GUIR at SemEval-2026 Task 8: Training-Free Multi-Query Fusion for Robust Conversational Retrieval

Pasha Abrishamchian **Ophir Frieder** **Nazli Goharian**
Georgetown University
Washington, D.C.
{pasha, ophir, nazli}@ir.cs.georgetown.edu

Abstract

We describe our SemEval-2026 Task 8 Subtask A system, which focuses on evaluating and improving the retrieval aspect of multi-turn Retrieval-Augmented Generation (RAG) conversations. We implement a training-free fusion approach that combines three distinct query representations to retrieve documents independently. The results from these three views are pooled and reranked using a MonoT5 cross-encoder. Our findings demonstrate that this fusion approach consistently outperforms single-strategy baselines, revealing that optimal retrieval strategies vary significantly at the query level, and establishing multi-query fusion as a baseline for multi-turn RAG systems.

1 Introduction

The prevalence of chat-based LLM systems has led to an increase in the utilization of Retrieval-Augmented Generation (RAG) methodologies to ensure that LLM-generated responses are trustworthy and grounded in relevant passages (Gao et al., 2023; Arabzadeh et al., 2025). To address the complexities of this environment, we participated in SemEval-2026 Task 8 (MTRAGEval) (Rosenthal et al., 2026b), specifically focusing on Subtask A: Retrieval.

The MTRAG benchmark consists of 110 conversations averaging 7.7 turns across four distinct domains: Government (Govt), Finance (FiQA), Technology (Cloud), and Wikipedia (ClapNQ) (Katsis et al., 2025; Rosenthal et al., 2026a). It evaluates RAG systems using realistic, human-generated conversations and introduces new challenges such as diversified domains, non-standalone information, false starts, and unanswerable queries. These are in addition to well-recognized challenges, including implicit context, co-references, ellipsis, topic switches, and clarifications (Dalton et al., 2020; Adlakha et al., 2022), all of which make conversational queries difficult for standard retrievers to process accurately.

In analyzing these challenges, we identify significant variation at both the query and domain levels, making a single, universal retrieval strategy inherently suboptimal. To address this, we propose a pooling-then-reranking approach (Ju et al., 2023; Rackauckas, 2024). By combining distinct conversational views, our architecture (Section 3) better captures exact lexical nuances, resolved co-references, and historical progression simultaneously. In unofficial, post-submission evaluations, our training-free, fusion pipeline demonstrated competitive performance, achieving scores that place it within approximately 7% of the top-performing systems, as announced by the MTRAGEval organizers. The main contributions described herein are threefold:

1. We demonstrate strategy heterogeneity at the query level for the MTRAG dataset, showing that no single query representation is universally optimal.

2. We present a domain-level analysis revealing how distinct linguistic and technical characteristics dictate retrieval success, specifically highlighting the acute challenges of jargon-heavy domains like FiQA.

3. We propose a multi-query fusion approach that combines different query strategies to exploit the unique benefits each provides, establishing a reliable and domain-robust baseline.

2 Background

Multi-turn conversational search relies heavily on accurate query transformation to bridge the gap between context-dependent dialogues and standalone ad-hoc retrievers. The dominant approach has been Conversational Query Rewriting (CQR),

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2583–2591

which transforms user utterances into context-independent queries (Anantha et al., 2021; Qian and Dou, 2022). With the advent of Large Language Models (LLMs), recent work has shifted toward prompt-based, generative disambiguation (Mo et al., 2023; Ye et al., 2023; Mo et al., 2024). To further align the rewriting process with the retriever's objective, recent advancements have introduced Reinforcement Learning (RL) techniques. Approaches like CONQRR (Wu et al., 2022), ChatR1 (Lupart et al., 2025), and ConvSearch-R1 (Zhu et al., 2025) utilize RL to dynamically optimize query reformulations based on downstream retrieval performance. However, these methods remain computationally expensive and prone to semantic drift, which can strip away the exact lexical matches required in highly specialized domains.

To mitigate the risks of semantic drift introduced by relying on a single rewritten query, the field has increasingly adopted multi-query generation and fusion strategies (Kostric and Balog, 2024). Studies have shown that fusing diverse query representations, such as combining the original user utterance with LLM-expanded queries, and applying cross-encoder reranking provides a more robust retrieval pipeline (Liu and Zhang, 2025; Rackauckas, 2024). Crucially, evidence from recent benchmarks demonstrates that "fusion-before-reranking" consistently outperforms pipelines that attempt to rerank prior to fusion (Chang et al., 2025). Building upon these findings, our methodology bypasses complex RL fine-tuning in favor of a training-free fusion approach, followed by cross-encoder reranking using MonoT5 (Nogueira et al., 2020).

3 System Overview

Our system consists of two primary components: the utilization of distinct query representations following the framework established by the original MTRAG dataset authors (Katsis et al., 2025; Rosenthal et al., 2026a), and a pooling-then-reranking fusion strategy.

3.1 Query Representation Strategies

We extract three distinct representations for every conversational turn, (1) **Last-Turn (LT):** User's raw input from the current turn, preserving original lexical choices and domain-specific terminology that are frequently altered, generalized, or lost entirely during LLM rewriting. This view is particularly effective for self-contained follow-up queries. (2) **Rewrite (RW):** Standalone query generated to resolve conversational dependencies, such as co-references, ellipsis, and implicit topic shifts. For our MTRAG pipeline, we utilized the official Mixtral-based baseline rewrites provided by the organizers (Rosenthal et al., 2026b). (3) **Concatenated Questions (QS):** Concatenate the current user utterance with all previous user questions in the dialogue session, exposing the raw, unaltered historical progression directly to the retrieval backbone without relying on generative intervention.

3.2 Multi-Query Fusion Architecture

For each conversational turn, we perform retrieval independently for each of the three query representations Q_{LT}, Q_{RW}, Q_{QS}. We retrieve the top-k candidates (with $k = 10$) for each representation, denoted as D^r_{LT}, D^r_{RW}, and D^r_{QS}, respectively. To maximize candidate diversity before reranking (Liu and Zhang, 2025; Chang et al., 2025), these candidate sets are then merged into a single pooled set:

$$D^r_{\text{pool}} = D^r_{LT} \cup D^r_{RW} \cup D^r_{QS}. \quad (1)$$

For final ranking, we rerank all documents in D^r_{pool} using a MonoT5 cross-encoder (Nogueira et al., 2020). To isolate the effect of multi-query fusion from reranker design, we use standard `castorini/monot5-{base,large}-msmarco-10k` checkpoints. A recent comprehensive evaluation of 22 reranking methods across 40 variants shows that MonoT5 consistently ranks among the top pointwise rerankers, on both established benchmarks and temporally novel queries (Abdallah et al., 2025), making it a strong, stable choice for this study.

The relevance score for each candidate document $d \in D^r_{\text{pool}}$ is computed using the rewritten query Q_{RW} as input to the reranker since it provides a standalone formulation of the user's information need by resolving co-reference and ellipsis. The model produces logits for the tokens `true` and `false`, and the final score $s(Q_{RW}, d)$ is obtained by applying a softmax over these two logits:

$$s(Q_{RW}, d) = \frac{e^{\text{logit}_{\text{true}}}}{e^{\text{logit}_{\text{true}}} + e^{\text{logit}_{\text{false}}}}. \quad (2)$$

The final ranked list is generated by sorting all candidates in D^r_{pool} in descending order of $s(Q_{RW}, d)$.

4 Experimental Setup

We design our experiments to evaluate both the performance of individual query strategies and the robustness of our multi-query fusion approach. We evaluate our system on the official MTRAGEval Subtask A (Retrieval) evaluation set (Rosenthal et al., 2026b). Following the task guidelines, we use the organizers' provided 512-token passage chunks (with a 100-token overlap). While the official shared task primary evaluation metric is nDCG@5, we expand our reporting to include MRR, nDCG@k, and Recall@k (where $k \in \{1, 3, 5, 10\}$). We compare our official submission against other teams using nDCG@5, but rely on this broader suite of metrics to fully capture early precision and overall retrieval performance across our analyses.

Our experimental methodology is structured into four distinct phases:

Single-Strategy Baselines & Official Submission: To demonstrate strategy heterogeneity at the query level, we first establish baselines by evaluating each query representation (LT, RW, QS) independently using the `.elser-2-elastic` index. We then compare these individual baselines against our official SemEval submission (constrained to a single run by the shared task guidelines) which utilized the multi-query fusion of all three views (LT+RW+QS) followed by cross-encoder reranking with MonoT5-base (Nogueira et al., 2020).

Domain-Level Stratification: To understand how linguistic and technical characteristics dictate retrieval success, we stratify all performance metrics across the four distinct MTRAG domains: Govt, FiQA, Cloud, and ClapNQ. This breakdown allows us to identify structural domain shifts and isolate where specific strategies succeed or fail. We utilize the `.elser-2-elastic` index for these experiments.

Multi-Query Fusion Over Combinations: We evaluate the pooling-then-rerank approach using different combinations of query representations (LT+RW, LT+QS, RW+QS) and compare the performance to the threeway fusion (LT+RW+QS). We utilize the `.elser-2-elastic` index for these experiments and explore the domain-level performance of the different query combinations.

Post-Submission Experiments: We conduct further experiments to test the upper bounds of our architecture. We test the `.elser-1-elastic` index compared to the newer `.elser-2-elastic` variant utilized in our official run. We also scale the reranker to MonoT5-large and compare the performance to MonoT5-base.

5 Results

We first evaluate our official SemEval submission against the shared task baselines. We then present a fine-grained domain analysis, isolate the impact of different query fusions, and explore the upper bounds of our architecture through post-submission experiments on reranker scaling and index variants.

5.1 Official Submission and Baselines

System	nDCG@5
Top Baseline (ELSER + GPT-OSS-20b)	0.4795
Our Submission (LT+RW+QS + MonoT5-Base)	*0.5387*
Best Performing System	0.5776

Table 2: Overall Subtask A retrieval performance compared to official SemEval baselines. We report our primary submission metric (nDCG@5). Full metrics across all cutoff thresholds (including MRR and Recall) are provided in Appendix A.

The results in Table 2 demonstrates that our system significantly outperforms the standard baseline by 12.3% relative (0.5387 vs. 0.4795), validating our hypothesis that diverse candidate pooling mitigates the semantic drift commonly associated with standalone query rewriting. As visualized comprehensively in Appendix A (Figure 2), individual query strategies exhibit high variance across different turns; however, the multi-query fusion consistently smooths out these fluctuations to outperform all single strategies.

Our submission placed 6th among competitors; however, based on the information we currently possess, our score is only 6.7% lower than the top performer. It is worth noting that the top performer utilized a 20B parameter model for rewriting purposes, whereas our system performs highly competitively using the much smaller, organizer-provided Mixtral 8x7B rewrites. As full architectural details of competing submissions are not yet publicly available, we restrict further external comparisons and focus our analysis on our system's internal dynamics.

5.2 Domain-Level Stratification

To better understand our system's behavior, we stratify its performance across the four MTRAG

Domain	MRR	R@1	R@3	R@5	R@10	nDCG@1	nDCG@3	nDCG@5	nDCG@10
ClapNQ (Wiki)	0.7051	0.3777	0.5807	0.6890	0.7751	0.6024	0.5967	0.6366	0.6734
Cloud (Tech)	0.6135	0.2078	0.4373	0.5546	0.6529	0.4884	0.4777	0.5143	0.5556
FiQA (Finance)	0.4550	0.1609	0.2615	0.3635	0.4999	0.3448	0.3021	0.3413	0.4028
Govt	0.6676	0.2533	0.5219	0.6636	0.7290	0.5524	0.5333	0.5903	0.6187

Table 1: Domain-wise performance breakdown of our official submission (fusion of LT, RW, and QS queries reranked with MonoT5-Base). The results highlight the system's strong performance in broad knowledge domains (ClapNQ) versus the challenges posed by dense, specialized terminology (FiQA).

domains in Table 1. We observe a consistent pattern across all evaluated metrics: our submitted system performs exceptionally well in broad, knowledge-intensive domains like ClapNQ (e.g., 0.6366 nDCG@5) and specialized, lexical-heavy domains like Govt (e.g., 0.5903 nDCG@5). In these domains, the combination of raw queries and LLM rewrites successfully bridges both lexical matching and semantic intent.

Conversely, the system universally faces challenges in the FiQA domain (e.g., 0.3413 nDCG@5). This steep drop in performance highlights the acute difficulty of conversational retrieval in highly technical domains characterized by dense financial jargon, acronyms, and implicit reasoning dependencies. In such environments, off-the-shelf retrieval models struggle to align colloquial user questions with formal financial texts. This finding is critical as it suggests that future conversational RAG pipelines cannot rely solely on generic fusion strategies; they must incorporate domain-specific adaptations, such as specialized vocabulary modeling or targeted reasoning steps, to effectively handle complex enterprise sectors.

5.3 Multi-Query Fusion Over Combinations

Table 3 breaks down the nDCG@5 performance for partial fusions (combinations of two query views) versus our full fusion. For the comprehensive set of results across all configurations and metric thresholds, refer to Appendix A.

We observe that strategy dominance varies significantly by domain, reflecting the distinct linguistic characteristics of each dataset. When examining partial fusions, no single pair of queries universally dominates. For instance, in the ClapNQ domain, combining the raw user input with the LLM rewrite (LT+RW) yields the best partial performance (0.6249 nDCG@5). This suggests that for general fact-seeking questions, maintaining the original entity references alongside a decontextualized rewrite provides sufficient semantic coverage. In the highly specialized Govt domain, combining the standalone rewrite with the full conversational history string (RW+QS) proves superior (0.5910 nDCG@5). The Govt domain often features intricate policy keywords where the implicit context built up over multiple turns is crucial; the full query string helps retrieve documents that match the broader conversational intent rather than just the immediate turn.

As shown in Table 3, the single-query baseline (Rewrite Only) consistently underperforms any fusion combination across all domains, which highlights the importance of having multiple query representations, and how standard LLM rewriting is susceptible to omitting subtle nuances necessary for accurate retrieval. This query-level and domain-level heterogeneity underscores the necessity of our full three-way fusion (LT+RW+QS) for a generalizable conversational retrieval system. While a two-view fusion might peak in an isolated scenario depending on the specific nature of the conversation, the three-way fusion effectively smooths out these variances.

5.4 Post-Submission Experiments

Reranker Scaling. Figure 1 demonstrates the impact of scaling from MonoT5-Base (220M parameters) to MonoT5-Large (770M parameters) on the fully fused candidate pool across all domains on the ELSER v2 index. The transition yields consistent improvements across all metrics, pushing our overall nDCG@5 from 0.5387 to 0.5490 (+1.9%). Notably, the improvements are more pronounced at lower cutoff thresholds, indicating enhanced early-rank precision rather than broad recall gains.

The magnitude of these gains varies by domain complexity. In the challenging FiQA domain, our Base reranker struggles to consistently identify the most relevant financial documents. Applying the Large reranker to the exact same fusion candidate pool yields a relative improvement of +8.5% relative. This confirms that while our training-free

Query Combination	ClapNQ	Cloud	FiQA	Govt	Overall
Rewrite Only (No Rerank)	0.5564	0.4207	0.3367	0.4923	0.4626
LT + QS	0.6067	0.4960	**0.3753**	0.5416	0.5170†
LT + RW	0.6249	0.5088	0.3494	0.5670	0.5284†
RW + QS	0.6186	0.5056	0.3400	**0.5910**	0.5319†
LT + RW + QS	**0.6366**	**0.5143**	0.3413	0.5903	**0.5387**†

Table 3: Domain-wise Combinations (nDCG@5) comparing partial and full query fusions under the MonoT5-Base reranker. Results utilize the ELSER v2 index as per our official submission. Bold indicates the best result per column. For the Overall column, statistical significance is computed using a paired t-test ($p < 0.05$). The symbol † denotes a significant improvement over the `Rewrite Only` baseline.

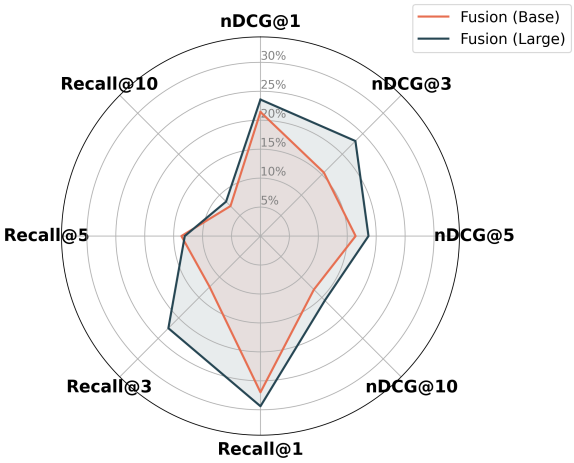

Figure 1: Impact of scaling the reranker capacity from MonoT5-Base to MonoT5-Large on the fully fused (LT+RW+QS) candidate pool across all domains on the ELSER v2 index.

multi-query fusion successfully surfaces relevant documents into the candidate pool, even in complex domains, a higher-capacity cross-encoder is required to accurately distinguish true relevance among dense, domain-specific texts.

Index Variant Analysis. We additionally evaluated our full fusion architecture on the organizers' original `.elser-1-elastic` index, motivated by the fact that the MTRAG corpus and relevance judgments were produced using this index during dataset creation (Katsis et al., 2025). Restricting our comparison to directly equivalent configurations (LT, QS, Rewrite, and LT+RW+QS under both reranker sizes), the results reveal a nuanced picture rather than a simple generational ordering.

At the aggregate level, the v2 index produces marginally stronger individual query baselines and a slightly higher Base fusion score (0.5387 vs. 0.5317 nDCG@5). However, at the domain level, neither index universally dominates. The v2 index holds a clear advantage in the Cloud domain, suggesting it better handles the precise technical terminology of technical documentation. Conversely, the v1 index retains consistent advantages in ClapNQ. Under the Large reranker, both FiQA and Govt are more closely aligned with the v1 encoding space.

Critically, scaling the reranker to MonoT5-Large fully reconciles these index-level differences at the aggregate level, with both v1 and v2 converging to an identical overall nDCG@5 of 0.5490. This demonstrates that our multi-query fusion architecture is robust to index-level variation: by generating a diverse and overlapping candidate pool, the pipeline provides sufficient coverage for a high-capacity cross-encoder to compensate for any retrieval-level discrepancies. The comprehensive per-domain metrics for the ELSER v1 index are provided in Appendix A (Table 4).

6 Conclusion

We described our training-free, multi-query fusion system for SemEval-2026 Task 8 (MTRAGEval), demonstrating that relying on a single query representation is suboptimal for conversational retrieval. By pooling candidates from the user's raw input, conversational history, and an LLM rewrite, our architecture softens domain-level vulnerabilities, particularly in jargon-heavy sectors like Finance. Our approach outperformed the baseline by 12.3% relative nDCG@5, establishing that robust multi-turn RAG requires diverse candidate generation paired with high-capacity cross-encoder reranking.

References

Abdelrahman Abdallah, Bhawna Piryani, Jamshid Mozafari, Mohammed Ali, and Adam Jatowt. 2025. How good are LLM-based rerankers? An empirical analysis of state-of-the-art reranking models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5693–5709, Suzhou, China. Association for Computational Linguistics.

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain Conversational Question Answering with Topic Switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-Domain Question Answering Goes Conversational via Question Rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. Association for Computational Linguistics.

Negar Arabzadeh, Ziheng Chen, Fabio Petroni, Federico Siciliano, Fabrizio Silvestri, and Giovanni Trappolini. 2025. IR-RAG @SIGIR25: The second edition of the workshop on information retrieval's role in RAG systems. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sigir '25, pages 4168–4171, New York, NY, USA. Association for Computing Machinery.

Yu-Cheng Chang, Guan-Wei Yeo, Quah Eugene, Fan-Jie Shih, Yuan-Ching Kuo, Tsung-En Yu, Hung-Chun Hsu, Ming-Feng Tsai, and Chuan-Ju Wang. 2025. CFDA & CLIP at TREC iKAT 2025: Enhancing Personalized Conversational Search via Query Reformulation and Rank Fusion. *Preprint*, arXiv:2509.15588.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC CAsT 2019: The Conversational Assistance Track Overview. *Preprint*, arXiv:2003.13624.

Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2023. *Neural Approaches to Conversational Information Retrieval*, volume 44 of *The Information Retrieval Series*. Springer International Publishing, Cham.

Jia-Huei Ju, Sheng-Chieh Lin, Ming-Feng Tsai, and Chuan-Ju Wang. 2023. Improving Conversational Passage Re-ranking with View Ensemble. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2077–2081, Taipei Taiwan. ACM.

Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. 2025. mtrag: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. *Transactions of the Association for Computational Linguistics*, 13:784–808.

Ivica Kostric and Krisztian Balog. 2024. A Surprisingly Simple yet Effective Multi-Query Rewriting Method for Conversational Passage Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2271–2275, Washington DC USA. ACM.

Lingyuan Liu and Mengxiang Zhang. 2025. Exp4Fuse: A Rank Fusion Framework for Enhanced Sparse Retrieval using Large Language Model-based Query Expansion. *Preprint*, arXiv:2506.04760.

Simon Lupart, Mohammad Aliannejadi, and Evangelos Kanoulas. 2025. ChatR1: Reinforcement Learning for Conversational Reasoning and Retrieval Augmented Question Answering. *Preprint*, arXiv:2510.13312.

Fengran Mo, Abbas Ghaddar, Kelong Mao, Mehdi Rezagholizadeh, Boxing Chen, Qun Liu, and Jian-Yun Nie. 2024. CHIQ: Contextual History Enhancement for Improving Query Rewriting in Conversational Search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2268, Miami, Florida, USA. Association for Computational Linguistics.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023. ConvGQR: Generative Query Reformulation for Conversational Search. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4998–5012, Toronto, Canada. Association for Computational Linguistics.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pre-trained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

Hongjin Qian and Zhicheng Dou. 2022. Explicit Query Rewriting for Conversational Dense Retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4725–4737, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zackary Rackauckas. 2024. RAG-Fusion: A New Take on Retrieval-Augmented Generation. *International Journal on Natural Language Computing*, 13(1):37–47.

Sara Rosenthal, Yannis Katsis, Vraj Shah, Lihong He, Lucian Popa, and Marina Danilevsky. 2026a. Mtragun: A benchmark for open challenges in multi-turn rag conversations.

Sara Rosenthal, Vraj Shah, Yannis Katsis, and Marina Danilevsky. 2026b. Semeval-2026 task 8: Mtrageval: Evaluating multi-turn rag conversations. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California. Association for Computational Linguistics.

Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, Hannaneh Hajishirzi, Mari Ostendorf, and Gaurav Singh Tomar. 2022. CONQRR: Conversational Query Rewriting for Retrieval with Reinforcement Learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10000–10014, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing Conversational Search: Large Language Model-Aided Informative Query Rewriting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5985–6006, Singapore. Association for Computational Linguistics.

Changtai Zhu, Siyin Wang, Ruijun Feng, Kai Song, and Xipeng Qiu. 2025. ConvSearch-R1: Enhancing Query Reformulation for Conversational Search with Reasoning via Reinforcement Learning. *Preprint*, arXiv:2505.15776.

A Comprehensive Reranking Performance Results

This appendix provides the complete set of retrieval and reranking metrics across all evaluated query strategies and domains. Table 5 presents the results using the ELSER v2 index (the backbone of our official submission), while Table 4 details the results using the original ELSER v1 index. Single-query strategies (LT, QS, RW) report without reranking to demonstrate baseline retrieval capacity, while fusion strategies (e.g., LT+QS, LT+RW+QS) report the performance after pooling the top-10 candidates from each view and reranking with MonoT5 (Base and Large).

Extended Analysis of Index Stability and Reranker Capacity

While the main text highlights the broad improvements gained by our fusion architecture, these extended tables provide deeper insights into the specific mechanical behaviors of our pipeline across different conversational environments.

Cross-Index Stability (v1 vs. v2): A key concern in retrieval is the sensitivity of the system to the underlying index. Table 4 shows the multi-query fusion effectively regularizes this discrepancy. When comparing the full fusion (LT+RW+QS) under the MonoT5-Large reranker, the v1 index achieves an overall nDCG@5 of 0.5490, matching the performance of the v2 index. This strongly indicates that pooling diverse candidate sets makes the pipeline significantly more resilient to index-level variations.

The Necessity of High-Capacity Reasoning: Across virtually every domain and fusion combination, scaling the cross-encoder from MonoT5-Base (220M parameters) to MonoT5-Large (770M parameters) yields consistent improvements. Overall, the full three-way fusion on ELSER v2 improves from 0.5387 to 0.5490 nDCG@5. The most dramatic gains are observed in domains with severe lexical mismatch. In the FiQA domain specifically, the Large reranker applied to the LT+QS pool improves nDCG@1 from 0.3966 (Base) to 0.4483 (Large), and the full LT+RW+QS fusion improves nDCG@5 from 0.3413 (Base) to 0.3702 (Large). Notably, LT+QS under the Large reranker achieves the best overall FiQA performance across nearly all metrics, suggesting that for dense financial queries, preserving the raw conversational context alongside the last turn is more beneficial than including a potentially drift-prone LLM rewrite. This confirms that while multi-query fusion successfully surfaces relevant documents into the top-k pool, a higher-capacity reasoning model is strictly required to accurately sort true positive documents among dense, domain-specific texts.

Domain-Specific Limits of Decontextualization: The extended metrics in Table 5 reveal edge cases where the full three-way fusion is not strictly optimal at every metric. In the Cloud domain, combining the Last Turn and Rewrite (LT+RW) under the Base reranker achieves the highest nDCG@1 (0.5), outperforming the full three-way fusion (0.4884 nDCG@1). This suggests that for highly technical troubleshooting queries, RW provides sufficient semantic coverage for top-ranked precision, and the addition of QS does not consistently improve early-rank precision. Conversely, in the Govt domain, the RW+QS combination under the Large reranker achieves the highest nDCG@5 (0.6041), outperforming even the full three-way fusion (0.5976 nDCG@5). This strongly indicates that for intricate official and legal domain discussions, the full conversational history string is the dominant signal of relevance, and the marginal addition of the raw Last Turn (LT) introduces more noise than signal. In the FiQA domain, the LT+QS combination under the Large reranker dominates across all metrics, further reinforcing that domain-specific fusion design may be the most principled path forward for highly specialized enterprise domains.

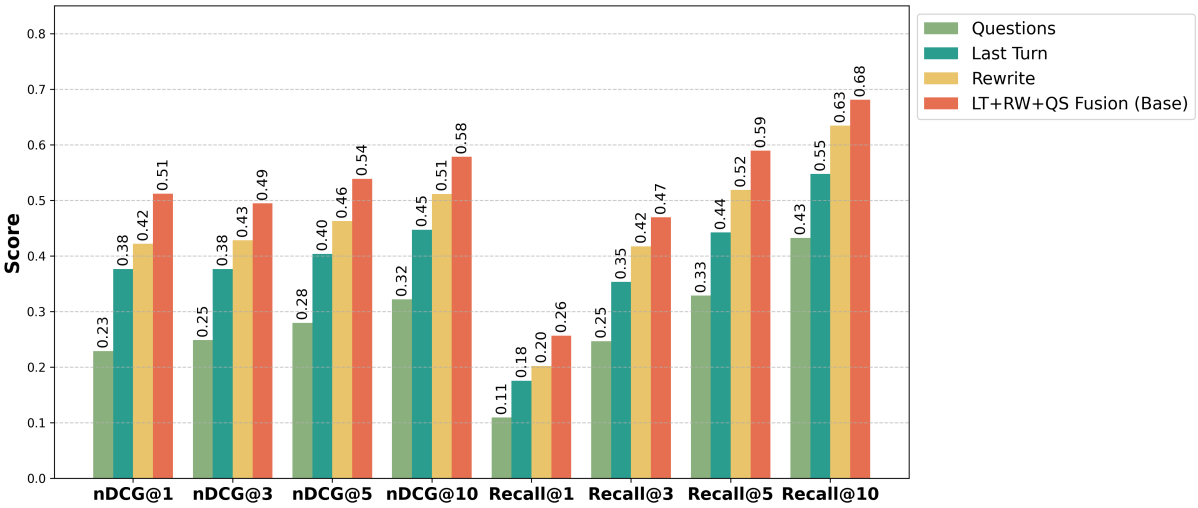

Figure 2: Visual comparison of all query representations and fusion strategies across domains using the ELSER v2 index, demonstrating that while no single strategy dominates universally at the query level, our multi-query fusion establishes the most robust baseline overall.

Domain	Strategy	MRR	R@1	R@3	R@5	R@10	nDCG@1	nDCG@3	nDCG@5	nDCG@10
Overall	LT	0.4856	0.1787	0.3440	0.4292	0.5389	0.3916	0.3702	0.3978	0.4440
	QS	0.3152	0.1105	0.2174	0.3010	0.4089	0.2349	0.2237	0.2564	0.3009
	Rewrite	0.5615	0.2199	0.4042	0.4962	0.6349	0.4518	0.4256	0.4578	0.5162
	LT+RW+QS (Base)	0.6170†	0.2487	0.4685†	0.5828†	0.6742†	0.4970	0.4899†	0.5317†	0.5706†
	LT+RW+QS (Large)	**0.6445**†‡	**0.2710**†	**0.4936**†‡	**0.5862**†	**0.6822**†	**0.5361**†	**0.5185**†‡	**0.5490**†	**0.5895**†‡
ClapNQ	LT	0.5069	0.2518	0.3988	0.4795	0.5841	0.4337	0.4093	0.4381	0.4802
	QS	0.3932	0.1695	0.3163	0.4070	0.5408	0.3012	0.3057	0.3413	0.3968
	Rewrite	0.6288	0.3100	0.5173	0.5984	0.7544	0.5301	0.5133	0.5439	0.6069
	LT+RW+QS (Base)	0.7138	**0.3928**	0.6054	**0.7161**	**0.7968**	**0.6145**	0.6119	**0.6551**	**0.6885**
	LT+RW+QS (Large)	**0.7139**	0.3767	**0.6227**	0.7151	0.7938	**0.6145**	**0.6208**	0.6517	0.6841
Cloud	LT	0.4893	0.1479	0.2960	0.4114	0.5172	0.3837	0.3489	0.3864	0.4307
	QS	0.3399	0.1139	0.2115	0.2816	0.3672	0.2791	0.2424	0.2630	0.2998
	Rewrite	0.5265	0.1865	0.3370	0.4255	0.5633	0.4070	0.3784	0.4098	0.4687
	LT+RW+QS (Base)	0.5635	0.1807	0.3909	**0.5167**	0.6048	0.4302	0.4297	0.4708	0.5086
	LT+RW+QS (Large)	**0.5804**	**0.2145**	**0.4171**	0.5109	**0.6055**	**0.4535**	**0.4541**	**0.4823**	**0.5242**
FiQA	LT	0.4442	0.1451	0.2399	0.3190	0.4626	0.3621	0.2799	0.3043	0.3640
	QS	0.2173	0.0589	0.1178	0.1681	0.2385	0.1552	0.1302	0.1471	0.1737
	Rewrite	**0.5140**	**0.1796**	0.2845	0.3764	0.4669	**0.4483**	0.3291	0.3597	0.3969
	LT+RW+QS (Base)	0.4607	0.1537	0.2572	0.3563	0.4669	0.3621	0.3051	0.3396	0.3878
	LT+RW+QS (Large)	0.5105	0.1782	**0.3032**	**0.3822**	**0.5086**	0.4138	**0.3515**	**0.3794**	**0.4302**
Govt	LT	0.4885	0.1648	0.3976	0.4648	0.5632	0.3810	0.4066	0.4270	0.4703
	QS	0.2873	0.0897	0.1992	0.3064	0.4330	0.1905	0.1950	0.2442	0.2964
	Rewrite	0.5633	0.1981	0.4360	0.5395	0.6919	0.4286	0.4481	0.4834	0.5493
	LT+RW+QS (Base)	0.6706	0.2430	0.5405	0.6568	0.7487	0.5333	0.5448	0.5903	0.6291
	LT+RW+QS (Large)	**0.7161**	**0.2849**	**0.5595**	**0.6587**	**0.7527**	**0.6095**	**0.5825**	**0.6161**	**0.6563**

Table 4: Comprehensive performance metrics using the original `.elser-1-elastic` index. The table reports MRR, Recall, and nDCG across multiple thresholds for both independent query representations and multi-query fusion pipelines. Bold indicates the best result per metric. Statistical significance on the Overall results is computed using a paired t-test ($p < 0.05$). The symbol † denotes a significant improvement over the `Rewrite` baseline, and ‡ denotes a significant improvement of the Large reranker over the Base reranker.

Domain	Strategy	MRR	R@1	R@3	R@5	R@10	nDCG@1	nDCG@3	nDCG@5	nDCG@10
	LT	0.4859	0.1754	0.3531	0.4420	0.5472	0.3765	0.3763	0. 4035	0.4470
	QS	0.3332	0.1092	0.2465	0.3284	0.4324	0.2289	0.2485	0. 2797	0.3218
	Rewrite	0.5504	0.2020	0.4174	0.5185	0.6343	0.4217	0.4280	0.4626	0.5115
	LT+QS (Base)	0.6100†	0.2419†	0.4519	0.5660†	0.6460	0.4940†	0.4753†	0.5170†	0.5520†
	LT+QS (Large)	0.6349†	0.2571†	0.4772†	0.5624†	0.6513	**0.5241**†	0.5016†	0.5288†	0.5669†
Overall	LT+RW (Base)	0.6133†	0.2510†	0.4574†	0.5774†	0.6608†	0.5030†	0.4844†	0.5284†	0.5646†
	LT+RW (Large)	0.6258†	0.2593†	0.4990†	0.5745†	0.6626†	0.5120†	0.5172†	0.5390†	0.5759†
	RW+QS (Base)	0.6167†	0.2536†	0.4689†	0.5811†	0.6746†	0.5030†	0.4905†	0.5319†	0.5714†
	RW+QS (Large)	0.6358†	**0.2629**†	0.5033†	0.5785†	0.6689†	0.5181†	0.5206†	0.5435†	0.5822†
	LT+RW+QS (Base)	0.6258†	0.2565†	0.4692†	**0.5893**†	0.6808†	0.5120†	0.4944†	0.5387†	0.5783†
	LT+RW+QS (Large)	**0.6412**†	0.2613†	**0.5113**†	0.5863†	**0.6876**†‡	0.5211†	**0.5273**†	**0.5490**†	**0.5920**†‡
	LT	0.4623	0.2127	0.4022	0.4839	0.5936	0.3494	0.3912	0. 4198	0.4641
	QS	0.3679	0.1384	0.3251	0.4098	0.5649	0.2289	0.3048	0. 3377	0.3971
	Rewrite	0.6188	0.3010	0.5317	0.6331	0.7402	0.5060	0.5192	0.5564	0.6020
	LT+QS (Base)	0.6827	0.3556	0.5687	0.6604	0.7235	0.5783	0. 5743	0.6067	0.6353
	LT+QS (Large)	0.7009	0.3586	0.5677	0.6538	0.7355	0.6145	0. 5791	0.6093	0.6445
ClapNQ	LT+RW (Base)	0.6919	0.3657	0.5687	0.6769	0.7576	0.5904	0. 5846	0.6249	0.6599
	LT+RW (Large)	**0.7134**	0.3767	0.5992	0.6853	0.7667	**0.6265**	0.6120	**0.6389**	0.6728
	RW+QS (Base)	0.6674	0.3663	0.5626	0.6775	0.7546	0.5663	0. 5735	0.6186	0.6516
	RW+QS (Large)	0.7021	0.3617	0.5968	0.6769	0.7576	0.6024	0. 5992	0.6262	0.6590
	LT+RW+QS (Base)	0.7051	**0.3777**	0.5807	**0.6890**	0.7751	0.6024	0.5967	0.6366	0.6734
	LT+RW+QS (Large)	0.7130	0.3647	**0.6112**	0.6853	**0. 7781**	0.6145	**0.6152**	0.6371	**0.6750**
	LT	0.4797	0.1511	0.3503	0.4300	0.5409	0.3605	0.3852	0. 3998	0.4458
	QS	0.3563	0.1038	0.2316	0.3352	0.4153	0.2558	0.2599	0. 2930	0.3241
	Rewrite	0.4984	0.1540	0.3743	0.4739	0.5978	0.3488	0.3881	0.4207	0.4752
	LT+QS (Base)	0.5928	0.1962	0.4218	0.5407	0.6343	0.4651	0. 4605	0.4960	0.5346
	LT+QS (Large)	0.5969	0.2203	0.4585	0.5395	0.6307	0.4651	0. 4849	0.5072	0.5465
Cloud	LT+RW (Base)	0.6135	0.2136	0.4257	0.5395	0.6292	**0. 5000**	0.4730	0.5088	0.5461
	LT+RW (Large)	0.6122	**0.2320**	0.4726	0.5504	0.6310	0. 4884	0.5008	0.5222	0.5583
	RW+QS (Base)	**0.6151**	0.2078	0.4234	0.5416	0.6467	0. 4884	0.4669	0.5056	0.5485
	RW+QS (Large)	0.6027	0.2223	0.4625	0.5446	0.6409	0.4651	0. 4881	0.5124	0.5548
	LT+RW+QS (Base)	0.6135	0.2078	0.4373	0.5546	0.6529	0.4884	0.4777	0.5143	0.5556
	LT+RW+QS (Large)	0.6125	0.2262	**0.4833**	**0.5576**	**0.6570**	0.4767	**0.5054**	**0.5253**	**0. 5700**
	LT	0.4453	0.1365	0.2148	0.3190	0.4411	0.3448	0.2728	0. 3071	0.3556
	QS	0.2667	0.0776	0.1408	0.1925	0.2744	0.2069	0.1650	0. 1826	0.2162
	Rewrite	0.4773	0.1408	0.2909	0.3506	0.4539	0.3621	0.3255	0.3367	0.3786
	LT+QS (Base)	0.4981	0.1839	0.2816	0.4023	0.5101	0.3966	0. 3270	0.3753	0.4232
	LT+QS (Large)	**0.5561**	**0.1925**	**0.3240**	**0.4059**	**0.5402**	**0.4483**	**0.3774**	**0.4011**	**0.4552**
FiQA	LT+RW (Base)	0.4515	0.1609	0.2471	0.3865	0.4941	0.3448	0. 2928	0.3494	0.3991
	LT+RW (Large)	0.4859	0.1595	0.3182	0.3671	0.5027	0.3621	0. 3541	0.3623	0.4177
	RW+QS (Base)	0.4596	0.1609	0.2802	0.3549	0.4855	0.3448	0. 3152	0.3400	0.3966
	RW+QS (Large)	0.4921	0.1609	0.3154	0.3642	0.4826	0.3621	0. 3529	0.3618	0.4123
	LT+RW+QS (Base)	0.4550	0.1609	0.2615	0.3635	0.4999	0.3448	0.3021	0.3413	0.4028
	LT+RW+QS (Large)	0.5050	0.1652	**0.3240**	0.3743	0.5315	0.3793	0.3626	0.3702	0.4348
	LT	0.5321	0.1873	0.3929	0.4867	0.5744	0.4286	0.4146	0. 4469	0.4851
	QS	0.3237	0.1079	0.2551	0.3335	0.4290	0.2190	0.2406	0. 2765	0.3186
	Rewrite	0.5794	0.1968	0.4321	0.5570	0.6802	0.4476	0.4451	0.4923	0.5430
	LT+QS (Base)	0.6285	0.2216	0.4783	0.6025	0.6694	0.5048	0. 4910	0.5416	0.5715
	LT+QS (Large)	0.6573	0.2425	0.5056	0.5954	0.6630	0.5429	0. 5226	0.5533	0.5838
Govt	LT+RW (Base)	0.6403	0.2406	0.5116	0.6351	0.7021	0.5238	0. 5202	0.5670	0.5957
	LT+RW (Large)	0.6451	0.2441	0.5413	0.6211	0.6944	0.5238	0. 5456	0.5713	0.6011
	RW+QS (Base)	0.6649	0.2533	0.5362	0.6621	**0.7386**	0. 5524	0.5410	0.5910	0.6234
	RW+QS (Large)	**0.6898**	**0.2743**	**0.5667**	0. 6470	0.7246	**0.5809**	**0.5777**	**0.6041**	**0.6378**
	LT+RW+QS (Base)	0.6676	0.2533	0.5219	**0.6636**	0.7290	0.5524	0.5333	0.5903	0.6187
	LT+RW+QS (Large)	0.6833	0.2616	0.5587	0.6486	0.7275	0.5619	0.5668	0.5976	0.6312

Table 5: Comprehensive per-domain and overall retrieval performance for all individual query representations (no reranking) and fusion strategies (MonoT5-Base and MonoT5-Large reranker) evaluated on the ELSER v2 index. Single-query strategies are reported without reranking. Bold indicates the best result per domain per metric. For the Overall metrics, statistical significance is computed using a paired t-test ($p < 0.05$). The symbol † denotes a significant improvement over the `Rewrite` baseline, and ‡ denotes a significant improvement of the full 3-way fusion over the best 2-way fusion.