QurSciOnto-Bench: A Cross-Lingual Benchmark for Ontology-Guided and Multihop Scientific Tafsir Retrieval

Anonymous ACL submission

Abstract

Retrieval over Islamic scholarly corpora remains difficult without structural grounding. To evaluate ontology-aware retrieval and reranking in scientific Quranic exegesis, we construct QurSciOnto-Bench over the QurSci-Onto scientific Tafsir corpus, containing 438 Arabic and English queries (factoid, conceptual, and multihop). We utilize GPT-4o-mini as the relevance oracle to label 6,515 query-chunk pairs on a 5-point Likert scale, as open-source judge models exhibited unstable evaluation behavior in this domain. We evaluate multiple retrieval configurations, including BM25, bi-encoders, hybrid RRF, Cross-Encoders, MaxSim Token Interaction (MTI), and Autoregressive Reranking (ARR), with ontology routing. We show that ontology-family filter substantially improves retrieval for scientific Quranic exegesis, improving bi-encoder Hit@3 from 0.531 to 0.792 and MRR from 0.503 to 0.759. These gains are larger than those obtained through reranking methods alone, suggesting that for morphologically complex Arabic corpora, structural routing contributes more to retrieval effectiveness than reranking architecture alone. Among reranking approaches, the Cross-Encoder achieves the strongest overall semantic reranking performance (nDCG@10=0.898). ARR fine-tuning with the Qwen2.5-3B model degrades performance, while the Qwen2.5-7B model yields only marginal gains over the pretrained baseline. Furthermore, stratified analysis of 146 multihop queries, evaluated against 3,278 query-chunk pairs, demonstrates that pretrained ARR maintains stable performance across both intra-node and inter-node multihop queries, with task-specific fine-tuning yielding modest additional gains.

1 Introduction

Retrieval over Islamic scholarly exegesis (Tafsir) remains challenging for modern retrieval systems. Arabic morphology, orthographic variation, and dense domain-specific terminology complicates semantic matching between Arabic and English queries against Arabic scholarly corpora (Oshallah et al., 2025). These challenges are particularly difficult for lexical retrieval pipelines operating over scientific Islamic terminology (Al-Rasheed et al., 2025). In scientific Quranic exegesis, retrieval is further complicated by hierarchically related scientific concepts, overlapping terminology between adjacent ontology nodes, and cross-verse reasoning requirements.

Scientific Quranic exegesis remains relatively underexplored as a retrieval benchmark setting (Al-Azani et al., 2026). Prior work has primarily focused on ontology construction (Suryana et al., 2018) and lexical question answering (Alan et al., 2025), without systematically evaluating modern retrieval and reranking approaches on ontology-structured Islamic scholarly corpora. In particular, autoregressive reranking (ARR) (Rozonoyer et al., 2026), which has shown strong ranking capabilities in recent retrieval research, has not previously been evaluated in an Islamic scholarly retrieval setting.

Building on the QurSci-Onto framework (Rashid et al., 2026) we construct the benchmark containing 438 Arabic and English queries spanning factoid, conceptual, and multihop retrieval settings, including a specialized multihop subset of 146 queries stratified into intra-node and inter-node reasoning structures. Using these queries we construct 6,515 labeled query-chunk pairs using GPT-4o-mini and evaluate lexical (BM25), dense (bi-encoder), hybrid (RRF), and reranking approaches including Cross-Encoders, MaxSim Token Interaction (MTI), and Autoregressive Reranking (ARR) under ontology routing with ontology-family filtering (a metadata filtering step that restricts retrieval to relevant ontology families[1] as described in Section 4). The

[1] An ontology family is a named scientific concept cluster in

Table 1: Comparison of related Quranic NLP and retrieval frameworks with our ontology-guided retrieval and reranking framework.

Resource / Framework	Primary Focus	Tafsir Grounding	Ontology-Guided Retrieval	Multihop Evaluation	Scientific Tafsir Focus	Cross-Lingual Retrieval
Quranic Arabic Corpus	Morphology / Syntax	✗	✗	✗	✗	✗
Semantic Quran	Multilingual RDF Linking	✗	✓	✗	✗	✗
QurAna / PropBank-style Resources	Anaphora / Semantic Roles	✗	✗	✗	✗	✗
Scientific Miracle Ontology	Scientific Concept Modeling	✗	✓	✗	✗	✓
OntologyRAG-Q	General Tafsir QA / RAG	✓	✓	✗	✗	✗
QurSciOnto-Bench (Ours)	**Scientific Tafsir Retrieval**	✓	✓	✓	✓	✓

multihop subset is further stratified into intra-node queries (where supporting evidence comes from a single ontology family) and inter-node queries (where evidence spans multiple ontology families), containing 3,278 query-chunk pairs. We make the following contributions:

- QurSciOnto-Bench, a multilingual retrieval benchmark for scientific Quranic exegesis containing 438 Arabic and English queries with 6,515 graded relevance labels.
- A comparative analysis of lexical retrieval, dense retrieval, hybrid retrieval, and modern reranking methods under ontology routing settings.
- A systematic evaluation of autoregressive reranking (ARR) on an Islamic scholarly retrieval corpus.
- An empirical analysis of LLM judge reliability for Arabic domain-specific relevance labeling across open-source and proprietary models.
- A stratified multihop evaluation separating intra-node and inter-node reasoning structures for ontology-aware retrieval analysis.

Code to reproduce all experiments is anonymously available[2].

2 Literature Review

Table 1 summarizes related Quranic NLP and retrieval frameworks. Early work on Quranic NLP focused primarily on ontology construction lexical search, and RAG.

Quranic NLP and Information Retrieval. Early work on Quranic NLP focused primarily on ontology construction and lexical search. Suryana et al. (2018) surveys ontology development for the Quran, identifying open challenges in semantic representation of Quranic concepts. Alan et al. (2025) demonstrates that integrating RAG with LLMs improves reliability in religious question answering by grounding generated responses in verified Islamic scholarly sources. Al-Rasheed et al. (2025) evaluates RAG pipelines for Arabic lexical information retrieval, finding that domain-specific embeddings outperform general-purpose models on Arabic morphological tasks. Similar domain-specific retrieval effects have been observed beyond the Islamic domain: Cho et al. (2026) finds that retrieval configuration choices substantially affect performance in materials science RAG, while Morand et al. (2026) shows that Cross-Encoder distillation preserves ranking quality at reduced cost. The closest prior work, OntologyRAG-Q (Al-Azani et al., 2025), introduces ontology-guided retrieval for Quranic exegesis and evaluates it as part of a generation-oriented QA pipeline using metrics such as BLEU and BERTScore. However, prior Quranic and Islamic-domain retrieval work lacks large-scale graded relevance evaluation, multihop retrieval analysis, and controlled comparison across lexical, dense, hybrid, and reranking architectures independent of generation quality. QurSciOnto-Bench addresses this gap directly.

Dense Retrieval and Reranking. Dense retrieval using bi-encoders (Karpukhin et al., 2020) encodes queries and documents independently into fixed vectors compared by dot product, enabling efficient approximate nearest-neighbor search at corpus scale. Multilingual extensions such as E5 (Wang et al., 2024) extend this to cross-lingual settings using contrastive pretraining on multilingual pairs. Cross-Encoders (Nogueira and Cho, 2020) jointly process query-document pairs through a transformer, often achieving stronger ranking accuracy through full attention interaction at the cost of precomputation. ColBERT (Khattab and Zaharia, 2020) introduces scoring via MaxSim and late interaction by retaining token-level embeddings. This allows near Cross-Encoder ranking quality while preserving much of the efficiency of bi-encoder re-

QurSci-Onto (e.g., BEE_BIO), grouping records that share the same scientific topic nodes like BEE_BIO_01, BEE_BIO_02.

[2] https://anonymous.4open.science/r/QurSci-Onto-bench-E3E4

trieval through offline document indexing. Reciprocal Rank Fusion (Cormack et al., 2009) combines rankings from multiple retrieval systems without requiring score normalization, providing robust hybrid retrieval improvements.

Generative and Autoregressive Retrieval. Generative retrieval (Tay et al., 2022) eliminates the explicit index by training a model to directly generate document identifiers (docIDs) conditioned on a query. Neural Corpus Indexer (Wang et al., 2023) extends this with prefix-aware decoding and hierarchical docID construction. Rozonoyer et al. (2026) proposes Autoregressive Reranking (ARR), providing a theoretical analysis showing that autoregressive models are capable of modeling ranking dependencies beyond those captured by dual encoders. The work also introduces SToICaL, which extends token-level supervision with item-level calibrated loss designed to better model ranking behavior beyond standard next-token prediction. ARR has not been previously evaluated on Islamic scholarly corpora; we present the first such evaluation on an ontology-structured scientific exegesis retrieval.

3 Benchmark and Retrieval Data Construction

The QurSci-Onto dataset (Rashid et al., 2026) is a hierarchical ontology framework for scientific Quranic exegesis constructed from two scientific Tafsir sources: *Tafsir Ayat al-Kawniyah* (El-Naggar, 2001) and *I'jaz al-'Ilmi fi al-Quran*. The framework integrates Quranic verses, thematic categories, scientific concepts, exegesis summaries, and Tafsir indices across three hierarchical layers consisting of 194 Ayah ontology records, 260 exegesis records, and 73 scientific ontology nodes.

Tafsir index records were digitized using Surya-OCR (Paruchuri and Team, 2025). Quranic verses were replaced post-OCR with verified text from an authenticated Quranic corpus to eliminate digitization errors.

3.1 Corpus Indexing

We maintain two ChromaDB (Chroma, 2023) vector stores: an ontology database containing 398 ontology-linked retrieval records, and an exegesis database containing 855 chunk-level Tafsir records. All text is encoded with intfloat/multilingual-e5-large (Wang et al., 2024)[3].

[3]https://huggingface.co/intfloat/e5-large-v2

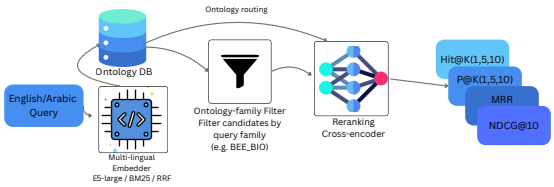

Figure 1: Ontology-family routing and filtering pipeline used in Group A experiments (route_top_k=1).

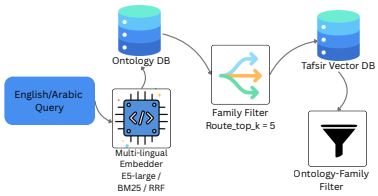

Figure 2: Ontology-guided Tafsir Retrieval setup for Group B and Group C experiments (route_top_k=5).

3.2 Ontology DB Construction

We construct an ontology-enriched retrieval database by joining the exegesis layer with Ayah ontology and scientific ontology metadata through TafsirID associations. At this stage, only Tafsir summaries, page-indexed references, and ontology metadata are retained; full OCR-extracted Tafsir text is not included.

Because a single TafsirID can be associated with multiple ontology families and scientific nodes, the join expands 260 Tafsir records into 398 ontology-linked retrieval records.

Exegesis Layer $\bowtie$ Ayah Ontology $\bowtie$ Scientific Ontology $\rightarrow$ 398 ontology-linked records.

3.3 Exegesis DB Construction

We construct a separate chunk-level exegesis database containing full scientifically relevant Tafsir documents required for chunk-level reranking experiments. Of the 260 Tafsir records, 29 TafsirIDs are selected due to their association with the 73 scientific ontology nodes in QurSci-Onto, then recursively segment the full Arabic Tafsir documents into 855 chunk-level records using a RecursiveCharacterTextSplitter.

Scientific Ontology Nodes $\rightarrow$ Select TafsirIDs $\rightarrow$ Recursive Chunking $\rightarrow$ Exegesis DB

3.4 Automated Relevance Annotation

We label query-chunk pairs using GPT-4o-mini on a 5-point Likert scale (1 = irrelevant, 2 = weakly related, 3 = contextually related, 4 = strongly relevant, 5 = direct answer evidence), producing 6,515 labeled pairs across 438 queries.

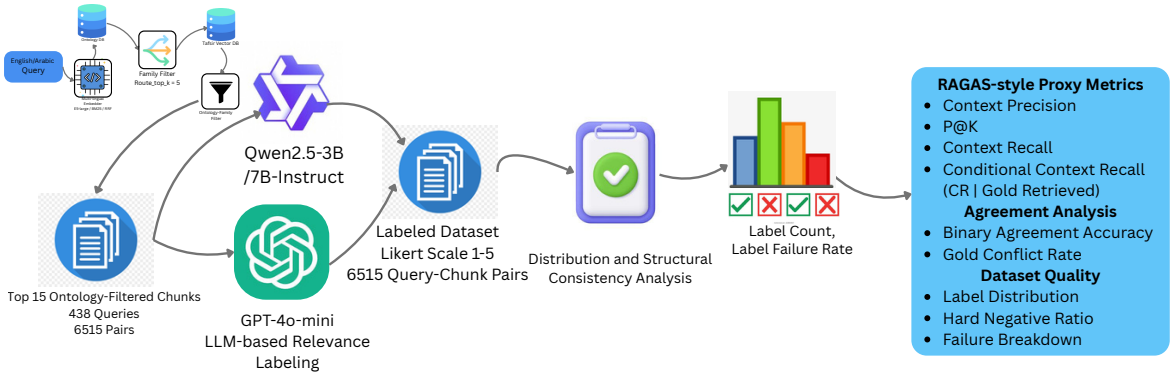

Figure 3: Construction pipeline for GPT-4o-mini relevance labeling and evaluation dataset generation used in Group B experiments.

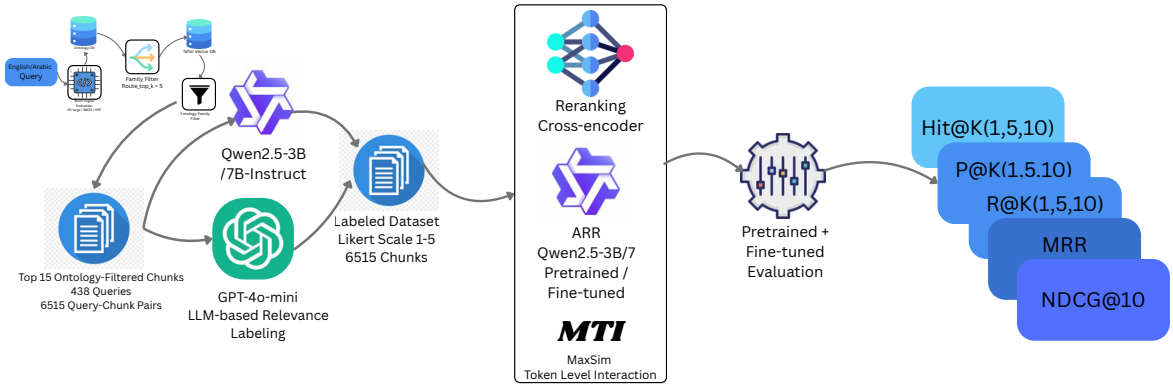

Figure 4: Lexical, dense, hybrid, and reranking pipelines evaluated under ontology-familly filtering configuration (route_top_k=5) in Group C.

3.5 Multihop Queries Labeling

Of the 438 queries, 146 are multihop queries (73 Arabic, 73 English). Each multihop query is decomposed into 2–5 independently retrievable sub-queries using GPT-4o-mini. Candidates are retrieved separately per sub-query with ontology-family filtering enabled, merged using RRF (k=60), deduplicated, and capped at 30 candidates per query. These candidates are then labeled by GPT-4o-mini on 5-point Likert scale.

Multihop queries are classified post-labeling based on where relevant evidence is found:

- **Intra-node**: supporting label-4/5 evidence remains within the same ontology family ($n = 42$), such as retrieval confined to BEE_BIO-related ontology nodes.
- **Inter-node**: supporting evidence spans multiple ontology families ($n = 104$), such as linking BEE_BIO and HIVE_ARCH ontology families within a single reasoning chain.

4 Ontology-Guided Retrieval and Reranking Framework

4.1 Ontology Routing and Filtering

For Group A retrieval experiments, queries are encoded using multilingual-e5-large and searched against the ontology database. A global candidate pool of 200 ontology-linked retrieval records is retrieved using cosine similarity. In the baseline setting, retrieval is performed without structural filtering. In the ontology-family setting, only candidates whose scientific ontology family matches the query family (e.g., BEE_BIO) are retained we call this ontology-family routing (route_top_k=1). The filtered candidates are optionally reranked using a Cross-Encoder, and the final top-10 results are evaluated using node-match metrics.

4.2 Two-Stage Ontology-Guided Tafsir Retrieval

Given a query, we first encode it with the bi-encoder and search the ontology database to retrieve the most semantically similar ontology records. Ontol-

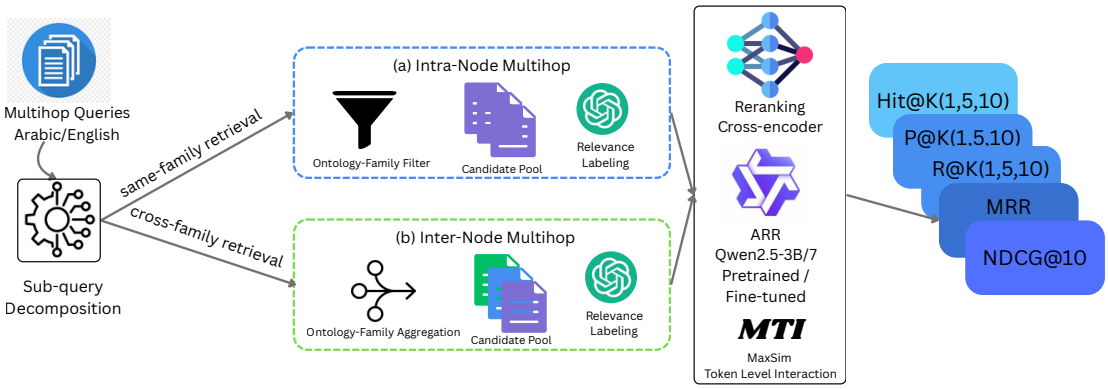

Figure 5: Specialized multihop retrieval pipeline separating intra-node and inter-node reasoning settings for topology-aware reranking evaluation.

ogy families are then ranked according to how frequently they appear among the retrieved records, and the top-5 ontology families (route_top_k=5) are selected. This routing step restricts retrieval to topically relevant ontology families before retrieval from the larger exegesis database.

In the second stage, the query embedding retrieves a pool of 200 chunks from the exegesis database. An ontology-family filter retains only chunks whose node_id belongs to the routed (route_top_k=5) ontology families, removing structurally unrelated candidates as shown in Figure 2. The top-15 chunks by cosine similarity are returned for reranking evaluation with Cross-Encoder, MTI, and ARR models.

4.3 Reranking

Cross-Encoder. We evaluate BGE-M3-reranker as a Cross-Encoder baseline. The Cross-Encoder processes each (query, chunk) pair jointly by concatenating them as input, enabling full transformer attention between query and document tokens for precise relevance estimation at higher inference cost than bi-encoder retrieval.

MaxSim Token Interaction (MTI). We implement a lightweight MaxSim Token Interaction (MTI) baseline inspired by ColBERT-style MaxSim interaction (Khattab and Zaharia, 2020) using frozen multilingual-e5-large token embeddings instead of full ColBERT token-level training. Unlike standard dense retrieval, MTI retains token-level embeddings for both queries and documents rather than pooling them into a single vector representation. Document embeddings are precomputed and cached offline. At query time, relevance is computed via MaxSim: for each query token, the maximum cosine similarity over all document tokens is computed, and these similarities are summed across query tokens. This enables token-level interaction while remaining computationally cheaper than Cross-Encoder inference through offline document indexing and precomputed document embeddings.

Autoregressive Reranking. Unlike pairwise reranking methods, ARR reformulates ranking as constrained autoregressive generation over document identifiers. Following (Rozonoyer et al., 2026), each chunk is assigned a hierarchical docID derived from ontology metadata in the format category.subcategory.leaf. Category and subcategory identifiers are derived from ontology-family metadata, while leaf identifiers are assigned using deterministic SHA-1 content hashing. A prefix trie is constructed over valid docID sequences to constrain autoregressive generation to valid candidate identifiers only.

Candidate chunks are presented to Qwen2.5-7B-Instruct in sliding windows of size 5 with stride 3. For each window, a catalog prompt lists docIDs paired with truncated chunk text. Each docID is scored by the mean token log-probability:

$$\text{score}(d \mid q) = \frac{1}{T} \sum_{t=1}^{T} \log P(w_t \mid q, w_{<t})$$

When a candidate appears in multiple overlapping windows, its final score is the mean log-probability across all windows in which it is visible. Higher scores indicate stronger semantic relevance between the query and candidate chunk.

4.4 Multihop Retrieval Configuration

For multihop experiments, each query is decomposed into 2–5 independently retrievable sub-

queries using GPT-4o-mini. Candidates are retrieved separately per sub-query with ontology-family filtering enabled, then merged and deduplicated prior to reranking. The candidate pool is expanded to 30 chunks compared to 15 in standard retrieval settings, to improve cross-family evidence coverage. Cross-Encoder, MTI, and ARR rerankers are evaluated under both intra-node and inter-node configurations.

5 Experiments

We organize experiments into three sequential groups. Group A evaluates ontology routing and filtering using node-match retrieval. Group B evaluates the reliability of open-source models as relevance judges, motivating GPT-4o-mini as the labeling strategy. Group C provides the primary semantic reranking evaluation under fine-grained Likert-scale relevance labels. In Group A, unfiltered BM25 and bi-encoder retrieval serve as standard retrieval baselines. Ontology-filtered retrieval and reranking, evaluated in Group C under graded relevance labels, represent the proposed contributions of this work.

5.1 Group A: Node-Match Retrieval

Group A retrieves records from the Ontology DB, as shown in Figure 1. A retrieval is considered correct when the returned record shares the same ontology node_id as the query ground-truth ontology node. We evaluate BM25, Bi-encoder, and Bi-encoder + BM25 RRF retrieval systems under both filtering ON and filtering OFF settings, with Cross-Encoder reranking. We additionally evaluate BM25 Stemmed to measure the effect of lexical normalization on sparse retrieval performance. Filtering configurations are described in Section 4.1. We report Hit@3, Hit@5, Hit@10, MRR, P@3, P@5, P@10, and nDCG@10.

5.2 Group B: Judge Reliability and Annotation

For each query, candidate chunks are first retrieved using the two-stage ontology-guided retrieval pipeline described in Section 4.2 and illustrated in Figure 2. The top-15 retrieved chunks are then evaluated by judge models using a structured relevance annotation prompt. This process produces 6,515 labeled query-chunk pairs across 438 queries, forming the relevance benchmark used in Group C evaluation, as shown in Figure 4. Label definitions and query construction are described in Section 3.4.

5.3 Group C: Semantic Reranking

Group C evaluates reranking performance using the labeled query-chunk pairs constructed in Group B. These query-chunk pairs are already stored in bi-encoder retrieval order; reranking is therefore performed over identical candidate pools using Cross-Encoder, MTI, and ARR models.

All systems operate with ontology-family filter enabled. All 438 queries rerank the top-15 filtered chunks.

We evaluate Bi-encoder retrieval together with MTI, BGE-M3-reranker, pretrained ARR (Qwen2.5-7B-Instruct), and ARR fine-tuned with SToICaL supervision (Section 4.3). ARR fine-tuning uses LoRA adaptation (Hu et al., 2021) with rank $r = 8$, $\alpha = 16$, AdamW optimizer with learning rate 1×10^{-4}, gradient accumulation over 4 steps, norm clipping at 1.0, and a single training epoch over 6,515 labeled pairs on a single A100 40GB GPU.

We report R@1, R@3, R@5, MRR, nDCG@10, and Hit@k ($k = 1, 3, 5$), where Hit@k is the fraction of queries with at least one relevant document (label ≥ 4) in the top-k.

SToICaL (Rozonoyer et al., 2026) scales each training instance using a rank-aware weight $w(r) = 1/r^{\alpha}$, where r is the document position in the relevance-sorted candidate list and $\alpha = 2.0$. The final objective is a weighted cross-entropy loss over target docID token sequences, placing greater emphasis on highly relevant documents.

5.4 Implementation Details

BM25 uses rank-bm25 with standard English tokenization and character-level trigrams for Arabic. All dense embeddings use intfloat/multilingual-e5-large (1,024 dimensions) stored as dense vectors in ChromaDB. ARR inference runs Qwen2.5-7B-Instruct in 4-bit quantization. All ARR inference experiments use greedy decoding.

6 Results and Discussion

6.1 Ontology Routing and Filtering

Enabling ontology-family filter improves bi-encoder Hit@3 from 0.531 to 0.792 and MRR from 0.503 to 0.759 over ontology routing, as shown in Table 2. These gains substantially exceed those obtained through Cross-Encoder reranking alone (Hit@3=0.595 unfiltered), indicating that structural routing constrains the search space to relevant thematic neighborhoods in a way that reranking can-

Table 2: Group A: Node-Match Retrieval on All 438 Queries.

System	Filter	nDCG@10	Hit@3	Hit@5	Hit@10	MRR	P@3	P@5	P@10
BM25	✗	0.413	0.418	0.511	0.596	0.359	0.184	0.143	0.091
+ Cross-Encoder rerank	✗	0.477	0.504	0.593	0.652	0.422	0.219	0.163	0.101
BM25 Stemmed	✗	0.508	0.493	0.626	0.772	0.425	0.221	0.178	0.114
Bi-encoder	✗	0.563	0.531	0.664	0.799	0.503	0.239	0.189	0.132
+ Cross-Encoder rerank	✗	0.570	0.595	0.712	0.805	0.501	0.262	0.200	0.130
Bi-encoder + BM25 RRF	✗	0.543	0.566	0.671	0.769	0.485	0.243	0.188	0.122
+ Cross-Encoder rerank	✗	0.567	0.591	0.710	0.792	0.500	0.261	0.197	0.128
BM25	✓	0.661	0.680	0.735	0.795	0.625	0.342	0.236	0.140
+ Cross-Encoder rerank	✓	0.697	0.737	0.780	0.817	0.665	0.370	0.246	0.143
Bi-encoder	✓	0.789	0.792	0.861	0.913	0.759	0.421	0.289	0.171
+ Cross-Encoder rerank	✓	0.786	0.817	0.867	0.913	0.753	0.431	0.289	0.169
Bi-encoder + BM25 RRF	✓	0.766	0.771	0.851	0.906	0.732	0.410	0.287	0.166
+ Cross-Encoder rerank	✓	0.785	0.815	0.865	0.913	0.753	0.432	0.288	0.169

Table 3: Preliminary judge-model analysis on 50 queries with 15 retrieved chunks per query (750 query–chunk pairs)

Judge Model	Binary Agreement	Structural Failures	Gold-Chunk Conflict Rate
Qwen2.5-3B	0.76	363	75.0%
Qwen2.5-7B	0.76	294	83.9%
GPT-4o-mini	**0.78**	**0**	**50.0%**

not replicate independently. Lexical baselines reveal a pronounced cross-lingual disparity. BM25 Stemmed applies standard English stemming and character-level trigrams for Arabic, as Arabic morphological complexity makes word-level stemming unreliable without a dedicated morphological analyzer. BM25 Stemmed achieves Hit@3=0.652 for English but only 0.333 for Arabic, reflecting the morphological complexity of Arabic corpora. Dense embeddings bridge this gap under ontology filtering, where Arabic bi-encoder Hit@3 rises from 0.488 to 0.776. Under filtering, the binary node-match ceiling is reached quickly. Adding RRF and Cross-Encoder reranking improves filtered Hit@3 only marginally (0.792→0.821), suggesting that node-level correctness cannot adequately reward fine-grained within-family semantic reordering. This motivates the graded Likert-scale evaluation introduced in Group C.

6.2 LLM Judge Reliability

Table 3 presents preliminary judge-model analysis on 50 queries with 15 retrieved chunks per query. Both open-source models exhibited substantial structural failures during JSON generation, with Qwen2.5-3B and Qwen2.5-7B failing structurally on 363 and 294 of 750 query–chunk pairs respectively. Beyond formatting instability, both models produced high gold-chunk conflict rates (75.0% and 83.9%), meaning they frequently assigned irrelevant scores to chunks whose ontology node_id directly matches the query ground-truth node, indicating unreliable semantic judgment in this domain. GPT-4o-mini produced zero structural failures, achieved the highest binary agreement (0.78), and yielded the lowest conflict rate (50.0%). Consequently, GPT-4o-mini was selected as the primary relevance judge for all Group C labeling experiments.

6.3 Reranking Under Fine-Grained Relevance Labels

Table 4 shows reranking performance under graded relevance labels. The Cross-Encoder achieves the strongest overall performance (nDCG@10=0.898, MRR=0.811), confirming that full-attention interaction provides meaningful within-family semantic reordering beyond bi-encoder retrieval order.

MTI produces limited gains (nDCG@10: 0.852→0.857), as multilingual-e5-large token embeddings are not trained with a token-level contrastive objective.

ARR using Qwen2.5-7B achieves competitive performance (nDCG@10=0.885, MRR=0.801), with SToICaL fine-tuning yielding only marginal additional gains (nDCG@10=0.888). The 3B model regresses slightly after fine-tuning (nDCG@10: 0.872→0.870), indicating insufficient capacity for effective ranking distillation.

6.4 Multihop Retrieval Performance

Table 5 reveals substantial differences between intra-node and inter-node multihop retrieval. Inter-node queries achieve consistently stronger baseline performance (bi-encoder MRR: 0.806 vs 0.583).

Table 4: Group C: Semantic reranking evaluation on all 438 queries using GPT-4o-mini relevance labels. The bi-encoder with ontology filtering serves as the retrieval baseline; all rerankers operate on the same top-15 filtered candidate pool. Best results per metric in **bold**.

Method	R@1	R@3	R@5	nDCG@10	MRR	Hit@1	Hit@3	Hit@5	P@1	P@3	P@5
Bi-encoder baseline	0.107	0.274	0.409	0.852	0.738	0.625	0.840	0.892	0.625	0.566	0.522
+ Cross-Encoder (BGE-M3)	**0.132**	**0.333**	**0.491**	**0.898**	**0.811**	**0.732**	**0.878**	**0.913**	**0.732**	**0.663**	**0.610**
MTI	0.112	0.281	0.415	0.857	0.756	0.655	0.844	0.897	0.655	0.576	0.529
ARR (Qwen2.5-3B)	0.112	0.282	0.442	0.872	0.772	0.680	0.840	0.901	0.680	0.598	0.561
ARR (Qwen2.5-3B + SToICaL)	0.108	0.278	0.439	0.870	0.765	0.664	0.840	0.901	0.664	0.595	0.555
ARR (Qwen2.5-7B)	0.127	0.317	0.474	0.885	0.801	0.710	0.876	0.924	0.710	0.635	0.582
ARR (Qwen2.5-7B + SToICaL)	0.128	0.323	0.470	0.888	0.808	0.719	0.883	0.922	0.719	0.643	0.583

Table 5: Detailed multihop reranking performance across pretrained and fine-tuned rerankers under global, inter-node, and intra-node evaluation settings.

Subset	Model	R@1	R@3	R@5	MRR	nDCG@10	Hit@1	Hit@3	Hit@5	P@1	P@3	P@5
Global	Bi-encoder	0.062	0.207	0.300	0.742	0.813	0.561	0.938	0.986	0.561	0.586	0.524
	MTI	0.070	0.188	0.294	0.754	0.818	0.623	0.863	0.952	0.623	0.566	0.527
	Cross-Encoder (pretrained)	0.082	0.243	**0.380**	0.836	0.853	0.719	0.952	0.979	0.719	0.700	**0.668**
	Cross-Encoder (fine-tuned)	0.074	0.186	0.312	0.780	0.819	0.650	0.904	0.965	0.650	0.538	0.531
	ARR (Qwen2.5-7B)	**0.104**	0.253	0.368	**0.912**	0.878	**0.849**	0.979	0.993	**0.849**	0.726	0.638
	ARR (+ SToICaL)	0.101	**0.257**	0.376	0.906	**0.880**	0.835	**0.986**	**1.000**	0.835	**0.732**	0.641
Inter-node	Bi-encoder	0.072	0.227	0.324	0.806	0.841	0.653	**0.980**	0.990	0.653	0.660	0.586
	MTI	0.077	0.205	0.322	0.806	0.833	0.682	0.923	0.980	0.682	0.625	0.582
	Cross-Encoder (pretrained)	0.083	0.243	**0.384**	0.858	0.867	0.750	0.971	0.990	0.750	0.740	**0.707**
	Cross-Encoder (fine-tuned)	0.064	0.175	0.301	0.752	0.806	0.605	0.894	0.961	0.605	0.535	0.538
	ARR (Qwen2.5-7B)	**0.101**	0.247	0.363	**0.922**	0.882	**0.865**	**0.980**	0.990	**0.865**	0.743	0.659
	ARR (+ SToICaL)	0.099	**0.251**	0.373	0.918	**0.884**	0.855	**0.980**	**1.000**	0.855	**0.753**	0.663
Intra-node	Bi-encoder	0.037	0.157	0.243	0.583	0.744	0.333	0.833	0.976	0.333	0.404	0.371
	MTI	0.052	0.145	0.226	0.627	0.782	0.476	0.714	0.880	0.476	0.420	0.390
	Cross-Encoder (pretrained)	0.082	0.243	0.371	0.781	0.818	0.642	0.904	0.952	0.642	0.603	0.571
	Cross-Encoder (fine-tuned)	0.098	0.214	0.342	0.848	0.851	0.761	0.928	0.976	0.761	0.547	0.514
	ARR (Qwen2.5-7B)	**0.112**	0.269	0.381	**0.886**	0.869	**0.809**	0.976	**1.000**	**0.809**	**0.682**	**0.585**
	ARR (+ SToICaL)	0.107	**0.270**	**0.383**	0.876	**0.870**	0.785	**1.000**	**1.000**	0.785	**0.682**	**0.585**

Unlike general reranking where the Cross-Encoder leads overall, ARR (Qwen2.5-7B) achieves the strongest multihop ranking performance, reaching an MRR of 0.887 on intra-node and 0.922 on inter-node queries. Cross-Encoder fine-tuning improves intra-node MRR (0.781 $\rightarrow$ 0.848) but sharply degrades inter-node MRR (0.858 $\rightarrow$ 0.752) and global performance (MRR: 0.836 $\rightarrow$ 0.780), indicating overspecialization toward localized semantic matching. Pretrained ARR and ARR+SToICaL maintain stable performance across both settings.

7 Conclusion

We presented QurSciOnto-Bench, a cross-lingual retrieval benchmark for scientific Quranic exegesis. We demonstrated that ontology routing with filtering is the most influential retrieval factor, improving bi-encoder Hit@3 from 0.531 to 0.792 and MRR from 0.503 to 0.759, gains that substantially exceed those obtained from reranking alone. Among rerankers, the Cross-Encoder achieves the highest nDCG@10 (0.898), while pretrained ARR leads on multihop reasoning (MRR: 0.887–0.922). Fine-tuning ARR and Cross-Encoders yields only modest or specialized gains, often at the cost of cross-family generalization. The evaluated open-source Qwen2.5 models (3B and 7B) were unreliable as relevance judges in this domain. These findings suggest that structural routing is particularly valuable for retrieval in morphologically complex and ontology-structured scholarly domains.

8 Limitations

This work has several limitations. First, GPT-4o-mini relevance labels have not been validated against human annotations; systematic label bias cannot be ruled out and human validation is left for future work. Second, all experiments are single-run without statistical significance testing. ARR fine-tuning gains of +0.001 to +0.002 nDCG@10 are within the range of run-to-run variance and should be interpreted with caution. Third, our MTI imple-

mentation uses multilingual-e5-large token embeddings without per-token contrastive training, representing a lower bound on true late-interaction performance. Fourth, the benchmark covers a single Tafsir source; generalization to other exegetical traditions, Hadith collections, or Arabic scholarly genres requires dedicated benchmark construction and evaluation.

9 Ethics Statement

QurSci-Onto and Tafsir Ayat Al Konia are published scholarly works used strictly for retrieval research. No personal data was collected or processed. No human annotators were employed in the current version; relevance labeling was performed automatically using GPT-4o-mini. Relevance labels do not represent theological judgments. Researchers applying these methods to religious corpora should be aware that automated relevance scoring may not capture scholarly or doctrinal nuance. Human expert review is advisable before deployment in any application involving religious interpretation.

References

Sadam Al-Azani, Shadi Abudalfa, and Hussein Samma. 2026. A comprehensive review of artificial intelligence in qur'anic research: Trends, methods, challenges, and future directions. *Information Processing Management*, 63(7, Part B):104870.

Sadam Al-Azani, Maad Alowaifeer, Alhanoof Alhunief, and Ahmed Abdelali. 2025. OntologyRAG-Q: Resource development and benchmarking for retrieval-augmented question answering in qur'anic tafsir. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15540–15558, Suzhou, China. Association for Computational Linguistics.

Raghad Al-Rasheed, Abdullah Al Muaddi, Hawra Aljasim, Rawan Al-Matham, Muneera Alhoshan, Asma Al Wazrah, and Abdulrahman AlOsaimy. 2025. Evaluating RAG pipelines for Arabic lexical information retrieval: A comparative study of embedding and generation models. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 155–164, Abu Dhabi, UAE. Association for Computational Linguistics.

Ahmet Yusuf Alan, Enis Karaarslan, and Omer Aydin. 2025. Improving llm reliability with rag in religious question-answering: Mufassirqas. *Preprint*, arXiv:2401.15378.

Zen Han Cho, Matthew Osvaldo, Sayan Doloi, Maloy Das, Jun Ci Goh, Bo Sheng Tan, Jiali Wang, Yujia Li, Xingchi Xiao, Amrita Joshi, and Leonard Wei Tat Ng. 2026. Large language models in materials science: assessing rag evaluation frameworks through graphene synthesis. *RSC Adv.*, 16:11306–11313.

Chroma. 2023. Chroma: the ai-native open-source embedding database.

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*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Zaghloul El-Naggar. 2001. *Tafsir al-Ayat al-Kawniyah fi al-Qur'an al-Karim [Scientific Exegesis of the Cosmic Verses in the Quran]*. Dar al-Ma'rifah, Beirut, Lebanon. 4 Volumes.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

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.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *Preprint*, arXiv:2004.12832.

Victor Morand, Mathias Vast, Basile Cooten, Laure Soulier, Josiane Mothe, and Benjamin Piwowarski. 2026. Reproducing and comparing distillation techniques for cross-encoders.

Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage re-ranking with bert. *Preprint*, arXiv:1901.04085.

Islam Oshallah, Mohamed Basem, Ali Hamdi, and Ammar Mohammed. 2025. Cross-language approach for quranic qa. In *Proceedings of Tenth International Congress on Information and Communication Technology*, pages 385–396, Singapore. Springer Nature Singapore.

Vikas Paruchuri and Datalab Team. 2025. Surya: A lightweight document ocr and analysis toolkit. https://github.com/VikParuchuri/surya. GitHub repository.

Ibad-ur-Rehman Rashid, Junaid Hussain, and Sadam Al-Azani. 2026. QurSci-onto: A hierarchical ontology and dataset for scientific exegesis in the Quran. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script*, pages 149–165, Rabat, Morocco. Association for Computational Linguistics.

Benjamin Rozonoyer, Chong You, Michael Boratko, Himanshu Jain, Nilesh Gupta, Srinadh Bhojanapalli, Andrew McCallum, and Felix Yu. 2026. Autoregressive ranking: Bridging the gap between dual and cross encoders. *Preprint*, arXiv:2601.05588.

N. Suryana, Mohd Azmi, and Fandy Setyo Utomo. 2018. Quran ontology: Review on recent development and open research issues. *Journal of Theoretical and Applied Information Technology*, 96:568–581.

Yi Tay, Vinh Q. Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. 2022. Transformer memory as a differentiable search index. *Preprint*, arXiv:2202.06991.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Hao Sun, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, Xing Xie, Hao Allen Sun, Weiwei Deng, Qi Zhang, and Mao Yang. 2023. A neural corpus indexer for document retrieval. *Preprint*, arXiv:2206.02743.

A Tables

B Human Validation and Agreement Analysis

Human agreement analysis using Cohen's κ on a stratified subset of query-chunk pairs will be reported in the final version.

C Error analysis

Qualitative error analysis of ARR rank changes is presented in Table 8.

D Figures

E Prompt Templates

E.1 Query Generation Prompt

We use Gemini 3 Pro to generate bilingual factoid, conceptual, and multihop retrieval queries from each scientific ontology node.

Act as an expert evaluator for an Agentic RAG system focused on scientific Quranic Exegesis.

You need to generate high-quality test queries in both English and Arabic for the following scientific ontology node.

=== NODE CONTEXT ===

Node ID: {NODE_ID}

Scientific Concept: {SCIENTIFIC_CONCEPT}

Broad Categories: {BROAD_CATEGORIES}

Quranic Themes: {THEMES}

Associated Ayahs: {AYAHS}

Ayah Translations: {TRANSLATIONS}

Tafsir Summaries: {TAFSIR_SUMMARIES}

==========

Generate the following query types:

1. Factoid Query: A direct question retrieving a specific scientific fact or phenomenon explicitly mentioned in the Tafsir summaries.

2. Conceptual Query: A question explaining the broader meaning, purpose, or mechanism of the scientific concept.

3. Multihop Query: A complex question requiring reasoning across ontology layers by connecting Quranic themes, categories, or scientific concepts.

Return ONLY valid JSON with English and Arabic variants for all query types.

Table 6: Detailed Retrieval Performance Breakdown by Language and Query Type

Configuration	Filter	Hit@3	Hit@5	Hit@10	MRR	nDCG@10	P@3	P@5	P@10
Language: English (n=219)									
BM25	✗	0.616	0.735	0.821	0.554	0.614	0.287	0.220	0.133
BM25 Stemmed	✗	0.652	0.762	0.904	0.583	0.657	0.312	0.235	0.144
BM25 + Cross-Encoder	✗	0.662	0.767	0.835	0.566	0.627	0.301	0.222	0.139
BM25	✓	0.780	0.826	0.872	0.732	0.761	0.401	0.275	0.158
BM25 + Cross-Encoder	✓	0.799	0.840	0.876	0.725	0.758	0.409	0.274	0.157
Bi-encoder	✗	0.575	0.707	0.840	0.546	0.605	0.263	0.210	0.142
Bi-encoder + Cross-Encoder	✗	0.625	0.753	0.849	0.542	0.611	0.283	0.218	0.142
Bi-encoder	✓	0.808	0.872	0.917	0.780	0.804	0.441	0.302	0.173
Bi-encoder + Cross-Encoder	✓	0.821	0.876	0.917	0.767	0.797	0.441	0.298	0.173
Bi-encoder + BM25 RRF	✓	0.826	0.885	0.922	0.795	0.816	0.449	0.309	0.174
Bi-encoder + BM25 RRF + Cross-Encoder	✓	0.821	0.876	0.917	0.767	0.795	0.441	0.295	0.173
Language: Arabic (n=219)									
BM25	✗	0.219	0.287	0.369	0.163	0.211	0.080	0.065	0.049
BM25 Stemmed	✗	0.333	0.488	0.639	0.266	0.358	0.129	0.120	0.083
BM25 + Cross-Encoder	✗	0.347	0.420	0.470	0.279	0.327	0.138	0.104	0.063
BM25	✓	0.579	0.643	0.716	0.516	0.560	0.283	0.196	0.121
BM25 + Cross-Encoder	✓	0.675	0.721	0.757	0.604	0.636	0.331	0.217	0.128
Bi-encoder	✗	0.488	0.621	0.757	0.459	0.520	0.216	0.168	0.122
Bi-encoder + Cross-Encoder	✗	0.566	0.671	0.762	0.459	0.529	0.242	0.182	0.118
Bi-encoder	✓	0.776	0.849	0.908	0.737	0.772	0.400	0.276	0.167
Bi-encoder + Cross-Encoder	✓	0.812	0.858	0.908	0.740	0.775	0.421	0.279	0.164
Bi-encoder + BM25 RRF	✓	0.716	0.817	0.890	0.669	0.716	0.372	0.264	0.159
Bi-encoder + BM25 RRF + Cross-Encoder	✓	0.808	0.853	0.908	0.739	0.774	0.421	0.280	0.165
Query Type: Factoid (n=146)									
BM25	✗	0.390	0.500	0.589	0.307	0.376	0.148	0.115	0.073
BM25 Stemmed	✗	0.431	0.575	0.794	0.340	0.451	0.164	0.134	0.097
BM25 + Cross-Encoder	✗	0.486	0.595	0.650	0.352	0.429	0.191	0.146	0.085
BM25	✓	0.664	0.719	0.801	0.597	0.643	0.324	0.219	0.128
BM25 + Cross-Encoder	✓	0.732	0.773	0.815	0.627	0.668	0.353	0.230	0.132
Bi-encoder	✗	0.431	0.589	0.760	0.386	0.471	0.175	0.152	0.108
Bi-encoder + Cross-Encoder	✗	0.513	0.664	0.780	0.387	0.484	0.205	0.161	0.107
Bi-encoder	✓	0.746	0.849	0.910	0.686	0.739	0.385	0.275	0.160
Bi-encoder + Cross-Encoder	✓	0.780	0.856	0.917	0.696	0.747	0.406	0.272	0.160
Bi-encoder + BM25 RRF	✓	0.732	0.842	0.904	0.686	0.735	0.381	0.273	0.155
Bi-encoder + BM25 RRF + Cross-Encoder	✓	0.773	0.849	0.917	0.694	0.745	0.404	0.269	0.159
Query Type: Conceptual (n=146)									
BM25	✗	0.390	0.513	0.636	0.316	0.390	0.148	0.123	0.089
BM25 Stemmed	✗	0.479	0.664	0.821	0.386	0.491	0.187	0.164	0.111
BM25 + Cross-Encoder	✗	0.458	0.595	0.691	0.369	0.448	0.175	0.142	0.101
BM25	✓	0.664	0.726	0.780	0.589	0.629	0.308	0.206	0.128
BM25 + Cross-Encoder	✓	0.691	0.753	0.808	0.591	0.643	0.317	0.213	0.128
Bi-encoder	✗	0.438	0.623	0.801	0.423	0.502	0.182	0.156	0.123
Bi-encoder + Cross-Encoder	✗	0.520	0.684	0.794	0.400	0.497	0.200	0.171	0.116
Bi-encoder	✓	0.746	0.828	0.904	0.724	0.756	0.381	0.254	0.159
Bi-encoder + Cross-Encoder	✓	0.773	0.842	0.904	0.683	0.734	0.388	0.258	0.150
Bi-encoder + BM25 RRF	✓	0.767	0.835	0.897	0.715	0.747	0.383	0.258	0.152
Bi-encoder + BM25 RRF + Cross-Encoder	✓	0.773	0.842	0.897	0.682	0.731	0.388	0.258	0.149
Query Type: Multihop (n=146)									
BM25	✗	0.472	0.520	0.561	0.453	0.471	0.255	0.190	0.110
BM25 Stemmed	✗	0.568	0.636	0.698	0.546	0.581	0.310	0.235	0.132
BM25 + Cross-Encoder	✗	0.568	0.589	0.616	0.546	0.554	0.292	0.201	0.117
BM25	✓	0.712	0.760	0.801	0.686	0.709	0.394	0.282	0.161
BM25 + Cross-Encoder	✓	0.787	0.815	0.828	0.776	0.780	0.440	0.294	0.169
Bi-encoder	✗	0.726	0.780	0.835	0.700	0.715	0.360	0.261	0.164
Bi-encoder + Cross-Encoder	✗	0.753	0.787	0.842	0.715	0.728	0.381	0.268	0.167
Bi-encoder	✓	0.883	0.904	0.924	0.865	0.870	0.495	0.338	0.191
Bi-encoder + Cross-Encoder	✓	0.897	0.904	0.917	0.882	0.877	0.500	0.335	0.196
Bi-encoder + BM25 RRF	✓	0.815	0.876	0.917	0.793	0.816	0.468	0.328	0.193
Bi-encoder + BM25 RRF + Cross-Encoder	✓	0.897	0.904	0.924	0.883	0.877	0.502	0.335	0.199

Notes: P@k = Precision at k. Filter indicates ontology-family filter. ✗= filter disabled, ✓= filter enabled. All values represent exact truncation to three decimal places without rounding.

Table 7: Complete multi-hop retrieval performance across all architectures. Intra-node queries (n=42) require evidence from a single ontology family; inter-node queries (n=104) span multiple families. Best per metric in **bold**.

Method	Type	Intra-node (n=42)					Inter-node (n=104)				
		R@1	R@3	R@5	MRR	nDCG@10	R@1	R@3	R@5	MRR	nDCG@10
Bi-encoder baseline	Global	0.062	0.208	0.301	0.742	0.786	—	—	—	—	—
	Stratified	0.037	0.158	0.244	0.584	0.734	0.072	0.228	0.324	0.806	0.808
+ MTI	Global	0.070	0.188	0.295	0.755	0.794	—	—	—	—	—
	Stratified	0.053	0.145	0.226	0.627	0.736	0.077	0.206	0.322	0.807	0.817
Cross-Encoder (pretrained)	Global	0.083	0.244	0.381	0.836	0.842	—	—	—	—	—
	Stratified	0.082	0.244	0.371	0.782	0.816	0.083	0.243	0.384	0.858	0.852
Cross-Encoder (fine-tuned)	Global	0.074	0.186	0.313	0.780	0.811	—	—	—	—	—
	Stratified	0.098	0.214	0.342	0.849	0.826	0.064	0.175	0.301	0.753	0.805
ARR (pretrained Qwen-7B)	Global	0.104	0.254	0.368	0.912	0.856	—	—	—	—	—
	Stratified	0.112	0.269	0.382	0.887	0.844	0.101	0.247	0.363	**0.922**	**0.861**
ARR (fine-tuned + SToICaL)	Global	0.102	0.257	0.377	0.906	0.857	—	—	—	—	—
	Stratified	0.107	0.270	0.383	0.877	0.843	0.100	**0.252**	0.374	0.918	**0.860**

Notes: R@k = Recall at k. Global row shows aggregate over all 146 queries; Stratified rows show intra/inter-node breakdown.

E.2 Relevance Labeling Prompt

We use GPT-4o-mini to assign graded relevance labels to query-chunk pairs using a 5-point Likert scale.

> You are an expert evaluator for Quranic Tafsir retrieval.
>
> Evaluate how relevant the Tafsir chunk is to the user query.
>
> The query may be in English or Arabic; evaluate semantic relevance accurately for both languages.
>
> Scoring rubric (Likert 1–5):
>
> 5 = direct scientific evidence strongly addressing the query
>
> 4 = mostly relevant with strong supporting evidence
>
> 3 = contextually related but no direct evidence
>
> 2 = weak relation / tangential
>
> 1 = irrelevant or only general history
>
> Return ONLY valid JSON with keys: `reasoning`, `relevance_score`, and `confidence`.
>
> Query: {QUERY}
>
> Tafsir Chunk: {CHUNK}

E.3 Multihop Query Decomposition Prompt

For specialized multihop experiments, GPT-4o-mini decomposes each multihop query into independently retrievable sub-queries.

> You are an expert at decomposing complex multihop Quranic queries into simpler sub-queries.
>
> Given a complex query requiring reasoning across multiple Quranic concepts or narratives, decompose it into 2–5 simpler independent sub-queries.
>
> Each sub-query should:
>
> 1. Be retrievable independently
> 2. Avoid redundancy with other sub-queries
> 3. Cover part of the overall reasoning chain
> 4. Preserve the original query language
>
> Return ONLY valid JSON with the following structure:
>
> `{"sub_queries": [...], "reasoning": "..."}`
>
> Original Query:
>
> {MULTIHOP_QUERY}

E.4 Unified Multihop Judgment Prompt

For multihop evaluation, retrieved chunks are judged jointly against the original multihop query and its decomposed sub-queries.

> You are an expert evaluator for multi-hop Quranic Tafsir retrieval.
>
> A complex query was decomposed into simpler sub-queries and retrieved together.
>
> Evaluate how relevant each Tafsir chunk is to solving the overall multihop query.
>
> Scoring rubric (Likert 1–5):
>
> 5 = directly supports the reasoning chain
>
> 4 = mostly relevant and supports multiple sub-queries
>
> 3 = contextually related to at least one sub-query
>
> 2 = weakly related
>
> 1 = irrelevant
>
> Original Multihop Query:
>
> {QUERY}
>
> Sub-queries:
>
> {SUB_QUERIES}

Table 8: Qualitative examples of ARR+SToICaL rank changes. Positive shifts indicate promotion toward earlier ranks; negative shifts indicate demotion.

Category	Type	Lang.	Label	Shift	Query	Evidence Chunk
Strong promotion	Multihop	AR	3	$3 \rightarrow 6$	كيف يستخدم القرآن موضوع التنوع البيولوجي لتسليط الضوء على البرمجة الجينية وراء مجتمعات النحل الدقيقة للغاية	البيضة إلى الحشرة الكاملة. هذه الدقة العلمية الفائقة فى الإشارة إلى ما وهب الله (تعالى) النحل من ذكاء فطرى وعلم وهبى تحاول المعارف المكتسبة
Noise suppression	Multihop	EN	2	$4 \rightarrow 8$	How does the Quranic theme of natural processes describe the advection and wind-driving mechanisms that initiate cloud formation?	قُلْ سِيرُوا فِي الْأَرْضِ فَانْظُرُوا كَيْفَ بَدَأَ الْخَلْقَ ثُمَّ اللَّهُ يُنْشِئُ النَّشْأَةَ الْآخِرَةَ إِنَّ اللَّهَ عَلَىٰ كُلِّ شَيْءٍ قَدِيرٌ
Partial win	Multihop	AR	3	$13 \rightarrow 8$	كيف يستخدم القرآن العملية البيولوجية المستمرة لتطور الجنين كإعجاز علمي ونبوي للبشرية؟	ؤية بعض مراحل الأجنة فى بيض الدجاج المخصب، وظن أن البيضة تحتوى فرخلبيضة إلى الحشرة الكاملة. هذه الدقة العلمية الفائقة فى الإشارة إلى ما وهب الله (تعالى) النحل من ذكاء فطرى وعلم وهبى تحاول المعارف المكتسبة
Toxic promotion	Factoid	EN	2	$12 \rightarrow 8$	According to the exegesis, where are bees granted the freedom to choose the locations of their homes?	البيضة إلى الحشرة الكاملة. هذه الدقة العلمية الفائقة فى الإشارة إلى ما وهب الله (تعالى) النحل من ذكاء فطرى وعلم وهبى تحاول المعارف المكتسبة اليوم الكشف عن
Loss of nuance	Multihop	AR	3	$3 \rightarrow 7$	كيف يسلط موضوع الإعجاز العلمي النبوي الضوء على ظاهرة التطبق الكثافي في كل من حدود المياه العذبة-المالحة والمياه البحرية المتجاورة؟	اسم منطقة توزيع ماء النهر تعمل على مزيد من الفصل بين الأنواع الثلاثة من الماء الموجود فى مصبات الأنهار وهى: الماء العذب،

Retrieved Chunks:

{CHUNKS}

Return structured JSON judgments for all chunks.

QUERY–CHUNK LABELING SNAPSHOT

Example evaluation dataset snapshot detailing both English and Arabic query variants alongside GPT-4o-mini oracle relevance classifications.

Rk	Node IDs	Query Type	Cosine Similarity	Label	Language	Retrieved Evidence	LLM Judgment Summary	Ontology Family
1	BEE_BIO_01 HIVE_ARCH_01	Factoid	0.577	5	English	"...الإشارة إلى ما وهب الله النحل من ذكاء فطرى وعلم وهبى تحاول المعارف المكتسبة اليوم الكشف عن..."	Directly answers query about the innate intelligence, awareness, and instinctive skills of bees.	BEE_BIO
1	BEE_BIO_01 HIVE_ARCH_01	Factoid	0.728	5	Arabic	...هذه الدقة العلمية الفائقة فى الإشارة إلى ما وهب الله النحل من ذكاء فطرى وعلم وهبى تحاول المعارف المكتسبة...	Directly discusses the innate intelligence and divine knowledge granted to the bee community.	BEE_BIO
4	BEE_BIO_01 HIVE_ARCH_01	Conceptual	0.633	5	English	"...ذكر ربنا قدرته البالغة فى الإيحاء إلى الشغالات من نحل العسل أن تتخذ من الجبال بيوتا ومن الشجر..."	Discusses divine capability and guidance, matching the query intent on the manifestation of power.	BEE_BIO
4	BEE_BIO_01 HIVE_ARCH_01	Conceptual	0.658	4	Arabic	...وقوة على الطيران بسرعات فائقة حتى تغطى أكبر مساحة ممكنة تجنى الرحيق وحبوب اللقاح من أزهار...	Details physical adaptations and flight capabilities linked to divinely endowed instincts and perception.	BEE_BIO
2	BEE_BIO_03 BEE_SENSE_01	Multihop	0.062	3	English	"﴿ وَمِنَ ٱلنَّاسِ وَٱلدَّوَآبِّ وَٱلْأَنْعَامِ مُخْتَلِفٌ أَلْوَانُهُۥ كَذَٰلِكَ... ﴾ والع..."	Mentions general biological diversity and colors, but completely lacks any bee society design specifics.	BEE_BIO
5	BEE_BIO_01 HIVE_ARCH_01	Multihop	0.047	5	Arabic	﴿ وَأَوْحَىٰ رَبُّكَ إِلَى ٱلنَّحْلِ أَنِ ٱتَّخِذِى... ﴾ ...ومنحها القدرات الفطرية على تنظيم مجتمعاتها تنظيما مبهرا دقيقا تتوزع فيه...	Primary cross-family link. Connects the thematic biological order to the precise genetic programming query.	BEE_BIO
11	BEE_BIO_06 PROPOLIS_01	Multihop	0.028	1	English	...يا بني آدم لا يفتننكم الشيطان كما أخرج أبويكم من الجنة ينزع عنهما لباسهما ليريهما سوءاتهما...	Completely irrelevant out-of-domain text segment. Lacks any relationship to biological systems or honeybee communities.	BEE_BIO

Figure 6: English and Arabic query variants sample

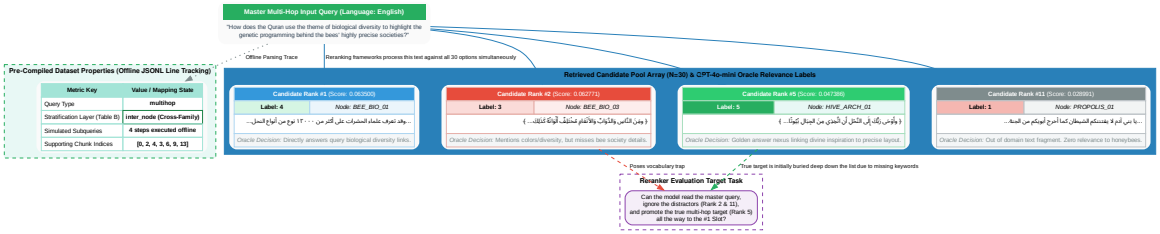

Figure 7: Inter-Node Multihop Retrieval and Reranking Example

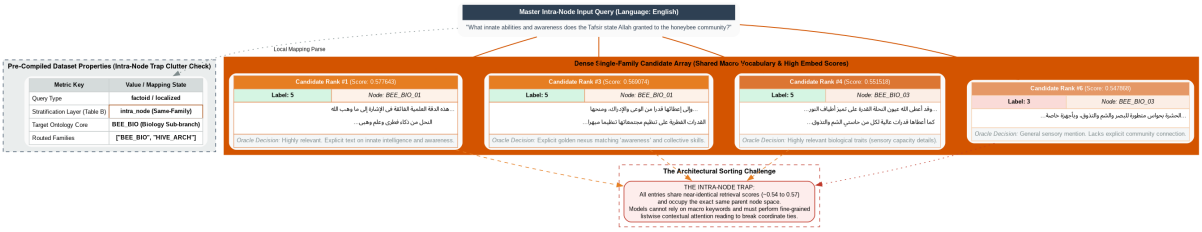

Figure 8: Intra-Node Multihop Candidate Retrieval and Oracle Labeling