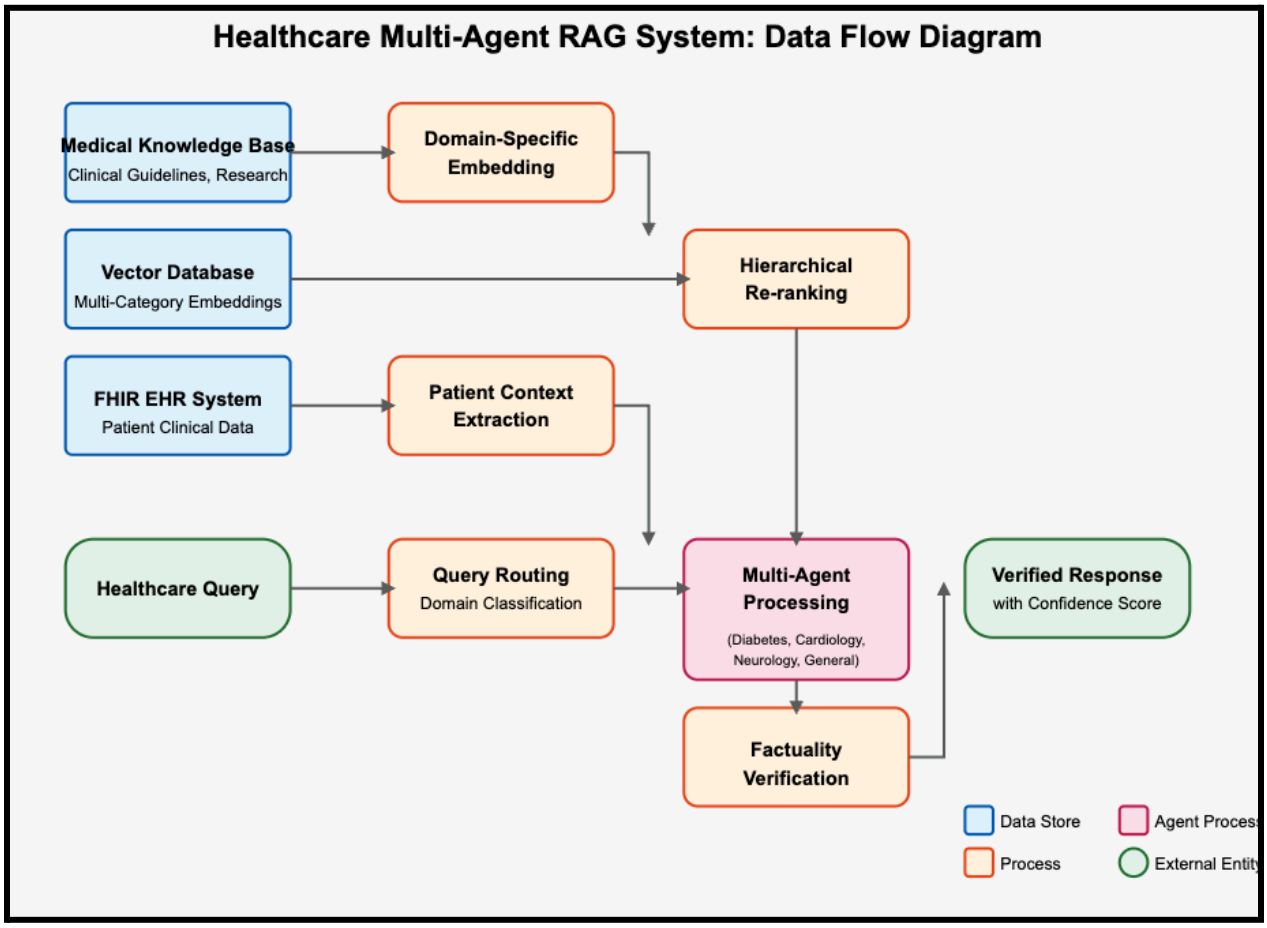
PWC Healthcare Multi-Agent RAG System:



1. **Data Sources** (left side):
    - Medical Knowledge Base containing clinical guidelines and research
    - Vector Database with multi-category medical embeddings
    - FHIR EHR System storing patient clinical data
2. **Core Processes** (center):
    - Domain-Specific Embedding: Converts medical text into specialized vector representations
    - Patient Context Extraction: Pulls relevant patient information from EHR
    - Query Routing: Directs queries to appropriate specialized agents
    - Hierarchical Re-ranking: Multi-stage retrieval and ranking process
3. **Agent Processing** (right center):
    - Multi-Agent Processing: Specialized agents for different medical domains (Diabetes, Cardiology, Neurology, General Practice)
    - Factuality Verification: Validates information against trusted medical sources
4. **Input/Output** (bottom):

- Healthcare Query: Entry point for user questions
- Verified Response: Final output with confidence scoring

The flow shows how a healthcare query moves through domain classification, is processed by specialized agents with support from retrieval systems, undergoes factuality verification, and ultimately produces a verified response. The parallel processes enable specialized knowledge integration while maintaining clinical accuracy.

KEY FEATURES PWC Healthcare-focused RAG system with multi-agent architecture.

1. **Domain-Specific Embeddings**:
   - Uses BioBERT/ClinicalBERT for healthcare-specific embeddings
   - Includes medical entity recognition and relationship enrichment
2. **Hierarchical Re-Ranking**:
   - Multi-stage pipeline: BM25 → lightweight embeddings → deep cross-attention → domain rules
   - Specialized medical knowledge rules for different domains
3. **Hallucination Reduction**:
   - Factuality verification against trusted medical sources
   - Confidence scoring and source attribution
   - Annotation of generated content with confidence levels
4. **Multi-Agent Architecture**:
   - Specialized agents for different medical domains (diabetes, cardiology, etc.)
   - Query routing based on medical specialization
   - Response combination with expert weighting
5. **Clinical Workflow Integration**:
   - FHIR-compliant data exchange
   - Patient context extraction
   - Clinician notification system
6. **Vector Database Optimization**:
   - Multi-vector representation (symptoms, treatments, outcomes)
   - Context-aware routing to specialized indexes

The system is designed to be extensible and modular, allowing for easy integration of new specialized agents or medical domains. You can use the example_usage() function to see how the system works with a sample query about diabetes management.