Ghulam Ishaq Khan Institute of Engineering Sciences & Technology

NatLearn: Multilingual translator and AI study buddy

by

Muhammad Azam - 2020284

Jawad Ali - 2020189

Rajab Ali - 2020403

Supervisor: Dr. Khurram Jadoon

Co Supervisor: Dr. Muhammad Hanif

Faculty of Computer Science and Engineering

Certificate of Approval

It is certified that the work presented in this report was performed by **Muhammad Azam, Jawad Ali, and Rajab Ali** under the supervision of **Dr. Khurram Jadoon**. The work is adequate and lies within the scope of the BS degree in Artificial Intelligence at Ghulam Ishaq Khan Institute of Engineering Sciences and Technology.

Dr. Khurram Jadoon
(Advisor)

Dr. Muhammad Hanif
(Co-Advisor)

Dr. Qadeer Ul Hasan
(Dean)

Abstract

NatLearn is a specialized software solution designed to address language barriers in educational settings. This project focuses on developing an application that converts lectures from unfamiliar languages into a student's preferred language, facilitating comprehension and accessibility of educational content. Central to NatLearn is an intelligent chatbot trained on lecture materials, offering personalized assistance by answering questions, providing clarifications, and delivering additional explanations on-demand. Emphasizing adaptation to diverse learning speeds and styles, NatLearn aims to augment rather than replace traditional classroom learning, catering particularly to international students, language learners, and those grappling with technical terminology or dialects. By democratizing access to knowledge, NatLearn strives to ensure no student is disadvantaged due to linguistic challenges, promoting independent learning and comprehension of complex concepts.

Acknowledgements

We express our heartfelt gratitude to all those who contributed to the successful completion of this final year project. First and foremost, we extend our deepest appreciation to our supervisor Dr.Khurram Khan Jadoon and our co-supervisor Dr. Muhammad Hanif, whose guidance, support, and insightful feedback were invaluable throughout the entire duration of this project. Your expertise and encouragement have been instrumental in shaping our work. We would also like to thank the faculty and staff of GIK Institute for providing us with the necessary resources, facilities, and academic environment conducive to our research and development efforts. Thank you all for being part of this endeavor and for your invaluable contributions.

Muhammad Azam

Jawad Ali

Rajab Ali

Table of Contents

List of Figures

Chapter 1: Introduction

1.1 Problem Statement

In the era of globalization and digital interconnectedness, accessing educational content in multiple languages has become increasingly essential for learners worldwide. However, the current landscape lacks a comprehensive solution that seamlessly integrates both video translation and contextualized query resolution. This gap poses significant challenges for individuals seeking to learn from educational resources not available in their native language and inhibits their ability to interact effectively with the content.

Traditional methods of language translation often rely on static subtitles or manual transcription, which can be time consuming, costly, and may not capture the nuances of the spoken word accurately. Moreover, existing translation tools lack integration with supplementary features such as chatbots, which could enhance the learning experience by providing on demand clarification and additional context.

Furthermore, while chatbots have become prevalent in various domains, their integration within educational platforms for contextualized query resolution remains underexplored. Users often encounter limitations in receiving precise and relevant responses tailored to the content they are engaging with, especially in multilingual environments.

Hence, the overarching problem addressed by this project is the absence of a comprehensive web application that allows users to upload lectures in any language and seamlessly translate them into their desired language, while also providing integrated chatbot functionality for contextualized query resolution. This project seeks to bridge this gap by developing a robust solution that empowers learners to access educational content in their preferred language and receive real-time assistance through a responsive chatbot, thereby enhancing the overall learning experience.

1.2 Motivation

The motivation behind the development of this web application stems from the recognition of the significant challenges faced by individuals seeking to access educational content in languages other than their own. In an increasingly interconnected world, where knowledge knows no boundaries, language barriers should not impede the pursuit of learning opportunities. Therefore, the primary motivation for this project is to democratize access to educational resources by providing a platform that facilitates seamless translation of lectures into multiple languages.

Furthermore, the integration of a chatbot within the web application aims to address the common frustration experienced by learners when seeking clarification or additional information while engaging with educational content. By leveraging natural language processing (NLP) techniques, the chatbot will offer immediate assistance and contextualized responses, thereby enhancing the learning experience and promoting greater comprehension.

Moreover, the project is motivated by the potential impact it can have on fostering cross-cultural exchange and collaboration. By breaking down language barriers, the web application encourages knowledge sharing and collaboration among individuals from diverse linguistic backgrounds, ultimately contributing to a more inclusive and interconnected global community.

Additionally, the project serves as an opportunity to explore and apply cutting-edge technologies, such as machine learning and artificial intelligence, in the field of education. By harnessing the power of these technologies, the web application aims to deliver innovative solutions to address longstanding challenges in language translation and educational accessibility.

Overall, the motivation behind this project is rooted in the belief that every individual should have equal access to educational resources, regardless of linguistic or cultural differences. By developing a comprehensive web application that integrates language

translation and chatbot functionality, we aspire to empower learners worldwide to overcome language barriers and unlock the full potential of educational content.

1.3 Objectives

This project aims to develop a web application that facilitates a comprehensive learning experience by addressing language barriers and providing in-video content support. The specific objectives are:

- Lecture Translation: Develop a system that allows users to upload lectures in any language and translate them into their preferred language. This will break down language barriers and allow users to access and understand educational content regardless of the source language.
- Context-Aware Chatbot Support: Implement a chatbot that can answer user queries related to the content of the translated lecture video. The chatbot should leverage Natural Language Processing (NLP) to understand the context of the video and provide relevant answers based on the translated content.
- Integrated Web Application Platform: Build a web application that incorporates both lecture translation and chatbot functionalities. This application should include user account management features.

1.4 Scope

The primary software product to be developed is referred to as the "NatLearn". NatLearn will convert lectures from unfamiliar languages into a language of the student's preference, allowing students to understand the content. The application will feature an intelligent chatbot, which will answer questions from the lecture content, offer clarifications, and provide additional explanations on-demand. The chatbot is designed to adapt to different learning speeds and styles, ensuring each student can grasp the material at their individual pace. It will not work as a general-purpose

translator for conversations outside the context of educational lectures. It will not replace traditional classroom learning but rather augment the learning experience for students facing language barriers. NatLearn is envisioned as a solution in the education sector, particularly catering to international students, those learning in foreign languages, or individuals grappling with dialects or technical terms unfamiliar to them. It ensures educational content is accessible to students irrespective of the language of instruction, thus democratizing knowledge. The chatbot acts as a personalized learning companion, promoting independent learning, and helping students understand complex ideas without requiring external help. By catering to different learning speeds, it ensures no student is left behind due to linguistic challenges.

Chapter 2: Literature Survey

The increasing globalization of education means that students, more than ever, are seeking opportunities across borders. However, language remains a significant barrier to this cross-border exchange of knowledge. Over the years, several studies and analyses have been conducted to address this issue.

Language Barriers in Education: Numerous studies have shown that language is a significant obstacle to effective learning. Kachru (1985) posited that English, often used as a medium of instruction in many countries, isn't the native tongue for the vast majority of the global population. This inherently places non-native speakers at a disadvantage, making it harder to grasp complex ideas presented in a foreign language.

Machine Translation in Education: Machine translation has come a long way since its inception. Koehn (2010) offers an insight into the evolution of these systems, emphasizing the shift from rule-based to neural machine translation. The rise of Neural Machine Translation (NMT) has greatly improved the

accuracy and fluency of translations, making it a viable tool for translating educational content.

The literature review delves into two primary domains: intelligent chatbots and language translation. While significant research has been conducted in each area independently, there exists a noticeable gap in the exploration of their integration. For instance, M. Aleedy and colleagues developed a chatbot proficient in English to Arabic translation using deep learning methodologies, trained on bilingual datasets. Similarly, Baharuddin et al. introduced an educational chatbot incorporated within multimedia learning materials, crafted through the Multimedia Development Lifecycle approach, exhibiting favorable performance as measured by the BLEU metric. Additionally, F. Qaseem, M. Ghaleb, and HS Mahdi utilized Dialogflow to construct an interactive English learning chatbot. OpenAI's ChatGPT also merits mention, although its precision in data dissemination does not specifically cater to lecture contexts. Concurrently, considerable strides have been made in speech translation endeavors; Xuan-Quy and collaborators devised a system automating the conversion of speech into educational video content. Meanwhile, Hirofumi Inaguma et al. employed a specific architecture for Multilingual End-to-End Speech Translation, facilitating seamless cross-linguistic communication. Google Translate, a landmark achievement in multilingual translation developed in 2006, while proficient, lacks nuanced domain-specific knowledge. A synthesized overview of these studies is provided in Table 1.

Table 1: Literature review table

Title	Publication Year	Link	Chatbot	Speech translation	Focuses on	Drawback
Towards Deep Learning-powered … [1]	2022	https://shorturl.at/huwL0	✓		To develop a chatbot to help language learners	Specific to Arabic language only
The Utilization of Artificial Intelligenc … [2]	2023	https://shorturl.at/nEM38	✓		To develop a chatbot based learning media	Limitation of Chatbot's response accuracy
Dialog chatbot as an interactive online … [3]	2023	https://shorturl.at/fpCDF	✓		Using a chatbot, Dialog Flow, to enhance the learning of (ESP)	Specific to English language only
ChatGPT [4]	2022	https://chat.openai.com/	✓		Generating human-like text responses to a wide range of prompts and inquiries.	The generated output of ChatGPT is not a lecture specific
AI-powered moocs: Video lecture generation [5]	2021	https://shorturl.at/yKV67		✓	To automatically create a video lecture with the instructor's voice and face without recording the video.	

Direct Speech to Speech Translation Using Machine Learning [6]	2019	https://shorturl.at/corT7		✓	To translation language from source to target using sequence-to-sequence model	Poor accuracy in educational terminologies
Google translator [7]	2006	https://shorturl.at/tORX3		✓	To translate text, documents and websites from one language into another.	Lacks domain knowledge
NatLearn	2024		✓	✓	Speech translation and intelligent Chatbot technology	

Although significant progress has been made in the realms of multilingual translation and intelligent chatbots individually, there remains a noticeable absence of applications that effectively integrate both modules for educational purposes. Currently, there is a lack of solutions that seamlessly translate lecture videos while enabling students to pose questions related to the content afterward, rather than in real-time. NatLearn aims to address this gap by serving as a multilingual language translation tool and AI study companion. It offers the capability to translate content into up to 20 languages, facilitating comprehension for non-native speakers. Additionally, NatLearn allows users to engage with the material by posing questions related to the translated video content post-viewing. This innovative approach aims to enhance accessibility and comprehension in educational settings, filling a critical void in the existing landscape of educational technology.

Chapter 3: Design (System Requirements/Specifications)

3.1 Design Approach and Problem Description

For the development of Natlearn, a meticulous design approach was employed, leveraging state-of-the-art design models to tackle the intricate challenges inherent in educational software. Natlearn's primary objective is to transcend language barriers and enrich personalized learning experiences by seamlessly integrating multi-lingual translation and an intelligent chatbot. This design paradigm is rooted in addressing the pressing need for educational resources that cater to diverse linguistic backgrounds and offer tailored assistance to learners.

3.2 Justification and Evidence

The adoption of the Agile methodology was a deliberate choice, driven by its efficacy in decomposing large tasks into manageable increments, aligning seamlessly with the project's overarching objectives. Each developmental cycle, or sprint, meticulously followed the Agile framework[1], ensuring

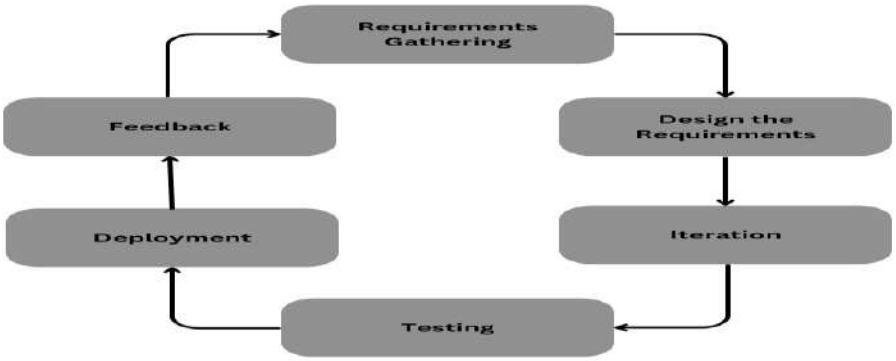

Figure 1: Agile development method

[1] Figure 1: Agile development process

focused attention on specific sub-tasks to facilitate incremental progress and foster continuous feedback loops. The early stages of the project were dedicated to gathering both functional and non-functional requirements, laying a robust foundation for subsequent development phases.

- 3.2.1 Functional Requirements

  Natlearn's functionality is defined by a set of comprehensive functional requirements (FRs) that span user registration, authentication, lecture translation, multimedia support, language selection, output presentation, intelligent chatbot capabilities, and user feedback mechanisms. Each requirement is meticulously crafted to address specific user needs and enhance the overall user experience.

- 3.2.2 Non-Functional Requirements

  Natlearn also adheres to a stringent set of non-functional requirements (NFRs) to ensure optimal performance, usability, security, and compatibility. These include provisions for performance, usability, security, and compatibility with the latest versions of the Chrome browser.

In the design stage, Unified Modeling Language (UML) diagrams and mockups were instrumental in architecting the software for both modules. These design choices were meticulously justified, providing a visual blueprint of the system's architecture and functionality. The incorporation of UML diagrams and mockups served as invaluable tools for conceptualizing and

refining the design of Natlearn. [See Figure [2] , Figure [3], Figure [4], Figure[5] for UML diagrams.]

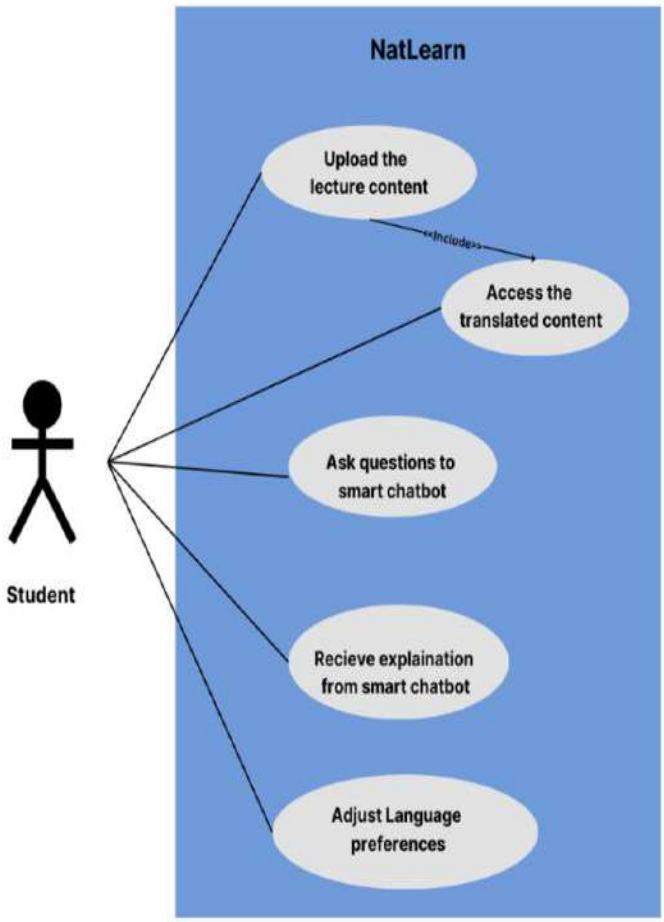

Figure 2: Use Case diagram

[2] Figure 2: Use Case diagram
[3] Figure 3: Sequence diagram
[4] Figure 4: Activity diagram
[5] Figure 5: Class diagram

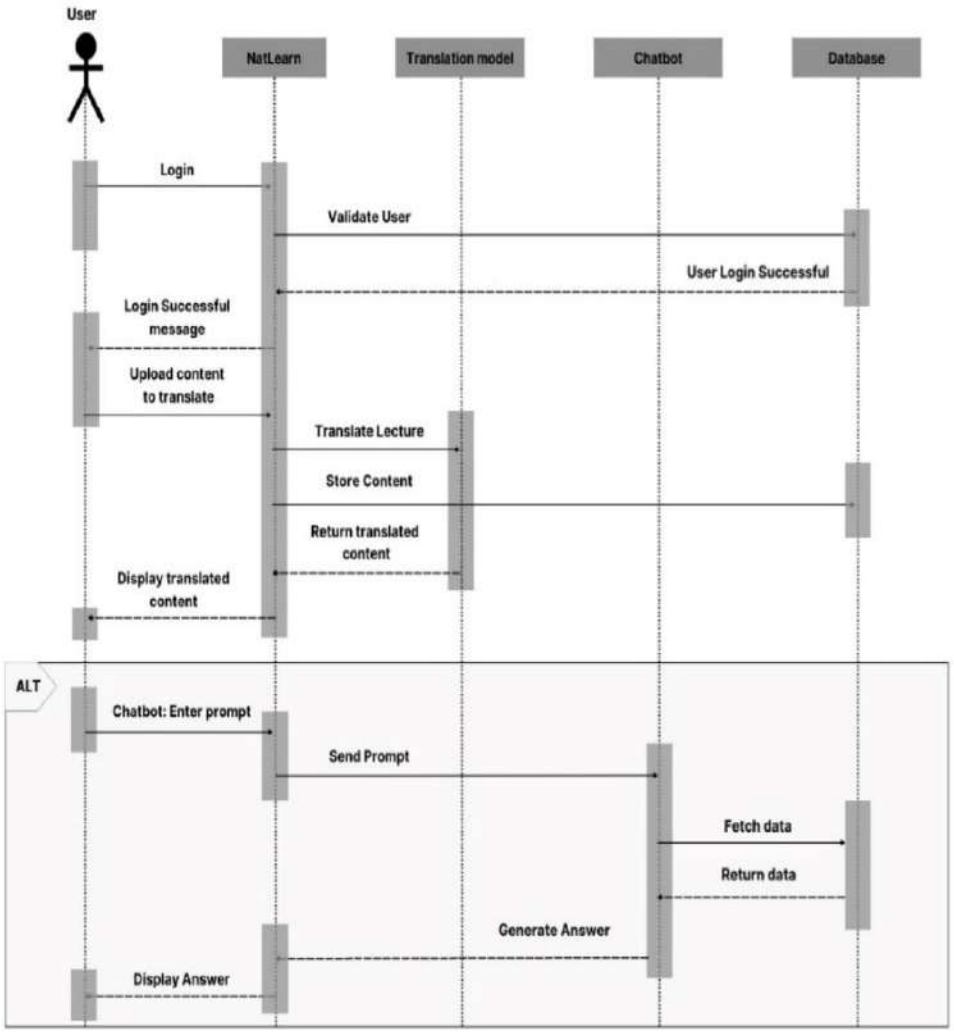

Figure 3: Sequence diagram

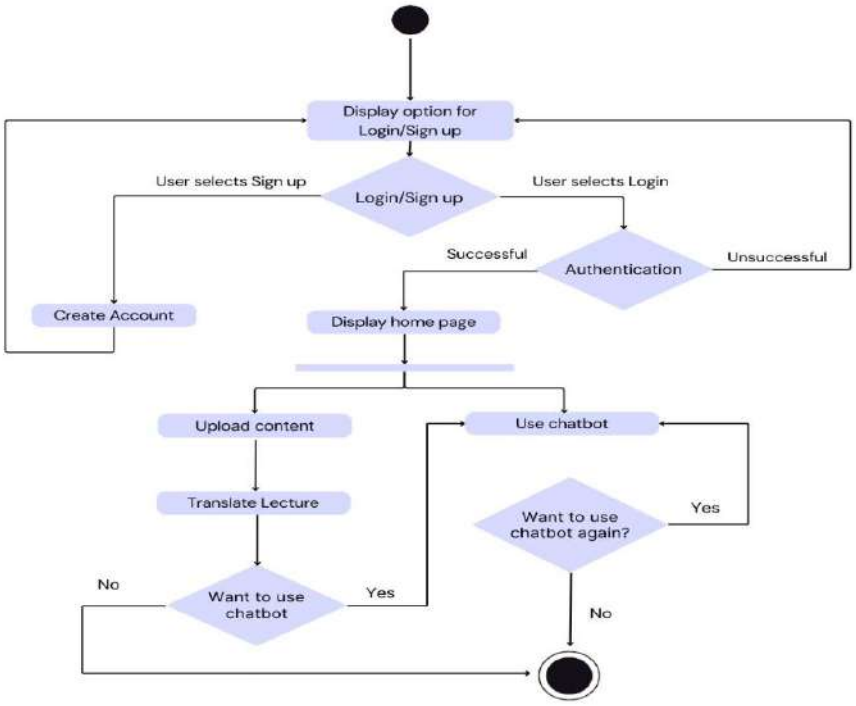

Figure 4: Activity diagram

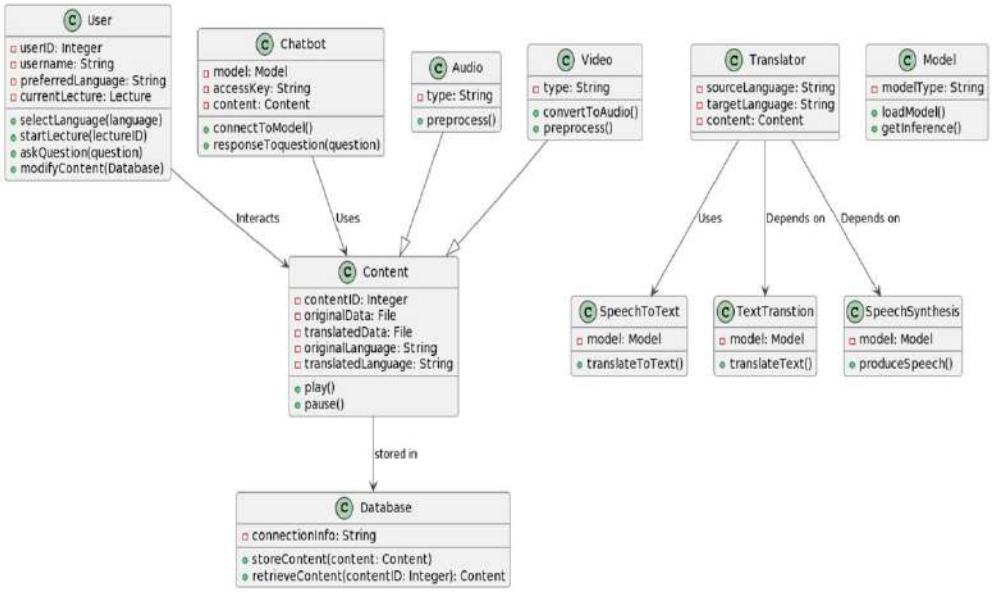

Figure 5: Class diagram

In the first cycle, backend functionality for language translation was developed using various models, such as neural machine translation. The integration of the frontend with this backend ensured seamless user interaction. Rigorous testing was conducted to verify basic functionality and ensure reliability.

During the second cycle, the backend for the chatbot module was developed, employing the Retrieval-Augmented Generation (RAG) [8] technique to facilitate intelligent conversation. Integration of the two modules was followed by comprehensive testing to evaluate their combined functionality and interoperability.

The development of Natlearn exemplifies a systematic approach to software design and implementation, characterized by the adoption of Agile methodology and the utilization of state-of-the-art design models. Strong data and evidence were provided throughout the project to justify design choices and ensure the successful integration of multi-lingual translation and an intelligent chatbot. The deployment of the software on the Azure cloud platform marks a significant milestone in realizing the project's objectives, paving the way for future enhancements and advancements in educational technology.

Chapter 4: Proposed Solution (Methodology)

NatLearn proposes a comprehensive solution whose aim is to enrich students educational experiences by granting them access to content in their native language. The project is structured into two distinct modules, each of which addresses a fundamental aspect of natural language processing: Multilingual

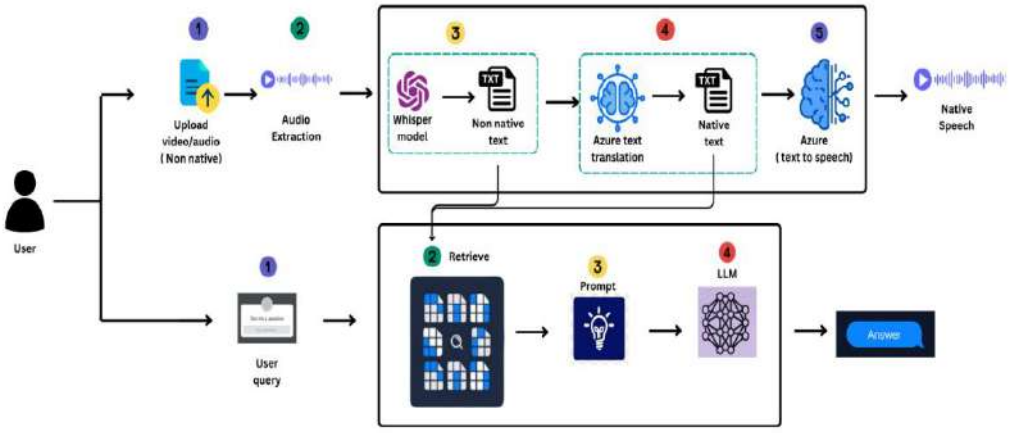

Figure 6: High Level block diagram

video translation and context specific chatbot. Both of these modules are designed to address the complex challenges within the realm of language understanding. The multilingual video translation module focuses on facilitating synchronized communication across different languages through state of art AI models and algorithms, whereas the chatbot module acts as an intelligent conversational agent capable of engaging users in natural language interactions. The high level diagram of NatLearn is shown in Figure[6].

Below, we provide a detailed explanation of each module:

[6] Figure 6: High Level block diagram

4.1 Multilingual Video Translation

The methodology for the speech translation module within NatLearn comprises six sub-modules. Herein, we delineate each one:

4.1.1 Audio Extraction

This initial step involves acquiring video content in a non-native language, serving as the primary data source. The video undergoes data preprocessing utilizing the ffmpeg module to extract audio, subsequently saved in '.wav' format. The following parameters govern this conversion process:

- **"-ab", "192k"**: Specifies the audio bitrate for the output file, set to 192 kbps to ensure quality audio encoding.
- **"-ar", "44100"**: Determines the audio sampling rate for the output file, set to 44100 Hz, adhering to the standard sampling rate for CD-quality audio.

4.1.2 Speech to Text Conversion and Text Alignment

Following audio extraction, the extracted audio segments, along with the source language name, are segmented to fit the input size of the speech-to-text model. These segments undergo speech-to-text conversion and subsequent text alignment. Here, the state-of-the-art open-source AI model, 'Whisper' from OpenAI, is employed for converting audio to text. The output text segments are merged to form the final transcribed text, which is stored in '.txt' format for retrieval by the chatbot. The text segments from the Whisper model are further processed by the Whisperx model, a variant of Whisper, to obtain segment-level timestamps for the transcribed text. These timestamped transcripts facilitate synchronization with audio in subsequent phases.

4.1.3 Text Translation and Profanity Filtration

The time stamped transcribed text, along with the target language name, undergoes translation using Microsoft Azure's Text Translation Model, facilitating conversion to the desired language. Subsequently, the translated text undergoes profanity filtration to eliminate any undesirable content stemming from either the speech-to-

text or text translation processes. The resulting translated text is stored in '.txt' format for retrieval by the chatbot.

4.1.4 Speech Synthesis

The translated text is segmented into smaller units suitable for the input size of the speech synthesis model. These segments, accompanied by the specified gender voice, are then fed into Microsoft Azure's [9] speech synthesis model to generate realistic human-like voices. Each synthesized segment is saved in '.wav' format.

4.1.5 Synchronization

In this phase, the length of each synthesized segment (denoted as 't1') is computed and compared against the difference of the corresponding index of the timestamped transcribed text ('t2') from phase 2. If the length of the synthesized segment exceeds that of the timestamped transcribed text, the audio segment is accelerated. The mathematical expression governing this process is as follows:

```
if (t1 > t2):
        speedup_factor = ceil (t1 / t2)
        sped_up_audio = audio.speedup(speedup factor)
```

The 'sped_up_audio' ensures that the length of the audio file remains within the bounds of the original audio and video files. Once this step is completed for all segments, the synthesized segments are merged to form the final translated audio file.

4.1.6 Audio Insertion over Video

In this crucial step, the final synthesized audio file is seamlessly integrated into the original video file. Upon completion, the synchronized translated video file is saved, representing the culmination of the entire process.

4.2 Context-Specific Chatbot

The chatbot component of NatLearn comprises five distinct parts, each serving a specific function. The detailed explanation of each part is as follows:

4.2.1 Documents Loading and Chunking

Initially, all text files corresponding to translated videos are loaded and laoded into a singular document. This document is subsequently divided into smaller chunks, each consisting of 200 tokens with an overlapping region of 10 tokens.

4.2.2 Vector Embedding Creation and Vector Database Storage

Following chunking, the document undergoes conversion into vector embeddings utilizing an embedding model. Specifically, Azure's OpenAI embedding model is employed to ensure enhanced performance. The resulting embedding vectors are then stored within vector databases, chosen for their superior performance over traditional relational and NoSQL databases concerning vector data. The FAISS vector database is utilized for local storage of vector embeddings.

4.2.3 Large Language Model Initialization

Large Language Models (LLMs) refer to those AI models that are trained on vast amounts of textual data, known as corpus, to produce human-like text in response to prompt. For NatLearn, Microsoft Azure's GPT 3.5 Turbo is initialized to facilitate answer generation upon receiving prompts.

4.2.4 Prompt Engineering and RAG Chain

Custom prompt is crafted to ensure that we have maximum control over the responses generated by the Large Language Model in response to user queries. At the core of NatLearn's chatbot component lies the Retrieval-Augmented Generation [8](RAG) technique, offering an alternative to fine-tuning AI models for contextual answer

generation. RAG combines two key components: retrieval and generation. In the retrieval part, relevant context and information is extracted from a large database or knowledge base. Whereas in the generation part, LLM is provided the relevant context and information along with the query to generate relevant responses. RAG chain refers to the sequence of steps involved in the process of generating text using retrieval-augmented generation models. Custom prompt, initialized LLM, and initialized retrieval object from the vector database collectively construct the RAG chain.

4.2.5 Answer Generation

Upon initialization of the RAG chain, the user's prompt is passed to this chain to generate relevant responses. This generated response is also checked for profanity using profanity filtration modules and then this profanity proof response is presented to the user.

Chapter 5 Implementation

NatLearn will be developed using a combination of programming languages such as Python's Django framework for backend development and JavaScript for frontend development. Frameworks like Hugging Face [11] and LangChain will be utilized for implementing AI models and NLP techniques. The following provides detailed information about the implementation of the frontend, backend, video translation, and chatbot components:

5.1 Backend Development**:** Python's Django framework is chosen for backend development due to its scalability, and built-in features such as ORM (Object-Relational Mapping), and authentication which streamlines the development process, enabling efficient creation of backend services.

5.2 Frontend Development: HTML, CSS, JavaScript, and the Bootstrap framework are used for the frontend development of NatLearn. This choice is made to ensure cross-platform compatibility and an intuitive user interface. Bootstrap facilitates rapid development by offering responsive-ready templates for different UI components, enhancing the user experience across different devices and screen sizes.

5.3 Video Translation: The video translation process involves several modules and techniques:

a. **Ffmpeg:** Ffmpeg is utilized to convert video file to audio format, enabling further processing of the audio stream.

b. **OpenAI Whisper Module:** OpenAI's Whisper module [12] is employed to transcribe original audio to text, providing accurate transcriptions of the audio content.

c. **WhisperX Module:** WhisperX module is used to obtain timestamped transcripts of the original audio, this allows precise synchronization of translated text with the original video.

d. **Microsoft Azure Text Translation Model:** Microsoft Azure's [9]text translation model is integrated to convert text from the source language to the target language.

e. **Microsoft Text-to-Speech Model:** Microsoft's text-to-speech model is used to generate realistic human voices from translated text.

f. **Pydub Module:** Pydub [13] is utilized for post-processing the generated audio clips, enabling tasks such as speeding up audio playback and merging audio clips.

g. **Moviepy Module:** Moviepy module [14] is used to overlay translated audio over the original video clip, which enables the creation of fully translated video content synchronized with the original visuals.

Constraints and Tools/Techniques:

Finetuning instead of RAG: In order to address the challenge of dynamic data handling, where each user's translated video content varies across different categories, NatLearn opted for finetuning a Large Language Model (LLM) for individual users. However, this approach proved to be time and memory consuming. To mitigate these issues, NatLearn adopted the Retrieval-Augmented Generation (RAG) technique, which allows for updates solely to the existing knowledge base, represented by our vector database. This approach minimizes the computational overhead associated with adapting LLMs for each user's unique requirements.

Hallucinations from LLMs: One of the major issues with LLMs is that they face "hallucinations," where the model generates text containing information or context not present in the input data or that is inaccurate. NatLearn addresses this challenge through two key measures. Firstly, we prioritize LLMs based on benchmark scores, selecting models that demonstrate superior performance. Secondly, NatLearn implements a profanity filtering technique to identify and remove any harmful or inappropriate content from the generated responses, ensuring the integrity of the output.

Resource Limitations and Response Time Constraint: NatLearn optimizes resource utilization to achieve scalability and performance. To enhance response time and scalability, NatLearn deploys AI model endpoints on cloud computing platforms, specifically Microsoft Azure.

Chapter 6 Results & Discussion

NatLearn's implementation was subjected to careful review of benchmarks score and experimentation done to the model's generalization ability by NatLearn team and open evaluation platforms. The following section presents a comprehensive discussion of the results obtained and their implications.

6.1 Speech to text Validation:

Evaluation of speech-to-text models, also known as Automatic Speech Recognition (ASR) models, is crucial for assessing their accuracy and performance. The primary metric used for this evaluation is the Word Error Rate (WER), calculated as the total number of mistakes (insertions, deletions, or replacements of words) divided by the total number of words in the ground truth. A WER value of zero indicates perfect prediction by the model.

The Whisper model by OpenAI underwent evaluation on FLEURS datasets, comprising short utterances read in controlled, low-noise environments. However, for real-world scenarios, Deepgram's validation on messier datasets revealed challenges, particularly in transcribing Hindi language content. Furthermore, consulting the 'Open ASR Leaderboard ranks' by Huggingface positioned Whisper at 6th place. Notably, Whisper's selection was influenced by its capability to transcribe over 45 different languages, enhancing accessibility for diverse user bases. Figure[7], Figure[8] and Figure[9] are showing results.

[7] Figure 7: Whisper evaluation
[8] Figure 8: Whisper's evaluation on Custom Messier dataset
[9] Figure 9: Official ASR leaderboard

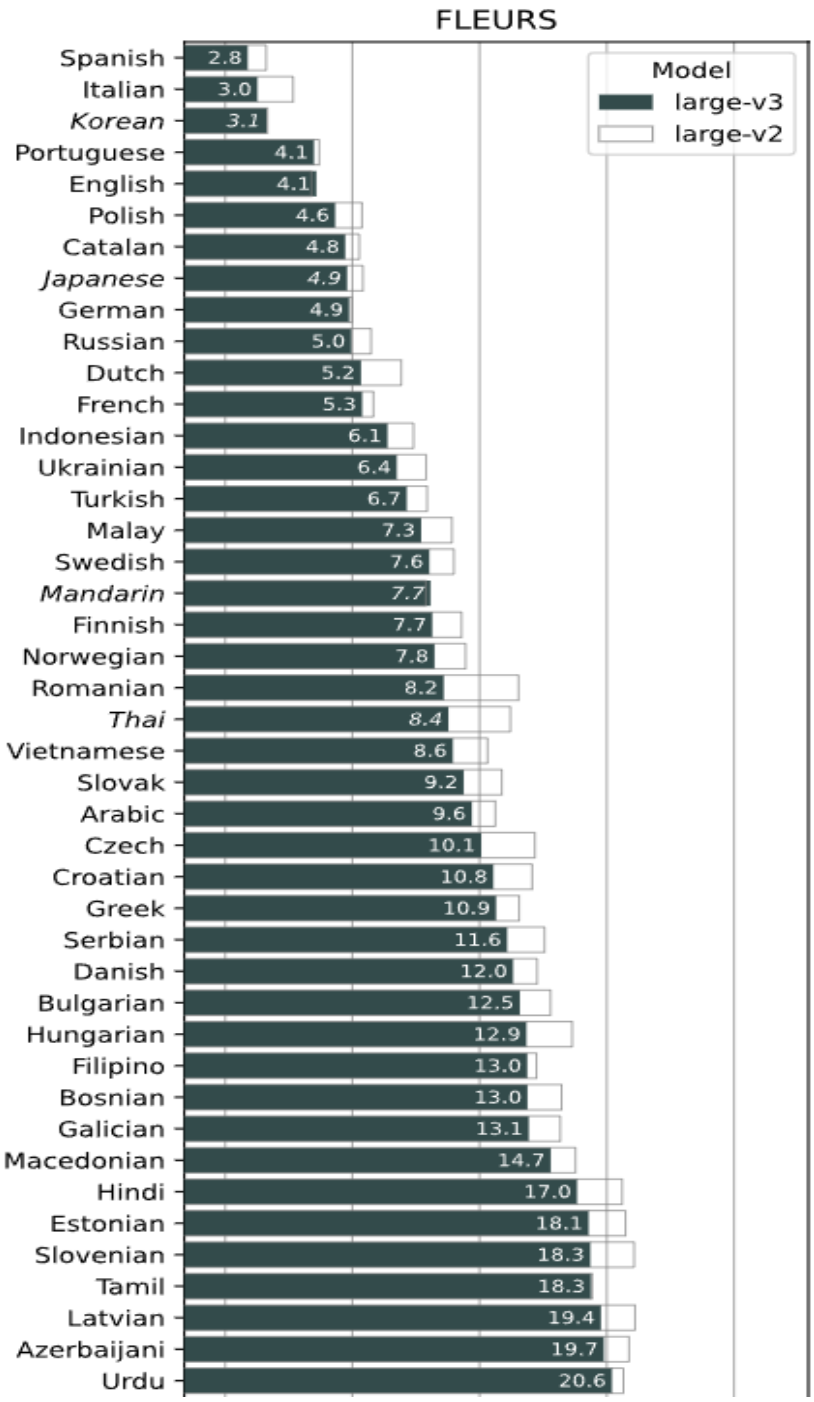

Figure 7: Official Whisper WERs evaluated on Fleurs dataset

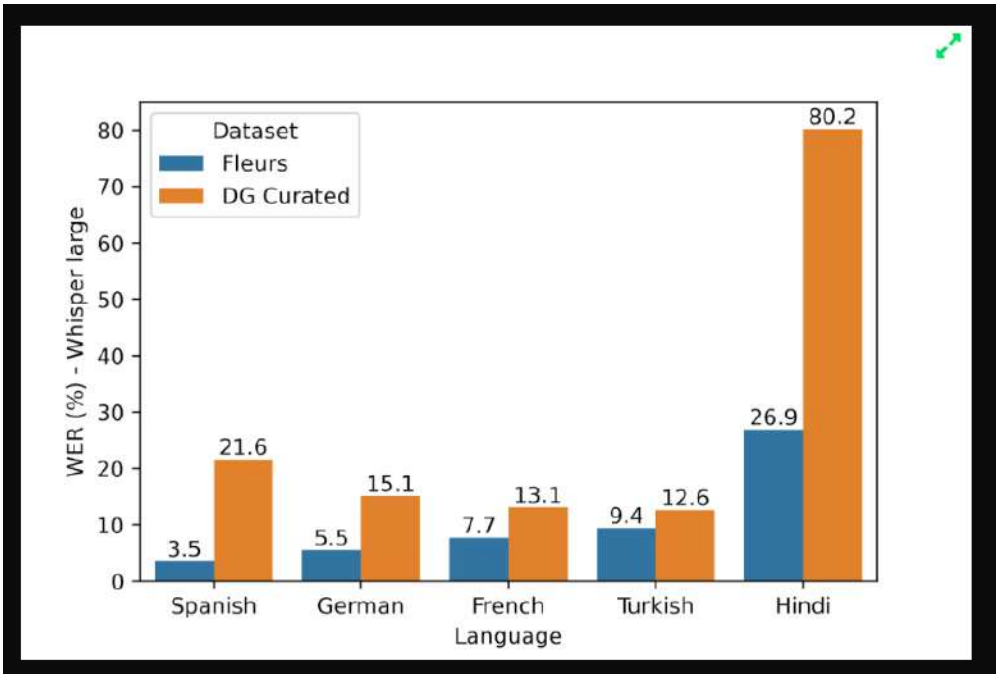

Figure 8: Whisper's evaluation on Custom Messier dataset

Leaderboard | Metrics | Request a model here!

model	Average WER	RTF (1e-3)	ANI	Earnings22	Gigaspeech	LS Clean	LS Other	SPGISpeech
nvidia/canary-1b	6.67	9.8	14	12.25	18.19	1.49	2.94	2.86
nvidia/parakeet-tdt-1.1b	6.95	1.7	15.9	14.65	9.55	1.39	2.62	3.42
nvidia/parakeet-rnnt-1.1b	7.04	2.4	17.1	14.15	9.96	1.46	2.48	3.11
nvidia/parakeet-ctc-1.1b	7.58	0.6	15.6	13.69	18.27	1.83	3.54	4.2
nvidia/parakeet-rnnt-0.6b	7.63	2	17.66	14.9	18.87	1.62	3.86	3.47
openai/whisper-large-v3	7.7	7.45	16.01	11.3	18.82	2.83	3.91	2.95

Figure 9: Official ASR leaderboard

6.2 Text Translation Model evaluation:

Evaluation of machine translation models primarily relies on the BLUE score and the model's language coverage. The BLUE score, short for Bilingual Evaluation Understudy, assesses precision-based features by comparing machine translations to reference translations. Higher scores (ranging from 0 to 1) indicate greater similarity between machine and human translations.

Initially, the NLLB Model by Meta demonstrated a 44% improvement in BLEU score relative to the previous state-of-the-art model, providing translations for over 200 languages. However, its performance on lengthy sentences in real-world scenarios was suboptimal. Consequently, Microsoft Azure's text translation model emerged as a preferable choice. Evaluated on the Global Voices corpus, encompassing news segments from over 50 languages, Azure's model exhibited the lowest standard deviation across languages, indicating consistent translation quality. This consistency, coupled with its performance surpassing competitors like Google Cloud Platform and AWS, led to its selection for integration into NatLearn. Figure[10] shows Metrics score for text translation.

[10] Figure 10 : Metrics score for text translation

Metric Scores (BLEU, METEOR, NIST)

	GCP (mean)	AWS (mean)	Azure (mean)	GCP (sd)	AWS (sd)	Azure (sd)
Persian BLEU	0.090	0.073	0.081	0.172	0.156	0.165
Arabic BLEU	0.209	0.195	0.200	0.233	0.225	0.223
Russian BLEU	0.240	0.211	0.221	0.253	0.239	0.243
Persian METEOR	0.254	0.294	0.244	0.209	0.225	0.206
Chinese BLEU	0.361	0.224	0.277	0.212	0.182	0.204
Arabic METEOR	0.405	0.501	0.400	0.221	0.235	0.219
Russian METEOR	0.435	0.508	0.419	0.222	0.236	0.221
Chinese METEOR	0.567	0.529	0.522	0.176	0.168	0.167
Persian NIST	0.847	0.888	0.810	0.854	0.870	0.840
Persian Combined	1.191	1.255	1.135	1.235	1.251	1.211
Arabic NIST	1.436	1.509	1.420	1.031	1.055	1.011
Russian NIST	1.554	1.594	1.500	0.981	0.992	0.973

Figure 10: Metrics score for text translation

6.3 Text to Speech Model Evaluation:

Evaluation of text to speech (TTS) models is essential to ensure natural and coherent synthesis of text into speech.Initially Meta's MMS model which is based on the wav2vec architecture was tested due to its extensive pre-training on a vast dataset covering multiple languages. However, its performance, particularly on less-resourced languages like Urdu and Hindi, was found to be unsatisfactory.

As a result, Azure's text to speech model was adopted for its high-quality speech production. The model can generate speech across more than 70 languages which makes it a suitable choice for NatLearn. According to the

Mean Opinion Score (MOS) metrics, the model's output had negligible difference from natural human recordings.

6.4 Embedding model evaluation:

The selection process for embedding model involved relying on the Massive Text Embedding Benchmark (MTEB) Leaderboard.

Initially, the Multilingual-E5-large-instruct model was chosen for its generalization capabilities which was trained on a mixture of multilingual datasets, ranking 12th on the MTEB leaderboard[11]. However, due to resource limitations, Microsoft Azure's text-embedding-ada-002 model was selected, ensuring efficient utilization of computational resources at hand.

Rank	Model	Model Size (GB)	Embedding Dimensions	Max Tokens	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)	Pair Classification Average (3 datasets)	Reranking Average (4 datasets)	Retrieval Average (15 datasets)
6	echo-mistral-7b-instruct-last	14.22	4096	32768	64.68	77.43	46.32	87.34	58.14	55.52
7	mxbai-embed-large-v1	0.67	1024	512	64.68	75.64	46.71	87.2	60.11	54.39
8	UAE-Large-V1	1.34	1024	512	64.64	75.58	46.73	87.25	59.88	54.66
9	text-embedding-3-large		3072	8191	64.59	75.45	49.01	85.72	59.16	55.44
10	voyage-lite-01-instruct		1024	4000	64.49	74.79	47.4	86.57	59.74	55.58
11	Cohere-embed-english-v3.0		1024	512	64.47	76.49	47.43	85.84	58.01	55
12	multilingual-e5-large-instruc	1.12	1024	514	64.41	77.56	47.1	86.19	58.58	52.47
13	GIST-large-Embedding-v0	1.34	1024	512	64.34	76.01	46.55	86.7	60.05	53.44
14	bge-large-en-v1.5	1.34	1024	512	64.23	75.97	46.08	87.12	60.03	54.29
15	Cohere-embed-multilingual-v3.		1024	512	64.01	76.01	46.6	86.15	57.86	53.84
16	GIST-Embedding-v0	0.44	768	512	63.71	76.03	46.21	86.32	59.37	52.31

Figure 11 Massive Text Embedding Benchmark (MTEB) Leaderboard

[11] Figure 11 : MTEB Leaderboard

6.5 LLM Evaluation

The evaluation of Large Language Models (LLMs) for NatLearn was critical to select a model that can generate relevant responses while minimizing the occurrence of hallucinations.

To assess LLM performance, we referred to the 'LLM Hallucination Index[12],' a benchmark report that provides a comprehensive measurement of LLM hallucinations based on their propensity to hallucinate across three common task types - question & answer without RAG, question and answer with RAG, and long-form text generation. Initially, the top-ranked open-source model for the "Q&A with RAG" category was identified as "zephyr-7b-beta." However, upon testing this model, it was found to be unreliable as it produced answers in different languages even when not instructed to do so, likely due to its training on multilingual data. Consequently, "gpt-3.5-turbo-1106," ranked 3rd in the "Q&A with RAG" category, was adopted.

The performance of "gpt-3.5-turbo-1106" proved satisfactory for NatLearn's requirements.

[12] Figure 12: Hallucination Index

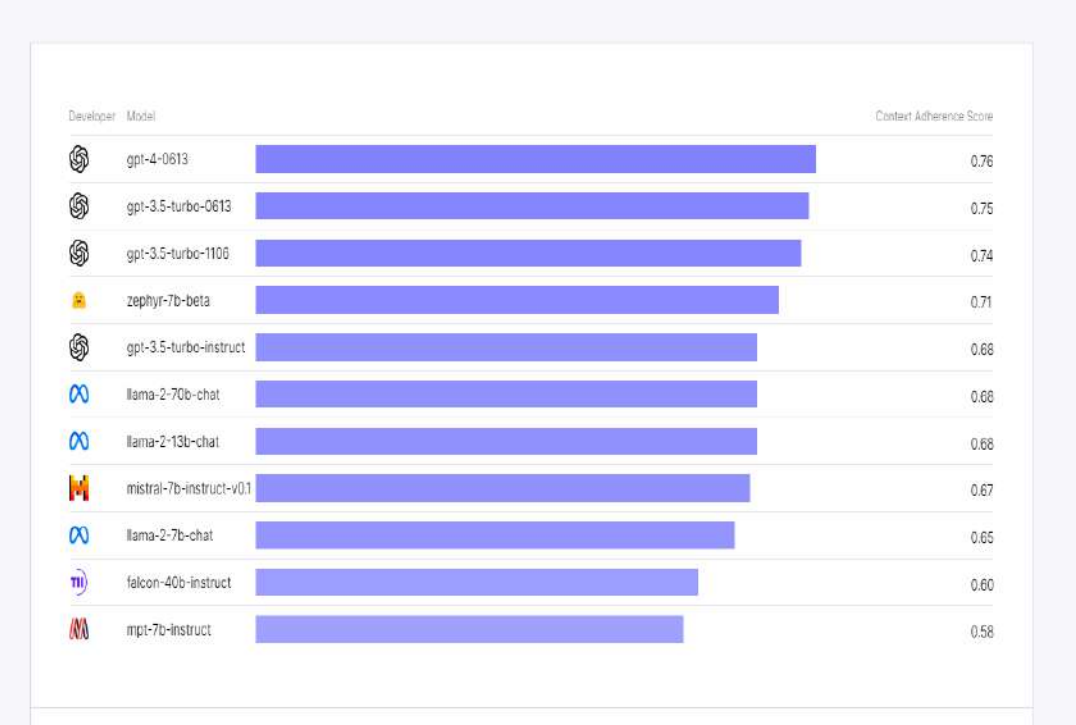

Figure 12 LLM Hallucination Index

Chapter 7: Conclusion and Future work

7.1 Conclusion

The development of the web application presented in this project represents a significant step towards addressing the challenges associated with accessing educational content in multiple languages and providing real-time assistance through integrated chatbot functionality. Through the implementation of automated video translation, intuitive user interfaces, and sophisticated natural language processing algorithms, the web application aims to democratize access to educational resources and foster a more inclusive learning environment. The project has successfully achieved its objectives by creating a user-friendly platform that allows users to upload lectures in any language and seamlessly translate them into their desired language, thereby breaking down language barriers and expanding access to knowledge. The integration of a chatbot enhances the learning experience by providing contextualized assistance and answering user queries related to the translated content, further facilitating comprehension and engagement. Moreover, the project has demonstrated the potential of cutting edge technologies, such as machine learning and artificial intelligence, to address longstanding challenges in language translation and educational accessibility. By leveraging these technologies, the web application not only delivers accurate and high-quality translations but also adapts to user interactions in real-time, enhancing the overall usability and effectiveness of the platform. In conclusion, the development of this web application represents a significant contribution to the field of education technology, offering a scalable and versatile solution for individuals seeking to access educational content in multiple languages and receive personalized assistance through integrated chatbot functionality. By

empowering learners worldwide to overcome language barriers and unlock the full potential of educational resources, the project contributes to the advancement of inclusive and accessible education on a global scale.

7.2 Future Work

While the current iteration of the web application has achieved its primary objectives, there are several avenues for future work and potential enhancements to further improve its functionality, usability, and impact. The following areas represent potential directions for future development and research:

- **Enhanced Translation Quality**: Continuously improving the quality and accuracy of translation through the exploration of advanced machine translation models, incorporating domain-specific terminology, and implementing user feedback mechanisms to refine translation algorithms.
- **Expanded Language Support**: Increasing the range of supported languages to cater to a more diverse user base and ensuring comprehensive coverage of less commonly spoken languages and dialects.
- **Advanced Chatbot Capabilities**: Enhancing the capabilities of the integrated chatbot by incorporating advanced natural language understanding and generation techniques, enabling more complex dialogue management and supporting a wider range of user queries and interactions.
- **Personalization and Adaptation**: Implementing personalized learning experiences by leveraging user data and preferences to tailor content recommendations, adaptive learning paths, and targeted interventions based on individual learning styles and goals.

- **Multimodal Learning Resources:** Integrating support for multimedia content beyond video lectures, such as interactive simulations, virtual reality experiences, and augmented reality overlays, to provide more engaging and immersive learning experiences.
- **Community Engagement and Collaboration:** Facilitating community engagement and collaboration among users through features such as discussion forums, collaborative annotation tools, and peer-to-peer language exchange opportunities, fostering a sense of belonging and shared learning experiences.
- **Accessibility and Inclusivity**: Enhancing accessibility features to accommodate users with disabilities, including support for screen readers, alternative text for multimedia content, and adjustable font sizes and contrast ratios to ensure a barrier-free learning environment for all users.
- **Integration with Learning Management Systems (LMS):** Seamlessly integrating the web application with existing learning management systems used in educational institutions, allowing for easy integration of translated content into course curricula and providing instructors with insights into student engagement and comprehension.
- **Continuous Evaluation and Improvement:** Conducting ongoing evaluation and iterative improvement of the web application through user testing, usability studies, and analytics-driven insights, ensuring that the platform remains responsive to evolving user needs and technological advancements.
- **Research and Innovation:** Exploring emerging technologies and research trends in the fields of natural language processing, machine

learning, and educational technology to identify opportunities for innovation and differentiation in the development of future iterations of the web application. By pursuing these avenues for future work, the web application can continue to evolve as a leading platform for language-agnostic education, empowering learners worldwide to overcome linguistic barriers and access high-quality educational content in their preferred language.

References

[1] E. A. &. S. M. Moneerh Aleedy, "Springer link," 2022. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-05675-8_11.

[2] Baharuddin, M. D. Mendoza, O. Y. Hutajulu and H. Fibriasari, "EBSCO Logo," 2023. [Online]. Available: https://openurl.ebsco.com/EPDB%3Agcd%3A5%3A1741317/detailv2?sid=ebsco%3Aplink%3Ascholar&id=ebsco%3Agcd%3A174669701&crl=c.

[3] M. G. H. S. M. Fawaz Qasem, "Dialog chatbot as an interactive," 2023. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/SJLS-10-2022-0072/full/pdf?title=dialog-chatbot-as-an-interactive-online-tool-in-enhancing-esp-vocabulary-learning.

[4] OpenAI, "ChatGPT," OpenAI, 2022. [Online]. Available: https://chat.openai.com/. [Accessed 2024].

[5] N.-B. L. T.-M.-T. N. Xuan-Quy Dao, "AI-Powered MOOCs: Video Lecture Generation," 2021.

[6] H. Inaguma, "Multilingual End-to-End Speech Translation," in *IEEE*, 2019.

[7] Google, 2006. [Online]. Available: https://translate.google.com/. [Accessed 2024].

[8] "Q&A with RAG," [Online]. Available: https://python.langchain.com/docs/use_cases/question_answering/. [Accessed 2024].

[9] "Microsoft Azure: Cloud Computing Services," Microsoft, [Online]. Available: https://azure.microsoft.com/en-us.

[10] "What is a large language model (LLM)?," Cloudflare, Inc., [Online]. Available: https://www.cloudflare.com/learning/ai/what-is-large-language-model/#:~:text=Large%20language%20models%20(LLMs)%20are,massive%20data%20sets%20of%20language.. [Accessed 2024].

[11] "Hugging Face," [Online]. Available: https://huggingface.co/. [Accessed 2024].

[12] "Introducing Whisper," [Online]. Available: https://openai.com/research/whisper. [Accessed 2024].

[13] "pydub," [Online]. Available: https://pypi.org/project/pydub/. [Accessed 2024].

[14] "moviepy," [Online]. Available: https://pypi.org/project/moviepy/. [Accessed 2024].