Informatique et systèmes
de communication ISC
Haute Ecole d'Ingénierie
Hochschule für Ingenieurwissenschaften

Rapport D&D Assistant

Your All-in-One resource for D&D Content

Author: Marco Caporizzi
Date:05.01.2025
Supervisor: Francesco Carrino

Abstract

The D&D Assistant aims to revolutionize the way newcomers and seasoned players navigate the complexities of Dungeons & Dragons. Set in the context of an engaging yet often intricate tabletop role-playing game, this assistant addresses the challenges of mastering game mechanics, rules, and lore, especially for those new to the experience.

The assistant serves as a locally hosted, AI-powered guide capable of answering a wide range of questions about D&D. Leveraging resources such as the System Reference Document (SRD) [1] and trusted online platforms, able to work with multiple file type (PDF, XLS and WEB-BASED) it ensures accurate, context-aware, and efficient responses. Designed to enhance the experience for both Dungeon Masters and adventurers, the assistant eliminates guesswork and enables seamless gameplay.

The project aims to deliver instantaneous, flawless assistance in multiple languages, catering to diverse player needs. Advanced features such as memory integration, voice-to-text interaction, and sentiment analysis promise an intuitive and interactive user experience. Future developments may include full voice-to-voice communication, positioning the assistant as a central character in the game's narrative.

By combining cutting-edge AI technologies with a deep understanding of D&D mechanics, the D&D Assistant aspires to become the ultimate tool for enhancing immersion, reducing mistakes, and elevating gameplays for all players.

1 Goals

The envisioned assistant aims to enhance gameplay with several goals.

Factual Assistance ensures the assistant can guide players during interrogations by delivering precise and referenced answers.

Multi-language Assistance enables seamless communication, responding in the language of the query for global accessibility and vocal answers. Complementing these, the goals of Routing and Self-Reflection focus on optimizing the assistant's decision-making and adaptability, ensuring it directs queries efficiently and continuously evaluates its performance to refine its contributions to the game experience. Together, these goals create a transformative, interactive journey for players.

2 Conception

2.1 Tools

Ollama [5] is an AI-powered platform designed to make large language models (LLMs) more accessible and efficient for users and developers. It focuses on simplifying interactions with LLMs by providing easy-to-use tools for model deployment, management, and integration into applications. Ollama's infrastructure emphasizes performance and scalability, catering to both individuals and enterprises that require natural language understanding and generation capabilities for a variety of tasks such as customer support, document analysis, and creative writing. As seen in figure 1.

Figure 1 - Ollama

LangChain [6] is a powerful framework designed to facilitate the development of applications that leverage large language models (LLMs). It provides developers with tools to build robust, multi-functional workflows by connecting LLMs to external data sources, APIs, and custom logic. LangChain is particularly well-suited for tasks involving natural language understanding, generation, and reasoning. As seen in figure 2.

🦜🔗 **LangChain**

Figure 2 - LangChain

Chroma [7] is an open-source, high-performance vector database designed to store, manage, and query embeddings (numerical representations of data). It is tailored for AI and machine learning applications, particularly those involving natural language processing, computer vision, or any task requiring similarity search and nearest neighbor retrieval. As seen in figure 3.

Chroma

Figure 3 - Chroma

2.2 Execution

To achieve the project's goals, the approach was to start with a simple foundation and incrementally add features to the working agent.

The first step was selecting the right tools. The focus was on ensuring the solution was free to use, locally operable as much as possible, and consistently reliable. For the model, Ollama 3.2 was chosen due to its lightweight nature (just 2.0GB) and local installation capability, meeting the criteria for simplicity and accessibility.

To enable multilingual support, preference was given to models trained in French and German. For this, bge-m3, a multilingual embedding model with versatile functionality, was selected.

For the Retrieval-Augmented Generation (RAG) system, Chroma DB was chosen because it is free, easy to set up, and effective for managing vectorized data.

The first step of implementation was document retrieval: the focus was on pdf, csv and web-based retrieval.

The second step of implementation was conversational memory: the ability to remember past conversation with the user.

The third step of implementation was sourcing the retrieve info: To enhance transparency, metadata was added to responses, ensuring users could trace the source of the information provided.

The fourth step of the implementation was the multilingual embedding: choosing an embedding which was trained on French and German.

The fifth step of the implementation was the self-reflection: A self-evaluation step was added, where responses were wrapped in a prompt designed to assess and refine their quality before being presented to the user.

The sixth step of the implementation was the routing: A routing mechanism was implemented to evaluate user queries and direct them through different processing workflows based on their intent and content.

Finally, vocal response capability was added to provide spoken answers, enhancing user accessibility and interaction.

2.3 Diagram of the concept

Here we have all the tasks combined to achieve the goal as seen in figure 4.

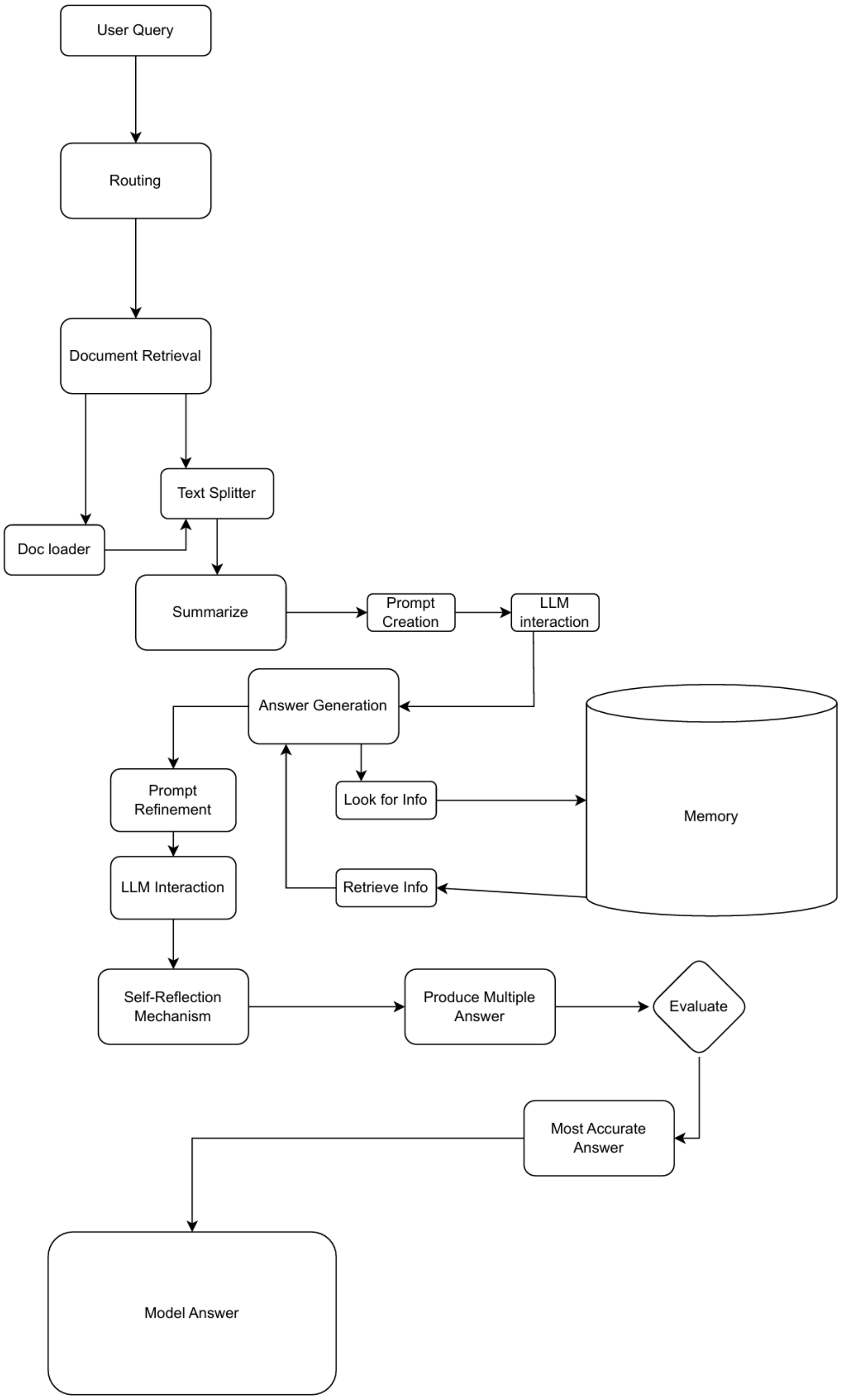


Figure 4 - Diagram

3 Analyze

This section addresses the technical results and management outcomes.

3.1 Results

To gather results and prove that my project works I decided to ask the same 3 questions in german, French and English.

Here is my English query:

"Tell me about racial traits in general"

Here is the generated answer:

“The racial traits in Dungeons & Dragons are unique abilities and characteristics that define a specific race. In general, racial traits can include: * Ability score increases (e.g., +2 Constitution for dwarves) * Age and lifespan * Alignment tendencies (e.g., lawful for dwarves) * Size and speed * Proficiency with certain tools or skills (e.g., battleaxe for dwarves) * Language abilities Each race has its own set of racial traits, which can influence gameplay and character development. The provided context from the Dungeons & Dragons player's handbook shows that dwarves have several unique traits, including Darkvision, Dwarven Resilience, Stonecunning, and proficiency with battleaxe and artisan's tools. It is not possible to provide a comprehensive overview of all racial traits in a single response, as there are many different races in the game. However, this context provides a good starting point for understanding the racial traits of dwarves.

[Sources: docs_en/races_eng_dem.pdf (page 2), docs_en/races_eng_dem.pdf (page 0), docs_en/races_eng_dem.pdf (page 1), docs_en/races_eng_dem.pdf (page 3)]’’

Here is my french query :

’’ Parle moi des traits raciaux en général ’’

Here is my generated answer:

’’ Bonjour ! En tant qu'expert de Donjons et Dragons, je vais analyser le contexte pour vous fournir des informations détaillées sur les traits raciaux en général.

Selon la documentation fournie, les traits raciaux sont des caractéristiques communes aux membres d'une race qui peuvent influencer le comportement, l'apparence et les capacités de votre personnage. Les traits raciaux peuvent inclure l'âge, l'alignement, l'augmentation de caractéristique, la catégorie de taille, la vitesse, les langues et les variantes raciales.

Il est important de noter que chaque race a ses propres traits raciaux uniques qui peuvent varier en fonction du peuple et du monde dans lequel ils vivent. Dans le cas des elfes, par exemple, leur valeur de Dextérité augmente de 2, ce qui suggère une certaine agilité et rapidité.

Il est également important de noter que les traits raciaux ne sont pas fixés et peuvent varier en fonction de votre personnage. Vous pouvez choisir d'augmenter certaines caractéristiques ou de modifier certains traits pour créer un personnage unique.

En résumé, les traits raciaux sont une partie importante des règles du jeu Donjons et Dragons et peuvent influencer le comportement et les capacités de votre personnage. Il est essentiel de comprendre ces traits pour créer un personnage cohérent et amusant.

Je ne peux pas fournir plus d'informations sur les traits raciaux généraux, car la documentation fournie n'en contient que quelques informations. Cependant, je vous conseille de consulter le livre de règles ou des ressources en ligne pour obtenir une compréhension plus approfondie des traits raciaux et de comment les utiliser dans votre jeu.

Je vous souhaite bonne chance avec votre personnage elfe ! [Sources: docs_fr/races_fr_demo.pdf (page 0), docs_fr/races_fr_demo.pdf (page 0)]''

Here is my german query:

"Sag mir etwas über die Volksmerkmale im Allgemeinen."

Here is my generated answer:

Basierend auf den bereitgestellten Informationen über die Volksmerkmale im Allgemeinen kann ich folgende Schlussfolgerungen ziehen: Die meisten Völker haben eine mittelgroße Größe, wobei einige kleinere Unterarten existieren. Die Bewegungsrate bestimmt die maximale Strecke, die ein Charakter auf Reisen und im Kampf zurücklegen kann. Die Gesinnung ist auch ein wichtiger Aspekt, der nicht bindend ist, aber den Spielercharakter besser ausarbeiten kann. Einige Völker haben eine bestimmte Sprache, die von ihren Angehörigen gesprochen wird. Es gibt jedoch keine expliziten Informationen über Volksmerkmale im Allgemeinen, sondern nur spezifische Merkmale für das elfische Volk und einige allgemeine Informationen zu Attributswerterhöhung, Alter und Größe. [Sources: docs_de/races_ger_demo.pdf (page 0), docs_de/races_ger_demo.pdf (page 0)]

3.2 Key Findings

This part elaborates on some good practices for effective answer generations.

3.2.1 Prompting

One of the critical aspects of generating effective answers is the quality of the prompt. Initially, I used a basic and general prompt structured as follows:

"Use the following pieces of retrieved context to answer the question. If you don't know the answer, just say that you don't know. Use three sentences maximum and keep the answer concise."

While functional, this approach often lacked depth and engagement. By refining the prompt to include flattery and clarity, I was able to achieve significantly better results. For instance, consider the following improved version:

"Analyze the context thoroughly to answer the question. If the answer is not found, acknowledge it. As an expert in Dungeons & Dragons, provide detailed insights when possible, using up to 10 sentences maximum."

This revised prompt encourages a more thoughtful response by emphasizing thorough analysis and providing clear guidelines for detailed answers. The inclusion of phrases such as "analyze thoroughly" and "as an expert" also sets a higher standard for the generated content, resulting in more accurate and nuanced answers.

3.2.2 Chunks

The size of the chunks used for processing has a significant influence on the quality of the output. Chunking affects the model's ability to understand, process, and generate coherent responses based on the provided input. Maintaining logical continuity is crucial to ensure that data from different categories or classes are not conflated or misinterpreted.

To preserve context and enhance the model's comprehension, I opted for larger chunks when creating embeddings. Larger chunks allow for more complete information to be captured, reducing the risk of losing critical context that could otherwise lead to fragmented or inaccurate responses. This approach aligns with best practices and recommendations for embedding-based models, ensuring both clarity and accuracy in the output.

3.2.3 Self-Reflection

Self reflection was implemented to better the accuracy of the response.

Allowing a check between the first given answer and the context.

For example, when working with larger data

Les nains sont une race de créatures présentées dans le contexte de Donjons et Dragons. Ils sont décrits comme des races robustes et courageuses, avec des traits physiques distinctifs tels que leur taille, leur vitesse et leur robustesse.

Lorsqu'ils sont nains des collines, ils bénéficient d'une augmentation de caractéristique de Sagesse, ainsi qu'un maximum de points de vie supplémentaire qui augmente de 1 avec chaque niveau gagné. Ils sont également dotés de sens aiguisés et d'une grand sagacité.

Les nains des collines sont également dotés d'une ténacité naine, ce qui signifie qu'ils peuvent augmenter leur maximum de points de vie de 1 à chaque fois qu'ils gagnent un niveau. Cela les rend plus résistants aux attaques et aux dommages.

Il convient de noter que la source fournie est un document en PDF intitulé "SRD CC v5.1 FR-2-8.pdf" qui semble être une version française de la 5e édition de la règle du jeu Donjons et Dragons, mais sans information sur l'auteur ou la date d'édition spécifiques. [Sources: docs_fr_self/SRD_CC_v5.1_FR-2-8.pdf (page 2), docs_fr_self/SRD_CC_v5.1_FR-2-8.pdf (page 1), docs_fr_self/SRD_CC_v5.1_FR-2-8.pdf (page 5), docs_fr_self/SRD_CC_v5.1_FR-2-8.pdf (page 0)]

Here is the reflected answer:

Hmmm... La réponse fournie semble avoir quelques inexactitudes et lacunes, notamment en ce qui concerne les nains.

D'après le texte fourni, il ressort que les nains sont considérés comme des créatures de taille P, avec une robustesse remarquable et une augmentation de caractéristique dans leur valeur de Sagesse de 1 point. Cependant, la réponse ne précise pas si les nains ont d'autres traits spécifiques ou si elles partagent certains traits avec d'autres races, comme le halfelin.

Il serait donc utile de compléter la réponse en précisant les caractéristiques unique à chaque race et en fournissant des informations plus détaillées sur les nains. [Sources: docs_fr_self/SRD_CC_v5.1_FR-2-8.pdf (page 2), docs_fr_self/SRD_CC_v5.1_FR-2-8.pdf (page 1), docs_fr_self/SRD_CC_v5.1_FR-2-8.pdf (page 5), docs_fr_self/SRD_CC_v5.1_FR-2-8.pdf (page 0)]

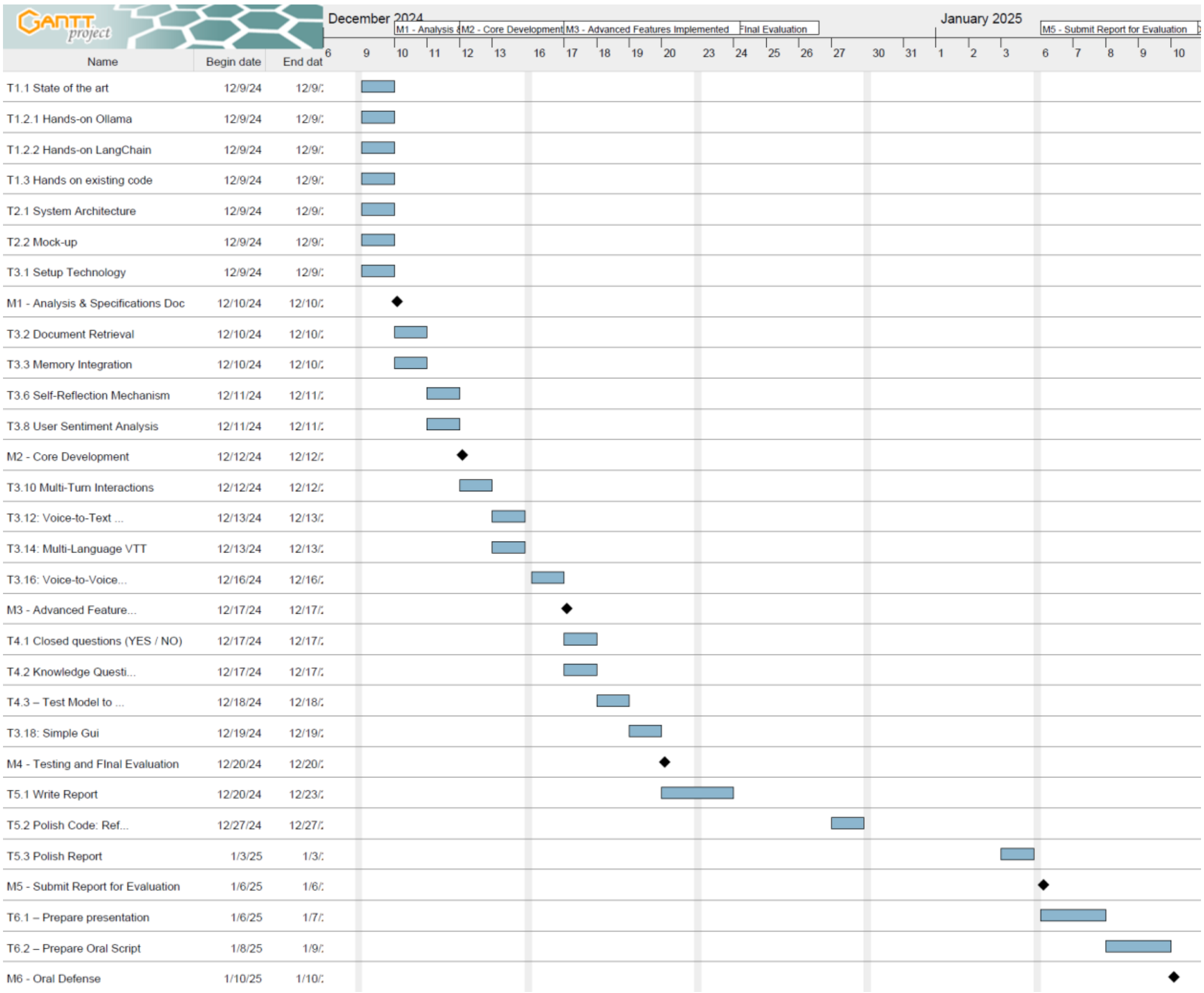
3.2.4 Routing

Routing can be implemented through various strategies, tailored to the needs of the system. For a multilingual agent, embedding plays a pivotal role in ensuring accurate and contextually relevant responses.

In this case, routing was implemented by first detecting the language of the user's query. Based on the detected language, the system dynamically adjusted the embedding process by modifying the prompts and selecting the appropriate documents for generating answers. This approach ensures that the agent delivers precise and culturally appropriate responses while maintaining multilingual capability.

3.3 Management

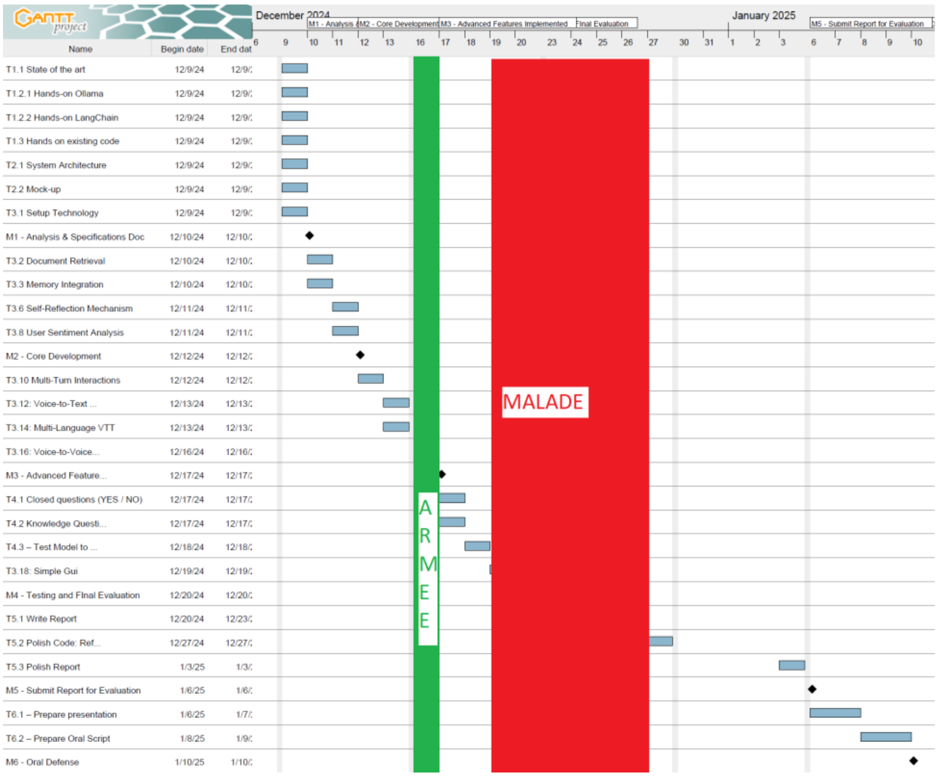
This figure 5 was what was planned.



Name	Begin date	End dat
T1.1 State of the art	12/9/24	12/9/
T1.2.1 Hands-on Ollama	12/9/24	12/9/
T1.2.2 Hands-on LangChain	12/9/24	12/9/
T1.3 Hands on existing code	12/9/24	12/9/
T2.1 System Architecture	12/9/24	12/9/
T2.2 Mock-up	12/9/24	12/9/
T3.1 Setup Technology	12/9/24	12/9/
M1 - Analysis & Specifications Doc	12/10/24	12/10/
T3.2 Document Retrieval	12/10/24	12/10/
T3.3 Memory Integration	12/10/24	12/10/
T3.6 Self-Reflection Mechanism	12/11/24	12/11/
T3.8 User Sentiment Analysis	12/11/24	12/11/
M2 - Core Development	12/12/24	12/12/
T3.10 Multi-Turn Interactions	12/12/24	12/12/
T3.12: Voice-to-Text ...	12/13/24	12/13/
T3.14: Multi-Language VTT	12/13/24	12/13/
T3.16: Voice-to-Voice...	12/16/24	12/16/
M3 - Advanced Feature...	12/17/24	12/17/
T4.1 Closed questions (YES / NO)	12/17/24	12/17/
T4.2 Knowledge Questi...	12/17/24	12/17/
T4.3 – Test Model to ...	12/18/24	12/18/
T3.18: Simple Gui	12/19/24	12/19/
M4 - Testing and FInal Evaluation	12/20/24	12/20/
T5.1 Write Report	12/20/24	12/23/
T5.2 Polish Code: Ref...	12/27/24	12/27/
T5.3 Polish Report	1/3/25	1/3/
M5 - Submit Report for Evaluation	1/6/25	1/6/
T6.1 – Prepare presentation	1/6/25	1/7/
T6.2 – Prepare Oral Script	1/8/25	1/9/
M6 - Oral Defense	1/10/25	1/10/

Figure 5 - Gantt Planned

This figure 6 was what actually happened.



Name	Begin date	End dat
T1.1 State of the art	12/9/24	12/9/:
T1.2.1 Hands-on Ollama	12/9/24	12/9/:
T1.2.2 Hands-on LangChain	12/9/24	12/9/:
T1.3 Hands on existing code	12/9/24	12/9/:
T2.1 System Architecture	12/9/24	12/9/:
T2.2 Mock-up	12/9/24	12/9/:
T3.1 Setup Technology	12/9/24	12/9/:
M1 - Analysis & Specifications Doc	12/10/24	12/10/:
T3.2 Document Retrieval	12/10/24	12/10/:
T3.3 Memory Integration	12/10/24	12/10/:
T3.6 Self-Reflection Mechanism	12/11/24	12/11/:
T3.8 User Sentiment Analysis	12/11/24	12/11/:
M2 - Core Development	12/12/24	12/12/:
T3.10 Multi-Turn Interactions	12/12/24	12/12/:
T3.12: Voice-to-Text ...	12/13/24	12/13/:
T3.14: Multi-Language VTT	12/13/24	12/13/:
T3.16: Voice-to-Voice...	12/16/24	12/16/:
M3 - Advanced Feature...	12/17/24	12/17/:
T4.1 Closed questions (YES / NO)	12/17/24	12/17/:
T4.2 Knowledge Questi...	12/17/24	12/17/:
T4.3 – Test Model to ...	12/18/24	12/18/:
T3.18: Simple Gui	12/19/24	12/19/:
M4 - Testing and FInal Evaluation	12/20/24	12/20/:
T5.1 Write Report	12/20/24	12/23/:
T5.2 Polish Code: Ref...	12/27/24	12/27/:
T5.3 Polish Report	1/3/25	1/3/:
M5 - Submit Report for Evaluation	1/6/25	1/6/:
T6.1 – Prepare presentation	1/6/25	1/7/:
T6.2 – Prepare Oral Script	1/8/25	1/9/:
M6 - Oral Defense	1/10/25	1/10/:

Figure 6 - Actual Gantt

This change of planification resulted in trade-off: The focus went into multilingual and some advanced feature implementation. Multi-turn interaction and other voice feature were discarded.

4 Conclusion

The final product is able to answer, and source based on document in 3 languages.

4.1 Limitations

One limitation encountered during this project was the rapidly evolving LangChain ecosystem. The differences between versions often introduced breaking changes or inconsistencies, making implementation and development less straightforward. This required frequent updates and adjustments to align with the latest version's functionality and compatibility.

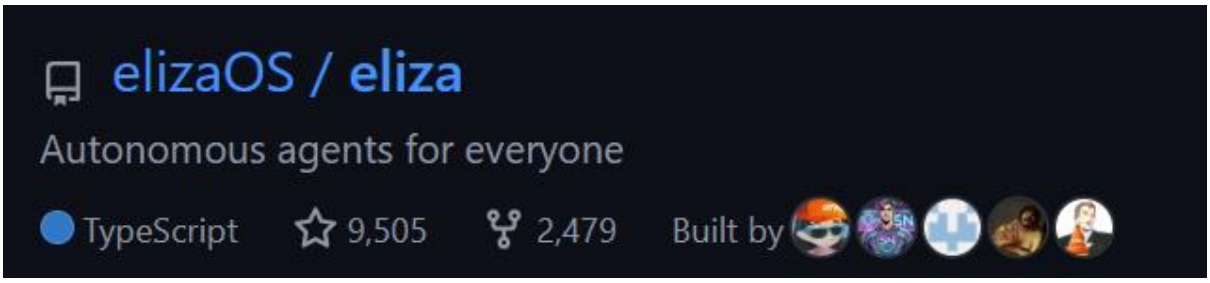
4.2 Improvement

This project can be improved in different ways.

The latest Ollama model, released in mid-December, offers significant performance improvements over version 3.2, including faster response times and enhanced functionality. Upgrading to this version would bring immediate technical benefits.

NVIDIA’s recent release of ChatRTX enables Retrieval-Augmented Generation (RAG) workflows with RTX acceleration. This allows for faster responses, robust multilingual support, and handling multiple files simultaneously. Additionally, it includes voice assistance and operates locally, ensuring improved privacy—a valuable feature for sensitive applications.

The popularity of agent-based frameworks surged in December, driven by the Agent x Crypto trend. This led to the emergence of highly versatile agent solutions. Notably, the most-starred GitHub repository during this period showcased an easy-to-deploy agent framework, indicating significant community interest and adoption. Leveraging these advancements could simplify agent integration and expand the project's capabilities.



elizaOS / eliza

Autonomous agents for everyone

● TypeScript ☆ 9,505 2,479 Built by

Figure 7 - Github Top Month

5 References

[1] https://media.wizards.com/2023/downloads/dnd/SRD_CC_v5.1.pdf

[2] https://media.wizards.com/2023/downloads/dnd/SRD_CC_v5.1_FR.pdf

[3] https://www.dndbeyond.com/

[4] https://www.aidedd.org/

[5] https://ollama.com/

[6] https://www.langchain.com/

[7] https://www.trychroma.com/

6 Figure

Figure 1 - Ollama

Figure 2 – LangChain

Figure 3 – Chroma Database

Figure 4 - Diagram

Figure 5 – Gantt Planned

Figure 6 – Actual Gantt

Figure 7 – Github Top Month