



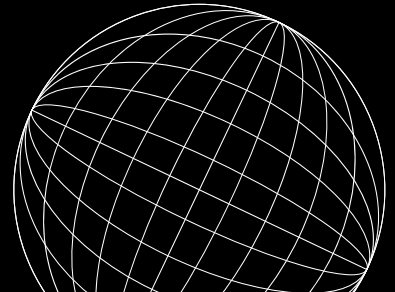
# **TLDW: Too Long Didn't Watch**

Theme: Digital Process Manual Creation

Team: HardToSpell

Member: Akash Paul

Email: [iakashpaul@gmail.com](mailto:iakashpaul@gmail.com)

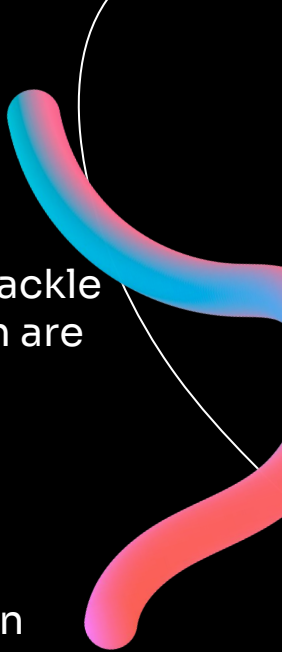
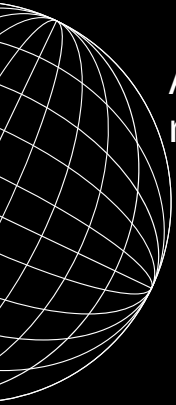


# Challenges with video-understanding

Similar to TLDR for lengthy blog posts. The most straightforward way to tackle the case of condensing instruction videos/manuals are figuring out which are your scene segments & to make the same relatable with your transcripts.

Here NLP techniques can help in identifying intent but mapping it with relevant scene may not be so straightforward.

A lot of content is also self referential so knowing the context also helps in making better linked manuals

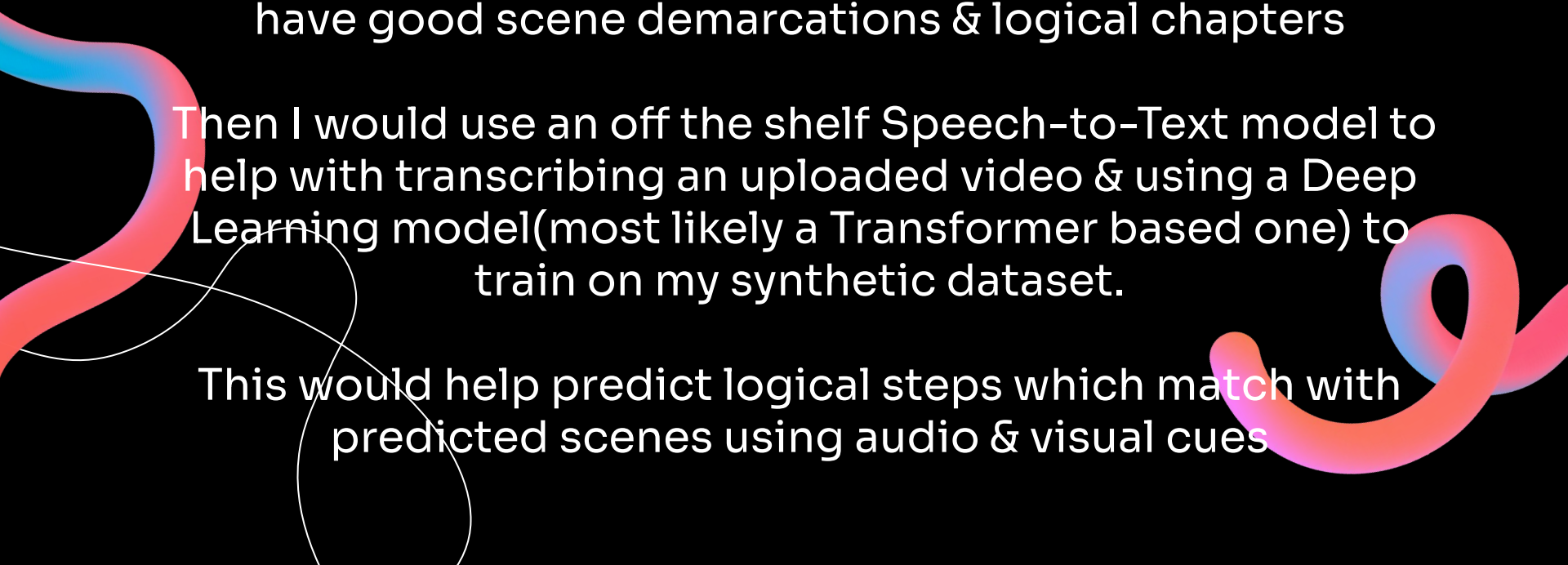
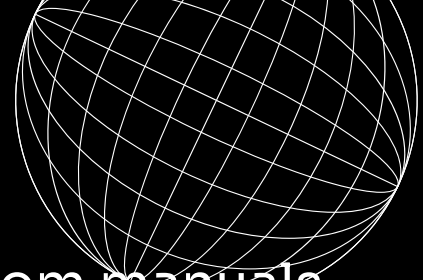


# Proposed Solution

My solution is to create a synthetic dataset from manuals on established YouTube channels of various topics, which have good scene demarcations & logical chapters

Then I would use an off the shelf Speech-to-Text model to help with transcribing an uploaded video & using a Deep Learning model (most likely a Transformer based one) to train on my synthetic dataset.

This would help predict logical steps which match with predicted scenes using audio & visual cues



# Planned Steps



## Dataset creation

A synthetic dataset will be created using opencv+FFMPEG for training a lightweight model to discern between NLP on transcripts & scene changes




## Model Training

Mostly in Python for the backend & use my own GPU for training



## Deployment

As a containerized web-application to output a word &/or Markdown document of all the logical steps that were detected with screen grabs to accompany it



**THANKS**