NeuCLIRTech: Chinese Monolingual and Cross-Language Information Retrieval Evaluation in a Challenging Domain

Anonymous ACL submission

Abstract

Applying deep research agents to scientific research has been one of the most promising yet challenging tasks in retrieval-augmented generation. While large language models (LLMs) have substantially improved due to pretraining on more scientific literature, search engines still rely heavily on lexical matches because of the complex terminologies. Recent retrieval benchmarks such as BRIGHT have started exploring these challenging domains to evaluate reasoning models, but are limited to coding or mathematics problems. Furthermore, these benchmarks only consider English, while realistic deep research agents are expected to gather literature from any language regardless of the one the user used to express the input. In this work, we introduce NeuCLIRTech, a retrieval evaluation benchmark that supports both Chinese and English-to-Chinese literature search on a wide range of scientific domains. 110 queries were created by Chinese native speakers with at least a master's degree in each discipline. More than 35,000 labels from these human experts support reproducibility and deep analysis of recall. NeuCLIRTech is the first retrieval evaluation benchmark for technical cross-language retrieval with deep human judgments. All data are publicly available on Huggingface Datasets (released upon publication).

1 Introduction

Scientific discovery with deep research agents is an emerging application of retrieval-augmented generation (Lála et al., 2023; Ding et al., 2025). Scientific discovery systems require a retrieval model capable of retrieving useful literature from a collection of scientific papers based on the user's input. Retrieval accuracy is a strong indicator of system effectiveness because retrieval failure results in insufficient source information for the downstream generation model (Samuel et al., 2026). Therefore, incorporating effective scientific retrieval models as part of the agent is critical.

Recent retrieval benchmarks, like BRIGHT (Su et al., 2024), incorporate technical documents as retrieval candidates for reasoning-intensive tasks. However, these domains are limited to coding or mathematics problems (Su et al., 2024) inherited from natural language understanding tasks such as GSM8K (Cobbe et al., 2021) and LiveCodeBench (Jain et al., 2024). Prior works in patent search also required retrieval models to identify technical documents, but queries are usually broad because users seek prior art or similar implementations of a given patent application (Montecchi et al., 2013). This work addresses the need to identify specific technical information.

Prior work in technical document retrieval primarily focuses on English (Ajith et al., 2024; Cohan et al., 2020), ignoring deep research agents that gather information in different languages, or users less competent in English. This work introduces NeuCLIRTech, a benchmark for evaluating retrieval effectiveness over technical documents across languages. Specifically, NeuCLIRTech consists of Chinese technical abstracts with queries developed by Chinese-native graduate students in the United States with at least a master's degree in one of the disciplines in the document collection.

This work presents our novel approach to developing a large, deeply judged retrieval benchmark over technical documents. Our approach overcomes the challenge of building a reusable retrieval benchmark without the most effective retrieval models available to create the judgment pools. When evaluating state-of-the-art multilingual retrievers and rerankers, we show that NeuCLIRTech provides a challenging task and is suitable to be used to evaluate systems developed after the benchmark was created.

2 Related Work

Benchmarks for Cranfield-style evaluation of information retrieval effectiveness have been avail-

Key	Original	Google Translation
doc_id	csl-121374	csl-121374
Title	(S)-2-氨基-1,1-二苯基-1-丙醇的合成和外消旋Droxidopa前体的拆分	(S) -2-amino-1,1-two phenyl-1-propane synthesis and external anti-rotation DroxIDOPA foretophylline is split
Abstract	(S)-2-氨基-1,1-二苯基-1-丙醇是一种合成多种手性助剂的重要中间体,也用于外消旋屈昔多巴前体化合物的拆分.从价廉易得的L-丙氨酸出发,通过四步反应制得,总收率55.6%L-丙氨酸经甲酯化,苄氧羰基保护制得的L-2-苄羰基氨基丙酸甲酯与苯基溴化镁反应制得(S)-2-苄氧羰基氨基-1,1-二苯基-1-丙醇.是在5%Pd/C催化加氢下脱除苄氧羰基得到标题化合物.该制备方法涉及的中间体及目标化合物易于纯化,总收率高且重现性好.我们用制得的氨基醇能成功地拆分外消旋苏式屈昔多巴前体化合物3-3,4-二苄氧苯基)-N-苄氧羰基丙氨酸.	(S) -2-amino-1,1-two phenyl-1-propionol is an important intermediate of a synthetic multi-hand assistant. It is also used for demolition Divided. From the price of low-priced L-Alanine, obtained through four-step reaction system, the total revenue is 55.6%L-alanine through methyl, and the L-2-cymbal base protected by the oxygen oxygenyl The reaction of aminopenate metropolis and phenyl bromide reaction (s) -2-苄oxygenyl amin amino-1,1-two phenyl-1-propyol. The divisor oxygen cymbal group gets the title compound. The intermediate and target compounds involved in the preparation method are easy to purify, the total yield is high and the reproducibility is good. Daba pre-body compound 3-3,4-two-苄oxygenyl) - 苄oxygenyl alanine.
Keywords	拆分剂, (S)-2-胺基-1,1-二苯基-1-丙醇, β-氨基醇, 屈昔多巴	Disassembly, (S) -2-aminel-1,1-two phenyl-1-1-propylene, β-amino alcohol, Koshidaba
Category	工学	Engineering
Discipline	化学/化学工程与技术	Chemistry and Chemical Engineering

Figure 1: Example document from the CSL dataset.

able for decades (Cleverdon, 1960; Harman, 1992). Cross-language benchmarks first appeared in the late 1990s (Schäuble and Sheridan, 1997). This retrieval evaluation methodology is well established and widely accepted.

Benchmarks for retrieval of technical documents are rarer, yet still available monolingually (usually with English queries and documents) (Hersh and Voorhees, 2009). Cross-language benchmarks for technical documents include the Kando et al. (1999) NACSIS benchmark. However, this benchmark is old and smaller than NeuCLIRTech, with just 21 Japanese queries over approximately 330,000 documents (in a combination of English and Japanese or either of the languages individually). The documents are technical papers published by 65 Japanese scientific associations.

While it tends to focus on high recall settings, patent retrieval is similar to technical document retrieval, among other retrieval tasks such as literature search (Lála et al., 2023). Like technical documents, patents have a distinct style that is unlike other genres. It also uses specialized domain-specific vocabulary. NTCIR ran several years of a patent retrieval task (Fujii et al., 2004, 2005, 2007) focusing on search for patent invalidation. The benchmark includes Japanese and English documents and queries, and supports cross-language retrieval. The CLEF Intellectual Property Track, which ran from 2009 to 2013, ran several patent-related tasks, including search for prior art and passage retrieval over patent applications and abstracts in English, French, and German (Roda et al., 2009; Piroi, 2010; Piroi et al., 2011, 2012, 2013).

3 Benchmark Creation

3.1 Documents

The NeuCLIRTech documents are abstracts from the Chinese Scientific Literature (CSL) dataset (Li et al., 2022) with 396,209 journal abstracts from 1,980 academic Chinese journals spanning 67 general disciplines, where Engineering, Science, Agriculture, and Medicine dominate. Articles were published between 2010 and 2020. Abstracts were originally obtained from the National Engineering Research Center for Science and Technology Resources Sharing Service (NSTR) in China.

The documents are distributed as JSONL formatted file, a list of JSON objects, one per line. Each line represents a document. Each document JSON structure consists of the following fields:

doc_id: the assigned id
title: title of the article
abstract: abstract of the article
keywords: keywords assigned to article
category: category assigned by NSTR
category_eng: human translation into English
discipline: discipline assigned by NSTR
discipline_eng: human translation into English

In addition to the documents their native language, NeuCLIRTech includes translations into English of all document fields obtained from the online *Google Translate* service in June 2023.

3.2 Queries

Queries were created by 22 graduate students and one postdoc from the Johns Hopkins University and the University of Maryland, College Park earning degrees in fields found in the document set.

Creators were hired based on their Chinese language skills and their familiarity with scientific research. They were paid USD$25 per hour for their work. During an application 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 a Chinese 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 some of the documents were related to their area of research and to verify their ability to understand technical documents written in Chinese. A Chinese speaker was one of the interviewers. Of the 22 graduate students, twelve were Ph.D. students and the remaining were Masters students.

Once hired, each annotator attended a three-hour online training session. They learned how to develop traditional TREC-style information needs scoped at the level of a scientific research paper. TREC information needs, called *topics*, consist of a short title, a sentence-length description, and a paragraph-length narrative. During ideation, query creators write a description of what they will search for and a narrative that discusses what is and is not considered relevant to the topic. During the search phase, the query creator uses a search engine to identify a set of documents to judge. The creator counts the number of relevant documents in the search results. Queries with too many relevant documents are revised to be more specific or discarded. During the revision phase, the description and narrative may be altered to reflect changes in the way documents were judged. Finally, the creator writes a title that summarizes the topic in three to five words. During training, each person worked independently to create their first topic. Two NeuCLIR Organizers reviewed their ongoing work to ensure that the topics had a suitable level of specificity and creators made proper use of the tools. Including training time, assessors were asked to spend up to a total of ten hours creating five to eight topics. Three assessors created five topics each, eleven assessors created six topics, seven assessors created seven topics, and two assessors created eight topics, yielding a total of 146 topics.

An organizer reviewed each English title, description, and narrative to ensure that the topic was sufficiently descriptive, was grammatically correct, and contained no spelling errors. In some cases, assessors were asked to revise topics that appeared to be too vague or not understandable. After any revision, assessors verified that the human translation incorporated any changes. The translations did not undergo any external quality control.

3.3 Relevance Judgments

The documents that assessors judged for relevance came from the top ranked documents of runs submitted by teams participating in the TREC NeuCLIR track of the year in which the query was created. For queries numbered 1 to 199, the top twenty documents of each run were assessed for relevance, while for queries numbered 200 and above, the top thirty-five documents were assessed. Documents were presented in random order, and assessors were asked to judge all documents for each of the topics they created.

Documents received a relevance score based on the answers assessors gave to two questions described in Appendix D. This led to document relevance scores of 3, 1, and 0 for very valuable, somewhat valuable, and not relevant respectively.

Before judging any documents, assessors participated in a second one-hour online training session on how to assess documents for relevance. The instructions on how to judge relevance appear in Appendix E.2. Relevance grades for the Technical Documents were designed for a searcher interested in scientific abstracts. 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 selected. They were asked to evaluate, based on its abstract, whether they would plan to read the paper being judged for inclusion in their related work section.

Assessor progress was tracked; when very few or an excessive number of relevant documents were found, or too little time was spent completing the task, they were asked to rejudge the pool. Some assessors ran out of time, leading to seven unjudged topics. Eight topics with fewer than three relevant documents were removed, as were sixteen topics for which more than 20% of the pool was judged to be somewhat or very valuable. Four other topics were removed because the assessor experienced technical difficulties while completing the task.

NeuCLIRTech includes 110 queries with 35,962 relevance judgments. Each query consists of the

Table 1: Retrieval nDCG@20 (nDCG) and Judged@20 (Jdg) on both monolingual and cross-language tasks ordered by CLIR nDCG.

	Monolingual		Cross-Lang.	
	nDCG	Jdg	nDCG	Jdg
BM25 w/ DT	–	–	0.237	1.00
BM25 w/ QT	0.290	0.99	0.274	1.00
Bi-Encoders				
BGE-M3 Sparse	0.307	0.75	0.044	0.13
e5 Large	0.233	0.55	0.151	0.43
RepLlama	0.269	0.59	0.245	0.56
Arctic-Embed Large v2	0.359	0.79	0.262	0.67
MILCO	0.253	0.56	0.264	0.66
JinaV3	0.382	0.79	0.305	0.68
PLAID-X	0.356	0.82	0.362	0.93
Qwen3 0.6B Embed	0.402	0.81	0.377	0.77
Fusion	0.438	0.92	0.431	0.96
Qwen3 4B Embed	0.469	0.84	0.450	0.83
Qwen3 8B Embed	0.480	0.87	0.472	0.87
Pointwise Rerankers on Fusion				
Mono-mT5XXL	0.456	0.94	0.407	0.93
SEARCHER Reranker	0.351	0.78	0.419	0.86
Jina Reranker	0.489	0.92	0.446	0.90
Qwen3 0.6B Rerank	0.494	0.94	0.485	0.94
Qwen3 8B Rerank	0.521	0.95	0.508	0.94
Rank1	0.531	0.88	0.512	0.88
Qwen3 4B Rerank	0.522	0.94	0.526	0.94
Listwise Rerankers on Fusion				
RankZephyr 7B	0.434	0.92	0.432	0.96
FIRST Qwen3 8B	0.539	0.94	0.520	0.93
RankQwen-32B	0.541	0.94	0.526	0.94
Rank-K (QwQ)	0.542	0.94	0.533	0.94

concatenation of the topic's title and description. The benchmark includes two tab-separated query files for the most common use cases (one for the original English queries, the other for human translations of those queries into Chinese), each line of which contains a query ID and the query text. A JSONL formatted query file includes many more details, such as the Google Translations of each field (to facilitate cross-language experiments on machine-translated queries).

4 Results

Table 1 summarizes monolingual and cross-language (CLIR) retrieval results. The only experiment that makes use of machine translated (MT) documents is BM25 w/DT (document translation). Notice that using document translation to cross the language barrier is among the lowest performing CLIR approaches; alternative approaches to crossing the language barrier are more effective because of the challenging vocabulary.

Qwen3-8B (Zhang et al., 2025) with FAISS is the strongest first-stage retriever in both tasks. MILCO (Nguyen et al., 2025), an effective multilingual LSR model on the Chinese newswire benchmark (Anonymous, 2026), performs below BM25 with query translation, which tokenizes the text using spaCy (Honnibal et al., 2020). For consistency with NeuCLIRBench (Anonymous, 2026), we use rank-fusion of three systems (Qwen3-8B, PLAID-X, and MILCO) as the first-stage retrieval results for rerankers, despite the fact the fusion performance is lower than that of Qwen3-8B.

Not all rerankers can outperform the initial retrieval results. Mono-mT5XXL, which is able to maintain roughly the same effectiveness in the monolingual task, has reduced effectiveness in CLIR, highlighting the difficulty of the cross-language task.

BGE-M3 Sparse and e5 Large are particularly ineffective on CLIR with low nDCG@20 values. Since it is cost-prohibitive to judge all documents for all queries, ineffective systems tend to have low Judged@20. We further investigate the top-20 documents retrieved from these two systems in Topics 5 and 7, where both systems retrieved zero or one judged document in the top 20. The average rank of a top-20 in the *Fusion*, Qwen3 4B or 8B Embed results is at least 700 (documents not appearing in one of these runs count as rank 1000);[1] this further confirms that the unjudged documents from these two particularly bad systems are likely to be non-relevant.

5 Conclusion

NeuCLIRTech is to our knowledge the first modern cross-language information retrieval benchmark that focuses on technical documents, enabling new CLIR retrieval models to be evaluated on a very challenging out-of-domain task.

While the baseline system ordering on NeuCLIRTech is similar to that of NeuCLIRBench (the newswire benchmark variant), NeuCLIRTech poses an additional challenge over CLIR on newswire because a cross-language technical task is out-of-domain for many current approaches. The fact that MILCO, a recently developed LSR model, still suffers from low performance on NeuCLIRTech demonstrates that this technical-domain and cross-language challenge requires further investigation.

[1] All rankings are released in Huggingface, allowing readers to reproduce this analysis.

Limitation

Because relevance judgments are made only for top documents retrieved by dozens of systems at the time of creation, systems that were not included as submitted runs to the TREC NeuCLIR track may have their effectiveness underestimated when they retrieve unjudged relevant documents that are scored as non-relevant. Researchers using NeuCLIRTech should report Judged@20 as an indicator of possible nDCG@20 underestimation.

References

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

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. 2024. Litsearch: A retrieval benchmark for scientific literature search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15068–15083.

Anonymous. 2026. NeuCLIRBench: A modern evaluation collection for monolingual, cross-language, and multilingual information retrieval.

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.

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.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Zijian Chen, Ronak Pradeep, and Jimmy Lin. 2025. Accelerating listwise reranking: Reproducing and enhancing FIRST. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3165–3172, New York, NY, USA. Association for Computing Machinery.

C.W. Cleverdon. 1960. The aslib cranfield research project on the comparative efficiency of indexing systems. *Aslib Proceedings*, 12(12):421–431.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2270–2282.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Cash Costello, Eugene Yang, Dawn Lawrie, and James Mayfield. 2022. Patapasco: A Python framework for cross-language information retrieval experiments. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II*, page 276–280, Berlin, Heidelberg. Springer-Verlag.

Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for ir with contextual neural language modeling. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 985–988.

Hang Ding, Yilun Zhao, Tiansheng Hu, Manasi Patwardhan, and Arman Cohan. 2025. Scirag: Adaptive, citation-aware, and outline-guided retrieval and synthesis for scientific literature. *arXiv preprint arXiv:2511.14362*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The Faiss library. *IEEE Transactions on Big Data*.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2004. Overview of patent retrieval task at NTCIR-4. In *Proceedings of NTCIR-4*.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2005. Overview of patent retrieval task at NTCIR-5. In *NTCIR Conference on Evaluation of Information Access Technologies*.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2007. Overview of the patent retrieval task at NTCIR-6. In *Proceedings of the 6th NTCIR Workshop Meeting*.

Donna Harman. 1992. Overview of the first text retrieval conference (TREC-1). In *Proceedings of The First Text REtrieval Conference*, volume 500-207 of *NIST Special Publication*, pages 1–20. National Institute of Standards and Technology (NIST).

William Hersh and Ellen Voorhees. 2009. Trec genomics special issue overview. *Information Retrieval*, pages 1–15.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.

Nasreen Jaleel, James Allan, W. Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark Smucker, and Courtney Wade. 2004. UMass at TREC 2004: Novelty and hard. In *Proceedings of the Thirteenth Text REtrieval Conference*. NIST.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint 2310.06825*.

Noriko Kando, Kazuko Kuriyama, and Toshihiko Nozue. 1999. Nacsis test collection workshop (ntcir-1) (poster abstract). In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 299–300, New York, NY, USA. Association for Computing Machinery.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*.

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*, pages 3917–3923, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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)*.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2022. *Pretrained transformers for text ranking: Bert and beyond*. Springer Nature.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023. Fine-tuning LLaMA for multi-stage text retrieval. *arXiv:2310.08319*.

Tiziano Montecchi, Davide Russo, and Ying Liu. 2013. Searching in cooperative patent classification: Comparison between keyword and concept-based search. *Advanced Engineering Informatics*, 27(3):335–345.

Thong Nguyen, Yibin Lei, Jia-Huei Ju, Eugene Yang, and Andrew Yates. 2025. MILCO: Learned sparse retrieval across languages via a multilingual connector. *arXiv preprint 2510.00671*.

Thong Nguyen, Sean MacAvaney, and Andrew Yates. 2023. A unified framework for learned sparse retrieval. In *European Conference on Information Retrieval*, pages 101–116. Springer.

Florina Piroi. 2010. CLEF-IP 2010: Retrieval experiments in the intellectual property domain. In *CLEF 2010 LABs and Workshops, Notebook Papers, 22–23 September 2010, Padua, Italy*, volume 1176 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Florina Piroi, Mihai Lupu, and Allan Hanbury. 2013. Passage retrieval starting from patent claims: A CLEF-IP 2013 task overview. In *CLEF (Working Notes)*, volume 1179 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Florina Piroi, Mihai Lupu, Allan Hanbury, Alan P. Sexton, Walid Magdy, and Igor V. Filippov. 2012. CLEF-IP 2012: Retrieval experiments in the intellectual property domain. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*, volume 1178 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Florina Piroi, Mihai Lupu, Allan Hanbury, and Veronika Zenz. 2011. CLEF-IP 2011: Retrieval in the intellectual property domain. In *CLEF 2011 Labs and Workshop, Notebook Papers, 19–22 September 2011, Amsterdam, The Netherlands*, volume 1177 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. RankZephyr: Effective and robust zero-shot listwise reranking is a breeze! *arXiv:2312.02724*.

Revanth Gangi Reddy, JaeHyeok Doo, Yifei Xu, Md Arafat Sultan, Deevya Swain, Avirup Sil, and Heng Ji. 2024. FIRST: Faster improved listwise reranking with single token decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8642–8652, Miami, Florida, USA. Association for Computational Linguistics.

Giovanna Roda, John Tait, Florina Piroi, and Veronika Zenz. 2009. CLEF-IP 2009: Retrieval experiments in the intellectual property domain. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September*

30 – October 2, 2009, Revised Selected Papers, volume 6241 of *Lecture Notes in Computer Science*, pages 385–409. Springer.

Saron Samuel, Alexander Martin, Eugene Yang, Andrew Yates, Dawn Lawrie, Ian Soborof, Laura Dietz, and Benjamin Van Durme. 2026. Beyond relevance: On the relationship between retrieval and rag information coverage. *arXiv preprint arXiv:2603.08819*.

Peter Schäuble and Paraic Sheridan. 1997. Cross-language information retrieval (CLIR) track overview. In *Proceedings of The Sixth Text REtrieval Conference, TREC 1997, Gaithersburg, Maryland, USA, November 19-21, 1997*, volume 500-240 of *NIST Special Publication*, pages 31–43. National Institute of Standards and Technology (NIST).

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task LoRA. *arXiv preprint 2409.10173*.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S Siegel, Michael Tang, et al. 2024. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. *arXiv preprint arXiv:2407.12883*.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937.

Qwen Team. 2025. QwQ-32B: Embracing the power of reinforcement learning.

Feng Wang, Yuqing Li, and Han Xiao. 2025. jina-reranker-v3: Last but not late interaction for document reranking. *arXiv preprint 2509.25085*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. 2025. Rank1: Test-time compute for reranking in information retrieval. In *Second Conference on Language Modeling*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Eugene Yang, Dawn Lawrie, and James Mayfield. 2024b. Distillation for multilingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2368–2373.

Eugene Yang, Dawn J. Lawrie, James Mayfield, Douglas W. Oard, and Scott Miller. 2024c. Translate-distill: Learning cross-language dense retrieval by translation and distillation. In *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part II*, volume 14609 of *Lecture Notes in Computer Science*, pages 50–65. Springer.

Eugene Yang, Andrew Yates, Kathryn Ricci, Orion Weller, Vivek Chari, Benjamin Van Durme, and Dawn Lawrie. 2025b. Rank-k: Test-time reasoning for listwise reranking. *arXiv preprint 2505.14432*.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of Lucene for information retrieval research. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 1253–1256.

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel F Campos. 2025. Arctic-embed 2.0: Multilingual retrieval without compromise. In *Second Conference on Language Modeling*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Appendix

A Model Descriptions

A.1 Sparse Retrieval

Sparse retrieval methods perform lexical matches between the query and the document. Based on the lexical matches, a score is produced for each retrieved document. We conduct experiments with BM25 using the Patapsco framework (Costello et al., 2022). Patapsco is a framework built on top of Pyserini (Lin et al., 2021) specifically for experiments in many languages. Given the reliance

on lexical matches, the input query and documents must be in the same language. Some experimental settings such as tokenenization and token normalization are language specific, while others are language agnostic. The language agnostic settings include the BM25 hyperparameter settings where $k1 = 0.9$ and $b = 0.4$ following standard conventions. The RM3 query expansion technique (Jaleel et al., 2004) added ten words based on the top ten documents, weighting terms from the original query and the expansion terms equally.

For the monolingual retrieval task, all characters were normalized. SpaCy (Honnibal et al., 2020) version 0.0.31 was used to perform tokenization. For Chinese, no stemming was applied. For the cross-language task, machine translation is used to cross the language barrier. BM25-DT indicates that the documents are translated into English and the spaCy tokenizer and stemmer are used to process the queries and documents, which have stopwords removed. BM25-QT uses machine translation to translate the English query into Chinese and then the Chinese mononlingual setting is followed.

A.1.1 Multi-Dense Vector Retrieval

We use PLAID-X (Yang et al., 2024c,b), a cross-language variant of the ColBERT model. PLAID-X is fine-tuned from an XLM-RoBERTa Large model with cross-language distillation from an mT5 reranker (described below) using MS-MARCO v1 training queries and passages. Each document is separated into passages using a sliding window of size 180 tokens with a stride of 90. At inference time, document scores are aggregated from the passage scores using MaxP (Dai and Callan, 2019).

A.1.2 Learned Sparse Retrieval

Learned Sparse Retrieval (LSR) methods represent queries and documents as sparse vectors where each dimension is associated with a term in a vocabulary. Some LSR methods produce vectors containing only terms from the input text, while others use a Masked Language Modeling head to also perform term expansion (Nguyen et al., 2023). To efficiently use these representations for first-stage retrieval, they can be stored in an inverted index. We conduct experiments with several LSR methods using Anserini (Yang et al., 2017), which is built on Lucene.

- BGE-M3-Sparse (Chen et al., 2024) is an LSR method trained as part of the multi-granularity M3-Embedding model. It is based on an XLM-RoBERTa backbone with 0.6B parameters. This method does not perform expansion, so it can only be used for monolingual retrieval.

- MILCO (Nguyen et al., 2025) is a multilingual LSR method that performs both query expansion and document expansion. MILCO performs Chinese retrieval by producing sparse vectors that are tied to an English vocabulary, so the model implicitly translates the non-English input. It is based on an XLM-RoBERTa backbone with 0.6B parameters.

A.1.3 Dense Retrieval

Dense Retrieval (DR) methods represent queries and documents as a single dense vector in a latent space (Lin et al., 2022). To efficiently use these representations for first-stage retrieval, they can be stored in a vector index that supports approximate nearest neighbor search, such as FAISS (Douze et al., 2025). We use fast scan product quantization with FAISS where the number of codes is set to half the embedding dimensionality and codes are 4 bits each.

- Arctic Embed v2 (Yu et al., 2025) is a multilingual model based on a XLM-RoBERTa backbone with 0.6B parameters. We use the `snowflake-arctic-embed-l-v2.0` checkpoint, which is the largest available Arctic Embed model.

- E5 (Wang et al., 2024) is a multilingual model based on an XLM-RoBERTa backbone with 0.6B parameters. We use the `multilingual-e5-large-instruct` checkpoint.

- Jina v3 (Sturua et al., 2024) is a multilingual model based on an XLM-RoBERTa backbone with 0.6B parameters.

- Qwen3-Embedding (Zhang et al., 2025) is a family of embedding models trained with synthetic data on top of Qwen3 backbones (Yang et al., 2025a). We use all three model sizes: 0.6B, 4B, and 8B.

- RepLlama (Ma et al., 2023) is a dense retrieval model based on LLaMA-2-7B. We use the `repllama-v1-7b-lora-passage` checkpoint trained on MS MARCO passages.

A.1.4 First-Stage Retrieval Fusion

We use reciprocal rank fusion (RRF) (Cormack et al., 2009) to combine the rankings produced by three strong first-stage retrieval models from the three different families described above: PLAID-X, MILCO, and Qwen-Embedding 8B. Following common practice and Cormack et al. (2009), we set $k = 60$ and do not use per-method weights.

A.1.5 Rerankers

We rerank the top-100 results from the first-stage retrieval fusion using a range of pointwise and listwise reranking models. Pointwise models take a query-document pair as input and output a relevance score. Listwise models take a query and set of documents as input. Some listwise models output one relevance score for each document, whereas others output an ordering of their input documents without producing relevance scores. For models that output an ordering, we use the partial sorting algorithm from RankGPT (Sun et al., 2023) to produce a ranking of all 100 results, where we rerank 20 documents at a time with a sliding window stride of 10 from the bottom of the top 100 candidate documents.

- FIRSTQwen-8B (Chen et al., 2025) is a listwise reranking model built on top of Qwen3 8B (Yang et al., 2025a) and trained with the FIRST objective (Reddy et al., 2024).
- jina-reranker-v3 (Wang et al., 2025) is a reranking model built on top of Qwen3 0.6B that can perform pointwise or listwise reranking. In both configurations, the model outputs one relevance score for each document. We use it in a pointwise configuration.
- mT5 (Bonifacio et al., 2022) is a pointwise reranking model built on top of a multilingual T5 backbone. We use the `mt5-13b-mmarco-100k` checkpoint with 13B parameters.
- Qwen3-Reranker (Zhang et al., 2025) is a family of pointwise reranking models built on top of Qwen3. We use all three model sizes: 0.6B, 4B, and 8B.
- Rank1 (Weller et al., 2025) is a pointwise reranking model fine-tuned to perform reasoning before reranking. Rank1 is built on top of the Qwen2.5 base model (Yang et al., 2024a). We use the 32B variant in our experiments.
- Rank-K (Yang et al., 2025b) is a listwise reranking model that performs reasoning before reranking using the QwQ 32B reasoning model (Team, 2025).
- RankQwen-32B is a zero-shot listwise reranking system that uses Qwen3-32B[2] without any training to rerank the top-100 documents. This system *does not* use any reasoning.
- RankZephyr-7B (Pradeep et al., 2023) is a listwise reranking model built on top of Mistral 7B (Jiang et al., 2023).
- SEARCHER (Agarwal et al., 2025) is a pointwise reranking model built on top of `Mistral-Nemo-Base-2407`[3] with 12B parameters. It was trained on Persian, Russian, and Chinese translations of MS MARCO passages (Bajaj et al., 2018), keeping the queries in English. Documents in the training batch were a mixture of all three languages.

B Computation Resources Required

In order to run the experiments reported in Table 1, we used a combination of several GPU types. We used AMD MI250X GPUs to obtain first-stage retrieval results with the bi-encoder models, which required approximately 500 GPU hours. We used NVIDIA H100 GPUs to obtain results with the reranking models, which took approximately 150 GPU hours.

C Licensing and Intended Use

Since documents are derived from the CSL collection, we use the Apache 2.0 license, which is the same as the original license.[4] Queries and relevance judgments are under CC-BY, allowing commercial uses with proper attribution.

D Judging Documents for Graded Relevance

Assessors answered up to two questions about each document to determine its relevance grade:
1. Does this document contain central information?

Yes Information in the abstract is related to their search topic.

[2] `https://huggingface.co/Qwen/Qwen3-32B`
[3] `https://huggingface.co/mistralai/Mistral-Nemo-Base-2407`
[4] `https://github.com/ydli-ai/CSL`

No (0 points) There is no information in the abstract related to their search topic.

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

The second question is answered if the first is *yes*: 2. How valuable is the most important information in this document?

Very Valuable (3 points) One would definitely read the paper associated with this abstract when writing the related work section for this research topic.

Somewhat Valuable (1 point) If one had enough time one would read the paper because it might contain information that could appear in the related work section, but confidence about that is low.

Not that Valuable (0 points) 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.

E Instructions for Assessors

E.1 Query Development Instructions

The NeuCLIR track is working on Cross-Language Information Retrieval (CLIR, get it?). In CLIR, the queries are in a different language than the documents. To support this research, you will be composing search topics for academic questions in your field study.

A search topic is a statement of a need for information. It is different from a query. A query is what is typed into the search box; in contrast, the topic is the need that the user is trying to articulate to the search engine. If I'm looking for an e-bike so I can learn what kinds of e-bikes there are to buy and how much they cost, I might type the query "e-bike models prices". The topic is what I want to find out about. The query is a short version of my topic that I think the search engine can work with.

For this task, we will be searching for information in abstracts of theses and articles written in simplified Chinese. Each of you is working on the project because of your facility in Chinese and your knowledge of a discipline represented in our documents. Your search topics should be something like what you use Google Scholar to find when writing the background section or a related work section for a paper.

You will explore the document collection to find interesting documents that inspire a search topic related to your field of study. You'll then scope that topic so that it doesn't have too many or too few relevant documents in the collection. Since you are being asked to create multiple topics, each topic should have a distinct focus; thus your imagined research project would be different for each topic. Then you'll compose a topic statement in English that describes your search need. Finally, you will translate that description into Chinese.

Each of you will be responsible for creating 5-7 topics over the course of ten hours by 26 June. If you create 7 topics in fewer than 10 hours, please save your remaining hours until August so that you will have more time to do the second task.

The goal of this process is to create a diverse set of search topics, each with a reasonable number of relevant documents.

E.1.1 Topic development process

1. To start, brainstorm a few ideas for research topics. Think about possible research projects that you and your colleagues are working on and what papers you might write about them. Your description will describe the information that you would want to include in the related work/background section of this hypothetical research paper. These documents were *written between 2010 and 2020*, so they will not include the most recent information.

2. On the SEARCH tab enter queries in English and/or Chinese to explore the collection to see if there is anything about the topic. If you find something you can move to the next step. You might have to change your topic idea a few times to find a good one.

3. On the DESCRIBE tab compose the **description**: a one-sentence statement or question about what you are searching for. For example, "What government agencies were affected by the WannaCry ransomware attack?" or "I am looking for studies on the detector design of the Beijing Electron-Positron Collider II (BEPCII)."

4. Write a first draft of the **narrative** paragraph. That paragraph should give more details, such as the kinds of information you would accept or would not accept as relevant, what details about the topic you are interested in finding, or anything you think is out of bounds for your topic. You will revise this in a later step, so

don't worry if it's not perfect. It can even be a comma-separated list at this point.

(a) Example narrative: The search should return information on detector studies of the Beijing Electron-Positron Collider II (BEPCII). If the result discusses the design or hardware components of BEPCII, but does not include information on detecting, it would be considered as less relevant. If the result includes a specific detector design, but does not mention the electron positron collider, it would be considered irrelevant. Additionally, if the result mainly discusses the software component of the BEPCII, it would be considered as irrelevant.

5. Lastly, give your topic a **title**, which should be two to five words describing the topic succinctly, like you might type into a web search engine. For example, "Beijing electron positron collider detector" might be your title for the last topic in number 3 above.

6. Return to the SEARCH tab, compose a single query for your topic in Chinese and English or just Chinese. It might be a translation of your title or it might be something else; it's up to you. Examine the 30 of the documents retrieved and judge them based on the value of the most valuable information in the document, as follows. First, find the information in the abstract that is closest to the topic you are searching for. Then answer these questions:

   - Question 1: Does this document contain central information?
     - **Yes**: There is information in the abstract related to your search topic
     - **No**: There is no information in the abstract related to your search topic.
     - **Unable to judge**: the document has not viewable in the document viewer panel
   - Question 2: How valuable is the most important information in this document?
     - **Very Valuable**: you would definitely read the paper associated with this abstract when writing the related work/background section for this research topic.
     - **Somewhat Valuable**: If you had enough time you would read the paper, because it might have something that could appear in your related work/background section, but your confidence about that is low.
     - **Not that Valuable**: You are unlikely to read the paper because you don't expect to find information that you would cite in your related work/background section.

7. If the number of very valuable and somewhat valuable abstracts is either less than 1, or greater than 20, the topic needs to be revised. Broaden or narrow the **description** sentence if possible and go back to step (6). If you don't see a way to revise, press the "Skip Task" button at the top of the interface to begin a new topic.

8. Otherwise, go back and revise your **narrative** to reflect your final topic. Make sure to mention edge cases you eliminated by narrowing the topic!

9. Update the **title** if necessary.

10. Go to the FINISH tab and translate the title and description into Chinese. Create a translation that is a natural expression of the idea in Chinese. Write the translation in Simplified Chinese.

Developing a topic takes about an hour, depending on how long the revision process takes. We would like you to develop 5-7 topics that are neither too narrow or too broad in 10 hours.

E.1.2 Tool Screenshots

After brainstorming to determine a potential topic is Step 1, the assessor used the search tab shown in Figure 2 to search the collection of documents for information that is relevant to their topic idea. Then the assessor moved on the describe tab shown in Figure 3 and completed Steps 3 to 5 before returning to the search tab to judge 30 documents. The counter shown in Figure 2 tracked their progress on the task. If they ever needed to revise their judgments on the documents, the modify tab in Figure 4 could be used. The task was completed on the Finish tab shown in Figure 5 where translations were entered.

Figure 2: Screenshot of the interface that searches the collection. A query in English or Chinese can be entered. Each ranked document is clicked on the link on the left to display the contents in the middle panel. A document is judged by answering the questions in the right panel.

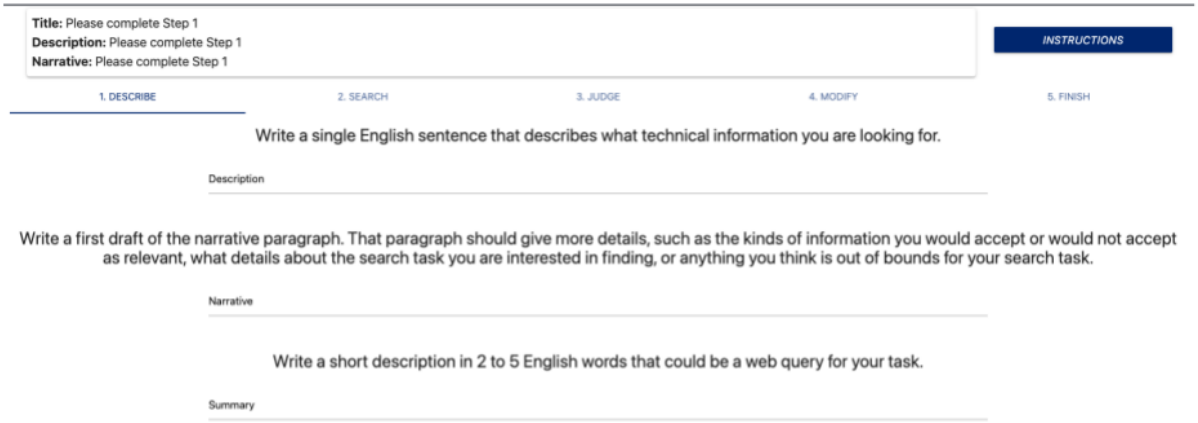

Figure 3: Screenshot of the interface where the assessor records information about the topic.

E.2 Relevance Assessment Instructions

E.2.1 Overview

The NeuCLIR task is concerned with Cross-Language Information Retrieval (CLIR): using search queries in one language to search for documents in another language. In our case, the searches are in English, and the documents are in Chinese. What makes this task unique is that the documents are abstracts of articles and theses, and therefore contain more different vocabulary than newswire.

You may remember that about a month ago you developed a set of **search topics**. Each topic represents an "information need," a gap in knowledge leading us to search. A topic is more than just a query – it defines what you were looking for and what kinds of information you would find relevant.

In this second part of the task, you will be assessing documents for relevance. For each topic, you will receive a **pool** of documents that were retrieved by TREC participant systems. You will judge each document as to whether it is relevant to your topic.

You will judge documents for the topics you created.

E.2.2 Tour

After you accept a task, you should see Figure 6 at the top of your browser.

On the first screen you will see the English description of the topic, the two questions that you must answer about the document, the contents of the document, a counter, and a button to "Save your progress." Once you answer the questions, the "NEXT" button will turn blue so that you can go to

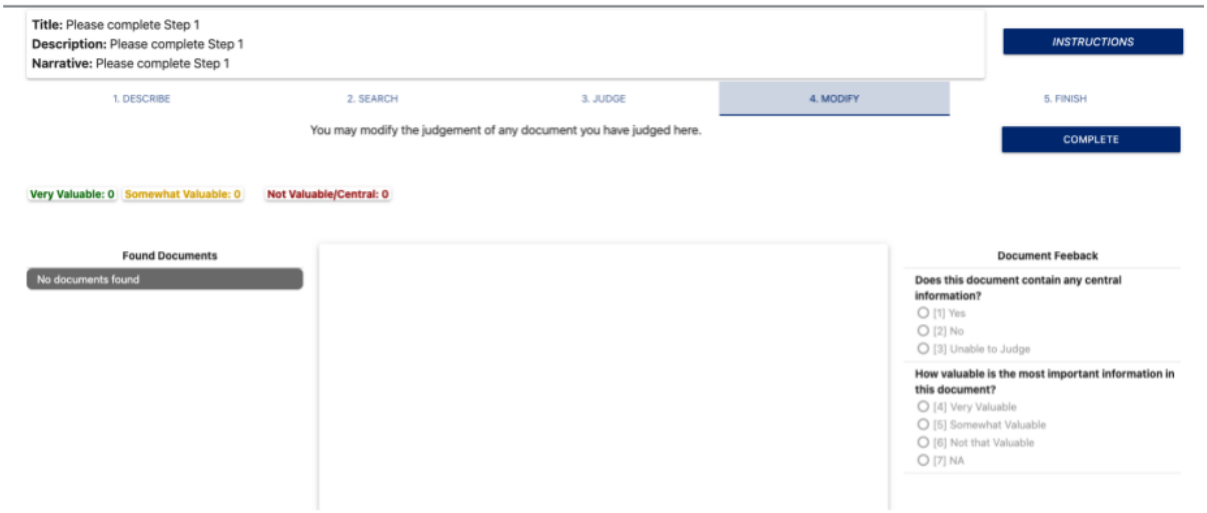

Figure 4: Screenshot of the interface where the assessor can modify the judgments of documents that have already been judged.

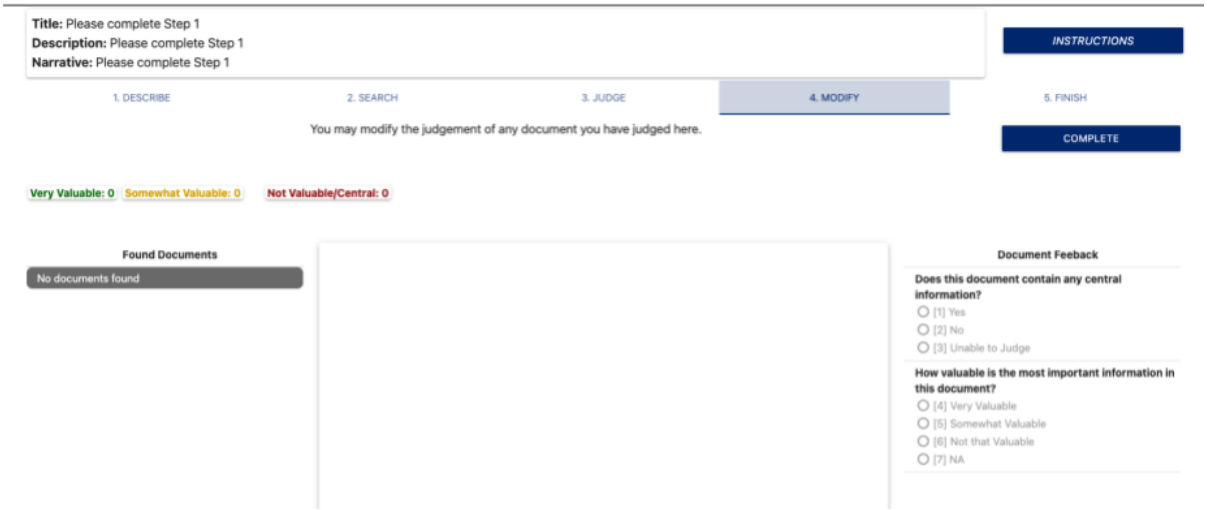

Figure 5: The screenshot of the final tab where assessors complete the task by entering translations of the title and description in Chinese. They had to rate how easy it was to find central information and optionally could add a comment about the topic.

the next document.

You begin on the JUDGE tab. This is where you will see each document in turn. The MODIFY tab is used to review or change the judgment for a document. Once you are done judging the pool, you will be brought to the FINISH tab. If you want, you may leave a comment. Then click SUBMIT. Only submitted tasks will be recorded. If you click on the FINISH tab before you judge the entire pool, you will have to confirm that you want to submit the task early. Please do not do this without first consulting with Dawn Lawrie.

The numerals next to the buttons can be used to select the radio buttons in lieu of a mouse click. For example, if I wanted to judge something "Very Valuable," I could type 1 followed by 4. In this way, your mouse can stay on the next button if the documents aren't too long.

E.2.3 Relevance

What does it mean for a document to be relevant to your topic? Imagine that you are writing the background section or the related work section for a scientific paper. You will use the best documents in the pool as citations for your paper.

You will answer two questions about each document.

- Question 1: Does this document contain central information?

Figure 6: The screenshot of the judge tab where assessors record relevance judgments for all the top ranked documents requiring assessment. The title, description, and narrative for the topic was always visible while judging documents. Below the relevance questions, the document was displayed. Once a document was judged, the next button would become active. A counter at the bottom tracked their progress.

  - **Yes**: There is information in the abstract related to your search topic
  - **No**: There is no information in the abstract related to your search topic.
  - **Unable to judge**: the document has not viewable in the document viewer panel

- Question 2: How valuable is the most important information in this document?
  - **Very Valuable**: you would definitely read the paper associated with this abstract when writing the related work/background section for this research topic.
  - **Somewhat Valuable**: If you had enough time you would read the paper, because it might have something that could appear in your related work/background section, but your confidence about that is low.
  - **Not that Valuable**: You are unlikely to read the paper because you don't expect to find information that you would cite in your related work/background section.

E.2.4 Tips

It is best to complete the judgments for the entire topic without closing your browser. However, there is a "Save your Progress" button. This saves your work to the cookies in your current browser; it won't be accessible from other browsers or if you clear your cookies. Thus, it will usually save your work, but is not robust.

Judge each document **independently**. It must be valuable (or not) on its own. That also means that if you see valuable information repeated, it's still valuable and should be judged the same.

How to judge each document is completely up to you, as the user of the search system and author of the imagined paper. However, it's critical that you try to be **consistent**. If you judge something to be very valuable once, apply that same standard to other documents. If some other piece of information is not relevant or not that valuable, judge it the same way the next time you see it.

It's important to make steady progress through your pool. If you're not sure how to judge a document, use your best judgment. You may write down the number of the document so that you can revisit it on the modify tab later.

The vast majority of documents you see will not be relevant. Aside from those, many documents may be topically related, but not be very valuable. Typically there are only a few somewhat valuable or very-valuable documents. However, recall that when creating the topic you found between 1 and 20 documents about your topic. These documents

are in the pool and we expect that they would continue to be valuable, so when I say a few, it could be a few dozen documents.

You will find that documents that are obviously not relevant take very little time to identify. Make sure you at least skim to the end of each document to make sure there isn't anything topical buried in there. If you want to look for specific terms, use Find in your browser.

Once you reach ten hours of working on these tasks, finish the one you are on and stop.

F Data Contents

The data we collected from assessors was converted into scores of 0, 1 or 3 for each document. The data is distributed as a file in the standard qrels format with a judgment per line. This judgment consists of a "<topic id> 0 <document id> <score>.