MyVision: A Retrieval-Augmented Generation System to Provide High School Students with Accurate Academic Guidance

Samuele Mazzei[1], Lorenzo Zambotto[1], Gabriele Tealdo[1] and Alberto Macagno[1]

[1]University of Trento, Italy

Abstract

This paper introduces the development and evaluation of a Retrieval-Augmented Generation (RAG) system designed to assist prospective students in navigating university options. The system provides accurate academic guidance by retrieving and synthesizing information on undergraduate and single-cycle master's degree programs, as well as library resources, from the University of Trento and the University of Verona. The RAG pipeline utilizes a streamlined toolchain, incorporating a Markdown parser for efficient data handling and the Llama3-8b-8192 Large Language Model (LLM) for query processing. The system's performance was assessed through both automated evaluation, using the Llama3-70b LLM as a reference, and blinded human evaluation. The results demonstrate the system's potential for providing relevant and accurate information to students. The evaluation also highlighted areas for further development, including enhanced retrieval mechanisms and expanded LLM testing. Future work aims to broaden the system's scope to include more degree levels and universities, ultimately creating a comprehensive platform to support students in their academic decision-making journey.

Keywords

Retrieval-Augmented Generation (RAG), Natural Language Processing (NLP), Large Language Models (LLMs), Dataset Creation, Academic Guidance

1. Introduction

Choosing a university path is one of the most complex and significant decisions for students nearing the end of high school. This, combined with the overwhelming amount of new information encountered when browsing various and often inconsistent university websites, creates confusion and a sense of being lost, leading to wasted time and uncertainty. These challenges stem from both the dispersion of available information and the lack of intuitive tools to guide students through the decision-making process. We first addressed these problems during the Qualitative Research course, part of the Master's degree in Human-Computer Interaction at the University of Trento (Unitn), laying the foundation for the development of myVision. This platform aims to integrate an AI-powered chatbot that provides relevant information about partner universities and online counseling services within a single interface. As part of the course, our group project offered hands-on experience with qualitative research tools and techniques to address the following question: "What role could a social robot play in effectively assisting university students in discovering and engaging with social events within the University of Trento community?" A survey conducted within the Department of Psychology and Cognitive Science (DIPSCO) was instrumental in identifying a significant need among students for improved online educational guidance. This need arises primarily from the fact that online resources, particularly university websites, are often counterintuitive and fragmented. As a result, students frequently experience frustration and discomfort while searching for essential information to make informed decisions about their academic paths. Similar difficulties are evident in the area of counseling services, where students struggle to locate and access appropriate resources, which are often scattered. In another course we took, Social Interaction, we further explored this topic by analyzing potential user interactions with such an interface. This course guided us in examining the scientific literature, which reinforced our findings and provided a solid research foundation. During the course, we identified a study closely aligned with our own, titled A Guidance and Counseling Mobile Application (GC Mobile App) for Educational Institutions [1]. In this study, the researchers evaluated an educational app called GC Mobile and concluded that it enhanced the counseling process by leveraging technology to provide a scalable, accessible, and confidential platform for student guidance. As the authors noted, "The GC Mobile App allows students to see a counselor anytime and from any location without having to visit them in the office." This highlights the potential of similar platforms, such as myVision, to improve the accessibility and effectiveness of counseling

samuele.mazzei@studenti.unitn.it (S. Mazzei);
lorenzo.zambotto@studenti.unitn.it (L. Zambotto);
gabriele.tealdo@studenti.unitn.it (G. Tealdo);
alberto.macagno@studenti.unitn.it (A. Macagno)

CEUR Workshop Proceedings (CEUR-WS.org)

services. In this paper, we aim to lay the foundation for the development of our chatbot by focusing on the academic offerings of two Italian universities: the University of Trento (Unitn) and the University of Verona (Univr). These institutions were selected due to their geographical proximity and the presence of interdisciplinary and interuniversity courses, which offer significant opportunities for prospective students interested in studying in these areas. The system was developed with a specific focus on post-diploma university orientation, considering only bachelor's degree programs and single-cycle master's degrees. This approach addresses the needs of recent high school graduates by providing an innovative tool to explore available academic options in a simple and immediate way. Furthermore, we included information about the universities' libraries to provide new students with access to a valuable resource that can support their studies.

2. Background literature/Related works

For our project to be successful, we need it to be as accurate as possible. For this reason, we chose to create our chatbot using Retrieval-Augmented Generation (RAG). By doing so, we address the limitations inherent in traditional methods and standalone Large Language Models (LLMs) [2], such as limited context and possible hallucinations. In the literature, we found several studies that support our choice. Dieing et al. (2024) describe a system for study program orientation that provides personalized recommendations using a Mixtral LLM paired with a RoBERTa embedding model. Their RAG approach retrieves data from a government website and achieves an average response accuracy above 0.75 [3]. Saha and Saha (2024) report that a GPT-3.5–based chatbot enhances support for international graduate students by combining generative capabilities with precise retrieval from social media sources [4]. Dakshit (2024) explored the use of RAG in higher education, focusing on applications as virtual teaching assistants and teaching aids. Faculty perspectives gathered in the study highlighted the benefits of RAG in supporting teaching processes, such as the generation of study guides, quizzes, and assignment questions, while also assisting students by providing precise answers to academic queries. Faculty members emphasized the importance of integrating broader data sources and advanced functionalities, including the ability to process mathematical content and image-based inputs, to improve the system's effectiveness [5]. The potential of RAG-powered systems lies in their ability to provide accurate, contextually relevant, and personalized support by combining retrieval mechanisms with generation capabilities [2]. A retrieval component first searches for relevant information from a curated set of academic resources, ensuring the content is accurate and domain-specific. The generation component then synthesizes this information to produce coherent and contextually appropriate responses [6]. This dual approach not only improves the reliability of responses but also enables the system to adapt to individual learning styles and paces, making it a valuable tool for personalized education. These findings align with the goals of myVision, particularly in creating a chatbot that integrates multiple functions—academic guidance, counseling services, and information retrieval—into a cohesive platform. Drawing from the studies mentioned above, we plan to leverage RAG's strengths to ensure that myVision not only meets students' informational needs but also provides reliable, context-aware responses to enhance their educational journey.

3. Dataset

To collect the documents for our task, we accessed the course websites of Unitn[1] and Univr[2] to gather the necessary data. Since the main objective of this project is to provide orientation for high school students, we selected undergraduate degrees and single cycle master's degrees. For Unitn, we obtained data from the "Prospective Student" section, which is divided into three parts: "Course Programme," providing an overview of the degree; "Course Content," listing all courses offered over the years along with their respective ECTS credits, and in some cases, detailed course descriptions; and "Application," which contains enrollment information. For Univr, we collected similar information. After selecting a degree, we retrieved the "Overview" section under the "Find out more" option, the study plan from the "Modules" section, and enrollment details from the "How to apply" option. All collected data of the courses was converted into Markdown format with the help of an extension of ChatGPT-4 called Markdown converter[3]. ChatGPT-4 does not always structure the data in the same way, so we manually adjusted the formatting when discrepancies were too large. We also collected data on the libraries of both universities. In this case, the data were gathered manually to ensure a consistent file structure and order. The collected library data included: a general overview, with information on access, location, staff, and available spaces; the services offered by the libraries; and the opening hours.

We used Markdown language for several reasons, including efficiency and flexibility. This format allows for a clear structuring of data through the use of headings, enabling the RAG to subsequently divide the information

[1] https://www.unitn.it/en/ateneo/1819/programmes-of-study
[2] https://www.univr.it/en/degree-programmes
[3] https://chatgpt.com/g/g-lnlmekbGd-markdown-converter

into well-defined and interconnected sections. This optimization facilitates the retrieval process, making it easier to identify and associate relevant information. Another advantage of Markdown is its ability to include tables, which are clearer and more understandable as responses for users. Moreover, compared to more complex formats such as JSON, Markdown simplifies the separation of data into meaningful nodes, ensuring that information is correctly linked to specific courses. Thanks to the titles of the Markdown files, metadata such as the university and course name could also be added, making the system more organized and functional. For this reason, each degree program was saved separately using the format "uniname-degree-name.md" for courses and "uniname-library-name.md" for libraries. Finally, the Markdown format is more practical during the dataset creation phase, as it allows for the use of tools like scrapers to quickly extract text from web pages. This process simplifies and accelerates the assembly of necessary information while ensuring greater consistency and quality of the data. In total, we collected data for 29 degrees from Unitn and 41 degrees from Univr, resulting in 70 course documents. Additionally, we collected data from 5 libraries from Unitn and 34 libraries from Univr, resulting in 39 library documents. This yielded a total of 109 documents.

Additionally, all data were translated into English when the English version of the site did not contain sufficient information compared to its Italian counterpart, as the answers provided by our RAG system were more accurate due to the embedding model introduced during the course. The English version of the embedding model is trained and tested on more data and has access to a larger corpus than the Italian version, which typically results in better training, improved generalization, and richer language representations [7]. To verify this, we consulted the literature and found a paper titled "Retrieval-augmented generation in multilingual settings" [8], which confirms our hypothesis.

4. Experiments

The objective of this study was to develop and evaluate a document retrieval system designed to query information from university course descriptions and library details. The system's performance was assessed based on its accuracy in retrieving relevant and contextually appropriate information. For this purpose, we utilized Groq[4] as the provider for Large Language Models (LLMs). Specifically, two models were employed: Llama3-8b-8192 (8 billion parameters) served as the primary LLM for query processing, while Llama3-70b (70 billion parameters) functioned as the reference ("golden") model during evaluation.

[4]https://groq.com/

4.1. RAG Pipeline

The experimental workflow started with a corpus of structured Markdown documents, detailing university courses and library information (as described in Section 3). These documents were processed using MarkdownNodeParser, a tool that segmented the documents into nodes while preserving the inherent hierarchical structure and relationships. Subsequently, these nodes were converted into vector embeddings using the BAAI/bge-m3 model[5], which generates semantic representations of the text. GPU acceleration (NVIDIA T4 instance via Google Colab[6]) was employed for efficient embedding generation and dataset indexing. To reduce the computational cost associated with repeated generation of embeddings and indexes, a caching system was implemented. This system cached generated embeddings and indexes, stored on Google Drive, thereby preventing redundant computations and reducing operational overhead.

The initial system design involved storing the BAAI/bge-m3 embeddings in ChromaDB[7], a vector database, and implementing a hybrid retrieval mechanism utilizing both sparse (keyword-based) and dense (semantic) indexing methods for enhanced retrieval combining lexical precision and semantic relevance. However, technical challenges prevented the successful implementation of this hybrid approach. Consequently, the retrieval process relied solely on the BM25 algorithm[8], a standard keyword-based method employing lexical matching, to retrieve relevant document sections, in our case the top 15-k nodes, based on user queries.

A graphical representation of the whole pipeline is shown in Figure 1.

4.2. Evaluation

Evaluation of the system's performance employed a dual approach: automated assessment using the Llama3-70b model and blinded human evaluation, ensuring objectivity. Both methods assessed the quality of the generated answers and, for the automated part, the suitability of the retrieved context.

For the automated evaluation of generated answers, the Llama3-70b model assessed relevance and correctness relative to the user query. It assigned a score on a 1-to-5 scale, which was subsequently normalized to a 0-to-4 scale for direct comparison with human scores. The model also generated a textual justification explaining its assessment, highlighting aspects like completeness or accuracy. Due to API call limitations with standard eval-

[5]https://huggingface.co/BAAI/bge-m3
[6]https://colab.research.google.com/
[7]https://www.trychroma.com/
[8]https://docs.llamaindex.ai/en/stable/examples/retrievers/bm25_retriever/

uation frameworks, custom requests were implemented to facilitate this automated assessment process.

Automated context assessment focused on the text passages retrieved by the BM25 algorithm before answer generation. The Llama3-70b model evaluated the context based on two criteria: (1) the relevance of the retrieved context to the subject matter of the user's query, and (2) the degree to which the context contained sufficient information to fully answer the query. These assessments contributed to a final context alignment score presented on a 0-to-4 scale.

The prompts used by the Llama3-70b model were adapted from the correctness evaluation[9] and context relevancy evaluation[10] modules available within the LlamaIndex framework. These prompt templates are included as an attachment at the end of this paper for full transparency and reproducibility.

In parallel, human annotators independently evaluated the final generated answers. They assessed relevance and correctness on a 0-to-4 scale and provided qualitative notes detailing their reasoning, pointing out strengths or weaknesses such as omissions or inaccuracies.

This comprehensive evaluation process utilized a dataset of 71 question-answer pairs. These pairs were randomly selected from a larger pool generated across all 109 source documents (covering both university courses and libraries). Notably, 10 of these 71 pairs were specifically designed to query information contained within the library documents, ensuring assessment of the system's performance on that subset of data. Overall, the system demonstrated comparable performance across both evaluation methodologies. It achieved an average normalized accuracy score of 83.63% (SD = 16.45%) in the AI evaluation and 79.22% (SD = 28.34%) in the human evaluation. This similarity in overall scores suggests reasonably consistent performance, although individual query evaluations could differ between the AI and human assessors, underscoring the value of the dual approach. Notably, the context evaluation score was 76.36% (SD = 20.89%). Some random test pairs results are shown in Tables 1, 2 3, 4 and 5.

Detailed implementation procedures, including data processing scripts, model configurations, and complete evaluation results, are documented in the associated Jupyter notebook.

[9] https://github.com/run-llama/llama_index/blob/main/llama-index-core/llama_index/core/evaluation/correctness.py

[10] https://github.com/run-llama/llama_index/blob/main/llama-index-core/llama_index/core/evaluation/context_relevancy.py

5. Discussion

5.1. Advantages

A significant advantage of the implemented system lies in its rapid deployment capability, stemming from the simplified toolchain. The streamlined setup process enabled quick deployment, facilitating efficient testing and development cycles. This ease of use facilitated the integration of various components, reducing the learning curve and making the system accessible even for individuals with limited prior experience.

Another notable benefit was the availability of multiple components, particularly the Markdown parser, which proved invaluable. The parser effectively handled document processing, ensuring accurate interpretation and formatting of content. This feature enhanced the system's overall functionality, enabling seamless handling of structured documents and consequently improving the user experience.

Despite certain challenges, the system achieved relatively high accuracy in its responses. However, document retrieval remains an area for improvement, presenting an opportunity for optimization to further enhance precision and relevance. Nevertheless, the current results demonstrate promising potential, indicating that the fundamental approach is sound and can be further refined with additional efforts.

5.2. Limitations

A primary difficulty encountered was the extensive documentation, which contained a wealth of information requiring considerable time for comprehension and analysis. Understanding the optimal implementation and optimization strategies demanded significant effort due to the complexity of the available options, which necessitated careful evaluation.

Another challenge arose from the numerous potential "blocks," such as different retrievers and rerankers, that could be integrated into the workflow. The wide array of choices required extensive experimentation to determine the most effective combination, leading to increased development time and complexity.

The necessity of a GPU to support computationally demanding embedding models presented another hurdle. While Google Colab offered an accessible environment for initial development, it occasionally failed to provide adequate hardware resources for intensive tasks. This issue was eventually resolved by transitioning to a local PC equipped with a dedicated graphics card, which provided a more stable and powerful development environment.

A particularly limiting factor was the API rate-limiting imposed on the LLM provider. While high-level methods offered precise functionality, they required multiple

API calls per query, resulting in significant costs and increased response times. To mitigate this, a delay was implemented between successive API calls, which, although effective in managing costs, considerably slowed down the evaluation process. Furthermore, the inability to modify built-in API functions to define specific rate limits led to challenges such as unnecessary calls and system crashes.

5.3. Other Attempts

One of the most complex approaches attempted was the creation of agents capable of responding to specific questions for each document to enhance response accuracy. However, we ultimately discarded this idea due to the excessive response times, which rendered the approach impractical for real-time applications.

Another challenge was to implement a more comprehensive, state-of-the-art evaluation system, such as Ragas. While this approach showed theoretical promise, API limits prevented us to use more sophisticated evaluation systems.

In conclusion, while the project encountered several challenges, the overall results were promising, demonstrating the potential of the approach. Future efforts should focus on optimizing document retrieval, improving workflow efficiency, and addressing hardware and API limitations to further enhance the system's performance and usability.

6. Release

The source code of the RAG pipeline and the dataset used are available both on the project's Google Drive page[11] and Github repository[12].

7. Conclusions

In this paper, we presented the development of a Retrieval-Augmented Generation (RAG) system designed to provide students with accurate academic guidance, specifically focusing on university course and library information. The system leverages a streamlined toolchain, incorporating a Markdown parser for efficient data handling and the Llama3-8b-8192 LLM for query processing. While the system demonstrates promising results, there are areas for enhancement.

Future work will concentrate on several key improvements. Firstly, we aim to enhance the evaluation framework to provide a more comprehensive assessment of the RAG model's performance, incorporating metrics for contextual relevance, accuracy, and adaptability. Secondly, the integration of reranking mechanisms will be explored to prioritize retrieved results based on relevance and quality. Thirdly, to ensure robust and scalable performance, we plan to test the model with a wider range of LLMs, such as Gemini, Claude and others.

Beyond these technical refinements, the myVision service will be expanded to serve a broader audience, including bachelor's degree graduates and students interested in specialized master's programs, and to include more universities. We envision the chatbot as a core component of a larger platform that will offer a dedicated user interface, informative podcasts, and direct interaction with student advisors. Ultimately, this work lays the groundwork for a powerful tool to aid students in navigating their academic journeys.

[11] https://drive.google.com/drive/folders/1nzWP-TS8gUNAt6e4Dlaa85PAokw45aZ_
[12] https://github.com/Samu01Tech/myVision-universities-RAG

References

[1] K. Ukaoha, J. Ndunagu, F. Osang, et al., A guidance and counseling mobile application (gc mobile app) for educational institutions, NIPES-Journal of Science and Technology Research 2 (2020).

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.

[3] T. I. Dieing, M. Scheffler, L. Cohausz, Enhancing chatbot-assisted study program orientation, in: Proceedings of DELFI Workshops 2024, Gesellschaft für Informatik eV, 2024, pp. 10–18420.

[4] B. Saha, U. Saha, Enhancing international graduate student experience through ai-driven support systems: A llm and rag-based approach, in: 2024 International Conference on Data Science and Its Applications (ICoDSA), IEEE, 2024, pp. 300–304.

[5] S. Dakshit, Faculty perspectives on the potential of rag in computer science higher education, in: Proceedings of the 25th Annual Conference on Information Technology Education, 2024, pp. 19–24.

[6] H. Modran, I. C. Bogdan, D. Ursuțiu, C. Samoila, P. L. Modran, Llm intelligent agent tutoring in higher education courses using a rag approach, Preprints 2024 2024070519 (2024).

[7] M.-L. M.-F. Multi-Granularity, M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation (2024).

[8] N. Chirkova, D. Rau, H. Déjean, T. Formal, S. Clinchant, V. Nikoulina, Retrieval-augmented generation in multilingual settings, arXiv preprint arXiv:2407.01463 (2024).

A. Correctness Evaluation Prompt

You are an expert evaluation system for a question answering chatbot. You are given the following information:
- a user query, and
- a generated answer

You may also be given a reference answer to use for reference in your evaluation. Your job is to judge the relevance and correctness of the generated answer. Output a single score that represents a holistic evaluation. You must return your response in a line with only the score. Do not return answers in any other format. On a separate line provide your reasoning for the score as well.

Follow these guidelines for scoring:
- Your score has to be between 1 and 5, where 1 is the worst and 5 is the best.
- If the generated answer is not relevant to the user query, you should give a score of 1.
- If the generated answer is relevant but contains mistakes, you should give a score between 2 and 3.
- If the generated answer is relevant and fully correct, you should give a score between 4 and 5.

Example Response:
4.0
The generated answer has the exact same metrics as the reference answer, but it is not as concise.

B. Context Relevancy Evaluation Prompt

Your task is to evaluate if the retrieved context from the document sources are relevant to the query. The evaluation should be performed in a step-by-step manner by answering the following questions: 1. Does the retrieved context match the subject matter of the user's query? 2. Can the retrieved context be used exclusively to provide a full answer to the user's query? Each question above is worth 2 points, where partial marks are allowed and encouraged. Provide detailed feedback on the response according to the criteria questions previously mentioned. After your feedback provide a final result by strictly following this format: '[RESULT] followed by the float number representing the total score assigned to the response'
Query: \n {query_str}
Context: \n {context_str}
Feedback:

C. Tables

ID	Question
5	What types of materials are generally not available for loan at the University of Trento libraries?
10	What kind of materials are available at the University Language Centre Library at the University of Verona?
43	What are the admission requirements for the Environmental Engineering program, and what further study opportunities are available after graduation?
60	What are the two different orientations that students can choose within the Business Administration and Law degree program, and where can students undertake internships?
70	What are the integrated courses and their modules in the first year of the Medicine and Surgery program, and what are their associated credit values (CFU)?

Table 1
Sample Questions

ID	Gold Answer	RAG Answer
5	Generally, the following materials are not available for loan at the University of Trento libraries: journals; reference material, such as encyclopedias, dictionaries, bibliographies; course reserves; legal codes; works published before 1900; degree theses; microfilm and microfiches; material which requires binding or repairing; and any other work Libraries decide to make available for the local consultation only.	Materials that are generally not available for loan at the University of Trento libraries include: * Journals * Reference material, such as encyclopedias, dictionaries, bibliographies * Course reserves * Legal codes * Works published before 1900 * Degree theses * Microfilm and microfiches * Material which requires binding or repairing * Any other work Libraries decide to make available for the local consultation only.
10	The University Language Centre Library at the University of Verona has approximately 5000 volumes, 57 periodical titles, 550 DVDs, 800 CD-Roms / Audio CDs, and 750 VHS tapes. The materials cover various linguistic disciplines, including Italian, English, German, French, Spanish, Russian, Arabic, Chinese, and Modern Greek, with some DVDs including other European and non-European languages.	Materials available for loan (20 days) include: * All books in the library, except: + Encyclopedias, dictionaries, bibliographies (available for consultation on the 2nd floor) + Journals (available in the periodicals section, basement) + Special consultation materials (request via email: prestito.frinzi@ateneo.univr.it)
43	Admission requires an **Italian high school diploma or equivalent foreign qualification, and passing a TOLC test**. The program has limited enrolment and seats allocated to Italian, EU, and non-EU citizens residing in Italy. Graduates have direct access to the **Master's program in Environmental and Land Engineering** at the University of Trento, and can also access other Master's programs in engineering that meet admission requirements. Transfers from other degree programs or universities are also possible, with applications due by February 3, 2025.	Admission requirements for the Environmental Engineering program: No specific admission requirements mentioned. Further study opportunities after graduation: * Direct access to the Master's program in Environmental and Land Engineering at the University of Trento. * Access to other Master's programs in engineering disciplines that meet admission requirements. * Graduates can continue their education with a Master's Degree in Civil Engineering or related fields such as Energy Engineering or Environmental Meteorology. * Alternative options include enrolling in specialized programs at other universities to further enhance expertise in industrial engineering disciplines.
60	Students can choose between two orientations: **one towards private professions and the other towards public professions**. Those choosing the private professions orientation will complete their mandatory internship mainly within **professional firms or auditing companies**. Students choosing the public professions orientation can complete their internship in **public entities and companies**.	The two different orientations that students can choose within the Business Administration and Law degree program are: 1. Orientation towards private professions, mainly within professional firms or auditing companies. 2. Orientation towards public professions, it can take place in public entities and companies. Students can undertake internships in either of these orientations.
70	The first year of the program includes the following integrated courses and modules with their corresponding credit values: **Physics and Informatics: Medical Physics (6 CFU), Informatics for Medicine (4 CFU). Chemistry and Biochemistry: Chemistry and Biochemical Basics (6 CFU), Biochemistry (6 CFU). General and Molecular Biology: General Biology (6 CFU), Molecular Biology (6 CFU). Histology and Embryology: (9 CFU). Human Anatomy 1: (7 CFU). English Language: (3 CFU)**.	Here are the integrated courses and their modules in the first year of the Medicine and Surgery program, along with their associated credit values (CFU): 1. Physics and Informatics: * Medical Physics (6 CFU) * Informatics for Medicine (4 CFU) 2. Chemistry and Biochemistry: * Chemistry and Biochemical Basics (6 CFU) * Biochemistry (6 CFU) 3. General and Molecular Biology: * General Biology (6 CFU) * Molecular Biology (6 CFU) 4. Histology and Embryology: * (no modules listed, 9 CFU) 5. Human Anatomy 1: * (no modules listed, 7 CFU) 6. English Language: * (no modules listed, 3 CFU) Note that some modules may not have specific titles or descriptions, but they are still part of the integrated course.

Table 2
Gold Answer vs. RAG Answer

ID	AI Evaluation Score (1-5)
5	5.0 The generated answer is almost identical to the reference answer, with the only difference being the formatting of the list. The generated answer is fully correct and relevant to the user query, providing a comprehensive list of materials that are not available for loan at the University of Trento libraries.
10	3.0 The generated answer is somewhat relevant to the user query, as it provides information about the materials available at the University Language Centre Library, but it only focuses on the loanable materials and their conditions, whereas the reference answer provides a broader overview of the library's collection, including the number of volumes, periodical titles, and multimedia materials. The generated answer lacks the scope and detail of the reference answer, making it only partially correct.
43	3.5 The generated answer is relevant to the user query, as it addresses both the admission requirements and further study opportunities. However, it lacks specific details, such as the Italian high school diploma or equivalent foreign qualification, and the TOLC test, which are mentioned in the reference answer. Additionally, the generated answer provides some correct information, like direct access to the Master's program in Environmental and Land Engineering, but also includes some vague or incorrect information, such as "no specific admission requirements mentioned" and "alternative options include enrolling in specialized programs at other universities".
60	4.5 The generated answer is highly relevant and correct, as it accurately identifies the two orientations within the Business Administration and Law degree program and provides the correct information about where students can undertake internships. The generated answer is very similar to the reference answer, with only minor differences in wording and structure. The generated answer is clear and concise, making it easy to understand for the user.
70	4.5 The generated answer is highly relevant and correct, providing a clear and organized list of integrated courses and their modules in the first year of the Medicine and Surgery program, along with their associated credit values (CFU). The answer is well-structured and easy to follow, with minor differences in formatting compared to the reference answer. The only minor issue is the lack of module titles or descriptions for Histology and Embryology and Human Anatomy 1, but the generated answer provides a note explaining this. Overall, the generated answer is very close to the reference answer in terms of content and accuracy.

Table 3
AI Evaluation Scores

ID	AI Context Score (0-4)
5	1. Does the retrieved context match the subject matter of the user's query? The retrieved context is highly relevant to the user's query. It provides detailed information about the loan service offered by the University of Trento libraries, including the types of materials that are available for loan and those that are not. The context specifically mentions the materials that are not available for loan, which directly answers the user's query. I would give this a score of 2.0. 2. Can the retrieved context be used exclusively to provide a full answer to the user's query? The retrieved context provides a comprehensive answer to the user's query. It lists the specific types of materials that are not available for loan, including general reference works, journals, course reserves, and others. The context is detailed and exhaustive, providing a complete answer to the user's question. I would give this a score of 2.0. Final Result: [RESULT] 4.0
10	1. Does the retrieved context match the subject matter of the user's query? The retrieved context is related to the University of Verona's library services, which is relevant to the user's query about the materials available at the University Language Centre Library. However, the context provides a broad overview of the library's services, including document consultation, loan services, Leganto, Affluences app, document retrieval, interlibrary services, and location. While it touches on the topic of materials available for loan, it does not provide a direct answer to the user's query. Therefore, I would give 1.5 points out of 2. 2. Can the retrieved context be used exclusively to provide a full answer to the user's query? The retrieved context does not provide a comprehensive list of materials available at the University Language Centre Library. It mentions that all books in the library are available for loan, except for general reference works and books labeled "CONS," but it does not provide a detailed list of materials. Therefore, the context cannot be used exclusively to provide a full answer to the user's query. I would give 0.5 points out of 2. Final Result: [RESULT] 2.0
43	1. Does the retrieved context match the subject matter of the user's query? The retrieved context partially matches the subject matter of the user's query. The query asks about the admission requirements for the Environmental Engineering program and further study opportunities after graduation. The context provides information about various programs, including Environmental Engineering, but it does not directly address the admission requirements. However, it does provide information about further study opportunities, which is relevant to the query. Score: 1.5 2. Can the retrieved context be used exclusively to provide a full answer to the user's query? The retrieved context cannot be used exclusively to provide a full answer to the user's query. While it provides some relevant information about further study opportunities, it does not address the admission requirements for the Environmental Engineering program. Additional information would be needed to fully answer the query. Score: 1 Final Result: [RESULT] 2.5
60	1. Does the retrieved context match the subject matter of the user's query? The retrieved context is highly relevant to the subject matter of the user's query. The context provides detailed information about the Business Administration and Law degree program, including its objectives, curriculum, and career opportunities. The context specifically mentions the two orientations within the program, which aligns with the user's query. However, the context is quite lengthy and covers a wide range of topics, making it challenging to extract the exact information related to the query. Score: 1.8/2.0 2. Can the retrieved context be used exclusively to provide a full answer to the user's query? The retrieved context provides a comprehensive overview of the Business Administration and Law degree program, including the two orientations mentioned in the query. However, the context does not directly answer the question about where students can undertake internships. Although the context mentions internships and provides some information about the internship experiences, it does not explicitly state where students can undertake them. Score: 1.5/2.0 Final Result: [RESULT] 3.3/4.0
70	1. Does the retrieved context match the subject matter of the user's query? The retrieved context provides a detailed breakdown of the integrated courses and their modules in the first year of the Medicine and Surgery program, along with their associated credit values (CFU). This matches the subject matter of the user's query, which asks for the same information. Therefore, I would give this a score of 2. 2. Can the retrieved context be used exclusively to provide a full answer to the user's query? The retrieved context provides a comprehensive answer to the user's query, covering all the integrated courses and their modules in the first year of the Medicine and Surgery program, along with their associated credit values (CFU). The context is well-structured and easy to follow, making it possible to extract the required information quickly. Therefore, I would give this a score of 2. Final Result: [RESULT] 4.0

Table 4
AI Context Scores

ID	Human Evaluation Score (0-4)	Human Evaluation Notes
5	4/4	The RAG answer is complete, accurate, and well-structured, matching the gold answer point-for-point without omissions or errors. It includes all types of non-loanable materials as listed in the gold reference.
10	1/4	The RAG answer does not address the actual question, which is about the types of materials available at the University Language Centre Library. Instead, it describes general loan policies and exceptions for borrowing, likely from a different library (e.g., Frinzi Library). It is partially related to library materials but fails to provide the specific, relevant content about the Language Centre Library's collection.
43	2/4	The RAG answer provides an accurate and detailed overview of postgraduate study opportunities, including direct access to the relevant Master's program and other engineering-related fields, which aligns well with the Gold answer. However, it entirely omits the admission requirements, including the essential TOLC test and diploma criteria, as well as the program's limited enrolment structure. This missing information is critical to the question, resulting in a response that is only partially complete.
60	4/4	The RAG answer accurately identifies the two orientations—private professions and public professions—and correctly associates each with the corresponding internship opportunities. The phrasing is slightly different but conveys the same meaning as the Gold answer. The response is complete, accurate, and fully aligned with the reference.
70	4/4	The RAG answer correctly lists all integrated courses, associated modules, and their respective CFU values, matching the Gold answer in structure and content. The note acknowledging that some courses do not have module breakdowns is appropriate and does not detract from the accuracy. The response is complete, well-organized, and fully aligned with the source material.

Table 5
Human Evaluation

D. Figures

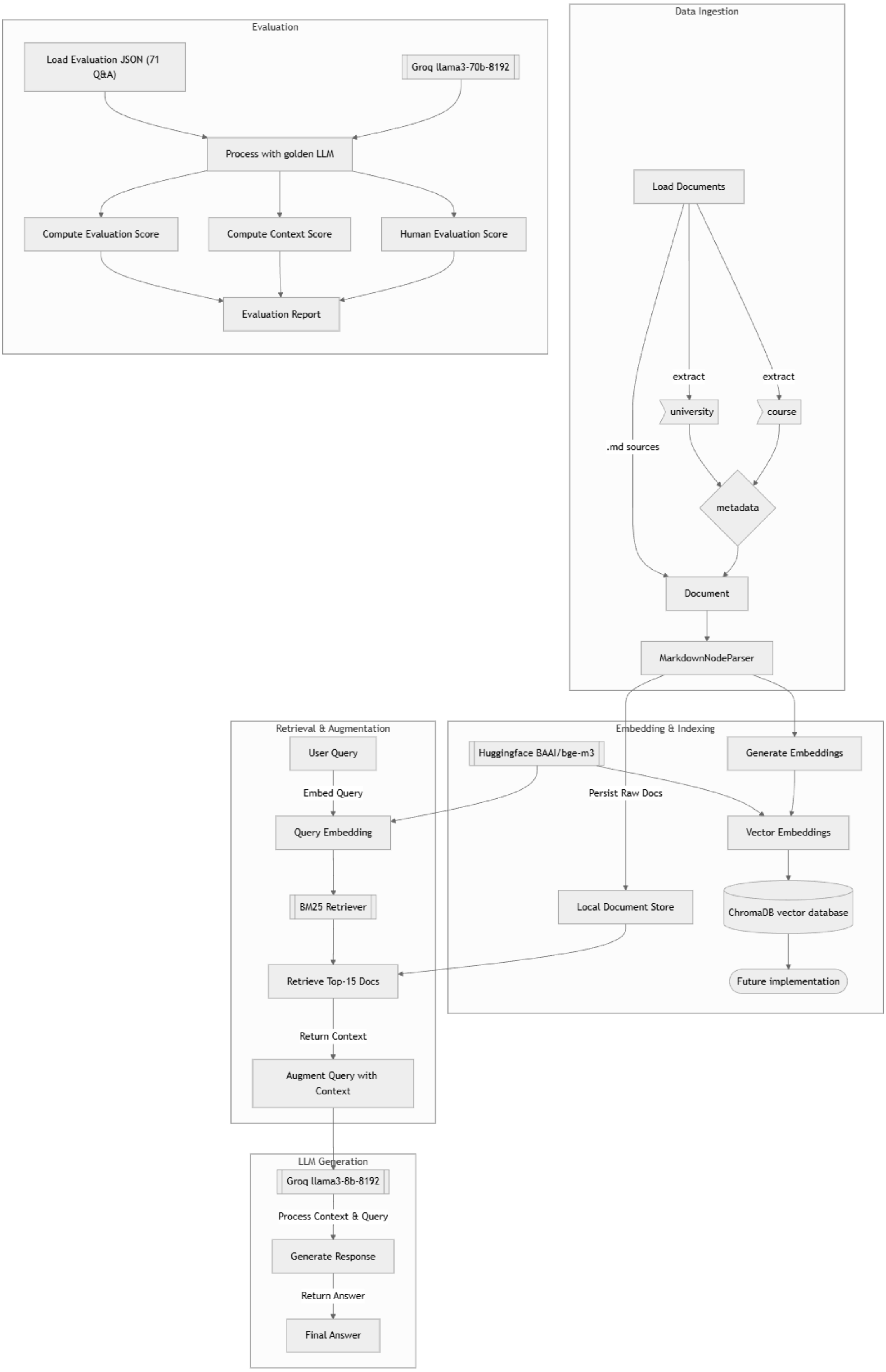

Figure 1: RAG Pipeline Diagram