The Effects of a Chatbot Solving Mathematical Equations using NLP

Jananan Gugagayanan

Submitted in accordance with the requirements for the degree of Masters by Research

York St John University

School of Science, Technology, and Health

December 2022

Contents

Declaration

The candidate confirms that the work submitted is their own and that appropriate credit has been given where reference has been made to the work of others.

Abstract

For many years, the field of chatbots have made an evolutionary expansion. However, chatbots have yet to be challenged against any higher-level academic knowledge. What this means is that chatbots have yet to be challenged with higher level information such as find the area of a circle and round the answer to the first significant value. The definition of higher-level academic knowledge for this project means that the experiment will be using college based (A-level) mathematical questions, this is because this level of academics provides a heavier amount of knowledge like calculus symbols for NLP to be examined with in the experiment phase of the thesis. Natural Language Processing (NLP) is a method in artificial intelligence that enables the chatbot to take a big dataset and divide the dataset into little tokens (IMB, 2023). In addition to having the dataset being broken down into pieces of tokens, the accuracy of how NLP answers mathematical questions is a vital principle that is in high demand by users who are using higher level mathematics. In this research paper, the concept of chatbots solving mathematical questions will be explored as well as analysing the accuracy levels when challenged with mathematical topics. The thesis has experimented the NLP chatbot with the following mathematical topics: trigonometry and logarithms. In addition to the NLP chatbot, there has been a pre-programmed chatbot called MATH that will also be used in the experiment to compare the accuracy of answering questions.

The experiment results confirms that NLP chatbot has a better accuracy for answering the selected topics than math library. This is based upon the number of questions each chatbot got correct when answering the questions. Trigonometry was the sole topic to have more than one question answered from both chatbots.

Abbreviation Key

Glossary of abbreviated words

Abbreviations	Meanings
NLP	Natural Language Processing
AI	Artificial Intelligence
NLU	Natural Language Understanding
AIML	Artificial Intelligence Markup Language
ML	Machine Learning
SD	Standard Deviation
M	Mean Value
P	Practical Significance

List of Figures

Chapter 1: Overview

Introduction

Do chatbots have the capability of answering higher level mathematical questions? According to Mcaulay (2022) chatbots have been perceived to tackle human related tasks such as being customer support service. Even though the accomplishments for AI chatbots have been stated, AI scientists have yet to develop a method to allow chatbots to understand higher level mathematics. This thesis highlights the issue that is occurring amongst chatbots and higher-level mathematics and postulates a solution. The beginning of the thesis will introduce the concept of a chatbot, then evaluates the issues that chatbots have with higher level mathematics and finally proposes a solution that will allow chatbots to digest the higher-level mathematics information and answer the questions.

Background

A chatbot is an artificial intelligence agent that enables the user to have discussions through text input (Brush and Scardina, 2022). The evolution of chatbots continues as AI scientists are finding new computations and methods that would allow the chatbots to be more productive than viewing the chatbots as virtual conversationalists. The timeline of chatbots can be seen in Figure 1 (timeline editor, 2022) which illustrates how chatbots evolved in the past.

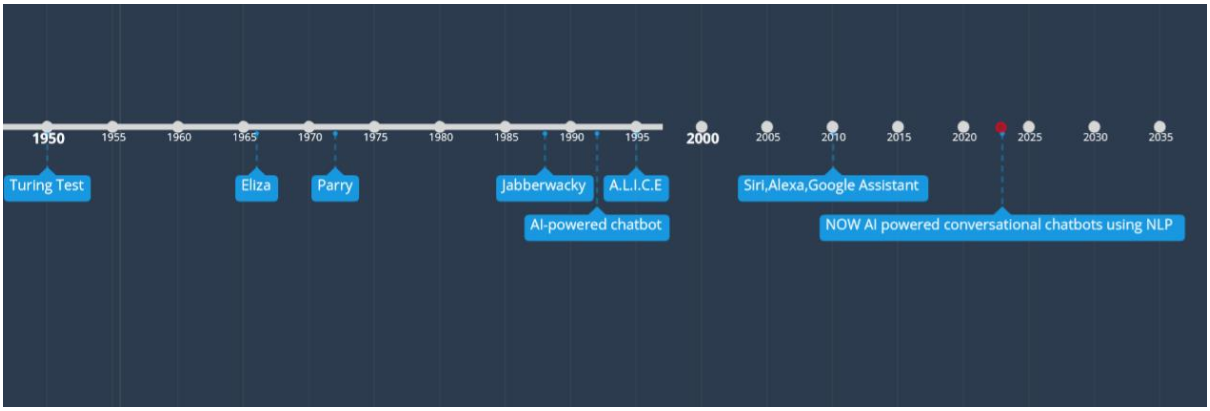

Figure 1: Timeline of chatbots from capacity

One of the current problems that artificial intelligence is facing is chatbots not being able to answer college based mathematical questions. This problem stems from the fact that there are many non-mathematicians who are counting on AI programs to answer questions such as **how many meters are in centimetres?** (Science daily, 2022). Chatbots are one of the AI programs that mathematicians request help from because chatbots can respond to answers quicker and provide detailed commentary for their workout methods (Lee et al 2022).

The importance of this problem is high because future generations in the next 10 years are going to be depending on AI powered chatbots to solve higher level mathematics such as calculus (Aisi, 2021). What this means is that, as the years progress more technologies are being developed to tackle specific jobs such as banks and healthcare (Engati Team, 2022). The ability for a chatbot to solve higher level mathematical questions is in high demand because it would lead to having less human error when it comes to mathematical calculations and the field of mathematics and artificial intelligence will go a step further (Leonard, 2022).

However, there are less chatbots that are fulling the promise of dealing with mathematical notations. One possible reason is because AI scientists have yet to develop a technique or procedure in which a chatbot can digest a mathematical notation (Sundaram, 2022). With the delay of this

issue still on the back burner list, the field of artificial intelligence is making the world of mathematics a harder and tougher process.

One of the reasons that machine learning is worthy of dealing with mathematical notations is that the chatbot is self-learning (Prasanna, 2022). What this means is that machine learning can study the information about the notations by itself without the programmer feeding the information into the chatbot. Another possible reason is that machine learning chatbots are becoming more popular because of the self-learning system which is why there is a huge demand for them to manage mathematical notations (Fintelics, 2021).

On the other hand, machine learning chatbots may not be suited for some users due to the following issues. Automation requirements for machine learning is an issue, what this means is that automation is mechanical tune up for chatbots for them to function properly. Some programmers enter automation into machine learning such as adapting the fetching process of the data, even though it can be helpful for other chatbots but altering the fetching process of machine learning can cause the chatbot to have issues. The other issue with machine learning is the type of data that is being entered into their system (prov, 2022), what this means is that there are types of data that can classified as good data or bad data. The judgement of how a dataset can be seen as good or bad is based on the type of data that the programmer chooses and how the machine learning algorithm responds to it. For example, the dataset $Sin\ x = \frac{1}{4}, \frac{1}{3}$ is considered good data as it is related to mathematics which machine learning can respond to with ease. Whereas a dataset in a foreign language is bad data as the machine learning algorithm would not know how to process the data. Specific maths notations are an issue for machine learning chatbots because the vast majority of chatbots are programmed using Natural Language Processing (NLP). Because NLP has the function to swap data characters with programmable characters, mathematical notations does not possess a programmable character as of late. Moreover, relentless maintenance is a burden because each time the chatbot has a programmable error the repairing time would take too long (Verstegen, 2022). A programmable error is when a chatbot displays a message that requires a programmer's assistance (vocabulary, 2022). An example of a programmable error would that the chatbot would display an error message to say that the program is not understanding the calculus notation. The frequency of having a programmable error would occasionally occur because after a specific time, programs would need to be updated to manage different concepts of questions. The size of the issue will depend on the expected outcome from the user, what this means is that the issue of not showing the full calculation method is a smaller issue compared to a bigger issue of the chatbot not being able to answer simultaneous equations.

Over the course of time, the evolution of chatbots have provided usages to different users. Some users prefer new models of chatbots compared to the models that have been previously created (Jovic, 2022). Even though there are chatbots that have been created in the past, there is always room to invent better versions. To examine the idea of building better versions of chatbots, a specific field of chatbot would be the ideal target point as the intended field of chatbots have had less advancements.

Selecting Chatbot

The reason for selecting a chatbot is that in the field of artificial intelligence, chatbots have the potential to further the advancements in both AI and higher-level mathematics (Greenburg, 2020). What this means is that mathematics is a straightforward topic in which chatbots can easily follow and execute. The fact that there has not been any indication that an AI chatbot dealing with mathematical notations on the horizon, this is the opportunity to develop one and evaluate if an AI chatbot with an NLP engine has the functions to digest mathematical notations.

The difference between the created chatbots online and this version is that it will be using NLP to tackle mathematical questions. Reason for being a difference is that the online ones only use math libraries as there mode of process when handling questions from the users. Another reason is that this version specifically targets the higher-level topics (Trigonometry and logarithms) whereas the online version tackles the simpler topics such as timetables and fractions.

Aim

The problem is that NLP chatbots have little experience with mathematical notations as they have never encountered the notations before, the focus of this research project is chatbots digesting mathematical notations from college-based exam questions using natural language processing. Natural language processing is an AI technique used in chatbots, the process behind NLP is that the chatbot compares the dataset from the text file to the user's question and outputs data to the user's request. NLP is important to chatbots because the AI technique is vital to the chatbot system. What this means is that NLP plays a crucial role in matching datasets and displaying the correct information. Chatbots are famous for using natural language processing as the chatbot stores information about any topic and as the user asks a question, the chatbot will fetch the data from the text file and output an answer (towards data science, 2019). The aim of this research project is to answer mathematical questions, the idea behind a mathematical question is that the mathematical question provides a challenge to the chatbot. For example, **find the value of** $\boldsymbol{log(b^5) + log\left(\frac{4}{b}\right)}$ is an example of a mathematical question as this question enables the chatbot to take the coefficients of the question (in this case, the chatbot takes the $\boldsymbol{logb(b^5) + log(\frac{4}{b})}$) and preform the calculation. After the chatbot calculated the question, the chatbot will output the answer.

The second focus is to compare the accuracy with a MATH library chatbot. The reason for comparing the accuracy is to determine if NLP's answering system matches a preprogramed mathematical tool.

Scope

NLP chatbots are not being challenged with mathematical notations from college-based exam questions. The target of this thesis is using A-level maths exam questions to evaluate the NLP chatbot's ability to digest mathematical notations.

The scope of the thesis is restricted to past papers and mark schemes from college-based exam boards. The selecting process will have a maximum duration of 3 weeks and the process will end after the 3 weeks have expired. Each mark scheme will be used in the feedback sheets to judge how well the NLP or MATH library chatbots answered the question.

Hypotheses

The research question is Does an NLP chatbot have the ability to digest mathematical notations? Accordingly, the thesis hypothesis is as follows:

H0: NLP chatbots can easily digest mathematical formulas and notations than a MATH library version.

H1: There will be a difference in the accuracy rates amongst the MATH library chatbot and an NLP chatbot.

Research Contribution

The contribution in this thesis is to show the advancement of NLP being able to solve mathematical questions from higher level education. An accuracy test will be used with the higher-level mathematical topic's trigonometry and logarithms will determine the measurement of how well NLP can solve questions. A MATH library chatbot was also created in this thesis and the experiment was that both chatbots will be asked a serious of mathematical questions from the selected topics to analyse which chatbot answered the most correct questions. In order to obtain statistical significance, there will be a three-column system labelled : **correct**, **failed** and **not able to answer.** What this means is that the column labelled correct will highlight the questions that both chatbots got correct, the failed column will highlight the questions where both chatbots got the answer wrong and the column labelled not able to answer highlights the questions that gave an error message saying that the chatbots were not able to give an answer (Feyiseye, 2019). The difference between a failed answer and a not able to answer is that a failed answer is an answer that is irrelevant to the asking question (Dictionary, 2023) and a not able to answer response is when the chatbot throws an error message to show that it was unable to find an answer to the question. The statistical significance of the experiment will come from the correct answers that NLP has answered in comparison with MATH library.

Chapter 2: Literature Review

2.1 Introduction

For many years, artificial intelligence has made little contribution to interpreting mathematical equations because AI does not have the correct tools to help ease the mathematical issues such as digesting long equations (University of Cambridge, 2022). Chatbots are a solution to this issue because chatbots enhance the capability of correctly calculating equations (Dsouza, 2021). What this means is that AI chatbots can help mathematicians reduce the workload such as having the chatbot solving the calculus equation whilst the mathematician solves the circumference of a circle. From that point, chatbots have been accomplished the target of solving higher level mathematical problems. The chapter starts with a discussion about the nature of a chatbot and clarifying how this is different to a calculator as some users would argue that a calculator can overwrite the capabilities that a chatbot can do. The reason that a calculator can overwrite a chatbot is that the calculator itself can take a shorter time to calculate an equation than a chatbot. For example, the equation $log3(7-3)$ was asked on both a chatbot and a calculator the calculator would have a slight quicker process in answering the equation than a chatbot. This is because the calculator is a device that has been created to responded within seconds. The next chapter discusses the concept of a good, bad, and sensible question and how they may affect the outcome of a chatbot. Then, the chapter continues with the background of the thesis and then discusses the types of algorithms that a chatbot currently possess and examining each algorithm. We then talk about NLP and discuss in detail the problem NLP is having with mathematical notations and propose a solution with the math library. The chapter closes with a summary of the discussed ideas mentioned in the previous chapters and how a specific algorithm will be selected for the project.

2.2 Chapter Background

The ongoing problem that NLP is facing with mathematical equations is that it has not encountered notations and symbols of a maths question (the sequence, 2021). NLP not being able to digest mathematical equations highlights the problem as NLP has no experience with mathematical equations. Calculus integration symbol, arithmetic notations and algebraic notations are the examples which imply NLP's incapability with higher level mathematical equations which leads to serious problems in the ongoing future. Currently, NLP can take a sentence or a paragraph and transforms the letters and words into specialised numerical values for the processing phase (Gomathy et al 2022). Afterwards, the specialised numerical values turn back into words and sentences the chatbot can output the answer.

The intention of getting NLP to solve mathematical equations would be an advancement as AI scientists can rely less on pre-programmed maths modules and have more faith in NLP. For example, NLP solving a calculus question would give AI scientists a big relief because the scientists would need to code in the calculus question whereas NLP allows the traditional written method. Companies that specialise in higher level mathematics would be invested in the chatbot as the capability of NLP solving higher level maths equations would be a huge persuasion (Daggett, 2022). Not only would the chatbot bring more help to employees who are using it, but the positive feedback would mean that the chatbots can be sold commercially to other companies that require a chatbot solving higher level maths issues.

Even with the chances of selling mathematical chatbots commercially, it will still take a long time due to a long maintenance work on the chatbots. What this means is that chatbots are still new to the field of mathematics and with the assumptions that mathematical calculations are easy for the chatbot to follow, the notations of the equations will be an issue. For example, if the chatbot were

to solve the equation: **Hence find** $\int \frac{8sin0}{5+tan^3}$ **60** the integration sign would throw the chatbot off its process at first because it has not encountered the notation before.

2.3 NLP

Natural Language Processing is a keyword recognition chatbot as both AI tools mirror each other very well in terms of functionality. NLP have been popular because computer scientists consider this algorithm to be straightforward to program and understand. Sangeetha et al 2021 states that NLP can understand international languages, which is a huge asset to educational institutions as the chatbot is able to communicate with students who do not speak English properly.

As well as NLP being able to break language barriers, the algorithm itself has been a success in the field of mathematics. Deborah Ferreira et al 2020 paper mentions that NLP accepts mathematical text from a user and the algorithm can output findings like the user's input. Even though Ferreira's statement is none different that previous researcher's statements, the idea that NLP allows mathematical text-based data is a big step in the field of mathematics. What this means is that NLP is starting to recognise mathematical text language such as equations and formulas. In addition to NLP being able to accept mathematical terminology, NLP has been identified as being able to recognise student writing (Scott A. Crossley et al 2020). This statement implies that human teachers are not able to understand the wording from the student's response to either an exam question or a quiz question. Because of the tools that NLP possess, the students can submit their answers to the chatbot and the chatbot can convert the writing to text and use NLP to check if their answers are correct. In addition to NLP being able to highlight written text, NLP has been able to deal with specific formats of mathematical notations. Sean Welleck et al 2021 has stated that there was an evaluation for natural language processing to solve multiple choice algebraic problems. With NLP under the evaluation for algebraic problems, the field of AI is getting one step closer to handling mathematical notational problems. If NLP is capable of handling algebraic problems then, the solution to solve other mathematical notational topics such as calculus will be developed sooner than expected. Other mathematical topics like linear algebra and vectors have been present with NLP (Widdows et al 2022). What this means is that NLP has been able to welcome another type of algebra and another maths topic that consists of mathematical notations. With NLP being able to adapt to newer mathematical notations, it shows that AI are evidently on the fast track to conquering the higher level of mathematics.

Building on the premise that NLP has started to get head around mathematical equations, other studies have confirmed that chatbots utilizing NLP have had a significant impact of student's examination performances in topics such as algebra (Kai-Chih Pai et al 2021). This statement solidifies the advancement that both chatbots and NLP has arisen to as NLP is now understanding the mathematics rules for how each equation operates. For example, if NLP were introduced to a calculus equation NLP would then understand the mechanics behind the calculus equation upon how to solve questions based around the equation.

Throughout this chapter the thesis has highlighted the nature of NLP, the issue NLP has with mathematical notations and exploring how mathematical notations can be translated for chatbots. With the issues and difficulties highlighted, the next section explains in depth the issue NLP has with mathematical notations.

2.4 Translation of the Notation of Mathematics in Chatbots

The reason that maths notations such as $\int$ has been an issue in chatbots is that maths symbols does not have a special symbol to transform into during a process. This is main issue with NLP – the engine that majority chatbot's run on when taking in questions from users, the reason that mathematical notations is an issue for NLP is that it has never encountered mathematical notations before. For example, numerical values like 1,2,3 and letters like x, y, z all have special values within AI that will allow the chatbot to allow the process stage to commence. Lomesh Mahajan et al 2022 proposes that using equations such as $f(x) = 3x^3 + 2sinx - \frac{4}{x^2}$ is the starting point in solving this issue. This is because AI scientists can use this information and further investigate newer methods to allow chatbots to manage maths notations.

The achievement for capitalising mathematical notations would require a quicker waiting time than expected. The idea stems from the fact that AI scientists have already developed similar methods to overcome a portion of this issue. Andres Zarza Davila 2021 issued that Unicode have been a little advancement in computer science with notations such as ∃. With some notations being able to have a special Unicode symbol, AI scientists can act on this discovery and be able to develop more Unicode symbols for mathematical notations. On the other hand, there is much work to be done with NLP accepting mathematical notations as the computations are more complex. The underlining issue with NLP is being able to understand the notations of mathematics, an example of mathematic notations is $\int$. The main question in this thesis is what issues NLP are is having with mathematical notations. The next section goes into detail as to the real reasons NLP has an issue with mathematical notations.

2.5 NLP issues with Mathematical Notations

Even though Unicode has the possibility of developing symbols based upon mathematical notations, the Unicode itself would have a challenging time communicating with NLP. Reason being is that both NLP and Unicode have unique encrypting methods that allow the datasets to change their format whilst a program is running. Antreas Dionysiou et al 2021 claims that the exchange amongst Unicode's would be a breach in the usability aspect, Dionysiou makes a compelling case as the program would return many error messages. In addition to having more errors, if the user asked a mathematical question it would only be formed in the context of a sentence rather than a context of a mathematical question (Acharya et al 2022). For example, if this calculus question was shown to NLP:

$$\frac{dy}{dx} = \frac{12x^2 + x - 16\sqrt{x}}{4x\sqrt{x}}$$

the NLP program will form the question into the sentence:

dy/dx = 12x^2 + x - 16 sqrt x/ 4xsqrt x.

This is the main issue NLP is facing as the program is not recognising the mathematical notations which is causing the NLP is output incorrect answers.

On top of the translation with Unicode, NLPs remaining issues with mathematical notations circulates around the design of their scripts and mathematical identifiers (Pankaj Dadure et al 2022). Dadure's statement suggests that the design of NLP scripts have not been planned out well which is causing both issues and delays with current projects. The fact that the mathematical identifiers have signalled a cause of concern indicates that the identifiers are out of date, causing a big halt with mathematical AI projects. In addition to having old identifiers, the current tools that NLP obtains

would not have the ability to divide the big equations and notations in specific mathematical topics such as trigonometry (Sophie McIntyre, 2021). McIntyre's statement is true because NLP is comfortable digesting textbook information as NLP provides the tools to deal with the words and paragraphs. The fact that McIntyre highlighted the main mathematical issue within NLP is a clear indication that AI needs to step up and act on these issues. In addition to McIntyre's statement, there has been further accusations that NLP issues with mathematical notations is all to do with the limited training experience NLP has been receiving (Faldu et al 2021). What this means is that NLP has been conditioned to cooperate with the English literature, mathematical notations are a newcomer for NLP as the notations are datasets that NLP has not yet mastered. Words like computers and computer is an example of what NLP can manage for searching sentences but, notations such as $\sqrt[3]{7}, \sum$ are unrecognisable for NLP. To summarise both 2.4 and 2.5, the translation of mathematical notations has little tools for chatbots and NLP issues is based upon the lack of experience with mathematical notations and symbols. Amongst NLP and MATH library, which method digested the notations well?

2.5.1 NLP Breaking Down Maths Questions

As mentioned above, NLP has not much experience when it comes to mathematical notations. However, when faced with mathematical questions NLP has obtained a process for handling these types of questions. Mathematical word problems are the closest types of questions NLP has experienced whenever this topic cross its path (Raiyan et al 2023). The process that NLP uses to break down mathematical word problems is that it takes the question as a whole, then it breaks the question down into pieces known as tokens (Berbatova et al 2023). Each token is then passed from the input system to the main function, in the main function NLP extrapolates the appropriate tokens that corresponds to the question being asked. NLP then matches the tokens from the questions with the data that is being stored in their database and if both the tokens and data match each other then NLP will output the answer (Arivazhagan et al 2023). Is NLP better than MATH library for accurate answering of the selected topics? The next section will discuss more about how the input system works and how the mathematical calculations occur in NLP.

2.5.2 Input System and The Operations of Mathematics

As briefly mentioned above, the input system divides the question up to appropriate amounts. Let's say that there was a trigonometry question : Solve the equation $sin0 * tan0 + 2sin0 = 3cos0$, NLP would carefully divide up the load into small tokens which will look like : Solve|the|equation|:|$sin0$|*|$tan0$|+|$2sin0$|=|$3cos0$ (Patil et al 2023). After the question gets divided up into the tokens, they are then sent through the mathematical operation of NLP. In this phrase, what NLP is doing is that it takes the tokens that is requires in order for it to complete the task of answering the trigonometry question. What this means is that NLP will extrapolate the $sin0 * tan0 + 2sin0 = 3cos0$ from the question and then starts the calculation process. This is where NLP calculates the question by comparing the coefficients to the stored dataset and if the stored dataset matches the coefficients, it is a match (Lin, 2022). And once the match has been connected, NLP will fetch the data and output the answer. This process can be thought of as a scanner scanning a barcode item and if the two items match it will output the data on the screen to confirm the findings.

As the issues have been explained, the next section proposes a solution to the ongoing issue.

2.5.3 Possible solutions

As mentioned before, there have been some success with NLP getting a handle with mathematical notations. One of the successes orbits around a method called MATH which is a group of data that includes answers and solutions (Dan Hendrycks et al 2021). Hendrycks' accomplishment does give a starting point for this issue as his team have introduced a dataset that enables the NLP program to channel and calculate the questions.

In addition to using MATH, there has been other breakthrough in which a model uses a transformer to digest the mathematical notations (Kimia Noorbakhsh et al 2021). Noorbakhsh and their team have stated that the transformer uses a technique like NLP that allows the notations to be converted into a sequence. Applying a transformer does in fact mean that AI is taking another step towards the advancement of tackling mathematical notations as the scientists have managed to include other resources.

As well as both the transformer and the MATH library, MathBERT is another solution that could possibly be used for this issue (Zhang et al 2022). To explain this, MathBERT operates like MATH library in which this tool also has pre-programmed formulas and equations (Liang et al 2022). Having another pre-programmed maths tool for the chatbots is positive because it would mean that the MathBERT would have the latest equations and formulas. It would also mean that promise of a chatbot digesting mathematical notations is in working progress because it would mean that, currently the MathBERT can digest equations for Pythagoras and the Fibonacci sequence. This is the building block that can easily be modified to tackle bigger topics like calculus and logarithms.

2.5.4 MATH library chosen over MathBERT.

Amongst the MATH library, MathBERT and the transformer the next section will further investigate into the MATH library and discuss the positives and negatives of this tool. The decision to focus on the MATH library stems from the principle that this tool is slightly better than MathBERT in terms of functionality (Meadows et al 2022). What this means is that the MATH library possesses tools that have been said to combat topics such as trigonometry and logarithms whereas MathBERT has not reached that level of solving bigger topics. An example of this is that the MATH library is able to manage equations like $log4 - 6^2$ and $sin30tan50 + 5sin0$ whereas MathBERT is not able to do that. This is important to the study because the MATH library tools will challenge NLP and see which AI method has a better accuracy for answering higher level maths questions.

2.5.5 MATH

MATH is a python mathematical library that consists of recorded mathematical codes. AI scientists favour these recorded codes because it minimizes the workload for them to conduct experiments (Liyanapathirana, 2022). For example, if an AI scientist wanted to use a logarithm equation the scientist would import the math library into their program and have the library use the equation. The MATH library is different to other types of maths algorithms because the library consists of mathematical formulas that have been pre-recorded. Vayadande et al 2022 has confirmed that the math library is subjected to performing calculations.

In addition to performing calculations, it has been said that the math library possesses a tool called sympy that can be used for topics such as calculus and algebra (Rajagopalan, 2021). With sympy at the MATH library's disposal, it would mean that the library itself is able to contribute towards the field of mathematics. The sympy tool would also provide an easier procedure to develop other equations for other topics of mathematics such as logarithms. However, each programming tool does have its errors and sympy is no different as sympy rejects any coefficients that consist of a negative value like n^{-y^6} (Abumosameh, 2022). This is an issue that AI scientists need to focus on as

most mathematical equations contain a negative coefficient therefore, the sympy tool needs to be altered slightly to become a tool that can evolve the mathematics world. Even though the math library can perform calculations, NLP has a better history even though NLP does not have much experience with mathematical equations.

2.5.6 NLP more suitable than MATH library

NLP has an algorithm called monomial naïve bayes which is an extremely popular algorithm used amongst chatbots. Reason for its popularity is that the algorithm takes the user's input and compares it with the data stored and outputs the result (upGrad, 2022). The comparison procedure can be applied to mathematical questions as the user can simply type in their question to be answered. In comparison to the math library, the library requires the user to know some coding skills to work. For example, if a user had a trigonometry question and used the math library the user would need to know how to code trigonometry equations.

Another reason is why NLP is better than MATH library is that NLP does not limit itself to the type of information being processed. What this means is that NLP can answer a calculus question and the math library cannot because the library has not reached that level of higher mathematics (Jain et al 2022). This implies that the MATH library does not possess the tools to manage topics such as calculus which causes the MATH library to provide no resources for calculus.

In addition to having no limits, NLP is engineered with a tool called a transformer deep learning model. Which is said to be able to intake a big data set and study the given dataset in order to output the appropriate answer (Zong et al 2022). The transformer deep learning model is an additional reason NLP is better than MATH library, this is because the transformer is able to intake long equations and formulas from the maths questions and have the transformer study the information before performing calculations and solving the problem. MATH library does not have this type of tool which may cause some of its answering functions to have errors such as not calculating the logarithm question properly. Which topic challenged NLP and MATH library the most?

The debate on NLP is better than math library has been stated, our attention diverts back to chatbots. The word chatbots has been mentioned numerous times but, there is no clear definition as to what a chatbot is and how the issue of mathematical notations links to them.

2.5.7 Accuracy of Chatbot

In this Morden day and age, chatbot's accuracy reputation has become more important. This is because chatbots have been bragged about having an accurate outcome to any answer that has been sent to them (Chang lin et al 2023). NLP is the system in which chatbots are having its accuracy fame being mentioned the most because this system is able to accurately respond an answer to the user's question (Hirosawa et al 2023). Mathematical topics are the types of questions that users are asking chatbots and have reported that the accuracy of the answers have been dependable (Shahriar et al 2023), this implies that the accuracy of chatbots answering mathematical questions has been a positive outcome. This leads to a big investment in chatbots because of the fact that many users have responded positively to answering of their questions. Overall, which chatbot produced a positive accurate rate?

The next section clarifies the definition of a chatbot and explains the link towards the ongoing issue of this thesis.

2.6 Chatbot

As mentioned before, a chatbot is an artificial intelligence program that can hold conversations with a human being. The problem at hand is that chatbots do not have any experience with mathematical notations. The proposed solution is to create an NLP chatbot that can digest the notations to solve the inputted questions. Others may point out that a calculator can do the same thing as the device is able to take mathematical notations and coefficients and output answers (Alnfiai, 2022). The next section will discuss the difference between the two creations.

2.7 Types of chatbots

Artificial Intelligent chatbots come in six different versions: Rule-based, Machine Learning, Data driven, Information retrieval, Keyword recognition-based and the hybrid model (Engati Team, 2021).

With the names of each chatbot method, machine learning is a concept that AI scientists lean towards when the thought of mathematical chatbots come to mind. Mathematical equations should be a straightforward dataset that a machine learning algorithm can follow as each maths equation has a specific procedure to follow. Because the maths equations and machine learning algorithm follow a specific repetitive set of instructions, the chatbot will find the process simple and easy to understand. In comparison with rule based chatbots, this type of algorithm requires hardcoding the chatbot to solve a maths equation from the programmer. Hardcoding is forcing a program to perform a specific task without the user having the option to choose what the program does when the program runs. Currently, chatbots are operating with button based and voice algorithms to solve maths problems. This is due to the simplicity of these chatbots and the fact that more users are using apps with causes the button chatbot to contain a high popularity.

2.8 Chatbot Algorithms

There are six main algorithms that have made chatbots establish their popularity: machine learning, rule based, data driven, information retrieval based, keyword recognition based and hybrid. Further details surrounding these algorithms will be mentioned in the ongoing sections.

2.8.1 Rule Based

Ruled based chatbots are essentially chatbots that have been forced to answer questions from the user (Jagdish Singh et al 2019). For example, the chatbot program will be forced to say **"Hello, I am a virtual chatbot. How are you"** if a user enters **"Hi/Hello/Greetings".**

Because the ruled based chatbot is limited to answering specific data, some users find working with a ruled based chatbots challenging. The researchers (Krishna Kumar Nirala et al 2022) have stated that due to the ruled based chatbot only accepting specific data from the user, AI scientists confirmed that ruled based chatbots are a challenge to work with today. Nirala has also stated that ruled based chatbots will become obsolete because AI scientists have reached a point where the AI scientists are not able to adapt the ruled based chatbot into a greater improved version. Ruled based chatbots would have catastrophic issues when dealing with mathematical equations. This is due to the chatbot only recognising how to solve one equation and not being able to solve another equation. For example, the ruled based chatbot was able to recognise Pythagoras Theorem equation for finding the hypotenuse but did not understand a logarithm equation. Joshua Grossman et al 2019 confirms the concept of ruled based chatbots not being able to solve many mathematical equations. Grossman's paper illustrates the concept that the creator of the rule based chatbot would need to rope in other technicians to the ruled based chatbot to solve other maths equations. As well as a ruled based chatbot being limited to outputting data, data driven chatbots have been making their statement as this type of chatbot is able to complete tasked jobs.

2.8.2 Data driven Chatbot.

Data driven chatbots main motivation is to complete daily tasks. What makes this chatbot different from the competition is that data driven chatbots are said to be like virtual assistants (Indicative, 2022). Data driven chatbots can also be thought of as voice chatbots as these chatbots can process information through the sound of a user's voice. Data driven chatbots would be extremely useful for students as they can verbally ask the chatbot their question and in response the chatbot can verbally reply to them (Mafra et al 2022). In addition to data driven chatbot being able to help students, the chatbot has been able to help students with special needs such as dyslexia (Kuhail et al 2022). This again highlights the fact that data driven chatbots can cover a wide range of learning issues. Having a data driven chatbot is good, information retrieval based chatbots are better due to working with databases.

2.8.2.1 Similarities and Differences between Ruled based chatbots and Data driven chatbots.

The big difference between ruled based chatbots and data driven chatbots is that ruled based are based on expectations created by the programmer and data driven can be seen as voice chatbots (Remus, 2022). Ruled based requires the user to type in their queries whereas data driven chatbots allows the user to verbally input their queries via microphone. Data driven chatbots has a wide range of information that can be extracted for the user's queries and ruled based chatbots are restricted to answering queries from the programmer's input (Labaska et al 2007).

The only similarity that both chatbots have is that they are an evolution from one another. What this means is that after a certain amount of time since the ruled based chatbot was invented, there were speculations as to getting a chatbot to verbally accept data as well as typing the data. This then caused the data driven chatbot to be created.

2.8.3 Information Retrieval Based Chatbot

Information retrieval based chatbots are used to created pre-programmed responses from databases (Fainchtein, 2020). The process that the information retrieval chatbot uses is that the chatbot is trained using text-based datasets and as a user requests data, the chatbot would find the data like the user request (Zaid et al 2022). Information retrieval chatbots are mostly used in areas like education and psychology (Xu et al 2020) as areas like these would create pre-programmed responses to help the needs of the user. It is important that the information retrieval based chatbot selects the appropriate information for the user's request otherwise the chatbot would receive a bad reputation (Zhu et al 2021). Having an information retrieval based chatbot is a handy tool but, a hybrid chatbot is more promising as this tool is combining two different chatbots as one.

2.8.3.1 Similarities and Differences between Data driven and Information Retrieval

The main difference is that data driven uses an internet type database for getting data and information retrieval uses a database filled with responses that the programmer created. As said before, data driven uses microphone to input text and information retrieval uses the classic keyboard system to input text.

The similarity between these chatbots is that both use a type of database to search for information.

2.8.4 The Hybrid Chatbot Model

As mentioned before, a hybrid chatbot is a multitasking chatbot. The word "hybrid" emphasises the evolving era of chatbots as AI scientists have discovered a method of merging two separate chatbot into one. The ability to refresh itself with updated content is an ability unique to hybrid chatbots (Kevin Jetten, 2021) because each time a user interacts with the chatbot, the special ability can take in the conversation and renew itself with better information.

Hybrid chatbots would be of an interest to college mathematical students as the students would ideally like a chatbot that would upgrade itself with latest information. Gabriel Edrick Acuna et al 2021 mentions that AI scientists are aiming to create a hybrid chatbot for students to use. This is statement emphases that AI scientists are on their way with developing stronger, better chatbots that college maths students can use for their college work or revision work. With the promise of developing more hybrid chatbots, it will take a long time to role the chatbot out to public as AI scientists will have to conduct multiple tests to perfect the outcome. Meanwhile, machine learning chatbots have been making a name for themselves as this type of chatbot is able to study the delivered dataset by itself.

2.8.4.1 Similarities and differences between Information Retrieval based chatbot and Hybrid chatbot.

The main difference is that hybrid chatbot is a combination of two chatbots merged as one and information retrieval chatbot uses a database for its functionality. Information retrieval chatbot has pre-programmed responses for the user to ask and hybrid chatbots are able to simultaneously complete multiple tasks.

There are no similarities amongst the two chatbots because the hybrid version is more complex and more in-demand for consumers than the information retrieval-based version.

2.8.5 Machine Learning Chatbot

Machine Learning chatbot is self-learning chatbot that can answer questions that the user asks. Sagar Badlani et al 2021 explains that machine learning is trained to undertake specific data and output the correct data when requested. Badlani's paper also highlights machine learning being categorised under supervised learning, which is an AI skill which allows the chatbot to accurately match the input data from the user to the output data that the chatbot currently stores (David Petersson, 2018).

Because machine learning primarily functions with supervised learning, datasets such as facts about computer science would be a straightforward concept for the chatbot to follow. The requirement that the chatbot would need is to simply match the correct data between the user's input and the stored output. Anagha Shenoy et al 2022 supports the statement of how supervised learning in machine learning operates. Shenoy also mentions that machine learning increases the chatbot's ability to examine the user's input to output accurate data. This is due to the chatbot learning the user's different inputs and overtime the chatbot would have built up a better experience. As well as machine learning being able to recognise patterns in a dataset, the ability for machine learning to solve mathematical equations is on the horizon as machine learning is iterating the same skills as before. Gwo-Jen-Hwang et al 2021 have mentioned that AI scientists have made machine learning solve maths problems a priority, this statement highlights the fact that AI scientists have acknowledged the lack of focus on chatbots solving mathematical problems within machine learning. As mentioned before, chatbots would have a quicker time processing the mathematical equations as the AI scientists would have labelled each dataset into a category that the chatbot would recognise. With machine learning able to self-conducted, keyword recognition is more popular as this chatbot can deliver information to users in an efficient manner.

2.8.5.1 Similarities and differences between Machine learning and Hybrid

The main difference between hybrid and machine learning is that hybrid requires another chatbot model whereas machine learning self learns the data. Today, machine learning is used more often than hybrid as machine learning can complete more tasks than hybrid.

2.9.5 Keyword Recognition Based Chatbot

The main procedure for a keyword recognition chatbot is that when the user inputs a question, the chatbot will scan the user's input. The text file stored within their system will compare the two words and if there are two words that match the chatbot will output the result. NLP is an example of keyword recognition chatbots as this ability has the same definition. The mathematical equations and formulas is an aspect of information that Keyword recognition can adjust to very quickly as the process is comparing both datasets and outputting an answer. What this means is that the keyword chatbot can compare the coefficients and submit the appropriate answer.

2.9.5.1 Similarities and differences between Machine Learning Chatbot and Keyword Recognition Chatbot

The similarities amongst both chatbots are that they can self-learn the data being fed into their systems (ringcentral, 2021). What this means is that, both chatbots can learn the data themselves and not have the programmer to explain the data to them. Another similarity is that both chatbots function in the same manner as both chatbots can take the inputs from the user and output the appropriate answer.

The sole difference between the two chatbots is that keyword can give an accurate response to the user's query than machine learning.

Summary

Throughout this chapter we have discussed the several types of chatbots and how each chatbot would manage mathematical topics. There was a further discussion on machine learning and keyword recognition chatbots as the main purpose of this investigation is to determine with a chatbot can solve higher level mathematical questions. We also talked about the issues that NLP is currently facing with mathematical notations and provided a solution to the problem which was to evaluate the NLP chatbot against a pre-programmed maths chatbot called MATH library and compare the accuracies.

Chapter 3: Methodology

In this section, there will be an explanation about the blueprint of the experiment. The investigation for this thesis is to evaluate if a chatbot can answer higher level mathematical questions using NLP. The blueprint will consist of an NLP chatbot and a MATH library chatbot, both chatbots will be challenged with the same mathematical topics. The reason that the blueprint will answer my research question is to evaluate NLP's answering capabilities against a pre-programmed MATH library, by doing this the hypotheses for this thesis will come to a solid conclusion on whether a chatbot with an NLP engine can process and solve higher level mathematics.

Framework

The thesis research framework was an adaptation from the original design science model. The model centres itself on the interpretation of a thesis researcher and outlines the entire thesis framework.

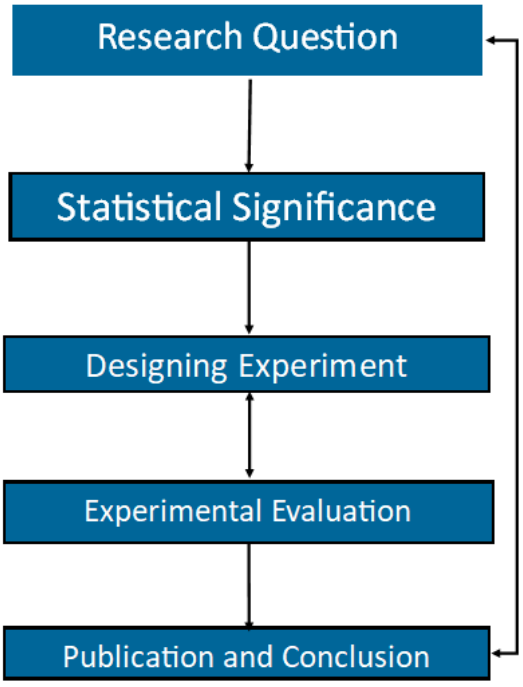

Figure 2: Design Science Model

Figure 2 shows the design science model having 5 different stages of the thesis. The first stage is clarifying the research question(s) of the thesis, the research question needs to be clearly explained and understandable for other researchers to acknowledge the purpose of the thesis.

The second stage is mentioning the statistical significance of the thesis. Clarifying how the statistical significance of the experiment will be identified.

The third stage is justifying the blueprint of the experiment. Explaining in detail how the experiment will be designed and how the experiment will be conducted. In addition to the detailed outlet of the experiment, the data sources (past papers) will also be clarified to state how the experiment will be generating the findings for the thesis.

The fourth stage is evaluating the data that has been created from the experiment. It is also important to iterate back to the design of the experiment to understand how and why the data was created. The analysis phrase also takes part in this stage and it is also where the statistical significance is highlighted to state the outcome of the hypothesis and research question.

The fifth and final stage is publishing the findings of the experiment. This stage is where the hypothesis gets the exposure of being approved or rejected based upon the outcome of the experiment. This is also where the research question fate is decided on whether the experiment proved the research question right or wrong. It is also a stage where the conclusion is drawn, where the thesis has stood before the experiment and how it has ended.

Data Sources

The data generated for this experiment are coming from England and Wales exam boards (AQA, Edexcel, OCR and WJEC). Specifically, the logarithms and trigonometry questions from the mentioned exam boards. This is because the experiment focuses on these two topics which will provide a better identification as to which chatbot is superior.

Description about Experiment

The NLP chatbot is evaluated first as this chatbot is facing all the selected topics. As soon as the main codes of NLP are running, the chatbot will ask for a question to be entered. A question randomly selected from the chosen past papers will be entered into the chatbot, the written style of the question needs to be exactly as written in the past paper. Once the question has been entered, the NLP chatbot will begin its answering process and output an answer to the question.

After the NLP chatbot experiment has ended, the MATH library chatbot begins with having its main codes running. Just like the NLP chatbot, MATH library will be subjected to facing all of the selected topics. Then, a question is randomly selected for the library to answer. Instead of typing the question word for word, the question needs to write in Python language (the programming language used in the experiment) because this is the format that the library accepts. And once the coefficients have been entered, the library will use its tools to answer the question.

The measurement for this experiment is how many questions both chatbots can answer correctly. The procedure for this measurement is that each time the chatbot outputs an answer, the decision to know if it is correct or not is based on the mark scheme from the past paper the question originated from. The reason that this is valid is because the data can easily be demonstrated as it is a clear indicator to measure the successfulness of how accurately the chatbots answered the questions.

Experiment

Designing Experiment

The experiment for this thesis is creating two chatbots: an NLP chatbot and a MATH library chatbot. Both NLP and MATH library will face both trigonometry and logarithms with the same number of questions. This is the correct design to answer my research question because the measurements are based upon both chatbots being challenged with higher level mathematics and seeing if the NLP chatbot can answer more questions that the MATH library one. Both chatbots will be monitored with how they approach both trigonometry and logarithm questions and the measurement will be determining the accuracy of both chatbots answering the questions by observing how many questions both chatbots got correct. This is the appropriate method for this experiment because it will demonstrate the capacity at which NLP is able to manage mathematical notations compared to a library filled with pre-programmed notations.

Once the NLP chatbot experiment was completed, the MATH library chatbot experiment begins. The procedure used to conduct the MATH library experiment was like the NLP chatbot, a past paper will be selected from any college exam board. From the chosen past paper, a question from trigonometry or logarithms must be selected. Once a question was selected, the question was translated into code for the MATH library to solve. For example, a past paper had the question **find the value of $log6\ 36$** the translation of the question into code would be: $\boldsymbol{math.log}\ (\mathbf{6}, \mathbf{36})$. Once the question was in code form, the MATH library chatbot can answer the question with the press of both the control and enter button at the same time. The dependent variable in this experiment is the MATH library solving the questions and the independent variable is the past paper questions.

Tools

For this experiment to work, college mathematical past papers would be required. The past papers will be available online for free. Once the past papers have been selected, the experiment will begin. What this means is that the experiment is not relying on the current year's college past papers, the experiment is branching out past years from 11 years ago. Not only does the experiment use old past papers, but it is also using all the college exam boards rather than using one exam board. Using Edexcel, AQA, OCR and WJEC provides the bigger range of collecting more exam questions than using only AQA.

Requirement Process

The past papers must be A-level mathematics and chosen from the exam boards: AQA, OCR, Edexcel and WJEC. The year of the past papers can range from 2021 to 2008 (depending on which exam board offers their oldest exam papers). From the past papers, one question from the listed topics (trigonometry or logarithms) must be chosen. The written method can be written exactly as the exam board has written the question or the question can be written differently. The MATH chatbot mirrors the exact topics but, the coefficients need to be used instead of the wording of the questions. After the question has been entered, the past paper that was selected requires the corresponding mark scheme. The chatbot needs to output a response regardless of the outcome because of the three-column system: correct, failed and not able to answer. What this means is that any response that chatbots give will be categorised by these three columns which will help the analysis of examining the results. For example, if the question solve the inequality $10x^2 + x - 2 > 0$ were to be asked the chatbot can output the correct answer of x=$-\frac{1}{2}$, x=$\frac{2}{5}$ which will be marked down as correct, a failed answer will be marked down if either chatbots gave an answer like x=4 or an error message that will be marked down as not able to answer. If by chance the chatbot does give out a partial answer, this will be marked down as a failed answer because for a question that requires more than one answer (like the question mentioned above) the chatbot needs to output both answers in order to be in the correct column.

Past Papers

AQA, Edexcel, OCR and WJEC are the college-based exam boards that will be used in the experiment. There will be a total of 100 questions for both NLP and MATH library. The process of selecting past papers is that the maths paper must contain a question from the mentioned topics (Trigonometry and Logarithms). Each past paper is required to have a complex equation for the chatbot to solve. The experiment was conducted based upon the availability which means that, the past papers would have to be free to use, the availability of previous past papers that dates to 2011 and backwards and created from a real exam board. The past papers are replacing human participants because they are available all the time, they are easily accessible online and there are numerous amounts of past papers that can be beneficial for the experiment.

Mark Schemes

In addition to using past papers for the experiment, mark schemes will also be used. The reason that mark schemes have been added to the experiment is because they will assist the outcome of each answer the chatbot gives. What this means is that, in the event that the chatbot throws an error message this will automatically be marked down as a not able to answer without question. But, if the chatbot does give out an answer the mark scheme can double check the chatbot's answer to clarify if the answer was correct or failed to answer. The double checking is a vital for the experiment because the chatbots can output a different trigonometry answer to a trigonometry question, this can cause many false positive answers to arise which are prohibited in this experiment. For example, the question solve the equation $1 - cos3x = \sin^2 x$ was the question that was asked to a MATH library chatbot and the chatbot outputted $sin\ x\ = \frac{1}{4}$ the mark scheme can double check the answer and declare that the MATH library chatbot has failed to answer this trigonometry question.

Proxy of Measurement

Past papers from college exam boards will be the proxy measurement for this thesis. The reason behind this decision is that the past papers will provide the proper testing tools to examine the chatbot's answering capabilities. What this means is that with human participants, they would take a long time trying to produce a question to ask as well as finding out if they are eligible to do the experiment. With past papers, not only are they available all year round they provide crucial tools such as testing both chatbot's answering functions and examining each word of the question carefully (Wilkinson, 2021). These are essential in the experiment as the selected topics will measure the success rate of answering questions. Not only that, but the testing routine will also allow both chatbots to get to grips with the type of questions that are going to asked before the experiment begins (Sophie, 2019). By doing this, both chatbots can familiarise itself with type of questions being asked and begin to understand the calculating process to obtain a better measurement.

Measures

The measurements for the success of the chatbot solving an A-level maths question will be broken down into three questions. The first question will ask which chatbot was used in the experiment, the reason for using this question is to monitor the which chatbot was used in the experiment. The second question asks which of the selected topics (Trigonometry and logarithms) was chosen. The reason for using this question is to highlight which of the selected topics the chatbot solved. The third and final question will simply ask if the chatbot answered the question, this question will have three options: correctly answered, failed to answer and not able to answer. The reason that this is a valid method for measuring the experiment is that the responses are nominal (lecture1,2022). What this means is that the three options can conclude on questions that the chatbot got correct, failed to answer or not able to answer (Nduwu, 2020). Not only that but, the data itself would be presented in an understanding way other researchers would acknowledge outcome of the experiment.

Examining Techniques

Accuracy, Precision, Recall and F1 are the types of potential examining techniques that can be used in the analysing phrase for this experiment. The reason that these 4 techniques have been selected is that each technique provides a unique outlook on how the to analyse the results from the experiment.

Accuracy

Accuracy is a self-explanatory technique. This method evaluates the on the dot aim for when a chatbot answers a question. The reason that this technique is in the race is because this technique will provide a better analysis for the experiment as it will dive into how accurate the chatbots were with the questions. Another reason is that allot of the time on the news and in the media, the concept of accuracy amongst chatbots has always been brought up and using this technique will correspond to the information being told in both the news and media.

Precision

Precision is a method in machine learning which guesses the chances if a false positive outcome is likely to occur (pathmind, 2023). What this means is that if there was low precision, it would indicate that there is something within the system that requires attention such as an updated library. If there was a high precision, then these concepts can be silenced as they can be interpreted as false alarms.

Recall

Recall is another method in machine learning that calculates the positive data from an experiment (C3.ai, 2023). This method is lowering the cases of false negatives and making sure that all the datasets are all positives. This would mean that the datasets have no false negatives but could consist of having false positives.

Difference between Precision and Recall

The difference between Precision and Recall is that Precision calculates the likelihood of a false positive scenario. And recall only calculates the positive data from the results and makes sure that the datasets have no false negatives.

F1

F1 is the final technique that measures a model's accuracy, it does this by combining the scores from precision and recall (Kundu, 2022). This technique also used to predict the number of times a model can make a positive prediction. This technique can only be used if the datasets have an equal amount of samples.

Why was Accuracy Selected?

After stating the reasons for the 4 techniques, I have decided to use accuracy as the analysis technique. This is because accuracy is more reliable as this technique reveals the true meaning of the experiment. It highlights the evidence of NLP chatbots in particular being accurate in answering mathematical notational questions.

Examining The Accuracy of Experiment

The accuracy method for examining the experiment will be observing the chatbot response volume (Visiativ, 2023). What this means is that the interaction between the mathematical question and the chatbot answering them will be observed to identify the accuracy of the chatbots answering of the question. The number of questions that the chatbots correct will highlight the accuracy of the experiment because it provides concrete evidence that the chatbots was able to accurately answer the mathematical questions.

Criteria for Accuracy

The criteria is that the chatbots must output an answer to the question being asked. And that the mark scheme will determine whether the answer the chatbots gave was correct or not. What this means is that, if the chatbot gives out an error message this will be marked down as **not able to answer** which will determine that the chatbot was not able to accurately answer the question. The

difference between **not able to answer** and **failed** is that the **failed** criteria highlights that the chatbot misunderstanding the question and outputting the wrong answer. And for the answers that the chatbots get right, those responses will fall in the **correct** category which highlights the accuracy of the chatbots success rate.

Metrics for the Accuracy

The metrics for the analysing the response volume is goal completion rate (Freshchat, 2023). The reason for implementing this type of metrics is to evaluate the goal of both chatbots answering mathematical questions. This is a sufficient method because the data generated from the experiment can be shown in an understandable manner, meaning that the data can identify which questions were accurately answered.

Validity

The type of validity used in this thesis is content validity. The reason for choosing this type of criteria is that the experiment can only use the selected topic questions and any other questions will impose an issue. What this means is that for this experiment trigonometry and logarithms are the selected topics for this experiment. Which means that both chatbots can only answer these questions to measure the accuracy of their answering capabilities. Topics outside the selected region such as timetables will not be valid because that type of data is irrelevant to the criteria. By implementing content validity into this thesis, the measurement for this experiment will be noticeably clear to other researchers. This is because they will acknowledge that higher level mathematics is an issue for NLP and content validity is a sensible and clear criterion for this experiment.

Internal Validity

The experiment is dependable because it is taking both an NLP chatbot and a math library chatbot and challenging them both with the selected topics. The measurement of the experiment is measuring how well the NLP chatbot manages and answers the selected topics that have mathematical notations. This leads to the answering of both the research question and sub questions because the measurements for the NLP chatbot experiment will provide the data to confirm if both the research questions and sub questions were proven or failed. What this means is that both the research question and the sub questions are asking if NLP can digest mathematical notations, the design of the experiment is specifically targeting those questions due to the selected topics. When the results are revealed, the data will conclude the answers to both the research question and sub questions.

External Validity

In terms of the real world, the research itself would have an ecological validity. What this means is that ecological validity states that a study can be transferred to another study that is anticipating a similar problem. My study focuses on an NLP chatbot managing mathematical notations from higher level mathematics and another researcher can transfer the fundamentals of my experiment and apply them to their research which is facing a similar situation such as NLP handling fractions.

Approach

The independent variable is the questions that are being entered into the chatbot. The dependent variable is the chatbot solving the mathematical questions. The reason for selecting the variables is because the independent variable will have questions that can be asked in different ways from different topics and the dependent variable is that the chatbot calculating the given question and outputting an answer.

Procedure

Once the past papers have been selected, the experiment will begin. Because past papers are being used in the experiment, the consent form will not be used. The Jupyter Notebook program will be loaded and once the program is loaded the NLP chatbot is then opened. After the required imports of the NLP libraries and codes have commenced, the program would reach the point where the chatbot is requesting an A-level maths question from the stated topics (Trigonometry and Logarithms) to be entered.

From the chosen past papers, one paper will be selected and the question needs be from a topic that has been selected for this thesis (Trigonometry and Logarithms). Once a question has been selected, it can be entered into the chatbot. After the question has been entered, the chatbot calculates the answer to the question. If in the event NLP outputs an error message saying "**unable to answer question"** this will automatically be marked down as a not able to answer and the experiment will continue straight after. If NLP does output an answer, the mark scheme is used to double check the answer and determine if it's correctly answer or failed to answer.

```
Solve, for -90 < A < 90, the equation 8 sin^3 A - 6 sin A = 1
Your answer is:
```

$a = \frac{1}{4}$.

Figure 3: NLP Chatbot Output

After the NLP chatbot experiment has been completed, then the MATH library chatbot program on Jupyter notebook is opened. For this experiment, trigonometry and logarithm questions are strictly used in the MATH library experiment. From the selected past papers, a question about trigonometry or logarithms must be chosen. Once the question has been selected, the question would then be translated into python (the main programming language in Jupyter notebook) code.

```
In [6]: print("The answer for show that 0 = 0.4+sin0 is",math.sin(0)+0.4)
```

Figure 4 : MATH library chatbot output

After the code has been entered, both the control and enter button were pressed at the same time and the chatbot would answer the question. After the MATH library finished answering all the trigonometry and logarithm questions, it would mean that both experiments have finished within the required period.

```
print("The answer for the question Solve log6 36 is:", math.log(6,36))
```

```
The answer for the question Solve log6 36 is: 0.5
```

Figure 5 : MATH library chatbot output logarithm question

After each question was asked during the experiment, the document below would record the outcome of the experiment. This survey type feedback sheet will record which chatbot was used in the experiment, the selected topic for the chatbot to answer and if the chatbot got the answer correct, failed to answer or was not able to answer.

Chatbot Outcome Sheet

Which chatbot was used in the experiment?

NLP ☐ MATH library ☐

Which topic was chosen to ask the chatbot?

Trigonometry ☐ Logarithms ☐

What was the outcome of the chatbot's response to the question?

Correctly Answered ☐ Failed to Answer ☐

Not Able to Answer ☐

Figure 6 : chatbot outcome sheet.

The overall time for both experiments was 1 hour and 45 minutes. After both experiments were completed, the chatbot outcome sheet was then filled out. The questions were based on the topic selected by the researcher. The chatbot outcome sheet will be filled out by the response of the mark scheme of the past papers. What this means is that the mark scheme of the chosen past paper is to determine the overall result of the experiment. Because the mark scheme has the official correct answer, the mark scheme dictates the official decision on whether the chatbot got the answer right or wrong. The images for the chatbot outcome sheet are in the appendix form figure 11. Once the chatbot outcome sheet has been filled out, the researcher would store the documents on the university account server for security reasons.

Summary

Throughout this chapter, the chatbot design has been mentioned and how the chatbot will be used throughout the study. The idea behind creating the chatbot was to combine the AI architecture with mathematical algorithms based on college level topics. With the creation of a mathematical chatbot, the quest to answer both the study question and hypotheses can begin. Finally, the above chapter has stated the route for discovering the answers.

Chapter 4 Results

In this section, there will be a presentation on the results from the experiment of NLP chatbot and MATH library chatbot answering mathematical questions.

Chatbot Results

The purpose of this experiment was to observe if NLP was able to have the higher number of correct questions than MATH library. In the experiment, a total of 100 questions from the mentioned exam boards and the mentioned topics were used.

Topics	Correct	Failed	Not Able to Answer
Trigonometry	2	20	3
Logarithms	3	12	10

Figure 7: Table for the answers of the NLP experiment

Figure 7 shows the results of the NLP experiment and the table shows that NLP was able to answer over 1 correct answer from both topics. Logarithms had more questions answered correctly then trigonometry which implies that NLP had slightly understood some of the logarithm questions. However, logarithms also racked up more not able to answer questions than trigonometry this implies that NLP had a hard time searching for an answer to output for logarithms. Trigonometry did obtain more failed answers than logarithms which implies that the trigonometry received answers that did not correspond to the appropriate question. This means that when a specific trigonometry question was asked, NLP must have confused the processing phrase of the question and outputted a different answer.

Topics	Correct	Failed	Not Able to Answer
Trigonometry	1	11	13
Logarithms	0	12	13

Figure 8: Table for the answers of MATH Library experiment

Figure 8 shows the results of the MATH library experiment. Trigonometry was able to record one correct answer and logarithms did not receive any correct answers. Logarithms did receive more incorrect answers than trigonometry which implies that the pre-programmed formulas for MATH library are outdated and require upgrades for future testing. Both topics in this experiment received the same number of not able to answer results, this implies that the trigonometry and logarithm questions troubled the MATH library because of the out-of-date pre-programmed formulas and equations.

NLP Chatbot Results

The bar chart below is illustrating the results from the NLP chatbot experiment. The construction of the bar chart originates from the past paper questions selected for trigonometry and logarithms. From the graph, it shows that trigonometry and logarithms had at least one question answered correctly by the NLP chatbot. With two topics having a positive outcome, the results show that NLP was able to answer mathematical questions. However, the chart also shows trigonometry having a higher number of incorrect answers as well as logarithms having the most not able to answer results.

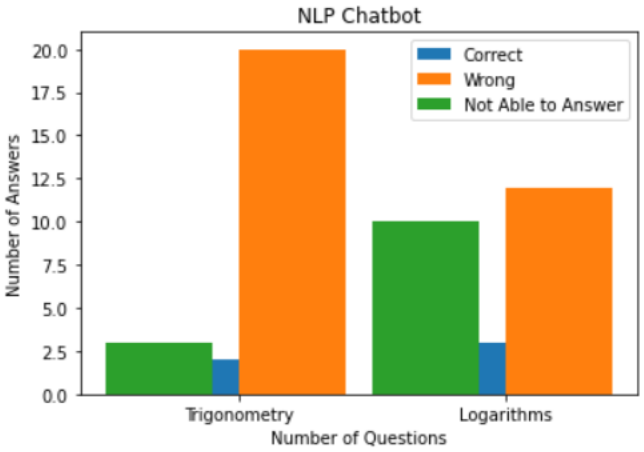

Figure 9: Bar chart of NLP Chatbot experiment

MATH library Chatbot Results

The bar chart below illustrates the second experiment which was asking a MATH library chatbot about trigonometry and logarithms. From the graph, it shows that logarithms had all its questions failed and trigonometry having only one successful outcome. This implies that the MATH library pre-programmed formulas and equations were able to manage one trigonometry question but, the tools for logarithms need further improvements. Both topics received the same amount of not able to answer results which again implies that the MATH library's resources need upgrading.

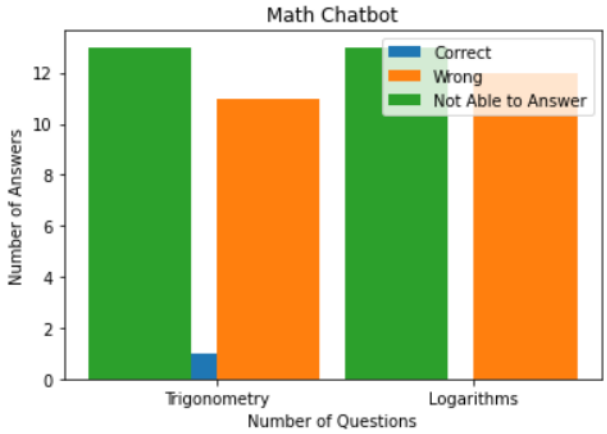

Figure 10: Bar chart of MATH library Chatbot experiment

Goal Completion Rate

In addition to displaying the outcomes of the experiment, the goal completion rate will be evaluated. The goal completion rate is a method that examines the chatbots achievement for the targeted goal which was to answer college based mathematical questions (Lishchynska, 2023). The formula for the goal completion rate is : $\frac{\text{the number of users who completed a set goal}}{\text{the number that activated a service}}$, this formula has been adapted to suit the experiment. The adapted formula is $\frac{\text{: the number of questions the chatbot answered correctly}}{\text{the number of questions used in the experiment}}$ because it is extrapolating the questions both chatbots got correct by the number of questions used in the experiment. For the NLP experiment, a total of 50 questions was used for both trigonometry and logarithms. Trigonometry answered 2 correct questions out of the 50 questions, which makes the calculation $\frac{2}{50}$ which equates to 0.04 and when multiplied by 100 the completion rate for trigonometry in NLP was 4%. Logarithms in NLP has 3 correct questions out of the 50, which makes the calculation $\frac{3}{50}$ which equates to 0.06 and when multiplied by 100 the completion rate for logarithms in NLP is 6%.

Completion Rate for NLP (%)	Trigonometry	Logarithms
	4%	6%

Figure 11: Table comparing Completion Rate amongst NLP Experiment

When both topic's completion rates are compared side by side, Logarithms beats Trigonometry by 2% because the NLP chatbot was able to answer more logarithm questions than trigonometry.

The MATH library chatbot also had 50 questions used for both trigonometry and logarithms. Trigonometry has 1 correct question out of the 50 questions which makes the calculation $\frac{1}{50}$ which equates to 0.02 which multiplied by 100 gives the completion rate of 2%. Logarithms had 0 correct answers out of the 50 questions as 11 out of the 50 failed and 13 out of the 50 were not able to answer, this causes logarithms in the MATH library to have a completion rate of 0%.

Completion Rate in MATH library (%)	Trigonometry	Logarithms
	2%	0%

Figure 12: Table comparing Completion Rate Amongst MATH library Experiment.

The comparison amongst the MATH library experiment shows that Trigonometry has the highest rate as the chatbot was able to answer at least one question. Logarithms has the lowest rate which shows that the tools for MATH library are out of date.

NLP Goal Completion Rate

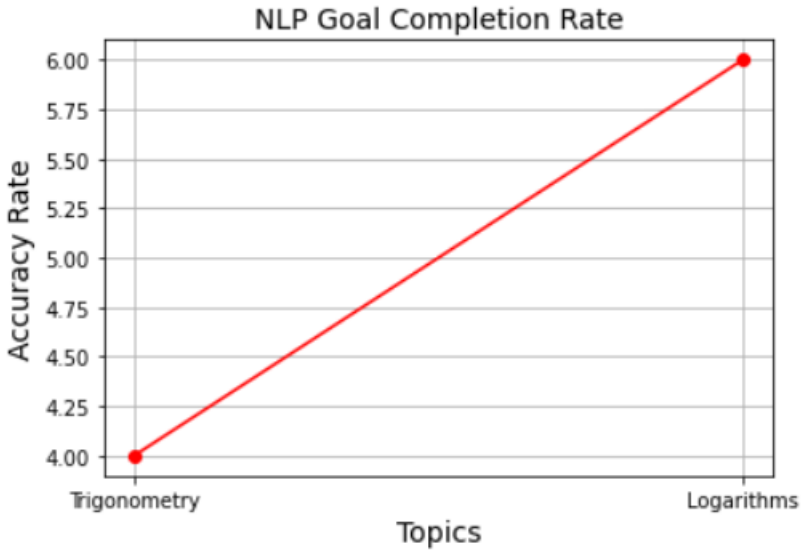

Figure 13: Line graph of NLP Goal Completion Rate

In addition to the bar charts, the second evaluation was investigating the accuracy of both chatbots. Goal completion rate was the method used to examine the chatbots response volume and figure 13 shows the NLP Goal Completion Rate being represented as a line graph. The reason for using a line graph was to imply the rate of which NLP correctly answered the questions used in the experiment. The graph shows a positive outcome as the line is drawn upwards which indicates that NLP has a positive accuracy rate.

MATH library Goal Completion Rate

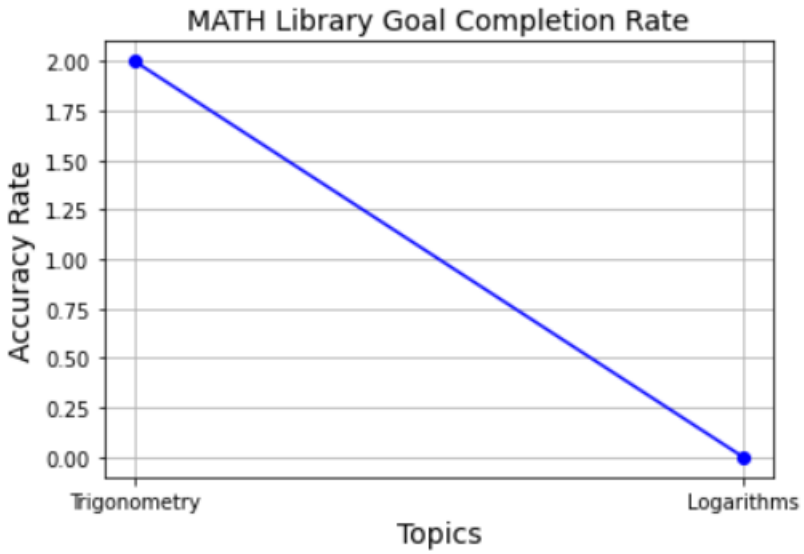

Figure 14: Line Graph for the MATH Library Goal Completion Rate

Figure 14 shows a line graph of the MATH library and having a negative rate. The reason it is having a negative rate is because MATH library was not able to answer any logarithm questions. Even though trigonometry has a 2% completion rate, the graph illustrates the line going down which shows that MATH library has a poor goal completion rate.

Summary

Throughout this section, the results were published by implementing the correct, failed and not able to answer system. The presentation of the results were displayed in the form of bar charts to illustrate how many questions from each topic were asked and which ones were successful. In addition to the bar charts, the line graphs provided a visual representation of the successfulness of chatbot's accuracy in answering the questions. In the next section, there will be a detailed discussion about the overall experiment.

Chapter 5 Analysis

In this section, there will be a thorough analysis of how the data from the experiment made an impact on both the hypothesis and the sub questions.

Data Analysis

The preparation before analysing the data was to make sure that all the selected questions were asked. It was important to check if all questions were asked because each question provides a better evaluation of the experiment. Microsoft Excel and Jupyter Notebook were the two software programs used in the analysis phase. Excel was used to gather all the answered questions and categorised them by the type of question and if the chatbot answered them or not. Jupyter Notebook was then used to create the bar charts for the NLP and MATH library outcomes of the experiment as well as, creating two lines graphs to illustrate the goal completion rate for both chatbots. The goal completion rate is a method used to analyse the accuracy rate of both chatbots answering the questions.

T-Test outcome

The t test is an evaluation process to analyse the statistical significance of two groups. In the experiment, both an NLP and a MATH library chatbot were used. Both Trigonometry and logarithm datasets from NLP and MATH library were used for the basis of the t-test. Reason being is that both topics were used in the two experiments. The t test compared the accuracy of both experiments to determine which chatbot was able to accurately answer the questions. All of the trigonometry and logarithm questions were used in the comparison as these two topics were used in both experiments.

Group	NLP	MATH library
MEAN	25	25
Standard Deviation	28.28427125	33.9411255
SEM	20	24
N	2	2
df	2	
p value	4.303	
t value	1	
Critial Value	4.30265273	

Figure 15 shows the T-Test 2 tailed Outcome.

Figure 15 displays the outcome of t test amongst the NLP and MATH library chatbots. A two tailed test was used in the t test as the thesis is comparing the accuracies for both NLP and MATH library. The p value is 4.303 because the t-distribution table highlights that for a 0.05 experiment that the results need to fall near to 4.303 (MedCalc, 2023). The mean value was calculated by adding up all the questions from the selected topics and dividing them by the number of topics used in the experiment. For example, NLP used 2 topics (Trigonometry and logarithms) which meant that the equation would be all the questions from the 2 topics added up and divided by 2 and for the MATH library it was adding up all the questions and dividing them by 2. The standard deviation was calculated by taking the mean and square rooting the value by 2. The degree of freedom (df) value is

2 because it is adding both the number of topics (2 +2) and taking away 2 which gave the value of 2. The t value was calculated with the formula:

$$T = \frac{mean\ 1 - mean2}{\frac{s(diff)}{\sqrt{(n)}}}$$

Which gave the value of 1. The critical value is 4.30265273 which was calculated by multiplying 0.05 with the degree of freedom which was 2. The t test value is 1 and when compared against the p value of 4.303 this means that the t test value rejects the null hypothesis. This means that there is statistical significance amongst the NLP chatbot and the MATH library chatbot. The mean value implies even though NLP and MATH library were given the same number of topics and the same amount of questions, NLP was able to answer more questions than MATH library because NLP was able to follow the calculation methods accurately. What this means is that the t test value is less than the p value of 0.05 which implies that there is statistical significance. The t test is significant because the t test tells us that there is a difference between NLP and MATH library, the difference can be identified by the standard deviations. Even though MATH library yield a bigger value than NLP, the interpretation is that NLP obtained less failed and not able to answer results than MATH library.

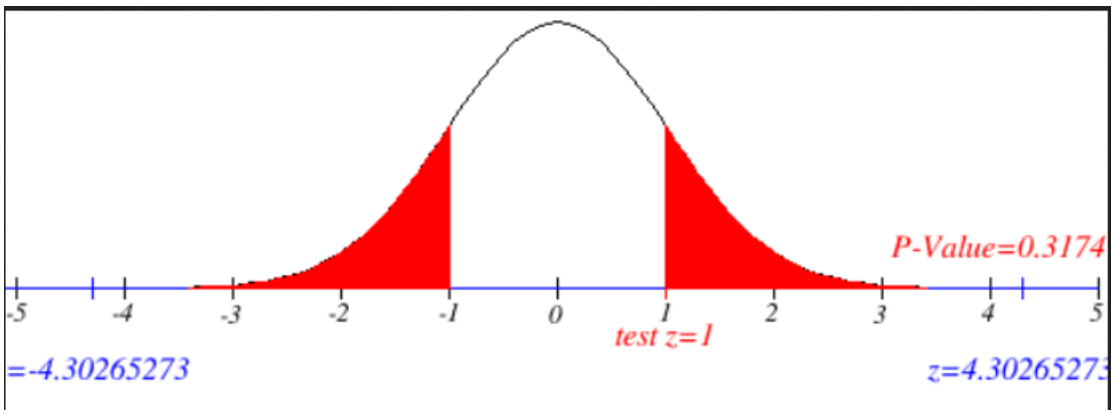

Figure 16: T-Test Graph

Figure 16 shows the graph of the t test and the red areas highlights the t value which is between -1 and 1. This is because the t value was calculated to be 1 from the values of the standard deviations and means from both NLP and MATH library.

Effect Size

In this experiment, both NLP and MATH library chatbot used 50 questions each from both trigonometry and logarithm questions created by college-based exam boards. NLP was modified to accept mathematical notations whereas MATH library was to import the trigonometry and logarithm equations and formulas libraries.

After the experiment had ended, both chatbots received the same mean number of 25 however both chatbots did receive different standard deviations. NLP's SD was 28.28427125 and MATH library's SD was 33.9411255, the reason that MATH library has a higher SD than NLP is due to the number of both failed and not able to answer questions as MATH library did have a high number of failed and not able to answer results.

Cohen'd NLP	24.11611652
Cohen'd MATH Library	24.26343044

Figure 17 shows Cohen's d equation

In order to calculate the effect size of the T-test, I implied Cohen's d which is a formula used to calculate the size difference of two groups (Scribbr, 2023). The method of using Cohen's d is to subtract the mean of the first group with the mean of the second group and divide the values with the SD from both groups. The figure above shows the results from Cohen's d and MATH library edges out NLP by 0.147313913. This shows that even though the sample sizes and means were exactly the same, the SD and Cohen's d implies that the effect size is small and has a limited practical application. What this means is that Cohen'd has a value of 0.147313913, which implies that even though there is a statistical significance amongst the means of NLP and MATH library the reality is that the difference amongst both chatbot means is negligible (ZACH, 2021). Meaning that the means of both chatbots are small, there is not much data within the means to be taken into the real world for further investigation.

Characteristics of Results

The purpose of this experiment was to observe if NLP was able to digest mathematical notation from higher level mathematics. Each time a chatbot outputted an answer, it was categorised into three columns: correct, failed and not able to answer. The experiment showed that NLP obtained the higher number of correct answers as well as a substantial amount of failed and not able to answer whereas, MATH library had one correct answer and the rest divided up amongst the failed and not able to answer columns. The characteristics shows us that NLP was able to digest more mathematical notations from the exam questions and correctly answered the questions accurately whereas, MATH library had trouble understanding the newer notations used in the exam questions.

Analysing Experiment Results

After the experiment was completed, the data informs us that NLP was able to answer more than one correct question amongst the selected topics than MATH library. This is because trigonometry and logarithms provided answerable questions for NLP than the MATH library, the data approves both hypotheses in the thesis. The hypotheses were to evaluate the accuracy between NLP and the MATH library and to clarify if NLP was able to answer one question from the selected topics, the data created from the experiment checks both boxes as NLP had a better accuracy than MATH library and NLP was able to answer over 1 question from 2 topics. As well as the correct and failed answers, NLP did have a high number of not able to answer results which implied that NLP was not able to calculate the correct answers as well as understanding the question. Logarithms also received some not able to answer results which again implies that NLP was having a hard time decoding the question and calculating the problem.

With the MATH library experiment, that type of chatbot was subjected to answering trigonometry and logarithm questions. And the overall report from that experiment was that logarithms was a topic that failed to have at least one question answered whereas trigonometry was about to have one question answered. However, both topics did receive the same number of not able to answer results which implies that the MATH library was not used to dealing with college-based exam questions. Which lead to many unanswered outcomes because the pre-programmed resources were not able to answer the exam-based questions.

Analysing Accuracy Rate Results

Alongside the observation of which topics had the most questions answered, the accuracy of both chatbots answering the questions was also under investigation. A goal completion rate method was used to observe which chatbot had a better completion rate of answering the maths questions accurately and correctly. The results show that NLP has best completion rate with trigonometry having a 4% completion rate and logarithms having a completion rate of 6% whereas, the MATH library had the poorest completion rate of 2% for trigonometry and 0% for logarithms. The data created from this method implies that NLP has a positive upwards completion rate due to answering more than one question from either topic. MATH library on the other hand, managed to have one question answered and the rest either failed or not able to answer as this chatbot has out of date formulas that require updating.

The research question was : Does an NLP chatbot have the ability to digest mathematical notations? From the NLP experiment alongside the MATH library experiment, the data approves the research question because NLP was able to digest the mathematical notations and had answered more questions than the MATH library.

Discussing Sub Questions

In this section the sub questions that were mentioned in the literature review will be answered based on the experiment. The first sub question was: Is NLP better than MATH library for accurate answering of the selected questions? The experiment confirms that NLP was better than MATH library as the accurate rate shows that NLP was able to answer 2 topics and MATH library only answered one.

Second question stated Amongst NLP and MATH library, which method digested the notations well? The experiment results show that NLP has a better digestion of notations as NLP was able to conquer 2 topics. Before the experiment, there was a test run during the building phrase for both chatbots and even in the test run NLP was on the ball when responding to the notations than MATH library.

Third question was Which topic challenged NLP and MATH library the most? Logarithms challenged the MATH library and NLP as this topic provided challenging questions for the library to solve and the calculations that NLP followed to solve the questions.

Fourth and final question was Overall, which chatbot produced a positive accurate rate? A goal completion rate method was implemented to further analyse the accuracy of both chatbots. From the line graphs, NLP provided a positive accurate with the graph showing a positive increase in growth. This implies that NLP was able to answer more questions amongst trigonometry and logarithms.

Discuss

Trigonometry

This topic was successfully in both NLP and MATH library testing. The data shows us that trigonometry is a mathematical topic that can be used in NLP and MATH libraries. The data also shows that trigonometry did receive a high number of failed answers for NLP and not able to answer results for the MATH library, this means that both chatbots has a challenging time with a majority of the questions. The data implies that the wording or the required calculations for the questions of trigonometry may have troubled both chatbot's answering systems.

In terms of the type of questions that trigonometry produced, this topic sprinkled itself across all three categories with the bad questions category at its highest. This indicates that the construction of trigonometry questions was dedicated towards a calculator and textbook examples rather than challenging NLP.

Logarithms

This topic managed to have 3 successful answered questions from the NLP experiment. This shows that logarithms have the potential to go move forward with NLP and AI. Even though all the logarithm questions failed in the MATH library experiment, this shows that the MATH library requires more attention if the AI scientists want to dominate logarithms with the MATH library. In addition, MATH library did receive a higher number of not able to answer questions than NLP which suggests that the logarithm equations and formulas within MATH library needs improvement.

Just like trigonometry, the logarithm questions were also deployed in all three categories with bad questions being the second highest. This also shows that most logarithm questions can be solved using a calculator despite the number of questions in the good question category.

Summary

In this chapter, there was a discussion on the published results from both experiments. After that, the sub questions were answered as well as reviewing each of the selected topics and analysing their participation in the experiments.

Chapter 6 Conclusion

This project was to investigate if a chatbot can solve A-level mathematical questions using NLP. The project also compared the calculation accuracy with a MATH library with the following chosen topics: Trigonometry and Logarithms. The intention of the project was to evaluate both NLP chatbot and MATH library chatbot to see which method was able to solve mathematical questions. A correct, failed and not able to answer system was implemented in the data collection process, this system was then used to determine the status of the hypotheses for the experiment.

For NLP, both topics have shown to have proven both hypotheses as NLP was able to answer at least one question. Compared with the MATH library, it can be concluded that NLP had a more accurate answering approach than MATH library. The hypotheses of this thesis were to compare the accuracy between NLP and MATH library and to have NLP answer a minimum of one question from the selected topics, the experiment and results can clarify the succession for both hypotheses due to having NLP answer more than one question and having a better accurate answering system than MATH library.

Alongside NLP answering questions, there was an additional analysis of examining the accuracy of the questions being answered. A goal completion rate method was implemented to examine and display the data generated from the experiment. From the experiment, it can be confirmed that NLP has a positive healthy goal completion rate due to having more than one correct answer in both topics. MATH library, however, has a negative goal completion rate as this chatbot was only able to answer one question from Trigonometry.

Limitations

The project conducted did have some limitations. For instance, the project only used 2 topics: Trigonometry and Logarithms as these topics were to consist of hard equations and formulas.

The other limitation was that the project was using A-level maths topics. With regards to GCSE and AS topics, A-level had a superior title that contained information that would be best used to evaluate against NLP and math library. What this means is that there have been creations within AI to tackle GCSE questions and AS questions but there have not been any resources within AI to educate A-level students.

As mentioned in the thesis, the concerns with chatbots in general is that they would need to be updated manually each time to keep up with the modern-day calculation trend. For example, if there was a new formula for a calculus question the chatbot would require a programmer to update the chatbot with the new formula. A chatbot can experience issues with students because without the correct information being used, many students would refuse to allow a chatbot to help them with their college work.

Future Work

Possible research ideas for this study to be used in the future could be NLP chatbots managing further maths questions, NLP chatbots solving all the questions in a different field of study such as physics or having an NLP chatbot solve GCSE maths questions. It would be most intriguing to see if NLP can manage college-based physics questions as there is many opportunities that would help college physics students have a better understanding of the subject. NLP digesting Schrodinger's equation and Eisenstein's relativity formulas is another interest as it would be fascinating for a chatbot to solve big equations such as the ones mentioned above. These ideas are related to the thesis because they all consist of the concept of having a higher-level equation and formula for an

NLP chatbot to follow, Schrodinger's equation and the GCSE maths questions all have instructions that NLP has yet to encounter.

This study used past papers as a method of collecting data, for other research opportunities like the physics idea human participants would be an idea. This would mean that the research would gather more information about the participant's experience as well as participating in the experiment. What this means is that, as well as obtaining the crucial data for the study having the participants express their opinion about the experiment would enlarge the experiment feedback report.

The study also focused more on NLP and MATH library, to render further opportunities testing transformer architecture against NLP would an idea. It is because transformer architecture has a process of digesting data as a whole rather than dividing the data in little tokens (datagen, 2023). This would an interesting topic as it shows the debate of is it easier to digest a block of data as a whole rather than splitting them up into tokens.

References

[1] Mcaulay (2022, Feb. 16). What Problem Does a Chatbot Solve?[Online].Available: <https://www.kindly.ai/blog/what-problem-does-a-chatbot-solve >

[2] Brush and Scardina (2022). Definition of Chatbot[Online].Available:

<https://www.techtarget.com/searchcustomerexperience/definition/chatbot>

[3] Sciencedaily. (2021, Dec. 1). Machine Learning Helps Mathematicians make new connections [Online]. Available: <https://www.sciencedaily.com/releases/2021/12/211201111925.htm >

[4] Lee (2022). Developing an AI-based chatbot for practicing responsive teaching in mathematics [Online]. [accessed 10th November 2022]. Available: <https://www.sciencedirect.com/science/article/abs/pii/S0360131522002172 >

[5] Leonard(2022). Understanding Chatbots [Online]. [accessed 10th November 2022]. Available: <https://fuzzymath.com/blog/understanding-chatbots/ >

[6] Yse(2019, Jan.15). Your Guide to Natural Language Processing (NLP)[Online]. [accessed 10th November 2022]. Available: <https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1 >

[7] AISI.(2022). Why Applied Maths is key to AI innovation[Online].Available: <https://aischoolofindia.com/uncategorized/need-for-applied-mathematics-for-ai-innovation/>

[8] Bharath(2021. Apr 21). Best Library to Simplify Math for Machine Learning [Online]. [accessed 10th November 2022].Available: <https://towardsdatascience.com/best-library-to-simplify-math-for-machine-learning-ed64cbe536ac >

[9] Paul's Online Notes.(2021). Section 3.3 : Differentiation Formulas[Online].Available: <https://tutorial.math.lamar.edu/classes/calci/diffformulas.aspx >

[10] Brtiannica (2020). Logarithm[Online].Available: <https://www.britannica.com/science/logarithm>

[11] Revision Maths (2020). Solving Trigonometric Equations[Online].Available: < https://revisionmaths.com/advanced-level-maths-revision/pure-maths/trigonometry/solving-trigonometric-equations >

[12] Stackoverflow (2020). Load Markdown File on Jupyter Notebook Cell[Online].Available: <https://stackoverflow.com/questions/64062225/load-markdown-file-on-a-jupyter-notebook-cell>

[13] Dynamics (2019). Introduction to Sympy and the Jupyter Notebook for engineering calculations[Online].Available: <https://dynamics-and-control.readthedocs.io/en/latest/0_Getting_Started/Notebook%20introduction.html >

[14] Gtribello (2019). Using markdown in Jupyter notebook[Online].Available: <https://gtribello.github.io/mathNET/assets/notebook-writing.html>

[15] The Sequence (2021. Nov 7). OpenAI New NLP Challenge: Mathematical Reasoning[Online].Available: <https://thesequence.substack.com/p/-openai-new-nlp-challenge-mathematical?s=r>

[16] Revision Maths(2019). Differential Equations [Online].Available: <https://revisionmaths.com/advanced-level-maths-revision/pure-maths/calculus/differential-equations >

[17] Tutorials point (2019). Jupyter Notebook – IpyWidgets[Online].Available: <https://www.tutorialspoint.com/jupyter/jupyter_notebook_ipywidgets.htm >

[18] Engati (2022, Mar 1). 6 types of chatbots – Which is best for your business? [Online].Available: <https://www.engati.com/blog/types-of-chatbots-and-their-applications?utm_content=types-of-chatbots-and-their-applications >

[19] Verstegen(2022, Aug 12). Breaking Down the pros and cons of a chatbot as a customer experience solution [Online]. Available: <https://www.chatdesk.com/blog/pros-and-cons-of-chatbots>

[20] Dilmegani (2020, Jul 4). Banking Chatbots in 2023 : Benefits, Use Cases and Best Practices [Online].Available: <https://research.aimultiple.com/banking-chatbot/>

[21] Vocabulary (2019). Programming Error [Online].Available: <https://www.vocabulary.com/dictionary/programming%20error>

[22] Jovic(2022, Mar 17). The Future is Now – 37 Fascinating Chatbot Statistics [Online].Available: <https://www.smallbizgenius.net/by-the-numbers/chatbot-statistics/#gref >

[23] Sundaram et al (2022, May 5). Why are NLP Models Fumbling at Elementary Math? A survey of Deep Learning based Word Problem Solvers [Online].Available: <https://deepai.org/publication/why-are-nlp-models-fumbling-at-elementary-math-a-survey-of-deep-learning-based-word-problem-solvers >

[24] Okonkwo (2021). Chatbots application in education : A systematic review [Online].Available: <https://www.sciencedirect.com/science/article/pii/S2666920X21000278>

[25] Indicative Team (2021). What is a Chatbot? [Online].Available: <https://www.indicative.com/resource/chatbot/ >

[26] Fainchtein (2020, Jun 28). Generative vs Retrieval Based Chatbots: A quick guide [Online].Available: <https://blog.cloudboost.io/generative-vs-retrieval-based-chatbots-a-quick-guide-8d19edb1d645 >

[27] Odoscope (2019). Data-driven versus rule-based-5 success tips for the optimal automation strategy [Online]. Available: <https://blog.odoscope.com/en/data-driven-versus-rule-based-5-tips >

[28] Prov (2022, Apr 19). 5 Common Machine Learning Problems and How to Solve Them [Online]. Available: <https://www.provintl.com/blog/5-common-machine-learning-problems-how-to-beat-them >

[29] Educationandcareernews (2019). How Chatbots are Solving Problems for College Students [Online]. Available: <https://www.educationandcareernews.com/empowering-our-educators/how-chatbots-are-solving-problems-for-college-students/>

[30] Khan(2020, Feb 26). 8 Benefits of Chatbots in education industry [Online]. Available: <https://botsify.com/blog/education-industry-chatbot/>

[31] Liyanapathirana(2018). The Python math Module: Everything You Need to Know [Online].Available: <https://realpython.com/python-math-module/ >

[32] Juventini (2020). Advantages and Disadvantages of Using Calculator [Online].Available: <https://www.scribd.com/document/345263938/Advantages-and-Disadvantages-of-Using-Calculator >

[33] Ring Central (2021 Jan 26). The Definitive Guide to Using a Deep Learning or Machine Learning Chatbot for Your Business [Online].Available: <https://www.ringcentral.com/gb/en/blog/the-definitive-guide-to-using-a-deep-learning-or-machine-learning-chatbot-for-your-business/>

[34] Prasana (2022 Apr 22). Advantages and Disadvantages of Machine Learning [Online].Available: <https://www.aplustopper.com/advantages-and-disadvantages-of-machine-learning/>

[35] Fintelics (2021 Aug 20). The Growing Popularity of Machine Learning Technology [Online].Available: <https://fintelics.medium.com/the-growing-popularity-of-machine-learning-technology-bd217bb00d4f>

[36] Dsouza (2021 Nov 24). All the Math You Need to Know in Artificial Intelligence [Online].Available: <https://www.freecodecamp.org/news/all-the-math-you-need-in-artificial-intelligence/>

[37] Chino (2022 Mar 17). Mathematical paradox demonstrates the limits of AI [Online].Available: <https://www.cam.ac.uk/research/news/mathematical-paradox-demonstrates-the-limits-of-ai>

[38] lecture1 (2020). Types of scales and levels of measurement [Online].Available: <https://web.pdx.edu/~newsomj/pa551/lecture1.htm>

[39] Ndukwu(2020 Feb 13). Levels of Measurement: Nominal, Ordinal, Interval and Ratio Scales [Online].Available: <https://www.kyleads.com/blog/nominal-ordinal-interval-ratio-scales/>

[40] Timeline maker (2022). Timeline Maker[Online]. Available: <https://time.graphics/editor>

[41] Taylor (2023 Mar 14). How to use past exam papers to revise effectively[Online]. [accessed 16th November 2022]. Available: < https://www.theuniguide.co.uk/advice/revision-help/how-to-use-past-exam-papers-to-revise-effectively >

[42] Sophie (2019 Apr 15). How can I use past papers most effectively? [Online]. [accessed 16th November 2022]. Available: < https://www.scienceandmathsrevision.co.uk/how-can-i-use-past-papers-most-effectively/ >

[43] Fadeni, Feyiseye(2019) A Method for Representing Knowledge and Improving Answer Prediction in Question-and-Answer Domain. Doctoral thesis, University of Huddersfield. [Online].[accessed 04th May 2023].Available: < https://eprints.hud.ac.uk/id/eprint/35105/1/FINAL%20THESIS%20-%20Fadeni.pdf >

[44] Dictinoary.com(2023) Definition & Meaning of the word failed.[Online].[accessed 4th May 2023].Available: < https://www.dictionary.com/browse/failed >

[45] Scribbr.com(2023) What is Effect Size and Why Does It Matter? (Examples).[Online].[accessed 11th May 2023].Available: < https://www.scribbr.co.uk/stats/effect-sizes/ >

[46] Visiativ(2023) Measuring Chatbot Effectiveness: 16 KPIs to Track.[Online][accessed 29th May 2023].Available: < https://www.visiativ.com/en/actualites/news/measuring-chatbot-effectiveness/>

[47] Freshchat(2023) Chatbot Analytics: 7 Essential Metrics to Track.[Online][accessed 30th May 2023].Available: < https://www.freshworks.com/live-chat-software/chatbots/chatbot-analytics/ >

[48] Publift (2023) What is Completion Rate and How to Measure It.[Online][accessed 30th May 2023].Available: < https://www.publift.com/blog/completion-rate >

[49] Lishchynska (2023, Jan 16) Chatbot Analytics: 14 Chatbot Metrics to Track in 2023.[Online][accessed 31st May 2023].Available:< https://botscrew.com/blog/chatbot-metrics/ >

[50] Data to fish (2022, Nov 12th) How to Plot a Line Chart in Python using Matplotlib.[Online][accessed 31st May 2023].Available: < https://datatofish.com/line-chart-python-matplotlib/ >

[51] MedCalc (2023).T-Distribution table (two-tailed). [Online].[accessed 1st June 2023]. Available:< https://www.medcalc.org/manual/t-distribution-table.php >

[52] IMB (2023). What is Natural Language Processing?.[Online].[accessed 8th June 2023].Available : < https://www.ibm.com/topics/natural-language-processing >

[53] Imathas (2023). Hypothesis Test Graph Generator . [Online].[accessed 13th June 2023].Available : < http://www.imathas.com/stattools/norm.html >

[54] pathmind (2023). Evaluation Metrics for Machine Learning – Accuracy, Precision, Recall and F1 defined.[Online].[accessed 5th July 2023].Available : < http://wiki.pathmind.com/accuracy-precision-recall-f1 >

[55] C3.ai(2023). Recall What is Recall? [Online].[accessed 5th July 2023].Available: < https://c3.ai/glossary/data-science/recall/ >

[56] Kundu (2022, Dec 16th). F1 Score in Machine Learning: Intro & Calculation.[Online].[accessed 5th July 2023].Available : < https://www.v7labs.com/blog/f1-score-guide#what-is-f1-score >

[57] ZACH (2021, Aug 31st). How to Interpret Cohen's d (With Examples).[Online].[accessed 7th July 2023].Available: < https://www.statology.org/interpret-cohens-d/ >

[58] Datagen (2023).What is the Transformer Architecture and How Does It Work?.[Online].[accessed 12th July 2023].Available: < https://datagen.tech/guides/computer-vision/transformer-architecture/ >

Labaka, G., Stroppa, N., Way, A. and Sarasola, K., 2007. Comparing rule-based and data-driven approaches to Spanish-to-Basque machine translation. In *Proceedings of Machine Translation Summit XI: Papers*.

Grossman, J., Lin, Z., Sheng, H., Wei, J.T.Z., Williams, J.J. and Goel, S., 2019. MathBot: Transforming online resources for learning math into conversational interactions. *AAAI 2019 Story-Enabled Intelligence*.

Singh, J., Joesph, M.H. and Jabbar, K.B.A., 2019, May. Rule-based chabot for student enquiries. In *Journal of Physics: Conference Series* (Vol. 1228, No. 1, p. 012060). IOP Publishing.

Ferreira, D. and Freitas, A., 2020. Natural language premise selection: Finding supporting statements for mathematical text. *arXiv preprint arXiv:2004.14959*.

Crossley, S.A., Karumbaiah, S., Ocumpaugh, J., Labrum, M.J. and Baker, R.S., 2020. Predicting math identity through language and click-stream patterns in a blended learning mathematics program for elementary students. *Journal of Learning Analytics*, *7*(1), pp.19-37.

Sangeetha, S., Sahithya, C., Rasiga, M.R. and Shalini, N., 2021. Chatbot for Personal Assistant Using Natural Language Processing. *International Journal of Research in Engineering, Science and Management*, *4*(3), pp.96-97.

Pai, K.C., Kuo, B.C., Liao, C.H. and Liu, Y.M., 2021. An application of Chinese dialogue-based intelligent tutoring system in remedial instruction for mathematics learning. *Educational Psychology*, *41*(2), pp.137-152.

Badlani, S., Aditya, T., Dave, M. and Chaudhari, S., 2021, May. Multilingual Healthcare Chatbot Using Machine Learning. In *2021 2nd International Conference for Emerging Technology (INCET)* (pp. 1-6). IEEE.

Hwang, G.J. and Tu, Y.F., 2021. Roles and Research Trends of Artificial Intelligence in Mathematics Education: A Bibliometric Mapping Analysis and Systematic Review. *Mathematics*, *9*(6), p.584.

Jetten, K.J.H., 2021. *A hybrid chatbot that uses contextual sensors to influence responses* (Master's thesis).

Acuna, G.E., Alvarez, L.A., Miraflores, J. and Samonte, M.J., 2021, June. Towards the Development of an Adaptive E-Learning System with Chatbot Using Personalized E-Learning Model. In *2021 The 7th International Conference on Frontiers of Educational Technologies* (pp. 120-125).

Zarza Davila, A. and Glineur, F., " Proof of concept of an interactive theorem prover system using natural language input.

Dionysiou, A. and Athanasopoulos, E., 2021, November. Unicode Evil: Evading NLP Systems Using Visual Similarities of Text Characters. In *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security* (pp. 1-12).

McIntyre, S., 2021. *Mind Map Automation: Using Natural Language Processing to Graphically Represent a Portion of a US History Textbook* (Doctoral dissertation, University Honors College Middle Tennessee State University).

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D. and Steinhardt, J., 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Noorbakhsh, K., Sulaiman, M., Sharifi, M., Roy, K. and Jamshidi, P., 2021. Pretrained Language Models are Symbolic Mathematics Solvers too!. *arXiv preprint arXiv:2110.03501*.

Welleck, S., Liu, J., Bras, R.L., Hajishirzi, H., Choi, Y. and Cho, K., 2021. Naturalproofs: Mathematical theorem proving in natural language. *arXiv preprint arXiv:2104.01112*.

Zhu, Y., Nie, J.Y., Zhou, K., Du, P. and Dou, Z., 2021, March. Content selection network for document-grounded retrieval-based chatbots. In *European Conference on Information Retrieval* (pp. 755-769). Springer, Cham.

Rajagopalan, G., 2021. Regular Expressions and Math with Python. In *A Python Data Analyst's Toolkit* (pp. 77-99). Apress, Berkeley, CA.

Faldu, K., Sheth, A., Kikani, P., Gaur, M. and Avasthi, A., 2021. Towards tractable mathematical reasoning: Challenges, strategies, and opportunities for solving math word problems. *arXiv preprint arXiv:2111.05364*.

Nirala, K.K., Singh, N.K. and Purani, V.S., 2022. A survey on providing customer and public administration-based services using AI: chatbot. *Multimedia Tools and Applications*, pp.1-32.

Shenoy, A., Bhoomika, M. and Annaiah, H., 2022. Design of chatbot using natural language processing. *Knowledge Engineering for Modern Information Systems: Methods, Models and Tools*, p.60.

Mahajan, L. and Bhagatb, S., 2022. An artificial neural network for the prediction of the strength of supplementary cementitious concrete.

Vayadande, K., Pate, S., Agarwal, N., Navale, D., Nawale, A. and Parakh, P., 2022. Modulo Calculator Using Tkinter Library. *EasyChair Preprint*, (7578).

Jain, D., Patel, S., Vadhvani, P., Tambili, N. and Malviya, V., INTERNAL ARCHITECTURE OF CHATBOT AND VARIOUS MODULES USED-A REVIEW.

Widdows, D., Zhu, D. and Zimmerman, C., 2022. Near-Term Advances in Quantum Natural Language Processing. *arXiv preprint arXiv:2206.02171*.

Abumosameh, M.D., 2022. *Python Code for Sturm-Liouville Eigenvalue Problems* (Doctoral dissertation, Carleton University).

Mafra, M., Nunes, K., Castro, A., Lopes, A., Oran, A.C., Braz Junior, G., Almeida, J., Paiva, A., Silva, A., Rocha, S. and Viana, D., 2022. Defining Requirements for the Development of

Useful and Usable Chatbots: An Analysis of Quality Attributes from Academy and Industry. In *International Conference on Human-Computer Interaction* (pp. 479-493). Springer, Cham.

Kuhail, M.A., Alturki, N., Alramlawi, S. and Alhejori, K., 2022. Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, pp.1-46.

Zaid, S., Malik, A. and Fatima, K., 2022. Jewelry Shop Conversational Chatbot. *arXiv preprint arXiv:2206.04659*.

Xu, B. and Zhuang, Z., 2022. Survey on psychotherapy chatbots. *Concurrency and Computation: Practice and Experience*, *34*(7), p.e6170.

Moreno-Guerrero, A.J., Marín-Marín, J.A., Dúo-Terrón, P. and López-Belmonte, J., Chatbots in Education: A Systematic Review of the Science Literature. *Artificial Intelligence in Higher Education*, pp.81-94.

GEETHA, D., ARTIFICIAL INTELLIGENCE CHATBOT USING PYTHON.

Mandal, S., Acharya, S. and Basak, R., 2022. Solving Arithmetic Word Problems Using Natural Language Processing and Rule-Based Classification. *International Journal of Intelligent Systems and Applications in Engineering*, *10*(1), pp.87-97.

Zhang, M., Baral, S., Heffernan, N. and Lan, A., 2022. Automatic Short Math Answer Grading via In-context Meta-learning. *arXiv preprint arXiv:2205.15219*.

Liang, Z., Zhang, J., Wang, L., Qin, W., Lan, Y., Shao, J. and Zhang, X., 2022, July. MWP-BERT: Numeracy-augmented pre-training for math word problem solving. In *Findings of the Association for Computational Linguistics: NAACL 2022* (pp. 997-1009).

Meadows, J. and Freitas, A., 2022. A Survey in Mathematical Language Processing. *arXiv preprint arXiv:2205.15231*.

Alnfiai, M., 2022. TapCalculator: nonvisual touchscreen calculator for visually impaired people preliminary user study. *Journal on Multimodal User Interfaces*, *16*(2), pp.143-154.

Zong, M. and Krishnamachari, B., 2022. Solving math word problems concerning systems of equations with gpt-3. In *Proceedings of the Thirteenth AAAI Symposium on Educational Advances in Artificial Intelligence*.

Lin, B., 2022. Knowledge management system with nlp-assisted annotations: A brief survey and outlook. *arXiv preprint arXiv:2206.07304*.

Lin, C.C., Huang, A.Y. and Yang, S.J., 2023. A review of ai-driven conversational chatbots implementation methodologies and challenges (1999–2022). *Sustainability*, *15*(5), p.4012.

Hirosawa, T., Harada, Y., Yokose, M., Sakamoto, T., Kawamura, R. and Shimizu, T., 2023. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: A pilot study. *International journal of environmental research and public health*, *20*(4), p.3378.

Shahriar, S. and Hayawi, K., 2023. Let's have a chat! A Conversation with ChatGPT: Technology, Applications, and Limitations. *arXiv preprint arXiv:2302.13817*.

Raiyan, S.R., Faiyaz, M.N., Kabir, S.M.J., Kabir, M., Mahmud, H. and Hasan, M.K., 2023. Math Word Problem Solving by Generating Linguistic Variants of Problem Statements. *arXiv preprint arXiv:2306.13899*.

Berbatova, M. and Ivanov, F., 2023. An Improved Bulgarian Natural Language Processing Pipeline.

Arivazhagan, N. and Van Vleck, T.T., 2023. Natural language processing basics. *Clinical Journal of the American Society of Nephrology*, *18*(3), pp.400-401.

Patil, R., Boit, S., Gudivada, V. and Nandigam, J., 2023. A Survey of Text Representation and Embedding Techniques in NLP. *IEEE Access*.

Appendix

Figure 1 shows the timeline of the chatbots.

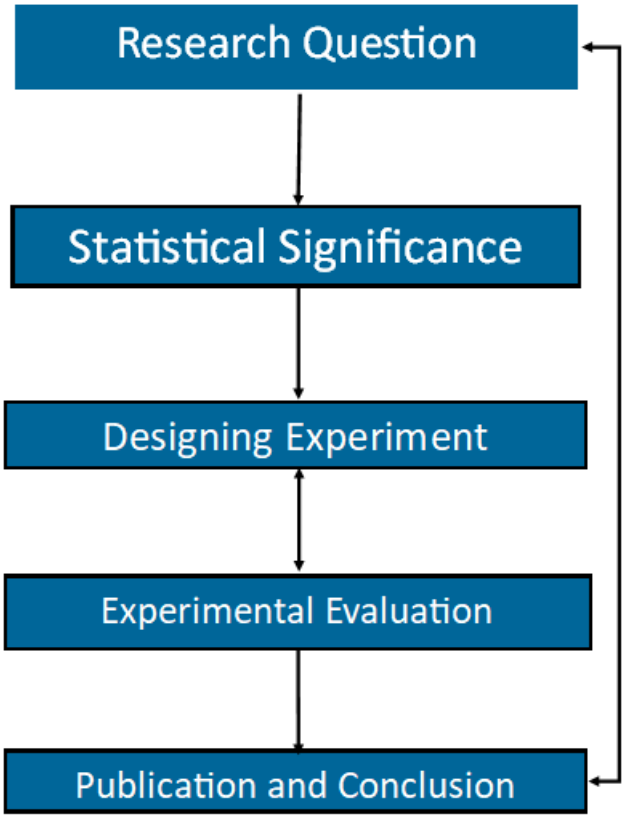

Figure 2 Design Science Model

Solve, for -90 < A < 90, the equation 8 sin^3 A - 6 sin A = 1
Your answer is:

$a = \frac{1}{4}$.

Figure 3 shows the chatbot answering a question.

```
In [6]: print("The answer for show that 0 = 0.4+sin0 is",math.sin(0)+0.4)
```

Figure 4 shows an example of a question being asked by the math library chatbot.

```
print("The answer for the question Solve log6 36 is:", math.log(6,36))
```

```
The answer for the question Solve log6 36 is: 0.5
```

Figure 5 shows the MATH library answering a logarithms question.

Chatbot Outcome Sheet

Which chatbot was used in the experiment?

NLP ☐ MATH library ☐

Which topic was chosen to ask the chatbot?

Trigonometry ☐ Logarithms ☐

What was the outcome of the chatbot's response to the question?

Correctly Answered ☐ Failed to Answer ☐

Not Able to Answer ☐

Figure 6 shows chatbot outcome sheet

Topics	Correct	Failed	Not Able to Answer
Trigonometry	2	20	3
Logarithms	3	12	10

Figure 7 shows the table of the NLP Experiment

Topics	Correct	Failed	Not Able to Answer
Trigonometry	1	11	13
Logarithms	0	14	13

Figure 8 shows the table of MATH library Experiment.

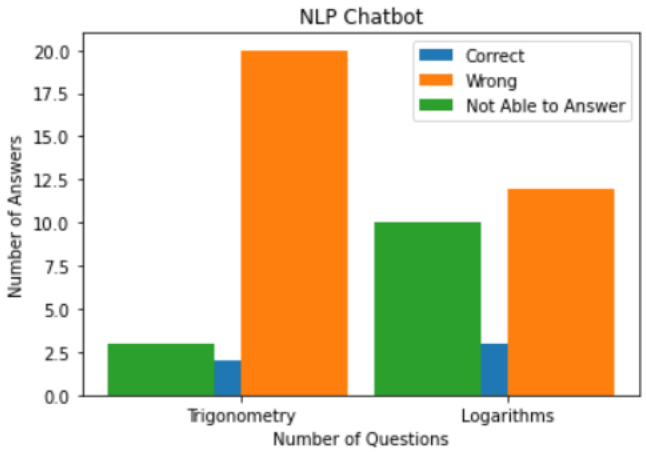

Figure 9 shows the bar chart of the NLP Chatbot Experiment

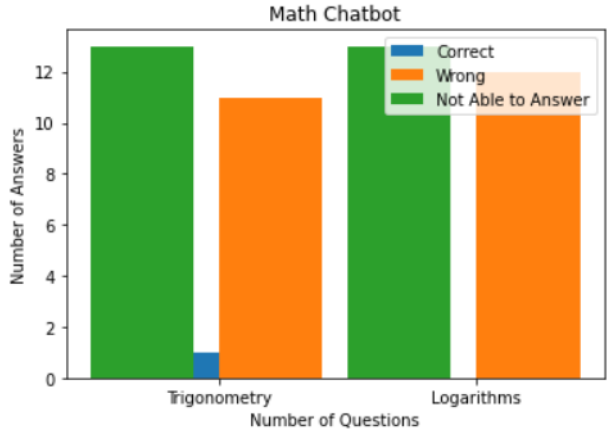

Figure 10 shows the bar chart of the MATH Library Experiment

Completion Rate for NLP (%)	Trigonometry	Logarithms
	4%	6%

Figure 11 shows NLP Goal Completion Rate

Completion Rate in MATH library (%)	Trigonometry	Logarithms
	2%	0%

Figure 12 shows MATH library Goal Completion Rate

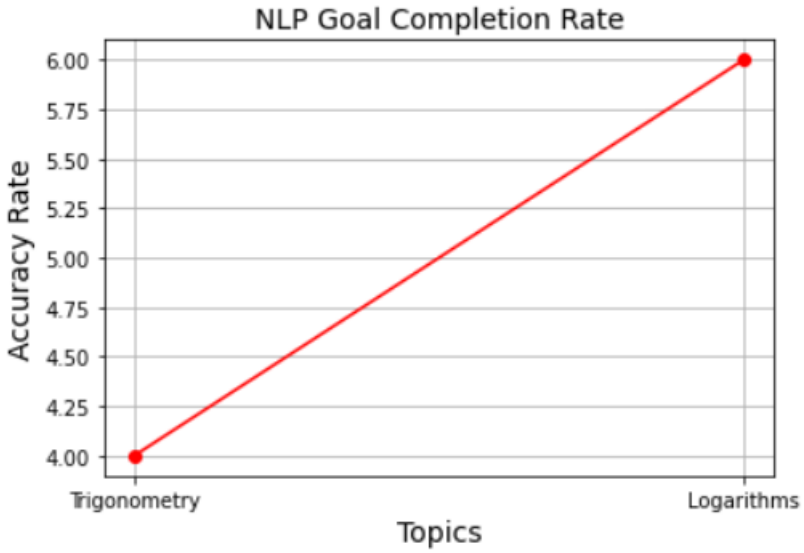

Figure 13 shows a line graph of NLP's Goal Completion Rate.

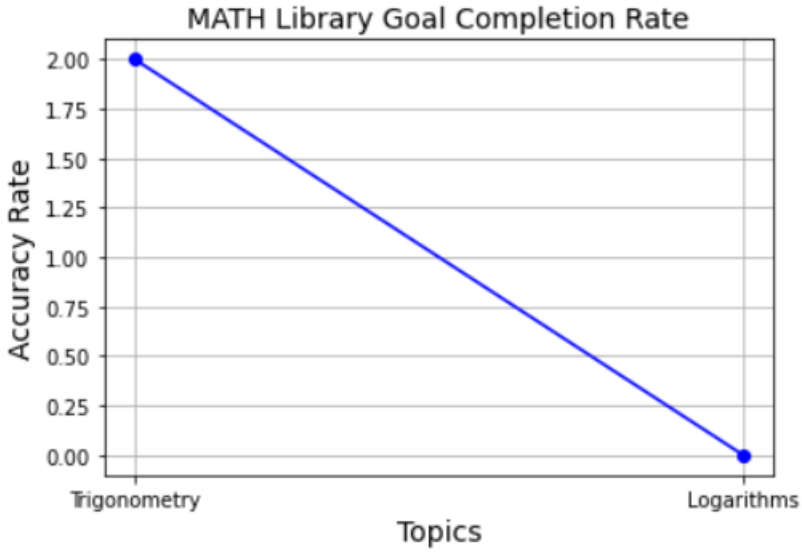

Figure 14 shows a line graph of MATH Library's Goal Completion Rate

Group	NLP	MATH library
MEAN	25	25
Standard Deviation	28.28427125	33.9411255
SEM	20	24
N	2	2
df	2	
p value	4.303	
t value	1	
Critial Value	4.30265273	

Figure 15 : T-Test outcome

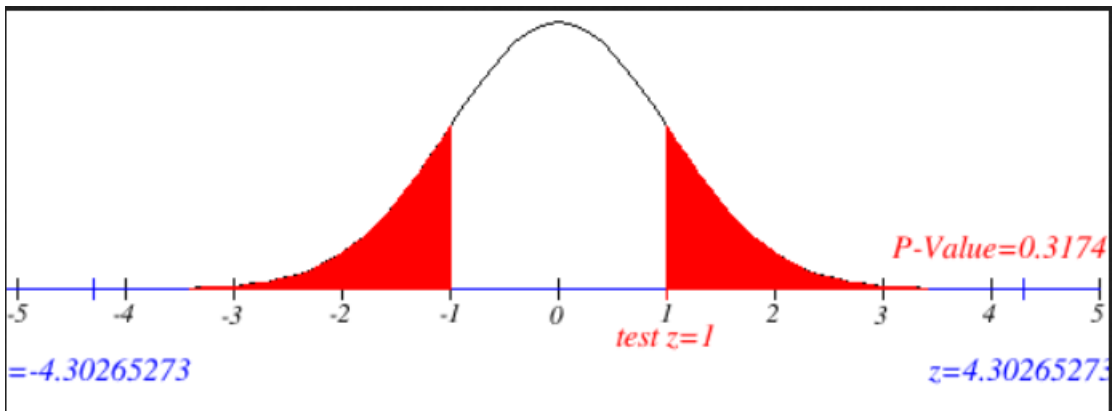

Figure 16: T-Test Graph

Cohen'd NLP	24.11611652
Cohen'd MATH Library	24.26343044

Figure 17 : Cohen's d Results