

Development of a Chatbot with AnythingLLM for Scientific Knowledge and Applications

Veniamin Malefioudakis
PhD Candidate

Department of Electronics Engineering
Hellenic Mediterranean University
2026

03.06.2026

🎯 Learning Objectives

Participants will understand:

- The **architecture of AnythingLLM**
- How to **create and configure Knowledge Bases**
- Use of **semantic search and context retrieval**
- How to design **prompts and domain-specific workflows**
- Integration of **AnythingLLM with APIs**

**What we may consider as Artificial
Intelligence (AI)?**

Definition try

- *Systems that can receive, learn, reason, plan, act based on models of the real or virtual world (Philipp Slusallek, DFKI, [AI: Intelligent, Machines, Smart Policies – Conference Summary](#), OECD Digital Economy Papers, 2018, pp 11)*
- *Artificial intelligence is the simulation of human intelligence processes by machines ([Ania Kubow, freeCodeCamp 2023](#))*

What Can AI Do and What Can It Not Do?

Can do:	Cannot (up to now) do:
Information retrieval	Develop emotional intelligence
Pattern recognition	Make ethical decisions
Text generation	Reason like humans
Decision support	Human supervision is required

AI Assistants



AI assistants act as:

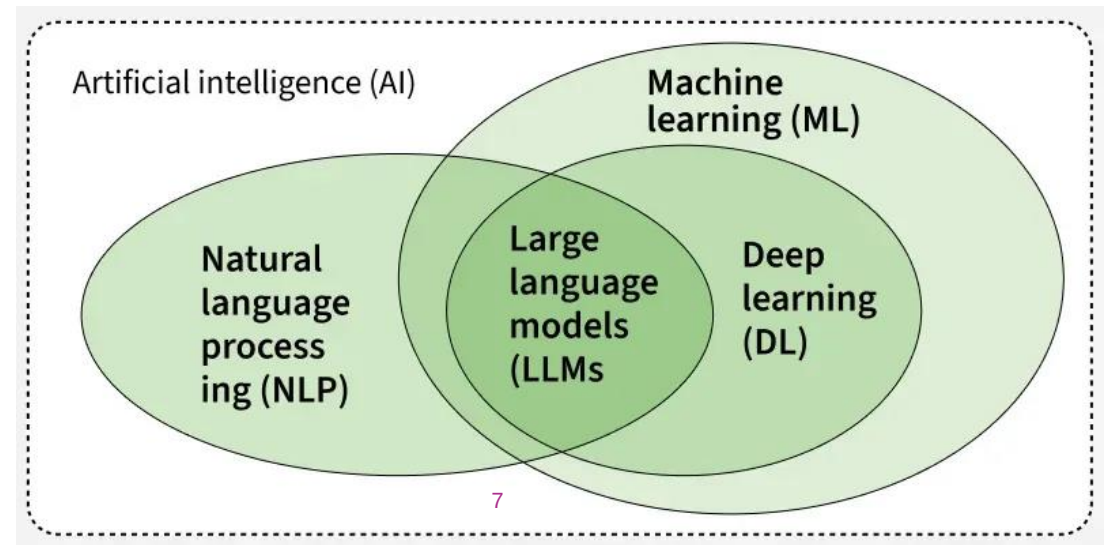
- digital knowledge support systems
- tools for information search and summarization
- interactive communication systems

Important principle:

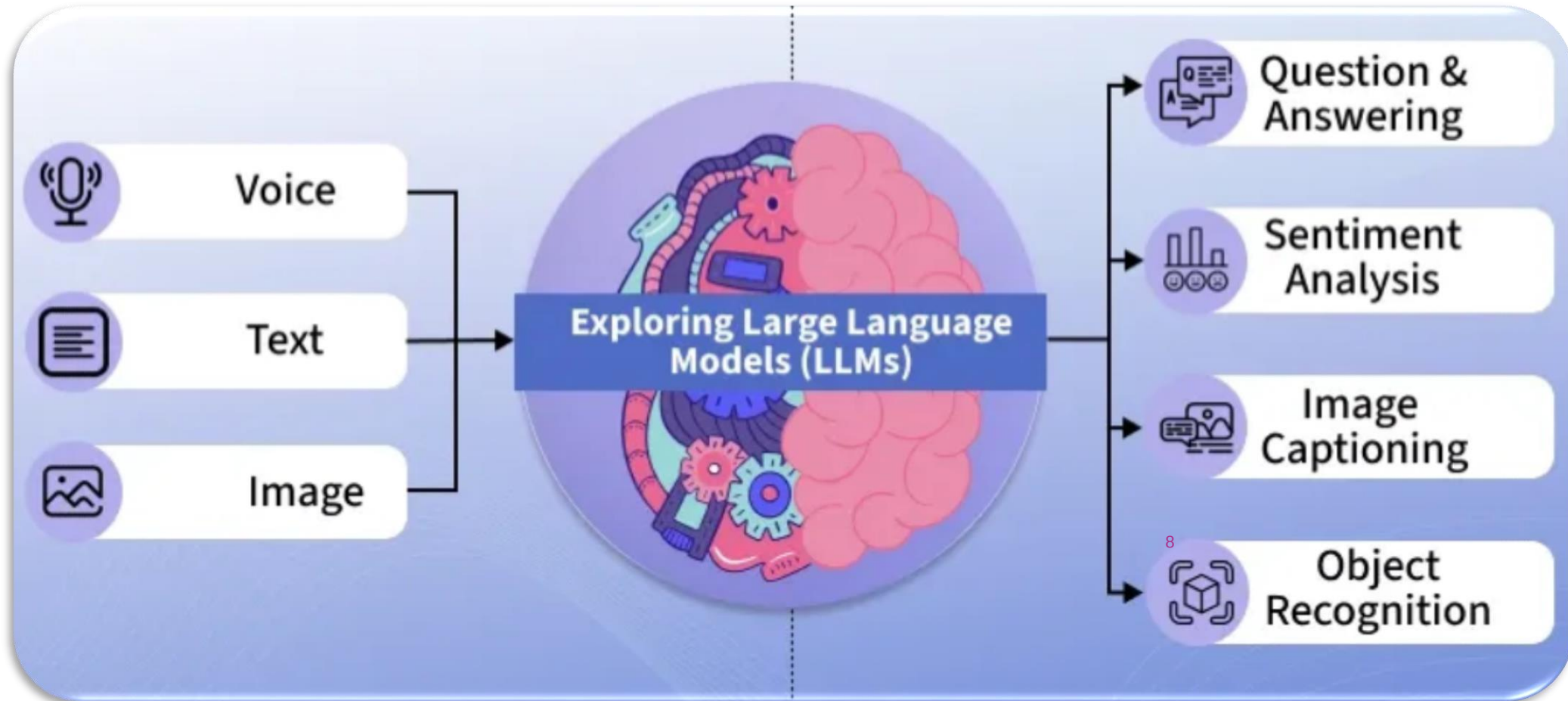
- AI assists human work – it does not replace professional decision-making.

Large Language Models – LLMs

- Large Language Models (LLMs) are artificial intelligence systems trained on massive amounts of text to understand and generate natural language
- **learn** statistical patterns of human language so they can recognize concepts, relationships, and context within a text
- **answer** questions, summarize information, translate, and compose text in a human-like manner
- **do** not “think” like humans, but predict which word or phrase is most likely to follow in a given context
- **require** reliable knowledge sources and verification mechanisms when used in sensitive areas such as healthcare



Large Language Models – LLMs



What are LLMs, and what can they do?



ChatGPT is a generative AI application developed by OpenAI. It is based on the multilingual and multimodal large language model GPT-5.1 (Generative Pre-trained Transformer – Generative Pre-trained Transformer version 5.1) and is designed to generate human-like text based on the inputs (prompts) it receives.



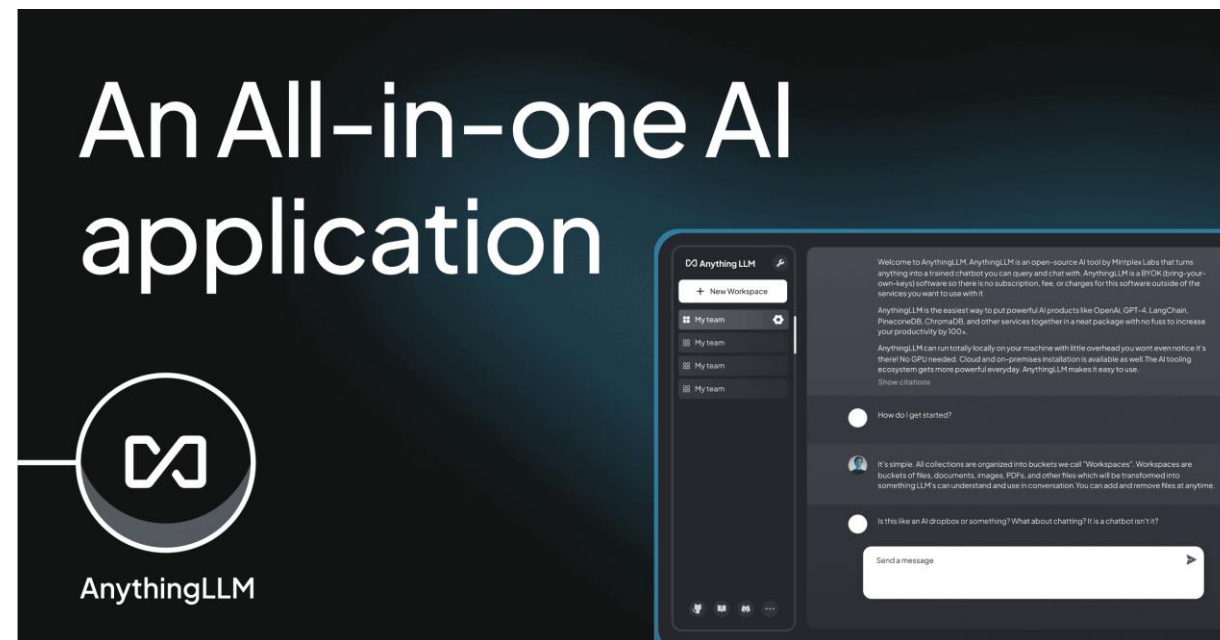
DeepSeek is also a Generative TL application developed by DeepSeek. It is based on the DeepSeek-R1 Large Language Model, which also features reasoning capabilities (checking, evaluating, and improving responses), while its training cost is lower than that of other Large Language Models (approximately 1/10).



Claude is also a series of LLMs (Claude 3 Opus (flagship), Claude 3.5 Sonnet (best cost-performance ratio), and the lightweight Claude 3 Haiku). It has been trained to function as a helpful, honest, and harmless conversational AI assistant. It can be used for various tasks, such as text composition, learning, coding, solving math problems, and image analysis.

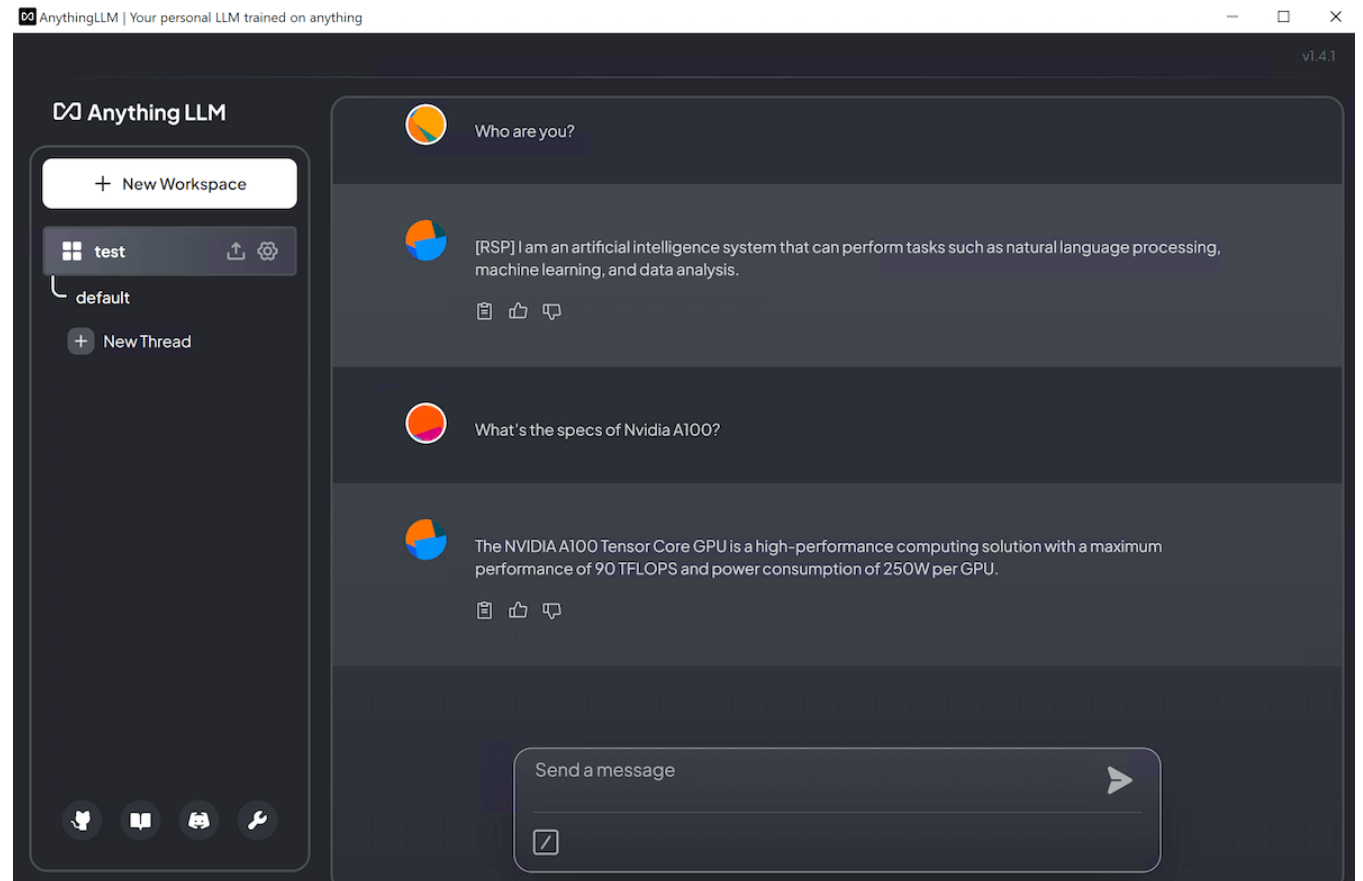
What is AnythingLLM?

- This is an open-source platform that allows for customization
- Integration of Artificial Intelligence (AI) and knowledge
- Provides mechanisms for organizing and searching documents
- Supports access control based on user roles
- Configuration allows for secure system settings



What is AnythingLLM?

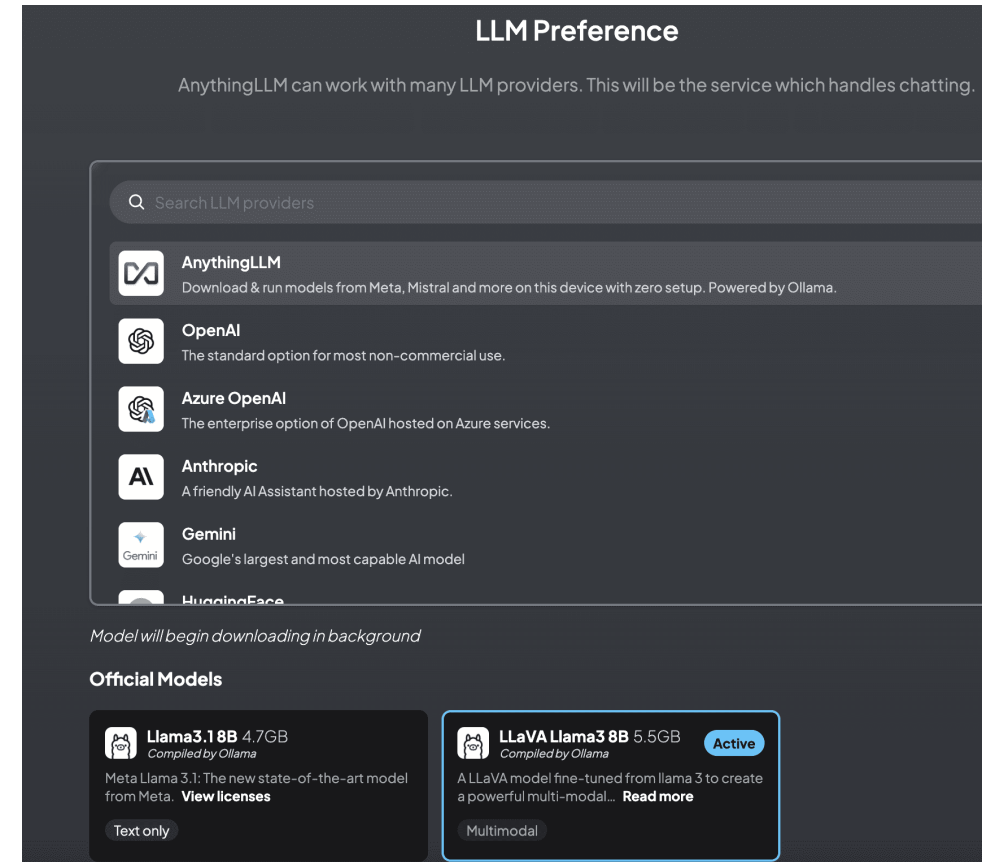
- Supports on-premises or cloud deployment
- Integration with APIs (OpenAI, Anthropic, DeepSeek, Ollama, Gemini)
- Integration of knowledge bases from files, URLs, or databases
- Focus on privacy, modularity and



AnythingLLM in Education

- Smart tools for research and clinical support
- **A tool to support, not replace**
- Connection to scientific sources
- Adaptation to healthcare needs
- Knowledge-based approach

7/21/2026



Why not just ChatGPT, DeepSeek... etc;

- Does not verify sources
- Lacks specialized knowledge
- There is a risk of inaccurate answers if sources and cross-checking are not required
- Source traceability is not ensured without citation mechanisms
- Behavior and source verification is limited to simple chatbots



Retrieval-Augmented Generation (RAG)

RAG is a technique in which the system first searches for information from reliable sources and then generates a response based on

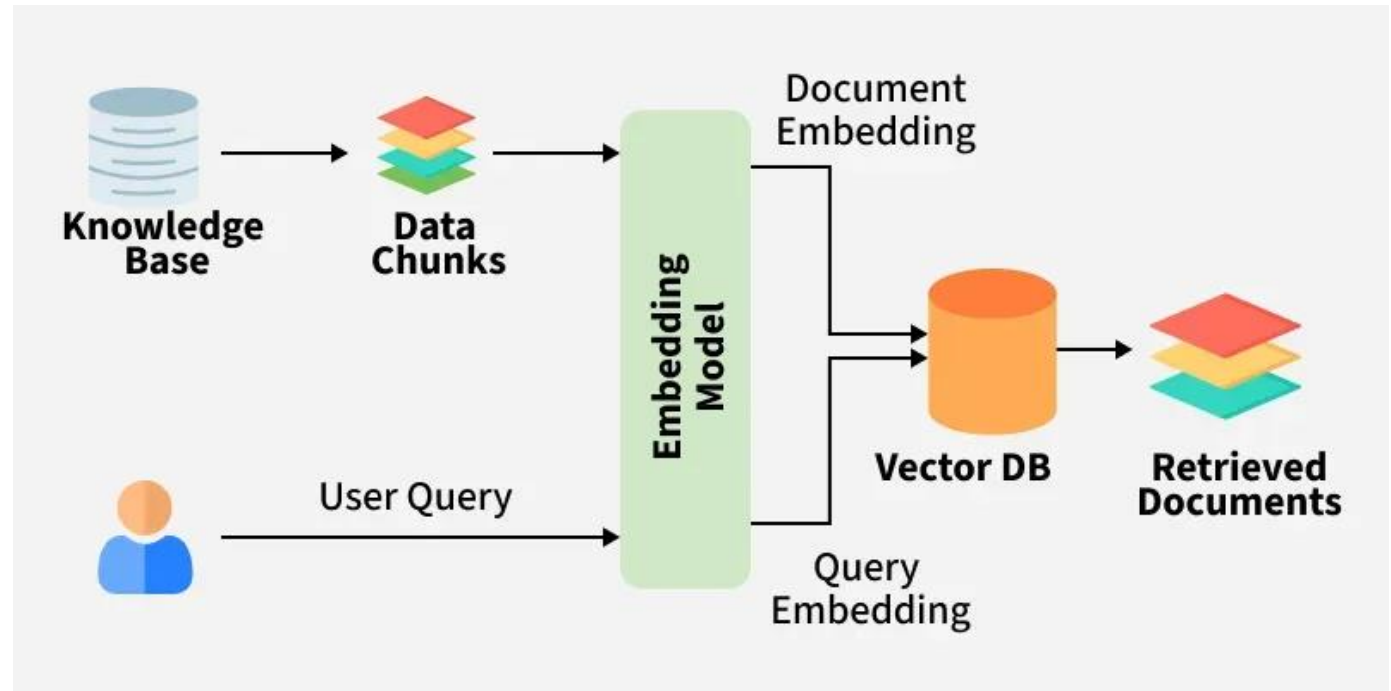
↓ The process begins with a search of reliable sources

↓ An answer is then generated

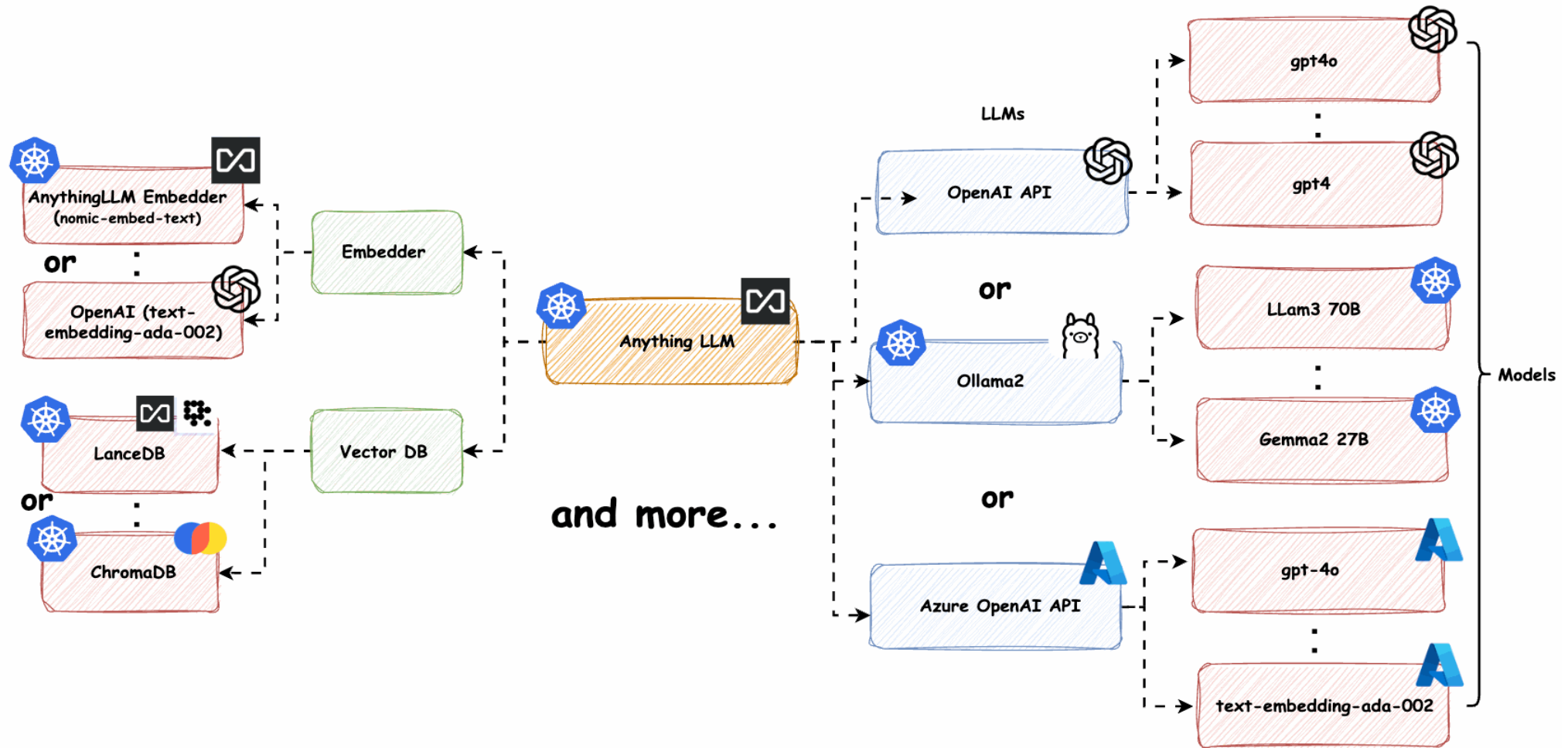
↓ The answer is based on evidence-based knowledge

↓ The methodology reduces the risk of errors

↓ Credibility is enhanced through

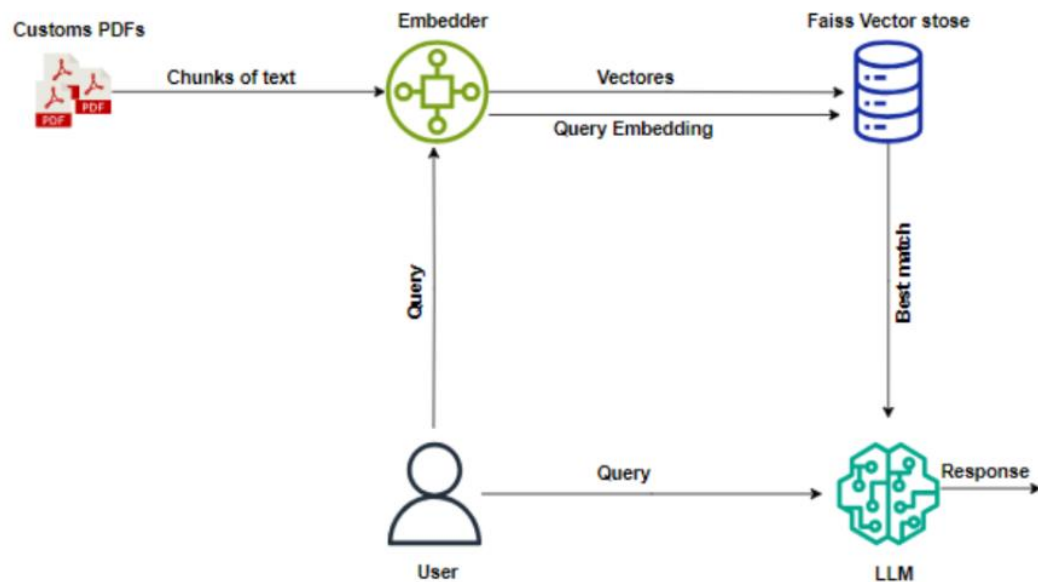


RAG flow in AnythingLLM



Knowledge base

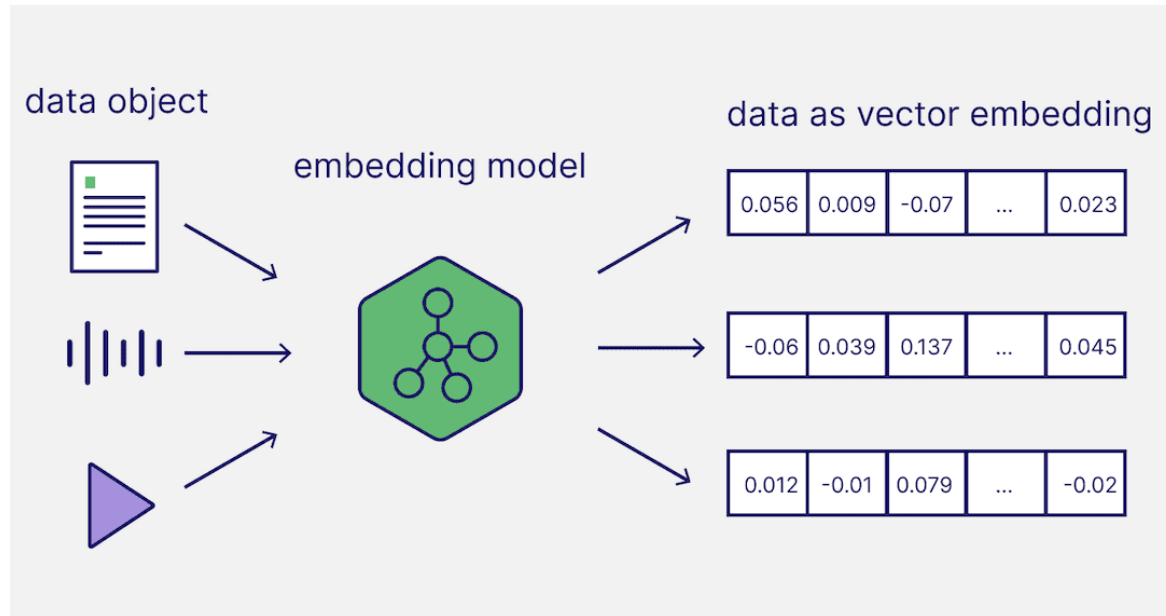
the collection of files and documents that the system uses to provide well-founded responses



- PDFs and articles
- Search guidelines are provided
- Health data is handled with care
- The database is updated periodically
- All documents are managed centrally

Embeddings

are mathematical representations of text that enable meaning-based search



- Documents are broken down into meaningful segments
- The meaning is stored for searching
- Searching is faster
- Similarities are identified based on meaning
- The results are more accurate

Why is Prompt Engineering important?

AI doesn't read minds; it reads text (prompts). If your prompt is vague, you will most likely receive a vague response that may contain inaccuracies.

As a teacher, write a few words about active learning in education.

Active learning is an educational approach that engages students in the learning process through activities, discussion, and real-world applications. Instead of passively listening to lectures, students take an active role in constructing their understanding.

There are three main types of active learning strategies:

- **Collaborative Learning:** Students work in pairs or small groups to solve problems, share ideas, and learn from each other.
- **Problem-Based Learning:** Students explore real-life problems and develop solutions by researching, discussing, and applying their knowledge.
- **Interactive Activities:** Methods such as quizzes, debates, case studies, and simulations encourage participation and deeper understanding.

The main benefits of active learning include improved motivation, better retention of information, development of critical thinking skills, and stronger communication abilities.

Teachers can apply active learning by using open-ended questions, encouraging student participation, providing feedback, and connecting lessons to real-world situations.

In summary, active learning helps students become more engaged, responsible, and successful learners.



Vague prompt => vague response with potential inaccuracies

ChatGPT has not received information regarding:
the type of text (context)
the objective
the writing style
the tone
the audience
the response format and length

COSTAR framework: Context, Objective, Style, Tone, Audience, Response Format

Improved prompting based on COSTAR:

As a teacher, create a 5-step study plan for high school students to improve their learning and exam performance, in simple language, actionable, with bullet points and tips, maximum more than 200 words.

5-Step Study Plan for High School Students

✓ 1. Set Clear Goals

- Decide what you want to achieve (e.g., improve grades in math).
- Break your goals into small, realistic targets.

📅 2. Plan Your Time

- Create a weekly study schedule and stick to it.
- Use a planner or calendar to organize your tasks.
- Allocate specific time slots for each subject.

📖 3. Study Smart

- Take active notes and summarize key concepts.
- Use techniques like mind maps, flashcards, and practice questions.
- Avoid last-minute cramming; study in short, focused sessions.

💡 4. Review and Practice

- Review your notes regularly.
- Solve past exams and sample questions.
- Teach the material to someone else to reinforce understanding.

💖 5. Take Care of Yourself

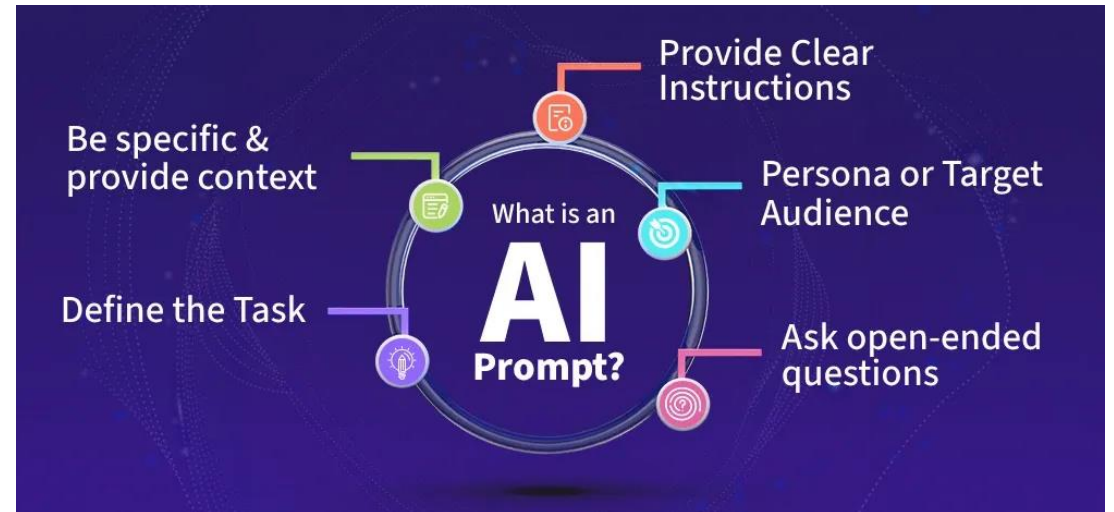
- Get enough sleep, eat healthy, and take breaks.
- Stay positive and celebrate small wins!
- Ask for help when you need it.

★ **Tip:** Consistency is the key! Small daily efforts lead to big results.

Prompt

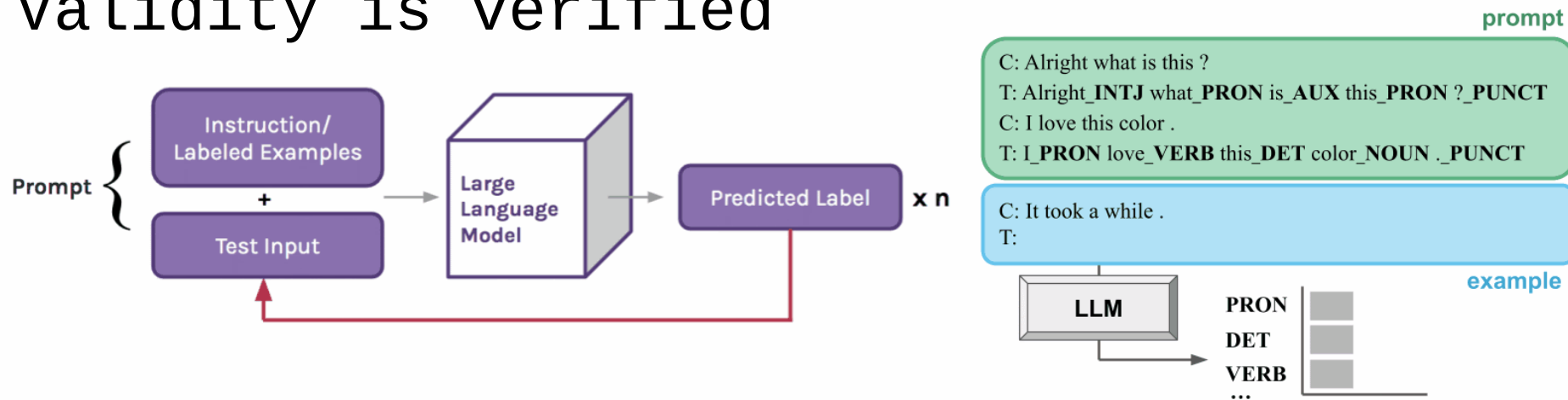
is the natural language instruction that the user gives to the system regarding how to generate a response

- System and user roles are defined
- Behavioral boundaries are established
- Sources are cited in responses
- Writing style is adapted
- The structure of responses is checked

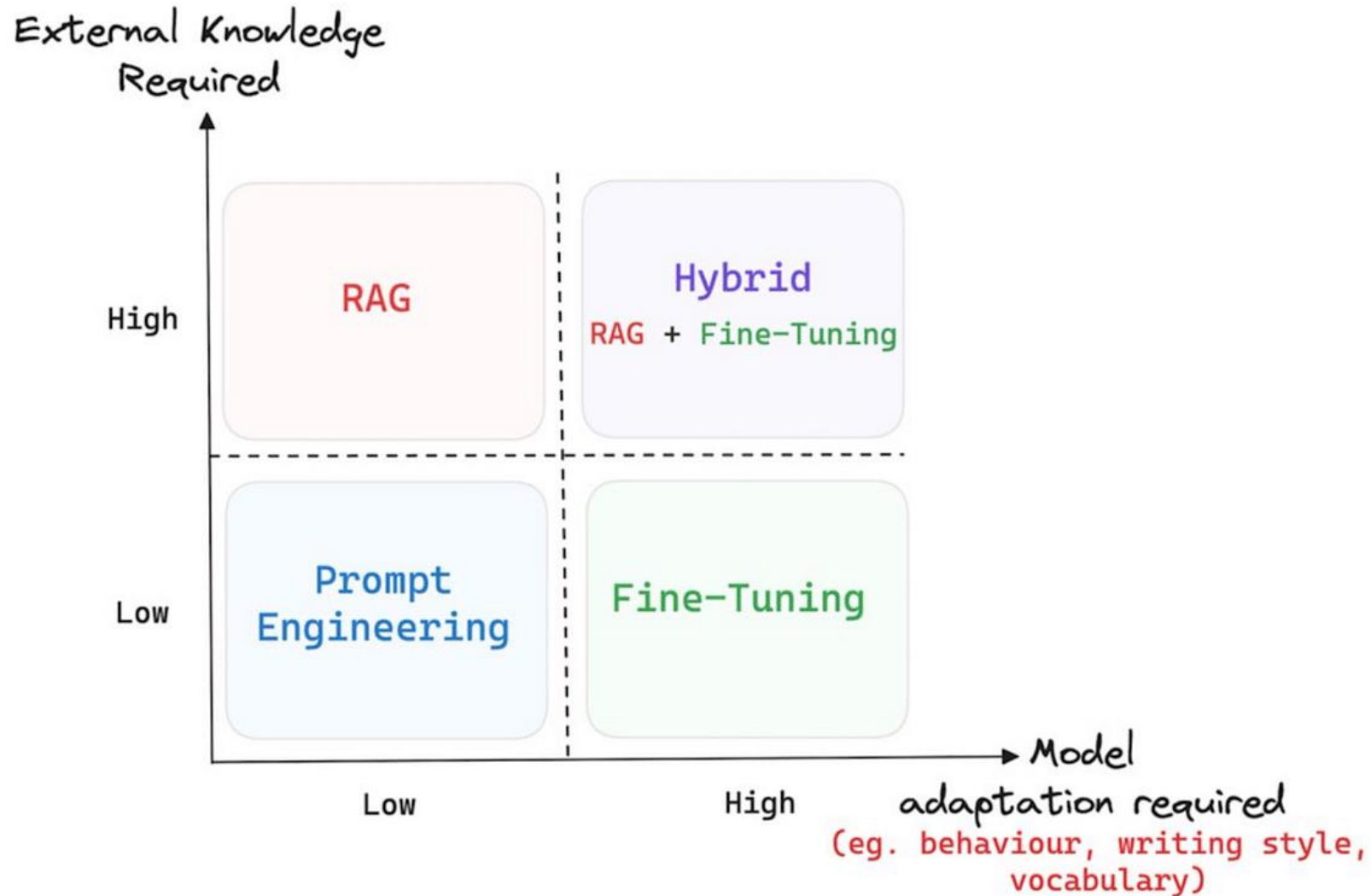


Prompt

- A clear question is formulated
- Bibliographic references are requested
- The purpose of the answer is defined
- The format of the answer is specified
- Validity is verified



General Comparison



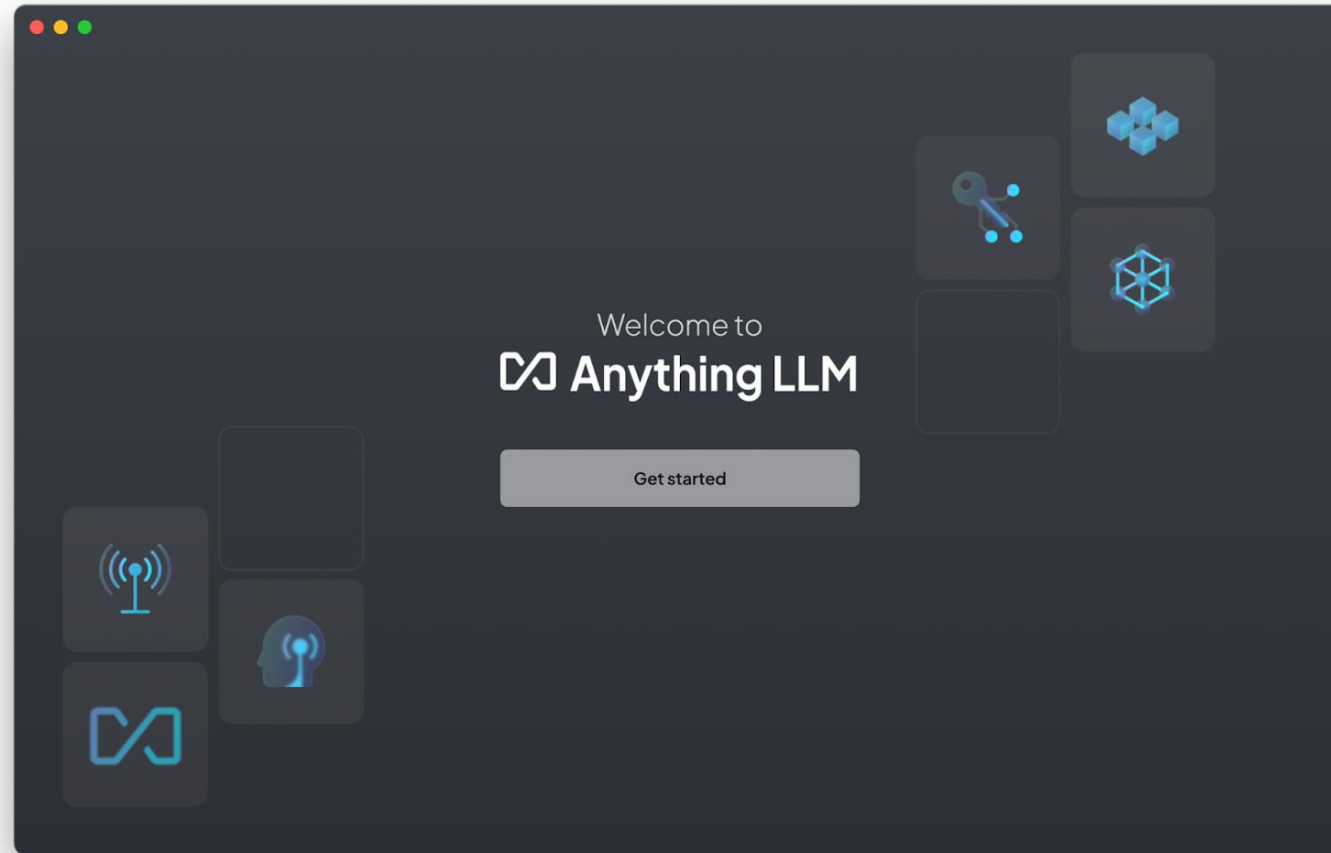
Scenario: Programming Education

- Use of curriculum-aligned resources
- Comparison of coding examples
- Summarization of concepts
- Provision of references
- Support for learner training

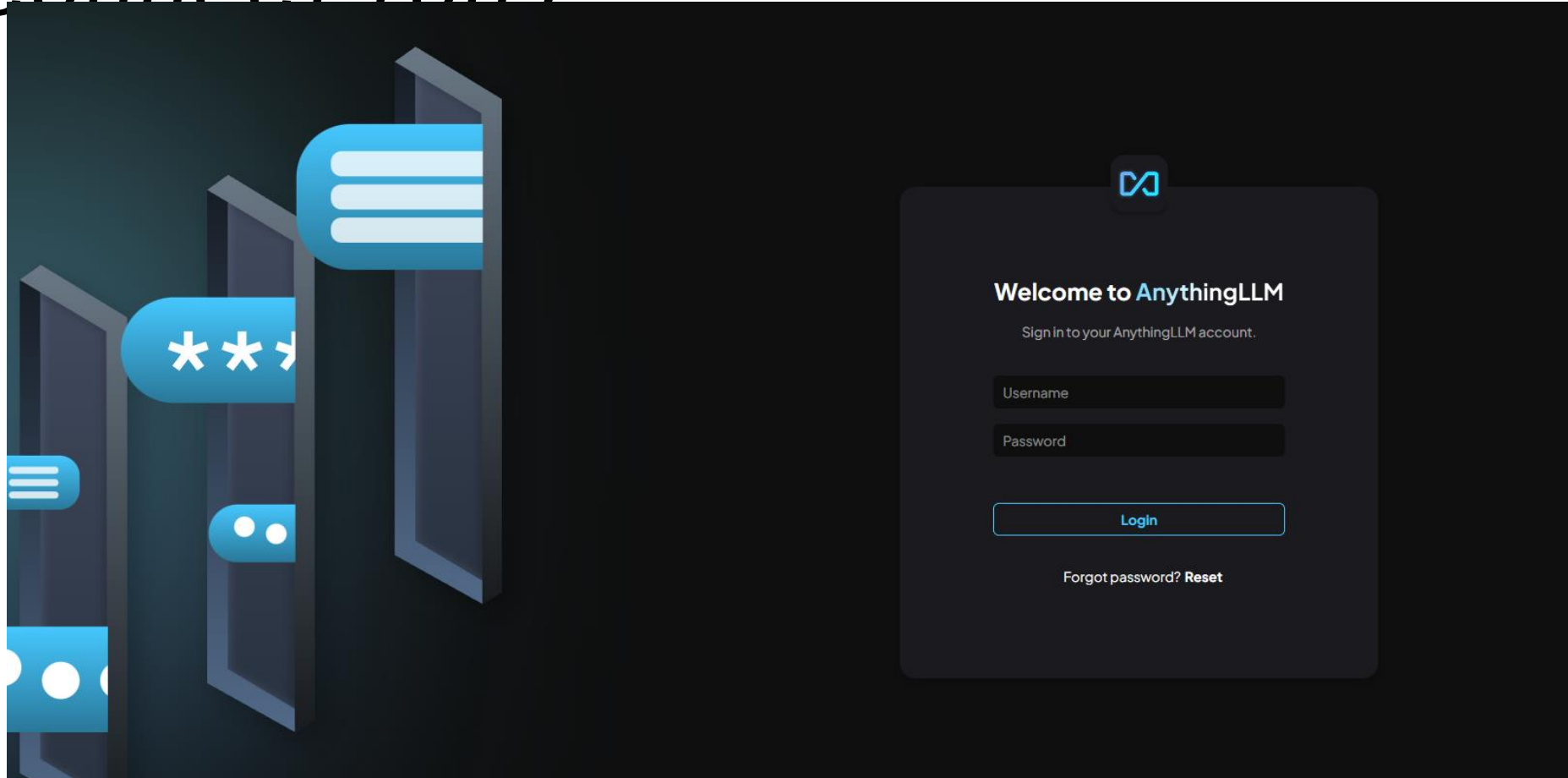
Scenario: Digital Skills Development

- Identification of updated guidelines
- Use of international evidence
- Comparison of educational data
- Support for continuous updates
- Improvement of knowledge

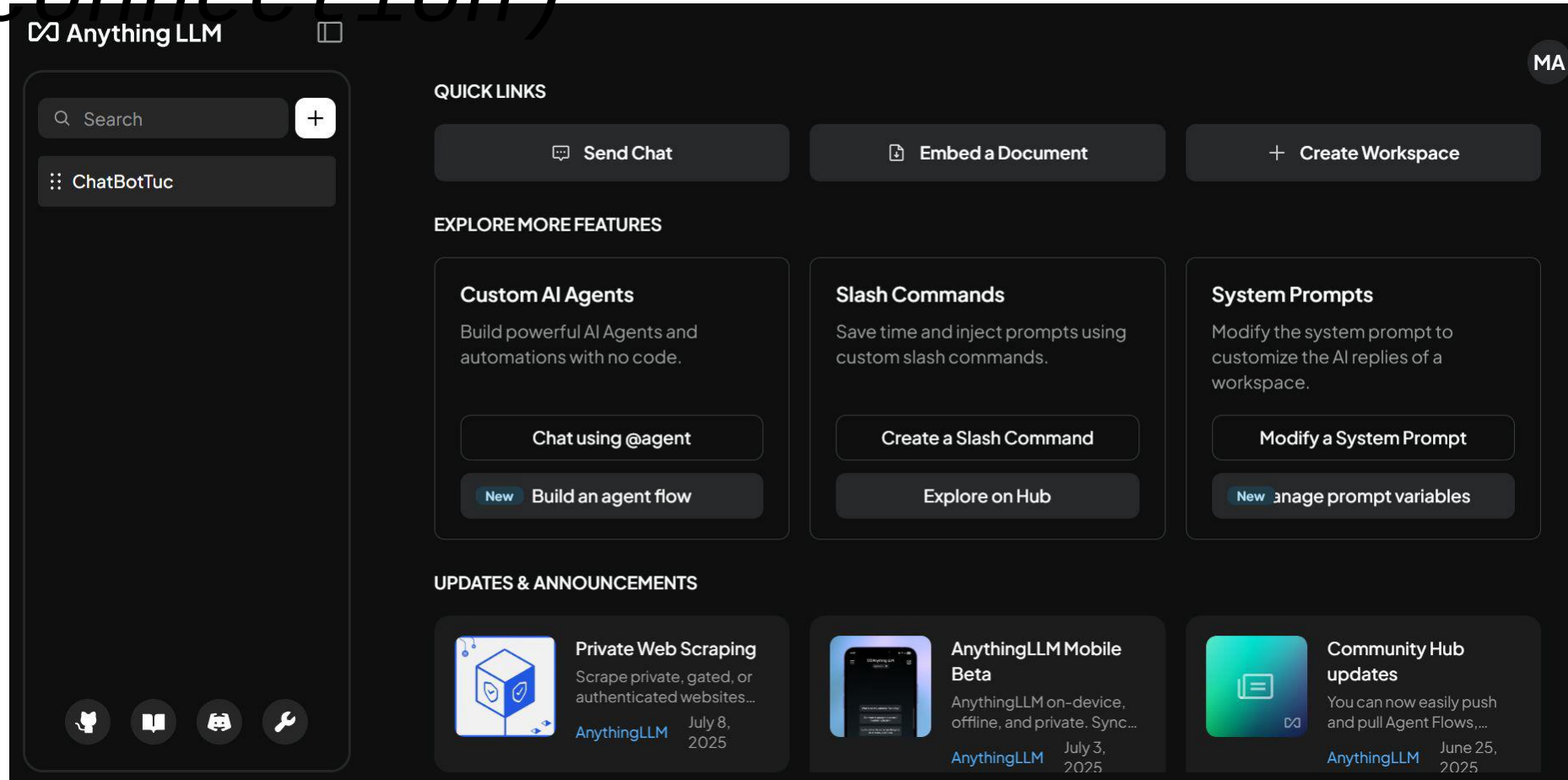
A practical example for using AnythingLLM



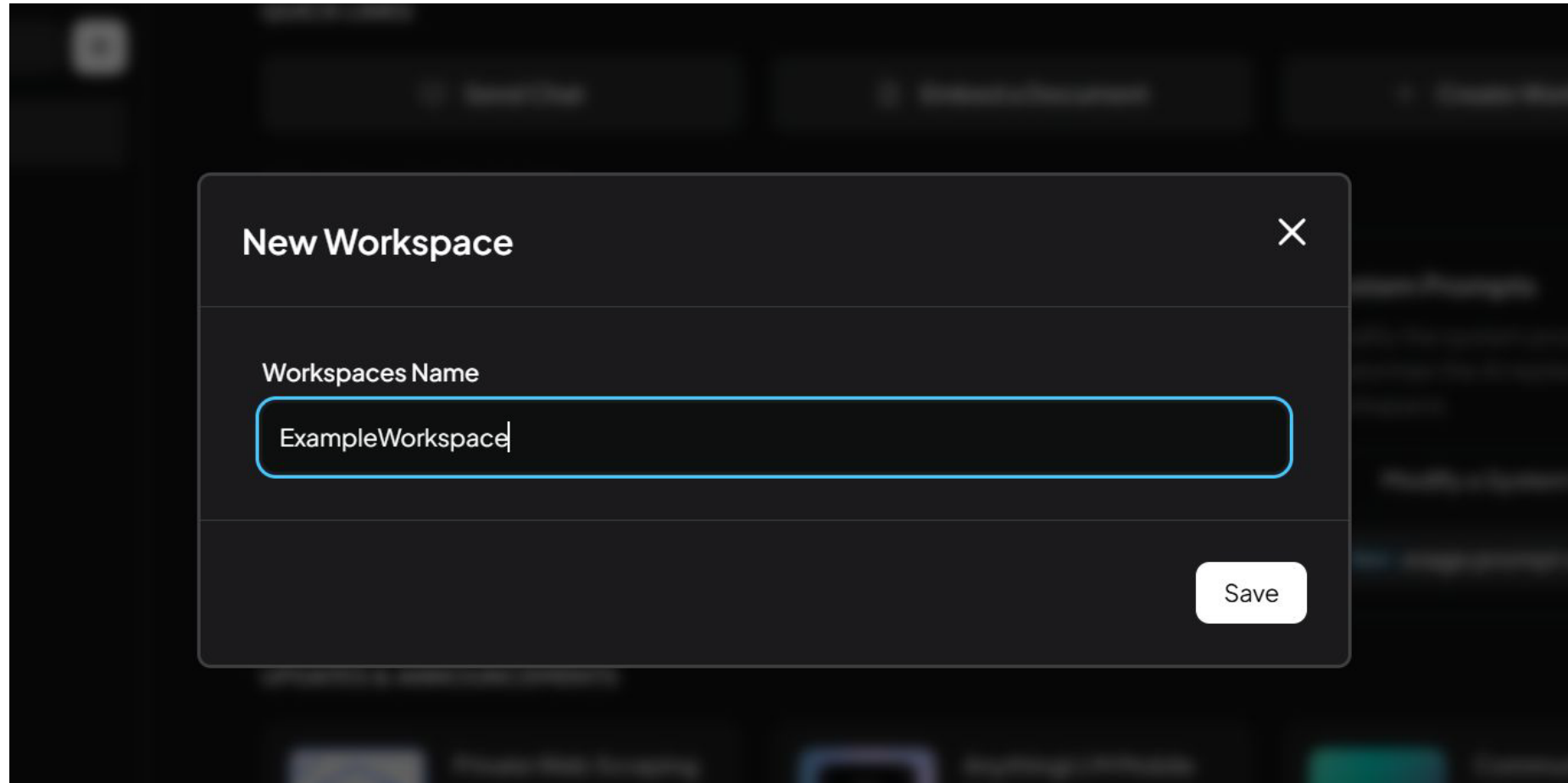
Basic steps – AnythingLLM (*Initial Environment before connection*)



Basic steps – AnythingLLM (*Initial Environment before connection*)



Basic steps – Anyt[icon]ingLLM (*Create workspace*)



A screenshot of a 'New Workspace' dialog box from the Anyt[icon]ingLLM application. The dialog box is dark-themed with a title bar that says 'New Workspace' and a close button (X) in the top right corner. Below the title bar, there is a label 'Workspaces Name' followed by a text input field containing the text 'ExampleWorkspace'. At the bottom right of the dialog box, there is a 'Save' button.

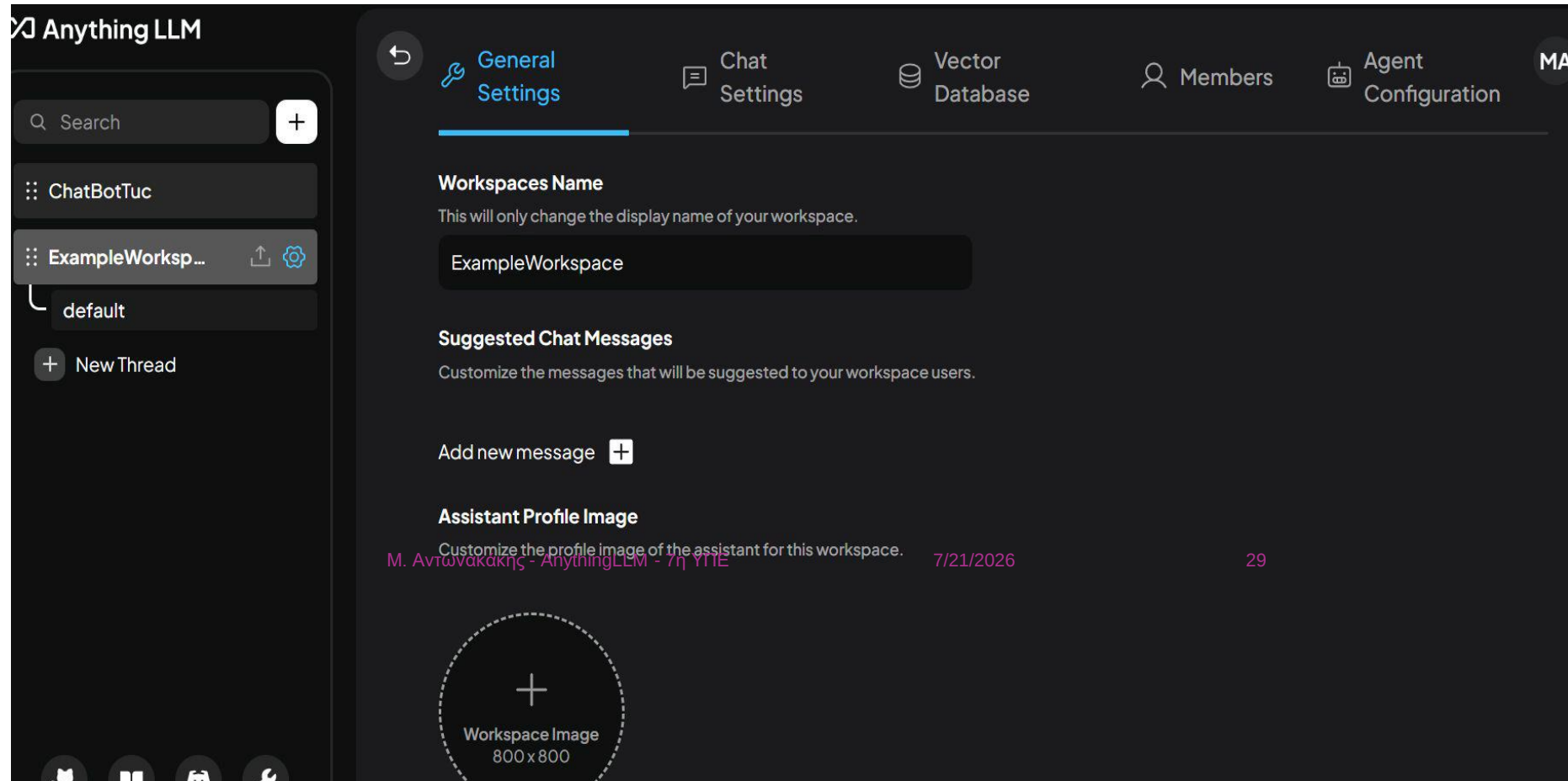
New Workspace

Workspaces Name

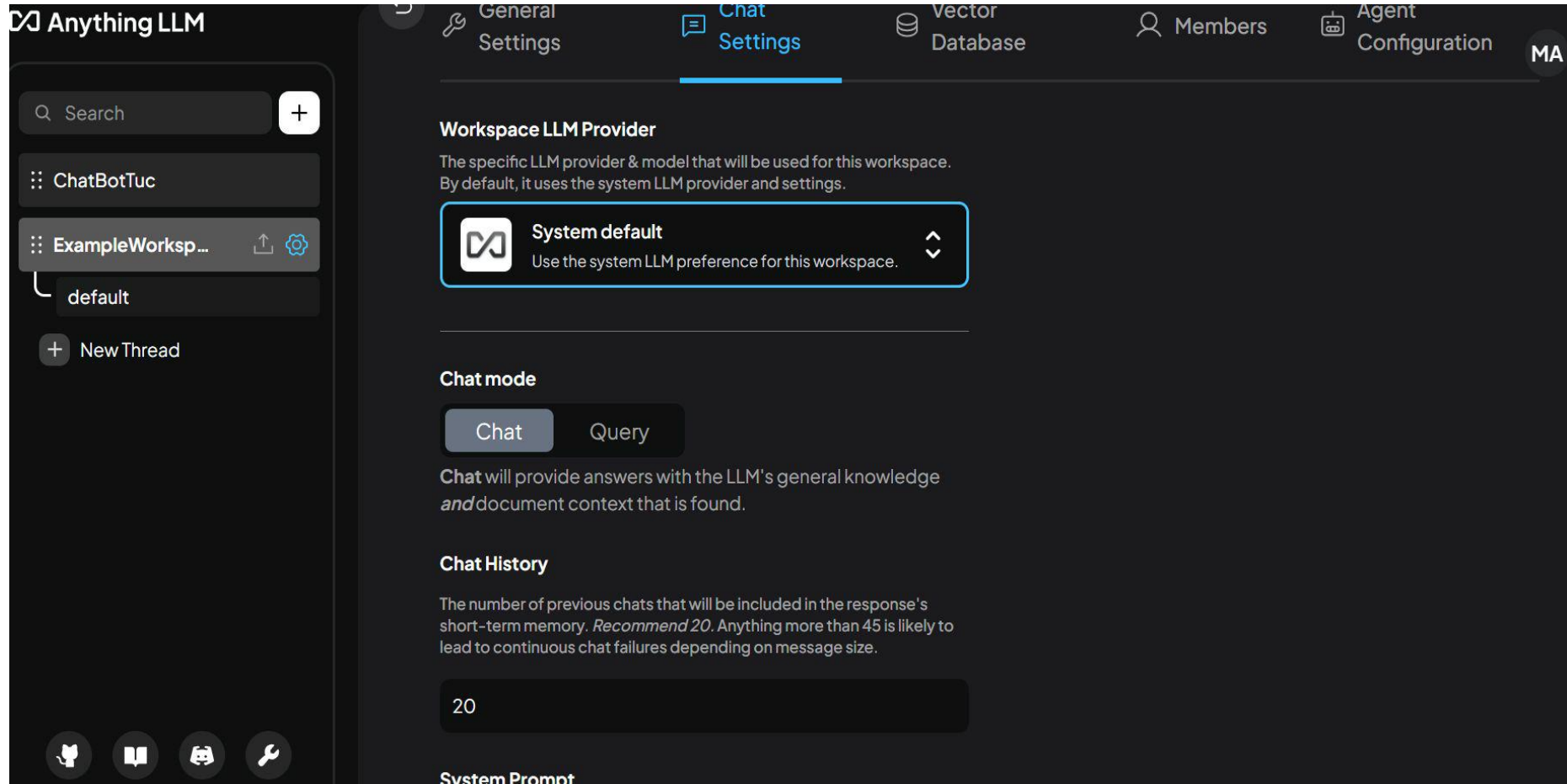
ExampleWorkspace

Save

Basic steps – AnythingLLM (*Chatbot parametrization*)



Basic steps – AnythingLLM (*Chatbot parametrization*)



Basic steps – Anyt[icon]gLLM (*Chatbot parametrization*)

Chat History

The number of previous chats that will be included in the response's short-term memory. *Recommend 20.* Anything more than 45 is likely to lead to continuous chat failures depending on message size.

20

System Prompt

The prompt that will be used on this workspace. Define the context and instructions for the AI to generate a response. You should provide a carefully crafted prompt so the AI can generate a relevant and accurate response.

You can insert **prompt variables** like: {time} , {date} , {datetime} , +5 [more...](#)

[View History](#)

Given the following conversation, relevant context, and a follow up question, reply with an answer to the current question the user is asking. Return only your response to the question given the above information following the users instructions as needed.

Query mode refusal response

When in **query** mode, you may want to return a custom refusal response when no context is found.

There is no relevant information in this workspace to answer your query.

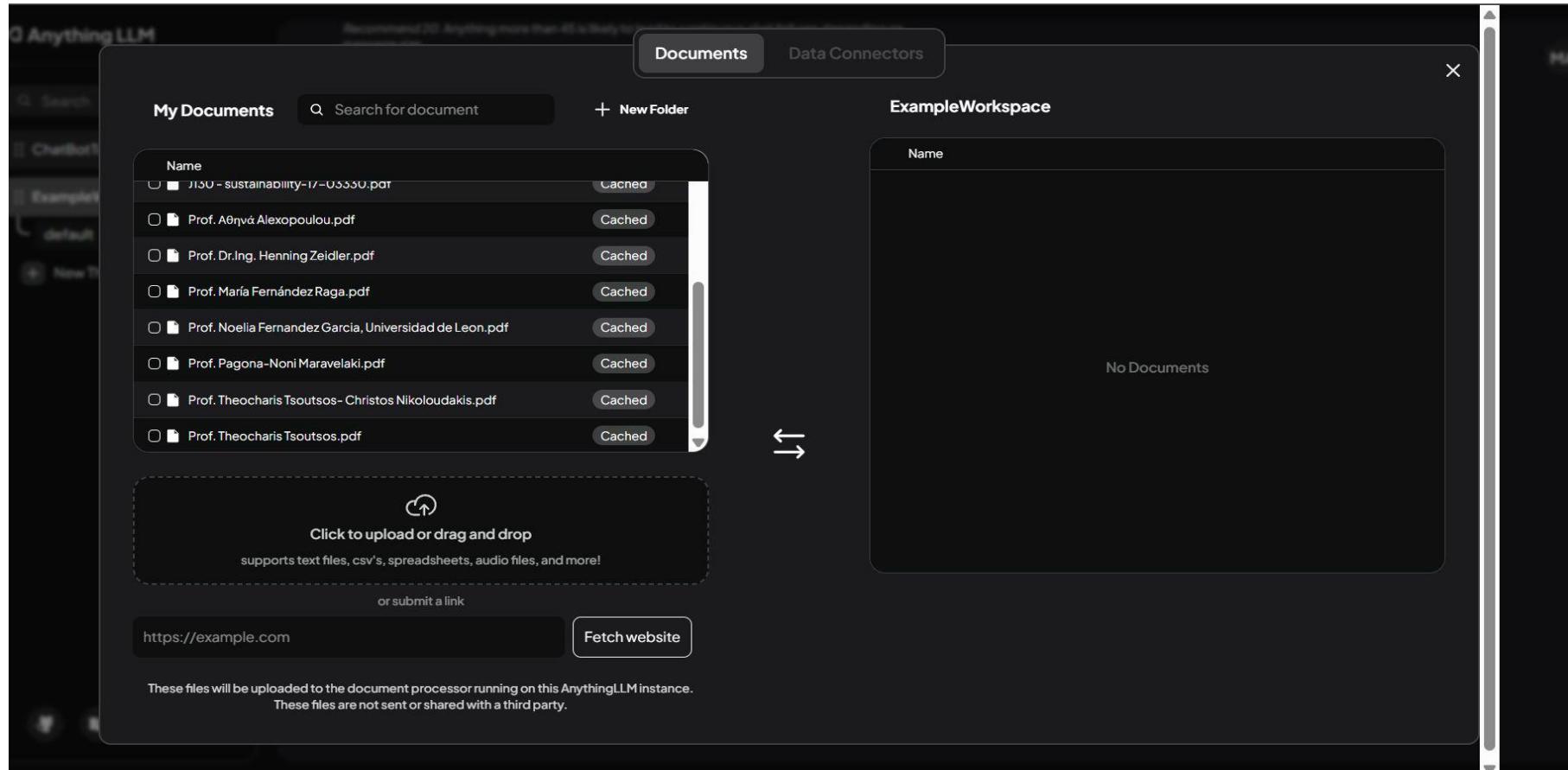
LLM Temperature

This setting controls how "creative" your LLM responses will be. The higher the number the more creative. For some models this can lead to incoherent responses when set too high.

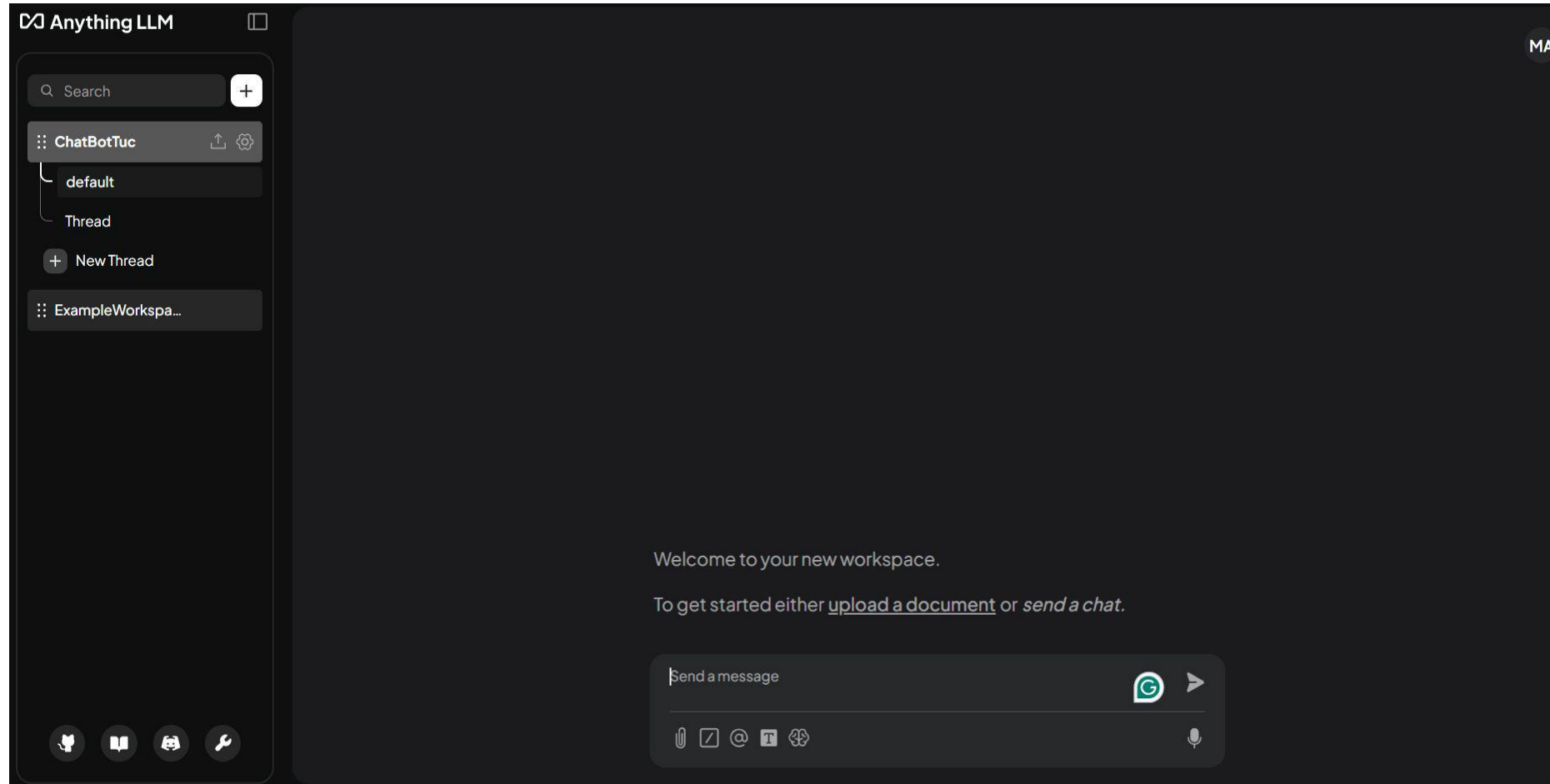
Most LLMs have various acceptable ranges of valid values. Consult your LLM provider for that information.

0.7

Basic steps – AnythingLLM (*Upload files*)



Basic steps – AnythingLLM (*Ready to use*)



Live Demo

- <https://anythingllm-twcc.onrender.com/workspace/hi-eurica-pro>

Thank you for your
attention!

Questions?