People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

KASDI MERBAH UNIVERSITY - OUARGLA

Faculty of New Technologies of Information and Telecommunication

Department of Computer Science and Information Technology

MASTER

Domain : Computer Science

Field : Industrial Computing

Submitted by : Bencherif Atika and Arif Nour el imane

Thesis :

Developement of a conversational agent for the university using LLM models

---

Evaluation Date : 21/06/2024

Before the Jury :

Abderrahim Mohammed El Amine	Supervisor	UKM Ouargla
Adel Zga	President	UKM Ouargla
Bachir Said	Examiner	UKM Ouargla

Academic year: 2023/2024

Acknowledgment

We wish to thank first and foremost our Allah, the Almighty, the greatest of all. We would like to express our sincere gratitude to the following individuals and organizations that have contributed to the completion of our master's degree thesis: First and foremost, we would like to express our deepest appreciation to our thesis advisor, **Pr. Abderrahim Mohammed El Amine**, for his guidance, expertise, and unwavering support throughout this research journey. His valuable insights, constructive feedback, and dedication have played a pivotal role in shaping this thesis. We would like to acknowledge the support and encouragement from our families, especially our parents and siblings, for their love, patience, and understanding. Their constant support and belief in us have been a constant source of motivation. We extend our appreciation to our friends and colleagues who have provided assistance, insightful discussions, and encouragement throughout this thesis undertaking. Their camaraderie and intellectual exchange have been invaluable. Furthermore, we are thankful to the research participants who generously shared their time, knowledge, and experiences, without whom this study would not have been possible. In conclusion, the completion of this master's degree thesis would not have been possible without the support, guidance, and contributions of all these individuals and organizations. We are deeply indebted to each and every one of them.

Dedication

"

الحمد لله حبا وشكرا وامتنانا، ما كنت لأفعل هذا لولا فضل الله فالحمد لله على البدء والختام. ها أنا اليوم أهدي نجاحي هذا الى أبطال بيت المقدس الذين غيروا العالم بخطواتهم الصادقة الى الصامدين على الثغور المدافعين عن شرف الأمة وفدوى الحق بأعمارهم(اخواننا في غزة)لمن رضاها يخلق لي التوفيق.

إلى تلك الانسانة العظيمة التي لطالما تمنت أن تقر عينها برؤيتي في يوم كهذا التي كانت الأم و الأخت والصديقة داعمتي الأولى (أمي الحبيبة)أطال الله في عمرها الى من أحمل اسمه بكل فخر إلى ذلك الرجل العظيم الذي علمني الحياة بأجمل شكل و بذل كل ما بوسعه ولم يبخل (أبي الغالي)أطال الله في عمره الى من رزقت بهم سندا لي اخوتي (صلاح الدين ، خديجة، طه العدنان، ليان)الى كل الأهل و العائلة الكريمة الذين يفرحهم نجاحنا بالأخص (خالتي مروة و سارة) الى رفقاء الروح الذين شاركوني خطوات هذا الطريق الى من ساندوني بكل حب عند ضعفي (أماني ،إسراء).

و أحب أن أختم الاهداء الى صاحبة الفضل العظيم صديقة الرحلة والنجاح الى من وقفت بجانبي كلما أوشكت أن أتعثر التي شاركتني لحظات التعب و الفرح طيلة المشوار(عاتكة)ولله الشكر كله أن وفقني لهذه اللحظة لم تكن الرحلة قصيرة ولم تكن الأمور يسرة ولكن بحول الله ها أنا قد وصلت لمشارف التخرج فالحمد لله رب العالمين.

"

عريف نور الإيمان

Dedication

"

(وَآخِرُ دَعْوَاهُمْ أَنِ الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِين)

الحمد لله حبا وشكرا وامتنانا على البدء والختام لم تكن الرحلة قصيرة ولا ينبغي لها ان تكون لم يكن الحلم قريبا لا الطريق كان محفوفا بالتسهيلات لكنني فعلتها فالحمد لله الذي يسر البدايات وبلغنا النهايات بفضلة وكرمه إلى الذي زين أسمي بأجمل الألقاب، من دعمني بلا حدود وأعطاني بلا مقابل إلى من علمني أن الدنيا كفاح وسلاحها العلم والمعرفة إلى من غرس في روحي مكارم الأخلاق داعمي الأول في مسيرتي وسندي وقوتي وملاذي بعد الله الى فخري واعتزازي والدي العزيز (احمد بن شريف) الى من جعل الله الجنة تحت اقدامها واحتضنني قلبها قبل يدها وسهلت لي الشدائد بدعائها الى القلب الحنون والشمعة التي كانت لي في الليالي المظلمات سر قوتي ونجاحي والدتي الحبيبة(بن عطية فضيلة) الى خيرة أيامي وصفوتها كانوا لي سندا وداعمين ومشجعين دائما أزاحوا عن طريقي المتاعب ممهدين الطريق زارعين الثقة والاصرار بداخلي سندي وكتفي الذي استند عليه دائما اخواني وإخوتي حفظهم الله إلى أصدقاء السنين وأصحاب الشدائد وملهمين نجاحي الى من رسموا بسمتي وقت الصعاب إلى من ذكروني بمدى قوتي واستطاعتي إلى الشموع التي تنير لي الطريق دوما صديقاتي (امنة ايمان جيهان) الى صديقة الرحلة والنجاح التي شاركتني لحظات التعب والفرح طيلة المشوار (نور) واخيراً من قال أنا لها "نالها" وأنا لها ان أبت رغما عنها اتيت بها ماكنت لأفعل لولا توفيق من الله ... لكل من كان عونا وسندا في هذا الطريق أهديكم هذا الإنجاز وثمرة نجاحي الذي لطالما تمنيته ها أنا اليوم أتممت أول ثمراته، راجية من الله تعالى أن ينفعني بما علمني وأن يعلمني ما أجهل ويجعله حجة لي لا علي.

"

بن شريف عاتكة

Abstract

A chatbot is a computer program designed to facilitate conversations between humans and machines. It can be used across various platforms, such as messaging apps and virtual assistants. Over the years, chatbots have evolved significantly, transitioning from being mere entertainment to performing important tasks. When creating a chatbot, several design considerations should be taken into account, including its purpose, target audience, communication channels, conversational flow, and the need for testing and iterative improvements to ensure accuracy and user-friendliness. Based on their domain, model, and conversation style, chatbots can be categorized into various types, including customer service, sales, informational, personal assistant, entertainment, health, and educational chatbots. Each type serves a specific function and caters to the needs of different user groups. Despite technological advancements, chatbot technology still faces several challenges. These challenges include contextual understanding, seamless integration with backend systems, personalization for individual users, ensuring security, and gaining user acceptance and trust. This dissertation aims to develop an information chatbot that can answer different questions related to university studies. The latter can be installed on the university home page.

Keywords : Chatbot, Large Language models (LLMs), Machine Learning (ML), Natural Language Processing (NLP), Artificial Intelligence (AI).

Résumé

Un chatbot est un programme informatique conçu pour faciliter les conversations entre les humains et les machines. Il peut être utilisé sur diverses plateformes, telles que les applications de messagerie et les assistants virtuels. Au fil des ans, les chatbots ont connu une évolution significative, passant d'un simple divertissement à l'exécution de tâches importantes. Lors de la création d'un chatbot, plusieurs considérations de conception doivent être prises en compte, notamment son objectif, son public cible, les canaux de communication, le flux de conversation, ainsi que la nécessité de tests et d'améliorations itératives pour assurer la précision et la convivialité pour l'utilisateur.

Selon leur domaine, leur modèle et leur style de conversation, les chatbots peuvent être classés en différentes catégories, notamment les chatbots de service client, de vente, d'information, d'assistant personnel, de divertissement, de santé et d'éducation. Chaque type remplit une fonction spécifique et répond aux besoins de différents groupes d'utilisateurs.

Malgré les avancées technologiques, la technologie des chatbots fait encore face à plusieurs défis. Ces défis incluent la compréhension contextuelle, l'intégration transparente avec les systèmes en arrière-plan, la personnalisation pour les utilisateurs individuels, la garantie de la sécurité et l'obtention de l'acceptation et de la confiance des utilisateurs.

Ce mémoire a pour objectif de développer un chatboot d'information pouvant répondre à différentes questions en relation avec les études universitaires. Ce dernier peut être installé dans la page d'accueil des universités.

Mots clés : Chatbot, Modèles de grands langages (LLM), Apprentissage automatique (ML), Traitement du langage naturel (NLP), Intelligence artificielle (AI).

ملخص

الشاتبوت هو برنامج حاسوبي مصمم لتيسير المحادثات بين البشر والآلات. يمكن استخدامه عبر مجموعة من المنصات، مثل تطبيقات المراسلة والمساعدين الافتراضيين. على مر السنين، تطورت الشاتبوتات بشكل كبير، انتقلت من كونها مجرد تسلية إلى أداء مهام مهمة. عند إنشاء شاتبوت، يجب مراعاة العديد من الاعتبارات التصميمية، بما في ذلك الغرض منه، والجمهور المستهدف، وقنوات التواصل، وسير المحادثة، وضرورة الاختبار والتحسينات التكرارية لضمان الدقة وسهولة الاستخدام.

بناء على مجالها ونموذجها وأسلوب المحادثة، يمكن تصنيف الشاتبوتات إلى أنواع مختلفة، بما في ذلك شاتبوتات خدمة العملاء، والمبيعات، وتقديم المعلومات، والمساعد الشخصي، والترفيه، والصحة، والتعليم. تقدم كل نوع وظيفة محددة وتلبي احتياجات فئات مستخدمين مختلفة.

على الرغم من التقدم التكنولوجي، تواجه تكنولوجيا الشاتبوت تحديات عديدة. من هذه التحديات الفهم السياقي، والتكامل السلس مع أنظمة الخلفية، والتخصيص للمستخدمين الفرديين، وضمان الأمان، وكسب قبول المستخدم والثقة. تهدف هذه المذكرة إلى تطوير برنامج دردشة (شاتبوت) معلوماتي يمكنه الإجابة على الأسئلة المختلفة المتعلقة بالدراسات الجامعية. يمكن تثبيت هذا الأخير على الصفحة الرئيسية للجامعة.

كلمات مفتاحية : شاتبوت، نماذج اللغات الكبيرة, التعلم الآلي، معالجة اللغة الطبيعية، الذكاء الاصطناعي.

Contents

Acknowledgment . . . . **I**

Dedication . . . . **II**

Abstract . . . . **IV**

Résumé . . . . **V**

ملخص . . . . **VI**

General introduction **1**

1 Natural Language Processing and Artificial Intelligence **3**

1.1 Introduction 4

1.2 Natural Language Processing 4

1.2.1 NLP Fields 4

1.2.2 NLP Challenges 5

1.3 Artificial Intelligence 5

1.3.1 Machine learning 5

1.3.2 Artificial Neural Network (ANN) 6

1.3.3 Deep learning 6

1.3.4 The Transformer Architecture 7

1.3.5 Large Language Models (LLMs) 9

1.3.5.1 Large Language Model Families 9

1.3.5.2 Llama2 10

1.3.5.3 Three scenarios are possible for using LLMs 12

1.4 Conclusion 15

2 ChatBot : Overview and Related Works **16**

2.1 Introduction 17

2.2 ChatBot : Overview 17

2.2.1 What is a Chatbot? 17

2.2.2 Evolution 18

2.2.3 Chatbot characteristics 20

2.2.4 Design principles 20

2.2.5 Chatbot Structure 21

2.2.6 Chatbot Tools 22

2.2.7 Chatbot challenges 23

2.3 ChatBot : Related Works . . . 24
2.3.1 Summary of related works . . . 26
2.4 Conclusion . . . 27

3 ChatBot for student needs: Design and Development . . . 28
3.1 Introduction . . . 29
3.2 Types of Information and Support Students Seek . . . 29
3.3 Analysis of Student Needs . . . 29
3.4 Student Expectations and Preferences for Interacting with a Chatbot . . . 30
3.5 Chatbot Features . . . 31
3.6 Chatbot Design and Development . . . 31
3.6.1 Use of LLMs . . . 31
3.6.2 Necessary Tools . . . 32
3.7 Using RAG technique . . . 32
3.7.1 Data collection . . . 32
3.7.2 RAG Process . . . 33
3.7.2.1 Setting Up the Environment and Installing Libraries . . . 33
3.8 Fine-Tuning Process . . . 35
3.9 Comparison of Llama-2 Model Answers Using RAG and Fine-Tuning Techniques 40
3.9.1 Methodology . . . 41
3.9.2 Results . . . 41
3.9.3 Analysis . . . 43
3.9.4 Discussion of Findings . . . 43
3.9.5 Conclusions and Recommendations . . . 43
3.10 Validation Description . . . 43
3.11 Challenges . . . 44
3.12 Conclusion . . . 44

General Conclusion . . . 45

Bibliography . . . 46

List of Figures

1.1 A simple artificial neural network architecture[24] 7
1.2 The Transformer - model architecture[7] 8
1.3 Populaire LLM Familiers [9] 10
1.4 Training of Llama 2-Chat [10] 11
1.5 Prompt engineering [25] 13
1.6 A representative instance of the RAG process applied to question answering[26] 14

2.1 An overview of the properties of chatbots [13] 20

3.1 Diagram Explained RAG Process 35
3.2 Diagram Explained Fine-tuning Process 39

List of Tables

2.1 Evolution of Chatbot [2] 19
2.2 Summary of related works 26

3.1 Dataset 33
3.2 Data Collection Informations 36
3.3 Comparison Table of Llama-2 Model Answers Using RAG and Fine-Tuning Techniques. 41

Abbreviations

AI	*Artificial Intelligence*
ANN	*Artificial Neural Network*
DP	*Deep Learning*
LLM	*Large Language Model*
ML	*Machine Learning*
NLP	*Natural Language Processing*
NLU	*Natural Language Understanding*
NLG	*Natural Language Generation*
RAG	*Retrieval Augmented Generation*
RLHF	*Reinforcement Learning with Human Feedback*

General Introduction

The evolution of chatbots has undergone significant transformations from simple rule based systems to advanced natural language processing (NLP) models. Initially designed to handle basic queries and automate repetitive tasks, chatbots have now become more sophisticated due to the progress in artificial intelligence (AI) and machine learning (ML). Presently, they can engage in complex conversations and offer personalized experiences [1]. The rising popularity of voice assistants and messaging apps has further contributed to their widespread adoption in various industries, such as customer service, sales, and marketing. Utilizing data analytics and sentiment analysis, chatbots are now capable of understanding human emotions and responding in a more human-like manner.

As technology continues to advance, chatbots are becoming indispensable tools for businesses' digital transformation, enhancing customer experiences, and improving overall efficiency [2].In today's digital age, these intelligent conversational agents have become invaluable tools in the field of universities, offering a wide range of benefits that enhance the overall educational experience. However, despite their widespread adoption in many parts of the world, Algerian universities have yet to fully embrace this transformative technology, highlighting a crucial gap that needs to be addressed. Chatbots play a pivotal role in modernizing and streamlining university operations. They provide instant and personalized assistance to students, faculty, and staff, improving communication channels and efficiency.

For Algerian universities, and especially in our university of kasdi merbah ouargla, integrating chatbots can lead to substantial advantages. Firstly, they can enhance student engagement by offering quick and accessible support for inquiries related to admissions, course information, schedules, and academic resources. This proactive approach fosters a positive student experience and contributes to higher retention rates. Moreover, chatbots can assist university administrators in managing administrative tasks more efficiently. They can automate routine processes such as handling registration forms, scheduling appointments, and providing updates on campus events. This automation not only saves time and resources but also reduces the risk of errors, ensuring smoother operations across departments.

Furthermore, the absence of chatbots in Algerian universities represents a missed opportunity to leverage technology for improving educational outcomes. By implementing chatbot systems tailored to the specific needs of Algerian higher education institutions, universities can optimize their services, increase accessibility, and adapt to the evolving expectations of students and stakeholders.

In this thesis, we delve into the significance of chatbots in the university of kasdi merbah ouargla settings, examining their potential benefits for Algerian universities and proposing strategies for their effective implementation. By exploring the role of chatbots in enhancing communication, efficiency, and user experience, we aim to shed light on the importance of embracing innovative technologies to propel Algerian higher education into the digital age. To achieve our objectives, we address the following outline:

Chapter 1 describes the main concepts of the two fields covered in our dissertation, namely the field of NLP and ML. Chapter 2 presents Chatbots, it outlines their developments, structures and implementations. It also introduces the currently most models used in their production, namely LLMs. Consequently, chapter 3 is devoted to the description of the subsequent implementation of our project.

Chapter 1

Natural Language Processing and Artificial Intelligence

1.1 Introduction

In the digital realm, chatbots act as sophisticated interfaces that mimic human conversation, blending AI and NLP seamlessly. In this chapter we will briefly introduce the two fields, namely NLP and AI.

1.2 Natural Language Processing

NLP is a specialized branch of AI that focuses on equipping computers with the capability to comprehend and interpret text and spoken language like how humans do. It involves developing algorithms and models that enable machines to extract meaning, context, and intent from human language, whether it is written or spoken.

NLP is considered one of the most challenging and complex fields within AI due to the intricacies involved in language understanding and processing. The nuances of human communication, including grammar, semantics, pragmatics, and context, present significant hurdles that NLP researchers and engineers strive to overcome. Despite its difficulties, NLP has been garnering increasing interest and attention with each passing day. As advancements in AI technologies continue to unfold, NLP plays a crucial role in developing intelligent virtual assistants, language translation systems, sentiment analysis tools, voice recognition systems, and many other applications that enhance human-computer interactions and make machines more human-like in their understanding and response capabilities[3].

The NLP domain includes several fields such as:

1.2.1 NLP Fields

The field of NLP includes several fields, we cite for example[21]

- Text Classification
- Topic Modelling
- Word Embedding
- Texte Generation
- Sentiment Analysis
- Text Similarity
- Auto-Correction
- Search Engines
- Chatbots

Among challenges of NLP we can cite[22][23]

1.2.2 NLP Challenges

Natural Language Processing (NLP) faces several intricate challenges due to the inherent complexity and dynamic nature of human language:

- **Language Complexity:** Natural languages exhibit complex structures, nuances, and multiple interpretations. Understanding and processing the intricacies of human language remains a significant challenge.
- **Dialect Variations:** Languages can vary significantly across regions and social groups, leading to dialectal differences that NLP models must accommodate for effective communication.
- **Ambiguity and Misunderstanding:** Ambiguous words, phrases, and context-dependent meanings can lead to misunderstandings in NLP systems, requiring context-aware understanding.
- **Handling Sarcasm and Irony:** Identifying sarcasm and irony in text poses challenges as they often require understanding the speaker's intent and context to interpret correctly.
- **Language Rhetoric:**NLP systems must grasp rhetorical devices used in language, such as metaphors, similes, and analogies, to fully comprehend text.
- **Abbreviations and Acronyms:** Texts often include various abbreviations and acronyms, which can be context-specific and challenging for NLP models to interpret accurately.
- **Ever-Changing Language:** : Languages evolve, incorporating new words, expressions, and meanings. Keeping NLP systems up-to-date with the latest language trends is a continual challenge.

Addressing these challenges requires the development of sophisticated NLP models that can adapt to language variations, context, and evolving language use. It involves leveraging machine learning algorithms, neural networks, and large-scale datasets to enhance the accuracy and understanding of NLP systems across diverse linguistic contexts. Additionally, research into context-aware language modeling and dialogue systems can help improve the naturalness and effectiveness of NLP interactions.

1.3 Artificial Intelligence

1.3.1 Machine learning

Machine learning is a subfield of AI that involves training algorithms to learn patterns and make predictions or decisions without being explicitly programmed. It relies on large datasets as input for learning algorithms, resulting in trained models that can perform various tasks.

ML models go through a lifecycle that includes collecting and organizing data, training the model, and using it for tasks or supporting other larger tasks. The trained model is a permanent file that can be used on previously unseen datasets or to inform and support other tasks. ML has practical applications in various fields, including NLP and computer vision[4].

1. **Data collection:** Large corpora of data, such as texts, images, and videos, are gathered as input for the ML algorithms.

2. **Data organization:** The collected data is organized and prepared for training the model, ensuring it is in a suitable format and quality.

3. **Model training:** The organized data is used to train the ML model, where algorithms learn patterns and make predictions or decisions without explicit programming.

4. **Model operation:** Once the model is trained, it can be used for various tasks or to support larger tasks, such as NLP or computer vision.

1.3.2 Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) consist of neurons organized in layers, including input, hidden, and output layers. Neurons receive inputs, process them using weighted connections, and produce outputs through activation functions. Connections between neurons have weights that determine signal strength. During training, weights are adjusted using algorithms like backpropagation to minimize errors and improve performance. ANNs are inspired by the biological nervous system and are used for tasks like pattern recognition, prediction, and system identification. They operate in a parallel and distributed manner, mimicking the brain's information processing. ANNs are effective in solving complex nonlinear problems and have applications in image processing, classification, and more. By combining many simple processing units, ANNs can generate complex behaviors and make accurate predictions. Overall, ANNs provide a powerful computational model for simulating human-like intelligence and processing vast amounts of data efficiently [5]. Figure 1.1 shows a simple artificial neural network architecture.

1.3.3 Deep learning

Deep learning is a subfield of ML that focuses on training ANNs with multiple layers, allowing them to learn hierarchical representations of data. It involves the use of deep neural networks, which are composed of multiple layers of interconnected neurons, enabling the network to learn complex patterns and relationships in the data. DL has been successful in various domains, including computer vision, NLP, and speech recognition, achieving state-of-the-art performance in tasks such as image classification, object detection, language translation, and speech synthesis[5].

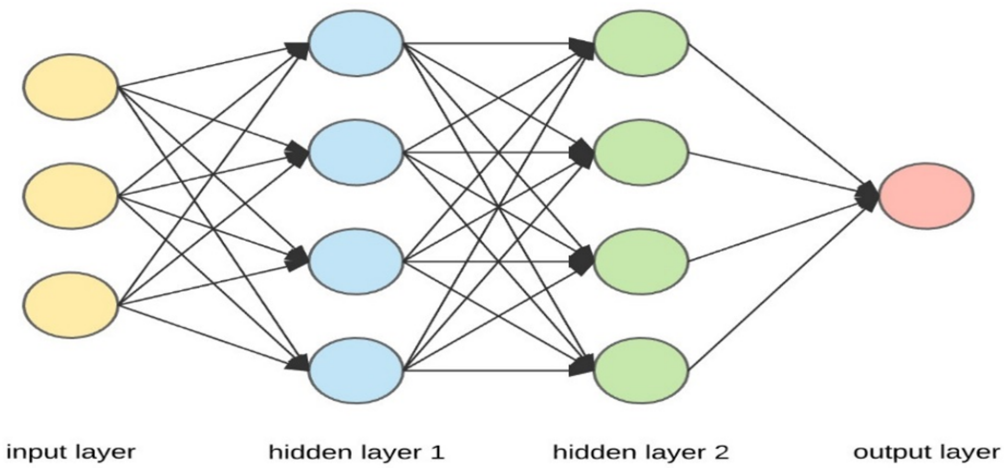

Figure 1.1: A simple artificial neural network architecture[24]

1.3.4 The Transformer Architecture

The Transformer is a neural network architecture that relies on self-attention mechanisms to process sequential data. It consists of multiple layers of self-attention and feedforward neural networks, allowing it to capture relationships between words in a sequence without the need for recurrent connections [6].

Key components of the Transformer include self-attention layers, which enable the model to weigh the importance of different words in a sequence based on their contextual relevance, and feedforward layers, which process the information extracted by the self-attention mechanism.

The Transformer architecture has been widely adopted in various NLP tasks, such as machine translation, text generation, and sentiment analysis, due to its ability to handle long-range dependencies effectively and its parallelizability, which makes it more efficient for training on large datasets [6]. Figure 2.1 shows a visualizes the architecture

- **Input Embedding:** The input text is tokenized into smaller units, such as words or sub-words, and each token is embedded into a continuous vector representation. This embedding step captures the semantic and syntactic information of the input.

- **Positional Encoding:** Positional encoding is added to the input embedding to provide information about the positions of the tokens because transformers do not naturally encode the order of the tokens. This enables the model to process the tokens while taking their sequential order into account.

- **Encoder:** Based on a neural network technique, the encoder analyses the input text and creates a number of hidden states that protect the context and meaning of text data. Multiple encoder layers make up the core of the transformer architecture. Self-attention mechanism and feed-forward neural network are the two fundamental sub-components of each encoder layer.

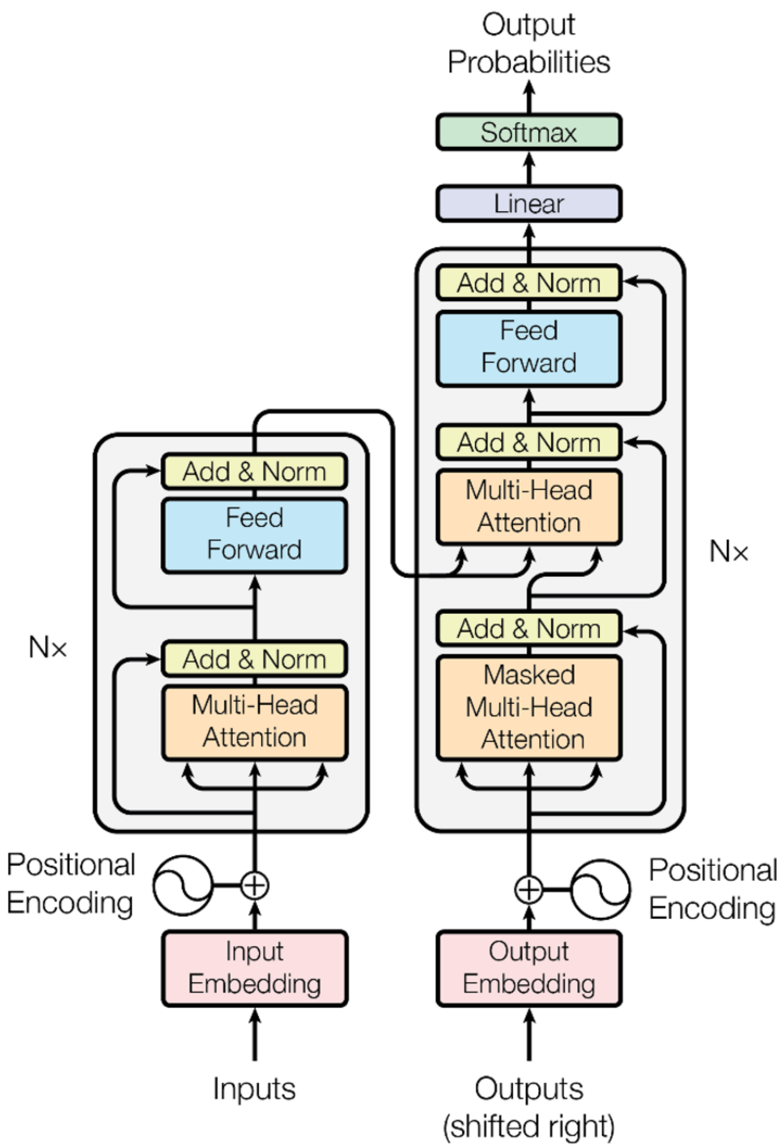

Figure 1.2: The Transformer - model architecture[7]

- **Self-Attention Mechanism:** Self-attention enables the model to weigh the importance of different tokens in the input sequence by computing attention scores. It allows the model to consider the dependencies and relationships between different tokens in a context-aware manner.

- **Feed-Forward Neural Network:** After the self-attention step, a feed-forward neural network is applied to each token independently. This network includes fully connected layers with non-linear activation functions, allowing the model to capture complex interactions between tokens.

- **Decoder Layers:** In some transformer-based models, decoder component is included in addition to the encoder. The decoder layers enable autoregressive generation, where the model can generate sequential outputs by attending to the previously generated tokens.

- **Multi-Head Attention:** Transformers often employ multi-head attention, where self-attention is performed simultaneously with different learned attention weights. This allows the model to capture different types of relationships and attend to various parts of the input sequence simultaneously.

- **Layer Normalization:** Layer normalization is applied after each sub-component or layer in the transformer architecture. It helps stabilize the learning process and improves the model's ability to generalize across different inputs.

- **Output Layers:** The output layers of the transformer model can vary depending on the specific task. For example, in language modeling, a linear projection followed by Soft max activation is commonly used to generate the probability distribution over the next token.

1.3.5 Large Language Models (LLMs)

Are AI models that can process and generate natural language text. They are trained on massive amounts of text data using DL techniques, allowing them to learn patterns and structures of language. LLMs are typically pre-trained on extensive corpora from the web, enabling them to learn complex patterns, linguistic nuances, and semantic relationships. Fine-tuning LLMs on specific downstream tasks has shown impressive performance in various benchmarks. The development of LLMs has been driven by advancements in deep learning methods, computational resources, and the availability of training data. The Transformer model, introduced in 2017, played a significant role in the development of LLMs by enabling the learning of longer-term dependencies in language and parallel training on multiple GPUs. LLMs have become a crucial tool in NLP, pushing the boundaries of what machines can achieve in understanding and generating human-like text [8].

1.3.5.1 Large Language Model Families

Large Language Model Families (LLMs) refer to transformer-based PLMs (Pre-trained Language Model)with tens to hundreds of billions of parameters. These models are divided into three main families: GPT, Llama, and Palm [9].

1. **The GPT Family:** comprise Generative Pre-Trained Transformers (GPT), which are decoder-only Transformer-based language models developed by Open AI. This family includes GPT-1, GPT-2, GPT-3, Instruct GPT, Chat GPT, GPT-4, CODEX, and Web GPT. While early GPT models like GPT-1 and GPT-2 are open-source, newer versions such as GPT-3 and GPT-4 are closed-source and accessible only through APIs .

2. **The Palm Family:** developed by Google, consists of the Pathways Language Models. The initial Palm model was announced in April 2022 but remained private until March 2023. This model is a transformer-based LLM with 540 billion parameters.

It undergoes pre-training on a high-quality text corpus comprising 780 billion tokens covering various natural language tasks and scenarios.

Palm utilizes the Pathways system for training, employing 6144 TPU v4 chips to ensure highly efficient training across multiple TPU Pods. The scaling benefits of Palm are evident in its state-of-the-art few-shot learning results across numerous language understanding and generation benchmarks. PaLM540B not only outperforms leading fine-tuned models on multi-step reasoning tasks but also achieves performance levels comparable to humans on the recently introduced BIG-bench benchmark .

3. **The Llama Family:** refers to a series of foundational language models released by Meta. Unlike GPT models, Llama models are open-source, meaning their model weights are made available to the research community under a noncommercial license. As a result, the Llama family is expanding quickly due to widespread use by various research groups. These groups use Llama models to create improved open-source LLMs that can compete with closed-source models or develop task-specific LLMs for critical applications.

   As shown in Figure 1.3, these popular LLM families illustrate the diverse approaches and capabilities of different language models in the field .

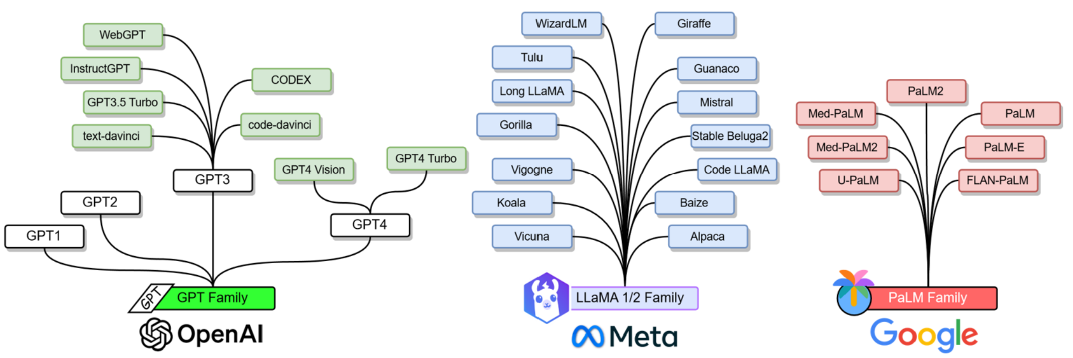

Figure 1.3: Populaire LLM Familiers [9]

1.3.5.2 Llama2

Llama 2 is a collection of pertained and fine-tuned (LLMs) optimized for dialogue use cases. It ranges in scale from 7 billion to 70 billion parameters. Llama 2-Chat, the fine-tuned version of Llama 2, outperforms open-source chat models on most benchmarks and is considered a suitable substitute for closed-source models. The models were trained between January 2023 and July 2023 and are the result of several months of research and iterative applications of alignment techniques. Llama 2-Chat is able to refer to attributes 100% of the time, for up to 20 turns, based on human evaluation. The models have undergone safety improvements and developers are encouraged to perform safety testing and tuning tailored to their specific applications. Llama 2 and Llama 2-Chat are released with the aim of enabling the community to contribute to the responsible development of LLMs [10].

The pre-training phase of the training of Llama 2-Chat (see Figure 1.4) is the fundamental stage where the models are trained extensively on a large corpus of text data to discover statistical patterns and language representations.

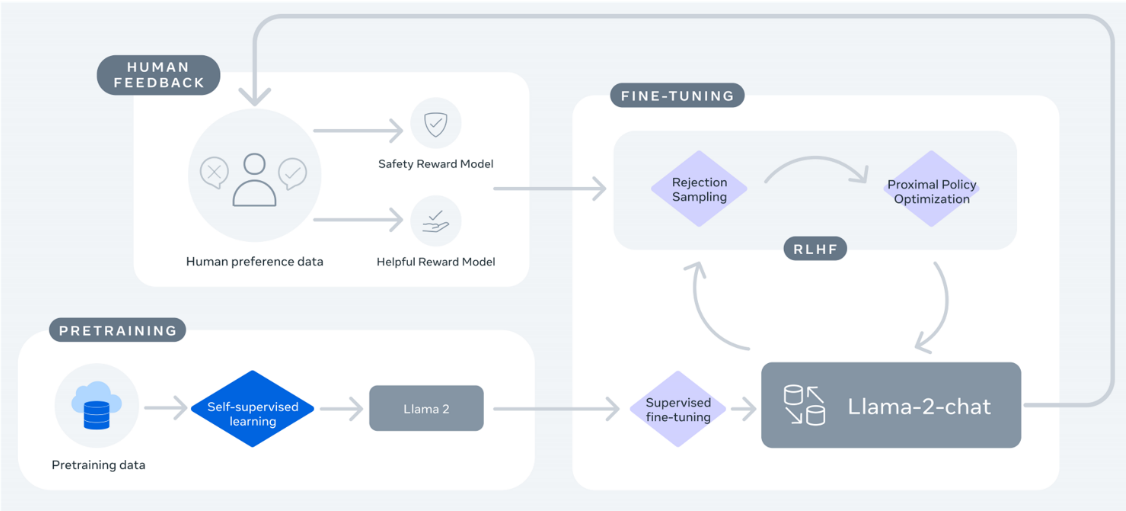

Figure 1.4: Training of Llama 2-Chat [10]

The Pretraining

The pre training phase is the initial stage in developing language models where the models undergo training on a large corpus of text data to learn statistical patterns and language representations. Here are the key points about pre training[10]:

- Llama 2 models were pre trained on a massive dataset, including 2 trillion tokens sourced from publicly available materials, aiming to capture a broad range of language patterns and knowledge.
- The training corpus for Llama 2 models was meticulously curated, emphasizing factual sources and efforts to filter out data from websites containing personal information.
- The pre training process involved utilizing an optimized autoregressive transformer and implementing several enhancements to boost performance, such as robust data cleaning, updated data mixes, increased training tokens, longer context length, and grouped-query attention (GQA) for improved inference scalability.
- The pre training data used for Llama 2 models has a cutoff point in September 2022, although some fine-tuning data is more recent, up to July 2023.

Fine-tuning

Fine-tuning is a crucial process used to enhance the performance of language models by training them on specific tasks or domains. It involves adjusting the parameters of a pertained model using additional data and specialized techniques. Here are the key points about fine-tuning[10]:

- **Supervised Fine-Tuning:** This process includes adversarial prompts and safe demonstrations to align the model with safety guidelines, ensuring responsible and ethical behavior.
- **Ghost Attention:** A technique employed during fine-tuning to regulate dialogue flow across multiple turns. Ghost Attention enhances dialogue control and overall performance of the model.
- **Autoregressive Objective:** In fine-tuning, an autoregressive objective is utilized, where the loss is minimized on tokens from the user prompt. Backpropagation occurs solely on answer tokens, contributing to improved model accuracy.
- **Reinforcement Learning with Human Feedback (RLHF):** RLHF is integrated into the fine-tuning phase to further align the model's behavior with human preferences and instructions. This iterative process refines the model's responses and enhances its usability in real-world applications.

1.3.5.3 Three scenarios are possible for using LLMs

a. Prompt Engineering

Prompt engineering is a critical technique in NLP that involves designing and optimizing prompts used to input information into models, aiming to enhance their performance on specific tasks. It focuses on designing prompts to guide model learning, enabling the generation of diversified text based on different contextual environments, and can be optimized and customized for various tasks and application scenarios. This innovative paradigm rooted in the development of LLMs has the potential to revolutionize the field of NLP by providing a more cost-effective and efficient means of training models, making it an increasingly important area of research[25] Prompt engineering can be categorized into two types (see figure 1.5): manual prompts and automated prompts.

- Manual prompts: include zero-shot prompting and few-shot prompting, relying on human expertise for manual configuration.
- Automated prompts: consist of discrete prompting and continuous prompting, designed using automatic algorithms.

Discrete prompts are human interpretable, while continuous prompting uses learning tokens interpretable by computers.

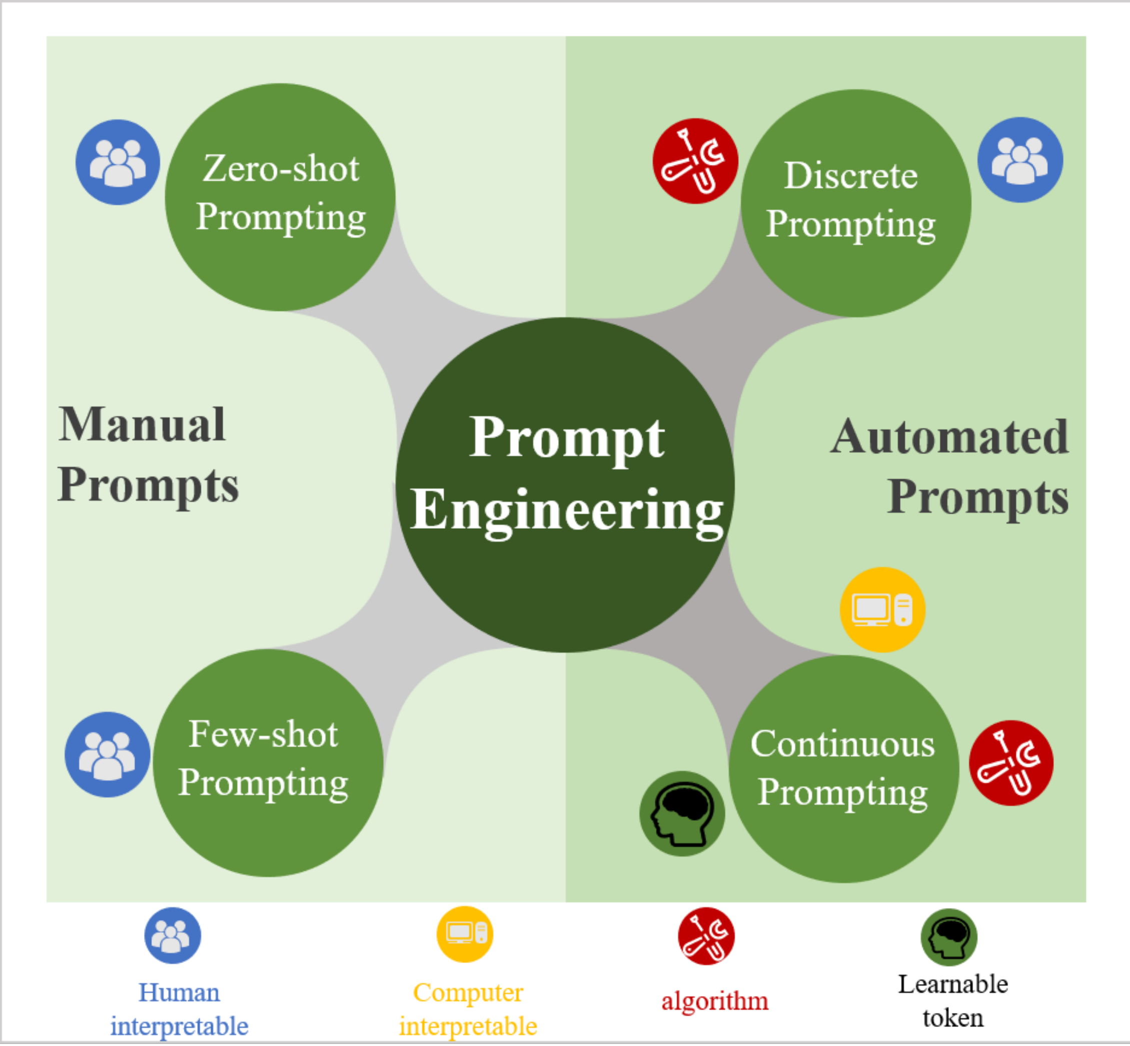

Figure 1.5: Prompt engineering [25]

b. RAG (Retrieval Augmented Generation)

Retrieval-Augmented Generation (RAG) is the process of optimizing the output of LLM by utilizing an external, reliable knowledge base in addition to the training data sources (see Figure 1.6). LLMs are trained on extensive datasets and employ billions of parameters to generate original outputs for tasks such as answering questions, translating languages, and completing sentences.

RAG extends the already formidable capabilities of LLMs to specific domains or an organization's internal knowledge base, all without the need to retrain the model. It is an economical approach to enhance LLM results, ensuring they remain consistent, accurate, and useful across various contexts. It aims to improve the quality of generated text by incorporating relevant information retrieved from a knowledge source before generating a response. This approach is often used in developing chatbots and conversational AI systems to provide more accurate and contextually relevant responses to user queries. RAG models typically consist of two components: a retriever, which retrieves relevant information from a knowledge source, and a generator, which generates text based on the retrieved information[26].

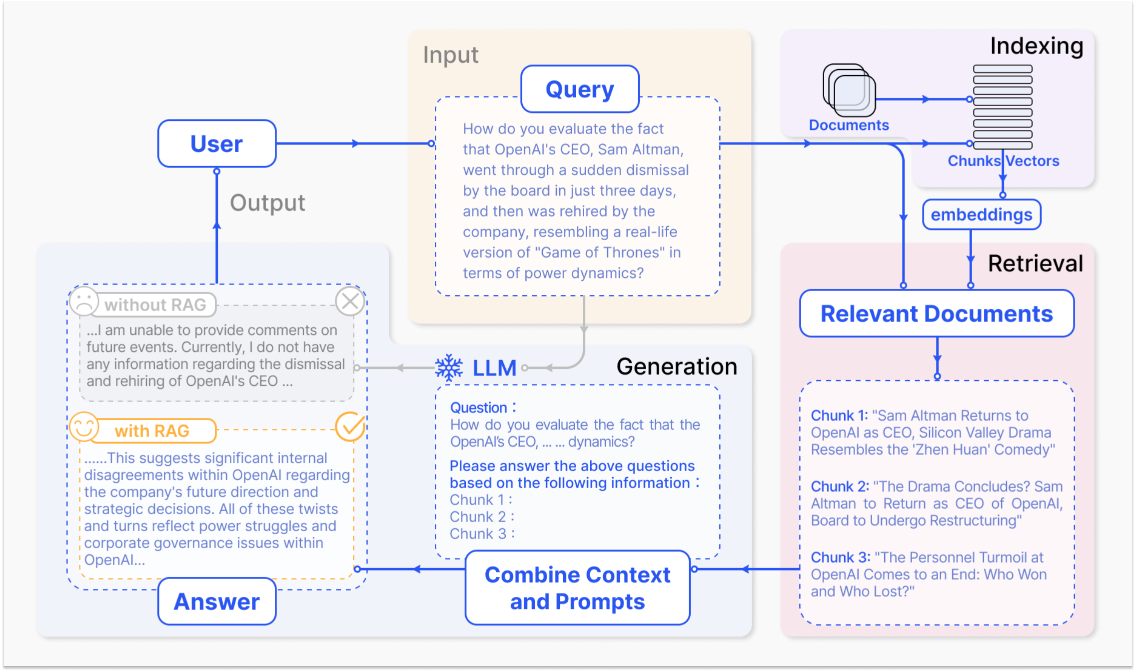

Figure 1.6: A representative instance of the RAG process applied to question answering[26]

A representative instance of the Retrieval-Augmented Generation (RAG) process applied to question answering consists mainly of three steps [26]:

1. **Indexing**
   - **Data Preparation:** Large amounts of text data (articles, code, manuals) are collected from various sources. The data is pre-processed to clean and standardize it, ensuring consistency for later steps.
   - **Chunking:** The data is divided into smaller, manageable segments called ”chunks.” These chunks can be sentences, paragraphs, or sections depending on the specific task.
   - **Encoding:** Each chunk is converted into a numerical representation using a technique like word embedding. This allows the system to understand the semantic meaning of the text.
   - **Vector Database Creation:** The encoded chunks and their corresponding metadata (source, identifier) are stored in a specialized database called a vector database. This database efficiently retrieves similar vectors based on semantic closeness.
2. **Retrieval**
   - **Query Formulation:** The user submits a question. The question may also be pre-processed to improve clarity and identify key terms.
   - **Semantic Matching:** The question is encoded using the same technique applied to the chunks during indexing. The encoded question is compared against the encoded chunks stored in the vector database.
   - **Top-k Retrieval:** A predefined number (k) of chunks with the highest

semantic similarity to the question are retrieved. These ”top k” chunks are considered the most relevant to answering the user’s query.

3. **Generation**
   - **LM Input Preparation:** The original question and the retrieved k chunks are combined and formatted appropriately for the Large Language Model (LLM). This formatting might involve highlighting key terms or structuring the information in a specific way.
   - **Answer Generation:** The LLM takes the combined input and uses its knowledge and understanding of language to generate a comprehensive and informative answer to the user’s question. The answer should be based on the retrieved chunks and address the specific context of the query.
   - **Answer Ranking (Optional):** In some cases, the LLM might generate multiple potential answers. A ranking step can be employed to select the most relevant and factually accurate answer based on additional criteria.

c. Fine-tuning

Fine-tuning in NLP involves adapting pre-trained models, like Llama, for specific tasks by further training them on new datasets. This process slightly adjusts the model’s weights to enhance its performance on the new task while retaining the pre-acquired knowledge. Fine-tuning improves task-specific performance efficiently without training the model from scratch, making it a practical approach for utilizing large pre-trained models in various NLP tasks [28].

1.4 Conclusion

In this chapter we briefly introduced the two fields, namely AI and NLP. LLMs like Llama 2 and Llama 2-Chat represent the pinnacle of AI, reshaping text generation with their advanced architectures. As chatbots evolve and LLMs push boundaries, the future promises seamless interactions and innovative solutions, enriching user experiences in the digital landscape.

Chapter 2

ChatBot : Overview and Related Works

2.1 Introduction

This chapter represents a thorough analysis of the subject of research chatbots. This review seeks to summarize the key conclusions, identify research gaps, and discuss how this study might advance the area.

2.2 ChatBot : Overview

2.2.1 What is a Chatbot?

A chatbot is a computer program or software application designed to simulate human conversation and engage in interactions with users through text or voice-based communication. The primary objective of a chatbot is to provide responses and assistance that are as natural and human-like as possible[11].

The functionality of a chatbot relies on two essential components: knowledge and understanding.

Knowledge: Chatbots are equipped with a knowledge base, which can be built using various methods. This knowledge base includes pre-programmed information, data from structured databases, or access to external sources of information. The more extensive and accurate the knowledge base, the better the chatbot can respond to user queries and provide relevant information.

Understanding: Chatbots use NLP and Machine Learning algorithms to comprehend and interpret user inputs. NLP allows the chatbot to extract meaning from text or voice messages, identify user intent, and generate appropriate responses. The chatbot continuously learns and improves its understanding through user interactions and feedback.

To evaluate the effectiveness of a chatbot, the Turing test is often used. Proposed by Alan Turing in 1950, the Turing test is a measure of a machine's ability to exhibit intelligent behavior indistinguishable from that of a human. If a chatbot can successfully pass the **Turing test**, it means it can provide responses and interactions so human-like that users cannot differentiate between chatting with the chatbot and a real person[12].

- A method for determining whether a computer is capable of thinking like a human being or not.
- One human functions as the questioner.
- The second human and the computer function as respondents.
- The questioner interrogates the respondents within a specific subject area.
- After a preset length of time or the number of questions, the questioner is then asked to decide which respondent was human and which was a computer.

Chatbots find applications in various fields, such as customer support, virtual assistants, information retrieval, and interactive entertainment. As advancements in AI and NLP continue, chatbots are becoming increasingly sophisticated, offering more seamless and natural conversations with users.

2.2.2 Evolution

From the realm of fiction to the realm of reality, chatbots boast a rich history that spans several decades. The concept of a machine emulating human behavior was first introduced through the Turing Test ("Can machines think?") proposed by Alan Turing in 1950. Subsequently, in 1966, Joseph Weizenbaum developed Eliza, the pioneering chatbot. Eliza relied on a rule-based system employing pattern matching to engage in simulated conversations resembling those with a therapist. [2] Although it used a simple pattern-matching algorithm and template-based responses, it lacked the conversational ability to truly mimic human-like interactions. Attempting to overcome Eliza's limitations, psychiatrist Kenneth Colby introduced PARRY, a chatbot with its personality.

Advancements in AI and the creation of the Artificial Intelligence Markup Language (AIML) led to the development of ALICE, the first chatbot to earn the title of the "most human computer". However, these early chatbots were constrained by their capabilities and primarily relied on basic rule-based systems. With the emergence of NLP and ML techniques in the 2000s, chatbots evolved into more sophisticated entities, possessing the ability to comprehend and respond to natural language queries. The internet era gave rise to chatbots like SmarterChild, accessible through messenger applications.

During the 2010s, chatbots like Apple Siri, Microsoft Cortana, Amazon Alexa, Google Assistant, and IBM Watson were introduced. Subsequently, the widespread application of chatbots led to the development of tools that enabled individuals with less technical expertise to create chatbots for their specific needs. The advancement of technologies such as NLP and ML further elevated the capabilities of chatbots. Recent research in reinforcement learning and neural networks introduced the concept of transformers in chatbot development. One of the most extensively used chatbots today for a myriad of purposes is ChatGPT, released in 2022. ChatGPT exhibits the ability to write and fix code, perform computations, help compile resumes, translate material, and perform numerous other tasks with sufficient proficiency to become a key tool in content development. This remarkable capability may one-day challenge search engines. Overall, a plethora of technological advancements have driven the progress of chatbots, creating increasingly intelligent and adaptable systems poised to revolutionize various industries[2].

As shown in Table 2.1, the evolution of chatbots is marked by significant milestones from the creation of Eliza to the development of ChatGPT.

Table 2.1: Evolution of Chatbot [2]

Sr.No.	Year	Chatbot	Remark
1	1965	ELIZA	One of the first chatbots, named Eliza, was created in the 1960s and employed a straightforward rule-based methodology to mimic a discussion between a therapist and a patient.
2	1972	PARRY	Parry's goal was to mimic a conversation with a patient who was paranoid and show how rule-based chatbots could be useful in the field of mental health
3	1988	Jabberwacky	Jabberwacky's goal was to develop a learning chatbot based on ML and NLP that could have engaging conversations with users.
4	1992	Dr. Sbaisto	With the use of a pre-recorded voice and straightforward responses, Dr. Sbaitso aimed to provide MS-DOS users with an easy-to-use and enjoyable conversational agent.
5	1995	ALICE	The goal of Alice was to develop a chatbot that could converse with people using NLP and ML and offer tailored responses based on the context and previous exchanges.
6	2001	SmaterChild	The goal of SmarterChild was to build an AI-powered chatbot that could engage in interesting discussions on a variety of messaging platforms, quickly respond to user questions, and access a huge collection of knowledge.
7	2010	Siri	The purpose of Siri was to create a virtual assistant that could perform tasks and answer questions for users using NLP, voice recognition, and a wide range of built-in functionalities.
8	2012	Google Now	Based on customers' search histories and other information, Google Now is an intelligent personal assistant that offers them support and information that is pertinent to their needs.
9	2015	Cortana	Microsoft created Cortana as a virtual assistant to aid users with a variety of tasks, such as creating reminders, sending emails, and providing answers.
10	2015	Alexa	Amazon has created Alexa, a virtual assistant that can respond to voice requests and carry out several functions like playing music, providing information, and operating smart home devices.
11	2016	Google Home	The Google Home smart speaker was created to give customers hands-free access to their smart home devices and a variety of services, including music streaming and voice-activated internet search.
12	2022	ChatGPT	A conversational AI language model called ChatGPT was created by OpenAI with the goal of understanding and producing text-based conversations that are human-like on a variety of subjects.

2.2.3 Chatbot characteristics

Understanding the key characteristics of a chatbot is crucial during its design phase. These attributes were identified by studying user expectations of chatbots . The research methodology employed included comparing past human-human conversations with interactions between humans and chatbots. As emphasized in a chatbot survey, the evolution of chatbots has shifted from simple pattern recognition and basic "question and answer" structures to more engaging conversations. This evolution highlights the expectation for advanced chatbots not only to provide answers but also to learn, improve with each interaction, and eventually demonstrate proficiency across various scenarios. To create an intelligent chatbot capable of achieving these capabilities, the essential characteristics and capabilities are outlined as follows[13]:

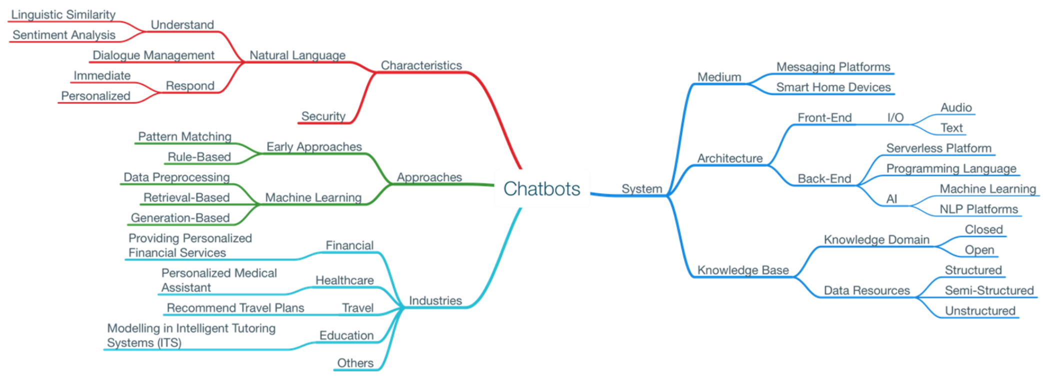

Figure 2.1: An overview of the properties of chatbots [13]

2.2.4 Design principles

A set of parameters for designing any chatbot are usually considered, including the purpose, dataset creation/procurement, response generation, and text processing. Advanced chatbots also include parameters like ML models and evaluation, which improve the accuracy and overall performance of the chatbot. It also makes them more dynamic and enables them to generate more personalized responses. The chatbot design/development process can be broadly viewed in six phases[2]:

1. **Purpose Identification:** In this phase, the main objective of the chatbot is determined, and different use cases for the design and development are identified.

2. **Dataset Creation/Procurement:** The collection or creation of the dataset used for training and improving the chatbot takes place. This dataset may include examples of expected questions and answers from users.

3. **Response Generation:** The chatbot needs a response generator capable of creating appropriate and logical responses to user questions Dl techniques and language models can be used to achieve this.

4. **Text Processing:** In this phase, NLP techniques are utilized to understand user inputs and extract meaning from them.

5. **Use of ML Models and Evaluation:** ML models and AI can be employed to enhance the chatbot's performance and increase its accuracy in interacting with users.

6. **Comprehensive and Dynamic Chatbot Development:** : This step aims to design a chatbot with high dynamism, capable of providing personalized and customized responses to meet user requirements.

   These stages represent a general framework for the process of designing and developing a chatbot, and the details and techniques used may vary depending on each case.

2.2.5 Chatbot Structure

Each chatbot requires a defined structure, which refers to the sequence of steps and tasks it performs in the backend. While structures may vary depending on the algorithm used, they generally align on three main steps.

Understanding the Query: This is the first crucial step where the chatbot processes the user's input or query. To achieve this, the chatbot leverages NLP techniques. NLP helps the chatbot analyze and comprehend the user's message, extract the intent, and identify key entities or keywords within the text. Understanding the query correctly is essential for the chatbot to provide relevant and accurate responses. For example:

- How to get the exact meaning .
- To understand the right ordre:
  - I m fine - am I fine .
  - France beat Germany - Germany beat France .
- Understand homonyms:
  - Like , where, right, execute .
- Understanding similar words:
  - Bye-by, right-write, steak-stick.
- How to find main keywords.
- Get rid of stopwords.
- Knowing when to focus on stopwords.
- Understand abbreviations.
- Understand sarcasm.

- Long phrases problems.
- Multi-lssue phrase.
- minding short memory issues.

Getting the Answer: Once the chatbot understands the user's query, it needs to retrieve the appropriate information to formulate a response. Depending on the chatbot's purpose, it may access a knowledge base, a database, or an external source of information to find the most relevant answer to the user's query. This step often involves using information retrieval techniques or querying data sources to obtain the required information. It consists of:

- Training the model
- Access to universal DBs
- Searching online
- Approaching closest answers
- To say: I don't know

Forming the Reply: After gathering the relevant information, the chatbot needs to craft a response that is coherent and appropriate for the user's query. This step involves generating the reply in natural language, ensuring it addresses the user's question or request accurately and understandably. The chatbot may utilize NLG techniques to create human-like responses.

By following these common steps of understanding the query, getting the answer, and forming the reply, chatbots can effectively engage in conversations with users, providing helpful and contextually appropriate responses. Different chatbot algorithms and frameworks may have additional layers and complexities in their structures, but the fundamental process of understanding, retrieving, and generating responses remains central to their operation.

2.2.6 Chatbot Tools

It is essential for chatbot developers to thoroughly study and comprehend key techniques from NLP that are integral for building effective chatbots. Neglecting or misinterpreting these tools could result in the creation of underperforming or unintelligent chatbots. We need the tools to do the following tasks[29]:

- Stemming
- Stop-Words
- Word Embedding
- Auto Correction

- Text Similarity
- Text Generation

2.2.7 Chatbot challenges

Several challenges can be considered[27]:

Dialect difference:

- Solved by train chatbot to all possible words in each dialect.
- Some chtbot start by classify the user dialect then depend on his dialect file.
- Dialect file be on line to be updated and edited by developers.
- Self-learning chatbots can teach it self.

Abbreviations:

- Developers need to list all available general and special abbreviations.
- Differs upon culture dialect community and time.
- Developers build a dictionary for each abbreviation. And its meaing.

Multi-topics:

- Can be detected using ML classifiers,depend on keywords existence.
- By comprehending the keywords chatbot can specify how it can reply.

Understanding homonyms:

- Each word meaning depend on context and current topic.
- Had to be solved to understand the phrase and form the answer.
- In case of equal probabilities, chatbot can ask the user.

Unknown words:

- Try auto-correct, as it might be miss-spelled word.
- Auto-correct depends on the topic: plase: plane, plate, please?
- Try languages lexicons.
- Ask the user.

2.3 ChatBot : Related Works

In [14], the authors utilized a web crawler to compile data for the knowledge base, extracting information from FAQ pages covering various university-related topics. They applied NLP techniques within the Preprocessor module to preprocess the data into a suitable format for the Training module. Two distinct algorithms were assessed to train the conversational model for the chatbot: a semantic similarity model and a seq2seq-based model. The system determines whether a query is domain-specific; if so, it directs domain-specific queries to the Preprocessor and chatbot Engine modules. The Engine module identifies the most relevant question to the user query and generates the corresponding response. Non-domain-specific queries are handled by the ALICE AIML server.

In [15]The authors utilizes Artificial Intelligence Markup Language (AIML) and Latent Semantic Analysis (LSA) to develop a chatbot tailored to provide efficient and accurate responses based on a dataset of frequently asked questions (FAQs). AIML handles template-based and general inquiries, such as welcome messages and common queries, while LSA manages service-related questions, ensuring timely and satisfactory responses. Within the AIML tag, multiple category tags represent distinct knowledge bases, organized using category and category tags. The pattern tag within AIML generates diverse responses, enhancing the bot's conversational dynamics. Additionally, the that tag enables the system to identify the most recent chatbot question, facilitating appropriate user interactions. Although the specific purpose of the topic tag within AIML is not explicitly outlined in the provided context, it likely contributes to organizing and categorizing responses for improved dialogue flow.

In [16]The authors proposes the use of AI and ML technology to develop a chatbot system for Matrusri Engineering College, which involves NLP and training the chatbot using appropriate ML methods. The program utilizes WordNet to select the closest matching response from the closest matching statement that matches the user input. The chatbot system is designed to communicate with users using natural language input and generate appropriate responses. The paper also mentions the use of a database created by a human expert to store the knowledge of the chatbot.

In [17]The proposed KBot system aims to enhance the end-to-end user experience by improving interactive question answering and performance, thus facilitating information retrieval, acquisition, intent classification, query understanding, and continuous learning. It enables users to explore myPersonality data through analytical queries (myPersonality is a social network dataset that has been processed and added to existing knowledge bases to extend the capabilities of the chatbot system) offering insights into various aspects like dominant political views, relationship status, and personality traits.

Designed to be scalable and flexible, KBot supports the addition of other knowledge bases, new languages, and diverse tasks, ensuring adaptability to different user needs. Leveraging ML and NLU techniques, including named entity recognition, factoid and recurrent questions, and dialogue management, enhances usability and performance. Integration of external APIs and support for multilingual and speech-to-text capabilities further augment the functionality and accessibility of the KBot system.

The paper employs NLP and NLU techniques to understand user intents and generate SPARQL queries. Named Entity Recognition (NER) is utilized to extract information about various entities like PERSON, GPE, ORG, etc., from user queries. Additionally, a machine learning-based classifier, specifically SVM (Support Vector Machine), is developed for intent classification. The KBot system integrates multiple knowledge bases such as DBpedia, Wikidata, and myPersonality to provide relevant answers to user queries.

DBpedia and **Wikidata** are utilized as knowledge bases in the chatbot system to retrieve relevant information. DBpedia is a community-driven project that extracts structured information from Wikipedia and presents it as linked data. On the other hand, Wikidata serves as a free and open knowledge base, providing structured data and acting as a central storage for the structured data of its Wikimedia sister projects.

In [18], a modular chatbot framework was implemented, comprising a web-based platform dedicated to training the chatbot in natural language comprehension. Additionally, a microservice was integrated into the framework to classify input text and extract relevant entities, facilitating more accurate responses.

The core of the chatbot's NLU engine was built upon (SVM) algorithms, which functioned as the primary classifier for text categorization. This approach enabled the chatbot to effectively analyze and interpret user queries, leading to more precise and contextually relevant interactions.

In [19], the proposed chatbot system utilizes AI techniques and NLP to recognize queries and provide answers to students. NLP techniques are employed for tokenizing and lemmatizing the data, which is crucial for text recognition. The model is trained using a combination of feed-forward ANN and the Keras Sequential Model. Preprocessing is applied to standardize the input text according to the system's requirements, and the appropriate context is recognized based on keywords used in the text. The model is trained on a dataset containing a mix of patterns and intents, consisting of 147 documents and 86 tags. Its accuracy is evaluated after 200 epochs, achieving an accuracy of 87.07%.

The study in [20] extensively utilized various transformer and semantic embedding models, such as AraBERT, CAMeLBERT, AraElectra-SQuAD, and AraElectra (Generator/Discriminator), to assess the performance of an Arabic chatbot. Two datasets were employed for evaluation purposes: one comprising 398 questions and the other containing 1395 questions, alongside 365,568 documents extracted from Arabic Wikipedia.

Through meticulous experimental procedures, involving the assessment of both manually curated questions and the entire question set using confidence and similarity metrics, the study demonstrated that the AraElectra-SQuAD model consistently surpassed other models, achieving notably high confidence and similarity scores across both datasets. Additionally, the research delved into the utilization of BERT-like language transformers, pre-trained on extensive Arabic text corpora, for developing an Arabic question-answering chatbot.

2.3.1 Summary of related works

Table 2.2: Summary of related works

Paper Name	Description	Observation
CollegeBot: A Conversational AI Approach to Help Students Navigate College [14]	Used a web crawler to gather data for the knowledge base, applied NLP techniques for data preprocessing, and evaluated two algorithms for training the conversational model.	Accuracy:80%
Chatbot for university related FAQs [15]	Employed AIML and LSA to develop a chatbot, with AIML managing general inquiries and LSA handling service-related questions.	
Smart College Chatbot Using ML and Python [16]	Leveraged AI, ML, and NLP technologies, utilized WordNet for response selection, and incorporated a human-curated database for chatbot knowledge.	
KBot: A Knowledge Graph Based Chatbot for Natural Language Understanding Over Linked Data [17]	Developed the KBot system using NLP and NLU techniques, integrated ML classifiers like SVM for intent classification, and utilized multiple knowledge bases for relevant answers.	Accuracy:85%
Development of an e-commerce Sales Chatbot [18]	Implemented a modular chatbot framework with SVM algorithms for text categorization and a microservice for entity extraction from user input.	
An Interactive Chatbot for College Enquiry [19]	Utilized AI, NLP, and ANN models for recognizing queries and providing accurate responses, trained on mixed data patterns, and evaluated accuracy metrics.	Accuracy:87.07%
Evaluation of an Arabic Chatbot Based on Extractive Question-Answering Transfer Learning and Language Transformers [20]	Utilized transformer and semantic embedding models for an Arabic chatbot, evaluated on Arabic Wikipedia datasets using confidence and similarity metrics, and explored BERT-like transformers for question-answering tasks.	

2.4 Conclusion

This chapter provides a comprehensive exploration of chatbots, covering their definition, evolution, characteristics, design principles, tools, challenges, and the landscape of development. It highlights key advancements, challenges, and opportunities, emphasizing the need for continuous improvement and adaptation to meet evolving user needs and contribute to the ongoing evolution of chatbot technology.

Chapter 3

ChatBot for student needs: Design and Development

3.1 Introduction

In this chapter, we explore the design and development of a chatbot specifically tailored to meet the diverse needs of students in the Kasdi Merbah Ouargla university environment. We relied on a range of tools and techniques in artificial intelligence and machine learning to develop this chatbot, enabling us to create a prototype that can effectively interact with users, understand their queries, and provide appropriate responses. We will summarize the main findings we have reached, review the challenges we faced during the research and development journey, and provide some future recommendations for improving and expanding the system.

3.2 Types of Information and Support Students Seek

1. Academic Information and Support
2. Exam Details: Information about exam dates, locations, and formats.
3. Study Resources: Access to study materials, past exams, and academic workshops:
   - They look for guidance subjects and learning materials.
   - Students seek information on specific university services.
   - Educational systems for teaching, learning, and searching for specific information.
   - Personalized learning support, instant answers to questions, and guidance in understanding research content.
   - Information related to research methodologies, process, instruments, population, sample group, variables, data collection, data analysis, and research report.

3.3 Analysis of Student Needs

Situations for which a chatbot can be useful

1. Chatbots can be useful in emergency situations to provide initial support to students and facilitate communication with universities.
2. Chatbots can act as virtual assistants or tutors for students, freeing up teachers from responding to repetitive questions.
3. Answering Academic and Administrative Questions:
   - Academic Questions: Assisting students in understanding complex concepts, providing additional study resources, and answering queries related to the curriculum.

- Administrative Questions: Providing information about exam dates, course registration details, and graduation procedures.

4. Providing Information About University Events:

    - Informing students about upcoming events, seminars, workshops, and other university activities.
    - Registering students for these events and sending reminders prior to the event date.

5. Supporting New Students:

    - Offering information about university resources such as libraries, health centers, and student centers.
    - Answering queries from new students about orientation and adapting to university life.

3.4 Student Expectations and Preferences for Interacting with a Chatbot

- Preferred Communication Mode: Text vs. Voice: Determine whether students prefer to interact through text messages or voice commands.

- Ease of use and user interface:

    – Simplicity: Evaluate preferences for a straightforward and easy-to-use interface that does not require learning new formats or complex commands designed to communicate with humans in their natural language, enhancing user experience and accessibility.

- The ability to operate as a 24/24support service, chatbots can provide timely and efficient assistance or information to users

- Instant responses to frequently asked questions, making it easier for students to obtain immediate answers.

- Quick response times and maintaining privacy while interacting with chatbots

- Students expect chatbots to provide efficient and accurate answers for university-related questions.

- Students seek chatbots that offer guidance, information retrieval, and a positive digital learning experience.

3.5 Chatbot Features

In the context of chatbot development, we refer to the methods, tools, and technical software used to build and effectively operate a chatbot. This includes software frameworks, libraries, and tools that assist in data analysis, processing, and user interaction. Common technologies used in chatbot development include:

- **LLM (Large Language Models):** (see chapter 1)
- **Natural Language Processing (NLP):** This fundamental technology enables the chatbot to understand human language in a natural manner.
- **Chatbot Development Platforms:** These platforms provide the necessary tools for building and operating chatbots:
  - **Hugging Face:** A platform that provides access to advanced AI models.
    * **Usage:** we utilized the LLaMA 2 language model and integrated it into the chatbot to leverage its capabilities in natural language processing and text generation based on trained interactions.
  - **Kaggle:** is a prominent platform for data science and machine learning, offering tools for model development and deployment.
    * **Usage:** Kaggle provides cloud-based Jupyter Notebooks with pre-installed libraries and GPU/TPU support, simplifying model prototyping and training.
  - **Chainlit:** To facilitate user interaction with our Lama model, we employed Chainlit, an open-source Python framework designed specifically for Conversational AI. This streamlined interface development and enabled rapid prototyping, expediting the creation of an interactive user experience.

3.6 Chatbot Design and Development

We followed a structured approach, encompassing concept ideation, training the model, scripting interactions, and designing the user interface, to create a user-friendly and efficient chatbot system.

3.6.1 Use of LLMs

Three scenarios are possible for using LLMs:

- **Prompt engineering :** is the process of crafting specialized prompts to guide the behavior of LLMs.
- **Fine tuning:** is the process of adapting a pretrained LLM to specific datasets or domains.

- **RAG (Retrieval Augmented Generation):** A technology that combines information retrieval and text generation to improve the quality of the system's final answers. In our case we integrated with LLaMA 2 to enhance the chatbot's ability to provide accurate and detailed informational answers. RAG allows for the retrieval of relevant information from the dataset before generating a response.

In our case, we used Llama 2 model with the three possible scenarios of using LLMs.

Model details :

- **Model:** Llama 2

- **Variation:** 7b-chat-hf (7b: 7B. hf: HuggingFace build)

- **Version:** V1

LlaMA 2 model is pretrained and fine-tuned with 2 Trillion tokens and 7 to 70 Billion parameters which makes it one of the powerful open source models. It is a highly improvement over LlaMA 1 model.

3.6.2 Necessary Tools

We will use Python to write our script, it is an open-source programming language known for its ease of reading and writing. It supports numerous libraries specialized in scientific computing, data processing, and machine learning. It is widely used in software development, particularly in artificial intelligence applications.

3.7 Using RAG technique

3.7.1 Data collection

In RaG technology, a dataset comprising nine university documents was utilized. the Table 3.1 representing the Kasdi Merbah Ouargla university documents with their sizes and related information:

Table 3.1: Dataset

Document Name	**Size**	**Format**
Canevas d'amendement OFFRE DE FORMATION L.M.D. LICENCE ACADEMIQUE	639Ko	PDF
HARMONISATION OFFRE DE FORMATION MASTER ACADEMIQUE Master Informatique Fondamentale 2016 – 2017	1.4 Mo	PDF
HARMONISATION OFFRE DE FORMATION MASTER PROFESSIONNALISANT Master Administration et Sécurité des Réseaux(ASR) 2016 – 2017	1.2 Mo	PDF
OFFRE DE FORMATION MASTER ACADEMIQUE Intelligence Artificielle et Science des Données2021 -2022	1.2 Mo	PDF
HARMONISATION OFFRE DE FORMATION MASTER ACADEMIQUE Master Informatique Industrielle (II) 2016 - 2017	1.3 Mo	PDF
example-description-invention-relative-product	408 Ko	PDF
Mechanisms-implementation-of-decision-1275 ; 2022-2023	990 Ko	PDF
Guide -project-obtaining-certificate-patent; 2022	12.2 Mo	DOCX

3.7.2 RAG Process

The RAG process is as fellow (see Figure 3.1):

3.7.2.1 Setting Up the Environment and Installing Libraries

Libraries Used

1. **pypdf:**[1] A library for processing PDF files.
2. **llama_index:**[2] A library for building a vector-based index to improve search and retrieval operations.
3. **sentence_transformers:**[3] A library for text transformation models used to convert sentences into vector representations.
4. **pydantic-settings:**[4] A library for managing application settings.
5. **langchain:**[5] A library for integrating language models and enhancing the performance of language models.

[1] https://pypdf.readthedocs.io/en/stable/
[2] https://docs.llamaindex.ai/en/stable/
[3] https://sbert.net/
[4] https://docs.pydantic.dev/latest/concepts/pydantic_settings/
[5] https://python.langchain.com/v0.2/docs/introduction/

6. **arabic-reshaper:**[6] A library for improving the display of Arabic texts.
7. **docx2txt:**[7] A library for extracting text from DOCX files.
8. **huggingface_hub:**[8] A library for accessing Hugging Face models.

- ## Document Loading:

  Relevant Kasdi Merbah Ouargla university program documents are read from a specific directory and loaded into the system.

- ## Language Model Settings:

  The language model from Hugging Face (Llama-2-7b-chat-hf) is defined for text analysis and response generation. Preparing the prompts for the model with a system prompt and a query wrapper prompt guiding the chatbot's responses.

- ## Embedding Model Setup:

  We use sentence transformers to create embeddings that convert texts into vectors. These vectors are numerical representations of textual data, facilitating comparison and retrieval.

- ## Context Service Creation:

  To connect the language models and embedding models, allowing them to be used integrally with the loaded data.

- ## Index Creation:

  A data structure that stores the vectors resulting from the embedding process, enabling quick search and retrieval.

- ## Query Engine Setup:

  With the query engine, we can now submit a question and receive an answer.

- ## Answer Generation:

  After building the index, it is converted into a query engine that can be used for search operations in the stored vectors. It retrieves relevant documents from the index and uses the Llama-2 model to generate an accurate and detailed response.

[6] https://pypi.org/project/arabic-reshaper/
[7] https://pypi.org/project/docx2txt/
[8] https://huggingface.co/docs/huggingface_hub/index

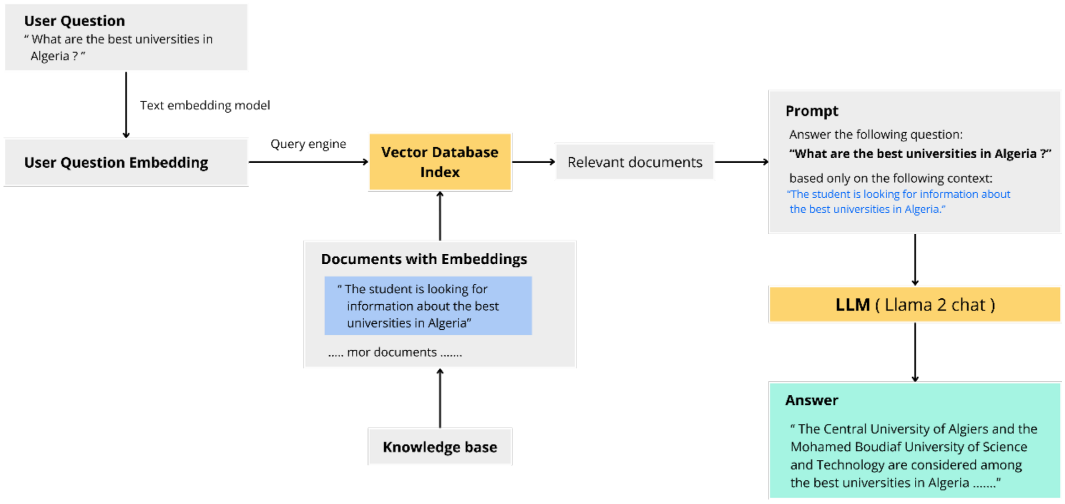

Figure 3.1: Diagram Explained RAG Process

- ### Sample Responses

  To demonstrate the effectiveness of the university chatbot in providing accurate and informative responses, below are examples of queries along with corresponding model-generated responses:

Query 1:

{ هل هناك نصائح عامة للطالب ؟}

Response 1:

"نعم، وهناك نصائح عامة للطالب في مواد دراسية، التخصصات الجامعية، والحياة الجامعية. للمعرفة عن كل نصائح وكيف يمكنك تطبيقها، يمكنك استعادة من الملخص أدوات ونصائح عامة للطالب."

Note: The assistant's answer is based on the information provided in the context and may not be comprehensive or accurate.

3.8 Fine-Tuning Process

- ### Data Collection:

  The first step is gathering information so that the chatbot's knowledge base is formed. The Gemini program was used to gather data for this project in order to generate a series of questions and responses. This information is necessary to train the model so that it can comprehend and appropriately reply to user inquiries. Examples :

Table 3.2: Data Collection Informations

id	question	context	answer	explanation	related_information
1	ما هي أفضل الجامعات في الجزائر؟	يبحث الطالب عن معلومات حول أفضل الجامعات في الجزائر.	تعتبر جامعة الجزائر المركزية وجامعة العلوم والتكنولوجيا محمد بوضياف من أفضل الجامعات في الجزائر. تتميز هاتان الجامعتان ببرامجها األكاديمية القوية وأعضاء هيئة التدريس المتميزين وفرص البحث المتنوعة.	تشتهر جامعة الجزائر المركزية بتاريخها العريق وسمعتها المرموقة، بينما تُعرف جامعة العلوم والتكنولوجيا محمد بوضياف بتخصصاتها العلمية والتكنولوجية المتقدمة.	موقع وزارة التعليم العالي والبحث العلمي الجزائري
2	كيف يمكنني التغلب على التوتر قبل ااالمتحانات؟	يشعر الطالب بالتوتر والقلق قبل ااالمتحانات.	هناك العديد من الطرق للتغلب على التوتر قبل ااالمتحانات، مثل: * مراجعة المواد بشكل جيد. * الحصول على قسط كا ف من النوم. * ممارسة ًالرياضة بانتظام. * تجنب الكافيين والمواد المنبهة األخرى. * التحدث إلى صديق أو أحد أفراد األسرة عن مخاوفك. * ممارسة تقنيات ااالسترخاء مثل اليوغا أو التأمل.	التوتر قبل ااالمتحانات أمر طبيعي، ولكن يمكن السيطرة عليه من خالل اتباع بعض النصائح البسيطة. من المهم أن تبدأ في مراجعة المواد مبكًرا وتضع خطة زمنية للدراسة. تأكد أي ًضا من الحصول على قسط كا ف من النوم وتناول نظاًما غذائي صحي. يمكن أن تساعدك ممارسة الرياضة بانتظام وتقنيات ااالسترخاء على تقليل التوتر والقلق.	نصائح للتغلب على التوتر قبل ااالمتحانات
3	أين يمكنني العثور على سكن جامعي في الجزائر؟	يبحث الطالب عن معلومات حول سكن جامعي في الجزائر	تتوفر العديد من الخيارات لسكن الجامعة في الجزائر، بما في ذلك: * سكنات الجامعة: توفر معظم الجامعات الجزائرية سكنات للطالب بأسعار معقولة. * السكن الخاص: يمكنك أي ًضا العثور على سكن خاص بالقرب من الجامعة. * مشاركة الشقة: يمكنك مشاركة شقة مع طالب آخرين لتقليل التكاليف.	يعتمد أفضل خيار لسكن الجامعة على احتياجاتك وميزانيتك. إذا كنت تبحث عن خيار رخيص، فإن سكنات الجامعة هي الخيار األفضل. إذا كنت تبحث عن المزيد من الخصوصية، يمكنك استئجار سكن خاص أو مشاركة شقة مع طالب آخرين	موقع وزارة التعليم العالي والبحث العلمي الجزائري

4	ما هي المنح الدراسية المتاحة للطالب الجزائريين؟	يبحث الطالب عن معلومات حول المنح الدراسية المتاحة للطالب الجزائريين	تتوفر العديد من المنح الدراسية للطالب الجزائريين، بما في ذلك: * المنح الدراسية الحكومية: تقدم الحكومة الجزائرية العديد من المنح الدراسية للطالب الجزائريين للدراسة في الجزائر أو في الخارج. * المنح الدراسية الدولية: تقدم العديد من الدول األجنبية منًحا دراسية للطالب الجزائريين للدراسة في جامعاتها. * المنح الدراسية الخاصة: تقدم العديد من المؤسسات الخاصة منًحا دراسية للطالب الجزائريين المتفوقين.	تختلف متطلبات المنح الدراسية اعتمادًا على المنحة المحددة. تأكد من مراجعة معايير األهلية ومتطلبات التقديم بعناية قبل التقديم للحصول على منحة دراسية.	موقع وزارة التعليم العالي والبحث العلمي الجزائري
5	صعب علي فهم المحاضرات في الجامعة، ماذا أفعل؟ (I find it difficult to understand lectures at university, what should I do?)	يواجه الطالب صعوبة في فهم محاضرات الجامعة.	فهمك للمحاضرات الجامعية: * قبل المحاضرة، قم بقراءة المادة مسبقاً سيساعدك هذا على فهم المصطلحات والمفاهيم األساسية التي سيتم مناقشتها في المحاضرة. * خذ مالحظات واضحة أثناء المحاضرة. ركز على النقاط الرئيسية والتعريفات المهمة. * ال تتردد في طرح األسئلة على المحاضر أو المعيد إذا لم تفهم شيئًا. * بعد المحاضرة، راجع مالحظاتك وأعد قراءة المادة. يمكنك أيًضا مناقشة المحاضرة مع زمالئك في الدراسة.	من الطبيعي أن تواجه صعوبة في فهم بعض المحاضرات الجامعية. المهم هو أن تكون استباقيًا وأن تطلب المساعدة عند الحاجة.	نصائح لتحسين فهم المحاضرات الجامعية

- **Data Preparation:**

  After the data is gathered, it needs to be formatted so that the machine learning model can use it. In this instance, the data was transformed into JSON format, which is commonly utilized for data sharing.

- **Embedding:**

  The process of converting textual data into numerical vectors that the model can understand is called embedding. In order for the model to comprehend and learn from the text data, this phase is essential.

- **Train-Test Split:**

  Training and testing subsets of the dataset are used to assess the model's performance. The testing set is used to assess the model's performance after it has been trained using the training set. Divide the dataset into training (90%) and validation (10%) sets.

- **Model Loading:**

  The Hugging Face platform loads the Llama-2 7b-chat model. Using the supplied dataset as a starting point, this pre-trained model can be further refined.

- **Model Training:**

  Using the training dataset, the pre-trained Llama-2 7b-chat model is refined during model training. This procedure modifies the model to fit the particular realm of inquiries pertaining to universities.

  Important Actions:

  - **Set Up Arguments and Training Parameters:** Setting up arguments and training parameters is an essential step in correctly adjusting the model's performance. These parameters comprise configurations that affect the model's learning process: the batch size=2, learning rate=5e-5, and number of epochs=10.
  - **Apply a Tokenization:** A tokenizer is a tool that divides text into more manageable chunks for the model to analyze, like words or tokens. To guarantee that the model understands the text data correctly, it is imperative to use an efficient tokenizer.
  - **Hire a Trainer to Oversee the Instructional Procedure:** The instrument in charge of managing the complete training procedure is the trainer. It ensures the model is efficiently optimized by using the given parameters and controlling the training cycle.

- **Query Processing** After training, the model can process user inquiries and produce relevant answers. This entails understanding the input query, using the model to identify pertinent responses, and providing the result.

- **Response Generation** Using the trained model, the final step entails producing answers to user inquiries. This guarantees that users will obtain relevant and reliable information regarding their university-related inquiries. The following diagram illustrates the elements explained and links each phase to the next in the fine-tuning process (see Figure 3.2) :

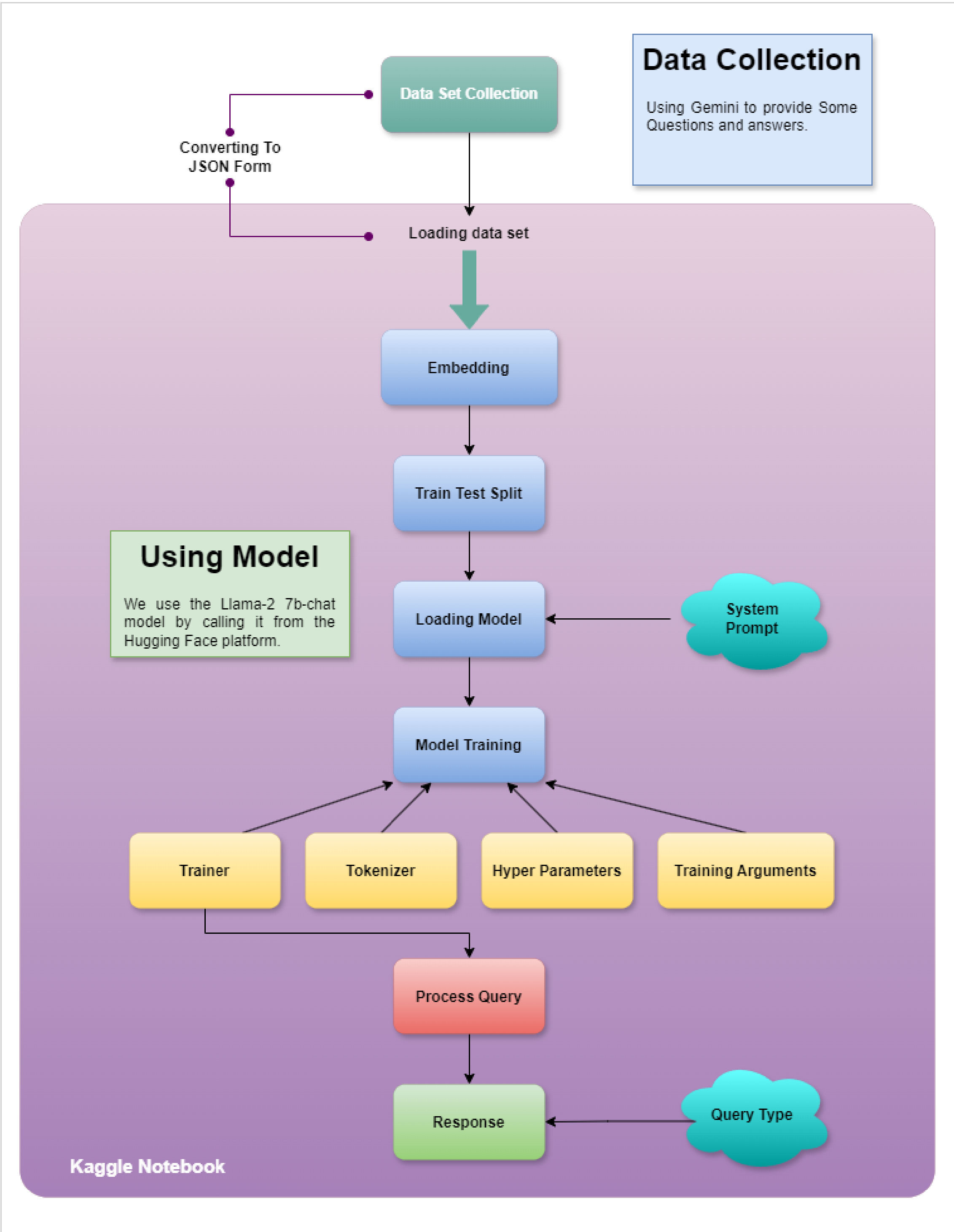

Figure 3.2: Diagram Explained Fine-tuning Process

- **Sample Responses**:

  To demonstrate the effectiveness of the university chatbot in providing accurate and informative responses, below are examples of queries along with corresponding model-generated responses:

Query 1:

" هل هناك نصائح عامة للطالب ؟"

Response 1:

"أعمل على تطوير أماكن تخلص من الهواجسة.

(Work on developing coping mechanisms for stress).

أبحث عن مهارات مناسبة للحفاظ على صحتك.

(Find suitable hobbies to take care of your health).

أستخدم زمن التخلص في المسار من الدراسة.

(Allocate time for relaxation and personal growth).

أعمل على ترفيه النجاح.

(Work on developing a growth mindset).

هذا هو نصائح عامة للطالب في الجامعة.

. (These are some general tips for students in university for the Jazairy| lifestyle.)"

3.9 Comparison of Llama-2 Model Answers Using RAG and Fine-Tuning Techniques

The purpose of this comparison is to evaluate the performance of the Llama-2 model using two different techniques: Retrieval-Augmented Generation (RAG) and Fine-Tuning. This analysis aims to assess the accuracy, relevance, completeness, and overall quality of the model's responses to student queries.

We conducted this preliminary evaluation by generating a set of representative questions and comparing the model answers using the two distinct techniques: RAG and Fine-Tuning.

By comparing the responses generated by these two techniques, we aimed to identify which method produced more accurate, relevant, complete, and high-quality answers. This preliminary evaluation provided us with insights into the trengths and weaknesses of each technique and helped us determine the most effective approach for our chatbot's development.

The results of this comparison are presented in the following sections, detailing the methodology, comparison table, and analysis of findings. Although we could not involve students in this phase, the insights gained from this preliminary evaluation are valuable for the continued improvement and refinement of our Kasdi Merbah Ouargla university chat system.

3.9.1 Methodology

The comparison was conducted using a set of predefined queries related to university services and requirements. The responses generated by the Llama-2 model using RAG and Fine-Tuning were evaluated based on the following criteria:

Accuracy: How correct and precise the answer is.

Relevance: The relevance of the answer to the query.

Completeness: Whether the answer covers all aspects of the query.

Quality: The overall quality of the response.

3.9.2 Results

Below is a table summarizing the comparison of responses from the Llama-2 model using RAG and Fine-Tuning techniques:

Table 3.3: Comparison Table of Llama-2 Model Answers Using RAG and Fine-Tuning Techniques.

Query	RAG Technique Answer	Fine-Tuning Technique Answer	Accuracy	Relevance	Completeness	Quality
ماهي التخصصات الموجودة في جامعة ورقلة؟	تخصصات الماستر في جامعة ورقلة هي: *تخصصاتالجامعة (takhassasat al-jami'a): كتب ومهندسين في مختلف المجالات العلمية، مثل الهندسة (الحاسوب، الميكانيكية)، الطب، الصيدلة، الحقوق، إدارة الأعمال، التربية، الآداب (اللغة الإنجليزية، اللغة العربية، التاريخ). *مواد الدراسة (ad- muwad dirasa): كتب ومهندسين فيمختلف المجالات العلمية	1. Computer Science 2. Electrical Engineering 3. Mechanical Engineering 4. Materials Science and Engineering 5. Chemical Engineering 6. Civil Engineering 7. Environmental Engineering 8. Geological Engineering 9. Mathematics 10. Physics	Moderately Accurate	Moderately Relevant	Mostly Complete	Moderate Quality
هل توفر الجامعة الجزائرية سكن جامعي للطلاب؟	the University of Algiers offers dormitories for students.	توفر الجامعة سكن جامعي مناسب للطلاب الجديدين. كل طالب من جزائريعاني من نهاية التعلم عن صعوبة إيجاد سكن جامعي. ، توفر الجامعة سكن جامعي مناسب للطلاب الجديدين.	Partially Accurate	Partially Relevant	Partially Complete	Low Quality

هل من الممكن العمل أثناء الدراسة الجامعية؟	وهو ممكن للطلاب للتعلم والتميز في المجالاتالمختلفة. ويمكن الطلاب أن يعملوا في المناقشات الدراسية والمناقشات التعليمية والمناقشات المهنية والمناقشات الاقتصادية. ويمكنهم أيضا ً أن يعملوا في المشاركة في المهنيات الدراسية والمهنيات التعليمية	In this situation, you are asking about the possibility of working while studying in university, using the Algerian dialect. Here's a possible response: (It's possible to work while studying in university, but it depends on the student's circumstances, Algerian dialect.)	Highly Accurate	Highly Relevant	Complete	High Quality
ما هي صعوبة العيش في الإقامة الجامعية في الجامعة الجزائرية؟	The difficulty of living in university dormitories in Algeria is mainly due to the lack of privacy and personal space. Students often share small rooms with several roommates, which can be challenging for those who value their privacy and independence. Additionally, the dormitories may not provide adequate security and safety measures, which can be a concern for students who are away from their families for the first time. Furthermore, the dormitories may not have the necessary facilities and amenities to ensure a comfortable living environment, such as cleaning services, laundry facilities, and recreational spaces. Finally, the dormitories may have strict rules and regulations that can be difficult to follow, which can be frustrating for students who value their freedom and autonomy.	the difficulties of living in a university dorm in Algeria may include: 1. Limited privacy 2. Lack of home comforts 3. Rules and regulations 4. Language barriers 5. Cultural differences 6. Limited security 7. Limited amenities 8. Noise and distractions 9. Limited storage space 10. Homesickness	Highly Accurate	Highly Relevant	Complete	High Quality

3.9.3 Analysis

- **Accuracy Analysis:** The Fine-Tuning technique generally provided more accurate responses, particularly for queries requiring detailed information.
- **Relevance Analysis:** Both techniques performed well in maintaining relevance to the queries, with Fine-Tuning slightly ahead due to its tailored responses.
- **Completeness Analysis:** Fine-Tuning often resulted in more complete answers, addressing all aspects of the queries more thoroughly than RAG.
- **Quality Analysis:** Overall, the quality of responses from Fine-Tuning was rated higher due to its comprehensiveness and detail. o **Case Studies: Fine-Tuning**

Example Query: ماهي التخصصات الجامعية؟

RAG Answer: تخصصات الماستر في جامعة ورقلة هي ....

Answer: The specializations of the Master's ...

In this example, the Fine-Tuning technique provided a more detailed and specific answer, which was rated higher in terms of completeness and quality.

3.9.4 Discussion of Findings

The Fine-Tuning technique generally provided more accurate and complete answers, likely due to its ability to learn specific patterns and details from the training data. RAG, while faster, sometimes lacked the depth and specificity required for more complex queries.

3.9.5 Conclusions and Recommendations

In conclusion, Fine-Tuning demonstrated superior performance in terms of accuracy and completeness. For future improvements, we recommend further optimizing the Fine-Tuning process and conducting more extensive user testing to refine the chatbot's capabilities.

3.10 Validation Description

In our validation of the university chat, our initial plan was to introduce the chat system to students and assess its effectiveness by collecting their questions and analyzing the corresponding answers provided by the chat. This would have allowed us to gather valuable feedback from real users and evaluate the chat's performance in a practical, real-world setting. However, due to time constraints, we were unable to carry out this comprehensive evaluation with the students.

3.11 Challenges

- Resource Management: Training a large model like LLaMA 2 requires significant computational resources, so Google Colab was used to leverage GPUs.
- Obtaining sufficient data: The challenge: the available data was not sufficient to train an accurate and robust model.
- Answering in Arabic: Most available models are primarily trained on English texts, limiting their ability to handle Arabic texts effectively.
- Integration with Existing Systems: The challenge: integrating the chatbot with the university's existing systems.
- Model Testing and Improvement: The challenge: ensuring that the model provides accurate and useful answers. Issues such as providing inaccurate or unexpected answers may arise.

3.12 Conclusion

In this chapter, we explored the complex design and development of a chatbot specifically tailored to meet the diverse needs of students in the university environment. We find that the prototype we developed showed somewhat acceptable results in meeting students' needs and interacting with them effectively. However, there is considerable room for improvement by adopting new training strategies, enhancing the models used, and expanding the knowledge base it relies on. This will take time. Additionally, the chatbot's performance can be improved by enhancing its integration with other university systems, which will contribute to providing more comprehensive and accurate services.

General Conclusion

Thanks to advancements in NLP and AI , chatbots are now capable of providing smooth and effective interactive experiences, thereby enhancing the educational process and supporting students in innovative ways. By exploring various aspects of NLP, we have found that these technologies aid in understanding, analyzing, and generating human language, allowing chatbots to interact more effectively and accurately with users.

We also examined LLMs, which represent some of the most advanced technologies in AI and NLP. These models enable chatbots to understand texts with a high level of accuracy and generate natural responses, as seen with Llama 2. We covered an overview of chatbots, their evolution, characteristics, design principles, architecture, tools used in their development, and the challenges they face.

In this project, we focused on enhancing the user experience by designing a chatbot based on RAG and Fine-Tuning techniques. These techniques allowed the chatbot to provide accurate and immediate responses while maintaining privacy and adapting to changing data. This dynamic approach ensures the chatbot remains up-to-date and effective, contributing to a better educational experience for students.

We hope this thesis has provided valuable insights and a deep understanding of NLP and AI technologies and their crucial role in developing university chatbots. We look forward to this chatbot being a starting point for further development and innovation in this field, ultimately contributing to a better and more integrated educational experience for students in the future.

Bibliography

[1] A. S. Lokman et M. A. Ameedeen, « Modern Chatbot Systems: A Technical Review », in Proceedings of the Future Technologies Conference (FTC) 2018, K. Arai, R. Bhatia, et S. Kapoor, Éd., Cham: Springer International Publishing, 2019, p. 1012-1023. Doi: 10.1007/978-3-030-02683-7_75.

[2] Ishita Shah, Shrihari Jhawar, Aashna Khater, Asher Jacob, et Dr. Girish Potdar, « Chatbot Development Through the Ages: A Survey », IJSRCSEIT, p. 262-271, mai 2023, Doi: 10.32628/CSEIT2390329.

[3] A. Ullah, G. Qi, S. Hussain, I. Ullah, et Z. Ali, « The Role of LLMs in Sustainable Smart Cities: Applications, Challenges, and Future Directions ». arXiv, 7 février 2024. Doi: 10.48550/arXiv.2402.14596.

[4] Kretschmer, Martin, Thomas Margoni, and Pinar Oruc. "Copyright Law and the Lifecycle of Machine Learning Models." IIC-International Review of Intellectual Property and Competition Law (2024): 1-29.

[5] Dongare, A. D., R. R. Kharde, and Amit D. Kachare. "Introduction to artificial neural network." International Journal of Engineering and Innovative Technology (IJEIT) 2.1 (2012): 189-194.

[6] Speech and Language Processing. Daniel Jurafsky & James H. Martin. Copyright Transformers and Large Language Models© 2023. Allrights reserved. Draft of January 5, 2024.

[7] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[8] « What Are Large Language Models (LLMs)? | IBM ». Consulté le: 10 juin 2024. [En ligne]. Available in: https://www.ibm.com/topics/large-language-models

[9] S. Minaee et al., « Large Language Models: A Survey ». arXiv, 20 février 2024. Doi: 10.48550/arXiv.2402.06196.

[10] H. Touvron et al., « Llama 2: Open Foundation and Fine-Tuned Chat Models ». arXiv, 19 juillet 2023. Doi: 10.48550/arXiv.2307.09288.

[11] By Jack Cahn Thesis Advisor: Dr. Boon Thau Loo Engineering Advisor: Dr. Jean Gallier https://www.academia.edu/37082899/CHATBOT_Architecture_Design_and_Deveopment

[12] Chen, J., Wang, Y., Shen, Y., & Xu, H. (2020). Natural Language Understanding for Chatbots: A Survey. Frontiers of Information Technology & Electronic Engineering, 21(3), 340-353.

[13] P. Suta, X. Lan, B. Wu, P. Mongkolnam, et J. H. Chan, « An Overview of Machine Learning in Chatbots », IJMERR, p. 502-510, 2020, Doi: 10.18178/ijmerr.9.4.502-510.

[14] Daswani, M., Desai, K., Patel, M., Vani, R., & Eirinaki, M. (2020). CollegeBot: A conversational AI approach to help students navigate college. In HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22 (pp. 44-63). Springer International Publishing.

[15] B. R. Ranoliya, N. Raghuwanshi, et S. Singh, « Chatbot for university related FAQs », in 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), sept. 2017, p. 1525-1530. doi: 10.1109/ICACCI.2017.8126057.

[16] H. K. K., A. K. Palakurthi, V. Putnala, et A. Kumar K., « Smart College Chatbot using ML and Python », in 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), juill. 2020, p. 1-5. doi: 10.1109/ICSCAN49426.2020.9262426.

[17] M. M. Khan, « Development of An e-commerce Sales Chatbot », in 2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET), déc. 2020, p. 173-176. doi: 10.1109/HONET50430.2020.9322667.

[18] A. Ait-Mlouk et L. Jiang, « KBot: A Knowledge Graph Based ChatBot for Natural Language Understanding Over Linked Data », IEEE Access, vol. 8, p. 149220-149230, 2020, doi: 10.1109/ACCESS.2020.3016142.

[19] W. Hassan et A. A. Elsayed, « An Interactive Chatbot for College Enquiry », Journal of Computing and Communication, vol. 2, n 1, p. 20-28, janv. 2023, doi: 10.21608/jocc.2023.282081.

[20] Alruqi, T. N., & Alzahrani, S. M. (2023). Evaluation of an Arabic chatbot based on extractive question-answering transfer learning and language transformers. AI, 4(3), 667-691.

[21] What is natural language processing (NLP)?; https://www.ibm.com/topics/natural-language-processing,(consulted19/05/2024).

[22] Inés Roldós, Major Challenges of Natural Language Processing (NLP); https://i2group.com/articles/the-10-biggest-issues-facing-natural-language-processing, (consulted 19/05/2024).

[23] Shaip, What is NLP? How it Works, Benefits, Challenges, Examples, https://www.fastsimon.com/natural-language-processing-uses-challenges/, (consulted 19/05/2024).

[24] Sebaitre, Zakaria. (2019). Object Detection & Tracking Through Deep Learning in the Autonomous Driving Context. 10.13140/RG.2.2.18180.55684.

[25] J. Wang et al., « Prompt Engineering for Healthcare: Methodologies and Applications ». arXiv, 23 mars 2024. doi: 10.48550/arXiv.2304.14670.

[26] Y. Gao et al., « Retrieval-Augmented Generation for Large Language Models: A Survey ». arXiv, 27 mars 2024. doi: 10.48550/arXiv.2312.10997.

[27] Shaip, What is NLP? How it Works, Benefits, Challenges, Examples,https://www.fastsimon.com/natural-language-processing-uses-challenges/, (consulted 19/05/2024).

[28] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, et N. Smith, « Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping ». arXiv, 14 février 2020. doi: 10.48550/arXiv.2002.06305.

[29] Pavel Surmenok, Natural Language Pipeline for Chatbots, https://medium.com/@surmenok/natural-language-pipeline-for-chatbots-897bda41482, Nov 6, 2016, (consulted 19/05/2024).