Chatbot Development Using LangChain: A Case Study to Foster Creativity and Critical Thinking

Chatbot Development Using LangChain: A Case Study to Foster Creativity and Critical Thinking

Laura Farinetti*
Dipartimento di Automatica e Informatica (DAUIN)
Politecnico di Torino
Torino, Italy
laura.farinetti@polito.it

Lorenzo Canale
Centro Ricerche, Innovazione Tecnologica e Sperimentazione (CRITS)
Rai-Radiotelevisione Italiana
Torino, Italy
lorenzo.canale@rai.it

ABSTRACT

Critical thinking and creativity are fundamental skills for engineers and computer scientists. The emergence of Large Language Models (LLMs) able to create chatbots that use natural language is an opportunity for educators to foster these skills. The well-known risk of generative AI for potential misinformation offers fertile ground to practice critical thinking.

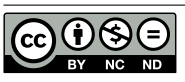
This paper describes a hands-on experience within a database course, where students had to develop a chatbot using the LangChain framework, and to evaluate it from different points of view. The students were free to choose the domain of their chatbot. The learning goal was twofold: on the one hand, to make them practice with state-of-the-art technologies, and on the other hand to stimulate critical analysis on their output. The paper discusses the students' evaluation of the chatbots under several metrics, including document retrieval, syntax and grammar accuracy, semantic relevance and information reliability. Students' assessments were also compared to the teachers' ones, to gain an insight on the critical attitude of the students and to offer a ground for discussion.

The experience was stimulating and appreciated by the students. The final results highlight that the majority of students successfully produced chatbot responses that were grammatically and syntactically correct, and that consistently extracted pertinent sections from documents, yielding semantically relevant outputs. Despite these achievements, a significant portion of students expressed reservations about the reliability of the chatbot's responses to prompts, gaining awareness of LLMs' capability to generate responses that make sense to humans but may be potentially misleading.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; *Information retrieval query processing*; • **Computing methodologies** → **Natural language processing**; *Question answering*; *Critical Thinking*; • **General and reference** → **Computing education**.

*Both authors contributed equally to this research.



ITiCSE 2024, July 8–10, 2024, Milan, Italy

ACM ISBN 979-8-4007-0600-4/24/07.
https://doi.org/10.1145/3649217.3653557

KEYWORDS

Large Language Models, LangChain framework, Chatbot development, Natural language interfaces, Information retrieval, DataBase education, Creativity and Critical Thinking

ACM Reference Format:
Laura Farinetti and Lorenzo Canale. 2024. Chatbot Development Using LangChain: A Case Study to Foster Creativity and Critical Thinking. In *Proceedings of the 2024 Innovation and Technology in Computer Science Education V. 1 (ITiCSE 2024), July 8–10, 2024, Milan, Italy*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3649217.3653557

1 INTRODUCTION AND LITERATURE

Generative Artificial Intelligence (AI) is gaining increasing attention in education, and experiments on using ChatGPT or other Learning Language Models (LLMs) are reported more and more frequently in almost every educational domain [7, 11, 12, 15]. Engineering and computer science education is naturally oriented to state-of-the-art technologies, and therefore the effort of using LLMs as educational tools is growing. The use cases cover different educational challenges, ranging from creation of artifacts, evaluation and grading [5, 13, 22, 23].

One of the most interesting challenges is to exploit generative AI to foster fundamental skills such as creativity and critical thinking, which represent key competences for the jobs of tomorrow [1, 17, 20]. Generative AI naturally requires a critical attitude for users to exploit outcomes in an effective and ethical way, and it is therefore a very useful potential tool for students [6, 8, 10, 14, 21].

This paper describes an experiment that involved students at the third year of a B.S. in Engineering, attending a course on databases.

The traditional content of a database course, i.e. SQL language, offers limited opportunity to involve creativity, and critical thinking often consists in distinguishing a correct solution from an incorrect one, in general by using tools that directly report errors, thus not requiring specific skills. The introduction of a lab activity focused on the creation of a chatbot using LLMs and on its evaluation, offered the opportunity to add this learning component. The aim of this experience was not to assess the utility of the chatbot for learning, as many studies have already done this (e.g., [2, 3, 16]), but to expose students to the technologies for creating chatbots, especially involving newer advancements like LLMs. The professional profile of the graduate in Media Engineering is different from computer scientists, as it is mainly oriented to the management of the innovation processes in digital production companies. Therefore, the lab experience does not focus on the technical details of LLMs, but on the use of an established framework like Langchain and functional

models from Huggingface to assemble the components, and on the evaluation of the effectiveness of the developed chatbots.

Similar works have been presented; for instance, in [19], students were asked to create a chatbot and to assess its effectiveness. In [4], the authors demonstrated that evaluating a chatbot improves engagement and motivates students. However, the evaluation of responses from Large Language Models remains an open challenge in the scientific community [9].

Some authors [18] argue that LLMs should be used as scientific reasoning engines rather than knowledge databases. In line with this, in the reported experience LLMs were not used as a knowledge base, but to elaborate the chatbot answer after the text segments are retrieved via embedding similarity.

2 EXPERIMENTAL SETTING

The hands-on experience was part of the database course for the students at the third year of a B.S. degree in Media Engineering. The curriculum in Media Engineering is highly interdisciplinary, and it combines technical skills (programming languages, 2D and 3D computer graphics, computer animation, virtual reality) with communication and marketing expertise regarding media, creative and cultural industries.

With respect to the more "traditional" database courses that our university offers to Management or Computer Science Engineers, where the focus is on database design and on SQL, this course also included a module about information retrieval and natural language interfaces that involved generative AI. In the teachers' intention, this was a way to expose the students to alternative and state-of-the-art approaches to knowledge representation and retrieval with respect to relational databases. The rationale behind this choice also considered that:

- *Vector stores* allow indexing documents with vector embeddings for vector search. This functionality is integrated into more traditional databases like MongoDB.
- *Large Language Models* play an essential role in accessing information through natural language interfaces. They enable a direct connection between human language and structured database queries, simplifying interaction and opening new ways for information retrieval.

80 students attended the course on database, but participation to the hands-on experience was not compulsory; its weight was 15% of the final grade of the course, and 60 students participated in the activity. Students were divided in teams of two, and the total number of teams that started the activity was 30.

The student teams' assignment was to create a chatbot using LangChain[1], a powerful framework designed for applications powered by language models. The teams were free to choose the topic of their chatbot and the type of questions to ask to measure performance. The only constraint was the topic had to be specific enough to potentially demonstrate better performance with respect to generalist chatbot such as ChatGPT. This freedom ensured higher motivation for the students and the activation of a creative approach. Most of the teams (30%) developed a chatbot related to sport or entertainment, others on tourism (17%), on services or tutorials (17%) - e.g. how to take care of houseplants, on cuisine (13%), on arts and literature (13%), and on animals (10%).

From the technical point of view, students had to follow a list of key steps to implement the chatbot:

(1) **Document Loading[2]:** Students uploaded documents using the document loader, capable of handling both PDF documents and website content. Each team was required to include a minimum of 3 and a maximum of 6 documents, including at least one PDF document and one web page.
(2) **Document Chunking via Text Splitter[3]:** The text splitter plays a crucial role in dividing documents into chunks. This step is essential for retrieval augmented generation, ensuring the extraction of the most relevant content segments.
(3) **Embedding Model Definition and Vectorstore Creation[4]:** A model was defined to create embeddings, i.e. text chunks that are stored as vectors in a Vectorstore to facilitate efficient retrieval. For the teams that chose documents in a language other than English, an additional step was required: the text chunks were automatically translated into English before being converted into embeddings, using Google Translate. This step was added to overcome the current limitations of smaller-sized Large Language Models, which perform better in English.
(4) **Large Language Model Integration[5]:** For each chatbot query, chunks most similar in cosine similarity were retrieved through embeddings. The large language model integrates the text of these retrieved chunks to generate the final response.

The entire process used Python, both for the development of the code for answer generation and for the creation of a Telegram channel as an interface connected to the Python code. This channel enabled students to interact with the chatbot through a user-friendly interface.

Large language models, despite being powerful in accessing information through natural language interfaces, are at-risk to provide inaccurate information due to imprecision inherent in the models. Additionally, even the embeddings used in the process might not always represent information accurately, leading to potential errors in the portions of documents retrieved by the Vectorstore. Given this challenge, students were specifically asked to perform an evaluation step to mitigate the risk of misinformation. The learning goal of this hands-on experience was therefore twofold: on the one hand, to practice using emerging technologies and on the other hand to activate a critical mind for evaluating the reliability of results.

The lab sessions were organized around the following phases, where each phase was guided by a dedicated notebook[6]:

(1) **Setup and preparation:** Students (a) familiarized with Colab Notebooks, (b) installed the necessary Python packages, (c) created a non-functioning Telegram bot obtaining a development key not yet integrated with Langchain, (d) selected

[1] https://python.langchain.com

[2] https://python.langchain.com/docs/modules/data_connection/document_loaders/
[3] https://python.langchain.com/docs/modules/data_connection/document_transformers/
[4] https://python.langchain.com/docs/modules/data_connection/text_embedding/
[5] https://python.langchain.com/docs/modules/model_io/llms/
[6] All notebooks are available in the following Github repository: https://github.com/Loricanal/chatbot_develpoment_langchain.

documents to extract information through Langchain, with the requirement to upload a minimum of 3 documents and a maximum of 6, including at least one web page and one PDF document, and (e) created 10 questions they assumed the chatbot could answer, manually writing the responses to be used as ground truth for later evaluation.

(2) **Initial chatbot configuration:** Students started the creation of the basic version of the chatbot (Python code without Telegram integration). They started the process by (a) selecting a preliminary method for document chunking, (b) using a single embedding model, and (c) integrating a single large language model. This initial chatbot configuration represented the groundwork for subsequent iterations, providing a baseline for comparison and evaluation. The goal was to create an initial version of the chatbot from start to finish.

(3) **Different configurations testing:** Students systematically tested 3 different embedding models and 3 different large language models, for a total of 9 combinations, by evaluating for each of them the chatbot responses to the 10 questions selected in the *Setup and preparation* phase. Their primary objective was to determine the optimal combination that would enhance the chatbot overall performance. The evaluation criteria used different metrics (retrieval accuracy and quality of the language model output in terms of semantic relevance and grammar syntax accuracy) that will be discussed later.

(4) **Integration of Telegram interface:** After the development of the chatbot logic and functionality, students integrated these features into a Telegram channel. Python code managed the functionalities of the Telegram channel, providing a user-friendly interface for real-time interaction with the chatbot. The goal was to extend the chatbot accessibility beyond coding environments, making it available on a widely used messaging platform.

Most of the notebooks required students to answer questions related to their experience too, to gather information about their satisfaction and about the main obstacles they encountered in the technical development. The analysis of these data is part of the evaluation of the experience and it is discussed in the next sections.

3 EVALUATION

The evaluation of the experiment includes several aspects, that will be analyzed in the next paragraphs. Specifically, the evaluation considers both the technical performance of the chatbot and the impact of the learning experience on the students.

3.1 Student Participation

30 student teams started the hands-on experience, and 25 of them completed all the tasks. 3 teams quit after the second phase, i.e. they tested a single configuration of the chatbot, while other 2 teams did not complete their submissions in full: one team failed to provide the code used for testing different configurations, and another did not respond to the feedback questions meant to assess the main technical challenges and satisfaction. On the other hand, 4 teams exceeded the requirements, by testing a higher number of large language models or by using more than 10 predefined questions for the chatbot. One team was remarkably active in providing assistance to other teams in testing their configurations.

In the following analysis, only the 25 teams that completed all the tasks are considered.

3.2 Chatbot Performance

This section summarizes the testing activities performed on the developed chatbots and the main results, discussed in Section 4.

3.2.1 Evaluation metrics. The performance evaluation of the chatbots was based on four key metrics:

- *Retrieval Recall:* For each question, students were requested to assign a binary value (0 or 1) to indicate whether the chatbot successfully retrieved relevant documents. The embedding model ability to represent the semantic content of the text plays a crucial role in determining the success of this retrieval process, while the LLM is not involved.
- *Syntax Grammar Accuracy:* Students used a binary value (0 or 1) to assess the quality of the syntax, looking for grammar errors in the responses generated by the large language model.
- *Semantic Relevance:* Students used a binary value (0 or 1) to assess whether a response was pertinent to the question and had an internal coherence, not being too dispersive or redundant.
- *Information Reliability*: For each of the 10 questions, students assigned a binary code (0 or 1), deciding whether the answers were verified considering the documents provided, or they introduced potential misinformation, thus assessing the reliability of each answer.

3.2.2 Best configuration selection. Student teams explored various combinations of embedding models and large language models from Hugging Face repository. Table 1 shows a summary of the models that were tested by at least two teams.

The configuration chosen most frequently by the teams uses *all-mpnet-base-v2* as embedding model and *flan-alpaca-large* as large language model (see Figure 1). Students chose the best embedding model based on *Retrieval Recall*, assessing the effectiveness of information retrieval. Concerning the large language model, the decision was made by considering the average between *Semantic Relevance* and *Information Reliability*, seeking a balance between the semantic appropriateness of the answers and the reliability of the information provided.

3.2.3 Student Assessment Results. Each student team assessed their own chatbot according to the described metrics and reported their results.

The histogram in Figure 2 summarizes the student assessment using the proposed metrics. Considering *Syntax Grammar Accuracy*, several teams achieved a high syntax and grammar accuracy rate, with 19 out of 25 teams obtaining the score of 100%. This shows that the majority of the chatbots were successful in generating responses without grammatical errors. Additionally, the majority of teams achieved a *Retrieval Recall* rate of 50% or higher, with 5 teams reaching 100%. Similar performance was reached for *Semantic Relevance*: most of the teams considered more than 70% of the retrieved answers to be correct. In contrast, for *Information*

Table 1: Embedding models and large language models selected from the Hugging Face repository and tested by at least two student teams. The column "Count" shows the number of teams that tested each model.

Model	Count
Embedding Models	
all-mpnet-base-v2	19
paraphrase-multilingual-mpnet-base-v2	10
bert-base-nli-mean-tokens	6
LaBSE	4
nq-distilbert-base-v1	4
stsb-xlm-r-multilingual	3
multi-qa-mpnet-base-cos-v1	3
msmarco-distilbert-dot-v5	3
all-MiniLM-L6-v2	3
all-MiniLM-L12-v2	3
gtr-t5-base	2
multi-qa-mpnet-base-dot-v1	2
multilingual-e5-base	2
paraphrase-distilroberta-base-v1	2
paraphrase-multilingual-MiniLM-L12-v2	2
Large Language Models	
flan-alpaca-large	17
flan-t5-base	15
flan-alpaca-base	9
flan-t5-large	6
flan-t5-small	4
long-t5-tglobal-base	4
flan-t5-text2sql-with-schema	3
t5-base-e2e-qg	2
mt5-base	2
parrot_paraphraser_on_T5	2

Reliability, the majority of teams (13 out of 25) indicated that no answers were reliable, thus acquiring awareness of LLMs capability to generate responses that make sense to humans but may be potentially misleading.

3.2.4 Teacher Assessment Results. To get an insight of the critical skills of the students, the teachers assessed all the chatbots on *Semantic Relevance* and *Information Reliability*, comparing their scores with those of the students'. Figure 3 shows that, in general, students' evaluations were more positive than the teachers'. Only in two cases teachers assigned a higher score for *Information Reliability*. Besides, the teachers' assessments reinforce the observation that *Information Reliability* tends to be generally perceived as lower than *Semantic Relevance*.

3.3 Student Experience

This section discusses the results of the hands-on experience from the educational point of view through the analysis of the students' responses to the questions included in the notebooks, designed to get feedback both on the technical challenges and on the level of satisfaction.

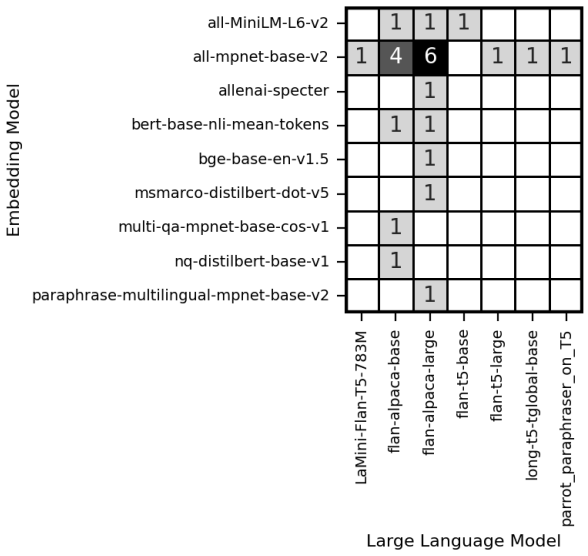


Figure 1: Heatmap that shows the best-performing configurations chosen by the student teams. The values in the cells are the number of teams that selected that configuration.

3.3.1 Technical Challenges. The following list results from the analysis of the students' responses to questions related to specific steps of the chatbot creation process. It is noteworthy that these challenges did not hinder the overall progress, as most teams employed effective strategies to overcome them.

- Document Loading Issues
  The majority of teams (18 out of 25) did not encounter any problem during this phase, and for the others teams the issues belong to two main categories:
  (a) *PDF format problems* - 2 teams initially chose PDF documents containing lots of images and/or with complex visual layout. They solved the problem looking for other PDF documents that contain mostly text;
  (b) *Websites problems:* 5 teams had issues in importing text from websites due to privacy restrictions or for the complexity of the web pages. They modified their approach by converting these pages into PDF documents.
- Document Chunking Issues
  11 teams out of 25 did not encounter problems during this phase; the remaining teams had issues with the length of segments being too long for the maximum token length of the embedding model chosen in the next phase. The solution was to perform additional splitting for documents that exceeded the token limit. 3 teams emphasized that careful selection of the length of segments and their overlap affected the *Retrieval Recall* value of the chatbot.
- Model Choice Issues
  During this phase, the only issue was the selection of functional models from HuggingFace; the main challenge was to

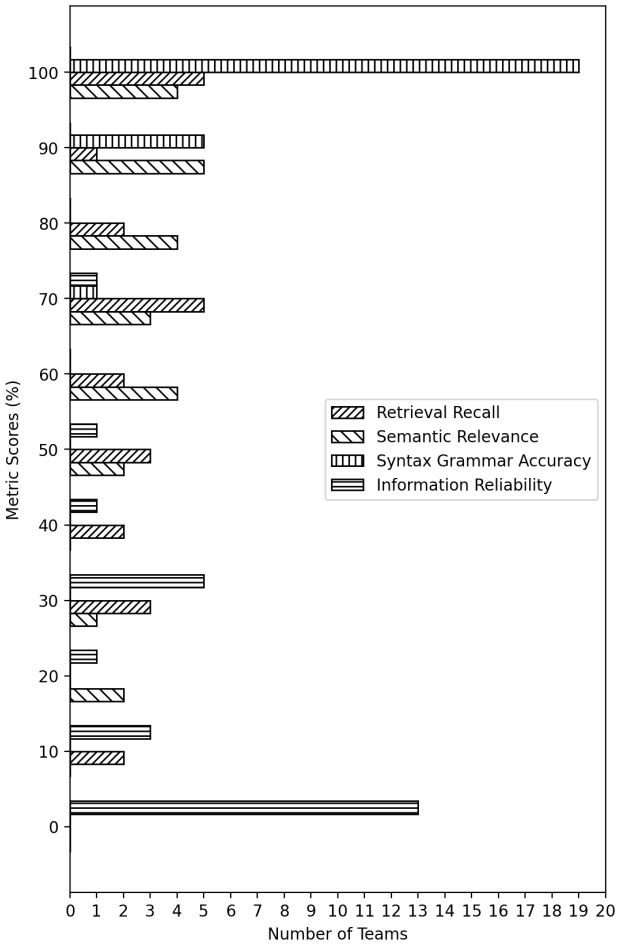


Figure 2: Histogram that shows the teams' scores for the different metrics. The Y-axis represents the percentage score for each metric, where, as an example, a value of 20 for *Retrieval Recall* indicates that the chatbot successfully retrieved relevant documents in 20% of the cases (e.g. for 2 questions out of 10). The X-axis represents the number of teams that reached a given score. For example, the bar at Y=20 has X=3, which means that three teams achieved a *Retrieval Recall* rate of 20%.

ensure compatibility with system memory limitations (especially for large language models), as all models needed to operate without GPU support.

3.3.2 Student Satisfaction. To assess the students' satisfaction with the lab sessions, they were asked to share comments, concerns, criticisms and curiosities about the learning experience. Since this was optional, only 14 teams gave their feedback.

The feedback is summarized in Table 2, where answers are classified according to the satisfaction level in "Positive engagement and interest", "Experience with pros and cons, and suggestions for improvement" and "Criticisms and suggestions".

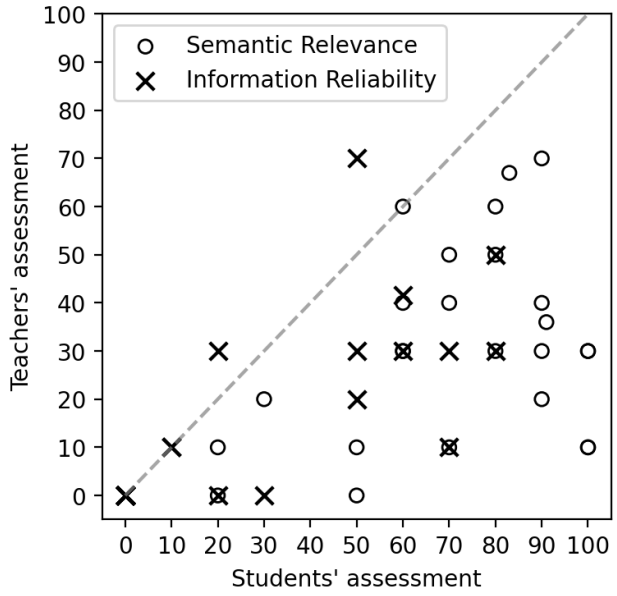


Figure 3: The scatterplot shows the comparison between students' and teachers' evaluations. Each point on the plot represents a specific evaluation instance, with the X-axis reporting the student assessments and the Y-axis the teachers' ones. This visual representation highlights the alignment or divergence between the two assessments for both Semantic Relevance and Information Reliability metrics.

4 DISCUSSION

The evaluation of the experience from the educational point of view covers different aspects, outlined in the following sections.

4.1 Impact on technical skills

Most of the teams (25 out of 30) were able to reach the final goal, i.e. to deliver a functional chatbot. The quality of their work was evaluated by the teachers based on the number of lab tasks fulfilled, code delivery, submission of the feedback responses, extra work and collaborative effort. The assigned scores varied from 1 to 5, where most of the teams (73%) got a 4.

The learning objective related to the technical skills was therefore reached by most of the students. They experienced a hands-on activity on the implementation process of a chatbot based on state-of-the-art AI techologies, and they had to face a series of technical challenges requiring understanding of new theoretical aspects and programming skills.

4.2 Impact on critical thinking and creativity

The experience was expressly designed to foster critical thinking, by (a) asking the student to assess their chatbot under different metrics that require a subjective judgement: *Retrieval Recall*, *Semantic Relevance* and *Information Reliability*, and (b) by asking them to compare *Semantic Relevance* and *Information Reliability*, encouraging them to gain awareness of LLMs capability to generate responses

that make sense to humans but may be potentially misleading. The continuous interaction between teachers and students during the whole experience made this process evident and outlined the effort of the students to critically examine all their outcomes.

Besides, the comparison between the students' and the teachers' assessment of the chatbots (see Figure 3) offered the chance to reflect together in the classroom on evaluation criteria and strategies, with open mind and critical attitude.

Students had to adopt a creative approach to select the topic of their chatbot and to write the list of 10 questions to test it. It is always difficult to evaluate creativity, but looking at the questions and by interviewing the students about their choices, the impression is that most of them appreciated the task for its creative potential, and that they really tried to create something original.

4.3 Lessons learned

Considering the students' feedback (through questionnaires and interaction during the whole semester) and the teachers' observations, the main positive aspects are:

- Innovative and intriguing experience, because it exposed students to new technologies such as large language models.
- Engaging and stimulating experience, appreciated mainly because of the detailed exploration of the chatbot development phases.
- Inspirational experience, able to arouse interest in further studies on natural language processing.

On the other hand, several areas for improvement are present:

- Need to balance the difficulty level of the laboratory with the students' technical knowledge; since the programming skills of the students can be very different, a solution could be fill the gap as a preliminary step of the lab.
- Need for more guidance, more explanations, and better support for specific technical aspects.

The main challenge for the future edition of the course is to find the best balance between challenges and appropriate support for the students.

5 CONCLUSION AND FUTURE WORK

This paper described a laboratory experience for students at the third year in a B.S. in Media Engineering as part of a database course, designed to improve both technical and soft skills. The evaluation results are satisfactory as regards student participation and engagement, and they demonstrate that the learning objectives were reached for most of the students. The students' suggestions will impact the future editions of the course: the teachers will try to better balance the difficulty level of the laboratory with the students' preliminary technical knowledge, possibly by offering different level of challenges, to avoid frustration on the one hand and lack of stimuli on the other hand.

Other actions that could improve the educational experience and/or the quality of its evaluation, and that will be implemented in the next editions of the course, are:

- Add another activity for students, i.e. peer evaluation of the chatbots, so as to provide teams with valuable feedback from other users.

Table 2: The table reports and classifies the teams' feedback about the lab experience.

Message	Details
Positive engagement and interest	
Challenging yet intriguing	Despite technical challenges, the team was intrigued by the world of embedding models and large language models.
Engaging	The team found the laboratory very interesting and engaging, thanks to the step-by-step guide into the development phases of chatbot.
Innovative and stimulating, with interest in exploring further	The team found the work extremely innovative and stimulating, expressing interest for further refining the training of the chatbot and the desire to study the process more thoroughly.
Technically fascinating	The team expressed interest in exploring the programming aspect of creating a chatbot, and it found fascinating how different combinations of models can produce different responses.
Inspirational	The team was grateful for the lab experience, as it was able to create new stimuli and to inspire further exploration into natural language processing.
Experience with pros and cons, and suggestions for improvement	
Well-structured experience, but need for more guidance	The team acknowledged that the lab was well-structured but it suggested a more rigorous guidance, especially given the complexity of the programming requirements.
Interesting, but need to better clarify technical aspects (*2 teams*)	The teams considered the activity very useful, but they suggested to dedicate more time to clarify and explore the technical aspects related to implementation. This reflects both positive engagement and a suggestion for improvement.
Interesting but too demanding time constraints	The team considered the activity very interesting, but reputed the available time to be insufficient for developing the chatbot, highlighting the complexity of the task.
Criticisms and suggestions	
Insufficient explanations (*2 teams*)	The teams faced confusion and struggled to understand technical aspects regarding Python programming and/or the Colab notebooks, suggesting the need for clearer explanations.
Insufficient support	The team suggested the need for more teaching assistants in the laboratory for a more timely support.
Insufficient programming skills (*3 teams*)	The teams did not fully appreciate the lab experience because they considered it too difficult for their preliminary programming skills.
Misalignment with course content	The team appreciated the text data retrieval theme of the lab activity, but it felt that it is not coherent with a database course content (focused on SQL), thus causing initial disorientation and requiring extra work.

- In all the notebooks, add questions to highlight the processes that involve creativity and critical thinking in each of the phases of the chatbot development. One of the limits of this experience, in fact, was the ex post assessment of the impact on creativity and critical thinking, where an ongoing assessment would have been preferable.
- Evaluate the possible impact of the topic chosen by the students on the quality of the chatbot responses.
- Assess how different types of input data influence response quality, considering factors such as text length or document layout.

REFERENCES

[1] Aoife Ahern, Caroline Dominguez, Ciaran McNally, John J. O'Sullivan, and Daniela Pedrosa. 2019. A literature review of critical thinking in engineering education. *Studies in Higher Education* 44, 5 (2019), 816–828. https://doi.org/10.1080/03075079.2019.1586325 arXiv:https://doi.org/10.1080/03075079.2019.1586325

[2] Tarek Ait Baha, Mohamed El Hajji, Youssef Es-Saady, and Hammou Fadili. 2023. The impact of educational chatbot on student learning experience. *Education and Information Technologies* (Sept. 2023). https://doi.org/10.1007/s10639-023-12166-w

[3] Fabio Clarizia, Francesco Colace, Marco Lombardi, Francesco Pascale, and Domenico Santaniello. 2018. Chatbot: An Education Support System for Student. In *Cyberspace Safety and Security*, Arcangelo Castiglione, Florin Pop, Massimo Ficco, and Francesco Palmieri (Eds.). Springer International Publishing, Cham,

291–302.

[4] Stephen Crown, Arturo Fuentes, Robert Jones, Rajiv Nambiar, and Deborah Crown. 2011. Anne G. Neering: Interactive chatbot to engage and motivate engineering students. *Computers in Education Journal* 21, 2 (2011), 24–34.

[5] Marian Daun and Jennifer Brings. 2023. How ChatGPT Will Change Software Engineering Education. In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1* (, Turku, Finland,) *(ITiCSE 2023)*. Association for Computing Machinery, New York, NY, USA, 110–116. https://doi.org/10.1145/3587102.3588815

[6] World Economic Forum. 2023. Jobs of Tomorrow: Large Language Models and Jobs. *White Paper* Published on 18 September 2023 (2023). https://www.weforum.org/publications/jobs-of-tomorrow-large-language-models-and-jobs/

[7] Stefania Giannini. 2023. Reflections on generative AI and the future of education. *White Paper* Published on 18 September 2023 (2023). https://teachertaskforce.org/sites/default/files/2023-07/2023_Giannini-UNESCO_Generative-AI-and-the-future-of-education_EN.pdf

[8] Ying Guo and Daniel Lee. 2023. Leveraging ChatGPT for Enhancing Critical Thinking Skills. *Journal of Chemical Education* (2023).

[9] Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating Large Language Models: A Comprehensive Survey. arXiv:2310.19736 [cs.CL]

[10] Sabrina Habib, Thomas Vogel, Xiao Anli, and Evelyn Thorne. 2024. How does generative artificial intelligence impact student creativity? *Journal of Creativity* 34, 1 (2024), 100072. https://doi.org/10.1016/j.yjoc.2023.100072

[11] Samuli Laato, Benedikt Morschheuser, Juho Hamari, and Jari Björne. 2023. AI-Assisted Learning with ChatGPT and Large Language Models: Implications for Higher Education. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*. 226–230. https://doi.org/10.1109/ICALT58122.2023.00072

[12] Fatemeh Mosaiyebzadeh, Seyedamin Pouriyeh, Reza Parizi, Nasrin Dehbozorgi, Mohsen Dorodchi, and Daniel Macêdo Batista. 2023. Exploring the Role of ChatGPT in Education: Applications and Challenges. In *Proceedings of the 24th Annual Conference on Information Technology Education* (, Marietta, GA, USA,) *(SIGITE '23)*. Association for Computing Machinery, New York, NY, USA, 84–89. https://doi.org/10.1145/3585059.3611445

[13] Gustavo Pinto, Isadora Cardoso-Pereira, Danilo Monteiro, Danilo Lucena, Alberto Souza, and Kiev Gama. 2023. Large Language Models for Education: Grading Open-Ended Questions Using ChatGPT. In *Proceedings of the XXXVII Brazilian Symposium on Software Engineering* (, Campo Grande, Brazil,) *(SBES '23)*. Association for Computing Machinery, New York, NY, USA, 293–302. https://doi.org/10.1145/3613372.3614197

[14] Janet Rafner, Roger Beaty, James Kaufman, Todd Lubart, and Jacob Sherson. 2023. Creativity in the age of generative AI. *Nature Human Behaviour* (11 2023). https://doi.org/10.1038/s41562-023-01751-1

[15] Parsa Rajabi, Parnian Taghipour, Diana Cukierman, and Tenzin Doleck. 2023. Exploring ChatGPT's Impact on Post-Secondary Education: A Qualitative Study. In *Proceedings of the 25th Western Canadian Conference on Computing Education* (, Vancouver, BC, Canada,) *(WCCCE '23)*. Association for Computing Machinery, New York, NY, USA, Article 9, 6 pages. https://doi.org/10.1145/3593342.3593360

[16] Jaakko Rajala, Jenni Hukkanen, Maria Hartikainen, and Pia Niemel. 2023. Call Me Kiran – ChatGPT as a Tutoring Chatbot in a Computer Science Course". In *Proceedings of the 26th International Academic Mindtrek Conference* (, Tampere, Finland,) *(Mindtrek '23)*. Association for Computing Machinery, New York, NY, USA, 83–94. https://doi.org/10.1145/3616961.3616974

[17] Branden Thornhill-Miller, Anaëlle Camarda, Maxence Mercier, Jean-Marie Burkhardt, Tiffany Morisseau, s Bourgeois-Bougrine, Florent Vinchon, Stephanie Hayek, Myriam Augereau-Landais, Florence Mourey, Cyrille Feybesse, Daniel Sundquist, and Todd Lubart. 2023. Creativity, Critical Thinking, Communication, and Collaboration: Assessment, Certification, and Promotion of 21st Century Skills for the Future of Work and Education. *Journal of Intelligence* 11 (03 2023), 54. https://doi.org/10.3390/jintelligence11030054

[18] Daniel Truhn, Jorge S. Reis-Filho, and Jakob Nikolas Kather. 2023. Large language models should be used as scientific reasoning engines, not knowledge databases. *Nature Medicine* 29, 12 (Dec. 2023), 2983–2984. https://doi.org/10.1038/s41591-023-02594-z

[19] Patchara Vanichvasin. 2021. Chatbot Development as a Digital Learning Tool to Increase Students' Research Knowledge. *International Education Studies* 14, 2 (2021), 44–53.

[20] Stéphan Vincent-Lancrin, Carlos González-Sancho, Mathias Bouckaert, Federico de Luca, Meritxell Fernández-Barrerra, Gwénaël Jacotin, Joaquin Urgel, and Quentin Vidal. 2019. *Fostering Students' Creativity and Critical Thinking*. 360 pages. https://doi.org/https://doi.org/10.1787/62212c37-en

[21] Florent Vinchon, Todd Lubart, Sabrina Bartolotta, Valentin Gironnay, Marion Botella, s Bourgeois-Bougrine, Jean-Marie Burkhardt, Nathalie Bonnardel, Giovanni Corazza, Vlad Petre Glăveanu, Michael Hanchett Hanson, Zorana Ivcevic, Maciej Karwowski, James Kaufman, Takeshi Okada, Roni Reiter-Palmon, and Andrea Gaggioli. 2023. Artificial Intelligence & Creativity: A Manifesto for Collaboration. *The Journal of Creative Behavior* 57 (06 2023). https://doi.org/10.1002/jocb.597

[22] Xiaoming Zhai. 2023. ChatGPT for Next Generation Science Learning. *XRDS* 29, 3 (apr 2023), 42–46. https://doi.org/10.1145/3589649

[23] Yong Zheng. 2023. ChatGPT for Teaching and Learning: An Experience from Data Science Education. In *Proceedings of the 24th Annual Conference on Information Technology Education* (, Marietta, GA, USA,) *(SIGITE '23)*. Association for Computing Machinery, New York, NY, USA, 66–72. https://doi.org/10.1145/3585059.3611431