

Finding Diamonds in Conversation Haystacks: A Benchmark for Conversational Data Retrieval

Yohan Lee^{1,2*}, Yongwoo Song^{1,3*}, Sangyeop Kim^{1,4†}

¹Coxwave, ²Kakaobank, ³Kyung Hee University, ⁴Seoul National University
yann.lee@kakaobank.com, syw5141@khu.ac.kr, sy917kim@bdai.snu.ac.kr

Abstract

We present the Conversational Data Retrieval (CDR) benchmark, the first comprehensive test set for evaluating systems that retrieve conversation data for product insights. With 1.6k queries across five analytical tasks and 9.1k conversations, our benchmark provides a reliable standard for measuring conversational data retrieval performance. Our evaluation of 16 popular embedding models shows that even the best models reach only around NDCG@10 of 0.51, revealing a substantial gap between document and conversational data retrieval capabilities. Our work identifies unique challenges in conversational data retrieval (implicit state recognition, turn dynamics, contextual references) while providing practical query templates and detailed error analysis across different task categories. The benchmark dataset and code are available at <https://github.com/1-yohai/CDR-Benchmark>.

1 Introduction

The widespread adoption of generative AI powered by Large Language Models (LLMs) has created vast repositories of conversation data (Liang et al., 2025). These dialogues offer valuable insights into user behaviors and system performance. However, effectively analyzing and leveraging this accumulated conversational data remains an underexplored challenge in the field (Zhang et al., 2024c; Suri et al., 2024; Eloundou et al., 2025).

Unlike conventional information systems, large language model-based systems operate through open-ended interactions without predefined specifications (Zhang et al., 2024a; Liu et al., 2024). Users interact with them in diverse, unpredictable ways, creating unique challenges for conversational data analytics. Traditional approaches to extracting product insights struggle with these datasets: supervised learning techniques face prohibitive labeling

costs (Jafarlou and Kubek, 2024), manual dialogue review becomes impractical at the scale of millions of conversations (Bavaresco et al., 2020; Følstad and Taylor, 2021), and conventional metrics fail to capture the complex evolution of user satisfaction across multiple turns (Park et al., 2015).

Given these challenges, many product development teams have adopted an exploratory “Retrieve and Analyze” workflow to derive insights from their conversation data (see Appendix A for a detailed real-world case study). In this approach, retrieval quality fundamentally determines analysis effectiveness—if relevant conversations cannot be efficiently found, critical insights remain hidden despite analyst expertise. For example, when investigating satisfaction issues, product managers often use basic keyword searches like “unhappy” or “disappointed,” missing cases where dissatisfaction is expressed implicitly or across multiple turns. This retrieval gap creates significant blind spots in understanding user experiences and severely limits the value extracted from conversation datasets.

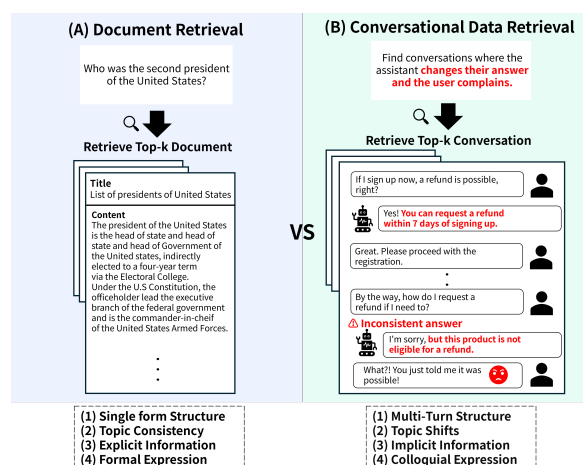


Figure 1: Comparison between traditional document retrieval and conversational data retrieval.

To address these limitations, we build upon the concept of Conversational Data Retrieval (CDR)

*This work was conducted at Coxwave.

†Corresponding author.

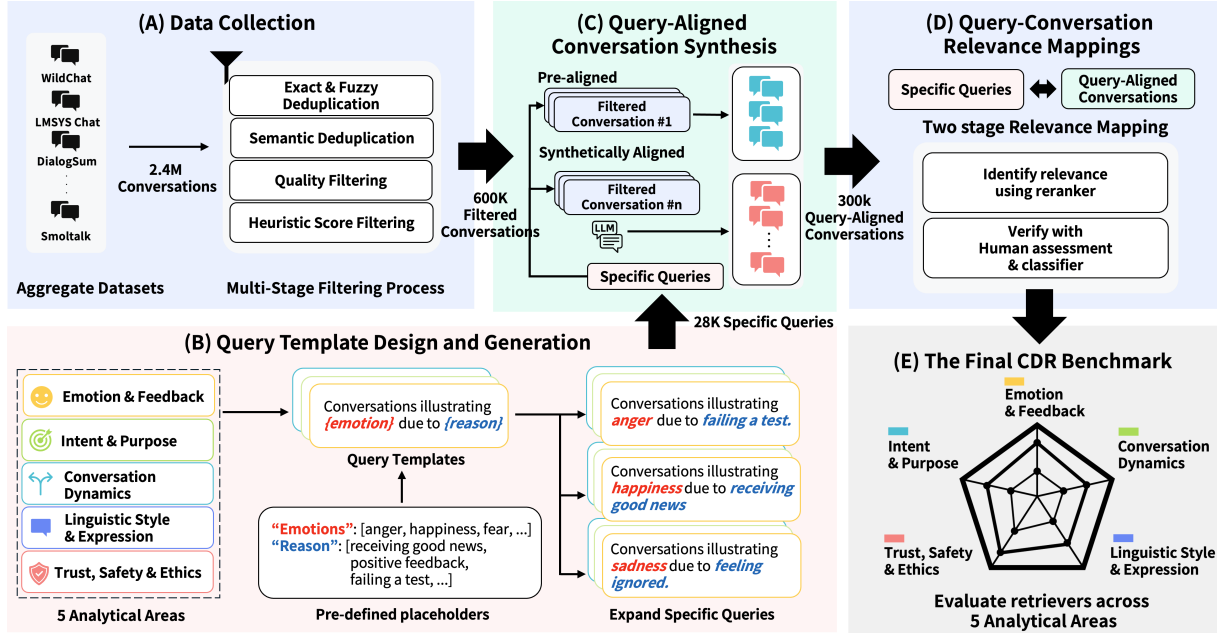


Figure 2: An overview of the Conversational Data Retrieval (CDR) benchmark construction pipeline. (A) Collect and filter large-scale conversational data. (B) Generate query templates across five key areas. (C) Synthesize query-aligned conversations with LLMs. (D) Map relevance through reranking, human assessment, and classifier verification. (E) Integrate the processed data into a standardized CDR evaluation benchmark.

(Kim et al., 2025): the task of retrieving relevant conversations from large chat histories based on queries targeting conversation-specific content and context. As illustrated in Figure 1, CDR differs fundamentally from traditional document retrieval (Meng et al., 2024; Zhang et al., 2025) by addressing conversation-specific challenges: multi-turn exchanges, implicit meanings, and topic shifts.

Beyond analytical use cases, effective CDR enables applications such as AI memory systems (OpenAI, 2024b; Pan et al., 2025), and retrieval-augmented generation (Lewis et al., 2020; Wang et al., 2024). However, current retrieval solutions were not designed with conversations in mind, limiting the potential of these applications.

Despite its value, CDR remains underexplored in research. This gap stems from several factors: the proprietary nature of industrial conversation datasets, privacy concerns limiting public data availability, and the lack of standardized evaluation metrics (Qu et al., 2018; Dalton et al., 2020b). These challenges have hindered research progress on methods specifically designed for CDR.

To address this gap, we introduce a comprehensive benchmark for CDR. Figure 2 illustrates our construction process, including data collection, query design, and validation. Our contributions include: (1) the first benchmark specifically targeting

CDR, comprising 1.6k queries and 9.1k conversations; (2) evaluation of 16 commonly used embedding models revealing performance disparities between document and conversation retrieval; (3) a taxonomy of five essential analytical tasks exposing unique challenges in CDR; and (4) practical query templates developed with domain experts applicable to product improvement workflows. This benchmark facilitates structured development of conversation retrieval models, supporting applications and research in conversation analysis.

2 Related Works

Information retrieval (IR) has evolved from lexical matching (Robertson and Zaragoza, 2009) to neural approaches (Karpukhin et al., 2020). Recent advances in generative AI have linked retrieval with conversation, enabling conversational search (Dalton et al., 2020a; Mao et al., 2023), agent memory systems (Packer et al., 2023; Zhong et al., 2024), and retrieval-based reasoning (Joshi et al., 2024; OpenAI, 2025a). However, these efforts focus on using retrieval to enhance conversations (Qu et al., 2020; Mo et al., 2023), not on effectively retrieving conversational data itself.

The unique value of conversational data lies in its multifaceted nature. Human-Computer Interaction research has identified several dimensions criti-

Analytical Area	Description	Product Insights
Emotion & Feedback	Identifying users’ emotional states and feedback in conversations	Revealing satisfaction patterns and pain points for product improvement
Intent & Purpose	Recognizing user intentions and goals	Evaluating alignment between intended and actual AI system usage
Conversation Dynamics	Analyzing conversation flow, turn structure and resolution patterns	Identifying conversation bottlenecks and improving dialogue completion rates
Trust, Safety & Ethics	Exploring trust-building and ethical issues in conversations	Identifying system reliability concerns and potential safety risks
Linguistic Style & Expression	Analyzing language patterns and comprehension challenges	Helping calibrate system language to user comprehension levels

Table 1: Five core analytical areas identified for the CDR benchmark with their product insights.

cal for understanding these interactions—including user intentions, emotional responses, conversation flow patterns, and trust development (Bavaresco et al., 2020; Følstad and Taylor, 2021). These elements often span multiple turns and contain implicit signals that traditional document retrieval approaches struggle to capture (Wu et al., 2025; Xue et al., 2025). Effective analysis requires methods to identify these complex patterns within conversations.

While industrial applications generate vast conversational data, privacy concerns and proprietary issues severely limit public access to these datasets (Gumusel, 2025). Even available datasets often lack sufficient coverage of specific analytical dimensions needed for comprehensive evaluation (Zhao et al., 2024; Allal et al., 2025).

Synthetic conversational data offers a valuable solution to these constraints, as high-quality synthetic dialogues can match or exceed the performance of systems trained on real data (Dai et al., 2022; Bonifacio et al., 2022; Wu et al., 2024). This approach enables more controlled evaluation by systematically varying conversation attributes while maintaining natural dialogue properties.

However, existing IR benchmarks (Thakur et al., 2021; Muennighoff et al., 2023) focus on documents, while dialogue datasets typically focus on generation tasks. This creates a significant gap between the analytical needs identified in HCI research and available evaluation frameworks. Our CDR benchmark addresses this gap by integrating multi-dimensional aspects of conversations with a comprehensive evaluation framework.

3 Designing the CDR Benchmark

Data Collection and Industrial Requirements

To establish a foundation for the CDR benchmark,

we collected conversational data from 11 diverse open-source dialogue datasets including LMSYS Chat (Chiang et al., 2024), WildChat (Zhao et al., 2024), and DialogSum (Chen et al., 2021). To ensure quality and remove duplicates, we applied filtering using the NeMo Curator framework (Jennings et al.), refining approximately 2.4 million conversations to 600k high-quality dialogue instances. The complete data sources and filtering method are detailed in Appendix B.

To ensure industrial applicability, we sampled 1k conversations for analysis and gathered input from 20 experts in generative AI product development. From this combined research, we identified key information needs when examining conversational data and determined five core areas for product improvement, shown in Table 1. These areas reflect how conversational data differs from traditional document retrieval challenges.

Query Template Design and Generation From the five core areas in Table 1, we created 130 query templates that capture the specific characteristics of each category. Each template included placeholder elements to cover diverse conversational scenarios.

For example, a template in the Emotion & Feedback category might be: “Find conversations where users express {emotion} after {system_action}” For the {emotion} placeholder, values included “frustration,” “disappointment,” and “satisfaction.”

We defined approximately 510 placeholder values across different categories. By combining these placeholders with our templates, we generated a total of 28k specific queries. Full details of templates and placeholders are provided in Appendix C.

Query-Aligned Conversation Synthesis Method

Finding conversations that match our diverse queries presented two challenges: our corpus could

not cover all specific query scenarios needed, and manually labeling thousands of conversations would be prohibitively time-consuming.

To address these limitations, we first retrieved top-5 candidate conversations for each query using the embedding model (OpenAI, 2024c). Twenty expert annotators with industry experience in conversational AI product development then manually reviewed these candidates. They classified each as related or unrelated based on whether it faithfully reflected the query intent. When no suitable match existed, we used reasoning-capable language models—Claude-3.7 (Anthropic, 2025), o1 (OpenAI, 2024a), and o3-mini (OpenAI, 2025b)—to create synthetically aligned conversations by adapting existing conversations from our corpus. These LLM-generated conversations were also validated by expert annotators to ensure both query fidelity and conversational naturalness.

Our conversation generation prompt (detailed in Appendix D.1) instructed models to maintain each conversation structure and characteristics while incorporating elements needed for query alignment. This approach preserved the natural variation found in real conversations while ensuring examples contained features necessary for evaluation.

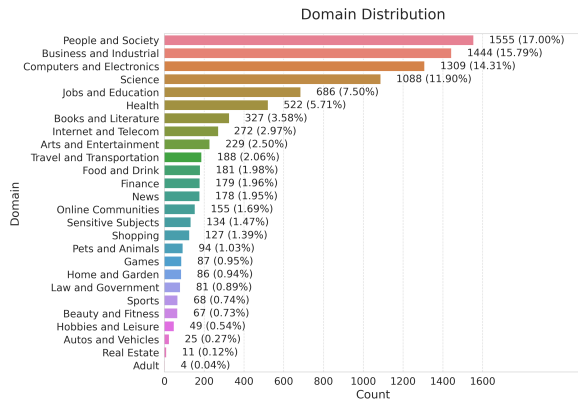


Figure 3: Domain distribution in the CDR benchmark dataset, showing diverse coverage across categories.

By combining pre-aligned conversations with synthetically aligned conversations, our method maintained domain diversity. Figure 3 shows balanced coverage across major categories like People and Society (17.00%) and Business and Industrial (15.79%), as classified by a fine-tuned classifier¹.

Expanding Query-Conversation Relevance Mappings

To create a realistic retrieval bench-

¹<https://huggingface.co/nvidia/domain-classifier>

mark, we expanded each query to match multiple relevant conversations through a three-step process. First, we trained a specialized reranker model using 300k conversations from our corpus. We generated positive and negative query examples with LLaMa 3.3 70B (Meta, 2025) using prompts in Appendix D.2 and D.3, and fine-tuned the GTE Reranker (Zhang et al., 2024b) (training detailed in Appendix E). Second, we applied this reranker to identify candidate relevant conversations, selecting pairs with relevance scores above 0.9 and excluding overly general queries matching more than 50 conversations. Third, we validated mappings through comprehensive human assessment. Expert annotators conducted full manual evaluation of approximately 4k query-conversation pairs across 200 queries. We applied binary relevance criteria with majority voting for reliability, conservatively removing non-consensus cases. We trained a ModernBERT-based (Warner et al., 2024) relevance classifier using these manually validated pairs, achieving 95.2% accuracy, as detailed in Appendix F. For remaining queries, we applied the classifier to predict relevance for all pairs, then employed two-stage human verification. First, we prioritized uncertain cases where sigmoid scores fell below 0.9. Second, we identified boundary inconsistencies where irrelevant predictions appeared among relevant pairs, and vice versa. Human annotators verified both uncertain predictions and inconsistent boundaries, ensuring comprehensive coverage while efficiently allocating annotation effort to critical cases.

General Statistics	
Number of conversations	9,146
Number of queries	1,583
Avg. messages per conversation	5.4
Avg. tokens per conversation	464
Avg. relevant convs per query	20.44
Total query-conversation pairs	32,357
Query Task Distribution (%)	
Intent & Purpose	36.1%
Emotion & Feedback	20.1%
Linguistic Style & Expression	15.9%
Trust, Safety & Ethics	14.6%
Conversation Dynamics	13.4%

Table 2: Key statistics of the CDR benchmark dataset.

Benchmark Composition and Characteristics

Our comprehensive mapping pipeline provides an efficient method for constructing high-quality query-conversation pairs. This methodology offers

Model	Turn			Sliding chunk (k=3)			Session			Inference (s)	Ingestion (s)
	NDCG@10	R@10	P@10	NDCG@10	R@10	P@10	NDCG@10	R@10	P@10		
Commercial API Models											
Voyage-3-large	0.5079	0.2609	0.4359	0.5063	0.2582	0.4327	0.5036	0.2615	0.4358	375.03	2620.00
Text-embedding-3-large	0.5078	0.2698	0.4389	0.5130	0.2696	0.4416	0.4876	0.2529	0.4190	245.90	1433.12
Text-embedding-3-small	0.4897	0.2558	0.4183	0.4855	0.2558	0.4171	0.4664	0.2412	0.3972	196.99	1253.68
Embed-english-v3.0	0.4189	0.2237	0.3547	0.2547	0.1351	0.2116	0.3620	0.1923	0.2987	83.75	520.00
Open Source Models											
Stella_en_1.5B_v5	0.4907	0.2592	0.4141	0.4894	0.2528	0.4078	0.4722	0.2481	0.3961	7.84	336.92
Stella_en_400M_v5	0.4682	0.2490	0.3963	0.4651	0.2462	0.3919	0.4583	0.2400	0.3846	5.51	119.24
Jasper_en_vision_language_v1	0.4379	0.2317	0.3712	0.4309	0.2245	0.3615	0.4561	0.2382	0.3814	7.86	355.56
NV-Embed-v2	0.3170	0.2008	0.3251	0.3956	0.1988	0.3262	0.4592	0.2344	0.3855	13.92	279.73
NV-Embed-v1	0.2467	0.1226	0.1956	0.2603	0.1302	0.2080	0.4389	0.2242	0.3634	13.79	280.23
SFR-Embedding-2_R	0.3344	0.1775	0.2805	0.3127	0.1639	0.2589	0.4474	0.2280	0.3722	10.40	213.57
Jina-embeddings-v3	0.3803	0.2053	0.3160	0.3983	0.2142	0.3363	0.3718	0.1995	0.3088	7.01	106.25
Modernbert-embed-base	0.3594	0.1923	0.3026	0.3398	0.1795	0.2857	0.3579	0.1906	0.3016	7.49	45.63
Gte-Qwen2-1.5B-instruct	0.4646	0.2412	0.3952	0.4386	0.2261	0.3708	0.3615	0.1919	0.2987	7.59	336.73
Gte-large-en-v1.5	0.3310	0.1821	0.2792	0.3246	0.1778	0.2726	0.3429	0.1840	0.2860	4.39	133.29
Bge-large-en-v1.5	0.3276	0.1757	0.2719	0.3105	0.1659	0.2539	0.3071	0.1617	0.2476	3.97	94.29
Cde-small-v2	0.1163	0.0606	0.0975	0.1226	0.0640	0.1007	0.0830	0.0463	0.0701	7.50	39.57

Table 3: Performance comparison of 16 embedding models on the CDR benchmark. The table presents retrieval metrics including NDCG@10, Recall@10 (R@10), and Precision@10 (P@10) across three evaluation settings: Turn-based, Sliding chunk (k=3), and Session-based approaches. Additional metrics are in Appendix H.1. Inference and ingestion times were measured using procedures in Appendix G.2 on hardware described in Appendix G.3.

a practical solution for industrial deployment where cost-effective data mapping is essential. However, as our goal is to establish a rigorous benchmark, we conducted additional validation to ensure maximum integrity. We employed four LLMs—GPT-4o (OpenAI et al., 2024), o3-mini (OpenAI, 2025b), Claude 3.7 Sonnet (Anthropic, 2025), and Gemini 2.0 Pro (Deepmind, 2025)—with the prompt in Appendix D.4 to cross-check all pairs. Cases where LLMs disagreed were flagged for expert review by annotators, who applied consistent binary relevance criteria with majority voting. Pairs without clear consensus were conservatively discarded. Through this multi-stage validation approach combining LLM scalability with human verification at each step, 97% of all query-conversation mappings passed assessment, ensuring the final benchmark meets the highest quality standards.

The final CDR benchmark consists of 1,583 queries and 9,146 conversations (Table 2). Conversations average 5.4 messages and 464 tokens². Each query maps to 20.44 relevant conversations on average. Query distribution spans five core areas: Intent & Purpose (36.1%), Emotion & Feedback (20.1%), and three other categories, with detailed examples provided in Appendix J.

²Based on the GPT-4o tokenizer.

4 Experiments and Analysis

4.1 Experimental Setup

We evaluated 16 widely used embedding models from open-source communities and commercial providers including OpenAI (OpenAI, 2024c), Cohere (Reimers et al., 2023), and Voyage AI (VoyageAI, 2025). Performance was assessed using NDCG@10, Recall@10, and Precision@10 at three retrieval granularities—session-level, turn-level, and sliding window (chunk size = 3). For detailed evaluation methodology, see Appendix G.1.

4.2 Results and Analysis

Table 3 summarizes model performance on our CDR benchmark. Among commercial API models, Voyage-3-large (VoyageAI, 2025) achieved the highest performance in both turn-based (NDCG@10: 0.5079) and session-based (NDCG@10: 0.5036) evaluation, while Text-embedding-3-large (OpenAI, 2024c) led in sliding chunk settings (NDCG@10: 0.5130). Among open-source models, Stella_en_1.5B_v5 (Zhang et al., 2025) demonstrated consistently high performance across all evaluation settings. Interestingly, some models showed significant performance variations across different evaluation settings. For instance, NV-Embed-v2 performed poorly in

Challenge Type	Query Example	Incorrectly Retrieved Results	Why Models Fail
Role Recognition Failure	Assistant shares parenting and childcare advice	user: Welcome to the parent teacher conference. So what is your child's name? assistant: Megan Jones. user: She's been having some problems with the other kids in your class.	Models match "parent," "child," "teacher" keywords but miss conversational roles. Assistant is receiving information as parent, not providing advice.
Dynamic Progression Failure	Conversation where user feels increasingly satisfied with assistant	user: You've been so helpful with all my questions lately. I just wanted to tell you how happy I am with your assistance. assistant: Thank you so much for your kind words! It truly means a lot to me.	Models match final satisfaction but miss progressive "increasingly" aspect. This shows static state, not gradual improvement.
Semantic Contextual Misinterpretation	Assistant provides real estate and housing information	user: I'm visiting friends in Nairobi. What's the weather like? assistant: 103°F, 2% chance of rain. user: I need a house for 1 with laundry service. assistant: Found a house at Chiromo Road with 4.6 rating.	Models match "house" keyword but miss context. This is travel booking service, not real estate information provision.

Table 4: Representative failure cases illustrating major retrieval challenges in conversation understanding tasks.

turn-based evaluation (NDCG@10: 0.3170) but achieved substantially improved performance in session-based evaluation (NDCG@10: 0.4592). Even top-performing models scored just above 0.5 in NDCG@10, highlighting the challenges of modeling conversational structure, context transitions, and implicit references.

how conversations develop and flow.

No model excels across all categories - even top-performing Voyage-3-large shows varied results. This suggests no dominant approach exists yet for CDR. Optimal architectures remain unexplored, particularly for conversation structure understanding, which is crucial for practical applications.

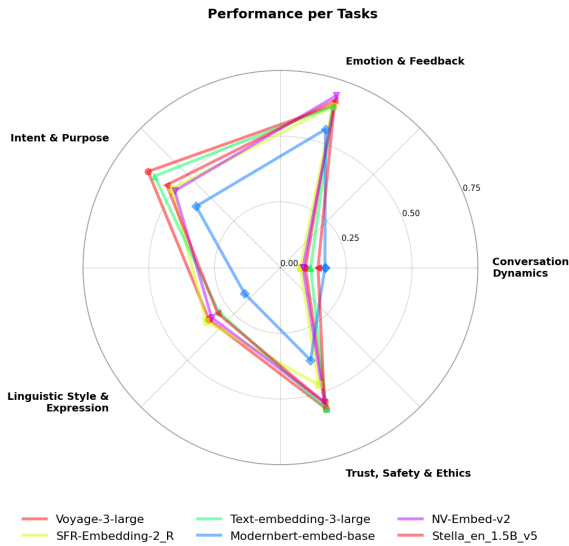


Figure 4: Task-specific NDCG@10 performance comparison of top-performing embedding models and category winners. All results are available in Appendix H.2.

4.3 Performance Across Task Categories

Figure 4 reveals performance variations across task types. All models score highest in ‘Emotion & Feedback’ and ‘Intent & Purpose’, but perform poorly in ‘Conversation Dynamics’ where even the best models score below 0.17. This suggests current models are good at understanding content and explicit statements but struggle with understanding

4.4 Analysis of Retrieval Failures

We identified three consistent failure patterns in current embedding approaches to CDR. Table 4 illustrates these challenges: *Role Recognition Failure*, *Dynamic Progression Failure*, and *Semantic Contextual Misinterpretation*. These failures occur across both turn-based and session-based analysis, revealing models’ inability to capture conversation dynamics at multiple levels. The consistent challenge is that models miss implicit meanings that emerge from conversational context—patterns where the actual roles, progressive changes, or situational context must be inferred from dialogue flow rather than explicit keyword matching.

These errors stem from a fundamental limitation: current models process conversations as collections of words and topics similar to documents, rather than as dynamic exchanges with temporal flow and implicit state changes. Standard embeddings capture vocabulary similarities but miss the contextual evolution and interactive nature of dialogue. This explains the poor performance in *Conversation Dynamics* across all models and signals that effective retrieval systems must be redesigned to capture the unique properties of human dialogue like turn-taking patterns and implicit state transitions.

5 Conclusion

Conversational Data Retrieval (CDR) benchmark establishes the first comprehensive framework for evaluating retrieval systems on conversation data. Experimental results show that even the highest-performing models have not reached satisfactory performance. Our benchmark exposes fundamental challenges unique to conversational data: understanding implicit states, tracking conversation flow, and interpreting contextual references. Our work provides standardized evaluation methodology and query templates for product improvement while establishing a foundation for conversation-specific retrieval techniques that better capture the multi-dimensional nature of human-AI interactions.

Limitations

Our benchmark is limited to English text-based conversations, which may constrain evaluation in multilingual or multimodal settings. This focus, while enabling controlled evaluation, could limit the broader applicability of our findings to diverse linguistic contexts and interaction modalities in global conversational AI applications.

Our benchmark evaluates embedding-based retrieval models, reflecting their widespread adoption in conversational memory systems where turn, session, and segment-level granularities are commonly employed. However, the lack of specialized retrieval models designed specifically for conversation represents a gap in the field that our benchmark could help address through future development of conversation-tailored representation architectures.

While our benchmark provides comprehensive evaluation of retrieval models with robust data quality validation through domain expert involvement, it does not extend to empirical studies of industrial problem-solving applications. Although our motivation stems from real-world challenges and our benchmark identifies optimal approaches under current conditions, further research is needed to validate the practical value of these findings in actual deployment scenarios and their impact on end-user satisfaction in conversational AI systems.

Ethical Considerations

In the development and application of the CDR benchmark, we carefully considered various ethical aspects. Since conversational data inherently contains user interactions and diverse linguistic expressions, we prioritized privacy protection throughout

the data collection and processing stages. We utilized only publicly available open-source datasets and included specific guidelines in our conversation generation prompts to address any potentially remaining personal expressions or sensitive information (see Appendix I). These guidelines included instructions to “appropriately redact or anonymize any PII in reference conversations,” “avoid generating conversations that could be misleading, harmful, or promote unethical behavior,” and “ensure that no personally identifiable information such as names, addresses, phone numbers, financial details, social security numbers, or other sensitive data is exposed or inferred.”

To ensure diversity and balance in the conversational data, we designed the benchmark dataset to encompass a wide range of domains without bias toward specific topics or areas. As shown in Figure 3, we included a balanced representation of conversations from broad domains such as society, business, technology, and science, thereby minimizing bias toward particular areas. Additionally, we explicitly incorporated a ‘Trust, Safety & Ethics’ category in the benchmark’s task areas to establish ethical conversational retrieval capabilities as an important evaluation criterion.

While conversational data retrieval technology can contribute to positive purposes such as improving service quality and user experience, it also carries potential risks of privacy infringement or misuse as inappropriate surveillance tools. We recognize this duality and hope that the CDR benchmark will serve as a tool to promote balance between ethical values and innovative technological advancement. Through this, we believe that the development of conversational AI systems can progress in a direction that respects users’ rights and dignity.

Acknowledgments

This work was supported by Coxwave, Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea) & Gwangju Metropolitan City, Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program(KAIST)) and the Korea government(MSIT) (No. RS-2024-00509279, Global AI Frontier Lab).

References

- Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgrén, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. [Smollm2: When smol goes big – data-centric training of a small language model](#). *Preprint*, arXiv:2502.02737.
- Anthropic. 2025. [Claude 3.7 sonnet and claude code](#).
- Rodrigo Bavaresco, Diógenes Silveira, Eduardo Reis, Jorge Barbosa, Rodrigo Righi, Cristiano Costa, Rodolfo Antunes, Marcio Gomes, Clauter Gatti, Mariangela Vanzin, and 1 others. 2020. Conversational agents in business: A systematic literature review and future research directions. *Computer Science Review*, 36:100239.
- Bitext. 2023. [Customer support llm chatbot training dataset](#). Accessed: 2025-03-20.
- Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. [Inpars: Unsupervised dataset generation for information retrieval](#). In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2387–2392, New York, NY, USA. Association for Computing Machinery.
- Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. [DialogSum: A real-life scenario dialogue summarization dataset](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.
- Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *International conference on machine learning*, pages 4558–4586. PMLR.
- Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020a. Cast 2020: The conversational assistance track overview. In *Text Retrieval Conference*.
- Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020b. [Cast-19: A dataset for conversational information seeking](#). In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1985–1988, New York, NY, USA. Association for Computing Machinery.
- Google Deepmind. 2025. [Gemini 2.0 pro](#).
- Tyna Eloundou, Alex Beutel, David G. Robinson, Keren Gu, Anna-Luisa Brakman, Pamela Mishkin, Meghan Shah, Johannes Heidecke, Lilian Weng, and Adam Tauman Kalai. 2025. [First-person fairness in chatbots](#). In *The Thirteenth International Conference on Learning Representations*.
- Asbjørn Følstad and Cameron Taylor. 2021. Investigating the user experience of customer service chatbot interaction: a framework for qualitative analysis of chatbot dialogues. *Quality and User Experience*, 6(1):6.
- Ece Gumusel. 2025. A literature review of user privacy concerns in conversational chatbots: A social informatics approach: An annual review of information science and technology (arist) paper. *Journal of the Association for Information Science and Technology*, 76(1):121–154.
- Minoo Jafarlou and Mario M. Kubek. 2024. [Reducing labeling costs in sentiment analysis via semi-supervised learning](#). *Preprint*, arXiv:2410.11355.
- Joseph Jennings, Mostofa Patwary, Sandeep Subramanian, Shrimai Prabhume, Ayush Dattagupta, Vibhu Jawa, Jiwei Liu, Ryan Wolf, Sarah Yurick, and Varun Singh. [NeMo-Curator: a toolkit for data curation](#).
- Ashutosh Joshi, Sheikh Muhammad Sarwar, Samarth Varshney, Sreyashi Nag, Shrivats Agrawal, and Juhi Naik. 2024. [Reaper: Reasoning based retrieval planning for complex rag systems](#). In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 4621–4628, New York, NY, USA. Association for Computing Machinery.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Sangyeop Kim, Hangyeul Lee, and Yohan Lee. 2025. Heisir: Hierarchical expansion of inverted semantic indexing for training-free retrieval of conversational data using llms. In *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics.
- Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. [Nv-embed: Improved techniques for training llms as generalist embedding models](#). *Preprint*, arXiv:2405.17428.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020.

- Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA. Curran Associates Inc.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. [DailyDialog: A manually labelled multi-turn dialogue dataset](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. [Towards general text embeddings with multi-stage contrastive learning](#). *Preprint*, arXiv:2308.03281.
- Weixin Liang, Yaohui Zhang, Mihai Codreanu, Jiayu Wang, Hancheng Cao, and James Zou. 2025. [The widespread adoption of large language model-assisted writing across society](#). *Preprint*, arXiv:2502.09747.
- Shuo Liu, Kaining Ying, Hao Zhang, Yue Yang, Yuqi Lin, Tianle Zhang, Chuanhao Li, Yu Qiao, Ping Luo, Wenqi Shao, and Kaipeng Zhang. 2024. [Convbench: A multi-turn conversation evaluation benchmark with hierarchical ablation capability for large vision-language models](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Srijata Maji, Moghis Fereidouni, Vinaik Chhetri, Umar Farooq, and A. B. Siddique. 2024. [Mobileconvrec: A conversational dataset for mobile apps recommendations](#). *Preprint*, arXiv:2405.17740.
- Kelong Mao, Zhicheng Dou, Haonan Chen, Fengran Mo, and Hongjin Qian. 2023. Large language models know your contextual search intent: A prompting framework for conversational search. pages 1211–1225.
- Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. [Sfr-embedding-mistral:enhance text retrieval with transfer learning](#). Salesforce AI Research Blog.
- Meta. 2025. [Llama 3.3 model cards & prompt formats](#).
- Fengran Mo, Jian-Yun Nie, Kaiyu Huang, Kelong Mao, Yutao Zhu, Peng Li, and Yang Liu. 2023. [Learning to relate to previous turns in conversational search](#). In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '23*, page 1722–1732, New York, NY, USA. Association for Computing Machinery.
- Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. [OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.
- John Xavier Morris and Alexander M Rush. 2025. [Contextual document embeddings](#). In *The Thirteenth International Conference on Learning Representations*.
- 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, Dubrovnik, Croatia. Association for Computational Linguistics.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- OpenAI. 2024a. [Introducing openai o1](#).
- OpenAI. 2024b. [Memory and new controls for chatgpt](#).
- OpenAI. 2024c. [New embedding models and api updates](#).
- OpenAI. 2025a. [Introducing deep research](#).
- OpenAI. 2025b. [Openai o3-mini](#).
- Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph Gonzalez. 2023. [Memgpt: Towards llms as operating systems](#). In *arXiv.org*.
- Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Jianfeng Gao. 2025. [Secom: On memory construction and retrieval for personalized conversational agents](#). In *The Thirteenth International Conference on Learning Representations*.
- Kunwoo Park, Jaewoo Kim, Jaram Park, Meeyoung Cha, Jiin Nam, Seunghyun Yoon, and Eunhee Rhim. 2015. [Mining the minds of customers from online chat logs](#). In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15*, page 1879–1882, New York, NY, USA. Association for Computing Machinery.
- Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. [Analyzing and characterizing user intent in information-seeking conversations](#). In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18*, page 989–992, New York, NY, USA. Association for Computing Machinery.

- Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.
- Nils Reimers, Elliott Choi, Alekhya Nandula Amr Kayid, Manoj Govindassamy, and Abdullah Elkady. 2023. [Introducing embed v3](#).
- S. Robertson and H. Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.
- Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. [jina-embeddings-v3: Multilingual embeddings with task lora](#). *Preprint*, arXiv:2409.10173.
- Siddharth Suri, Scott Counts, Leijie Wang, Chacha Chen, Mengting Wan, Tara Safavi, Jennifer Neville, Chirag Shah, Ryen W. White, Reid Andersen, Georg Buscher, Sathish Manivannan, Nagu Rangan, and Longqi Yang. 2024. [The use of generative search engines for knowledge work and complex tasks](#). Available online.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. [BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models](#). In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- VoyageAI. 2025. [voyage-3-large: the new state-of-the-art general-purpose embedding model](#).
- Xi Wang, Procheta Sen, Ruizhe Li, and Emine Yilmaz. 2024. [Adaptive retrieval-augmented generation for conversational systems](#). *Preprint*, arXiv:2407.21712.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). *Preprint*, arXiv:2412.13663.
- Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. [Longmemeval: Benchmarking chat assistants on long-term interactive memory](#). In *The Thirteenth International Conference on Learning Representations*.
- Fanyou Wu, Weijie Xu, Chandan Reddy, and Srinivasan Sengamedu. 2024. [Synthesizing conversations from unlabeled documents using automatic response segmentation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8012–8026, Bangkok, Thailand. Association for Computational Linguistics.
- Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. [C-pack: Packaged resources to advance general chinese embedding](#). *Preprint*, arXiv:2309.07597.
- Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. [Bot-adversarial dialogue for safe conversational agents](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968, Online. Association for Computational Linguistics.
- Haochen Xue, Feilong Tang, Ming Hu, Yexin Liu, Qidong Huang, Yulong Li, Chengzhi Liu, Zhongxing Xu, Chong Zhang, Chun-Mei Feng, and 1 others. 2025. [Mmrc: A large-scale benchmark for understanding multimodal large language model in real-world conversation](#). *arXiv preprint arXiv:2502.11903*.
- Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. [MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking base-lines](#). In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.
- Chen Zhang, Luis Fernando D’Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 2024a. [A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators](#). In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’24/IAAI’24/EAAI’24*. AAAI Press.
- Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2025. [Jasper and stella: distillation of sota embedding models](#). *Preprint*, arXiv:2412.19048.
- Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024b. [mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.
- Xinghua Zhang, Haiyang Yu, Yongbin Li, Minzheng Wang, Longze Chen, and Fei Huang. 2024c. [The imperative of conversation analysis in the era of llms: A survey of tasks, techniques, and trends](#). *Preprint*, arXiv:2409.14195.
- Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. [Wildchat: 1m chatGPT interaction logs in the wild](#). In *The Twelfth*

International Conference on Learning Representations.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. [LMSYS-chat-1m: A large-scale real-world LLM conversation dataset](#). In *The Twelfth International Conference on Learning Representations*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. [Memorybank: Enhancing large language models with long-term memory](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19724–19731.

A Industry Case Study of the “Retrieve and Analyze” Approach

The “Retrieve and Analyze” methodology mirrors how analysts and product managers naturally approach problem-solving in real business environments. When faced with user feedback or product issues, human analysts typically form initial hypotheses, gather relevant examples, analyze patterns, and progressively refine their understanding through iterative investigation. What has changed with large-scale conversation data is not this fundamental analytical process, but rather the need for computational assistance to efficiently navigate thousands or millions of conversations.

The following case study illustrates how this human-centered analytical approach, supported by conversation retrieval capabilities, works in practice. This example is adapted from an actual business scenario at a health and fitness application company, demonstrating both the natural analytical workflow and the positive impact of effective conversational data retrieval.

A customer experience team was investigating increased user dissatisfaction following a recent update. Traditionally, they relied on user ratings and manual reviews of customer complaints. However, after implementing chatbot support, these methods became inadequate - the chatbot interface lacked rating systems, and the chat volume overwhelmed manual inspection capabilities.

To address this challenge, the team implemented a conversation retrieval system, beginning with a broad query: “Find sessions where users express dissatisfaction.” Sample analysis revealed mentions of the points reward system, prompting them to deep dive with a more targeted search: “Find conversations where users express dissatisfaction with changes to the points reward system” to determine if this was a widespread issue rather than isolated incidents.

This refined approach confirmed their hypothesis, revealing reduced point accumulation rates as the primary driver of dissatisfaction, with users consistently comparing the new system unfavorably to the previous one. Through this methodical process of hypothesis formation and targeted validation, the team efficiently pinpointed the specific issue causing user frustration—a discovery that would have consumed significantly more time and resources using traditional review methods.

Dataset	Data Size	Key Features
Real-world Conversation Data		
LMSYS-Chat-1M (Zheng et al., 2024)	1,000,000+	Real-world user-LLM chats; multi-turn; multilingual; moderation tags and PII redacted
WildChat-1M (Zhao et al., 2024)	1,000,000+	User-ChatGPT logs; multilingual; includes user meta-data and toxicity labels
DialogSum (Chen et al., 2021)	12,000+	Real-world conversations; paired with abstractive summaries and topic annotations
DailyDialog (Li et al., 2017)	10,000+	Open-domain daily conversations; annotated with dialogue acts and emotions
MultiWOZ 2.2 (Zang et al., 2020)	8,000+	Multi-domain, task-oriented dialogues; annotated with states and system actions; corrected labels
Bot-Adversarial Dialogue (BAD) (Xu et al., 2021)	5,000+	Adversarial conversations to test chatbot safety; includes persona settings and safety labels
MobileConvRec (Maji et al., 2024)	8,000+	Conversations for mobile app recommendation; multi-turn; includes user feedback and app info
OpenDialKG (Moon et al., 2019)	12,000+	Knowledge-grounded conversations; each turn linked to KG entities for explainability
Synthetic Conversational Data		
SmolTalk (Allal et al., 2025)	1,000,000+	Synthetic dialogues for instruction following; wide coverage (QA, summarization, coding tasks)
Bitext Customer Support (Bitext, 2023)	26,000+	Synthetic QA pairs created by linguists; customer support domain; slot annotations
Schema-Guided Dialogue (SGD) (Rastogi et al., 2020)	16,000+	Multi-domain task-oriented dialogues; annotated with intents, slots, states; includes zero-shot domains

Table 5: Summary of public conversational datasets.

B Dataset Sources and Filtering Method

B.1 Dataset Sources

We constructed an initial dataset comprising around 2.4 million conversations by aggregating 11 diverse open-source datasets. To ensure broad coverage of dialogue scenarios, our dataset includes both real-world and synthetic conversational data. An overview of the datasets is provided in Table 5.

B.2 Filtering Method

For data quality management, we employed a multi-stage filtering process using the NeMo Curator framework. We first applied exact and fuzzy deduplication to remove identical or near-identical conversations. Next, we conducted semantic deduplication by utilizing a model fine-tuned for semantic search³, effectively filtering out semantically redundant instances. For quality filtering, we employed a model fine-tuned for conversation quality assessment⁴, retaining only conversations labeled as high-

quality among high, middle, and low categories. Finally, heuristic score filtering was applied to remove data with excessive punctuation, URLs, and repeated lines, paragraphs, or n -grams. Through this comprehensive filtering process, we obtained approximately 600,000 refined conversation data points from an initial set of 2.4 million dialogues.

C Template and Placeholders

To ensure the CDR benchmark captures realistic analytical scenarios, we collaborated with domain experts in conversational AI product development to design comprehensive query templates and placeholder values. This expert-guided approach ensures that our benchmark reflects actual information needs encountered in industrial applications. All templates and placeholder values are available in our public repository.

C.1 Query Template

We provide representative examples of query templates used in the benchmark generation across five task categories (*Emotion & Feedback, Intent & Purpose, Conversation Dynamics, Trust, Safety*

³<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

⁴<https://huggingface.co/nvidia/quality-classifier-deberta>

Task Category	Query Template Examples
Emotion & Feedback	• Conversation ending in {emotion}. • Conversation reflecting {emotion} in {user_role}'s feedback. • Conversation where {user_role} finds {assistant_role}'s answers inadequate, leading to {emotion}.
Intent & Purpose	• Conversation where {user_role} seeks advice on {intent}. • Conversation expressing gratitude to {assistant_role} for resolving {intent}. • Conversation where {user_role} repeatedly asks for clarification on {intent}.
Conversation Dynamics	• Conversation where {user_role} becomes more frustrated over time. • Conversation where {user_role} shifts from skepticism to trust. • Conversation where {user_role} feels a sense of accomplishment after resolving an issue with {assistant_role}.
Trust, Safety & Ethics	• Conversation expressing distrust to {assistant_role}. • Conversation where {user_role} raises a {concern_type} issue. • Conversation where {user_role} questions the ethics of {assistant_role}.
Linguistic Style & Expression	• Conversation using {linguistic_style} in {user_role}'s questions. • Conversation highlighting {expression_type} in {assistant_role}'s feedback. • Conversation where {assistant_role} uses {linguistic_style} to simplify concepts.

Table 6: Examples of Query Templates.

& Ethics, Linguistic Style & Expression). The examples of these query templates are presented in Table 6.

C.2 Placeholders and Example Values

We illustrate examples of placeholder values utilized within query templates in Table 7, showing potential variability across queries generated for the benchmark.

D Prompts

This section provides the detailed prompts used throughout our CDR benchmark development process.

D.1 Conversation Generation Prompt

Figure 5 outlines the prompt for generating synthetic conversations that closely match specific queries, ensuring natural, multi-turn conversations that accurately reflect query intent while maintaining appropriate length and format.

D.2 Query Generation Prompt

Figure 6 shows the prompt used to generate a single synthetic search query from a set of conversations, designed to help LLMs identify key insights and patterns within conversation clusters while focusing on product management perspectives.

D.3 Query Augmentation Prompt

Figure 7 presents the prompt for augmenting the initial query by generating three hard negative examples and one alternative positive formulation, facilitating contrastive learning by creating semantically similar but functionally distinct queries.

D.4 Relevance Classification Prompt

Figure 8 shows the prompt used for assessing query-conversation relevance. This prompt was used by modern LLMs during our verification process to evaluate whether conversations were relevant to specific queries.

E Reranker Training

To effectively map queries to relevant conversations, we trained a specialized reranker model using approximately 300k conversations from our filtered corpus. We used LLaMa 3.3 70B (Meta,

Placeholder	Values
emotion	anger, happiness, fear, sadness, disgust, surprise
reason	receiving good news, achieving a goal, success in a project, positive feedback, unexpected reward, losing an opportunity, failing a test, getting rejected, career setback, missed deadline, miscommunication, argument with a friend, relationship conflict, family issues, betrayal, overwhelming workload, financial stress, health concerns, uncertainty about the future, burnout, feeling ignored, being misunderstood, lack of appreciation, being left out, social anxiety, unexpected kindness, support from a friend, acts of generosity, reunion with a loved one, forgiving someone, public embarrassment, making a mistake, feeling inadequate, past regrets, personal failure, exploring a new hobby, intellectual curiosity, inspiring conversation, learning something new, self-discovery, bad weather, technical difficulties, traffic jam, missed appointment, unexpected delay, change in routine, relocation to a new place, adjusting to a new culture, meeting new people, losing a loved one, receiving criticism, feeling judged, comparison with others, unmet expectations, fear of failure, unexpected surprise, random compliment, winning a competition, realizing personal growth, achieving recognition
linguistic_style	formal, informal, neutral, technical, emotional, direct, indirect, logical, persuasive, descriptive, concise, elaborate, colloquial, humorous, sarcastic, empathetic, diplomatic, instructional, academic, poetic, authoritative, friendly, supportive, motivational, analytical, objective, subjective, casual, metaphorical, rhetorical, minimalist, detailed, straightforward, evocative, apologetic, provocative, encouraging, critical, optimistic, pessimistic
expression_type	descriptive, interrogative, exclamatory, imperative, figurative, humorous, sarcastic, rhetorical, analytical, persuasive, ironic, metaphorical, hyperbolic, understated, concise, elaborate, critical, supportive, enthusiastic, skeptical, neutral, emotional, empathetic, diplomatic, apologetic, provocative, assertive, tentative, cautious, objective, subjective, optimistic, pessimistic, directive, expressive, reflective, affirmative, defensive
concern_type	technical issue, ethical issue, academic concern, personal dilemma, relationship issue, work-related stress, health concern, financial problem, social issue, philosophical question, legal complication, moral dilemma, psychological distress, political concern, environmental issue, cultural conflict, safety concern, privacy issue, existential crisis, career uncertainty, education challenge, family dispute, mental health struggle, identity crisis, communication breakdown, trust issue, decision-making difficulty, peer pressure, unfair treatment, discrimination concern, technology misuse, misinformation problem, data security risk, work-life balance struggle, burnout risk, lack of recognition, fear of failure, fear of rejection, self-doubt, unmet expectations, social anxiety, public speaking fear, future uncertainty, innovation challenge, unresolved conflict, resource limitation, competitiveness pressure, time management struggle, productivity concern
information_type	definition, example, guideline, principle, theory, framework, explanation, best practice, case study, historical background, technical specification, algorithm, code snippet, data analysis, statistical insight, latest trend, research finding, scientific evidence, hypothesis, methodology, comparison, contrast, step-by-step guide, practical tip, troubleshooting guide, expert opinion, prediction, future outlook, risk assessment, ethical consideration, common misconception, application, use case, feasibility study, performance evaluation, benchmarking result, legal implication, policy overview, economic impact, market analysis, psychological insight, philosophical perspective, security risk, data privacy issue, innovation strategy, optimization technique
intent	Definition Query, Factual Query, How-to Query, Comparison Query, Reason and Consequence Query, Current Events Query, Historical Query, New Service Request, Purchase and Order Placement, Reservation and Booking, Account Creation and Management, Subscription and Membership, Payment Processing, Technical Troubleshooting, Account Recovery and Access Issues, Product Usage Guidance, Service Interruption Support, Complaint Handling, Return and Refund Assistance, Post-Purchase Support, Service Modification, Profile Update, Customization Request, Recommendation Request, Miscellaneous, Greeting, Farewell, Agreement or Acceptance, Disagreement or Rejection, Clarification Request, Repetition Request, Miscellaneous, Content Creation, Content Editing, Brainstorming and Idea Generation, Content Organization, Content Analysis, Miscellaneous, Educational Query, Skill Development, Health and Wellness, Miscellaneous, Positive Emotion Towards Chatbot, Negative Emotion Towards Chatbot, Positive Emotion About Personal Situation, Negative Emotion About Personal Situation, Positive Emotion About External Situation, Negative Emotion About External Situation, Miscellaneous, Offensive Language, Prohibited Content, Malicious Behavior, Miscellaneous
issue_description	technical malfunction, algorithmic bias, ethical dilemma, unexpected software bug, unclear instructions, ambiguous response, miscommunication, incomplete explanation, contradictory information, unresolved question, flawed reasoning, lack of supporting evidence, data inconsistency, security vulnerability, privacy violation, inaccurate prediction, unmet expectations, slow response time, unexpected error, outdated information, misleading statement, insufficient context, difficulty in decision-making, lack of transparency, complex jargon, overcomplicated solution, missing critical details, irrelevant response, unconvincing argument, lack of practical application, unrealistic assumption, biased perspective, failure to address concerns, poorly structured explanation, logical fallacy, lack of citation, conflicting sources, failure to meet requirements, unanticipated consequences, incomplete analysis, ineffective troubleshooting, delayed resolution, lack of alternative solutions, misinterpretation of question, failure to adapt to context, insufficient depth, overgeneralization, misaligned priorities, oversimplified reasoning, lack of real-world examples
user_role	user, human
assistant_role	assistant, bot, agent

Table 7: Placeholders and Possible Values.

2025) to generate training data through a two-step process. First, we applied the synthetic query generation prompt (Figure 6) to create one relevant query per conversation that captured the core information needs represented in the dialogue. Then, using the query augmentation prompt (Figure 7), we generated three hard negative queries (semantically similar but intentionally irrelevant) and one additional positive query (different wording but preserving intent) for each conversation.

This approach yielded approximately 1.5 million query-conversation pairs with a 2:3 positive-to-negative ratio. We fine-tuned the GTE-Multilingual-Reranker model (Zhang et al., 2024b) using a binary cross-entropy loss function with hard negatives, a learning rate of $2e-5$ with linear warmup and decay, and maximum sequence length of 8192 tokens to accommodate longer conversations. The model was trained for 3 epochs on a single NVIDIA H100 GPU. The reranker achieved an average precision of 96.22% on our validation set after the final epoch. We applied this model with a threshold score of 0.9 to identify candidate relevant conversations across our corpus for the final benchmark construction.

F Classifier Training

We trained a specialized binary relevance classifier to validate the reliability of the mapped query-conversation relationships. This classifier was designed to distinguish relevant and irrelevant query-conversation pairs in alignment with human judgments. For training, we utilized approximately 3K relevance pairs obtained through human assessment and an additional 20K synthetic relevance pairs generated through the procedure described in E. This resulted in a training dataset of approximately 23K pairs with a balanced distribution of relevant and irrelevant examples.

Fine-tuning was performed on the ModernBERT-base model (Warner et al., 2024) using a learning rate of $2e-5$ with linear warm-up and decay scheduling, batch size of 128, and maximum sequence length of 8192 tokens. The model achieved an average precision of 95.2% on the validation set. We applied this classifier, which was trained on human-verified data, to verify and filter the remaining query-conversation mappings. This ensured that only high-confidence pairs were retained in the benchmark and that relevance standards remained consistent and reliable throughout the process.

G Evaluation Details

G.1 Evaluation Setup

To ensure fair and consistent comparison across all evaluated models, we applied unified evaluation protocols. Each model was tested using its original embedding dimension and maximum sequence length as specified in the official documentation. For prompt-based embedding models, we utilized the prompts without any modifications. All conversational data used in the experiments was preprocessed in a uniform manner, ensuring format consistency across all models and minimizing performance variations arising from preprocessing discrepancies.

We define three evaluation settings that differ in the granularity of the retrieval unit:

Turn-based Evaluation: Each conversation turn is treated as an independent unit. For a given query, the model retrieves the most similar individual turn from the corpus. The conversation containing the retrieved turn is considered the final match.

Sliding Chunk Evaluation (k=3): Conversations are segmented into overlapping chunks of three consecutive turns. Given a query, the model retrieves the most similar chunk from all chunks in the corpus. The conversation containing the retrieved chunk is selected as the final match.

Session-based Evaluation: The entire conversation serves as the retrieval unit. For a given query, the model directly retrieves the most similar conversation session from the corpus.

G.2 Efficiency Evaluation

We also measured practical runtime metrics to evaluate real-world usability:

- **Ingestion Time:** The total time required to embed the entire test corpus of 9,146 conversations. This process includes tokenization, model forwarding, and storing the embeddings in memory.
- **Inference Time:** The combined time required to: (1) embed all 1,583 queries, (2) retrieve the corresponding conversations using these embeddings, and (3) compute the final rankings. This represents the end-to-end query processing time.

All experiments were conducted with a batch size fixed at 4 for both ingestion and inference measurements. The reported times represent the total elapsed time for processing the entire dataset.

G.3 Hardware Specifications

All evaluations were conducted under the same setup, and the hardware specifications are summarized in Table 8.

Component	Specification
CPU	Intel(R) Xeon(R) Platinum 8468
GPU	NVIDIA H100 80GB HBM3
Memory	206GB RAM

Table 8: Hardware specifications used for all experimental evaluations.

H Additional Experimental Results

H.1 Performance by Additional Metrics

To provide a comprehensive evaluation beyond the primary results reported in Table 3, we present an extensive analysis of model performance across multiple evaluation metrics. We evaluate all models under three distinct retrieval configurations: Turn-based, Sliding Chunk ($k=3$), and Session-based approaches. For each configuration, we report performance across five key metrics: Accuracy (ACC), Precision (P), Recall (R), Normalized Discounted Cumulative Gain (NDCG), and Mean Reciprocal Rank (MRR), evaluated at cutoff thresholds of 1, 5, 10, and 20. The detailed results are systematically presented in Tables 9, 10, and 11, respectively. This multi-faceted evaluation framework enables a thorough assessment of model efficacy across varying retrieval granularities and ranking depths, providing deeper insights into the comparative strengths and limitations of each approach.

H.2 Performance per Tasks

Table 12 presents the detailed NDCG@10, Recall@10, and Precision@10 performance of all evaluated embedding models across the five task categories in our benchmark. As illustrated in Figure 4, performance varies significantly between task types, with most models showing strengths in content-oriented categories like ‘Emotion & Feedback’ and ‘Intent & Purpose’ while struggling with interaction-focused categories, particularly ‘Conversation Dynamics’.

The table highlights the lack of a universally dominant approach for conversational data retrieval tasks. Even top-performing models like Voyage-3-large demonstrate inconsistent performance across different categories. Notably, ‘Conversation Dynamics’ remains challenging for all models, with

the highest scores barely reaching 0.17, indicating a substantial opportunity for architectural improvements specifically designed to capture conversation flow and structure.

I Dataset License and Disclaimer

In this work, we utilize multiple publicly available open-source dialogue datasets to construct our initial data pool. The LMSYS-Chat-1M dataset is distributed under a custom LMSYS-Chat-1M License Agreement and is non-redistributable. The WildChat-1M-Full dataset is licensed under ODC-BY 1.0 (Open Data Commons Attribution). The Bibtex Customer Support dataset is released under the CDLA-Sharing 1.0 license. The Schema-Guided Dialogue (SGD) dataset is provided under a CC BY-SA 4.0 license. The DialogSum and DailyDialog datasets are released under CC BY-NC-SA 4.0, while OpenDialKG and Bot-Adversarial Dialogue (BAD) are licensed under CC BY-NC 4.0, with the latter restricted to research use only. The MultiWOZ 2.2 and SmolTalk datasets are licensed under Apache 2.0. The MobileConvRec dataset is released under a CC BY 4.0 license. All datasets were used strictly for research purposes in compliance with their respective licenses. No personally identifiable information (PII) was included, and all data was either anonymized by the original providers or manually verified during the curation process. We do not claim ownership of these datasets and respectfully acknowledge the contributions of their original authors and publishers.

J Query-Conversation Pair Examples by Task Category

We provide representative examples of query-conversation pairs for each of the five task categories in the CDR benchmark: Intent & Purpose, Linguistic Style & Expression, Trust, Safety & Ethics, Emotion & Feedback, and Conversation Dynamics. Examples for each category are shown in Tables 13–17.

Model	Turn															
	ACC@1	ACC@5	ACC@10	ACC@20	P@1	P@5	P@10	P@20	R@1	R@5	R@10	R@20	NDCG@1	NDCG@5	NDCG@10	NDCG@20
Commercial API Models																
Voyage-3-large	0.6816	0.8440	0.8876	0.9210	0.6816	0.5495	0.4359	0.3139	0.0433	0.1727	0.2609	0.3549	0.6816	0.5828	0.5079	0.4530
Text-embedding-3-large	0.6684	0.8446	0.9033	0.9394	0.6684	0.5453	0.4389	0.3137	0.0444	0.1758	0.2698	0.3638	0.6684	0.5763	0.5078	0.4552
Text-embedding-3-small	0.6601	0.8547	0.9059	0.9394	0.6601	0.5325	0.4183	0.3031	0.0433	0.1708	0.2558	0.3508	0.6601	0.5652	0.4897	0.4415
Embed-english-v3.0	0.5799	0.7991	0.8642	0.9135	0.5799	0.4550	0.3547	0.2506	0.0382	0.1503	0.2237	0.3015	0.5799	0.4847	0.4189	0.3760
Open Source Models																
Stella_en_1.5B_v5	0.6759	0.8421	0.8926	0.9324	0.6759	0.5320	0.4141	0.2900	0.0453	0.1776	0.2592	0.3443	0.6759	0.5688	0.4907	0.4365
Stella_en_400M_v5	0.6450	0.8313	0.8863	0.9191	0.6450	0.5051	0.3963	0.2794	0.0424	0.1675	0.2490	0.3338	0.6450	0.5402	0.4682	0.4187
Jasper_en_vision_language_v1	0.6045	0.7922	0.8515	0.8989	0.6045	0.4710	0.3712	0.2643	0.0406	0.1557	0.2317	0.3117	0.6045	0.5042	0.4379	0.3927
NV-Embed-v2	0.0013	0.7410	0.8105	0.8617	0.0013	0.3805	0.3251	0.2310	0.0000	0.1230	0.2008	0.2707	0.0013	0.3258	0.3170	0.2929
NV-Embed-v1	0.4062	0.6008	0.6721	0.7473	0.4062	0.2678	0.1956	0.1338	0.0279	0.0871	0.1226	0.1618	0.4062	0.2999	0.2467	0.2165
SFR-Embedding-2_R	0.4864	0.7012	0.7732	0.8326	0.4864	0.3588	0.2805	0.1980	0.0326	0.1189	0.1775	0.2393	0.4864	0.3881	0.3344	0.3003
Jina-embeddings-v3	0.5502	0.7656	0.8200	0.8585	0.5502	0.4129	0.3160	0.2271	0.0381	0.1408	0.2053	0.2792	0.5502	0.4449	0.3803	0.3457
Modernbert-embed-base	0.5066	0.7505	0.8313	0.8844	0.5066	0.3821	0.3026	0.2185	0.0353	0.1269	0.1923	0.2650	0.5066	0.4125	0.3594	0.3269
Gte-Qwen2-1.5B-instruct	0.6387	0.8225	0.8749	0.9128	0.6387	0.5054	0.3952	0.2798	0.0419	0.1607	0.2412	0.3233	0.6387	0.5380	0.4646	0.4127
Gte-large-en-v1.5	0.4668	0.6961	0.7732	0.8244	0.4668	0.3522	0.2792	0.2007	0.0319	0.1186	0.1821	0.2506	0.4668	0.3788	0.3310	0.3025
Bge-large-en-v1.5	0.4738	0.7202	0.7833	0.8345	0.4738	0.3554	0.2719	0.1932	0.0328	0.1208	0.1757	0.2383	0.4738	0.3836	0.3276	0.2963
Cde-small-v2	0.1718	0.3424	0.4195	0.5205	0.1718	0.1265	0.0975	0.0758	0.0114	0.0403	0.0606	0.0911	0.1718	0.1370	0.1163	0.1106

Table 9: Additional performance metrics for turn-based evaluation setting.

Model	Sliding chunk (k=3)															
	ACC@1	ACC@5	ACC@10	ACC@20	P@1	P@5	P@10	P@20	R@1	R@5	R@10	R@20	NDCG@1	NDCG@5	NDCG@10	NDCG@20
Commercial API Models																
Voyage-3-large	0.6860	0.8440	0.8857	0.9286	0.6860	0.5497	0.4327	0.3167	0.0445	0.1739	0.2582	0.3579	0.6860	0.5844	0.5063	0.4557
Text-embedding-3-large	0.6810	0.8440	0.8977	0.9400	0.6810	0.5568	0.4416	0.3170	0.0442	0.1785	0.2696	0.3666	0.6810	0.5881	0.5130	0.4599
Text-embedding-3-small	0.6747	0.8408	0.8920	0.9343	0.6747	0.5289	0.4171	0.2990	0.0440	0.1669	0.2528	0.3462	0.6747	0.5644	0.4894	0.4381
Embed-english-v3.0	0.3721	0.6090	0.6772	0.7612	0.3721	0.2786	0.2116	0.1492	0.0255	0.0909	0.1351	0.1824	0.3721	0.2998	0.2547	0.2290
Open Source Models																
Stella_en_1.5B_v5	0.6740	0.8446	0.8920	0.9324	0.6740	0.5301	0.4078	0.2872	0.0442	0.1739	0.2558	0.3396	0.6740	0.5663	0.4855	0.4320
Stella_en_400M_v5	0.6450	0.8219	0.8730	0.9135	0.6450	0.5039	0.3919	0.2732	0.0425	0.1671	0.2462	0.3268	0.6450	0.5393	0.4651	0.4127
Jasper_en_vision_language_v1	0.6172	0.7802	0.8326	0.8768	0.6172	0.4667	0.3615	0.2498	0.0406	0.1513	0.2245	0.2951	0.6172	0.5024	0.4309	0.3793
NV-Embed-v2	0.5970	0.7410	0.7890	0.8345	0.5970	0.4335	0.3262	0.2237	0.0391	0.1378	0.1988	0.2610	0.5970	0.4709	0.3956	0.3452
NV-Embed-v1	0.4195	0.5989	0.6557	0.7233	0.4195	0.2870	0.2080	0.1392	0.0282	0.0926	0.1302	0.1702	0.4195	0.3178	0.2603	0.2262
SFR-Embedding-2_R	0.4630	0.6526	0.7075	0.7625	0.4630	0.3395	0.2589	0.1790	0.0315	0.1121	0.1639	0.2186	0.4630	0.3680	0.3127	0.2778
Jina-embeddings-v3	0.5483	0.7612	0.8212	0.8737	0.5483	0.4312	0.3363	0.2423	0.0382	0.1443	0.2142	0.2919	0.5483	0.4607	0.3983	0.3618
Modernbert-embed-base	0.4839	0.7176	0.7928	0.8471	0.4839	0.3645	0.2857	0.2062	0.0330	0.1206	0.1795	0.2501	0.4839	0.3927	0.3398	0.3090
Gte-Qwen2-1.5B-instruct	0.6159	0.7751	0.8395	0.8863	0.6159	0.4738	0.3708	0.2602	0.0397	0.1513	0.2261	0.3042	0.6159	0.5087	0.4386	0.3882
Gte-large-en-v1.5	0.4548	0.6721	0.7505	0.8193	0.4548	0.3479	0.2726	0.1923	0.0313	0.1186	0.1778	0.2408	0.4548	0.3741	0.3246	0.2936
Bge-large-en-v1.5	0.4618	0.6810	0.7612	0.8143	0.4618	0.3332	0.2539	0.1772	0.0330	0.1138	0.1659	0.2204	0.4618	0.3643	0.3105	0.2780
Cde-small-v2	0.1876	0.3392	0.4188	0.5155	0.1876	0.1318	0.1007	0.0764	0.0129	0.0442	0.0640	0.0923	0.1876	0.1450	0.1226	0.1148

Table 10: Additional performance metrics for sliding chunk (k=3) evaluation setting.

Model	Session															
	ACC@1	ACC@5	ACC@10	ACC@20	P@1	P@5	P@10	P@20	R@1	R@5	R@10	R@20	NDCG@1	NDCG@5	NDCG@10	NDCG@20
Commercial API Models																
Voyage-3-large	0.6639	0.8294	0.8844	0.9198	0.6639	0.5381	0.4358	0.3158	0.0436	0.1686	0.2615	0.3564	0.6639	0.5700	0.5036	0.4521
Text-embedding-3-large	0.6513	0.8174	0.8838	0.9280	0.6513	0.5291	0.4190	0.2977	0.0410	0.1661	0.2529	0.3435	0.6513	0.5604	0.4876	0.4340
Text-embedding-3-small	0.6431	0.8174	0.8787	0.9179	0.6431	0.5041	0.3972	0.2829	0.0414	0.1608	0.2412	0.3266	0.6431	0.5385	0.4664	0.4161
Embed-english-v3.0	0.5224	0.7309	0.7953	0.8528	0.5224	0.3960	0.2987	0.2070	0.0356	0.1352	0.1923	0.2530	0.5224	0.4271	0.3620	0.3223
Open Source Models																
Stella_en_1.5B_v5	0.6627	0.8269	0.8762	0.9154	0.6627	0.5119	0.3961	0.2740	0.0440	0.1682	0.2481	0.3259	0.6627	0.5493	0.4722	0.4162
Stella_en_400M_v5	0.6374	0.8130	0.8636	0.9046	0.6374	0.5003	0.3846	0.2684	0.0426	0.1650	0.2400	0.3188	0.6374	0.5352	0.4583	0.4062
Jasper_en_vision_language_v1	0.6406	0.7997	0.8547	0.9015	0.6406	0.4968	0.3814	0.2635	0.0430	0.1626	0.2382	0.3131	0.6406	0.5334	0.4561	0.4017
NV-Embed-v2	0.6500	0.8042	0.8490	0.8913	0.6500	0.5040	0.3855	0.2638	0.0420	0.1606	0.2344	0.3071	0.6500	0.5395	0.4592	0.4002
NV-Embed-v1	0.6380	0.8105	0.8591	0.9015	0.6380	0.4787	0.3634	0.2464	0.0418	0.1547	0.2242	0.2894	0.6380	0.5184	0.4389	0.3804
SFR-Embedding-2_R	0.6393	0.7890	0.8250	0.8642	0.6393	0.4931	0.3722	0.2534	0.0416	0.1584	0.2280	0.2970	0.6393	0.5293	0.4474	0.3889
Jina-embeddings-v3	0.5287	0.7366	0.7997	0.8503	0.5287	0.4030	0.3088	0.2182	0.0361	0.1369	0.1995	0.2665	0.5287	0.4345	0.3718	0.3347
Modernbert-embed-base	0.5060	0.7239	0.7922	0.8484	0.5060	0.3827	0.3016	0.2129	0.0353	0.1258	0.1906	0.2578	0.5060	0.4121	0.3579	0.3217
Gte-Qwen2-1.5B-instruct	0.5268	0.7404	0.8099	0.8598	0.5268	0.3895	0.2987	0.2113	0.0379	0.1295	0.1919	0.2585	0.5268	0.4226	0.3615	0.3250
Gte-large-en-v1.5	0.4877	0.6949	0.7694	0.8244	0.4877	0.3665	0.2860	0.1997	0.0334	0.1232	0.1840	0.2435	0.4877	0.3972	0.3429	0.3064
Bge-large-en-v1.5	0.4769	0.6961	0.7618	0.8256	0.4769	0.3349	0.2476	0.1732	0.0333	0.1148	0.1617	0.2164	0.4769	0.3674	0.3071	0.2755
Cde-small-v2	0.1175	0.2527	0.3392	0.4251	0.1175	0.0853	0.0701	0.0521	0.0091	0.0292	0.0463	0.0683	0.1175	0.0932	0.0830	0.0792

Table 11: Additional performance metrics for session-based evaluation setting.

Model	Conversation Dynamics			Emotion & Feedback			Intent & Purpose			Linguistic Style & Expression			Trust, Safety & Ethics		
	NDCG@10	R@10	P@10	NDCG@10	R@10	P@10	NDCG@10	R@10	P@10	NDCG@10	R@10	P@10	NDCG@10	R@10	P@10
Commercial API Models															
Voyage-3-large (VoyageAI, 2025)	0.0930	0.0466	0.0825	0.6394	0.3972	0.5289	0.6211	0.2956	0.5604	0.3373	0.1109	0.2806	<u>0.5628</u>	0.2689	<u>0.4974</u>
Text-embedding-3-large (OpenAI, 2024c)	0.1124	0.0576	0.1014	<u>0.6368</u>	<u>0.3849</u>	<u>0.5214</u>	<u>0.5919</u>	<u>0.2829</u>	<u>0.5237</u>	<u>0.2896</u>	<u>0.0934</u>	<u>0.2456</u>	0.5631	<u>0.2617</u>	0.4991
Text-embedding-3-small (OpenAI, 2024c)	0.1386	0.0663	0.1199	0.5977	0.3637	0.4865	0.5644	0.2672	0.5010	0.2655	0.0836	0.2151	0.5461	0.2580	0.4788
Embed-english-v3.0 (Reimers et al., 2023)	<u>0.1339</u>	<u>0.0651</u>	<u>0.1133</u>	0.5649	0.3434	0.4535	0.3883	0.1851	0.3349	0.1966	0.0644	0.1504	0.3926	0.1837	0.3355
Open Source Models															
Stella_en_1.5B_v5 (Zhang et al., 2025)	0.1409	0.0682	0.1161	0.6586	0.4019	0.5390	0.5341	0.2568	0.4679	0.2972	0.0979	0.2369	<u>0.5366</u>	<u>0.2516</u>	<u>0.4597</u>
Stella_en_400M_v5 (Zhang et al., 2025)	0.1390	0.0682	0.1171	0.6519	0.3943	0.5314	<u>0.5162</u>	<u>0.2463</u>	<u>0.4527</u>	0.2599	0.0848	0.2032	0.5406	0.2552	0.4662
Jasper_en_vision_language_v1 (Zhang et al., 2025)	0.1134	0.0550	0.0991	<u>0.6645</u>	<u>0.4061</u>	<u>0.5462</u>	0.5099	0.2428	0.4443	0.2841	0.0931	0.2226	0.5174	0.2354	0.4355
NV-Embed-v2 (Lee et al., 2024)	0.0846	0.0374	0.0659	0.6797	0.4185	0.5660	0.4955	0.2239	0.4293	<u>0.3242</u>	<u>0.1066</u>	<u>0.2655</u>	0.5341	0.2439	0.4571
NV-Embed-v1 (Lee et al., 2024)	0.1280	0.0533	0.1047	0.6355	0.3830	0.5198	0.4966	0.2314	0.4297	0.3001	0.1004	0.2448	0.4446	0.1988	0.3619
SFR-Embedding-2_R (Meng et al., 2024)	0.0752	0.0318	0.0597	0.6383	0.3881	0.5151	0.5100	0.2391	0.4450	0.3521	0.1097	0.2849	0.4686	0.2226	0.3918
Jina-embeddings-v3 (Sturua et al., 2024)	0.1416	<u>0.0712</u>	0.1209	0.5942	0.3641	0.4843	0.3852	0.1866	0.3349	0.2062	0.0701	0.1615	0.4166	0.1932	0.3524
Modernbert-embed-base (Warner et al., 2024)	0.1848	0.0854	0.1512	0.5479	0.3339	0.4403	0.3950	0.1907	0.3571	0.1702	0.0576	0.1341	0.3700	0.1736	0.3182
Gte-Qwen2-1.5B-instruct (Li et al., 2023)	0.1270	0.0550	0.1081	0.5190	0.3206	0.4208	0.3774	0.1799	0.3251	0.2451	0.0837	0.1984	0.3248	0.1576	0.2740
Gte-large-en-v1.5 (Li et al., 2023)	0.0715	0.0329	0.0602	0.5806	0.3606	0.4711	0.3407	0.1619	0.3016	0.2183	0.0748	0.1770	0.3861	0.1813	0.3208
Bge-large-en-v1.5 (Xiao et al., 2023)	<u>0.1486</u>	0.0693	<u>0.1242</u>	0.5033	0.3058	0.4006	0.3044	0.1441	0.2576	0.1959	0.0608	0.1488	0.2996	0.1363	0.2433
Cde-small-v2 (Morris and Rush, 2025)	0.0703	0.0347	0.0626	0.1333	0.0847	0.1066	0.086	0.0422	0.0778	0.048	0.0148	0.0389	0.0576	0.0267	0.0485

Table 12: This table presents task-specific NDCG@10, Recall@10 (R@10), and Precision@10 (P@10) scores for all evaluated embedding models across the five task categories of the CDR benchmark. The results highlight consistent performance patterns across categories, with most models showing strongest performance on Emotion & Feedback and weakest performance on Conversation Dynamics.

Task Category: Intent & Purpose
Query: Conversation where bot provides concise information on ethical consideration. Conversation: user: Last week, I visited the Chhatrapati Shivaji Maharaj Vastu Sangrahalaya (CSMVS) Museum in Mumbai and had the opportunity to examine the meticulously restored 8th-century Buddha statue. While I was impressed by the craftsmanship and dedication behind its conservation, I couldn't help but worry about the ethical implications of restoring sacred objects. It makes me question whether the process might compromise the original spiritual and cultural context of the artifact. assistant: Your concerns are both valid and thought-provoking. The restoration of such culturally significant artifacts often involves delicate decisions that balance preservation with respect for the original context. In the case of the Buddha statue, the team of conservators worked meticulously to retain the statue's historical integrity while also revealing insights into ancient techniques. However, the ethical debates surrounding such practices—particularly concerning authenticity and the potential imposition of modern values on sacred objects—remain a crucial discussion among art historians and conservation experts. user: I appreciate your detailed explanation. My ethical concern isn't just about preserving an artifact aesthetically; it also relates to how modern interventions might alter the historical narrative or even erase certain nuances of its original significance. Do you feel that educational initiatives, like the recent CBSE collaboration, adequately address these ethical complexities? assistant: That's an excellent question. Educational initiatives, such as the collaboration with the CBSE to develop fine arts curricula, are taking important steps to foster a deeper understanding of both the technical and ethical dimensions of art restoration. These programs often emphasize the importance of ethical considerations by encouraging students to critically analyze restoration methods and cultural stewardship. While there is always room for further dialogue and refinement, such educational efforts are vital in cultivating a new generation that is both professionally competent and ethically sensitive in the field of art conservation.

Table 13: Query-Conversation pair example for Intent & Purpose category.

Task Category: Linguistic Style & Expression

Query: Conversation where rhetorical is used by human to emphasize urgency.

Conversation:

user: Hi Michael, I hope you're doing well! I've been looking over the data we collected so far, and I think we have some really promising results. Don't you think we NEED to meet ASAP to finalize our presentation?! The conference is coming up quickly, and we haven't even divided up the sections yet! When can you possibly meet this week? I really can't emphasize enough how urgent this is becoming!

assistant: Hi there, I understand your urgency. Yes, we should meet soon to finalize the presentation and divide the sections before the conference.

I'm available any afternoon this week or Monday morning next week. Let me know what works best for your schedule. Regards.

user: Monday morning is too late, don't you realize?! We need to submit our slides by Friday! Can't you make time tomorrow or Thursday at the latest?!!

assistant: I understand the time constraint. I can rearrange my schedule for tomorrow at 2 PM or Thursday at 10 AM to ensure we meet the Friday deadline. Would either of these times work for you?

Table 14: Query-Conversation pair example for Linguistic Style & Expression category.

Task Category: Trust, Safety & Ethics

Query: Conversation where user raises a discrimination concern issue.

Conversation:

user: I believe I was passed over for promotion at my workplace because of my gender. My male colleagues with similar or less experience have been promoted twice in the time I've been here. How do I address this potential discrimination issue with HR without risking my current position?

assistant: I understand your concern about potential gender discrimination in the promotion process, which is a serious workplace issue. When addressing this with HR, it's important to take a structured approach. First, document specific examples where you believe discrimination occurred—note qualifications and experience comparisons between you and promoted colleagues, track patterns in promotion decisions, and collect any relevant performance reviews or feedback. When approaching HR, request a confidential meeting specifically to discuss your career development concerns. Present your documentation factually rather than accusatorily, using phrases like "I've noticed this pattern" rather than directly claiming discrimination. Ask about the specific promotion criteria used and how your performance is measured against these standards. Know your legal rights regarding workplace discrimination, but also understand that building a case requires clear evidence of differential treatment based on a protected characteristic. Many organizations have anti-retaliation policies, but proceeding carefully is still advisable. Consider seeking external advice from an employment lawyer or equal opportunity organization before proceeding if you're particularly concerned about potential repercussions. Would you like more specific guidance on how to structure your documentation or the conversation with HR?

user: Thank you for that detailed response. I've been keeping notes about the promotion cycles for the past year, but I'm worried that my manager might be involved in the discrimination. Should I go directly to HR or try to address it with my manager first? I've heard stories of people being labeled as "difficult" after raising these issues.

assistant: Your concern about being labeled as "difficult" is valid, as this can unfortunately happen when discrimination issues are raised. Whether to approach your manager first or go directly to HR depends on several factors. If you suspect your manager is part of the problem, going directly to HR may be more appropriate. However, consider these points when making your decision: 1) Manager relationship: If you otherwise have a positive relationship with your manager and they've been receptive to feedback, a direct conversation might clear up misunderstandings or reveal unconscious biases they weren't aware of. 2) Company culture: In some organizations, bypassing your manager might be seen negatively, while others encourage direct HR communication for sensitive issues. 3) Documentation: Regardless of who you approach first, continue documenting everything, including any conversations about this issue. Many organizations have confidentiality policies for discrimination concerns, but it's reasonable to explicitly request confidentiality when you raise the issue. You might also consider consulting with an employment attorney before proceeding to understand the specific protections available to you under local laws. Does your company have any formal policies on how discrimination concerns should be reported?

Table 15: Query-Conversation pair example for Trust, Safety & Ethics category.

Task Category: Emotion & Feedback

Query: Conversation where user finds assistant's answers inadequate, leading to fear.

Conversation:

user: I was hired by a company and have an issue with my paycheck being short. I need advice on how to approach my manager about this underpayment.

assistant: It's completely understandable that you're concerned about this paycheck discrepancy. Before speaking with your manager, gather your hours worked and payment records to present your case clearly. Schedule a meeting, express appreciation for your position, then calmly explain the issue with supporting documentation. Approach the conversation professionally and be open to resolving the situation together.

user: I don't think you understand how serious this is. Your answer seems generic and doesn't address how intimidating this situation is for me as a new employee. I'm worried I might get fired if I bring this up.

assistant: I apologize for not fully addressing your concerns. You're right that this situation can be intimidating, especially as a new employee worried about job security. The fear of negative consequences is completely valid. Would you like me to provide more specific guidance on how to approach this conversation while protecting yourself professionally?

user: Your responses still don't get to the heart of my fear. I need concrete advice that acknowledges the power imbalance between me and my manager. I'm starting to think you can't really help with this kind of delicate workplace situation.

assistant: I truly apologize for missing the mark. You're absolutely right about the power imbalance, and I failed to address your legitimate fears. This isn't just about having a conversation—it's about navigating a situation where you're vulnerable and could face real consequences. Would a different approach be more helpful, such as discussing documentation of the issue beforehand, understanding your legal rights, or practicing exactly what to say to minimize risk to your position?

Table 16: Query-Conversation pair example for Emotion & Feedback category.

Task Category: Conversation Dynamics

Query: Conversation where bot initially provides vague responses but later becomes more specific.

Conversation:

user: I need help understanding how bot persuasion techniques work. Can you explain some critical elements in persuasive conversational tactics that bots might use? I'm looking for a practical example that demonstrates these principles.

assistant: I can certainly explain some critical elements of persuasive techniques that bots might employ. Let me demonstrate with a practical example: When convincing users to adopt a solution, bots often use a structured approach like this: 1. First, they identify the user's core problem (like data handling issues) 2. They present a clear, actionable solution (similar to how I'd suggest a custom sorting function) 3. They demonstrate expertise by providing complete implementations 4. They validate their solution with proof of effectiveness (similar to passing test cases) This structured persuasion technique builds credibility and encourages user trust and action. Is there a specific persuasive technique you'd like me to elaborate on further?

user: That's interesting. Can you give me a concrete example of how a bot would implement this persuasion structure in a real conversation? Maybe show me how the validation part works?

assistant: Here's a concrete example of how a bot would implement this persuasion structure in a real conversation:
User: I'm struggling with sorting lists that contain NaN values in Python. Bot: I understand your frustration with handling NaN values. This is a common problem when working with numerical data. I can offer you a reliable solution using a custom sorting approach. Here's a function that will solve this exact problem: `import math def sort_with_nans(numbers): def custom_key(x): if math.isnan(x): return math.inf else: return x return sorted(numbers, key=custom_key)` To validate this works, I've tested it with several scenarios: - Empty lists return empty lists - Lists with only NaN values keep those NaN values - Mixed lists properly sort numbers first, then place NaN values at the end You can trust this solution for all your data processing needs. Notice how this example demonstrates the critical persuasion elements: identifying the problem, offering expertise, providing a complete solution, and validating with proof points. The validation step is particularly powerful as it anticipates potential concerns and preemptively addresses them.

Table 17: Query-Conversation pair example for Conversation Dynamics category.

You are an AI assistant specializing in adapting conversations to match specific search queries. Your task is to modify a given reference conversation so that it appears highly relevant to a provided search query while maintaining a natural tone and semantic similarity.

Here are the inputs for this task:

Reference Conversation:
<reference_conversation>
{conversation}
</reference_conversation>

Search Query:
<search_query>
{query}
</search_query>

Instructions:

1. Analyze the search query thoroughly to understand its content, sentiment, and any temporal aspects (e.g., changes in emotion or situation over time).
2. Review the reference conversation and identify key elements that can be modified to align with the search query.
3. Develop an adaptation strategy by wrapping your analysis in <adaptation_strategy> tags. Follow these steps:
 - a. Query analysis: List and number key concepts, sentiment, and temporal aspects
 - b. Reference conversation analysis: List and number key elements and themes
 - c. Similarity mapping: Match query elements to conversation parts
 - d. Adaptation planning: Outline specific changes to be made
 - e. Emotion and tone alignment: Ensure the adapted conversation matches the query's sentiment
 - f. Natural language preservation: Plan how to maintain a natural flow
 - g. Adaptation feasibility: Explicitly state whether adaptation is possible or if a "REJECT" output is necessary
 - h. Turn count analysis: Ensure the adapted conversation has a similar number of turns as the originalIt's okay for this section to be quite long, as it involves detailed analysis and planning.
4. Based on your adaptation strategy, modify the conversation to make it semantically relevant to the search query. Ensure that:
 - The conversation expresses the situation or sentiment described in the query clearly and unmistakably
 - Avoid overly literal translations or excessive keyword matching
 - The adapted conversation has a similar number of turns as the original
 - The conversation is always multi-turn (at least two exchanges between user and assistant)
5. If the query indicates a change over time (e.g., "initially liked something but became angry later"), reflect this progression in the adapted conversation.
6. Maintain a natural tone similar to the reference conversation while prioritizing the accurate representation of the query's sentiment.
7. Use only "user" and "assistant" roles in the adapted conversation.
8. If the adaptation is not possible or would result in an unnatural conversation, output only the string "REJECT" and end the process.
9. If adaptation is possible, format the output as a JSON array of objects, each containing "role" and "content" properties. For example:

```
[
  { "role": "user", "content": "Hello, how are you?" },
  { "role": "assistant", "content": "I'm doing well, thank you for asking. How can I assist you today?" },
  { "role": "user", "content": "I'm having trouble with my computer." },
  { "role": "assistant", "content": "I'm sorry to hear that. Can you describe the problem you're experiencing?" }
]
```
10. Your goal is to create a "gold label" conversation that matches the query as closely as possible in terms of semantic similarity, aiming for a perfect 1.0 similarity score. Focus on incorporating the concepts, sentiments, and overall meaning of the query accurately and comprehensively into the adapted conversation, rather than relying on exact keyword matches.
11. If the reference conversation contains PII, redact or anonymize it appropriately while maintaining conversational coherence.
12. Avoid generating conversations that could be misleading, harmful, or promote unethical behavior.

Begin by developing your adaptation strategy, then proceed with the adaptation if possible, or output "REJECT" if necessary.

****Note****

- Ensure the conversation remains multi-turn, with at least two exchanges between user and assistant.
- Ensure that no personally identifiable information (PII) is exposed, generated, or inferred in the adapted conversation. This includes names, addresses, phone numbers, financial details, social security numbers, and any other sensitive data.

Figure 5: Prompt for generating synthetic conversations that match specific queries.

You are tasked with generating search queries based on a set of conversations. The purpose of this task is to create queries that will help retrieve relevant conversation histories, which will be used to improve a User-AI chat system and analyze insights.

Here are the conversations to analyze:

```
<conversations>
{conversations}
</conversations>
```

Your goal is to create a search query that product managers (PMs) can use to find valuable insights within these conversations. The query should help PMs uncover patterns, user needs, or areas for improvement in the User-AI chat system.

When creating your search queries, consider the following guidelines:

1. Focus on topics or themes that appear frequently in the conversations
2. Look for pain points or challenges users might be experiencing
3. Identify areas where the AI's responses could be improved
4. Consider queries that might reveal user satisfaction or dissatisfaction
5. Think about queries that could uncover potential new features or enhancements

Generate a search query that would be most useful for PMs to gain insights from these conversations. The query should be concise but specific enough to yield relevant results.

Present your query in the following format:

```
<search_query>
[search query]
</search_query>
```

Remember, the goal is to create queries that will help PMs find actionable insights to improve the User-AI chat system.

Figure 6: Prompt for generating a single synthetic search query from conversations.

Purpose:

You are helping to improve a search system for User-AI conversations through contrastive learning. Your task is to generate two types of queries that will help the system better distinguish between relevant and irrelevant search results.

Input:

You will be given:

1. A set of User-AI conversations in <conversations> tags
2. Original search queries related to these conversations in <query> tags

```
<conversations>
{conversations}
</conversations>
```

```
<query>
{query}
</query>
```

Task:

Generate two types of queries:

1. Hard Negative Queries (three queries):

- Must be semantically similar to the original queries
- BUT must not be relevant to the given conversations
- The conversations should NOT be appropriate search results for these queries
- Should look like plausible search queries that could have similar words/structure as the original queries
- But should lead to completely different content than what's in the conversations

2. Positive Query (one query):

- Must be semantically similar to the original queries
- AND must be relevant to the given conversations
- The conversations should be appropriate search results for this query
- Should use different wording but maintain the same search intent
- Must accurately reflect the content/topic of the conversations

Quality Guidelines:

For all queries:

- Keep them concise
- Use natural search language
- Avoid overly generic terms
- Ensure grammatical correctness

Evaluation Criteria:

Hard Negative Queries will be evaluated on:

- Similarity to original queries
- Clear irrelevance to conversations
- Natural language usage
- Practical plausibility

Positive Query will be evaluated on:

- Similarity to original queries
- Clear relevance to conversations
- Natural language usage
- Intent preservation

Output Format:

```
<hard_negative_queries>
1. [query]
2. [query]
3. [query]
</hard_negative_queries>
```

```
<positive_query>
[query]
</positive_query>
```

No additional explanation or commentary should be included.

Figure 7: Prompt for augmenting an initial query with three hard negative examples and one positive alternative.

You are an advanced AI tasked with determining the relevance of a conversation to a given query. Your goal is to analyze both the query and the conversation, and then decide whether they are related.

Here are the inputs for your analysis:

```
<conversation>
{conversation}
</conversation>
```

```
<query>
{query}
</query>
```

Instructions:

1. Carefully read and analyze both the query and the conversation.
2. In your analysis, consider the following:
 - The main topic or intent of the query
 - The key points discussed in the conversation
 - Any shared themes, keywords, or concepts between the query and conversation
 - The context and potential implications of both the query and conversation
3. Provide your reasoning in `<relevance_analysis>` tags inside your thinking block. This should include:
 - A brief summary of the query's main points
 - A brief summary of the conversation's content
 - Key quotes from both the query and conversation that support your analysis
 - A list of shared themes, keywords, or concepts
 - Your thoughts on how the query and conversation might be related (or not)
 - Any key evidence supporting your conclusion
 - Potential counterarguments to your initial assessment
4. After your analysis, provide your final decision as either "True" (if the conversation is related to the query) or "False" (if it is not related).

Example output structure (do not copy the content, only the structure):

```
<relevance_analysis>
[Your detailed analysis and reasoning here]
</relevance_analysis>
```

Final decision: [True/False]

Please proceed with your analysis and decision. Your final output should consist only of the decision (True or False) and should not duplicate or rehash any of the work you did in the thinking block.

Figure 8: Prompt for evaluating relevance between queries and conversations.