RAG System Tuning Levers

> **Priority-sorted environment variables for optimizing Onyx RAG system performance and behavior**

This document focuses on the most impactful environment variables that directly affect RAG (Retrieval-Augmented Generation) system behavior, quality, and performance. Variables are organized by priority and impact level.

Table of Contents

Critical RAG Impact Variables

> **Variables with fundamental impact on RAG system behavior**

CHUNK_SIZE

Property	Value
Default	`512`
Type	Integer (tokens)
Priority	CRITICAL
RAG Impact	Fundamental control over retrieval granularity and context balance
Tuning Range	`256-1024`
Considerations	Smaller chunks = more precise retrieval but less context. Larger chunks = more context but may dilute relevance
Recommended Values	Precise: `256-384`, Balanced: `512`, Contextual: `768-1024`

HYBRID_SEARCH_ALPHA

Property	Value
Default	`0.6`
Type	Float (0.0-1.0)
Priority	CRITICAL

Property	Value
RAG Impact	Controls semantic vs keyword search balance - core retrieval strategy
Tuning Range	`0.0-1.0`
Considerations	0.0 = pure keyword, 1.0 = pure semantic. Most content benefits from hybrid approach
Recommended Values	Keyword-focused: `0.2-0.4`, Balanced: `0.5-0.7`, Semantic-focused: `0.8-1.0`

DOCUMENT_ENCODER_MODEL

Property	Value
Default	`sentence-transformers/all-MiniLM-L6-v2`
Type	String
Priority	CRITICAL
RAG Impact	Determines semantic understanding quality and computational requirements
Tuning Range	Model-dependent
Considerations	Larger models = better understanding but higher resource usage
Common Alternatives	Fast: `all-MiniLM-L6-v2`, Balanced: `all-mpnet-base-v2`, High-quality: `all-MiniLM-L12-v2`

ENABLE_CONTEXTUAL_RAG

Property	Value
Default	`false`
Type	Boolean
Priority	CRITICAL
RAG Impact	Dramatically improves document understanding and context preservation
Resource Cost	Very High
Considerations	Significant performance impact but major quality improvement
Recommended Use	Enable for high-quality requirements, disable for speed

NUM_RERANK_CHUNKS

Property	Value
Default	`15`
Type	Integer

Property	Value
Priority	CRITICAL
RAG Impact	Directly controls recall vs precision balance in results
Tuning Range	`5-50`
Considerations	More chunks = better recall but higher computational cost
Recommended Values	Fast: `5-10`, Balanced: `15-25`, High-recall: `30-50`

High RAG Impact Variables

> **Variables with significant impact on RAG quality and performance**

NUM_RETURNED_HITS

Property	Value
Default	`50`
Type	Integer
Priority	HIGH
RAG Impact	Initial retrieval pool size affects downstream reranking quality
Tuning Range	`10-200`
Considerations	Larger pools improve reranking quality but increase processing time
Recommended Values	Fast: `20-30`, Balanced: `50-75`, Comprehensive: `100-150`

SEARCH_DISTANCE_CUTOFF

Property	Value
Default	`null` (no cutoff)
Type	Float (0.0-1.0)
Priority	HIGH
RAG Impact	Quality gate for retrieved results - filters low-relevance matches
Tuning Range	`0.3-0.9`
Considerations	Higher values = stricter filtering. May reduce recall if set too high
Recommended Values	Permissive: `0.3-0.5`, Balanced: `0.6-0.7`, Strict: `0.8-0.9`

CHUNK_OVERLAP

Property	Value

Property	Value
Default	`0`
Type	Integer (tokens)
Priority	HIGH
RAG Impact	Preserves context across chunk boundaries, reduces information loss
Tuning Range	`0-128`
Considerations	Overlap increases storage and processing but improves context continuity
Recommended Values	None: `0`, Light: `32-64`, Heavy: `96-128`

EMBEDDING_BATCH_SIZE

Property	Value
Default	`32`
Type	Integer
Priority	HIGH
RAG Impact	Affects embedding generation speed and memory usage
Tuning Range	`8-128`
Considerations	Larger batches = faster processing but more memory usage
Recommended Values	Resource-limited: `8-16`, Balanced: `32-64`, High-throughput: `64-128`

ENABLE_RERANKING_REAL_TIME_FLOW

Property	Value
Default	`false`
Type	Boolean
Priority	HIGH
RAG Impact	Improves result ranking quality but adds latency
Resource Cost	Medium-High
Considerations	Better result quality vs increased response time
Recommended Use	Enable for accuracy-critical applications

NUM_INDEXING_WORKERS

Property	Value
Default	`1`

Property	Value
Type	Integer
Priority	HIGH
RAG Impact	Affects document processing speed and system resource utilization
Tuning Range	`1-8`
Considerations	More workers = faster indexing but higher CPU/memory usage
Recommended Values	Single-core: `1`, Multi-core: `2-4`, High-performance: `4-8`

MINI_CHUNK_SIZE

Property	Value
Default	`150`
Type	Integer (tokens)
Priority	HIGH
RAG Impact	Enables fine-grained information extraction for precise retrieval
Tuning Range	`50-300`
Considerations	Smaller = more precise but higher overhead
Recommended Values	Precise: `50-100`, Balanced: `150-200`, Contextual: `250-300`

Medium RAG Impact Variables

> **Variables with moderate but meaningful impact on RAG behavior**

NORMALIZE_EMBEDDINGS

Property	Value
Default	`true`
Type	Boolean
Priority	MEDIUM
RAG Impact	Improves search consistency with cosine similarity
Considerations	Usually beneficial unless using specific embedding models that expect unnormalized vectors

DOC_EMBEDDING_CONTEXT_SIZE

Property	Value

Property	Value
Default	`512`
Type	Integer (tokens)
Priority	MEDIUM
RAG Impact	Controls context window for document embedding generation
Tuning Range	`256-1024`
Considerations	Must not exceed embedding model's maximum context length

ENABLE_INFORMATION_CONTENT_CLASSIFICATION

Property	Value
Default	`false`
Type	Boolean
Priority	MEDIUM
RAG Impact	Improves search precision through content type understanding
Resource Cost	Medium
Considerations	Adds processing overhead but can improve relevance

BATCH_SIZE_ENCODE_CHUNKS

Property	Value
Default	`8`
Type	Integer
Priority	MEDIUM
RAG Impact	Affects indexing speed and GPU utilization
Tuning Range	`4-32`
Considerations	Larger batches improve GPU efficiency but require more memory

VESPA_TIMEOUT

Property	Value
Default	`120`
Type	Integer (seconds)
Priority	MEDIUM
RAG Impact	Controls search operation timeout - affects user experience

Property	Value
Tuning Range	`30-300`
Considerations	Balance between accommodating complex queries and responsiveness

ENABLE_LARGE_CHUNK_SEARCH

Property	Value
Default	`false`
Type	Boolean
Priority	MEDIUM
RAG Impact	Enables search over larger context windows for comprehensive results
Considerations	Provides more context but may reduce precision

AGENTIC_ENABLED

Property	Value
Default	`false`
Type	Boolean
Priority	MEDIUM
RAG Impact	Enables advanced query understanding and multi-step reasoning
Resource Cost	High
Considerations	Sophisticated query processing but significant resource requirements

DISABLE_LLM_CHUNK_FILTER

Property	Value
Default	`false`
Type	Boolean
Priority	MEDIUM
RAG Impact	Speed vs quality tradeoff for chunk filtering
Considerations	Set to `true` for faster processing, `false` for better quality

---

Quick Tuning Reference

Performance Priority Tuning

Goal	Key Variables	Recommended Settings
Maximum Speed	`CHUNK_SIZE`, `ENABLE_CONTEXTUAL_RAG`, `DISABLE_LLM_CHUNK_FILTER`	`384`, `false`, `true`
Maximum Quality	`CHUNK_SIZE`, `ENABLE_CONTEXTUAL_RAG`, `NUM_RERANK_CHUNKS`	`512-768`, `true`, `25-50`
Balanced	`HYBRID_SEARCH_ALPHA`, `NUM_RETURNED_HITS`, `CHUNK_OVERLAP`	`0.6`, `50`, `64`

Content Type Optimization

Content Type	Key Adjustments
Technical Docs	`CHUNK_SIZE=768`, `HYBRID_SEARCH_ALPHA=0.4` (more keyword-focused)
General Knowledge	`CHUNK_SIZE=512`, `HYBRID_SEARCH_ALPHA=0.7` (more semantic)
Code/Structured	`CHUNK_SIZE=384`, `CHUNK_OVERLAP=64`, keyword-focused search

Use Case Configurations

Precision-Focused Setup

```
# For applications requiring highly relevant results
CHUNK_SIZE=384
CHUNK_OVERLAP=64
HYBRID_SEARCH_ALPHA=0.8
NUM_RERANK_CHUNKS=30
SEARCH_DISTANCE_CUTOFF=0.7
ENABLE_RERANKING_REAL_TIME_FLOW=true
ENABLE_INFORMATION_CONTENT_CLASSIFICATION=true
```

Speed-Optimized Setup

```
# For applications prioritizing fast response times
CHUNK_SIZE=384
CHUNK_OVERLAP=0
HYBRID_SEARCH_ALPHA=0.5
NUM_RERANK_CHUNKS=10
NUM_RETURNED_HITS=30
ENABLE_CONTEXTUAL_RAG=false
DISABLE_LLM_CHUNK_FILTER=true
```

High-Quality Setup

```
# For applications requiring comprehensive, high-quality results
CHUNK_SIZE=768
CHUNK_OVERLAP=96
HYBRID_SEARCH_ALPHA=0.7
NUM_RERANK_CHUNKS=40
NUM_RETURNED_HITS=100
ENABLE_CONTEXTUAL_RAG=true
ENABLE_RERANKING_REAL_TIME_FLOW=true
ENABLE_LARGE_CHUNK_SEARCH=true
```

Resource-Constrained Setup

```
# For deployments with limited computational resources
CHUNK_SIZE=384
NUM_INDEXING_WORKERS=1
EMBEDDING_BATCH_SIZE=16
NUM_RERANK_CHUNKS=10
NUM_RETURNED_HITS=30
ENABLE_CONTEXTUAL_RAG=false
BATCH_SIZE_ENCODE_CHUNKS=4
```

Tuning Guidelines

Step-by-Step Optimization Process

1. **Start with Core Settings**
   - Set `CHUNK_SIZE` based on your content type and precision needs
   - Configure `HYBRID_SEARCH_ALPHA` based on search strategy preference
   - Adjust `NUM_RERANK_CHUNKS` based on quality vs speed requirements
2. **Fine-tune Retrieval**
   - Optimize `NUM_RETURNED_HITS` for your reranking pool
   - Set `SEARCH_DISTANCE_CUTOFF` if quality filtering is needed
   - Add `CHUNK_OVERLAP` if context preservation is important
3. **Advanced Features**
   - Enable `ENABLE_CONTEXTUAL_RAG` if quality is paramount
   - Consider `ENABLE_RERANKING_REAL_TIME_FLOW` for better ranking
   - Evaluate `ENABLE_INFORMATION_CONTENT_CLASSIFICATION` for mixed content
4. **Performance Optimization**
   - Scale `NUM_INDEXING_WORKERS` based on available CPU cores
   - Increase `EMBEDDING_BATCH_SIZE` for better throughput
   - Adjust timeouts based on query complexity patterns

Monitoring and Validation

- **Quality Metrics**: Monitor precision, recall, and relevance scores
- **Performance Metrics**: Track response times, throughput, and resource usage
- **User Feedback**: Collect qualitative feedback on result quality and relevance

---

This document focuses on the most impactful RAG tuning parameters. For complete configuration options, refer to ONYX_ENVIRONMENT_VARIABLES.md