# Edvent
# Speech to Presentation Automation


| Medha Koppam | Rohan Mehta | Rudransh Kulkarni | Shruthi Pai | Srinivas K S |
|---|---|---|---|---|
| Dept of CSE | Dept of CSE | Dept of CSE | Dept of CSE | Dept of CSE |
| PES University | PES University | PES University | PES University | PES University |
| medhakoppam@gmail.com | rohanmehta394@gmail.com | rudranshklkrn@gmail.com | shruthipai1924@gmail.com | srinivasks@pes.edu |



***Abstract* - Through this research paper, we showcase the usage of cutting-edge technology to transform education, empowering students to unleash their true potential, learn, explore, and excel. By harnessing the power of visionary Generative AI and Deep Learning techniques, we present an innovative end-to-end solution that revolutionizes learning. Our state-of-the-art approach leverages Transfer Learning, Fine-tuning and Large Language Models to seamlessly convert a teacher's audio input into a captivating and comprehensive presentation. This dynamic presentation includes immersive visuals, concise text summaries, interactive simulations, self-assessment tools, and curated external resources. Our pioneering AI-powered solution goes beyond conventional learning methods, capturing the essence of a teacher's delivery to provide students with concise and relevant summaries that truly resonate. In addition, our system generates hyper-realistic images that take students on an immersive visual journey, igniting their curiosity and deepening their understanding. With curated resources and thought-provoking questions, our platform empowers students to delve deep into each subject, fostering a holistic and dynamic learning experience. By embracing this AI-driven educational paradigm, students can enjoy personalized learning experiences, access knowledge anytime and anywhere, and gain a deep understanding of complex concepts. Not only does this groundbreaking system empower students, but it also liberates teachers from the arduous task of content creation, allowing them to devote more time to personalized student engagement.**



***Keywords* - Education, Generative AI, Deep Learning, Teaching and Learning, Large Language Models**


## I. INTRODUCTION

With the advancement of technology, it has become more convenient to digitize and automate various aspects of our lives, including air conditioners and databases. In the field of education, it is important to stay updated with the latest technological trends to keep students engaged and interested in different subjects.

Research suggests that the human brain processes visuals much faster than text, with people retaining around 65% of information when it is paired with relevant images, compared to only 10% retention with text alone. Studies have shown that visual thinking is deeply ingrained in the brain, as even when people are prompted to use verbal thinking, they create visual images to accompany their thoughts.

According to a recent study released by America's Promised Alliance, the graduation failure rate in the U.S. stands at 30%. In the 2022 CBSE 12th Board Exams in India, 1.34 lakh students scored above 90%, and 92.7% cleared the exam. However, in the JEE Main examination, the top 10 percentile of students had an average score of 110/300 (36.66%). This indicates that some students who performed well in board exams struggled to score well in concept-based exams like JEE Main, possibly due to a lack of strong foundation through visual learning.

Unfortunately, in developing countries like India, access to quality education is limited, and many schools lack basic resources and internet services. As a result, students often struggle to visualize and understand various topics, making education theoretical and challenging to grasp. This puts them at a disadvantage in their educational journey.

A government survey in India revealed that only 22% of schools had internet facilities in the academic year 2019-20. Many students find it difficult to grasp fundamental concepts when taught using conventional textual methods and may benefit from visual aids.

Promoting digital content through automated solutions, such as AI-generated presentations with relevant graphics, important text, 3D models, lecture summaries, assessments, and educational links, can be a suitable approach to address this issue. These resources can be provided in an easy-to-use, editable .pptx format, customized according to the institution's preferences. This comprehensive approach to learning benefits all students and also saves time and effort.

This paper discusses the multi-step process of automating speech to presentation using various models, including GPT 3.5, Stable Diffusion, and Langchain. The paper explores different fine-tuning methods applied to generate optimal results.

To summarize, through this paper, we would like to propose an AI based end-to-end approach that generates immersive educational presentations making learning efficient and effective for students and saves hours of teachers' time in curating these materials. This gives teachers and institutions more time and effort to spare which they can then use to personally hand-hold students with their learning.

## II. LITERATURE REVIEW

### *A. Semi-supervised NMF Models for Topic Modelling in Learning Tasks [1]*

*a) Proposed Methodology:* A non-negative matrix is divided into two non-negative matrices using the Non-negative Matrix Factorization (NMF) model of matrix factorization. Typically, a document-term matrix serves as the topic modelling input matrix, with each row denoting a document and each column denoting a term (word) in the vocabulary. The two matrices are randomly initialised at the

beginning of the process, and the next two steps are alternated between: Modernise the document-topic matrix. A topic-term matrix update

*b) Pros:* In order to find latent topics in text data, NMF imposes non-negativity requirements on the factorization. The model only chooses a tiny portion of the words in the documents to represent the themes because the NMF also enforces sparsity constraints. Computing effectiveness: NMF is computationally effective and is capable of handling huge datasets as compared to other topic modelling techniques like Latent Dirichlet Allocation (LDA).

*c) Cons:* The initialization of the factorization can provide varied results and have an impact on the topic modeling's quality because NMF is sensitive to it. a challenge in estimating the quantity of themes. It is more difficult to assess the effectiveness of the topic modelling using NMF since it lacks a probabilistic interpretation, in contrast to Latent Dirichlet Allocation (LDA). Limited capacity to scale to huge datasets and processing of lengthy documents

d) *Learnings:* Non-negative Matrix high-dimensional vectors are factored into a low-dimensional form using the linear algebraic model of factorization. NMF makes use of the vectors' non-negative nature in a manner akin to principal component analysis (PCA).

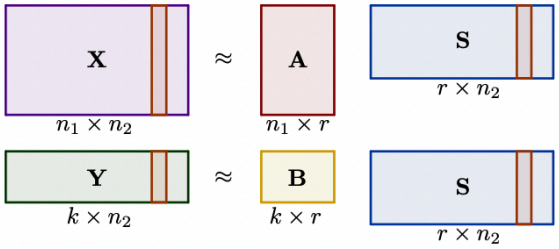


Figure 1: Given the number of classes k, and a desired dimension r, SSNMF is formulated as a joint factorization of a data matrix X ∈ R n1×n2 ≥0 and a label matrix Y ∈ R k×n2 ≥0 , sharing representation factor S ∈ R r×n2 ≥0 .

## *B. VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning [2]*

*a) Proposed Methodology:* GPT-2 and other pre-trained language models (PLMs) are trained using data from a single modality. They use a PLM as the caption decoder and feed the PLM visual data using encoder-decoder attention, which is essential for the PLMs to quickly adjust. They strive to carefully balance visual information from the encoder and linguistic knowledge contained in the PLM with the design of the encoder-decoder focus.

*b) Datasets:* On three datasets—MS COCO, Conceptual Captions, and IU X-ray—they tested their model. Each of the 123,287 photos in MS COCO has 5 distinct captions annotating it. For the validation set and test set, we adhere to the Karpathy divide. In comparison to COCO, the Conceptual Captions dataset has a far higher diversity of images, with about 3.3M for training and 28K for validation. They used the public validation data instead of the private test data as the test data was not publicly accessible, and they randomly selected 5000 unique image-caption pairings from the training set to serve as the validation set.

*c) Pros:* Via a minimal quantity of in-domain image-text input, the PLM can be swiftly adjusted via a self-resurrecting encoder-decoder attention method.

*d) Cons:* As in-domain training data grow, the gap between baseline models and VisualGPT gradually closes.

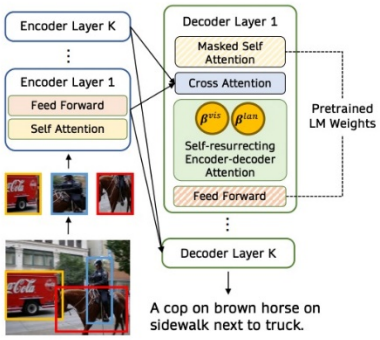


Figure 2: This VisualGPT model transfers the knowledge from a pre-trained language model to the caption decoder. A self- resurrecting encoder-decoder attention is designed to connect the multi-level visual features and caption decoder.

## *C. Automated News Summarisation Using Transformers [3]*

*a) Summary*: For the purpose of summarization, pre-trained language models based on the transformer architecture were implemented. Their research led them to the conclusion that carefully calibrated transformers based on pre-trained language models produced excellent results and produced a sound and fluent summary for a specific text document. The T5 model beat all other models, according on the computation of ROUGE scores for the predictions given by each of the models for comparative studies.

| Models | Evaluation Metrics | | |
|---|---|---|---|
| | ***ROUGE-1*** | ***ROUGE-2*** | ***ROUGE-L*** |
| Pipeline - BART | 0.38 | 0.28 | 0.38 |
| BART modified | 0.40 | 0.28 | 0.40 |
| **T5** | **0.47** | **0.33** | **0.42** |
| PEGASUS | 0.42 | 0.29 | 0.40 |

Figure 3: Evaluation and Comparison of mean ROUGE Scores

*b) Pros:* High-quality summaries: T5 is able to produce high-quality summaries that are loyal to the original text because it has been pre-trained on a huge corpus of text data and optimised for specific tasks like summarization. Versatility: Because T5 is a text-to-text transfer model, it may be honed to do a variety of text-to-text activities, such as summarising, translating, answering questions, and more. Support for several languages: T5 is a useful tool for summarising content in various languages because it can be adjusted for a variety of languages. - T5 can be fine-tuned on certain domains, such as news stories, academic papers, or legal documents, to produce summaries that are suited to the particular topic.

*c) Cons:* Computing power: T5 requires a lot of computing power to train and fine-tune due to the high number of parameters it has. For businesses with few resources, this could be a restricting constraint. Training data relevance and quality: These factors have a significant impact on T5's performance. The performance of the model may deteriorate if the training data are inaccurate, noisy, or not representative of the target domain. Similar to other neural language models, T5 may have trouble with uncommon or infrequent words that are poorly represented in the training data. This can lead to inaccurate summaries or mistakes.

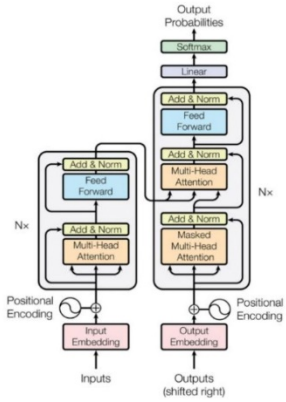


Figure 4: This figure represents the working of the T5 transformer

*D. High-Resolution Image Synthesis with Latent Diffusion Models [4]*

*a) Proposed Methodology:* The method begins with a pixel-space examination of diffusion models that have already been trained. Learning can be divided into two stages, broadly speaking, as with any likelihood-based model: The first stage is perceptual compression, which eliminates high-frequency information while learning just a small amount of semantic diversity. The actual generative model picks up on the semantic and conceptual organisation of the input (semantic compression) in the second stage. The approach thus seeks to first identify a space where we may train diffusion models for high-resolution picture synthesis that is perceptually identical but computationally more suited. To gradually process/diffuse information in the information (latent) space, use ClipText for text encoding and Scheduler. Using the array of information that has been processed, the autoencoder decoder paints the final image. Improved computing efficiency using the latent diffusion model

*b) Pros:* Pay attention to the crucial, meaningful portions of the data. Train in a space with fewer dimensions and higher computational efficiency

*c) Cons:* Training progress is sluggish for LDM-1,2 due to small down sampling variables. After relatively few training steps, excessively high values result in stagnant fidelity

*E. Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures [5]*

*a) Proposed Methodology:* In this approach, a NeRF model is tuned to display 2D feature maps in the latent space Z of stable diffusion. In addition to a volume density, LatentNeRF produces four pseudo-color channels (c1, c2, c3, c4) that correspond to the four latent features that stable diffusion operates over. Because of the spatial radiance field and rendering equation, employing NeRF to represent the world implicitly imposes spatial consistency between different viewpoints. However, it is not trivial that Z can be represented by a NeRF with spatial consistency. Guidelines for sketch shape: The various text prompts can direct the shape towards more precise geometries that more closely resemble the text prompt. A linear approximation is sufficient to predict plausible RGB colors given a single four-channel latent super pixel, via the following transformation, Fig 5, which was calculated using pairs of RGB images and their corresponding latent codes over a collection of natural images.

$$\begin{pmatrix} \hat{r} \\ \hat{g} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} 0.298 & 0.187 & -0.158 & -0.184 \\ 0.207 & 0.286 & 0.189 & -0.271 \\ 0.208 & 0.173 & 0.264 & -0.473 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix}$$

Figure 5: Transformation to predict RGB colors

*b) Pros:* Instead of using the high-resolution images directly, LDM, a particular type of diffusion model, is used. LDM is taught to denoise latent codes of a pre-trained auto-encoder.

*c) Cons:* Similar to the majority of research that use diffusion models, the outcomes exhibit stochastic behaviour.

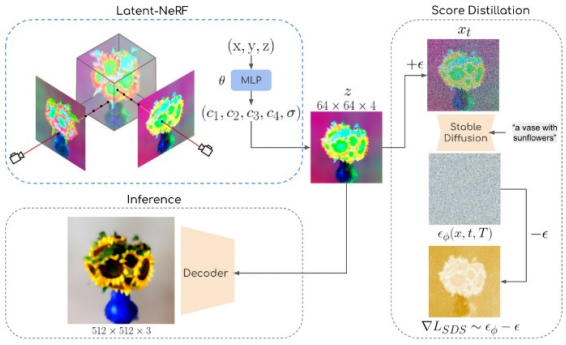


Figure 6: Represents working of the Latent Nerf model

*F. Learning to Ask: Neural Question Generation for Reading Comprehension[6]*

*a) Proposed Methodology:* The way a human would approach the problem is partially what the model is based on. People typically focus on certain elements of the input sentence and associate context information from the paragraph when posing a natural question. They use an RNN encoder-decoder architecture to describe conditional probability, and they employ the global attention technique to direct the model's attention when creating each word during decoding. Here, we examine two variants of our models, one that only encodes the sentence and the other that encodes information at both the sentence and paragraph levels.

*b) Datasets:* They extract sentences from the SQUAD dataset and pair them with the queries. They use the sentence-question pairs to train the models. Over 100k questions have been asked about the 536 articles in the dataset, which contains the questions.

*c) Pros:* The method is entirely data-driven (there are no manually produced rules), and it considers information from the reading material at the paragraph rather than the sentence level.

*d) Cons:* The IR performs badly and the paragraph-level model does not reach the optimal performance across all inquiry types.

## III. WORKFLOW

*1) Data preparation*: For image generation, the publicly available Stable Diffusion model was used which works wonders, however we needed it to be fine-tuned to our specific use-case and therefore, we gathered images relevant to the educational domain through crawlers and also from renowned middle school textbooks and captioned them accordingly to the domain they fall into.

*2) Speech to text:* The first step was to convert the given speech as input into text so that it can be used to generate summaries, images, questions and answers. We achieved this through the Assembly AI API, which takes the .mp3 file

as input and converts the speech to give us a .txt file as output which was then used further.

*3) Topic modelling:* Once the text file was ready, we then proceeded onto pre-processing the text file to add a new line after each sentence so that the topic detection happens smoothly. We then used the non-negative matrix factorisation model for topic modelling and keyword generation for each topic on the processed text file. The NMF model is particularly effective for discovering latent topics in text data because it enforces non-negativity and sparsity constraints on the factorization process.

*4) Segregation of text under each topic:* Since, the topics under each chapter will not be mentioned in the speech input, we then needed to segregate the sentences in the text file under each topic that was detected through the nmf model. The keywords that were generated for each topic were used to achieve this. We ensured that each sentence comes under the most related topic so that there is no loss of data.

*5) Data pre-processing for summariser:* We then grouped the sentences under each topic to form a paragraph to pass it onto the summariser. If the number of sentences under a particular topic exceeded a threshold (>30 sentences), it was broken down to form more than one paragraph under that topic to pass it to the summariser and ensure that there is no loss of data.

*6) Text summarisation:* The GPT 3.5 turbo model was then used for text summarization. Additionally, we adjusted a few parameters, including temperature, max_tokens, top_p, frequency penalty, and presence penalty, which produced precise findings relevant to our area of interest, the education sector. The summarized text was then stored in a dictionary with the topic as its key to be put in the slide.

*7) Image generation for each slide:* Summarized slide by slide text was fed into our fine-tuned Stable Diffusion Dreambooth model which spews out one or more images per slide. Using Dreambooth we had fine-tuned some parameters to match the image generation to our specific application which was in this case the educational domain. These images contain valuable information relevant to the topic on that particular slide of the presentation adding more value to the information in text that is already present there.

*8) Question and answer generation:* Lastly, an extensive slide containing questions and answers related to the topic discussed in the presentation is added to help solidify the learnings. We use a Langchain model to generate closed passage questions and answers.

*9) External links generation:* To encourage further learning, we used the beautiful soup package to parse the web to gather some educational external links which will be added to the presentation. This way there will be a holistic approach to learning the chapter.

*10) Ppt formation:* Finally, all the above generated summaries for each topic and the corresponding images for the points on the slides are put in the ppt. Additionally, question and answers and further reading material is put at the end of the ppt.

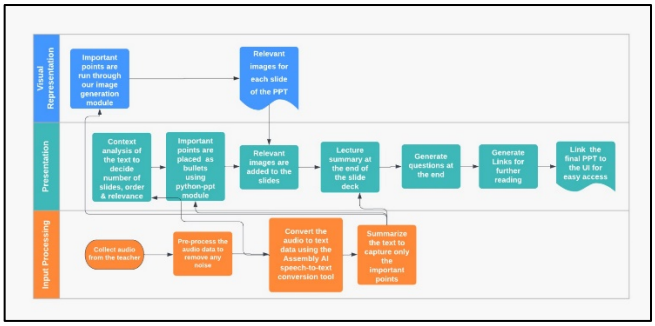


Figure 7: Swimlane diagram representing the workflow of the model

## IV. PROPOSED METHODOLOGY

*A. Text Summarization:*

After evaluating multiple models for text summarization, including BART and T5 Transformers, it was found that GPT-3.5 Turbo emerged as the superior choice. While BART lacked coherence and had issues with repetition and grammar, T5 Transformers provided a more structured and coherent output. However, it focused on extractive summarization, lacking the ability to analyze context and present educational content effectively. In contrast, GPT-3.5 Turbo excelled in abstractive summarization, generating concise, relevant, and tailored summaries specifically suited to the Ed tech domain, meeting the needs of students more effectively.

*B. Image Generation:*

Our primary goal while generating lecture relevant images was simplicity and unambiguity. Since students from each image have to process only one particular piece of information it is important that the image generated by the chosen model needs to be unambiguous and straight to the topic and it also needs to be realistic to an extent and not artistic or cartoon-styled.

Our main three image generation models in consideration were Stable Diffusion, Midjourney and DALL-E2.

After considering and trying out the various models, we decided to use Stable Diffusion for image generation for lectures due to multiple reasons:

- Stable Diffusion, unlike the other two, is free to use and the model's weights have been released publicly. This makes it easier for us to fine-tune the model by training certain layers to tend to data pertaining to the educational context and their respective images.
- When the three models were tested to assess the quality of generated faces from a set of real faces using the Frechet Inception Distance (FID), Stable Diffusion performed significantly better as compared to the other two models in consideration.
- In terms of realism, Stable Diffusion does a better job as compared to the other two. DALL-E2 and Midjourney have a specific style of image generation tending to a more artistic and creative crowd which do not convey the relevant information from an educational standpoint. The lack of publicly available weights also makes it impossible to fine-tune the model to our liking.

## V. MODELS

*A. NMF*

The famous unsupervised machine learning approach known as Non-negative Matrix Factorization (NMF) is utilised for topic modelling. Because it imposes non-

negativity and sparsity requirements on the factorization process, it is very useful for identifying latent topics in text data. Each row in the term-document matrix used by the NMF model corresponds to a term (word), and each column to a document. The values in the matrix show how often or how significant each term is in each document. This matrix is to be factorised into two non-negative matrices, a term-topic matrix and a topic-document matrix, using NMF.

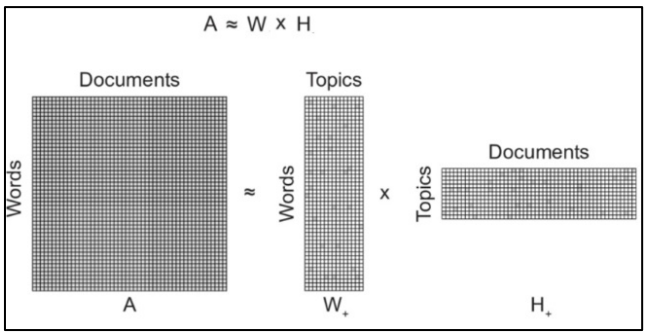


Figure 8: Figure represents the dimension of input and output matrices

*1) Loading and Preprocessing the Data:* We first load the data, which reads the text data from a file, then preprocess it using scikit-learn's TfidfVectorizer to perform operations like tokenization, stop word removal, and TF-IDF (Term Frequency-Inverse Document Frequency) feature extraction. Words are given weights by TF-IDF depending on their rarity across the entire corpus and their frequency inside each individual document.

*2) NMF Model Initialization and Fitting:* The stated number of topics (num_topics) are then initialised in an NMF model. The document-topic and topic-word associations are represented by two lower-rank matrices using the matrix factorization technique known as NMF. The fit method is then used to fit the model to the TF-IDF matrix.

*3) Mathematical approach:* Let's consider a document-term matrix A of size m × n, where m represents the number of documents in the corpus and n represents the number of unique terms in the vocabulary. Each entry A(i, j) represents the frequency or occurrence of term j in document i.

*a*) We aimed to decompose the document-term matrix A into two nonnegative matrices, W (m × r) and H (r × n), where r represents the desired number of topics. The columns of W are the topic vectors, and the rows of H are the term vectors associated with each topic.

*b) Matrix Factorization:* Finding W and H that are close to the document-term matrix A is the aim of NMF. By minimising the objective function based on an appropriate distance metric, such as the Kullback-Leibler divergence or the Euclidean distance, this can be accomplished. Here, the distance metric is the Kullback-Leibler divergence (KL divergence).

*c) Objective function*: minimize D(A || WH), subject to $W \geq 0$ and $H \geq 0$

$$\text{minimize } \Sigma[A(i,j) * \log(A(i,j) / (WH)(i,j)) - A(i,j) + (WH)(i,j)], \text{ for all i and j} \quad \ldots(1)$$

where D(A || WH) represents the Kullback-Leibler divergence between X and the matrix product WH.

*d) Optimization:* The optimization problem can be solved using iterative update rules that alternate between updating W and H. The multiplicative update rule is commonly used in NMF for topic modelling.

For each iteration t until convergence:

- Update H:
  - $H(t+1) = H(t) * ((W^T * (A / (W * H))) / (W^T * 1))$ …(2)
- Update W:
  - $W(t+1) = W(t) * ((A / (W * H(t+1))) * H(t+1)^T / (1 * H(t+1)^T))$ …(3)

  where * denotes element-wise multiplication, / denotes element-wise division, and 1 represents a matrix of ones with appropriate dimensions. These update rules iteratively refine W and H to minimize the KL divergence between A and WH.

Once the NMF model converges, the resulting W matrix represents the document-topic matrix, where each entry W(i, j) represents the strength of the association between document i and topic j. Similarly, the H matrix represents the topic-term matrix, where each entry H(i, j) represents the importance of term j in topic i.

*e) Dimensionality Reduction:* NMF performs dimensionality reduction by approximating the TF-IDF matrix with lower-rank matrices. It captures the underlying latent topics within the data, reducing the dimensionality from the original term space to the topic space. The number of topics is a parameter that is specified.

*B. GPT 3.5 Turbo for text summarization*

This project also presents our implementation of the GPT-3.5 model for text summarization in an automated speech to presentation system tailored for educational purposes. We specifically focus on the fine-tuning of GPT-3.5's parameters to optimize its performance and adapt it to our specific needs. By leveraging the power of GPT-3.5 and customizing it for educational content, we aim to enhance teachers' efficiency in creating presentations by generating accurate and informative summaries. This paper discusses the fine-tuning process, evaluates the results, and highlights the impact of our customization.

*1) Implementation of GPT-3 for Text Summarization:*

To utilize GPT-3.5 for text summarization, we utilize OpenAI's turbo model. For each topic and its corresponding paragraphs, we first assign the system a role of a helpful assistant for text summarisation and then generate summaries by providing which grade is the chapter taken from and the text as a prompt to the turbo model. The generated response represents a summary of the input text. However, the out-of-the-box GPT-3.5 model may not produce summaries tailored to educational content. Therefore, we employ fine-tuning techniques to optimize the model's parameters for our specific needs.

*2) Fine-tuning GPT-3 Parameters:*

During the fine-tuning process, we experimented with various parameters to improve the quality and relevance of the generated summaries. Some of the key parameters we focused on include:

*a) Temperature:* By adjusting the temperature parameter, we control the randomness of the generated text. Higher values (e.g., 0.7) introduce more randomness, while lower values (e.g., 0.2) produce more deterministic responses. To ensure that the model provides succinct summaries on a diverse range of topics, a higher temperature value was set for greater randomness.

*b) Max_tokens:* We limit the length of the generated summary by setting the max_tokens parameter. This ensures that the summaries remain concise and within the desired length.

*c) Top-p (Nucleus Sampling):* By setting the top-p parameter (e.g., 0.9), we control the diversity of the generated text. It helps to avoid repetitive or irrelevant information by sampling from a subset of the most likely tokens.

*d) Frequency Penalty and Presence Penalty:* These parameters allow us to fine-tune the model's behaviour regarding the repetition and relevance of generated text. By adjusting the penalties (e.g., frequency_penalty=0.5, presence_penalty=1.0), we encourage the model to generate more unique and contextually relevant summaries.

*3) Results:*

The fine-tuning of GPT-3.5's parameters has a significant impact on the quality and relevance of the generated summaries. By customizing the parameters for educational content, we observed improvements in terms of accuracy, coherence, and informativeness. The summaries became more concise, capturing the key points of the input text effectively and framing it in a way that's easier to understand. The fine-tuning process allowed us to optimize GPT-3.5 for our specific ed-tech oriented needs, enabling teachers to generate presentation slides with accurate and relevant content more efficiently.

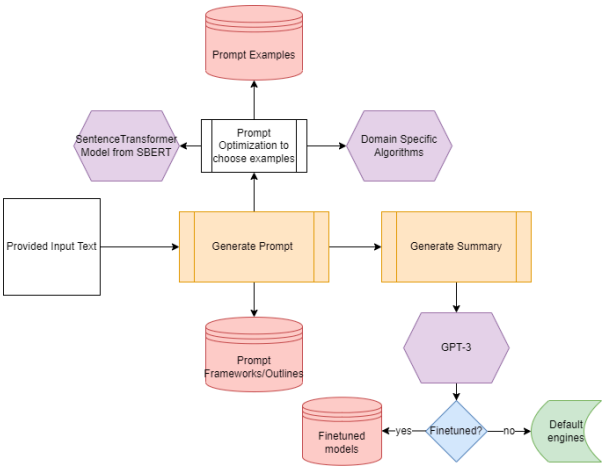


Figure 9: State-of-the-art GPT 3 summariser

*4) Conclusion:*

Through our implementation of the GPT-3.5 model and the fine-tuning of its parameters, we have demonstrated the effectiveness of customized text summarization for educational presentations. The optimization process significantly improves the quality and relevance of the generated summaries, making GPT-3.5 a powerful tool for educators. The fine-tuning techniques we employed have enhanced the model's performance, allowing it to generate concise and informative summaries tailored to our specific domain. Future work can explore further parameter tuning and investigate the integration of GPT-3.5 into other educational applications.

## *C. Stable Diffusion*

As mentioned, for the image generation of the respective topic, per slide of the presentation, we have used the Stable Diffusion model with the Dreambooth fine-tuning implementation. Stable Diffusion utilizes the Latent Diffusion Model which trains by removing successive applications of gaussian noise on training images. It has been trained over 5 billion image-text pairs containing high-resolution, hyper-real images as a part of their LAION-5B dataset. This dataset contains LAION-high-resolution, LAION-2B-en and LAION-aesthetics v2 5+. This makes it appropriate for our use-case to generate simple, realistic and unambiguous images making it easy for students to receive relevant and accurate information from each image. We have also experimented with specific parameters during the fine-tuning process along with specific images gathered in-house to guide the model to generate images relevant to the education sector and the particular lecture. This was done using Google's Dreambooth technique which will be explained in the further sections.

*1) Fine-tuning methods:*

Since Stable Diffusion's training weights are released publicly there are multiple ways to fine-tune the model to match a specific use-case.

End-user fine-tuning can be primarily implemented in three ways on Stable Diffusion:

- An 'embedding' can be trained using images specific to the use case. The model can then generate visually similar images when the embedding name is used in the prompt. This can be used to mimic certain visual styles.
- A relatively smaller neural network can be integrated into parts of the larger neural network to make the model sway in a particular direction as specified by the user. These smaller neural networks are known as 'Hypernetworks' and they work by modifying certain features of the image like face shape, nose and so on.
- DreamBooth can be used to fine-tune the model to generate accurate and personalized outputs that can generate a specific subject, following training via a set of images which depict the subject. Overall, Dreambooth is a powerful tool that can be used to create personalized, creative, and easy-to-use images. It is a valuable tool for artists, designers, and anyone else who wants to create unique and original images.

*2) Our Dreambooth fine-tuning implementation:*

Dreambooth is a technique that can be used to fine-tune a diffusion model like Stable Diffusion to generate images of a specific subject or style. It works by taking a few images of the subject or style as input and then training the diffusion model to generate images that are similar to the input images. It also allows us to tune certain specific parameters prior to generation which we will talk about further.

- The Stable Diffusion weights can be downloaded from the HuggingFace website to be used in your respective code.
- A specific model of Stable Diffusion as per the particular use case needs to be loaded into the code.
  Model used in our case: CompVis/stable-diffusion-v1-4
- A path for images needs to be mentioned for images of the concept for training which is set to any particular path as per your file structure.
- Before beginning the process of fine-tuning, we are required to set a class name specific to the images being trained on for the fine-tuning. In our case, while training the model on a set of topics from the subject Geography, we used classes such as landforms, solar system and so on which guide the model to a particular end-point when used in the text prompt.
- Model weights generally take around 4-5 GB and can be saved locally or with the notebook.

- Specific flags need to be chosen based on the system memory and speed for ideal output of the training.
- The specific parameters used in our case for the best results were as follows (for training):
    - Flag: fp16
    - Seed: 1337
    - Resolution: 512
    - Train Batch Size: 1
    - Gradient Accumulation Steps: 1
    - Learning Rate: 5e-6
    - Number of Class Images: 50
    - Sample Batch Size: 4
    - Max Train Steps: 1000
- A number of specific parameters need to be input while entering the prompt to generate images from this fine-tuned Stable Diffusion Model.
    - Prompt and Negative Prompt: A prompt along with an optional negative prompt can be entered to generate images as per our liking. The prompt can optionally also include the "Class" name as mentioned above to emphasize the fine-tuned weights.
    - Number of Samples: Number of samples required by the model can also be entered as per the user's requirement. In our case, it varied from slide to slide depending on the requirement for images for particular topics.
    - Guidance scale: Guidance scale also needs to be set as a measure of how closely the model must follow the given prompt and in our specific use-case a value of 7.5 seemed to work best in generating accurate images.
    - Number of inference steps: Number of inference steps that worked out ideally for us were 50.
    - Dimensions: Height and width of the output image can also be chosen but we stuck to the regular 512x512 as Stable Diffusion is trained on images of this size and hence provides best results when output is of the same size too.

*3) Training of Stable Diffusion:*

On LAION-5B, which had more than 5 billion image-text pairs, Stable Diffusion was trained. The three LAION-5B subsets that it was trained on were laion-high-resolution, laion2B-en, and laion-aesthetics v2 5+. Laion-aesthetics v2 5+ was a subset of 600 million photographs that, according to a predictor, had scores of over 50%. It also, to the best of its ability, excluded images with watermarks and low resolution. The model was first trained on Laion2B-en and Laion-high-resolution before being moved to Laion-aesthetics v2 5+. During the final rounds, the amount of text conditioning was decreased by about 10% in order to improve Classifier-Free Diffusion Guidance.

On 256 NVidia GPUs, this was trained over the course of 150,000 GPU hours.

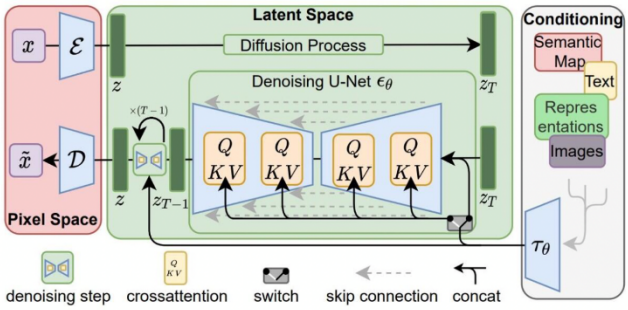


Figure 10: Stable diffusion architecture

*4) Results:*

The summarized text is passed as points to the stable diffusion model to generate images. The output is the slides with the points and the generated images for each slide.

## *D. Langchain for Question answer pairs*

Question answering pairs are a type of active learning strategy that involves students asking and answering questions on a topic. This strategy can be used to improve students' understanding of a topic by encouraging them to think critically about the material and to generate their own questions. In this research paper, we will implement a question answering pairs model using Langchain. Langchain is a toolkit for natural language processing that makes it easy to build and deploy question answering models.

Our model will take in a text document as input and generate question answer pairs from it. The model will use a large language model (LLM) to generate the questions. The LLM will be trained on a massive dataset of text and code, which will give it the knowledge it needs to generate comprehensive and informative questions.

The QAGenerationChain class has a number of parameters that can be used to control the generation of question-answer pairs, such as the length of the questions, the number of questions to generate, and the type of answers to generate.

The question will be a string, and the answer will be a dictionary with the following keys:

- text: The text of the answer.
- span: The start and end indices of the answer in the text.

*1) Internal Working of the Concept*

The internal working of the question answering pairs concept is as follows:

a) The LLM first reads the text document and identifies the important concepts in the document.
b) The LLM then uses these concepts to generate a set of questions.
c) The LLM then ranks the questions in order of difficulty.
d) The LLM then presents the questions to the student.
e) The student answers the questions.
f) The LLM then uses the student's answers to improve its understanding of the text document.
g) The LLM then repeats steps 2-6 until the student has mastered the material in the text document.

This concept is effective because it allows students to learn by doing. By generating their own questions and answering them, students are forced to think critically about the material and to apply what they have learned. This process helps students to retain the information that they have learned and to develop a deeper understanding of the topic.

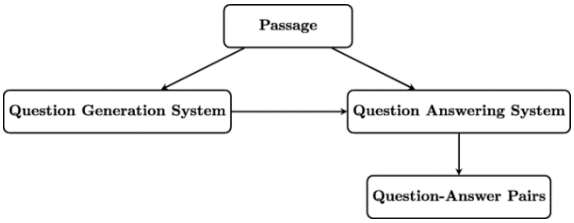


Fig 11. This figure represents the working of the QA pairs

## VI. RESULTS

We have compared various summarisation models while keeping in mind various factors such as the domain of interest and informativeness of the summary as shown in Fig. 12.

| *Models* | *Evaluation metrics* | | | |
|---|---|---|---|---|
| | *Rouge - 1* | *Rouge - 2* | *Rouge - L* | *chrF* |
| Bart - large | 0.30 | 0.11 | 0.27 | tensor(0.3149) |
| T5 - base | 0.25 | 0.10 | 0.23 | tensor(0.2249) |
| Bert | 0.21 | 0.09 | 0.19 | tensor(0.5732) |
| GPT 3 - Davinci | 0.39 | 0.16 | 0.35 | tensor(0.5297) |
| GPT 3.5 - Turbo | **0.59** | **0.32** | **0.54** | tensor(0.8020) |

Figure 12: Evaluation and comparison of summarizers

Similarly, various image generation models were also compared to generate the most appropriate image with respect to the education domain as shown in Figure 14.

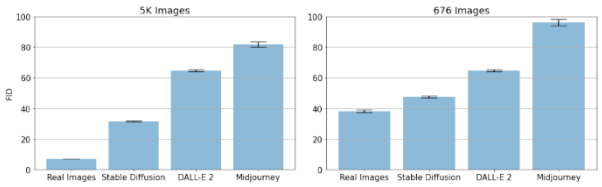


Figure 13: Frechet Inception Distance (FID) scores of image generation models

Overall, the results of various models are shown with respect to the input below:

Input**:**

Because of harsh climate, less people live in the mountain areas. A plateau may have one or more sides with steep slopes. The Tibet Plateau, figure five one p. 31 is the highest plateau in the world with a height of 4000 to 6000 meters above the mean sea level. As a result, many of the mining areas in the world are located in the plateau areas. The African plateau is famous for gold and diamond mining in India, huge reserves of iron, coal and manganese are found in the chhota Nagpur plateau. In the plateau areas there may be several waterfalls as the river falls from a great height in India, the Hundred Falls in the chhota Nagpur Plateau on the River Subarnarekha and the Jog Falls in Karnataka are examples of such waterfalls. Some plains are extremely level, others may be slightly rolling and undulating. Most of the plains are formed by rivers and their tributaries. Some of the largest plains made by the rivers are found in Asia and North America. For example, in Asia, these plains are formed by the Ganga and the Brahmaputra in India and the Yangtze in China. In India too, the Indo Gangetic plains are the most densely populated regions of the country. Life is difficult in mountainous areas.

Outputs:

A. NMF for topic detection

Topic: plateau

Keywords: ['plateau', 'plains', 'india', ‘areas’]

B. GPT 3.5 Turbo for Text Summarisation

Summary: Mountain areas have harsh climates, so not many people live there. A plateau is a flat area with steep slopes on one or more sides. The Tibet Plateau is the highest plateau in the world and it is located at a height of 4000 to 6000 meters above sea level. Many mining areas are found on plateaus because they have valuable resources like gold, diamonds, iron, coal, and manganese. Waterfalls can also be found in plateau areas where rivers fall from great heights. Some plains are very flat while others have gentle hills. Most plains are formed by rivers and their smaller streams. Some of the largest plains are found in Asia and North America, created by rivers like the Ganga and Brahmaputra in India and the Yangtze in China. These plains, especially the Indo Gangetic plains in India, are highly populated regions. Life in mountainous areas can be tough because of the challenging conditions there.

C. Stable diffusion for image generation and final .pptx formation

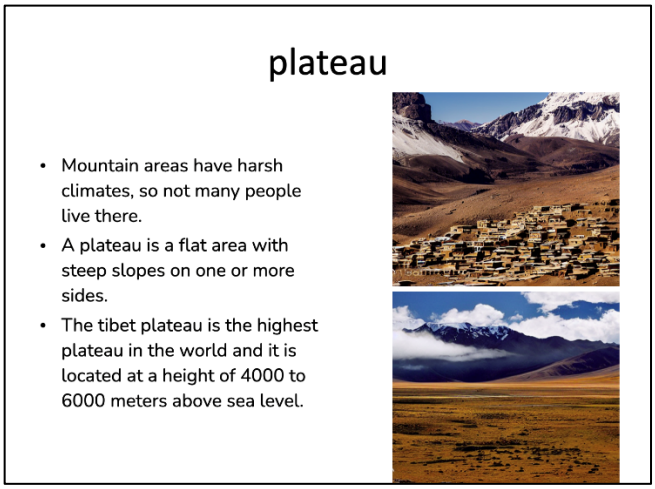


Figure 14: Result slide 1

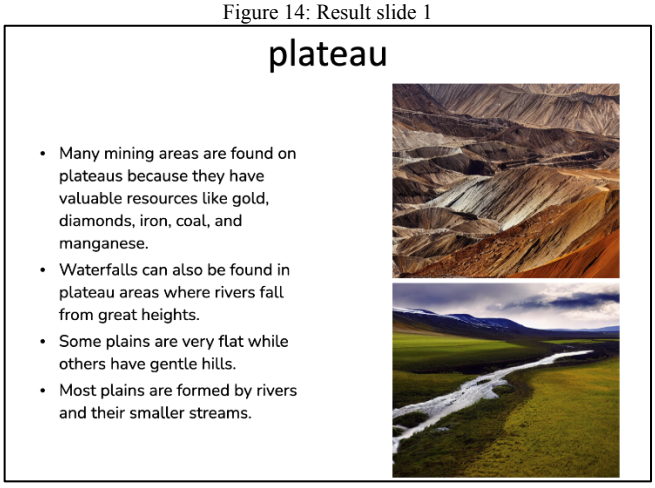


Figure 15: Result slide 2

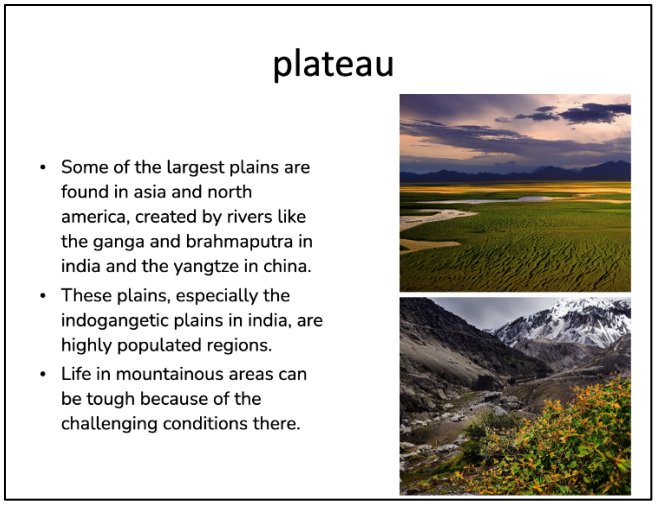


Figure 16: Result slide 3

D. Langchain for Question answer pairs

[{'question': 'What is the highest plateau in the world?', 'answer': 'The Tibet Plateau'}]

## VII. CONCLUSION

Based on the output generated by our models in tandem with some fine-tuning, it was clearly visible that developments in the digital space for education is a growing sector and will boom further. This end-to-end generative solution tackles that primary issue making it easier to convey quality information/knowledge from the educator (be it of any kind) to a student trying to educate themselves in a particular topic. Through our research in the education field, we have found that there are primarily four kinds of learners, Visual Learners, Auditory (or aural) Learners, Kinesthetic (or hands-on) Learners and Reading (or writing) Learners. Through this solution we attempt to successfully cater to all of them. This gives them not only a personalized learning experience but also a holistic one where they can comprehend complex topics through multiple facets of delivery. Summarizing text through cutting edge models, makes sure to keep all the essential points intact while reducing the readers' time manifold. Images generated by training on millions of data points, give an accurate description of what the educator wants to convey with minimal effort from them and also deliver an easily comprehensible picture to the viewer/student. Towards the end of each lecture/slide deck, some questions relevant and important to that particular lecture are generated giving an opportunity for the students to assess themselves on what they have learnt. These answers can be fed into the system and our AI solution will check for similarity giving the student the results of their assessment making it easier for them to fill in their knowledge holes.

This solution is entirely black-boxed to the user. The educator only needs to record the audio file of their class and feed it into our solution. It generates everything aforementioned using AI and builds out an end-to-end lecture material for that particular class in an easily shareable PowerPoint Presentation format. This solution largely benefits both sides of the party, saving the educators plenty of time and effort which they could devote to personally attending to the students instead of curating these lecture modules with all these features manually. On the other hand, students can pick up the entire lecture content in a concise but highly informative format, saving them time as well as giving them a complete understanding of the subject.

Overall, our future plans revolve around reducing the burden on a teacher in any facility as well as providing students quality digital material in any part of the world which they previously might not have had the access to due to various circumstances.

We would like to thank PES University for their help and support.

## VIII. FUTURE WORK

In the future, our research aims to expand the capabilities of our AI solution to enhance the learning experience even further. One of our key objectives is to incorporate 3D images and interactive visuals into the generated lecture material. This will enable students to explore subjects in a more immersive and engaging manner, providing a deeper understanding of complex concepts. By leveraging cutting-edge technologies, such as virtual reality and augmented reality, we can create a dynamic learning environment that fosters curiosity and active participation.

Additionally, we plan to introduce answer verification functionality to our AI solution. This feature will enable students to receive immediate feedback on their assessments, allowing them to identify and address any knowledge gaps more effectively. By incorporating advanced natural language processing techniques, we can compare student answers with model answers and provide insightful feedback, further facilitating the learning process. Furthermore, we recognise the importance of tailoring educational content to the specific needs of different student audiences. Therefore, in addition to the summaries of the teachers content, if the teacher chooses to, we intend to develop algorithms that can adapt the generated material based on factors such as the age group, academic level, and desired difficulty of the students to generate extra facts and information about the topic . This customisation will enable educators to provide personalised as well as a complete learning experience that caters to the unique requirements and preferences of their students.

## IX. REFERENCES

1. Semi-supervised NMF Models for Topic Modelling in Learning Tasks Jamie Haddock, Lara Kassab, Sixian Li, Alona Kryshchenko, Rachel Grotheer, Elena Sizikova, Chuntian Wang, Thomas Merkh, R. W. M. A. Madushani, Miju Ahn, Deanna Needel, Kathryn Leonard
2. VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning Jun Chen , Han Guo , Kai Yi , Boyang Li , Mohamed Elhoseiny King Abdullah University of Science and Technology (KAUST) Carnegie Mellon University Nanyang Technological University
3. Automated News Summarisation Using Transformers Anushka Gupta , Diksha Chugh , Anjum , Rahul Katarya
4. High-Resolution Image Synthesis with Latent Diffusion Models Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser Bjorn Ommer
5. Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures Gal Metzer, Elad Richardson, Patashnik, Raja Giryes, Daniel Cohen
6. Learning to Ask: Neural Question Generation for Reading Comprehension Xinya Du, Junru Shao, Claire Cardie
7. Transformer-based End-to-End Question Generation, Luis Enrico Lopez, Diane Kathryn Cruz, Jan Christian Blaise Cruz, Charibeth Cheng

8. Algorithms for Non-negative Matrix Factorization Daniel D. Lee, H. Sebastian Seung
9. Attention is all you need. In Advances in neural information processing systems Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017).
10. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer Colin Raffel, Moam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu.
11. Identifying Effective Classroom Practices Using Student Achievement Data Kane, T.J., Taylor, E.S., Tyler, J.H., & Wooten, A.L.
12. Application in Info-Tech Education Based on Visual Learning Wang Yan-rong; Gu Yue-sheng; Xu Wu
13. A Digital Reality Learning Environment with Instant Assessment on Learning with Body and Visual Interaction Chiu-Chen Yen; Chia-Ying Lee; Gwo-Dong Chen; Jen-Hang Wang; Su-Hang Yang
14. Research on Visual Performance Evaluation Model of E-commerce Websites Fan ZHANG; Cuiqin LAN; Tao Wang; Feng GAO; Enmao Liu
15. Visual and Verbal Communications: Similarities and Differences Khatuna Kacharava, Nino Kemertelidze
16. https://www.gartner.com/en/documents/3694517
17. Transformers for Natural Language Processing: Build, train, and fine-tune deep neural network architectures for NLP with Python, Hugging Face, and OpenAI's GPT-3, ChatGPT, and GPT-4 Denis Rothman; Antonio Gulli
18. Diffusion Models Beat GANs on Image Synthesis Prafulla Dhariwal Alex Nichol