



UNIVERSITY OF WEST ATTICA



---

FACULTY OF ENGINEERING  
Department of Electrical & Electronics Engineering  
Department of Industrial Design & Production Engineering  
MSc in Intelligence and Deep Learning

## **Master of Science Thesis**

**(Development and Evaluation of AI-Powered Educational Assistant)**

**Student: PAPAZOGLOU, ALEXANDER Registration Number: mscaidl-0048**

**MSc Thesis Supervisor: LELIGKOU, ELENIAIKATERINI**

**Professor**

**ATHENS-EGALEO, SEPTEMBER,2024**



This MSc Thesis has been accepted, evaluated and graded by the following committee:

Supervisor	Member	Member
ELENI AIKATERINI LELIGKOU	CHARALAMBOS PATRIKAKIS	Dr. DIMITRIOS KOGIAS
Professor	Professor	Lecturer
Engineering School	Engineering School	Engineering School
University of West Attica	University of West Attica	University of West Attica

**Copyright ©** All rights reserved.

**University of West Attica and (Name and Surname of the student) Month,  
Year**

You may not copy, reproduce or distribute this work (or any part of it) for commercial purposes. Copying/reprinting, storage and distribution for any non-profit educational or research purposes are allowed under the conditions of referring to the original source and of reproducing the present copyright note. Any inquiries relevant to the use of this thesis for profit/commercial purposes must be addressed to the author.

The opinions and the conclusions included in this document express solely the author and do not express the opinion of the MSc thesis supervisor or the examination committee or the formal position of the Department(s) or the University of West Attica.

### **Declaration of the author of this MSc thesis**

I, Papazoglou Alexander, Stylianos, with the following student registration number: mscaidl0048, postgraduate student of the MSc program in “Artificial Intelligence and Deep Learning”, which is organized by the Department of Electrical and Electronic Engineering and the Department of Industrial Design and Production Engineering of the Faculty of Engineering of the University of West Attica, hereby declare that:

I am the author of this MSc thesis and any help I may have received is clearly mentioned in the thesis. Additionally, all the sources I have used (e.g., to extract data, ideas, words or phrases) are cited with full reference to the corresponding authors, the publishing house or the journal; this also applies to the Internet sources that I have used. I also confirm that I have personally written this thesis and the intellectual property rights belong to myself and to the University of West Attica. This work has not been submitted for any other degree or professional qualification except as specified in it.

Any violations of my academic responsibilities, as stated above, constitutes substantial reason for the cancellation of the conferred MSc degree.

I wish to deny access to the full text of my MSc thesis until 30/11/2024, following my application to the Library of UNIWA and the approval from my supervisor.

The author

ALEXANDER PAPAZOGLOU



ΑΛΕΞΑΝΔΡΟΣ ΠΑΠΑΖΟΓΛΟΥ



This work is dedicated to my parents!



I would like to express my deepest gratitude to my thesis advisor, Leligou Aikaterini, for their invaluable guidance, support, and encouragement throughout this project. Your expertise and insights have been instrumental in shaping this work.

I am also thankful to my colleagues and peers who provided thoughtful feedback and helped refine my ideas during the development process.

Finally, I would like to thank my family and friends for their unwavering support and understanding during this journey. Your encouragement has been a source of strength throughout this endeavour.

## **Abstract**

In order to improve student learning in college courses, this thesis investigates the creation and application of J.A.R.V.I.S. (Just An Resourceful Virtual Instructor and Study-aid), an AI-powered instructional assistant. Through the use of sophisticated Natural Language Processing (NLP) methods, such as refining models like as LLaMA-3 and incorporating pre-made solutions like Livechat AI, J.A.R.V.I.S. offers accurate, context-sensitive support in response to student enquiries.

The research is organised around two main strategies: the first is optimising the LLaMA-3 model using Alpaca and Unsloth to produce a highly specialised teaching assistant that can produce responses that are rich in context and information. The second method uses Livechat AI to incorporate the J.A.R.V.I.S. chatbot into a specially designed website, with scalability, userfriendliness, and real-time communication as top priorities. The system's performance was assessed using two main criteria: its feasibility in real-world deployment and its capacity to manage intricate educational issues.

The expected result is a notable enhancement in the understanding and involvement of students, since J.A.R.V.I.S. effectively provides precise and pertinent answers customised to meet individual learning requirements. By showcasing the use of cutting-edge natural language processing techniques to improve learning outcomes and providing insights into the advantages and practical difficulties of implementing artificial intelligence in educational settings, this study advances the area of educational technology.

## **Keywords**

LLM, Educational AI, Chatbot, Transformers, NLP, Livechat-AI

## Table of Contents

<b>CHAPTER I: Introduction.....</b>	<b>10</b>
<b>1.1 Research Background.....</b>	<b>10</b>
<b>1.2 Research Objectives.....</b>	<b>10</b>
<b>1.3 Structure .....</b>	<b>11</b>
<b>CHAPTER II: Literature Review &amp; Foundational Knowledge .....</b>	<b>12</b>
<b>2.1 History of Natural Language Processing ( NLP) .....</b>	<b>12</b>
<b>2.2 Recent Advances in Natural Language Understanding .....</b>	<b>12</b>
2.2.1 Challenges and Ethical Considerations .....	14
2.2.2 Future Directions in Natural Language Processing (NLP).....	14
2.2.3 The Transformer Architecture .....	15
2.2.4 LLMs and Their Effects .....	15
2.2.5 AI Marketplaces and Model Integration .....	16
2.2.5.1 Rise of AI Marketplaces.....	16
2.2.6 Advancements in Natural Language Processing (NLP).....	16
2.2.7 Artificial Intelligence and Computational Linguistics Implications.....	17
2.2.8 NLP Technology and Its Applications .....	17
2.2.9 Challenges and Future Directions .....	17
2.2.10 Fine-Tuning with Alpaca.....	18
2.2.11 Speed Optimization with Unsloth .....	18
<b>2.3 NLP in Educational Technology.....</b>	<b>20</b>
2.3.1 The Role of NLP in Personalized Learning.....	20
2.3.2 NLP for Education Challenges .....	22
2.3.3 Case Studies of NLP in Education.....	22
<b>2.4 Contextual Assistance and Adaptive Learning.....</b>	<b>23</b>
2.4.1 Adapting Learning Systems.....	23
2.4.2 Feedback Mechanisms.....	24
<b>2.5 Empowering Content Creation and Curation .....</b>	<b>26</b>
<b>2.6 Ethical Considerations And Challenges .....</b>	<b>26</b>
<b>2.7 Commercial AI Solutions for Education .....</b>	<b>27</b>
2.7.1 Overview of Livechat AI .....	27
2.7.2 Application in Educational Settings.....	28
2.7.3 EDU NLP Conclusion.....	28
2.8 LLaMA-3 & Livechat-AI.....	29
<b>CHAPTER III: System Architecture and Deployment .....</b>	<b>32</b>
<b>3.1 First Approach: Fine-Tuning LLaMA-3 with Alpaca and Unsloth .....</b>	<b>32</b>
3.1.1 System Overview.....	32
3.1.2 Fine-Tuning Process .....	32
3.1.3 Technical Challenges and Solutions .....	36



3.1.4 Results and Evaluation .....	37
3.1.5 Summary of the First Approach .....	37
<b>3.2 Second Approach: Livechat AI Integration/ Custom Website build .....</b>	<b>37</b>
3.2.1 System Overview .....	38
3.2.2 Creating the website .....	38
3.2.3 Advantages in Technology.....	41
3.2.4 Summary of Second Approach.....	42
<b>4 CHAPTER IV: Demonstrations and Comparative Analysis.....</b>	<b>43</b>
<b>4.1 Introduction .....</b>	<b>43</b>
<b>4.2 First Approach: Fine-Tuned LLaMa-3 .....</b>	<b>43</b>
4.2.1 Code/Model Functionality .....	43
<b>4.3 Second Approach: Livechat AI Integration.....</b>	<b>45</b>
4.3.1 J.A.R.V.I.S Chatbot on the Website.....	45
4.3.2 Technical Analysis of Livechat AI .....	47
<b>4.4 Comparative Commentary .....</b>	<b>47</b>
<b>4.5 Summary .....</b>	<b>48</b>
<b>5 CHAPTER V: Conclusion and Recommendations.....</b>	<b>49</b>
<b>5.1 Summary of Findings.....</b>	<b>49</b>
<b>5.2 Recommendations .....</b>	<b>49</b>
<b>5.3 Future Work.....</b>	<b>50</b>
<b>5.4 Final Thoughts.....</b>	<b>50</b>
<b>Bibliography-References-Online sources .....</b>	<b>51</b>



Figure 1: Alpaca: A Strong, Replicable Instruction-Following Model .....	18
Figure 2: Unsloth: Easy finetuning for AI and LLMs .....	19
Figure 3: Educational NLP .....	20
Figure 4: Benefits of ED. NLP .....	21
Figure 5: Contextual AI .....	23
Figure 6: Feedback Loop .....	24
Figure 7: Cycle between NLP and feedback in Intelligent Tutoring Systems .....	24
Figure 8: Livechat AI .....	27
Figure 9: Installing essential packages .....	31
Figure 10 : Model loading and Configuration .....	32
Figure 11: Data Prep .....	33
Figure 12: Training the Model .....	34
Figure 13: General Home Page .....	37
Figure 14: Learning Material .....	37
Figure 15: JARVIS Section .....	38
Figure 16: Example 1 LLaMa .....	40
Figure 17: Example 2 : LLaMa .....	41
Figure 18: Example 3 LLaMa .....	42
Figure 19 : Explaining Word Embeddings .....	43

## CHAPTER I: Introduction

### 1.1 Research Background

Recently, artificial intelligence (AI) and natural language processing (NLP) have been evolving very fast, creating new opportunities in EdTech, which focuses more on student involvement and personalization. To cater for the many diverse learning needs, more and more learners are opting for AI-powered educational assistants. These assistants can provide personalized materials, instant responses, and interesting ways of learning. There are several AI breakthroughs such as Google's BERT, Open AI's range of models, known as the GPT series, and Meta's LLaMA, that have tremendously contributed towards the development of engaging AI-powered educational tools. Similarly, Livechat AI has also shown massive potential in changing the perspective of education as it can give learners individual context efficiently and on a scale.

This thesis presents the whole process of creating the AI-based instructional aide power system, by first trying out LLaMA-3 fine tuning and ending up with the Livechat AI integration. LLaMA-3 was tested because it could provide the sought-after depth of context in every anticipated response. On the other hand, Livechat AI was picked because it had great potential in terms of scalability and easy integration, making it possible to provide timely assistance to learners. The two approaches were put to the test the basis of their feasibility.

### 1.2 Research Objectives

The following objectives have been formulated in this thesis:

- **Consider Usage of LLaMA-3 as a Direct Instructional Assistant:** This research aims to assess through the fine-tuning of LLaMA-3 whether the model is well-equipped to engage with pedagogical context and give a pertinent and detailed response.
- **Assessing the Use of LivechatAI as an Educational Chatbot:** the given objective examines how efficient, how scalable and how good user experience does LivechatAI provide as an educational application.

- **Specific assessment and recommendations:** Looking at the two in a more holistic approach, it will most certainly reveal the merits that each one has and thus helping us find the best fit for our educational scenarios.

### 1.3 Structure

The structure is as follows:

- **Chapter 1** consists of the objectives, the research background and the structure in general.
- **Chapter 2** dives into some basic NLP tasks, compares various tools for the complete achievement of tasks and justifies the choice of LLaMA-3 and Livechat AI.
- **Chapter 3** gives an account of the features, structures and implementation procedures of the two approaches, as well as the rationale for their choice and not any others.
- **Chapter 4** undertakes a comparative analysis with the description of two techniques in which one focuses one metric – performance and scalability amid its educational appropriate applicability.
- **In the last chapter**, a summary is made, which provides a view on applying AI in education as well as possible areas of focus for future studies.

## CHAPTER II: Literature Review & Foundational Knowledge

This chapter critically examines the studies which are relevant to the technologies and methodologies applied in this thesis. It contains the improvement of Natural Language Processing (NLP) with a special focus on models such as LLaMA-3, Alpaca, Unsloth and actual use of commercial AI tools such as Livechat AI in educational interactive classes.

### 2.1 History of Natural Language Processing ( NLP)

The area of Natural Language Processing (NLP) has grown much in the last decade, most notably because of the great interest in automation and model-based developments. With these breakthroughs, NLP is now rewriting how machines understand and engage with human language—imperatives in providing a backbone for tools like automated translation services and AI-driven conversational platforms. Drawing heavily from the paradigms of machine learning and deep learning, NLP now takes on tasks way beyond basic question-answering or translation—its focus areas being NLU, where machines interpret written or spoken input, and NLG, which enables them to generate human-like responses in context. [1] With the exponential rise in NLP applications, from sophisticated chatbots to sentiment analysis, personal medicine, and many more, our way of interacting with technology is changing. Long-standing challenges persist in this field: limited AI hardware infrastructure, lack of high-quality training data, complex linguistic issues such as understanding homonyms, or generating polysemy. [2]

### 2.2 Recent Advances in Natural Language Understanding

The advancements in Natural Language Processing (NLP) in 2024 can be divided into three key subsections, each highlighting a specific area of development:

- ✦ **Natural Language Generation and Automated Content Creation:** NLG, a part of NLP, has become a key technology to create automated content. NLG turns structured data into natural human-sounding text. More and more organizations use it to produce news stories financial reports, and other content types. This not only streamlines the content creation process but also ensures consistency and accuracy, proving beneficial in journalism, finance, and other data-intensive sectors.
  
- ✦ **Named Entity Recognition and Data Classification:** The role of Named Entity Recognition, or NER, in 2024 has been far more impactful. NER systems are adept at classifying and annotating diverse data parameters in unstructured data, like identifying person names, organizations, dates, and numerical values. With such an advancement NLP facilitates more efficient data extraction workflows, which ends up enhancing data processing and analysis across various industries.
  
- ✦ **LLMs integration in Complex NLP Tasks:** The LLMs integration has been one of the biggest changes in dealing with complex NLP tasks. Such models, with a core of advanced machine-learning algorithms, are increasing the capability of systems to understand and manipulate human language more accurately and with more context. The volumes of unstructured data have grown, driven by LLMs, with an increasing rate, hence fostering the growing importance of NLP in customer service, marketing, and data analytics. This trend points out that applications of NLP are going to be much subtler and sophisticated moving away from the traditional approaches towards deeper learning and understanding of human communication patterns.[3][4]

### 2.2.1 Challenges and Ethical Considerations

Despite these advances, the NLP field is beset by serious challenges, especially regarding the computational resources required to train large-scale models like BERT and T5. The amount of computational power needed for pre-training and fine-tuning of these models becomes a barrier to accessibility, especially for smaller organizations and research groups. Another regular source of debate must be the ethical considerations of these models, which include issues of biases within the training data and even environmental impact due to the training of large models. Furthermore, the dynamic nature of NLP technology has meant that current researchers are working on the limitations of pretraining on static corpora by continuously adapting models to dynamic language use and incorporating knowledge from external sources. There is also active research in new transformer variants, attention mechanisms, and model architectures for better efficiency and performance of NLP tasks.[5]

### 2.2.2 Future Directions in Natural Language Processing (NLP)

Natural language processing (NLP) is expected to undergo profound changes in the coming years due to developments in semantic and cognitive technologies. These advances are expected to improve human-like speech and text comprehension and enable more intuitive and intelligent applications across a range of domains. The incorporation of more complex NLP techniques, such as linguistics, semantics, statistics, and machine learning, is necessary for machines to understand the nuances of human communication, including not only individual words but also the context and nuances of language. [3][6]

- One of the most central areas of NLP is chatbots, in which the ultimate objective is to develop fast, intelligent, and friendly platforms. The prowess of chatbots is to understand and respond to longer-form, more complex questions in a diversity of circumstances and live is key to their future. That means that to completely comprehend the meaning of human language, NLP must be combined with other cognitive technologies.

- Another new concept in NLP is that of an invisible or zero interface. Such direct communication between humans and machines will be done through NLP and its ability to process and respond to various modes of human language input: text, voice, or hybrid. The latter mode is critical to applications centered on direct interaction between humans and machines, as with the case of Amazon's Echo.
- Smart search capabilities represent a major avenue for the continuation of the development of NLP. In our search functions, NLP usage now often takes a more conversational shape where one can speak with search engines as one would in a normal conversation. Second, the shift from the keyword based to conversational search, as in the case for integration of NLP into google drive, such that you can make more natural language queries to your file in google drive.

The last capability of NLP which promises is its ability to extract intelligence from unstructured information. The ability to extract meaningful insight from very large volumes of text, especially from very complex documents such as annual reports, legal and compliance documents is critical when NLP is able to discern the subtleties of text. [3]

### 2.2.3 The Transformer Architecture

By taking advantage of self attention mechanisms, the RNNs in the model introduced by Vaswani et al. in 2017 break the normality of NLP systems in the sense that a transformer model represents a radical change within the framework. It is easy to interpret the context in a given natural language because the design of the transformer architecture enforces better integration of information amongst words contained in each sentence irrespective of the location of the words in each sentence. The transformer employs an encoder decoder architecture—encoder takes input sequences and decoder generates output sequences. With no global structure, transformers rely on a type of self attention mechanism, which weights the importance of a word against another to produce better language understanding and generation. [7][8]

### 2.2.4 LLMs and Their Effects

Consequently, irrespective of where LLMs are applied (anywhere else), one of their primary impacts is in NLP, where a new state of the art model is here almost monthly. GPT-3 from OpenAI is the largest (175 billion parameters) and most such generator of coherent, contextually relevant, engaging text we know of thus far. If it is left to its own devices it will fill pages with

human like writing. [7] [9] [10]. A newer addition from Meta AI's would be the LLaMA-3, which aims to bridge the gap between the great performance of models like GPT-3, while also being more computationally consumed. However, the LLaMA-3 models are especially well suited to the educational environment, and they also excel at creative tasks. [9][11]

## **2.2.5 AI Marketplaces and Model Integration**

### **2.2.5.1 Rise of AI Marketplaces**

Natural Language Processing (NLP) has contributed significantly to the growth of AI Marketplaces. These platforms provide access to prebuilt AI models for various NLP tasks, such as sentiment analysis and entity recognition. This expansion is driven by the enhanced flexibility and integration capabilities of large language models (LLMs). [12]

In recent times, Artificial Intelligence (AI) has changed more rapidly, and it has for developers, become easier and possible to build solutions using pre-prepared models. It is about solving many natural language processing tasks: natural language understanding, natural language generation and speech recognition. Special deep learning models are built with huge training amounts of data for the purpose of high accuracy in the tasks like semantic analysis, word disambiguation, and part of speech tagging. [3][12]

In general, the whole idea of AI era within AI marketplaces enables all the NLP and other AI affiliated industries to embrace it. It fast enables organizations to work with existing models instead of creating them from scratch. It seems to be good for applications dealing with human language, and for processing mined models, reusing them, and feeding in functional extensions.[12]

## **2.2.6 Advancements in Natural Language Processing (NLP)**

The Natural Language Processing (NLP) sector has made incredible strides with the help of deep learning models and complex language models, especially in language generation and understanding. We have gone from highly specific architectures to more generalized and agile models which have been shown in the transition from BERT to T5. We have bettered the machines' proficiency in understanding and processing human languages , thus also improving a bunch of the NLP tasks like sentiment analysis, machine translation, entity recognition example.



### 2.2.7 Artificial Intelligence and Computational Linguistics Implications

With great progress made in natural language processing (NLP), artificial intelligence and computational linguistics have been greatly influenced as well. Recent development and application of models such as GPT-3 and its successors demonstrate potential of large pre-trained language models use in many domains from human language generation to chatbots and intelligent decision-making. These advancements have enhanced our capacity to process and analyse textual data while also fostering a greater understanding of human communication patterns.

### 2.2.8 NLP Technology and Its Applications

Previously, NLP technology has been integrated into many applications that change the world today. As an example, intelligent OTTs are shifting to be able not only to perform interaction, but interaction that actually looks and seems natural to humans, with a focus mainly in customer service interactions. NLP today is one indispensable machine intelligence product which provides analysis and necessary data. Consequently, the most recent developments, especially of voice user interfaces and multilingual natural language processing make it possible for people with different linguistic and cultural backgrounds to communicate naturally. [3][12]

Furthermore, considering how its use can vary from Transformers /Attention mechanisms all the way to Information Retrieval and Search Engines is nothing short of incredible. NLP's services span industries, from healthcare (e.g. diagnosing diseases via medical texts) to entertainment (e.g. personalizing movie recommendations) to law enforcement (e.g. analysing crime reports). What makes it one of the most exciting and impactful fields in AI is its ability to bridge the gap between human language and machine understanding. [10] [13]

### 2.2.9 Challenges and Future Directions

Despite the advancements achieved, challenges such as ethical issues, the high computational cost necessitated by the training of large models, and the requirement for neutrality in information processing do remain. Going forward, the trends will probably be to make NLP models that are more resource efficient (time, space, and financial), ethical, and accessible. The evolution of AI and NLP will not only continue but the evolution will start to incorporate genai (generative ai) which will extend the powerful GenAI NLP and disrupt the way we will work with machines. Fine-tuning is another key component when it comes to adapting LLMs to a specific domain. [3] [12]

### 2.2.10 Fine-Tuning with Alpaca

The Alpaca instruction-tuning method by Stanford AI Lab utilizes domain-specific datasets to enhance the domain-specific task performance of LLMs like LLaMA-3. [14]

# Stanford Alpaca



Figure 1: Alpaca: A Strong, Replicable Instruction-Following Model

### 2.2.11 Speed Optimization with Unsloth

After fine-tuning, it is time for speed optimization with Unsloth, which prepares the model for deployment in a production environment. Unsloth is a cost-efficient model reduction method that maintains both the efficiency and accuracy of heavy language models. [9]



**Figure 2: Unsloth: Easy finetuning for AI and LLMs**

Unsloth focusses on determining the most efficient type of computation in the model and minimising the number of active parameters during inference. This helps the model become more responsive to requests, as the required inputs and outputs require less processing. This optimisation is especially significant in educational settings where quick response times are required to keep students engaged. [9]

## 2.3 NLP in Educational Technology

The introduction of natural language processing (NLP) into educational technology has recently received increased research interest, with a particular emphasis on the construction of AI-Powered Educational Assistant systems. These technologies seek to help students study by providing targeted, on-demand support. [15] [16]

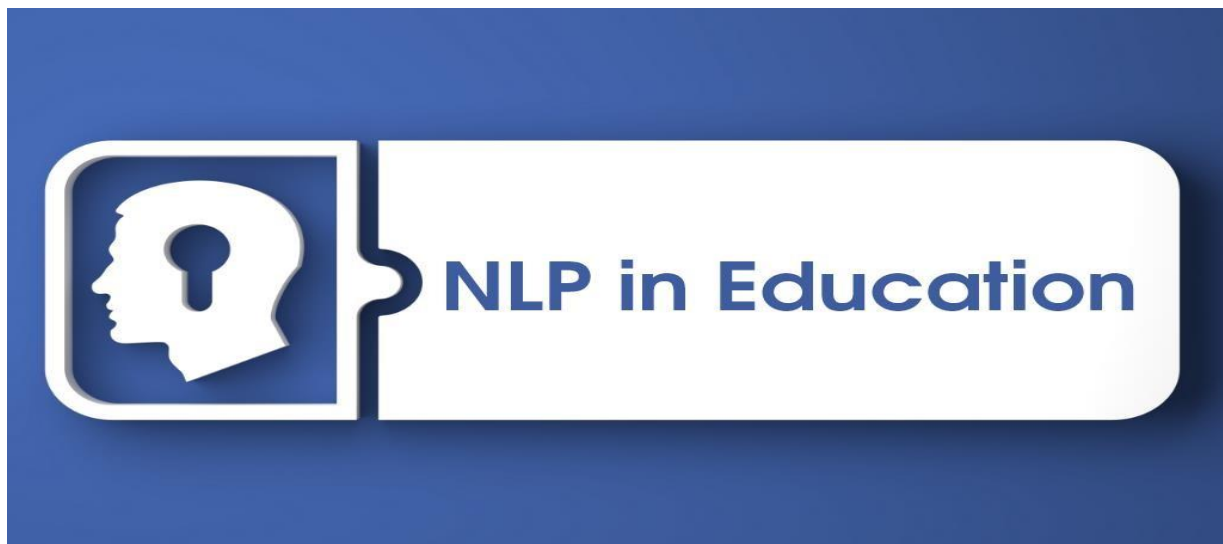


Figure 3: Educational NLP

### 2.3.1 The Role of NLP in Personalized Learning

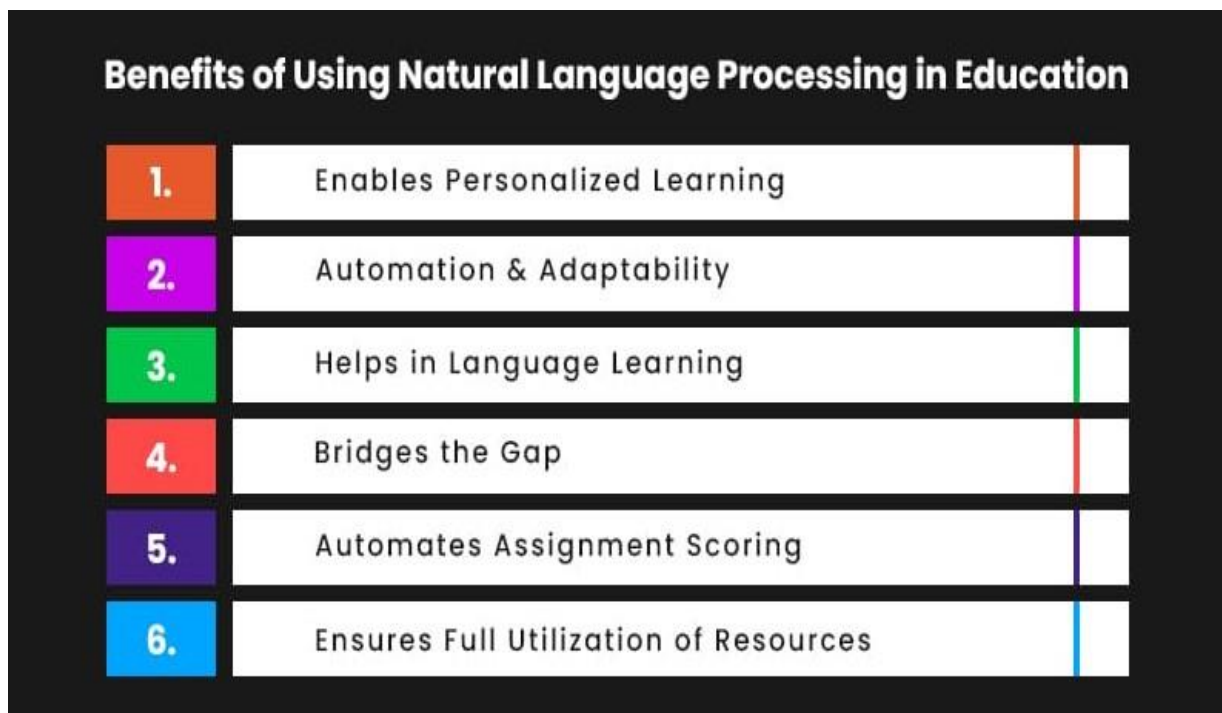
It is called a personalised learning strategy if the curriculum basis is created based on the needs and preferences of the students involved. Because it's able to understand and respond with contextually relevant information when it encounters a student's question, NLP is essential to personalised learning. [17]

For example, a NLP driven system can grade a student's question to understand what level of comprehension is their current level and provide responses appropriate for them. In addition to

closing those learning gaps, this technique not only closes specific learning gaps, but it also promotes ongoing learning by responding to the needs of the student. [12][18]

Among the top 4 most famous ways that NLP is used in Education are:

- Automated Grading and Feedback
- Intelligent Tutoring Systems (ITS)
- Chatbots for Student Support
- Personalized Learning and Adaptive Systems



**Figure 4: Benefits of ED. NLP**

It's essential though, to recognize that the organised education era is over. Thus, due to NLP, education will be tailored for every learner to face and practice, in order to meet his or her needs. There is matchmaking based on learners' performance, desired subjects, and learning modalities. This was followed by a rise in interest and understanding. When students are given content that

is appropriate for their abilities, they are more likely to receive and retain the information. This personalised approach also develops a sense of empowerment and ownership, which motivates and drives dedication. This customisation creates an individual learning path for each person. Consider always having a knowledgeable companion by your side to clear up confusion, explain complex concepts, and guide you through difficult situations. This goal is realised by NLP-driven conversational agents, sometimes known as chatbots. These online coaches communicate in a realistic manner while offering rapid guidance. We can clearly see that chatbots have significantly improved when it comes to real-time support. Just like students do with a real advisor, ask questions and get custom-to their need's answers, the same thing is possible with the chatbots. That way, NLP (chatbots) can enhance their communicative / problem managing skills while also motivating them to learn more. [12]

### 2.3.2 NLP for Education: Challenges

Although it seems very promising, there are still a lot of issues when it comes to Educational NLP. Getting the system to understand a wide range of questions is a big one, especially when those questions require complex domain specific terminology or are poorly worded. [17] Getting these models into existing educational platforms is another challenge. Large models like LLaMA-3 are hard to deploy in resource constrained environments because of their high resource requirements. And as such this is how Unsloth and other tools step into play, because finding middle ground when it comes to operational efficiency and model complexity is a top priority. [9] [12] [19]

### 2.3.3 Case Studies of NLP in Education

NLP use cases have been spotted all around in education. A good example is iTalk2Learn, which by enhancing instant feedback towards the students, manages to show how language acquisition can be uplifted by NLP. [4] Furthermore there are some great AI platforms like Quizlet, which, depending on the topic we choose to feed into the NLP, can produce questions and clarifications.[17][20] These examples show the benefits of NLP in creating dynamic and adaptive learning environments. But they also highlight how important it is to develop and use the models carefully so that the AI gives students accurate and helpful advice. [2] [21] With the worthy mentions stated above, it is clear that when it comes to producing dynamic & learning environments, NLP can truly be helpful. But it's also really important to tread carefully when it comes to the creation and utilization of these models, since they can sometimes hallucinate, and we want the best available advice for our students. [11][20]

## 2.4 Contextual Assistance and Adaptive Learning

By incorporating NLP into the section of EDU Tech , a great deal of increased attention has occurred, especially when it comes to creating AI-fuelled teaching assistants. By using these solutions there is an intention to uplift the learning experience, by providing custom-tailored, on-demand aid to students.

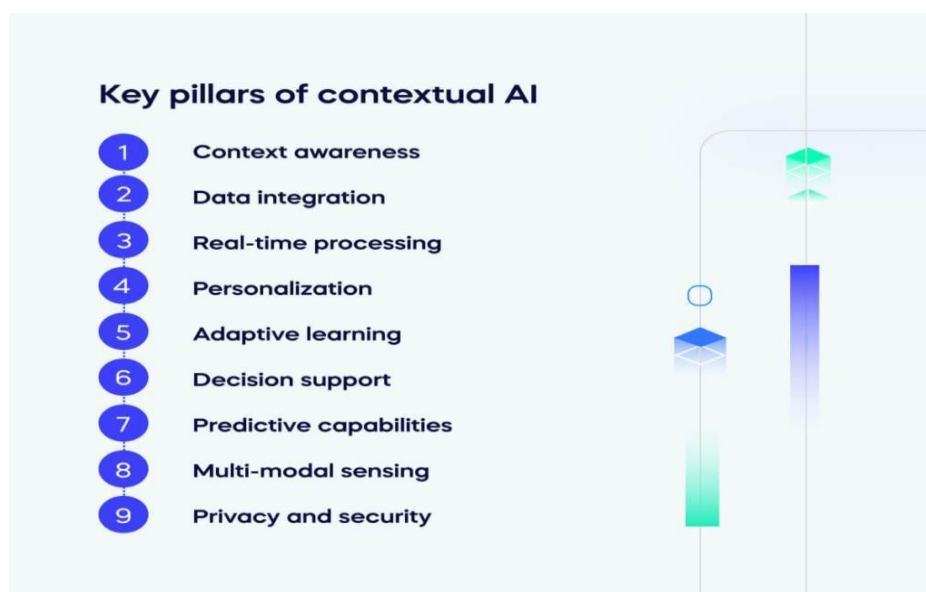


Figure 5: Contextual AI

### 2.4.1 Adapting Learning Systems

When it comes to the students' necessities and overall performance, there is a need to change the course material & speed which aligns with them, that's why adaptive learning systems are made. With the assistance of these systems, we don't only track student's interactivity & progress using real-timed data, but also manage to decide on what the next step will be towards their progress utilizing the AI. [6] When the dust settles, Adaptive learning, in an AI-fuelled educational context, is nothing more that the continuous adaptation and learning of the AI, thanks to its interlinkage with the students. Let's make a hypothetical scenario, where a student can't seem to be able to work around a certain topic. By providing to-the-point clarifications and additional assets, the AI assistant can considerably amplify the student's comprehension. To produce custom-tailored learning programs that are ideal for the various needs of students, it is of most importance to achieve a great level of adaptability. [18] [22]

## 2.4.2 Feedback Mechanisms

Just like in adaptive learning, feedback methods are required to enhance the AI's overall performance and optimise its responses. In educational contexts, teachers' and students' feedback can be obtained. Students can assess how beneficial the AI's responses are, while teachers can provide insights into how the AI's answers align with the course objectives. [4] Once feedback mechanisms are incorporated within AI, then that enables it to evolve over time and thereby support students even more effectively. To make sure that AI remains accurate and applicable in a learning environment, an iterative improvement strategy is required. [18] [23]

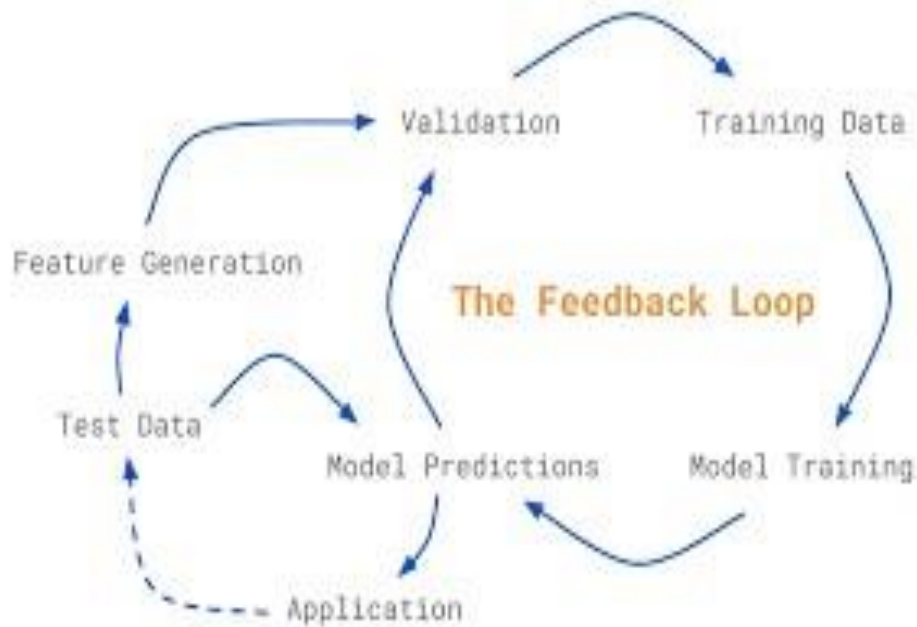
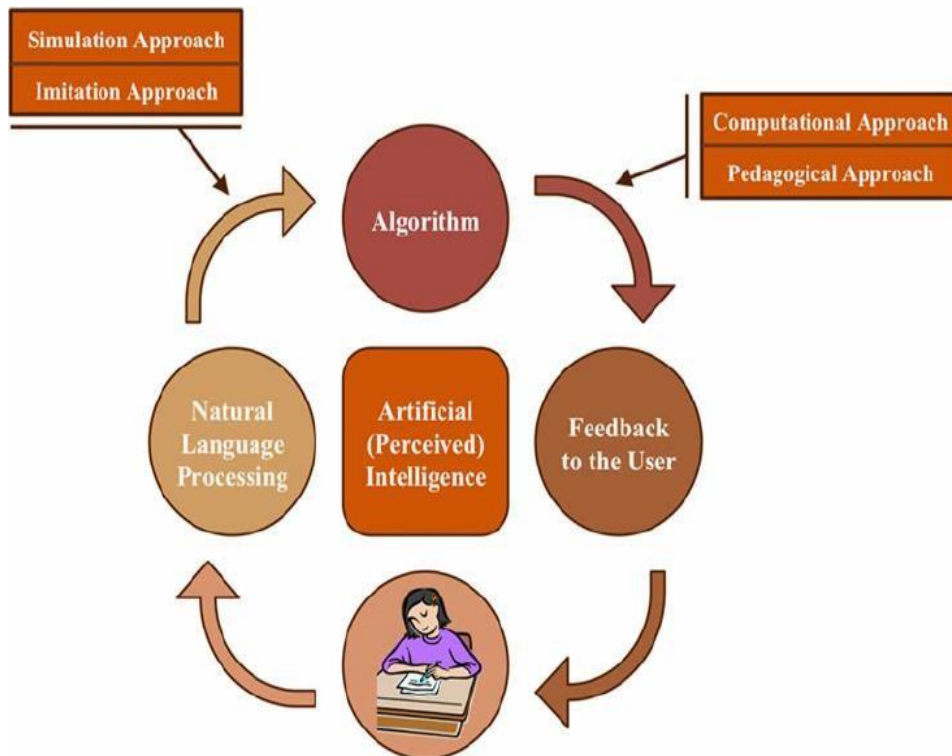


Figure 6: Feedback Loop





**Figure 7: Cycle between NLP and feedback in Intelligent Tutoring Systems**

There are a lot of ways to go about redesigning and generating constructive feedback. Take into consideration that one of the most important aspects of school, assessment, is being revolutionised by NLP. NLP-powered automation is replacing manual grading, which is notoriously time-consuming and potentially subjective. Written assignments, assessments and even programming exercises are graded by algorithms. NLP algorithms quickly assign accurate grades by examining syntax, semantics and content. This effectiveness allows teachers to spend more time on individualised instruction without compromising fairness and consistency of assessment. This change makes the best use of class time and ensures fair assessment. The foundation of improvement is giving constructive criticism. By providing indepth evaluations of students' work, NLP elevates the feedback process to the level of an art form. These algorithms not only identify errors but also offer useful suggestions for enhancements. Learners gain from regular, objective feedback that is readily available and creates an atmosphere that encourages skill growth. The recurring nature of this feedback loop encourages a growth mindset where setbacks are seen as opportunities to improve. As they get education tailored to their unique areas of need, learners experience a sense of continuous progress. [12][23]

## 2.5 Empowering Content Creation and Curation

Something that also should be stated is the huge enhancement of the speed & efficacy of instructional content creation that has been seen when NLP was used to automatically produce content. Even teachers can generate various content, such as test questions and clarification texts, just by employing NLP-enabled systems. NLP algorithms generate lesson content by assessing current resources and learning objectives, saving teachers time while ensuring highquality teaching materials. Teachers might also focus on tailoring the content to the needs of each student. In addition, because there is so much information available online, which can be overwhelming for students, content curation can be challenging. Especially when it comes to considering the preferences, progress, and learning style of every individual, NLP provided techniques for recommendation-making that do just that. These algorithms examine huge data sets to identify correlations between success and student interactions. In this way, students receive content that is tailored to their own interests and learning styles. As students interact with the recommended content, the algorithm improves its suggestions, creating a positive feedback loop. This constant refinement ensures that students always receive relevant and helpful information. [12] [24]

## 2.6 Ethical Considerations And Challenges

When it comes to the Ethical side of things, there are two main sections that require our utmost focus:

- **Data Privacy And Security:** The fact that NLP systems collect data to enhance their functionality gave rise to even more ethical concerns regarding data security and privacy. Data encryption both in transit & in storage, clear data use policies and strict access controls are critical security components. It is highly important to maintain a balance between the protection of students' private information and using using the data for development to maintain trust. [12]
- **Bias Mitigation:** Dealing with prejudices that can unintentionally creep into algorithms is necessary for the ethical use of NLP in eLearning. Biased training data can affect advice on content, grading and language comprehension. Continuous monitoring and testing of NLP systems is critical to detect and correct biases early. Educators and

developers need to improve algorithms, actively seek out diverse perspectives and create thorough guidelines to ensure fairness and inclusivity - this commitment to reducing bias will ensure technology continues to be a tool for equitable learning experiences.

## 2.7 Commercial AI Solutions for Education

Commercial AI systems such as Livechat AI offer a more realistic deployment strategy, especially in resource-limited situations, although custom solutions such as the refined LLaMA3 allow for deep customization and potentially higher accuracy.

### 2.7.1 Overview of Livechat AI

Livechat AI is a well-known commercial chatbot platform for a number of areas, including education. It offers several pre-built features, including natural language understanding, easily customizable response templates, and an easy-to-use connection to the web platform.



Figure 8: Livechat AI

The main benefit of Livechat AI is its scalability. The platform is ideal for extensive educational implementations since it is built to accommodate a high volume of users at the same time. Furthermore, the user-friendly interface of Livechat AI enables quick customization and deployment, thus cutting down on time and resources needed for implementation. [17]

### **2.7.2 Application in Educational Settings**

Livechat AI has been effectively integrated into various educational systems, allowing students to receive real-time guidance. The platform's capacity to answer frequent queries, provide fast response, and connect students to relevant resources makes it a useful learning support tool. Compared to custom-built solutions such as the fine-tuned LLaMA-3, Livechat AI is easier to install. While it may not have the deep contextual understanding of a fine-tuned model, its ease of use and scalability make it a viable option for educational institutions wishing to deploy AI-assisted learning solutions fast and efficiently.

### **2.7.3 EDU NLP Conclusion**

By enhancing the way EDU content is being created-delivered and evaluated, NLP is making important changes in online Learning. Not only it's able to tailor learning specifically to the individuals' needs, but further allows learners of different backgrounds to join in by removing language barriers and facilitating access to worldwide learning. These advances bring to light critical ethical considerations, including the need to ensure data privacy and work towards technologies that are free from bias. If used wisely, the synergy of NLP and eLearning holds incredible potential to shape a more adaptable, inclusive and engaging future of education. The impact of NLP on education is not a passing trend, but a fundamental shift in a rapidly evolving landscape. NLP is helping us to rethink education by enabling personalized experiences, breaking down language barriers and driving content creation and recommendations. The ethical challenges remind us that we need to use these innovations responsibly and steer the technology in the right direction. [12]

## 2.8 LLaMA-3 & Livechat-AI

### 2.8.1 LLaMA-3 Consideration

The rationale behind deciding to use LLaMA-3 for our use case instead of other advanced and capable models can be broken down into 3 reasons:

- I. **Harmonizing Computational Efficiency and Performance:** Although models like GPT-4 and GPT-3 are excellent at producing contextualised, high-quality information, their large computing requirements make them impractical for educational institutions with inadequate infrastructure. The good thing about LLaMA-3 is that it arranges a certain compromise, with which a good performance is established without requiring large resource amounts. Thanks to that, LLaMA-3 proves to be an ideal choice for educational applications, where most of the times processing power as well as funds may be hard to find. On top of that, as a consequence of its effective fine-tuning abilities, LLaMA-3 may also support certain instructional tasks that would otherwise require far more resource-rigorous models.
  
- II. **Customization Abilities:** While models like BERT and RoBERTa are optimised for text categorisation, LLaMA-3 is intended for more dynamic and generative features, making it perfect for AI-powered instructional assistants. LLaMA-3 allows for more individualisation than task-specific models like T5, as it can adjust responses to specific course content. With the ability to manage dynamic interactions, LLaMA-3 can provide a more immersive learning environment.
  
- III. **Scalability and Adaptability for Educational Needs:** LLaMA-3 may not be the greatest model, but it is capable of providing meaningful and contextualised answers. In an educational environment, this is crucial for answering open-ended questions or processing complex queries. Techniques such as Alpaca-based fine-tuning and unsloth optimization, further improve the efficiency of the model and allow it to scale without

excessive computational overhead, so that it can be adapted to different educational requirements.

## 2.8.2 Livechat AI Considerations

Following a somewhat similar “train of thought” of looking after a scalable-affordable yet productive solution, we ended up choosing Livechat AI as a second candidate because it fitted the criteria we wanted:

**Effortless deployment and integration:** It offers a pre-made solution with advanced NLP functionality that is ideal for schools with limited technological resources. Not only that but it also integrates smoothly into existing platforms with minimal effort, as opposed to custom models that require extensive setup and maintenance. One important thing that sets it aside from other state-of-the-art platforms, is its simplicity in allowing instant and extensive adoption in EDU institutions.

**Scalability and Live Features:** Livechat AI was developed utilising a cloud-based architecture that enables for large-scale deployments while managing multiple user interactions at the same time. This scalability is particularly useful in institutions where a large number of students require AI support at the same time. Even though models like BERT or GPT-4 can accomplish the same task, they often require more resources to run in real time. On the other hand, this one is designed for real-time use in high-demand situations, providing a smooth experience without straining the institution's infrastructure.



**Operational efficiency and cost-effectiveness:** It offers a more affordable option when compared to the deployment of a custom-trained model such as GPT-4, which entails continuous maintenance and maybe elevated operating expenses. Its subscription-based or usage-based pricing models can be more affordable, particularly for educational institutions that need to optimize budgets. Livechat AI's pre-configured NLP models and intuitive user interface mean that less technical expertise is required to maintain it, reducing operational overhead and allowing staff to focus on supporting student learning rather than managing complex AI systems.

## CHAPTER III: System Architecture and Deployment

### 3.1 First Approach: Fine-Tuning LLaMA-3 with Alpaca and Unsloth

In this section we discuss the technical steps needed in fine-tuning the LLaMA-3 model with Alpaca, followed by optimisation with Unsloth. The purpose of this technique was for us to develop a highly accurate and contextually aware AI model that might help with educational queries.

#### 3.1.1 System Overview

LLaMA-3 (Large Language Model Meta AI), an advanced language model created by Meta AI, is intended for high-performance NLP jobs. The model was fine-tuned with Alpaca, an approach for refining big models for specific educational objectives, and optimised with Unsloth to improve computational efficiency, resulting in faster inference and less memory consumption.

#### 3.1.2 Fine-Tuning Process

The fine-tuning process involved multiple stages, each crucial for adapting LLaMA-3 to educational queries: [25]

- **Initial Setup and Installations:**

Installing necessary packages, such as Xformers and Unsloth, which are necessary for memory management and attention mechanism optimisation during training, was the first step in the process. To ensure compatibility with the computational infrastructure, these packages were installed within the Jupyter environment using Python package management commands.



```
▼ First step installations

[1] 1 %%capture
    2 import torch
    3 major_version, minor_version = torch.cuda.get_device_capability()
    4 # Must install separately since Colab has torch 2.2.1, which breaks packages
    5 !pip install "unsloth[colab-new] @ git+https://github.com/unslothai/unsloth.git"
    6 if major_version >= 8:
    7
    8     !pip install --no-deps packaging ninja einops flash-attn xformers trl peft accelerate bitsandbytes
    9 else:
   10
   11     !pip install --no-deps xformers trl peft accelerate bitsandbytes
   12 pass
```

Figure 9: Installing essential packages

- **Load and configure model:**

Using the FastLanguageModel class provided by Unsloth, the LLaMA-3 model was loaded, with the configuration set to use 4-bit quantization for memory efficiency. By doing this we managed to significantly reduce the memory requirements and that way the model could work relatively well across low-memory GPUs. We included LoRA adapters in order to enrich the training part. With the assistance from these adapters, we managed to fine-tune a small portion of the model's parameters, and as such, reduced the overall heavy computational load while maintaining our desired performance.

```
✓ We now add LoRA adapters so we only need to update 1 to 10% of all parameters!
```

```
1 model = FastLanguageModel.get_peft_model(  
2     model,  
3     r = 16, # Choose any number > 0 ! Suggested 8, 16, 32, 64, 128  
4     target_modules = ["q_proj", "k_proj", "v_proj", "o_proj",  
5                     "gate_proj", "up_proj", "down_proj",],  
6     lora_alpha = 16,  
7     lora_dropout = 0, # Supports any, but = 0 is optimized  
8     bias = "none",    # Supports any, but = "none" is optimized  
9     # [NEW] "unsloth" uses 30% less VRAM, fits 2x larger batch sizes!  
10    use_gradient_checkpointing = "unsloth", # True or "unsloth" for very long context  
11    random_state = 3407,  
12    use_rslora = False, # We support rank stabilized LoRA  
13    loftq_config = None, # And LoftQ  
14 )
```

Unsloth 2024.8 patched 32 layers with 32 QKV layers, 32 O layers and 32 MLP layers.

Figure 10 : Model loading and Configuration

- **Data Preparation:**

The refined dataset was formed with the sanitized Alpaca dataset, featuring 52,000 instances of instruction-response pairs. This data set was selected due to its pertinence to teaching tasks commonly found in educational environments. For the creation of the Alpaca dataset, a revised prompt template was used to guarantee that every question-answer pair is correctly structured and includes a token at the end of each sequence to help with error identification during inference.

```
▼ Data Prep

We now use the Alpaca dataset from yahma, which is a filtered version of 52K of the original Alpaca dataset.

1 alpaca_prompt = """Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.
2
3 ### Instruction:
4 {}
5
6 ### Input:
7 {}
8
9 ### Response:
10 {}"""
11
12 EOS_TOKEN = tokenizer.eos_token # Must add EOS_TOKEN
13 def formatting_prompts_func(examples):
14     instructions = examples["instruction"]
15     inputs       = examples["input"]
16     outputs      = examples["output"]
17     texts = []
18     for instruction, input, output in zip(instructions, inputs, outputs):
19         # Must add EOS_TOKEN, otherwise your generation will go on forever!
20         text = alpaca_prompt.format(instruction, input, output) + EOS_TOKEN
21         texts.append(text)
22     return { "text" : texts, }
23 pass
24
25 from datasets import load_dataset
26 dataset = load_dataset("yahma/alpaca-cleaned", split = "train")
27 dataset = dataset.map(formatting_prompts_func, batched = True,)
```

Downloading readme: 100% ██████████ 11.6k/11.6k [00:00<00:00, 83.7kB/s]

Downloading data: 100% ██████████ 44.3M/44.3M [00:00<00:00, 117MB/s]

Generating train split: 100% ██████████ 51760/51760 [00:00<00:00, 81709.38 examples/s]

Figure 11: Data Prep

- **Training Configuration:**

Huggingface's TRL library was utilised to control the training procedure through the SFTTrainer class. The training parameters included stages of gradient accumulation, a stack size of two, and a learning rate of  $2e-4$  to maximise GPU memory use. The 60-step technique was developed to ensure rapid adaptation and prevent overfitting. During training, memory usage was monitored closely to stay within hardware limits, and a record of the peak memory usage was maintained. Careful monitoring was necessary to maintain a balance between model performance and training efficiency.

```

  ✓
  4s
  ▶
  1 from trl import SFTTrainer
  2 from transformers import TrainingArguments
  3 from unsloth import is_bfloat16_supported
  4
  5 trainer = SFTTrainer(
  6     model = model,
  7     tokenizer = tokenizer,
  8     train_dataset = dataset,
  9     dataset_text_field = "text",
 10    max_seq_length = max_seq_length,
 11    dataset_num_proc = 2,
 12    packing = False, # Can make training 5x faster for short sequences.
 13    args = TrainingArguments(
 14        per_device_train_batch_size = 2,
 15        gradient_accumulation_steps = 4,
 16        warmup_steps = 5,
 17        max_steps = 60,
 18        learning_rate = 2e-4,
 19        fp16 = not is_bfloat16_supported(),
 20        bf16 = is_bfloat16_supported(),
 21        logging_steps = 1,
 22        optim = "adamw_8bit",
 23        weight_decay = 0.01,
 24        lr_scheduler_type = "linear",
 25        seed = 3407,
 26        output_dir = "outputs",
 27    ),
 28 )

```

Map (num\_proc=2): 100%  51760/51760 [00:53<00:00, 1280.55 examples/s]

max\_steps is given, it will override any value given in num\_train\_epochs

Figure 12: Training the Model

### 3.1.3 Technical Challenges and Solutions

Several challenges were encountered during the fine-tuning and subsequent deployment of the LLaMA-3 model:

- **High Computational Requirements:** In contexts with limited resources, the model's high parameter count and the requirement for precision in fine-tuning demanded significant processing power. With the implementation of 4-bit quantization and LoRA adapters we achieve a reduction in the memory demand, thus allowing for the model to function well on mainstream GPUs like T4. And with Unsloths' help we also manage to lower even more the computational overhead without significantly losing accuracy.

- **Implementing/Integrating models:** Since LLaMA-3 is quite large as a model and requires lots of resources, its integration into an online platform proved to be quite difficult and we couldn't get the responses we wanted. To work around the problem, we fine-tune the original model and bring it down to our computational level. Although scaling remained difficult, this modular strategy improved load management and decreased latency

### 3.1.4 Results and Evaluation

We adjusted LLaMA-3 using various inputs from the PDFs to observe its performance. The outcomes are displayed below:

**Response Quality:** We received thorough and precise responses, indicating the model's high level of proficiency in elucidating intricate ideas. The adjustment of the model to specific educational tasks and offering relevant and insightful responses was achieved through the refinement process.

**Increased Efficiency:** The model improved its performance in inference tasks by utilizing Unsloth optimizations and 4-bit quantization, making it suitable for deployment in environments with limited computational resources due to reduced memory usage.

### 3.1.5 Summary of the First Approach

With the use of Alpaca and Unsloth, the fine-tuned LLaMA-3 proved to be quite competent and showed promise of advanced fine-tuning methods. Sadly though, due to the scalability / computational power issues we encountered, the focus was shifted towards other options. It was crystal clear that a more straightforward and simple solution was needed.

## 3.2 Second Approach: Livechat AI Integration/ Custom Website build

In this section we describe how the Livechat AI platform was incorporated into a custom website we created, with the J.A.R.V.I.S chatbot being used to provide real-time help to students.

### 3.2.1 System Overview

One of the main reasons we chose Livechat AI was because it has superb NLP capabilities, and it was really easy to integrate it into the website that we created with the Website.com platform. Just like with Livechat AI, Website.com offered an ease of use like no other, allowing for instant incorporation of external API's and gave to the user the absolute freedom to change the platform fast and with no issues. [17]



Livechat AI operates as a cloud-based service. As a result, it does not require the same level of regional data processing. Sources such as LLaMA-3 serve as a model. This enables it to grow easily, managing a high number of concurrent user interactions without any issues.

The platform provides extensive customization options, allowing for the adaptation of its NLP models to specific domains, such as education. Custom intents and entities can be defined to ensure the chatbot responds accurately to domain-specific queries.

### 3.2.2 Creating the website

In order to make sure that the user would have the best possible experience, without struggling to figure out how to use the chatbot, we used Website.com to develop the platform. It was simple, on-point and most importantly it allowed seamless integration of Livechat AI. [26]

#### 3.2.2.1 Website Design and Structure

The website was designed with a user-centric focus. Improved usability by making the chatbot feature easily accessible for students. The site features a straightforward and uncluttered appearance, easy-to-navigate navigation, and a prominent chatbot interface. [26]

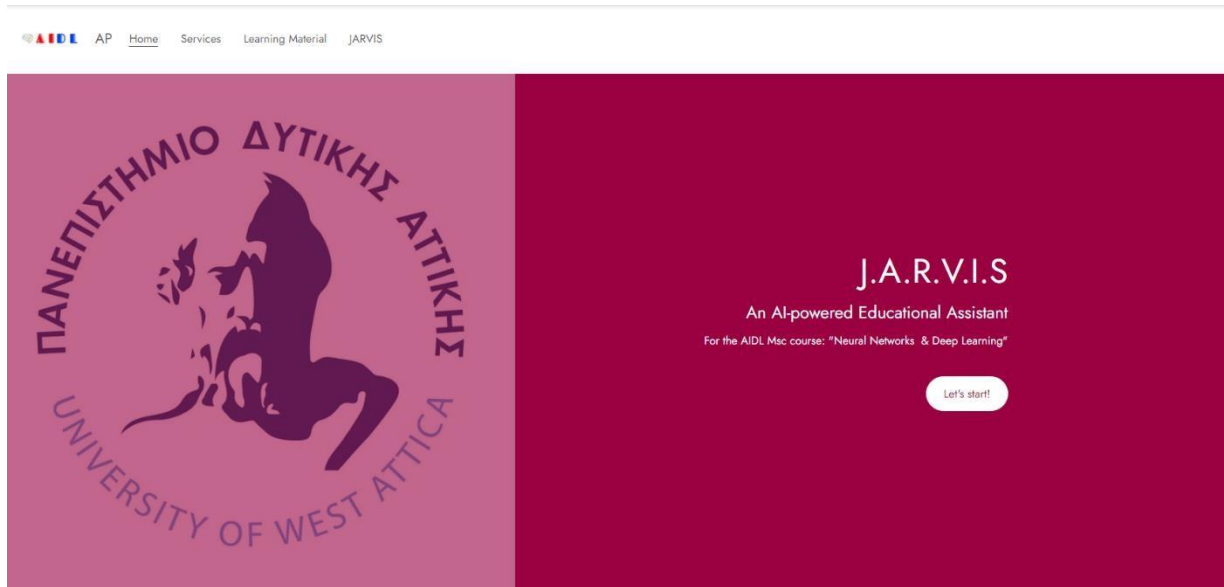


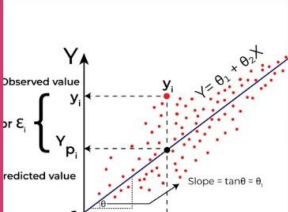
Figure 13: General Home Page

The website included course materials to expand the chatbot's skills and allow learners to engage with information independently or with AI support.

## LEARNING MATERIAL

Get access to PDFs containing all the material taught in the course of Neural Networks and Deep Learning. Enhance your understanding of key concepts and theories with comprehensive learning materials at your fingertips. Below you can see the files and dive right in!

- [Lesson 1 : Introduction](#)
- [Lesson 2 : Linear Regression](#)
- [Lesson 3 : Logistic Regression](#)
- [Lesson 4 : Neural Networks](#)
- [Lesson 5 : Network Optimization](#)
- [Lesson 6 : CNNs](#)
- [Lesson 7 : Advanced CV Techniques & Transfer Learning](#)
- [Lesson 8 : Word Embeddings](#)
- [Lesson 9 : Recurrent Neural Networks](#)
- [Lesson 10 : Object Detection / Image Segmentation](#)
- [Lesson 11 : GAN](#)



A scatter plot showing observed values  $y_i$  and predicted values  $\hat{y}_i$  against input  $x_i$ . A regression line is shown with the equation  $Y = \theta_0 + \theta_1 X$  and slope  $\text{slope} = \tan \theta = \theta_1$ .

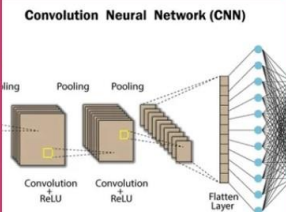


Diagram of a CNN architecture showing layers: Convolution ReLU, Pooling, Convolution ReLU, Pooling, and Flatten Layer.

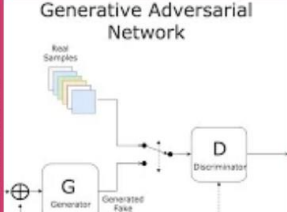


Diagram of a GAN architecture showing a Generator (G) that takes noise as input to produce generated data, which is then evaluated by a Discriminator (D) against real samples.

**Figure 14: Learning Material**

### 3.2.2.2 Creating JARVIS

Using the API integration tools that Livechat AI provided, it was relatively easy to integrate the chatbot into the website. Students could get help anytime they wanted it just by clicking on the “JARVIS” section of the website. The chatbot was created to deliver quick and accurate answers. This was achieved by utilizing Livechat AI's adaptable cloud-based infrastructure, which adjusted automatically based on user needs. [17] [26]



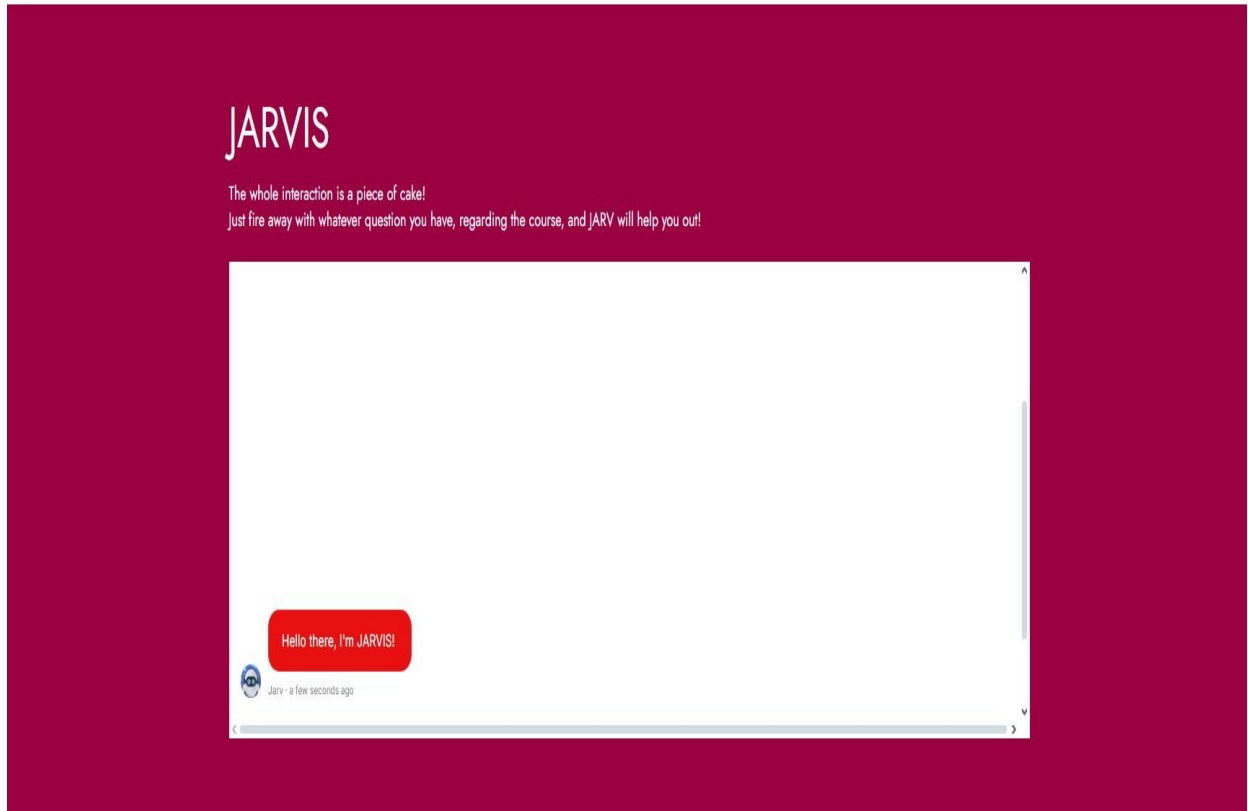


Figure 15: JARVIS Section

### 3.2.3 Advantages in Technology

When we combined Livechat AI with our custom website, we found advantages that the optimised LLaMa-3 was lacking. One of the most spectacular aspects of it was how easy it was to put it to action. In contrast to LLaMA-3, Livechat AI integration on the Website.com platform was simple and did not require complicated setup or extensive optimization. Its plug-and-play functionality made deployment effortless, with no need for additional customization. Compared to the resource-intensive LLaMA-3 paradigm, livechat AI is more scalable because it is designed to manage numerous users at once. This was particularly important in an educational setting where multiple students may interact with the chatbot at the same time. It's really important to mention how cost efficient this viewpoint was. Thanks to Website.com's affordable hosting and with Livechat AI's subscription model updates and support, we can reduce the long-term costs. [17] [26]

### **3.2.4 Summary of Second Approach**

For the educational platform, integrating Livechat AI with a custom website built on Website.com turned out to be a more workable and scalable approach. As it was previously stated, although LLaMA-3 performed really well, we ended up not choosing it for our use case, since we did not have the computational power or the infrastructure to use it at an acceptable and satisfying level. Thus, we went along with Livechat AI which was far simpler to use, had guaranteed scalability along with an affordable package for our use case. That way we could secure that a stable, robust and trustworthy assistant / platform would be used by the students.

## 4 CHAPTER IV: Demonstrations and Comparative Analysis

### 4.1 Introduction

In this chapter we explore the two approaches that were used in the thesis: the fine-tuned LLaMA-3 model and the Livechat AI chatbot, which was integrated into a custom educational website. Through carefully chosen visual examples, code samples, and in-depth commentary, each approach is examined based on practical performance, emphasizing how well each one supports educational objectives and where limitations may appear.

### 4.2 First Approach: Fine-Tuned LLaMa-3

#### 4.2.1 Code/Model Functionality

The LLaMA-3 model, which can be customised to answer educational questions, was created to cover complicated topics and provide students with exciting learning experiences. In this section we included several examples that demonstrate its capabilities. [27]

In one example, a prompt asking the model to explain Recurrent Neural Networks (RNNs) yields a comprehensive answer outlining its applications in speech recognition and natural language processing, as well as how RNNs handle sequences by keeping previous input information. This example underscores LLaMA-3's proficiency in addressing complex topics, breaking them down in ways that enhance understanding. The model makes abstract ideas more relatable by demonstrating practical applications, which can make challenging subjects more accessible to students.

```
▼ Inference
Let's run the model! You can change the instruction and input - leave the output blank!

1 # alpaca_prompt = Copied from above
2 FastLanguageModel.for_inference(model) # Enable native 2x faster inference
3 inputs = tokenizer(
4 [
5     alpaca_prompt.format(
6         "Answer the question in detail,as if you explain it to someone who has no knowledge about it and provide examples.", # instruction
7         "What is a Neural Network and in what architectures are they commonly used?", # input
8         "", # output - leave this blank for generation!
9     )
10 ], return_tensors = "pt").to("cuda")
11
12 outputs = model.generate(**inputs, max_new_tokens = 64, use_cache = True)
13 tokenizer.batch_decode(outputs)

[ '<|begin_of_text|>Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.\n\n## Instruction:\nAnswer the question in detail,as if you explain it to someone who has no knowledge about it and provide examples.\n\n## Input:\nWhat is a Neural Network and in what architectures are they commonly used?\n\n## Response:\nA neural network is a type of artificial intelligence that is modeled after the human brain and nervous system. It consists of interconnected nodes, also known as neurons, that are arranged in layers. Each neuron is connected to several other neurons, and information flows through the network by way of these connections.\n\nNeural networks are commonly used']
```

Figure 16: Example 1 LLaMa

Another prompt (see **example 2 picture**) requests the model to generate a 10-question quiz on neural networks, and the model provides a diverse set of questions covering essential concepts, network types, and layer functions. This feature illustrates LLaMA-3's capacity to automate aspects of test creation, potentially saving educators time while supplying meaningful, relevant content for assessments. By producing varied questions that probe a student's understanding of neural networks from multiple angles, the model supports an interactive and comprehensive learning experience.

```
[ ] 1 inputs = tokenizer(  
2  [  
3    alpaca_prompt.format(  
4      "Please create a 10 question test regarding neural networks.", # instruction  
5      "", # input  
6      "", # output - leave this blank for generation!  
7    )  
8  ], return_tensors = "pt").to("cuda")  
9  
10 outputs = model.generate(**inputs, max_new_tokens = 264, use_cache = True)  
11 tokenizer.batch_decode(outputs)
```

```
[ '<|begin_of_text|>Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.\n\n### Instruction:\nPlease create a 10 question test regarding neural networks.\n\n### Input:\n\n\n### Response:\n1. What is a neural network?\n2. What is the difference between a feedforward neural network and a recurrent neural network?\n3. What is a hidden layer in a neural network?\n4. What is the backpropagation algorithm used for in a neural network?\n5. What is a convolutional neural network?\n6. What is a pooling layer in a convolutional neural network?\n7. What is a fully connected layer in a neural network?\n8. What is the purpose of a loss function in a neural network?\n9. What is gradient descent in a neural network?\n10. What is the difference between a supervised neural network and an unsupervised neural network?<|end_of_text|>']
```

Figure 17: Example 2 : LLaMa

A third example (see **example 3 picture**) wants a basic explanation of neural networks for folks with no prior knowledge. The model's response gives a comprehensible overview by drawing analogies to real-world applications. This remark emphasises the model's ability to adjust explanations to meet the requirements of different students, making it an effective instructional tool for students with various degrees of expertise of the subject. LLaMA-3 serves as a foundational guide for novices, offering straightforward, sample-based guidance to assist individuals in gaining assurance and essential comprehension in technical fields.

```
We can also use a TextStreamer for continuous inference - so you can see the generation token by token, instead of waiting the whole time!
```

```
1 # alpaca_prompt = Copied from above
2 FastLanguageModel.for_inference(model) # Enable native 2x faster inference
3 inputs = tokenizer(
4 [
5     alpaca_prompt.format(
6         "Answer the question in detail,as if you explain it to someone who has no knowledge about it and provide examples.", # instruction
7         "What is a Neural Network and in what architectures are they commonly used?", # input
8         "", # output - leave this blank for generation!
9     )
10 ], return_tensors = "pt").to("cuda")
11
12 from transformers import TextStreamer
13 text_streamer = TextStreamer(tokenizer)
14 _ = model.generate(**inputs, streamer = text_streamer, max_new_tokens = 128)
```

```
<|begin_of_text|>Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appr
### Instruction:
Answer the question in detail,as if you explain it to someone who has no knowledge about it and provide examples.
### Input:
What is a Neural Network and in what architectures are they commonly used?
### Response:
A neural network is a type of machine learning algorithm that mimics the functioning of the human brain. It consists of interconnected nodes,
Neural networks are commonly used in a variety of applications such as image recognition, natural language processing, and predictive modeling
```

Figure 18: Example 3 LLaMa

When it comes to performance, LLaMA-3 stands out in delivering thorough, relevant answers, which makes it an attractive option for detailed educational tasks. Nevertheless, the extensive computing needs and complex implementation demands of this method pose substantial challenges. Its demand for resources suggests that it may be most suitable for specialized environments with abundant resources that require sophisticated, AI-driven teaching support.

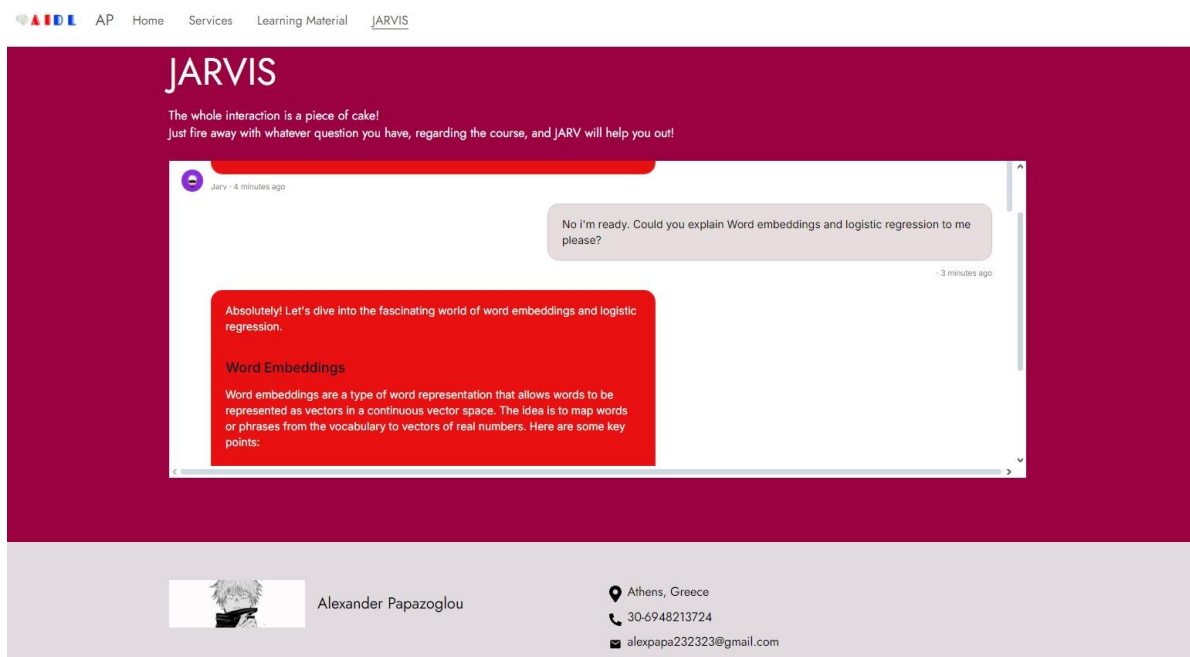
## 4.3 Second Approach: Livechat AI Integration

### 4.3.1 J.A.R.V.I.S Chatbot on the Website

The AI-powered J.A.R.V.I.S , was integrated into our website to provide students with immediate support, focussing on ease of use and accessibility. Thanks to Livechat AI's

sophisticated API, the chatbot was easily installed as a separate page of the website, allowing students to access it anytime they needed to.

In one illustrative interaction, a student asks the chatbot to explain word embeddings and logistic regression. The chatbot provides a detailed explanation of word embeddings, illustrating how they convert words into vectors within a continuous vector space and connecting the idea to practical uses in the real world. The chatbot adeptly helps the student navigate each topic step-by-step when they inquire more about logistic regression. This capacity for responsive, adaptive guidance illustrates the chatbot's potential to explain challenging concepts in a progressive, understandable way, helping students build knowledge in manageable steps without feeling overwhelmed.



**Figure 19 : Explaining Word Embeddings**

The chatbot's user-friendly interface and reliable real-time responses make it particularly beneficial for students who might need immediate clarification without having to navigate complex menus. This streamlined interaction design helps students stay focused on learning without unnecessary delays. Additionally, its conversational flow, which allows for follow-up questions, makes it an effective educational tool that can support a wide range of queries and adapt its guidance to each user's needs. Student feedback emphasised the chatbot's simplicity and intuitive design. The technology gave straightforward responses and was especially useful

*Alexander Papazoglou / mscaidl-0048.*

for people without a technological background, allowing them to get the necessary information without prior understanding of the platform. In this approach, the chatbot provides dependable, accessible assistance, making the learning process easier and more entertaining.

### 4.3.2 Technical Analysis of Livechat AI

The technical qualities of Livechat AI make it a valuable instructional tool. Scalability is one of its most notable features, allowing the system to manage a high amount of requests without sacrificing response time or quality. This makes it an ideal fit for educational platforms where many users might be seeking help at the same time. In terms of deployment, Livechat AI's builtin API integration tools make it straightforward to integrate with existing platforms. Its cloudbased infrastructure allows for rapid implementation without the technical complexity that models like LLaMA-3 demand. This ease of deployment, coupled with the simplicity of maintenance in a cloud setting, allows institutions to integrate Livechat AI efficiently and at scale. In conclusion, Livechat AI shows cost-efficiency, especially in comparison to the expenses needed for creating, teaching, and upkeeping a personalized AI model such as LLaMA-3. The Livechat AI's subscription model includes updates and maintenance, which lowers operational costs in the long run and makes it a financially feasible choice for educational institutions looking to implement AI support tools.

## 4.4 Comparative Commentary

This analysis explores the benefits and drawbacks of using the Livechat AI chatbot and LLaMA3 model in educational settings, with a focus on their performance, scalability, and practical applications. LLaMA-3 is particularly effective for scenarios that demand detailed analysis and nuanced responses, such as advanced courses or specialized training programs. Its ability to generate thorough, contextually accurate answers makes it well-suited for these purposes. However, its complex setup and significant resource requirements can be limiting factors, making it challenging for institutions with smaller budgets or limited technical support to implement.

Its intensive processing demands may also reduce its practicality in broader applications where a more accessible and affordable option is necessary. In contrast, Livechat AI, though it may not reach LLaMA-3's level of depth, is highly efficient in managing multiple queries simultaneously. This feature makes it ideal for settings that prioritize quick, straightforward responses, such as classrooms or tutoring centers, where prompt assistance is key. Its real-time responsiveness and ease of use give it a clear advantage in situations where immediate, straightforward answers are needed. Livechat AI's simplicity and flexibility make it an appealing option for institutions

focused on achieving wide, accessible coverage without sacrificing reliability. In terms of deployment and maintenance, these two options present different requirements. Implementing LLaMA-3 demands considerable computing power and specialized technical skills to maintain and update its functionality. This level of upkeep, including regular fine-tuning, makes it more appropriate for settings where its capacity for detailed responses is essential and worth the investment. On the other hand, Livechat AI's deployment is more straightforward and managed through a subscription-based model that includes automatic updates and optimizations, thus minimizing the need for ongoing technical intervention. Its cloud-based architecture reduces maintenance demands, making it a feasible option for educational institutions with limited IT resources. On the other hand, when it comes to cost-effectiveness, LLaMA-3 requires specialized hardware and continuous technical support, contributing to a high cost of ownership. This investment is significant, even though the model's precise responses and accuracy can provide valuable insights in the right setting. Livechat AI, on the other hand, provides a more cost-effective solution by including regular upgrades and customer assistance in its subscription plan. This methodology keeps prices modest, making it an enticing option for institutions looking to embrace AI without incurring the high upfront and continuing expenditures of a custom solution.

## 4.5 Summary

This section analyzed the capabilities and performance of the Livechat AI chatbot and the refined LLaMA-3 model within an educational setting. Through detailed commentary and visual examples, we assessed how each tool meets specific educational needs. While the LLaMA-3 model excels in generating detailed, in-depth responses that support deep learning, its high technical and computational demands make it less practical for larger-scale or resource-limited environments. Meanwhile, Livechat AI offers a scalable, accessible solution that delivers accurate, real-time responses, positioning it well for wider educational use. Though it may not match LLaMA-3's depth, its ease of use and cost-efficiency make it a compelling choice for educational institutions aiming to integrate AI into their platforms effectively.



## 5 CHAPTER V: Conclusion and Recommendations

### 5.1 Summary of Findings

In this thesis, except for creating a website and integrating a live ai assistant into it , we managed to explore the method of fine-tuning a pretrained model, like LLaMA-3 , as well as witnessing how we can use to our advantage a pre-built online platform and adjust it to our use case with our own data. There is far more that can be done in this specific AI domain, but as a first step we managed to get the grasp of it. LLaMA-3 showed an impressive ability to deliver detailed, context-sensitive responses, making it a powerful tool for tackling complex questions and supporting advanced learning needs. However, the model's high computational requirements and complex setup present challenges, especially when considering large-scale implementation. Livechat AI, on the other hand, proved to be more practical for widespread educational use, providing quick and accurate responses to general questions while requiring fewer resources. Despite not providing the same insights as LLaMA-3, due to its scalability and ease of deployment, it continues to be a viable option in an educational environment.

### 5.2 Recommendations

Based on the findings, the following recommendations are proposed:

- **Utilize AI-Powered Livechat for Extensive Educational Support:** Livechat AI is a suitable choice for companies seeking to incorporate AI-powered assistance across various scenarios. It works particularly well in settings where timely, accurate information is crucial because of its simple setup, low cost, and ability to process high query rates.
- **Reserve LLaMA-3 for Specialized Educational Programs:** LLaMA-3's fine-tuned model is best suited for specialized educational programs where in-depth, context-rich responses are essential. This strategy can be particularly useful in higher education or specialised training, when the emphasis is on depth and institutions have the means to accommodate more complex technologies.

- **Consider a Hybrid Model:** Combining Livechat AI and LLaMA-3 could provide a balanced solution. In this configuration, Livechat AI could easily handle common requests, while LLaMA-3 covers more complicated issues, combining depth and scalability in educational support.

### 5.3 Future Work

Looking ahead, there are topics for additional research:

**Improving Scalability of Custom AI Models:** New methodologies may make models such as LLaMA-3 more accessible for broader deployments. Optimizations or model distillation could help lessen the computational demand, making these models more feasible for broader applications.

**Enhancing Contextual Comprehension in Pre-Made AI Solutions:** Pre-built platforms like Livechat AI can significantly improve their effectiveness by adopting more advanced natural language processing techniques to match the capabilities of sophisticated models like LLaMA-3. This enhancement would enable these systems to gain a deeper understanding of user inquiries while still ensuring the quick response times that are essential for seamless, real-time interactions.

### 5.4 Final Thoughts

The use of AI has a significant potential to improve accessibility and overall learning experiences in educational settings. Educational institutions can develop more effective and interesting learning environments by carefully selecting and implementing various AI technologies to satisfy specific requirements. This thesis provides critical information on the benefits and drawbacks of employing AI in education, setting the framework for future improvements in this rapidly evolving sector.

## Bibliography-References-Online sources

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008. Retrieved from <https://arxiv.org/abs/1706.03762>
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171-4186. Retrieved from <https://aclanthology.org/N19-1423/>
3. Anna Harazim. (2024). Advancements in natural language processing (NLP). *Iteo Blog*. Retrieved from <https://iteo.com/blog/post/advancements-in-natural-language-processingnlp/>
4. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12, 2493-2537. Retrieved from <https://www.jmlr.org/papers/volume12/collobert11a/collobert11a.pdf>
5. Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence Unleashed: An Argument for AI in Education*. Pearson Education. Retrieved from [https://www.researchgate.net/publication/299561597\\_Intelligence\\_Unleashed\\_An\\_argument\\_for\\_AI\\_in\\_Education](https://www.researchgate.net/publication/299561597_Intelligence_Unleashed_An_argument_for_AI_in_Education)
6. Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press. Retrieved from [https://icog-labs.com/wpcontent/uploads/2014/07/Christopher\\_D.\\_Manning\\_Hinrich\\_Schütze\\_Foundations\\_Of\\_Statistical\\_Natural\\_Language\\_Processing.pdf](https://icog-labs.com/wpcontent/uploads/2014/07/Christopher_D._Manning_Hinrich_Schütze_Foundations_Of_Statistical_Natural_Language_Processing.pdf)
7. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1-67. Retrieved from <https://jmlr.org/papers/v21/20-074.html>

8. Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – Where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 39, *Springer Open*. Retrieved from <https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-019-01710>
9. Unknown Author. (2024, April 22). *Unsloth: Making AI language model training faster and more efficient*. Retrieved from <https://didyouknowbg8.wordpress.com/2024/04/22/unsloth-making-ai-language-modeltraining-faster-and-more-efficient/>
10. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901. Retrieved from <https://papers.nips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64aAbstract.html>
11. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220-229. Retrieved from <https://dl.acm.org/doi/10.1145/3287560.3287596>
12. Sameeksha Medewar. (2024). The role of natural language processing in eLearning. *eLearning Industry*. Retrieved from <https://elearningindustry.com/the-role-of-naturallanguage-processing-in-elearning>
13. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227-2237. Retrieved from <https://aclanthology.org/N18-1202/>
14. Center for Research on Foundation Models. (2023). *Alpaca: Instruction-tuning for LLaMA models*. Stanford University. Retrieved from <https://crfm.stanford.edu/2023/03/13/alpaca.html>

15. **Meta** AI Research. (2024). *LLaMA-3: A new standard for language models*. Meta AI. Retrieved from <https://ai.meta.com/blog/meta-llama-3/>
16. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*. Retrieved from <https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-MultitaskLearners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe>
17. LiveChat. (2024). *LiveChat AI integration for educational platforms*. Retrieved from <https://livechatai.com/education-ai-assistant>
18. Laya, B., Anuraag, A., & Koushik, P. (2024). Transformative advances in PDF information management: Leveraging transformer models for contextual query-answering. *International Journal of Scientific Research in Engineering and Management*, 7(2), 112120. Retrieved from <https://ijsrem.com/download/transformative-advances-in-pdfinformation-management-leveraging-transformer-models-for-contextual-query-answering/>
19. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645-3650. Retrieved from <https://aclanthology.org/P19-1355/>
20. Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Center for Curriculum Redesign. Retrieved from [https://www.researchgate.net/publication/332180327\\_Artificial\\_Intelligence\\_in\\_Education\\_Promise\\_and\\_Implications\\_for\\_Teaching\\_and\\_Learning](https://www.researchgate.net/publication/332180327_Artificial_Intelligence_in_Education_Promise_and_Implications_for_Teaching_and_Learning)
21. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*. Retrieved from <https://arxiv.org/abs/1907.11692>
22. Olney, A., Graesser, A., & Person, N. (2012). *Tutorial Dialog in Natural Language*. *International Journal of Artificial Intelligence in Education*,. Retrieved from [https://bpbusw2.wpmucdn.com/blogs.memphis.edu/dist/d/2954/files/2019/10/olney\\_advances\\_its\\_20](https://bpbusw2.wpmucdn.com/blogs.memphis.edu/dist/d/2954/files/2019/10/olney_advances_its_20)

[10.pdf](#)

23. Jurafsky, D., & Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall. Retrieved from <https://web.stanford.edu/~jurafsky/slp3/10.pdf>
24. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*. Retrieved from <https://arxiv.org/abs/2108.07258>
25. Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., ... & Jégou, H. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. Retrieved from <https://arxiv.org/abs/2302.13971>
26. Website.com. (2024). *Create your own custom website*. Retrieved from <https://www.website.com>
27. JARVIS AP. (2024). Llama mscaidl-0048.ipynb. *Google Colab Notebook*. Retrieved from [https://colab.research.google.com/drive/14hyz\\_vi2QFJb2EIth1kIFFUMqWyvyrKb](https://colab.research.google.com/drive/14hyz_vi2QFJb2EIth1kIFFUMqWyvyrKb)