

A RAG implementation for OpenACS

Antonio Pisano

This will NOT be an AI talk...

- ...but rather a talk about integrating an AI tool in an OpenACS application.
- Still, I will say a few words on what RAG is

RAG in a nutshell

Large Language Model

- A computing model to simulate conversation
- Good at translating, paraphrasing and integrating texts
- Trained on a very large dataset
- It embeds some actual knowledge, but will also make up stuff
- Can be instructed via prompts to make answers more relevant or specific

Information Retrieval system

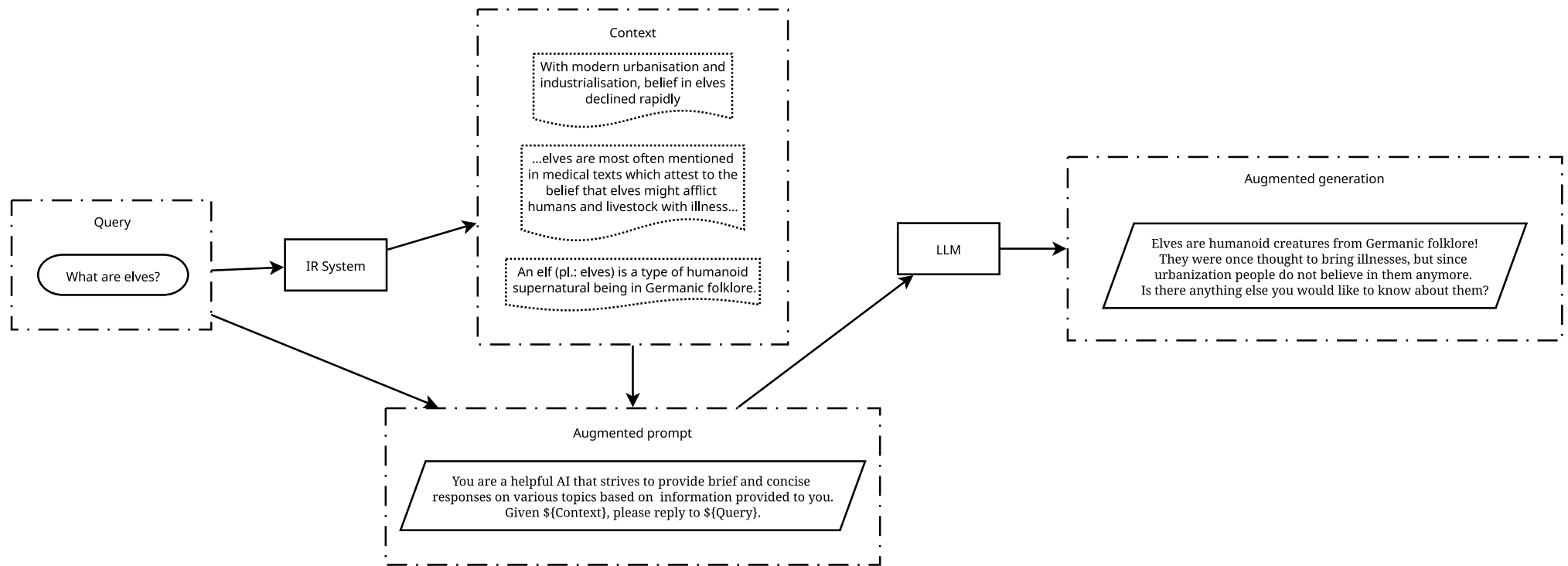
- Creates an index from a corpus of documents
- Compares a user query with documents in the index and retrieves the most relevant
- In classical IR, relevance is defined by some distance metric computed using Natural Language Processing techniques

LLM + IR = Retrieval Augmented Generation

- The combination of an LLM and an IR system
- The query is first used to extract relevant context from a document corpus
- The query and the context are combined into a prompt
- The prompt is used to elicit a relevant and accurate response from the LLM

(Optional) Semantic Information Retrieval

- Some language models are able to compute semantic vector representations of a document, or embeddings
- An information retrieval system can use vector distance over embeddings as a distance metric to provide semantically relevant results
- Although not mandatory for RAG, semantic retrieval can increase the relevance of the augmented context



Some considerations

RAG vs LLM fine tuning

- Fine tuning an LLM is
 - Computing-intensive
 - Tricky → the final result will not necessarily be a better model for our purpose
- RAG is
 - Cheap → The LLM is just a drop-in component
 - Easier to control → By controlling the context we feed to the model via our prompt, we can normally expect responses to be accurate and relevant to the query

Why would I need my own RAG system?

- Many LLM applications on the market are already some form of RAG, which normally has access to the whole Internet
- A custom RAG application makes sense when your corpus:
 - is very specific
 - is very authoritative
 - is not available on the Internet
 - is proprietary or privacy-sensitive

Ingredients of a RAG system on OpenACS

Ollama

- Open source tool to run Large Language Models locally and on consumer hardware
- Extensive library of models ready to download, including:
 - Meta Llama
 - Microsoft Phi
 - Alibaba Qwen
 - Deepseek
 - ... and various embedding models
- ChatGPT-compatible web API




OpenACS search

- Information Retrieval implementation
- Packages implementing the Full Text Search Service Contract provide callbacks to convert their objects into text representations
- The actual indexing happens via “drivers” implemented for Oracle and Postgres
- On Postgres, current driver uses the DBMS full text search capabilities

OpenACS search (2)

- Examples of packages supporting search are:
 - File Storage
 - XoWiki-based packages
 - Forums
- The drivers enforce access control → only documents we can see will be retrieved

OpenACS search (3)



OpenACS – The Toolkit for Online Communities : Search : Results

Search results for **conference 2025**

1. OpenACS and Tcl/Tk **Conference 2025**: July 10th - July 11th, **2025**
2025: Deadline for submissions of abstracts (max. 2 pages, min. 250 words); June 19th, **2025**: Notification of acceptance; tentative program June 30th, **2025**: Registration ends July 9th, **2025**: Meet & greet July 10th - July 11th, **2025**: **Conference**
</conf2025/info/index>
2. Save the Date: European OpenACS and Tcl/Tk **Conference 2025**
conferences hosted in Vienna, we are proud to announce that the joint European OpenACS and Tcl/Tk **Conference 2025**
https://openacs.org/forums/message-view?message_id=7395322

« ‹ 1 › »

Try your query on: [Ask.com](#) [Bing](#) [DuckDuckGo](#) [Google](#) [Twitter](#) [Yahoo!](#)

<https://openacs.org/search/>

pgvector

- Postgres module to store and retrieve vector data types
- Implements indexing and distance queries on vector spaces
- Can be used to store and retrieve embeddings produced by embedding models

Implementation

Wrapping the Ollama API

- The ollama API is simple and is made to resemble OpenAI
- Generated responses are typically returned via a stream
 - Gustaf Neumann introduced streaming HTTP proxy functionality for ns_http in NaviServer 5.0 release
 - Previous implementation used tcl http package streaming functionality as a workaround

Implementing a new search driver

- Content segmentation
 - Computing the embeddings works best on smaller content (~1000 words)
 - The single documents are segmented in chunks that are indexed separately
 - The chunks overlap slightly to avoid truncating sentences
 - We trust the LLM to tolerate the noise due to this naive segmentation

Document

Tcl (pronounced "tickle" or "TCL";[8] originally Tool Command Language) is a high-level, general-purpose, interpreted, dynamic programming language. It was designed with the goal of being very simple but powerful.[9] Tcl casts everything into the mold of a command, even programming constructs like variable assignment and procedure definition.[10] Tcl supports multiple programming paradigms, including object-oriented, imperative, functional, and procedural styles.

It is commonly used embedded into C applications,[11] for rapid prototyping, scripted applications, GUIs, and testing.[12] Tcl interpreters are available for many operating systems, allowing Tcl code to run on a wide variety of systems.

Because Tcl is a very compact language, it is used on embedded systems platforms, both in its full form and in several other small-footprint versions.[13]

The popular combination of Tcl with the Tk extension is referred to as Tcl/Tk (pronounced "tickle teak"[citation needed] or "tickle TK") and enables building a graphical user interface (GUI) natively in Tcl. Tcl/Tk is included in the standard Python installation in the form of Tkinter.

Document segments

Tcl (pronounced "tickle" or "TCL";[8] originally Tool Command Language) is a high-level, general-purpose, interpreted, dynamic programming language. It was designed with the goal of being very simple but powerful.[9] Tcl casts everything into the mold of a command, even programming constructs like variable assignment and procedure definition.[10] Tcl supports multiple programming paradigms, including object-oriented, imperative, functional, and procedural styles.

programming paradigms, including object-oriented, imperative, functional, and procedural styles.

It is commonly used embedded into C applications,[11] for rapid prototyping, scripted applications, GUIs, and testing.[12] Tcl interpreters are available for many operating systems, allowing Tcl code to run on a wide variety of systems. Because Tcl is a very compact language, it is used on embedded systems platforms, both in its full form and in several other small-footprint versions.[13]

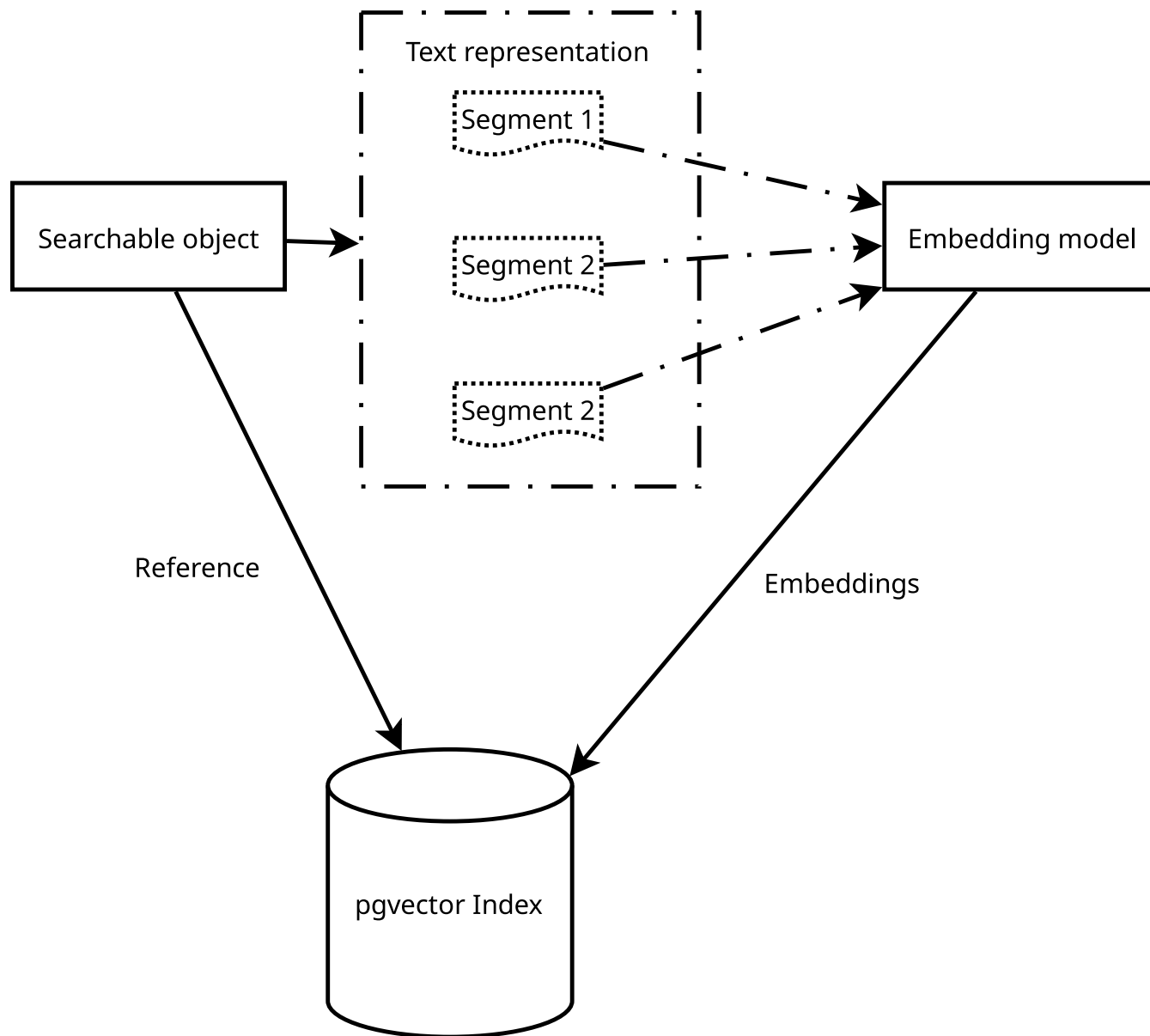
The popular combination of Tcl with the Tk extension is referred to as Tcl/Tk

Because Tcl is a very compact language, it is used on embedded systems platforms, both in its full form and in several other small-footprint versions.[13]

The popular combination of Tcl with the Tk extension is referred to as Tcl/Tk (pronounced "tickle teak"[citation needed] or "tickle TK") and enables building a graphical user interface (GUI) natively in Tcl. Tcl/Tk is included in the standard Python installation in the form of Tkinter.

Implementing a new search driver (2)

- Embedding index
 - Ollama used as a backend to compute content embeddings
 - The embeddings are computed in batches to reduce request overhead
 - pgvector Postgres extension used for storage and retrieval



Packages as knowledge base

- Our RAG package can be instantiated multiple times
- Under any such instance, subnodes mounting searchable packages are treated as “knowledge base”
- The RAG package instance will use results from this packages to provide replies

Server-Sent Events Notifications

- Indexing of documents must be asynchronous
 - Big documents = many chunks = multiple HTTP requests
 - Generating the embeddings is an intensive operation
 - Waiting time in the order of tens of seconds (or more)
- A new SSE delivery backend has been implemented for OpenACS
 - An SSE channel is open on every page (when new backend is enabled)
 - OpenACS notifications the user subscribed to are broadcast on the channel as JSON messages, produced via tDOM
 - The client renders the notifications via the Notification web API

A little tour of the final product

Conversation UI: a reply without access to the context

You:

What is bullet?

LLM:

A "bullet" can refer to different things depending on the context:

1. **Projectile:** In a general sense, a bullet is a small, pointed projectile designed to be fired from a gun.
2. **Programming language:** In computer programming, a bullet (•) is often used as a bullet point in lists.
3. **Social media:** On social media platforms like Twitter, Instagram, or Facebook, a "bullet" can refer to a post or a comment.
4. **Punctuation mark:** In writing, the bullet (•) is a punctuation mark used to indicate a list item or a section header.

Which context were you thinking of?

You:

Message (required)

What is bullet?

Model (required)

llama3.2:latest

Search in the knowledge base? (required)

No

Search the web? (required)

No

OK

Conversations list and knowledge base

[Main Site](#) : [Ollama](#)

Knowledge Base

- File Storage - </ollama/file-storage/> - [Request notifications](#)

Conversations

[≤](#) [1 \(current\)](#) [2](#) [≥](#)

New conversation

Title	First Message	Last Message	N. Messages	
tcl-socket-implementation	12/18/24 10:56 AM	12/18/24 10:57 AM	3	Delete
"Socket Example in Tcl"	12/18/24 11:11 AM	12/18/24 11:11 AM	2	Delete

[Main Site](#) : [Ollama](#) : File Storage

[Add File](#) [Create a URL](#) [New Folder](#) [Upload Compressed Folder](#) [Modify permissions on this folder](#)

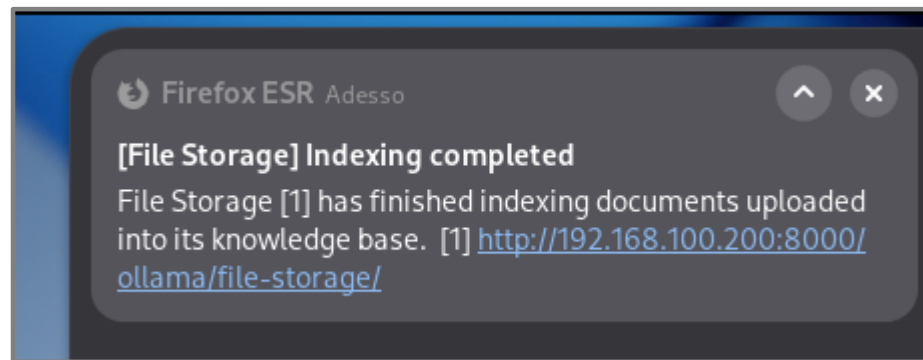
<input type="checkbox"/>	Type ↑↓	Name ↑↓	Size ↑↓	Last Modified ↑↓			
<input type="checkbox"/>	PDF	Bullet_User_Manual.pdf bullet_user_manual.pdf	2.8 MB	06/10/25 12:32 PM	properties	New	Download

[Move](#) [Copy](#) [Delete](#) [Download as a ZIP file](#)

[Download an archive of the contents of this folder](#)

Note: This may take a while, please be patient.

SSE notifications



Conversation UI: a reply with RAG

You:

What is bullet?

References:

- [bullet_user_manual.pdf](#)
- [bullet_user_manual.pdf - 2](#)
- [bullet_user_manual.pdf - 3](#)
- [bullet_user_manual.pdf - 4](#)
- [bullet_user_manual.pdf - 5](#)

LLM:

Bullet Physics is a professional open source collision detection, rigid body and soft body dynamics engine free for commercial use under the ZLib license.

You:

Message (required)

What is bullet?

Model (required)

llama3.2:latest

Search in the knowledge base? (required)

Yes

Search the web? (required)

No

OK

Conversation UI: clicking on the reported context display the segment contents

bullet_user_manual.pdf

can press the F3 key in most of the Bullet examples to save a 'testFile.bullet'. You can read .bullet information about .bullet serialization is at the Bullet wiki at <http://bulletphysics.org/mediawiki-1.5> bulletphysics.org 13 General Tips Avoid very small and very large collision shapes The minimum object size should be manipulated, reduce the internal simulation frequency accordingly, using the third argument of btDynamicsWorld (a gravity of 9.8m/s2) requires a frequency of at least 300Hz (1./300.). It is recommended to keep the simulation frequency high. Simulation becomes unstable when a heavy object is resting on a very light object. It is best to keep multiple static triangle meshes into one Many small btBvhTriangleMeshShape pollute the broadphase. Use at least 60 hertz (1/60 second). For safety and stability, Bullet will automatically subdivide the variable timestep (if stepSimulation is set to true). When the timestep is smaller then the internal substep, Bullet will interpolate the timestep (if stepSimulation is set to true): the internal timestep and substeps are disabled, and the actual timestep is simulated. Don't build a ragdoll out of btHingeConstraint and/or btConeTwistLimit for knees, elbows and arms. Don't use a large gap is noticeable, please compensate the graphics representation. ©Erwin Coumans Bullet 2.83 Physics SDK. The number of vertices in a btConvexHullShape is limited. It is better for performance, and too many vertices in a triangle mesh Keep the size of triangles reasonable, say below 10 units/meters. Also degenerate triangles are bad. btQuickProf bypasses the memory allocator If necessary, disable the profiler when checking for memory leaks. #define BT_NO_PROFILE 1 in Bullet/src/LinearMath/btQuickProf.h ©Erwin Coumans Bullet 2.83 Physics SDK. By default, there is only one friction value for one rigidbody. You can achieve per shape or per triangle friction by using the CF_CUSTOM_MATERIAL_CALLBACK to the collision flags or the rigidbody, and register a global material callback. This matches the triangleId/partId of the striding mesh interface. An easier way to use custom materials is by using the Solvers Bullet uses its btSequentialImpulseConstraintSolver by default. You can use a different constraint solver in Bullet/src/BulletDynamics/MLCPSolvers. See the source code of examples/vehicles/VehicleDemo. For certain types of objects, you can register a friction function in the constraint solver for certain body types. USER_DEFINED_FRICTION_MODEL in Demos/CcdPhysicsDemo.cpp. ©Erwin Coumans Bullet 2.83 Physics SDK. For collision detection. We implement from scratch a rigid body and collision detection pipeline that runs 100% on GPUs (NVIDIA 680 or newer). A simple OpenCL example is disabled by default in the example browser. If you have a GPU, you can enable_experimental_opengl Note that there are many reasons why the OpenCL kernels fail, and you can find more background about the OpenCL collision detection and rigid body pipeline in the Bullet/docs folder. CRC Press. This book is also available from Amazon. ©Erwin Coumans Bullet 2.83 Physics SDK. More information on our website at <http://bulletphysics.org> for a discussion forum, a wiki with frequently asked questions and a blog. [http://en.wikipedia.org/wiki/Bullet_\(software\)](http://en.wikipedia.org/wiki/Bullet_(software)) Authoring Tools • Dynamica Maya plugin and Bullet C++ • Bullet physics support: <http://www.blender.org> • COLLADA physics standard: <http://www.khronos.org/collada> • discussed voronoi simplex solver for GJK • Collision Detection in Interactive 3D Environments, Gino Baraff • algorithms, very useful to understand Bullet • Physics Based Animation, Kenny Erleben <http://www.khronos.org/collada>

[See full size](#)

Similarity: 0.48458996874392135

Close

Settings: various parameters can be customized

General

default_generation_model (default:)

llama3.2

i Default model for text generation.

ollama_host (default: <http://localhost:11434>)

<http://192.168.122.1:11434>

i Ollama host instance.

Indexing

embedding_batch_size (value is default)

50

i Number of documents to be sent simultaneously to the Ollama backend to generate embeddings.

embedding_model (value is default)

all-minilm

i Model used to generate embeddings. See <https://ollama.com/search?c=embedding> for possible values.

indexing_chunk_overlap (value is default)

100

i By how many words will chunk overlap when indexing content.

indexing_chunk_size (value is default)

1000

i Number of words for each chunk when splitting the content for indexing.

Settings (2): various parameters can be customized

Rag

rag_context_template (value is default)

Use the following context as your learned knowledge, enclosed within <context></context> XML tags.

<context>

\$context


</context>

When answering the user:

- If you don't know the answer, simply state that you don't know.
- If you're unsure, seek clarification.
- Avoid mentioning that the information was sourced from the context.
- Respond in accordance with the language of the user's question.

Given the context information, address the query.

Query: \$query

 Template that will be added to the chat prompt in order to instruct the LLM to use additional context when providing a reply. It can contain the variables "context" (the actual extra content) and "query" (the question to the model), which will be substituted.

rag_top_k (value is default)

5

 Max number of entries, sorted by decreasing relevance, to retrieve when fetching context for Retrieval Augmented Generation.

websearch_p (default: 0)

1

 Allow to search the web when querying the model?

Search

similarity_threshold (value is default)

0.9

 Minimal similarity threshold used to decide if a document is relevant. This value expresses a cosine similarity in the 0 (highest) to 2 (smallest) range.

Streaming response in action

You:

Message (required)

Model (required)

llama3.2:latest



Search in the knowledge base? (required)

Yes





Search the web? (required)

No



OK

This website is maintained by the OpenACS community. Any problems, email [webmaster](#) or [submit](#) a bug report.
(Powered by Tcl , Next Scripting , NaviServer 5.0.0a , IPv4)

Acknowledgments

- The RAG workflow implemented in this project is similar in principle to that in Open WEBUI (<https://docs.openwebui.com/>)
- Many thanks to Gustaf Neumann as always for quickly addressing my needs for new NaviServer features :-)

Thanks for watching!

- My contacts
 - antonio@elettrotecnica.it
 - <https://github.com/Elettrotecnica>
- Links
 - <https://openacs.org/>
 - <https://ollama.com/>
 - <https://github.com/pgvector/pgvector>
 - <https://github.com/Elettrotecnica/openacs-ollama>