DisastIR: A Comprehensive Information Retrieval Benchmark for Disaster Management

Anonymous ACL submission

Abstract

Effective disaster management requires timely access to accurate and contextually relevant information. Existing Information Retrieval (IR) benchmarks, however, focus primarily on general or specialized domains, such as medicine or finance, neglecting the unique linguistic complexity and diverse information needs encountered in disaster management scenarios. To bridge this gap, we introduce **DisastIR**, the first comprehensive IR evaluation benchmark specifically tailored for disaster management. DisastIR comprises 9,600 diverse user queries and more than 1.3 million labeled query-passage pairs, covering 48 distinct retrieval tasks derived from six search intents and eight general disaster categories that include 301 specific event types. Our evaluations of 30 state-of-the-art retrieval models demonstrate significant performance variances across tasks, with no single model excelling universally. Furthermore, comparative analyses reveal significant performance gaps between general-domain and disaster management-specific tasks, highlighting the necessity of disaster management-specific benchmarks for guiding IR model selection to support effective decision-making in disaster management scenarios. All source codes and DisastIR are available at this repository.

1 Introduction

Natural disasters and technological crises cause severe threats to human lives, infrastructure, and the environment, necessitating timely and effective management responses (Dong et al., 2020; Yin et al., 2023; Liu et al., 2024). In such critical scenarios, stakeholders, including emergency responders, government agencies, and the general public, require rapid access to reliable and contextually relevant information to make informed decisions (Jayawardene et al., 2021; Abbas and Miller, 2025). Information Retrieval (IR) systems thus play a critical role in disaster management, where rapid, accurate access to relevant information can significantly impact emergency response outcomes and decision-making efficacy (Basu and Das, 2020; Kumar et al., 2023; Langford and Gulla, 2024).

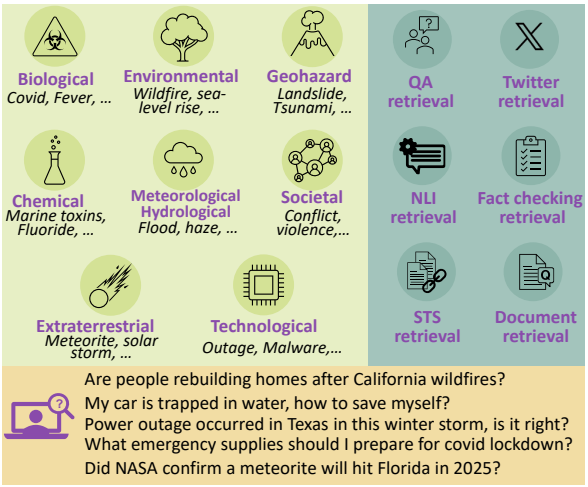

Figure 1: Examples of user queries across diverse search intents and event types during disaster management.

Information needs during real-world disasters are highly diverse (Figure 1), including intents such as question answering, rumor verification, social media monitoring, and evidence retrieval (Purohit et al., 2014; Imran et al., 2015; Zubiaga et al., 2018). These varied intents require tailored retrieval behavior (Asai et al., 2022; Su et al., 2022; Lee et al., 2024b) and understanding of "relevance" (Dai et al., 2022). In addition, different types of disasters (Figure 1), such as geohazards, biological threats, and technological failures, differ significantly in terminology, phrasing, and discourse styles (Andharia, 2020; UNDRR, 2020; Bromhead, 2021). This complexity presents significant challenges for retrieval systems aiming to serve real-world disaster response scenarios.

However, existing retrieval benchmarks primarily target general-domain tasks, such as BEIR (Thakur et al., 2021), or focus on specific domains like medicine (Wang et al., 2024a) and finance (Tang et al., 2024). They are not designed to re-

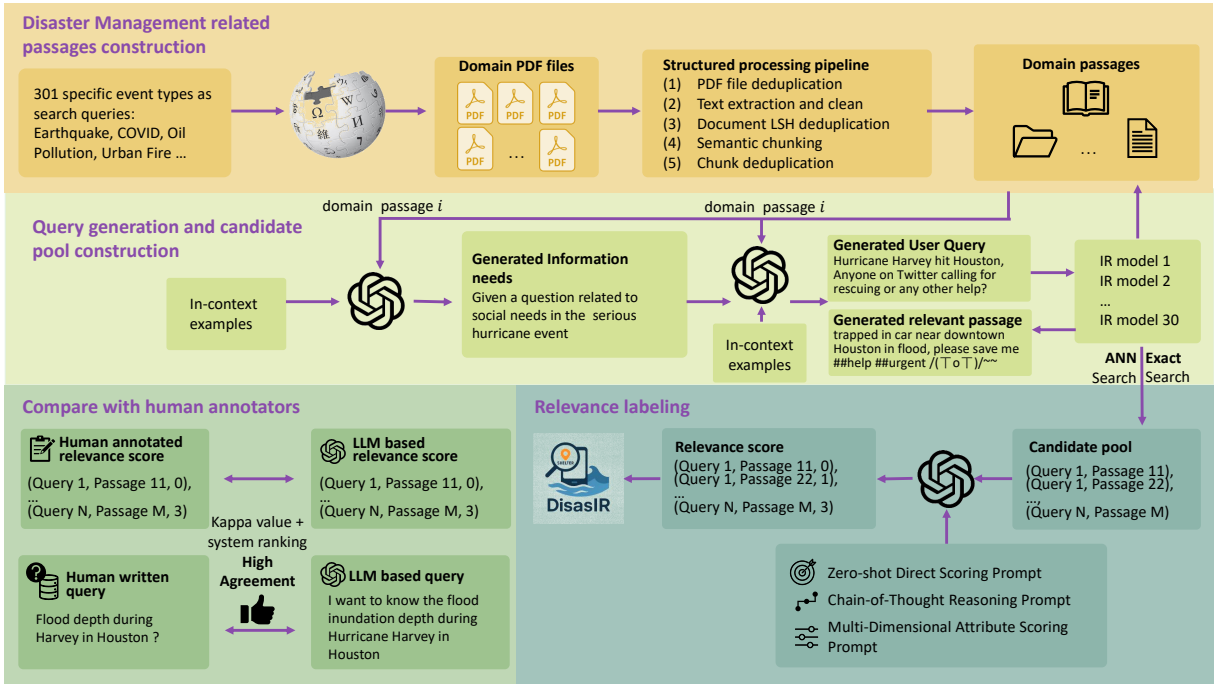

Figure 2: Proposed framework to develop DisastIR from scratch.

flect the search task diversity and domain-specific demands of disaster management scenarios. As a result, current IR evaluation benchmarks offer limited guidance for selecting retrieval models in disaster management applications.

To address this gap, we present DisastIR, the first comprehensive IR benchmark tailored to disaster management. DisastIR evaluates retrieval models across 48 distinct tasks, defined by combinations of six real-world search intents and eight general disaster event types, covering a total of 301 specific event types (see Section 3.2).

DisastIR is built on a systematically constructed disaster management-specific corpus, developed through extensive web crawling, semantic chunking, and deduplication (Section 3.3). To simulate realistic information needs, we use a large language model (LLM)[1] to generate diverse, contextually grounded user queries (Section 3.4). Candidate passages are aggregated from multiple state-of-the-art (SOTA) retrieval models (Section 3.5), and query-passage pairs are annotated using LLMs with three different designed prompts whose outputs are ensembled for robust relevance labeling (Section 3.6).

To ensure annotation quality and evaluation reliability, we validate LLM-generated relevance labels against human annotations, observing substantial agreement (average Cohen's kappa = 0.77; see Section 4.2). We also compare LLM-generated and human-written queries across all 48 tasks (Section 4.3) and find highly consistent evaluation results (Kendall's τ = 0.93), supporting the use of synthetic queries and relevance labels in DisastIR.

Using DisastIR, we benchmark 30 open-source retrieval models of varying sizes, architectures, and backbones under both exact and approximate nearest neighbor (ANN) search settings (Section 5). Our results show that no single model consistently outperforms others across all disaster management-related retrieval tasks (Section 6.2). We also observe substantial performance gaps between general-domain benchmarks (e.g., MTEB (Muennighoff et al., 2022)) and DisastIR (Section 6.3), highlighting the need for a domain-specific benchmark to guide reliable and effective retrieval model selection in disaster management scenarios.

The contributions of this work are as follows:

(1) We release DisastIR, the first IR benchmark tailored to disaster management. It includes a systematically constructed evaluation corpus of 239,704 passages and 9,600 user queries, with over 1.3 million annotated query-passage pairs across 48 retrieval tasks spanning diverse search intents and disaster event types.

(2) We conduct a comprehensive evaluation of 30 open-source retrieval models under both exact and ANN search settings, offering practical guidance for model selection based on task requirements and computational constraints

[1]The LLM used in this work is GPT-4o-mini.

in disaster management scenarios.

(3) We empirically demonstrate substantial performance gaps between general-domain and disaster management-specific retrieval, underscoring the necessity of disaster management-specific IR evaluation benchmarks.

2 Related work

Existing IR benchmarks target mainly general-purpose or specialized domains, such as medicine and finance. BEIR (Thakur et al., 2021) evaluates zero-shot retrieval models across 18 tasks, such as fact verification, QA, and scientific document ranking. Instruction-based benchmarks like FollowIR (Weller et al., 2024), InstructIR (Oh et al., 2023), and MAIR (Zhang et al., 2024b) reformulate IR tasks using natural language instructions. Some domain-specific IR benchmarks, such as MIRAGE (Wang et al., 2024a) and FinMTEB (Tang et al., 2024), focus on biomedical and financial domains. While effective in their respective domains, they fail to capture the linguistic and contextual patterns in disaster management areas.

Despite the critical role of information retrieval in disaster management, existing benchmarks are limited in scope, scale, and task diversity. Prior datasets—such as the FIRE IRMiDis track (Basu et al., 2017) and event-specific corpora from disasters in Nepal, Italy, and Indonesia (Khosla et al., 2017; Basu and Das, 2019; Kumar et al., 2023)—primarily focus on Twitter microblogs, targeting short-text retrieval or keyword matching with narrow task coverage. Case-based systems like Langford and Gulla (2024) use proprietary data for concept-based retrieval in search and rescue planning. These benchmarks typically rely on single-source or scenario-specific data and lack support for realistic, multi-intent retrieval. In contrast, DisastIR provides a large-scale, multi-intent, and multi-source benchmark covering diverse disaster types and information needs, enabling comprehensive evaluation in real-world contexts.

3 DisastIR: Disaster Management Information Retrieval Benchmark

3.1 Overview

The construction of DisastIR follows a four-stage pipeline, as illustrated in Figure 2: (1) disaster management corpus construction, (2) user query generation, (3) candidate pool development, and (4) relevance labeling. DisastIR is built upon a large-scale, high-quality corpus of disaster management-related passages covering diverse event types. User queries are generated by prompting an LLM with these domain passages as context, targeting different search intents. Relevance scores for each query-passage pair are then assigned by the LLM.

3.2 Evaluation Task

To evaluate how well retrieval models address diverse user intents and disaster contexts, DisastIR defines six search intents and eight general disaster event types, resulting in 48 distinct retrieval tasks.

Specifically, 301 specific event types are identified spanning eight general categories: Biological (Bio), Chemical (Chem), Environmental (Env), Extraterrestrial (Extra), Geohazard (Geo), Meteorological & Hydrological (MH), Societal (Soc), and Technological (Tech) (UNDRR, 2020). See Figure 1 for examples of specific event types belonging to each general disaster event type.

Six distinct search intents are included, inspired by prior benchmarks such as BEIR (Thakur et al., 2021), BERRI (Asai et al., 2022), MEDI (Su et al., 2022), and MAIR (Sun et al., 2024): question-answer (QA) retrieval, Twitter retrieval, Fact Checking (FC) retrieval, Natural Language Inference (NLI) retrieval, and Semantic Textual Similarity (STS) Retrieval. For QA, we further distinguish between retrieving relevant passages (QA) and retrieving relevant documents (QAdoc), following common practice in prior work (Kwiatkowski et al., 2019; Khashabi et al., 2021; Xu et al., 2024).[2] Due to token limitations in many retrieval models – especially encoder-based ones – it is often infeasible to encode full documents directly. To address this, we prompt an LLM to summarize each document and include the summary in the corpus as a proxy for the original document.

3.3 Domain knowledge corpus construction

To construct the domain knowledge corpus, we perform a large-scale web crawling using 301 disaster event types as search queries, collecting domain-specific PDF documents from publicly available sources. A structured pipeline is then applied to convert raw PDFs into clean, retrieval-ready passages: (1) exact-URL deduplication, (2) text extraction and preprocessing, (3) document-level near-

[2] A passage refers to a single chunk with limited token length, while a document denotes a full source file, which may be segmented into multiple passages.

duplicate removal using locality-sensitive hashing (LSH), (4) semantic chunking, and (5) embedding-based near-duplicate filtering. The full pipeline is described in Appendix A.

3.4 User Query Generation

A key challenge in constructing domain-specific IR evaluation datasets is generating user queries that reflect real information needs (Rahmani et al., 2024a). With the advent of LLMs, it is now feasible to synthesize high-quality, diverse, and contextually grounded queries by prompting models with domain-specific passages (Alaofi et al., 2023; Rajapakse and de Rijke, 2023; Rahmani et al., 2024a).

In this work, we propose a two-stage few-shot prompting strategy to generate user queries based on disaster management passages. In the first stage, an LLM is prompted to brainstorm diverse information need statements grounded in the content of the given passage. In the second stage, given a randomly selected information need and the associated passage, the LLM generates a user query and a directly relevant passage as shown below:

$$LLM_{query}\big(\underbrace{LLM_{info}(P_{IN}, P_{seed})}_{\text{information need}},\ P_{QG},\ P_{seed}\big) \longrightarrow (q,\ psg) \quad (1)$$

where LLM_{info} and LLM_{query} are LLMs prompted to generate retrieval information needs statements and the query-passage pair respectively, P_{IN} and P_{QG} are prompts for information needs and query generation, P_{seed} is the domain passage, q is the synthesized user query, and psg is the corresponding relevant passage.

To ensure generated queries align with the core characteristics and objectives of each search intent, we design intent-specific prompts for both stages of query generation. The full prompt templates for each intent are provided in Appendix B.

For each search task, we generate 200 unique user queries by prompting an LLM with randomly sampled domain-specific passages, resulting in 9,600 queries. The final corpus combines disaster management-related passages from Section 3.3 with generated passages to reflect various search intents. Some tasks, such as Twitter, NLI, and FC retrieval, require passage types with distinct styles and semantics. Including generated passages ensures the corpus can support realistic evaluation across diverse retrieval scenarios.[3]

[3]Relevance scores of query-generated passage pairs are also evaluated instead of directly giving them the highest relevance score.

3.5 Assessment Candidate Pool Development

Given the large size of the corpus, annotating all possible (query, passage) pairs is impossible (Thakur et al., 2021). Following prior work, we construct a candidate pool for each query using existing retrieval models. Inspired by TREC's standard practice, where top-ranked passages from multiple systems are aggregated to form the candidate set, we adopt a similar strategy in DisastIR.

Specifically, for each query, we collect the top 10 retrieved passages from 30 retrieval models under two retrieval settings: exact and ANN search settings (detailed in Section 5). These models also serve as baselines for performance evaluation, following practices in recent work (Rahmani et al., 2024b; Wang et al., 2024b). The candidate pool for each query is formed by taking the union of passages retrieved under both settings.

3.6 Relevance Labeling

Once query-passage pairs are prepared, we annotate them using an LLM. Recent studies have shown that LLMs can reliably produce relevance judgments that align closely with human annotations (Rahmani et al., 2024a,b, 2025; Wang et al., 2024b). Furthermore, Wang et al. (2024b); Rahmani et al. (2024c) demonstrate that ensembling relevance scores from multiple prompts or LLMs yields more robust and calibrated annotations.

To this end, we design three diverse prompts for each search intent and use a single LLM to generate relevance scores. The prompts, inspired by Thomas et al. (2024); Farzi and Dietz (2024); Rahmani et al. (2025), are: (1) zero-shot direct scoring—a single-pass judgment; (2) chain-of-thought reasoning—a multi-step prompt mimicking human-style reasoning; and (3) multi-dimensional attribute scoring—relevance decomposed into interpretable sub-criteria. For each search intent, relevance is defined to align with its specific objectives, reflecting the varying interpretations of "relevance" across different task types (Dai et al., 2022). Full prompt templates are provided in Appendix C.

Relevance scores are assigned on a 4-point scale (0 to 3) for all intents, except STS, which follows a 6-level scale as in Agirre et al. (2013); Cer et al. (2017). The final score for each pair is computed by averaging scores from three prompts.

	QA	QAdoc	Twitter	FC	NLI	STS
Bio	26651 (133.3)	25335 (126.7)	35182 (175.9)	23987 (119.9)	25896 (129.5)	27065 (135.3)
Chem	26885 (134.4)	26032 (130.2)	34186 (170.9)	24592 (123.0)	27856 (139.3)	26787 (133.9)
Env	26685 (133.4)	25930 (129.7)	33243 (166.2)	25805 (129.0)	25207 (126.0)	27048 (135.2)
Extra	26807 (134.0)	25598 (128.0)	33202 (166.0)	24363 (121.8)	26399 (132.0)	27313 (136.6)
Geo	27140 (135.7)	26573 (132.9)	35503 (177.5)	27864 (139.3)	28210 (141.1)	29816 (149.1)
MH	28422 (142.1)	27256 (136.3)	33924 (169.6)	26670 (133.4)	27052 (135.3)	28702 (143.5)
Soc	27116 (135.6)	23353 (116.8)	33834 (169.2)	27850 (139.3)	26997 (135.0)	27074 (135.4)
Tech	28044 (140.2)	27071 (135.4)	33388 (166.9)	26759 (133.8)	28394 (142.0)	26920 (134.6)

Table 1: Number of labeled query-passage pairs and pairs per query (in parentheses) of each search task in DisastIR.

	Count	Avg	Median	Min	Max
Query	9,600	33.75	19	2	281
Passage	239,704	197.17	224	6	2,536

Table 2: Statistics of number of query and passage and their token lengths. Tokenization is based on the `cl100k_base` tokenizer (used in GPT-4 / GPT-3.5).

4 DisastIR Benchmark Analysis

4.1 Query and Passage Characteristics

Query and Passage Lengths. As shown in Table 2, the average query length is 33.75 tokens, with a median of 19, and a long tail extending to 281 tokens. This variation reflects the diversity of search intents, from short entity-style queries to detailed information needs typical in real-world disaster management scenarios. Passages are much longer on average (197.17 tokens), with a median of 224, and some exceeding 2,500 tokens. This wide distribution captures the diversity of disaster management-related texts, including both brief updates and detailed descriptions like event summaries or emergency protocols.

Labeled Query-Passage Pairs. Table 1 summarizes the distribution of labeled query-passage pairs. In total, we obtained 1,341,986 labeled pairs, with each query linked to an average of 140 passages.

As shown in Table 1, Twitter-related search tasks tend to have a higher average number of query-passage pairs per query. The candidate pool for each query is built by merging the top 10 passages retrieved by 30 different models. This larger pool in Twitter tasks suggests greater divergence in model outputs, indicating lower agreement among retrieval models when ranking passages in social media contexts within disaster management scenarios. Additional analyses of labeled query-passage pairs are provided in Appendix D.

4.2 LLM-based vs. Human Labeling

Since relevance scores in DisastIR are judged by LLM, it is vital to evaluate their consistency with human annotations. Thus, we construct the LVHL dataset (**L**LM-based **V**s. **H**uman **L**abeling) by sampling disaster management-related query-passage pairs with human-labeled relevance scores from several open-source datasets. MS MARCO (Bajaj et al., 2016) and TriviaQA (Joshi et al., 2017) are for QA, ALLNLI (sentence-transformers, 2021) and XNLI (Conneau et al., 2018) for NLI, Climate-Fever (Diggelmann et al., 2021) for FC, and STSB (Cer et al., 2017) for STS. Appendix E provides details on the construction of LVHL.

The LLM-based relevance scores for each query-passage pair in LVHL are computed as described in Section 3.6. Since most human-annotated relevance scores in LVHL are binary, we follow Wang et al. (2024b) and binarize the LLM scores into two levels: relevant (score > 0) and not relevant (score = 0), to enable meaningful comparison.

To assess agreement between LLM-based and human relevance labeling, we compute Cohen's kappa for each search intent. All datasets yield kappa scores above 0.6 (Figure 6), with an average of 0.77, indicating substantial agreement. These suggest that LLM-generated relevance scores align well with human judgments and can reliably substitute for manual annotation in DisastIR.

4.3 LLM vs. Human-generated User Query

To evaluate whether LLM-generated queries can serve as a reliable alternative to human-authored ones for retrieval benchmarking, we construct LVHQ (**L**LM **V**s. **H**uman-generated **Q**uery), a comparison set spanning all 48 retrieval tasks. For each task, both an LLM-generated and a human-written query are created based on the same domain passage. All query-passage pairs are annotated using the same method as in DisastIR. Appendix F provides full details on the construction of LVHQ.

Model Name	Param Size	Size Bin	Exact ↑ QA	QAdoc	TW	FC	NLI	STS	Ex. Avg	ANN Avg	Drop (%)
inf-retriever-v1	7B	XL	74.23	67.82	69.33	68.91	53.10	**77.70**	**68.52**	**66.90**	2.36
SFR-Embedding-Mistral	7B	XL	71.76	67.58	**70.42**	**70.62**	50.86	73.61	67.47	66.75	1.08
inf-retriever-v1-1.5b	1.5B	XL	70.64	64.42	66.11	66.53	53.86	76.19	66.29	65.54	1.13
NV-Embed-v2	7B	XL	**74.77**	**69.74**	43.18	68.64	**58.73**	77.01	65.34	64.45	1.36
multilingual-e5-large	560M	Large	67.29	64.25	63.75	60.26	51.02	75.06	63.61	62.08	2.41
multilingual-e5-large-instruct	560M	Large	68.39	64.90	63.24	67.21	49.38	64.31	62.91	62.06	1.34
e5-small-v2	33M	Small	65.87	63.00	60.89	62.04	47.09	74.83	62.29	60.91	2.22
e5-base-v2	109M	Medium	65.77	63.07	58.37	62.28	45.54	74.64	61.61	60.23	2.25
e5-large-v2	335M	Large	60.15	63.42	56.20	62.29	50.96	74.99	61.34	60.20	1.85
NV-Embed-v1	7B	XL	68.36	63.02	56.84	60.04	48.31	67.86	60.74	59.24	2.47
granite-embedding-125m	125M	Medium	64.90	61.04	47.14	62.78	48.04	71.94	59.31	58.67	1.06
llmrails-ember-v1	335M	Large	64.82	63.34	45.93	61.06	44.56	73.99	58.95	58.35	1.02
arctic-embed-m-v2.0	305M	Medium	61.44	62.50	47.80	58.04	42.34	65.28	56.23	54.75	2.64
mxbai-embed-large-v1	335M	Large	64.57	62.96	40.58	58.47	40.22	68.81	55.93	55.42	0.92
gte-base-en-v1.5	137M	Medium	60.66	56.06	47.06	52.49	39.88	71.27	54.57	53.71	1.57
gte-large-en-v1.5	434M	Large	67.70	58.55	40.31	53.09	34.81	67.30	53.63	53.09	1.01
snowflake-arctic-embed-l-v2.0	568M	Large	55.34	59.43	38.75	60.42	41.19	63.38	53.09	51.90	2.23
bge-base-en-v1.5	109M	Medium	49.67	53.97	47.03	58.47	35.58	64.43	51.52	50.71	1.57
bge-large-en-v1.5	335M	Large	57.02	54.72	32.71	55.22	35.19	65.14	50.00	48.99	2.03
gte-Qwen2-7B-instruct	7B	XL	62.47	39.12	41.40	29.98	47.28	70.98	48.54	47.07	3.03
bge-small-en-v1.5	33M	Small	47.38	45.51	27.82	52.13	27.20	56.86	42.82	42.07	1.75
snowflake-arctic-embed-m-v1.5	109M	Medium	25.73	30.56	18.31	48.24	43.04	64.95	38.47	36.25	5.79
snowflake-arctic-embed-s	33M	Small	36.05	27.46	18.17	42.52	37.27	66.59	38.01	33.59	11.63
snowflake-arctic-embed-l	335M	Large	40.82	30.41	15.32	32.70	34.62	56.82	35.11	31.35	10.72
Linq-Embed-Mistral	7B	XL	35.61	31.41	26.55	39.46	27.30	45.69	34.34	33.52	2.38
thenlper-gte-base	109M	Medium	9.22	5.35	38.54	60.80	42.85	46.64	33.90	31.78	6.24
snowflake-arctic-embed-m	109M	Medium	33.35	14.25	8.64	35.30	38.93	56.88	31.23	28.83	7.67
all-mpnet-base-v2	109M	Medium	15.06	9.77	16.17	46.30	27.38	37.23	25.32	25.05	1.08
e5-mistral-7b-instruct	7B	XL	21.65	19.51	19.48	31.04	20.39	39.57	25.27	24.05	4.82
gte-Qwen2-1.5B-instruct	1.5B	XL	13.98	22.21	19.61	23.90	18.00	31.20	21.48	21.27	0.98

Table 3: Performances of 30 evaluated IR models in DisastIR. Models are ranked by their overall performance under exact search (highest to lowest) in DisastIR. "Size Bin" indicates its model parameter size bin category (small, medium, large, and extra large as defined in Appendix G). "TW" represents Twitter. Overall performance across all queries under exact and ANN search are in "Ex. Avg" and "ANN Avg" columns. "Drop" shows the percentage decrease from exact to ANN average scores. **Bold** indicates the highest value, and underline indicates the second-highest.

We evaluate all selected baseline models using LVHQ under exact search for both human- and LLM-generated queries (see Section 5 for evaluation setup up). Model performance, measured by NDCG@10, shows highly consistent results across the two query types, with a Kendall's τ of 0.9264, indicating strong agreement in model evaluations.

5 Experimental Setup

5.1 Models

DisastIR is adopted to comprehensively evaluate open-source IR models and support the selection of suitable IR models for real-world disaster management applications. Models are chosen based on three criteria: (1) strong performance on the MTEB retrieval benchmark; (2) inclusion in widely adopted embedding families such as `BGE` (Chen et al., 2024; Xiao et al., 2024), `E5` (Wang et al., 2022, 2023), `Snowflake Arctic` (Merrick, 2024), and `GTE` (Li et al., 2023; Zhang et al., 2024a), which are commonly used as baselines and in downstream IR tasks (Sun et al., 2024; Xu et al., 2024; Lee et al., 2024b,a; Cao, 2025; Park et al., 2025); and (3) coverage of different types of search tasks adapted from Wang et al. (2024b).

We select 30 models with parameter sizes ranging from 33 million to 7 billion. Detailed descriptions of these models and their implementations are provided in Appendix G.

5.2 Evaluation

We evaluate model performance under two retrieval settings – exact and ANN – using Normalized Discounted Cumulative Gain at rank 10 (NDCG@10)

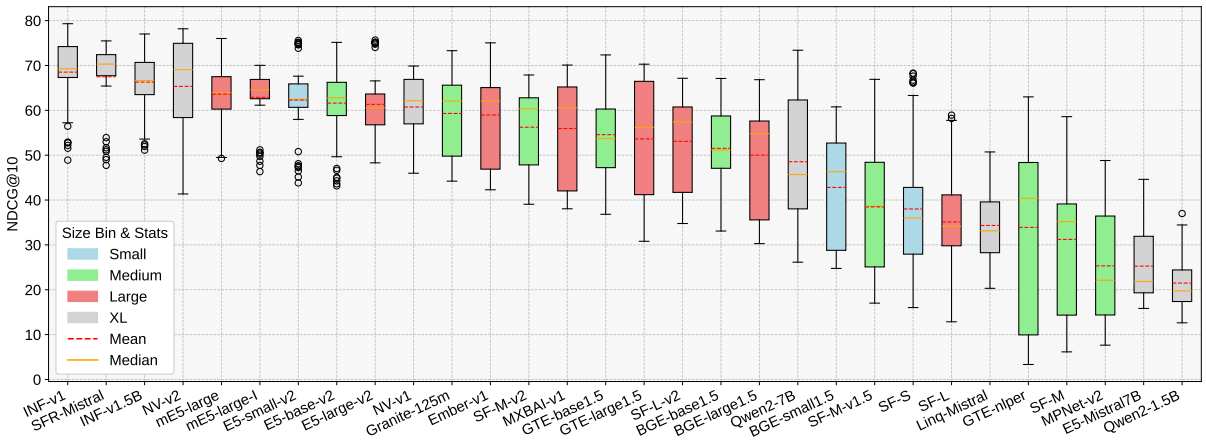

Figure 3: Distribution of evaluated models' performances across all 48 tasks. The full name of each model in the X axis is listed in Model Name column in Table 3.

as the primary metric, consistent with prior works.

(1) Exact Brute-force Retrieval. Following prior work such as BEIR (Thakur et al., 2021), InstructIR (Oh et al., 2023), FollowIR (Weller et al., 2024), and MAIR (Zhang et al., 2024b), we compute similarity scores between each user query and all passages in the corpus, retrieving the top k most similar ones. This setting reflects model performance under ideal retrieval conditions.

(2) Approximate Nearest Neighbor (ANN) Retrieval. For large-scale corpora, brute-force retrieval is computationally infeasible. A common solution is a multi-stage architecture, where an ANN search retrieves a candidate set of passages, which are then re-ranked for final output (Tu et al., 2020; Macdonald and Tonellotto, 2021). To reflect real-world large-scale disaster information retrieval scenarios, we also evaluate model performance during the candidate generation stage using ANN search. We adopt Usearch (Vardanian, 2023), a high-performance, memory-efficient library based on the HNSW (hierarchical navigable small world) algorithm (Malkov and Yashunin, 2018), to retrieve top k passages per query using precomputed embeddings. For fair comparison, k is set to match the value used in exact search.

6 Evaluation Results

6.1 Overall Performance

Table 3 summarizes the overall performance of all 30 evaluated models across all queries in DisastIR, with detailed results for each search task provided in Appendix H. The `inf-retriever-v1` model ranks highest in both exact and ANN search settings, followed closely by `SFR-Embedding-Mistral` (1.52% and 0.22% lower, respectively). Among non-XL models, `multilingual-e5-large` performs best, reaching 92.83% and 92.79% of the top model's performance. Notably, **the lightweight `e5-small-v2` model (33M parameters) achieves 90.91% and 91.05% of the top model's performance, despite being 212 times smaller in size**.

The `Snowflake-arctic-embed-s` model shows the largest performance drop (11.63%) under ANN search compared to exact search (Table 3). Most models exhibit drops within 5%; only five exceeded this margin, four of which belong to snowflake family, indicating strong robustness when switching from exact to ANN search in DisastIR. Among the top-performing models analyzed above, `SFR-Embedding-Mistral` has the smallest drop (1.08%), while `multilingual-e5-large` has the largest (2.41%). All subsequent analyses are based on exact search; analyses under ANN search can be conducted similarly.

6.2 Performance across all 48 Tasks

Figure 3 presents the performance distribution of all evaluated models across all 48 search tasks. While `inf-retriever-v1` achieves the highest average NDCG@10, its median performance is lower than that of `SFR-Embedding-Mistral`, and it exhibits greater variability across tasks, as reflected by a larger interquartile range (IQR). This suggests that `inf-retriever-v1` is less stable across diverse search tasks in DisastIR.

As shown in Figure 5, **no single model consistently outperforms others across all 48 tasks**. Instead, top performance is distributed

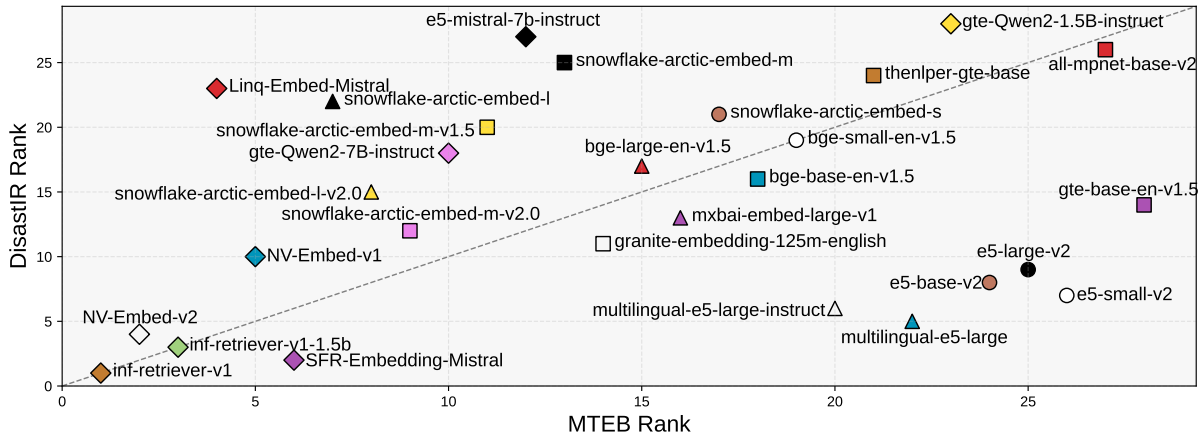

Figure 4: Comparison between DisastIR and MTEB model rankings. Legend shapes indicate the model size bin: ♦ XL, ▲ Large, ■ Medium, ● Small.

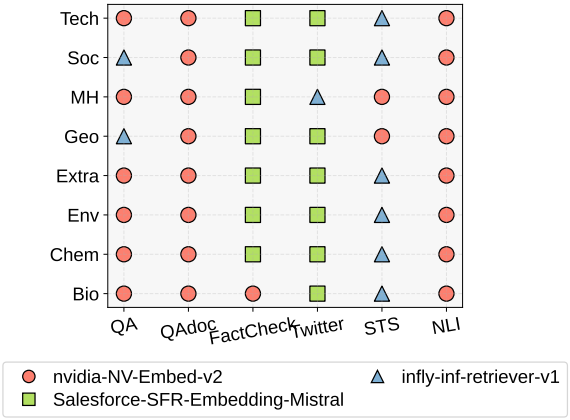

Figure 5: Best-performing models in each search task.

among three models: `inf-retriever-v1`, `SFR-Embedding-Mistral`, and `NV-Embed-v2`. `inf-retriever-v1` achieves the best results in only nine tasks, primarily in STS. **This highlights the complexity and diversity of disaster management-related retrieval tasks and reinforces the need for domain-specific IR models in real-world disaster management scenarios.** Appendix I provides additional analyses of model performance across 48 tasks.

6.3 Comparison with General Domain

Figure 4 compares model rankings in DisastIR and MTEB. Ranking value of each model is based on overall performance in DisastIR and official retrieval scores from the MTEB English leaderboard. The Spearman correlation between the two rankings is 0.225 ($p = 0.251$), indicating no significant correlation. This suggests that **strong performance on general-domain benchmarks does not guarantee effectiveness in disaster management-related retrieval**. For example, models like `Linq-Embed-Mistral` and `snowflake-arctic-embed-l` perform well in MTEB but poorly in DisastIR, while models from the `E5` family show the opposite trend.

Although `inf-retriever-v1` ranks highest in both DisastIR and MTEB, detailed analysis in Section 6.2 show that it achieves top performance in fewer than 20% of DisastIR tasks, mostly those related to STS. Furthermore, when computational resources are limited and large models are impractical to serve, relying solely on MTEB rankings for model selection, such as choosing `snowflake-arctic-embed-l`, may fail to retrieve critical or relevant content. **These discrepancies underscore the necessity of a domain-specific benchmark like DisastIR to guide retrieval model selection across different disaster management-related search tasks.**

7 Conclusion

In this work, we introduce and publicly release DisastIR, the first comprehensive retrieval benchmark for evaluating model performance in disaster management contexts. DisastIR consists of 9,600 user queries and more than 1.3 million labeled query-passage pairs, spanning 48 retrieval tasks defined by six search intents and eight general disaster event types, covering 301 specific event types.

Using DisastIR, we evaluate 30 SOTA open-source retrieval models under both exact and ANN search settings. Our findings provide practical guidance for selecting appropriate IR models based on task type and computational constraints, supporting timely and effective access to critical information in disaster management scenarios.

Limitations

While DisastIR represents a significant step toward domain-specific evaluation in disaster information retrieval, several aspects merit further enhancement. DisastIR currently focuses on English-language resources. Expanding DisastIR to multilingual settings would enable broader applicability. Furthermore, tables and figures in domain-specific PDF files may contain useful domain knowledge. Further study could consider extracting this critical information for evaluation set development.

Ethics Statement

DisastIR is designed to support disaster management by improving the evaluation and selection of retrieval models. All data used in the benchmark are sourced from publicly available materials, and no personally identifiable information is included. All contents generated by LLMs are evaluated by a human expert to ensure no offensive content is included in the DisastIR. We recognize potential risks associated with the misuse of retrieval models in disaster contexts, such as the spread of disinformation during crises. To mitigate these risks, DisastIR is intended solely for evaluation purposes and is released for research use only.

References

Reem Abbas and Todd Miller. 2025. Exploring communication inefficiencies in disaster response: Perspectives of emergency managers and health professionals. *International Journal of Disaster Risk Reduction*, 120:105393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. * sem 2013 shared task: Semantic textual similarity. In *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*, pages 32–43.

Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. 2021. Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing. In *Proceedings of the International AAAI conference on web and social media*, volume 15, pages 923–932.

Marwah Alaofi, Luke Gallagher, Mark Sanderson, Falk Scholer, and Paul Thomas. 2023. Can generative llms create query variants for test collections? an exploratory study. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 1869–1873.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic qa corpora generation with roundtrip consistency. *arXiv preprint arXiv:1906.05416*.

Janki Andharia. 2020. *Disaster studies*. Springer.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2022. Task-aware retrieval with instructions. *arXiv preprint arXiv:2211.09260*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, and 1 others. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Mrittunjay Basu and Dipankar Das. 2019. Extracting resource needs and availabilities from microblogs for aiding post-disaster relief operations. *International Journal of Disaster Risk Reduction*, 33:370–385.

Mrittunjay Basu and Dipankar Das. 2020. Neural relational inference for disaster multimedia retrieval. *Multimedia Tools and Applications*, 79(45):33691–33710.

Mrittunjay Basu, Dipankar Das, Richard McCreadie, Bhaskar Srivastava, and Takeshi Sakaki. 2017. Overview of the fire 2017 track: Information retrieval from microblogs during disasters (irmidis). In *Proceedings of the FIRE 2017 Working Notes*.

Helen Bromhead. 2021. Disaster linguistics, climate change semantics and public discourse studies: a semantically-enhanced discourse study of 2011 queensland floods. *Language Sciences*, 85:101381.

Hongliu Cao. 2025. Enhancing negation awareness in universal text embeddings: A data-efficient and computational-efficient approach. *arXiv preprint arXiv:2504.00584*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

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.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755*.

Thomas Diggelmann, Ori Yoran, Gabriela Csurka, Iryna Gurevych, and Pierre Massé. 2021. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Shangjia Dong, Amir Esmalian, Hamed Farahmand, and Ali Mostafavi. 2020. An integrated physical-social analysis of disrupted access to critical facilities and community service-loss tolerance in urban flooding. *Computers, Environment and Urban Systems*, 80:101443.

Naghmeh Farzi and Laura Dietz. 2024. Best in tau@ llmjudge: Criteria-based relevance evaluation with llama3. *arXiv preprint arXiv:2410.14044*.

Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *ACM computing surveys (CSUR)*, 47(4):1–38.

Vimukthi Jayawardene, Thomas J Huggins, Raj Prasanna, and Bapon Fakhruddin. 2021. The role of data and information quality during disaster response decision-making. *Progress in disaster science*, 12:100202.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, and Chris Callison-Burch. 2021. Gooaq: Open question answering with diverse answer types. *arXiv preprint arXiv:2104.08727*.

Shubham Khosla, Bodhisattwa Prasad Majumder, Tanmoy Mitra, and Dipankar Das. 2017. Microblog retrieval for post-disaster relief: Applying and comparing neural ir models. In *Proceedings of the SIGIR 2017 Workshop on Neural Information Retrieval (Neu-IR)*.

Tanmay Kumar, Mrittunjay Basu, and Dipankar Das. 2023. Taqe: Tweet retrieval-based infrastructure damage assessment during disasters. *Multimedia Tools and Applications*, 82(1):727–755.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Katherine Langford and Jon Atle Gulla. 2024. Improving search and rescue planning and resource allocation through case-based and concept-based retrieval. *Journal of Intelligent Information Systems*.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024a. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, and 1 others. 2024b. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*.

Zhenyu Lei, Yushun Dong, Weiyu Li, Rong Ding, Qi Wang, and Jundong Li. 2025. Harnessing large language models for disaster management: A survey. *arXiv preprint arXiv:2501.06932*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Zhewei Liu, Natalie Coleman, Flavia Ioana Patrascu, Kai Yin, Xiangpeng Li, and Ali Mostafavi. 2024. Artificial intelligence for flood risk management: A comprehensive state-of-the-art review and future directions. *International Journal of Disaster Risk Reduction*, page 105110.

Craig Macdonald and Nicola Tonellotto. 2021. On approximate nearest neighbour selection for multi-stage dense retrieval. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3318–3322.

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

Luke Merrick. 2024. Embedding and clustering your data can improve contrastive pretraining. *arXiv preprint arXiv:2407.18887*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Alice Oh, Kalpesh Krishna, Eric Wallace, Yichong Zhao, Patrick Lewis, and Antoine Bosselut. 2023. Instructir: Making dense retrievers follow instructions. *Preprint*, arXiv:2305.14252.

Chanhee Park, Hyeonseok Moon, Chanjun Park, and Heuiseok Lim. 2025. Mirage: A metric-intensive benchmark for retrieval-augmented generation evaluation. *arXiv preprint arXiv:2504.17137*.

Hemant Purohit, Carlos Castillo, Fernando Diaz, Amit Sheth, and Patrick Meier. 2014. Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday*.

Hossein A Rahmani, Nick Craswell, Emine Yilmaz, Bhaskar Mitra, and Daniel Campos. 2024a. Synthetic test collections for retrieval evaluation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2647–2651.

Hossein A Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles LA Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. 2025. Judging the judges: A collection of llm-generated relevance judgements. *arXiv preprint arXiv:2502.13908*.

Hossein A Rahmani, Xi Wang, Emine Yilmaz, Nick Craswell, Bhaskar Mitra, and Paul Thomas. 2024b. Syndl: A large-scale synthetic test collection for passage retrieval. *arXiv preprint arXiv:2408.16312*.

Hossein A Rahmani, Emine Yilmaz, Nick Craswell, and Bhaskar Mitra. 2024c. Judgeblender: Ensembling judgments for automatic relevance assessment. *arXiv preprint arXiv:2412.13268*.

Thilina C Rajapakse and Maarten de Rijke. 2023. Improving the generalizability of the dense passage retriever using generated datasets. In *European Conference on Information Retrieval*, pages 94–109. Springer.

sentence-transformers. 2021. sentence-transformers-all-nli dataset. https://huggingface.co/datasets/sentence-transformers/all-nli. Accessed: 2025-04.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. 2022. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*.

Weiwei Sun, Zhengliang Shi, Jiulong Wu, Lingyong Yan, Xinyu Ma, Yiding Liu, Min Cao, Dawei Yin, and Zhaochun Ren. 2024. Mair: A massive benchmark for evaluating instructed retrieval. *arXiv preprint arXiv:2410.10127*.

Tianxiang Tang, Diyi Yang, and 1 others. 2024. Do we need domain-specific embedding models? an empirical investigation. *arXiv preprint arXiv:2409.18511*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Sebastian Riegler, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM)*, pages 3577–3589.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1930–1940.

Zhengkai Tu, Wei Yang, Zihang Fu, Yuqing Xie, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2020. Approximate nearest neighbor search and lightweight dense vector reranking in multi-stage retrieval architectures. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pages 97–100.

UNDRR. 2020. Hazard definition and classification review technical report. Technical report, United Nations Office for Disaster Risk Reduction, Geneva, Switzerland. Supported by BMZ and USAID. Chair: Professor Virginia Murray.

Ash Vardanian. 2023. USearch by Unum Cloud. DOI: https://doi.org/10.5281/zenodo.7949416.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Luyu Wang, Sewon Min, Eric Wallace, Ledell Wu, Xi Victoria Lin, Daniel Khashabi, Bill Yuchen Lin, and Hannaneh Hajishirzi. 2024a. Benchmarking retrieval-augmented generation for medicine. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Shuai Wang, Ekaterina Khramtsova, Shengyao Zhuang, and Guido Zuccon. 2024b. Feb4rag: Evaluating federated search in the context of retrieval augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 763–773.

Orion Weller, Jason Phang, Arman Cohan, and Kyle Lo. 2024. Followir: Instruction-following models are zero-shot retrievers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 566–592.

Cheng Wen, Xianghui Sun, Shuaijiang Zhao, Xiaoquan Fang, Liangyu Chen, and Wei Zou. 2023. Chathome: Development and evaluation of a domain-specific language model for home renovation. *arXiv preprint arXiv:2307.15290*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Yanqiao Zhu, May D Wang, Joyce C Ho, Chao Zhang, and Carl Yang. 2024. Bmretriever: Tuning large language models as better biomedical text retrievers. *arXiv preprint arXiv:2404.18443*.

Kai Yin, Chengkai Liu, Ali Mostafavi, and Xia Hu. 2024. Crisissense-llm: Instruction fine-tuned large language model for multi-label social media text classification in disaster informatics. *arXiv preprint arXiv:2406.15477*.

Kai Yin, Jianjun Wu, Weiping Wang, Der-Horng Lee, and Yun Wei. 2023. An integrated resilience assessment model of urban transportation network: A case study of 40 cities in china. *Transportation Research Part A: Policy and Practice*, 173:103687.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, and 1 others. 2024a. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669*.

Yu Zhang, Zhenghao Jiang, Liangming Pan, Yuxuan Zhang, Daxin Jiang, and Maosong Sun. 2024b. Mair: A multidomain benchmark for instruction-following information retrieval. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *Acm Computing Surveys (Csur)*, 51(2):1–36.

A Structural PDF File Processing Pipeline

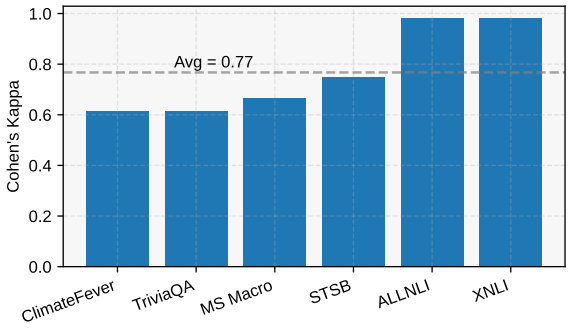

Figure 6: Cohen's kappa scores between LLM-based and human-annotated relevance labels across all LVHL datasets, as described in Section 4.2

All disaster management-related data (in PDF format) is obtained from publicly available sources with no personally identifiable information. Hence, explicit consent was not required. PDF files are collected using `googlesearch-python` (v1.3.0) and processed with `PyMuPDF` (v1.24.10) for content extraction. The extracted PDFs are then processed into text chunks through the following steps:

(1) Exact-URL Deduplication. The URL of each downloaded PDF is recorded, and duplicate documents are removed by identifying identical download links.

(2) Text Extraction and Preprocessing. Each PDF file is converted into plain text, where tables and figures are removed following the work of (Wen et al., 2023).

(3) Locality-Sensitive Hashing (LSH) Deduplication. After cleaning, we apply LSH-based near-duplicate detection to identify and remove documents with highly overlapping content.

(4) Semantic Chunking. Cleaned documents are segmented into semantically coherent text chunks. Each chunk is constrained to fewer than 256 tokens to optimize retrievability while maintaining semantic integrity.

(5) Embedding-based Near Deduplication. To further eliminate redundancy at the passage level, dense embeddings are computed for all chunks. An ANN index is built to retrieve the top-k nearest chunks, and pairs with cosine similarity above 0.9 are removed.

B Prompt Templates for Query Generation

Prompts for query generation based on disaster management-related passage under different search intents for QA, QAdoc, Twitter, FC, NLI, STS are in Tables 4, 5, 6, 7, 8, and 9.

C Prompt Templates for Relevance Labeling

This section presents the prompt templates used for LLM-based relevance judgments across six search intents, employing three prompting strategies: Zero-shot Direct Scoring, Chain-of-Thought Decomposed Reasoning, and Multi-Dimensional Attribute Scoring. For QA, QAdoc, and Twitter tasks, we adapt templates from Thomas et al. (2024); Farzi and Dietz (2024), as shown in Table 10. Based on these templates, we design relevance prompts for FC, NLI, and STS tasks, shown in Tables 11, 12, and 13, respectively. For STS, we adopt only Zero-shot Direct Scoring, as our preliminary experiments show it yields higher agreement with human labels (Cohen's kappa). The estimated cost of generating 9,600 user queries and labeling over 1.3 million query-passage using GPT-4o-mini API is about $1,400.

D Additional Analyses of Labeled query-passage pairs

As shown in Table 14, in certain retrieval tasks, such as NLI_Bio, NLI_Geo, the number of query-passage pairs assigned the highest relevance score is smaller than the number of user queries. This indicates that some queries do not have any passage in their candidate pool that is judged as fully relevant. For each query, we prompt an LLM to generate a directly relevant passage based on the associated domain passage and include it in the labeling pool (Section 3.4). However, the labeling results in Table 15 show that not all generated passages are considered fully relevant. This suggests that, even when guided by task-specific prompts, LLMs may produce passages that only partially address the query or fail to capture its key intent.

Many recent works have tried to employ LLM to generate synthetic training data to improve the quality of retrievers (Wang et al., 2023; Rajapakse and de Rijke, 2023; Xu et al., 2024; Lee et al., 2024b). This finding underscores the importance of consistency filtering (Alberti et al., 2019) to improve retrieval models' performance, as LLM will generate irrelevant pairs. This aligns with prior research highlighting the need for consistency filtering when leveraging LLM-generated data to train retrievers (Dai et al., 2022; Xu et al., 2024; Lee et al., 2024b).

E LVHL Dataset Construction

We use the names of 301 specific disaster event types as queries to search for disaster management-related user queries within each selected open-source dataset listed in Table 16. For each dataset, we first filter queries by keyword matching and then prompt an LLM to further remove queries that are irrelevant to disaster management. From the remaining queries, we randomly select up to 400 queries per dataset. The corresponding passage and relevance score in each source dataset are also included. This process results in the final query-passage pairs along with the human-annotated relevance scores used in the LVHL dataset for evaluating the agreement of LLM-based and human-annotated relevance scores. [4]

F LVHQ Dataset Construction

We sample 48 domain passages developed in Section 3.3, ensuring one passage per retrieval task and keeping all sampled passages different from those used in developing DisastIR. For each passage, a domain expert in the disaster management field is asked to read the passage and write a realistic user query that reflects a practical information need based on the content, resulting in 48 human-authored queries. The Human expert is given the same instructions for the query written (shown in Tables 4, 5, 6, 7, 8, 9) as those given to LLM to ensure fair comparison. In parallel, for the same set of passages, we also generate 48 queries using LLM in the same way as described in Section 3.4.

Each query, both human-authored and LLM-generated, is used to retrieve relevant passages from DisastIR corpus. As we have validated the agreement of LLM-based and Human-annotated relevance score in Section 4.2, all query-passage pairs are labeled in the same way as described in Section 3.5 and Section 3.6.

[4] LVHL is used solely to evaluate agreement between LLM and human annotations. It is not suitable for benchmarking retrieval models in the disaster management area, as most queries are drawn from training sets of the source datasets.

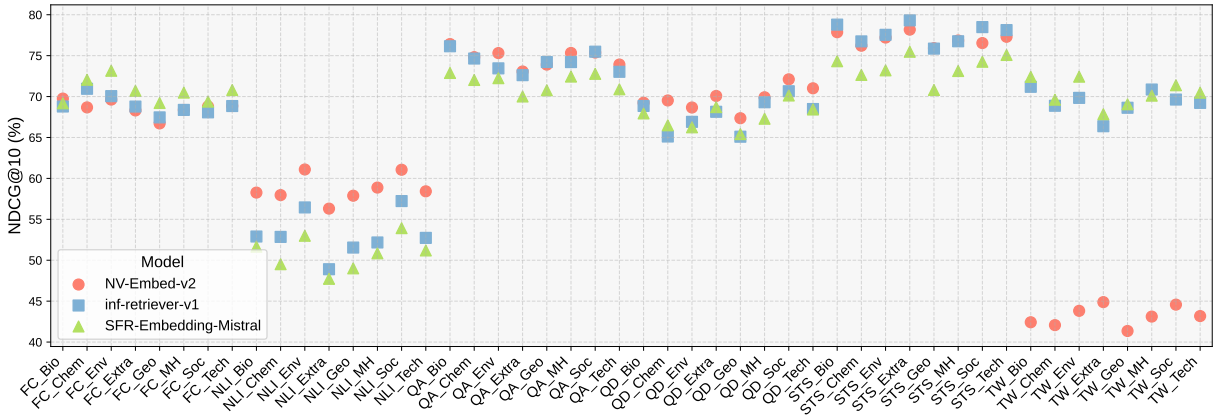

Figure 7: Performance of three top models across different tasks

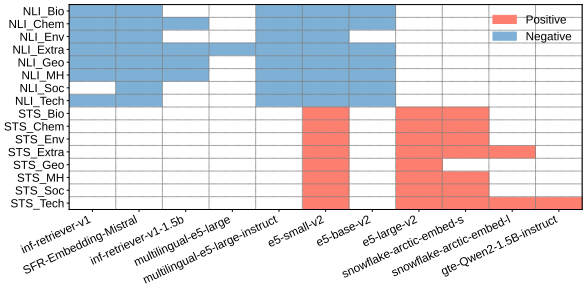

Figure 8: Distribution of outliers of evaluated models' performances

G Information of Evaluated Models and Model Implementation

Detailed information on all selected models is summarized in Table 17. The HuggingFace links and licenses of these models are in Table 18. The model parameter size is categorized as four levels: small (<109M), medium (109M - 305M), large (305M-1B), and extra large (XL) (> 1B).

For each model, we follow official implementation guidelines to generate normalized query and passage embeddings. All evaluations are conducted in a zero-shot setting, with input sequences truncated to 512 tokens and a task-specific instruction prepended to each query. All models are run on a single NVIDIA A6000 GPU using HuggingFace Transformers, following the configurations specified in the official implementations.

H Performance of Evaluated Models

Performance of all evaluated models in all 48 search tasks in DisastIR is shown in Tables 19, 20, 21, 22, and 23.

I Additional Analyses of Model Performance across 48 Tasks

`NV-Embed-v2` achieves the best performance on all QAdoc and NLI-related tasks (See Table 3 and Figure 5 in the main content). However, as shown in Figure 7, its poor results on Twitter-related tasks significantly lower its overall performance in DisastIR. This reflects its limitation in handling the informal, noisy, and contextually ambiguous nature of social media content. Given the importance of Twitter as a real-time, crowd-sourced information source during disasters (Alam et al., 2021; Yin et al., 2024; Lei et al., 2025), this weakness raises concerns about its reliability in real-world disaster response scenarios.

All three models perform poorly on NLI-related tasks, with the best achieving only an average score of 58.73 (Figure 7). Further analysis of outliers in the box plot (See Figure 3 in the main content) reveals that tasks causing significant performance drops consistently involve NLI search intents (Figure 8). This reveals a key limitation of current open-source SOTA retrievers, that they struggle with the complex reasoning required for NLI tasks in disaster contexts. Such limitations may lead to incorrect results or failure to retrieve critical information, which can negatively impact decision-making in disaster situations.

Information Need Generation Stage

"Brainstorm a list of useful text retrieval tasks where the goal is: Given a user question, retrieve passages that directly answer the question. Here are a few examples: Given a question about evacuation procedures during a flood, retrieve a passage that explains the steps involved. Given a question about the cause of infrastructure failure in a disaster, retrieve a passage identifying the cause. Given a question about relief funding timelines, retrieve a passage providing the relevant information. Guidelines: Each task description should be one sentence that clearly describes the user question and the kind of answer passages to be retrieved. Focus on real-world domains like disaster planning, relief logistics, early warning systems, community impact, government response, etc. Your output should be a JSON list of 3 strings, each describing a distinct and useful text retrieval task. Only output the list. Be creative. No explanations or additional content."

User Query Generation Stage

"You have been assigned a retrieval task: `{task}`. Your mission is to write one text retrieval example for this task in JSON format. The JSON object must contain the following keys:

- user_query: a string, a random user search query specified by the retrieval task.

- positive_document: a string, a relevant document for the user query.

Please adhere to the following guidelines:

- The user_query should be `{query_length}`, `{clarity}`.

- All documents must be created independent of the query. Avoid copying the query verbatim. It is acceptable if some parts of the positive_document are not topically related to the query.

- All documents should be `{num_words}` long.

- Do not provide any explanation in any document on why it is relevant or not relevant to the query. The query and documents must be realistic and inspired by real-world content (e.g., disaster management). All generated content should be in English no matter the provided content language is.

- Both the query and documents require `{difficulty}` level education to understand.

Your output must always be a JSON object only, do not explain yourself or output anything else. Be creative!"

Table 4: Prompt templates for user query generation in QA-related tasks. The `clarity` placeholder takes values: clear, understandable with some effort, and ambiguous. The `difficulty` placeholder includes: elementary school, high school, college, and PhD. For `query_length`, possible values are: less than 10 words, 5 to 20 words, less than 20 words, at least 50 words, and at least 150 words. The `num_words` placeholder takes values such as: at least 100 words, at least 200 words, at most 50 words, and 50 to 150 words.

Information Need Generation Stage

"Brainstorm a list of document retrieval tasks, where the goal is: Given a user query, retrieve documents that provide useful and relevant answers. Here are a few examples to get you started: Given a query about emergency evacuation procedures, retrieve a document that outlines the proper steps. Given a query asking how heatwaves affect public health, retrieve a document discussing the medical or environmental impacts. Given a query about funding for post-disaster recovery, retrieve a document describing the financial aid process. Given a query on how early warning systems reduce disaster risk, retrieve a document explaining their function and benefits. Guidelines: Each task should be a single sentence describing what the query is and what kind of document should be retrieved in response. Tasks should span a broad range of information needs, from facts to procedures to causal relationships. Focus on disaster management-related themes such as risk mitigation, emergency logistics, climate impact, institutional roles, and infrastructure damage. Your output should be a JSON list of 20 strings, each describing a distinct and useful text retrieval task. Only output the list. No explanations."

User Query Generation Stage

"You have been assigned a document retrieval task: `{task}`. Your mission is to write one example for this task in JSON format. The JSON object must include:

- `user_query`: a single, well-formed, natural language query that clearly asks for information based on the assigned task.

Guidelines: The query can be answered by the content provided in the following given paragraph. The query should reflect a realistic information need in the disaster management domain. Avoid generic or overly broad questions—make the query specific and grounded in actual scenarios (e.g., logistics, policies, actions). Use language inspired by real-world usage, such as what a policymaker, journalist, or emergency planner might ask. Output only a single JSON object with a `user_query` field. No extra formatting, documents, or explanations. Be clear, informative, and realistic!"

Table 5: Prompt templates for the user query generation for QAdoc-related search task. The `clarity` placeholder takes values: clear, understandable with some effort, and ambiguous. The `difficulty` placeholder includes: elementary school, high school, college, and PhD. For `query_length`, possible values are: less than 10 words, 5 to 20 words, less than 20 words, at least 50 words, and at least 150 words. The `num_words` placeholder takes values such as: at least 100 words, at least 200 words, at most 50 words, and 50 to 150 words.

Information Need Generation Stage

"Brainstorm a list of entity retrieval tasks where the goal is to retrieve tweets that mention or provide relevant information about one or more entities (e.g., organizations, people, locations, events) found in the query. Here are a few examples to inspire your thinking: Given a query referencing "UNICEF," retrieve tweets about their emergency relief efforts. Given a query about "Cyclone Mocha," retrieve tweets reporting its impact or aftermath. Given a query that mentions "World Health Organization," retrieve tweets that discuss their role in disaster health responses. Given a query with "Manila," retrieve tweets about disaster conditions or relief actions in that location. Guidelines: Each task should be a single sentence describing a situation where an entity is referenced and tweets related to that entity should be retrieved. Focus on disaster management-related entities such as emergency response agencies, international organizations, locations, events, or key figures. Encourage diversity in topics: ground response, aid distribution, weather events, infrastructure failure, etc. Your output should be a JSON list of about 3 strings, each describing a different NER Twitter retrieval task. No explanations or extra formatting. Be concise, diverse, and realistic."

User Query Generation Stage

"You have been assigned an entity-tweet retrieval task: `{task}`. Your mission is to write one example for this task in JSON format. The JSON object must include:

- `query`: a sentence that mentions one or more disaster management-related entities.
- `positive_tweet`: a tweet that provides relevant and informative content about the mentioned entity or entities.

Guidelines: The query should clearly mention a recognizable entity tied to the disaster domain. The positive_tweet should be informal, observational, or emotional—realistic Twitter-style language providing relevant details about the entity. All content should be inspired by disaster management-related themes such as rescue missions, weather events, humanitarian aid, response coordination, etc. Your output must be a single JSON object only. No explanation, no formatting beyond JSON. Keep it realistic and natural in tone."

Table 6: Prompt templates for the user query generation for Twitter-related search task. The `clarity` placeholder takes values: clear, understandable with some effort, and ambiguous. The `difficulty` placeholder includes: elementary school, high school, college, and PhD. For `query_length`, possible values are: less than 10 words, 5 to 20 words, less than 20 words, at least 50 words, and at least 150 words. The `num_words` placeholder takes values such as: at least 100 words, at least 200 words, at most 50 words, and 50 to 150 words.

Information Need Generation Stage

"Brainstorm a list of fact-checking retrieval tasks where the goal is: Given a claim, retrieve documents that either support or refute the claim, while distinguishing them from topically similar documents that do not address the claim's veracity. Here are a few examples to guide your ideas: Given a claim about the effectiveness of early warning systems during floods, retrieve documents that either support or refute the claim. Given a claim about the number of people displaced by a recent earthquake, retrieve evidence that verifies or challenges it. Given a claim about the government's relief distribution timeline, retrieve text that affirms or contradicts the stated timeline. Given a claim about the relationship between climate change and disaster frequency, retrieve relevant supporting or refuting content. Guidelines: Each task should be one sentence and describe what the claim is about and what kind of evidence is needed (support or refute). Base the topics on real-world domains such as disaster management, humanitarian aid, policy, climate, health impacts, etc. The tasks should vary in specificity and format (e.g., statistical claim, causal claim, factual assertion). Your output should be a JSON list of about 3 strings, each string representing a distinct fact-checking retrieval task. Output only the list. No explanations. Be creative and diverse in topic!"

User Query Generation Stage

"You have been assigned a fact-checking retrieval task: `{task}`. Your mission is to write one fact-checking retrieval instance in JSON format. The JSON object must contain:

- `claim`: a short, factual or semi-factual statement (assertion) related to the task.
- `positive_document`: a paragraph that supports or refutes the claim.

Guidelines: The claim should be `{query_length}`, `{clarity}`. The positive document must clearly support or refute the claim—either is acceptable. The claim should be clear, concise, and specific—not overly vague or too broad. Use examples from realistic disaster management-related content: climate events, emergency response, humanitarian relief, damage estimates, etc. All positive documents should be `{num_words}` long. All generated content should be in English no matter the provided content language is. Both the claim and documents must be understandable with `{difficulty}` level education. Output only a single JSON object. No additional text. Be precise and creative!"

Table 7: Prompt templates for the user query generation for fact-checking related search task. The `clarity` placeholder takes the values: clear, understandable with some effort, and ambiguous. The `difficulty` placeholder includes: elementary school, high school, college, and PhD. The `query_length` placeholder accepts values such as: less than 10 words, 5 to 20 words, at least 10 words, at least 20 words, and at least 50 words. The `num_words` placeholder includes: at most 15 words, at most 50 words, 50 to 150 words, at most 100 words, and at least 100 words.

Information Need Generation Stage

"Brainstorm a list of Natural Language Inference (NLI) retrieval tasks. In these tasks, the objective is: Given a premise sentence from a paragraph (e.g., about disaster management), retrieve a hypothesis sentence that is logically entailed by the premise. Here are a few examples to inspire your creativity: Given a sentence describing a government emergency response, retrieve a hypothesis that reflects an outcome or implication of that action. Given a statement about climate-induced hazards, retrieve a hypothesis summarizing the likely impact. Given a factual description of infrastructure damage, retrieve a hypothesis about the services affected. Given a claim about disaster preparedness strategies, retrieve a hypothesis that is logically supported by it. Guidelines: Each task description should be one sentence and should clearly specify the type of premise and the nature of the entailed hypothesis. Tasks should be generalizable across topics but inspired by domains such as climate, crisis response, risk, logistics, etc. Be diverse in topic and formality: from news-like to academic to conversational. Your output should be a JSON list of about 3 strings, each describing a different NLI retrieval task. Output only the list of task descriptions, no explanations."

User Query Generation Stage

"You have been assigned an NLI retrieval task: `{task}`. Your mission is to write one example for this task in JSON format. The JSON object must include:

- `premise`: a sentence drawn or inspired from a paragraph (e.g., about disaster management).
- `entailed_hypothesis`: a sentence that must logically follow from the premise.

Guidelines: The premise should be `{query_length}`, `{clarity}`. Use realistic examples from domains like climate risk, emergency response, infrastructure, health, or logistics, etc. Ensure the entailed hypothesis is non-trivial and clearly follows from the premise. Avoid word-for-word overlap between the sentences unless necessary for clarity. `entailed_hypothesis` should be `{num_words}` long. All generated content should be in English no matter the provided content language is. All contents require `{difficulty}` level education to understand and should be diverse in terms of topic and length. Output a single JSON object only. Do not explain yourself or add anything else. Be creative and accurate!"

Table 8: Prompt templates for the user query generation for NLI-related search task. The `clarity` placeholder takes values: clear, understandable with some effort, and ambiguous. The `difficulty` placeholder includes: elementary school, high school, college, and PhD. The `query_length` placeholder accepts values such as: less than 10 words, 5 to 20 words, at least 20 words, at least 50 words, and at least 150 words. The `num_words` placeholder includes: less than 10 words, 5 to 20 words, at least 20 words, at least 50 words, and at most 50 words.

Information Need Generation Stage

"Brainstorm a list of similar sentence retrieval tasks where the goal is: Given a sentence, retrieve other sentences that express the same or very similar meaning (paraphrases or semantically equivalent expressions). Here are a few examples to inspire your ideas: Given a sentence describing the impact of a flood, retrieve other sentences that paraphrase or closely restate the same impact. Given a sentence about the steps taken during emergency evacuation, retrieve sentences that express the same process using different wording. Given a sentence about climate-related disasters increasing in frequency, retrieve other sentences conveying the same trend. Given a factual statement about relief distribution, retrieve sentences that express the same fact using alternate phrasing. Guidelines: Each task should be written in one sentence and describe the kind of sentence (source) and what type of similar sentences should be retrieved. Focus on disaster management-related themes such as risk, policy, action, climate, or aid—but vary topics for diversity. Your output should be a JSON list of about 3 strings, each one describing a distinct similar sentence retrieval task. Output the list only. No explanations. Be creative and precise."

User Query Generation Stage

"You have been assigned a similar sentence retrieval task: `{task}`. Your mission is to write one example for this task in JSON format. The JSON object must contain:

- `query`: a single sentence that expresses a specific idea.
- `positive`: a sentence that expresses the same meaning as the query (semantic equivalence or high similarity).

Guidelines: The query should be `{query_length}`, `{clarity}`. The query and positive should be semantically equivalent, possibly using different wording or structure. Avoid copy-paste or trivial rewordings—be realistic and diverse. Use examples inspired by real-world disaster management-related content: emergency protocols, environmental impact, infrastructure, humanitarian response, etc. All positive documents should be `{num_words}` long. All generated content should be in English no matter the provided content language is. Both the query and documents must be understandable with `{difficulty}` level education. Output only a single JSON object. No explanation. Make it high-quality and realistic!"

Table 9: Prompt templates for the user query generation for STS-related search task. The `clarity` placeholder takes values: clear, understandable with some effort, and ambiguous. The `difficulty` placeholder includes: elementary school, high school, college, and PhD. The `query_length` placeholder accepts values such as: less than 10 words, 5 to 20 words, at least 50 words, and at most 50 words. The `num_words` placeholder includes: less than 10 words, 5 to 20 words, at least 50 words, and at most 50 words.

[1] Zero-shot Direct Scoring

You are a search quality rater evaluating passage relevance based on detailed instructions and outputting JSON.

Given a query and a passage, provide a score (0–3): 0 = Irrelevant, 1 = Related, 2 = Highly relevant, 3 = Perfectly relevant.

Important: 1 if somewhat related but not completely, 2 if important info + extra, 3 if only refers to topic.

Query: `{query}`, **Passage:** `{passage}`. Consider intent, content match (M), trustworthiness (T), then decide final score (O).
Output MUST be JSON: `{"final_score": <0-3>}`

[2] Chain-of-Thought Decomposed Reasoning

PHASE 1 – Answer Presence Prediction:

Instruction: Given a passage and a query, predict whether the passage includes an answer to the query by producing either "Yes" or "No". **Query:** `{query}`, **Passage:** `{passage}`, Answer:

Output your prediction as JSON: `{"has_answer": "Yes"}` or `{"has_answer": "No"}`

PHASE 2 – Fine-grained Criterion Scoring:

System prompt: Please assess how well the provided passage meets specific criteria in relation to the query. Use the following scoring scale (0–3) for evaluation: 0: Not relevant at all / No information provided. 1: Marginally relevant / Partially addresses the criterion. 2: Fairly relevant / Adequately addresses the criterion. 3: Highly relevant / Fully satisfies the criterion.
Your output MUST be JSON: `{"criterion_score": <0-3>}`

User prompt: Please rate how well the given passage meets the `{criterion_name}` criterion in relation to the query. The output should be a single score (0–3) indicating `{criterion_definition}`. **Query:** `{Query}`, **Passage:** `{Passage}`, Output JSON: `{"criterion_score": <integer_score_0_to_3>}`

Note: placeholders for `{criterion_name}` *and* `{criterion_definition}` *are:*

`{Exactness: "how precisely the passage answers the query", Coverage: "proportion of content discussing the query"}`.

PHASE 3 – Final Relevance Scoring:

System prompt when Phase 1 = "Yes": You are a search quality rater. Provide a final relevance score (2 or 3):
2 = Highly relevant, 3 = Perfectly relevant. Your output MUST be JSON: `{"relevance_score": <2_or_3>}`

System prompt when Phase 1 = "No": You are a search quality rater. Provide a final relevance score (0 or 1).
0 = Irrelevant, 1 = Related. Your output MUST be JSON: `{"relevance_score": <0_or_1>}`

User prompt when Phase 1 = "Yes": The passage is relevant. Rate how relevant (2 or 3). **Query:** `{Query}`, **Passage:** `{Passage}`, Output JSON: `{"relevance_score": <2_or_3>}`

User prompt when Phase 1 = "No": The passage is irrelevant. Rate how irrelevant (0 or 1).**Query:** `{Query}`, **Passage:** `{Passage}`, Output JSON: `{"relevance_score": <0_or_1>}`

[3] Multi-Dimensional Attribute Scoring

PHASE 1 – Sub-criterion Scoring:

System prompt: Please assess how well the provided passage meets specific criteria in relation to the query. Use the following scoring scale (0–3): 0 = Not relevant at all / No information provided. 1 = Marginally relevant / Partially addresses the criterion. 2 = Fairly relevant / Adequately addresses the criterion. 3 = Highly relevant / Fully satisfies the criterion.
Your output MUST be JSON: `{"criterion_score": <0-3>}`

User prompt: Please rate how well the given passage meets the `{criterion_name}` criterion in relation to the query. The output should be a single score (0–3) indicating `{criterion_definition}`.
Query: `{query}`, **Passage:** `{passage}`, Output JSON: `{"criterion_score": <integer_score_0_to_3>}`

Note: placeholders for `{criterion_name}` *and* `{criterion_definition}` *are:*

`{Exactness: "how precisely the passage answers the query", Coverage: "proportion of content discussing the query", Topicality: "subject alignment between passage and query", Contextual Fit: "presence of relevant background information"}`.

PHASE 2 – Final Relevance Aggregation:

System prompt: You are a search quality rater evaluating overall relevance. Given a query, passage, and sub-scores, provide a final score (0–3): 3 = Perfectly relevant, 2 = Highly relevant, 1 = Related, 0 = Irrelevant. Your output MUST be JSON: `{"final_relevance_score": <0-3>}`

User prompt: Please rate how relevant the passage is based on the given sub-scores.**Query:** `{query}`, **Passage:** `{passage}`, Exactness: LLM score, Coverage: LLM score, Topicality: LLM score, Contextual Fit: LLM score.
Output your final rating as JSON: `{"final_relevance_score": <integer_score_0_to_3>}`

Table 10: LLM relevance judgment prompt templates for QA, QAdoc, and Twitter-related search tasks.

[1] Zero-shot Direct Scoring

System prompt: You are a search quality rater evaluating evidence for fact-checking and outputting JSON.

User prompt: Given a claim and a passage, score relevance (0–3): 0 = Irrelevant, 1 = Related (no help), 2 = Relevant (unclear/mixed), 3 = Direct support/refutation.

Important: 1 if related but no help, 2 if important info + noise, 3 if clearly supports/refutes.

Claim: `{Query}`, **Passage:** `{Passage}`. Consider intent, support/refutation (M), trustworthiness (T), then decide final score (O). Output MUST be JSON: `{"final_score": <0_to_3>}`

[2] Chain-of-Thought Decomposed Reasoning

PHASE 1 – Answer Presence Prediction:

Instruction: Given a passage and a claim, predict whether the passage includes information that supports or refutes the claim by producing either "Yes" or "No".

Claim: `{Query}`, **Passage:** `{Passage}`. Output JSON: `{"has_answer": "Yes"}` or `{"has_answer": "No"}`

PHASE 2 – Fine-grained Criterion Scoring:

System prompt: Please assess how well the provided passage serves as evidence for evaluating the claim. Use the following scoring scale (0–3): 0 = Not relevant at all / No information provided. 1 = Marginally relevant / Partially addresses the criterion.2 = Fairly relevant / Adequately addresses the criterion. 3 = Highly relevant / Fully satisfies the criterion. Your output MUST be JSON: `{"criterion_score": <0_to_3>}`

User prompt: Please rate how well the given passage meets the `{criterion_name}` criterion in relation to the claim. The output should be a single score (0–3) indicating `{criterion_definition}`.

Claim: `{Query}`, **Passage:** `{Passage}`. Output JSON: `{"criterion_score": <integer_score_0_to_3>}`

Note: placeholders for `{criterion_name}` *and* `{criterion_definition}` *are:*

`{Exactness: "how precisely the passage supports or refutes the claim", Coverage: "the extent to which the passage discusses content directly relevant to the claim",}`. Each criterion evaluated independently.

PHASE 3 – Final Relevance Scoring:

System prompt when Phase 1 = "Yes": You are a rater evaluating evidence for fact-checking. Score (2 or 3): 2 = Highly relevant (some support/refutation), 3 = Perfectly relevant (direct support/refutation). Your output MUST be JSON: `{"relevance_score": <2_or_3>}` *System prompt when Phase 1 = "No":* You are a rater evaluating evidence for fact-checking. Score (0 or 1): 0 = Irrelevant, 1 = Related (no support/refutation). Your output MUST be JSON: `{"relevance_score": <0_or_1>}`

User prompt when Phase 1 = "Yes": Passage is relevant. Rate its relevance (2 or 3). **Claim:** `{Query}`, **Passage:** `{Passage}`. Output JSON: `{"relevance_score": <2_or_3>}`

User prompt when Phase 1 = "No":

Passage irrelevant for evidence. Rate its relevance (0 or 1). **Claim:** `{Query}`, **Passage:** `{Passage}`. Output JSON: `{"relevance_score": <0_or_1>}`

[3] Multi-Dimensional Attribute Scoring

PHASE 1 – Sub-criterion Scoring:

System prompt: Please assess how well the provided passage serves as evidence for evaluating the claim according to the following criteria. Use a score (0–3): 0 = Not relevant at all / No information provided. 1 = Marginally relevant / Partially addresses the criterion. 2 = Fairly relevant / Adequately addresses the criterion. 3 = Highly relevant / Fully satisfies the criterion. Your output MUST be JSON: `{"criterion_score": <0_to_3>}`

User prompt: Please rate how well the given passage meets the `{criterion_name}` criterion in relation to the claim. The output should be a single score (0–3) indicating `{criterion_definition}`.

Claim: `{Query}`, **Passage:** `{Passage}`. Output JSON: `{"criterion_score": <0_to_3>}`

Note: placeholders for `{criterion_name}` *and* `{criterion_definition}` *are:*

`{Exactness: "how precisely the passage supports or refutes the claim", Coverage: "the extent to which the passage discusses content directly relevant to the claim", Topicality: "how closely the subject matter aligns with the claim topic", Contextual Fit: "how much relevant background/context is provided to verify the claim"}`.

PHASE 2 – Final Relevance Aggregation:

System prompt: You are a search quality rater evaluating evidence relevance for fact-checking. Given a claim, passage, and sub-scores, provide a final score (0–3): 3 = Perfectly relevant (direct support/refutation), 2 = Highly relevant (helps verification), 1 = Related (topic match, no verification aid), 0 = Irrelevant. Your output MUST be JSON: `{"final_relevance_score": <0_to_3>}`

User prompt: Please rate how relevant the given passage is to the claim based on the given scores.

Claim: `{Query}`, **Passage:** `{Passage}`, Exactness: LLM score, Coverage: LLM score, Topicality: LLM score, Contextual Fit: LLM score. Output JSON: `{"final_relevance_score": <0_to_3>}`

Table 11: LLM relevance judgment prompt templates for FC-related search tasks.

[1] Zero-shot Direct Scoring

System prompt: You are a rater evaluating entailment. Output only the final score in JSON.

User prompt: Given a premise and hypothesis, score entailment (0–3): 0 = Not entailed/Contradicted, 1 = Related not entailed, 2 = Mostly entailed, 3 = Perfectly entailed.

Important: 1 if related but not inferable, 2 if captures important implied content but not fully, 3 if clearly/fully supported.

Premise: `{Query}`, **Hypothesis:** `{Passage}`. Consider premise implications, logical following (E), info gaps, then decide final score (O). Output MUST be JSON: `{"final_score": <0_to_3>}`

[2] Chain-of-Thought Decomposed Reasoning

PHASE 1 – Answer Presence Prediction:

Instruction: Given a hypothesis and a premise, predict whether the hypothesis is entailed by the premise by producing either "Yes" or "No".

Premise: `{Query}`, **Hypothesis:** `{Passage}`.Output JSON: `{"has_answer": "Yes"}` or `{"has_answer": "No"}`

PHASE 2 – Fine-grained Criterion Scoring:

System prompt: Please assess how well the given hypothesis meets the `{criterion_name}` criterion in relation to the premise. Use a single score (0–3) indicating `{criterion_definition}`.

Premise: `{Query}`, **Hypothesis:** `{Passage}`. Output JSON: `{"criterion_score": <0_to_3>}`

Note: placeholders for `{criterion_name}` *and* `{criterion_definition}` *are:* `{Exactness: "how precisely the hypothesis is entailed by the premise", Coverage: "the extent to which the hypothesis reflects core information from the premise".`

PHASE 3 – Final Relevance Scoring:

System prompt when Phase 1 = "Yes": You are a rater evaluating entailment. Provide a final score (2 or 3): 2 = Mostly entailed, 3 = Perfectly entailed. Your output MUST be JSON: `{"relevance_score": <2_or_3>}`

System prompt when Phase 1 = "No": You are a rater evaluating entailment. Provide a final score (0 or 1): 0 = Not entailed/Contradicted, 1 = Related but not entailed. Your output MUST be JSON: `{"relevance_score": <0_or_1>}`

User prompt when Phase 1 = "Yes": Hypothesis is entailed. Rate how well (2 or 3). **Premise:** `{Query}`, **Hypothesis:** `{Passage}`. Output JSON: `{"relevance_score": <2_or_3>}`

User prompt when Phase 1 = "No": Hypothesis is not entailed. Rate how (0 or 1). **Premise:** `{Query}`, **Hypothesis:** `{Passage}`. Output JSON: `{"relevance_score": <0_or_1>}`

[3] Multi-Dimensional Attribute Scoring

PHASE 1 – Sub-criterion Scoring:

System prompt: Please assess how well the hypothesis is entailed by the premise according to the following criteria. Use a score (0–3): 0 = Not relevant at all / No information provided. 1 = Marginally relevant / Partially addresses the criterion. 2 = Fairly relevant / Adequately addresses the criterion. 3 = Highly relevant / Fully satisfies the criterion. Your output MUST be JSON: `{"criterion_score": <0_to_3>}`

User prompt: Please rate how well the given hypothesis meets the `{criterion_name}` criterion in relation to the premise. The output should be a single score (0–3) indicating `{criterion_definition}`.

Premise: `{Query}`, **Hypothesis:** `{Passage}`. Output JSON: `{"criterion_score": <0_to_3>}`

Note: placeholders for `{criterion_name}` *and* `{criterion_definition}` *are:* `{Exactness: "how precisely the hypothesis is entailed by the premise", Coverage: "the extent to which the hypothesis reflects core information from the premise", Topicality: "how closely the subject matter of the hypothesis aligns with that of the premise", Contextual Fit: "how well the hypothesis fits within the context or background established by the premise"}`.

PHASE 2 – Final Relevance Aggregation:

System prompt: You are a search quality rater evaluating entailment. Given a premise, hypothesis, and sub-scores, provide a final score (0–3): 3 = Perfectly entailed, 2 = Mostly entailed, 1 = Related but not entailed, 0 = Not entailed/Contradicted. Your output MUST be JSON: `{"final_relevance_score": <0_to_3>}`

User prompt: Please rate how well the hypothesis is entailed by the premise based on the given scores.

Premise: `{Query}`, **Hypothesis:** `{Passage}`, Exactness: LLM score, Coverage: LLM score, Topicality: LLM score, Contextual Fit: LLM score. Output JSON: `{"final_relevance_score": <0_to_3>}`

Table 12: LLM relevance judgment prompt templates for NLI-related search tasks.

[1] Zero-shot Direct Scoring for STS

System prompt: You are a semantic-similarity rater. Output JSON with a score (0–5) where 0 = unrelated, 5 = semantically equivalent.

User prompt: Rate the semantic similarity 0–5: 0 = unrelated | 1 = slight | 2 = partial | 3 = moderate | 4 = high | 5 = equivalent

Sentence A: `{input1}` **Sentence B:** `{input2}`

Output JSON: `{"final_similarity_score": <0_to_5>}`

Table 13: LLM relevance judgment prompt templates for STS-related search tasks.

Search Intent	Event Type	rel=0	rel=1	rel=2	rel=3	rel=4	rel=5
FC	Bio	17987	3034	1952	1014	0	0
	Chem	16345	3845	2746	1656	0	0
	Env	16739	4243	3235	1588	0	0
	Extra	16867	3748	2436	1312	0	0
	Geo	20151	4126	2549	1038	0	0
	MH	18671	4283	2601	1115	0	0
	Soc	20757	3287	2707	1099	0	0
	Tech	18651	3905	2958	1245	0	0
NLI	Bio	19031	5388	1284	193	0	0
	Chem	19391	5981	2151	333	0	0
	Env	16130	6823	2041	213	0	0
	Extra	19372	5437	1378	212	0	0
	Geo	20557	6098	1404	151	0	0
	MH	19151	6100	1587	214	0	0
	Soc	19597	5570	1612	218	0	0
	Tech	21240	5540	1430	184	0	0
QA	Bio	19843	3073	3099	636	0	0
	Chem	18823	3411	3662	989	0	0
	Env	18073	3838	4036	738	0	0
	Extra	19293	3480	3300	734	0	0
	Geo	19478	3699	3373	590	0	0
	MH	20241	3819	3689	673	0	0
	Soc	19832	3206	3551	527	0	0
	Tech	19803	4005	3731	505	0	0
QAdoc	Bio	18615	4424	2112	184	0	0
	Chem	18254	4871	2604	303	0	0
	Env	16710	5900	3125	195	0	0
	Extra	17120	5394	2749	335	0	0
	Geo	19182	5157	2075	159	0	0
	MH	19610	4898	2537	211	0	0
	Soc	15153	4981	2937	282	0	0
	Tech	19247	5042	2569	213	0	0
STS	Bio	9189	8880	5229	2303	1350	114
	Chem	8201	8400	5780	2670	1605	131
	Env	7276	8600	6381	2938	1744	109
	Extra	9625	8639	5151	2459	1322	117
	Geo	8815	10232	6709	2771	1182	107
	MH	8590	8809	6436	3072	1680	115
	Soc	9402	8349	5114	2575	1520	114
	Tech	7846	9316	5606	2686	1345	121
Twitter	Bio	29385	3650	1904	243	0	0
	Chem	27326	4259	2312	289	0	0
	Env	24757	5062	3058	366	0	0
	Extra	26604	4168	2084	346	0	0
	Geo	28491	4482	2257	273	0	0
	MH	27086	4255	2252	331	0	0
	Soc	28238	3472	1861	263	0	0
	Tech	26349	4457	2219	363	0	0

Table 14: Distribution of qrels scores rel=0 through rel=5 for each search task in DisastIR. “rel” represents relevance score. Only STS-related search task is labeled in 6 levels, with others labeled in 4 levels.

Search Intent	Event Type	rel=0	rel=1	rel=2	rel=3	rel=4	rel=5
FC	Bio	1	11	19	169	0	0
	Chem	0	5	15	180	0	0
	Env	0	2	14	184	0	0
	Extra	1	6	24	169	0	0
	Geo	1	9	21	169	0	0
	MH	0	4	26	170	0	0
	Soc	1	2	17	180	0	0
	Tech	0	6	16	178	0	0
NLI	Bio	8	33	78	81	0	0
	Chem	7	39	81	73	0	0
	Env	10	29	96	65	0	0
	Extra	20	40	79	61	0	0
	Geo	6	51	89	54	0	0
	MH	10	44	67	79	0	0
	Soc	5	52	77	66	0	0
	Tech	13	39	84	64	0	0
QA	Bio	1	3	33	163	0	0
	Chem	0	3	30	167	0	0
	Env	0	2	36	162	0	0
	Extra	0	4	27	169	0	0
	Geo	2	1	38	159	0	0
	MH	0	2	38	160	0	0
	Soc	0	5	30	165	0	0
	Tech	0	2	48	150	0	0
QAdoc	Bio	3	38	96	63	0	0
	Chem	6	45	85	64	0	0
	Env	1	35	102	62	0	0
	Extra	4	34	77	85	0	0
	Geo	2	44	81	73	0	0
	MH	4	36	74	86	0	0
	Soc	2	19	95	84	0	0
	Tech	3	28	79	90	0	0
STS	Bio	0	0	4	8	79	109
	Chem	0	0	1	3	80	116
	Env	0	0	1	7	88	104
	Extra	0	0	0	5	85	110
	Geo	0	0	0	4	94	102
	MH	0	0	0	6	88	106
	Soc	0	0	0	11	87	102
	Tech	0	0	2	8	83	107
Twitter	Bio	0	10	94	96	0	0
	Chem	1	19	98	82	0	0
	Env	0	12	103	85	0	0
	Extra	0	14	101	85	0	0
	Geo	0	9	95	96	0	0
	MH	4	9	74	113	0	0
	Soc	0	16	104	80	0	0
	Tech	0	19	94	87	0	0

Table 15: Distribution of query-generated relevant document relevance scores rel-0 through rel-5

Name	Intent	Link	#
MS Macro	QA	https://huggingface.co/datasets/microsoft/ms_marco	400
TriviaQA	QA	https://huggingface.co/datasets/sentence-transformers/trivia-qa-triplet	400
ALLNLI	NLI	https://huggingface.co/datasets/sentence-transformers/all-nli	400
XNLI	NLI	https://huggingface.co/datasets/mteb/xnli/viewer/en	400
STSB	STS	https://huggingface.co/datasets/sentence-transformers/stsb	400
ClimateFever	FC	https://huggingface.co/datasets/tdiggelm/climate_fever	400

Table 16: Overview of selected open-source datasets in LVHL. "#" represents the number of selected queries in the corresponding dataset.

Model Name	Param Size	Size Bin	Base Model	Embed. Size	MTEB Rank	Arch.
inf-retriever-v1	7B	XL	gte-Qwen2-7B-instruct	3584	1	decoder
NV-Embed-v2	7B	XL	Mistral-7B-v0.1	4096	2	decoder
inf-retriever-v1-1.5b	1.5B	XL	gte-Qwen2-1.5B-instruct	1536	3	decoder
Linq-Embed-Mistral	7B	XL	E5-mistral-7b-instruct	4096	4	decoder
NV-Embed-v1	7B	XL	Mistral-7B-v0.1	4096	5	decoder
SFR-Embedding-Mistral	7B	XL	E5-mistral-7b-instruct	4096	6	decoder
snowflake-arctic-embed-l	335M	Large	e5-large-unsupervised	1024	7	encoder
snowflake-arctic-embed-l-v2.0	568M	Large	gte-multilingual-mlm-base	1024	8	encoder
snowflake-arctic-embed-m-v2.0	305M	Medium	bge-m3-retromae	768	9	encoder
gte-Qwen2-7B-instruct	7B	XL	Qwen2-7B	3584	10	decoder
snowflake-arctic-embed-m-v1.5	109M	Medium	BERT-base-uncased	768	11	encoder
e5-mistral-7b-instruct	7B	XL	Mistral-7b	4096	12	decoder
snowflake-arctic-embed-m	109M	Medium	e5-unsupervised-base	764	13	encoder
granite-embedding-125m-english	125M	Medium	RoBERTa	768	14	encoder
bge-large-en-v1.5	335M	Large	–	1024	15	encoder
mxbai-embed-large-v1	335M	Large	–	1024	16	encoder
snowflake-arctic-embed-s	33M	Small	e5-unsupervised-small	384	17	encoder
bge-base-en-v1.5	109M	Medium	–	768	18	encoder
bge-small-en-v1.5	33M	Small	–	384	19	encoder
multilingual-e5-large-instruct	560M	Large	xlm-roberta-large	1024	20	encoder
thenlper-gte-base	109M	Medium	EBRT-base	768	21	encoder
multilingual-e5-large	560M	Large	xlm-roberta-large	1024	22	encoder
gte-Qwen2-1.5B-instruct	1.5B	XL	Qwen2-1.5B	1536	23	decoder
e5-base-v2	109M	Medium	bert-large-uncased	1024	24	encoder
e5-large-v2	335M	Large	bert-base-uncased	768	25	encoder
e5-small-v2	33M	Small	MiniLM	384	26	encoder
all-mpnet-base-v2	109M	Medium	microsoft/mpnet-base	768	27	encoder
gte-base-en-v1.5	137M	Medium	EBRT-base	768	28	encoder
gte-large-en-v1.5	434M	Large	EBRT-large	1024	–	encoder
llmrails-ember-v1	335M	Large	–	1024	–	encoder

Table 17: Information of all evaluated models. "–" means no publicly available information is available.

Model Name	Link	License
inf-retriever-v1	https://huggingface.co/infly/inf-retriever-v1	apache-2.0
NV-Embed-v2	https://huggingface.co/nvidia/NV-Embed-v2	cc-by-nc-4.0
inf-retriever-v1-1.5b	https://huggingface.co/infly/inf-retriever-v1-1.5b	apache-2.0
Linq-Embed-Mistral	https://huggingface.co/Linq-AI-Research/Linq-Embed-Mistral	cc-by-nc-4.0
NV-Embed-v1	https://huggingface.co/nvidia/NV-Embed-v1	cc-by-nc-4.0
SFR-Embedding-Mistral	https://huggingface.co/Salesforce/SFR-Embedding-Mistral	cc-by-nc-4.0
snowflake-arctic-embed-l	https://huggingface.co/Snowflake/snowflake-arctic-embed-l	apache-2.0
snowflake-arctic-embed-l-v2.0	https://huggingface.co/Snowflake/snowflake-arctic-embed-l-v2.0	apache-2.0
snowflake-arctic-embed-m-v2.0	https://huggingface.co/Snowflake/snowflake-arctic-embed-m-v2.0	apache-2.0
gte-Qwen2-7B-instruct	https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B-instruct	apache-2.0
snowflake-arctic-embed-m-v1.5	https://huggingface.co/Snowflake/snowflake-arctic-embed-m-v1.5	apache-2.0
e5-mistral-7b-instruct	https://huggingface.co/intfloat/e5-mistral-7b-instruct	mit
snowflake-arctic-embed-m	https://huggingface.co/Snowflake/snowflake-arctic-embed-m	apache-2.0
granite-embedding-125m-english	https://huggingface.co/ibm-granite/granite-embedding-125m-english	mit
bge-large-en-v1.5	https://huggingface.co/BAAI/bge-large-en-v1.5	apache-2.0
mxbai-embed-large-v1	https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1	apache-2.0
snowflake-arctic-embed-s	https://huggingface.co/Snowflake/snowflake-arctic-embed-s	mit
bge-base-en-v1.5	https://huggingface.co/BAAI/bge-base-en-v1.5	mit
bge-small-en-v1.5	https://huggingface.co/BAAI/bge-small-en-v1.5	mit
multilingual-e5-large-instruct	https://huggingface.co/intfloat/multilingual-e5-large-instruct	mit
thenlper-gte-base	https://huggingface.co/thenlper/gte-base	mit
multilingual-e5-large	https://huggingface.co/intfloat/multilingual-e5-large-instructt	apache-2.0
gte-Qwen2-1.5B-instruct	https://huggingface.co/Alibaba-NLP/gte-Qwen2-1.5B-instruct	mit
e5-base-v2	https://huggingface.co/intfloat/e5-base-v2	mit
e5-large-v2	https://huggingface.co/intfloat/e5-large-v2	mit
e5-small-v2	https://huggingface.co/intfloat/e5-small-v2	apache-2.0
all-mpnet-base-v2	https://huggingface.co/sentence-transformers/all-mpnet-base-v2	apache-2.0
gte-base-en-v1.5	https://huggingface.co/Alibaba-NLP/gte-base-en-v1.5	apache-2.0
gte-large-en-v1.5	https://huggingface.co/Alibaba-NLP/gte-large-en-v1.5	mit
llmrails-ember-v1	https://huggingface.co/llmrails/ember-v1	mit

Table 18: HuggingFace model links and licenses for all evaluated models.

	QA	QAdoc	Twitter	FC	NLI	STS	Avg.
Alibaba-NLP-gte-Qwen2-1.5B-instruct							
Biological	12.93	19.75	19.16	25.18	18.82	34.19	21.67
Chemical	14.74	20.20	19.57	25.92	15.82	26.73	20.50
Environmental	14.03	26.03	23.66	21.48	19.75	34.39	23.23
Extraterrestrial	13.13	21.89	15.62	27.12	17.38	30.09	20.87
Geohazard	14.68	17.35	18.51	21.79	15.51	34.44	20.38
Meteorological&hydrological	12.63	19.33	23.63	22.10	19.47	30.22	21.23
Societal	14.51	29.00	17.56	24.04	18.59	22.57	21.05
Technological	15.20	24.16	19.15	23.58	18.63	37.00	22.95
Avg.	13.98	22.21	19.61	23.90	18.00	31.20	21.48
Alibaba-NLP-gte-Qwen2-7B-instruct							
Biological	62.58	37.35	43.67	27.64	45.76	71.38	48.06
Chemical	64.42	38.19	41.06	31.31	46.84	70.43	48.71
Environmental	62.58	38.36	41.83	29.77	48.23	71.06	48.64
Extraterrestrial	58.55	34.42	37.55	31.45	46.54	73.40	46.99
Geohazard	61.12	36.31	40.13	28.31	47.12	68.06	46.84
Meteorological&hydrological	62.22	39.15	42.08	26.14	45.62	70.57	47.63
Societal	66.77	47.63	43.10	31.95	51.06	72.05	52.09
Technological	61.51	41.51	41.80	33.27	47.08	70.86	49.34
Avg.	62.47	39.12	41.40	29.98	47.28	70.98	48.54
Alibaba-NLP-gte-base-en-v1.5							
Biological	60.86	54.84	46.25	53.77	42.53	70.95	54.87
Chemical	60.99	55.23	43.96	53.63	37.30	70.23	53.56
Environmental	62.02	55.77	47.11	51.32	43.30	71.72	55.21
Extraterrestrial	59.48	53.38	49.06	53.30	37.65	72.35	54.20
Geohazard	60.10	54.26	45.96	49.82	36.83	70.45	52.90
Meteorological&hydrological	59.98	58.31	47.26	51.44	39.13	70.93	54.51
Societal	61.99	59.05	48.04	53.11	42.88	71.83	56.15
Technological	59.83	57.66	48.85	53.55	39.39	71.68	55.16
Avg.	60.66	56.06	47.06	52.49	39.88	71.27	54.57
Alibaba-NLP-gte-large-en-v1.5							
Biological	68.22	59.69	43.28	56.01	36.94	67.79	55.32
Chemical	70.28	59.30	38.93	56.44	33.85	68.50	54.55
Environmental	67.62	57.23	41.34	52.00	38.74	66.61	53.92
Extraterrestrial	65.28	56.57	42.45	51.24	32.11	68.85	52.75
Geohazard	65.65	52.61	35.16	49.45	30.81	64.75	49.74
Meteorological&hydrological	67.20	59.98	38.81	51.14	33.15	66.41	52.78
Societal	69.23	62.26	41.72	53.73	38.01	68.27	55.54
Technological	68.15	60.74	40.77	54.70	34.90	67.23	54.41
Avg.	67.70	58.55	40.31	53.09	34.81	67.30	53.63
BAAI-bge-base-en-v1.5							
Biological	51.58	53.85	50.86	58.41	37.78	65.45	52.99
Chemical	47.97	50.77	48.75	58.64	34.71	67.10	51.32
Environmental	50.63	51.44	47.75	60.93	41.00	63.61	52.56
Extraterrestrial	50.50	57.93	45.21	59.49	33.18	65.59	51.98
Geohazard	47.81	52.76	44.28	56.91	33.09	62.52	49.56
Meteorological&hydrological	49.27	53.06	47.10	59.05	34.13	62.09	50.78
Societal	50.00	59.17	45.32	55.97	36.82	62.81	51.68
Technological	49.56	52.77	46.99	58.39	33.90	66.24	51.31
Avg.	49.67	53.97	47.03	58.47	35.58	64.43	51.52
BAAI-bge-large-en-v1.5							
Biological	56.35	55.10	36.71	55.50	35.62	66.83	51.02
Chemical	56.24	51.49	33.57	56.73	34.65	66.02	49.78
Environmental	57.20	51.53	33.76	54.54	38.85	65.74	50.27
Extraterrestrial	58.77	57.02	33.03	58.30	32.68	66.38	51.03
Geohazard	55.15	53.62	30.65	53.17	35.16	63.51	48.54
Meteorological&hydrological	56.45	54.24	31.36	53.73	33.20	62.50	48.58
Societal	58.61	58.52	32.33	54.25	35.96	63.75	50.57
Technological	57.39	56.25	30.28	55.56	35.44	66.37	50.21
Avg.	57.02	54.72	32.71	55.22	35.19	65.14	50.00

Table 19: Performance of the first six evaluated models under six search intents and eight event types under the exact search setting. Part I

	QA	QAdoc	Twitter	FC	NLI	STS	Avg.
BAAI-bge-small-en-v1.5							
Biological	47.81	47.68	31.38	53.01	28.93	60.77	44.93
Chemical	45.25	42.32	30.68	55.45	27.44	57.61	43.12
Environmental	51.27	45.90	28.36	52.48	31.09	53.90	43.83
Extraterrestrial	48.32	45.50	26.64	53.35	25.57	58.43	42.97
Geohazard	46.74	42.91	25.79	50.88	25.91	55.30	41.25
Meteorological&hydrological	46.89	46.96	26.52	52.61	28.11	55.46	42.76
Societal	47.90	47.18	26.05	45.50	25.78	54.76	41.20
Technological	44.82	45.61	27.12	53.74	24.75	58.68	42.45
Avg.	47.38	45.51	27.82	52.13	27.20	56.86	42.82
Linq-AI-Research-Linq-Embed-Mistral							
Biological	35.16	31.88	27.70	43.16	28.81	50.09	36.13
Chemical	39.05	28.91	27.52	44.70	20.32	39.50	33.33
Environmental	42.55	33.30	30.97	36.88	30.97	50.49	37.53
Extraterrestrial	33.98	33.26	25.64	44.49	28.29	42.68	34.72
Geohazard	35.77	28.05	23.63	33.08	27.89	47.75	32.70
Meteorological&hydrological	33.17	28.16	28.04	32.61	28.43	44.46	32.48
Societal	35.09	36.96	23.60	38.97	30.61	39.84	34.18
Technological	30.12	30.78	25.34	41.78	23.06	50.72	33.63
Avg.	35.61	31.41	26.55	39.46	27.30	45.69	34.34
Salesforce-SFR-Embedding-Mistral							
Biological	72.89	67.94	72.41	69.18	51.66	74.32	68.07
Chemical	72.03	66.49	69.62	72.05	49.51	72.65	67.06
Environmental	72.25	66.24	72.43	73.16	52.99	73.20	68.38
Extraterrestrial	70.01	68.71	67.85	70.71	47.73	75.48	66.75
Geohazard	70.77	65.41	69.06	69.21	49.01	70.81	65.71
Meteorological&hydrological	72.44	67.29	70.11	70.49	50.84	73.12	67.38
Societal	72.78	70.13	71.38	69.38	53.92	74.26	68.64
Technological	70.89	68.42	70.48	70.80	51.20	75.08	67.81
Avg.	71.76	67.58	70.42	70.62	50.86	73.61	67.47
Snowflake-snowflake-arctic-embed-l							
Biological	41.34	30.98	16.73	33.28	35.21	54.92	35.41
Chemical	40.77	30.26	16.39	34.55	33.24	56.39	35.27
Environmental	43.56	34.71	17.68	34.42	40.28	56.81	37.91
Extraterrestrial	41.09	27.96	14.48	31.93	32.31	58.93	34.45
Geohazard	36.34	28.20	12.92	33.83	31.08	53.90	32.71
Meteorological&hydrological	37.75	25.59	14.99	33.70	34.90	57.75	34.11
Societal	42.99	35.02	16.49	28.39	36.37	57.69	36.16
Technological	42.70	30.53	12.85	31.51	33.57	58.20	34.89
Avg.	40.82	30.41	15.32	32.70	34.62	56.82	35.11
Snowflake-snowflake-arctic-embed-l-v2.0							
Biological	53.33	58.87	38.42	60.63	41.22	62.01	52.41
Chemical	56.40	57.57	40.62	61.11	39.01	63.87	53.10
Environmental	57.52	59.91	45.04	62.53	45.87	63.73	55.77
Extraterrestrial	54.74	56.70	36.38	60.14	37.86	62.82	51.44
Geohazard	54.21	57.26	34.76	59.82	39.80	60.19	51.01
Meteorological&hydrological	51.69	59.76	36.28	61.42	40.67	62.49	52.05
Societal	57.78	63.46	41.87	57.77	44.56	64.80	55.04
Technological	57.02	61.94	36.67	59.97	40.49	67.14	53.87
Avg.	55.34	59.43	38.75	60.42	41.19	63.38	53.09
Snowflake-snowflake-arctic-embed-m							
Biological	31.68	13.54	8.09	38.04	39.39	54.55	30.88
Chemical	33.35	14.43	9.76	35.38	36.26	58.17	31.23
Environmental	35.53	14.87	9.18	35.83	43.61	57.19	32.70
Extraterrestrial	33.40	14.31	8.29	35.18	39.67	58.59	31.57
Geohazard	31.14	14.32	6.15	35.78	36.75	54.06	29.70
Meteorological&hydrological	30.96	14.49	9.29	35.75	39.04	56.34	30.98
Societal	35.17	14.34	10.72	31.87	40.57	57.82	31.75
Technological	35.55	13.70	7.63	34.58	36.14	58.32	30.99
Avg.	33.35	14.25	8.64	35.30	38.93	56.88	31.23

Table 20: Performance of evaluated models under six search intents and eight event types under the exact search setting. Part II

	QA	QAdoc	Twitter	FC	NLI	STS	Avg.
Snowflake-snowflake-arctic-embed-m-v1.5							
Biological	20.41	30.10	17.56	48.74	42.93	63.69	37.24
Chemical	23.39	29.15	20.20	48.55	41.74	65.45	38.08
Environmental	28.64	33.80	21.74	51.30	49.00	65.53	41.67
Extraterrestrial	25.43	24.07	17.35	46.65	41.79	66.04	36.89
Geohazard	26.06	30.31	17.01	46.63	42.02	63.49	37.59
Meteorological&hydrological	22.61	28.98	17.43	48.37	42.51	64.42	37.39
Societal	29.55	35.48	17.74	47.59	42.61	64.06	39.51
Technological	29.77	32.56	17.47	48.12	41.71	66.91	39.42
Avg.	25.73	30.56	18.31	48.24	43.04	64.95	38.47
Snowflake-snowflake-arctic-embed-m-v2.0							
Biological	60.58	62.53	48.09	57.58	42.68	64.84	56.05
Chemical	61.22	61.01	47.12	59.78	40.78	64.72	55.77
Environmental	63.96	62.03	49.99	60.78	47.75	65.53	58.34
Extraterrestrial	60.83	61.17	46.86	58.68	39.05	65.74	55.39
Geohazard	60.38	60.30	47.90	57.05	43.33	63.02	55.33
Meteorological&hydrological	60.47	63.37	46.95	58.29	40.69	63.79	55.59
Societal	62.72	65.79	47.63	54.61	43.17	66.73	56.78
Technological	61.34	63.79	47.86	57.56	41.24	67.88	56.61
Avg.	61.44	62.50	47.80	58.04	42.34	65.28	56.23
Snowflake-snowflake-arctic-embed-s							
Biological	34.80	26.88	18.93	44.11	36.25	66.25	37.87
Chemical	33.78	28.03	20.01	45.40	35.16	66.72	38.18
Environmental	39.09	30.43	23.33	42.75	43.67	68.17	41.24
Extraterrestrial	35.15	21.60	16.03	41.94	34.80	66.06	35.93
Geohazard	35.78	25.98	16.63	41.53	36.37	63.31	36.60
Meteorological&hydrological	34.13	28.33	16.74	43.00	38.08	66.33	37.77
Societal	37.91	30.80	17.34	39.73	38.71	67.63	38.69
Technological	37.77	27.63	16.34	41.68	35.15	68.29	37.81
Avg.	36.05	27.46	18.17	42.52	37.27	66.59	38.01
ibm-granite-granite-embedding-125m-english							
Biological	65.94	61.42	50.92	62.06	46.86	71.23	59.74
Chemical	64.39	57.94	52.68	63.52	46.68	71.24	59.41
Environmental	65.87	60.32	47.60	64.84	52.70	73.30	60.77
Extraterrestrial	63.75	58.73	44.83	62.05	47.34	72.81	58.25
Geohazard	63.08	59.78	44.29	60.60	46.28	70.89	57.49
Meteorological&hydrological	65.83	61.04	45.60	62.71	45.98	71.02	58.70
Societal	66.36	65.53	44.21	62.98	50.31	71.93	60.22
Technological	63.95	63.52	47.03	63.49	48.21	73.06	59.88
Avg.	64.90	61.04	47.14	62.78	48.04	71.94	59.31
infly-inf-retriever-v1							
Biological	76.15	68.85	71.20	68.79	52.90	78.79	69.45
Chemical	74.65	65.12	68.88	70.95	52.85	76.74	68.20
Environmental	73.45	66.91	69.84	70.05	56.45	77.55	69.04
Extraterrestrial	72.61	68.14	66.38	68.78	48.90	79.31	67.35
Geohazard	74.22	65.10	68.63	67.46	51.55	75.86	67.14
Meteorological&hydrological	74.21	69.31	70.86	68.37	52.17	76.76	68.61
Societal	75.49	70.65	69.64	68.07	57.23	78.49	69.93
Technological	73.03	68.48	69.23	68.84	52.73	78.12	68.41
Avg.	74.23	67.82	69.33	68.91	53.10	77.70	68.52
infly-inf-retriever-v1-1.5b							
Biological	72.23	65.55	67.42	65.01	53.59	76.99	66.80
Chemical	71.18	63.09	65.04	67.03	51.93	75.97	65.71
Environmental	70.06	63.56	65.57	68.21	57.26	76.07	66.79
Extraterrestrial	67.37	61.83	63.32	65.93	51.12	77.02	64.43
Geohazard	71.49	62.52	65.70	65.56	52.29	75.73	65.55
Meteorological&hydrological	70.52	65.69	67.86	67.18	52.54	75.66	66.58
Societal	71.90	66.62	66.65	66.51	57.99	76.24	67.65
Technological	70.39	66.48	67.29	66.79	54.18	75.83	66.83
Avg.	70.64	64.42	66.11	66.53	53.86	76.19	66.29

Table 21: Performance of evaluated models under six search intents and eight event types under the exact search setting. Part III

	QA	QAdoc	Twitter	FC	NLI	STS	Avg.
intfloat-e5-base-v2							
Biological	67.85	62.93	58.90	62.08	47.03	74.82	62.27
Chemical	66.55	59.44	58.58	63.27	45.05	74.71	61.27
Environmental	67.45	63.11	60.69	64.54	49.66	74.99	63.41
Extraterrestrial	63.48	63.63	55.96	63.26	43.63	75.01	60.83
Geohazard	65.17	60.34	56.01	60.72	43.17	74.34	59.96
Meteorological&hydrological	66.14	62.64	59.13	62.21	44.88	74.08	61.52
Societal	64.92	67.05	59.00	60.68	46.63	74.02	62.05
Technological	64.58	65.38	58.65	61.44	44.27	75.15	61.58
Avg.	65.77	63.07	58.37	62.28	45.54	74.64	61.61
intfloat-e5-large-v2							
Biological	59.31	63.11	56.68	62.40	51.02	75.08	61.27
Chemical	59.77	59.78	54.60	63.05	49.49	75.10	60.30
Environmental	62.61	63.25	57.49	64.79	54.51	75.21	62.98
Extraterrestrial	58.42	63.93	54.57	61.76	48.29	75.67	60.44
Geohazard	59.25	62.26	54.94	60.25	50.44	74.02	60.19
Meteorological&hydrological	60.07	63.55	57.59	62.98	49.77	75.12	61.51
Societal	60.04	66.56	56.92	60.85	53.09	74.75	62.04
Technological	61.74	64.95	56.80	62.20	51.07	74.97	61.95
Avg.	60.15	63.42	56.20	62.29	50.96	74.99	61.34
intfloat-e5-mistral-7b-instruct							
Biological	18.64	19.90	22.40	32.79	22.57	43.06	26.56
Chemical	25.67	19.23	19.34	36.47	16.04	35.54	25.38
Environmental	24.96	17.65	21.97	28.46	22.05	41.30	26.06
Extraterrestrial	21.42	21.71	18.92	34.04	21.21	37.67	25.83
Geohazard	21.45	17.30	16.98	25.16	19.57	41.11	23.59
Meteorological&hydrological	20.67	15.84	20.54	24.92	20.85	37.62	23.41
Societal	21.53	24.83	16.94	31.62	23.46	35.64	25.67
Technological	18.86	19.62	18.73	34.85	17.34	44.61	25.67
Avg.	21.65	19.51	19.48	31.04	20.39	39.57	25.27
intfloat-e5-small-v2							
Biological	66.38	63.19	61.91	61.38	48.09	75.04	62.67
Chemical	66.99	60.81	61.30	63.53	47.24	75.17	62.51
Environmental	67.61	64.30	62.38	65.73	50.80	75.57	64.40
Extraterrestrial	63.65	60.90	57.98	60.98	45.13	74.96	60.60
Geohazard	65.73	61.10	59.06	60.21	43.82	74.60	60.75
Meteorological&hydrological	66.89	63.88	61.80	62.53	46.50	74.76	62.73
Societal	64.53	65.51	61.41	60.10	47.63	74.69	62.31
Technological	65.19	64.31	61.32	61.88	47.51	73.84	62.34
Avg.	65.87	63.00	60.89	62.04	47.09	74.83	62.29
intfloat-multilingual-e5-large							
Biological	68.63	62.63	64.36	60.27	49.91	74.90	63.45
Chemical	66.87	62.82	63.95	61.04	49.49	74.90	63.18
Environmental	66.74	63.98	64.06	62.20	54.65	74.76	64.40
Extraterrestrial	65.81	63.71	60.26	59.62	49.28	75.41	62.35
Geohazard	66.65	62.97	63.71	58.47	49.68	75.10	62.76
Meteorological&hydrological	68.57	63.87	63.41	60.45	50.40	74.53	63.54
Societal	67.61	67.61	64.44	58.97	54.07	74.86	64.59
Technological	67.47	66.39	65.80	61.08	50.71	76.02	64.58
Avg.	67.29	64.25	63.75	60.26	51.02	75.06	63.61
intfloat-multilingual-e5-large-instruct							
Biological	70.03	63.24	64.67	66.51	50.24	64.07	63.13
Chemical	67.34	63.58	61.41	68.56	48.65	63.33	62.15
Environmental	68.93	64.34	63.38	69.14	50.64	62.96	63.23
Extraterrestrial	66.19	65.08	61.14	68.44	49.74	66.73	62.89
Geohazard	68.16	62.03	62.75	64.45	46.34	62.08	60.97
Meteorological&hydrological	68.59	66.05	63.42	66.52	50.51	64.22	63.22
Societal	69.62	68.27	63.92	66.58	51.19	65.31	64.15
Technological	68.25	66.61	65.26	67.51	47.72	65.81	63.53
Avg.	68.39	64.90	63.24	67.21	49.38	64.31	62.91

Table 22: Performance of evaluated models under six search intents and eight event types under the exact search setting. Part IV

	QA	QAdoc	Twitter	FC	NLI	STS	Avg.
llmrails-ember-v1							
Biological	64.90	62.60	49.37	60.23	45.66	74.47	59.54
Chemical	65.00	60.86	47.16	62.76	43.98	75.04	59.14
Environmental	65.68	61.87	47.82	62.25	48.17	74.18	59.99
Extraterrestrial	65.52	65.06	44.55	62.47	42.29	74.28	59.03
Geohazard	63.42	60.96	44.08	59.80	43.70	72.78	57.46
Meteorological&hydrological	64.02	64.12	45.55	60.74	42.42	72.11	58.16
Societal	65.11	66.53	43.32	59.41	46.03	74.16	59.09
Technological	64.87	64.73	45.61	60.78	44.23	74.90	59.19
Avg.	64.82	63.34	45.93	61.06	44.56	73.99	58.95
mixedbread-ai-mxbai-embed-large-v1							
Biological	64.38	62.94	43.89	57.80	40.49	69.47	56.50
Chemical	64.29	60.87	42.16	59.77	39.28	70.09	56.08
Environmental	65.73	60.71	42.67	59.05	43.81	69.91	56.98
Extraterrestrial	65.50	65.11	38.38	60.94	38.15	69.16	56.21
Geohazard	62.01	60.48	40.11	56.35	40.05	67.06	54.34
Meteorological&hydrological	64.65	62.91	39.82	57.32	38.05	67.04	54.97
Societal	65.72	65.82	39.30	58.65	41.65	68.60	56.62
Technological	64.26	64.83	38.31	57.85	40.30	69.16	55.78
Avg.	64.57	62.96	40.58	58.47	40.22	68.81	55.93
nvidia-NV-Embed-v1							
Biological	69.88	63.74	57.69	62.45	47.40	68.75	61.65
Chemical	69.31	56.86	57.02	59.12	45.99	66.66	59.16
Environmental	69.14	63.86	57.85	61.98	52.46	68.36	62.28
Extraterrestrial	65.64	62.43	53.94	57.93	45.98	69.16	59.18
Geohazard	66.26	62.29	55.90	58.56	48.33	65.33	59.44
Meteorological&hydrological	69.42	63.50	58.28	61.66	47.69	68.60	61.52
Societal	68.68	66.72	57.45	58.88	50.42	67.42	61.59
Technological	68.59	64.78	56.57	59.77	48.24	68.62	61.10
Avg.	68.36	63.02	56.84	60.04	48.31	67.86	60.74
nvidia-NV-Embed-v2							
Biological	76.42	69.25	42.42	69.75	58.26	77.86	65.66
Chemical	74.82	69.52	42.07	68.68	57.96	76.22	64.88
Environmental	75.31	68.66	43.82	69.63	61.09	77.21	65.95
Extraterrestrial	73.06	70.08	44.88	68.30	56.31	78.18	65.14
Geohazard	73.91	67.35	41.35	66.72	57.88	75.91	63.85
Meteorological&hydrological	75.33	69.91	43.11	68.37	58.88	76.85	65.41
Societal	75.38	72.11	44.57	68.78	61.06	76.54	66.41
Technological	73.91	71.01	43.18	68.85	58.42	77.31	65.45
Avg.	74.77	69.74	43.18	68.64	58.73	77.01	65.34
sentence-transformers-all-mpnet-base-v2							
Biological	14.00	7.64	17.18	43.65	27.58	35.38	24.24
Chemical	13.09	9.93	14.53	48.81	24.14	35.84	24.39
Environmental	12.45	10.01	18.17	47.32	30.32	35.48	25.62
Extraterrestrial	16.24	8.11	14.82	48.79	27.48	42.02	26.24
Geohazard	14.96	7.75	13.30	43.82	27.70	36.33	23.98
Meteorological&hydrological	16.52	7.84	16.39	44.34	28.07	36.99	25.02
Societal	20.10	14.49	18.41	45.98	27.31	39.06	27.56
Technological	13.16	12.41	16.55	47.68	26.44	36.75	25.50
Avg.	15.06	9.77	16.17	46.30	27.38	37.23	25.32
thenlper-gte-base							
Biological	10.09	5.90	45.04	61.49	44.70	49.42	36.11
Chemical	9.73	5.62	40.67	62.27	41.06	53.75	35.52
Environmental	6.03	4.75	38.62	63.00	48.95	40.21	33.59
Extraterrestrial	10.42	4.54	33.58	59.45	41.39	44.64	32.34
Geohazard	7.96	3.33	34.20	58.54	41.68	43.00	31.45
Meteorological&hydrological	9.16	5.31	38.46	60.76	40.55	43.60	32.97
Societal	10.00	7.88	39.76	60.39	44.40	48.17	35.10
Technological	10.39	5.44	38.00	60.50	40.08	50.33	34.13
Avg.	9.22	5.35	38.54	60.80	42.85	46.64	33.90

Table 23: Performance of evaluated models under six search intents and eight event types under the exact search setting. Part V