Section 1 : Topic Submission Form

This form should be submitted by the mentioned deadline.

Name: Sunayna Talreja

Student Number: **1140492**

Course: MS AI ML

Fill your topic/s below

Project Title/Area 1: Psychology AI Assistant with accessibility for people with Hearing disability using Large Language Model

Dataset: Indian Language Signed Animated Videos (Contains alphabets and numbers of ISL)

ISL Dataset (Contains ISL Library of all possible words)

nlp-mental-health-conversations (Mental health consultation conversations)

isear-dataset(Emotion relation to the sentence)

Model : GPT-4o or Gemini 1.5 Pro 002

Description: Therapy chatbots currently rely on text and voice interactions. However, effective communication of emotions requires more. To bridge this gap, we aim to develop a platform that accommodates individuals with hearing disabilities, enabling them to express themselves in their preferred language and receive responses that foster a sense of freedom and empowerment.

Project Title/Area 2: Video transcript Summarization in Indian Languages using Large Language Model

---

Dataset: EnglishToIndianLanguageTranslation (English to Indic Languages translations)

Opus Dataset (Language translator dataset)

YoutubeVideoToCaptions (This dataset consist of youtube video ids and the corresponding text summary caption)

Model : GPT-4o or Gemini 1.5 Pro 002

Description: Currently, video summarization tools fall short in providing transcripts in Indian languages. By bridging this gap, we can offer a seamless reading experience in all major Indian languages, revolutionizing the way people consume video content. This innovation will bring convenience and accessibility to a vast audience, breaking language barriers and enhancing overall understanding.

Project Title/Area 3: Question answer generation and correction on Indian Penal Code Large Language Model

---

Dataset: Indian Penal Code (IPC Data document)

IPC Sections Information

Model : GPT-4o or Gemini 1.5 Pro 002

Description: Introducing a pioneering application that generates and corrects Q&A pairs for the Indian Penal Code document. This innovative tool empowers users to effortlessly access and comprehend the complexities of the IPC, facilitating a deeper understanding of the law and its applications. With its accuracy and ease of use, this application is poised to revolutionize legal education and research.

Fill in this section if a member of staff has agreed to be your supervisor:

Member of Staff:_____

If you have found a supervisor then you and the member of staff who agreed to supervise your project should sign below.

Sunayna Talreja	_____
Student Signature	Supervisor Signature
04 Oct 2024	_____
Date	Date

Section 2 : Topic Selection Research

Table 1 : Topic 1

Title	Link to the Paper	Understanding of the Dataset	Understanding the Methodology Used	Dataset Link
Therapy Chatbot Powered by Artificial Intelligence: A Cognitive Behavioral Approach	**Therapy Chatbot Powered by Artificial Intelligence: A Cognitive Behavioral Approach**	Dataset defines the emotional relation of the input sentences	Dataset which characterises the user input statement is sent to Dialogue engine to return the response.	International Survey on Emotion and Antecedents and Reactions (ISEAR) dataset
Artificial Intelligence Enabled Mobile ChatbotPsychologist using AIML and Cognitive BehavioralTherapy	Artificial intelligence will change the future of psychotherapy: A proposal for responsible, psychologist-led development	datasets may be obtained via Spacy orMITIE, FastText, or any similar service. Because there is notraining data readily available, the user of this method will berequired to generate the data from scratch in order to use it	In order to create the virtual assistant, the RASA platform was used. RASA Open Source is a machine learningframework that may be used to automate conversationalassistants that are based on data from either text or speech.	NA

3D Avatar Approach for Continuous Sign Movement Using Speech/Text	3D Avatar Approach for Continuous Sign Movement Using Speech/ Text	Custom dataset for 50 daily used ISL words	The English text is converted to its corresponding ISL text. The words in the ISL sentence have been identified to generate corresponding sign movements. For the conversion from English to ISL, we use the Natural Language Toolkit (NLTK)	NA
Real Time Sign Language Translation Systems: A review study	Real Time Sign Language Translation Systems: A review study	Custom Data, collected by acquiring from video camera	It is a game-learning application for students in learning applications, and the k-Nearest-Neighbour method was implemented as a classification method.	NA
LLM Based 3D Avatar Assistant	LLM Based 3D Avatar Assistant	custom dataset that contains texts and their corresponding emotions	Speech recognition Emotion analysis Intent classification Encryption 3d waveform representation	NA

Table 2 : Topic 2

Title	**Link to the Paper**	**Understanding of the Dataset**	**Understanding the Methodology Used**	**Dataset Link**

Video-to-Text Summarization using Natural Language Processing	Video-to-Text Summarization using Natural Language Processing	The dataset contains all types of videos (including news debates, sports commentary, funny videos, TV shows etc.) The videos were downloaded from open resources like YouTube.	The text thus obtained from the video is put through a summarizer (pre-trained language models from Hugging Face library) and a summary is obtained. Basically, there are 2 types of summarizers: - Extractive(traditional) and Abstractive(advanced).	No present
Video summarization using deep learning techniques: a detailed analysis and investigation	Video summarization using deep learning techniques: a detailed analysis and investigation	TVSum, SumMe, OVP,Footnote1 Youtube Youtube High-light Videos with extracted details of thumnails, matlab files and transcripts are used	Feature based vs technique clustering based vs technique shot selection based vs technique event based vs technique Trajectory based vs technique	SumMe and TVSum Datasets Cosum Dataset (others not found)

Statistical machine translation for Indic languages	Statistical machine translation for Indic languages	Samanantar and OPUSDataset for fifteen IL-EN and EN-IL pairs (both directions), including the Indo-Aryan andDravidian groups.	Statistical MT (SMT) is one of the popular methods proposed to solve these problems. It isa way of translation wherein a statistical-based learning algorithm is applied to a large bilingualcorpus that helps the machine learn the translation. This method also enables the machine totranslate sentences not encountered by the machine during its training and testing.	samanantar opus
A Voyage on Neural Machine Translation for Indic Languages	A Voyage on Neural Machine Translation for Indic Languages	Pretrained models are finetune by IITB-En-Hi, ILCI, CVIT Mann Ki Baat , PMIndia-Corpus and Samanantar introduced by IndicNLP	Multilingual MT aims to create a universal model that can translate between two languages. Incorporating low-resource language pairs into multilingual models might be a way to access extra data from other, presumably related languages.	samanantar IITB EN-Hi PM India Corpus Mann ki baat ILCI CVIT

Table 3 : Topic 3

Title	Link to the Paper	Understanding of the Dataset	Understanding the Methodology Used	Dataset Link

GLQA: A Generation-based Method for Legal Question Answering	GLQA: A Generation-based Method for Legal Question Answering	Vietnamese legal documents dataset	Nearest Neighbor Search (NNS) to find the top-k applicable law articles. Then the top-k law articles are concatenated with the question, and input to the generator to produce the final answer.	Not found
Automatic question-answer pair generation using Deep Learning	**Automatic question-answer pair generation using Deep Learning**	SQuADRUn, which combines traditional SQuAD with 50,000 unanswerable questions.	extract answer from the sentence generate question based on the extracted answer generate answer for each question formed in *step2* compare the answers from step 1 and step 3	SQuADRUn_Dataset
Opinerium: Subjective Question Generation Using Large Language Models	Opinerium: Subjective Question Generation Using Large Language Models	custom dataset comprising 40,000 news articles along with human-generated questions	Automatic question generation systems aim to generate a question consisting of a sequence of words based on a given context represent words from a vocabulary set *V*	Not found

Subjective Answers Evaluation Using Machine Learning and Natural Language Processing	Subjective Answers Evaluation Using Machine Learning and Natural Language Processing	Dataset consisted of 35classes and 850 instances	Cosine Similarity functions which take two sentences or word vectors and return their Similarity.	Not found