instability in R@1000. While 0.92 Kendall's τ is still extremely high, it indicates less stability than measuring the top of the ranked list, i.e., measuring nDCG@20. However, we anticipate no reusability issues for this collection. Future CLIR research can and should evaluate on this collection.

4 REPORT GENERATION PILOT TASK

CLIR solves one problem—it ranks documents relative to a query in another language; but it creates another—someone has to read all those documents! This is not just a matter of the time and effort required—some searchers may also not be able to read documents in their original language. The goal of the TREC 2024 NeuCLIR Report Generation task is to address both of these challenges by creating concise focused reports (i.e., multi-document summaries) in the language of the report request (which in our case is English). Each report is based on documents from a single NeuCLIR collection (Chinese, Persian or Russian). These reports are evaluated based on the degree to which they use correctly cited references to documents in the specified collection to answer questions that the report requester wished answered using the procedure proposed by Mayfield et al. [15].

4.1 Documents

The Report Generation task used the same collections as the News CLIR tasks, which contain CommonCrawl News articles in Chinese, Russian or Persian.

4.2 Report Requests

Assessors began with topics created for the CLIR tasks. They worked from each of these topics to create a report request by adding a background section describing why the report was needed and a detailed problem statement that described what the report should contain. These added sections were generally based on the topic description and narrative fields of the CLIR topic from which the report request was derived. In 2024, relevant documents were assumed to contain answers to the nugget questions used to evaluate the generated reports; thus, report requests were intended to ask for the same information as the original topic as described in that topic's title, description, and narrative fields.

A report request consists of a request ID, a collection ID, a background section, a problem statement, and a length limit (in Unicode characters). The background and problem statement fields of a report request are expressed in unstructured text. Here is an example:

```
Background: I am a Hollywood reporter writing an article
about the highest grossing films Avengers: Endgame and
Avatar.
Problem statement: The article needs to include when
each of these films was considered the highest grossing
film and any manipulations undertaken to bring
moviegoers back to the box office with the specific
```

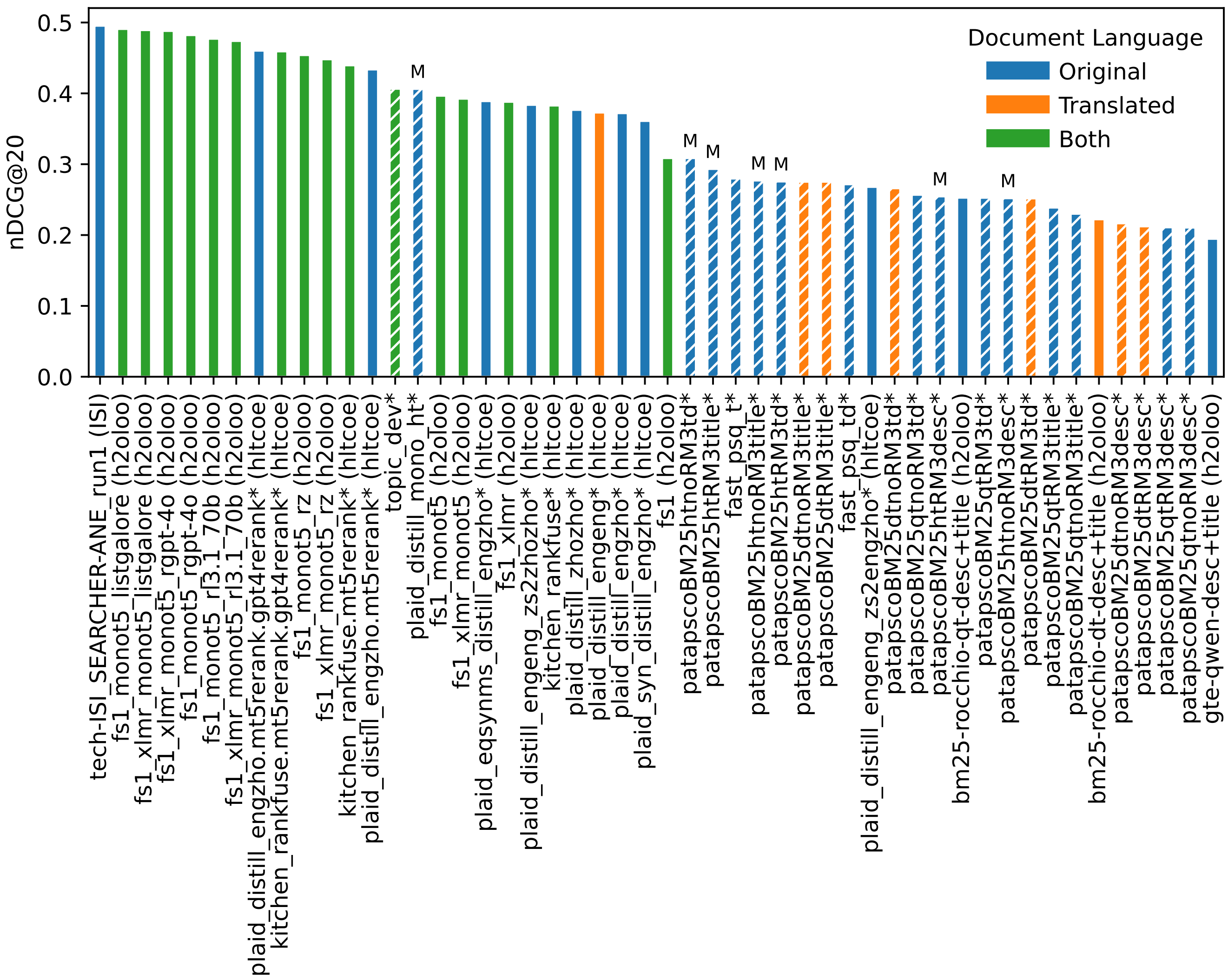


Figure 5: Technical Document CLIR task nDCG@20. Coordinator runs are marked with slashes. Monolingual runs (i.e., using human-translated topics) are marked with "M" at the top of the bar.

```
goal of increasing the money made on the film.
Limit: 2000
```

4.3 The Assessment Process

NIST assessors created the ground truth data that was used to evaluate Report Generation runs. They did this using a process that largely follows the evaluation design described in Mayfield et al. [15]. A summary of the assessment statistics is presented in Table 5. The first products of this assessment process were created together with the report requests. After drafting a report request, the assessor manually wrote an example report. Using their report request and their example report, they then decided on the questions that a report responsive to the report request would need to answer. At this time they also recorded any answers to each question that were known to them from their research on the topic while writing their example report. Answers were expressed in English rather than in the language of the document. After an initial quality assurance check these elements were finalized and retained for later use during evaluation. Only the report request was provided to participating teams. In total, we developed 59 report requests. We assign all requests to all three languages, resulting in 177 requests.

After runs for the Report Generation task were received, assessors began "nugget judgment." This process was timed to follow completion of relevance judgment for the CLIR and MLIR tasks because it was useful to have the fullest possible set of relevant documents available. The initial nugget questions developed immediately following the report request creation process were then reviewed again and, when necessary, revised to ensure they were atomic and could be answered with phrases. The questions were

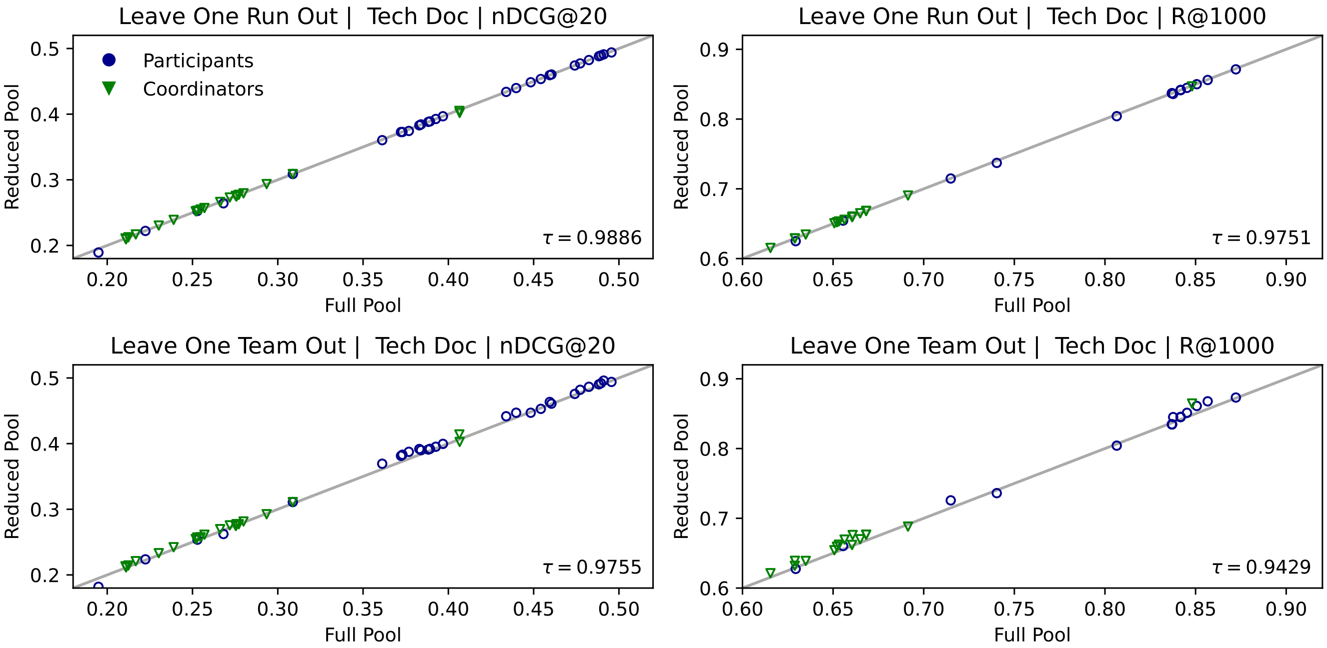


Figure 6: Leave-One-Out and Leave-One-Team-Out Experiments on the Technical Document CLIR task.

Table 5: Report Generation Assessment Statistics.

	Chinese	Persian	Russian
# of Report Request Developed	59	59	59
# of Report Request Assessed	21	20	21
Avg. # of Nugget Questions per Request	13.71	13.80	13.76
Avg. # of Nugget Questions w/o Answers in that language per Request	2.57	3.45	0.86
Avg. # of Uncaptured but Supported Crucial Nuggets per Assessed Report	1.57	1.89	0.55
Avg. # of Uncaptured but Supported Topical Nuggets per Assessed Report	1.97	3.83	3.90

also marked as “ok” or “vital” during this review process. A second quality assurance check was then performed.

Assessors then examined all relevant documents (based on relevance judgments for the associated CLIR topic) and all cited documents (i.e., all documents cited in any submitted report). The set of nugget answers was then revised to match the revised questions, and extended based on additional relevant and cited documents that had been found. Each answer was linked to all known documents that contained that answer.

Concurrently with this answer revision process, the assessors also judged all citations in each report sentence. Citations were judged based on whether the facts expressed in the sentence containing the citation were found in the cited document. A sentence citation was scored as full, partial, or no support. Reports could cite up to two documents per report sentence; however, no attempt was made to determine if two partial scores when combined would fully support the information in the sentence. This was a limitation of the assessment process, which collected all report sentences citing a particular document for the assessor to judge. While sentences and citations were presented to assessors without the context of the report from which they had been extracted, assessors could access the entire report if they needed it to fully understand the information in a sentence. At the end of this assessment phase, a list of known correct answers to questions had been recorded and all citations had been assessed.

The final assessment phase required the assessor to determine whether the sentence answered one or more nugget questions. If it did answer a question, the particular answer contained in the sentence was selected or the answer was marked as “other answer.” If the sentence did answer one of the identified questions, a sentence could be scored as “other crucial nugget to the request” indicating that an additional nugget question/answer pair should have been created (assuming the LLM did not hallucinate the information); “topical nugget” indicating the information was on topic but not necessary for the report; “irrelevant nugget” indicating a fact not responsive to the report request was included in the report; and

"No nugget found" indicating the sentence contained no facts or its facts had previously been expressed in the report. When performing this assessment phase, assessors were not told whether a sentence under consideration contained citations. This was meant to prevent biasing the question answering assessment with effects from the presence or absence of suitable citations. The number of supported sentences (supported by the citations) that answer a crucial or topical nugget that should have been created are summarized in Table 5.

Both assessment processes were performed using the tool described in Yang et al. [22]. This tool facilitated linking answers to documents, assessing citations, assessing whether report sentences contained facts, and whether those facts were answers to a nugget question.

Once citations were judged and answers to nugget questions had been identified, those results were used to compute a score for each sentence.

4.4 Additional Resources

4.4.1 Retrieval Service. To support teams primarily focusing on the report generation task, we provided a PLAID-X [24] search service through a web API that used an English-trained model [23][6] for all languages. To minimize the resources needed to host the service, we included the ability to remove documents in other than the requested language.. The user can request up to 100 documents for each query. The service retrieves ten times the number of documents the user requested to ensure enough documents in the requested language are retrieved.

4.4.2 Development Data. Report generation topics were taken from the MLIR topic set. NIST assessors were asked to generate report requests and sample reports for these topics. Some topics were generated by more than one assessor, both because assessors working on different languages shared topics, and to support cross-assessor studies. Track coordinators selected a single report request for each topic for inclusion in the track data. No software was available to support this process, so assessors entered everything free-form in a shared document. This led to a need for topic curation, as citation and question formats differed across assessors. Track coordinators manually curated the report requests, and converted them to JSON format. A total of forty-seven report requests were released as development data, statistics on which appear in Table 6. Because the questions and answers were not normalized, questions were issued with the following caution: "Warning: these questions and answers are not all representative of the way questions will be asked and answered in the pilot, because some questions are compound and some answers do not include citations."

4.5 Participation

We received 51 runs for the Report Generation task from four participating teams, including 17 runs for each of the three document languages:

- Johns Hopkins University HLTCOE (4 runs per language) [25]
- University of Waterloo (4 runs per language) [17]
- University of Amsterdam (7 runs per language) [9]
- IDA/CCS (2 runs per language) [4]

The track coordinators did not prepare baseline runs for this pilot task.

4.6 Results

Tables 13, 14, and 15 show ARGUE report generation scores for Chinese, Persian, and Russian respectively. Figure 8 indicates how the scores in those tables were derived. Scores were mapped to a zero-to-one scale by scoring categories labeled (−) as 0.0 and the single category labeled (+), Supported Relevant Nuggets, as 1.0. Neutral labels (0) were ignored by removing them from both numerators and denominators of the ARGUE precision calculation.

4.6.1 Citation Precision. is the proportion of citations that accurately refer to a relevant document. There are two ways a document might be considered relevant for this purpose: the relevance label is derived from nugget annotation (that is, any document that contains a mention of a nugget is treated as relevant), or by using the qrels for ranked retrieval (shown in Table 7). Despite the apparent difference in the definition of relevance (or approach for acquiring such annotations), Pearson's rank correlations under the two definitions are high – 0.9 for all three languages.

4.6.2 Nugget Recall and Support. *Nugget recall* is the proportion of nugget questions correctly answered by at least one report sentence. Multiple report sentences correctly answering a single nugget question are counted only once. In contrast, *nugget support* is the proportion of the report's nuggets for which the sentence that answers a nugget question is supported by its cited documents. Both measures require the nugget mentioned in the report to be supported by the cited documents to be credited.

4.6.3 Sentence Support. Similar to Nugget Support, *sentence support* is the proportion of sentences in a report supported by its cited documents. This includes support for information that is not captured by nuggets (which is unlikely to be relevant to the report request). Such information is allowed in a report without penalty to partially address disagreements between systems and assessors over which nuggets are crucial to a report.

4.6.4 Scores. The top submitted runs score close to 0.9 on the main ARGUE measure. However, top citation precision scores are only around 0.3, suggesting that systems can still improve on the inclusion of relevant information in the report. Top nugget recall is less than 0.5, indicating that there is also room for improvement in nugget/fact coverage of the topic.

5 FUTURE DIRECTIONS

The NeuCLIR track is retiring after TREC 2024. In 2025, we will transition to the new RAGTIME track, where the primary task will be report generation from multilingual news content in Arabic, Chinese, English, and Russian.

5.1 What's New in RAGTIME 2025?

RAGTIME will include a new test collection, with the size of each of the four language components balanced across those languages. Report Generation in RAGTIME will differ from our early work with Report generation in NeuCLIR in three important ways. First,

[6] https://huggingface.co/hltcoe/plaidx-large-eng-tdist-mt5xxl-engeng

Table 6: Development report request statistics.

	Chinese	Persian	Russian
# Reports	19	8	20
Avg. # Characters / Reports	1865.2	2256.0	1894.1
Avg. # Sentences / Reports	15.0	16.8	16.6
Avg. # Citations / Reports	10.0	11.1	16.9
Avg. # Unique Citations / Reports	6.3	5.9	7.8
Avg. # Questions / Reports	11.5	12.4	14.4

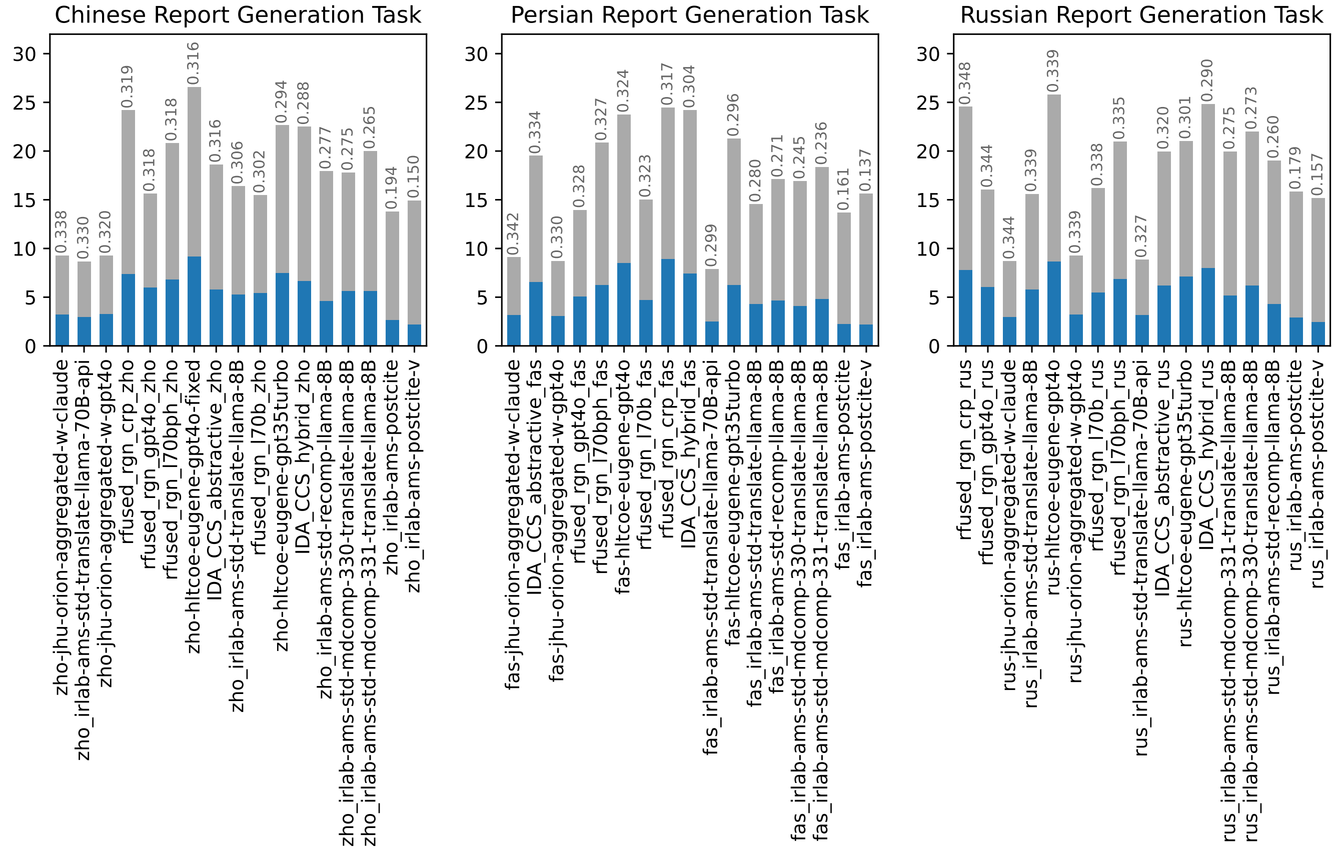


Figure 7: Bar chart of the number of relevant citations (in blue) across all cited documents (in gray) for each submitted run. Each bar represents a run. Values at the top of each bar are the document-level precision of the generated report, which are reported in Tables 13, 14, and 15.

the report request in RAGTIME is generated *de novo* for the task, rather than having been developed based on a topic originally developed for MLIR (as was the case in NeuCLIR). Second, the RAGTIME report generation task is not limited to documents in a single language. Instead, systems will be asked to draw on content in all four RAGTIME languages. Notably, one of the languages is English, the same language as the report request and the report. Thus RAGTIME involves the additional challenge of integrating same-language and cross-language sources, which was not present even in the NeuCLIR MLIR task (because the NeuCLIR test collection did not include English). Third, RAGTIME will also include an emphasis on automating parts of the evaluation, with an eye toward fostering reusability of the test collection.

Our present plans for RAGTIME also include four changes from NeuCLIR that are intended to facilitate the entry of new participants. First, the RAGTIME collections are somewhat smaller than those of NeuCLIR, with about 1 million documents per language. This change is intended to deemphasize the scalability of the retrieval component to allow participating teams to focus more on report

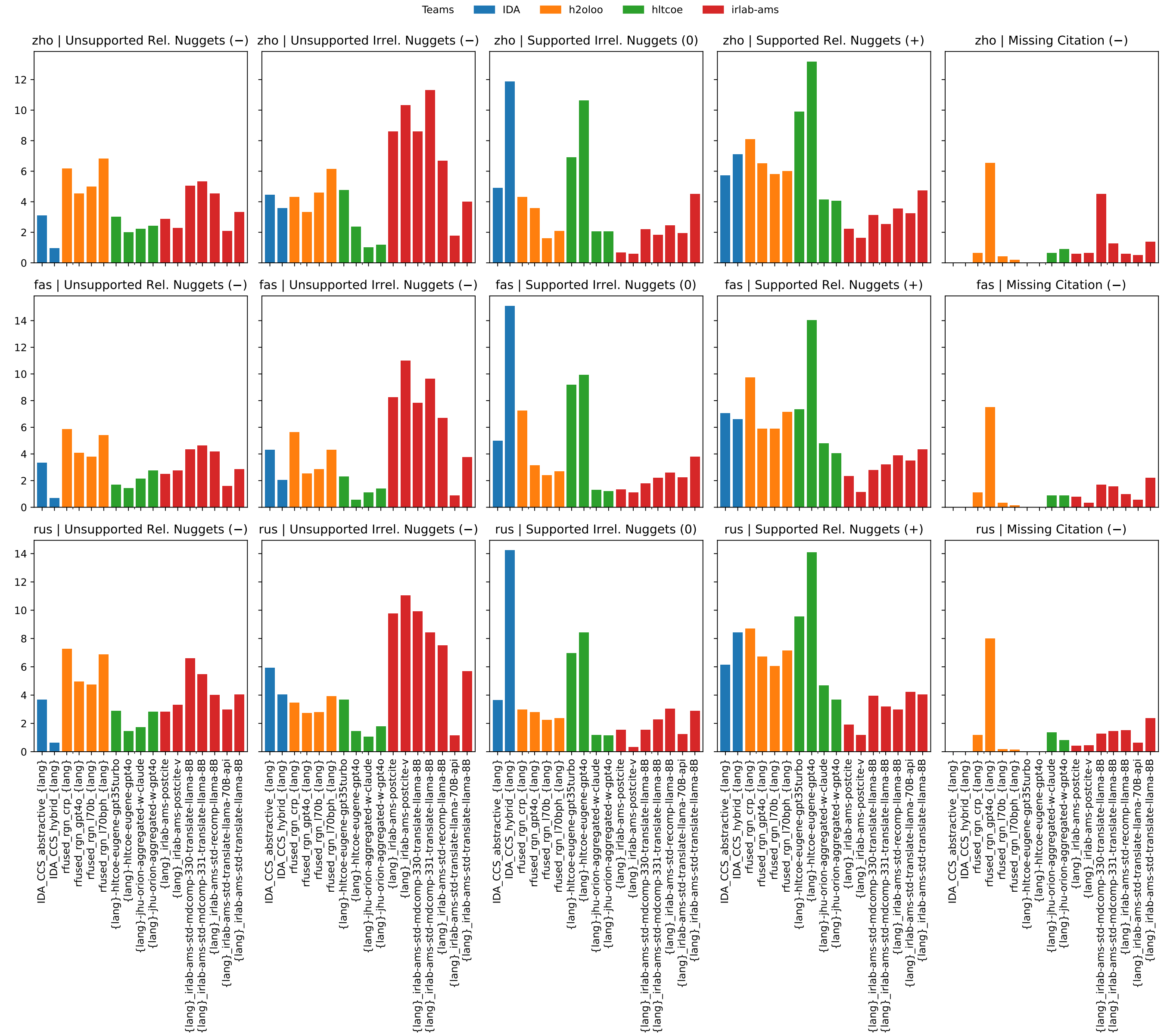


Figure 8: Components of ARGUE scores for each Report Generation run.

Table 7: Relevance agreement between ranked retrieval (qrels) and nugget annotation.

	From Nugget Annotation					
	Chinese		Persian		Russian	
From qrels	NRel.	Rel.	NRel.	Rel.	NRel.	Rel.
Non-Relevant	475	372	480	310	455	424
Relevant	37	169	24	249	35	278

generation. Second, because two of the source languages will be new, the TREC 2025 RAGTIME track will include an early-summer dry run to give participating teams access to some manually annotated development data for their final systems. Third, there will be a monolingual English task for teams that do not want to work with non-English content. Fourth, the RAGTIME output formats are consistent with those of other TREC RAG tracks (BioGen, DRAGUN, and RAG) to simplify participation in several of those tracks.

We are not currently planning future work with the NeuCLIR Technical Documents collection because that collection lacks the degree of topical convergence across documents that we believe would be needed for use in a Report Generation task. We believe that the collection is suitable in its present form for evaluating the ability of a CLIR system to handle technical vocabulary.

5.2 What's Not Changing?

There will be some continuity along with these changes. Most notably, RAGTIME will continue to report results for CLIR and MLIR for teams who have interest in those tasks. This benefits reusability of the RAGTIME data through pool enrichment, because CLIR and MLIR runs may find documents that no Report Generation system chose to include in its report. We expect this will be of particular interest for Arabic, where we expect RAGTIME to produce the largest available CLIR test collection in that language. Notably, however, the form and content of the topics will differ in RAGTIME. Specifically, while there will be a title field (which can be used as a short web-style query), the traditional TREC description and narrative fields will be replaced with the report request, and relevance judgments will be based on that report request. Thus the RAGTIME CLIR and MLIR tasks may help to push the state of the art for retrieval over very long queries.

6 CONCLUSION

In this third and final iteration of the TREC NeuCLIR track, we have completed our development of the first set of large CLIR test collections with judgment pools augmented by modern transformer-based neural information retrieval systems. This NeuCLIR news collection includes more than 100 topics for each of the three languages. As in prior years, we continue to see strongest effectiveness from neural CLIR systems. The NeuCLIR test collection also supports MLIR evaluation, which continues to be a more challenging task for systems than is CLIR.

NeuCLIR 2024 was the second and final year of the Chinese Technical Documents CLIR task, which has produced a new test collection that now also has more than 100 topics. In contrast to 2023 when no training data specific to this task had been available, we saw the emergence of relatively strong CLIR systems for these more challenging documents that are not reliant on machine translation of the full document set.

Finally, NeuCLIR 2024 has also served as an incubator for a new Report Generation task, which has a bright future. Results on the NeuCLIR Report Generation task are promising, but reveal areas that need significant further research. Experience with the design of the NeuCLIR task has informed the design of the TREC 2025 RAGTIME track.

In short, the three years of the TREC NeuCLIR achieved what it set out to do, and more. While the test collections built by the track are one clear legacy with enduring value, it is the research community that has and will make use of those test collections that we expect will be the track's most lasting legacy.

REFERENCES

[1] Shantanu Agarwal, Joel Barry, and Scott Miller. 2025. ISI's SEARCHER II System for TREC's 2024 NeuCLIR Track. In *The Thirty-Third Text REtrieval Conference (TREC 2024) Proceedings.*

[2] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268 [cs.CL]

[3] Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. mMARCO: A Multilingual Version of the MS MARCO Passage Ranking Dataset. arXiv:2108.13897 [cs.CL]

[4] John M. Conroy, Razieh Fathi, Daria Smylova, and Y. Kelly Wu. 2025. Cross-lingual Hybrid and Abstractive Summarization Automatic with Attribution for the NeuCLIR 2024 Pilot Report Generation Track. In *The Thirty-Third Text REtrieval Conference (TREC 2024) Proceedings.*

[5] Cash Costello, Eugene Yang, Dawn Lawrie, and James Mayfield. 2022. Patapsco: A Python Framework for Cross-Language Information Retrieval Experiments. In *Proceedings of the 44th European Conference on Information Retrieval (ECIR).*

[6] Kareem Darwish and Douglas W Oard. 2003. Probabilistic structured query methods. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* 338–344.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/V1/N19-1423

[8] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-Strength Natural Language Processing in Python. (2020). https://doi.org/10.5281/zenodo.1212303

[9] Jia-Huei Ju and Andrew Yates. 2025. IRLab-AMS at TREC 2024 NeuCLIR Track. In *The Thirty-Third Text REtrieval Conference (TREC 2024) Proceedings.*

[10] Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W Oard, Luca Soldaini, and Eugene Yang. 2022. Overview of the TREC 2022 NeuCLIR Track. In *Proceedings of The Thirty-First Text REtrieval Conference.*

[11] Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W Oard, Luca Soldaini, and Eugene Yang. 2023. Overview of the TREC 2023 NeuCLIR Track. In *Proceedings of The Thirty-Second Text REtrieval Conference.*

[12] Dawn Lawrie, James Mayfield, Douglas W. Oard, and Eugene Yang. 2022. HC4: A New Suite of Test Collections for Ad Hoc CLIR. In *Proceedings of the 44th European Conference on Information Retrieval (ECIR).*

[13] Yudong Li, Yuqing Zhang, Zhe Zhao, Linlin Shen, Weijie Liu, Weiquan Mao, and Hui Zhang. 2022. CSL: A Large-scale Chinese Scientific Literature Dataset. In *Proceedings of the 29th International Conference on Computational Linguistics.* International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 3917–3923. https://aclanthology.org/2022.coling-1.344

[14] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021).* 2356–2362.

[15] James Mayfield, Eugene Yang, Dawn Lawrie, Sean MacAvaney, Paul McNamee, Douglas W Oard, Luca Soldaini, Ian Soboroff, Orion Weller, Efsun Kayi, et al. 2024. On the Evaluation of Machine-Generated Reports. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1904–1915.

[16] Salar Mohtaj, Behnam Roshanfekr, Atefeh Zafarian, and Habibollah Asghari. 2018. Parsivar: A Language Processing Toolkit for Persian. In *International Conference on Language Resources and Evaluation.* https://api.semanticscholar.org/CorpusID:21715688

[17] Ronak Pradeep, Yuvan Sooryakumar, Zijian Chen, Antea Abilekaj, Eric Wang, Nathan Kuissi, Ryan Nguyen, Jie Min, Yidi Chen, and Jimmy Lin. 2025. $h_2$oloo @ TREC BioGen, Lateral Reading, NeuCLIR, RAG and ToT 2024: Weaving Dual Encoders, Rerankers, and RAG. In *The Thirty-Third Text REtrieval Conference (TREC 2024) Proceedings.*

[18] Shigehiko Schamoni, Felix Hieber, Artem Sokolov, and Stefan Riezler. 2014. Learning Translational and Knowledge-based Similarities from Relevance Rankings for Cross-Language Retrieval. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.* https://www.cl.uni-heidelberg.de/~riezler/publications/papers/ACL2014short.pdf

[19] Shuo Sun and Kevin Duh. 2020. CLIRMatrix: A massively large collection of bilingual and multilingual datasets for Cross-Lingual Information Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics, Online, 4160–4170. https://doi.org/10.18653/v1/2020.emnlp-main.340

[20] Jianqiang Wang and Douglas W. Oard. 2012. Matching Meaning for Cross-Language Information Retrieval. *Information Processing & Management* 48, 4 (2012), 631–653. https://doi.org/10.1016/j.ipm.2011.09.003

[21] Jinxi Xu and Ralph Weischedel. 2000. Cross-Lingual Information Retrieval using Hidden Markov Models. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.* 95–103.

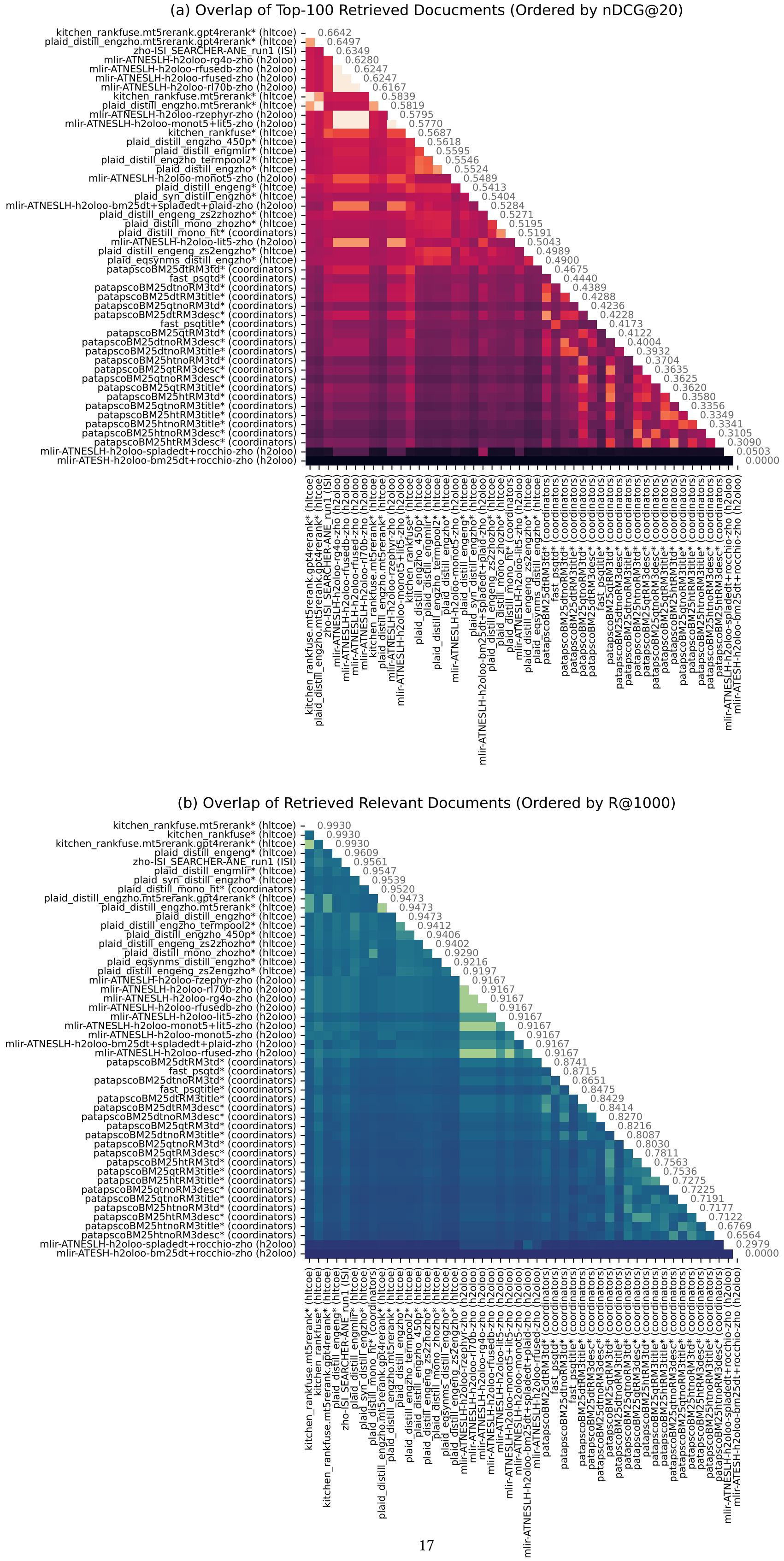
[22] Eugene Yang, Dawn Lawrie, Hoa Dang, Ian Soboroff, and James Mayfield. 2025. Nugget-based Annotation Protocol and Tool For Evaluating Long-form Retrieval-Augmented Generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*. https://doi.org/10.1145/3726302.3730156

[23] Eugene Yang, Dawn Lawrie, and James Mayfield. 2024. Distillation for Multilingual Information Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2368–2373.

[24] Eugene Yang, Dawn Lawrie, James Mayfield, Douglas W Oard, and Scott Miller. 2024. Translate-Distill: Learning Cross-Language Dense Retrieval by Translation and Distillation. In *Advances in Information Retrieval: 46th European Conference on IR Research, ECIR 2024*.

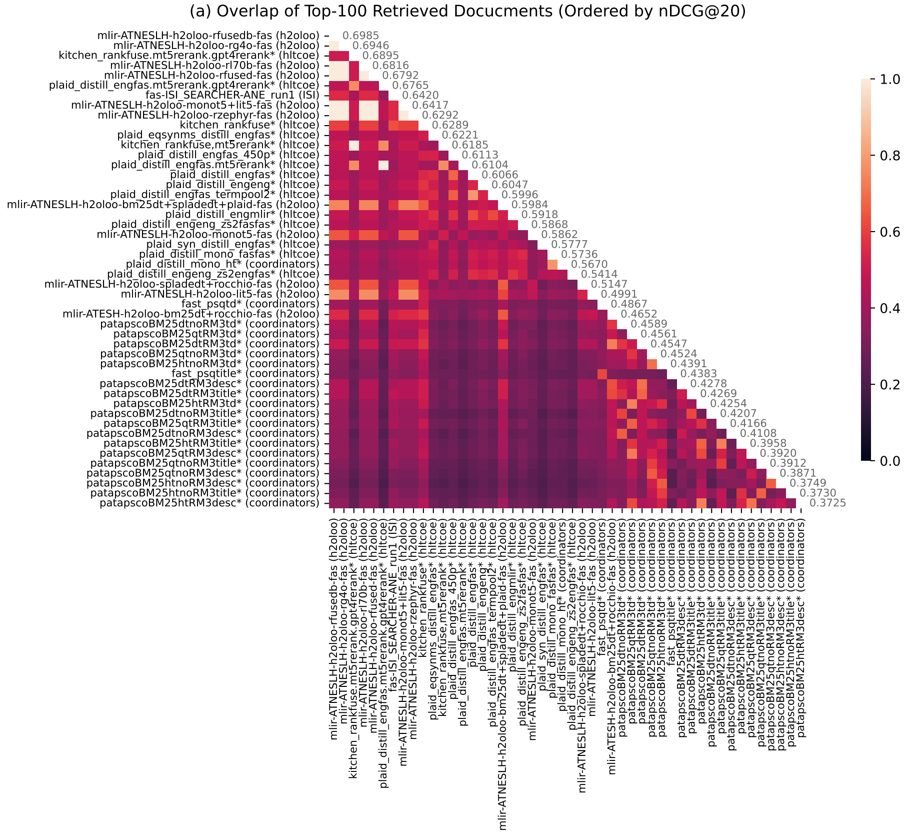
[25] Eugene Yang, Dawn Lawrie, Orion Weller, and James Mayfield. 2025. HLTCOE at TREC 2024 NeuCLIR Track. In *The Thirty-Third Text REtrieval Conference (TREC 2024) Proceedings*.

[26] Eugene Yang, Suraj Nair, Dawn Lawrie, James Mayfield, Douglas W Oard, and Kevin Duh. 2024. Efficiency-Effectiveness Tradeoff of Probabilistic Structured Queries for Cross-Language Information Retrieval. *arXiv preprint arXiv:2404.18797* (2024).

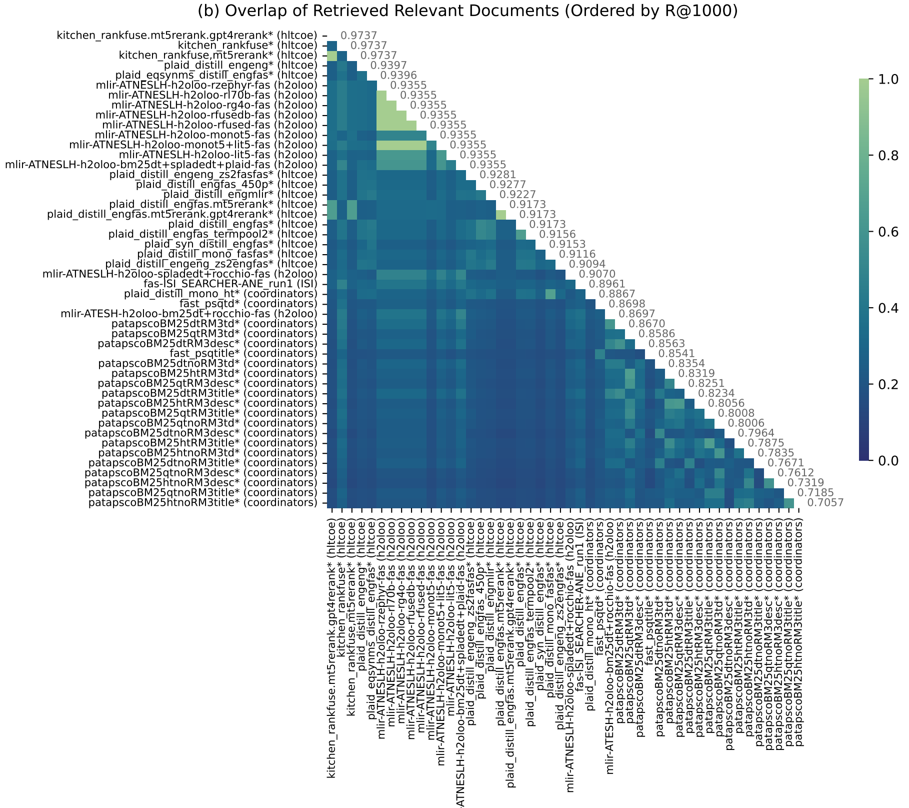

(a) Overlap of Top-100 Retrieved Docucments (Ordered by nDCG@20)

(b) Overlap of Retrieved Relevant Documents (Ordered by R@1000)

Figure 9: Overlap of documents retrieved by systems that participated in Chinese. * indicates manual runs.



(a) Overlap of Top-100 Retrieved Docucments (Ordered by nDCG@20)



(b) Overlap of Retrieved Relevant Documents (Ordered by R@1000)

Figure 10: Overlap of documents retrieved by systems that participated in Persian. * indicates manual runs.

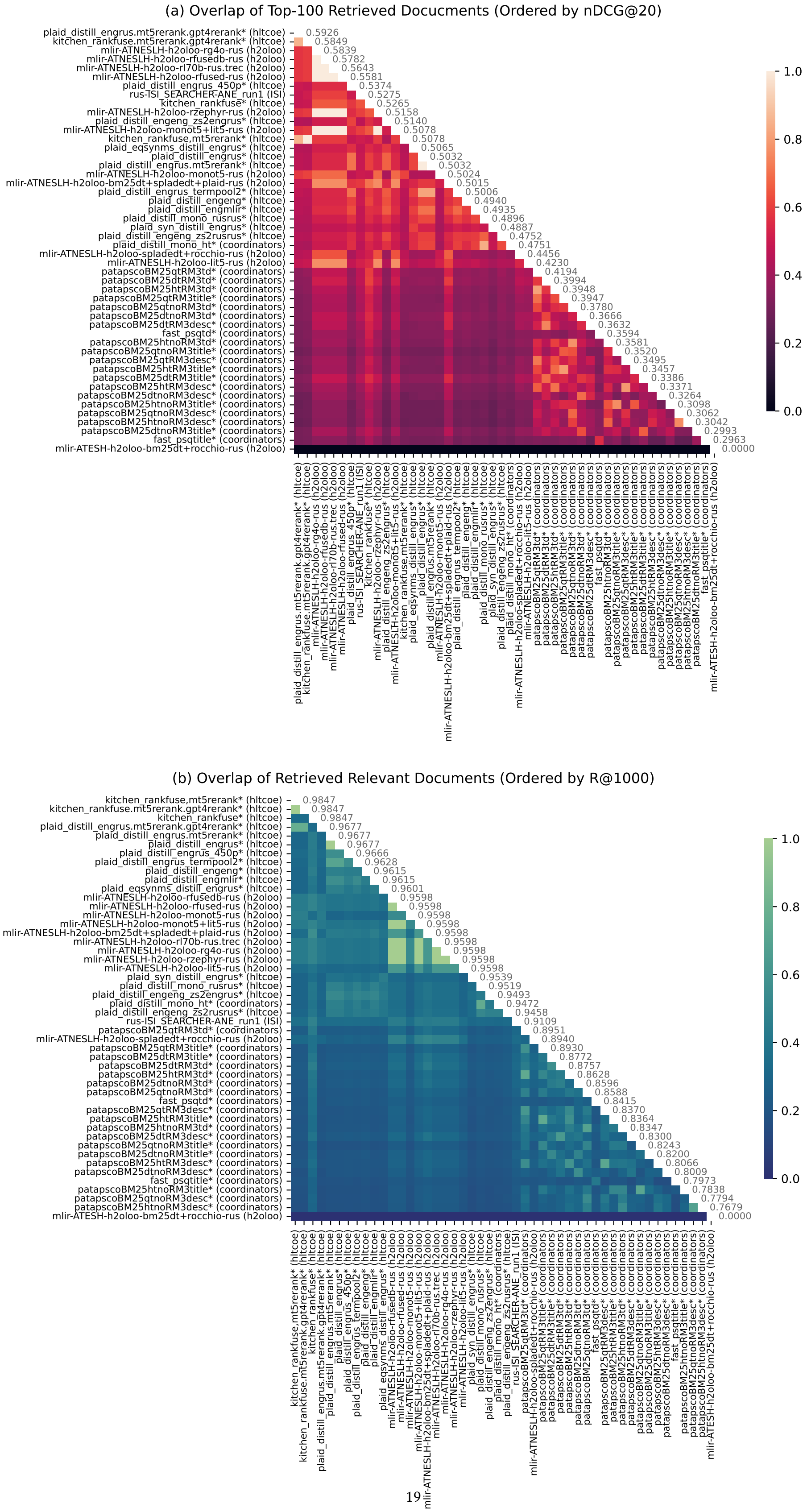


Figure 11: Overlap of documents retrieved by systems that participated in Russian. * indicates manual runs.

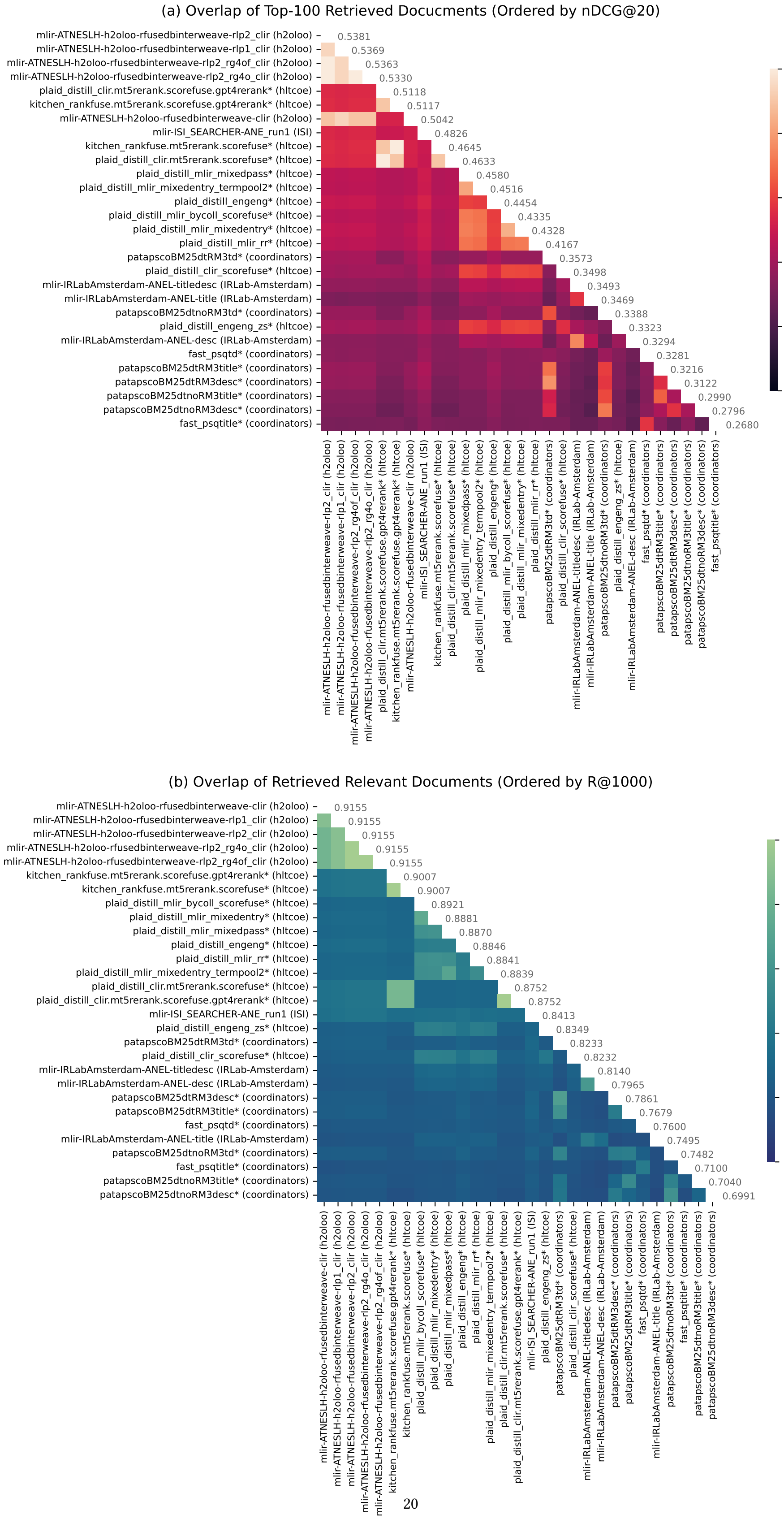
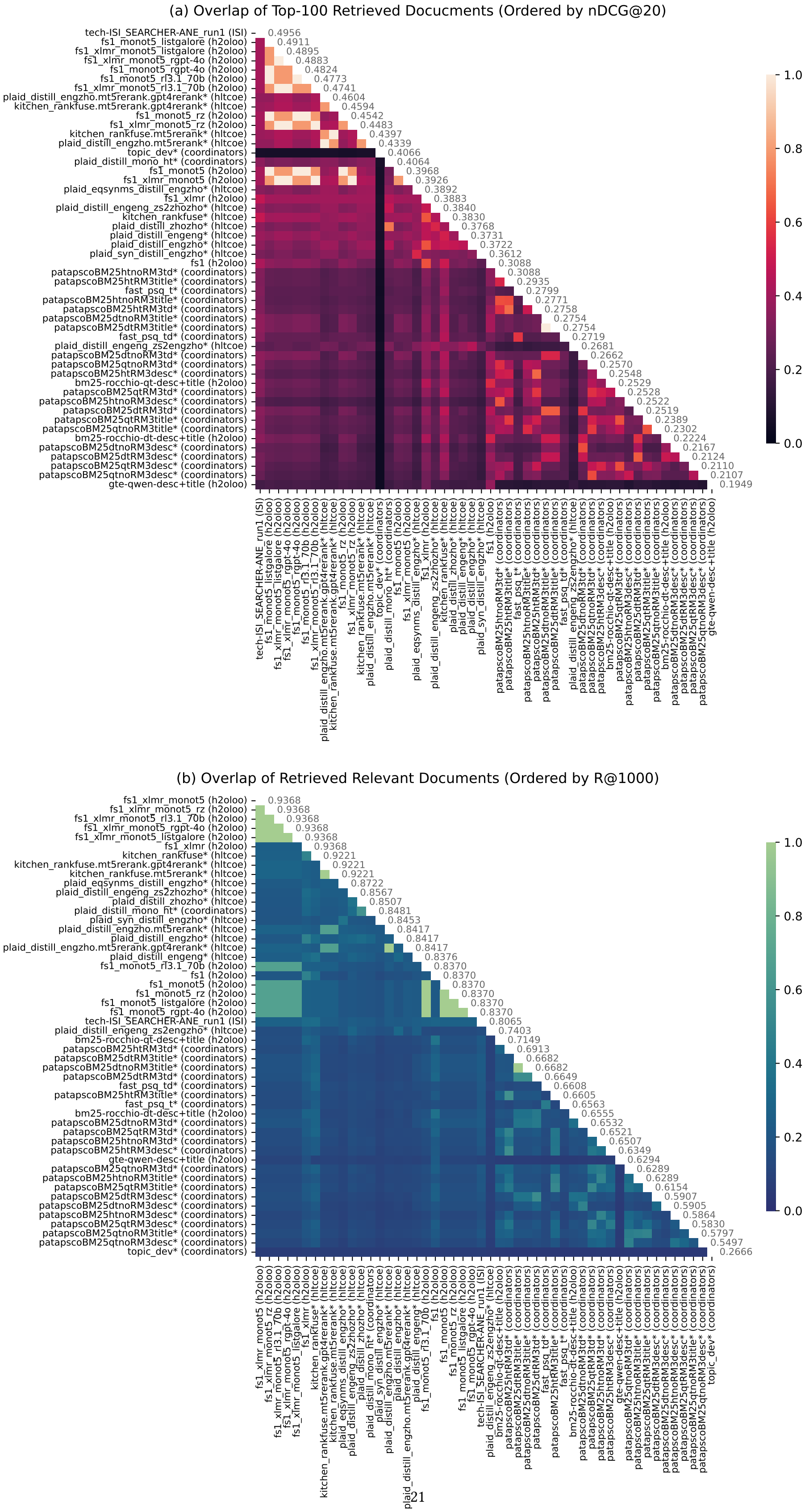


Figure 12: Overlap of documents retrieved by systems that participated in MLIR runs. * indicates manual runs.



(a) Overlap of Top-100 Retrieved Docucments (Ordered by nDCG@20)

(b) Overlap of Retrieved Relevant Documents (Ordered by R@1000)

Figure 13: Overlap of documents retrieved by systems that participated in the Technical Document Task. * indicates manual runs.

Table 8: Chinese Results. Monolingual runs, which use human translations of the queries, are shown in green. * indicates manual runs. Column "JFD" indicates whether the run is judged at full depth, which is 100. Other runs were judged to depth 50.

Team	Run Name	JFD	DS	QS	nDCG@20	RBP	AP	R@100	R@1k
hltcoe	kitchen_rankfuse.mt5rerank.gpt4rerank*	✓	Orig+DT	Orig+GT	0.664	0.524	0.578	0.836	0.993
hltcoe	plaid_distill_engzho.mt5rerank.gpt4rerank*	✓	Orig	Orig	0.650	0.503	0.541	0.808	0.947
ISI	zho-ISI_SEARCHER-ANE_run1	✓	Orig	Orig	0.635	0.480	0.540	0.881	0.956
h2oloo	mlir-ATNESLH-h2oloo-rg4o-zho	✓	Orig+DT	Orig+GT	0.628	0.487	0.537	0.814	0.917
h2oloo	mlir-ATNESLH-h2oloo-rfusedb-zho	✓	Orig+DT	Orig+GT	0.625	0.488	0.533	0.814	0.917
h2oloo	mlir-ATNESLH-h2oloo-rfused-zho	✓	Orig+DT	Orig+GT	0.625	0.483	0.530	0.814	0.917
h2oloo	mlir-ATNESLH-h2oloo-rl70b-zho	✓	Orig+DT	Orig+GT	0.617	0.482	0.526	0.814	0.917
hltcoe	kitchen_rankfuse.mt5rerank*	✓	Orig+DT	Orig+GT	0.584	0.450	0.502	0.836	0.993
hltcoe	plaid_distill_engzho.mt5rerank*	✓	Orig	Orig	0.582	0.445	0.485	0.808	0.947
h2oloo	mlir-ATNESLH-h2oloo-rzephyr-zho	✓	Orig+DT	Orig+GT	0.580	0.455	0.486	0.814	0.917
h2oloo	mlir-ATNESLH-h2oloo-monot5+lit5-zho	✓	Orig+DT	Orig+GT	0.577	0.455	0.488	0.814	0.917
hltcoe	kitchen_rankfuse*	✓	Orig+DT	Orig+GT	0.569	0.436	0.497	0.852	0.993
hltcoe	plaid_distill_engzho_450p*	✗	Orig	Orig	0.562	0.438	0.488	0.795	0.941
hltcoe	plaid_distill_engmlir*	✗	Orig	Orig	0.560	0.422	0.472	0.812	0.955
hltcoe	plaid_distill_engzho_termpool2*	✗	Orig	Orig	0.555	0.416	0.471	0.825	0.941
hltcoe	plaid_distill_engzho*	✓	Orig	Orig	0.552	0.421	0.472	0.811	0.947
h2oloo	mlir-ATNESLH-h2oloo-monot5-zho	✓	Orig+DT	Orig+GT	0.549	0.440	0.460	0.793	0.917
hltcoe	plaid_distill_engeng*	✗	DT	Orig	0.541	0.420	0.467	0.823	0.961
hltcoe	plaid_syn_distill_engzho*	✓	Orig	Orig	0.540	0.387	0.454	0.815	0.954
h2oloo	mlir-ATNESLH-h2oloo-bm25dt+spladedt+plaid-zho	✓	Orig+DT	Orig+GT	0.528	0.389	0.439	0.754	0.917
hltcoe	plaid_distill_engeng_zs2zhozho*	✗	Orig	GT	0.527	0.392	0.445	0.824	0.940
hltcoe	plaid_distill_mono_zhozho*	✗	Orig	GT	0.520	0.382	0.433	0.789	0.929
coordinators	plaid_distill_mono_ht*	✓	Orig	HT	0.519	0.402	0.452	0.831	0.952
h2oloo	mlir-ATNESLH-h2oloo-lit5-zho	✓	Orig+DT	Orig+GT	0.504	0.376	0.410	0.776	0.917
hltcoe	plaid_distill_engeng_zs2engzho*	✓	Orig	Orig	0.499	0.380	0.407	0.761	0.920
hltcoe	plaid_eqsynms_distill_engzho*	✓	Orig	Orig	0.490	0.377	0.426	0.759	0.922
coordinators	patapscoBM25dtRM3td*	✓	DT	Orig	0.468	0.368	0.395	0.688	0.874
coordinators	fast_psqtd*	✓	Orig	Orig	0.444	0.337	0.363	0.701	0.871
coordinators	patapscoBM25dtnoRM3td*	✓	DT	Orig	0.439	0.338	0.358	0.667	0.865
coordinators	patapscoBM25dtRM3title*	✗	DT	Orig	0.429	0.324	0.361	0.671	0.843
coordinators	patapscoBM25qtnoRM3td*	✓	Orig	GT	0.424	0.321	0.316	0.602	0.803
coordinators	patapscoBM25dtRM3desc*	✗	DT	Orig	0.423	0.327	0.348	0.648	0.841
coordinators	fast_psqtitle*	✗	Orig	Orig	0.417	0.314	0.331	0.677	0.848
coordinators	patapscoBM25qtRM3td*	✓	Orig	GT	0.412	0.333	0.348	0.661	0.822
coordinators	patapscoBM25dtnoRM3desc*	✗	DT	Orig	0.400	0.301	0.312	0.609	0.827
coordinators	patapscoBM25dtnoRM3title*	✗	DT	Orig	0.393	0.300	0.307	0.604	0.809
coordinators	patapscoBM25htnoRM3td*	✓	Orig	HT	0.370	0.290	0.289	0.524	0.718
coordinators	patapscoBM25qtRM3desc*	✗	Orig	GT	0.363	0.272	0.304	0.594	0.781
coordinators	patapscoBM25qtnoRM3desc*	✗	Orig	GT	0.363	0.257	0.274	0.545	0.723
coordinators	patapscoBM25qtRM3title*	✗	Orig	GT	0.362	0.275	0.298	0.571	0.754
coordinators	patapscoBM25htRM3td*	✓	Orig	HT	0.358	0.302	0.306	0.560	0.756
coordinators	patapscoBM25qtnoRM3title*	✗	Orig	GT	0.336	0.250	0.250	0.489	0.719
coordinators	patapscoBM25htRM3title*	✓	Orig	HT	0.335	0.279	0.287	0.541	0.728
coordinators	patapscoBM25htnoRM3title*	✗	Orig	HT	0.334	0.265	0.258	0.487	0.677
coordinators	patapscoBM25htnoRM3desc*	✗	Orig	HT	0.311	0.226	0.246	0.483	0.656
coordinators	patapscoBM25htRM3desc*	✗	Orig	HT	0.309	0.242	0.261	0.517	0.712
h2oloo	mlir-ATNESLH-h2oloo-spladedt+rocchio-zho	✗	Orig+DT	Orig+GT	0.050	0.037	0.036	0.112	0.298
h2oloo	mlir-ATESH-h2oloo-bm25dt+rocchio-zho	✗	Orig+DT	Orig+GT	0.000	0.000	0.000	0.000	0.000

Table 9: Persian Results. Monolingual runs, which use human translations of the queries, are marked as green. * indicates manual runs. Column "JFD" indicates whether the run is judged at full depth, which is 100. Other runs were judged to depth 50.

Team	Run Name	JFD	DS	QS	nDCG@20	RBP	AP	R@100	R@1k
h2oloo	mlir-ATNESLH-h2oloo-rfusedb-fas	✓	Orig+DT	Orig+GT	0.698	0.572	0.611	0.790	0.936
h2oloo	mlir-ATNESLH-h2oloo-rg4o-fas	✓	Orig+DT	Orig+GT	0.695	0.576	0.616	0.790	0.936
hltcoe	kitchen_rankfuse.mt5rerank.gpt4rerank*	✓	Orig+DT	Orig+GT	0.690	0.570	0.583	0.782	0.974
h2oloo	mlir-ATNESLH-h2oloo-rl70b-fas	✓	Orig+DT	Orig+GT	0.682	0.562	0.594	0.790	0.936
h2oloo	mlir-ATNESLH-h2oloo-rfused-fas	✓	Orig+DT	Orig+GT	0.679	0.569	0.599	0.790	0.936
hltcoe	plaid_distill_engfas.mt5rerank.gpt4rerank*	✓	Orig	Orig	0.676	0.569	0.572	0.766	0.917
ISI	fas-ISI_SEARCHER-ANE_run1	✓	Orig	Orig	0.642	0.517	0.550	0.781	0.896
h2oloo	mlir-ATNESLH-h2oloo-monot5+lit5-fas	✓	Orig+DT	Orig+GT	0.642	0.518	0.560	0.790	0.936
h2oloo	mlir-ATNESLH-h2oloo-rzephyr-fas	✓	Orig+DT	Orig+GT	0.629	0.544	0.557	0.790	0.936
hltcoe	kitchen_rankfuse*	✓	Orig+DT	Orig+GT	0.629	0.485	0.547	0.789	0.974
hltcoe	plaid_eqsynms_distill_engfas*	✓	Orig	Orig	0.622	0.472	0.552	0.791	0.940
hltcoe	kitchen_rankfuse,mt5rerank*	✓	Orig+DT	Orig+GT	0.619	0.500	0.526	0.782	0.974
hltcoe	plaid_distill_engfas_450p*	✗	Orig	Orig	0.611	0.480	0.543	0.774	0.928
hltcoe	plaid_distill_engfas.mt5rerank*	✓	Orig	Orig	0.610	0.494	0.516	0.766	0.917
hltcoe	plaid_distill_engfas*	✓	Orig	Orig	0.607	0.469	0.536	0.775	0.917
hltcoe	plaid_distill_engeng*	✗	DT	Orig	0.605	0.474	0.523	0.780	0.940
hltcoe	plaid_distill_engfas_termpool2*	✗	Orig	Orig	0.600	0.466	0.528	0.763	0.916
h2oloo	mlir-ATNESLH-h2oloo-bm25dt+spladedt+plaid-fas	✓	Orig+DT	Orig+GT	0.598	0.468	0.513	0.776	0.936
hltcoe	plaid_distill_engmlir*	✗	Orig	Orig	0.592	0.451	0.511	0.759	0.923
hltcoe	plaid_distill_engeng_zs2fasfas*	✗	Orig	GT	0.587	0.455	0.529	0.770	0.928
h2oloo	mlir-ATNESLH-h2oloo-monot5-fas	✓	Orig+DT	Orig+GT	0.586	0.490	0.508	0.778	0.936
hltcoe	plaid_syn_distill_engfas*	✓	Orig	Orig	0.578	0.444	0.493	0.758	0.915
hltcoe	plaid_distill_mono_fasfas*	✗	Orig	GT	0.574	0.453	0.513	0.747	0.912
coordinators	plaid_distill_mono_ht*	✓	Orig	HT	0.567	0.440	0.510	0.714	0.887
hltcoe	plaid_distill_engeng_zs2engfas*	✓	Orig	Orig	0.541	0.410	0.473	0.741	0.909
h2oloo	mlir-ATNESLH-h2oloo-spladedt+rocchio-fas	✗	Orig+DT	Orig+GT	0.515	0.407	0.428	0.705	0.907
h2oloo	mlir-ATNESLH-h2oloo-lit5-fas	✓	Orig+DT	Orig+GT	0.499	0.410	0.438	0.763	0.936
coordinators	fast_psqtd*	✓	Orig	Orig	0.487	0.376	0.419	0.718	0.870
h2oloo	mlir-ATESH-h2oloo-bm25dt+rocchio-fas	✗	Orig+DT	Orig+GT	0.465	0.385	0.392	0.663	0.870
coordinators	patapscoBM25dtnoRM3td*	✓	DT	Orig	0.459	0.362	0.365	0.645	0.835
coordinators	patapscoBM25qtRM3td*	✓	Orig	GT	0.456	0.385	0.369	0.659	0.859
coordinators	patapscoBM25dtRM3td*	✓	DT	Orig	0.455	0.384	0.384	0.663	0.867
coordinators	patapscoBM25qtnoRM3td*	✓	Orig	GT	0.452	0.380	0.355	0.638	0.801
coordinators	patapscoBM25htnoRM3td*	✓	Orig	HT	0.439	0.358	0.337	0.587	0.784
coordinators	fast_psqtitle*	✓	Orig	Orig	0.438	0.337	0.371	0.645	0.854
coordinators	patapscoBM25dtRM3desc*	✗	DT	Orig	0.428	0.349	0.350	0.663	0.856
coordinators	patapscoBM25dtRM3title*	✗	DT	Orig	0.427	0.337	0.347	0.624	0.823
coordinators	patapscoBM25htRM3td*	✓	Orig	HT	0.425	0.358	0.328	0.616	0.832
coordinators	patapscoBM25dtnoRM3title*	✗	DT	Orig	0.421	0.318	0.330	0.596	0.767
coordinators	patapscoBM25qtRM3title*	✗	Orig	GT	0.417	0.349	0.337	0.601	0.801
coordinators	patapscoBM25dtnoRM3desc*	✗	DT	Orig	0.411	0.325	0.329	0.624	0.796
coordinators	patapscoBM25htRM3title*	✓	Orig	HT	0.396	0.326	0.304	0.571	0.787
coordinators	patapscoBM25qtRM3desc*	✗	Orig	GT	0.392	0.344	0.313	0.626	0.825
coordinators	patapscoBM25qtnoRM3title*	✗	Orig	GT	0.391	0.327	0.303	0.574	0.718
coordinators	patapscoBM25qtnoRM3desc*	✗	Orig	GT	0.387	0.328	0.300	0.576	0.761
coordinators	patapscoBM25htnoRM3desc*	✗	Orig	HT	0.375	0.292	0.274	0.528	0.732
coordinators	patapscoBM25htnoRM3title*	✗	Orig	HT	0.373	0.312	0.287	0.534	0.706
coordinators	patapscoBM25htRM3desc*	✗	Orig	HT	0.372	0.309	0.285	0.573	0.806

Table 10: Russian Results. Monolingual runs, which use human translations of the queries, are marked as green. * indicates manual runs. Column "JFD" indicates whether the run is judged at full depth, which is 100. Other runs were judged to depth 50.

Team	Run Name	JFD	DS	QS	nDCG@20	RBP	AP	R@100	R@1k
hltcoe	plaid_distill_engrus.mt5rerank.gpt4rerank*	✓	Orig	Orig	0.593	0.553	0.532	0.784	0.968
hltcoe	kitchen_rankfuse.mt5rerank.gpt4rerank*	✓	Orig+DT	Orig+GT	0.585	0.544	0.519	0.779	0.985
h2oloo	mlir-ATNESLH-h2oloo-rg4o-rus	✓	Orig+DT	Orig+GT	0.584	0.521	0.516	0.820	0.960
h2oloo	mlir-ATNESLH-h2oloo-rfusedb-rus	✓	Orig+DT	Orig+GT	0.578	0.518	0.511	0.820	0.960
h2oloo	mlir-ATNESLH-h2oloo-rl70b-rus.trec	✓	Orig+DT	Orig+GT	0.564	0.512	0.494	0.820	0.960
h2oloo	mlir-ATNESLH-h2oloo-rfused-rus	✓	Orig+DT	Orig+GT	0.558	0.512	0.491	0.820	0.960
hltcoe	plaid_distill_engrus_450p*	✗	Orig	Orig	0.537	0.443	0.458	0.796	0.967
ISI	rus-ISI_SEARCHER-ANE_run1	✓	Orig	Orig	0.527	0.450	0.464	0.817	0.911
hltcoe	kitchen_rankfuse*	✓	Orig+DT	Orig+GT	0.526	0.434	0.447	0.803	0.985
h2oloo	mlir-ATNESLH-h2oloo-rzephyr-rus	✓	Orig+DT	Orig+GT	0.516	0.489	0.461	0.820	0.960
hltcoe	plaid_distill_engeng_zs2engrus*	✓	Orig	Orig	0.514	0.427	0.413	0.747	0.949
h2oloo	mlir-ATNESLH-h2oloo-monot5+lit5-rus	✓	Orig+DT	Orig+GT	0.508	0.475	0.450	0.820	0.960
hltcoe	kitchen_rankfuse,mt5rerank*	✓	Orig+DT	Orig+GT	0.508	0.458	0.457	0.779	0.985
hltcoe	plaid_eqsynms_distill_engrus*	✓	Orig	Orig	0.507	0.412	0.425	0.767	0.960
hltcoe	plaid_distill_engrus.mt5rerank*	✓	Orig	Orig	0.503	0.423	0.428	0.779	0.968
hltcoe	plaid_distill_engrus*	✓	Orig	Orig	0.503	0.423	0.428	0.779	0.968
h2oloo	mlir-ATNESLH-h2oloo-monot5-rus	✓	Orig+DT	Orig+GT	0.502	0.448	0.445	0.782	0.960
h2oloo	mlir-ATNESLH-h2oloo-bm25dt+spladedt+plaid-rus	✓	Orig+DT	Orig+GT	0.501	0.430	0.423	0.794	0.960
hltcoe	plaid_distill_engrus_termpool2*	✗	Orig	Orig	0.501	0.416	0.420	0.783	0.963
hltcoe	plaid_distill_engeng*	✗	DT	Orig	0.494	0.421	0.430	0.757	0.962
hltcoe	plaid_distill_engmlir*	✗	Orig	Orig	0.493	0.415	0.422	0.762	0.961
hltcoe	plaid_distill_mono_rusrus*	✗	Orig	GT	0.490	0.414	0.416	0.787	0.952
hltcoe	plaid_syn_distill_engrus*	✓	Orig	Orig	0.489	0.390	0.398	0.757	0.954
hltcoe	plaid_distill_engeng_zs2rusrus*	✗	Orig	GT	0.475	0.405	0.402	0.775	0.946
coordinators	plaid_distill_mono_ht*	✓	Orig	HT	0.475	0.403	0.396	0.762	0.947
h2oloo	mlir-ATNESLH-h2oloo-spladedt+rocchio-rus	✗	Orig+DT	Orig+GT	0.446	0.398	0.387	0.677	0.894
h2oloo	mlir-ATNESLH-h2oloo-lit5-rus	✓	Orig+DT	Orig+GT	0.423	0.420	0.375	0.763	0.960
coordinators	patapscoBM25qtRM3td*	✓	Orig	GT	0.419	0.377	0.338	0.670	0.895
coordinators	patapscoBM25dtRM3td*	✓	DT	Orig	0.399	0.353	0.322	0.634	0.876
coordinators	patapscoBM25htRM3td*	✓	Orig	HT	0.395	0.349	0.324	0.630	0.863
coordinators	patapscoBM25qtRM3title*	✗	Orig	GT	0.395	0.345	0.309	0.651	0.893
coordinators	patapscoBM25qtnoRM3td*	✓	Orig	GT	0.378	0.340	0.305	0.635	0.859
coordinators	patapscoBM25dtnoRM3td*	✓	DT	Orig	0.367	0.311	0.283	0.642	0.860
coordinators	patapscoBM25dtRM3desc*	✗	DT	Orig	0.363	0.319	0.291	0.578	0.830
coordinators	fast_psqtd*	✓	Orig	Orig	0.359	0.306	0.289	0.626	0.842
coordinators	patapscoBM25htnoRM3td*	✓	Orig	HT	0.358	0.315	0.283	0.601	0.835
coordinators	patapscoBM25qtnoRM3title*	✗	Orig	GT	0.352	0.317	0.277	0.588	0.824
coordinators	patapscoBM25qtRM3desc*	✗	Orig	GT	0.350	0.294	0.284	0.581	0.837
coordinators	patapscoBM25htRM3title*	✓	Orig	HT	0.346	0.299	0.280	0.588	0.836
coordinators	patapscoBM25dtRM3title*	✗	DT	Orig	0.339	0.297	0.274	0.620	0.877
coordinators	patapscoBM25htRM3desc*	✗	Orig	HT	0.337	0.292	0.280	0.554	0.807
coordinators	patapscoBM25dtnoRM3desc*	✗	DT	Orig	0.326	0.273	0.232	0.566	0.801
coordinators	patapscoBM25htnoRM3title*	✗	Orig	HT	0.310	0.284	0.253	0.535	0.784
coordinators	patapscoBM25qtnoRM3desc*	✗	Orig	GT	0.306	0.279	0.244	0.542	0.779
coordinators	patapscoBM25htnoRM3desc*	✗	Orig	HT	0.304	0.270	0.244	0.543	0.768
coordinators	patapscoBM25dtnoRM3title*	✗	DT	Orig	0.299	0.272	0.235	0.579	0.820
coordinators	fast_psqtitle*	✗	Orig	Orig	0.296	0.247	0.235	0.539	0.797
h2oloo	mlir-ATESH-h2oloo-bm25dt+rocchio-rus	✗	Orig+DT	Orig+GT	0.000	0.000	0.000	0.000	0.000

Table 11: MLIR Results. The run used as the first stage retrieval for the reranking task is marked in bold. * indicates manual runs. Results may change after TREC conference. Column “JFD” indicates whether the run is judged at full depth, which is 100, otherwise 50.

Team	Run Name	JFD	DS	QS	nDCG@20	RBP	AP	R@100	R@1k
h2oloo	mlir-ATNESLH-h2oloo-rfusedbinterweave-rlp2_clir	✓	Orig+DT	Orig+GT	0.545	0.590	0.460	0.702	0.917
h2oloo	mlir-ATNESLH-h2oloo-rfusedbinterweave-rlp1_clir	✓	Orig+DT	Orig+GT	0.544	0.595	0.450	0.694	0.917
h2oloo	mlir-ATNESLH-h2oloo-rfusedbinterweave-rlp2_rg4of_clir	✓	Orig+DT	Orig+GT	0.544	0.603	0.467	0.702	0.917
h2oloo	mlir-ATNESLH-h2oloo-rfusedbinterweave-rlp2_rg4o_clir	✓	Orig+DT	Orig+GT	0.540	0.604	0.472	0.702	0.917
hltcoe	kitchen_rankfuse.mt5rerank.scorefuse.gpt4rerank*	✓	Orig+DT	Orig+GT	0.521	0.630	0.457	0.663	0.903
hltcoe	plaid_distill_clir.mt5rerank.scorefuse.gpt4rerank*	✓	Orig	Orig	0.520	0.625	0.442	0.656	0.878
h2oloo	mlir-ATNESLH-h2oloo-rfusedbinterweave-clir	✓	Orig+DT	Orig+GT	0.511	0.564	0.414	0.678	0.917
ISI	mlir-ISI_SEARCHER-ANE_run1	✓	Orig	Orig	0.490	0.568	0.468	0.736	0.844
hltcoe	kitchen_rankfuse.mt5rerank.scorefuse*	✗	Orig+DT	Orig+GT	0.475	0.562	0.438	0.663	0.903
hltcoe	plaid_distill_clir.mt5rerank.scorefuse*	✗	Orig	Orig	0.474	0.560	0.431	0.656	0.878
hltcoe	plaid_distill_mlir_mixedpass*	✓	Orig	Orig	0.468	0.502	0.416	0.656	0.889
hltcoe	plaid_distill_mlir_mixedentry_termpool2*	✓	Orig	Orig	0.462	0.510	0.420	0.652	0.886
hltcoe	plaid_distill_engeng*	✓	DT	Orig	0.453	0.524	0.419	0.678	0.887
hltcoe	plaid_distill_mlir_mixedentry*	✓	Orig	Orig	0.444	0.494	0.402	0.654	0.890
hltcoe	plaid_distill_mlir_bycoll_scorefuse*	✓	Orig	Orig	0.438	0.493	0.389	0.639	0.894
hltcoe	plaid_distill_mlir_rr*	✓	Orig	Orig	0.428	0.479	0.383	0.642	0.886
IRLab-Amsterdam	mlir-IRLabAmsterdam-ANEL-titledesc	✓	Orig	Orig	0.354	0.415	0.303	0.554	0.818
hltcoe	plaid_distill_clir_scorefuse*	✗	Orig	Orig	0.353	0.431	0.315	0.577	0.827
IRLab-Amsterdam	mlir-IRLabAmsterdam-ANEL-title	✓	Orig	Orig	0.352	0.414	0.285	0.485	0.754
coordinators	patapscoBM25dtRM3td*	✓	DT	Orig	0.350	0.431	0.283	0.507	0.827
hltcoe	plaid_distill_engeng_zs*	✓	Orig	Orig	0.343	0.392	0.304	0.568	0.838
coordinators	patapscoBM25dtnoRM3td*	✓	DT	Orig	0.337	0.402	0.239	0.500	0.753
IRLab-Amsterdam	mlir-IRLabAmsterdam-ANEL-desc	✓	Orig	Orig	0.335	0.388	0.276	0.516	0.800
coordinators	fast_psqtd*	✓	Orig	Orig	0.322	0.401	0.273	0.516	0.765
coordinators	patapscoBM25dtRM3title*	✓	DT	Orig	0.315	0.389	0.253	0.466	0.772
coordinators	patapscoBM25dtRM3desc*	✓	DT	Orig	0.306	0.374	0.243	0.458	0.790
coordinators	patapscoBM25dtnoRM3title*	✓	DT	Orig	0.298	0.354	0.213	0.439	0.710
coordinators	patapscoBM25dtnoRM3desc*	✓	DT	Orig	0.279	0.339	0.195	0.458	0.705
coordinators	fast_psqtitle*	✓	Orig	Orig	0.263	0.307	0.205	0.455	0.716

Table 12: Technical Document Task Results. Monolingual runs, which use human translations of the queries, are shown in green. * indicates manual runs. All runs are judged to depth 40.

Team	Run Name	DS	QS	nDCG@20	RBP	AP	R@100	R@1k
ISI	tech-ISI_SEARCHER-ANE_run1	Orig	Orig	0.496	0.468	0.350	0.638	0.807
h2oloo	fs1_monot5_listgalore	Orig+DT	Orig+GT	0.491	0.481	0.355	0.573	0.837
h2oloo	fs1_xlmr_monot5_listgalore	Orig+DT	Orig+GT	0.489	0.485	0.370	0.592	0.937
h2oloo	fs1_xlmr_monot5_rgpt-4o	Orig+DT	Orig+GT	0.488	0.485	0.372	0.592	0.937
h2oloo	fs1_monot5_rgpt-4o	Orig+DT	Orig+GT	0.482	0.473	0.353	0.573	0.837
h2oloo	fs1_monot5_rl3.1_70b	Orig+DT	Orig+GT	0.477	0.473	0.345	0.573	0.837
h2oloo	fs1_xlmr_monot5_rl3.1_70b	Orig+DT	Orig+GT	0.474	0.467	0.353	0.592	0.937
hltcoe	plaid_distill_engzho.mt5rerank.gpt4rerank*	Orig	Orig	0.460	0.469	0.340	0.614	0.842
hltcoe	kitchen_rankfuse.mt5rerank.gpt4rerank*	Orig+DT	Orig+GT	0.459	0.464	0.340	0.644	0.922
h2oloo	fs1_monot5_rz	Orig+DT	Orig+GT	0.454	0.440	0.318	0.573	0.837
h2oloo	fs1_xlmr_monot5_rz	Orig+DT	Orig+GT	0.448	0.449	0.328	0.592	0.937
hltcoe	kitchen_rankfuse.mt5rerank*	Orig+DT	Orig+GT	0.440	0.431	0.327	0.644	0.922
hltcoe	plaid_distill_engzho.mt5rerank*	Orig	Orig	0.434	0.430	0.320	0.614	0.842
coordinators	topic_dev*	Orig+DT	Other	0.407	0.462	0.227	0.267	0.267
coordinators	plaid_distill_mono_ht*	Orig	HT	0.406	0.403	0.284	0.593	0.848
h2oloo	fs1_monot5	Orig+DT	Orig+GT	0.397	0.379	0.276	0.573	0.837
h2oloo	fs1_xlmr_monot5	Orig+DT	Orig+GT	0.393	0.383	0.287	0.592	0.937
hltcoe	plaid_eqsynms_distill_engzho*	Orig	Orig	0.389	0.389	0.291	0.578	0.872
h2oloo	fs1_xlmr	Orig+DT	Orig+GT	0.388	0.391	0.285	0.628	0.937
hltcoe	plaid_distill_engeng_zs2zhozho*	Orig	GT	0.384	0.369	0.267	0.568	0.857
hltcoe	kitchen_rankfuse*	Orig+DT	Orig+GT	0.383	0.365	0.269	0.580	0.922
hltcoe	plaid_distill_zhozho*	Orig	GT	0.377	0.373	0.265	0.558	0.851
hltcoe	plaid_distill_engeng*	DT	Orig	0.373	0.376	0.264	0.534	0.838
hltcoe	plaid_distill_engzho*	Orig	Orig	0.372	0.367	0.268	0.563	0.842
hltcoe	plaid_syn_distill_engzho*	Orig	Orig	0.361	0.361	0.264	0.541	0.845
h2oloo	fs1	Orig+DT	Orig+GT	0.309	0.305	0.216	0.531	0.837
coordinators	patapscoBM25htnoRM3td*	Orig	HT	0.309	0.307	0.187	0.418	0.651
coordinators	patapscoBM25htRM3title*	Orig	HT	0.293	0.291	0.197	0.418	0.660
coordinators	fast_psq_t*	Orig	Orig	0.280	0.274	0.164	0.365	0.656
coordinators	patapscoBM25htnoRM3title*	Orig	HT	0.277	0.283	0.175	0.388	0.629
coordinators	patapscoBM25htRM3td*	Orig	HT	0.276	0.280	0.178	0.397	0.691
coordinators	patapscoBM25dtnoRM3title*	DT	Orig	0.275	0.285	0.191	0.426	0.668
coordinators	patapscoBM25dtRM3title*	DT	Orig	0.275	0.285	0.191	0.426	0.668
coordinators	fast_psq_td*	Orig	Orig	0.272	0.267	0.163	0.387	0.661
hltcoe	plaid_distill_engeng_zs2engzho*	Orig	Orig	0.268	0.272	0.191	0.455	0.740
coordinators	patapscoBM25dtnoRM3td*	DT	Orig	0.266	0.268	0.166	0.397	0.653
coordinators	patapscoBM25qtnoRM3td*	Orig	GT	0.257	0.253	0.151	0.380	0.629
coordinators	patapscoBM25htRM3desc*	Orig	HT	0.255	0.267	0.164	0.372	0.635
h2oloo	bm25-rocchio-qt-desc+title	Orig	GT	0.253	0.248	0.175	0.398	0.715
coordinators	patapscoBM25qtRM3td*	Orig	GT	0.253	0.259	0.160	0.355	0.652
coordinators	patapscoBM25htnoRM3desc*	Orig	HT	0.252	0.260	0.150	0.366	0.586
coordinators	patapscoBM25dtRM3td*	DT	Orig	0.252	0.245	0.165	0.402	0.665
coordinators	patapscoBM25qtRM3title*	Orig	GT	0.239	0.250	0.155	0.354	0.615
coordinators	patapscoBM25qtnoRM3title*	Orig	GT	0.230	0.240	0.142	0.355	0.580
h2oloo	bm25-rocchio-dt-desc+title	DT	Orig	0.222	0.222	0.146	0.378	0.656
coordinators	patapscoBM25dtnoRM3desc*	DT	Orig	0.217	0.223	0.131	0.340	0.590
coordinators	patapscoBM25dtRM3desc*	DT	Orig	0.212	0.215	0.148	0.365	0.591
coordinators	patapscoBM25qtRM3desc*	Orig	GT	0.211	0.214	0.128	0.315	0.583
coordinators	patapscoBM25qtnoRM3desc*	Orig	GT	0.211	0.212	0.116	0.311	0.550
h2oloo	gte-qwen-desc+title	Orig	GT	0.195	0.193	0.122	0.338	0.629

Table 13: Chinese Report Generation Task Scores. Values in parentheses are standard deviations across topics.

Team	Run ID	ARGUE Score	Citation Precision	Nugget Recall	Nugget Support	Sentence Support
hltcoe	zho-hltcoe-eugene-gpt4o-fixed	0.726 (0.263)	0.859 (0.266)	0.327 (0.219)	0.298 (0.169)	0.840 (0.136)
IDA	IDA_CCS_hybrid_zho	0.637 (0.310)	0.795 (0.261)	0.236 (0.164)	0.166 (0.120)	0.803 (0.234)
hltcoe	zho-hltcoe-eugene-gpt35turbo	0.587 (0.284)	0.813 (0.291)	0.250 (0.160)	0.248 (0.171)	0.720 (0.190)
hltcoe	zho-jhu-orion-aggregated-w-claude	0.528 (0.251)	0.900 (0.249)	0.177 (0.123)	0.238 (0.201)	0.662 (0.200)
hltcoe	zho-jhu-orion-aggregated-w-gpt4o	0.496 (0.277)	0.899 (0.234)	0.182 (0.145)	0.237 (0.184)	0.636 (0.226)
irlab-ams	zho_irlab-ams-std-translate-llama-70B-api	0.464 (0.295)	0.927 (0.174)	0.181 (0.123)	0.243 (0.150)	0.574 (0.227)
IDA	IDA_CCS_abstractive_zho	0.456 (0.226)	0.873 (0.187)	0.230 (0.144)	0.184 (0.132)	0.601 (0.206)
h2oloo	rfused_rgn_crp_zho	0.432 (0.206)	0.895 (0.190)	0.261 (0.138)	0.292 (0.182)	0.555 (0.152)
h2oloo	rfused_rgn_l70b_zho	0.376 (0.246)	0.850 (0.246)	0.236 (0.187)	0.249 (0.172)	0.433 (0.223)
irlab-ams	zho_irlab-ams-std-translate-llama-8B	0.360 (0.171)	0.861 (0.196)	0.209 (0.157)	0.164 (0.112)	0.563 (0.193)
h2oloo	rfused_rgn_gpt4o_zho	0.348 (0.181)	0.894 (0.216)	0.246 (0.150)	0.171 (0.132)	0.594 (0.180)
h2oloo	rfused_rgn_l70bph_zho	0.346 (0.212)	0.894 (0.233)	0.260 (0.185)	0.286 (0.181)	0.399 (0.225)
irlab-ams	zho_irlab-ams-std-recomp-llama-8B	0.277 (0.177)	0.777 (0.227)	0.158 (0.119)	0.164 (0.174)	0.384 (0.181)
irlab-ams	zho_irlab-ams-postcite	0.181 (0.241)	0.546 (0.297)	0.102 (0.137)	0.135 (0.187)	0.229 (0.235)
irlab-ams	zho_irlab-ams-std-mdcomp-330-translate-llama-8B	0.166 (0.138)	0.759 (0.271)	0.135 (0.113)	0.104 (0.086)	0.269 (0.174)
irlab-ams	zho_irlab-ams-std-mdcomp-331-translate-llama-8B	0.144 (0.108)	0.744 (0.270)	0.151 (0.103)	0.089 (0.061)	0.236 (0.145)
irlab-ams	zho_irlab-ams-postcite-v	0.121 (0.140)	0.423 (0.293)	0.105 (0.131)	0.115 (0.143)	0.156 (0.159)

Table 14: Persian Report Generation Task Scores. Values in parentheses are standard deviations across topics.

Team	Run ID	ARGUE Score	Citation Precision	Nugget Recall	Nugget Support	Sentence Support
hltcoe	fas-hltcoe-eugene-gpt4o	0.872 (0.078)	0.918 (0.115)	0.303 (0.167)	0.311 (0.134)	0.919 (0.061)
IDA	IDA_CCS_hybrid_fas	0.681 (0.320)	0.853 (0.149)	0.271 (0.197)	0.183 (0.139)	0.845 (0.193)
hltcoe	fas-hltcoe-eugene-gpt35turbo	0.646 (0.290)	0.846 (0.215)	0.215 (0.162)	0.201 (0.136)	0.792 (0.202)
irlab-ams	fas_irlab-ams-std-translate-llama-70B-api	0.566 (0.261)	0.852 (0.211)	0.197 (0.127)	0.272 (0.122)	0.710 (0.270)
IDA	IDA_CCS_abstractive_fas	0.525 (0.271)	0.940 (0.106)	0.282 (0.182)	0.215 (0.140)	0.621 (0.252)
hltcoe	fas-jhu-orion-aggregated-w-claude	0.519 (0.173)	0.945 (0.121)	0.213 (0.114)	0.268 (0.156)	0.650 (0.174)
h2oloo	rfused_rgn_l70b_fas	0.463 (0.234)	0.931 (0.097)	0.231 (0.163)	0.270 (0.162)	0.551 (0.222)
hltcoe	fas-jhu-orion-aggregated-w-gpt4o	0.449 (0.265)	0.941 (0.089)	0.220 (0.165)	0.272 (0.182)	0.565 (0.243)
h2oloo	rfused_rgn_crp_fas	0.431 (0.224)	0.889 (0.169)	0.295 (0.197)	0.253 (0.156)	0.586 (0.169)
h2oloo	rfused_rgn_l70bph_fas	0.420 (0.224)	0.940 (0.129)	0.240 (0.147)	0.328 (0.208)	0.504 (0.185)
irlab-ams	fas_irlab-ams-std-translate-llama-8B	0.337 (0.219)	0.795 (0.181)	0.182 (0.130)	0.183 (0.135)	0.553 (0.246)
h2oloo	rfused_rgn_gpt4o_fas	0.304 (0.162)	0.934 (0.153)	0.230 (0.149)	0.189 (0.119)	0.572 (0.207)
irlab-ams	fas_irlab-ams-std-recomp-llama-8B	0.292 (0.220)	0.781 (0.216)	0.133 (0.137)	0.155 (0.165)	0.412 (0.225)
irlab-ams	fas_irlab-ams-postcite	0.188 (0.189)	0.452 (0.253)	0.108 (0.167)	0.105 (0.129)	0.266 (0.238)
irlab-ams	fas_irlab-ams-std-mdcomp-330-translate-llama-8B	0.179 (0.105)	0.703 (0.244)	0.171 (0.173)	0.118 (0.083)	0.290 (0.101)
irlab-ams	fas_irlab-ams-std-mdcomp-331-translate-llama-8B	0.159 (0.159)	0.665 (0.282)	0.150 (0.138)	0.107 (0.116)	0.257 (0.181)
irlab-ams	fas_irlab-ams-postcite-v	0.087 (0.114)	0.389 (0.262)	0.058 (0.083)	0.058 (0.086)	0.147 (0.147)

Table 15: Russian Report Generation Task Scores. Values in parentheses are standard deviations across topics.

Team	Run ID	ARGUE Score	Citation Precision	Nugget Recall	Nugget Support	Sentence Support
hltcoe	rus-hltcoe-eugene-gpt4o	0.808 (0.208)	0.904 (0.133)	0.339 (0.167)	0.420 (0.192)	0.874 (0.139)
IDA	IDA_CCS_hybrid_rus	0.615 (0.360)	0.768 (0.288)	0.296 (0.221)	0.235 (0.150)	0.799 (0.212)
hltcoe	rus-hltcoe-eugene-gpt35turbo	0.602 (0.305)	0.799 (0.293)	0.313 (0.207)	0.309 (0.198)	0.715 (0.242)
hltcoe	rus-jhu-orion-aggregated-w-claude	0.519 (0.284)	0.902 (0.208)	0.255 (0.161)	0.323 (0.189)	0.673 (0.274)
irlab-ams	rus_irlab-ams-std-translate-llama-70B-api	0.472 (0.229)	0.871 (0.209)	0.255 (0.169)	0.393 (0.243)	0.566 (0.231)
h2oloo	rfused_rgn_l70b_rus	0.469 (0.246)	0.903 (0.198)	0.278 (0.203)	0.319 (0.212)	0.527 (0.270)
hltcoe	rus-jhu-orion-aggregated-w-gpt4o	0.415 (0.267)	0.904 (0.198)	0.247 (0.165)	0.287 (0.187)	0.495 (0.278)
h2oloo	rfused_rgn_crp_rus	0.414 (0.223)	0.914 (0.185)	0.293 (0.209)	0.304 (0.193)	0.498 (0.220)
IDA	IDA_CCS_abstractive_rus	0.403 (0.296)	0.851 (0.241)	0.298 (0.227)	0.235 (0.197)	0.486 (0.297)
h2oloo	rfused_rgn_l70bph_rus	0.403 (0.202)	0.894 (0.210)	0.298 (0.193)	0.363 (0.237)	0.469 (0.220)
h2oloo	rfused_rgn_gpt4o_rus	0.309 (0.174)	0.927 (0.139)	0.284 (0.170)	0.203 (0.155)	0.536 (0.238)
irlab-ams	rus_irlab-ams-std-translate-llama-8B	0.265 (0.190)	0.889 (0.178)	0.216 (0.162)	0.188 (0.122)	0.436 (0.221)
irlab-ams	rus_irlab-ams-std-recomp-llama-8B	0.227 (0.194)	0.719 (0.261)	0.152 (0.141)	0.165 (0.137)	0.380 (0.227)
irlab-ams	rus_irlab-ams-std-mdcomp-330-translate-llama-8B	0.185 (0.149)	0.737 (0.264)	0.157 (0.147)	0.164 (0.136)	0.264 (0.168)
irlab-ams	rus_irlab-ams-std-mdcomp-331-translate-llama-8B	0.185 (0.134)	0.739 (0.269)	0.168 (0.111)	0.153 (0.115)	0.292 (0.145)
irlab-ams	rus_irlab-ams-postcite	0.126 (0.175)	0.480 (0.230)	0.085 (0.136)	0.084 (0.118)	0.216 (0.200)
irlab-ams	rus_irlab-ams-postcite-v	0.074 (0.122)	0.439 (0.245)	0.050 (0.091)	0.070 (0.095)	0.091 (0.120)