An Intent–Background–Constraint Structured Representation for Underspecified Queries in Long-Term Conversational Memory Retrieval

Yuma Masuda[1,*], Tetsuya Takiguchi[1], Yasuo Ariki[1], Yusuke Iwasaki[2], Motoki Kojima[2] and Masaki Yamamoto[2]

[1]Graduate School of System Informatics, Kobe University, Kobe, Japan
[2]DENSO TEN Limited, Kobe, Japan

Abstract

Long-term conversational agents must retrieve user-specific evidence distributed across past dialogue, but underspecified queries often leave intent, background, and constraints unstated. We propose an intermediate structured query representation with three fields: Intent, Background, and Constraints (IBC). IBC is inferred from the query alone to describe information to retrieve, not reconstruct concrete facts from dialogue history. For dense retrieval, we prepend all three fields to the preserved original query; we call this realization Full IBC. We evaluate on custom query–answer–evidence (QAE) data derived from LiHua-World and on LongMemEval, comparing the Original query, Free rewrite, Full IBC, and element-specific prefixes. On LiHua-World, Full IBC significantly outperformed both baselines, increasing Recall@10 from 0.311 to 0.363, MRR@10 from 0.293 to 0.374, and nDCG@10 from 0.239 to 0.292. Element-wise ablations indicated that Intent primarily drove this gain. On LongMemEval, however, Full IBC reduced nDCG@10 from 0.583 to 0.420; Background-only produced the smallest degradation, whereas Constraints-only and Full IBC degraded substantially more. The dataset-level directions on both benchmarks were reproduced by two additional dense retrievers. On LongMemEval, human analysis found that 95.5% of generated IBC content was supported by the original query, but 83.3% added no information that discriminated the gold evidence. These results separate the IBC representation from uniform natural-language concatenation: semantically supported expansion may lack cues that identify the evidence, motivating query-adaptive use of individual elements.

Keywords

Long-term conversational memory, Structured query representation, Query expansion, Dense retrieval

1. Introduction

Long-term conversational agents must retrieve relevant information from extensive dialogue histories. User utterances, however, are often short and omit the intent, background, and constraints needed to identify the target, which challenges similarity-based dense retrieval [1]. In this study, *underspecified queries* are queries that lack explicit retrieval cues and whose interpretation or answer depends on user-specific information distributed across past conversations.

We formulate IBC as an intermediate query representation that separates the retrieval objective, relevant context, and applicable conditions before instantiating them as a natural-language prefix for dense retrieval. Background and Constraints do not reconstruct concrete facts from past dialogue; instead, they describe the types of information that may constitute the retrieval target. This decomposition enables analysis of the relationship between generated information and retrieval behavior.

We investigate the following research questions:

- **RQ1**: Does the IBC prefix improve retrieval for underspecified queries in long-term conversational memory?
- **RQ2**: Does the IBC prefix generalize to an external benchmark?

IJCKG 2026: 15th International Joint Conference on Knowledge Graphs, November 19–21, 2026, Bangkok, Thailand
*Corresponding author.
267x076x@stu.kobe-u.ac.jp (Y. Masuda); takigu@kobe-u.ac.jp (T. Takiguchi); ariki@kobe-u.ac.jp (Y. Ariki); yuusuke.iwasaki.j7p@jpgr.denso.com (Y. Iwasaki); motoki.kojima.j3h@jpgr.denso.com (M. Kojima); masaki.yamamoto.j3h@jpgr.denso.com (M. Yamamoto)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-06

CEUR-WS.org/Vol-4316/pd-short-4.pdf

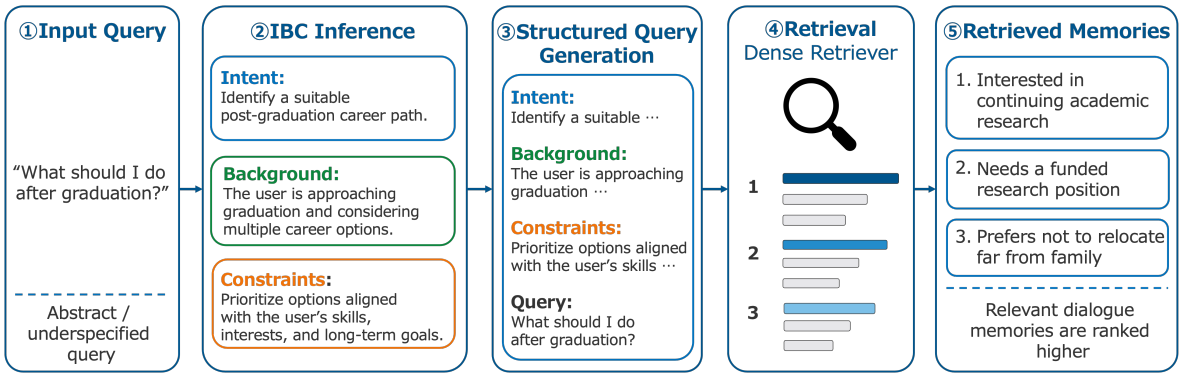


Figure 1: IBC-structured query generation and long-term conversational memory retrieval.

- **RQ3**: How does each IBC element affect retrieval performance?

We address RQ1 using QAE data derived from LiHua-World [2], RQ2 using LongMemEval [3], and RQ3 through element-wise ablation on both datasets.

Our contributions are: (1) an IBC representation that makes implicit retrieval requirements explicit; (2) evaluation of its natural-language prefix on LiHua-World QAE and LongMemEval; and (3) element-wise and human analyses demonstrating a mismatch between semantic validity and retrieval utility.

2. Related Work

2.1. Long-Term Conversational Memory Retrieval

LiHua-World is a long-term daily-life dialogue dataset introduced with MiniRAG [2], whereas LongMemEval evaluates extraction, multi-session and temporal reasoning, knowledge updates, and abstention [3]. HAConvDR addresses history-aware conversational dense retrieval [4]. Our IBC inference instead uses the current query without retrieving or rewriting from dialogue history.

2.2. Conversational Query Rewriting and Retrieval-Oriented Query Expansion

CANARD frames question-in-context rewriting as producing a self-contained question from conversational history [5]; CONQRR learns standalone rewrites with a retrieval-oriented reward [6]. IBC instead infers the retrieval requirements from the current query alone, without resolving omitted facts against history. Query2doc expands a query with an LLM-generated pseudo-document [7], whereas ExpandR jointly optimizes LLM-guided expansion and dense retrieval [8]. IBC is the intermediate structured representation, whereas Full IBC is the particular inference-time natural-language realization evaluated in this study, rather than pseudo-evidence generation or retriever training. It retains the original query and permits field-wise analysis. The Free rewrite control in our experiments is not an implementation of these published systems, so our results do not compare their effectiveness directly.

3. Structured Query Representation Based on IBC

Figure 1 illustrates the proposed workflow. The workflow comprises five stages: ① the input query is received; ② IBC is inferred from the query alone; ③ the three elements and original query form a structured query; ④ the dialogue history is retrieved and ranked; and ⑤ the retrieved memories are returned.

3.1. IBC Definition and Inference

Intent denotes the retrieval objective that the user seeks to achieve, **Background** denotes situations, events, or relations that may constitute the retrieval target, and **Constraints** denote conditions to consider during retrieval or answer generation. Background and Constraints describe the types of information to retrieve rather than reconstructing concrete past facts. IBC is not intended as a comprehensive taxonomy of pragmatics, but as an operational representation that links generated content to retrieval behavior.

We use the OpenAI API with GPT-5.1 to infer IBC from only the original query. The output is JSON with the keys `intent`, `background`, and `constraints`. A condensed version of the inference instructions is shown below.

```
Given a user query, infer concise, query-grounded:
- INTENT: the user's retrieval objective.
- BACKGROUND: relevant situations, events, or relations.
- CONSTRAINTS: applicable requirements or limitations.
Do not invent unsupported specific facts.
Return JSON with keys "intent", "background", and
"constraints".
Query: {query}
```

3.2. Prefix Expansion

Full IBC concatenates the three fields without selection or weighting while preserving the original query:

```
Intent: {intent}
Background: {background}
Constraints: {constraints}
Query: {original query}
```

For an expanded query $\tilde{q}$ and a chunk c, the retrieval score is

$$s(\tilde{q}, c) = \cos(\mathbf{e}_{\tilde{q}}, \mathbf{e}_c), \tag{1}$$

where $\cos$ denotes cosine similarity.

4. Datasets

4.1. Custom QAE with Underspecified Queries Derived from LiHua-World

We used the OpenAI API with GPT-5.1 to generate QAE instances from LiHua-World dialogue logs, with queries containing fewer surface cues, ellipsis, or implication. The dataset contains 227 queries and 1,234 256-token chunks, with 2.34 gold chunks per query on average. One annotator manually checked all 227 instances to determine whether each query was sufficiently underspecified, whether the query and answer were consistent, and whether the gold evidence contained the information required for the answer. The IBC model receives only the query, without the answer or gold evidence.

4.2. LongMemEval

From the 500 questions in LongMemEval's `longmemeval_m_cleaned.json` [3], we exclude 30 abstention questions and 51 without a target turn, leaving 419 shared by both conditions. The retrieval unit is a conversation turn, and the target turns provided by LongMemEval are treated as gold relevance judgments.

5. Experimental Setup

We evaluate only the ranking of gold evidence, without response generation. We use the OpenAI API with GPT-5.1 for IBC inference and Free rewrite. We evaluate three dense retrievers: BAAI/bge-base-en-v1.5 [9], all-MiniLM-L6-v2 [10], and e5-base-v2 [11]. For all models, embeddings are L2-normalized and ranked by cosine similarity. For LiHua-World, we retrieve the top 10 results from a 256-token index constructed from one year of dialogue logs; for LongMemEval, we retrieve individual turns and retain the top 50.

The main experiment compares **Original query**, **Free rewrite** (a one-sentence rewrite that preserves the information need and concrete terms without adding information), and **Full IBC**. For element-wise analysis, we reuse each IBC output and compare Intent-only, Background-only, and Constraints-only conditions with Full IBC. The LiHua-World main experiment uses 227 queries, its statistical ablation uses the 221 valid under every condition, and LongMemEval uses 419.

We report Recall@10, MRR@10, and nDCG@10 [12] for LiHua-World, and Recall@1, Recall@10, and nDCG@10 for LongMemEval. For LiHua-World, we use paired two-sided randomization tests with Holm correction applied separately to the main, single-element, and supplementary two-element comparison families, together with paired-bootstrap 95% confidence intervals. For LongMemEval, we use a two-sided Wilcoxon signed-rank test for Original query versus Full IBC and bootstrap intervals for element-wise comparisons.

6. Results

6.1. Main Results

Table 1 presents the main results. On LiHua-World, Full IBC significantly outperformed Original query and Free rewrite on all metrics after Holm correction; Free rewrite and Original query did not differ. This comparison does not isolate structure from query length or preservation of the original query. On LongMemEval, Full IBC underperformed Original query on every metric, and the nDCG@10 difference was significant ($p = 1.57 \times 10^{-23}$). Thus, the IBC prefix improved retrieval for the underspecified LiHua-World QAE but did not generalize when uniformly applied to the external benchmark. As a retriever robustness check, we repeated the Original query–Full IBC comparison with all-MiniLM-L6-v2 and e5-base-v2. Both alternative retrievers reproduced the dataset-level direction observed with BAAI/bge-base-en-v1.5: Full IBC improved LiHua-World but underperformed the original query on LongMemEval.

Table 1
Main results with BAAI/bge-base-en-v1.5. Bold marks the highest value in each column.

(a) LiHua-World ($n = 227$)			
Condition	Recall@10	MRR@10	nDCG@10
Original query	0.311	0.293	0.239
Free rewrite	0.313	0.292	0.241
Full IBC	**0.363**	**0.374**	**0.292**
(b) LongMemEval ($n = 419$)			
Condition	Recall@1	Recall@10	nDCG@10
Original query	**0.434**	**0.878**	**0.583**
Full IBC	0.325	0.704	0.420

6.2. Element-Wise Ablation

Table 2 presents the element-wise results. Because dense retrieval is nonlinear, these ablations measure the end-to-end behavior of each prefix formulation rather than an additive causal contribution of each

element. On LiHua-World, Intent-only significantly outperformed Original query on every metric, but did not differ significantly from Full IBC, suggesting that the Full IBC improvement was driven primarily by Intent. Background-only had a limited effect, and Constraints-only did not outperform Original query. In the supplementary two-element comparison, adding Constraints to Intent significantly reduced MRR@10 and nDCG@10. On LongMemEval, every IBC condition underperformed Original query, and all paired-bootstrap 95% confidence intervals for the nDCG@10 differences were below zero. Background-only produced the smallest degradation, whereas Constraints-only and Full IBC degraded substantially.

Table 2
Element-wise IBC ablation with BAAI/bge-base-en-v1.5. Bold marks the highest value in each column.

(a) LiHua-World, shared queries ($n = 221$)

Condition	Recall@10	MRR@10	nDCG@10
Original query	0.319	0.301	0.245
Intent-only	0.369	**0.384**	0.298
Background-only	0.325	0.352	0.272
Constraints-only	0.295	0.270	0.220
Full IBC	**0.372**	**0.384**	**0.300**

(b) LongMemEval ($n = 419$)

Condition	Recall@1	Recall@10	nDCG@10	Δ nDCG@10
Original query	**0.434**	**0.878**	**0.583**	—
Intent-only	0.351	0.738	0.457	−0.126
Background-only	0.432	0.802	0.521	−0.061
Constraints-only	0.327	0.718	0.431	−0.151
Full IBC	0.325	0.704	0.420	−0.163

7. Analysis

Intent mainly improved LiHua-World by making the information need of underspecified queries explicit. Many LongMemEval queries were already specific, so adding Intent may merely paraphrase or dilute existing signals.

Background provided limited benefit on LiHua-World and the smallest degradation on LongMemEval. It may include evidence-related entities, events, relations, or temporal context, but did not outperform Original query.

Constraints-only expansion did not outperform the original query on either dataset. Answer policies such as "refer to past conversations" may not help identify the gold evidence and can introduce noise as natural-language prefixes. Constraints may instead be useful as filters, reranking criteria, or generation controls.

7.1. Human Analysis

The 419 questions were presented in random order to one annotator. Query support was judged against the original query, whereas evidence discrimination was judged with reference to the gold answer and evidence; question IDs, query types, retrieval scores, and performance labels were hidden. Of the generated IBC outputs, 400 (95.5%) were supported by the original query and contained no contradictions, whereas 349 (83.3%) added no new information that discriminated the gold evidence. The two judgments were not mutually exclusive, and 340 outputs (81.1%) satisfied both. These findings show that even semantically valid expansion does not improve retrieval unless it adds cues that distinguish the evidence.

8. Discussion and Future Work

Our findings concern the uniform natural-language realization of IBC, rather than the utility of the structured representation itself.

LiHua-World evaluates underspecified queries over fixed-length chunks with multiple evidence pieces; LongMemEval tests existing questions over individual turns. The custom QAE design may therefore favor IBC, and its results cannot be generalized directly to long-term conversational memory retrieval as a whole.

This study uses one IBC generator, three dense retrievers, and one IBC generation per query. The consistent dataset-level trend across the three dense retrievers mitigates dependence on a single embedding model, but does not establish generality to other retriever families. Human analysis used one annotator without agreement measurement, and element lengths were not controlled, so ablations reflect both content and input-length differences.

8.1. Potential KG-Based Retrieval

A future KG-based realization could index dialogue history as a user-centric personal memory KG rather than assuming that all relevant memories already exist in a general-purpose KG. User, entity, event, and time-varying state nodes, such as goals, preferences, skills, and experiences, could be connected by typed relations and provenance links to their source dialogue turns. These nodes could optionally be aligned with an external domain KG, although external knowledge would not itself constitute evidence about the user. Query-concept-seeded graph retrieval provides a related precedent [13].

Graph construction and retrieval should remain separate. Only dialogue-grounded facts would be inserted into the personal memory KG; query-only IBC output would instead be compiled into a retrieval plan. Intent would specify target node and relation types, Background would provide optional seeds or priority signals rather than mandatory path constraints, and Constraints would serve as evidence filters, ranking criteria, or answer-generation controls. Retrieved facts would then be mapped back to their provenance-linked dialogue turns.

For the query "What should I do after graduation?" in Figure 1, Intent would prioritize the user's goals, skills, preferences, values, and experiences. A dialogue-grounded graduation event could serve as an optional Background seed, but would not be required for retrieving every relevant fact. Career preferences expressed in other contexts would therefore remain eligible. Moreover, "after graduation" would primarily constrain the time horizon of the recommendation rather than the timestamp of the evidence. This remains a proposal: we did not construct or evaluate a personal memory KG in this study.

9. Conclusion

We proposed an Intent–Background–Constraint structured representation for implicit retrieval requirements in long-term conversational memory. Retrieval improved on LiHua-World QAE, mainly through Intent, but uniform IBC prefixes degraded LongMemEval, showing that semantic validity does not ensure retrieval utility. Future work will selectively transform individual elements into retrieval queries, filters, and ranking constraints.

Declaration on Generative AI

During the preparation of this work, the authors used the OpenAI API with GPT-5.1 to generate IBC representations, Free rewrite queries, and candidate QAE instances from the LiHua-World dialogue logs. The candidate instances were manually checked as described in Section 4.1. OpenAI Codex was used for CEURART formatting assistance and for editing and revising the manuscript text. After using these tools and services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 6769–6781. doi:10.18653/v1/2020.emnlp-main.550.

[2] T. Fan, J. Wang, X. Ren, C. Huang, MiniRAG: A lightweight RAG system with small language models, in: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2026, pp. 37121–37144. doi:10.18653/v1/2026.acl-long.1721.

[3] D. Wu, H. Wang, W. Yu, Y. Zhang, K.-W. Chang, D. Yu, LongMemEval: Benchmarking chat assistants on long-term interactive memory, in: The Thirteenth International Conference on Learning Representations, OpenReview.net, 2025. URL: https://openreview.net/forum?id=pZiyCaVuti.

[4] F. Mo, C. Qu, K. Mao, T. Zhu, Z. Su, K. Huang, J.-Y. Nie, History-aware conversational dense retrieval, in: Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, 2024, pp. 13366–13378. doi:10.18653/v1/2024.findings-acl.792.

[5] A. Elgohary, D. Peskov, J. Boyd-Graber, Can you unpack that? Learning to rewrite questions-in-context, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, pp. 5918–5924. doi:10.18653/v1/D19-1605.

[6] Z. Wu, Y. Luan, H. Rashkin, D. Reitter, H. Hajishirzi, M. Ostendorf, G. S. Tomar, CONQRR: Conversational query rewriting for retrieval with reinforcement learning, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2022, pp. 10000–10014. doi:10.18653/v1/2022.emnlp-main.679.

[7] L. Wang, N. Yang, F. Wei, Query2doc: Query expansion with large language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2023, pp. 9414–9423. doi:10.18653/v1/2023.emnlp-main.585.

[8] S. Yao, P. Huang, Z. Liu, Y. Gu, Y. Yan, S. Yu, G. Yu, ExpandR: Teaching dense retrievers beyond queries with LLM guidance, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2025, pp. 19036–19054. doi:10.18653/v1/2025.emnlp-main.963.

[9] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, J.-Y. Nie, C-Pack: Packed resources for general chinese embeddings, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, 2024, pp. 641–649. doi:10.1145/3626772.3657878.

[10] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.

[11] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, F. Wei, Text embeddings by weakly-supervised contrastive pre-training, arXiv preprint arXiv:2212.03533 (2022). doi:10.48550/arXiv.2212.03533.

[12] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Transactions on Information Systems 20 (2002) 422–446. doi:10.1145/582415.582418.

[13] B. Jiménez Gutiérrez, Y. Shu, Y. Gu, M. Yasunaga, Y. Su, HippoRAG: Neurobiologically inspired long-term memory for large language models, in: Advances in Neural Information Processing Systems 37, 2024. doi:10.52202/079017-1902.