

# **VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY**

**(An Autonomous Institute Affiliated to University of Mumbai)**

## **Department of Computer Engineering**



Project Report on

## **CareerLens: Career Counseling Meet**

## **Summarizer**

Submitted in partial fulfillment of the requirements of the degree

## **BACHELOR OF ENGINEERING IN COMPUTER ENGINEERING**

By

**Deven Bhagtani D17B / 05**

**Piyush Chugeja D17B / 10**

**Sakshi Kirmathe D17B / 24**

**Manraj Singh Viridi D17B / 63**

Project Mentor

**Dr. (Mrs.) Nupur Giri**

**University of Mumbai**

**(A.Y. 2024 - 25)**

# **VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY**

**(An Autonomous Institute Affiliated to University of Mumbai)**

## **Department of Computer Engineering**



### **Certificate**

This is to certify that the Mini Project entitled **“CareerLens: Career Counseling Meet Summarizer”** is a bonafide work of **Deven Bhagtani (D17B - 05), Piyush Chugeja (D17B - 10), Sakshi Kirmathe (D17B - 24), & Manraj Singh Viridi (D17B - 63)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **“Bachelor of Engineering”** in **“Computer Engineering”**.

**Dr. (Mrs.) Nupur Giri**  
Mentor, Head of Department

**Dr. (Mrs.) J. M. Nair**  
Principal

# Mini Project Approval

This Mini Project entitled “**CareerLens: Career Counseling Meet Summarizer**” by **Deven Bhagtani (D17B - 05), Piyush Chugeja (D17B - 10), Sakshi Kirmathe (D17B - 24), & Manraj Singh Virdi (D17B - 63)** is approved for the degree of **Bachelor of Engineering** in **Computer Engineering**.

## Examiners

1. ....  
(Internal Examiner name & sign)

2. ....  
(External Examiner name & sign)

**Date:** 23<sup>rd</sup> October 2024

**Place:** Chembur, Mumbai

# Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

-----

(Deven Bhagtani - D17B / 05)

-----

(Piyush Chugeja - D17B / 10)

-----

(Sakshi Kirmathe - D17B / 24)

-----

(Manraj Singh Viridi - D17B / 63)

Date: 23<sup>rd</sup> October 2024

## Acknowledgement

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to the Head of the Computer Department **Dr. (Mrs.) Nupur Giri** (Project Guide) for her kind help and valuable advice during the development of project synopsis and for her guidance and suggestions.

We are deeply indebted to our Principal **Dr. (Mrs.) J.M. Nair**, for giving us this valuable opportunity to do this project.

We express our hearty thanks to them for their assistance without which it would have been difficult to finish this project synopsis and project review successfully.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support and selfless help throughout the project work. It is a great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Engineering.

We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement several times.

# Index

| Chapter No.      | Title   | Page Number |
|------------------|---|-------------|
| <b>Chapter 1</b> | <b>Introduction</b>   | <b>1</b>    |
| 1.1              | Introduction  | 1           |
| 1.2              | Motivation  | 1           |
| 1.3              | Drawback of the existing system   | 2           |
| 1.4              | Problem Definition  | 2           |
| 1.5              | Relevance of the Project  | 2           |
| 1.6              | Methodology used  | 3           |
| <b>Chapter 2</b> | <b>Literature Survey</b>  | <b>4</b>    |
| 2.1              | Research Papers<br>Abstract of the research paper<br>Inference drawn from the paper   | 4           |
| <b>Chapter 3</b> | <b>Requirements Of Proposed System</b>  | <b>6</b>    |
| 3.1              | Functional Requirements   | 6           |
| 3.2              | Non-Functional Requirements   | 6           |
| 3.3              | Constraints   | 7           |
| 3.4              | Hardware & Software Requirements  | 7           |
| 3.5              | Techniques utilized   | 7           |
| 3.6              | Tools utilized  | 8           |
| 3.7              | Project Proposal  | 9           |
| <b>Chapter 4</b> | <b>Proposed Design</b>  | <b>10</b>   |
| 4.1              | Block diagram representation  | 10          |
| 4.2              | Modular diagram representation  | 11          |
| 4.3              | Design of the proposed system<br>a. Data Flow Diagrams<br>b. Flowchart for the proposed system<br>c. Screenshot of implementation | 12          |
| <b>Chapter 5</b> | <b>Proposed Results and Discussions</b>   | <b>18</b>   |
| 5.1              | Evaluation of Models  | 18          |

|                  |   |           |
|------------------|---|-----------|
| 5.2              | Insights from Evaluation  | 19        |
| 5.3              | Graphs of training loss vs validation loss for different models | 19        |
| <b>Chapter 6</b> | <b>Plan Of Action For the Next Semester</b>                     | <b>21</b> |
| 6.1              | Work done till date   | 21        |
| 6.2              | Plan of action for project II                                   | 22        |
| <b>Chapter 7</b> | <b>Conclusion</b>   | <b>23</b> |
| <b>Chapter 8</b> | <b>References</b>   | <b>24</b> |
| <b>Chapter 9</b> | <b>Appendix</b>   | <b>25</b> |
| 9.1              | List Of Figures   | 25        |
| 9.2              | List Of Tables  | 25        |
| 9.3              | Xerox of project review sheet                                   | 26        |
| 9.4              | Draft of research paper   | 27        |

# Chapter I: Introduction

## 1.1 Introduction

With the rise of remote work in 2020-2021, online meetings have become integral to professional interactions. Platforms like Zoom and Google Meet report millions of daily participants, signaling a significant shift toward virtual communication [1, 2, 3, 4]. This trend extends to career counseling, where virtual meetings offer the opportunity to guide individuals in professional development. However, these sessions need accurate documentation to ensure valuable insights are captured, which is challenging with traditional methods. Abstractive summarization, particularly when using pre-trained large language models (LLMs), addresses this challenge by providing concise, actionable summaries of large transcripts [5, 6].

CareerLens enhances the efficiency of career counseling by automatically transcribing and summarizing conversations using a fine-tuned LLaMA model. By leveraging QLoRA [7] for efficient finetuning, the system ensures minimal memory usage while maintaining high performance. This approach not only saves time but also allows counselors and clients to focus on essential takeaways without getting lost in details. The system's interactive capabilities further enhance understanding, allowing users to ask questions about the summarized content, thus promoting more informed decision-making in career guidance [6, 8].

## 1.2 Motivation

The CareerLens project is motivated by the rising demand for efficient and effective communication in career counseling. Since many people find it difficult to express their job goals and obstacles, counselors must offer clear direction. Traditional note-taking techniques may hinder the process, resulting in miscommunication and loss of important information. CareerLens wants to give counselors and clients useful insights by automating the summarizing of counseling sessions. In order to help users make educated choices regarding their professional pathways, the initiative aims to improve counseling overall. In the end, We want to close the communication gap between customers and counselors in order to promote a more productive and encouraging atmosphere for professional growth.



### **1.3 Drawback of existing systems**

Current systems are mostly involved with general meeting functions, and often lack the particular capabilities required for successful career counseling. These systems have limitations in the ability to give specific guidance and insights for people's career paths since they do not provide any analysis specifically designed for career counseling. Furthermore, many applications don't offer live transcription; instead, they rely on captions or transcriptions from the meeting app, which lowers accuracy and speed. In addition, the lack of real-time question-and-answer features restricts efficient user-application communication. These systems fail to effectively arrange and retrieve meeting data because of their inflexible user interfaces and limited analytics, which leaves them without follow-up methods to monitor users' progress after sessions.

### **1.4 Problem Definition**

In today's fast-paced digital age, career counseling sessions have become essential for guiding people down successful career pathways. However, traditional methods for taking notes, summarizing, and determining outcomes from these meetings are difficult and susceptible to mistakes, which causes delays in the retrieval of information and the making of decisions. This may have a negative impact on how well counseling is delivered. In addition, the lack of modern technologies for automatically analyzing and summarizing meeting information limits counselors' capacity to deliver timely and relevant insights to their clients. It can be difficult to extract useful insights because of the naturally difficult process of documenting and reviewing these sessions, which frequently results in inefficiency. As a result, it can be challenging for both clients and counselors to keep track of progress, pinpoint important topics for discussion, and make decisions based on counseling sessions.

### **1.5 Relevance of the project**

The relevance of the project lies in its potential to transform the way career guidance is delivered. By integrating advanced technologies like AI and natural language processing, the project addresses existing gaps in career counseling systems, such as real-time transcription and interactive Q&A capabilities. This innovative approach not only enhances the efficiency and effectiveness of counseling sessions but also provides users with tailored insights and actionable recommendations. Ultimately, the project aims to empower individuals to make informed career decisions and improve their professional journeys.

## 1.6 Methodology used

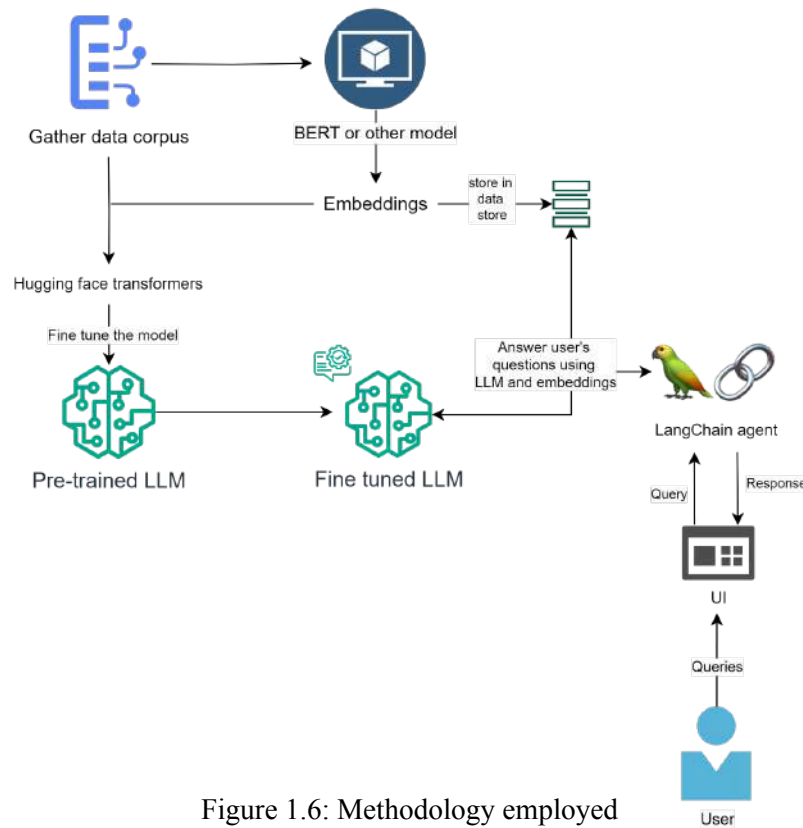


Figure 1.6: Methodology employed

Methodology used in the Career Counseling Meet Analyzer:

- **Fine-tune Llama Model:** Adjust the Llama model parameters to optimize its capabilities for accurately analyzing and summarizing career counseling sessions, ensuring relevant outputs.
- **Transcription Collector:** Implement a mechanism to collect real-time captions from Google Meet sessions, incorporating multilingual support to cater to a diverse audience and enhance accessibility.
- **Create User Dashboard:** Develop an intuitive user interface that allows users to easily interact with the system, view generated summaries, and access key insights from their counseling sessions.
- **Build & Test:** Construct the entire system, followed by a comprehensive testing phase to verify that all components function correctly and meet the desired performance standards.
- **Deploy Model to Cloud:** Move the trained model to a cloud environment, enabling easy access and scalability. The model will be accessible via APIs, allowing seamless integration into various applications.
- **Conduct Final Testing:** Execute thorough testing of the developed system in its cloud environment to ensure all functionalities work and provide a reliable user experience.

# Chapter II: Literature survey

## 2.1 Research Papers Review

The domain of meeting summarization has gained pace significantly due to the high usage of virtual communication technologies. Numerous methodologies have been proposed to address the issue of summarizing meeting transcripts, and each methodology identifies its merits and demerits. Nonetheless, there is much still lacking for many systems that are currently applied, such as content dealing with multilingual content, summarizing in real-time, and accuracy for complex conversational contexts.

New innovations have resulted in the establishment of even more sophisticated techniques like AI-based multilingual summarization systems [9], tasked with the ever-increasing needs of running multilingual meetings. Such systems utilize Latent Semantic Analysis (LSA) besides advanced NLP techniques to establish important points and action items. The method, however, tends to over-simplify the complex conversational interactions because of the inherent use of LSA-like techniques for dimensionality reduction. Transformer models, on the other hand, can recognize complex relations within dialogue without trading off contextual richness. The summaries are thus concise yet comprehensive in multiple languages. Some of the works focus on abstractive summarization using Transformer-based architectures [5]. For example, the use of Pointer Generator Networks with coverage mechanisms has significantly improved conventional models in terms of reducing word repetition and enhancing the readability of generated summaries. However, such methods often rely on extensive datasets for training, such as news summaries, which can be limiting when applied to more domain-specific contexts like career counseling. Unlike these approaches, our methodology fine-tunes Meta’s LLaMA 3.1 on a focused dataset, ensuring that the model adapts to the subtleties of career counseling conversations and reduces hallucinations or the generation of unrelated content.

To further enhance our model’s efficiency, we leverage QLoRA for fine-tuning, a method that significantly reduces memory usage while maintaining performance on large-scale models [7]. This allows us to fine-tune LLaMA on a bit quantized pretrained model, facilitating the efficient handling of a specialized dataset for career counseling. Moreover, this approach ensures scalability, enabling even resource-constrained systems to handle models with billions of parameters. Furthermore, we prioritize the use of LLaMA 3.1 over other closed-source models such as GPT-4 due to cost and privacy

considerations. Open-source models like LLaMA provide competitive performance, even in zero-shot settings, and avoid the privacy risks and high costs associated with API-based fine-tuning of closed models [6].

Hybridization of extractive and abstractive techniques, as seen in Jotter [11] hybridizes the techniques to gain another promising avenue. By combining BERT embeddings with sequence-to-sequence models, Jotter is able to balance efficiency with extractive accuracy and human-like readability of abstractive summaries. This two-tiered approach excels at coherence but may introduce computational overhead, limiting its use in real time. A thorough analysis by Kumar and Kabiri [10] points out the key challenges in the meeting summarization domain, specifically the challenge of extracting relevant information from large dialogue datasets. It also intimates domain-specific models and evaluation metrics because most of the approaches currently in use are benchmarked on generic datasets like AMI and ICSI. Our work directly addresses that gap because career counseling involves sensitive and domain-specific information requiring special summarization techniques. The fine-tuned models will carry customized evaluation metrics considering accuracy and actionable wisdom, rather than mere ROUGE or BLEU scores.

In more specialized domains, like mental health counseling, one can see that there are even systems, like ConSum[12], implementing filtering and structuring dialogues based on domain knowledge techniques. The use of PHQ-9 by ConSum for utterance filtering showcases the possibility of externalized knowledge in the summarization process. Motivated by this concept, our system integrates user roles and speaker information directly within the model's output. This integration facilitates role-aware insights that improve the contextual significance of the summaries, rendering them more applicable in professional environments.

Thus our system builds on the foundation laid down by the earlier work for meeting summarization in which it offers multilingual capabilities, real-time summarization, and optimization tailored specifically for the job of career counseling. We employ transformer-based frameworks like MeetSum [5], but we fine-tune our models using QLoRA [7] that adapt to the unique framework of career counseling interactions. In addition, our focus on efficient deployment aligns with the real-time objectives emphasized in prior work, yet also resolves the existing challenges pertaining to computational burden and latency.

# Chapter III: Requirement Gathering for the Proposed System

## 3.1 Functional Requirements

Functional requirements specify the specific functionalities and features that the system must provide to meet the needs of stakeholders and users. Functional requirements for our system are:

- **Real-time Transcription:** The system must capture and transcribe audio from Google Meet sessions in real-time, providing accurate and timely captions for users.
- **Summarization of Meetings:** The system should generate concise and contextually relevant summaries of the counseling sessions, highlighting key discussion points and action items.
- **Interactive Q&A Module:** Users should be able to ask questions related to the session's content, and the system must provide accurate answers based on the transcribed text.
- **Integration with Existing Tools:** The system must be able to integrate with popular video conferencing tools, such as Google Meet, to facilitate seamless operation.
- **Multilingual Support:** The transcription feature should support multiple languages, allowing users from diverse linguistic backgrounds to access the system.

## 3.2 Non-Functional Requirements

Non-functional requirements define the quality attributes and constraints that the system must adhere to. In our project, non-functional requirements includes:

- **Performance:** The system should provide real-time transcription with minimal latency, ensuring that users receive timely and accurate captions during counseling sessions.
- **Scalability:** The architecture must support scalability to handle an increasing number of users and concurrent sessions without degrading performance.
- **Reliability:** The system should maintain a high level of reliability, ensuring that transcriptions and summaries are consistently accurate and available to users.
- **Security:** The system must implement robust security measures to protect user data, including encryption for stored data and secure authentication for user access.
- **Compatibility:** The application should be compatible with multiple devices and operating systems, including desktops, tablets, and mobile devices.

### 3.3 Constraints

- **Data Availability:** The effectiveness of the model is dependent on the availability of high-quality training data, which may be limited or require extensive preprocessing.
- **Technical Limitations:** There may be limitations related to the performance of the Llama model, such as processing speed and memory usage, which can affect real-time transcription capabilities.
- **Budget Constraints:** Financial limitations may restrict the resources available for cloud services, infrastructure, and technology needed to implement and maintain the system.
- **Integration Issues:** Challenges may arise in integrating the new system with existing platforms or tools used by counselors and clients, affecting overall usability and functionality.
- **Multilingual Support Complexity:** Providing effective multilingual support can increase the complexity of the system, requiring additional resources for language models and testing.

### 3.4 Hardware & Software Requirements

| Hardware Requirements  | Software Requirements                 |
|--|---------------------------------------|
| Computer System: <ul style="list-style-type: none"><li>a. CPU i7 or higher</li><li>b. RAM 16GB or more</li><li>c. GPU - 16GB or more</li></ul> | Code Editor - VS Code or Google Colab |
| Webcam or Camera   | Llama 3.1 Model                       |
| Microphone   | Browser (Any)                         |
|  | Visualization Libraries (charts)      |

Table 3.4: Hardware & Software Requirements

### 3.5 Techniques utilized till date

- **Speech-to-Text Technology:** Implemented to capture real-time audio from career counseling meetings and convert it into text transcripts. This ensures accurate documentation of discussions.
- **Natural Language Processing (NLP):** Utilized through the LLM module (Llama 3) to analyze the transcribed text. Key tasks include extracting insights, summarizing content, and responding to user inquiries based on the meeting discussions.
- **Real-time Data Collection:** Enabled through the integration of live data streaming from the meeting, facilitating immediate analysis and interaction during the

counseling sessions.

- **Interactive Q&A Mechanism:** Developed to allow users to engage with the system in real-time, providing them with instant answers to questions related to the topics discussed in the meeting.
- **Data Visualization:** Employed to present summaries, insights, and action items in an easy format on the dashboard, enhancing user comprehension and decision-making.
- **User Interface Design:** Focused on creating an intuitive and user-friendly interface for both the live session and the post-meeting dashboard to ensure a seamless experience for users.
- **Feedback Loop Mechanism:** Established to gather user feedback on the effectiveness of the system, enabling continuous improvement and adaptation to meet user needs.

### **3.6 Tools utilized till date**

- **Programming Languages:**
  - **Python:** The primary programming language for backend development, particularly for integrating AI and data processing functionalities.
  - **JavaScript:** Utilized for frontend development to create interactive web elements and manage user interactions.
- **Natural Language Processing Libraries:**
  - **Hugging Face Transformers:** Used for implementing the LLM module (Llama 3), enabling advanced NLP capabilities such as text summarization and question answering.
- **Speech Recognition:**
  - **Google Speech-to-Text API:** Employed to convert audio from live meetings into text, facilitating real-time transcription.
- **Data Visualization Tools:**
  - **Matplotlib and Seaborn:** Libraries used for creating visualizations and graphs to represent insights on the dashboard effectively.
- **Web Development Frameworks:**
  - **React or Angular:** For frontend development, ensuring a responsive and dynamic user interface.
- **Database Management:**
  - **SQLite or PostgreSQL:** Used to store session transcripts, user interactions, and any other relevant data for analysis and reporting.

- Development and Testing Platforms:
  - Google Colab: Used for prototyping, testing various algorithms, and model fine-tuning in a collaborative environment.
  - Jupyter Notebook: Helpful for exploratory data analysis and visualizations during the development phase.
- Dataset Sourcing:
  - Kaggle: Employed for sourcing datasets relevant to career counseling and for model fine-tuning to improve performance.
- Version Control:
  - Git and GitHub: Utilized for source code management and collaboration, enabling tracking of changes and versioning throughout the development process.

### **3.7 Project Proposal**

The CareerLens: Career Counseling Meet Analyzer aims to revolutionize career counseling by utilizing advanced AI and natural language processing technologies. This innovative tool will capture and transcribe real-time discussions during counseling sessions, generating concise summaries and actionable insights. By integrating interactive question-and-answer features, users can easily extract relevant information, enhancing the overall effectiveness of the counseling process. The project addresses existing gaps in traditional counseling methods, ensuring a more efficient and supportive experience for both counselors and clients, ultimately leading to improved career outcomes.



# Chapter IV: Proposed Design

## 4.1 Block diagram of system

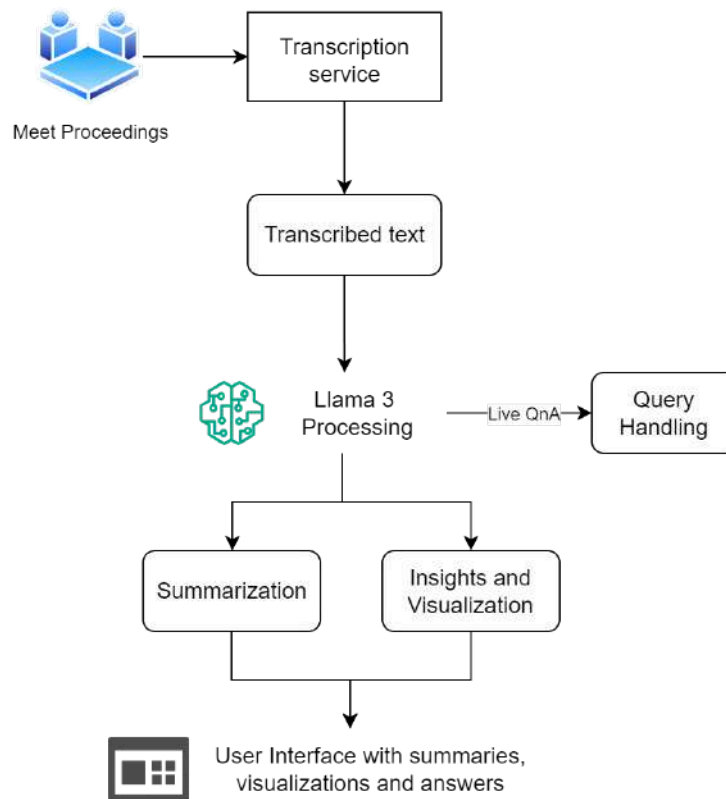


Figure 4.1: Block diagram of the system

The block diagram displays the workflow of the CareerLens system, which combines transcription, language processing, and visualization to provide users with an effortless experience.

1. **Meet Proceedings:** This is a representation of the direct conversation or meeting that occurs during a career counseling session.
2. **Transcription service:** A transcription service records and transforms the meeting's spoken information into transcribed text in real time.
3. **Processing by Llama 3:** The transcribed text is fed into the fine-tuned Llama 3 model for advanced processing, allowing the system to analyze and understand the meeting content.
4. **Query Handling:** The system supports live Q&A functionality, enabling users to ask questions related to the session topics and receive instant responses.
5. **Summarization:** The Llama 3 model also automatically summarizes the counseling session by extracting important ideas and conclusions into brief summaries.
6. **Visualizations and Insights:** The system generates visual summaries and key insights, highlighting important data points and patterns from the meeting.
7. **User Interface:** A dashboard displays summaries, insights, and visualizations, allowing users to interact easily with the system and get answers to their questions.

## 4.2 Modular design of the system

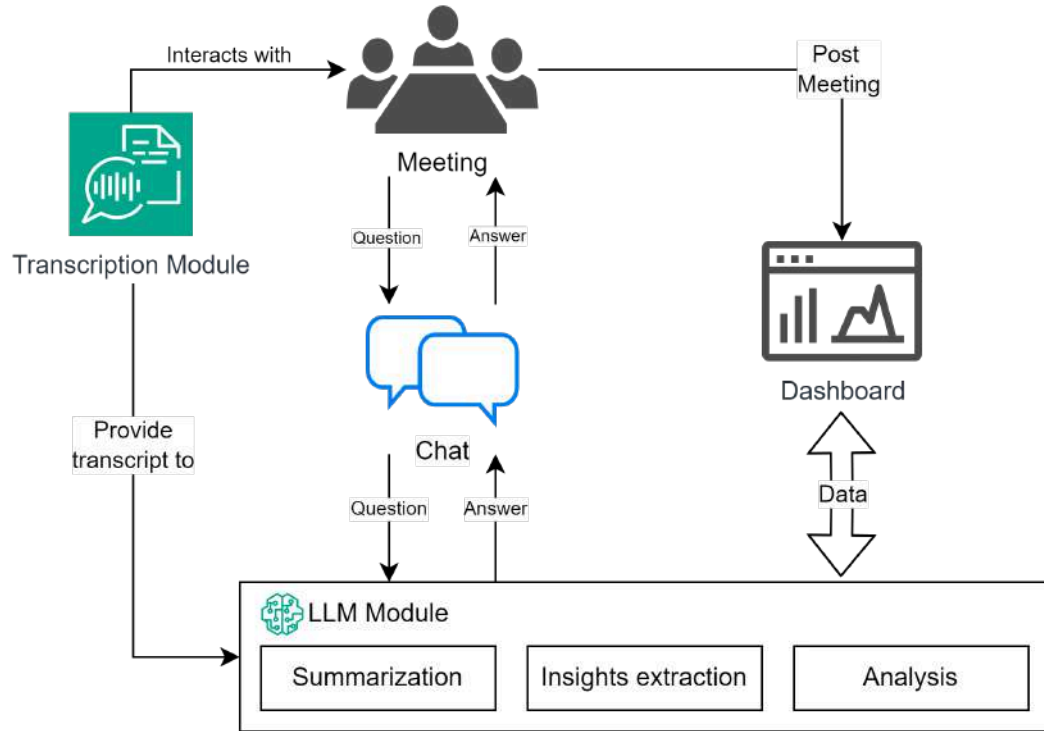


Figure 4.2: Modular diagram of the system

The modular design of our system encompasses several key components, each responsible for specific functionalities to ensure the overall effectiveness and efficiency of the system.

1. **Meeting:** The system collects data in real-time from career counseling meetings, capturing all discussions among participants. This live data collection enables immediate analysis and interaction, enhancing the overall counseling experience.
2. **Transcription Module:** This module captures the live conversation during the meeting, converting spoken words into real-time transcripts. These transcripts are essential for the subsequent analysis and are transmitted to the LLM module for deeper insights.
3. **LLM Module (Llama 3):** As the core AI component, this module analyzes the transcripts to perform various tasks. It extracts key insights, summarizes important points, and provides responses to questions raised during the session, enhancing the understanding of the discussions.
4. **Chat Interaction:** This feature facilitates real-time question-and-answer sessions, allowing users to engage with the system to receive relevant responses on the topics discussed in the meeting. It fosters an interactive environment and ensures participants' inquiries are addressed promptly.
5. **Dashboard:** After the meeting concludes, the system displays a comprehensive dashboard featuring detailed summaries, insights, visualizations, and actionable items. This presentation allows users to review and analyze the key takeaways from the session effectively.

## 4.3 Design of the proposed system

### a. Data Flow Diagrams

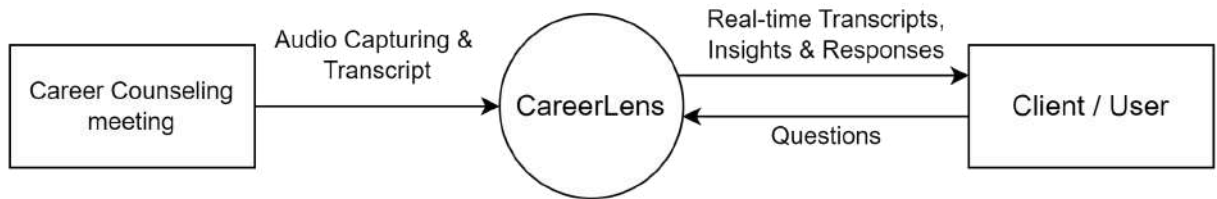


Figure 4.3.a.a: Level 0 DFD

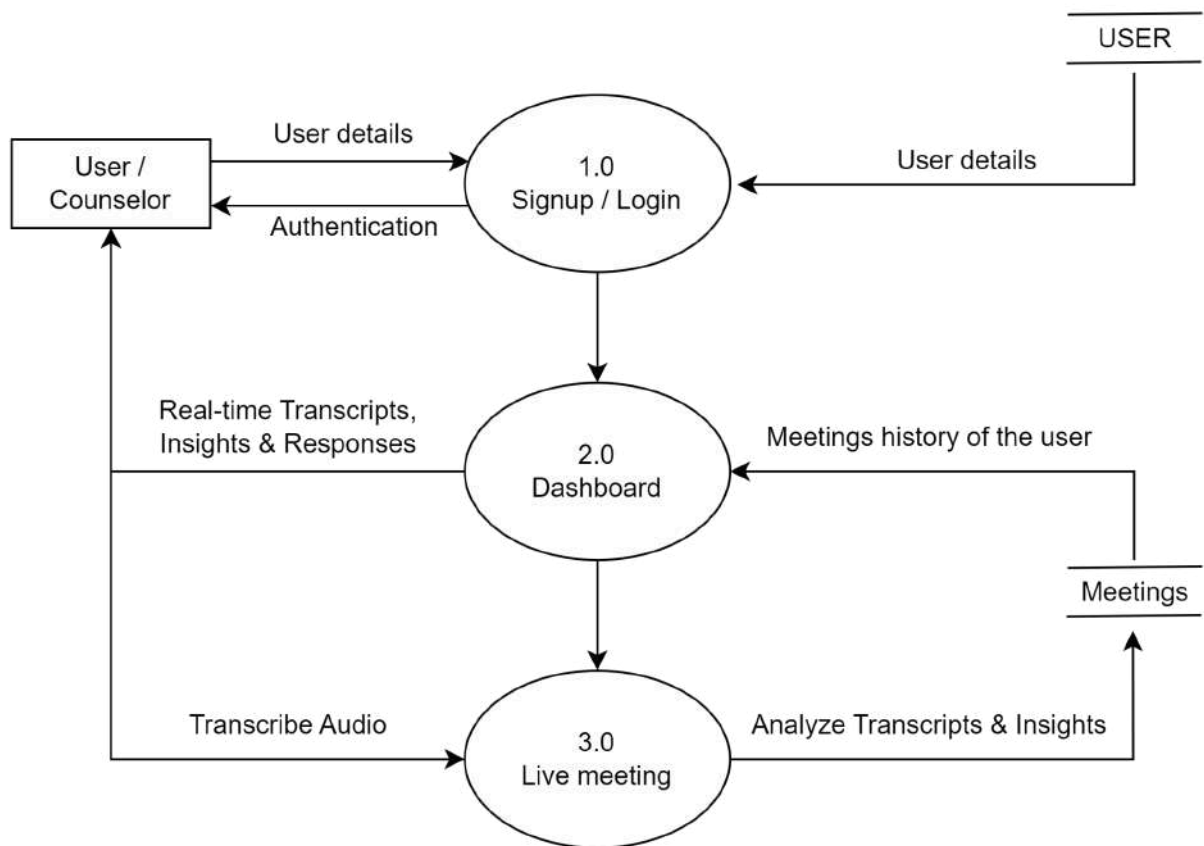


Figure 4.3.a.b Level 1 DFD

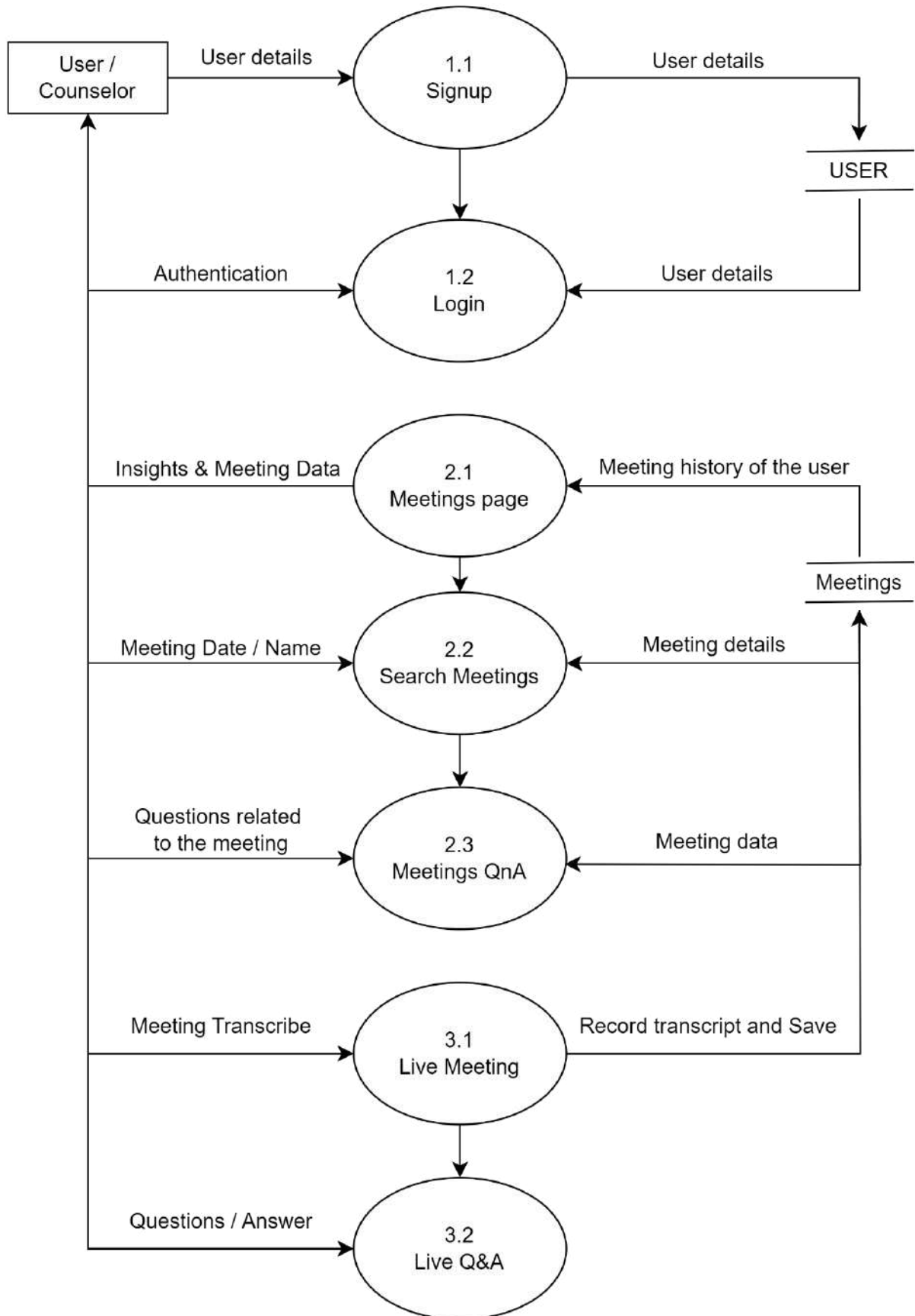


Figure 4.3.a.c: Level 0 DFD

**b. Flowchart for the proposed system**

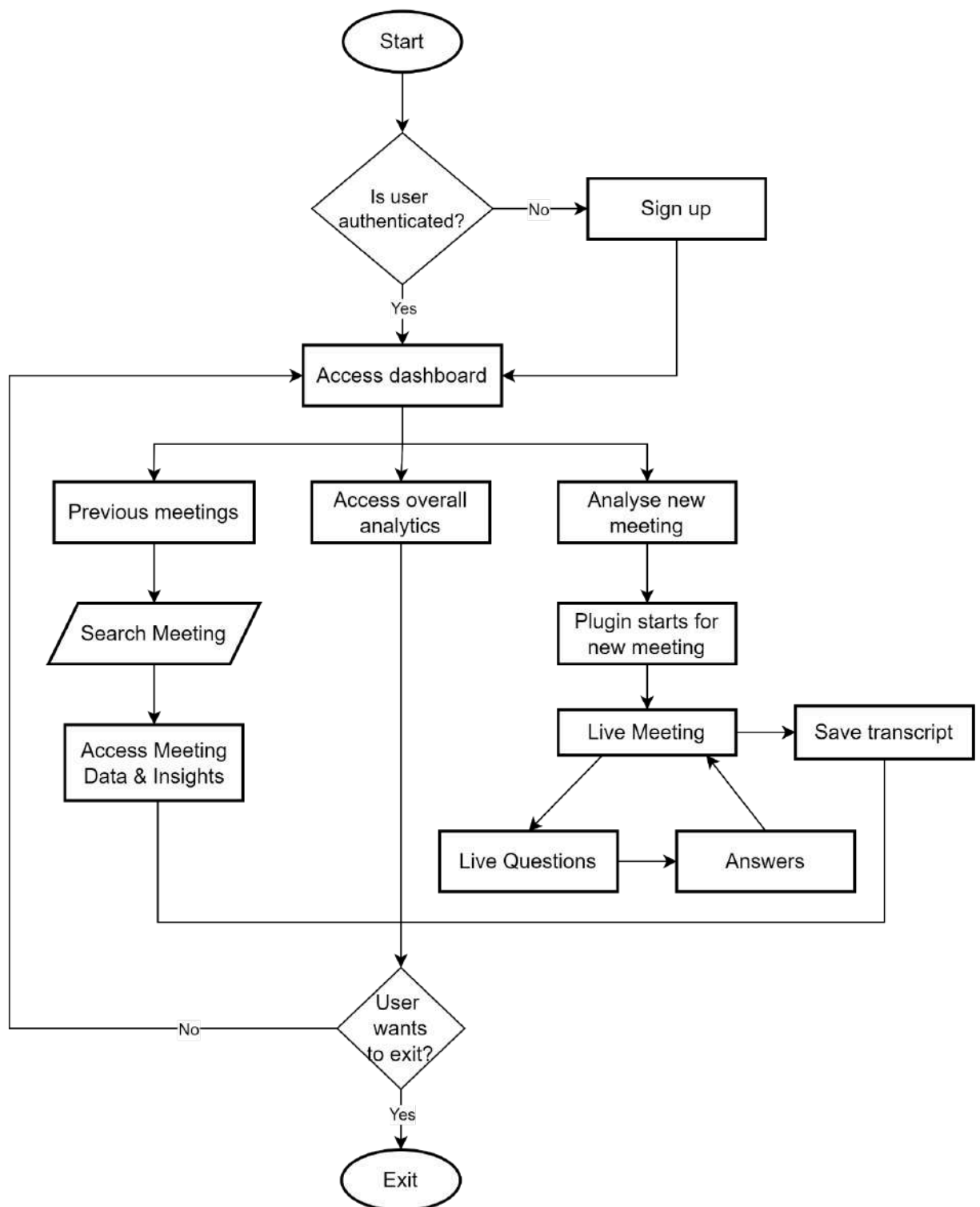


Figure 4.3.b: Flowchart of the system

### c. **Outputs**

For a given transcript as follows:

Dr. Kirmathe: Hi, Piyush! How are you today

Piyush: Hi, Dr. Kirmathe. I'm good, thank you. I'm really looking forward to our conversation today. I've been thinking a lot about my career direction lately.

Dr. Kirmathe: I'm glad to hear that, Piyush. Let's dive right in. Tell me a bit about what you're currently doing and what's been on your mind regarding your career.

Piyush: Well, I'm currently in my final year of engineering. I've enjoyed certain aspects of the coursework, like coding and problem-solving, but I'm also considering career paths outside of traditional software development. I want to leverage my technical background but move into a role that involves more creativity and strategy.

Dr. Kirmathe: That's a very insightful perspective. It sounds like you have a strong technical foundation and you're looking for ways to apply it in a more dynamic field. Do you have any specific roles in mind?

Piyush: Yes, I've been exploring roles like product management and UI/UX design. Both seem to combine technical skills with creativity and problem-solving, which really appeals to me. But I'm unsure about how to transition into these areas since my experience is mostly technical.

Dr. Kirmathe: Those are great options, and they do indeed align well with your engineering background. Product management involves working with teams to develop products from concept to launch, which requires a mix of technical knowledge, communication skills, and an understanding of market needs. UI/UX design, on the other hand, is more focused on creating intuitive and appealing interfaces. It's a highly creative field where you can directly influence user experience.

Piyush: That's exactly what I'm interested in—using my skills to make an impact on how users interact with products. I'm just not sure where to start in terms of building the necessary skills.

Dr. Kirmathe: That's perfectly normal. Let's break ...

Output in JSON is:


```
{
  "summary": "Piyush, a software engineering student, seeks advice on transitioning from software development to roles that combine technical skills with creativity and problem-solving. Dr. Kirmathe suggests exploring product management and UI/UX design, emphasizing the importance of building a portfolio, communication skills, and empathy. Piyush plans to start with online courses and personal projects to build his portfolio, and Dr. Kirmathe offers to support him along the way.",
  "action_items": [
    "Start by learning about product lifecycle, customer research, and project management. Look for internships or part-time roles to gain hands-on experience....."
  ],
  "insights": [
    "Both product management and UI/UX design require excellent communication skills.",
    "Empathy is important for both roles, helping product managers understand customer needs and UI/UX designers put themselves in the user's shoes..."
  ],
  "speakers": [
    "Dr. Kirmathe",
    "Piyush"
  ]
}
```

The screenshot displays the 'Career Switch' transcript viewing interface on the Career Lens platform. The interface is divided into several sections:

- Left Sidebar:** Contains navigation links such as 'Add a Meeting', 'My Meetings', 'Shared with Me', 'Account & Settings', and 'Archive'. Below these is a 'Recent Meetings' list showing multiple entries for 'Meetings with Nupur Ma'am'.
- Header:** Features the 'CAREER LENS' logo and a 'Search in Transcript' input field.
- Main Content Area:**
  - Title:** 'Career Switch'.
  - Metadata:** 'Recorded by msv, 10/23/2024, 1:35:59 AM, Last updated 3 days ago'.
  - Tools:** Includes 'Quick Prompts' and buttons for 'Quick Summary' and 'Bullet Points'.
  - Search:** A text input field with the placeholder 'Ask anything about this meeting.' and an 'Ask' button.
  - Transcript:** A list of conversation turns with speaker avatars and timestamps:
    - Dr. Kirmathe (00:00):** 'Hi, Piyush! How are you today'.
    - Piyush (00:01):** 'Hi, Dr. Kirmathe. I'm good, thank you. I'm really looking forward to our conversation today. I've been thinking a lot about my career direction lately.'
    - Dr. Kirmathe (00:02):** 'I'm glad to hear that, Piyush. Let's dive right in. Tell me a bit about what you're currently doing and what's been on your mind regarding your career.'
    - Piyush:** 'Well, I'm currently in my final year of engineering. I've enjoyed certain aspects of the'.

- Right Sidebar:**
- Speaker Stats:** Displays progress bars for 'Dr. Kirmathe' (55.56%) and 'Piyush' (44.44%).
- Notes:** A text area with the placeholder 'Add your notes here'.
- Comments:** A text area with the placeholder 'Add your comments here' and a 'Save' button.

Figure 4.3.c.1: Transcript Viewing


< Meeting Transcription

---

Add a Meeting

My Meetings

Shared with Me

Account & Settings

Archive

**Action Items**

Copy 
 Delete

- ☐ Piyush to research product management and UI/UX design roles further, focusing on the necessary skills and qualifications.
- ☐ Dr. Kirmathe to provide guidance on how to transition into product management and UI/UX design, including potential courses or training programs.
- ☐ Piyush to explore networking opportunities in the product management and UI/UX design fields to gain insights from professionals in these areas.
- ☐ Dr. Kirmathe to recommend resources or tools for Piyush to develop his skills in product management and UI/UX design.
- ☐ Piyush to create a plan for transitioning into a product management or UI/UX design role, including setting specific goals and timelines.

**Summary**


Copy 
 Delete

In this session, the discussion focused on Piyush's career direction and exploration of roles that combine technical skills with creativity and strategy. Key insights and strategies include: a narrative overview of Piyush's current situation and career aspirations, details about the exploration of roles like product management and UI/UX design, analysis of the challenges in transitioning into these areas, and key quotes or moments that stood out, such as Piyush's interest in using his skills to make an impact on user experience.


**Insights**

Copy 
 Delete

Piyush's Interest in using his technical skills to make an impact on user experience is a key insight into his career aspirations. The exploration of roles like product management and UI/UX design is a strategic move for Piyush to leverage his technical background in a more dynamic field. The challenge of transitioning into product management and UI/UX design requires careful planning and research. Networking opportunities in the product management and UI/UX design fields can provide valuable insights and guidance for Piyush. Developing skills in product management and UI/UX design requires a commitment to ongoing learning and professional development.



**Manraj Singh Virdi**  
 d2021.manrajsingh.virdi@ves.ac.in



CAREER

LENS

+

Add a Meeting

+

My Meetings

+

Shared with Me

+

Account & Settings

+

Archive

Recent Meetings

Meetings with Nupur Ma'am

Meetings with Nupur Ma'am

Meetings with Nupur Ma'am

Meetings with Nupur Ma'am

Meetings with Nupur Ma'am

Meetings with Nupur Ma'am

Meetings with Nupur Ma'am

Meetings with Nupur Ma'am

Meetings with Nupur Ma'am

Meetings with Nupur Ma'am

Search in Transcript

Career Switch

Recorded by msv, 10/23/2024, 1:35:59 AM, Last updated 3 days ago

Tools

Quick Prompts

Quick Summary

Bullet Points

Why is piyush considering a career switch?

Ask

Answer

Piyush is considering a career switch because he wants to leverage his technical background in a more dynamic field that involves more creativity and strategy, and he's interested in roles that combine technical skills with problem-solving and creativity.

Transcript

D

Dr. Kirmathe

00:00

Hi, Piyush! How are you today

P

Piyush

00:01

Hi, Dr. Kirmathe. I'm good, thank you. I'm really looking forward to our conversation today. I've been thinking a lot about my career direction lately.

Speaker Stats

Dr. Kirmathe

55.56%

Piyush

44.44%

Notes

Add your notes here

Comments

Add your comments here

Save



# Chapter V: Proposed Results and Discussions

## 5.1 Evaluation of Models

| Model Configuration   | GLEU   | ROUGE  | BLEU   | BERT Score | Eval Loss | Inference Time |
|---|--------|--------|--------|------------|-----------|----------------|
| batch_size: 2<br>gradient_accumulation_steps: 8<br>bnb_4bit_quant_type: fp16<br>learning_rate: 5e-5 | 0.2054 | 0.3804 | 0.1299 | 0.9168     | 1.9762    | 24.3s          |
| batch_size: 4<br>gradient_accumulation_steps: 4<br>bnb_8bit_quant_type: fp16<br>learning_rate: 5e-5 | 0.2165 | 0.3899 | 0.1304 | 0.9151     | 1.8698    | 26s            |
| batch_size: 4<br>gradient_accumulation_steps: 8<br>bnb_4bit_quant_type: nf4<br>learning_rate: 4e-5  | 0.1788 | 0.3578 | 0.1073 | 0.9104     | 1.8417    | 20.67s         |
| batch_size: 4<br>gradient_accumulation_steps: 4<br>bnb_4bit_quant_type: fp16<br>learning_rate: 3e-5 | 0.0805 | 0.1880 | 0.0305 | 0.8624     | 2.0341    | 66s            |

Table 5.1.a: Model Evaluation Results

| Sr No. | Evaluation Metric | Purpose  | Expected Output   |
|--------|-------------------|--|---|
| 1.     | GLEU              | Evaluates the fluency and accuracy of generated responses or summaries by comparing them to reference outputs  | GLEU score (0-1), higher is better                          |
| 2.     | ROUGE             | Measures overlap of n-grams between system-generated and reference summaries                                   | ROUGE score (precision, recall, F1)                         |
| 3.     | Perplexity        | Evaluates the quality of generated text by comparing it to one or more reference texts, focusing on precision. | BLEU score (0-1), higher is better                          |
| 4.     | BERT Score        | Compares semantic meaning between system output and reference using contextual embeddings                      | BERT score (semantic similarity)                            |
| 5.     | Eval Loss         | Measures the difference between predicted and true values during model training or evaluation                  | Loss value (lower is better)                                |
| 6.     | Inference Time    | Measures the time taken by the model to generate a response or summary in real-time                            | Time in seconds (lower is better for real-time performance) |

Table 5.1.b: Evaluation Metrics

## 5.2 Insights from Evaluation

- **Higher Batch Size Shows Improved Scores:**
  - Models with a batch size of 4 tend to perform better across most metrics (GLEU, ROUGE, BLEU, and BERT scores) compared to those with a batch size of 2.
  - For example, the model with a batch size of 4 and gradient accumulation steps of 4 achieved the highest GLEU score (0.2165) and ROUGE score (0.3899).
- **Trade-offs Between Inference Time and Accuracy:**
  - The model with the best BLEU score (0.1304) had a relatively moderate inference time (26 seconds). On the other hand, models with lower scores (e.g., the one with a BLEU score of 0.0305) had significantly higher inference times (66 seconds).
  - This highlights a trade-off between model efficiency (inference time) and overall performance.
- **Optimal Configuration:**
  - Based on the evaluation loss and other metrics, the model with a batch size of 4, 8-bit quantization, and a learning rate of  $5e-5$  appears to provide the best balance between performance and time efficiency.
- **Lower Learning Rate Leads to Lower Performance:**
  - The model with a learning rate of  $3e-5$  had the lowest scores across most metrics, indicating that a lower learning rate might not be suitable for this task.

## 5.3 Graphs of training loss vs validation loss for different models

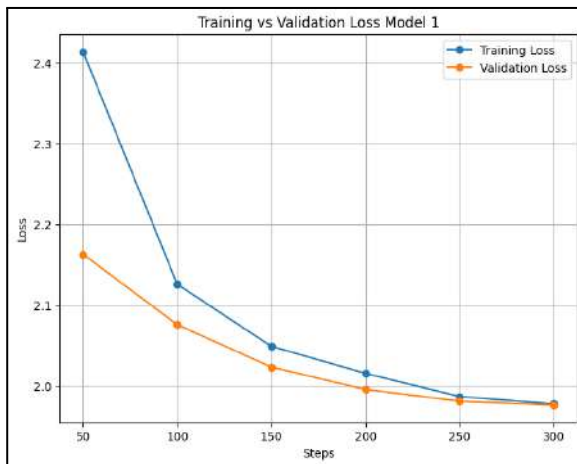


Figure 5.2.a: Training Loss of Model 1

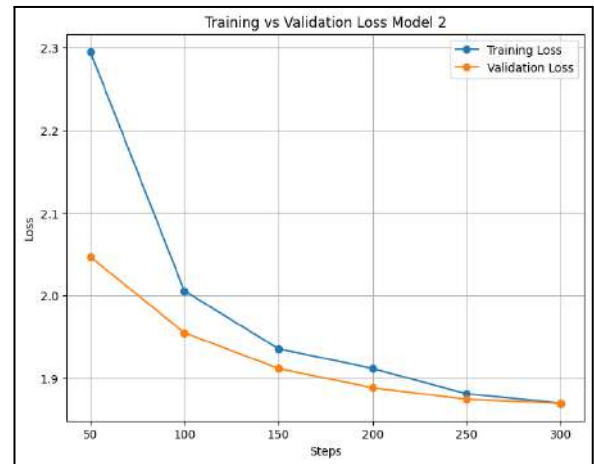


Figure 5.2.b: Training Loss of Model 2

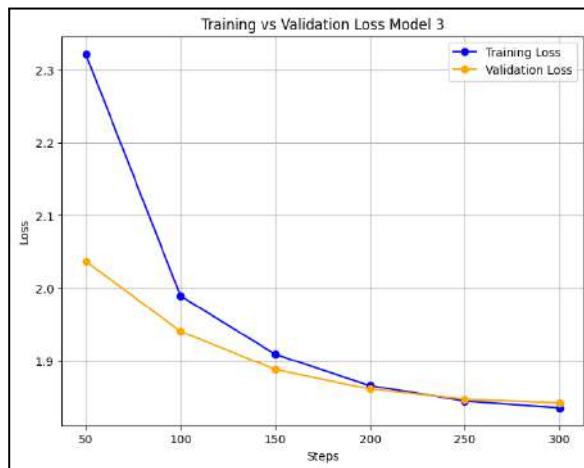


Figure 5.2.c: Training Loss of Model 3

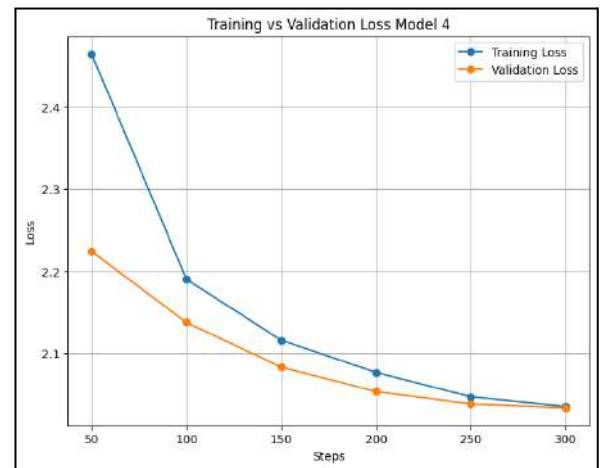


Figure 5.2.d: Training Loss of Model 4

The plots visually represent the relationship between the training and validation loss across different training steps. Analyzing such plots is crucial for understanding model performance. Ideally, both training and validation loss should decrease over time. The graphs indicate that training and validation loss both reduce over time, meaning that the model learnt and understood the data well.

# Chapter VI: Plan Of Action For the Next Semester

## 6.1 Work done till date

This section outlines the key tasks completed so far, including the fine-tuning of the Llama model and the development of the dashboard UI with core functionalities.

- **Fine-tuning Llama Model using QLoRA and PEFT**

We have successfully fine-tuned the Llama model using QLoRA (Quantized Low-Rank Adaptation) and PEFT (Parameter-Efficient Fine-Tuning). These methods have allowed us to improve the model's capability to extract relevant insights and summarize key points during career counseling sessions efficiently. The focus was on optimizing performance while maintaining low resource consumption, ensuring the model is both effective and scalable.

- **UI for Dashboard**

The user interface for the dashboard has been designed and implemented with various features for an enhanced user experience:

- **Side Bar:** This section allows users to manage their profiles, view meeting details, and navigate between key sections of the system.
- **Home Page:** Includes a search functionality where users can search keywords within the meeting transcript. Key features on the home page include:
- **Ask Anything Search Box:** Users can ask questions related to the transcript, and the results are displayed on the result page.
- **Transcript Display:** The entire meeting transcript is displayed, searchable by keywords.
- **Search History:** A log of previous queries for easy reference.
- **Speaker Stats:** Provides insights into speaker activity during the meeting, including participation data.
- **Comments and Notes:** Users can leave comments or take notes during or after the meeting.
- **Result Page:** Displays results for queries made through the "Ask Anything" search box, providing detailed responses based on the transcript.
- **Functionalities:** Users can download or share both the full meeting transcript and the search results, making it easy to review and distribute key information.

## 6.2 Plan of action for project II

In the next phase of the CareerLens project, the following key tasks are planned:

- **Implementation of Live Transcription**

We will focus on enabling real-time transcription during career counseling sessions. This feature will allow the system to capture and process spoken content instantly, providing users with live transcripts as the session progresses. This will improve the system's responsiveness and enhance the user experience by offering immediate access to the session data.

- **Deployment of Fine-Tuned Llama Model**

The fine-tuned Llama model will be deployed into the production environment, integrating it with the system to generate real-time insights, summaries, and responses during live sessions. This model will help in extracting key points from the transcript and providing immediate feedback to users, further enhancing the system's interactive features.

- **Improving Dashboard-Backend Interaction**

The next step involves optimizing the interaction between the dashboard and the backend model, ensuring a smooth and efficient data flow. This will involve refining how the dashboard retrieves insights, visualizations, and results from the model, making the system more responsive to user queries and actions. Ensuring real-time communication between the UI and the model is crucial for user experience.

- **Multilingual Support**

We plan to incorporate multilingual capabilities to broaden the system's usability. This will allow transcription and analysis of career counseling sessions conducted in different languages, making CareerLens accessible to a wider audience. Supporting multiple languages will ensure that the system can cater to diverse user groups, providing tailored insights in various linguistic contexts.

## **Chapter VII: Conclusion**

The CareerLens project effectively integrates career counseling techniques with innovative technology to give users a more engaging and informative experience. The technology enables live data collecting that improves engagement among attendees during meetings by recording interactions in real time. While the LLM module uses advanced AI capabilities to analyze transcripts, extract important insights, and deliver relevant responses to participant questions, the Transcription module makes sure that every session is accurately recorded.

The integration of a Chat Interaction feature facilitates working together by allowing users to quickly obtain clarity and guidance on all kinds of topics discussed. In addition, the post-meeting Dashboard provides a thorough recap of the discussion, complete with actionable insights, extensive summaries, and visualizations that enable users to thoughtfully consider their career alternatives.

All things considered, CareerLens increases user experience and speeds up the counseling process, making career guidance more effective and accessible. Through the use of these advanced technologies, the project helps individuals make better decisions and, in the end, helps them pursue their careers with more confidence.

## Chapter VIII: References

- [1] Martin Thomas Falk and Eva Hagsten. 2021. When international academic conferences go virtual. *Scientometrics* 126, 1 (01 Jan 2021), 707–724. <https://doi.org/10.1007/s11192-020-03754-5>
- [2] Paris V. Stefanoudis, Leann M. Biancani, Sergio Cambronero-Solano, Malcolm R. Clark, Jonathan T. Copley, Erin Easton, Franziska Elmer, Steven H. D. Haddock, Santiago Herrera, Ilysa S. Iglesias, Andrea M. Quattrini, Julia Sigwart, Chris Yesson, and Adrian G. Glover. 2021. Moving conferences online: lessons learned from an international virtual meeting. *Proceedings of the Royal Society B: Biological Sciences* 288, 1961 (2021), 20211769. <https://doi.org/10.1098/rspb.2021.1769>
- [3] Jay Peters. [n. d.]. Google’s Meet teleconferencing service now adding about 3 million users per day — theverge.com. <https://www.theverge.com/2020/4/28/21240434/google-meet-three-million-users-per-day-pichai-earnings>. [Accessed 14-10-2024]
- [4] Tom Warren. [n. d.]. Zoom grows to 300 million meeting participants despite security backlash — theverge.com. <https://www.theverge.com/2020/4/23/21232401/zoom-300-million-users-growth-coronavirus-pandemic-security-privacy-concerns-response> [Accessed 14-10-2024].
- [5] Nima Sadri, Bohan Zhang, and Bihan Liu. 2021. MeetSum: Transforming Meeting Transcript Summarization using Transformers! arXiv:2108.06310 [cs.CL] <https://arxiv.org/abs/2108.06310>
- [6] Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan TN. 2023. Building Real-World Meeting Summarization Systems using Large Language Models: A Practical Perspective. arXiv:2310.19233 [cs.CL] <https://arxiv.org/abs/2310.19233>
- [7] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Fine Tuning of Quantized LLMs. arXiv:2305.14314 [cs.LG] <https://arxiv.org/abs/2305.14314>
- [8] Fei Ge. 2024. Fine-tune Whisper and transformer large language model for meeting summarization. Ph. D. Dissertation. UCLA.
- [9] M. Wyawahare, M. Shelke, S. Bhorge and R. Agrawal, "AI Powered Multilingual Meeting Summarization," 2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2024, pp. 86-91, doi: 10.1109/Confluence60223.2024.10463307.
- [10] Lakshmi Prasanna Kumar and Arman Kabiri. 2022. Meeting Summarization: A Survey of the State of the Art. arXiv:2212.08206 [cs.CL] <https://arxiv.org/abs/2212.08206>
- [11] S. S. Bhat, U. Ahmed Nawaz, S. M, N. Tantri and V. Vasudevan, "Jotter: An Approach to Summarize the Formal Online Meeting," 2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIE), Ballari, India, 2023, pp. 1-6, doi: 10.1109/AIKIE60097.2023.10390455.
- [12] Aseem Srivastava, Tharun Suresh, Sarah P. Lord, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Counseling Summarization Using Mental Health Knowledge Guided Utterance Filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*. Association for Computing Machinery, New York, NY, USA, 3920–3930. <https://doi.org/10.1145/3534678.3539187>

# Chapter IX: Appendix

## 9.1 List Of Figures

| Sr. no. | Figure No. | Name of the figure            | Page No. |
|---------|------------|-------------------------------|----------|
| 1.      | 1.6        | Methodology Employed          | 3        |
| 2.      | 4.1        | Block diagram of the system   | 10       |
| 3.      | 4.2        | Modular diagram of the system | 11       |
| 4.      | 4.3.a.a    | Level 0 DFD                   | 12       |
| 5.      | 4.3.a.b    | Level 1 DFD                   | 12       |
| 6.      | 4.3.a.c    | Level 2 DFD                   | 13       |
| 7.      | 4.3.b      | Flowchart of the system       | 14       |
| 8.      | 4.3.c.1    | Transcript Viewing Dashboard  | 16       |
| 9.      | 4.3.c.2    | Meet Analysis                 | 17       |
| 10.     | 4.3.c.3    | QnA                           | 17       |
| 11.     | 5.2.a      | Training Loss of Model 1      | 19       |
| 12.     | 5.2.b      | Training Loss of Model 2      | 19       |
| 13.     | 5.2.c      | Training Loss of Model 3      | 20       |
| 14.     | 5.2.d      | Training Loss of Model 4      | 20       |

## 9.2 List Of Tables

| Sr. no. | Table No. | Name of the Table                | Page No. |
|---------|-----------|----------------------------------|----------|
| 1.      | 3.4       | Hardware & Software Requirements | 7        |
| 2.      | 5.1       | Evaluation Results               | 18       |
| 3.      | 5.2       | Evaluation metrics used          | 18       |



### 9.3 Xerox of project review sheet

3

**Industry/Inhouse:** \_\_\_\_\_ **Project Evaluation Sheet 2024-25** **Class: D17B**

Title of Project(Group no): CareerLens : Meet Summarizer

Group Members: Deven Bhagatani (5), Piyush Chyaya (10), Sakshi Kirmath (24), Manraj Singh Viridi (63)

|                           | Engineering Concepts & Knowledge                                    | Interpretation of Problem & Analysis | Design / Prototype | Interpretation of Data & Dataset | Modern Tool Usage | Societal Benefit, Safety Consideration | Environment Friendly | Ethics | Team work | Presentation Skills | Applied Engg & Mgmt principles | Life-long learning | Professional Skills | Innovative Approach | Total Marks |
|---------------------------|---|--------------------------------------|--------------------|----------------------------------|-------------------|--|----------------------|--------|-----------|---------------------|--------------------------------|--------------------|---------------------|---------------------|-------------|
|                           | (5)   | (5)                                  | (5)                | (3)                              | (5)               | (2)                                    | (2)                  | (2)    | (2)       | (3)                 | (3)                            | (3)                | (5)                 | (5)                 | (50)        |
| Review of Project Stage 1 | 5   | 4                                    | 5                  | 2                                | 4                 | 2                                      | 2                    | 2      | 2         | 3                   | 2                              | 3                  | 4                   | 4                   | 44          |
| Comments:                 | 1) Develop Agent Framework<br>2) Setup data corpus for fine tuning. |                                      |                    |                                  |                   |  |                      |        |           |                     |                                |                    |                     |                     |             |

Dr. Nupur G.  Name & Signature Reviewer1

|                           | Engineering Concepts & Knowledge                 | Interpretation of Problem & Analysis | Design / Prototype | Interpretation of Data & Dataset | Modern Tool Usage | Societal Benefit, Safety Consideration | Environment Friendly | Ethics | Team work | Presentation Skills | Applied Engg & Mgmt principles | Life-long learning | Professional Skills | Innovative Approach | Total Marks |
|---------------------------|--|--------------------------------------|--------------------|----------------------------------|-------------------|--|----------------------|--------|-----------|---------------------|--------------------------------|--------------------|---------------------|---------------------|-------------|
|                           | (5)  | (5)                                  | (5)                | (3)                              | (5)               | (2)                                    | (2)                  | (2)    | (2)       | (3)                 | (3)                            | (3)                | (5)                 | (5)                 | (50)        |
| Review of Project Stage 1 | 5  | 4                                    | 5                  | 2                                | 4                 | 2                                      | 2                    | 2      | 2         | 3                   | 2                              | 3                  | 4                   | 4                   | 44          |
| Comments:                 | Change straight forward pipeline to agent based. |                                      |                    |                                  |                   |  |                      |        |           |                     |                                |                    |                     |                     |             |

Sanjay M.  Name & Signature Reviewer2

Date: 23rd August, 2024

03

**Industry/Inhouse:** \_\_\_\_\_ **Project Evaluation Sheet 2024-25** **Class: D17B**

Title of Project(Group no): CareerLens : Meet Summarizer

Group Members: Deven Bhagatani (5), Piyush Chyaya (10), Sakshi Kirmath (24), Manraj Singh Viridi (63)

|                           | Engineering Concepts & Knowledge  | Interpretation of Problem & Analysis | Design / Prototype | Interpretation of Data & Dataset | Modern Tool Usage | Societal Benefit, Safety Consideration | Environment Friendly | Ethics | Team work | Presentation Skills | Applied Engg & Mgmt principles | Life-long learning | Professional Skills | Innovative Approach | Total Marks |
|---------------------------|---|--------------------------------------|--------------------|----------------------------------|-------------------|--|----------------------|--------|-----------|---------------------|--------------------------------|--------------------|---------------------|---------------------|-------------|
|                           | (5)   | (5)                                  | (5)                | (3)                              | (5)               | (2)                                    | (2)                  | (2)    | (2)       | (3)                 | (3)                            | (3)                | (5)                 | (5)                 | (50)        |
| Review of Project Stage 1 | 5   | 4                                    | 4                  | 3                                | 5                 | 2                                      | 2                    | 2      | 2         | 3                   | 2                              | 3                  | 3+1=4               | 3                   | 44          |
| Comments:                 | Consider NVIDIA GPU's<br>Different models with different parameters to be considered. |                                      |                    |                                  |                   |  |                      |        |           |                     |                                |                    |                     |                     |             |

Sanjay M.  Name & Signature Reviewer1

|                           | Engineering Concepts & Knowledge   | Interpretation of Problem & Analysis | Design / Prototype | Interpretation of Data & Dataset | Modern Tool Usage | Societal Benefit, Safety Consideration | Environment Friendly | Ethics | Team work | Presentation Skills | Applied Engg & Mgmt principles | Life-long learning | Professional Skills | Innovative Approach | Total Marks |
|---------------------------|--|--------------------------------------|--------------------|----------------------------------|-------------------|--|----------------------|--------|-----------|---------------------|--------------------------------|--------------------|---------------------|---------------------|-------------|
|                           | (5)  | (5)                                  | (5)                | (3)                              | (5)               | (2)                                    | (2)                  | (2)    | (2)       | (3)                 | (3)                            | (3)                | (5)                 | (5)                 | (50)        |
| Review of Project Stage 1 | 5  | 4                                    | 4                  | 3                                | 5                 | 2                                      | 2                    | 2      | 2         | 3                   | 2                              | 3                  | 4                   | 3                   | 44          |
| Comments:                 | Evaluation parameters too to be made.<br>UI with working prototype<br>Quantization - 8 bits. |                                      |                    |                                  |                   |  |                      |        |           |                     |                                |                    |                     |                     |             |

Dr. Nupur G.  Name & Signature Reviewer2

Date: 26th September, 2024

# CareerLens

## Career Counseling Analysis with Open Source Llama 3.1

Dr. Nupur Giri  
nupur.giri@ves.ac.in  
Vivekanand Education Society's  
Institute of Technology  
Chembur, Mumbai, India

Manraj Singh Virdi  
d2021.manrajsingh.virdi@ves.ac.in  
Vivekanand Education Society's  
Institute of Technology  
Chembur, Mumbai, India

Sakshi Kirmathe  
d2021.sakshi.kirmathe@ves.ac.in  
Vivekanand Education Society's  
Institute of Technology  
Chembur, Mumbai, India

Deven Bhagtani  
d2021.deven.bhagtani@ves.ac.in  
Vivekanand Education Society's  
Institute of Technology  
Chembur, Mumbai, India

Piyush Chugeja  
d2021.piyush.chugeja@ves.ac.in  
Vivekanand Education Society's  
Institute of Technology  
Chembur, Mumbai, India

### ABSTRACT

This paper details the development and tuning of a multilingual meeting analysis tool capable of live transcription, summarization, and critical insights generation from meeting discussions. Based on Meta's LLaMA 3.1 this tool aims for producing concise summaries, important action items, and contextual insights right after the meetings conducted with users. The JSON-formatted output created by the model ensures easy integration with the user dashboard. Configured with an array of quantization and batch size parameters, the system has been fine-tuned for various deployment contexts. Furthermore, the tool facilitates user engagement across multiple languages by analyzing dialogues in real-time and offering tailored insights via a multilingual, user-centric dashboard. This paper additionally examines the underlying architecture, training setups, and comparative outcomes of distinct model configurations to promote efficiency and reduce the occurrence of hallucinations.

### KEYWORDS

multilingual analysis, real-time transcription, meeting summarization, llama fine-tuning

## 1 INTRODUCTION

Due to the increase in remote work during 2020-2021, online meetings and conferences have become more widespread than ever before [4]. People worldwide are now preferring online platforms for professional interactions as streaming digital content has become very accessible [13]. Major platforms like Zoom and Google Meet report between 3 to 300 million daily participants in online meetings [10, 14], which shows the significant trend towards virtual communication. Therefore, virtual meetings have become indispensable for conferences, educational institutions, corporate functions, and numerous counseling relationships.

Career counseling, in particular, is important in helping individuals navigate professional life and the processes associated with job searches. Regardless of the particular field of focus, counseling generally involves professionals who offer specialized guidance

and support to clients facing challenging decisions. Career counseling often encompasses topics such as career development, career changes, professional growth, and related issues [1]. These sessions must be tailored to the individual, requiring accurate records so that both counselor and client can refer back to the discussions and interactions. However, traditional methods of recording and documenting sessions are time-consuming and may lack accuracy, establishing the need for more efficient solutions [11]. While counselors and clients may record and transcribe sessions for future reference, these transcripts are often huge, making it time-consuming to analyze in detail. Condensing transcripts into brief summaries helps to extract insights, highlight actionable items for clients, and save time for both parties when reviewing material.

Such summarization can be achieved through two basic methods: extractive and abstractive. Extractive summarization assigns weights to different parts of the source material and selects the highest-weighted sections to generate a summary. In contrast, abstractive summarization reformulates and generates information to generate a summary that reflects the main elements of the content in a more comprehensible manner. Abstractive summarization is particularly useful for meeting transcripts, as it can provide an integrated summary that highlights all critical details of the conversation. Due to the sensitive nature of career counseling sessions, large datasets containing meeting transcripts and summaries are often limited. This data scarcity encourages the use of pre-trained large language models (LLMs) which enhance summarization efficiency and streamline workflows across a wide range of applications [5, 9]. By leveraging these advanced models, it is possible to create tools that transcribe and summarize career counseling sessions, thereby helping both clients and counselors derive actionable insights to support their respective professional journeys.

## 2 LITERATURE REVIEW

The domain of meeting summarization has gained pace significantly due to the high usage of virtual communication technologies. Numerous methodologies have been proposed to address the issue of summarizing meeting transcripts, and each methodology identifies its merits and demerits. Nonetheless, there is much still lacking for many systems that are currently applied, such as content dealing

with multilingual content, summarizing in real-time, and accuracy for complex conversational contexts.

New innovations have resulted in the establishment of even more sophisticated techniques like AI-based multilingual summarization systems [15], tasked with the ever-increasing needs of running multilingual meetings. Such systems utilize Latent Semantic Analysis (LSA) besides advanced NLP techniques to establish important points and action items. The method, however, tends to over-simplify the complex conversational interactions because of the inherent use of LSA-like techniques for dimensionality reduction. Transformer models, on the other hand, can recognize complex relations within dialogue without trading off contextual richness. The summaries are thus concise yet comprehensive in multiple languages.

Some of the works focus on abstractive summarization using Transformer-based architectures [11]. For example, the use of Pointer Generator Networks with coverage mechanisms has significantly improved conventional models in terms of reducing word repetition and enhancing the readability of generated summaries. However, such methods often rely on extensive datasets for training, such as news summaries, which can be limiting when applied to more domain-specific contexts like career counseling. Unlike these approaches, our methodology fine-tunes Meta’s LLaMA 3.1 on a focused dataset, ensuring that the model adapts to the subtleties of career counseling conversations and reduces hallucinations or the generation of unrelated content.

To further enhance our model’s efficiency, we leverage QLoRA for fine-tuning, a method that significantly reduces memory usage while maintaining performance on large-scale models [3]. This allows us to fine-tune LLaMA on a 4-bit quantized pretrained model, facilitating the efficient handling of a specialized dataset for career counseling. Moreover, this approach ensures scalability, enabling even resource-constrained systems to handle models with billions of parameters. Furthermore, we prioritize the use of LLaMA 3.1 over other closed-source models such as GPT-4 due to cost and privacy considerations. Open-source models like LLaMA provide competitive performance, even in zero-shot settings, and avoid the privacy risks and high costs associated with API-based fine-tuning of closed models[9]. Our approach is designed to strike a balance between performance, privacy, and efficiency, ensuring a practical solution for real-world career counseling summarization applications.

Hybridization of extractive and abstractive techniques, as seen in Jotter [2] hybridizes the techniques to gain another promising avenue. By combining BERT embeddings with sequence-to-sequence models, Jotter is able to balance efficiency with extractive accuracy and human-like readability of abstractive summaries. This two-tiered approach excels at coherence but may introduce computational overhead, limiting its use in real time.

A thorough analysis by Kumar and Kabiri [8] points out the key challenges in the meeting summarization domain, specifically the challenge of extracting relevant information from large dialogue datasets. It also intimates domain-specific models and evaluation metrics because most of the approaches currently in use are benchmarked on generic datasets like AMI and ICSI. Our work directly addresses that gap because career counseling involves sensitive and domain-specific information requiring special summarization

techniques. The fine-tuned models will carry customized evaluation metrics considering accuracy and actionable wisdom, rather than mere ROUGE or BLEU scores.

In more specialized domains, like mental health counseling, one can see that there are even systems, like ConSum[12], implementing filtering and structuring dialogues based on domain knowledge techniques. The use of PHQ-9 by ConSum for utterance filtering showcases the possibility of externalized knowledge in the summarization process. Motivated by this concept, our system integrates user roles and speaker information directly within the model’s output. This integration facilitates role-aware insights that improve the contextual significance of the summaries, rendering them more applicable in professional environments.

Thus our system builds on the foundation laid down by the earlier work for meeting summarization in which it offers multilingual capabilities, real-time summarization, and optimization tailored specifically for the job of career counseling. We employ transformer-based frameworks like MeetSum [11], but we add fine-tuned models that adapt to the unique framework of career counseling interactions. In addition, our focus on efficient deployment aligns with the real-time objectives emphasized in prior work, yet also resolves the existing challenges pertaining to computational burden and latency.

### 3 NOVELTY OF PROJECT

The CareerLens project offers an innovative method for improving career counseling by providing real-time transcription and multilingual summarization of online meetings, utilizing Meta’s fine-tuned LLaMA 3.1 model. In contrast to traditional approaches that usually depend on manual note-taking, which can result in errors, CareerLens automates the documentation process, producing actionable insights and clear summaries that are readily available in various languages. This innovative tool not only captures key conversations but also features a user-friendly dashboard that enables counselors and clients to interact with the content in a dynamic way. By combining advanced NLP techniques with a commitment to enhancing accessibility in career guidance, CareerLens marks a step forward in the field, tackling the challenges presented by the increasing use of virtual counseling sessions.

### 4 PROPOSED METHODOLOGY

As shown Figure 1, our methodology aims to fine-tune Meta’s LLaMA 3.1 model using the Huuuyeah MeetingBank dataset for the purpose of generating multilingual meeting summaries and actionable insights. The architecture is designed to transcribe and summarize real-time meeting dialogues, delivering structured outputs in JSON format. This section outlines the key stages of the methodology, from dataset preprocessing and model fine-tuning to evaluation.

#### 4.1 Dataset Preprocessing

The Huuuyeah MeetingBank dataset [7] is utilized for fine-tuning, which includes meeting transcripts from a variety of real-world conversations. The dataset is divided into a training set and a validation set. Each transcript is tokenized using the LLaMA tokenizer with a maximum sequence length of 512 tokens. A specific JSON structure

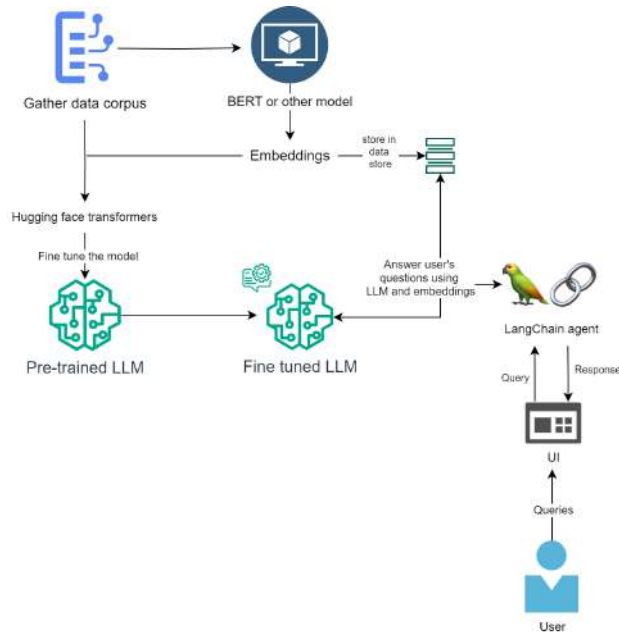


Figure 1: Fine Tuning Pipeline

is employed to format the model’s output, consisting of the meeting ID, summary, action items, insights, and speaker information.

To maintain efficiency, we employ prompt engineering techniques to guide the model in generating JSON-formatted outputs directly from the meeting transcripts. The prompts are crafted to ensure the model generates structured outputs only when valid transcription content is present, minimizing the risk of hallucinations, a common issue in large language models as noted in previous research [3, 9].

## 4.2 Model Fine-Tuning

The LLaMA 3.1 model is loaded with a 4-bit quantization configuration using BitsAndBytes [3] to optimize computational efficiency while maintaining model performance. The quantization type nf4 and bf16 are used to ensure compatibility with the NVIDIA CUDA environment.

Model fine-tuning is conducted using LoRA (Low-Rank Adaptation) [6] to reduce memory overhead while preserving accuracy in language modeling tasks. We configure the LoRA parameters with  $r = 8$ ,  $\alpha = 16$ , and a dropout rate of 0.05 for causal language modeling tasks, which has shown to be effective in several studies [3]. The model’s forward projection layers (q\_proj, k\_proj, v\_proj, o\_proj) and LM head are fine-tuned specifically to adapt to the characteristics of the meeting data.

## 4.3 Training and Evaluation

The fine-tuning process utilizes a batch size of 2 with gradient accumulation over 8 steps to manage GPU memory constraints. The training process runs for a total of 300 steps with a learning rate of  $3 \times 10^{-5}$ , leveraging the Paged AdamW optimizer [3].

Evaluation is conducted using the validation set to assess the model’s ability to generalize across unseen meeting transcripts. The evaluation prompt instructs the model to return concise summaries in JSON format, including action items, insights, and speaker roles. To ensure reproducibility and proper quantization, model checkpoints are saved periodically, allowing for mid-training evaluations.

The model is assessed using quantitative metrics such as BLEU and ROUGE scores to measure summarization accuracy, and qualitative analysis is performed by human evaluators to ensure coherence and factual correctness [11, 12].

## 4.4 JSON Output and Dashboard Integration

The output generated by the fine-tuned model is designed in a JSON format, optimized for easy integration into a user-friendly dashboard. The JSON fields include the meeting ID, a summary of key discussion points, action items for participants, and insights derived from the conversation. Additionally, the model identifies and categorizes speakers based on their roles, which is essential for meeting analysis in the context of career counseling and other professional settings [9, 12].

This methodology ensures that the model is capable of handling multilingual meeting transcripts, generating concise and relevant summaries that can be presented in real-time on a user dashboard. The use of LoRA and quantization techniques further enhances the model’s efficiency, allowing for deployment in resource-constrained environments without compromising on performance [3, 6].

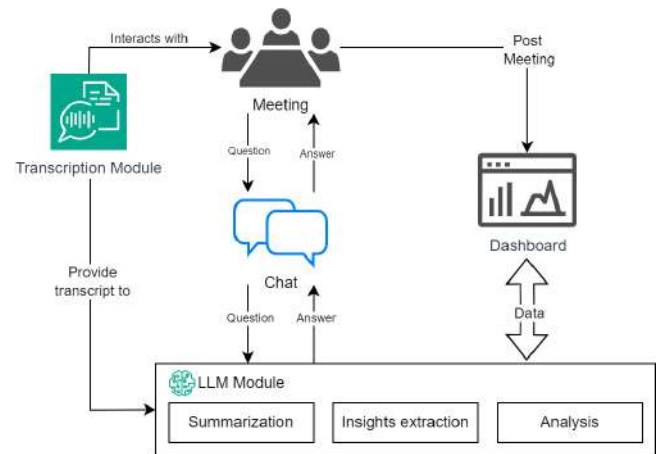


Figure 2: Flow of system

## 4.5 System Architecture

The overall system is integrated into a pipeline (2), where the transcription engine (powered by fine-tuned large language models such as LLaMA 3.1) captures live meeting dialogues. These dialogues are processed in real-time, with the model producing JSON-based summaries. The dashboard fetches these outputs and presents key insights, ensuring an interactive and multilingual user experience, a feature that is becoming increasingly essential for modern career counseling and other professional domains [11, 12].

**Table 1: Comparison of Model Configurations and Metrics**

| Model Configuration   | GLEU   | ROUGE  | BLEU   | BERT Score | Eval Loss | Inference Time |
|---|--------|--------|--------|------------|-----------|----------------|
| batch_size: 2<br>gradient_accumulation_steps: 8<br>bnb_4bit_quant_type: fp16<br>learning_rate: 5e-5 | 0.2054 | 0.3804 | 0.1299 | 0.9168     | 1.9762    | 24.3s          |
| batch_size: 4<br>gradient_accumulation_steps: 4<br>bnb_8bit_quant_type: fp16<br>learning_rate: 5e-5 | 0.2165 | 0.3899 | 0.1304 | 0.9151     | 1.8698    | 26s            |
| batch_size: 4<br>gradient_accumulation_steps: 8<br>bnb_4bit_quant_type: nf4<br>learning_rate: 4e-5  | 0.1788 | 0.3578 | 0.1073 | 0.9104     | 1.8417    | 20.67s         |
| batch_size: 4<br>gradient_accumulation_steps: 4<br>bnb_4bit_quant_type: fp16<br>learning_rate: 3e-5 | 0.0805 | 0.1880 | 0.0305 | 0.8624     | 2.0341    | 66s            |

## 5 OUTCOMES

The results of various model configurations, each tested on our dataset for real-time multilingual meeting analysis, are summarized in Table 1. The table presents the GLEU, ROUGE, BLEU, BERT Score, Evaluation Loss, and Inference Time for each configuration.

In the configuration with a batch size of 2 and gradient accumulation steps of 8 using 4-bit quantization with fp16, the system showed a solid GLEU score of 0.2054, a ROUGE score of 0.3804, and a BLEU score of 0.1299. This model configuration also demonstrated a competitive BERT score of 0.9168 and a relatively low evaluation loss of 1.9762, though the inference time was slightly longer at 24.3 seconds.

A configuration with a larger batch size of 4, using 8-bit quantization, achieved slightly better performance in terms of BLEU (0.1304) and a reduced evaluation loss of 1.8698. The GLEU and ROUGE scores also saw incremental improvements at 0.2165 and 0.3899, respectively.

For the model utilizing 4-bit nf4 quantization with a batch size of 4 and gradient accumulation steps of 8, the GLEU score dropped slightly to 0.1788, but the evaluation loss was reduced to 1.8417, which indicates better alignment between predicted and reference outputs. This model showed an inference time improvement as well, reducing it to 20.67 seconds.

Finally, the model with a batch size of 4 and gradient accumulation steps of 4 using 4-bit quantization with fp16 had a lower GLEU score (0.0805) and a higher evaluation loss (2.0341). Despite the model's lower scores, its inference time was significantly longer at 66 seconds, indicating potential inefficiencies in this configuration for real-time applications.

Overall, these results highlight the trade-offs between model performance and inference time, where larger batch sizes and finer quantization types can yield better accuracy and reduced evaluation loss, but at the cost of increased processing time.

## 6 CONCLUSION

In this work, we designed and tested a multilingual meeting analysis tool that is capable of transcribing, summarizing, and extracting insights from real-time meeting dialogues. We used the LLaMA 3.1 model fine-tuned with the MeetingBank dataset; we optimized the system using several different model configurations with the 4-bit quantization technique. The study highlights the applicability of the tool in generating concise and practical findings based on transcripts of meetings, which is a critical aspect in enriching documentation and follow-through actions concerning career counseling. Further, model output organization in JSON format makes the system especially suitable for smooth integration into dashboards whose orientation is user-centric.

## REFERENCES

- [1] [n. d.]. Career counseling - Infogalactic: the planetary knowledge core — infogalactic.com. [https://infogalactic.com/w/index.php?title=Career\\_counseling&oldid=722971075](https://infogalactic.com/w/index.php?title=Career_counseling&oldid=722971075). [Accessed 14-10-2024].
- [2] Sumedh S Bhat, Uzair Ahmed Nawaz, Sujay M, Nameesha Tantri, and Vani Vasudevan. 2023. Jotter: An Approach to Summarize the Formal Online Meeting. In *2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIE)*. 1–6. <https://doi.org/10.1109/AIKIE60097.2023.10390455>
- [3] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv:2305.14314 [cs.LG] <https://arxiv.org/abs/2305.14314>
- [4] Martin Thomas Falk and Eva Hagsten. 2021. When international academic conferences go virtual. *Scientometrics* 126, 1 (01 Jan 2021), 707–724. <https://doi.org/10.1007/s11192-020-03754-5>
- [5] Fei Ge. 2024. *Fine-tune Whisper and transformer large language model for meeting summarization*. Ph. D. Dissertation. UCLA.
- [6] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL] <https://arxiv.org/abs/2106.09685>
- [7] Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. MeetingBank: A Benchmark Dataset for Meeting Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 16409–16423. <https://doi.org/10.18653/v1/2023.acl-long.906>
- [8] Lakshmi Prasanna Kumar and Arman Kabiri. 2022. Meeting Summarization: A Survey of the State of the Art. arXiv:2212.08206 [cs.CL] <https://arxiv.org/abs/2212.08206>

- [9] Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan TN. 2023. Building Real-World Meeting Summarization Systems using Large Language Models: A Practical Perspective. arXiv:2310.19233 [cs.CL] <https://arxiv.org/abs/2310.19233>
- [10] Jay Peters. [n. d.]. Google's Meet teleconferencing service now adding about 3 million users per day — theverge.com. <https://www.theverge.com/2020/4/28/21240434/google-meet-three-million-users-per-day-pichai-earnings>. [Accessed 14-10-2024].
- [11] Nima Sadri, Bohan Zhang, and Bihan Liu. 2021. MeetSum: Transforming Meeting Transcript Summarization using Transformers! arXiv:2108.06310 [cs.CL] <https://arxiv.org/abs/2108.06310>
- [12] Aseem Srivastava, Tharun Suresh, Sarah P. Lord, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Counseling Summarization Using Mental Health Knowledge Guided Utterance Filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington DC, USA) (KDD '22). Association for Computing Machinery, New York, NY, USA, 3920–3930. <https://doi.org/10.1145/3534678.3539187>
- [13] Paris V. Stefanoudis, Leann M. Biancani, Sergio Cambroner-Solano, Malcolm R. Clark, Jonathan T. Copley, Erin Easton, Franziska Elmer, Steven H. D. Haddock, Santiago Herrera, Ilysa S. Iglesias, Andrea M. Quattrini, Julia Sigwart, Chris Yesson, and Adrian G. Glover. 2021. Moving conferences online: lessons learned from an international virtual meeting. *Proceedings of the Royal Society B: Biological Sciences* 288, 1961 (2021), 20211769. <https://doi.org/10.1098/rspb.2021.1769> arXiv:<https://royalsocietypublishing.org/doi/pdf/10.1098/rspb.2021.1769>
- [14] Tom Warren. [n. d.]. Zoom grows to 300 million meeting participants despite security backlash — theverge.com. <https://www.theverge.com/2020/4/23/21232401/zoom-300-million-users-growth-coronavirus-pandemic-security-privacy-concerns-response>. [Accessed 14-10-2024].
- [15] Medha Wyawahare, Madhuri Shelke, Siddharth Bhorge, and Rohit Agrawal. 2024. AI Powered Multilingual Meeting Summarization. In *2024 14th International Conference on Cloud Computing, Data Science Engineering (Confluence)*. 86–91. <https://doi.org/10.1109/Confluence60223.2024.10463307>