In22-S5-CS3501 Data Science and Engineering Project
Department of Computer Science and Engineering
University of Moratuwa

Youtube to Newsletter

Automated Summarization and Personalized Delivery of YouTube Video Content

Team Members:
220338N Kumara B. H. D. H
220352C Lakruwan R. W. S.
220358B Lakshani D. L. S.

Group ID: 16
Project ID : 4
Mentor: Dr. Thanuja Ambegoda
Teaching Assistant: Lasana Sanketh

1. Executive Summary:

This project proposes an automated AI-powered system that finds relevant videos and summarizes YouTube videos in the Artificial Intelligence domain. Leveraging expert YouTube channels , the system will extract video transcript or get transcript using YouTube API and apply large language models (LLMs) for summarization. These summaries are then compiled into structured , personalized newsletters delivered via email and more detailed summarised articles published on web-site.
The system uses an agentic, modular pipeline to automate video retrieval, transcription, summarization, delivery and publish the article. It reduces information overload, improves content accessibility, and helps professionals, researchers, and students stay updated efficiently.

2. Problem Statement:

The rapid growth of AI-related content on platforms like YouTube presents a significant challenge: valuable insights are often buried within lengthy videos, making it time-consuming for viewers to stay informed. Additionally, issues such as information overload, low content accessibility, and the absence of domain-specific filtering further hinder efficient knowledge consumption.
This project addresses these issues by filtering and summarizing relevant content from expert channels into concise, text-based newsletter, tailored to user preferences. It is useful for time constrained students, researchers and professionals finding updates in the rapidly growing AI field.

3. Data Description:

Data Sources: YouTube Data API, Whisper, LLM Summaries, User Input, User Interaction Logs.
Data Types:

- **Unstructured**
  - Raw video content , Video transcripts , Summaries
- **Structured**
  - Video metadata, User profile preferences , Engagement signals ,Personalized dataset cache

Key Features

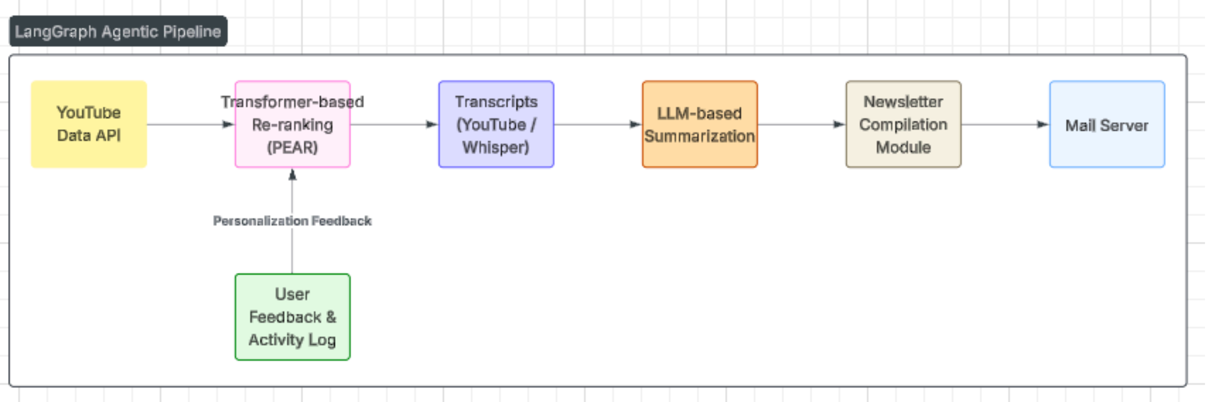
- Video title, description, publish date, channel name
- Transcript text, summary text
- View count, like count, comment count
- User interest tags (e.g., “LLMs,” “Computer Vision”)
- Engagement history per user
- Relevance scores for videos (from re-ranking model)

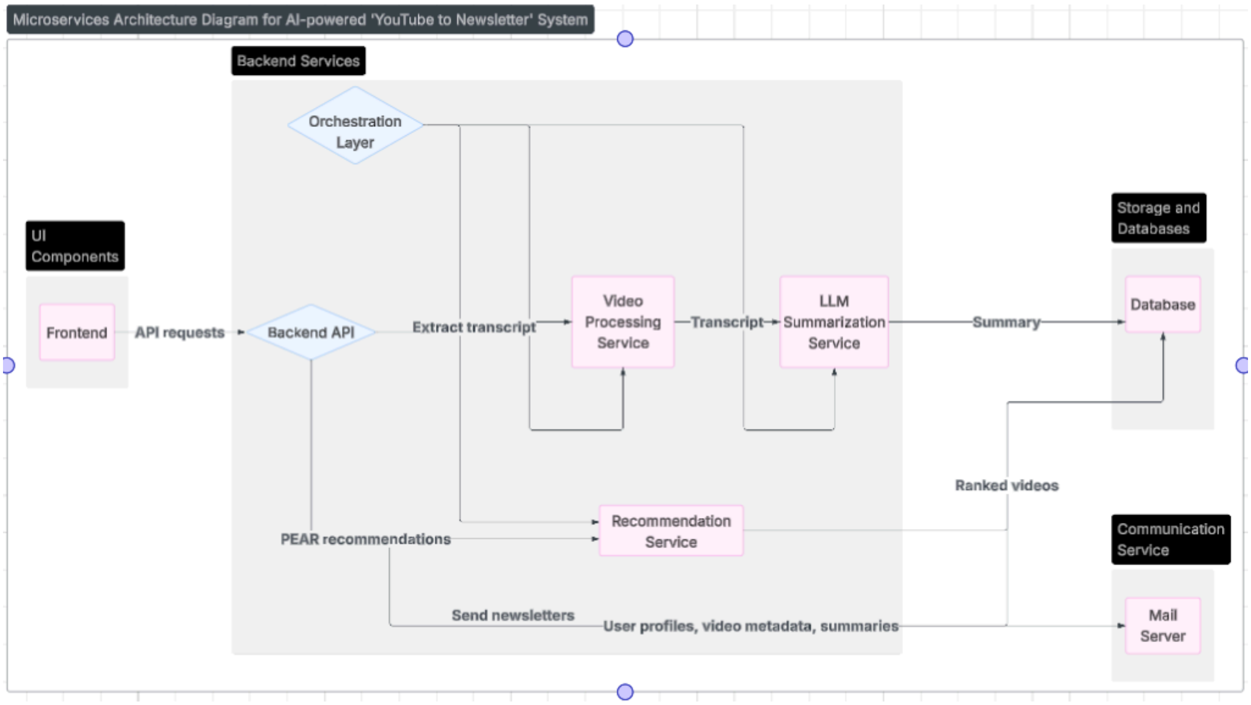
To solve the cold start problem, users first choose their favorite AI topics (like NLP or Robotics) when signing up. This helps create their initial profile. As they read newsletters and give feedback, we track their activity and build a personal dataset. Using a PEAR-style re-ranking system, we then recommend videos based on both overall quality and user preferences making the content more relevant and useful over time.

4. Methods:

We adopt a modular, agent-based workflow for automation:

- **Video Retrieval** - via YouTube Data API with keyword/channel filters.
- **Re-ranking & Recommendation** - Transformer-based video scoring model for relevance
- **Transcript Extraction** - YouTube captions or Whisper
- **Summarization** - Self-hosted LLMs for extractive summaries
- **Orchestration** - LangGraph agentic pipeline to manage workflow
- **Newsletter Delivery** - Mail server integration


LangGraph Agentic Pipeline
YouTube Data API
Transformer-based Re-ranking (PEAR)
Transcripts (YouTube / Whisper)
LLM-based Summarization
Newsletter Compilation Module
Mail Server
Personalization Feedback
User Feedback & Activity Log



5. Evaluation Plan

- **Transcript Accuracy** - Compare Whisper transcripts with actual captions
- **Summary Quality** - Human review, user feedback, ROUGE/BLEU metrics and LLM-as-a-judge method
- **Recommendation Accuracy** - Engagement score vs. user rating
- **Delivery System** - Success/failure rates in email delivery logs
- **Overall System** - The level of user interaction, failure recovery, and user satisfaction

6. Expected Outcomes and Success Criteria:

Expected outcomes:

- A functional end-to-end pipeline for generating AI newsletters from YouTube.
- A user interface to manage preferences, view transcripts, and provide feedback.
- Scalability and modularity to support future feature additions.

Success Criteria:

- Minimum 90% delivery success rate
- At least 75% of users rate summaries as relevant
- System runs autonomously with minimal manual intervention
- Deployment within 6–8 weeks with a working prototype

7. Division of work (individual responsibilities)

Kumara B.H.D.H**.**

- Development of the recommendation module.
- Integrate module.

Lakruwan R. W. S.

- Development of an LLM-based transcript generation module.
- Frontend Development.
- Generate a newsletter and manage the mail server.

Lakshani D. L. S.

- Development of an LLM-based transcript summarization module.
- Backend Development.

8. Preliminary Bibliography:

Li, Y., Zhu, J., Liu, W., Su, L., Cai, G., Zhang, Q., Tang, R., Xiao, X., & He, X. (2022). *PEAR: Personalized Re-ranking with Contextualized Transformer for Recommendation*. Tsinghua University & Huawei Noah’s Ark Lab. arXiv:2203.12267