

Caraman at SemEval-2026 Task 8: Three-Stage Multi-Turn Retrieval with Query Rewriting, Hybrid Search, and Cross-Encoder Reranking

David-Maximilian Caraman

Babeş-Bolyai University
Cluj-Napoca, Romania
david.caraman@stud.ubbcluj.ro

Gheorghe Cosmin Silaghi

Babeş-Bolyai University
Cluj-Napoca, Romania
gheorghe.silaghi@ubbcluj.ro

Abstract

We describe our system for SemEval-2026 Task 8 (MTRAGEval), participating in Task A (Retrieval) across four English-language domains. Our approach employs a three-stage pipeline: (1) query rewriting via a LoRA-fine-tuned Qwen 2.5 7B model that transforms context-dependent follow-up questions into standalone queries, (2) hybrid BM25 and dense retrieval combined through Reciprocal Rank Fusion, and (3) cross-encoder reranking with BGE-reranker-v2-m3. On the official test set, the system achieves nDCG@5 of 0.531, ranking 8th out of 38 participating systems and 10.7% above the organizer baseline. Development comparisons reveal that domain-specific temperature tuning for query generation, where technical domains benefit from deterministic decoding and general domains from controlled randomness, provides consistent gains, while more complex strategies such as domain-aware prompting and multi-query expansion degrade performance.

1 Introduction

Multi-turn conversational retrieval poses a fundamental challenge for information retrieval systems: users formulate follow-up questions that depend on prior conversational context through pronouns (“*How much does it cost?*”), ellipsis, and implicit topic continuations. A retrieval system that processes only the latest user utterance loses critical context leads to severe performance degradation in later conversation turns (Katsis et al., 2025).

SemEval-2026 Task 8 (Rosenthal et al., 2026b) provides a rigorous benchmark for this problem, spanning four English domains with 777 retrieval queries drawn from 110 multi-turn human conversations. Task A requires participants to produce passages that are relevant to the user’s final question, while evaluation being conducted only on the subset of answerable questions.

We address this challenge with a three-stage pipeline where each component targets a distinct failure mode: (i) **query rewriting** resolves conversational dependencies, (ii) **hybrid retrieval** captures both lexical and semantic matches, and (iii) **cross-encoder reranking** refines the final ranking through fine-grained query-passage interaction. The query rewriter is a Qwen 2.5 7B Instruct model (Yang et al., 2024) fine-tuned with LoRA (Hu et al., 2022) on the MTRAGEval gold rewrites, trained entirely on Apple Silicon using the MLX framework.¹

Development comparison studies presented in this paper surface instructive negative results: domain-aware prompting and multi-query expansion both *degraded* performance relative to simple rewriting; fine-tuning the cross-encoder reranker on task-specific data yielded only marginal gains; and larger candidate pools for reranking *decreased* quality, a counterintuitive finding we analyze in detail. Our code² and fine-tuned model³ are publicly available.

2 Background

2.1 Task Description

MTRAGEval (Rosenthal et al., 2026b), built on the MTRAG-UN benchmark (Rosenthal et al., 2026a), evaluates retrieval-augmented generation in multi-turn conversational settings. Task A (Retrieval) requires systems to return a ranked list of 10 passages per query from domain-specific corpora totaling 366,479 passages across 78,170 documents in four English domains: **ClapNQ** (Wikipedia, 183K passages), **Cloud** (IBM technical documentation, 72K), **FiQA** (financial forum discussions, 61K), and **Govt** (U.S. government policy documents,

¹<https://github.com/ml-explore/mlx>

²<https://github.com/davidcaraman/semEval2026-mtrag-retrieval>

³<https://huggingface.co/caraman/Qwen2.5-7B-mtrag-query-rewriter-final>

50K). Each corpus uses 512-token passages with 100-token overlap. The primary evaluation metric is nDCG@5 (Järvelin and Kekäläinen, 2002), with nDCG@10 and Recall@10 as secondary measures.

Relevance judgments are derived from human annotations. Following the MTRAG-UN taxonomy (Rosenthal et al., 2026a), the benchmark contains a substantial fraction of unanswerable and under-specified queries; these remain undisclosed in the test set and are excluded from scoring.

2.2 Related Work

Our system builds on three lines of work. **Conversational query rewriting** transforms context-dependent questions into standalone queries suitable for traditional retrievers (Vakulenko et al., 2021); we extend this with domain-specific temperature control for the rewriter’s generation. **Hybrid retrieval** combines lexical matching (BM25; Robertson and Zaragoza, 2009) with dense semantic search (Reimers and Gurevych, 2019) through fusion methods such as Reciprocal Rank Fusion (Cormack et al., 2009). **Cross-encoder reranking** (Nogueira and Cho, 2019) jointly encodes query-passage pairs for fine-grained relevance scoring, consistently outperforming bi-encoder approaches at the cost of higher latency.

3 System Overview

Our pipeline processes each conversational query through three sequential stages. We describe each stage’s design, the experimental evidence motivating it, and the key hyperparameters needed for replication.

3.1 Stage 1: Query Rewriting

Multi-turn queries frequently contain unresolved references that make them unintelligible to a retrieval system in isolation. For instance, given a conversation about IBM Cloud where the user asks “How much does it cost?”, the pronoun “it” must be resolved to produce the standalone query “What is the pricing for IBM Cloud services?”.

This stage takes the base Qwen 2.5 7B Instruct model, fine-tunes it with LoRA on gold rewrites from the MTRAGEval training set, and applies a fixed prompt template (Appendix A) that provides conversation history and instructs the model to produce a standalone query. At inference time, each query is rewritten with a domain-specific temperature identified through a systematic sweep on the

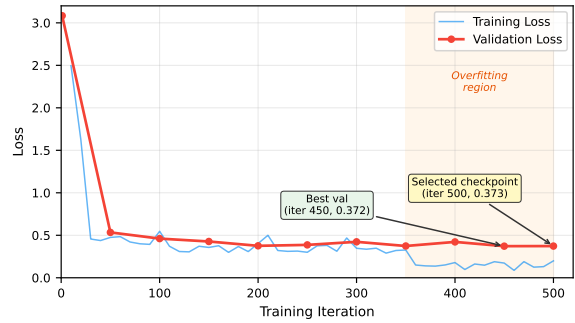


Figure 1: LoRA fine-tuning loss curve for the query rewriter.

holdout set (Section 5.1), producing one rewritten query per domain corpus.

Model and Training. We fine-tune Qwen 2.5 7B Instruct (Yang et al., 2024) using LoRA (Hu et al., 2022) with the parameters presented in Appendix B. These hyperparameters were selected through systematic experimentation over LoRA dropout (0.05–0.15), learning rate (5×10^{-6} – 1×10^{-5}), and the number of adapted layers (16–28), with the final configuration chosen based on holdout validation loss. Training uses the MLX framework on an Apple M4 Max (128 GB unified memory) with AdamW (lr= 10^{-5} , weight decay 0.01), effective batch size 16 (micro-batch 2, gradient accumulation 8), and gradient checkpointing for memory efficiency. The model trains for 500 iterations on 699 query rewriting examples derived from MTRAGEval gold rewrites, with 78 examples held out for validation.

Checkpoint Selection. Figure 1 shows the training dynamics. Validation loss decreases rapidly in the first 200 iterations ($3.084 \rightarrow 0.376$), then plateaus around 0.37–0.42. After iteration 350, training loss drops sharply to ~ 0.1 while validation loss oscillates, indicating the onset of overfitting. We select the final checkpoint at iteration 500 (validation loss 0.373), which is within 0.001 of the best validation loss (0.372 at iteration 450), confirming that the model has converged without significant overfitting despite the low training loss.

The prompt template provides up to 10 turns of conversation history followed by the current question, instructing the model to resolve pronouns, include necessary context, and produce a concise, search-friendly standalone query (see Appendix A for the full prompt).

Domain-Specific Temperature. Because optimal rewriting behavior varies across domains, our system generates separate rewritten queries for each domain using domain-specific temperatures, producing four queries per conversation turn, one per domain corpus. We conducted a systematic temperature sweep (0.0–1.0) on the development holdout to identify the optimal temperature per domain; the full results and analysis are presented in Section 5.1.

First-Turn Optimization. First-turn queries have no conversational history and are therefore already standalone. We validate empirically on 19 first-turn holdout queries that skipping the rewriter produces identical nDCG@5 (0.381) while reducing inference time by 29% (32 s vs. 45 s per domain).

3.2 Stage 2: Hybrid Retrieval

No single retrieval paradigm dominates across all domains: lexical matching captures exact technical terms critical for Cloud, while semantic similarity handles paraphrases prevalent in ClapNQ. We combine both approaches through score-level fusion.

BM25 (Lexical). We use BM25S (Lù, 2024; Robertson and Zaragoza, 2009) with English stop-word removal and no stemming. Omitting stemming preserves exact technical terms (e.g., “*Kubernetes*”, “*401(k)*”) that stemming would corrupt. Each query retrieves 50 candidates.

Dense (Semantic). Queries and passages are encoded with BGE-base-en-v1.5 (Chen et al., 2024) (768 dimensions) and searched against FAISS IndexFlatIP indices (Johnson et al., 2021) with L2-normalized embeddings (cosine similarity). Each query retrieves 50 candidates.

Reciprocal Rank Fusion. BM25 and dense result lists are merged using RRF (Cormack et al., 2009):

$$\text{score}_{\text{RRF}}(d) = \sum_{r \in \{B, D\}} \frac{1}{k + \text{rank}_r(d)} \quad (1)$$

where $k=60$, B is the BM25 ranking, and D is the dense ranking. RRF requires no learned weights and produces a single fused candidate list. Documents appearing in both lists receive contributions from both rankings, naturally boosting passages with cross-modal agreement.

| Reranker | Params | nDCG@5 |
|---------------------------|-------------|--------------|
| BGE-reranker-base | 110M | 0.307 |
| ms-marco-MiniLM-L-12-v2 | 33M | 0.375 |
| IBM Granite R2 | 149M | 0.391 |
| BGE-reranker-v2-m3 | 568M | 0.416 |

Table 1: Cross-encoder reranker comparison at $k=50$ candidates (nDCG@5, no query rewriting).

| Reranker | $k=30$ | $k=50$ | $k=100$ | $k=250$ | $k=500$ |
|-----------------|--------|-------------|---------|---------|---------|
| BGE-v2-m3 | .411 | .416 | .400 | .390 | .388 |
| IBM Granite R2 | .399 | .391 | .392 | .385 | .380 |
| ms-marco-MiniLM | .376 | .375 | .373 | .369 | .366 |
| BGE-base | .321 | .307 | .290 | .266 | .251 |

Table 2: Effect of candidate pool size on reranking quality (nDCG@5). Three of four models peak at $k=30$, but the selected model (BGE-v2-m3) peaks at $k=50$.

3.3 Stage 3: Cross-Encoder Reranking

The fused candidate list is reranked by BGE-reranker-v2-m3 (Chen et al., 2024), a 568M-parameter cross-encoder that jointly encodes each query-passage pair. Table 1 compares four reranker models at $k=50$ candidates. BGE-reranker-v2-m3 (nDCG@5 0.416) substantially outperforms both ms-marco-MiniLM-L-12-v2 (0.375) and BGE-reranker-base (0.307). We also evaluated IBM Granite Reranker R2 (Granite Team, IBM Research AI, 2025), which achieved nDCG@5 of 0.391; we selected BGE-v2-m3 for the final system as it achieved the highest score across all candidate pool sizes (Table 2).

Candidate Pool Size. Reranking candidates (k) involve a trade-off between recall and noise from borderline passages. Our sweep over $k \in \{30, 50, 100, 250, 500\}$ with four rerankers (Table 2) shows that all models degrade from their peak to $k=500$, but the optimal k is model-dependent: three of four rerankers peak at $k=30$, yet BGE-v2-m3 achieves its best nDCG@5 at $k=50$ (0.416 vs. 0.411 at $k=30$), indicating that stronger cross-encoders can exploit a moderately larger pool before noise dominates. We adopt $k=50$ for the final system.

We also evaluated fine-tuning BGE-reranker-v2-m3 on task-specific hard negatives (passages from ranks 20–100). The gain was marginal: nDCG@5 improved from 0.396 to 0.402 (+1.5%), while Recall@10 actually declined from 0.544 to 0.526. The pre-trained model is already well-calibrated for this mixed-domain distribution.

4 Experimental Setup

Data. We split the 110 human conversations from the MTRAGEval training set into 88 training conversations (613 queries) and 22 holdout conversations (164 queries: ClapNQ 27, Cloud 47, Govt 62, FiQA 28). All development results reported in this paper are on the holdout set. Query rewriting training data comprises 699 training and 78 validation examples extracted from gold rewrites. For the final submission, the query rewriter was retrained on all 777 examples (training and holdout combined) using the same hyperparameters selected during development, since the holdout set was no longer needed for model selection. Official competition results are on the separate test set released by the organizers.

Metrics. We report nDCG@5 (Järvelin and Kekäläinen, 2002) as the primary metric (matching the official evaluation), with nDCG@10 and Recall@10 as secondary measures, computed using `pytrec_eval` (Van Gysel and de Rijke, 2018).

Infrastructure. All experiments run on an Apple M4 Max with 128 GB unified memory. LoRA training uses the MLX framework;⁴ retrieval and reranking use PyTorch with MPS acceleration.⁵

5 Results and Analysis

5.1 Query Rewriting Analysis

To identify the optimal generation temperature for each domain, we performed a systematic end-to-end sweep over seven temperature values (0.0, 0.1, 0.2, 0.3, 0.5, 0.7, 1.0) on the 164-query development holdout set. For each temperature, we ran the full pipeline (rewriting → hybrid retrieval → reranking) and measured nDCG@5, ensuring that the selected temperatures optimize the final retrieval quality rather than an intermediate proxy. Table 3 presents the results.

Table 4 provides the full per-domain breakdown at each domain’s optimal temperature on the development holdout.

Query rewriting at the best uniform temperature ($t=0.2$) improves nDCG@5 from 0.371 (no rewriting) to 0.422, a **13.7%** relative gain, confirming that resolving conversational dependencies is the

| Temp | Overall | ClapNQ | Cloud | FiQA | Govt |
|------------|-------------|-------------|-------------|-------------|-------------|
| None | .371 | .535 | .432 | .275 | .296 |
| 0.0 | .416 | .550 | .473 | .321 | .358 |
| 0.1 | .408 | .542 | .472 | .290 | .354 |
| 0.2 | .422 | .563 | .468 | .321 | .371 |
| 0.3 | .416 | .544 | .441 | .346 | .373 |
| 0.5 | .403 | .524 | .463 | .277 | .362 |
| 0.7 | .390 | .527 | .438 | .280 | .343 |
| 1.0 | .380 | .522 | .440 | .233 | .338 |

Table 3: Temperature sweep for query rewriting (nDCG@5) on the development holdout. “None” denotes the no-rewriting baseline using last-turn queries. Bold indicates best per column.

| Domain | Temp | nDCG@5 | nDCG@10 | Recall@10 |
|---------------------|------|--------|---------|-----------|
| ClapNQ | 0.2 | 0.563 | 0.593 | 0.710 |
| Cloud | 0.0 | 0.473 | 0.518 | 0.599 |
| Govt | 0.3 | 0.373 | 0.438 | 0.551 |
| FiQA | 0.3 | 0.346 | 0.367 | 0.439 |
| Overall ($t=0.2$) | 0.2 | 0.422 | 0.467 | 0.569 |

Table 4: Per-domain results at domain-optimal temperatures on the development holdout (164 queries).

single most impactful intervention for multi-turn retrieval.

The optimal temperature varies substantially across domains. **Cloud** (technical documentation) achieves its best result at $t=0.0$ (0.473): its precise technical vocabulary (e.g., “*Kubernetes*”, “*VPC peering*”) means any generation randomness risks substituting incorrect technical terms, corrupting the lexical signal that BM25 relies on. **ClapNQ** (Wikipedia) peaks at $t=0.2$ (0.563), reflecting its well-structured language that benefits from minimal reformulation diversity. **FiQA** (financial forums) and **Govt** (government documents) both prefer $t=0.3$ (0.346 and 0.373), as their more ambiguous query patterns, including informal forum language in FiQA and policy-specific terminology in Govt, benefit from slightly more exploratory rewriting.

Performance degrades monotonically above $t=0.3$ across all domains, with $t=1.0$ reducing overall nDCG@5 to 0.380, barely above the no-rewriting baseline of 0.371. FiQA is particularly sensitive: at $t=1.0$, its nDCG@5 drops to 0.233, *below* the no-rewriting baseline (0.275), as high randomness corrupts its already-ambiguous financial jargon. This observed correlation between domain formality and optimal temperature, where technical domains favor deterministic generation while informal domains benefit from controlled ran-

⁴mlx v0.12+, mlx-lm v0.12+

⁵Key libraries: bm25s v0.2+, sentence-transformers v2.2.2+, faiss-cpu v1.7.4+, FlagEmbedding v1.2+, transformers v4.36+.

| Strategy | nDCG@5 | nDCG@10 |
|-------------------------------------|--------------|--------------|
| No rewriting (baseline) | 0.371 | 0.412 |
| Simple rewriting ($t=0.2$) | 0.422 | 0.467 |
| Domain-aware prompting | 0.350 | 0.399 |
| Multi-query expansion | 0.350 | 0.395 |

Table 5: Query formulation strategy comparison. Simple rewriting substantially outperforms more complex alternatives. Both domain-aware and multi-query strategies underperform even the no-rewriting baseline.

domness, confirms that query rewriting demands high precision and motivates the per-domain temperature configuration adopted in our final system.

5.2 Comparison Studies

Query Strategy Comparison. We tested two alternatives to simple rewriting. **Domain-aware prompting** injects domain metadata into the rewriter prompt (e.g., “*This is a technical cloud computing query*”). This degraded nDCG@5 to 0.350, *below* even the no-rewriting baseline of 0.371. The domain context caused the model to over-specialize queries, narrowing retrieval scope and introducing domain-specific jargon not present in the original question. **Multi-query expansion** generates three query variants and merges their retrieved results. This also scored 0.350, as the additional queries diluted the signal from the primary rewrite with noisy alternatives.

Table 5 provides the full comparison. Both results demonstrate that added complexity in query formulation does not compensate for a well-tuned simple rewriter.

5.3 Error Analysis

Manual inspection of failure cases reveals three dominant error patterns. First, **unanswerable queries** (~25% of the dataset) produce false positives when the system retrieves topically related but non-relevant passages, a limitation inherent to any retrieval-only system without answerability prediction. Second, **long conversation histories** (10+ turns) may exceed the rewriter’s context window (2048 tokens) in production settings; while the current dataset’s conversations remain within this limit, longer real-world interactions would require truncation of early turns that may contain critical referents. Third, FiQA is consistently the most challenging domain (nDCG@5 0.346 at best), attributable to its informal forum language, domain-specific financial jargon (“*401k rollover*”, “*FIRE*

movement”), and abbreviations that create vocabulary mismatch for both lexical and semantic retrievers.

The substantial gap between development hold-out performance (nDCG@5 0.422 at best uniform temperature) and official test set performance (0.531) suggests that domain-specific temperature tuning and the final pipeline configuration generalize well to unseen conversational patterns, and that our holdout set, which comprises only 22 conversations, may have been a conservative estimate of system capability.

6 Conclusion

We presented a three-stage retrieval pipeline for multi-turn conversational search that ranks 8th out of 38 systems on the MTRAGEval benchmark of SemEval 2026 Task 8 - Retrieval (Rosenthal et al., 2026b) with nDCG@5 of 0.531. Each component (query rewriting, hybrid retrieval, and cross-encoder reranking) is justified through systematic comparison studies.

Our analysis surfaces two key insights: domain-specific temperature tuning for query generation yields meaningful gains by adapting to domain vocabulary characteristics, and larger reranking candidate pools counter-intuitively degrade quality. Negative results with domain-aware prompting, multi-query expansion, and reranker fine-tuning reinforce that simplicity outperforms complexity when the base components are well-tuned.

Current limitations include English-only support, consumer hardware constraints that limited model sizes, and unweighted rank fusion. Future work could explore larger rewriter models (14B+), learned fusion weights, ensemble reranking, cross-lingual extension, and a systematic study of the relationship between domain formality and optimal generation temperature for query rewriting.

Acknowledgments

We thank the MTRAGEval organizers for creating a challenging and well-designed benchmark for multi-turn retrieval evaluation.

References

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In

- Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. [Reciprocal rank fusion outperforms Condorcet and individual rank learning methods](#). In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759.
- Granite Team, IBM Research AI. 2025. [Granite embedding R2 models](#). *arXiv preprint arXiv:2508.21085*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Kalervo Järvelin and Jaana Kekäläinen. 2002. [Cumulated gain-based evaluation of IR techniques](#). *ACM Transactions on Information Systems*, 20(4):422–446.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. [Billion-scale similarity search with GPUs](#). *IEEE Transactions on Big Data*, 7(3):535–547.
- 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.
- Xing Han Lù. 2024. [BM25S: Orders of magnitude faster lexical search via eager sparse scoring](#). *arXiv preprint arXiv:2407.03618*.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. [Passage re-ranking with BERT](#). *arXiv preprint arXiv:1901.04085*.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992.
- Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.
- Sara Rosenthal, Yannis Katsis, Vraj Shah, Lihong He, Lucian Popa, and Marina Danilevsky. 2026a. [MTRAG-UN: A Benchmark for Open Challenges in Multi-Turn RAG Conversations](#). *Preprint*, arXiv:2602.23184.
- 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.
- Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. [Question rewriting for conversational question answering](#). In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 355–363.
- Christophe Van Gysel and Maarten de Rijke. 2018. [Py trec_eval: An extremely fast python interface to trec_eval](#). In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 873–876. ACM.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bowen Zheng, Bowen Yu, and 1 others. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.

A Query Rewriting Prompt

The following system prompt is used for query rewriting. The user message concatenates up to 10 turns of conversation history followed by the current question.

You are a query rewriting assistant for information retrieval. Given a conversation history and a current question, rewrite the question to be completely standalone and self-contained.

Rules:

1. Resolve all pronouns (it, they, this, that) to their explicit referents
2. Include relevant context from the conversation that's needed to understand the query
3. Keep the rewritten query concise and search-friendly
4. Do not add information not present in the conversation
5. If the question is already standalone, return it unchanged

B Hyperparameters

Table 6 lists all hyperparameters needed to reproduce our system.

| Parameter | Value |
|--------------------------------|------------------------------------|
| <i>Query Rewriter (LoRA)</i> | |
| Base model | Qwen2.5-7B-Instruct |
| LoRA rank | 16 |
| LoRA alpha (α) | 32 |
| LoRA dropout | 0.15 |
| Target modules | q/k/v/o_proj, gate/up/down_proj |
| Number of layers | 28 (all) |
| Trainable params | 40.4M (0.53%) |
| Optimizer | AdamW |
| Learning rate | 1×10^{-5} |
| Weight decay | 0.01 |
| Batch size (micro) | 2 |
| Gradient accumulation | 8 |
| Effective batch size | 16 |
| Training iterations | 500 |
| Max sequence length | 2048 |
| Gradient checkpointing | Yes |
| Precision | bf16 |
| Seed | 42 |
| <i>BM25 Retrieval</i> | |
| Library | BM25S |
| Stopwords | English |
| Stemming | None |
| Candidates | 50 |
| <i>Dense Retrieval</i> | |
| Embedding model | BGE-base-en-v1.5 |
| Embedding dim | 768 |
| Index type | FAISS IndexFlatIP |
| Normalization | L2 (cosine sim.) |
| Candidates | 50 |
| <i>Hybrid Fusion</i> | |
| Method | RRF |
| RRF k | 60 |
| <i>Cross-Encoder Reranking</i> | |
| Model | BGE-reranker-v2-m3 |
| Parameters | 568M |
| Rerank candidates | 50 |
| Output size | 10 |
| Batch size | 8 |
| <i>Domain Temperatures</i> | |
| Cloud | 0.0 |
| ClapNQ | 0.2 |
| FiQA | 0.3 |
| Govt | 0.3 |

Table 6: Complete hyperparameter configuration for the final submission.