Problem Definition

Problem Visual media such as comics are largely inaccessible for people who are blind or have vision impairments. Existing tools either (a) read plain text aloud or (b) describe images; few provide a unified, language-flexible experience tailored to the narrative structure of comics. Ensuring equitable access to this content is an ethical imperative.

Solution Enable a user to upload one or more comic pages and receive an accurate, engaging audio narration in their preferred language. This narration must faithfully convey the story, dialogue, and scene transitions of the comic.

Why it matters This unlocks independent access to cultural and informational content and supports inclusion for visually impaired readers.

Inputs Comic page(s) as images, user language preference, and possibly pacing/voice settings

Outputs Synthesized speech audio description of panels and dialogue in the user's language

Subproblems

1. Creating a simple interface for users to upload pictures of comics and specify their preferred output language
2. Creating an accurate and engaging text description of the comic without leaving out parts or confusing the order of cells. (Might be achieved by using a multimodal large language model)
3. Translating the text description into the user-preferred language. (Via a translation model, trained by our team members)
4. Creating speech based on the translated text description
5. Presenting the created speech in an intuitive way to the user

Objectives

- **Organisational objectives**
  - Make comics independently consumable via audio in each user's preferred language.
  - Increase inclusion and autonomy of visually impaired readers in everyday media use.
  - Operate sustainably and predictably so access remains reliably available.
- **Leading objectives**
  - Achieve strong user sentiment: clarity of narration, perceived faithfulness to the story, and overall satisfaction.
  - Minimize friction: most users succeed with "upload → play" without assistance.
- **User-oriented objectives**
  - **What:** Faithful narration that preserves correct panel order, speaker turns, and key sound effects.
  - **How fast:** End-to-end processing per page within an agreed target; smooth playback with minimal delays.
  - **How clear:** Natural, intelligible speech; concise, unambiguous descriptions with minimal omissions or confusion.
- **Component (model) objectives**
  - **Description generation:** Description coherently describes the commic with minimal hallucinations.
  - **Translation:** Preserve meaning, names, and tone in the user's language;
  - **Text-to-speech:** Audio is easy to understand.

Measurements / KPIs

- **Organisational objectives**
  - **Independent Access Rate** — % sessions where users complete an audio comic without assistance. (Target >= 90%)
  - **Service Reliability** — % audio plays without failure; (Target >= 99%)
- **Leading objectives**
  - **Sentiment Score** — average user rating for clarity + faithfulness (1–5). (Target >= 4.3/5)
  - **First-Try Success** — % users who complete "upload → play" on the first attempt. (Target >= 95%)
- **User-oriented objectives**
  - **Narrative Mistakes** — % panels in incorrect order or incorrect speaker attribution. (Target >= 95%)
  - **Processing Speed** — end-to-end time per page from upload to ready-to-play. (Target <= 10s/page)
  - **Speech Intelligibility** — listener rating of ease of understanding (1–5). (Target >= 4.0/5)
- **Component (model) objectives**
  - **Description Quality** — panels affected by omissions or hallucinations. (Target <= 5%)
  - **Translation Adequacy** — human adequacy score preserving meaning/names (1–5). (Target >= 4.0/5)