Howard University-AI4PC at SemEval-2026 Task 8: Query Reformulation and Dense-Lexical Retrieval Fusion for Multi-Turn Retrieval-Augmented Generation

Sijan Shrestha and Saurav K. Aryal*
Howard University
AI4PC Lab
sijan.shrestha@bison.howard.edu
saurav.aryal@howard.edu

Abstract

We present a training-free hybrid retrieve-then-rerank system for multi-turn retrieval-augmented generation, submitted to all three subtasks of SemEval-2026 Task 8 (MTRAGEval): passage retrieval (Task A), generation with reference passages (Task B), and end-to-end RAG (Task C). Our system addresses the core multi-turn challenges—non-standalone questions, unanswerable queries, and shifting passage relevance—across four domain-specific corpora: ClapNQ, Cloud, FiQA, and Govt. Queries are reformulated through LLM-driven rewriting, decomposition into sub-queries, and Hypothetical Document Embeddings (HyDE). Retrieved candidates from dense vector search (BGE-base-en-v1.5) and BM25 lexical matching are fused via Reciprocal Rank Fusion and reranked by a cross-encoder (BGE-reranker-large). Llama-3.3-70B-Instruct generates extractive, context-grounded responses with built-in abstention for unanswerable queries. Using only open-source models without fine-tuning, the system achieves nDCG@5 of 0.4098 on Task A (22nd/38), a harmonic mean of 0.7462 on Task B (9th/26), and 0.5796 on Task C (**2nd/29**), coming within 1.1% of the top submission. We attribute the strong Task C result to the synergy between multi-signal query reformulation and faithful extractive generation.

1 Introduction

Retrieval-augmented generation (RAG) grounds LLM responses in retrieved passages, reducing hallucinations and improving factual accuracy (Aryal and Akomoize, 2025; Prioleau et al., 2025; Hagos et al., 2025). However, most RAG benchmarks evaluate single-turn interactions (Pradeep et al., 2025; Katsis et al., 2025; Aryal and Pant, 2025; Prioleau et al., 2025), overlooking the distinct challenges inherent to multi-turn conversations (Katsis et al., 2025):

- **Non-Standalone Questions:** Later turns rely on references from prior context (e.g., pronouns, ellipsis), requiring coreference resolution.
- **Unanswerable Questions:** Systems must abstain rather than hallucinate when retrieved context is insufficient.
- **Active Retrieval:** Relevant passages shift between turns as the conversation topic evolves.

The MTRAG benchmark (Katsis et al., 2025) provides the first end-to-end human-generated multi-turn RAG evaluation framework with 110 conversations (7.7 avg turns) across four document collections, yielding 842 evaluation tasks. It defines three subtasks—passage retrieval (Task A), generation with reference passages (Task B), and end-to-end RAG (Task C)—and was adopted as SemEval-2026 Task 8 (Rosenthal et al., 2026b). The test set is drawn from the MTRAG-UN benchmark (Rosenthal et al., 2026a), which extends MTRAG with additional challenges including underspecified and non-standalone questions across six domains.

In this paper, we describe our system submitted to **all three subtasks** under the team name *Howard University-AI4PC*. Our approach centers on two ideas: (1) *query reformulation* to transform context-dependent questions into standalone queries through LLM-driven rewriting, decomposition, and hypothetical document generation; and (2) *dense-lexical retrieval fusion* to combine dense vector search and lexical term matching via Reciprocal Rank Fusion, followed by cross-encoder reranking. All components use open-source models without task-specific fine-tuning.

2 Related Work

Multi-Turn RAG. Multi-turn conversational question answering has been explored through benchmarks such as QuAC (Choi et al., 2018), CoQA (Reddy et al., 2019), and FaithDial (Dziri

*Corresponding author

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2533–2539

et al., 2022). However, these typically keep retrieval fixed or focus on a single domain. MTRAG (Katsis et al., 2025) addresses these limitations by incorporating active retrieval, multiple domains, unanswerable questions, and long-form answers. MTRAG-UN (Rosenthal et al., 2026a) extends MTRAG with additional challenges including underspecified and non-standalone questions across six domains. RAD-Bench (Kuo et al., 2025) also evaluates multi-turn RAG but does not include passage-level retrieval evaluation.

Hybrid Retrieval. Dense retrieval using pretrained encoders such as BGE (Xiao et al., 2024) has shown strong performance on passage retrieval tasks. BM25 (Robertson and Zaragoza, 2009) remains competitive for lexical matching, particularly for entity-rich queries. Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) provides an effective unsupervised method to combine ranked lists from heterogeneous retrieval systems without requiring training data.

Query Reformulation. Query rewriting for conversational search transforms context-dependent utterances (Aryal and Pant, 2025; Aryal and Prioleau, 2023) into standalone queries (Katsis et al., 2025). Hypothetical Document Embeddings (HyDE) (Gao et al., 2023) improves zero-shot dense retrieval by using an LLM to generate a hypothetical answer passage, whose embedding is then used as the retrieval query. Query decomposition breaks complex questions into simpler sub-queries that can be searched independently (Bae, 2025; Tiwari et al., 2025; Aryal and Pant, 2025).

3 System Description

Our system consists of three pipelines corresponding to the three MTRAG subtasks. Figure 1 overviews the Task C pipeline, which subsumes Tasks A and B.

3.1 System Protocol

The entire pipeline is **training-free**: no model is fine-tuned at any stage. The pipeline spans two execution environments. *Locally* (Apple M3 Mac, MPS backend), the dense encoder (BGE-base-en-v1.5), FAISS index search, BM25 retrieval, and cross-encoder reranker (BGE-reranker-large) all run on-device. *Remotely*, a single LLM—Llama-3.3-70B-Instruct (Meta AI, 2024), served via the HuggingFace Inference API in `float16` precision

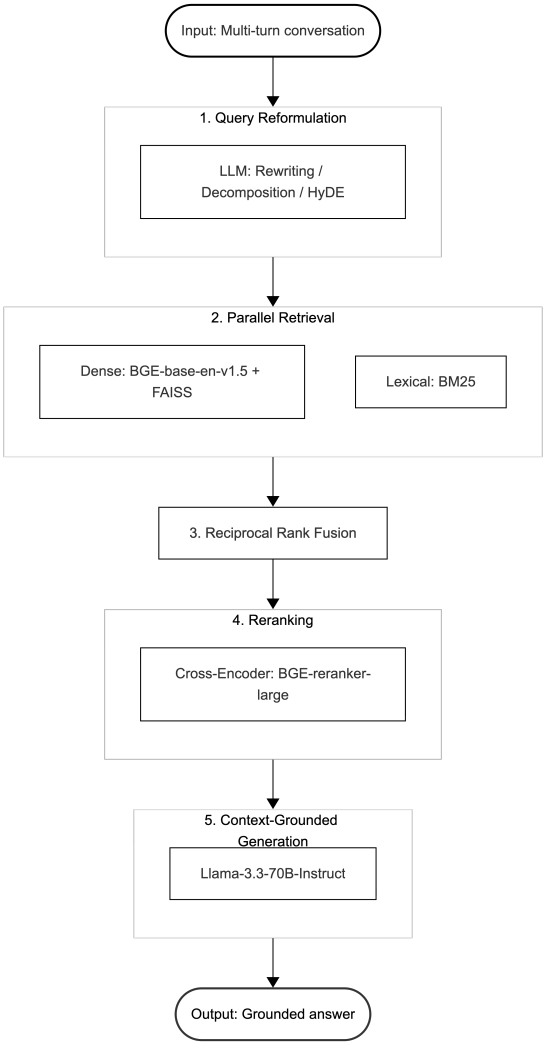

Figure 1: Task C end-to-end RAG pipeline architecture. Tasks A and B use subsets of this pipeline.

without quantization—handles all query reformulation (rewriting, HyDE, decomposition) and final answer generation. Each input produces exactly **one generation** per step. Query reformulation steps execute **sequentially**: rewriting runs first, followed by dense retrieval, BM25 retrieval, HyDE generation with its dense search, and query decomposition with sub-query searches. Cross-encoder reranking is applied **once per turn** after RRF fusion; there are no iterative retrieval loops. Table 1 provides the full system configuration.

3.2 Document Corpora

The MTRAG benchmark provides four document collections (Table 2), pre-processed into passages of 512 tokens with 100-token overlap using the ELSER tokenizer (Katsis et al., 2025).

Table 1: System configuration.

Parameter	Value
Generator / Query LLM	Llama-3.3-70B-Instruct (remote)
LLM serving	HF Inference API, float16
Embedding model	BAAI/bge-base-en-v1.5 (local)
Reranker model	BAAI/bge-reranker-large (local)
Local hardware	Apple M3 Mac (MPS)
Temp (rewrite / decomp / gen)	0.1
Temp (HyDE)	0.3
Max tokens (generation)	200 (Task B) / 256 (Task C)
Max tokens (rewrite)	256
Max tokens (HyDE)	150
Dense top-k (Task A / C)	50 / 100
BM25 k_1 / b / top-k	1.5 / 0.75 / 100
RRF k	60
HyDE top-k	50
Decomp sub-query top-k	33
Rerank pool / output	50 / 10 (Task A), 75 / 5 (Task C)
Generations per input	1
Retry / fallback	Raw last turn on rewrite failure

Table 2: MTRAG document corpora.

Corpus	Domain	Docs	Passages
ClapNQ	Wikipedia	4,293	183,408
FiQA	Finance	57,638	61,022
Govt	Government	7,661	49,607
Cloud	Technical docs	8,578	72,442
Total		**78,170**	**366,479**

3.3 Task A: Passage Retrieval

Our retrieval pipeline follows a two-stage *retrieve-then-rerank* architecture. In Stage 1, all passages and queries are encoded using BGE-base-en-v1.5 (Xiao et al., 2024) (768-dim), L2-normalized and indexed with FAISS (Johnson et al., 2019) (IndexFlatIP). We retrieve the top 50 candidates using the benchmark-provided query rewrites, which convert multi-turn questions into standalone queries (Katsis et al., 2025). In Stage 2, candidates are reranked using BGE-reranker-large (Xiao et al., 2024), a cross-encoder that computes fine-grained query-passage relevance scores, outputting the top 10 passages.

3.4 Task B: Generation with Reference Passages

Task B provides gold reference passages, isolating the generation component. Before generation, an LLM-based coreference resolution step rewrites the current question to be self-contained by resolving pronouns to their referents from conversation history. We use Llama-3.3-70B-Instruct with an extractive grounding prompt that instructs the model to: (1) answer using *only* the provided passages, (2) copy exact phrases when possible, (3) keep answers concise (1–3 sentences), and (4) respond with a fixed abstention phrase when context is insufficient. We use temperature = 0.1 and max tokens = 200.

Max Token Rationale. Task B uses 200 max tokens, matching the concise 1–3 sentence answers the benchmark expects; shorter outputs improve Bert-K-Precision by reducing extraneous content. Task C uses 256 tokens to accommodate noisier retrieved passages. Development-set experiments confirmed that exceeding 256 tokens reduced RB_{agg} through verbosity.

3.5 Task C: End-to-End RAG

Task C combines retrieval and generation. We extend Task A with three query reformulation strategies and hybrid retrieval:

Query Reformulation.

- **LLM-driven rewriting:** Llama-3.3-70B-Instruct rewrites the last user turn into a self-contained question, resolving pronouns and implicit references (temp = 0.1).
- **Query decomposition:** Complex questions are decomposed into 2–3 simpler sub-queries, each searched separately (temp = 0.1).
- **HyDE** (Gao et al., 2023): The LLM generates a short hypothetical answer passage (2–3 sentences), whose BGE embedding serves as an additional retrieval query (temp = 0.3).

Decomposition Rationale. We limit decomposition to 2–3 sub-queries because each incurs a separate retrieval call (top-33 each), so 3 sub-queries contribute $\sim$99 candidates—comparable to the dense budget of 100. Most MTRAG questions involve at most two or three information needs; over-decomposing into 4+ fragments risks losing the question's cohesion and retrieving passages that miss the comparative intent.

Retrieval and Fusion. Task C runs parallel retrieval: dense search (BGE + FAISS, top 100), BM25 lexical search (k_1=1.5, b=0.75, top 100), HyDE-based dense search (top 50), and decomposed sub-query searches (top 33 each). All lists are merged via RRF (Cormack et al., 2009) (k=60), the top 75 fused candidates are reranked by the cross-encoder, and the top 5 passages are passed to the Task B generator.

4 Experimental Setup

4.1 Data

We use the MTRAG benchmark (Katsis et al., 2025) for development: 110 human-generated conversations (842 tasks) across four domains. The test set is drawn from the MTRAG-UN benchmark (Rosenthal et al., 2026a), which contains new conversations emphasizing challenging phenomena including underspecified questions, non-standalone questions, and unanswerable queries across six domains. Our system is evaluated on the four original MTRAG domains. The test set comprises 507 evaluation tasks, of which 332 are answerable or partially answerable (evaluated for Task A retrieval) and all 507 are evaluated for Tasks B and C generation. Ground-truth answerability labels were not released with the test set.

4.2 Evaluation Metrics

Task A uses nDCG@k and Recall@k (official metric: nDCG@5). Tasks B and C use three IDK-conditioned metrics: RB_{agg} (harmonic mean of Bert-Recall, Bert-K-Precision, Rouge-L) (Adlakha et al., 2024), RB_{llm} (LLM judge adapted from RAD-Bench) (Kuo et al., 2025), and RL_{F} (RAGAS Faithfulness) (Es et al., 2024). All generation metrics are conditioned on an IDK classifier that first determines whether the response contains a substantive answer or an abstention. The official ranking metric for Tasks B and C is the harmonic mean of RB_{agg}, RL_{F}, and RB_{llm}.

5 Results

5.1 Task A: Retrieval

Table 3 presents retrieval results on the MTRAG-UN test set (Rosenthal et al., 2026a), evaluated on 332 answerable/partially answerable queries. We achieve nDCG@5 of **0.4098** (ranked 22nd of 38 submissions), improving over the BGE-base-en-v1.5 + query rewrite baseline (0.34) (Katsis et al., 2025) by 20%, demonstrating the value of cross-encoder reranking. Cloud achieves the highest nDCG@5 (0.4736), while FiQA is most challenging (0.2786) due to the informal, subjective nature of financial forum posts (Katsis et al., 2025).

The top submission achieved 0.5776 nDCG@5; our lower score reflects using a general-purpose encoder without domain adaptation. Our primary contribution is the strong end-to-end Task C performance through query reformulation and faithful generation (Section 6.1).

Table 3: Task A retrieval results by domain (test set).

Domain	N	nDCG@5	R@5	R@10
ClapNQ	83	0.4218	0.4667	0.5394
Cloud	86	0.4736	0.5046	0.6050
FiQA	58	0.2786	0.3165	0.4364
Govt	105	0.4205	0.4489	0.5605
Overall	**332**	**0.4098**	0.4446	0.5451

Table 4: Generation scores by domain for Tasks B and C.

	Task B			Task C		
Domain	RB_{ag}	RB_{lm}	RL_{F}	RB_{ag}	RB_{lm}	RL_{F}
ClapNQ	0.486	0.600	0.650	0.381	0.517	0.586
FiQA	0.506	0.716	0.726	0.272	0.473	0.654
Govt	0.513	0.653	0.696	0.372	0.543	0.624
Cloud	0.622	0.745	0.831	0.461	0.608	0.722

5.2 Tasks B and C: Generation

On Task B (reference passages), our system achieves a harmonic mean of **0.7462** (RB_{agg} = 0.6291, RL_{F} = 0.8540, RB_{llm} = 0.7937), ranking 9th of 26 submissions and surpassing the top baseline (gpt-oss-120b, 0.639) by 16.8%. The top-performing submission achieved 0.7827. The high RL_{F} (0.854) confirms that extractive grounding with conservative generation parameters effectively reduces hallucination. On Task C (end-to-end), the harmonic mean is **0.5796** (RB_{agg} = 0.4427, RL_{F} = 0.7507, RB_{llm} = 0.6310), **ranking 2nd of 29 submissions**, surpassing the top baseline (qwen-30b-a3b-thinking, 0.5366) by 8.0% and coming within 1.1% of the top submission (0.5861).

Table 4 shows per-domain generation scores. Cloud consistently achieves the highest scores, while FiQA is most challenging across both tasks. The drop from Task B to C (0.7462 $\rightarrow$ 0.5796) quantifies the impact of retrieval quality on generation.

6 Analysis

6.1 Why Task C Succeeds Despite Moderate Retrieval

Our most notable result is the gap between moderate Task A retrieval (22nd/38) and strong Task C end-to-end RAG (2nd/29). Three factors explain this. First, Task C draws candidates from *four* parallel channels (dense, BM25, HyDE, decomposition) fused via RRF, providing substantially higher re-

Table 5: Retrieval ablation on development set (842 queries). Rows are intermediate pipeline runs, not controlled single-variable ablations.

Configuration	nDCG@5	R@5	R@10
Dense only	0.262	0.227	0.290
Dense + BM25 (RRF)	0.334	0.296	0.360
Full pipeline	0.381	0.330	0.423

call diversity than Task A's single-channel pipeline. Each channel addresses different failure modes—BM25 captures exact entities, HyDE bridges vocabulary gaps, decomposition ensures coverage across multiple information needs—and RRF is robust to noise from any individual channel (Aryal and Adhikari, 2023; Aryal et al., 2023; Cormack et al., 2009). Second, extractive generation compensates for imperfect retrieval: even when the top-5 passages include noise, the 70B model identifies relevant passages while conservative temperature (0.1) prevents fabrication, maintaining high RL_F (0.751). Third, effective prompt-driven abstention on 27.6% of queries avoids hallucination penalties on the IDK-conditioned metrics.

Table 5 reports retrieval ablation results from intermediate development-set runs (842 queries). BM25 fusion provides the largest single retrieval gain (+27.5% relative over dense-only), followed by query reformulation (+14.1% over dense+BM25). Cross-encoder reranking drives the improvement from the development-set dense baseline (0.262) to the test-set score (0.41)—a 20% gain over the BGE baseline (0.34) (Katsis et al., 2025). Using a single 70B model (Meta AI, 2024) for all LLM tasks ensures consistent language understanding, with strong instruction-following enabling reliable abstention and extractive generation (Katsis et al., 2025).

6.2 Abstention Behavior

Our system abstains on 149/507 test queries (29.4%) for Task B and 140/507 (27.6%) for Task C, exceeding the ∼19% unanswerable proportion in the test set, suggesting some false abstentions on partially answerable queries. Abstention is prompt-driven (Ngueajio et al., 2025; Aryal and Prioleau, 2024): the generation prompt instructs the model to respond with a fixed phrase when context is insufficient, without a separate IDK classifier.

6.3 Latency, Sensitivity, and Limitations

Task A runs entirely locally (Apple M3 Mac) in 3–4 hours. Task C latency is dominated by up to 4 remote LLM calls per turn (rewrite, HyDE, decomposition, generation); no conversation-level caching was implemented. RRF k=60 and top-k allocations were selected on the development set; we found no meaningful sensitivity to k in the range 40–80 (Cormack et al., 2009). Key limitations include: no conversation-level caching, simple regex-based BM25 tokenization (no stemming), a single LLM for all tasks rather than specialized models, latency from sequential reformulation calls, and intermediate-run ablations rather than controlled single-variable experiments.

7 Conclusion

We presented a training-free hybrid retrieve-then-rerank pipeline (Aryal and Pant, 2025; Aryal et al., 2023; Prioleau et al., 2025) for multi-turn RAG, submitted to all three subtasks of SemEval-2026 Task 8 (Rosenthal et al., 2026b). Dense and BM25 retrieval are fused via RRF and refined by cross-encoder reranking, while a single open-source 70B model handles query reformulation and extractive generation without fine-tuning. The system ranks 2nd/29 on Task C (H-Mean = 0.5796), within 1.1% of the top submission, demonstrating that multi-signal query reformulation and faithful generation with appropriate abstention can compensate for moderate retrieval quality. Future work includes fine-tuning the dense encoder, conversation-level caching, a dedicated IDK classifier, and controlled ablation studies. All code and configurations are available at `https://github.com/shrsijan/mt-rag-benchmark.git`.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

Saurav Aryal and Mildness Akomoize. 2025. Howard university-ai4pc at semeval-2025 task 3: Logit-based supervised token classification for multilingual hallucination span identification using xgbod. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1790–1794.

Saurav Aryal and Kritika Pant. 2025. Howard university-ai4pc at semeval-2025 task 9: Using open-

weight bart-mnli for zero shot classification of food recall documents. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1919–1923.

Saurav Aryal and Howard Prioleau. 2023. Howard university computer science at semeval-2023 task 12: A 2-step system design for multilingual sentiment classification with language identification. In *Proceedings of the 17th international workshop on semantic evaluation (SemEval-2023)*, pages 2153–2159.

Saurav K Aryal and Howard Prioleau. 2024. Ad-hoc ensemble approach for detecting adverse drug events in electronic health records. *Journal of Computing Sciences in Colleges*, 40(3):238–249.

Saurav K Aryal, Ujjawal Shah, Howard Prioleau, and Legand Burge. 2023. Ensembling and modeling approaches for enhancing alzheimer's disease scoring and severity assessment. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1364–1370. IEEE.

Saurav Keshari Aryal and Gaurav Adhikari. 2023. Evaluating impact of emoticons and pre-processing on sentiment classification of translated african tweets.

Ho Bae. 2025. A study on enhancing zero-shot dense retrieval using query and hypothetical document embedding combination. *The Transactions of the Korea Information Processing Society*, 14(3):161–171.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2174–2184.

Gordon V Cormack, Charles LA 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.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar R Zaïane, Mo Yu, Edoardo M Ponti, and Siva Reddy. 2022. Faithdial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*, pages 150–158.

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.

Desta Haileselassie Hagos, Saurav Keshari Aryal, Patrick Ymele-Leki, and Legand L Burge. 2025. Ai-driven multimodal colorimetric analytics for biomedical and behavioral health diagnostics. *Computational and structural biotechnology journal*, 27:2219–2232.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. 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. mt RAG: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. *Transactions of the Association for Computational Linguistics*, 13:784–808.

Tzu-Lin Kuo, FengTing Liao, Mu-Wei Hsieh, Fu-Chieh Chang, Po-Chun Hsu, and Da-Shan Shiu. 2025. Radbench: Evaluating large language models' capabilities in retrieval augmented dialogues. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 868–902.

Meta AI. 2024. Llama 3.3.

Mikel K Ngueajio, Saurav Aryal, Marcellin Atemkeng, Gloria Washington, and Danda Rawat. 2025. Decoding fake news and hate speech: A survey of explainable ai techniques. *ACM Computing Surveys*, 57(7):1–37.

Ronak Pradeep, Nandan Thakur, Sahel Sharifymoghaddam, Eric Zhang, Ryan Nguyen, Daniel Campos, Nick Craswell, and Jimmy Lin. 2025. Ragnarök: A reusable rag framework and baselines for trec 2024 retrieval-augmented generation track. In *European Conference on Information Retrieval*, pages 132–148. Springer.

Howard Prioleau, Saurav K Aryal, and Jeremy Blackstone. 2025. Leveraging large language models for adverse drug event detection: A comparative study of token and span-based named entity recognition. In *Biocomputing 2026: Proceedings of the Pacific Symposium*, pages 205–218.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.

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.

Saharsha Tiwari, Saurav K Aryal, and Legand Burge. 2025. Enhancing geospatial reasoning in large language models: An optimized retriever approach using r-tree-based point-in-polygon and nearest neighbor search. In *International Conference on Information and Communication Technology for Intelligent Systems*, pages 509–523. Springer Nature Singapore Singapore.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.