clulab-retrieval at SemEval-2026 Task 8: A Comparative Analysis of Dense Retrievers and HyDE for Multi-Turn Conversational Retrieval

Hyungji Kim **Siva Rohit Kondapaneni** **Steven Bethard**
University of Arizona
{hyungjikim, sivarohit2002, bethard}@arizona.edu

Abstract

We present a comparative analysis of dense retrievers and retrieval strategies for multi-turn conversational retrieval in SemEval-2026 Task 8 (MTRAGEval). Our official submission employed a fine-tuned E5-based dense retriever (E5-FT, ∼110M parameters) with Hypothetical Document Embeddings (HyDE), achieving nDCG@5 of .3309, ranking 31 out of 38 systems. On the development set we also compared E5-FT versus BGE embeddings, dense-only versus hybrid retrieval strategies, and HyDE versus keyword extraction approaches. We found: (1) BGE (general-purpose, ∼110M) outperforms our domain-fine-tuned E5-FT (∼110M) by 30.5% on baseline retrieval, suggesting that model selection may matter more than domain-specific fine-tuning, (2) hybrid retrieval combining BM25 and dense methods provides complementary signals, with HyDE improving BM25 by 26.7% and dense retrieval by 4.0%, and (3) keyword-based query simplification degrades performance by 11-28% across domains, validating HyDE's approach of preserving semantic richness through passage-level text.

1 Introduction

Conversational retrieval systems face unique challenges compared to single-turn retrieval: queries often contain elliptical references, topic shifts, and contextual dependencies across multiple turns. The MTRAGEval benchmark (Katsis et al., 2025; Rosenthal et al., 2026a,b) addresses these challenges with 110 human-created conversations spanning 842 retrieval tasks across four domains.

Our approach leverages Hypothetical Document Embeddings (HyDE; Gao et al., 2023), which generates hypothetical answer passages that are embedded and used for retrieval instead of the original query. HyDE is particularly valuable in conversational settings where queries may be terse, ambiguous, or context-dependent.

We present a comparative analysis of dense retrievers and retrieval strategies for this task. Our official submission employed a fine-tuned E5-based dense retriever (E5-FT, ∼110M parameters) with HyDE, achieving nDCG@5 of .3309. We also explored alternative configurations on our development set, including BGE embeddings, hybrid retrieval (BM25+dense), and various HyDE application strategies. Our main contributions are:

- Systematic comparison of dense retrievers for conversational search: E5-FT (∼110M, fine-tuned, domain-specific) versus BGE (∼110M, general-purpose)
- Evaluation of retrieval strategies: dense-only versus hybrid (BM25+dense)
- Extensive HyDE ablation studies across models, query formulations, and retrieval methods
- Keyword extraction ablation showing that query simplification degrades retrieval performance
- Analysis of factors affecting performance in multi-turn conversational retrieval: training approach (general-purpose vs. domain-specific fine-tuning), query formulation sensitivity, and complementary retrieval signals

Our system achieved nDCG@5 of .3309 on the official test set, ranking 31 out of 38 systems and below the top baseline (ELSER + Rewrite: .4795) and the top-performing system (.5776). Extensive development set analysis revealed that (1) BGE outperforms our domain-fine-tuned E5-FT by 30.5%, suggesting model selection may matter more than domain-specific fine-tuning, (2) hybrid retrieval combining BM25 and dense methods provides strong complementary signals, and (3) keyword-based simplification degrades performance by 11-28% across domains. Our code is available at https://github.com/clulab/semeval2026-task8.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2787–2792
July 3-4, 2026 ©2026 Association for Computational Linguistics

2 Background

2.1 Task Overview

MTRAGEval (Rosenthal et al., 2026a,b) evaluates RAG systems on multi-turn conversational scenarios across four domains: ClapNQ (Wikipedia), Cloud (IBM technical), FiQA (financial), and Govt (government). All conversations and documents are in English. The benchmark contains 110 conversations with an average of 7.7 turns each, totaling 842 evaluation tasks.

The task provides three query formulations: **last-turn** (only the most recent user question), **questions** (concatenation of all user questions), and **rewrite** (standalone rewritten query incorporating necessary context). Each domain uses 512-token passages with 100-token overlap. Retrieval is evaluated using Recall@K and nDCG@K (K=1,3,5,10).

2.2 Related Work

Conversational retrieval: Prior work on conversational search (Yu et al., 2021) has explored query rewriting, context modeling, and multi-turn understanding. MTRAGEval adds realistic human-created conversations spanning diverse domains.

Hypothetical Document Embeddings: Gao et al. (Gao et al., 2023) introduced HyDE for web search, demonstrating that generating hypothetical answer passages can improve dense retrieval. Our work extends HyDE to multi-turn conversational settings and provides the first systematic comparison of HyDE's effectiveness on sparse (BM25) versus dense retrieval, showing substantially larger gains on sparse retrieval (+26.7% vs +4.0%).

Dense retrieval models: Dense retrieval methods like DPR (Karpukhin et al., 2020) and ColBERT (Khattab and Zaharia, 2020) have shown strong performance on single-turn retrieval benchmarks. Our comparative analysis of E5-FT (domain-specific, fine-tuned) versus BGE (general-purpose) provides insights about model selection for multi-turn conversational retrieval, suggesting that robustness across query formulations may be more important than domain-specific optimization.

3 System Overview

3.1 Hypothetical Document Embeddings

HyDE (Gao et al., 2023) bridges the semantic gap between queries and relevant passages by generating a hypothetical answer document, then using its embedding for retrieval. We use Gemini 2.5 Flash for generation with the prompt shown in Figure 1.

```
You are a helpful assistant who generates
hypothetical answer passages (1-3 sentences).
Given the conversation history and final user
query, write a short, factual paragraph that
directly and concisely answers the final user
query. This passage will be used to find
similar real documents.

Instructions:
- Produce a concise factual passage (1-3
  sentences) that answers the final query.
- Preserve named entities and numeric tokens
  exactly as they appear in the query.
- Do NOT add notes, explanations, or meta-
  comments; output ONLY the hypothetical
  passage.
- Do NOT include any preamble like "Here is
  the answer." Output only the passage text.

Conversation History:
{conversation_history}

Final Query:
{query}
```

Figure 1: HyDE prompt for Gemini 2.5 Flash.

3.2 Official Submission System

Our official submission pipeline consists of three steps: (1) generate a hypothetical answer passage using HyDE, (2) embed the HyDE passage with a fine-tuned dense retriever (E5-FT), and (3) retrieve passages via FAISS nearest neighbor search. We built separate FAISS indices (`IndexFlatIP`) for each domain using pre-encoded passages.

3.3 Development Experiments

Beyond our official submission, we conducted extensive experiments on our development set to compare: (1) dense retrievers (E5-FT vs BGE), (2) retrieval strategies (dense-only vs hybrid BM25+dense), (3) HyDE application variants (BM25 only, dense only, or both), (4) query formulations (last-turn, questions, rewrite) with optional conversation history, and (5) keyword extraction as an alternative to HyDE.

4 Experimental Setup

We evaluate on two sets:

- **Official test set**: 507 tasks from the SemEval evaluation phase (official metric: nDCG@5)
- **Development set**: 4-domain internal development set (ClapNQ, Cloud, FiQA, Govt) used for comparative analysis and ablation studies (metrics: nDCG@5, nDCG@10, Recall@10)

We explored the following configurations. The official submission settings are marked with †.

- Dense retrievers:
  - †**E5-FT**: Fine-tuned E5-based dense retriever, ∼110M parameters; base model `intfloat/e5-base-v2` (BertModel, 768-dim); trained on 170,176 domain-balanced query–passage pairs from the shared task using MultipleNegativesRankingLoss (in-batch negatives, cosine similarity, temperature scale 20); 2 epochs, batch size 16, learning rate 5×10^{-5}, linear schedule, max sequence length 256 tokens, FP16; `sivarohit2002/qwen06b_bi-e5-ft-weighted`
  - **BGE**: ∼110M-parameter general-purpose bi-encoder, `BAAI/bge-base-en-v1.5`
- Retrieval strategies:
  - †**Dense-only**: Using dense retriever with Top-K=10 from FAISS inner product similarity (`IndexFlatIP`); all passage and query embeddings are L2-normalized before indexing and search, making inner product equivalent to cosine similarity
  - **Hybrid**: Combining BM25 (Elasticsearch) with dense retrieval
  - **HyDE application**: Applied to BM25 only, dense only, or both
- Query formulations: **last-turn**, **questions**, †**rewrite** (with optional conversation history; provided by organizers)
- Query reformulation:
  - †**HyDE**: Gemini 2.5 Flash, temperature=0.7, max_tokens=200
  - **Keyword** Keyword extraction using Gemini 2.5 Flash (temperature=0, few-shot prompting), outputting canonical 2-6 word noun phrases, preserving named entities, acronyms, and numbers.
- General settings: All results from single runs executed on NVIDIA A100 (40GB)

5 Results

5.1 Official Submission Performance

Table 1 shows our official submission results on the test set. The following sections perform comparative analyses of alternative retrieval configurations on our development set.

5.2 Model Comparison: E5-FT vs BGE

Table 2 compares E5-FT vs. BGE. Although both models are ∼110M parameters (BERT-base scale), BGE's baseline performance (nDCG@10=.398) exceeds E5-FT's HyDE-augmented performance (nDCG@10=.344) by 15.7%. Because BGE and E5-FT differ in both base model and training objective, this comparison is observational rather than controlled; nonetheless, it suggests that general-purpose retrieval training may outperform domain-specific fine-tuning even at the same model scale.

Both models benefit from HyDE, but with different magnitudes: E5-FT improves by +12.8% while BGE improves by +4.0%. This suggests that stronger baseline models see smaller relative improvements from HyDE, though absolute performance remains higher.

System	nDCG@5
Baselines (from organizers)	
Top Performing System	.5776
Top Baseline (ELSER + Rewrite)	.4795
Our Submission	
clulab-retrieval (E5-FT + HyDE)	**.3309**

Table 1: Official submission results on the test set.

Domain	E5-FT (Dense-only)				BGE (Dense-only)			
	Baseline		+ HyDE		Baseline		+ HyDE	
	G10	R10	G10	R10	G10	R10	G10	R10
ClapNQ	.425	.515	.507	.588	.492	.591	.519	.623
Cloud	.191	.248	.205	.267	.343	.423	.369	.452
FiQA	.257	.337	.293	.390	.341	.418	.334	.415
Govt	.348	.472	.372	.516	.415	.518	.434	.533
Macro	**.305**	**.393**	**.344**	**.440**	**.398**	**.488**	**.414**	**.506**

Table 2: Comparison of models (E5-FT versus BGE) on the development set. nDCG@10 and R@10 are abbreviated as G10 and R10, respectively.

5.2.1 Model Size vs Domain Adaptation

Table 3 compares E5-FT, which has been fine-tuned to the task domains, to gte-Qwen2-1.5B, a ∼13.6× larger general-domain instruction-tuned model. Despite its larger size, gte-Qwen2-1.5B dramatically underperforms E5-FT on domain-specific

Embedding Model	FiQA (Finance)		Cloud (Technical)	
	Rewritten	HyDE	Rewritten	HyDE
`e5-ft`	**.257**	**.293**	**.191**	**.215**
`gte-Qwen2-1.5B`	.162	.186	.069	.114

Table 3: Comparison of model size vs. domain adaptation: development set nDCG@10 scores for domain-specific fine-tuned E5-FT (∼110M) and general domain instruction-tuned gte-Qwen2-1.5B (1.5B).

Component	Baseline		+ HyDE	
	nDCG@10	R@10	nDCG@10	R@10
Sparse (BM25)	.247	.320	.313	.393
Dense (BGE)	.398	.488	.414	.506

Table 4: Comparison of retrieval strategies on the development set with structured conversation history.

Configuration	Model	Apply	Baseline	+HyDE
Dense-only (nDCG@10)				
E5-FT, rewrite	E5-FT	Dense	.305	.344
BGE, rewrite	BGE	Dense	.398	.414
BGE, lastturn	BGE	Dense	.339	.365
Hybrid (nDCG@10)				
BGE, rewrite	BGE	BM25 only	.247	.313
BGE, rewrite	BGE	Dense only	.398	.414

Table 5: Ablation of HyDE across configurations on the development set.

retrieval, with particularly severe degradation on Cloud (-64% rewritten, -47% HyDE) and FiQA (-37% rewritten, -36% HyDE). This confirms that domain-adapted fine-tuning is more effective than instruction-tuning for specialized retrieval tasks, even when the instruction-tuned model has $\sim$13.6$\times$ more parameters. However, the comparison between E5-FT and BGE (Table 2) suggests that general-purpose retrieval training (BGE) can outperform domain-specific fine-tuning when the base model has broader coverage and better robustness.

5.3 Retrieval Strategy Comparison: Dense-Only vs Hybrid

Table 4 compares sparse, dense, and hybrid retrieval approaches. BM25 benefits substantially from HyDE (+26.7% nDCG@10), while dense retrieval shows smaller gains (+4.0%). This asymmetry suggests that hypothetical answer passages provide valuable lexical expansion for sparse retrieval, while dense retrievers already capture much of the semantic information that HyDE provides. The hybrid approach provides complementary signals: BM25 excels at exact matches (entities, technical terms, numeric values) while dense retrieval handles paraphrases and conceptual similarity.

5.4 HyDE Ablation

Table 5 compares configurations with and without HyDE. HyDE provides consistent improvements across all configurations tested. Weaker baselines show larger relative improvements (E5-FT: +12.8% vs BGE: +4.0%). BM25 shows the largest gains (+26.7%), suggesting HyDE's lexical expansion is particularly valuable for sparse retrieval. Query formulation affects baseline and HyDE performance (rewrite $>$ lastturn for both models). All HyDE results reported here are from single runs. We did not observe obvious instability across manual inspection of generated passages: for a given query, HyDE consistently produced factually similar hypothetical passages, suggesting limited within-query variance. However, we did not run systematic multi-seed experiments to quantify variance.

5.5 Keyword Extraction

Table 6 compares HyDE-style rewriting to an alternative rewriting: keyword extraction (converting queries into canonical 2-6 word noun phrases). Keyword extraction consistently underperforms both baseline rewriting and HyDE across all domains. Keywords from rewritten queries show 11-28% degradation in nDCG@10 compared to the rewritten baseline (ClapNQ: -11%, Cloud: -13%, FiQA: -17%, Govt: -13%). Even when keywords are extracted from HyDE-generated passages, performance remains substantially below both the rewritten baseline and HyDE approaches. The particularly large degradation on technical domains (Cloud: -13%, FiQA: -17%) suggests that domain-specific queries especially require full context and terminology that keywords alone cannot capture.

5.6 Discussion

5.6.1 General-Purpose Training vs. Domain-Specific Fine-Tuning

Our comparison of E5-FT ($\sim$110M parameters, domain-specific fine-tuning) and BGE ($\sim$110M parameters, general-purpose) shows that BGE outperforms E5-FT by 30.5% on baseline retrieval. Even when E5-FT uses HyDE, BGE without HyDE still exceeds it by 15.7%, and BGE+HyDE outperforms E5-FT+HyDE by 20.3%. Because these two models differ in both base model architecture and training objective, the comparison is observational rather than a controlled ablation of training approach.

We hypothesize that BGE's training on diverse retrieval tasks provides better robustness to varied query formulations and domain-specific language. E5-FT's domain-specific fine-tuning may be overfitting particular query patterns, reducing performance on queries outside the fine-tuning distribution. Confirming this hypothesis would require

	ClapNQ		Cloud		FiQA		Govt	
Query Method	**nDCG@10**	**R@10**	**nDCG@10**	**R@10**	**nDCG@10**	**R@10**	**nDCG@10**	**R@10**
Baseline: Standard Rewriting								
Rewritten Query	.425	.515	.191	.248	.257	.337	.348	.472
Strategy: Hypothetical Document Embeddings (HyDE)								
HyDE (from Last Turn)	.506	**.597**	**.215**	**.276**	**.293**	.370	.336	.487
HyDE (from Rewritten)	**.507**	.588	.205	.267	**.293**	**.390**	**.372**	**.516**
Strategy: Keyword Extraction (Ablation)								
Keywords (Last Turn)	.320	.410	.182	.228	.176	.227	.260	.365
Keywords (Rewritten)	.379	.477	.167	.218	.214	.285	.302	.438
Keywords (HyDE Last)	.335	.415	.150	.192	.202	.265	.269	.392
Keywords (HyDE Rewr)	.375	.472	.167	.210	.196	.256	.277	.399

Table 6: Keyword extraction ablation on development set using E5-FT (nDCG@10 and Recall@10).

fine-tuning from the same base model with and without domain-specific data.

5.6.2 Complementary Retrieval Signals

The large performance gap between BM25's HyDE improvement (+26.7%) and dense retrieval's improvement (+4.0%) highlights the complementary nature of sparse and dense retrieval. BM25 benefits from HyDE's lexical expansion—converting terse queries into passage-like text with richer vocabulary. Dense retrieval, already operating in semantic space, sees smaller gains as it can bridge vocabulary gaps without explicit lexical matching.

This complementarity suggests that hybrid retrieval strategies may be particularly valuable for conversational search, where queries vary widely in formulation (from single words to complete sentences) and context requirements.

5.6.3 Query Formulation Sensitivity

Our results show that both baseline performance and HyDE improvements vary substantially with query formulation. Rewritten queries (standalone, context-incorporated) outperform lastturn queries (contextless, potentially ambiguous) by 17% for E5-FT baseline and 35% for BGE baseline.

Interestingly, BGE shows more consistent performance across query types than E5-FT, suggesting better robustness to query formulation variance. This robustness is consistent with our hypothesis that general-purpose retrieval training may yield more stable representations across varied query formulations than domain-specific fine-tuning.

5.6.4 Error Analysis

Manual inspection of 50 challenging queries reveals common failure patterns:

HyDE hallucination: HyDE occasionally generates specific details not present in queries (e.g., inventing specific company names or dates when the query asks about "a company" or "recent events"). These hallucinated details can lead retrieval toward irrelevant but lexically-matching passages.

Topic ambiguity: When queries admit multiple interpretations (e.g., "What about security?" could mean cybersecurity, physical security, or financial security), HyDE must commit to one interpretation, potentially missing relevant passages addressing alternative interpretations.

Query formulation sensitivity: E5-FT shows a larger performance variance across query types than BGE. For example, concatenated question strings (questions variant) cause significant QwenFT performance degradation, while BGE remains relatively stable.

5.6.5 Reflections on Official Submission

Our official submission's performance (nDCG@5 of .3309) can be attributed to several factors identified through our comparative analysis:

1. **Suboptimal model selection**: E5-FT's baseline performance (dev nDCG@10 of .305) is 30.5% below BGE baseline (.398)
2. **Dense-only approach**: Our submission did not leverage BM25's complementary signal, which shows +26.7% improvement from HyDE
3. **Limited robustness**: E5-FT's sensitivity to query formulation may have hurt performance on varied test set queries

While we cannot evaluate BGE or hybrid retrieval on the official test set due to time constraints and data preprocessing differences, our development set analysis suggests these alternative configurations merit investigation in future work.

6 Conclusion

We presented a comparative analysis of dense retrievers and retrieval strategies for multi-turn conversational retrieval in SemEval-2026 Task 8. Our official submission (E5-FT + HyDE, dense-only) achieved nDCG@5=.3309.

Through extensive development set experiments we found:

BGE outperforms E5-FT, suggesting general-purpose training may matter more than domain-specific fine-tuning: BGE (∼110M parameters, general-purpose) outperforms E5-FT (∼110M parameters, domain-specific fine-tuned) by 30.5% on baseline retrieval. Because the two models differ in both base model and training objective, this is an observational finding; we hypothesize that breadth of retrieval training contributes to BGE's stronger and more robust performance

Hybrid retrieval provides complementary signals: BM25 and dense retrieval show asymmetric responses to HyDE (+26.7% vs +4.0%), suggesting they capture different aspects of relevance

HyDE consistently helps: Across all configurations tested, HyDE provides improvements ranging from +4.0% to +26.7%, with particularly strong gains on sparse retrieval

Keyword simplification hurts: Reducing queries to keyword phrases degrades performance by 11-28% across domains, demonstrating that dense retrievers benefit from preserving full semantic context rather than lexical precision alone

These findings suggest future research directions: investigating model robustness metrics beyond single-turn benchmarks, exploring optimal fusion strategies for hybrid retrieval, and developing HyDE variants that balance lexical expansion with semantic precision.

Acknowledgments

We thank the SemEval 2026 Task 8 organizers for creating the MTRAGEval benchmark and providing comprehensive evaluation infrastructure.

References

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

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

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of SIGIR*.

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.

Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2021. Few-shot conversational dense retrieval. In *Proceedings of SIGIR*.