Methodology

Step 1: Sentence Splitting (Segmentation)

To process the transcription effectively, it's important to break it into individual sentences and allow for better handling and analysis of each part. This can be achieved by splitting the text based on punctuation marks such as full stops (".") or Sinhala-specific punctuation like " ।" or "?". I can use libraries such as nltk or spaCy for this task, or, if these libraries don't manage the punctuation well for Sinhala, I can implement custom rules for the language's structure.

Output: List of separate sentences.

- Example Input: "ඔන්න අද අපි වොකලට් පුඩින් එකක් කරන්න යන්නේ. පළවෙනිටම කෝන් ෆ්ලා ගන්න."
- Example Output: ["ඔන්න අද අපි වොකලට් පුඩින් එකක් කරන්න යන්නේ", "පළවෙනිටම කෝන් ෆ්ලා ගන්න"]

Step 2: Tokenization (Breaking into Words)

To classify a sentence effectively, it's important to identify specific words such as action verbs, measurements, and other key terms that help categorize its meaning. This can be achieved by splitting each sentence into individual words, either by using spaces or libraries like NLTK. However, for Sinhala, tokenization may require custom rules, as many tools struggle with non-Latin languages.

Output: A list of words per sentence.

- Example Input: "කෝන් ෆ්ලා එකතු කරන්න"
- Example Output: ["කෝන්", "ෆ්ලා", "එකතු", "කරන්න"]

Step 3: Stop Word Removal

To improve the focus on key terms, it's important to remove common, less meaningful words, also known as stop words. This can be done by using a Sinhala stop word list that includes words like "අපි", "එක", "හා", and "ද". While libraries like NLTK provide stop word tools, a custom stop word list specifically designed for Sinhala works more effectively and guarantees that only the most relevant words are retained for further processing.

Output: Cleaner, more meaningful words.

- Example Input: ["අපි", "කෝන්", "ෆ්ලා", "එකතු", "කරන්න"]
- Example Output: ["කෝන්", "ෆ්ලා", "එකතු", "කරන්න"]

Step 4: Feature Extraction (Identify Sentence Type)

1. Introduction/Context

- Sentences that greet the audience, set up the video, or explain what's going to happen should be removed, as they are not part of the key instructions and only serve to set the mood.
- Examples:
  - "ආයුබෝවන්! කොහොමද ඔයාලට?"
  - "ඔන්න අද මං ඔයාලට වොකලට් පුඩින් රෙසිපි එකක් කියලා දෙන්න හදන්නේ."
- Algorithm Idea: Look for friendly greetings, introductions, and phrases like "අද මං" (today I'm), "ඔන්න" (here we go), and questions like "කොහොමද?" (how are you?).

2. Ingredients/Materials

- Sentences listing the ingredients or tools required for the process should be kept, as they are essential for preparation and must be included in the performance guide.
- Examples:
  - "කෝන් ෆ්ලා එකතු කරගන්නවා ග්‍රෑම් තිහක්."
  - "වොකලට් පව්ඩර් ග්‍රෑම් 20කුත් එකතු කරගන්න."
- Algorithm Idea: Look for numbers with units like grams (ග්‍රෑම්), tablespoons (මේස හැඳි), cups (කෝප්ප), and mentions of food or tools.

3. Action/Instruction Steps (Key Points)

- These are the most important steps that tell you exactly what to do. They are the core of the visual guide and should be kept, as they provide the essential instructions needed to complete the process.
- Examples:
  - "කෝන් ෆ්ලා එක්ක කිරි හොඳින් කලවම් කරගන්න."
  - "ලිපේ තියාගත්තා."
- Algorithm Idea: Look for action words like "කරගන්න" (mix), "එකතු කරන්න" (add), "තියාගන්න" (place), "මිශ්‍ර කරන්න" (stir).

4. Explanations/Descriptions

- These provide extra details about texture, appearance, or progress. While useful, they are not usually key instructions, so they should generally be removed. However, if a description is required for recognizing a step's success, it can be kept.
- Examples:
    - “මිශ්‍රණය ටිකක් රත් වෙද්දි කැටි ගැහුනු ගතිය නැතිවෙනවා.”
    - “මේක ඉස්සෙල්ලට වඩා ගොඩක් ඝන වෙලා තියෙනවා.”
- Algorithm Idea: Look for descriptive words like “ඉස්සෙල්ලට වඩා ”

5. Tips/Alternatives

- Tips and alternative methods are optional suggestions. Since they aren't essential steps, they should be removed as they don't contribute to the core instructions.
- Examples:
    - “පැණි රස කන්න ආස කෙනෙක් නම් තව පොඩ්ඩක් වැඩිපුර එකතු කරගන්න.”
    - “මිල්ක් චොකලට් හරි ඩාක් චොකලට් කපලා එකතු කරන්න.”
- Algorithm Idea: Look for words like "කෙනෙක් නම්" (if you like), "තව" (more), "අන්තිමට" (finally).

6. Engagement/Call to Action

- Engagement prompts, such as requests for likes, comments, or subscriptions, are not related to the instructions. These should be removed, as they do not contribute to the process itself.
- Examples:
    - “කමෙන්ට් එකක් දාන් යන්න අමතක කරන්න එපා.”
    - “සබ්ස්ක්‍රයිබ් බට්න් එක ක්ලික් කරන්න.”
- Algorithm Idea: Look for phrases related to subscribing, liking, and commenting

Summary of Algorithms and What to Do

Category	Algorithm Idea	Keep/Remove

Introduction/Context	Identify greetings, intro phrases (“අද මං”, “ඔන්න”)	REMOVE
Ingredients/Materials	Detect numbers + units + item names (“ග්‍රෑම්”, “කෝප්ප”)	KEEP
Action/Instruction	Spot action words (“කරගන්න”, “එකතු කරන්න”, “තියාගන්න”)	KEEP
Explanations	Look for descriptive words (“රත්”, “ඝන”)	REMOVE (Mostly)
Tips/Alternatives	Catch optional language (“නම්”, “තව”, “මොකද නම්”)	REMOVE
Engagement/CTA	Find social media calls (“සබ්ස්ක්‍රයිබ්”, “කමෙන්ට්”)	REMOVE

Step 5: Classification and Tagging

In this step, each sentence in the transcription is categorized by assigning one of the following labels: INTRO, INGREDIENT, ACTION, DESCRIPTION, TIP, or ENGAGEMENT. The goal is to identify which category best represents the content of each sentence. After labeling, each sentence is also tagged with a decision to either "keep" or "remove" based on its relevance to the core instructions. This classification helps in determining which sentences will contribute to the final performance guide and which ones are not necessary.

Step 6: Sentence Filtering

After the sentences are classified and tagged, the next step is to filter out the ones that are not essential. Only the sentences tagged as INGREDIENT and ACTION are kept, as these are the most important for the performance guide. All other sentences, such as those classified as INTRO, DESCRIPTION, TIP, or ENGAGEMENT, are removed because they do not provide direct instructions or essential information for the process at hand. This ensures that the final guide focuses solely on the key instructions needed for the task

Step 7: Sentence Summarization (Combining and Simplifying)

In this step, combine multiple sentences about the same step into a short, clear instruction.

1. Identify sentences that talk about the same action or ingredient.

   Example: Sentences about adding cornflour and chocolate powder.

2. Extract key details like quantities and measurements.

   Example:"ග්‍රෑම් 30", "මේස හැඳි තුනක්"

3. Combine them into a single, simplified sentence.

   Example:"එයට කෝන් ෆ්ලා ග්‍රෑම් 30 (මේස හැඳි තුනක්) සහ චොකලට් පව්ඩර් ග්‍රෑම් 20 (මේස හැඳි දෙකහමාරක්) එකතු කරගන්න"

Step 8: Sinhala Writing Style Conversion

Sinhala spoken and written language styles are different. The written style is more formal and instructional.

1. Replace informal phrases like "කරගන්න" with formal equivalents like "කර ගත යුතුයි".
2. Change sentence structure to match written Sinhala norms.

   Example:

   Spoken: "එයට කෝන් ෆ්ලා එකතු කරගන්න"

   Written: "එයට කෝන් ෆ්ලා එකතු කළ යුතුයි"

Step 9: Final Cleaning and Formatting

In this step, make sure that the text has consistent spacing and punctuation throughout the document. This includes guaranteeing that there are no extra spaces between words, proper use of full stops, commas, and other punctuation marks. Additionally, the text should follow proper Sinhala grammar rules, such that sentence structures are correct, and the flow of language is natural. This helps improve readability and makes the instructions clear and professional.