

Build An Intelligent Tutoring System

— Final Report —

Nikita Dmitrieff, Ryan El Khoury, Sukhjeet Atwal, Shang Shi, Nicole Farah
{nikita.dmitrieff19, ryan.el-khoury22, sukhjeet.atwal22, shang.shi18, nicole.farah21}@imperial.ac.uk

Supervisor: Dr Nicole Salomons
Module: COMP70048, Imperial College London

May 2nd, 2023

Contents

1	Introduction	2
1.1	Background and context of the project	2
1.2	Project Problem statement	2
2	Technical Decisions, Design, and Architecture	3
2.1	Specification of the functionality	3
2.1.1	Versatile Chatbot	3
2.1.2	Algebraic Expression Simplifier and Solver	4
2.1.3	Facial Expression Recognition	4
2.1.4	Handwriting Recognition	4
2.1.5	PDF Evaluation	5
2.1.6	Other Features	5
2.2	AI Used and Developed for the ITS	6
2.2.1	DNN for solving algebraic equations	6
2.2.2	Q&A Generation for student evaluation	7
2.2.3	Facial Recognition for student sentiment	8
2.2.4	Handwriting recognition	8
2.3	Main design choices and justifications	8
2.4	Architectural diagram	8
2.5	App Screenshots and Link to web-app for production	9
3	Evaluation	10
3.1	Quantitative Test of Performance	10
3.1.1	Facial Expression Recognition	10
3.1.2	Image to Text Transformer	10
3.1.3	Maths Equation Solver	11
3.1.4	DNN for solving expression	13
3.1.5	PDF Evaluation	13
3.2	Testing User Needs and Product Robustness	14
4	Conclusion	14
A	Prompt engineering examples	17

1 Introduction

An Intelligent Tuition System (ITS) can be broadly defined as a computerized system which consists of the following steps:

1. Delivers tuition to a student (eg learning material, hints, prompts, etc)
2. Evaluates the students' response to that material (eg testing learning, monitoring emotional response)
3. Adapts the delivery in Step 1 using the evaluation in Step 2.

There is considerable evidence that the usage of ITS improves learning outcomes. For example [7] in their meta study, found the use of ITS was associated with greater achievement in comparison with many forms of teaching with no use of ITS. The level of achievement was very similar to that of learning from individualized human tutoring or small-group instruction. The results were robust with respect to the subject under consideration, the school age of the individual, and the level of prior domain knowledge among other characteristics.

Individualized tutoring is considered the optimal form of pedagogic instruction. However, this is impractical in most cases. Given its performance, therefore, broadening access to ITSs and improving their quality and robustness should be a goal for educators everywhere.

This report details the steps taken to build an ITS. It firstly outlines the context of this problem, and highlights other platforms that exist in the current market. It will show how considerations of these factors led to the final ITS product named Filo. In part two of the report, the architecture of the website is explored. The design and implemented features - AI or otherwise - will be examined in depth. In the evaluation section of the report, the overall usability and reliability of the AI features are assessed through quantitative tests and feedback from children in the target age group.

1.1 Background and context of the project

Tutoring systems with supplementary curriculum resources for young children are widely available, with an extensive range of online tutoring systems and subject-specific websites that are well developed. There are numerous options for parents in terms of educational websites and apps to aid their children's learning experience in conjunction with their lessons in school. The incorporation of AI into these systems is already common, with adaptive learning technology providing personalised content suiting a student's individual needs.

Parents in the UK collectively spend a staggering £6bn every year on private tutors for their children, and 1 in 3 parents who manage to shell out the £53 weekly cost confirm they are doing so because they feel that their child isn't getting the help they need at school. Cost aside, it is clear that there is plenty of space in this market to provide something unique to parents that will support their child's learning.

Figure 1 shows examples of existing tutoring websites. Both Prodigy and Adventure Academy are game based tutoring sites; the child personalises a character and is free to roam around places on an extensive map. They can play educational minigames to earn coins which they can use to buy clothes or customise their own home. This approach clearly runs the risk of being too distracting to children; especially since there is a chat function that allows them to chat to friends. BrainPOPjr is aimed at slightly younger children than the target market, but is an important example to consider too. This site is clearly a platform for learning and children can choose the topics they want to focus on, watch short clips and complete quizzes. These types of sites, although well resourced, may seem too impersonal to children, and fail to keep them engaged. Upon considering other similar apps, it is possible to consider the full problem of this task and how it might be tackled.

1.2 Project Problem statement

Initially, it is important to consider the challenges that students and parents face when seeking an online tutoring aid and where there might be a gap in the current market. With the awareness that

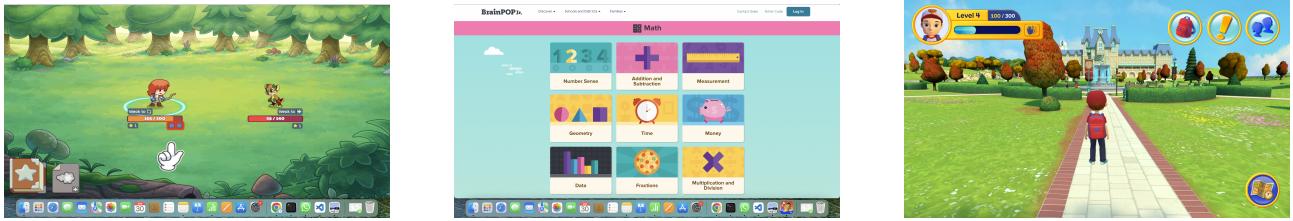


Figure 1: Examples of Existing Tutoring Interfaces available: (from left to right) Prodigy, BrainPOPjr, Adventure Academy

the software of websites such as the aforementioned in part 1.1 are advanced beyond the scope of this project, it was clear that the focus must be to incorporate AI that is not yet implemented on such sites, and how these might be used to aid an educational experience.

The majority of online tutoring apps aim to provide individual lessons to children in a subject of their choosing. It became clear, after some research, that these websites are not helpful enough when a student needs a specific answer, or help on a challenging question that lessons do not provide enough detail on. Websites like BBC bitesize are good for outlining key points of a subject matter, but every child is different, and something that is not stated explicitly may lead to confusion and shame in a student.

In response to this problem, Filo was created. Filo is a virtual learning environment for a target market of ages 9-13. It can be utilised by students in *conjunction* with their homework and textbooks. The platform is not intended to teach a curriculum, but rather provide support to a student who is independently learning. When a teacher is not available to answer a question, Filo provides suggestions and reasonable help. In addition to this, it offers multiple time management tools for students to help them prioritise their work. This is the uniqueness of the product. It is entirely agnostic to the subject matter. It is more properly viewed as a ‘Homework Helper and Study Buddy’. The AI is friendly, warm and is designed to be the child’s go-to platform when they are looking for assistance.

2 Technical Decisions, Design, and Architecture

2.1 Specification of the functionality

This section provides a detailed description of the app’s functions, how they were implemented and in what way they are useful. It will define how each function supports the app’s objective and user needs, as well as consider the limitations of these features too.

2.1.1 Versatile Chatbot

The app incorporates a Chatbot style architecture. The intended outcome of the personification of the app is that it will lead to increased engagement and reduced feelings of negative emotions from the user. The chatbot runs using the gpt-3.5-turbo model from OpenAI. It answers any school related questions in a simple, succinct way. If a child attempts to distract from the topic, the app will attempt to refocus the child. Crucial to achieving a suitable chatbot is prompt engineering.

The type of prompt engineering used is called ”system role-based instruction.” It involves using prompts included in the code to set the context and guide the model’s behavior throughout the conversation. Thus, it is possible to guide the model to generate more targeted and contextually relevant responses. For example, the prompts used in the app include content such as the emotional state of the user, the age of the user, engagement levels of the user, etc. This approach aims to make the model generate answers that are suitable for a young child of the given age and cater to them.

An essential consideration while prompt engineering in this project was the temperature of the LLM. The temperature determines how creative the model is. A lower temperature makes the model more deterministic, and a larger temperature imposes fewer limits on the available words for the model to use. After testing a few temperatures, it was obvious that a lower temperature would be needed

in this context. The lower temperature assures the quality of the answers, and in the case of an ITS and teaching students, the quality had to be prioritised over the diversity of language.

Specifically, prompt engineering experiments were performed to search for the best prompt and temperature that fit the purpose of this ITS. A range of questions and prompts and temperatures were selected. In the appendix some sample experimentation with prompting and temperature is detailed. Table 6

2.1.2 Algebraic Expression Simplifier and Solver

AI chatbots, such as the ChatGPT that has been implemented in other areas of the app, are famously inaccurate when given specific maths questions. This led to the creation of a separate code to power the 'numbers' section of the website.

This module has two main functions: simplifying mathematical expressions and resolving mathematical equations. It is capable of solving problems that involve basic mathematical operations such as addition, subtraction, multiplication, division, exponentiation, and brackets. However, it is limited to solving valid questions that involve simple mathematical symbols and operations, such as ' $2x+1=3$ ', and is unable to solve problems that contain words or complicated operations of exponentiation, such as '9 times 3', powers of fractions and powers of variables. Since the target group of children is not expected to ask more complicated questions, it is considered that the capability of this module can satisfy the user needs. For the problem entered, a step-by-step solution that involves both word explanations and mathematical expressions will be generated to assist children in their understanding of the problem-solving process.

The function to be activated is determined by whether or not the problem includes the equation symbol '='. If the symbol is present, the maths solver function is launched. Otherwise, the simplification function is activated. These tasks are accomplished through the deployment of a JavaScript model called mathsteps [2] is deployed for this function. The model utilises mathjs expression node to simplify the expression or solve the equation by searching through hardcoded documentation. To incorporate the JavaScript code into the Python environment in which the ITS is constructed, the js2py library is employed.

2.1.3 Facial Expression Recognition

The app has integrated **Facial Expression Recognition AI** which is capable of interpreting the emotional state of a child from their facial expression. The mascot - Filo the puppy - then mimics the emotion of the child on the screen. The reasoning behind implementing this feature was to make the app more engaging and personal for the child using it.

In addition to this, if the facial recognition feature finds that the child is expressing negative emotions for 80% of the last 50 seconds (with a photo taken every 5 seconds), Filo will pop up on their screen with an encouraging statement. The encouraging statement is once again powered by chatGPT with some prompt engineering to make it more suitable for the context. Before using the app, the child should enter their age - this is included in the prompt so that appropriate language is used for the child's age group. The prompt 'A child's age-year old child is feeling emotion' is given as additional context for the LLM.

There are clear privacy concerns around this particular aspect of the app. Taking electronic images of children should only be done with the appropriate safe-guarding procedures in place and with due regard to the relevant ethical and legal requirements. Were the project to be extended and detailed evaluation on target users be undertaken, this safe-guarding would need to be done with the highest priority.

Detail about how this AI was implemented is in section 2.2.

2.1.4 Handwriting Recognition

In the current market, applications are available to students that allow them to upload written mathematical problems. Photomath and Microsoft Math Solver are highly popular, given that mathematics done in the classroom is generally done in written form. In fact, it is recognised that physically writing

in mathematical notation helps the process of learning. In order to make the app as useful as possible and competitive in this market space, it was decided to incorporate handwriting recognition for mathematical formulae. Further detail about how Optical Character Recognition was implemented is in section 2.2.

Uploading the image via the website could be considered a limitation since most students take pictures on their phones and would need to spend extra time finding a way to transfer it to their computer. In future versions of the app, streamlining this procedure might be a priority.

2.1.5 PDF Evaluation

The app includes a PDF Evaluation feature that allows students to upload their notes and utilise the prompt bar in two ways: by asking a question about the uploaded file or requesting a multiple-choice question to test themselves. This feature is particularly useful for students who study independently or don't have access to someone who can quiz them. The student will have the ability to increase their engagement with the material and be quizzed on it, thus improve their chances of retaining that information.

Detail about how this AI was implemented is in section 2.2.

2.1.6 Other Features

There are a host of other tools that have been implemented on the ITS for coherence, usability and to enhance the uniqueness of the concept. One of such tools is a work/break countdown. At the bottom of each page throughout the website, a student can start, pause and reset the counter. This is based on the Pomodoro Technique which is a time management method based on 25 minutes of focused work followed by 5 minute breaks. Obviously, it would be expected that such a tool would be used by the upper end of the target age bracket, since it's mostly helpful for extensive, focused study.

The system incorporates two task management tools that enable students to prioritize tasks and access them throughout the completion process. These tools are the Pyramid and the Urgent-Important Matrix. The 1-3-5 Pyramid works as such: the page displays nine text boxes in a pyramid formation. At the top of the pyramid, the task with the highest priority should be entered. The second row of the pyramid is for tasks that will be started as soon as the resources to do so become available, and the bottom of the pyramid is for tasks that will be started in the near future. The Urgent-Important matrix similarly helps students manage their current tasks. Four boxes are available to be filled in with homework/tasks that are Urgent and Important, Urgent but Not Important, Important but Not Urgent, and Neither Urgent nor Important. Similarly to the Pomodoro Counter, these tools are likely for children at the upper end of the target age bracket. However, these study tools perpetuate the design motivation that this tutoring app is useful to be used in conjunction with current studies and homework.

Users of the website can also access their previous chats. When completing homework or studying a topic they have asked Filo about before, they can quickly navigate to their previous conversation about it and see the questions and answers. In addition to this, they can seamlessly continue this chat and it will be saved again.

A leaderboard was added in order to introduce a social element to the website. Users can see who has interacted with Filo the most, and the number of queries they have submitted. In further versions of the app, a 'friends' feature could be added, so that the leaderboard would reflect usernames familiar to the user. This is a common feature of other tutoring systems, and engages students since they can keep updated with their peers outside of the classroom.

Finally, there is a 'My Account' section. Included here are: the number of conversations, the number of entries, the average number of entries per conversation, the first conversation date, the last conversation date, the longest conversation length, current sentiment, and average happiness level. There is also the option to change the username and password of the account - or to delete the account altogether. These features may be of interest to older students, but are expected to be primarily used by parents of younger students to check on the progress of their child, and even track their average mood while using the app.

As mentioned earlier, all of these features are instrumental in creating a cohesive system that strengthens the initial motivation of the concept of Filo.

2.2 AI Used and Developed for the ITS

2.2.1 DNN for solving algebraic equations

Following [4] an attempt has been made to implement a DNN-based solver for algebraic equations. While it is relatively trivial to implement a symbolic solver for the sorts of equations in scope, this was considered a more suitable approach for the project given its novelty. It would allow for the exploration of the limitations of such techniques when applied to this space.

The underlying idea is to treat symbolic representation of an algebraic expression as a string. The shape of that string is determined by a set of well defined rules. Any algebraic expression can be represented as a tree with nodes consisting of operators, constants or variables. [10] provides an exposition of how algebraic expressions, trees and strings can be generated from each other.

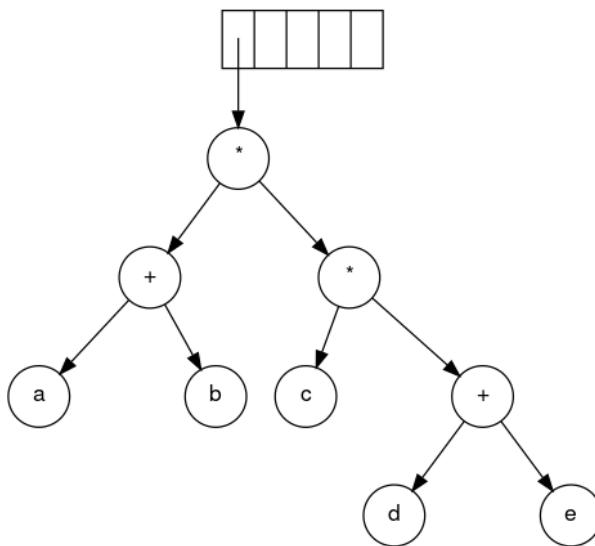


Figure 2: Steps to construct an expression tree for $*+*abc+de$, Source: By Abloomfi - Own work, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=31158993>

Using this sort of formulation, algebraic expressions can be generated by generating well formed binary trees. With the well formed binary trees, it is then possible to generate the appropriate strings. In the example above a tree is generated representing $(a + b) * c * (d + e)$ with a corresponding string representation $* + *abc + de$.

Given any algebraic expression, there is an equivalent string representation. Also, any given algebraic expression can be ‘simplified’. That is, methodically altered until it is in some reduced form using the usual rules of algebra. In this way then, a relationship between two strings can be defined: one pre-simplified and one post-simplified. The modeling of this transformation can then be cast as a ‘seq2seq’ type problem.

A large corpus of pre-simplified string / post simplified string pairs is required for training. In order to do this, well formed random trees, which are easier to generate than well formed algebraic expressions or well formed string representations, are generated. The algebraic expressions are passed to a website with incorporates an analytical solver (<https://www.cymath.com/>) which then is used to simplify the algebraic expressions. The website is then scraped for those expressions and then once again those expressions are represented as strings. In this way, pairs of expression strings and simplified expression strings are generated. Training a ‘seq2seq’ model on these expressions can then be attempted.

The sequences modelled are short with a maximum length of 20. With the absence of long dated dependencies, it was decided to use a simple RNN encoder-decoder model with attention. The archi-

ture follows [3] and has been modified to use algebraic string representations as opposed to Natural Language. A diagrammatic representation of the architecture is taken from [6].

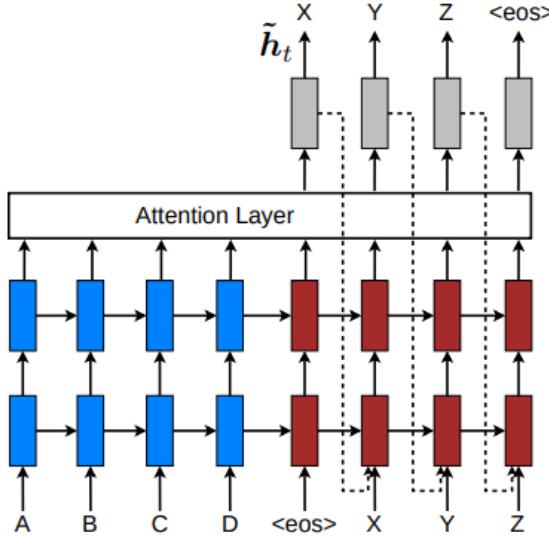


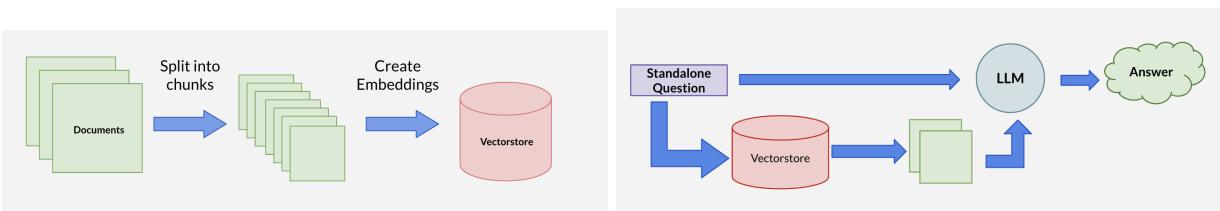
Figure 3: Encoder-Decoder architecture

Details of the training and performance of the model are reported in the evaluation section.

2.2.2 Q&A Generation for student evaluation

An important part of any tuition system is to have an ability to allow a student to evaluate their own progress. The obvious way to do this is to provide relevant testing. Since the start of the project, there has been further significant progress in the field of LLMs and this has been leveraged for the purposes of student evaluations.

A key challenge in producing evaluative testing is to generate testing relevant to the particular student and subject. Simply asking an LLM agent to generate a question-and-answer pair will not necessarily lead to optimal results as the LLM agent will not have access to the specific context in which the student is being assessed. In order to get around this issue, users are enabled with the ability to upload a relevant set of documents - eg an exam syllabus - and then generate evaluation tools germane to that material. The Langchain package has been used in order to do this efficiently. The schematic below shows how this process works in practice.



The uploaded documents are first split into chunks of a particular length. Those chunks are then embedded using pre-existing embeddings (OpenAI by default). These embedded chunks are stored in a Vectorstore. When the LLM is queried with a specific query, chunks from the Vectorstore which are semantically similar to the query are passed to the LLM for context. This process is known as "Retrieval Augmented Generation".

This technique affords the ability to build a number of different evaluation techniques, as well as allowing the user to directly query their uploaded documents. A qualitative evaluation of the QA generation potential of this utility is provided in the evaluation section.

2.2.3 Facial Recognition for student sentiment

As noted previously, having real-time facial expression analysis and providing appropriate feedback through the mascot could help to provide a more interactive experience for children. The application requires access to the camera, and images would be obtained from the real-time stream video every five seconds for analysis and outputting results.

The task is divided into two stages: face detection and emotion recognition. The former is achieved via YOLO (You Only Look Once) [8] techniques, while the latter is accomplished using a CNN model. YOLO is an advanced object detection algorithm that uses a deep neural network to achieve real-time detection with high accuracy. It works by dividing an image into a grid of cells and predicting the bounding boxes and class probabilities for each object within each cell. Once the cropped images of detected faces are obtained, they are passed on for emotion recognition. A VGG-19 [9] network is adopted for the expression classification task. VGG-19 is a robust CNN architecture which contains 19 layers that employ small 3×3 filters, pooling layers, and fully connected layers, making it suitable for various computer vision tasks. The model predicts probabilities for seven different classes (angry, disgust, fear, happy, sad, surprise, neutral), and the label with the highest probability is produced as the final prediction and used for the mascot's feedback.

2.2.4 Handwriting recognition

To improve user experience and speed up the process of manually typing up problems, a handwriting recognition module is developed. This module is able to convert handwritten text and maths symbols in images to strings ready for further processing.

To achieve this objective, a TrOCR (Transformer-based Optical Character Recognition with Pre-trained Models) model [5] was deployed for handwriting recognition. TrOCR utilises an encoder-decoder structure, where the encoder captures the representation of the images and the decoder generates text using the visual features and prior predictions as guidance. To improve the performance, TrOCR was initialised using pre-trained ViT-style models for encoder and pre-trained BERT-style models for decoder. This end-to-end CNN-free architecture was shown in the literature to outperform the traditional approach of using CNN for image interpretation and RNN for text generation.

2.3 Main design choices and justifications

One of the initial design choices for the Tutoring System was to make it a website. The addition of a log on page ensured the safety and security of the children using it. Most children now have access to computers and a large portion of homework is assigned through computer programmes. A website/app would allow for flexibility and easy access from any device with internet access and at any point can be brought up by the child so they can use it whilst doing homework or just studying.

The need for a mascot was agreed upon early. Personification of some of the features of the app was necessary to enhance the engagement of the target users, especially the younger ones, and reduce feelings of boredom or frustration. Having the mascot mirror the emotions of the user was thought to help with this. The mascot was designed to particularly appeal to a younger target age group with the main colour choice of green - a colour associated with tranquility, peace and concentration.

2.4 Architectural diagram

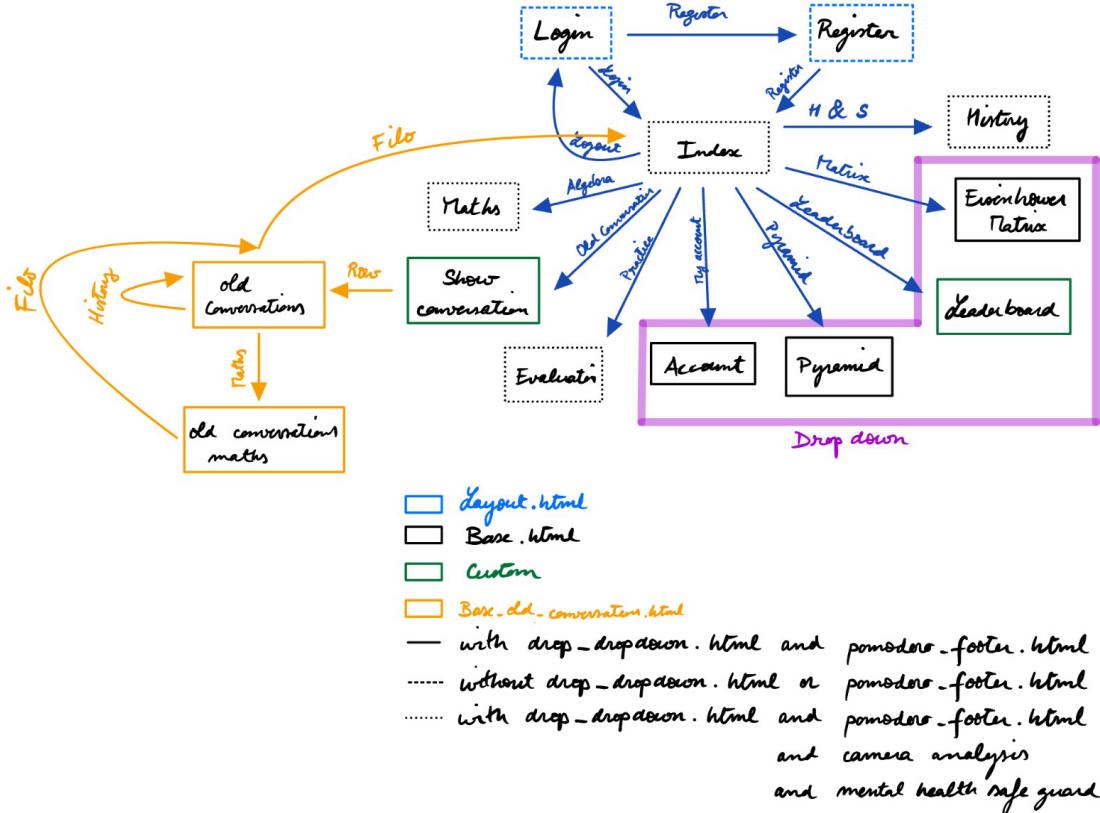


Figure 4: Architecture of Filo web app

Figure 5: Database schema

The application under discussion was developed using Python and Flask. In order to ensure smooth and efficient continuous integration/continuous development (CI/CD) as well as collaborative development, Gitlab was employed. Initially, the application was deployed using Heroku. However, due to a multitude of issues faced, the decision was taken to switch to an Infrastructure as a Service Provider (IAAP) instead of Platform as a Service (PAAS). A Google Cloud virtual machine was set up to host the application, which runs on port 5000. Additionally, an external IP address was provided to enable access to the application from any location.

2.5 App Screenshots and Link to web-app for production

A number of screenshots of Filo are provided, detailing a few of the different features, giving a sense for the overall feel of the web app. The app can be accessed via the following url: <https://34.142.18.175:5000/login> using the Username: Imperial and Password: filosegp.

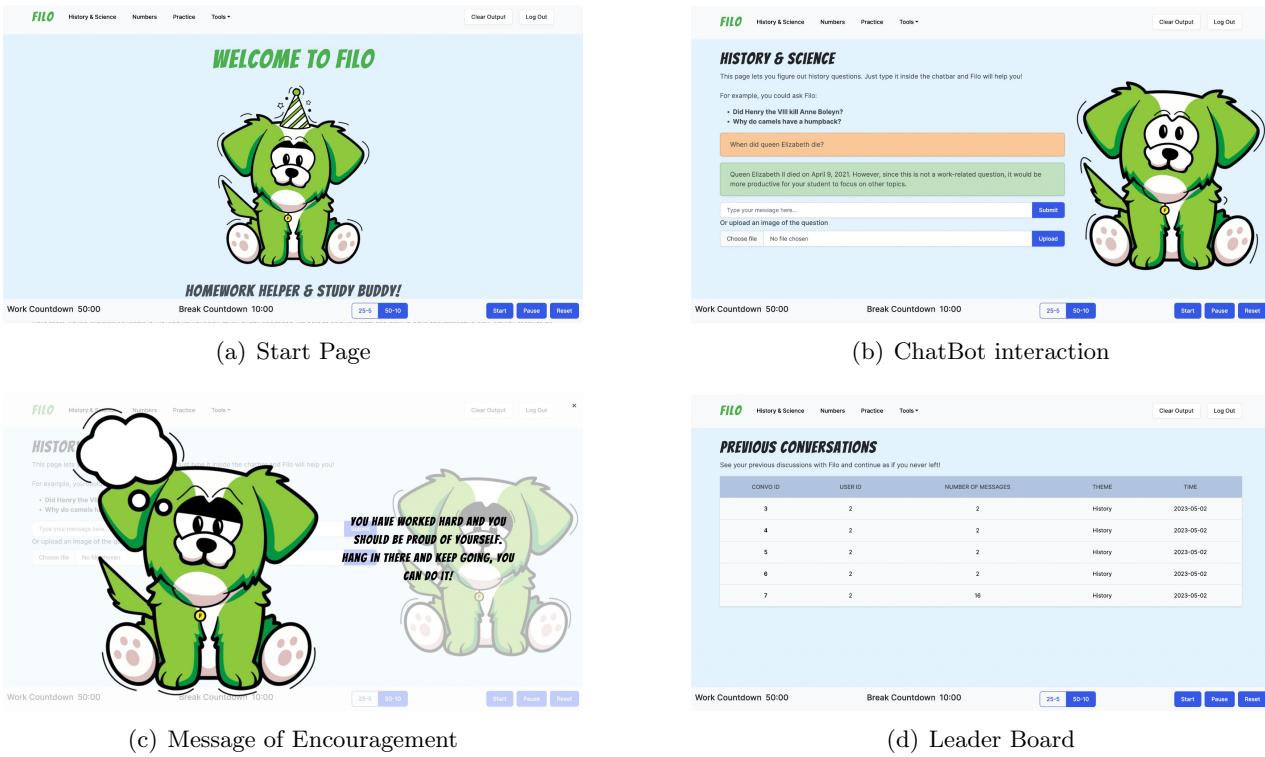


Figure 6: App Screen Shots

3 Evaluation

In this section, the final product was evaluated via both quantitative tests using either an online-sourced or internally-developed dataset, as well as qualitative assessments through user feedback from two 9-year-old children.

3.1 Quantitative Test of Performance

3.1.1 Facial Expression Recognition

The performance of the facial expression recognition module was examined through computing the confusion matrix, precision, recall and F1-score using 500 test samples from the FER2013 [1] dataset. This dataset consists of 48x48 pixel grayscale images of faces which is categorised into the same seven categories (angry, disgust, fear, happy, sad, surprise, neutral) that our model used to label the emotions. Note that the test samples are not balanced as shown in Table 1, therefore normalised confusion matrix and metrics per class are presented to evaluate the performance. Figure 7 demonstrates the normalised confusion matrix from the test results, and Table 1 presents the metrics. The model achieves a macro averaged accuracy of 0.76, with most of the misclassification occurring in the case between similar negative feelings, which have limited impact on the model's effectiveness. Overall, these results suggest that the model is performing well for the purpose of this module in the system.

3.1.2 Image to Text Transformer

A test set containing 50 samples was created to test the performance of the model. This set includes 10 handwritten equation images from the 5 members of this group, providing diversity in the handwriting style. An example image of a test sample is presented in Figure 8. For each maths equation expression, the text extracted was compared with the ground truth element-by-element, and the partial accuracy was computed accordingly. The average accuracy for all test samples is 0.78. This indicates that the function can potentially save children up to 78% of the time required to type out the questions.

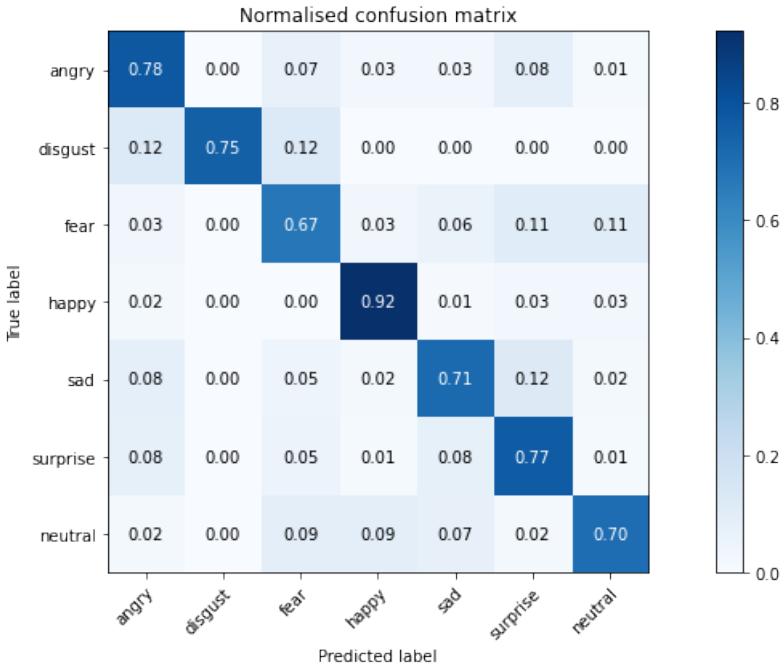


Figure 7: Normalised Confusion Matrix for facial expression recognition.

	precision	recall	f1-score	support
angry	0.73	0.78	0.75	72
disgust	1.00	0.75	0.86	8
fear	0.69	0.67	0.68	66
happy	0.90	0.92	0.91	115
neutral	0.82	0.71	0.76	108
sad	0.66	0.77	0.71	77
surprise	0.73	0.70	0.72	54
macro avg accuracy	0.79	0.76	0.77	500
weighted avg accuracy	0.78	0.77	0.77	500

Table 1: Metrics for the performance of facial expression recognition

3.1.3 Maths Equation Solver

The robustness and performance of the maths solver were examined by testing against edge cases, including questions that cover all basic mathematical operations, as well as invalid questions or questions not in the expected form. The results are presented in Table 2 for passed tests and Table 3 for failed tests. It was found that the model was capable of solving the questions of the correct form and is valid, while it produced empty responses or errors when the questions were not in the accepted form or not valid, as expected, with the only exception being the case '0x=0', where the model provided an incomplete solution. It should be noted that certain test cases fall outside the scope of the target audience, and they are primarily to assess the robustness of the model.

Test Cases	Results
0^0	Step 1, reduce exponent by zero: 0^0 becomes 1
$2(x+1)^5$	Step 1, collect and combine like terms: $2 * (x + 1) * 5$ becomes $10 * (x + 1)$ Step 2, distribute: $10 * (x + 1)$ becomes $10x + 10$
$((5+2*2)/3-1)^2x=20$	Step 1, simplify left side: $((5 + 2 * 2) / 3 - 1)^2 * x = 20$ becomes $4x = 20$ Step 2, divide from both sides: $4x = 20$ becomes $(4x) / 4 = 20/4$ Step 3, simplify fraction: $4x / 4 = 20/4$ becomes $x = 20/4$ Step 4, simplify fraction: $x = 20/4$ becomes $x = 5$

Table 2: Example Passed Tests

Test Cases	Results
Invalid equation: $2/0x = 3$	<code>TypeError: Undefined and null dont have properties (tried getting property 'ascii')</code>
Undefined expression: $1/0$	Empty response.
Invalid equation: $5=6$	<code>TypeError: Undefined and null dont have properties (tried getting property 'ascii')</code>
Word expression: 9 times 3	Step 1, collect and combine like terms: 9 times * 3 becomes 27 times
Power of fractions: $(5+2*3/4-1)^2x=20$	Empty response.
Power of fractions: $(11 / 2)^2$	Empty response.
No solution: $0x=1$	<code>TypeError: Undefined and null dont have properties (tried getting property 'ascii')</code>
Infinite solutions: $0x=0$	Step 1, find roots: $0x = 0$ becomes $x = 0$
Power of variables: $x^3=8$	Empty response.

Table 3: Example Failed Tests

$$5x + 1 = 6$$

Figure 8: Example of a handwritten equation test sample.

3.1.4 DNN for solving expression

The DNN training did not work as well as planned. No translation model between pre-simplified and post-simplified expressions was learned. There are multiple hypotheses for this failure.

- It was not possible to generate sufficiently large sample sizes in an efficient way. For the RNN architecture described, tens of thousands of pairs would be required. The webscraping methodology employed was not efficient enough to do this.
- Attention mechanism in the RNN was not sufficient to produce context for the RNN encoder-decoder. Adding just one element to the string entirely alters the structure of the tree formed. Perhaps it would have been more appropriate to use a Transformer Architecture.
- It is well known that Machine Learning techniques struggle with representations of mathematical concepts - eg LLMs. Perhaps this was a harder problem than anticipated.

It was disappointing to not be able to report more positively on this front. Significant time was spent on trying to make this work. With more time perhaps the issues could have been overcome. However, it was decided that other features should be developed in order to produce a viable product.

3.1.5 PDF Evaluation

A free PDF of a KS2 Science book published by Collins was downloaded. This was then used as source material to generate any type of questions desired by the user. The user could ask for a series of multiple questions and gauge and better their understanding. An example of the type of conversation between a user and the bot follows:

- User: Give me a random question and its answer on any topic in the book.
- Chatbot: Why do camels have a humpback? It helps them survive in the desert by storing fat and hence energy. A number of extensions to this functionality were left unimplemented in the final version of the app due to time constraints. Indeed, this required deep Javascript expertise. As this is a very important part of the app, details of these are provided below.

Firstly, keywords/phrases for a specific subject were generated. Using these keywords an LLM generated sets of ‘question’ and ‘answer’ pairs which included the keyword/phrase in the ‘answer’. These pairs could then be used in three different ways:

- Generate a list of syntactically similar words/phrases to the keyword from the relevant context. Mask the keyword/phrase in the ‘answer’ and ask the user to select the correct word from a list including the syntactically similar words/phrase along with the keyword in response to the ‘question’
- Mask the keyword/phrase in the ‘answer’ and ask the user to enter the correct word/phrase in the ‘answer’ in response to the ‘question’
- Ask the user to enter a response to a ‘question’

These three categories of evaluation are broadly increasing in difficulty for the student. In the first category, there is a simple correct/incorrect classification. For the second the classification is also straightforwardly correct or incorrect. In the third category, a semantic entailment metric is implemented to assert whether a user enters a correct or incorrect answer based on the ‘answer’ generated in the pair as ‘ground-truth’. Using the roberta-large-mnli model, an assessment is made as to whether the user answer entails the correct answer and whether the correct answer entails the user

answer. If the entailment is symmetric, the answer provided is asserted to be correct. If the ground truth answer entails the entered answer but not vice versa, the answer provided is partially correct.

This approach does require some fine tuning of the probabilistic thresholds for semantic entailment which were not completed in time. It also requires the search for key words to be robust. The search worked well for science based documents but performed poorly on subjects such as history. Improvements to generalise the search for key words was not completed on time.

3.2 Testing User Needs and Product Robustness

In order to evaluate the application on the target market, two 9-year-old children used the website for about ten minutes each. Feedback was also collected from their mother, to gain a perspective from a parent of younger users of the website.

There was certainly perceived engagement from both children, although consideration must be made for the novelty of trying something new. When asked about the design of the app, there was positive feedback about the colours and Filo, the mascot.

When informed that the mascot would reflect the emotion the user portrayed, Child One became very engaged, and purposefully smiled while inputting his query to the chatbot. This reinforced the expectation that the mascot would mostly appeal to and engage the younger audience.

Within a short period of time, it became clear that the chatbot needed to be more age friendly to younger users. Child One asked a straightforward question about Henry VIII, and the chatGPT based chatbot responded with some detail. For younger users of the app, picking out relevant details amongst a lot of text is difficult. However, limiting the number of words outputted by the chatbot depending on the age of the user is not an insignificant task, and would be a consideration for future versions of the application.

Child 2 attempted to ask the chatbot a question about Lego. There are measures in place intended to divert such questions, and the chatbot should output the following: 'I'm sorry, but we are here to focus on your work and learning. We can take a break later for playtime, but for now, let's focus on our studies. What subject would you like to work on?'. However, in this instance, it was unable to detect the digression in topic, and responded to the Lego question. Following this feedback, stricter measures were implemented however it is not 100% accurate at detecting irrelevant questions, and this is unfortunately a limitation of the app that was unable to be wholly overcome.

Upon discovery of the leaderboard, Child 2 requested the link to the website so that he could 'share it with [his] friends'. This positively reinforced the reason for implementing the feature, and led to discussion about future possibilities about its place in the application.

Lastly, feedback was given from the mother of the test subjects. When asked whether she felt that the app was suitable for her 9 year old children she said that the chatbot function might be slightly beyond their capabilities since their homework still included things like very simple arithmetic. Further, the 'number' portion of the app interprets the 'x' symbol as algebraic, rather than a multiplication sign, and this reported to likely be confusing for the younger children.

The user feedback was immensely helpful in reviewing the strengths and weaknesses of the final product. Although product robustness is hard to ensure when attempting to harness the powers of large models like ChatGPT for a specific purpose, the user feedback confirmed that larger purpose of creating a 'Homework Helper' was achieved.

4 Conclusion

In conclusion, after considering the application, though largely positive, some negative aspects were identified.

One of the challenges not overcome was in developing the DNN to aid with symbolic mathematical manipulation. It is recognised that LLMs struggle with this sort of task. Having a separate hand-crafted workflow in order to assist with even a small subset of these sort of problems would have provided a much more satisfactory outcome.



Figure 9: Images of children of target age using the app

As mentioned in section 3.1.5, several promising additions to the functionality for student evaluation were left unimplemented due to time constraints. They would have enhanced the product significantly.

Importantly, if the application were to be introduced to the public, there would likely need to be more work on product robustness. This is due to the fact that the app would be utilised by children and safeguarding must be the number one priority with this target market.

It might be helpful to have features split by age group in future versions of the application, such that children in the bottom of the age bracket receive straightforward answers, and are not able to ask the chatbot any irrelevant questions. Similarly, older children should not receive any encouraging pop-ups that they deem 'condescending' or 'babyish'.

Despite these observations, upon evaluation the app certainly delivers against the problem statement. It is a totally unique Intelligent Tutoring System that incorporates several AI elements to enhance the experience of the user. The AI technologies employed in conjunction with the tools implemented in the software wholly achieved the aim of 'Homework Helper, Study Buddy.' Though the evaluation was limited to two test subjects, the feedback from those users was overwhelmingly positive. There is no equivalent product on the market.

References

- [1] Retrieved April 2023 from <https://www.kaggle.com/datasets/msambare/fer2013>.
- [2] Matt Gaunt-Seo. mathsteps. <https://github.com/google/mathsteps>.
- [3] Sean Robertson |https://github.com/spro/practical_pytorch|. Nlp from scratch: Translation with a sequence to sequence network and attention. https://github.com/pytorch/tutorials/commits/main/intermediate_source/seq2seq_translation_tutorial.py, 2017.
- [4] Guillaume Lample and François Charton. Deep learning for symbolic mathematics, 2019.
- [5] Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models, 2021.
- [6] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation, 2015.
- [7] Wei Ma, Olusola O. Adesope, John C. Nesbit, and Qingqing Liu. Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4):901–918, 2014.
- [8] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- [9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [10] Wikipedia contributors. Binary expression tree — Wikipedia, the free encyclopedia, 2023. [Online; accessed 30-April-2023].

A Prompt engineering examples

Temp	Answer
0	Mathematics is an important subject that helps you develop problem-solving skills, logical reasoning, and critical thinking abilities. It is used in many fields such as science, engineering, finance, and technology. Learning math will also help you in your daily life, such as calculating change, measuring ingredients for cooking, and managing your finances.
0.2	Mathematics is an essential subject that helps you develop problem-solving skills, logical reasoning, critical thinking, and analytical skills. It is also used in many real-life situations, such as calculating change, measuring ingredients for cooking, and managing finances. Learning math will help you understand the world around you better and prepare you for future academic and career opportunities.
0.4	Mathematics is an essential subject that helps you develop critical thinking and problem-solving skills which are important in everyday life. It is also a fundamental subject that forms the basis for many other subjects like science, engineering, and economics. Additionally, math is used in many professions, such as architecture, medicine, and finance, so having a good understanding of math will help you in your future career.
0.6	Maths is a fundamental subject that helps you develop analytical and problem-solving skills. It is used in many areas of life, from calculating the amount of money you need to buy something, to designing buildings, to understanding scientific concepts. Additionally, many careers require a basic understanding of maths, and so learning it now will help you in the future.
0.8	Mathematics is a very important subject as it is applicable in our daily lives. By learning mathematics, you will develop your problem-solving skills, logical thinking, and analytical skills. Math is also necessary for many occupations and careers such as engineering, science, technology, finance, and many more.

Table 4: Question: 'Why do I have to learn maths', Prompt: 'You are tutoring a 10-year-old child that's been feeling a bit frustrated. Take an educational approach and answer only work related questions. '

Temp	Answer
0	Mathematics is an important subject that helps us understand and solve problems in our daily lives. It is used in many fields such as science, engineering, finance, and technology. Learning math also helps develop critical thinking skills, logical reasoning, and problem-solving abilities. It may seem difficult at times, but with practice and patience, you will be able to master it and use it to your advantage in the future.
0.2	Mathematics is a very important subject that you will use throughout your life. It helps you develop problem-solving skills, logical thinking, and critical reasoning abilities. You will use math in many areas of your life, such as managing your finances, cooking, measuring, and even playing games. It is also a fundamental subject that is required for many careers, such as engineering, science, and technology. So, learning math now will help you in the future and open up many opportunities for you.
0.4	Mathematics is an important subject that helps us understand the world around us. It is used in many different fields, such as science, engineering, finance, and technology. Learning math can help you develop problem-solving skills, logical reasoning, and critical thinking skills, which are important for success in many areas of life. Additionally, math is a subject that builds on itself, so the more you learn now, the easier it will be to learn more advanced concepts in the future. So, while it may
0.6	Mathematics is an essential subject that helps you develop critical thinking skills and problem-solving abilities. It is used in many aspects of our daily lives, such as cooking, shopping, and managing money. Moreover, many exciting and rewarding careers require a solid foundation in mathematics, such as engineering, science, technology, and finance. Learning maths may seem challenging at times, but it can also be fun and rewarding. With practice, you can become confident in your ability to solve problems and achieve your goals.
0.8	Mathematics is a very important subject that we use in our daily lives. We need it to make sense of the world around us, to solve problems, and to make informed decisions. For example, we use math when shopping, cooking, measuring, and even playing games. Math also helps us to develop critical thinking skills and logical reasoning which can be applied to many other areas of life. So, even if you don't plan to become a mathematician, learning math will benefit you in many ways

Table 5: Question: 'Why do I have to learn maths', Prompt: 'You are tutoring a 10-year-old child who is a little frustrated. Answer the child as a teacher might answer their question. '

Temp	Answer
0	Maths is important because it helps us understand and solve problems in our daily lives. It helps us with things like counting money, measuring ingredients for cooking, and even playing games. It also helps us develop critical thinking and problem-solving skills that we can use in many areas of our lives. While it may seem frustrating at times, learning maths can be fun and rewarding once you start to see how it applies to the world around you.
0.2	Mathematics is important because it helps us understand and solve problems in our daily lives. It is used in many fields such as science, engineering, finance, and technology. Learning math also helps develop critical thinking and problem-solving skills that can be useful in any career or situation. While it may be frustrating at times, practicing math can help you become more confident and successful in the future.
0.4	Math is important because it helps us solve problems in our daily lives, like figuring out how much money we need to buy something or how much time we have to finish a task. It also helps us understand the world around us, like how shapes fit together or how patterns work. Plus, learning math can be fun and rewarding!
0.6	You have to learn maths because it is a very important subject that is used in many areas of life. Maths helps you to solve problems, understand patterns and relationships, and make sense of the world around you. It is also a key skill that is required in many careers such as engineering, finance, science, and technology. By learning maths, you will be giving yourself the tools to succeed in the future.
0.8	Maths is important because it is a fundamental skill that you will use in many aspects of your life. It helps you develop problem-solving and critical thinking skills, which will be useful in many areas of work and everyday life. Maths can also be fun and interesting, and can help you understand the world around you better. It may be challenging at times, but with practice and patience, you can become better at maths and enjoy it more.

Table 6: Question: 'Why do I have to learn maths', Prompt: 'You are teaching a 10-year-old child. Answer succinctly and clearly. Consider that they are slightly frustrated'

Temp	Answer
0	Mathematics is an important subject that helps us understand and solve problems in our daily lives. It helps us with things like counting money, measuring ingredients for cooking, and even playing games. It also helps us develop critical thinking skills and problem-solving abilities that are useful in many areas of life. While it may seem frustrating at times, with practice and patience, you can become better at math and see how it can be useful in your life.
0.2	Mathematics is a very important subject that helps us understand and solve problems in our daily lives. It helps us with things like counting money, measuring ingredients for cooking, and even playing games. It also helps us develop critical thinking skills and problem-solving abilities that are useful in many areas of life. While it may seem frustrating at times, with practice and patience, you can become better at math and see how it can be useful in your life.
0.4	Math is an important subject that helps us in many different ways. It helps us to solve problems, make decisions, and understand the world around us. For example, we use math when we go shopping, cook food, plan a trip, or even play games. Learning math can also help you develop important skills like critical thinking, logic, and problem-solving, which will be useful throughout your life. While it can be frustrating at times, with practice and patience, you can become better at math and even find it enjoyable.
0.6	Maths is an important subject that helps us in many areas of our lives. It helps us to solve problems, think logically, and make sense of the world around us. For example, we use maths when we go shopping, cook, plan a trip, or even play games. Learning maths may seem challenging at times, but with practice and patience, it can become easier and even enjoyable. Remember, every new skill takes time and effort to develop, so don't give up!
0.8	Math is important because it helps us solve problems in our everyday lives, from figuring out how much money we have to spend to measuring ingredients for a recipe. It also helps us understand and analyze patterns in the world around us. Learning math may be frustrating at times, but with practice and patience, you can become confident and successful in math.

Table 7: Question: 'Why do I have to learn maths', Prompt: 'You are teaching a 10-year-old child. Answer succinctly and clearly. Be caring and kind. The child is a little frustrated.'

Temp	Answer
0	<p>Hi there! I understand that you might be feeling frustrated about learning math, but let me tell you that math is a very important subject that you will use throughout your life.</p> <p>Math helps us to solve problems, make decisions, and understand the world around us. For example, if you want to buy something at the store, you need to know how much money you have and how much the item costs. You also need to know how to add, subtract, multiply, and divide to make sure you get the right change back.</p> <p>Math is also used in many different careers, such as engineering, science, and finance. If you want to be successful in these fields, you need to have a good understanding of math.</p> <p>I know that</p>
0.2	<p>Hi there! I understand that you might be feeling frustrated about learning maths, but let me explain why it's important to learn it. Maths is a very important subject because it helps us understand and solve problems in our daily lives. For example, when you go to the store and want to buy something, you need to know how much money you have and how much the item costs. Maths helps you figure out how much change you should get back after you pay for something. Maths is also important for many careers, such as engineering, science, and finance. If you want to be a doctor, you need to know how to measure doses of medicine. If you want to be an architect, you need to know how to calculate the area'</p>
0.4	<p>Maths is a very important subject to learn because it helps us in our daily lives. It helps us to solve problems, make decisions and understand the world around us. For example, when you go shopping with your parents, you need to know how much money you have and how much things cost so that you can make sure you have enough money to buy what you need. Maths also helps us to understand things like time, measurements, and shapes. I know that sometimes maths can be frustrating, but with practice and patience, you can become really good at it. It's okay to make mistakes, because that's how we learn. I'm here to help you understand maths better and to make it more fun for you. Let's work together</p>
0.6	<p>Hello! I understand that learning math can sometimes be frustrating, but it's actually a very important subject that we all need to learn. Math helps us to understand the world around us and solve problems in our daily lives.</p> <p>For example, if you want to buy a toy that costs \$10 and you only have \$5, you need to figure out how much more money you need to save up. This is where math comes in handy. You can use addition and subtraction to calculate how much money you need to save up to buy the toy.</p> <p>Math also helps us to understand patterns and relationships. For instance, if you notice that your height is increasing every year, you can use math to predict how tall you might be when you're older</p>
0.8	<p>Hi there! I understand that you may be feeling frustrated about learning math. Let me try to explain why learning math is important. Math is a subject that helps us in our daily lives. For example, when you go to the grocery store with your parents, you use math to calculate the total cost of the items you want to purchase. When you want to share your toys with your friends, you use math to divide them equally. When you want to build something, math helps you measure and cut the materials accurately. Without math, it would be difficult to do many of the things we do every day. Additionally, math is an important subject for many careers like engineering, science, and finance. By learning math, you will have an advantage in</p>

Table 8: Question: 'Why do I have to learn maths', Prompt: 'You are a tutor. Explain concepts to a child, bearing in mind they are only 10-year-old. They are frustrated so be considerate of that.'

Temp	Answer
0	Maths is important because it helps us solve problems in our daily lives, like counting money or measuring ingredients for cooking. It also helps us understand how things work in the world around us, like how shapes fit together or how patterns repeat. Plus, it can be fun once you get the hang of it!
0.2	Maths is important because it helps us solve problems in our daily lives, like counting money or measuring ingredients for a recipe. It also helps us understand how things work in the world around us, like how shapes fit together or how patterns repeat. Plus, it can be fun and challenging!
0.4	Maths is important because it helps us solve problems in our daily lives, like figuring out how much money we have or how much time we need to finish our homework. It also helps us understand how things work in the world around us, like how shapes fit together or how patterns repeat. And who knows, you might even find it fun!
0.6	Maths helps us understand and solve problems in everyday life, like making sure we have enough money for things we want to buy. It also helps us develop our critical thinking skills and problem-solving abilities. Plus, it can be fun!
0.8	Maths is important because it helps us solve problems in our daily life, like telling time, measuring ingredients while cooking, or calculating change at the store. It also helps us understand how the world around us works, like the shapes of buildings or the patterns in nature. Finally, it can lead to interesting and rewarding careers in fields like engineering, finance, and science.

Table 9: Question: 'Why do I have to learn maths', Prompt: 'You are a tutor. Explain concepts to a child, bearing in mind they are only 10-year-old. They are frustrated so be considerate of that. Try to keep it to under 3 sentences '