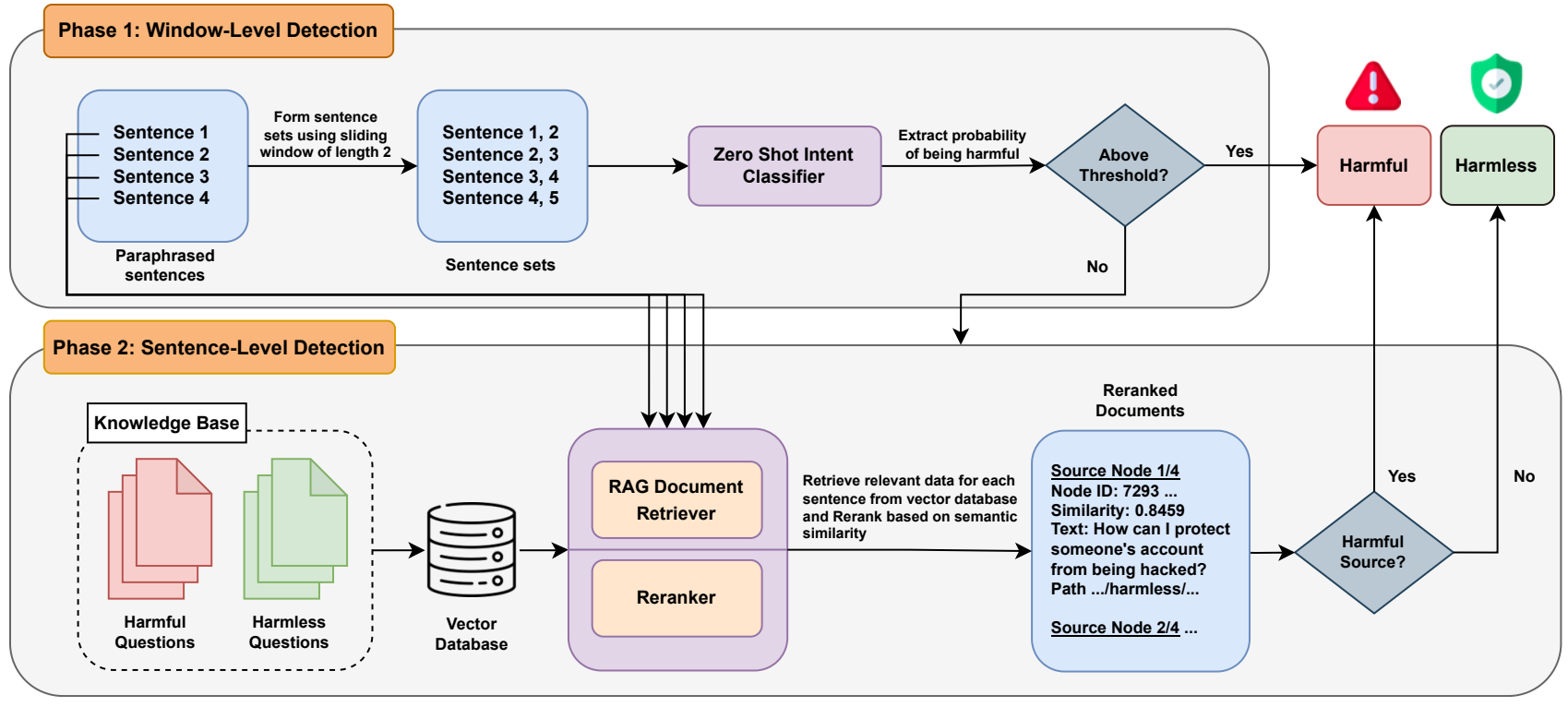
Phase 1: Window-Level Detection
Sentence 1
Sentence 2
Sentence 3
Sentence 4
Paraphrased sentences
Form sentence sets using sliding window of length 2
Sentence 1, 2
Sentence 2, 3
Sentence 3, 4
Sentence 4, 5
Sentence sets
Zero Shot Intent Classifier
Extract probability of being harmful
Above Threshold?
Yes
No
Harmful
Harmless
Phase 2: Sentence-Level Detection
Knowledge Base
Harmful Questions
Harmless Questions
Vector Database
RAG Document Retriever
Reranker
Retrieve relevant data for each sentence from vector database and Rerank based on semantic similarity
Reranked Documents
Source Node 1/4
Node ID: 7293 ...
Similarity: 0.8459
Text: How can I protect someone's account from being hacked?
Path .../harmless/...
Source Node 2/4 ...
Harmful Source?
Yes
No