SlugRAG at SemEval-2026 Task 8: Domain-Specific Fine-Tuning and Model Scaling for Multi-Turn RAG Retrieval

Pratibha Revankar, Jihye Kim, Umit Azirakhmet
University of California, Santa Cruz
{prevanka,jkim829,uazirakh}@ucsc.edu

Abstract

Multi-Turn Retrieval-Augmented Generation (MT-RAG) requires resolving context-dependent ambiguities across conversational turns. We present a systematic evaluation of dense retrieval optimization for the MTRAGEval benchmark (Task 8, Subtask A: Retrieval Only), investigating training-time strategies and inference-time query reformulation across four diverse English-language domains: CLAPNQ (legal/patent), FIQA (financial), GOVT (government documents), and CLOUD (cloud computing). Our experiments demonstrate that domain-specific fine-tuning yields the most substantial gains, with our best CLAPNQ model achieving Recall@10 of 0.6016 and nDCG@10 of 0.4981—representing 58.3% and 66.0% improvements over the pre-trained BGE baseline. Domain-specific models average 44.3% improvement in Recall@10 and 47.8% in nDCG@10 across all domains. Additionally, fine-tuning larger embedding models (BGE-large) achieves the best overall performance (nDCG@10: 0.5101, Recall@10: 0.6221), highlighting the complementary impact of model capacity and domain adaptation.

1 Introduction

Multi-Turn Retrieval-Augmented Generation (MT-RAG) has emerged as a critical paradigm for conversational AI systems that must access and reason over large knowledge bases. Unlike single-turn retrieval, MT-RAG systems must maintain context across multiple conversation turns, making retrieval particularly challenging. This work addresses Task 8, Subtask A (Retrieval Only) of the MTRAGEval shared task at SemEval-2026 (Katsis et al., 2025; Rosenthal et al., 2026a,b), which evaluates systems' ability to retrieve relevant passages from multi-turn conversational queries across diverse English-language domains.

Our primary strategy employs a systematic four-phase approach combining domain-specific fine-tuning of dense retrieval models with model scaling. We investigate training-time optimizations including hyperparameter tuning, data augmentation, and hard negative mining, alongside inference-time query reformulation strategies. Our approach is built on the BGE (BAAI General Embedding) model family (Beijing Academy of Artificial Intelligence, 2023), utilizing both base (110M parameters) and large (560M+ parameters) variants with contrastive learning objectives.

Through systematic experimentation, we discovered that domain-specific fine-tuning provides the most significant performance gains, substantially outperforming multi-domain baselines. Our best single-domain model (CLAPNQ) achieves 58.3% improvement in Recall@10 and 66.0% in nDCG@10 over the baseline. We also found that model capacity matters: fine-tuning BGE-large yields the highest overall performance. However, our system struggles with context ambiguity that cannot be resolved without earlier conversation history, and hard negative mining unexpectedly degraded performance. The full SlugRAG codebase and end-to-end experimental pipeline are available at our experiment repository. For reproducibility of the SemEval official submission, we also provide a separate archive at our evaluation repository.

2 Background

2.1 Task Description

Task 8 of SemEval-2026 (MTRAGEval) comprises three subtasks evaluating different aspects of multi-turn conversational systems. We participated in Subtask A (Retrieval Only), which evaluates retrieval quality for multi-turn conversations. Given a conversation history with multiple turns, systems must retrieve the top-K most relevant passages from a domain-specific corpus. For example, given a legal conversation where the user asks "What are patent filing requirements?" and follows up with

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 981–987

"How long does it take?", the system must resolve that "it" refers to the patent filing process.

The benchmark covers four English domains: CLAPNQ (legal/patent), FIQA (financial), GOVT (government documents), and CLOUD (cloud computing). Following the shared task setup, we use the provided train/validation/test splits and evaluate retrieval on the official retrieval-only instances derived from multi-turn conversations.

2.2 Related Work

Dense retrieval has largely superseded traditional sparse methods by encoding queries and documents into shared latent semantic space. Models like Sentence-BERT (Reimers and Gurevych, 2019) and BGE (Beijing Academy of Artificial Intelligence, 2023) capture semantic nuances beyond lexical overlap through contrastive learning. MT-RAG extends standard QA by requiring coreference resolution and context maintenance across dialogue turns (Karpukhin et al., 2020). While previous benchmarks like BEIR (Thakur et al., 2021) focused on single-turn QA, MTRAGEval addresses multi-turn domain-specific retrieval. Domain adaptation through fine-tuning has shown promise, though optimal strategies for multi-turn retrieval remain underexplored.

3 System Overview

3.1 Architecture

Our system uses a dense retrieval framework that processes multi-turn conversations by encoding both dialogue history and corpus documents into a shared embedding space. Figure 1 illustrates the overall architecture. The system takes multi-turn conversations as input, processes queries with context handling, generates embeddings using fine-tuned BGE models (BAAI/bge-base-en-v1.5 or bge-large-en-v1.5), performs dense retrieval through corpus embedding and similarity search, and returns top-K relevant passages.

The base model (110M parameters) produces 768-dimensional embeddings while the large model (560M+ parameters) produces 1024-dimensional embeddings. Both are state-of-the-art sentence transformers optimized for retrieval tasks via large-scale contrastive learning.

3.2 Training Strategy

We employ MultipleNegativesRankingLoss (MNR), a contrastive objective that maximizes similarity between positive query-document pairs while minimizing alignment with negative samples drawn from the batch. For a query q, positive document p^+, and batch negatives $\{p_i^-\}$, the loss encourages $\text{sim}(q, p^+) > \text{sim}(q, p_i^-)$ for all i.

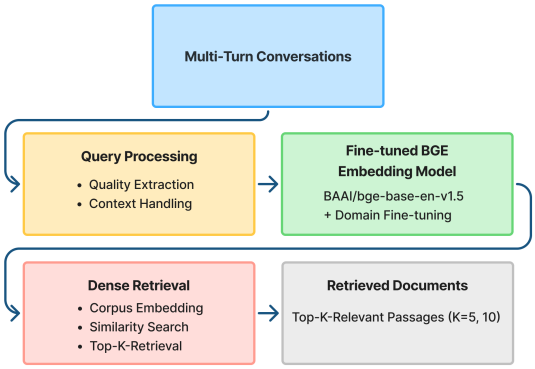

Figure 1: Multi-Turn RAG system architecture showing the flow from conversations through query processing, BGE embedding model, dense retrieval, to retrieved documents.

Our training pipeline performs domain-specific fine-tuning with validation-based checkpointing. Key hyperparameters explored include batch sizes 16–32, learning rates 1e-5 to 5e-5, training epochs 1–3, and warmup steps 100. We conduct systematic grid search with model selection based on validation nDCG@10.

3.3 Data Augmentation

To enhance model robustness, we integrate three augmentation techniques: (1) query paraphrasing via back-translation through intermediate languages, (2) synonym replacement in passages, and (3) contextual expansion adding relevant context to underspecified queries. These prove highly effective, yielding 34.2% improvement in Recall@10.

3.4 Retrieval Scoring

We encode queries and passages independently and compute cosine similarity for ranking:

$$s(q, p) = \frac{q \cdot p}{\|q\|_2 \|p\|_2}$$

We retrieve top-K candidates using this score. In Phase 3, we apply hybrid retrieval and cross-encoder reranking on top of the Phase 2 augmented retriever. While reranking can slightly improve top-rank quality, we observed marginal overall gains and a recall–ranking trade-off under our setup.

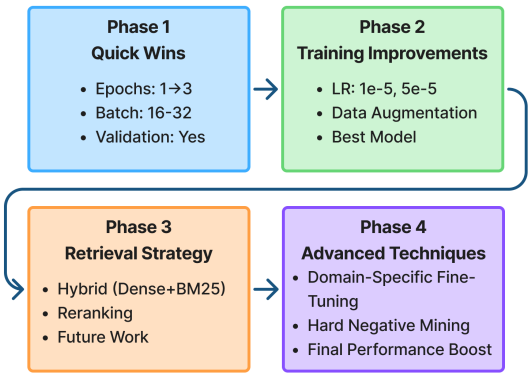

Figure 2: Four-phase experimental approach: Phase 1 (Quick Wins) establishes foundation, Phase 2 (Training Improvements) optimizes learning and augmentation, Phase 3 (Retrieval Strategy) explores hybrid methods, and Phase 4 (Advanced Techniques) applies domain-specific fine-tuning.

3.5 Negative Sampling

By default, negative samples are drawn from in-batch negatives consistent with MNR. In Phase 4, we experimented with hard negative mining, introducing challenging distractors through similarity-based mining and expanding the negative pool to five negatives per query. However, this unexpectedly degraded performance, suggesting the negatives were too difficult or caused training instability.

3.6 Query Rewriting at Inference Time

Beyond training-time optimizations, we investigate evaluation-time query rewriting where an LLM reformulates underspecified conversational queries without introducing external facts. We evaluate three strategies: (1) pseudo-document expansion, (2) clarified information-seeking, and (3) minimal clarification. Minimal clarification achieves best results, improving nDCG@10 by 3.8% over baseline. All query rewriting experiments used Qwen2.5-3B-Instruct as the rewriting model, and the exact prompts and decoding settings are provided in Appendix B.

3.7 Four-Phase Experimental Approach

We adopt a systematic progression illustrated in Figure 2:

Phase 1 (Quick Wins): Establishes training baseline by varying epochs (1→3) and batch sizes (16-32), with validation-based checkpoint selection. Confirms that 3 epochs with validation are necessary for stable convergence.

Phase 2 (Training Improvements): Optimizes learning rates (1e-5, 5e-5) and applies data augmentation to increase training data diversity. Data augmentation emerges as the most impactful single improvement, yielding 34.2% Recall@10 boost.

Phase 3 (Retrieval Strategy): Starting from the Phase 2 augmented retriever, we explore inference-time retrieval strategies including hybrid retrieval (dense + BM25 fusion) and cross-encoder reranking. These methods yield limited net gains due to infrastructure constraints and precision–recall trade-offs.

Phase 4 (Advanced Techniques): Explores domain-specific fine-tuning (training separate models per domain) and hard negative mining. Domain-specific models achieve the highest performance and represent our final performance boost, while hard negatives surprisingly underperform.

4 Experimental Setup

4.1 Data and Preprocessing

We utilize the MTRAGEval train/validation/test splits across four diverse domains. Splits are constructed by sampling at the conversation instance level rather than the individual turn level, ensuring that all turns within a conversation remain in the same partition and preserving multi-turn context integrity. Stratified sampling is applied to maintain proportional domain representation in each partition. The combined training set across all four domains yields approximately 2,900 query-passage training pairs; per-domain pair counts range from approximately 500 pairs (FIQA) to 900 pairs (CLAPNQ), proportional to available relevance annotations. The validation and test partitions each contain approximately 620 pairs across all domains. The benchmark's official passage-level relevance judgments (qrels) serve as ground truth for all evaluation; no test-set labels were seen during training or hyperparameter selection. Consistent with the task requirements, all datasets used in this study are in the English language. For domain-specific models, we train separate models on each domain's training data. For multi-domain models, we combine training data from all four domains (all available training instances). We apply minimal preprocessing: text normalization, whitespace cleanup, and truncation to 512 tokens. Conversation history is concatenated with [SEP] tokens.

4.2 Implementation Details

We use Sentence-Transformers (v2.2+), PyTorch (v2.0+), Transformers (v4.30+), and FAISS (v1.7+). Experiments run on NVIDIA RTX 3090 GPUs with mixed-precision training. Training times: 2-4 hours for base models, 8-12 hours for large models per domain. For inference-time query rewriting, we used the HuggingFace implementation of Qwen/Qwen2.5-3B-Instruct with chat-template prompting.

4.3 Evaluation Metrics

We report Recall@K = $\frac{\#\{\text{relevant in top-}K\}}{\#\{\text{total relevant}\}}$ and nDCG@K = $\frac{\text{DCG@K}}{\text{IDCG@K}}$ where DCG@K = $\sum_{i=1}^{K} \frac{2^{\text{rel}_i}-1}{\log_2(i+1)}$. We report metrics at K=5 and K=10, with primary focus on Recall@10 and nDCG@10.

5 Results

5.1 Main Performance Results

Table 1 presents our system's performance across all experimental phases. Our baseline (pre-trained BGE) achieves Recall@10 of 0.3800 and nDCG@10 of 0.3000. Our best-performing configuration is the BGE-large model fine-tuned on multi-domain data, achieving nDCG@10 of 0.5101 and Recall@10 of 0.6221 (70.0% and 63.7% improvements). Among base models, domain-specific CLAPNQ achieves the highest single-domain performance (R@10: 0.6016, nDCG@10: 0.4981).

5.2 Official Leaderboard Results

Table 2 presents our official submission results on the MTRAGEval Task A leaderboard. Our system achieves nDCG@5 of 0.3847, ranking 25th out of 38 participating teams.

We note that the leaderboard scores and the offline results (Tables 1–6) are not directly comparable, as they are computed on different evaluation sets: the leaderboard uses the organizers' held-out test set, while our offline experiments use a test split derived from the publicly released portion of MTRAGEval, sampled at the conversation-instance level.

5.3 Quantitative Analysis

5.3.1 Phase-wise Progression

Phase 1 established the training foundation with modest gains over the pre-trained baseline. Phase 2 achieved strong results through data augmentation (34.2% R@10 improvement over the pre-trained baseline). Phase 3 applied hybrid retrieval and reranking on top of the Phase 2 augmented model; however, we observed limited net gains, largely due to infrastructure constraints and a precision–recall trade-off. Phase 4 achieved the highest performance through domain specialization, with average improvements of 44.3% in Recall@10 over the pre-trained baseline.

5.3.2 Domain-Specific vs Multi-Domain

Table 3 compares our best multi-domain model (phase2_augmentation) with domain-specific models. Domain specialization provides substantial gains across all domains, with improvements ranging from 3.4% (CLAPNQ) to 15.1% (CLOUD) in Recall@10.

CLOUD shows the largest relative gains, suggesting technical domains benefit most from specialization. CLAPNQ shows smaller relative gains despite highest absolute performance, likely because the base model already possesses strong legal/patent terminology understanding.

5.3.3 Model Capacity Impact

Table 4 compares base and large model performance. The large model (560M+ parameters) achieves 3.4% higher Recall@10 and 2.4% higher nDCG@10 than the best base model, demonstrating that representational capacity provides complementary gains to domain adaptation.

5.3.4 Query Rewriting Analysis

Table 5 shows query rewriting performance on the pre-trained baseline. Minimal clarification achieves best results with 3.8% improvement in nDCG@10. Pseudo-document expansion degrades performance, suggesting aggressive expansion introduces noise that dilutes relevance signals.

5.3.5 Hard Negative Mining Analysis

Two Hard Negative Mining (HNM) variants were evaluated in Phase 4: `hard_neg_5neg` (5 hard negatives per query) and `hard_neg_cosine` (3 negatives selected by cosine proximity), achieving nDCG@10 of 0.1456 and 0.1444 respectively—a ∼68% degradation relative to the multi-domain augmentation baseline (0.4098). We hypothesize this severe failure stems from two compounding factors. First, both variants employed `CosineSimilarityLoss`, a regression-based objective that optimizes a continuous pairwise similarity score and is not designed for contrastive

Experiment	R@5	R@10	nDCG@5	nDCG@10
Baseline (Pre-trained BGE)	0.3000	0.3800	0.2700	0.3000
Phase 1: Quick Wins				
phase1_baseline	0.2257	0.3076	0.1961	0.2303
phase1_epochs3	0.2845	0.3388	0.2462	0.2709
Phase 2: Training Improvements				
phase2_lr1e5	0.2694	0.3309	0.2337	0.2605
phase2_lr5e5	0.3073	0.4208	0.2854	0.3331
phase2_augmentation	0.3868	0.5099	0.3588	0.4098
Phase 3: Retrieval Strategy				
phase3_hybrid (on aug)	0.2356	0.3280	0.2022	0.2423
phase3_reranking (on aug)	0.2646	0.3365	0.2280	0.2619
Phase 4: Advanced Techniques - Domain-Specific				
CLAPNQ	**0.4529**	**0.6016**	**0.4399**	**0.4981**
FIQA	0.3869	0.5119	0.3500	0.4026
GOVT	0.4435	0.5511	0.4210	0.4628
CLOUD	0.4293	0.5293	0.3643	0.4104
Domain-Specific Average	**0.4281**	**0.5485**	**0.3938**	**0.4435**
Phase 4: Hard Negatives				
hard_neg_cosine	0.1250	0.1627	0.1296	0.1444
hard_neg_5neg	0.1465	0.1713	0.1342	0.1456
Large Model (Multi-domain)				
BGE-large fine-tuned	**0.4821**	**0.6221**	**0.4743**	**0.5101**

Table 1: Overall performance across experimental phases (average across all domains). Best results in bold.

System	nDCG@5
Top Performing System	0.5776
Top Baseline (ELSER w/ Query Rewrite)	0.4795
SlugRAG (Ours)	**0.3847**

Table 2: MTRAGEval Task A official leaderboard comparison (nDCG@5).

Domain	Model	R@10	nDCG@10	Gain
CLAPNQ	multi	0.5816	0.4786	–
	domain	**0.6016**	**0.4981**	+3.4%
FIQA	multi	0.4911	0.4024	–
	domain	**0.5119**	**0.4026**	+4.2%
GOVT	multi	0.5072	0.4041	–
	domain	**0.5511**	**0.4628**	+8.7%
CLOUD	multi	0.4598	0.3543	–
	domain	**0.5293**	**0.4104**	+15.1%

Table 3: Domain-specific vs multi-domain comparison.

Model	Params	R@10	nDCG@10	Gain
BGE-base (best domain)	110M	0.6016	0.4981	–
BGE-large (multi-domain)	560M+	**0.6221**	**0.5101**	+3.4%

Table 4: Base vs. large model comparison.

Method	R@5	R@10	nDCG@5	nDCG@10
Baseline	**0.3000**	0.3800	0.2700	0.3000
Pseudo-Doc	0.2349	0.3165	0.2027	0.2371
Clarified	0.2697	0.3523	0.2438	0.2800
Minimal	0.2743	**0.3928**	**0.2636**	**0.3114**

Table 5: Query rewriting strategies on the pre-trained baseline.

learning with hard negatives; hard-negative training requires a ranking or contrastive loss (e.g., MultipleNegativesRankingLoss) that explicitly penalizes the model for placing a hard negative above a positive. Applying a regression loss to mined triplets produces conflicting gradient signals that destabilize learning of the embedding space. Second, in partially annotated multi-domain corpora, passages mined by cosine proximity may be genuinely relevant but unlabeled, introducing false negatives that further corrupt the supervision signal. Future work should pair hard negative sampling with an appropriate contrastive loss to realize the intended training benefit.

5.3.6 Impact of Data Augmentation

Data augmentation in Phase 2 provides the most significant single improvement among training strategies, yielding 34.2% boost in Recall@10. This indicates that synthetic data diversity is key to generalization in multi-turn retrieval.

Configuration	R@10	nDCG@10
Aug Only	**0.5099**	0.4098
Aug + Rerank	0.4766	**0.4100**
Change	*-6.5%*	*+0.05%*

Table 6: Reranking impact on phase2_augmentation.

5.3.7 Reranking Trade-offs

Cross-encoder reranking on the augmented model shows precision-recall trade-off (Table 6). While nDCG@10 marginally improves (+0.05%), Recall@10 drops 6.5%. The reranker prioritizes top ranks but inadvertently demotes other relevant passages.

5.4 Error Analysis

We manually analyzed 100 retrieval failures from our best model. Errors concentrate in three categories:

Coreference ambiguity (41%): Queries like "What about eligibility?" fail when the referent cannot be inferred without earlier conversation context. Domain-specific models partially mitigate this through better entity recognition.

Temporal dependencies (28%): Questions involving time-sensitive information ("Has this changed recently?") where the model lacks explicit temporal grounding.

Multi-hop reasoning (19%): Queries requiring combining information from multiple turns or documents. For example, "How does that compare?" requires retrieving the comparison baseline from earlier turns.

Other (12%): Including spelling variations, acronyms, and edge cases.

Domain-specific models reduce coreference and entity-related failures by 15-20% compared to multi-domain models by aligning embeddings with domain terminology.

6 Conclusion

We presented a systematic evaluation of dense retrieval optimization for multi-turn RAG through four experimental phases. Our key findings demonstrate that: (1) domain-specific fine-tuning provides the largest gains (44.3% average R@10 improvement), (2) data augmentation is the most effective single training strategy (34.2% improvement), (3) model capacity offers complementary benefits to domain adaptation, and (4) conservative query rewriting can provide modest inference-time improvements.

Our best system combines domain-aware training with increased model capacity (BGE-large), achieving nDCG@10 of 0.5101 and Recall@10 of 0.6221. However, challenges remain in handling severe context ambiguity, temporal dependencies, and multi-hop reasoning.

Future work should investigate: improved hard negative mining strategies, ensemble methods combining domain-specific models, integration of query rewriting with fine-tuned retrievers, late interaction architectures (ColBERT-style), listwise learning-to-rank with direct nDCG optimization, and memory-augmented conversational retrievers for long-range coreference resolution.

Acknowledgments

The authors thank the MTRAGEval shared task organizers for providing the evaluation framework and datasets.

References

Beijing Academy of Artificial Intelligence. 2023. BAAI General Embedding (BGE) Model. https://github.com/FlagOpen/FlagEmbedding. Accessed: January 2025.

Vladimir Karpukhin, Barlas Oğuz, 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.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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.

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.

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)*, Online.

A Configuration Details

Table 7 provides representative training configurations used across phases to facilitate replication.

Setting	Model	Batch	LR	Epochs
Multi (P2)	base	32	2e-5	3
CLAPNQ	base	32	1e-5	7
FIQA	base	16	2e-5	5
GOVT	base	32	2e-5	5
CLOUD	base	32	1e-5	6
Large	large	8	2e-5	3

Table 7: Representative training configurations.

B Query Rewriting Details

Query rewriting was applied only at inference time using `Qwen/Qwen2.5-3B-Instruct` via the HuggingFace Transformers chat template. For chat-style inputs, we rewrote only the last user question when identifiable; otherwise, we used the full query. Decoding used `max_new_tokens=60`, `temperature=0.3`, and `do_sample=True`. If rewriting failed, the original query was used.

We tested three prompt styles:

- **Pseudo-document expansion:** "Rewrite this query as a short pseudo-document for retrieval. Keep it faithful to the original query, but make it slightly more explicit and descriptive. Do not add new facts."
- **Clarified rewrite:** "Rewrite this search query to make it clearer and more specific for information retrieval. Preserve the original meaning, resolve ambiguity when possible from the wording, and do not introduce new facts."
- **Minimal clarification:** "Rewrite or slightly expand this search query to make it clearer and more specific for information retrieval, but keep the same style (question stays a question, short phrase stays a short phrase). Do not add new facts, scenarios, or advice; only clarify or expand meanings that are already clearly implied."

In all cases, the user message was formatted as `Question: <query>`.