



For Thai handwriting.

by Group 4 : พีบ้าสสุดหล่อ

Presented to :
Anantaporn Hanaskunatai

 ARTIFICIAL INTELLIGENCE



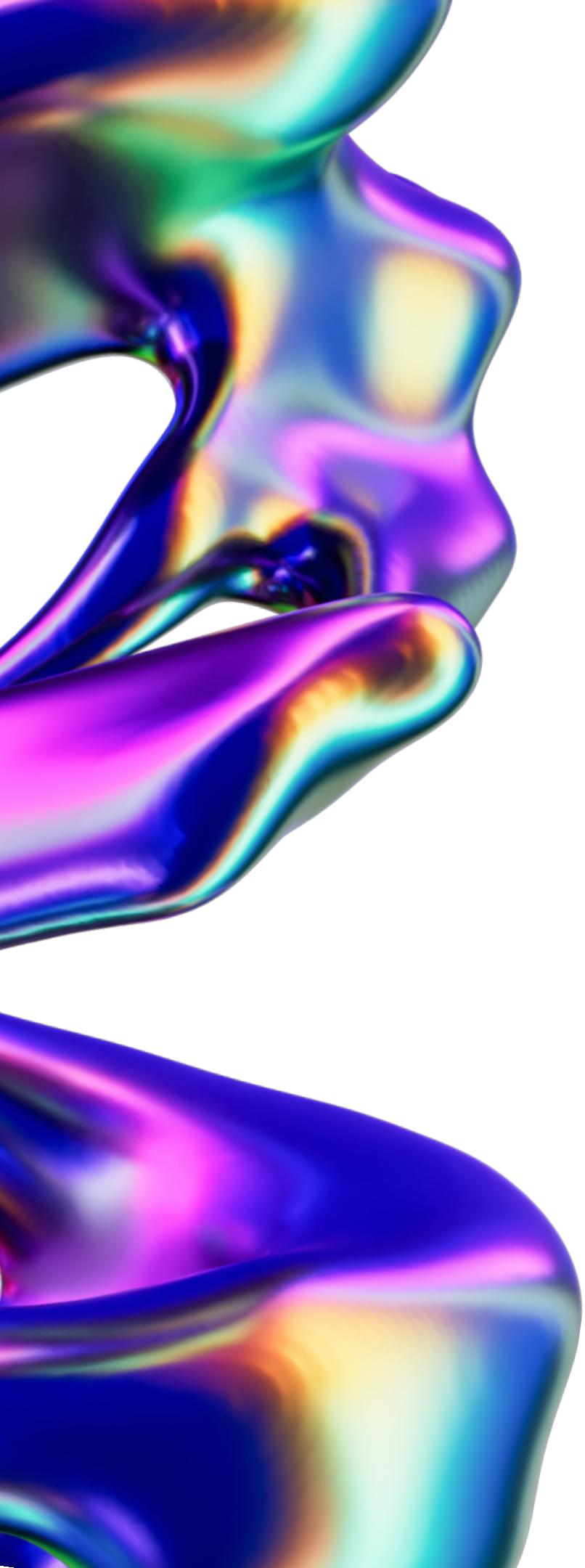
Objective

- Converted Thai handwritten into computer-readable text.
- Exploring AI and Machine Learning using LSTM.



Benefit

- Help people with converting handwritten documents to computer-readable documents.
- Turn complexed handwritten texts that is difficult to read by human into an easy-to-read format.



Model: LSTM

Long Short-Term Memory

- Pattern and Sequence Recognition
- Handling Diverse Data
- Handling Complex and Important Languages
- Mitigating the Vanishing Gradient Problem

Tools

- tesseract 5 : เทคนิคเดลสำหรับ OCR ลายมือภาษาไทย
- python 3.x.x : สร้าง script ไฟล์เพื่อสร้างข้อมูลและทดสอบ
- react.js + vite : พัฒนาในส่วนของหน้า Frontend และใช้จำลองหน้า browser
- python flask : พัฒนาในส่วนของ Server ในการติดต่อกับหน้า Frontend

Efficiency

```
avg cer : 2.2097902097902096  
avg wer : 1.4  
avg rate : 0.08625468336154787
```

model from handwriting image datasets

```
avg cer : 0.7902097902097902  
avg wer : 2.8  
avg rate : 0.4122534608248894
```

model from handwriting font datasets

Efficiency

$$CER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

- S is the number of substitutions characters
- D is the number of deletions characters
- I is the number of insertions characters
- C is the number of correct characters
- N is the number of characters in the reference ($N=S+D+C$)

Efficiency

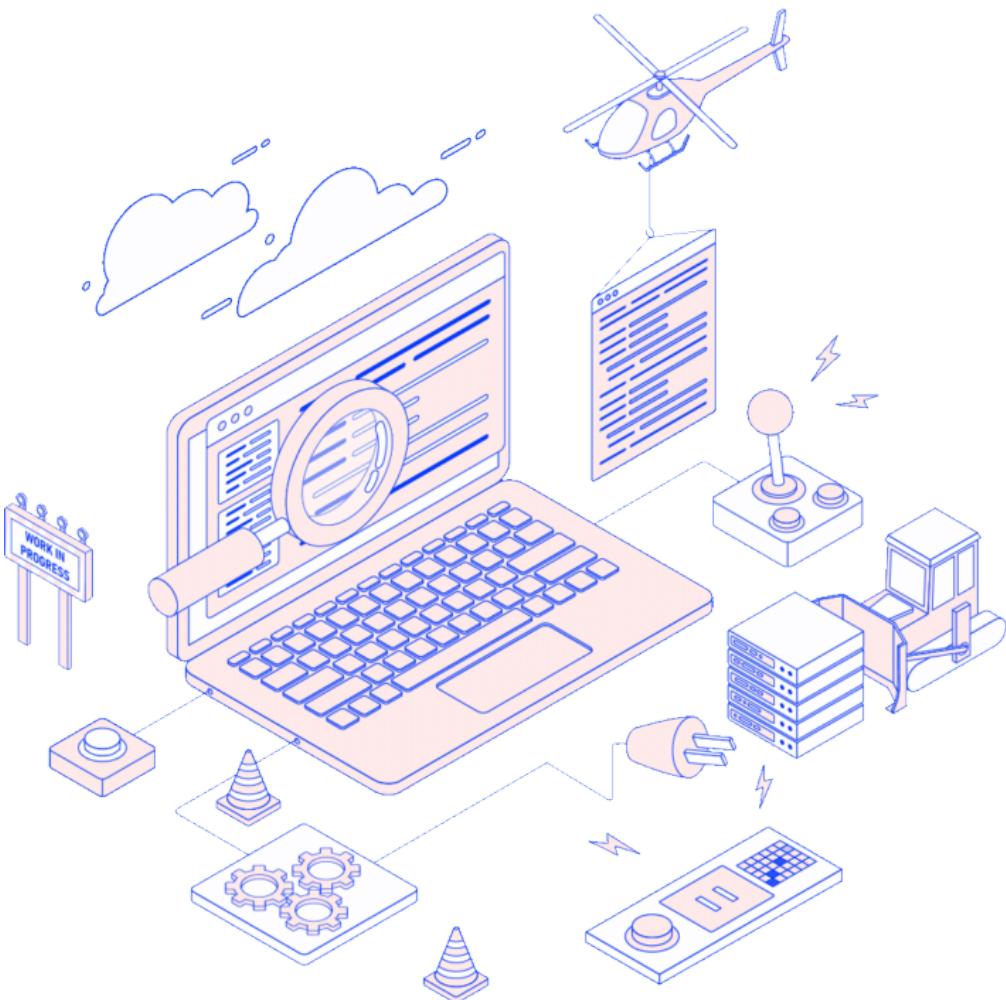
$$\text{WER} = \frac{S + D + I}{N}$$

- S is the number of substitutions word
- D is the number of deletions word
- I is the number of insertions word
- N is the number of characters in the reference ($N=S+D+C$)

Efficiency

$$\text{ratio} = \frac{2 * M}{T} * 100$$

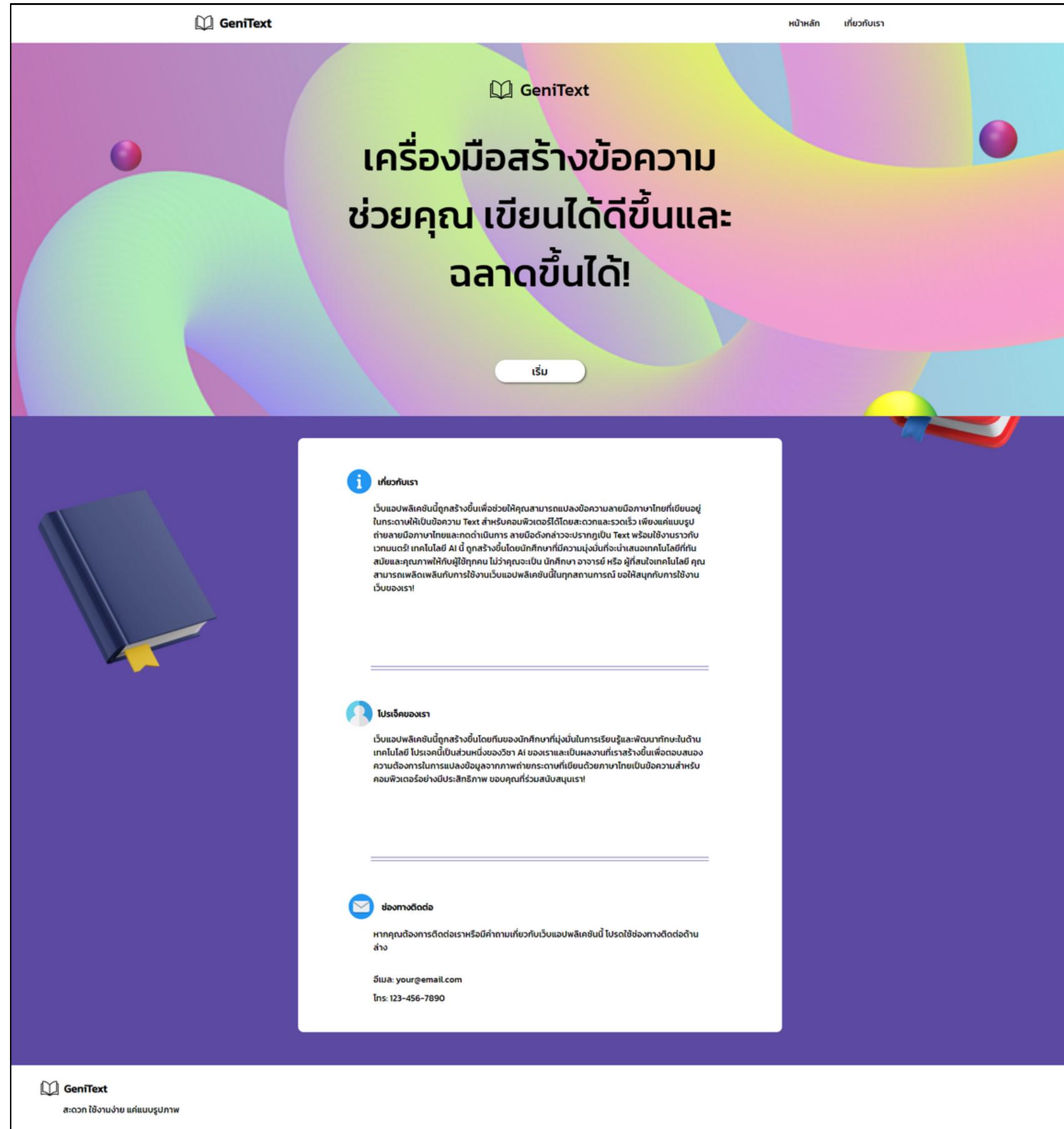
- T is the total number of characters in both strings
- M is the number of matches in the two strings.



Limitation

- Can convert only thai lang handwritten into computer-readable text.
- After generating, it cannot be exported as a pdf file or other file types.
- Can be uploaded only 2 of type .jpg and .png
- Can't to read images where the font is too small.
- Low accuracy low volume and cycles used for data training
- Only 4 handwriting fonts supported

Application Demo



Application Demo

 GeniText

หน้าหลัก เกี่ยวกับเรา

ใส่รูปภาพ
ใส่รูปภาพลายอัคบร
เพื่อให้จดจำและแปลงเป็นตัวอักษร

เป็นมนุษย์สุดประเสริฐเลิศคุณค่า
กว่าบรรดาผู้สัตว์เดร็จนา
จะฝ่าฟันพัฒนาวิชาการ
อย่างล้ำผ่านญาณเข่นฆ่าบีท่า!
ไม่ถือโหงโกรธแข่งชัดชัดด่า
หัดอภัยเหมือนกีฬาอัชณาสัย
ปฏิบัติประพฤติกฎกำหนดใจ
พุดจาให้ฉันฯ จำฯ นำฟังเมยฯ

ไฟล์ภาพที่เลือก:
IMG_Ai3.png

สร้างข้อความ

 GeniText
สะดวก ใช้งานง่าย แค่แนบรูปภาพ

ปัญหาและข้อเสนอ

ปัญหา : การสร้างโมเดลเพื่ออ่านลายมือภาษาไทยจากภาพใช้แรงงานค่อนข้างมาก
เนื่องจากการเทรนโมเดล tesseract จำเป็นต้องใช้ไฟล์ .box ประกอบด้วยตัวอักษร และระบบพิกัดจากของตัวอักษรภายในภาพ ดังนั้นต้องสร้างไฟล์ .box ที่พิกัดตรงกับภาพ 100% จึงจะทำให้โมเดลมีประสิทธิภาพ ซึ่งต้องใช้แรงงานจำนวนมากเพื่อสร้าง .box ที่มีคุณภาพ

วิธีการแก้ปัญหา

- เริ่มเทรนโมเดลจากการเทรนด้วย font ลายมือภาษาไทย เพราะ tesseract มี command text2image() สำหรับสร้างไฟล์ .box และไฟล์ภาพของ font ทำให้ลดการกำในส่วนของการจัดเตรียมข้อมูลเพื่อเทรนโมเดล

ข้อเสนอแนะ

- เทรนโมเดลด้วย font ให้ครบถ้วน font และค่อยทำการรวมโมเดลเข้าด้วยกัน เนื่องจาก การรวมโมเดลจะทำให้โมเดลมีขนาดใหญ่ขึ้น ทำให้ใช้ระยะเวลาในการเทรนเพิ่มขึ้นอีก จึงควรเทรนทีละ font ให้เรียบร้อย

The Team

64050149 ປັນທະວັນ ບ້ວສັງຍິ

64050202 ຮັບຮກມີ ແກ້ວຄຣີກາ

64050199 ຮເຫຕ ກູກົດມົຍ

64050061 ຜວນາກຣ ຈິຕຣໄພບູລຍ

64050188 ການຸພົງສ ເບ້າແກ້ວ

64050191 ກູມີຣີ ສຣີພໍາ

64050213 ວຽການຕ ແສນສຸຂສກຸລະຈລ

64050218 ວຽຍຸກຣ ພຣມພາດີ