QurSci-Onto: A Hierarchical Ontology and Dataset for Scientific Exegesis in the Quran

Ibad-ur-Rehman Rashid[1], **Junaid Hussain**[1], **Sadam Al-Azani**[2]
[1]Government Post Graduate College, Mansehra, Affiliated with Hazara University, Pakistan
[2]SDAIA-KFUPM Joint Research Center for Artificial Intelligence,
King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia
ibad@gcm.edu.pk, junaidbce@gmail.com, sadam.azani@kfupm.edu.sa

Abstract

This paper introduces resources for the computational study of scientific exegesis (Tafsir Ilmi): a structured ontology, a curated dataset of 194 scientifically relevant Quranic verses linked to 260 exegetical records from two authoritative Tafsir books, and an annotation framework that organizes scientific references by topic and sequential context. Existing Quranic resources treat verses as unstructured text, losing the logical order and causal relationships of scientific concepts documented in exegesis. To address this, we present QurSci-Onto, a three-layer ontology that categorizes verses by scientific domain, links them to authoritative Tafsir, and provides a framework for representing sequential processes through stage-based annotations. Our dataset includes page-level citations and covers 8 major scientific topics across 73 nodes. While the full corpus is tagged with broad categories and scientific topics, a specialized subset features granular node-level mappings to capture complex scientific narratives. We release QurSci-Onto as a foundational resource for Arabic semantic NLP and demonstrate that it enables significant improvements in semantic retrieval and enables multi-hop sequential reasoning capabilities over unstructured baselines.

1 Introduction

The application of large language models (LLMs) in sensitive domains from healthcare to law requires rigorous grounding in structured knowledge to mitigate hallucinations (Ji et al., 2023). This is particularly critical in Quranic Studies, where interpretive accuracy and theological nuance are paramount. While recent works have introduced general Quranic Question-Answering (QA) datasets (Malhas and Elsayed, 2020) and morphological ontologies (Dukes and Habash, 2010), a significant gap remains in the domain of Scientific Exegesis (Tafsir Ilmi).

Scientific narratives in the Quran are not merely thematic labels. They typically describe related or sequential processes and are defined by causality and time. For instance, the description of human embryonic development (Surah Al-Mu'minun 23:12-14) follows a strict biological chronology. Existing resources, however, treat these verses as unstructured text bags, losing the embedded logical order. There is a clear and urgent need to develop reliable, transparent, and ethically aligned resources that capture this structural granularity for scientific exegesis (Tafaseer[1]).

This work makes several key contributions:

1. QurSci-Onto: A Three-Layer Ontology: A hierarchical framework linking verses to scientific domains, authoritative Tafsir commentary with page-level citations, and process nodes representing sequential stages or static entities.

2. Scientific Tafsir Index: A curated dataset from two Tafsir books, providing structured exegetical records for RAG applications.

3. Hybrid Annotation Schema: A framework that distinguishes between dynamic processes (using `LogicalOrder` to preserve sequence) and static entities, enabling narrative reconstruction beyond simple keyword matching.

We validate the utility of this schema through a comparative semantic retrieval evaluation, demonstrating that ontological grounding significantly improves the retrieval of scientific concepts by bridging the lexical gap between modern terminology and classical Quranic text. The full annotated dataset, ontology files, and the retrieval codebase are released to support future research in low-resource Quranic NLP.[2]

The rest of this paper is structured as follows. Section 2 reviews related works in ontologies and

[1]Tafaseer is the plural of Tafsir and refers to multiple books of Quranic exegesis (Tafsir)

[2]Github Link: https://github.com/Ebad-urRehman/QurSci-Onto

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 149–165

RAG, Section 3 presents the QurSci-Onto framework, dataset attributes and construction pipeline, Section 4 present Dataset Statistics, Section 5 presents the semantic retrieval evaluation and results. Finally, Section 6 concludes the paper, followed by Ethics Statement, Practical Implications, and a discussion of limitations in Sections 7, 8 and 9.

2 Related Works

2.1 Quranic Ontologies

Previous efforts have largely focused on morpho-syntactic structures. Dukes and Habash (2010) established the Quranic Arabic Corpus, providing granular morphological tagging and dependency trees (Dukes and Buckwalter, 2010). Building on this foundation, Sharaf and Atwell (2012) developed QurAna to resolve pronominal anaphora, while Zaghouani et al. (2012) introduced a pilot PropBank annotation for semantic roles.

Sherif and Ngonga Ngomo (2015) developed the Semantic Quran ontology, enabling multilingual RDF representations. These ontologies primarily address linguistic features (parts of speech, dependencies) or broad thematic categorization like "Living Creatures" by Khan et al. (2013). Additionally, Al-yahya et al. (2010) proposed an ontological model for time nouns, categorizing temporal concepts, though without considering the dynamic sequentiality of processes.

Recent initiatives have expanded the scope of Quranic ontologies toward thematic and heritage modeling. Basharat et al. (2025) (preprint) introduced SemanticTafsir, a knowledge graph derived from Tafsir al-Tabari that models the text as a cultural heritage, focusing on narrator chains (Isnad) and historical context. Similarly, Ta'a et al. (2018) developed a thematic ontology to enhance knowledge retrieval, grouping verses into knowledge themes like 'Faith' or 'Worship'. Al-Azani et al. (2025) introduced OntologyRAG-Q, a Tafsir ontology across 15 books with over 4,200 QA pairs, setting a benchmark for retrieval-augmented generation in general Quranic studies. Tashtoush et al. (2017) proposed a thematic ontology for human and social relations, mapping concepts such as kinship and moral domains across Arabic, English, and Arabizi to support semantic search. Moogab et al., 2021 proposed the Scientific Miracle Ontology (SMO), which uses METHONTOLOGY to categorize scientific facts across multiple domains into Quranic and scientific concept classes. While OntologyRAG-Q provides comprehensive Tafsir grounding for general Quranic QA and the Scientific Miracle Ontology (SMO) categorizes static scientific facts, QurSci-Onto is the first to model sequential scientific concepts with explicit causal-temporal relations.

2.2 RAG in Low-Resource Domains

Domain Retrieval-Augmented Generation (RAG) has emerged as a standard for grounding LLMs (Lewis et al., 2020). However, standard RAG relies on semantic chunking, which often fragments coherent narratives. In the context of Arabic RAG, Al-Rasheed et al. (2025) evaluated various embedding models, but their focus remained on general information retrieval. Al-Azani et al. (2025) introduced the OntologyRAG-Q approach, a Tafaseer-specific RAG retrieval method that performs Ayah (verse) level chunking and enriches each chunk by incorporating relevant ontology information. However, existing retrieval methods do not preserve sequence or causal relations in scientific exegesis. Our ontology addresses this gap through causal-temporal semantic enrichment, enabling process-aware queries rather than static fact retrieval.

3 The QurSci-Onto Framework

3.1 Ontology Architecture

The QurSci-Onto framework is built upon three interconnected layers as shown in Figure 1. (1) Ayah Ontology, a relational layer that categorizes verses and maps them to exegesis and scientific domains; (2) Tafsir Index, a structured index of exegetical sources; and (3) Scientific Ontology, a causal-temporal knowledge graph.

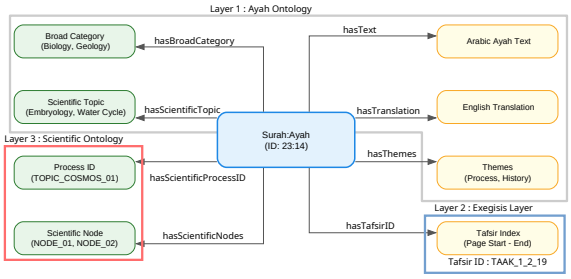

Figure 1: QurSci-Onto Architecture

Table 1: Comparison of QurSci-Onto with existing Quranic NLP resources

Resource / Framework	Primary Domain	Structure Type	Scientific Nodes	Causal Logic	RAG Focus
Quranic Arabic Corpus	Linguistic	Morphological Tree	✗	✗	Syntax
Semantic Quran	General	RDF Knowledge Graph	✗	✗	Multilingual Linking
OntologyRAG-Q	General Tafsir	Thematic Ontology	✗	✗	QA Retrieval
SemanticTafsir	Heritage/History	Knowledge Graph	✗	✗	Cultural Context
Noble Quran Ar ontology	Social/Human	Thematic Ontology	✗	✗	Semantic Search
Quranic Time Lexicon	Temporal (lexical)	Taxonomy of Nouns	✗	✗	Concept Classification
Scientific Miracle Ontology	Scientific	Static Concept Map	✓	✗	Fact Matching
QurSci-Onto (Ours)	**Scientific Exegesis**	**Relation and Process Oriented Graph**	✓	✓	**Narrative Reconstruction**

3.1.1 The Ayah Ontology

The Ayah Ontology serves as the primary annotated corpus, comprising 194 verses selected for their scientific relevance cited in the Ayah Index of Tafsir al-Ayat al-Kawniyah fi al-Quran al-Karim (El-Naggar, 2001). Each record includes the verse Arabic and its English translation, as provided in (Khan, 2019), categorized by broad scientific domains (e.g., Cosmology) and specific phenomena. Beyond simple thematic tagging, this layer acts as the framework's central bridge. It utilizes pointers `hasTafsirID` and `hasScientificConceptID` to link scripture directly to exegetical evidence and ontological nodes.

3.1.2 The Scientific Tafsir Index (Source)

We release a curated dataset that links verses to their Tafsir in specific sections in two authoritative books: Tafsir Ayat al-Kawniyah (4 volumes) and I'jaz al-'Ilmi fi al-Quran (1 volume). This resource serves as a structured index of 260 records, providing page-level citations and expert summaries extracted directly from the books' table of contents for scientific verses. This creates a reliable index for researchers, which can be useful in many RAG applications.

3.1.3 The Scientific Ontology

We introduce a hierarchical Scientific Ontology constructed through a data mining process applied to the source Tafsir Ayat al-Kawniyah. This schema maps scientific concepts spanning processes, natural phenomena, and physical entities to Quranic terminology. The Ontology contains 8 scientific topics decomposed into 73 nodes. Unlike existing morphological taxonomies, this schema models concepts as either static entities with causal relations, or dynamic sequences (defined by temporal progression), enabling systems to distinguish between static entities and evolving processes.

Figure 2 presents an example from the Cosmology topic (BigBang_CosmosExpansion), showing the sequential progression from *Ratq* (Singularity) through *Fatq* (Big Bang) to Musi'un (Cosmic Expansion).

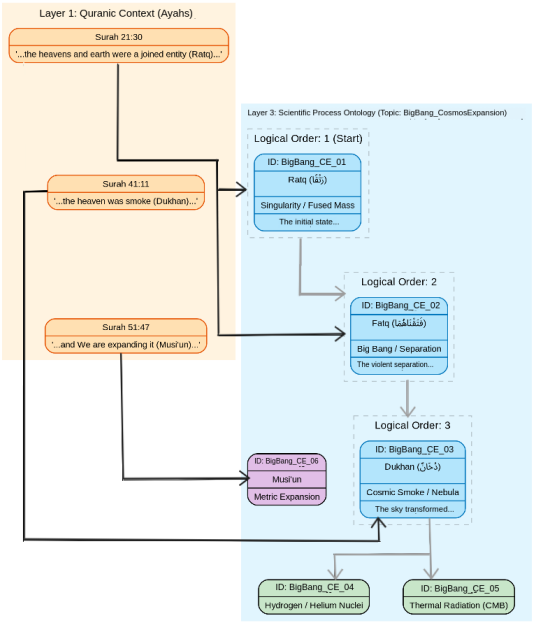

Figure 2: Sequential ontology mapping Quranic terms to Big Bang (cosmology). Nodes represent temporal stages from singularity (Ratq) to cosmic expansion (Musi'un), linked to verses through Tafsir interpretations.

This structure allows for the disambiguation of terms based on their sequential position within a process. For example, *Ratq* (رَتْقًا) is positioned as NODE_01 (BigBang_CosmosExpansion_01), the initial singularity state that precedes the separation event described by *Fatq* (BigBang_CosmosExpansion_02) derived from Tafsir.

Conversely, to model static entities, the ontology uses a central root node connected to interacting components rather than a sequence. In the 'Estuary' topic, the 'Barrier' (Barzakh) is defined as the root node (`ESTUARY_01`). Unlike the dynamic process of the Big Bang, this node does not

Table 2: Core attribute schema for QurSci-Onto. These fields enable hierarchical organization, multilingual support, and process-aware retrieval. Complete definitions with examples appear in Appendix B.

Layer	Attribute	Description
Ayah Ontology	`Surah:Ayah`	Verse citation (e.g., 23:14)
	`hasTafsirID`	Pointer to Layer 2 (e.g., TAAK_1_2_19)
	`hasText`	Original Arabic text in Uthmani script
	`hasTranslation`	English translation by Dr. Mustafa Khattab (Khan, 2019)
	`hasBroadCategories`	List of Broader Scientific categories (Biological, Cosmological)
	`hasScientificTopics`	List of specific topic described (Embryonic Development, Water Cycle)
	`hasThemes`	Themes of the Ayah like Moral or Ethical Reflection.
	`hasScientificConceptID`	Pointer to Layer 3 (Scientific Topic/Process ID, e.g., EMBRYO)
	`hasScientificNodes`	Pointer to Layer 3 granular nodes (EMBRYO_01, EMBRYO_02)
Scientific Tafsir Index	`TafsirID`	Unique identifier (e.g., TAAK_1_2_19)
	`SourceBook`	Authoritative exegesis book (Tafsir al-Ayat al-Kawniyah, I'jaz al-'Ilmi)
	`PageStart-PageEnd`	Page-level citations for auditability
	`ScientificTopicTextArabic`	Summary of scientific phenomenon in Arabic (extracted from book's Table of Contents)
	`ScientificTopicTextEnglish`	Summary of scientific phenomenon in English (translated from Arabic summary)
Scientific Ontology	`hasTafsirID`	Pointer to Layer 2 (e.g., TAAK_1_2_19)
	`hasTopicID`	Scientific topic identifier (e.g., BigBang_CosmosExpansion)
	`hasType`	Classification as Process (sequential stages) or Entity (static phenomena)
	`hasNodeID`	Atomic knowledge unit ID (e.g., BigBang_CosmosExpansion_01)
	`hasRelation`	Causal relationship (transforms_into, causes) with Parent Node.
	`hasParentNode`	Reference to parent node for hierarchical relationships
	`hasQuranicTermArabic`	Original Quranic terminology (e.g., رَتْقًا)
	`hasScientificKeywordsArabic`	Arabic Scientific words and concepts used in source book
	`hasScientificKeywords`	Modern scientific interpretation (e.g., Singularity)
	`hasLogicalOrder`	Sequential position within a process (1, 2, 3...)
	`hasTafsirSummary`	Concise summary of exegetical explanation linking Quranic term to scientific concept

change over time; instead, it links to 'Fresh Water' and 'Salt Water' to model their interaction (Figure 3). This captures the stable physical separation described in Surah 55:20.

3.2 Schema and Attributes

The QurSci-Onto schema is formally defined as a Directed Property Graph (DPG), where nodes represent discrete entities and edges represent typed, semantic relationships. This structural choice enables the explicit representation of hierarchical categorization, exegetical provenance, and causal-temporal sequences. The core attributes are defined in Table 2, with full technical definitions provided in Appendix B.

3.2.1 Core Entities (Nodes)

The graph is composed of three interconnected node types:

- **Ayah:** Represents a Quranic verse as a central entity linked to both exegetical sources and scientific concepts.
- **TafsirRecord:** Represents an entry from an authoritative exegesis book, providing the interpretive source for a scientific claim.
- **ScientificNode:** Represents an atomic unit of scientific knowledge, either a static entity like a mountain or a process stage like the nutfah stage in embryology. Scientific node (`hasScientificNodes`) instances are grouped under a Scientific process

Table 3: Taxonomy of Semantic Relations in QurSci-Onto, categorized by ontological function.

Category	Relation	Function & Example
Causal & Temporal	`causes / caused_by`	Direct causality (Gravity → Star Formation)
	`transforms_into`	Sequential evolution stages (Nutfah → Alaqah)
	`precedes`	Immediate temporal priority (Sperm Motility → Zygote)
	`determines`	Decisive factor for an outcome (XY Chromosomes → Gender)
Structural	`composed_of`	Part-Whole composition (Cosmic Smoke → Nuclei)
	`contained_in`	Spatial containment (Hailstones ⊂ Clouds)
	`surrounds`	Complete encasement (Membranes → Fetus)
	`covers`	Surface layering (Musculature → Bones)
Functional	`performs_function`	Teleological purpose (Propolis → Sterilization)
	`stabilizes`	Maintaining equilibrium (Pegs → Lithosphere)
	`facilitates`	Enabling an action (Navigation → Foraging)
	`regulated_by`	Control mechanism (Pycnocline → Fluid Dynamics)
Interaction	`interacts_with`	Dynamic engagement (Fresh Water ↔ Salt Water)
	`separates_from`	Divergence/Filtration (Scum → Runoff)
	`manifests_as`	Observable form (Charge Separation → Lightning)
Conceptual Mapping	`analogy_to`	Decodes metaphors (Mountains → Pegs)
	`synonym_to`	Terminological equivalence (Frontal Partition → Barrier)

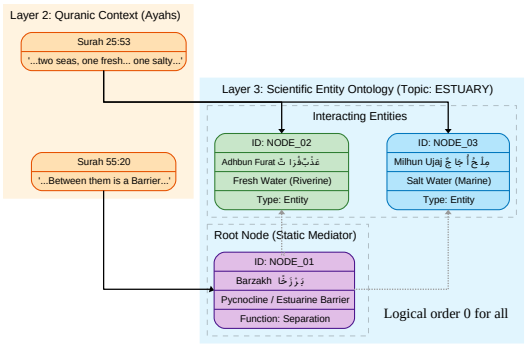

Figure 3: Static ontology for estuarine barrier phenomena. Central node (Barzakh/barrier) mediates between fresh and salt water entities, with verses mapped to physical components.

(`hasScientificConceptID`) like `EMBRYO` for Human Embryonic Development.

The semantic richness of the ontology is achieved through a set of predefined relationship types (edges) that capture exegetical logic. These relations are organized into five semantic categories:

- **Causal & Temporal:** Relations that capture dynamic processes, causality, and temporal sequences in scientific phenomena.
- **Structural:** Relations describing part-whole composition, spatial organization, and physical containment.
- **Functional:** Relations expressing purpose, regulation, and the functional roles of entities or processes.
- **Interaction:** Relations modeling dynamic engagements, separations, and manifestations between entities.
- **Conceptual Mapping:** Relations bridging Quranic terminology with scientific interpretation through metaphor and direct correspondence.

The relations between scientific nodes are formally defined to encode the causal and sequential narratives of scientific exegesis, as specified in Table 3.

3.3 Resource Construction Pipeline

The dataset was constructed in a sequential multi-stage pipeline (as shown in Figure 4), designed to transform raw PDF volumes into a structured knowledge graph.

Phase 1: Source Indexing & Extraction We began by curating 5 volumes of authoritative exegesis. We manually indexed scientifically relevant sections to create the Tafsir Index (Layer 2). This process involved extracting page ranges and topic summaries directly from the source indices, ensuring that every downstream data point could

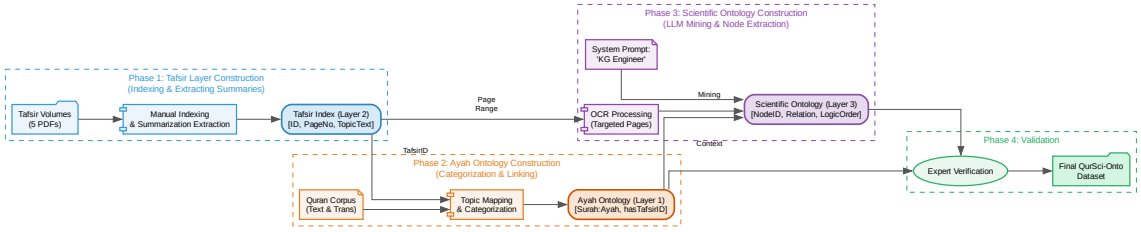

Figure 4: **Resource Construction Pipeline.** The process spans three phases: (1) `Source Indexing & Extraction`, (2) `Semantic Linking & Categorization`, and (3) `Automated Graph Mining`

be traced back to a specific physical page in the source text for auditability.

Phase 2: Semantic Linking & Categorization In this phase, we constructed the Ayah Ontology (Layer 1) by aligning the Quranic corpus with the Tafsir Index. Using the `hasTafsirID` as a pointer, we linked specific verses to their exegetical sources. We employed Large Language Models (LLMs) to sequentially process these entries, classifying them by scientific domain (e.g., Hydrology, Embryology), while validating outputs through manual cross-referencing with authoritative sources to ensure precise contextual analysis.

Phase 3: Automated Graph Mining To build the Scientific Ontology (Layer 3), we applied Optical Character Recognition (OCR) to the topic-specific page clusters identified in Phase 1. The extracted Arabic text was processed by Gemini 3 Pro using a specialized "Scientific Knowledge Graph Engineer" system prompt as shown in Appendix 14. The model was instructed to enforce strict data lineage, extracting:

- **Process Nodes:** Distinct stages of a scientific phenomenon (e.g., *Ratq*, or *Fatq*).
- **Causal Relations:** The logic connecting these nodes (e.g., `transforms_into`).
- **Textual Grounding:** The specific Arabic terminology (`hasScientificKeywordsArabic`) from the text that validates the concept, verified manually against the source.

Expert Validation The annotation process followed a two-stage protocol. Initial annotations were performed by a computational researcher adhering to a strict extraction schema. To ensure theological accuracy, the dataset underwent a secondary validation phase by domain experts from the Faculty of Islamic Studies. This review focused on verifying the general theological plausibility of the Tafsir and scientific concept mappings, ensuring that the interpretations remained within the bounds of accepted exegetical possibilities without introducing semantic distortions. Expert review identified and corrected instances where LLM-assisted extraction produced inaccurate mappings or interpretations.

4 Dataset Statistics

Dataset composition is shown in Table 4. The corpus comprises 194 scientifically relevant verses, all of which are broadly topic-categorized and linked to 260 Tafsir records from two authoritative books (5 volumes total). The Scientific Ontology decomposes 8 major topics into 73 nodes. Of the 194 verses, 36 are associated with topics in scientific ontology, with 24 of them having granular alignments to specific ontology nodes. Future work will expand node-level coverage across additional scientific domains.

Table 4: Statistics of the QurSci-Onto Dataset

Feature	Count
Total Annotated Ayahs	194
Total Tafsir Indices	260
Tafsir Books (Volumes)	2 (5)
Unique Scientific Ontology Topics	8
Scientific Ontology Nodes	73
Ayah with ScientificConceptID Mapped	36
Ayah with ScientificNodes Mapped	24

5 Semantic Retrieval Evaluation

We evaluated the information retrieval performance using a test set of 24 domain-specific queries targeting the scientific concepts and process nodes defined in the Scientific Ontology

Table 5: Selected evaluation queries representing different scientific domains with corresponding ground-truth verses.

Query	Ground Truth
Cosmology: big bang universe creation heavens earth joined separated ratq fatq	(21:30)
Embryology: embryo development stages nutfah alaqah mudghah sequential	(23:14)
Oceanography: two seas barrier barzakh salt fresh water mixing estuarine	(25:53)
Biology: honey bees instinct foraging nectar collection	(16:68)
Hydrology: wind driving clouds advection convergence	(30:48)
General: atmospheric processes weather clouds precipitation	(24:43)

layer. Both systems utilized OpenAI's text-embedding-3-small model and a FAISS-backed vector store, with the search space restricted to the 194 annotated verses to maintain a closed-set evaluation environment. The Enhanced Implementation utilizes Ontological Grounding to enrich verse embeddings and consistently outperformed the Baseline across all metrics as shown in Table 6. These results show that the structured ontology captures domain-specific relationships between modern scientific concepts and classical Quranic text. Detailed category-wise performance and statistical significance are provided below.

5.1 Ontological Grounding & Semantic Enrichment

The performance gains in the Enhanced system are attributed to the Ontological Grounding of each verse. While the Baseline system embeds only the raw English translation, the Enhanced system performs Semantic Enrichment by concatenating the translation with two specialized layers extracted from the Knowledge Graph:

- **Scientific Description Nodes** We inject modern scientific keywords (`hasScientificKeywords`) to align the verse with contemporary terminology. For example, this maps the Quranic phrase "joined together" to the specific scientific concept "Singularity/Fused Mass."

- **Tafsir Summaries** We append the exegetical summary (`hasTafsirSummary`). This provides the necessary interpretive context that bridges the gap between the literal text and its intended scientific meaning.

These two enrichments provide complementary semantic layers. Scientific keywords align the text with modern terminology, while Tafsir summaries preserve exegetical interpretation.

5.2 Statistical Significance

Statistical significance was determined using a paired t-test with a sample size of N=24 queries. As shown in Table 6, the Enhanced Implementation achieved significant improvements across the majority of key ranking and retrieval metrics. Notably, the results for P@5, P@10, and R@5 all achieved $p < 0.05$, while NDCG@10 [3] reached a high level of significance ($p < 0.01$).

The strong improvement in R@5 (+23.71%) shows that the Ontology-Guided system doesn't just rank known verses better, it actually finds relevant verses that the plain-text baseline misses. The marginal significance in MRR (p=0.0719) indicates that while relevant verses usually appear in the top 5, securing the Rank-1 position remains sensitive to the lexical gap between unconstrained query phrasing and the standardized terminology used in our semantic enrichment.

Table 6: Performance comparison of the Baseline Keyword vs. Ontology-Guided retrieval.

Metric	Baseline	Enhanced	Delta Abs	Imp. (%)	p-value
P@1$\diamond$	0.4583	0.5833	0.0125	+27.27%	0.083
P@3	0.3611	0.4167	0.0556	+15.38%	0.2127
P@5*	0.2417	0.2833	0.0417	+17.24%	**0.0218**
P@10*	0.1417	0.1625	0.0208	+14.71%	**0.0218**
R@5*	0.6736	0.8333	0.1597	+23.71%	**0.0334**
MRR$\diamond$	0.6286	0.7326	0.1040	+16.54%	0.0719
NDCG@10**	0.6281	0.7570	0.1289	+20.53%	**0.0105**

* $p < 0.05$ (Significant), $\diamond$ $0.05 \leq p < 0.1$ (Marginally Significant)

5.3 Category-Wise Analysis

Performance varied based on the density of the Scientific Ontology layer.

High Impact: Oceanography (+43.1%), Embryology (+31.9%) and Hydrology (+27.7%) saw the largest gains, where Ontological Grounding successfully bridged the gap between modern technical terms and the original text.

[3] P@k (precision at rank k), R@k (recall at rank k), MRR (mean reciprocal rank), and NDCG@k (normalized discounted cumulative gain at rank k) are standard information retrieval metrics.

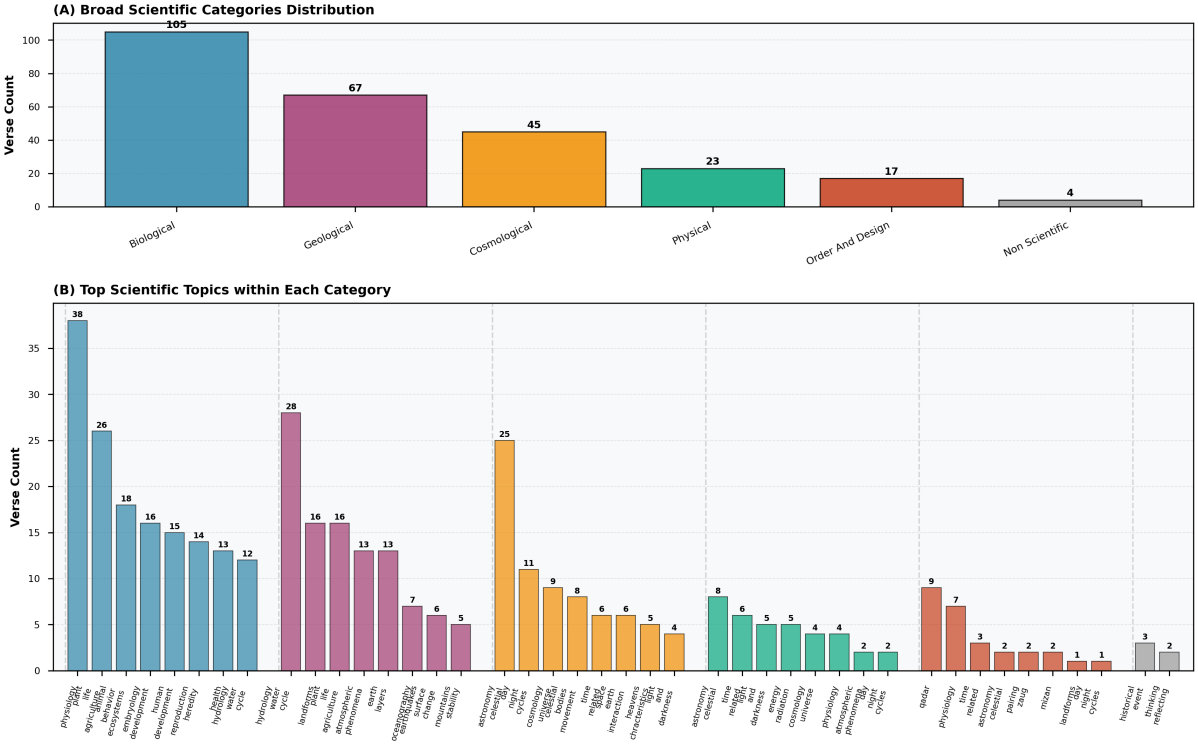

Figure 5: The distribution of verses across (a) Broad Scientific Categories and (b) Specific Scientific Topics in the QurSci-Onto dataset, illustrating the coverage and focus areas of our annotated corpus.

Low Impact: In Bee Biology (+0.0%), the Baseline already achieved a perfect NDCG (1.000) due to the highly specific vocabulary (e.g., "bees," "honey") unique to Surah An-Nahl.

General: General queries saw a modest +7.2% gain, as the baseline already retrieves these high-frequency concepts effectively due to the explicit lexical overlap between the query terms and the verse text. Table 7 presents the category-wise NDCG comparison.

Table 7: Category-wise NDCG comparison between Baseline and Ontology-Guided (Enhanced) retrieval.

Category	Queries	Baseline NDCG	Enhanced NDCG	Imp. (%)
Bee Biology	3	1.000	1.000	+0.0%
Cosmology	4	0.733	0.908	+23.8%
Embryology	5	0.501	0.660	+31.9%
General	5	0.478	0.512	+7.2%
Hydrology	4	0.656	0.838	+27.7%
Oceanography	3	0.541	0.775	+43.1%

5.4 Query Examples

Example queries targeting specific scientific concepts are listed in Table 5, with corresponding ground-truth verses for reference. For instance, the cosmology query “big bang universe creation heavens earth joined separated ratq fatq” targets verses describing cosmic origins (21:30), where the classical Arabic terms *ratq* and *fatq* (“joined” and “separated”) are mapped to modern concepts like “Singularity” and “Big Bang” through ontological grounding. This demonstrates the system’s ability to bridge modern scientific terminology with classical Quranic vocabulary.

6 Conclusion

We have developed a comprehensive resource for the computational analysis of scientific narratives in the Quran. QurSci-Onto systematically aligns authoritative exegesis with distinct scientific concepts to capture the logical progression of natural phenomena. Our expert-validated ontology provide a foundation for reliable, hallucination-resistant RAG systems in this sensitive domain.

This research contributes to the field of AI-driven Quranic NLP by introducing a framework for extracting causal relations and process-aware knowledge from exegesis texts. While situated in the Quranic domain, the methodology aligning scientific exegesis with structured modern ontologies has broader relevance for analyzing historical and classical literature. Future work will expand the ontology to cover additional Quranic verses and scientific domains.

7 Ethics Statement

Throughout this project, the Quranic text has been approached with the highest level of ethical care, reverence, and sensitivity. Ethical considerations were integrated into every stage. We explicitly state that the "Scientific Concepts" mapped in this ontology are extracted from and represent interpretive possibilities (Tafsir) and do not claim to be the final or exclusive meaning. The resources are released for academic research to facilitate the study of linguistic and semantic patterns.

8 Practical Implications

This work provides structured grounding resources for developing AI systems in sensitive theological domains. We identify the following primary areas where QurSci-Onto enables novel capabilities:

Reliable RAG with Provenance: Standard RAG systems often retrieve fragmented text chunks that lose context. By grounding generation in the Scientific Tafsir Index (Layer 2), developers can build Question-Answering systems that enforce citation-backed generation, restricting answers to authoritative page-level sources rather than generating plausible but unverified text.

Process-Aware Retrieval: Unlike standard vector retrieval based on semantic similarity alone, the ontology's explicit causal-temporal relations enable structured lookups that preserve sequential dependencies. This supports multi-hop queries tracing logical steps in scientific processes like stages of embryonic development or the water cycle, allowing systems to retrieve contextually connected stages rather than isolated facts.

Structured Educational Tools: The Scientific Ontology (Layer 3) facilitates the development of visual pedagogical tools. Unlike static keyword searches, these tools can render the logical progression of scientific concepts, like displaying the embryology sequence as a directed graph, helping students visualize the distinction between static entities and dynamic processes in the Quran.

Cross-Lingual Semantic Interoperability: The ontology provides a structured mapping between Classical Arabic morphology and modern scientific English concepts, addressing the semantic evolution from the 7th-century Quranic vocabulary to contemporary scientific discourse. This offers a standardized schema for future translation projects and digital heritage initiatives.

9 Limitations and Future Work

Our current ontology focuses on frequently discussed topics in classical Tafsir literature, establishing a replicable methodology. While all verses are linked to authoritative Tafsir sources with page-level citations, granular node-level annotations represent an initial subset for demonstration.

Additionally, the Scientific Ontology was constructed using LLM-assisted extraction. Although validated by experts, the mapping between Classical Arabic and modern scientific English remains interpretive. From a computational perspective, the retrieval evaluation relies on a limited set of scientific queries and may not generalize to all Quranic scientific Question-Answering (QA) scenarios. Currently, our retrieval evaluation demonstrates the utility of keyword and summary enrichment but does not yet leverage the full relational structure of the ontology for multi-hop process-aware or causal reasoning tasks.

Future work will expand the dataset's domain coverage and incorporate multiple Tafsir schools to broaden interpretative scope. We also plan to systematically evaluate the relational structure of the ontology for multi-hop reasoning and complex logical inferences. Specifically, we aim to develop reasoning-enhanced RAG models capable of utilizing the ontology's causal links to reconstruct scientific narratives and answer complex, process-driven queries.

Acknowledgments

The authors express their sincere gratitude to the co-authors for their invaluable contributions, guidance, and supervision throughout this research. We are particularly grateful to Professor Muhammad Amjad Khan and Dr. Abdul Majid for their rigorous review and expert verification of the dataset, ensuring its theological and scientific accuracy. Special thanks are due to Anwaar Shah for technical assistance in developing the annotation platform and to Warda Niaz for reviewing and validating the retrieval evaluation experiments. Finally, we thank the anonymous reviewers for their insightful feedback and constructive suggestions, which significantly improved the quality and clarity of this work.

References

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 15551--15569, 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.

Maha Al-yahya, Hend Al-Khalifa, Alia Bahanshal, Iman Al-Odah, and Nawal Al-Helwah. 2010. An ontological model for representing semantic lexicons: An application on time nouns in the holy quran. *The Arabian Journal for Science and Engineering*, 35.

Amna Basharat, Amna Binte Kamran, and Misbahur Rehman. 2025. Semantictafsir: Building a cultural heritage ontology and knowledge graph from the quranic exegesis of al-tabari. Manuscript submitted to the Semantic Web Journal; under review. Preprint (SWJ Tracking #: 3884-5098). Accessed: 2026-01-04.

Kais Dukes and Tim Buckwalter. 2010. A dependency treebank of the Quran using traditional Arabic grammar. In *2010 the 7th International Conference on Informatics and Systems (INFOS)*, pages 1--7. IEEE.

Kais Dukes and Nizar Habash. 2010. Morphological annotation of Quranic Arabic. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

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.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Hikmat Khan, Syed Saqlain, Muhammad Shoaib, and Muhammad Sher Ramzan. 2013. Ontology based semantic search in holy quran. *International Journal of Future Computer and Communication*, 2:570--575.

Imran Khan. 2019. The quran dataset. Kaggle. Contains Quranic text and translation by Dr. Mustafa Khattab.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wentau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401.

Rana Malhas and Tamer Elsayed. 2020. Ayatec: Building a reusable verse-based test collection for arabic question answering on the holy quran. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(6).

Sameha Abdullah Moogab, Ahmed Abdullah Al-Shalabi, and Ibrahim Ahmed Al-baltah. 2021. An ontological model for scientific miracle in the holy quran. In *2021 International Conference of Technology, Science and Administration (ICTSA)*, pages 1--6.

Abdul-Baquee M Sharaf and Eric Atwell. 2012. Qurana: Corpus of the Quran annotated with Pronominal Anaphora. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 130--137, Istanbul, Turkey. European Language Resources Association (ELRA).

Mohamed Sherif and Axel-Cyrille Ngonga Ngomo. 2015. Semantic quran: A multilingual resource for natural-language processing. *Semantic Web*, 6:339--345.

A Ta'a, Q.A. Abed, and M Ahmad. 2018. Al-quran ontology based on knowledge themes. *Journal of Fundamental and Applied Sciences*, 9(5S):800817.

Yahya M. Tashtoush, Majd R. Al-Soud, Reema M. AbuJazoh, and Manar Al-Frehat. 2017. The noble quran arabic ontology: Domain ontological model and evaluation of human and social relations. In *2017 8th International Conference on Information and Communication Systems (ICICS)*, pages 40--45.

Wajdi Zaghouani, Abdelati Hawwari, and Mona Diab. 2012. A pilot propbank annotation for quranic Arabic. In *Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature*, pages 78--83.

A General and Technical Category Definitions

To analyze the system's performance across different levels of terminological specificity, we classified evaluation queries into two distinct categories:

Note : Technical Categories involve Cosmology, Embryology, Oceanography, Biology, Hydrology.

B Dataset Details

Dataset details are shown in Figures below.

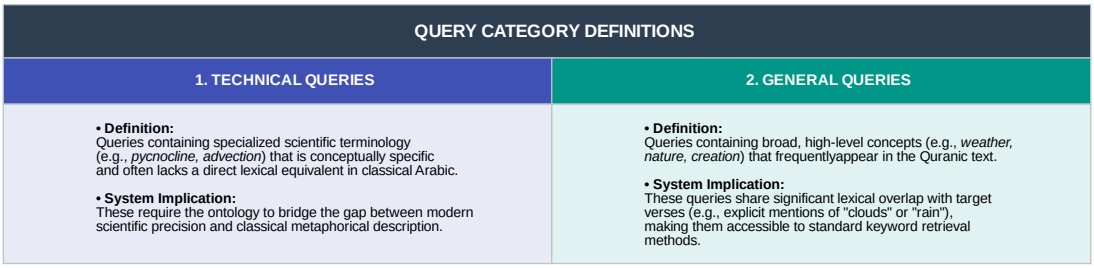

Figure 6: Explanation of General versus Technical Queries

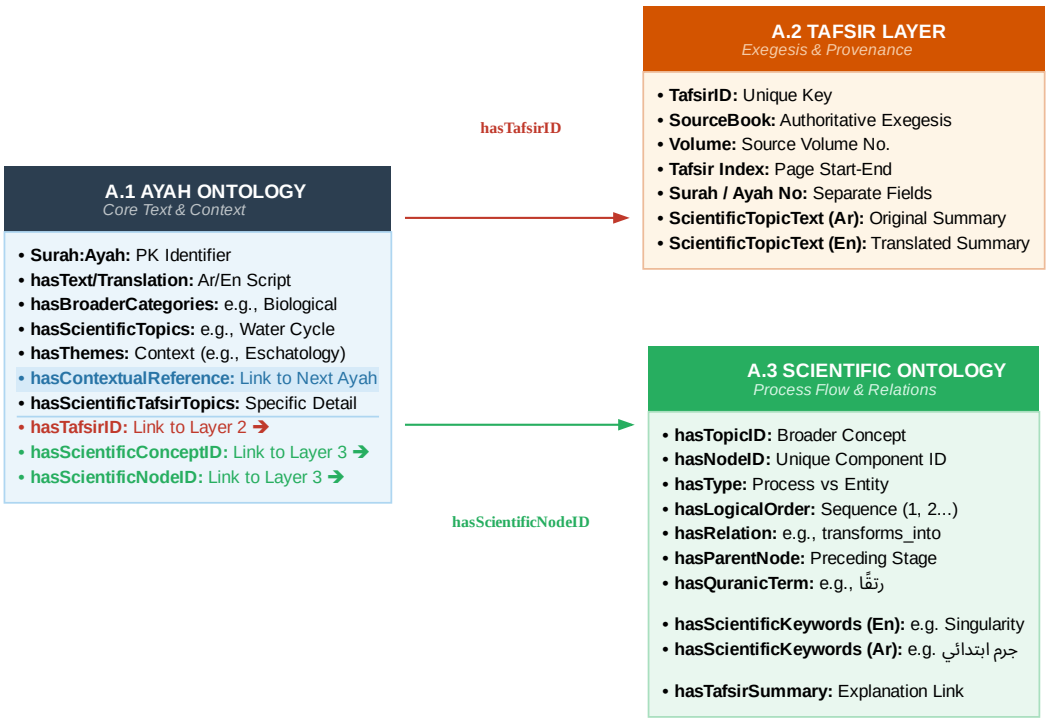

Figure 7: Highlevel overview of hierarchical ontology structure linking verses, exegesis, and scientific concepts.

A.1 AYAH ONTOLOGY DETAILS

• Surah and Ayah (Surah:Ayah)
The Ayah and Surah number of the record.

• Arabic Text (hasText)
The original Arabic script of the Ayah inferred from The Quran Dataset.

• English Translation (hasTranslation)
The English translation (Dr. Mustafa Khattab) inferred from The Quran Dataset.

• Broad Category (hasBroadCategories)
The broad scientific field like Biological, Cosmological, Geological.

• Scientific Topic (hasScientificTopics)
The specific phenomenon being described such as Human Embryonic Development, the water cycle, cosmic expansion.

• Context (hasThemes)
The context of the specific ayah or in which the Tafsir is related like Historical Event, Eschatology, Moral or Ethical Reflection.

• Related Ayah (hasContextualReference)
Used when a single scientific exegesis spans multiple continuous verses.
It links the current record to the subsequent Ayah ID (e.g., linking 25:45 to 46) to ensure the RAG system captures the full narrative flow.

• Specific Tafsir Topic (hasScientificTafsirTopics)
Captures the precise scientific phenomena discussed in the exegesis, offering more granularity than the broader *hasScientificTopics*.
Examples include *"Stages of cloud formation"*, *"Physics of storm clouds"*, or *"Formation of Hail"*.

• Exegesis Link (hasTafsirID)
A pointer to the Layer 2 scientific commentary (Tafsir Ilmi), enabling the retrieval of rich-text explanations that interpret the Ayah in scientific terms.

• Scientific Concept ID (hasScientificConceptID)
Pointer to the broader scientific topic in Layer 3 (e.g., "HYDRO_METEO", "EMBRYO"), categorizing the verse under a general phenomenon.

• Scientific Node ID (hasScientificNodes)
Pointer to the specific node (stage or entity) within the scientific graph (Layer 3). This enables RAG systems to retrieve verses based on sequential process flow and causal relations (e.g., ["HYDRO_METEO_04", "HYDRO_METEO_10"]).

Figure 8: Ayah Ontology schema attributes and their descriptions.

TABLE A.1: AYAH ONTOLOGY DATA SNAPSHOT
Selected samples (2:22, 2:26, 11:44, 10:5) illustrating domain diversity and metadata structure.

ID	Text / Translation (Truncated)	Broad Cat.	Sci. Topics	Tafsir Topics	Themes	Tafsir ID
2:22	ٱلَّذِى جَعَلَ لَكُمُ ٱلْأَرْضَ... Who has made the earth a place of settlement for you...	geological	hydrology_water_cycle	Terrestrial & Hydrological Cycle	process	TAAK_1_2_22
2:26	إِنَّ ٱللَّهَ لَا يَسْتَحْيِۦٓ... Surely Allah does not shy away from using the parable...	biological	animal_behavior ecosystems	Micro-Complexity Mosquito Structure	analogical ethical_theological	TAAK_1_2_26
11:44	وَقِيلَ يَٰٓأَرْضُ ٱبْلَعِى... And it was said, "O earth! Swallow up your water...	geological	hydrology_water_cycle earthquakes_surface	Volcanic Origin of Water & The Ark	process historical_event	TAAK_1_11_44
10:5	هُوَ ٱلَّذِى جَعَلَ ٱلشَّمْسَ... He made the sun a radiant source and the moon light...	physical cosmological	time_related light_and_darkness astronomy_celestial	Reflection from moon Real light from sun	signs_to_reflect knowledge	TAAK_1_10_5

Figure 9: Ayah Ontology Dataset Sample.

A.2 TAFSIR LAYER DETAILS

- **Tafsir Book (SourceBook)**
  The name of the authoritative exegesis book used (e.g., *Tafsir Ayat Al Konia*).
- **Volume (Volume)**
  The specific volume number of the source book where the explanation is found.
- **Tafsir Index (PageNoStart - PageNoEnd)**
  Page-level citations within the source volume to ensure auditability and provide grounding for RAG systems.
- **Surah Number (SurahNo)**
  The specific chapter number of the Quran being interpreted (e.g., Surah 2).
- **Ayah Number (Ayah)**
  The specific verse number within the Surah (e.g., Ayah 19).
- **Scientific Summary - Arabic (ScientificTopicTextArabic)**
  The original Arabic summary of the scientific topic extracted from the index (e.g., *"...تفصيل لأنواع الرياح المعروفة"*).
- **Scientific Summary - English (ScientificTopicTextEnglish)**
  An English Translation of ScientificTopicTextArabic (e.g., *"A detailed exposition on the known types of winds..."*).
- **Tafsir ID (TafsirID)**
  A unique identifier (e.g., *TAAK_1_2_19*) used as a pointer to link the Ayah ontology to this Tafsir record.

Figure 10: Scientific Tafsir Index schema attributes and their descriptions.

TABLE A.2: TAFSIR LAYER DATA SNAPSHOT
Selected samples (2:19, 2:26, 10:5, 11:44) showing provenance, indexing, and scientific summaries.

Tafsir ID (PK)	Source Book	Citation	S:A	Scientific Topic (English)	Scientific Topic (Arabic)	Vol
TAAK_1_2_19	Tafsir Ayat Al Konia	pp. 63-70	**2:19**	A detailed exposition on the known types of winds, rain-bearing clouds, and thunder...	تفصيل لأنواع الرياح المعروفة، وشرح تكوّن السحب...	1
TAAK_1_2_26	Tafsir Ayat Al Konia	pp. 79-86	**2:26**	Reference to the mosquito, highlighting its remarkable construction and complexity...	الإشارة إلى البعوضة، وهى من أبسط الحشرات...	1
TAAK_1_10_5	Tafsir Ayat Al Konia	pp. 333-340	**10:5**	Differentiation between light and illumination; sources of light from outer space...	التفريق الواضح بين الضياء والنور، وتحديد مصادر الضوء...	1
TAAK_1_11_44	Tafsir Ayat Al Konia	pp. 347-354	**11:44**	Conclusion that Earth's water came from volcanic vents; location of Noah's Ark...	استنتاج أن كل ماء الأرض أخرجه الله من باطنها...	1

Figure 11: Tafsir Index Dataset Sample.

A.3 SCIENTIFIC ONTOLOGY DETAILS

• Topic ID (hasTopicID)
Unique identifier for each scientific topic such as *BigBang_CosmosExpansion*.

• Node ID (hasNodeID)
Unique identifier for each component of the scientific topic such as *BigBang_CosmosExpansion_01*.

• Type (hasType)
Classification as either **Process** (sequential stages) or **Entity** (static phenomena).

• Quranic Term (hasQuranicTermArabic)
The original Arabic terminology used in the Quran like رتقًا for "Singularity/Fused Mass".

• Scientific Keywords English (hasScientificKeywords)
Modern scientific interpretation term in English (e.g., "Singularity/Fused Mass").

• Scientific Keywords Arabic (hasScientificKeywordsArabic)
Modern Arabic scientific terms used in the exegesis to explain the classical Quranic term such as جرم ابتدائي واحد.

• Logical Order (hasLogicalOrder)
Sequential position of the node within a process.

• Relation (hasRelation)
Relationship to parent node like *transforms_into, composed_of, characterizes, is_part_of*.

• Parent Node (hasParentNode)
Reference to the preceding stage or parent entity.

• Tafsir Summary (hasTafsirSummary)
Summary extracted from Tafsir explanation linking the Quranic term to scientific concept.

• Verse Reference (Surah:Ayah)
Linked Surah:Ayah citations.

• Exegesis Link (hasTafsirID)
A pointer to the Layer 2 scientific commentary (Tafsir Ilmi)

***Note** : All records are mined from Tafsir Ayat al-Kawniyah..*

Figure 12: Scientific Ontology schema attributes and their descriptions.

TABLE A.3: SCIENTIFIC ONTOLOGY DATA SNAPSHOT
Sample process chain (Big Bang & Expansion) illustrating the logical order, causal relations, and term grounding.

ID	Node ID	Relation	Parent Node	Ord.	Quranic Term	Scientific Keywords	Tafsir Summary (Truncated)
1	BigBang_Cosmos Expansion_01	*root_node*	-	1	رَتْقًا	Singularity / Fused Mass	Initial state of the universe as a single joined entity...
2	BigBang_Cosmos Expansion_02	*transforms_into*	...Expansion_01	2	فَفَتَقْنَاهُمَا	Big Bang / Cosmic Inflation	Violent separation or splitting of the initial fused mass.
3	BigBang_Cosmos Expansion_03	*originates_from*	...Expansion_02	3	دُخَانٌ	Cosmic Smoke / Nebula	Sky transformed into a gaseous state after separation.
4	BigBang_Cosmos Expansion_04	*composed_of*	...Expansion_03	0	*Tafsir-Derived*	Hydrogen & Helium Nuclei	Smoke contained protons/neutrons forming H and He nuclei.

Figure 13: Scientific Ontology Dataset Sample.

SYSTEM PROMPT FOR REPRODUCIBLE KNOWLEDGE EXTRACTION

SYSTEM ROLE:
You are an expert Scientific Knowledge Graph Engineer and Quranic Scholar. Your task is to extract structured scientific data from unstructured Arabic Tafsir text to build a verified Knowledge Graph (KG).

OBJECTIVE:
Analyze the provided Arabic text. Extract the Nodes (Concepts) and Edges (Relations) for ONE Target Topic only. Structure the data to capture sequences, composition, and crucially source provenance.

INPUT VARIABLES:

- **[TARGET TOPIC]:** The specific concept (e.g., "Embryology").
- **[SOURCE TEXT]:** The Arabic content (including Page/Source tags).

1. DETERMINE TOPIC CATEGORY (hasType)

- **Process:** Events happening over time (e.g., Embryology, Water Cycle). Key Feature: *Sequence*.
- **Entity:** Physical objects or structures (e.g., Cell, Cloud). Key Feature: *Composition*.

2. SELECT THE CORRECT RELATION (Edge Logic)

- **Time/Sequence:** `precedes, causes, transforms_into, originates_from`.
- **Structure:** `composed_of, is_part_of, contains`.
- **Function/Prop:** `has_property, performs_function, has_shape`.
- **Context:** `separates, interacts_with, analogy_to`.

3. STRICT PROVENANCE & LINGUISTIC VERIFICATION

- **hasQuranicTermArabic:** Must be the exact word from the Ayah (Scripture).
- **hasScientificKeywordsArabic:** You **MUST** extract the specific technical Arabic phrase. Do not translate; copy from text.

4. SCOPE OF INTERACTION

- Include external entities only if the target topic directly acts upon them. (e.g., if Target="Estuary", include "Fresh Water" and "Salt Water").

OUTPUT SCHEMA (CSV FORMAT ONLY)

```
"id, hasTopicID, hasNodeID, hasRelation, hasParentNode, hasLogicalOrder,
 hasQuranicTermArabic, hasScientificKeywords, hasTafsirSummary,
                Surah:Ayah, hasType, hasSourceBook
      hasScientificKeywordsArabic"
```

DEFINITIONS OF COLUMNS:

- **id:** Sequential Integer.
- **hasTopicID:** Unique Topic String (e.g., BEE_BIO).
- **hasNodeID:** Unique Node String (e.g., BEE_BIO_01).
- **hasRelation:** The edge type connecting this to the Parent.
- **hasParentNode:** The NodeID this connects to (NULL for Root).
- **hasLogicalOrder:** Integer (1, 2, 3...) for processes; 0 for entities.
- **hasQuranicTermArabic:** The specific word from the Verse.
- **hasScientificKeywords:** Modern English Scientific Concept.
- **hasTafsirSummary:** 1-sentence summary of the interpretation.
- **Surah:Ayah:** The verse citation (e.g., 16:69).
- **hasType:** Process, or Entity
- **hasSourceBook:** "Tafsir Ayat al-Kawniyah".
- **hasScientificKeywordsArabic:** The exact Arabic phrase found in the Tafsir text.

SAMPLE ANNOTATION (FEW-SHOT)

```
id, hasTopicID, hasNodeID, hasRelation, hasParentNode, hasLogicalOrder,
hasQuranicTermArabic, hasScientificKeywords, hasTafsirSummary, ...

1, BigBang, BigBang_01, root_node, NULL, 1, , Singularity/Fused Mass,
Text describes initial state as a joined entity of infinite density., 21:30, Process,

2, BigBang, BigBang_02, transforms_into, BigBang_01, 2, , Cosmic Inflation,
Text describes violent separation of the fused mass., 21:30, Process,
```

Figure 14: Complete System Prompt for reproducibility

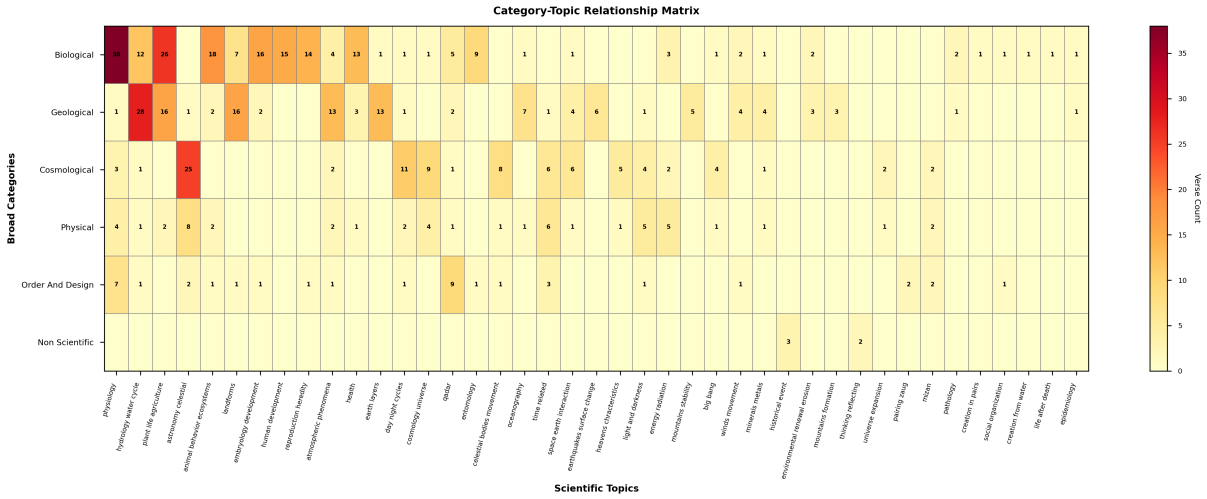

Figure 15: Cross-tabulation of topics and categories showing frequency distributions for each relation.