

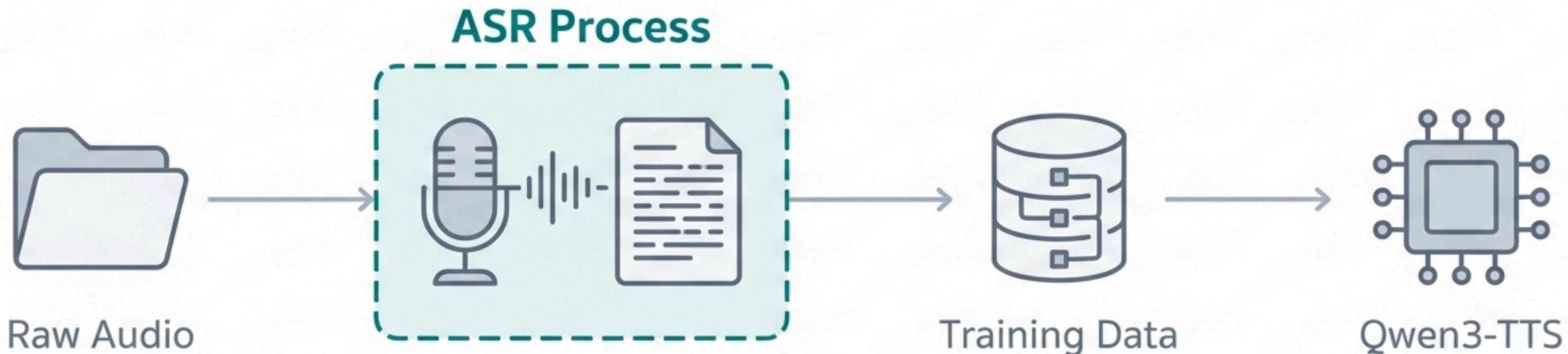
# รายงานสรุปโครงการ Fine-tuning Text-to-Speech ภาษาไทย

เจาะลึกกระบวนการสร้างโมเดล Qwen3-TTS และการเปรียบเทียบ ASR

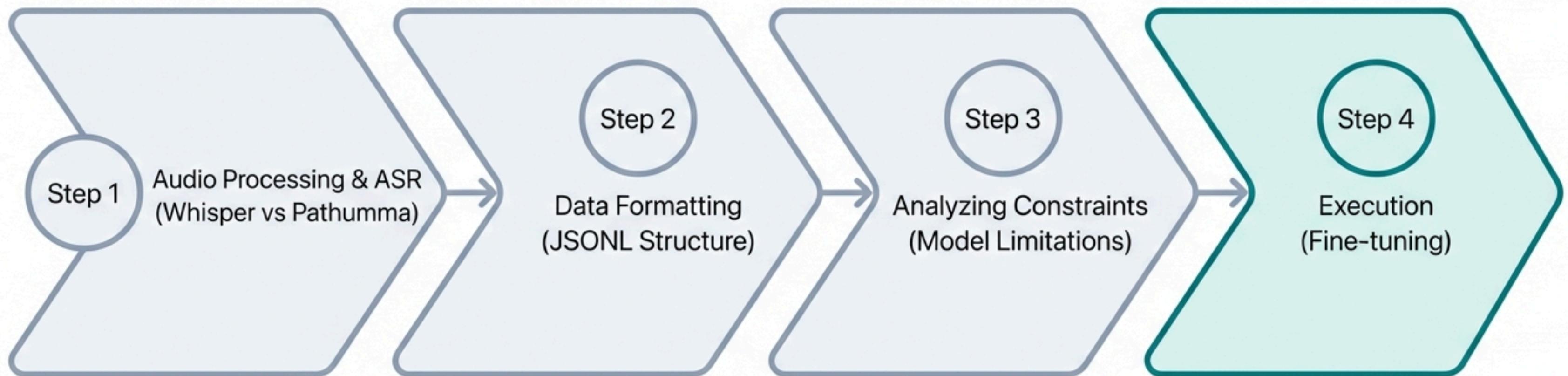


# ทีมฯและวัตถุประสงค์: การสร้างเสียงจากข้อมูลดิบ

- Objective: พัฒนาโมเดล Qwen3-TTS ให้สามารถสร้างเสียงภาษาไทยที่เป็นธรรมชาติ
- The Challenge: ข้อมูลต้นทางเป็นไฟล์เสียงจำนวนมาก แต่ **ไม่มี Transcript (บทพูด) กำกับ**
- The Solution: ต้องใช้กระบวนการ Speech-to-Text (ASR) เพื่อแปลงเสียงเป็นข้อความก่อน จึงจะนำไปฝึกโมเดลได้

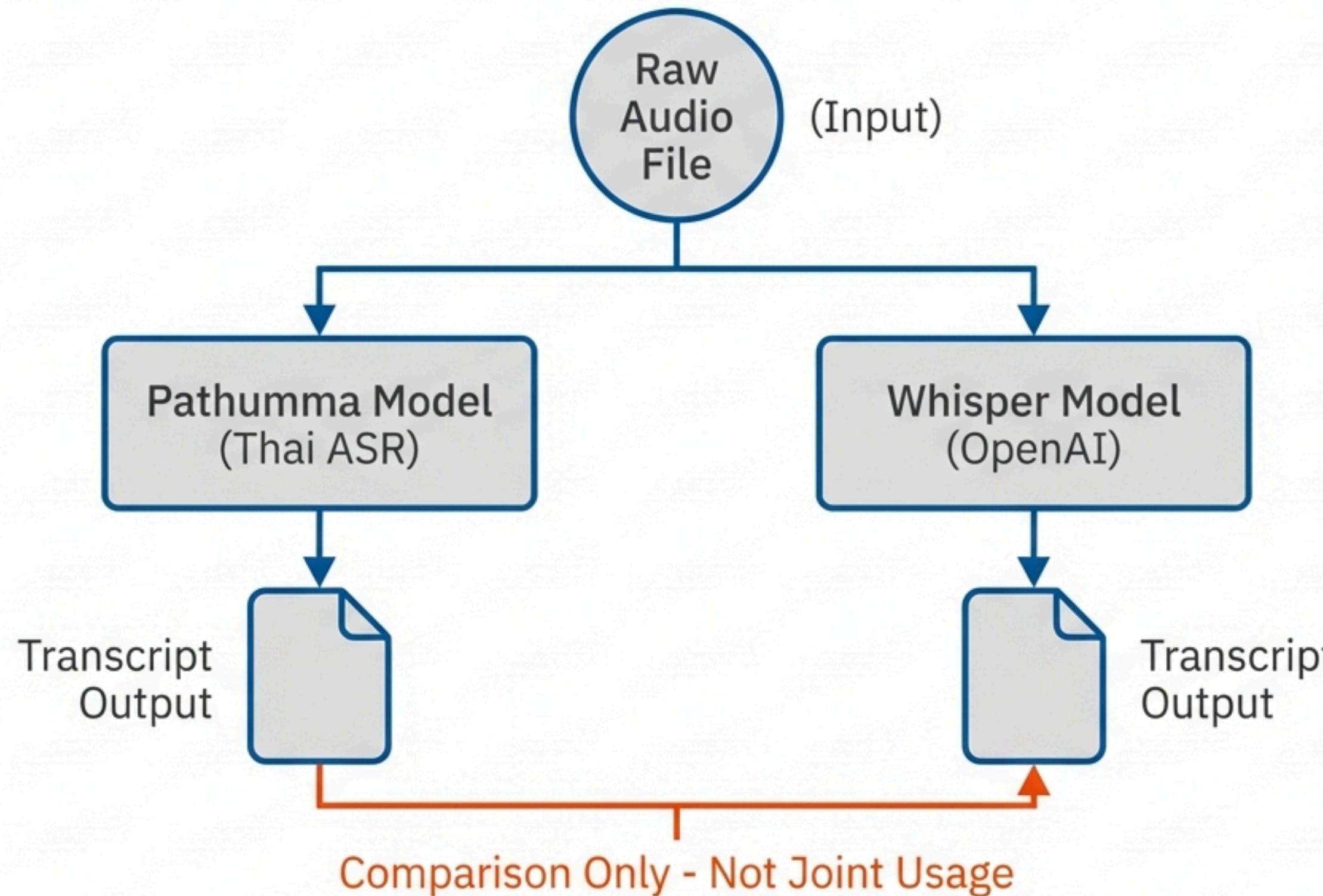


# ກາລົມການຄະດີ (Pipeline Overview)



## Step 1: การเลือกโมเดล ASR (Whisper vs. Pathumma)

เราสร้าง Dataset แยกกัน 2 ชุดเพื่อเปรียบเทียบคุณภาพ (ไม่ได้ใช้ร่วมกัน)



# Step 1: การเลือกโมเดล ASR (Whisper vs. Pathumma)

ทดสอบ 3 ภาษา: ไทย – อังกฤษ – เกาหลี



## Whisper Model

- **ภาษาไทย: ดีมาก**
  - รองรับสำเนียงภาษาไทยได้หลากหลาย
- **ภาษาเกาหลี: ดี**
  - รองรับภาษาเกาหลีได้ในระดับพื้นฐาน
- **ภาษาอังกฤษ: ดีมาก**
  - รองรับการสลับภาษา (ไทย–อังกฤษ–เกาหลี) ได้ดี
  - จัดการประโยคผสมหลายภาษา (Code-switching) ได้ดี

⚠️ จุดอ่อน: สะกดตามเสียงพูด ไม่ตรงมาตรฐาน และเสี่ยง Overfit กับคำ翦



## Pathumma Model

- **ภาษาไทย: ดีที่สุด**
  - ปรับแต่ง (Fine-tuned) มาเฉพาะภาษาไทยโดยตรง
  - ข้อมูลสะอาด ตรงมาตรฐานราชบัณฑิต
- **ภาษาเกาหลี: ไม่รองรับ**
  - ไม่มีการฝึกสอนด้วยข้อมูลภาษาเกาหลี
- **ภาษาอังกฤษ: พอดี**
  - เน้นเฉพาะภาษาไทย ความสามารถภาษาอังกฤษมีจำกัด
  - ประโยคผสมภาษาอังกฤษอาจมีความแม่นยำต่ำกว่า

✓ จุดแข็ง: สะกดถูกต้องตามมาตรฐาน ข้อมูลสะอาด เหมาะกับงานไทยโดยเฉพาะ

ภาษา / Language / 언어

Whisper Model

Pathumma Model

TH ไทย (Thai)

ดีมาก

รองรับสำเนียงหลากหลาย

ดีที่สุด

ปรับแต่งเฉพาะภาษาไทย

US อังกฤษ (English)

ดีมาก

รองรับการสลับภาษาได้ดี

พอดี

เน้นเฉพาะภาษาไทย

KR เกาหลี (Korean)

ดี

รองรับได้ในระดับพื้นฐาน

ไม่รองรับ

ไม่มีข้อมูลฝึกสอนภาษาเกาหลี

# ตัวอย่างข้อมูลจริง: การเปรียบเทียบ Output

Source Audio: data/SPEAKER\_06\_sent\_0000.wav

## Whisper Output

เราทำเต็มที่แล้ว เราภูมิใจในของเรา **อืม** แต่มันก็จะมีบางคอมเม้นท์...

## Pathumma Output

เราทำเต็มที่แล้ว เราภูมิใจในของเรา แต่�ันก็จะมีบางคอมเมนต์...

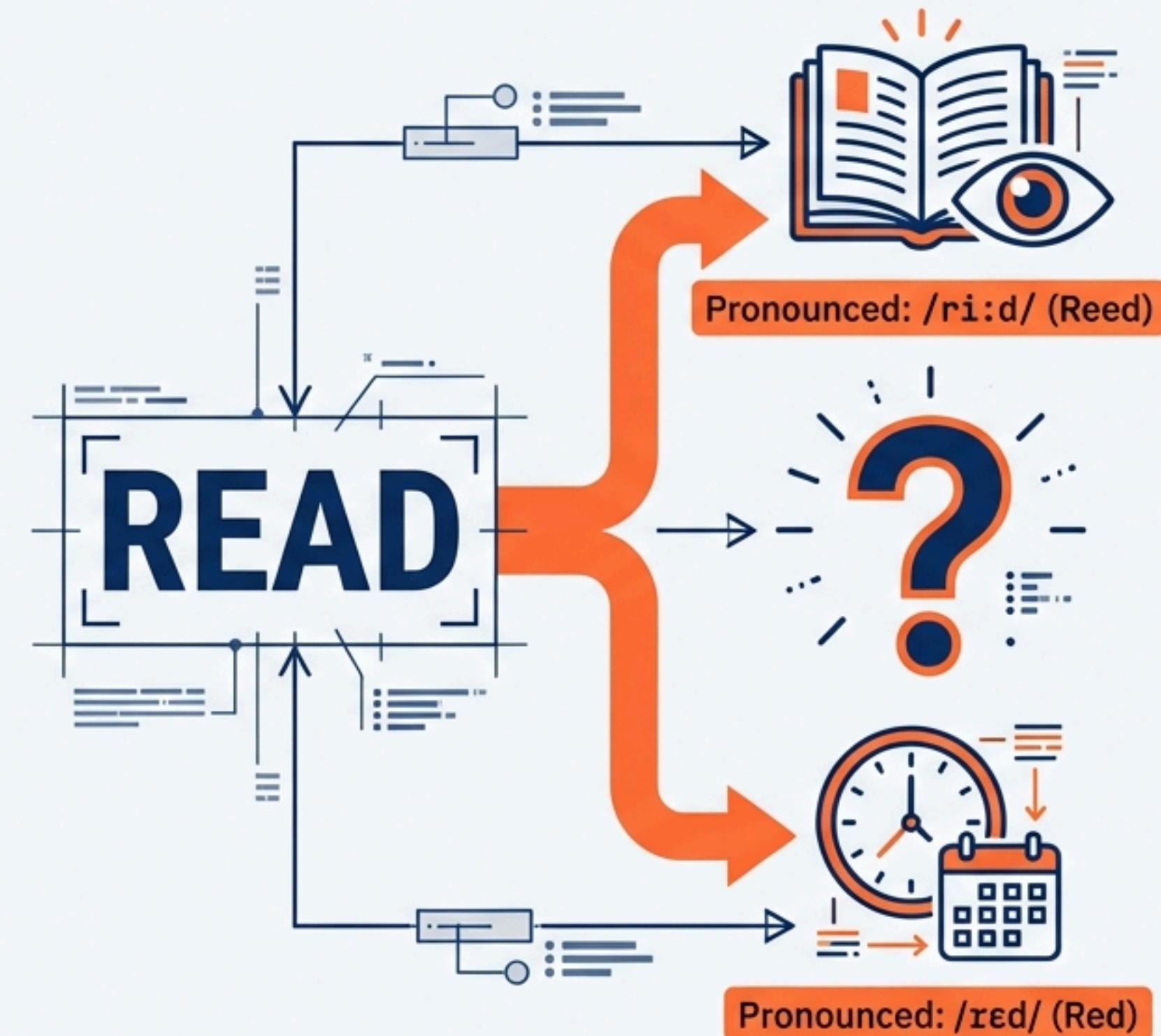
# กระบวนการประมวลผล Phoneme สำหรับ TTS Pipeline

เจาะลึกขั้นตอนการแปลงข้อมูล (Data Transformation) และการเพิ่มความทนทานของโมเดล (Robustness)



# วัตถุประสงค์ของการแปลง Grapheme เป็น Phoneme

- **Input:** ข้อความดิบ (Grapheme)
- **Problem:** ข้อความเขียนมีความ  
含混ในการอ่านเสียง (Ambiguity)
- **Solution:** สร้าง Input  
Representation ที่ระบุเสียงอ่าน  
ชัดเจนให้ TTS Model

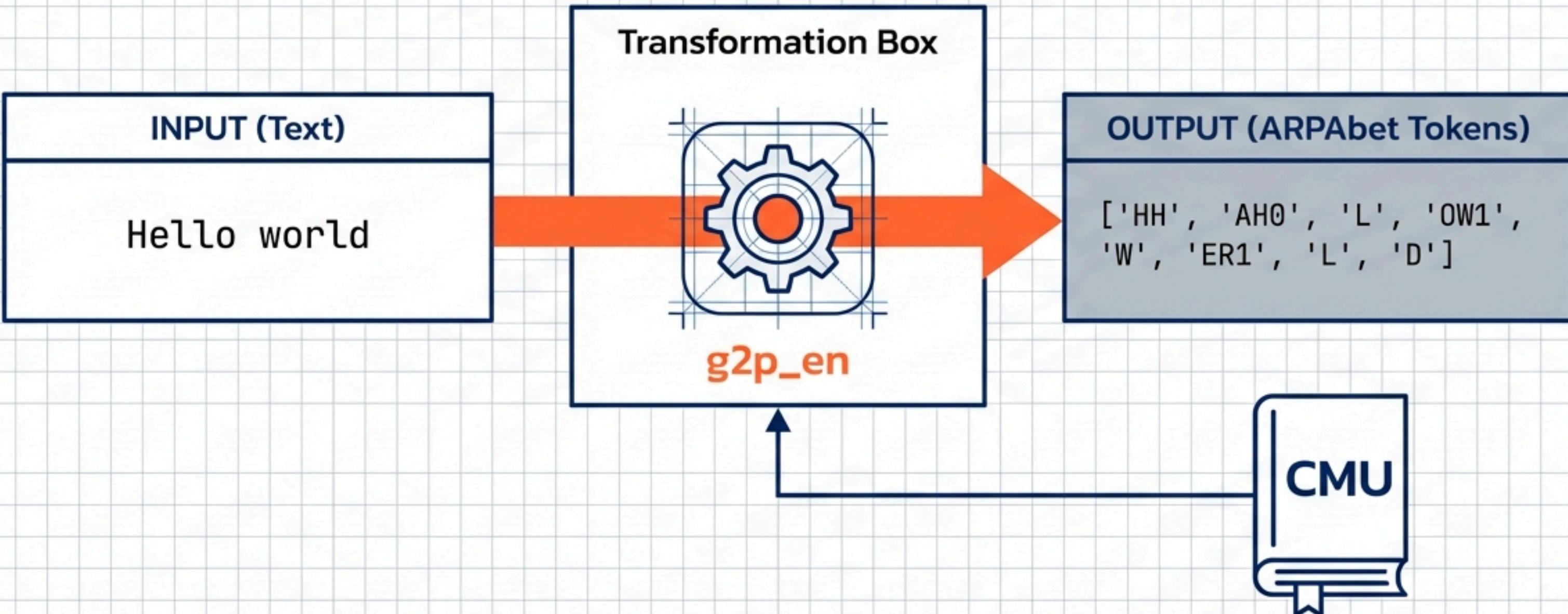


# 1.1 การแปลง G2P สำหรับภาษาอังกฤษ (English Phonemization)

Library: g2p\_en

กลไกการทำงาน: ใช้ระบบ Rule-based ร่วมกับพจนานุกรม CMU Dictionary

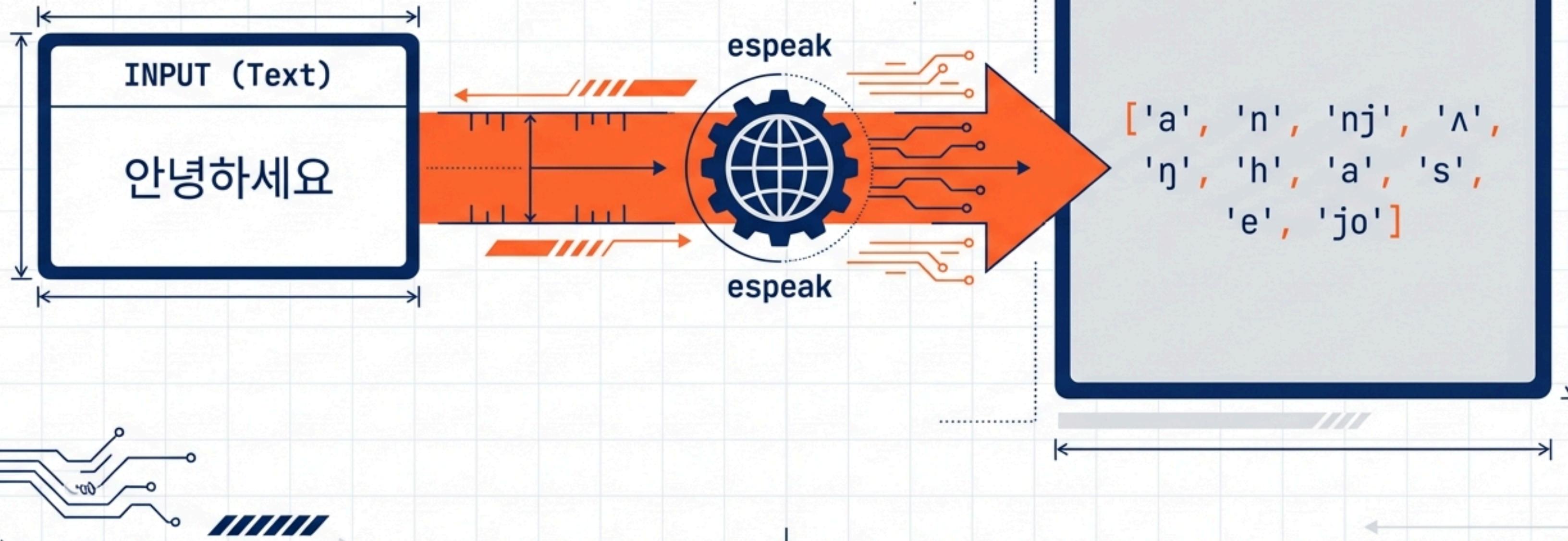
Format: ผลลัพธ์ที่ได้จะเป็น ARPAbet phoneme tokens



## 1.2 การแปลง G2P สำหรับภาษาเกาหลี (Korean Phonemization)

**Backend:** ใช้งาน espeak engine

**Format:** พาล์ปเป็นสัญลักษณ์สัต堪ศาสตร์ภาษา  
(IPA-based phoneme)



## 1.3 Silence Injection: การกำหนดขอบเขตเสียง

โมเดล TTS ต้องการ Boundary Control เพื่อความชัดเจนของการออกเสียง  
วิธีการ: แทรก Silence token ( '') ลงไว้ระหว่าง Phoneme tokens



Inject Silence



## 1.4 Sequence Padding: การปรับขนาดข้อมูลให้คงที่

**Neural Networks** ต้องการ Tensor ที่มีความยาวคงที่ (Fixed-length)

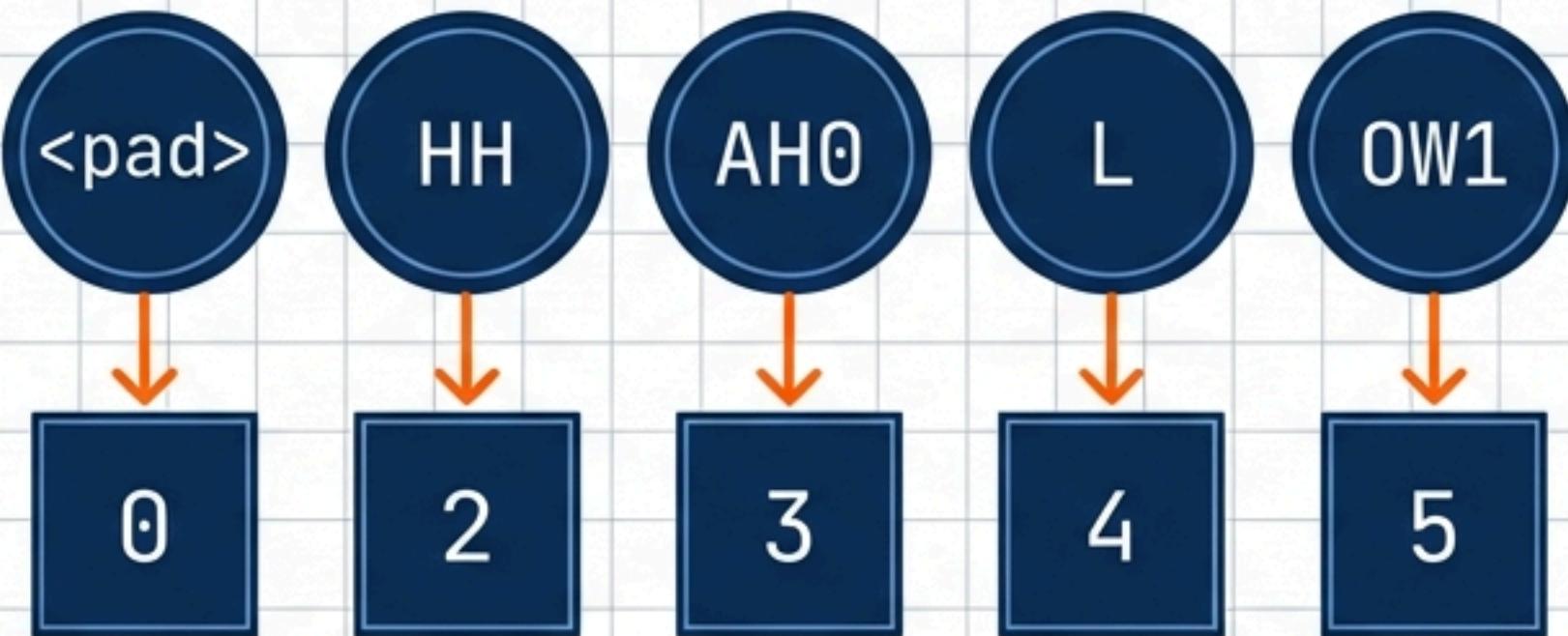
กำหนด  $\text{MAX\_LEN} = 100$

เงื่อนไข: หากข้อมูลสั้นกว่า 100 ให้เติม (Pad) ส่วนที่เหลือจนเต็ม



# 1.5 Token to Integer Mapping (Encoding)

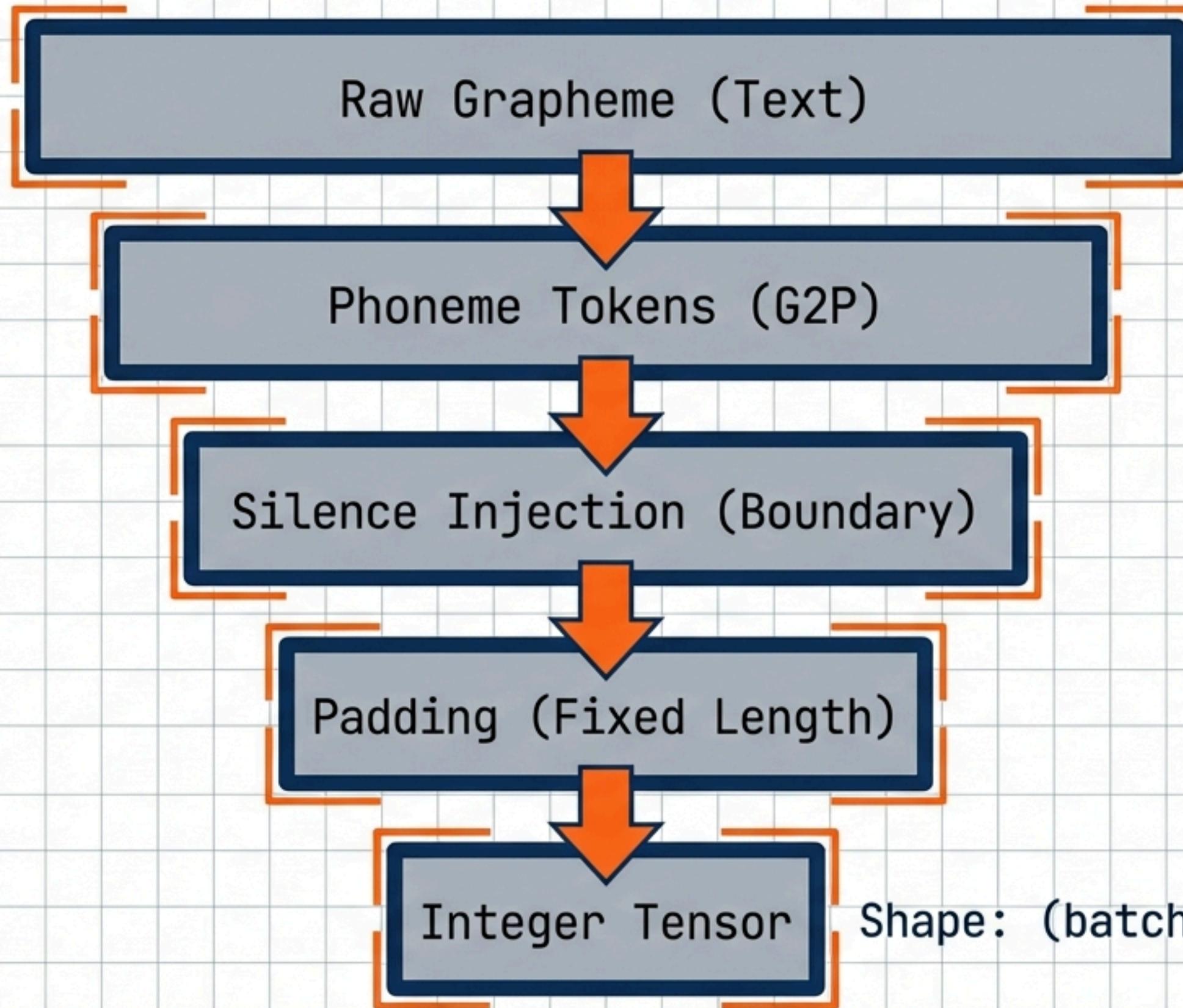
โมเดลไม่สามารถเข้าใจ String ได้โดยตรง ต้องแปลงเป็น Numeric ID  
กระบวนการ: Vocabulary Lookup



Input Tensor Result:  
[1, 2, 3, 4, 5, 1, 0, 0, 0, ...]

Ready for Embedding Layer

# สรุปกระบวนการ: จากข้อความสู่ Tensor (Technical Summary)





# Speech Processing Tools



**Speech to Text**  
Whisper / Pathumma



**แยก Speaker**  
pyannote.audio



**Clone เสียง**  
Fine-tune Qwen3 TTS



**แบ่งประโยค**  
snakers4/silero-vad



**ลด Noise**  
DeepFilterNet v3

01



**Download Video** FETCH SOURCE

ดาวน์โหลดไฟล์วิดีโอจาก URL หรือแหล่งข้อมูล เช่น YouTube, S3

02



**Convert** AUDIO EXTRACT

แปลงวิดีโอ → ไฟล์เสียง WAV / MP3 • ปรับ sample rate • mono/stereo

03



**Diarization** SPEAKER SPLIT

แยกเสียงตามผู้พูด (Speaker 1, 2, ...) • ระบุช่วงเวลาของแต่ละคน

04



**Summarize Speaker** PROFILE STATS

สรุปสถิติแต่ละ speaker • เวลา • จำนวน segment • ความยาวเสียง

05



**Export Preview** LISTEN & SELECT

ฟังเพื่อเลือก

Export ตัวอย่างเสียงแต่ละ speaker เพื่อฟังและตัดสินใจเลือก speaker ที่ต้องการ

06



**Get Sentence** SEGMENT SLICE

ตัดเสียงออกเป็น segment ระดับประโยค • align กับ timestamp

07



**Diarization + Select Speaker** FILTER TARGET

Diarize วีกครึ่งในระดับ segment • กรองเฉพาะ speaker ที่เลือกไว้

08



**Export Dataset** FINAL OUTPUT

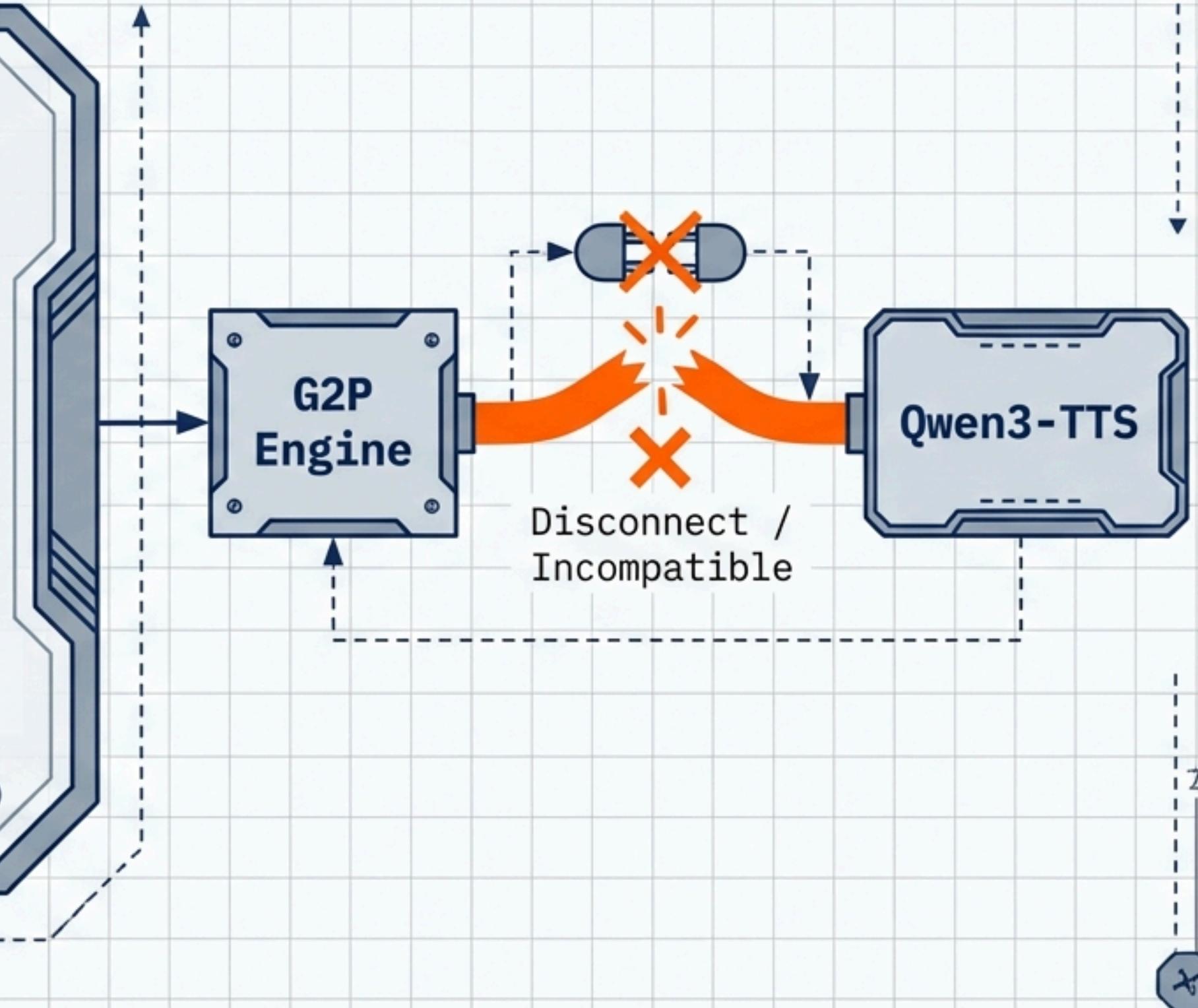
Export ไฟล์เสียง + transcript + metadata พร้อมใช้ train TTS / ASR

# ข้อจำกัดในทางปฏิบัติ: กรณีศึกษา Qwen3-TTS

**The Problem:** โมเดล Qwen3-TTS ไม่รองรับ Phoneme Input

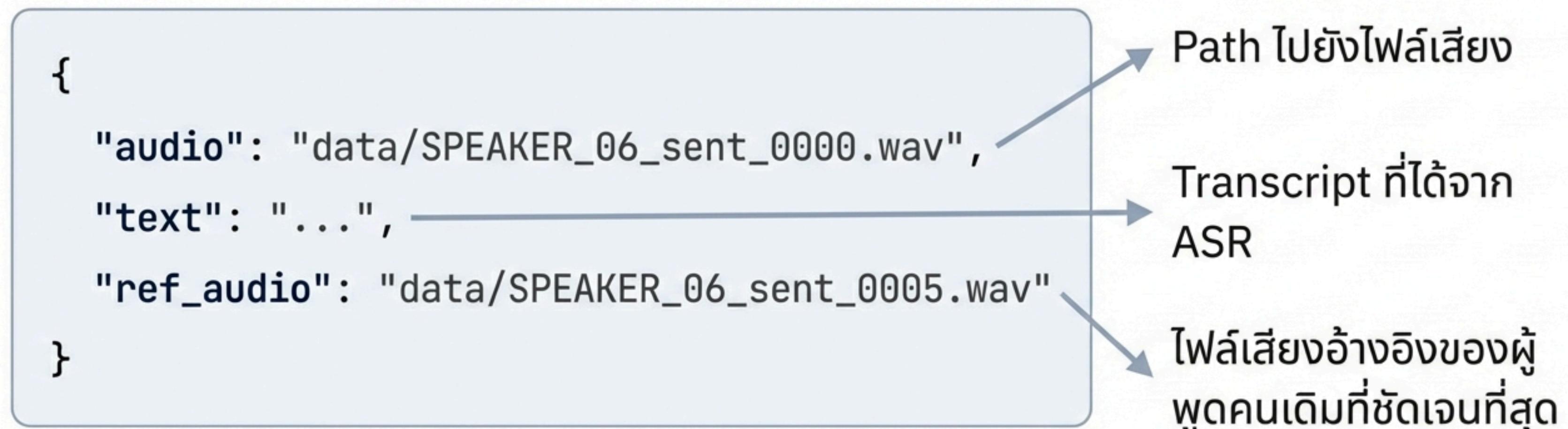
**Impact:** Pipeline ที่เตรียมไว้ (ใน 03\_phoneme.ipynb) ไม่สามารถนำมาใช้ Training ได้

**Consequence:** สูญเสียความสามารถในการควบคุมการอ่านเสียง (Pronunciation Control) อย่างละเอียด



## Step 2: โครงสร้างข้อมูล (Train Raw Format)

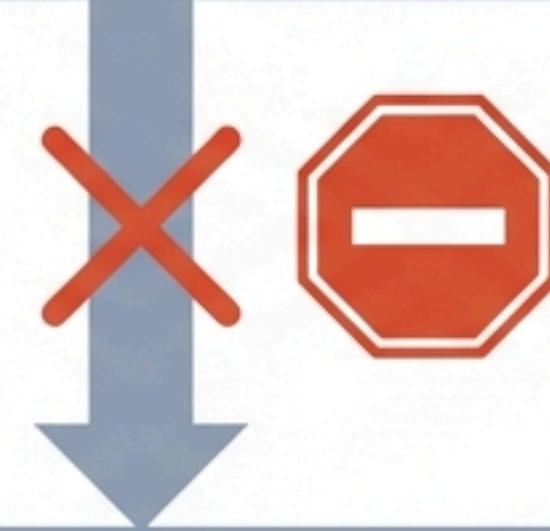
Format: **.jsonl**



# การจัดการระบุตัวตนผู้พูด (Speaker Identity Strategy)

แนวคิด: เพิ่ม Tag [Speaker Name] หน้าข้อความเพื่อระบุตัวตน

"[Lisa] เรากำเต็มที่แล้ว . . ."



Qwen3-TTS Model

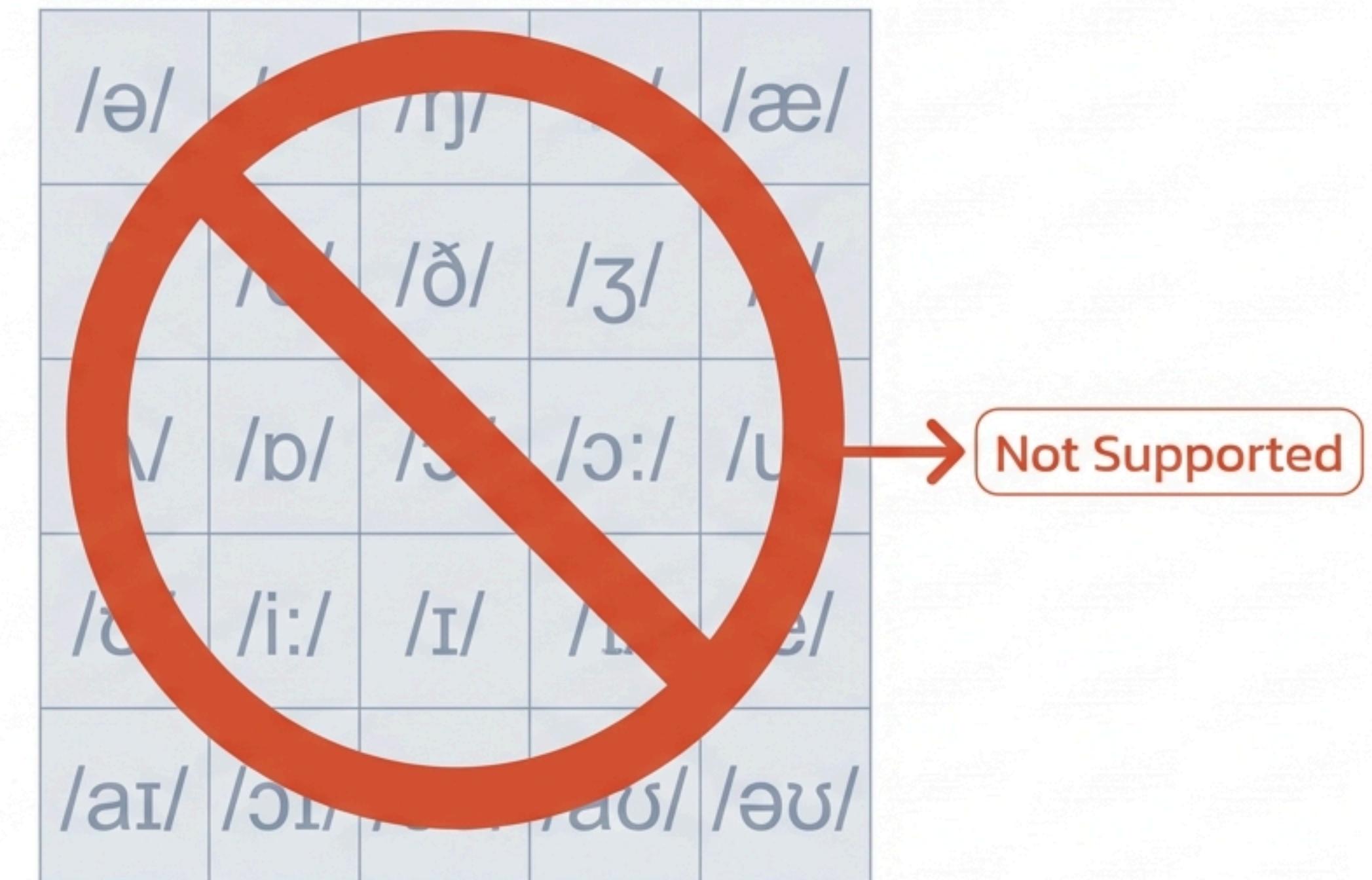


\*\*ข้อควรระวัง:\*\* แม้จะใส่ Tag ใน Field Text แต่ไม่สามารถ Qwen3-TTS ปัจจุบัน \*\*ไม่รองรับ Multi-speaker conditioning\*\*

# Step 3: ข้อจำกัดสำคัญ - ไม่รองรับ Phoneme

โมเดลรองรับเฉพาะ Raw Text  
Token เท่านั้น

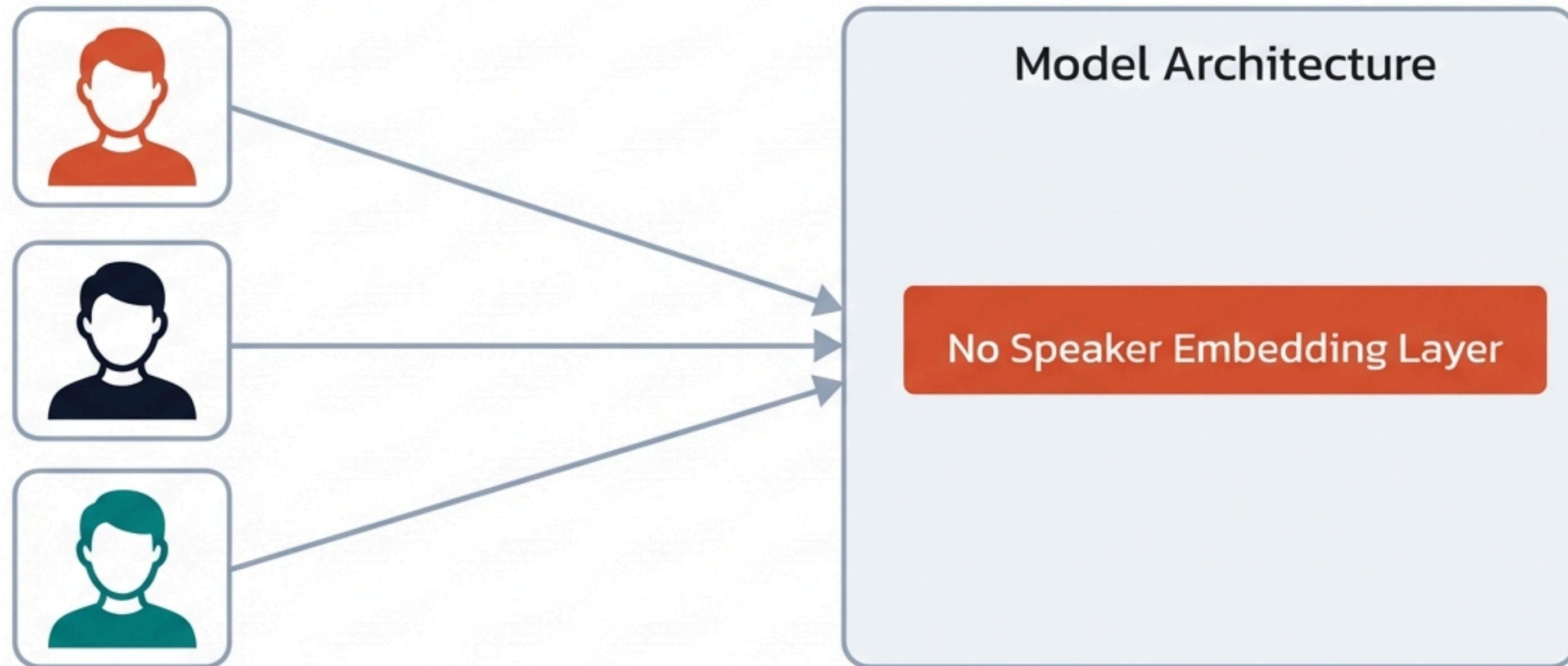
1. ไม่สามารถควบคุมการออกเสียง  
ระดับหน่วยเสียง (Phonetic  
Control)
2. หาก Transcript พิດ โมเดลจะ  
จำและเรียนรู้แบบพิດๆ กันที



# Step 3: ข้อจำกัดสำคัญ - ไม่รองรับ Multi-Speaker

โมเดลไม่มี Speaker Embedding Layer

โมเดลไม่รู้ว่า "คร" เป็นคนพูดในแต่ละไฟล์เสียง แม้ข้อมูล Input จะมาจากหลายคนก็ตาม



# Inference Data



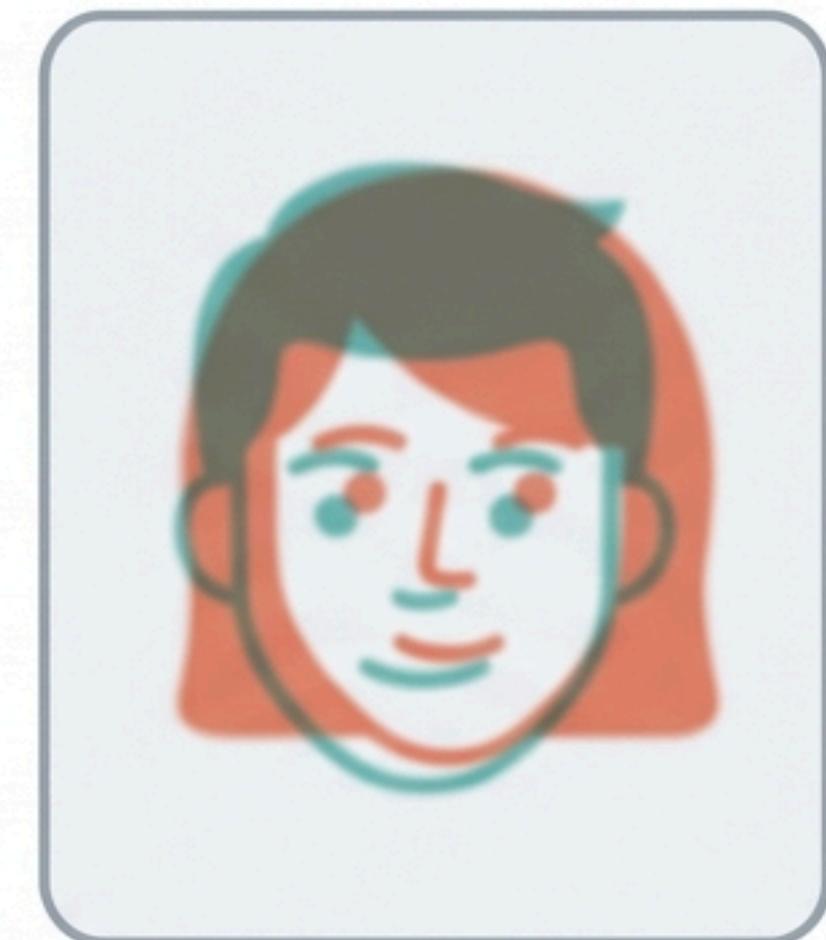
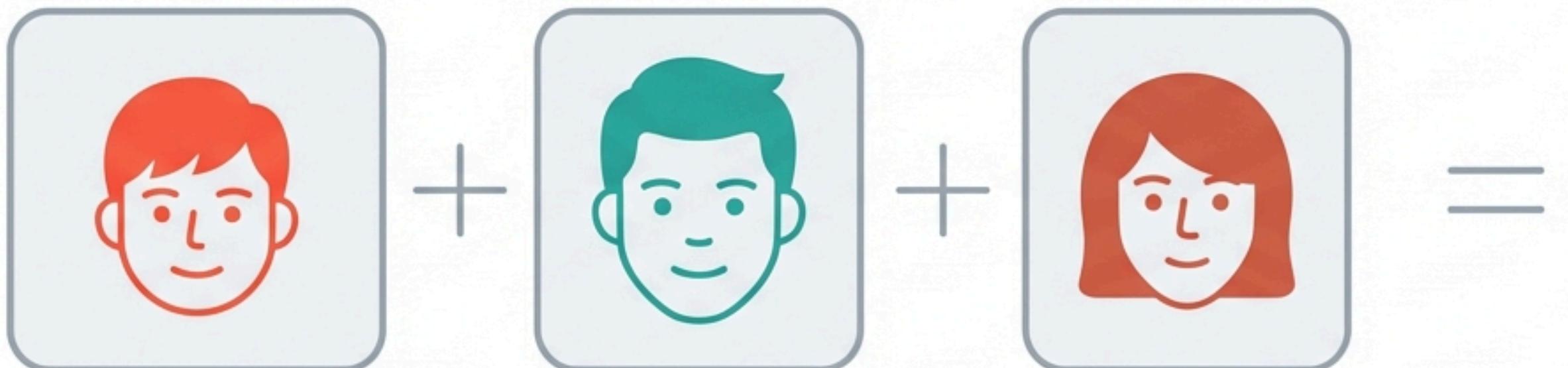
song1 = """ขออยู่ในชีวิตที่เหลือของเรอได้ไหม..  
อยากลืมตาแล้วได้พบร่องน้ำสุดท้าย..  
อยากรีบคนที่ได้นอนดูดาวข้างเรออีกหนึ่งวัน  
และเอนไปจุ่มพิตเรอซักล้านครั้ง  
อยู่กับฉันไปนานๆ... นะเรอ....."""



song2 = """ดอกกระเจียวบาน.. อีกไม่นานก็คงสิเจา..  
อ้ายก็รอเจ้าอยู่คือเก่า ไปเป็นผู้สาวผู้ใดหนอ..นาง  
นั่ง...คิดอด..บ่ได้นอนจนฟ้า sáng..  
ย่านความอักเสบแตกเม่ง  
เจ้าลืมทุกอย่างของสองเหา...  
  
ดอกกระเจียวบาน.. ผ่านหน้าแล้งเจ้าไปอยู่ไส..  
เข้าหน้าฝนบ่โドนเท่าไหร'.  
น้ำตาของอ้ายกะໄหลหย่าว...  
ใจสวอย..อย่าให้ค้อยถึงหน้าหนาว....  
จนดอกกระเจียวของอ้ายเหี้ยวเจา  
เจ้ายังบ่..คืนมา....."""

# ผลกระทบ: ปรากฏการณ์ 'Blended Voice'

- เสียงค่าเฉลี่ย: Tone ของทุกคนผสมกัน
- เอกลักษณ์หายไป: ไม่สามารถระบุตัวตนผู้พูดได้ชัดเจน



**Blended Voice**

## Step 4: แนวทางที่เลือกใช้จริง (Implementation Strategy)



Raw Text 100% (ตัด Phoneme ออก)



Clean Transcript (เลือกใช้ Pathumma)



Single Blended Voice



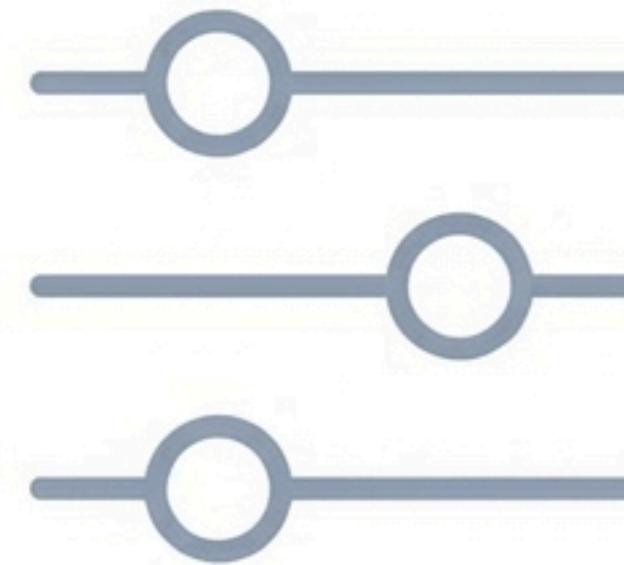
Speaker Tags (เพื่อการวิเคราะห์)

# ឧស្សាហ៍បច្ចុប្បន្ន (Technical Key Takeaways)



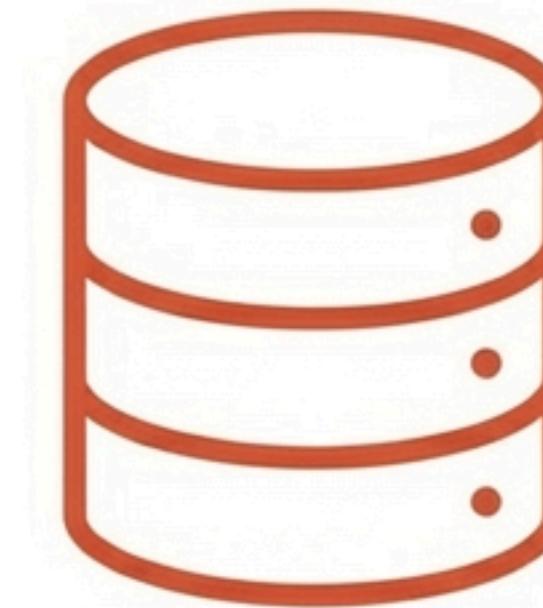
## Quality Dependency

ASR ធិន = TTS រៀនទូទិន



## Control Issue

មិនមែន Phoneme គុម Tone យក

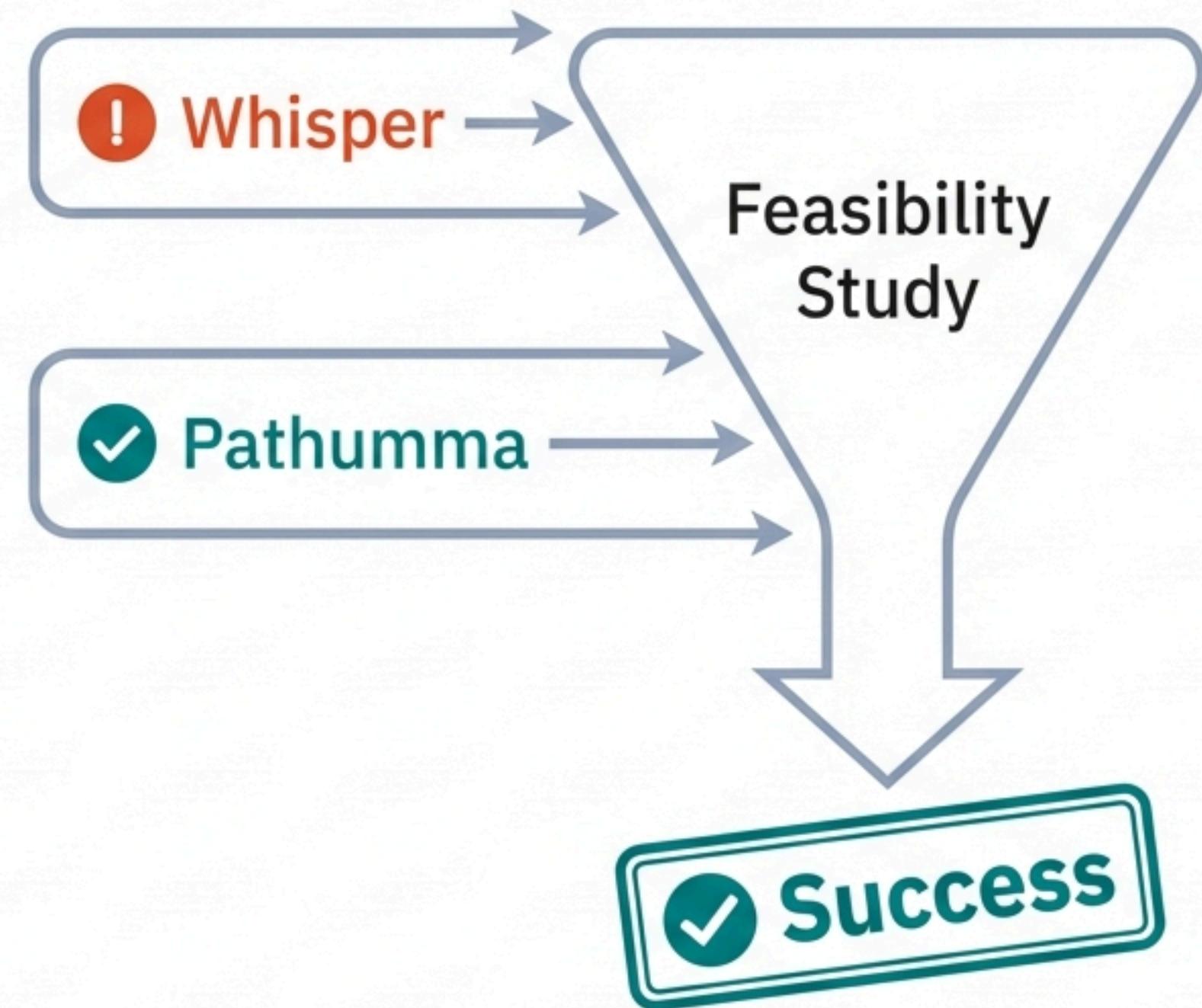


## Data Priority

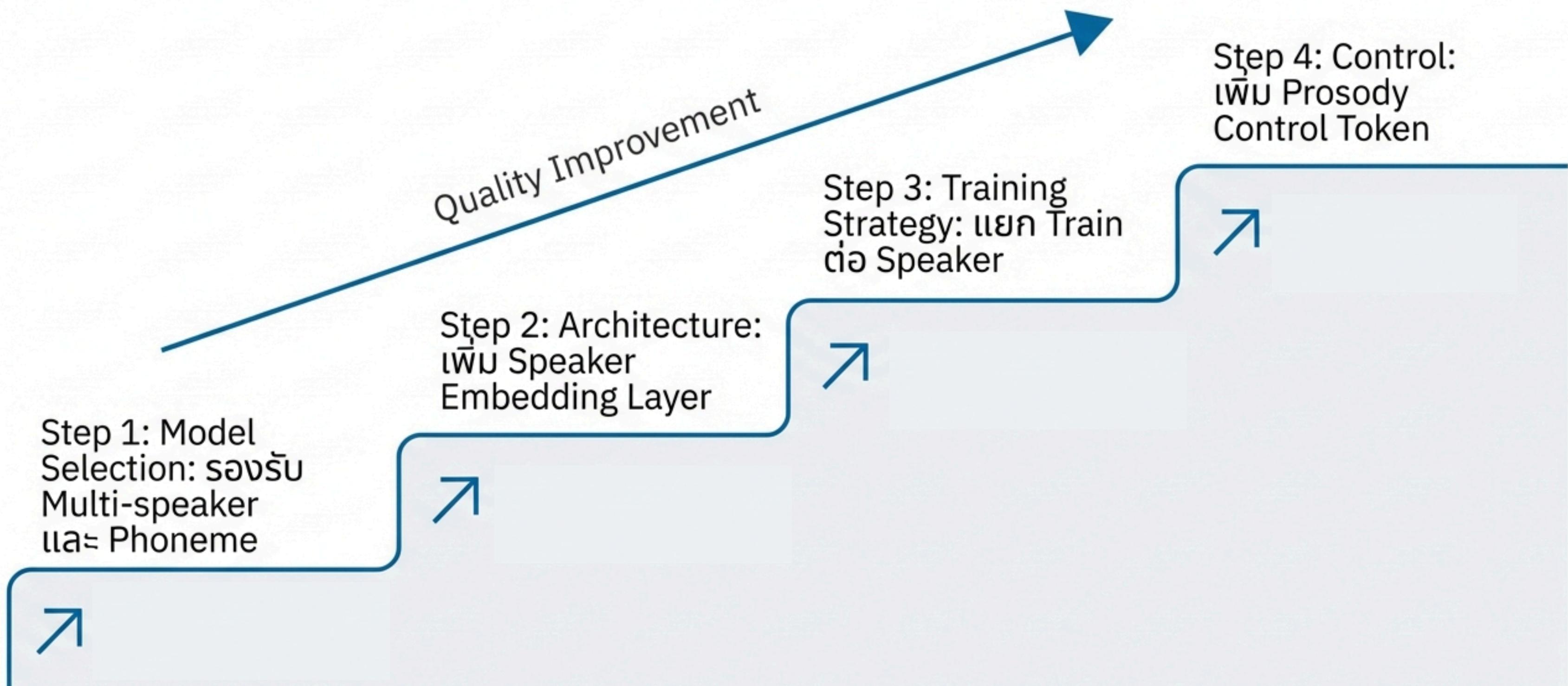
Data សំខាន់ > Hyperparameter

# สรุปภาพรวมโครงการ

- ⓘ เราไม่ได้ใช้ Whisper และ Pathumma ผสมกัน แต่ใช้เพื่อค้นหา Source Data ที่ดีที่สุด
- ✓ โครงการนี้สำเร็จใน阶段การประเมินศักยภาพ (Feasibility Study) ของการ Fine-tune ภาษาไทย
- ⚠ ข้อจำกัดเรื่อง Blended Voice เกิดจากสถาปัตยกรรมโมเดล ไม่ใช่คุณภาพข้อมูล



# ແບນທາງການພັດທະນາໃນອນໄຕ



# Step 1: Model Selection (การเลือกโมเดลพื้นฐาน)

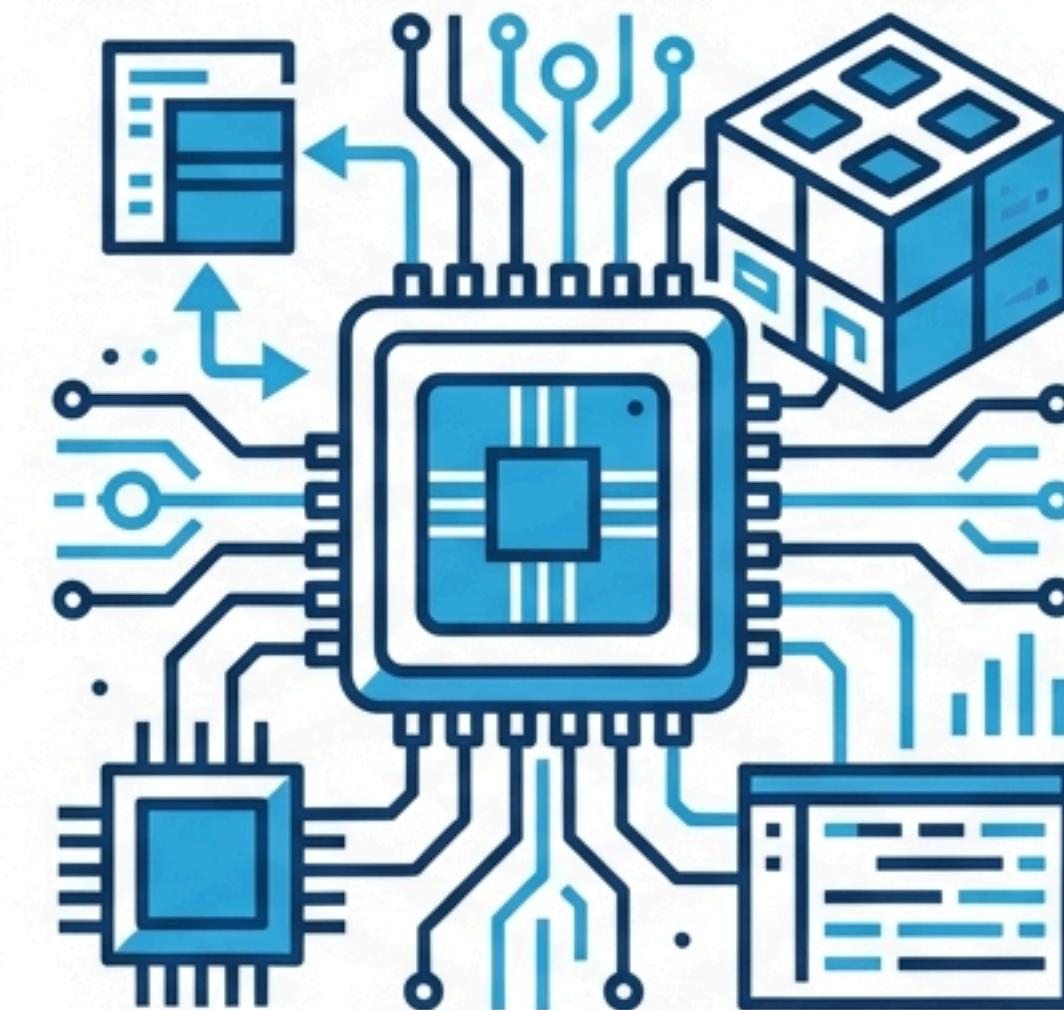
เป้าหมาย: รองรับ Multi-speaker และ Phoneme

Solution: Universal End-to-End Architecture

Model: Qwen3-TTS-Tokenizer-12Hz

Specs: 16 Codebooks / 2048 Size

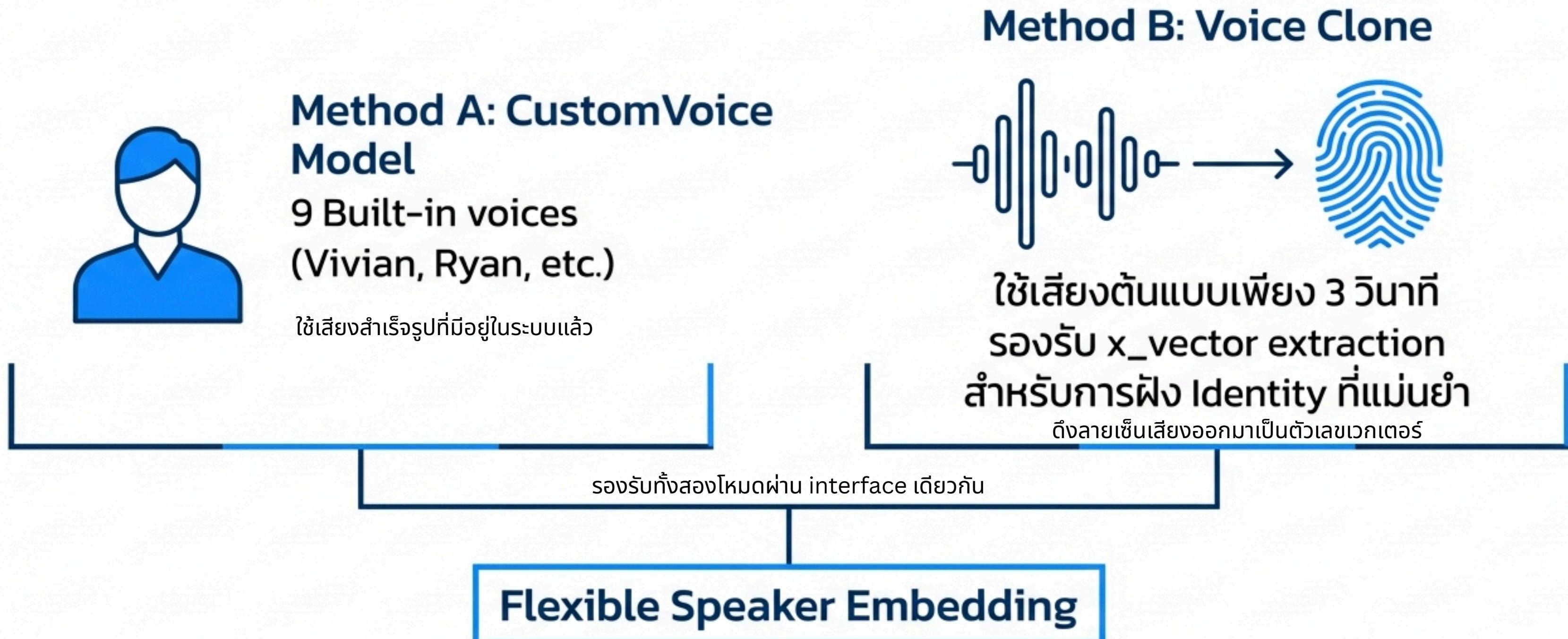
Feature: Discrete Multi-Codebook LM



Tokenizer ถูกออกแบบให้ Encode ลักษณะเสียงที่หลากหลายได้ตั้งแต่ต้น (Native Multi-speaker)  
โดยไม่ต้องใช้ Phoneme pipeline แบบเก่า

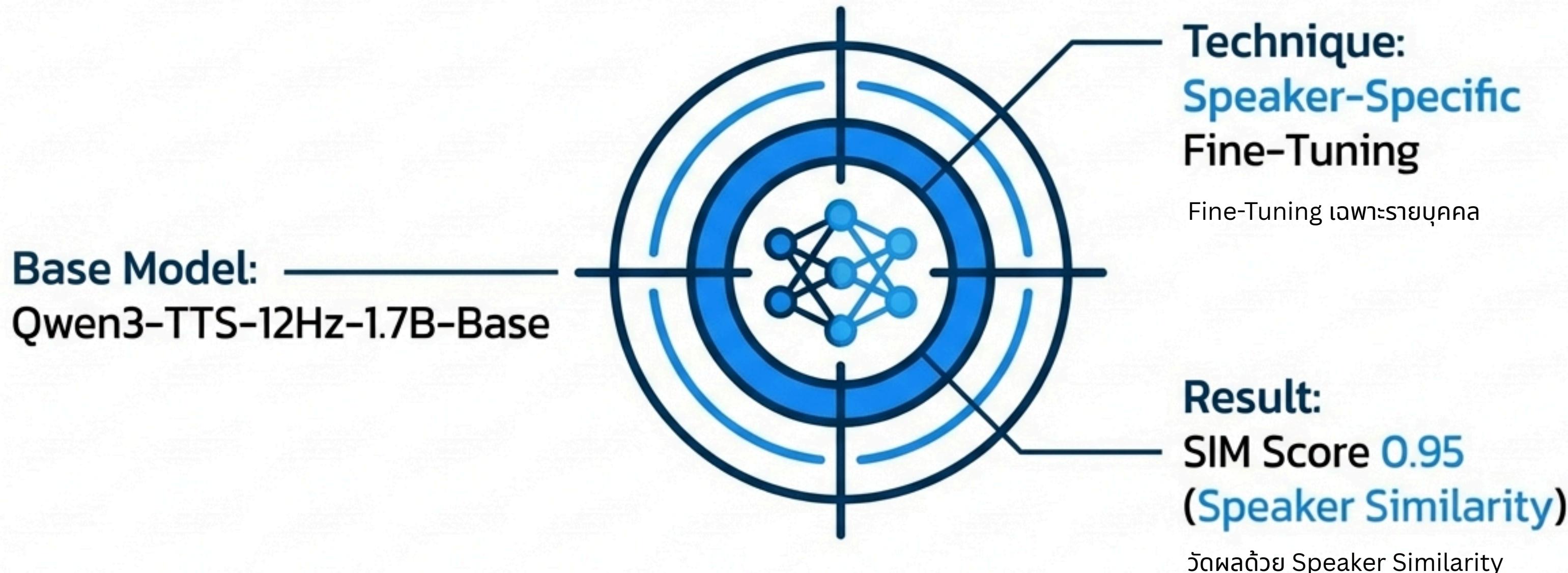
## Step 2: Architecture (สถาปัตยกรรมและการระบุตัวตน)

เป้าหมาย: เพิ่ม Speaker Embedding Layer



# Step 3: Training Strategy (กลยุทธ์การเทส)

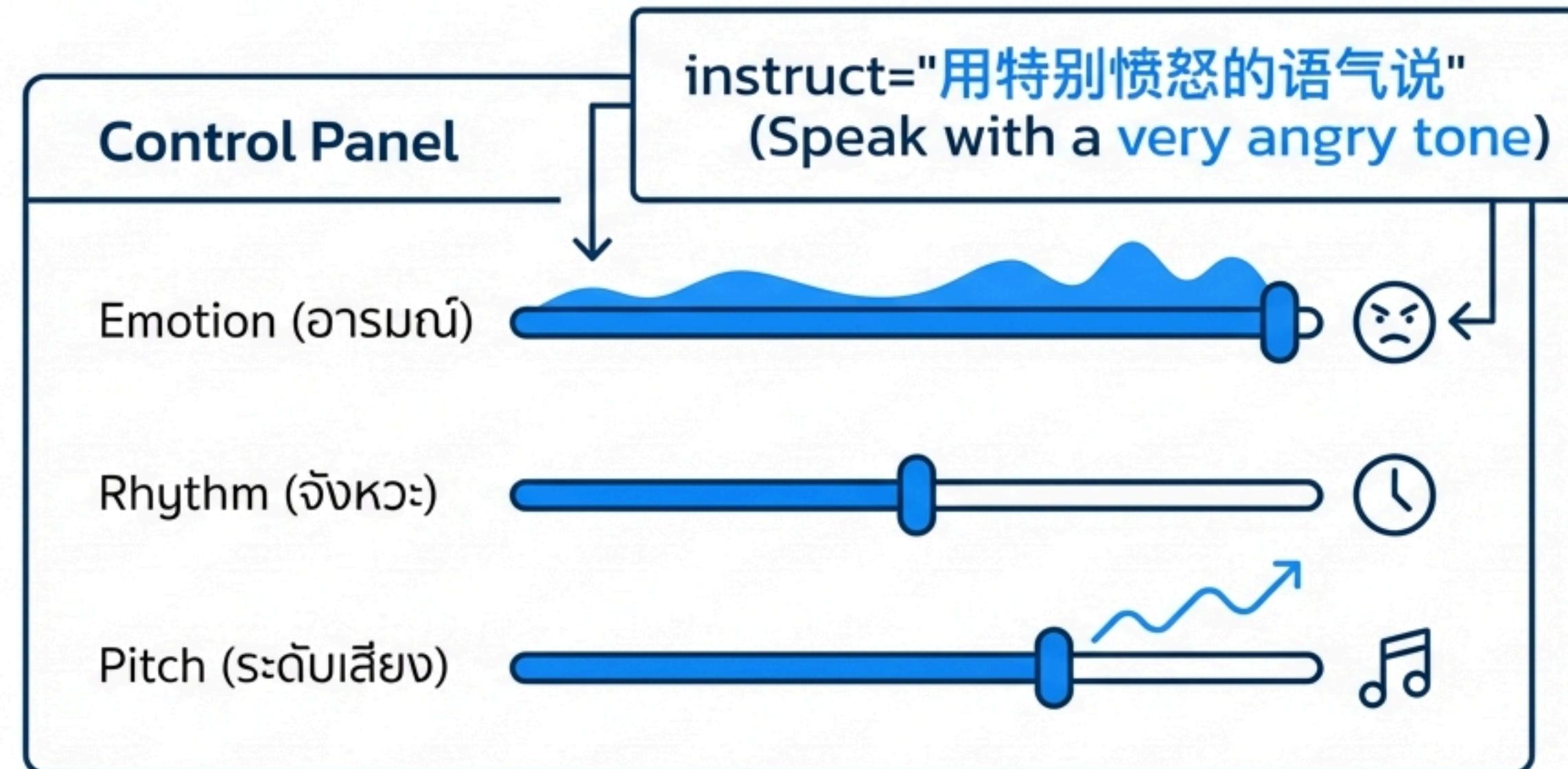
เป้าหมาย: แยก Train ต่อ Speaker



กระบวนการ Fine-tuning แยกรายบุคคล  
เพื่อความคมชัดและความเหมือนจริงสูงสุด (State-of-the-Art Similarity)

# Step 4: Control (การควบคุมขั้นสูง)

เป้าหมาย: เพิ่ม Prosody Control Token



Natural Language Understanding replaces rigid tokens.

แทนที่จะต้องกำหนดค่าตัวเลขหรือ token แบบตายตัว ระบบนี้ให้ผู้ใช้ สั่งงานด้วยภาษาธรรมชาติ

**สั่งงานด้วยภาษาธรรมชาติเพื่อให้ได้ Output ตามจินตนาการ ‘What you imagine is what you hear’**