MeerKAT+RAG: Evidence-Grounded Answers from Telescope Documentation

A document-aware RAG framework combining

hybrid retrieval, graph-assisted search,

and answer-support checks

F. AIT BENKHALI, F. JANKOWSKY, S. WAGNER

Abstract

We present **MeerKAT+RAG**, a domain-specific retrieval-augmented generation (RAG) framework for answering technical and operational questions about the MeerKAT-Extension radio telescopes. It makes answer verifiability a central architectural concern, transforming documentation into structured, provenance-aware chunks and combining dense semantic retrieval with lexical search via Reciprocal Rank Fusion (RRF), followed by cross-encoder reranking. A deterministic knowledge graph and bounded agentic retrieval support entity-centred and multi-step queries. Before delivery, answers undergo two-tier verification using deterministic claim and numeric checks, complemented by an optional language-model-based entailment assessment. The framework can withhold answers when evidence is insufficient. Evaluation on a 123-question benchmark informs the design and provides an initial assessment of its capabilities and limitations. MeerKAT+RAG prioritises traceable evidence and explicit verification over fluent but potentially unsupported responses.

Making Operational Knowledge Accessible

MeerKAT-Extension commissioning operations depend on precise technical specifications and procedures, including wind limits, antenna pointing ranges, receiver frequency bands, cooldown procedures, equipment identifiers, and safety interlocks etc. This knowledge is distributed across technical manuals & datasheets, embedded in prose, tables, lists, and figures. Three challenges limit reliable access to this information:

Lexical Search: can miss questions phrased differently from the source text

Semantic retrieval: can overlook exact identifiers or confuse similar specifications and numerical values.

Unconstrained LLMs: can generate plausible but unsupported answers, particularly when evidence is incomplete

In telescope operations, a plausible answer is not necessarily a reliable one. *Answers must be grounded in authoritative documentation, and insufficient evidence must remain a valid reason not to answer*

Objective: Develop an end-to-end RAG workflow that makes MeerKAT-Extension technical documentation queryable while preserving source provenance and providing explicit mechanisms to verify answer support.



System Architecture

MeerKAT+RAG separates **offline knowledge preparation** from **online question answering**. The core retrieval pipeline is extended by a knowledge graph, bounded agentic search, figure-captioning capabilities, and answer-support checks. These components are configuration-dependent rather than mandatory in every serving path. Current configuration: **Qwen3-Embedding-4B** for embeddings, **Qwen3-Reranker-4B** for reranking, and **MiniMax-M3** for generation, accessed through an OpenAI-compatible endpoint. These are implementation choices, not architectural requirements.

1. Ingestion & provenance: PDF/HTML sources are converted into Markdown with extracted tables, figures,and metadata. Document identity, revision, page context, and source-file hashes preserve links to the original material.

2. Structure-aware chunking: A section-aware chunker creates passages from prose, tables, lists, procedures, and code or configuration blocks. Stable chunk identifiers and document metadata support traceability across retrieval, agent logs, and evaluation.

3. Hybrid Indexing: Chunks are embedded and indexed in **Qdrant**, which also provides the full-text search used by the active lexical retrieval path. Dense semantic search captures paraphrases, while lexical matching helps retain exact identifiers and specifications. BM25 is used as an evaluation baseline, not as the production lexical implementatio.n

4. Evidence fusion and reranking: Reciprocal Rank Fusion (RRF) combines dense and full-text rankings without requiring their scores to be calibrated to a common scale. A cross-encoder reranks a bounded set of leading candidates, while the remaining candidates retain their RRF order.

5. Context-grounded generation: Selected passages and their source references are supplied to an OpenAI-compatible language model. Document, section, revision, content type, and page metadata support source tracing. Retrieval provides evidence for generation but does not guarantee correct interpretation.

6. Answer-support assessment: Two complementary checks are available

– **Deterministic checks:** examine textual overlap, numeric and unit-aware grounding, and reference validity. These checks are repeatable and inexpensive but cannot detect every semantic or logical error

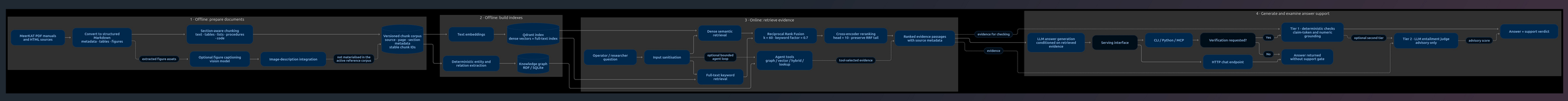
– **LLM entailment assessment:** an optional language-model judge provides an additional, advisory assessment of whether the evidence supports the answer

Knowledge Graph & Agentic Retrieval

KG and Agent Loop: A lightweight RDF knowledge graph, stored in SQLite, complements document retrieval through deterministic entity extraction and cue-anchored relations (seven node types and eight relation types). An optional agent combines five validated tools: graph-search, vector-search, hybrid-search, entity-extract, and lookup-chunk for multi-step retrieval, bounded to two decision iterations. Graph results use the same evidence representation as retrieved document chunks.

Vision and Figure Understanding: Integrated vision-model capabilities extend the framework beyond text retrieval by extracting information from figures and generating descriptions for downstream use.

Interfaces and Operational Design: MeerKAT+RAG supports CLI/Python, an OpenAI-compatible HTTP API, MCP, and Open WebUI. Stable chunk identifiers, recorded benchmark artefacts, validated tool inputs, bounded execution, and typed failures support traceability and reproducibility. These controls improve operational robustness but do not guarantee answer correctness.



Gefördert durch:

Bundesministerium
für Forschung, Technologie
und Raumfahrt

www.deutscheszentrumastrophysik.de

DEUTSCHES ZENTRUM ASTROPHYSIK