MedEureka: A Medical Domain Benchmark for Multi-Granularity and Multi-Data-Type Embedding-Based Retrieval

Yongqi Fan[◇], **Nan Wang**[◇], **Kui Xue**[♠], **Jingping Liu**[◇*], **Tong Ruan**[◇*]
[◇]School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China
[♠]Intelligent Healthcare, Shanghai Artificial Intelligence Laboratory, Shanghai, China
{johnnyfans,y80230069}@mail.ecust.edu.cn
xuekui@pjlab.org.cn, {jingpingliu,ruantong}@ecust.edu.cn

Abstract

Embedding-based retrieval (EBR), the mainstream approach in information retrieval (IR), aims to help users obtain relevant information and plays a crucial role in retrieval-augmented generation (RAG) techniques of large language models (LLMs). Numerous methods have been proposed to significantly improve the quality of retrieved content and many generic benchmarks are proposed to evaluate the retrieval abilities of embedding models. However, texts in the medical domain present unique contexts, structures, and language patterns, such as terminology, doctor-patient dialogue, and electronic health records (EHRs). Despite these unique features, specific benchmarks for medical context retrieval are still lacking. In this paper, we propose MedEureka, an enriched benchmark designed to evaluate medical-context retrieval capabilities of embedding models with multi-granularity and multi-data types. MedEureka includes four levels of granularity and six types of medical texts, encompassing 18 datasets, incorporating granularity and data type description to prompt instruction-fine-tuned text embedding models for embedding generation. We also provide the MedEureka Toolkit to support evaluation on the MedEureka test set. Our experiments evaluate state-of-the-art open-source and proprietary embedding models, and fine-tuned classical baselines, providing a detailed performance analysis. This underscores the challenges of using embedding models for medical domain retrieval and the need for further research. Our code and data are released in the repository: https://github.com/JOHNNY-fans/MedEureka.

1 Introduction

Embedding-based Retrieval (Huang et al., 2020; Li et al., 2021; He et al., 2023) has been a valuable research topic and has become a mainstream method for information retrieval. By transforming the text into semantic-rich dense vectors, the relevant text is retrieved by calculating the vector similarity. Siamese networks (Reimers and Gurevych, 2019), contrastive learning (Gao et al., 2021; Wang et al., 2022), and various effective negative sampling strategies (Chuang et al., 2022; Formal et al., 2022; Nishikawa et al., 2022; Wang and Dou, 2023) have been widely used in the training of embedding models to improve retrieval performence by bringing the query closer to positive samples and pushing it farther away from negative samples. A more advanced approach is based on the concept of prompting, using instruction-fine-tuned text embeddings (Su et al., 2023; Chen et al., 2024) to achieve more detailed vector representations. Many advanced embedding models trained on large corpora have been released (Wang et al., 2023; Li et al., 2023; OpenAI, 2024b; Wang et al., 2024), while evaluations for the retrieval ability of embedding models are already present in some generic-domain benchmarks (Thakur et al., 2021; Muennighoff et al., 2023; Xiao et al., 2023).

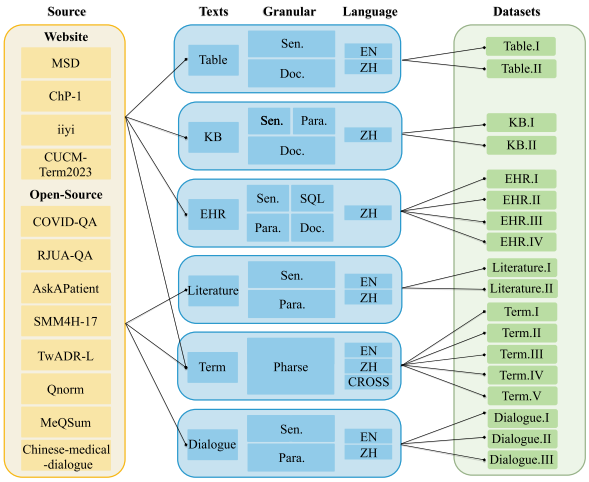

Figure 1: The overall architecture of the MedEureka, including data source, data type, granularity, language, and corresponding constructed task.

[*] Corresponding authors.

Findings of the Association for Computational Linguistics: NAACL 2025, pages 2825–2851
April 29 - May 4, 2025

In the medical domain, the evaluation of model retrieval capabilities is conducted separately for individual datasets (Xiao et al., 2023; Chen et al., 2020) due to the unique context and expertise. Meanwhile, there are numerous practical applications for medical information retrieval, such as terminology normalization (Xu et al., 2020), the retrieval and analysis of EHRs (Myers et al., 2024), and some RAG-based QA (Lozano et al., 2023; Huang et al., 2024) of medical literature and knowledge bases. However, there is a lack of domain-specific evaluations in the medical field to guide the selection of suitable embedding models for medical context retrieval in different scenarios.

In this paper, we introduce MedEureka, a benchmark specifically designed for multi-granularity medical-context retrieval using embedding models. As illustrated in Figure 1, MedEureka encompasses four levels of granularity: phrase, sentence, paragraph, and document. It also incorporates six types of medical texts: Table, Literature, Knowledge Base (KB), Term, Electronic Health Record (EHR), and medical Dialogue, along with a unique SQL case. The raw data for this benchmark were sourced from academic open-source projects and websites. Following data organization and classification, and with the support of AI-assisted annotation, we constructed 18 datasets for training, validation, and testing, each defined by distinct granularity levels and semantic labels.

We developed the MedEureka Toolkit to facilitate the evaluation of the MedEureka benchmark's test set and assessed the retrieval performance of various advanced embedding models, including both proprietary and open-source options. The overall performance is illustrated in radar plots, as shown in Figure 2. To evaluate these instruction fine-tuned embedding models, we generated prompts using granularity and semantic tags. Additionally, we trained several baseline embedding models on MedEureka using classical training methods based on both generic and medical Pre-trained Language Models (PLMs). Our experiments demonstrate that state-of-the-art embedding models trained on large-scale corpora exhibit strong medical retrieval capabilities, particularly for data types with minimal divergence from the generic domain or clear semantics, such as dialogues. However, challenges persist in handling more specialized medical content, such as electronic health records, medical literature, and terminology. Even with supervised methods, training is prone to bias due to the constraints of PLM parameter size and capabilities, as well as diverse text granularity and type settings. We also conducted a detailed analysis to provide insights and guidance, aiming to encourage the NLP community to pursue further research and collaboration in addressing the practical challenges highlighted by this benchmark.

2 Related Work

2.1 Advanced Embedding Models.

Embedding-based retrieval (EBR) has become a mainstream approach to information retrieval. Particularly, with the advent of LLMs, numerous high-quality corpora have been created, and many generalized embedding models have been released, achieving excellent performance, such as OpenAI text embeddings (OpenAI, 2022, 2024b), GTE (Li et al., 2023). More advanced models, e.g., Instructor-xl (Su et al., 2023) and E5 (Wang et al., 2023), incorporate instructions into training processes. By employing different prompts, these models (Wang Yuxin, 2023; Xiao et al., 2023) improve the quality and adaptability of embeddings across diverse scenarios. Additionally, recent advancements in diversifying training data, generating synthetic data, and leveraging LLMs as backbones have further contributed to the development of more generalized embeddings, such as multilingual (Wang et al., 2024), multi-granularity (Chen et al., 2024), and LLM-based (Li et al., 2023).

2.2 Training Strategies for Embedding

Various training strategies for embedding models have been proposed to achieve more detailed vector representations with the success of PLMs. Traditional method employs siamese network to generate semantically meaningful sentence embeddings (Reimers and Gurevych, 2019). Contrastive learning (Gao et al., 2021; Wang et al., 2022), along with numerous negative sampling strategies, has recently emerged as a prominent method for training embedding models. SNCSE (Wang and Dou, 2023) alleviates feature suppression by treating the negations of positive samples as soft negatives. DiffCSE (Chuang et al., 2022) introduces equivariant contrastive learning, InfoCSE (Wu et al., 2022) learns sentence representations by reconstructing segments of the original sentences, while RankCSE (Liu et al., 2023) incorporates ranking consistency and ranking distillation, thereby enhancing representation quality.

Task	ID	Source	Language	Annotation	Phrase	Sen.	SQL	Para.	Doc.	# Examples(train/dev/test)	# Avg Len (query/target)
Table	I	Website	zh	Auto & Human	-	✈	-	-	🏆	1,776 / 592 / 592	105.80 / 2811.27
	II	Website	en	Auto & Human	-	✈	-	-	🏆	806 /269 / 269	60.08 / 561.42
Literature	I	RJUA-QA	zh	Auto & Human	-	✈	-	🏆	-	360 / 120 / 120	204.11 / 264.25
	II	COVID-QA	en	Auto & Human	-	✈	-	-	🏆	979 / 327 / 326	12.77 / 6065.79
KB	I	Website	zh	Auto & Human	-	✈	-	🏆	-	985 / 330 / 328	188.76 / 192.09
	II	Website	zh	Auto & Human	-	✈	-	-	🏆	1967 / 657 / 656	158.36 / 1823.71
Term	I	Website	zh	Auto & Human	✈🏆	-	-	-	-	7,749 / 1,005 / 1,005	15.26 / 13.44
	II	Website	cross	Auto & Human	✈🏆	-	-	-	-	2,4786 /8,262 / 8,262	6.38 / 12.91
	III	AskAPatient	en	Auto	✈🏆	-	-	-	-	5,051 / 2,748 / 4,137	4.36 / 4.03
	IV	SMM4H-17	en	Auto	✈🏆	-	-	-	-	2,124 / 709 / 1,194	4.20 / 5.45
	V	TwADR-L	en	Auto	✈🏆	-	-	-	-	3,204 / 656 / 988	4.32 / 4.51
EHR	I	Self-built	zh	Auto & Human	-	✈	🏆	-	-	2,097 / 718 / 721	45.98 / 69.27
	II	Website	zh	Auto & Human	-	✈	-	🏆	-	1,709 / 570 / 570	48.17 / 191.72
	III	Website	zh	Auto & Human	-	-	✈	🏆	-	1,692 / 564 / 564	71.37 / 191.72
	IV	Website	zh	Auto & Human	-	✈	-	-	🏆	1229 / 410 / 410	72.62 / 976.30
Dialogue	I	Qnorm	zh	Auto	-	✈🏆	-	-	-	959 / 320 / 320	93.63 / 100.18
	II	MeQSum	en	Auto	-	✈	-	🏆	-	600 / 200 / 200	13.69 / 81.41
	III	Chinese-medical-dialogue	zh	Auto & Human	-	✈🏆	-	-	-	1,800 / 600 / 600	87.46 / 85.31

Table 1: Dataset statistics and descriptions. The columns represent the annotation method (auto-generated or human), the number of examples, the granularity, and the average length of both query and target.

2.3 Generic-domain Evaluation of Retrieval.

The evaluation of retrieval performance has been included in many embedding benchmarks, such as the STS benchmark (Reimers and Gurevych, 2019) and MTEB (Muennighoff et al., 2023). Additionally, prominent benchmarks like BEIR (Thakur et al., 2021) focus on retrieval capabilities across generic domains and include some medical datasets. Ajith et al. (2024) introduced a comprehensive benchmark for scientific literature retrieval, while (Zhang et al., 2021) evaluated dense retrieval in multilingual settings. However, there remains a gap in comprehensive, fine-grained, and multi-task evaluation of embedding model retrieval capabilities specifically within the medical domain.

2.4 Applications of EBR

Embedding-based retrieval has numerous application scenarios within the NLP field. Some general use cases include document retrieval (Schopf et al., 2022), question-answering systems (Liu et al., 2019), and recommendation systems (Agrawal et al., 2021). Particularly in the era of LLMs, EBR is widely used in Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Siriwardhana et al., 2023) technology to enhance the quality of generated content and reduce hallucinations. Furthermore, the medical field also presents numerous retrieval scenarios, such as clinical terminology normalization (Castano et al., 2016; Sarker et al., 2018; Xu et al., 2020), electronic health record (EHR) retrieval (Sivarajkumar et al., 2024; Myers et al., 2024), medical literature and knowledge base search (Lozano et al., 2023; Huang et al., 2024).

3 The MedEureka Benchmark

3.1 Data Collection

Collecting corpora of varying granularity and types poses significant challenges due to copyright and privacy protection concerns. We chose not to use simulation, self-building, or distillation techniques because they require specialized medical expertise. As a result, we invested considerable effort into finding academic open-source materials, identifying formal application pathways, and securing copyright-free medical data and knowledge. As shown in Fig 1, we obtained six different types of medical text data from open-source projects and websites to construct retrieval evaluation datasets with varying granularity and types. We discovered several datasets related to medical terminology (Limsopatham and Collier, 2016; Sarker et al., 2018), clinical evidence (Lyu et al., 2023), scientific literature (Möller et al., 2020) and doctor-patient dialogues (Ben Abacha and Demner-Fushman, 2019; Toyhom, 2019; DataTager, 2024) within open-source projects. We also identified several publicly accessible Chinese terminology base "Chinese Common Clinical Medical Terminology 2023 Edition" (CUCMTerm2023) and the first part of Chinese Pharmacopoeia (ChP-1)[1] from the website, which we stored locally after performing OCR and post-processing. In addition, we obtained some desensitized patient cases from open-source forums iiyi[2], and medical tables coming from an open-

[1] https://ydz.chp.org.cn/#/main
[2] https://bingli.iiyi.com/

Model	Language	Max Length	Hidden Size	Instruction	Publish Time
text-embedding-ada-002 (OpenAI, 2022)	multi-lingual	8,192	1,536	✗	2022.12
text-embedding-3-large (OpenAI, 2024b)	multi-lingual	8,192	3,072	✗	2024.01
gte-large-zh (Li et al., 2023)	zh	512	1,024	✗	2023.12
gte-large (Li et al., 2023)	en	512	1,024	✗	2023.12
Instructor-xl (Su et al., 2023)	en	512	1,024	✔	2023.01
bge-large-en-v1.5 (Xiao et al., 2023)	en	512	1,024	✔	2023.12
bge-large-zh-v1.5 (Xiao et al., 2023)	zh	512	1,024	✔	2023.12
bge-m3 (Chen et al., 2024)	multi-lingual	8,196	1,024	✗	2024.01
m3e-base (Wang Yuxin, 2023)	multi-lingual	512	768	✗	2023.06
m3e-large (Wang Yuxin, 2023)	zh	512	1,024	✗	2023.06
e5-mistral-7b-instruct (Wang et al., 2023)	en	32,768	4,096	✔	2023.12
multilingual-e5-large-instruct (Wang et al., 2024)	multi-lingual	512	1,024	✔	2024.02
gte-Qwen2-1.5B-instruct (Li et al., 2023)	multi-lingual	32,768	1,536	✔	2024.06
gte-Qwen2-7B-instruct (Li et al., 2023)	multi-lingual	32,768	1,536	✔	2024.06
bge-multilingual-gemma2 (Chen et al., 2024)	multi-lingual	8,192	3,584	✔	2024.07
bert-base-uncased (Devlin, 2018)	en	512	768	✗	2019.01
biobert-base-cased-v1.2♯ (Lee et al., 2020)	en	512	768	✗	2021.01
bert-base-chinese (Devlin, 2018)	zh	512	768	✗	2021.01
medbert-base-wwm-chinese♯ (Yang et al., 2021)	zh	512	768	✗	2021.05
bert-base-multilingual-uncased (Devlin et al., 2018)	multi-lingual	512	768	✗	2019.01

Table 2: Descriptions of the advanced embedding models and pre-trained language models (PLMs) used in training the baselines, detailing the language support, maximum length, hidden size, instruction support (analysed from the official example), and publish time. ♯ denotes those biomedical PLMs.

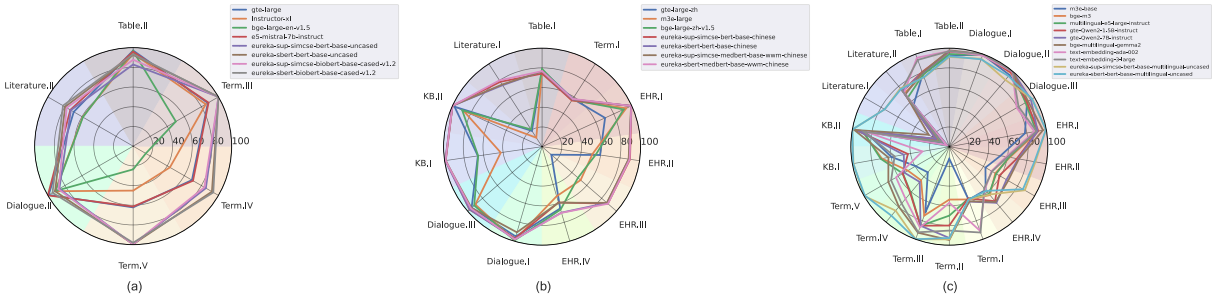

Figure 2: The radar charts illustrating the performance of different language-support types of embedding models on the MedEureka datasets (a): Performence of English embedding models. (b): Performance of Chinese embedding models. (c): Performance of multi-lingual embedding models.

source medical website MSD[3], via web crawler techniques.

3.2 Tasks in MedEureka

MedEureka focuses on evaluating retrieval capabilities in the medical domain, particularly for embedding models. Our main task is identifying the most relevant content to a query from a target base. This involves testing their ability to generate effective vector representations. Additionally, for embedding models that support instructions, we use the dataset's granularity and data type to prompt for better vector representations, the specific instruction prompts are shown in Appendix C.

Formally, the main task of the evaluation is: Given a query Q (with/without prompt) and a target T, the embedding model E generates the corresponding vector representations V_Q and V_T. By calculating the semantic distance between them, the top-K candidate information C with a higher score is retrieved, where K is a hyper-parameter to determine the number of recalls.

In the medical domain, common retrieval applications such as Knowledge-Base Question Answering (KBQA) systems, Table Question Answering (TQA) systems, Literature Question Answering systems, and medical dialogue systems (Raghavan et al., 2021; Luo et al., 2022; Pal et al., 2022; Xu et al., 2019) typically use the RAG method to retrieve relevant medical sources like Knowledge Bases, Tables, Literature, queries, and dialogues. Clinical data analysis (Xu et al., 2024), Clinical Decision Support Systems (CDSS) (Papadopoulos et al., 2022), and Diagnosis-Related Group (DRG)

[3] https://www.msdmanuals.cn/professional/pages-with-widgets/tables?mode=list

systems (Wang et al., 2020) often involve searching relational databases and standard terminology bases. From these practical scenarios, we abstract some seed retrieval tasks. As shown in Table 1, across the involved tasks, we consider six medical data types: **Table**, **Literature**, **KB**, **Term**, **EHR**, and **Dialogue**; four levels of granularity: **Phrase**, **Sentence**, **Paragraph**, and **Document**, as well as a special case involving **SQL**. The SQL-related task is designed to investigate the semantic similarity between SQL queries and natural language texts within the vector space of the embedding model. In this statistic table, the special mark ✈ denotes the granularity of the query, and ⛱ denotes the granularity of the target.

3.3 Construction and Annotation

As shown in Figure 1, using this collected corpus, we defined a total of 18 task datasets that involve retrieval between medical texts of varying granularity and types.

Table. For the tables in Markdown format obtained from MSD[4], we used an LLM to automatically generate questions based on the table content, followed by manual corrections. To offer a more intuitive demonstration of the automatic query generation process using LLMs, we provide a specific prompt example for the Table task. This example is presented in Figure C11 in the Appendix. We obtained both English and Chinese versions of these table sources. Using these questions as queries and the corresponding entire table as the target, we created **Table I** and **Table II**.

Literature. In RJUA-QA (Lyu et al., 2023), we identified several clinical references meticulously sourced from professional literature, guidelines, major textbooks, authoritative publications from PubMed, and the extensive clinical experience of seasoned practitioners with over a decade of expertise. Additionally, the dataset includes numerous virtual patient questions related to medical specialty diagnosis and examination advice. We used these questions as queries and expert-annotated references as retrieval targets, resulting in the creation of dataset **Literature.I**. In COVID-QA, we use questions as queries and corresponding context from the "paragraphs" field as the target, resulting in dataset **Literature.II**

KB. Using the Chinese Medicine Pharmacopoeia (ChP-1) obtained, we transformed it into a structured knowledge base containing 1,493 drug documents, each covering about ten drug attribute fields, through OCR techniques and post-processing. We then generated questions about drug attributes using a combination of automatic questioning by an LLM and manual review. These questions were used as queries, and we set two levels of targets: attribute level and document level, resulting in datasets **KB.I** and **KB.II**.

Term. In this part, we employ a terminology normalization task to find the corresponding standard term for a medical phrase from a large standard terminology database. The medical phrase serves as the query, and the standard terminology database as the target. For the Chinese, we have independently constructed a dataset **Term.I** for terminology normalization based on the synonyms and previously utilized phrases in CUCMTerm2023 corpus, which includes the same 4 term categories present in our established standard terminology database. Moreover, we extracted the Chinese-English translation pairs from the CUCMTerm2023 and constructed the cross-lingual dataset **Term.II**. For English terms, we adopt three reputable datasets AskAPatient, SMM4H-17, and TwADR-L and get **Term.III**, **Term.IV** and **Term.V**.

EHR. Based on typical patient cases from iiyi, we extracted and structured the EHR-related sections, which include fields such as the patient's chief complaint, symptoms, imaging study results, and findings from a complete checkup. We also use LLMs with human verification to generate queries based on the contents of one or two EHR fields. These queries' target includes the granularity of specific fields and the entire EHR. Specifically, we included an interesting pseudo-SQL code experiment in this section since hospital doctors often look up medical records by writing SQL queries. Thus, a total of four datasets **EHR.I**, **EHR.II**, **EHR.III**, and **EHR.IV** were obtained for this part.

Dialogue. Patient-dialogue datasets are common in open-source projects, and in this study, we consider three cases: the patient questions standardization dataset (Qnorm[4]), the patient questions summary dataset (MeQSum[5]), and the doctor-patient dialogues dataset (Chinese-medical-dialogue[6]). We extracted some corresponding queries and targets from them and then obtained the dataset **Dialogue.I**, **Dialogue.II** and **Dialogue.III**.

[4] https://huggingface.co/datasets/PandaVT/datatager_standard_med_question

[5] https://github.com/abachaa/MeQSum

[6] https://github.com/Toyhom/Chinese-medical-dialogue-data

Model	Table		Literature		KB		Term					EHR				Dialogue		
	I	II	I	II	I	II	I	II	III	IV	V	I	II	III	IV	I	II	III
Advanced instruction-fine-tuned Embedding Models (without instruction)																		
MRR@10																		
bm25†	69.45	64.01	26.76	62.22	26.11	81.88	39.88	-	35.98	15.67	14.73	76.35	**69.28**	7.05	**73.42**	70.09	66.17	54.90
gte-large-zh	68.26	-	47.90	-	53.62	**84.16**	**51.54**	-	-	-	-	52.15	50.54	7.78	51.86	87.03	-	**84.15**
gte-large	-	87.80	-	52.97	-	-	-	-	67.81	46.67	39.79	-	-	-	-	-	90.15	-
m3e-base	70.45	80.05	36.66	39.48	**54.59**	77.48	50.12	7.84	55.49	25.61	27.17	62.97	43.26	33.60	42.18	95.15	81.50	78.54
m3e-large	68.32	-	36.85	-	53.84	68.63	48.52	-	-	-	-	83.64	42.40	43.46	37.54	93.95	-	77.60
bge-m3	70.17	85.80	**48.90**	52.48	46.35	82.88	50.58	37.80	58.21	32.62	30.58	**84.03**	52.60	45.93	50.30	98.07	87.94	76.52
bge-m3 (sparse)†	**70.93**	77.86	37.75	**65.58**	48.01	82.73	47.67	1.64	31.94	11.42	11.55	82.13	60.88	**65.53**	60.47	82.57	79.79	67.37
text-embedding-ada-002	69.07	88.13	44.09	53.48	29.92	82.30	47.74	37.88	66.12	41.48	38.42	74.24	50.28	38.74	40.57	96.85	90.66	63.90
text-embedding-3-large	64.74	**88.95**	37.86	51.25	51.89	77.01	50.27	**65.77**	**70.65**	**50.83**	**42.51**	77.51	49.22	36.84	37.00	**98.49**	**91.68**	67.52
Exact HR@10																		
bm25†	77.53	79.55	0.83	81.60	19.51	85.21	48.36	-	48.10	20.10	22.27	85.58	**76.32**	9.04	**84.88**	83.75	83.00	64.17
gte-large-zh	77.87	-	**19.17**	-	65.85	**97.71**	57.81	-	-	-	-	70.32	59.47	12.41	65.37	95.00	-	**94.67**
gte-large	-	**97.03**	-	69.63	-	-	-	-	87.62	68.93	62.15	-	-	-	-	-	99.00	-
m3e-base	78.72	92.57	8.33	56.13	59.76	86.43	57.32	12.30	71.24	34.00	39.47	77.95	52.81	42.38	57.32	99.38	95.00	91.17
m3e-large	77.53	-	10.83	-	42.38	87.65	55.82	-	-	-	-	**93.90**	51.75	51.77	52.44	99.06	-	89.67
bge-m3	77.70	96.65	13.33	70.86	**69.51**	89.48	57.01	53.69	74.47	44.56	48.58	91.96	62.28	53.90	64.15	**100.00**	97.50	90.17
bge-m3 (sparse)†	79.22	89.96	6.67	**85.58**	46.95	85.21	54.43	2.01	45.44	15.66	17.21	90.98	71.40	**75.53**	76.59	92.50	89.50	80.00
text-embedding-ada-002	**96.11**	97.02	3.33	71.47	27.74	84.90	90.74	56.97	87.28	69.17	62.44	87.79	58.59	49.46	56.34	98.43	99.50	79.00
text-embedding-3-large	93.75	97.02	9.16	75.15	56.40	88.10	**92.73**	**85.18**	**92.45**	**75.29**	**68.21**	89.87	60.00	50.17	55.12	**100.00**	**100.00**	81.16
Advanced instruction-fine-tuned Embedding Models (with instruction)																		
MRR@10																		
Instructor-xl	-	82.06	-	36.28	-	-	-	-	64.29	25.02	25.45	-	-	-	-	-	82.41	-
bge-large-en-v1.5	-	85.90	-	43.04	-	-	-	-	26.95	7.30	10.44	-	-	-	-	-	81.09	-
bge-large-zh-v1.5	68.72	-	46.27	-	62.76	82.20	49.80	-	-	-	-	83.68	48.50	36.94	46.92	95.12	-	80.55
e5-mistral-7b-instruct	-	85.84	-	52.64	-	-	-	-	66.35	47.38	38.52	-	-	-	-	-	**94.23**	-
multilingual-e5-large-instruct	70.94	85.45	46.25	46.46	63.47	81.33	49.01	49.03	65.29	37.96	34.27	78.32	52.02	44.24	49.51	95.24	92.82	75.29
gte-Qwen2-1.5B-instruct	72.22	89.06	48.31	55.72	63.37	**83.97**	46.01	55.22	62.23	36.56	33.04	81.63	61.83	47.30	58.91	94.86	92.36	79.39
gte-Qwen2-7B-instruct	70.61	87.54	50.43	**57.05**	55.41	68.20	51.87	75.31	54.44	26.24	27.04	84.19	64.96	59.38	59.56	99.17	93.45	**87.69**
bge-multilingual-gemma2	**74.39**	**89.98**	**58.33**	47.39	**70.53**	78.25	**52.39**	**80.47**	**73.03**	**53.77**	**46.11**	**89.74**	**65.61**	**62.13**	**60.05**	**99.45**	93.26	79.75
Exact HR@10																		
Instructor-xl	-	94.05	-	59.20	-	-	-	-	85.13	43.10	45.14	-	-	-	-	-	91.50	-
bge-large-en-v1.5	-	96.65	-	58.90	-	-	-	-	50.04	14.82	23.58	-	-	-	-	-	91.00	-
bge-large-zh-v1.5	78.89	-	20.83	-	65.55	89.18	56.42	-	-	-	-	90.98	55.26	48.23	66.59	99.06	-	91.17
e5-mistral-7b-instruct	-	95.91	-	73.62	-	-	-	-	88.37	70.18	61.34	-	-	-	-	-	**100.00**	-
multilingual-e5-large-instruct	77.87	94.05	10.00	65.34	**70.73**	88.87	56.22	69.99	85.52	60.97	54.45	89.60	61.40	54.96	62.68	99.06	99.50	89.33
gte-Qwen2-1.5B-instruct	**80.74**	96.65	15.83	73.62	42.68	96.19	55.82	80.05	85.40	61.73	54.66	91.26	69.65	58.69	72.20	99.06	98.00	91.00
gte-Qwen2-7B-instruct	79.39	97.03	15.83	**73.93**	46.95	94.21	**59.31**	92.70	85.38	46.73	52.73	90.98	71.05	67.02	**74.39**	**100.00**	**100.00**	**95.67**
bge-multilingual-gemma2	80.07	**97.77**	**24.17**	69.33	62.80	**98.02**	59.00	**94.75**	**92.94**	**79.65**	**71.46**	**96.53**	**71.75**	**68.97**	73.17	**100.00**	**100.00**	92.33

Table 3: Comparison of advanced embedding model performance on MRR@10 and Exact HR@10 using cosine similarity as the distance metric. † indicates sparse retrieval, while all others are dense retrieval.

We used different construction methods for each of the three sources of data and the specific construction details are shown in Appendix B.

Specifically, for the human-machine collaborative portion of the dataset construction process, we use GPT-4o (OpenAI, 2024a) as the auxiliary LLM, and we implemented a dual annotation strategy, where two annotators independently reviewed and validated the data. For datasets where the data source already contains a clear mapping between query and target, such as the **Term** task, we do not manually correct the content of each sample, but only perform data cleaning and processing of the format. For datasets with a predefined query-target mapping, such as the **Term** task, we did not manually modify individual samples but focused on data cleaning and format processing. In contrast, for datasets involving query generation and automatic annotation using LLMs, such as the **Table** task, we performed manual corrections. First, we consulted a medical expert to define key validation criteria, emphasizing medically relevant terms and factual accuracy (e.g., numerical values). Then, a PhD student and a Master's student specializing in medical NLP independently reviewed the data. Identified risks and discrepancies were subsequently discussed and corrected, resulting in a refined dataset. Approximately 10% of the data required modification. This approach ensured a high level of accuracy and reliability in the dataset. We show sample data for each task in the Appendix Figure C2 to Figure C10.

3.4 Data Statistic

We present the dataset statistics in Table 1, describing the data type of the dataset, the corresponding ID (Roman numerals), the source, the language, the annotation method, the granularity of the query and target, the number of samples and the average length of the query and target.

3.5 Evaluation Metrics

For all retrieval tasks, we first selected Mean Reciprocal Rank (MRR) as the evaluation metric. Denoted as MRR@n, where n represents the number of top-retrieved items considered, MRR measures the average reciprocal rank of the first relevant item for each query. Higher MRR values indicate better ranking quality, meaning relevant items appear

Model	Table		Literature		KB		Term					EHR				Dialogue		
	I	II	I	II	I	II	I	II	III	IV	V	I	II	III	IV	I	II	III
bert-base-uncased*	-	1.86	-	7.36	-	-	-	-	17.50	7.54	7.59	-	-	-	-	-	3.50	-
biobert-base-cased-v1.2*	-	1.12	-	7.98	-	-	-	-	19.94	9.46	8.70	-	-	-	-	-	1.50	-
eureka-sup-simcse-bert-base-uncased	-	82.90	-	73.01	-	-	-	-	**100.00**	85.26	99.80	-	-	-	-	-	86.50	-
eureka-unsup-simcse-bert-base-uncased	-	12.27	-	16.87	-	-	-	-	35.15	24.25	15.08	-	-	-	-	-	18.50	-
eureka-sup-simcse-biobert-base-cased-v1.2	-	87.73	-	77.91	-	-	-	-	**100.00**	83.50	99.70	-	-	-	-	-	87.50	-
eureka-unsup-simcse-biobert-base-cased-v1.2	-	59.11	-	33.74	-	-	-	-	60.72	38.50	28.34	-	-	-	-	-	56.50	-
eureka-sbert-bert-base-uncased	-	92.57	-	79.75	-	-	-	-	**100.00**	92.36	99.49	-	-	-	-	-	91.50	-
eureka-sbert-biobert-base-cased-v1.2	-	**93.31**	-	**82.21**	-	-	-	-	99.93	**93.96**	98.08	-	-	-	-	-	**94.50**	-
bert-base-chinese*	1.69	-	0.00	-	0.00	0.61	30.74	-	-	-	-	3.05	1.75	0.18	0.24	5.00	-	29.83
medbert-base-wwm-chinese*	1.52	-	0.00	-	0.00	0.61	25.27	-	-	-	-	0.55	0.00	0.00	0.49	1.25	-	22.00
eureka-sup-simcse-bert-base-chinese	74.16	-	78.33	-	98.78	**100.00**	55.92	-	-	-	-	**99.72**	**90.18**	88.30	59.02	96.88	-	**97.50**
eureka-unsup-simcse-bert-base-chinese	53.55	-	0.00	-	10.06	64.18	51.74	-	-	-	-	93.96	46.84	44.50	18.05	82.50	-	69.67
eureka-sup-simcse-medbert-base-wwm-chinese	76.35	-	75.00	-	99.09	**100.00**	57.21	-	-	-	-	99.31	88.77	87.06	59.27	90.62	-	94.83
eureka-unsup-simcse-medbert-base-wwm-chinese	53.04	-	2.50	-	7.32	46.80	52.94	-	-	-	-	91.95	48.42	45.21	23.66	79.69	-	69.00
eureka-sbert-bert-base-chinese	76.52	-	74.17	-	**99.39**	**100.00**	57.01	-	-	-	-	99.60	88.95	87.23	69.76	97.81	-	97.17
eureka-sbert-medbert-base-wwm-chinese	**76.69**	-	**80.00**	-	99.09	**100.00**	**57.61**	-	-	-	-	99.46	89.82	**88.48**	**70.00**	**98.12**	-	95.67
bert-base-multilingual-uncased*	0.84	1.12	0.00	8.90	0.00	0.46	24.38	1.03	20.30	8.04	7.69	1.66	1.05	1.06	0.49	0.62	1.50	21.50
eureka-sup-simcse-bert-base-multilingual-uncased	74.83	91.08	**75.83**	**77.30**	98.78	**100.00**	**57.11**	**94.06**	**100.00**	84.76	**100.00**	99.58	**88.77**	**88.48**	**63.41**	**95.00**	88.50	**97.00**
eureka-unsup-simcse-bert-base-multilingual-uncased	61.15	69.89	0.00	40.49	14.94	77.29	53.93	7.36	65.43	38.10	32.59	95.84	48.95	46.10	21.22	82.19	73.50	67.00
eureka-sbert-bert-base-multilingual-uncased	**75.17**	**91.82**	**75.83**	76.38	**99.09**	**100.00**	**57.11**	92.64	**100.00**	**91.44**	99.09	**99.60**	87.72	86.35	58.54	94.38	**93.50**	**97.00**

Table 4: Comparison of training baseline models performance on Exact HR@10 using cosine similarity as the distance metric. "$*$" indicates direct retrieval with frozen base PLMs

closer to the top. While MRR emphasizes ranking importance, it can be lenient for tasks requiring the retrieval of multiple relevant items. To address this, we also introduced Exact Hit Rate, denoted as Exact HR@n, which measures the proportion of queries where all relevant items are ranked within the top n results.

Mathematically, MRR@n is defined as:

$$\text{MRR@n} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{\text{rank}_q} \tag{1}$$

where Q represents the set of queries, and rank_q is the rank position of the first relevant item retrieved for query q.

Exact HR@n is defined as:

$$\text{Exact HR@n} = \frac{1}{|Q|} \sum_{q \in Q, c \in C} \mathbb{I}(q, c) \tag{2}$$

where C represents the candidates retrieved by queries Q, and $\mathbb{I}$ is an indicator function that returns 1 if all relevant items are included in c for query q, and 0 otherwise.

4 Experiments

4.1 Baseline Models

We selected the traditional information retrieval method BM25, along with recent state-of-the-art embedding models and 8 advanced instruction-fine-tuned embedding models as baselines. Meanwhile, we chose two mainstream training methods, SimCSE (Gao et al., 2021) and SBERT (Reimers and Gurevych, 2019), to train BERT-base PLMs on MedEureka as training baselines. The basic information for these models is shown in Table 2.

4.2 Implementation Details

We use cosine similarity to measure distances between embedding vectors, incorporating prompts from Section 3.2 for models that support them. FAISS (Johnson et al., 2019) is employed for accelerated computation. Text exceeding model length limits is truncated accordingly. The specific LLM used in the data construction process is gpt-4o-2024-08-06. For training baseline models, we follow the original method's parameters, with a batch size of 64 per device, training for ten epochs on four H800 GPUs. Optimal checkpoints are selected based on validation performance, and the pooled CLS token of the PLMs is used as sentence representation. In particular, for supervised training, we distill negative samples with the powerful bge-multilingual-gemma2, which finds up to five negative samples for each query by setting a threshold.

4.3 Results and Analysis

4.3.1 Overall Results and Analysis.

We evaluated the overall experimental results regarding the performance of different models on various tasks. Figure 2 shows the overall performance of embedding models using radar charts. The results for "MRR@10" and "Exact HR@10" with cosine similarity are presented in Table 3. Additionally, we use line charts to depict the trends for all "Exact HR@n" results in Appendix Figure C13 and Figure C14, where n $\in \{5, 10, 20, 50, 100, 200, 500\}$.

It is evident that different models exhibit varying levels of performance. Some models show clear proficiency in certain tasks, making them well-suited to handle those tasks effectively. As expected, model accuracy significantly improves as the number of recalls increases. Both OpenAI's em-

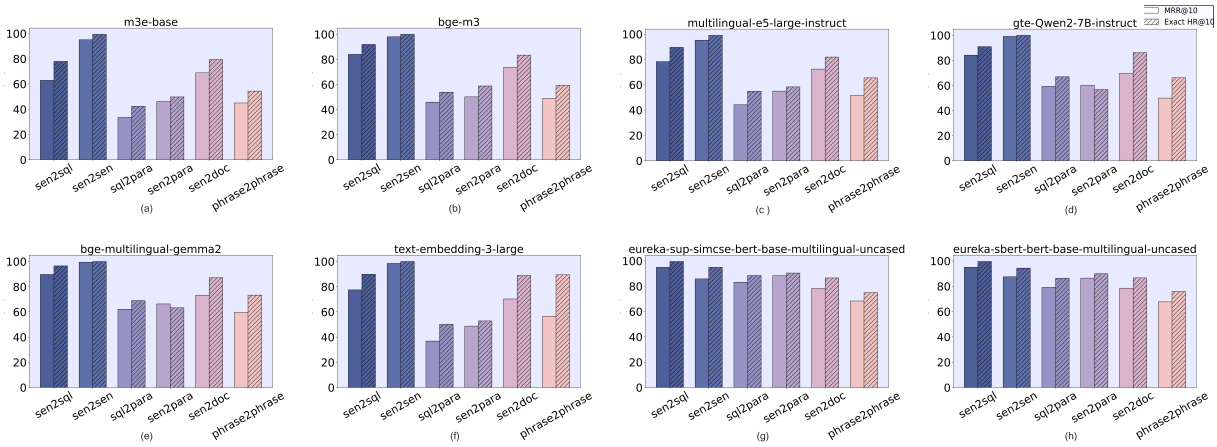

Figure 3: Performance comparison on six granularities across eight models: (a) m3e-base, (b) bge-m3, (c) multilingual-e5-large-instruct, (d) gte-Qwen2-7B-instruct, (e) bge-multilingual-gemma2, (f) text-embedding-3-large, (g) eureka-sup-simcse-bert-base-multilingual-uncased, (h) eureka-sbert-bert-base-multilingual-uncased.

bedding and the bge-multilingual-gemma2 models demonstrate strong performance and robust multilingual capabilities. However, there is no "hexagonal warrior" that is competent in every aspect.

Analyzed in terms of model size, hidden vector dimension, and supported context length, generally, larger models and higher vector dimensions yield better results, with LLM-based embedding models showing particular superiority. However, this improvement has limitations, such as the gte model exhibits fluctuations on certain tasks. Notably, longer context lengths tend to provide more complete semantics, which can enhance performance.

Table 4 shows the performance of training baselines. Intuitively, trained methods, particularly supervised training, show significant improvement across multiple tasks, such as Term, Literature, and EHR. An interesting phenomenon is that different training methods align with different dominant tasks and tend to exhibit bias, risking local optima when training on existing PLMs. In contrast, advanced embedding models, trained on large-scale data, demonstrate better generalization. Additionally, training based on PLMs in the medical field yields superior performance.

From the perspective of different metrics, we observe that almost all models perform worse on MRR@10 compared to HR@10 across all tasks. This reflects the influence of MRR, which penalizes models based on the ranking position of correctly recalled candidates. While the models can find the correct answers within a certain range, the lower MRR suggests that the models struggle with more suitable ranking. This highlights the importance of re-ranking to address this issue.

4.3.2 Analysis by Data Type.

To provide a more intuitive analysis of the model's performance across different types of medical texts, we used radar charts in Figure 2 to illustrate capabilities by data type, distinguishing between Chinese, English, and multilingual models.

With the accumulation of high-quality training corpora in the era of large models, advanced embedding models have achieved strong performance on many retrieval tasks. Specifically, models excel in Table and Dialogue data, though performance in Literature, EHR, and Terminology still has room for improvement. This observation underscores that while large volumes of medical knowledge-based data have been leveraged to train state-of-the-art models, challenges persist for datasets that mirror real-world medical scenarios. In areas such as terminology normalization, medical literature QA, and electronic health record retrieval, the semantic alignment between queries and targets remains incomplete.

4.3.3 Analysis by Text Granularity.

From the perspective of granularity, as shown in Figure 3, retrieval performance tends to improve when the query and target share the same granularity. However, when the granularity is too fine, such as with phrases, performance may degrade due to the limited semantic information. Fine-grained retrieval is also more challenging. For instance, retrieving a paragraph with a sentence is more difficult than retrieving a document. Additionally, for SQL statements, direct alignment with natural

language is challenging without specific training.

4.3.4 Error Analysis

We conducted error analyses on bge-multilingual-gemma2, which exhibits strong performance across a wide range of retrieval tasks. We randomly selected 200 error cases, which evenly covered all 18 tasks on average. The errors were categorized into three types: fine-grained errors, semantic ambiguity, and lack of professional knowledge. Figure C1 presents the distribution of these error cases using a bar chart, where an error case may belong to multiple categories due to the overlapping nature of some error types. Additionally, we provide specific error examples in Appendix Section A for further illustration.

The results indicate that most errors stem from challenges in fine-grained retrieval, particularly in accurately capturing key numerical details and specific symptoms, where subtle distinctions are often overlooked. The model also struggles with semantic ambiguity, especially in long texts, and demonstrates limitations in domain-specific knowledge, particularly when multiple specialized concepts or technical terms are intertwined.

5 Ethical Considerations

This paper proposes a new medical-domain retrieval evaluation benchmark **MedEureka** for Embedding Models. All of the datasets in MedEureka adhere to ethical guidelines and respect copyright laws. The entire data collection process is free of copyright issues and privacy issues, and there are three types of data sources, including license applications, the open source community, and public file cleaning and organizing. Meanwhile, the manual participation part in the dataset construction process was all done by the authors of this paper without any ethical issues.

6 Conclusion

We have taken a significant step forward by developing **MedEureka**, a multi-granularity and multi-data-type evaluation benchmark designed to advance the study of embedding models in information retrieval scenarios. MedEureka encompasses six distinct medical data types: Table, Literature, Knowledge Base (KB), Terminology, Electronic Health Records (EHR), and Dialogue. It also includes four different text granularities, including Phrase, Sentence, Paragraph, and Document as well as a special SQL form, resulting in a total of eighteen datasets. These datasets provide a comprehensive resource for evaluating embedding models within the medical domain. We assessed fifteen state-of-the-art embedding models, trained two types of baseline models, and provided performance results and analyses across various formats. Furthermore, we examined the impact of different data types and granularities on retrieval performance.

Limitations

Evaluating medical retrieval tasks is challenging, primarily due to limited access to specialized resources, necessitating reliance on open-source data. Access to private data, like complete EHRs and cutting-edge studies, remains difficult. Additionally, balancing the benchmark is challenging, as some datasets, such as English EHRs and professional knowledge bases, are unbalanced. Besides, we have only constructed a cross-language dataset consisting of Chinese and English retrievals. There is still an opportunity to expand the dataset by adding more languages, which would provide a more comprehensive evaluation of the multilingual capabilities of embedding models. Moreover, MedEureka focuses on evaluating whether embeddings retrieve relevant content, using two classical metrics: MRR and Exact HR. However, it does not assess the relevance ranking of the retrieved content, as there is no labeling of the order of relevant documents, which is challenging to implement. Consequently, metrics like nDCG cannot be used. These limitations highlight areas for future research and improvement.

7 Acknowledgments

We would like to express our sincere gratitude to the anonymous reviewers for their valuable feedback. Our thanks also go to the Chairs and the organizing staff for their dedicated efforts in facilitating this work. This paper was supported by the Shanghai Sailing Program (No. 23YF1409400), National Natural Science Foundation of China (No. 62306112), and Shanghai Pilot Program for Basic Research (No. 22TQ1400100-20).

References

Suraj Agrawal, Dwaipayan Roy, and Mandar Mitra. 2021. Tag embedding based personalized point of

interest recommendation system. *Information Processing & Management*, 58(6):102690.

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. 2024. Litsearch: A retrieval benchmark for scientific literature search. *arXiv preprint arXiv:2407.18940*.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28th - August 2*.

José Castano, María Laura Gambarte, Hee Joon Park, Maria del Pilar Avila Williams, David Pérez, Fernando Campos, Daniel Luna, Sonia Benítez, Hernán Berinsky, and Sofía Zanetti. 2016. A machine learning approach to clinical terms normalization. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 1–11.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Nan Chen, Xiangdong Su, Tongyang Liu, Qizhi Hao, and Ming Wei. 2020. A benchmark dataset and case study for chinese medical question intent classification. *BMC Medical Informatics and Decision Making*, 20:1–7.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. Diffcse: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218.

DataTager. 2024. Extract medical information dataset. urlhttps://github.com/PandaVT/DataTager.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From distillation to hard negative sampling: Making sparse neural ir models more effective. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2353–2359.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 6894–6910. Association for Computational Linguistics (ACL).

Yunzhong He, Yuxin Tian, Mengjiao Wang, Feier Chen, Licheng Yu, Maolong Tang, Congcong Chen, Ning Zhang, Bin Kuang, and Arul Prakash. 2023. Que2engage: Embedding-based retrieval for relevant and engaging products at facebook marketplace. In *Companion Proceedings of the ACM Web Conference 2023*, pages 386–390.

Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2553–2561.

Zhongzhen Huang, Kui Xue, Yongqi Fan, Linjie Mu, Ruoyu Liu, Tong Ruan, Shaoting Zhang, and Xiaofan Zhang. 2024. Tool calling: Enhancing medication consultation via retrieval-augmented large language models. *arXiv preprint arXiv:2404.17897*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Sen Li, Fuyu Lv, Taiwei Jin, Guli Lin, Keping Yang, Xiaoyi Zeng, Xiao-Ming Wu, and Qianli Ma. 2021. Embedding-based product retrieval in taobao search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3181–3189.

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*.

Nut Limsopatham and Nigel Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1014–1023.

Jiduan Liu, Jiahao Liu, Qifan Wang, Jingang Wang, Wei Wu, Yunsen Xian, Dongyan Zhao, Kai Chen, and Rui Yan. 2023. Rankcse: Unsupervised sentence representations learning via learning to rank. In *Proceedings of the 61st Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 13785–13802.

Ye Liu, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S Yu. 2019. Generative question refinement with deep reinforcement learning in retrieval-based qa system. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1643–1652.

Alejandro Lozano, Scott L Fleming, Chia-Chun Chiang, and Nigam Shah. 2023. Clinfo. ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*, pages 8–23. World Scientific.

Man Luo, Sharad Saxena, Swaroop Mishra, Mihir Parmar, and Chitta Baral. 2022. Biotabqa: Instruction learning for biomedical table question answering. *arXiv preprint arXiv:2207.02419*.

Shiwei Lyu, Chenfei Chi, Hongbo Cai, Lei Shi, Xiaoyan Yang, Lei Liu, Xiang Chen, Deng Zhao, Zhiqiang Zhang, Xianguo Lyu, et al. 2023. Rjua-qa: A comprehensive qa dataset for urology. *arXiv preprint arXiv:2312.09785*.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. Covid-qa: A question answering dataset for covid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

Skatje Myers, Timothy A Miller, Yanjun Gao, Matthew M Churpek, Anoop Mayampurath, Dmitriy Dligach, and Majid Afshar. 2024. Lessons learned on information retrieval in electronic health records: a comparison of embedding models and pooling strategies. *Journal of the American Medical Informatics Association*, page ocae308.

Sosuke Nishikawa, Ryokan Ri, Ikuya Yamada, Yoshimasa Tsuruoka, and Isao Echizen. 2022. Ease: Entity-aware contrastive learning of sentence embedding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3870–3885.

OpenAI. 2022. New and improved embedding model. Technical report.

OpenAI. 2024a. Hello gpt-4o. Technical report, OpenAI.

OpenAI. 2024b. New embedding models and api updates. Technical report.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Petros Papadopoulos, Mario Soflano, Yaelle Chaudy, Wilson Adejo, and Thomas M Connolly. 2022. A systematic review of technologies and standards used in the development of rule-based clinical decision support systems. *Health and Technology*, 12(4):713–727.

Preethi Raghavan, Diwakar Mahajan, Jennifer Liang, Rachita Chandra, and Peter Szolovits. 2021. emrkbqa: A clinical knowledge-base question answering dataset. *NAACL-HLT 2021*, page 64.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, et al. 2018. Data and systems for medication-related text classification and concept normalization from twitter: insights from the social media mining for health (smm4h)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283.

Tim Schopf, Daniel Braun, and Florian Matthes. 2022. Lbl2vec: An embedding-based approach for unsupervised document retrieval on predefined topics. *arXiv preprint arXiv:2210.06023*.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Sonish Sivarajkumar, Haneef Ahamed Mohammad, David Oniani, Kirk Roberts, William Hersh, Hongfang Liu, Daqing He, Shyam Visweswaran, and Yanshan Wang. 2024. Clinical information retrieval: A literature review. *Journal of Healthcare Informatics Research*, pages 1–40.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir:

A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Toyhom. 2019. Chinese-medical-dialogue-data. urlhttps://github.com/Toyhom/Chinese-medical-dialogue-data.

Hao Wang and Yong Dou. 2023. Sncse: Contrastive learning for unsupervised sentence embedding with soft negative samples. In *International Conference on Intelligent Computing*, pages 419–431. Springer.

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*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Qiong Wang, Zongcheng Ji, Jingqi Wang, Stephen Wu, Weiyan Lin, Wenzhen Li, Li Ke, Guohong Xiao, Qing Jiang, Hua Xu, et al. 2020. A study of entity-linking methods for normalizing chinese diagnosis and procedure terms to icd codes. *Journal of biomedical informatics*, 105:103418.

He sicheng Wang Yuxin, Sun Qingxuan. 2023. M3e: Moka massive mixed embedding model.

Xing Wu, Chaochen Gao, Zijia Lin, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022. Infocse: Information-aggregated contrastive learning of sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3060–3070.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Dongfang Xu, Zeyu Zhang, and Steven Bethard. 2020. A generate-and-rank framework with semantic type regularization for biomedical concept normalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8452–8464.

Hua Xu, Dina Demner Fushman, Na Hong, and Kalpana Raja. 2024. Medical concept normalization. In *Natural Language Processing in Biomedicine: A Practical Guide*, pages 137–164. Springer.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7346–7353.

Feihong Yang, Xuwen Wang, and Jiao Li. 2021. Exploration and research of bert model in chinese clinical natural language processing. https://github.com/trueto/medbert.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. tydi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137.

A Cases of Three Classical Error Types

Fine-grained error: This occurs when the embedding vectors fail to sufficiently distinguish key information during retrieval. For example, in response to the query "*What are the abnormalities of the gastrointestinal, genitourinary, and neurological systems that are manifested during the physical examination of a pregnant woman who is vomiting? How do these manifestations help diagnose the underlying etiology?*", the most relevant erroneously retrieved document correctly identified pregnancy and symptoms such as vomiting but mistakenly provided information related to early-stage pregnancy instead of addressing the requested abnormalities.

Semantic ambiguity: This type of error arises when the embedding vector captures only partial semantics while overlooking the overall meaning of the query. For instance, for the query "*What characterises the epidemiological distribution of rabies, Powassan encephalitis and West Nile virus encephalitis globally? Please describe in detail the main endemic regions for each condition.*", the most relevant retrieved tables incorrectly focused only on encephalitis and its endemic regions but failed to account for the specific diseases mentioned, leading to the retrieval of information about "some arboviral encephalitis" rather than the targeted conditions.

Lack of professional knowledge: This occurs when the embedding model struggles to encode specialized medical terminology, leading to the retrieval of content unrelated to the medical terms in the query. For example, in response to the query "*What are the results of anti-HAV IgM and anti-HAV IgG antibodies in serological testing for acute hepatitis A? How do these results help confirm the diagnosis of acute infection?*", the most relevant erroneously retrieved document contained no information about acute hepatitis A or the specified antibodies. Instead, it retrieved content related to hepatitis B and its corresponding antibodies.

B The specific construction details for Dialogue Task

For **MeQSum**, we only partitioned the dataset without applying any additional processing, resulting in a total of **1,000** test samples.

For **Chinese-Medical-Dialogue**, we first analyzed the dataset and identified data from six medical departments. We then clustered the data within each department based on query embedding vectors using the **bge-m3** model. The number of clusters obtained for each department is shown in Table B1:

Department	**Clusters**
Andriatria	18,298
Internal Medicine (IM)	51,826
Obstetrics and Gynecology (OAGD)	48,204
Oncology	21,360
Pediatrics	29,044
Surgery	34,434

Table B1: Cluster statistics for each medical department.

To ensure a more balanced data distribution, we randomly selected one sample from each cluster after clustering and then randomly sampled **500** instances per department, yielding a final test set of **3,000** samples.

For **Qnorm**, due to the high similarity among QA pairs, we applied a filtering process based on **relative edit distance**, setting a threshold of **0.85** to extract **1,599** challenging test samples.

C Supplementary materials for dataset and experiment

Figure C1 presents the distribution of these error cases

Figure C2 to Figure C10 presents the sample data each task.

Figure C11 illustrates the specific prompt used for automatic query generation, taking the Table task as an example.

Figure C12 illustrates the specific query prompts used for each task.

Figure C14 presents the performance and trend of embedding models on the Chinese task, depicted as a line chart from Exact HR@5 to Exact HR@500. The Exact HR@num represents the exact hit rate with num indicating the number of candidates. Similarly, the performance of embedding models on the English and cross-lingual tasks is illustrated in Figure C13.

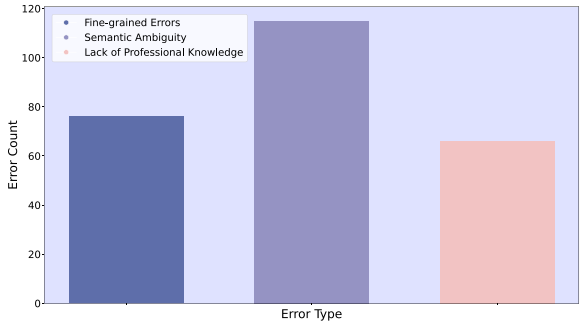

Figure C1: Distribution of these error cases on bge-multilingual-gemma2.

Model	Table		Literature		KB		Term					EHR				Dialogue		
	I	II	I	II	I	II	I	II	III	IV	V	I	II	III	IV	I	II	III
bert-base-uncased*	-	0.70	-	3.82	-	-	-	-	14.52	6.81	6.26	-	-	-	-	-	0.94	-
biobert-base-cased-v1.2*	-	0.36	-	4.10	-	-	-	-	16.13	8.33	6.70	-	-	-	-	-	0.88	-
eureka-sup-simcse-bert-base-uncased	-	67.69	-	54.18	-	-	-	-	**98.55**	74.84	**85.56**	-	-	-	-	-	68.28	-
eureka-unsup-simcse-bert-base-uncased	-	6.20	-	8.97	-	-	-	-	26.59	11.57	10.27	-	-	-	-	-	13.42	-
eureka-sup-simcse-biobert-base-cased-v1.2	-	75.34	-	60.75	-	-	-	-	98.45	74.11	85.37	-	-	-	-	-	73.83	-
eureka-unsup-simcse-biobert-base-cased-v1.2	-	35.31	-	16.82	-	-	-	-	45.57	18.83	18.99	-	-	-	-	-	39.98	-
eureka-sbert-bert-base-uncased	-	78.45	-	59.57	-	-	-	-	96.70	**77.42**	82.90	-	-	-	-	-	77.94	-
eureka-sbert-biobert-base-cased-v1.2	-	**84.52**	-	**65.33**	-	-	-	-	95.12	75.35	77.37	-	-	-	-	-	**81.25**	-
bert-base-chinese*	1.56	-	11.27	-	0.19	0.60	23.11	-	-	-	-	1.62	1.08	0.09	0.11	2.60	-	28.03
medbert-base-wwm-chinese*	0.92	-	3.72	-	0.15	0.33	19.08	-	-	-	-	0.20	0.06	0.02	0.43	0.27	-	19.62
eureka-sup-simcse-bert-base-chinese	63.29	-	83.80	-	92.37	**99.85**	47.98	-	-	-	-	94.89	85.65	**83.58**	45.83	89.82	-	**90.48**
eureka-unsup-simcse-bert-base-chinese	37.32	-	12.18	-	20.13	46.31	42.78	-	-	-	-	83.78	40.04	35.30	11.83	66.41	-	59.45
eureka-sup-simcse-medbert-base-wwm-chinese	64.28	-	79.41	-	93.88	99.77	50.03	-	-	-	-	94.77	84.73	80.69	44.60	80.32	-	86.61
eureka-unsup-simcse-medbert-base-wwm-chinese	34.91	-	20.86	-	17.60	28.81	44.41	-	-	-	-	81.81	40.15	37.70	14.19	64.00	-	57.67
eureka-sbert-bert-base-chinese	67.91	-	80.87	-	**95.74**	99.77	49.26	-	-	-	-	**94.92**	85.56	81.02	**55.80**	**92.45**	-	90.17
eureka-sbert-medbert-base-wwm-chinese	**69.61**	-	**85.62**	-	94.95	99.77	**50.69**	-	-	-	-	95.20	**85.82**	82.79	54.24	92.02	-	87.45
bert-base-multilingual-uncased*	0.52	0.76	3.08	5.90	0.04	0.47	18.06	0.48	16.17	7.10	6.35	0.83	0.88	0.43	0.15	0.16	1.12	20.24
eureka-sup-simcse-bert-base-multilingual-uncased	63.20	80.64	82.01	**59.78**	**95.42**	**100.00**	49.42	**83.44**	**98.55**	**75.09**	**86.34**	95.00	**85.62**	**83.21**	46.80	85.85	76.13	**89.73**
eureka-unsup-simcse-bert-base-multilingual-uncased	48.71	51.90	24.97	23.10	20.83	66.66	45.54	4.26	51.49	20.78	22.87	86.98	41.89	38.04	12.14	70.28	57.81	56.48
eureka-sbert-bert-base-multilingual-uncased	**64.18**	**83.58**	**83.58**	57.70	94.46	99.77	**49.71**	80.90	96.23	74.60	79.67	**95.05**	82.50	79.13	**47.41**	**87.55**	**76.42**	89.04

Table C1: Comparison of training baseline models performance on MRR@10 using cosine similarity as the distance metric. "$*$" indicates direct retrieval with frozen base PLMs

Table I

Query: 氮卓斯汀、色甘酸钠和奥洛他定作为鼻内肥大细胞稳定剂在初始剂量方面有何不同？请详细描述各自的使用年龄段和相应的初始剂量。

Target:

药物	每喷剂量	初始剂量
氮卓斯汀（Azelastine）*,†	0.1%溶液中137微克 0.15%溶液中205.5毫克	2-5年： 针对季节性过敏性鼻炎，每天两次，每次鼻孔喷0.1%溶液 6 个月至 5 岁： 每个鼻孔喷1次0.1%的溶液，每天两次，用于治疗常年性过敏性鼻炎 6-11 岁： 0.1%或0.15%的溶液，每个鼻孔1喷，每天2次，或每个鼻孔2喷，每天1次 ≥ 12岁： 0.15%的溶液，每个鼻孔1-2喷，每天2次
色甘酸钠	5.2mg	≥6岁: 每个鼻孔 1 喷，每天 3 或 4 次
奥洛他定*	665微克	6～11岁： 每个鼻孔 1 喷，每天 2 次 ≥ 12岁： 每个鼻孔 2 喷，每天 2 次
* 氮卓斯汀和奥洛帕坦是双重作用的肥大细胞稳定剂/抗组胺药。		
† 可使用氮卓斯汀/氟替卡松 (137 mcg/50 mcg) 的组合。初始剂量为每个鼻孔1喷，每天两次。		

Table II

Query: During cardiopulmonary resuscitation (CPR) for children of different ages, the compression techniques for newborns, 1-year-olds, and 8-year-olds vary. For newborns, it is recommended to use thumb compressions with hands encircling the chest, or two-finger compressions; for 1-year-olds, single-hand compressions are advised; and for 8-year-olds, two-hand compressions are recommended. Please explain in detail the specific implementation methods of these different compression techniques and the reasons for their applicability at different ages.

Target:

Age (yr)	Term neonate	< 12 mo	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Weight, typical (kg)	3.5	< 10	10	12	14	16	18	20	22	25	28	30	35	40	45	50	55	60
Compression techniques	Thumb compression, hands around chest (preferred) or 2 fingers	1 hand	2 hands															
Airway size (Portex) in cm	000	00	00	0	0	7	7	7	7	7	7	7	7	8	8	8	8	
3.5	5	5	6	6														
Masks in Laerdal sizes or equivalent	Circular 0/1	Rendell-Baker type # 1	Rendell-Baker type # 2	Dome cuff mask # 3	Dome cuff mask # 4													
Ventilation bag with reservoir for 100% O2 delivery	Infant 240 mL	Child 400–500 mL	Adult 1600 mL															
Laryngoscope blade size	Miller 0 Straight blade	1	1	1	2	2	2	2	2	2	2	3	3	3	3	3	3	3
	Straight blade (preferred) or curved blade	Curved or straight blade																
ETT size (Portex) in mm	3	3.5	4	4.5	4.5	5	5	5.5	5.5	6	6	6	6.5	6.5	6.5	6.5	7	
Uncuffed	Uncuffed	Cuffed																
Suction catheter	Direct oropharyngeal Through ETT	10 F	Pediatric tonsil suction8 Fr	Adult tonsil suction10 Fr														
ETT =endotracheal tube; Fr = French.																		
Courtesy of Dr. B. Paes and Dr. M. Sullivan, the Departments of Pediatrics and Medicine, St. Joseph's Hospital, The Children's Hospital, Hamilton Health Sciences Corporation, McMaster University, Hamilton, Ontario, Canada.																		

Figure C2: Sample data for the **Table** Task.

Literature I

Query: 患者：医生您好，我是一名50岁的男性患者。最近我出现了一些排尿方面的问题，包括进行性排尿困难、尿频和尿急的情况。我还注意到尿等待时间变长，尿流变细，尿不尽，尿分叉，而且夜间需要起床3-5次上厕所。这些症状困扰了我一段时间了，我想请问您有什么治疗方案可以帮助我缓解这些症状吗？谢谢。

Target:

1. 慢性前列腺炎患者病程长，通常3~6个月及以上。II型与III型前列腺炎的临床表现类似且具有多样性;症状在同一患者的不同阶段，以及不同患者之间存在差异，主要表现为以下症状。疼痛是慢性前列腺炎最主要的临床表现。最常见的是会阴区疼痛不适(63%)，疼痛还可见于睾丸(58%)、耻骨区(42%)及阴茎(32%);患者也可出现尿道、肛周、腹股沟、腰骶部及下背部的疼痛。与排尿症状相比，疼痛症状对患者生活质量的影响更高，而疼痛的严重程度和频率比疼痛的部位和类型影响更大，当疼痛发生于骨盆外时，患者疼痛症状往往较为广泛，其社会心理健康及生活质量也较骨盆内者差。射精时或射精后的疼痛不适(45%)也是慢性前列腺炎重要的非特异性临床表现。慢性前列腺炎的另一个重要临床表现是储尿期和排尿期症状，包括尿频、尿急、夜尿增多、排尿等待、排尿中断等。此外，约62%的慢性前列腺炎患者伴有性功能障碍，40%的患者可出现早泄，其疼痛程度与性功能障碍密切相关。
2. II型和III型：须详细询问病史，尤其是反复下泌尿道感染史，全面体格检查(包括直肠指检)，尿液和前列腺按摩液常规检查。推荐应用NIH慢性前列腺炎症状评分(NIHchronicprostatitissymptomindex，NIH-CPSI，见附录15-2)进行症状评分。推荐“两杯法”或“四杯法”(见附录15-3)进行病原体定位试验(表15-1)。为明确诊断需对类似症状的疾病进行鉴别。
3. 前列腺炎应采取个体化的综合治疗。II型：推荐以口服敏感抗生素治疗为主，疗程为4~6周，建议治疗2周后对患者进行阶段性的疗效评价。如抗生素疗效不满意者，可改用其他敏感抗生素。伴有下尿路刺激症状的患者推荐联合使用α受体阻滞剂、植物制剂和M受体阻滞剂等改善症状。IIIA型：可先口服抗生素2~4周，后续是否继续抗生素治疗取决于前期的疗效反馈。推荐结合使用α受体阻滞剂、植物制剂、非甾体抗炎镇痛药和(或)M受体阻滞剂等改善排尿症状和疼痛症状。IIIB型：推荐使用α受体阻滞剂、植物制剂、非甾体抗炎镇痛药和M受体阻滞剂等药物治疗。

Figure C3: Sample data for the **Literature** Task (1/2).

Literature II

Query: Where can published genomic sequences be found for the 2019-nCoV virus?

Target: "Note from the editors: novel coronavirus (2019-nCoV)\n\nhttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6988271/\n\nSHA: d958168df85240e544a918d843a14e887dc41d2b\n\nAuthors: nan\nDate: 2020-01-23\nDOI: 10.2807/1560-7917.es.2020.25.3.2001231\nLicense: cc-by\n\nAbstract: nan\n\nText: The situation has continued to evolve rapidly since then and just a few weeks later, as at 23 January, 614 laboratory-confirmed cases and 17 deaths have been reported [2] including some cases detected outside mainland China [3] . Meanwhile, on 7 January 2020, the novel coronavirus, currently named 2019-nCoV, was officially announced as the causative agent by Chinese authorities [3] . In order to support public health action, viral genome sequences were released by Chinese researchers on 10 January [4] and 2 days later, four further sequences were also made available on the Global Initiative on Sharing All Influenza Data (GISAID) (https://www.gisaid.org/). While more cases are being reported on a daily basis and there is evidence for some human-to-human transmission in China, a number of important questions remain unanswered. For example, there is no certainty about the source of the outbreak, the transmissibility of the virus as well as the clinical picture and severity of the disease.\n\nIn this issue of Eurosurveillance, we are publishing two articles on different aspects of the newly emerged 2019-nCoV. One is a research article by Corman et al. on the development of a diagnostic methodology based on RT-PCR of the E and RdRp genes, without the need for virus material; the assays were validated in five international laboratories [5] . Before this publication, a description of the assays had already been made publically available on a dedicated WHO webpage [6] to support rapid development of laboratory testing capacities. The other is a rapid communication where researchers based in Hong Kong report on their attempt to estimate the severity among hospitalised cases of 2019-nCoV infection through modelling based on publically available information, mainly from Wuhan health authorities [7] . It also points out the need for more detailed information to make an informed evaluation of the situation as basis for public health decision-making.\n\nToday, the WHO Director-General Tedros Adhanom Ghebreyesus, taking into consideration the deliberations of the members of the International Health Regulations (IHR) Emergency Committee on 2019-nCoV in their second meeting, decided not to declare a public health emergency of international concern. \n\nInternational health organisations such as the European Centre for Disease Prevention and Control (ECDC) and the WHO are monitoring the situation and provide regular updates. ECDC has set up a dedicated webpage on which updates and risk assessments with focus on Europe are available: https://www.ecdc.europa.eu/en/ novel-coronavirus-china.

Figure C4: Sample data for the **Literature** Task (2/2).

KB I

Query: 某女，3岁。症见发热头痛、腹胀满、咳嗽痰多、呕吐酸腐。医师处以小儿百寿丸。该药剂组成为：钩藤45g、炒僵蚕45g、胆南星(酒炙)75g、天竺黄75g、桔梗30g、木香75g、砂仁45g、陈皮75g、麸炒苍术75g、茯苓30g、炒山楂150g、六神曲(麸炒)45g、炒麦芽45g、薄荷45g、滑石150g、甘草30g、朱砂10g、牛黄10。该药剂中牛黄的具体制备方法是什么？在该药剂中，牛黄的含量测定方法是什么？

Target:
药名:小儿百寿丸。制法:以上十八味，除牛黄外，朱砂水飞成极细粉；其余钩藤等十六味粉碎成细粉；将牛黄研细，与上述粉末配研，过筛，混匀。每100g粉末加炼蜜100～120g制成大蜜丸，即得。

药名:小儿百寿丸。含量测定:木香照高效液相色谱法（通则0512)测定。色谱条件与系统适用性试验以十八烷基硅烷键合硅胶为填充剂；以甲醇-0.1%磷酸溶液（63：37)为流动相；检测波长为225nm。理论板数按去氢木香内酯峰计算应不低于14000。对照品溶液的制备取木香烃内酯对照品、去氢木香内酯对照品适量，精密称定，加甲醇制成每1ml各含30μg的混合溶液，即得。供试品溶液的制备取重量差异项下的本品，剪碎，混匀,取约3g,精密称定,精密加人甲醇50ml,密塞，称定重量，超声处理（功率200W，频率40kHz)45分钟，放冷，再称定重量，用甲醇补足减失的重量，摇匀，滤过，取续滤液，即得。测定法分别精密吸取对照品溶液与供试品溶液各10ul,注入液相色谱仪，测定，即得。本品每丸含木香以木香烃内酯（C5H2oO2）和去氢木香内酯(C1sH1sO2）的总量计,不得少于1.25mg。牛黄照高效液相色谱法(通则0512)测定(避光操作)。色谱条件与系统适用性试验以十八烷基硅烷键合硅胶为填充剂；以乙晴-1%冰醋酸溶液（95：5)为流动相；检测波长为450nm。理论板数按胆红素峰计算应不低于3000。对照品溶液的制备取胆红素对照品适量，精密称定，加二氯甲烷制成每1ml含15μg的溶液,即得。供试品溶液的制备取重量差异项下的本品，剪碎，取适量，精密称定，精密加人硅藻土适量（约为取样量的2倍），混合均匀后充分研磨成细粉，取细粉约1.5g（相当于本品0.5g),精密称定,置具塞锥形瓶中,加入10%草酸溶液（含0.15%十六烷基三甲基氯化铵）10ml，密塞，涡旋至充分混匀，精密加人水饱和的二氯甲烷50ml，密塞，称定重量，涡旋至充分混匀，超声处理（功率500W，频率53kHz）40分钟，放冷，再称定重量，用水饱和的二氯甲烷补足减失的重量，摇匀，离心，取二氯甲烷液，滤过，取续滤液，即得。测定法分别精密吸取对照品溶液与供试品溶液各5pl，注人液相色谱仪，测定，即得。本品每丸含牛黄以胆红素（C33H3N4O）计，不得少于2.2mg。

Figure C5: Sample data for the **KB** Task (1/2).

KB II

Query: 某女，45岁。症见肝肾阴亏，眩晕耳鸣，羞明畏光，迎风流泪，视物昏花。医师处以杞菊地黄丸。该药剂组成为：枸杞子40g菊花40g熟地黄160g酒萸肉80g牡丹皮60g山药80g茯苓60g泽泻6。在鉴别过程中，枸杞子的颜色变化和显色特征有哪些？

Target:
药名:杞菊地黄丸(浓缩丸)
处方:枸杞子40g菊花40g熟地黄160g酒萸肉80g牡丹皮60g山药80g茯苓60g泽泻60g\n制法:以上八味，取酒英肉26.7g、牡丹皮26.5g、山药粉碎成细粉；泽泻、茯苓加水煎煮二次，第一次3小时，第二次2小时，滤过，滤液合并并浓缩成相对密度为1.30～1.35（60～80℃）的稠膏；熟地黄切片，加水煎煮三次，第一次3小时，第二次2小时，第三次1小时，滤过，滤液合并并浓缩成相对密度为1.30～1.35（60～80℃）的稠膏；枸杞子以45%乙醇作溶剂，剩余的酒英肉与牡丹皮及菊花以70%乙醇作溶剂，浸渍24小时后，分别进行渗漉，收集漉液，合并上述漉液，回收乙醇浓缩成相对密度为1.30～1.35（60～80°℃)的稠膏，与上述细粉与稠膏混匀,制成浓缩丸,干燥，打光，即得。
性状:本品为棕色至棕黑色的浓缩丸；味甜而酸。
鉴别:⑸项下的供试品溶液及上述对照药材溶液和对照品溶液各5μl,分别点于同一硅胶G薄层板上，以甲苯-乙酸乙酯-冰醋酸（24：8：1)为展开剂，展开，取出，晾干，喷以10%硫酸乙醇溶液，在105℃加热至斑点显色清晰。供试品色谱中，在与对照药材色谱和对照品色谱相应的位置上，显相同的紫红色斑点。⑸取本品6g，研碎，加乙醚40ml.加热回流1小时，滤过，滤液挥去乙醚，残渣加丙酮1ml使溶解，作为供试品溶液。另取牡丹皮对照药材1g，同法制成对照药材溶液。再取丹皮酚对照品，加丙酮制成每1ml含1mg的溶液，作为对照品溶液。照薄层色谱法（通则0502)试验，吸取上述三种溶液各10μl,分别点于同一硅胶G薄层板上，使成条状，以环已烷-乙酸乙酯（3：1）为展开剂，展开，取出，晾十，喷以盐酸酸性5%三氯化铁乙醇溶液，加热至斑点显色清晰。供试品色谱中，在条斑。
检查:应符合丸剂项下有关的各项规定（通则0108）。
含量测定:照高效液相色谱法（通则0512）测定。色谱条件与系统适用性试验以十八烷基硅烷键合硅胶为填充剂；以乙晴为流动相A，以0.3%磷酸溶液为流动相B，按下表中的规定进行梯度洗脱；莫诺苷和马钱苷检测波长为240nm，丹皮酚检测波长为274nm；柱温为40℃。理论板数按莫诺苷、马钱苷峰计算均应不低于4000。
表格:|时间(分钟)|流动相A(%)|流动相B(%)||---|---|---||0~5|5→8|95→92||5～20|8|92||20~35|8→20|92→80||35~45|20→60|80→40||45~55|60|40|对照品溶液的制备取莫诺苷对照品、马钱苷对照品和丹皮酚对照品适量，精密称定，加70%甲醇制成每1ml中含莫诺苷与马钱苷各20ug、含丹皮酚45ug的混合溶液，即得。供试品溶液的制备取重量差异项下的本品，研细，取约0.3g，精密称定，置具塞锥形瓶中，精密加人70%甲醇25ml，密塞，称定重量，加热回流1小时，放冷，再称定重量，用70%甲醇补足减失的重量，摇匀，滤过，取续滤液，即得。测定法分别精密吸取对照品溶液与供试品溶液各10ul，注人液相色谱仪，测定，即得。本品每丸含酒英肉以莫诺苷（C17H26O11）和马钱苷（C17H2601o）的总量计，不得少于0.28mg；含牡丹皮以丹皮酚（CgHoO）计，不得少于0.20mg。
功能与主治:滋肾养肝。用于肝肾阴亏，眩晕耳鸣，羞明畏光，迎风流泪，视物昏花。
用法与用量口服:一次8丸，一日3次。
规格:每8丸相当于原药材3g
贮藏:密封。

Figure C6: Sample data for the **KB** Task (2/2).

Term I

Clinical_examination

Query: 二氧化碳容积图 Target: 二氧化碳图形

Disease_dignosis

Query: 腰椎间盘突出症 Target: 硬膜内型腰椎间盘突出症

Procedure_operation

Query: 左肾根治性切除 Target: 单侧肾切除术

Symptom_sign

Query: 延髓背外侧综合征 Target: 瓦伦贝格综合征

Term II

Clinical_examination

Query: optical coherence tomography Target: 光相干断层扫描

Disease_dignosis

Query: dysfunction after cardiac surgery Target: 心脏手术后功能障碍

Procedure_operation

Query: reduction of vertebral fracture Target: 脊椎骨折复位术

Symptom_sign

Query: autonomic dysreflexia Target: 自主神经反射障碍

Term III

Query: GI distress Target: Excessive upper gastrointestinal gas

Term IV

Query: accident i kept waking up Target: Middle insomnia

Term V

Query: Mental illness Target: Psychotic Disorders

Figure C7: Sample data for the **Term** Task.

EHR I

Query: 主要症状包含上腹痛、发热的病人,

Target:
SELECT * FROM Outpatient_Medical_Record WHERE Chief_Complaint LIKE '%上腹痛%' AND Chief_Complaint LIKE '%发热%'

EHR II

Query: 现病史中包含右股骨干粉碎性骨折的患者

Target:
现病史:患者源于1小时前因出车祸致右大腿肿痛畸形、活动受限，无皮破流血，伤时无昏迷、近事遗忘，伤后无明显头晕头痛、恶心呕吐，无咳嗽咯血，无胸闷胸痛及呼吸困难，无腹胀腹痛等症，伤后120急送我院，急诊医生予询问病史、查体及拍片检查等处理，拍片检查示“右股骨中段粉碎性骨折”予夹板外固定后拟“右股骨干粉碎性骨折”收住我科住院进一步治疗。入院症见患者神清，右大腿肿痛畸形、活动受限，不能站立行走，无口苦口干，无畏寒发热，无呼吸困难、无胸闷心悸、无自汗盗汗，伤前纳寐正常，二便自调。

现病史:该患者于2小时前不慎滑倒，伤及右大腿，导致肿痛，畸形，活动受限，到当地医院拍片后诊断为右股骨干粉碎性骨折。为求进一步手术治疗急来我院，经门诊检查并阅片后，以右股骨干粉碎性骨折收入院，现症：右大腿肿痛，活动受限，饮食、睡眠尚可，二便正常。

EHR III

Query: SELECT * FROM Outpatient_Medical_Record WHERE History_of_Present_lllness LIKE '%慢性肾功能不全%'

Target:
现病史:患者30余年前因腰痛查B超提示多囊肾，未予特殊治疗。3年前体检发现血肌酐260umol/L，长期于门诊服中药护肾治疗，一年余前因“多囊肾出血”于我院住院，复查肌酐369.4umol/L，予抗感染、止血等治疗后好转出院。一个月前患者因“慢性肾功能不全”于我院住院，出院时复查肌酐801umol/L，3天前再次出现肉眼血尿，色鲜红，时有血块，遂今至我院就诊，为求进一步治疗入住我科。刻下：患者解肉眼血尿，色鲜红，偶有刺痛，无尿频、尿急，头晕乏力感，腹部时有疼痛，无头痛，无胸闷心慌，无畏寒发热，无手足麻木，无腹胀，纳差，大便尚调，夜寐尚安。

现病史:患者于二十余天无明显诱因出现双下肢水肿，呈对称性压陷性水肿，伴乏力，偶有清晨双眼睑轻度浮肿，下午减轻，偶有心慌、喘气，无畏寒、发热，无头晕、头痛，无咳嗽、咳痰，无胸痛、胸闷，无恶心、呕吐，无腹胀、腹痛、腹泻，无大小便失禁及肢体活动障碍，在家未予治疗，于今日上午患者出现头晕，口服降压药物后好转，伴有后颈部、肩部胀痛不适，今来我院，门诊以\"慢性肾功能不全\"收住我科。起病以来，患者精神、食欲、睡眠欠佳，大小便正常，体力稍下降，体重无明显变化。

Figure C8: Sample data for the **EHR** Task (1/2).

EHR IV

Query: 查体中有周身皮肤微红、肿胀、充血等表现，血常规示白细胞、中性粒细胞比率等指标异常的患者

Target:
基本信息:男，53岁，农民
主诉:落冰水后全身僵硬麻木4小时。
现病史:患者诉缘于4小时前在冰上巡逻时不慎掉入冰水中，5分钟后获救成功，伤后全身皮肤苍白、冰凉、僵硬，厥冷，四肢麻木，无昏迷抽搐，无呼吸浅慢及呼吸困难，无恶心呕吐，无咳嗽、咳痰及咯血，自述胸痛，无胸闷、心悸，无二便失禁。伤后于当地复温，输液治疗（药名及剂量不详），自觉周身皮肤热、痒、灼痛，为求进一步诊治，来我院，经门诊以“冻伤”收住院。
既往史:既往体健，无手术史，外伤史及药物过敏史，否认“肝炎”、“结核”等传染病接触史。
查体:T:37.6℃，P:98次/分，R:20次/分，BP:130/80mmHg。发育正常，营养中等，神志清楚，合作。全身粘膜无苍白，无黄疸，皮肤弹性差，未见肝脏、蜘蛛痣。余见外科情况。全身浅表淋巴结未触及肿大。头颅外形正常。结膜无苍白，巩膜无黄染，角膜无混浊，双侧瞳孔等大等圆，直径约2.5mm，眼球运动正常，光反射存在。耳廓外形正常，外耳道无异常分泌物，乳突无压痛，鼻无畸形，鼻腔粘膜无充血、水肿，鼻中隔无偏曲，鼻翼无扇动，各副鼻窦无压痛。口唇无苍白，颊粘膜无溃疡、白斑，伸舌居中。咽后壁无红肿，悬雍垂居中，双侧扁桃体Ⅱ°肿大，表面无脓性分泌物，喉发音清晰。颈略抵抗，未见颈静脉怒张，颈动脉无异常搏动及杂音。气管居中，甲状腺无肿大。胸廓对称，无畸形。胸壁无静脉曲张，未及皮下气肿。胸式呼吸，双侧呼吸动度一致，肋间隙无增宽。语颤无增强及减弱，无捻发感及胸膜摩擦感。双肺叩诊清音，肺肝浊音介于右锁骨中线第五肋间。双侧呼吸音清晰，下野闻及细小湿性啰音。心前区无隆起，心尖搏动位于第五肋间左锁骨中线内侧1.5cm。无震颤及心包摩擦感。心浊音界无扩大。心律98次/分，律齐，各瓣膜区未闻及病理性杂音，未闻及心包摩擦音。无脉搏短拙，无奇脉及大动脉枪击音。无水冲脉，毛细血管搏动征（-）。腹部平坦，无腹壁静脉曲张，未见胃肠型及逆蠕动波，腹部无压痛、反跳痛、肌紧张，肝脾肋下未触及， Murphy征阴性，麦氏点无压痛，肝肾区无叩击痛，叩诊无移动性浊音，肠鸣音4-6次/分，未闻及气过水声及金属音。肛门及外生殖器未见异常。脊柱四肢无畸形，四肢关节活动正常。双侧肢体肌力正常，肌张力正常，角膜反射、腹壁放射、肱二、三头肌腱反射、膝腱反射正常存在，颈项无强直，双侧巴氏征、布氏征阴性。外科情况：周身皮肤微红、肿胀，充血，无水泡，皮肤痛温感觉略迟钝，无感觉过敏，肢体可持重。
辅助检查: 血常规示：白细胞10.03*10^9/L，中性粒细胞比率66.90%，淋巴细胞比率26.80%，中性粒细胞数6.71*10^9/L，淋巴细胞数2.69*10^9/L。血凝四项示：凝血酶原时间13.4Sec，国际标准化比率1.07，部分活化凝血酶原时间27.0Sec，纤维蛋白原2.88g/L，凝血酶时间13.6Sec。

Figure C9: Sample data for the **EHR** Task (2/2).

Dialogue I

Query:反胃很长一段时间了，平时出现呕吐，反胃，吃不下饭，恶心的东西看了也想吐，我想问这样持续久了会不会有大问题？病因是什么？谢谢,目前是天天吃药，但是没什么效果，感觉反胃越来越重了，会不会有严重的问题?

Target:
我长期经历反胃、呕吐、食欲不振和恶心，对恶心食物有强烈反应。目前每天服药但效果不佳，反胃症状似乎在加重。我想了解这种持续的反胃是否会导致严重健康问题，以及可能的病因是什么。

Dialogue II

Query:Is titanium dioxide an inactive ingredient in Equate acetaminophen or Tylenol acetaminophen?

Target:
SUBJECT: Acetaminophen inactive ingredients
MESSAGE: My wife is severely allergic to Titanium Dioxide. It is not listed in the inactive ingredients for Equate or Tylenol Acetaminophen on the package or on your page but other sources on the Internet claim it is used to coat the tablets to make them easier to swallow. Are they coated with Titanium Dioxide and if so why is it not listed as an inactive ingredient? Thanks, [NAME]

Dialogue III

Query: 继发性小儿癫痫应该怎样治疗,小孩子有继发性癫痫应该怎样治疗呢。

Target:
小儿继发性癫痫怎么康复,我的孩子现在12个半月，宝宝出世20多天，30天前后出现4次抿嘴，身体往后仰，头往右边，斗鸡眼的情况。7个多月出现撇嘴，每次都是单一的一个表情，一天二三十次，精神状况不好。然后经当地医生建议，做了24小时脑电图和脑部CT。确诊为大脑发育不良引起的继发性癫痫。确诊后开始吃德巴金3ml，每天两次。之后发作间隔由拉长，而且发作时精神状态良好，喊他都能回应，照样玩。这一个多月来发现发作表情更为夸张，握紧拳头，好像要使很大力气。怕药量不够，就是验了血液浓度，都在正常值内。最近发作间隔短，表情夸张，但精神状态良好。不过有感冒，咳嗽。

Figure C10: Sample data for the **Dialogue** Task.

query_generation

我将给你一份来源于专业医学诊疗手册的表格。首先我需要你理解该表格的空间布局以及包含的信息类型，并选择某一列作为提问的依据；其次，从你所选取的列中挑选出两到三行，在理解行内容的基础上，根据这些行对应单元格中的信息生成一个问题，要求提问要专业且具有逻辑性，严格涉及了所有筛选出的单元格的信息，输出应包含以下五个部分：

表名：表格的名称
列：选取的列
行：选取的行集
问题：生成的问题
答案: 根据问题以及单元格内容生成的回答

==========

以下是一个示例

【示例表格】
{example_table}

【示例输出】
```json
{{
"表名":"胸痛的病因",
"列":"有提示意义的临床表现",
"行":["急性心肌梗死（心血管）","胸主动脉夹层（心血管）","食管破裂"],
"问题":"急性心肌梗死的临床表现包括突发的压榨样胸痛向下颌或上肢放射，劳力性胸痛在休息后缓解，以及常有红色信号表现等。胸主动脉夹层则表现为突发的撕裂样胸痛向后背放射，可能伴有晕厥、卒中或下肢缺血，四肢脉搏或血压不相同，且常见于年龄>55岁和高血压患者。食管破裂的典型表现包括呕吐或器械检查后突发、严重的胸痛，听诊有皮下捻发音，以及多个红色信号表现。请结合这些信息，分析这三种病因的疼痛性质、伴随症状及常见的体征差异。",
"答案":"急性心肌梗死的疼痛性质为突发的压榨样胸痛，通常向下颌或上肢放射，并伴随劳力性胸痛在休息后缓解，常伴有红色信号表现，如异常生命体征和气短。胸主动脉夹层的疼痛为突发的撕裂样胸痛，向后背放射，常见于年龄>55岁和高血压患者，可能伴有晕厥、卒中或下肢缺血，四肢脉搏或血压不相同。食管破裂的疼痛通常在呕吐或器械检查后突发，表现为严重的胸痛，听诊时可发现皮下捻发音，并伴有多个红色信号表现，如低血压和气急。这些特征可以帮助区分这三种病因。"
}}
```

==========

【表格】
{query_table}

【输出】

Figure C11: The specific prompt for automatic query generation on Table task.

```
{
    "Literature_en": "Given a query for the experts, retrieve relevant scientific
articles.",
    "Literature_zh": "给定一个来自患者的问题，查询相关的医学片段。",
    "Table_en": "Given a query, retrieve the relevant medical table.",
    "Table_zh": "给定一个问题，查询相关的医学诊疗表格。",
    "KB_para_zh": "给定一个问题，查询中药药品说明书中的相关片段。",
    "KB_doc_zh": "给定一个问题，查询相关的中药药品说明书。",
    "AskAPatient_en": "Given a social media phrase, retrieve relevant medical
terminology",
    "SMM4H-17_en": "Given a social media phrase, retrieve relevant medical terminology",
    "TwADR-L_en": "Given a social media phrase, retrieve relevant medical terminology",
    "Disease_dignosis_zh": "给定一个短语，查询标准的疾病诊断术语。",
    "Clinical_examination_zh": "给定一个短语，查询标准的体格检查短语。",
    "Procedure_operation_zh": "给定一个短语，查询标准的手术操作短语。",
    "Symptom_sign_zh": "给定一个短语，查询标准的症状体征短语。",
    "Disease_dignosis_cross": "Given a phrase, retrieve normalized disease diagnosis
term.",
    "Clinical_examination_cross": "Given a phrase, retrieve normalized clinical
examination term.",
    "Procedure_operation_cross": "Given a phrase, retrieve normalized procedure operation
term.",
    "Symptom_sign_cross": "Given a phrase, retrieve normalized symptom sign term.",
    "EHR_query2sql_zh": "给定一个关于医疗电子病历的问题，查询相关的SQL语句。",
    "EHR_sql2para_zh": "给定一个SQL语句，查询相关的医疗电子病历段落。",
    "EHR_query2para_zh": "给定一个问题，查询相关的医疗电子病历段落。",
    "EHR_query2doc_zh": "给定一个问题，查询相关的医疗电子病历。",
    "Dialogue_qnorm_zh": "给定一个来自患者的问题，查询相关的问题。",
    "Dialogue_en": "Given a query for patients, retrieve relevant medical patient's
questions",
    "Dialogue_zh": "给定一个来自患者的问题，查询相关的回答。"
}
```


Figure C12: The specific query prompts for different tasks.

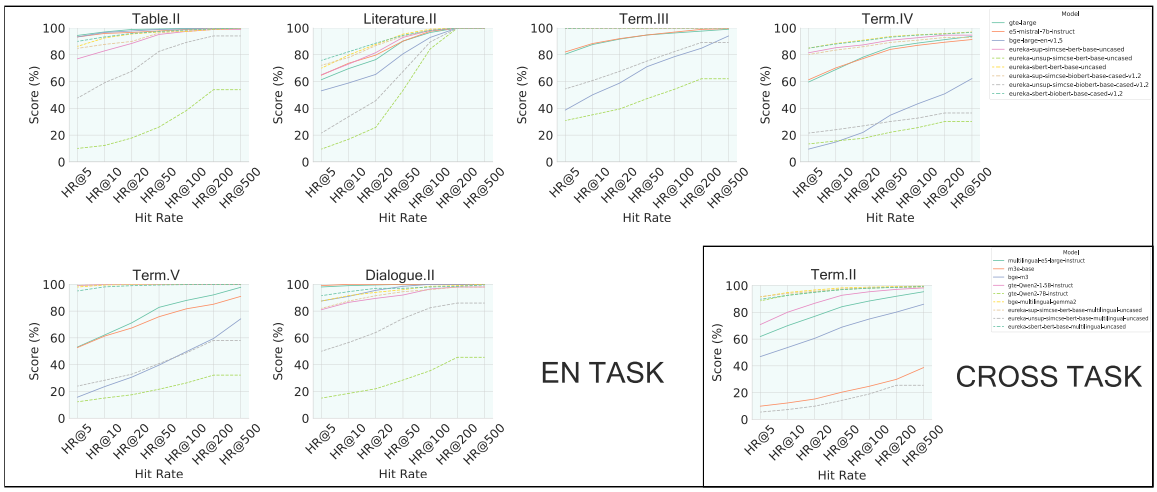

Figure C13: Line chart of performance across different numbers of recalled items in English and Cross-lingual datasets

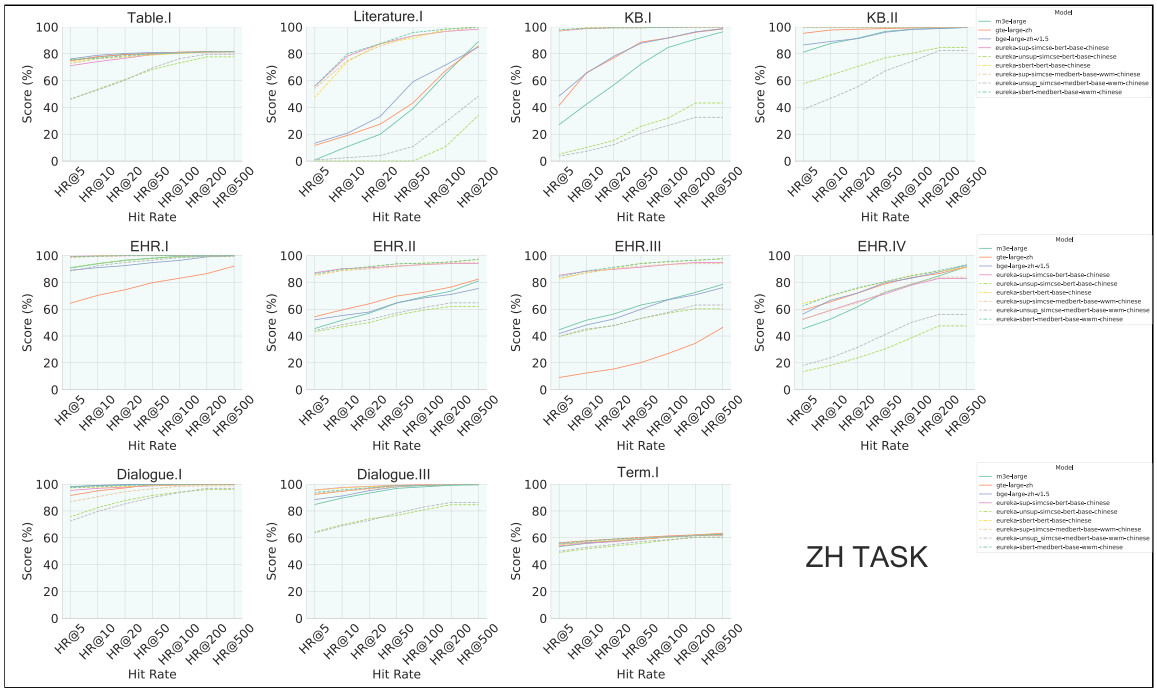

Figure C14: Line chart of performance across different numbers of recalled items in Chinese datasets(Since the corpus length of some datasets is less than 500, the hit rate range for certain datasets has been set from 5 to 200.)