RESTOR-AI-TION

ESTD 2023

Preserving the past, illuminating the future

Meet the Team

Research Intern
Mackenzie Austin

Software Engineer
Ramanuj Singh

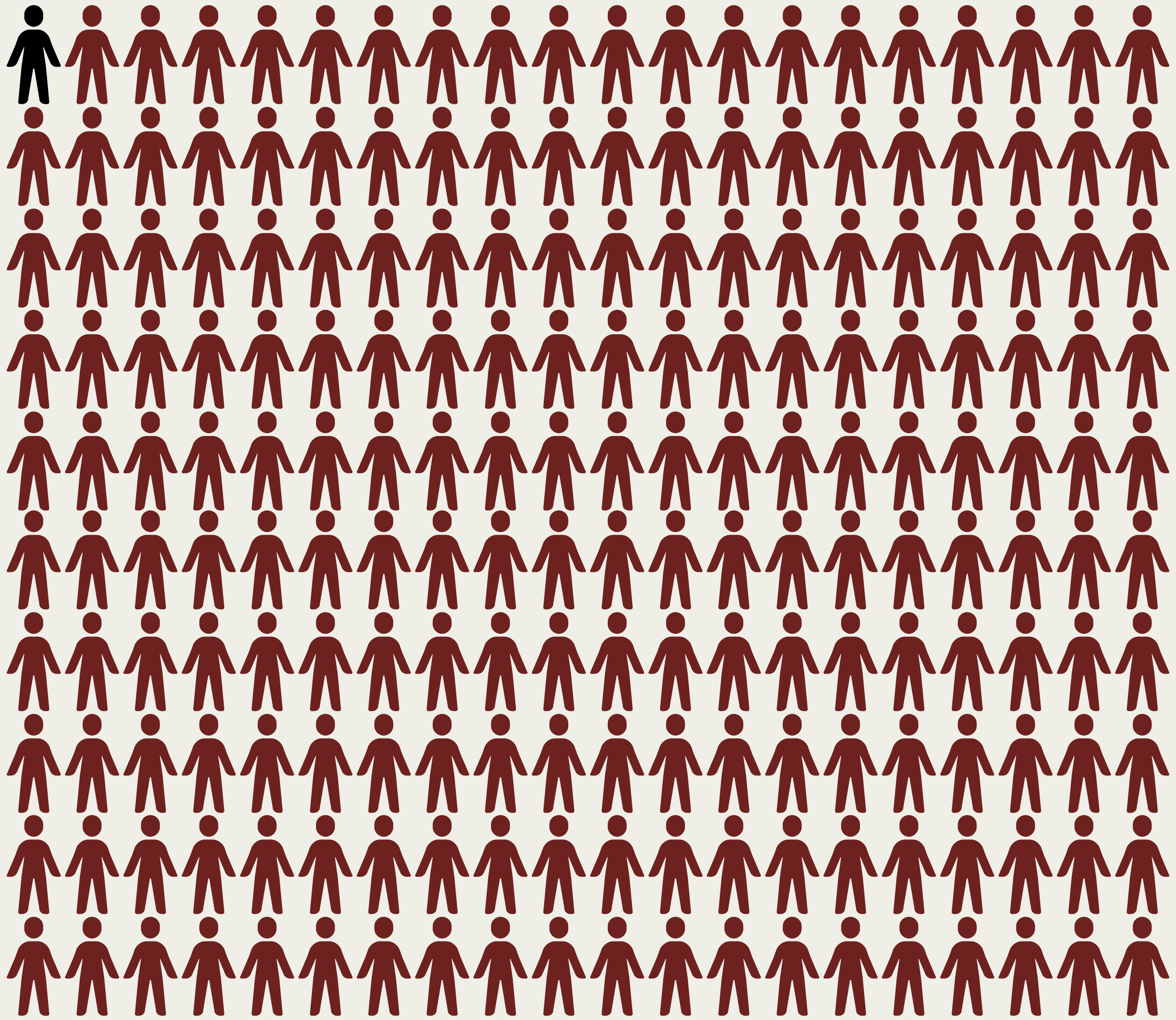
Software Engineer
Mahesh Arumugam

Engineering Director
Vinod Viswanathan

Lead Data Analyst
Jherson Fuentes

Navigating Ancient Texts

Spanning over a thousand years of Japanese history, premodern Japanese literature and historical documents were penned in Kuzushiji, a script now legible by less than 0.01% of modern Japanese speakers.



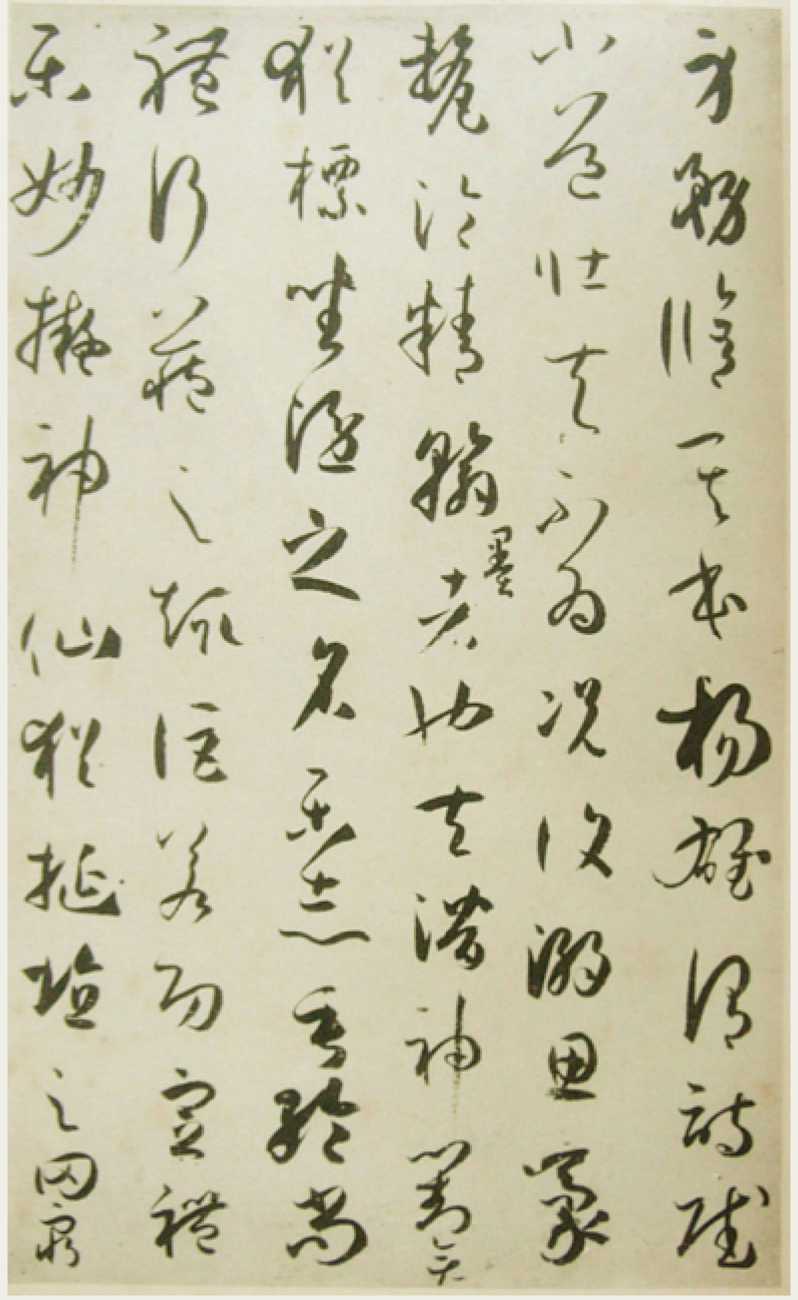
Navigating Ancient Texts

Preserve Japan's rich history and culture

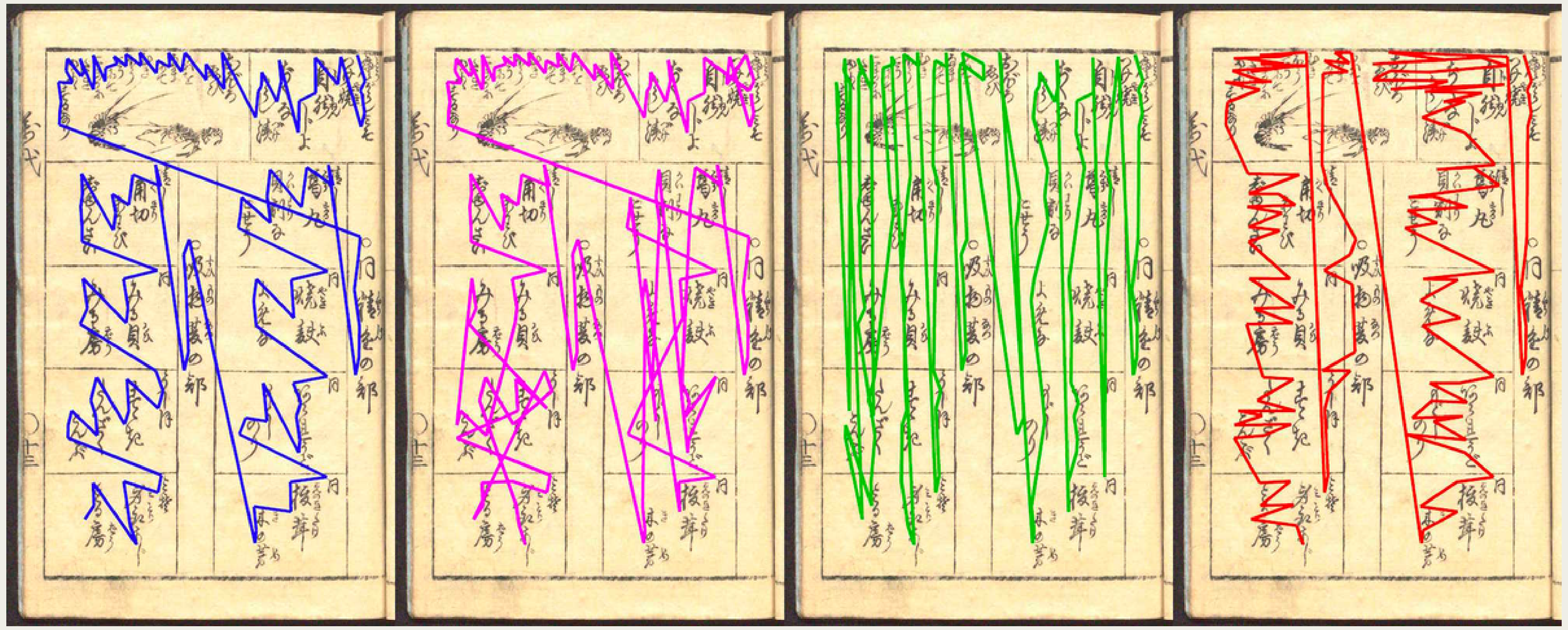
Ignite a renewed interest in ancient Japanese customs and practices

Enhance legibility and the overall quality of written texts

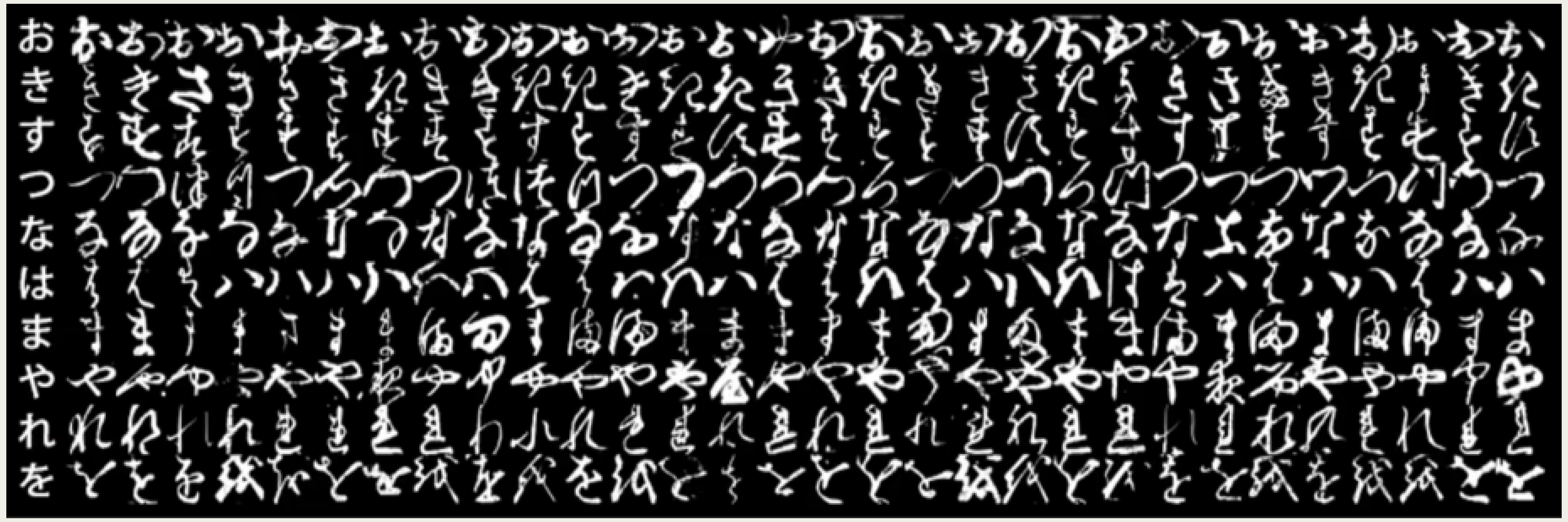
Navigating Ancient Texts



Cursive writing



Reading order may be non-linear



Ancient/archaic grammar

Understanding Our Users

Scholars

Fewer Resources

Educators

Ease of access

Everyday Individuals

Sate curiosity

Under the Hood: Data

KMNIST (k49)

Full representation of Kuzushiji Hiragana characters

- **49** classes. **48** Hiragana, and **1** Hiragana iteration mark.
- Imbalanced dataset of **270,912** images

KMNIST(kkanji)

Large dataset of 3832 Kanji characters

- **3832** Kanji characters
- Highly imbalanced, ranging from **1766** examples to **1** example per class
- **140,426** images

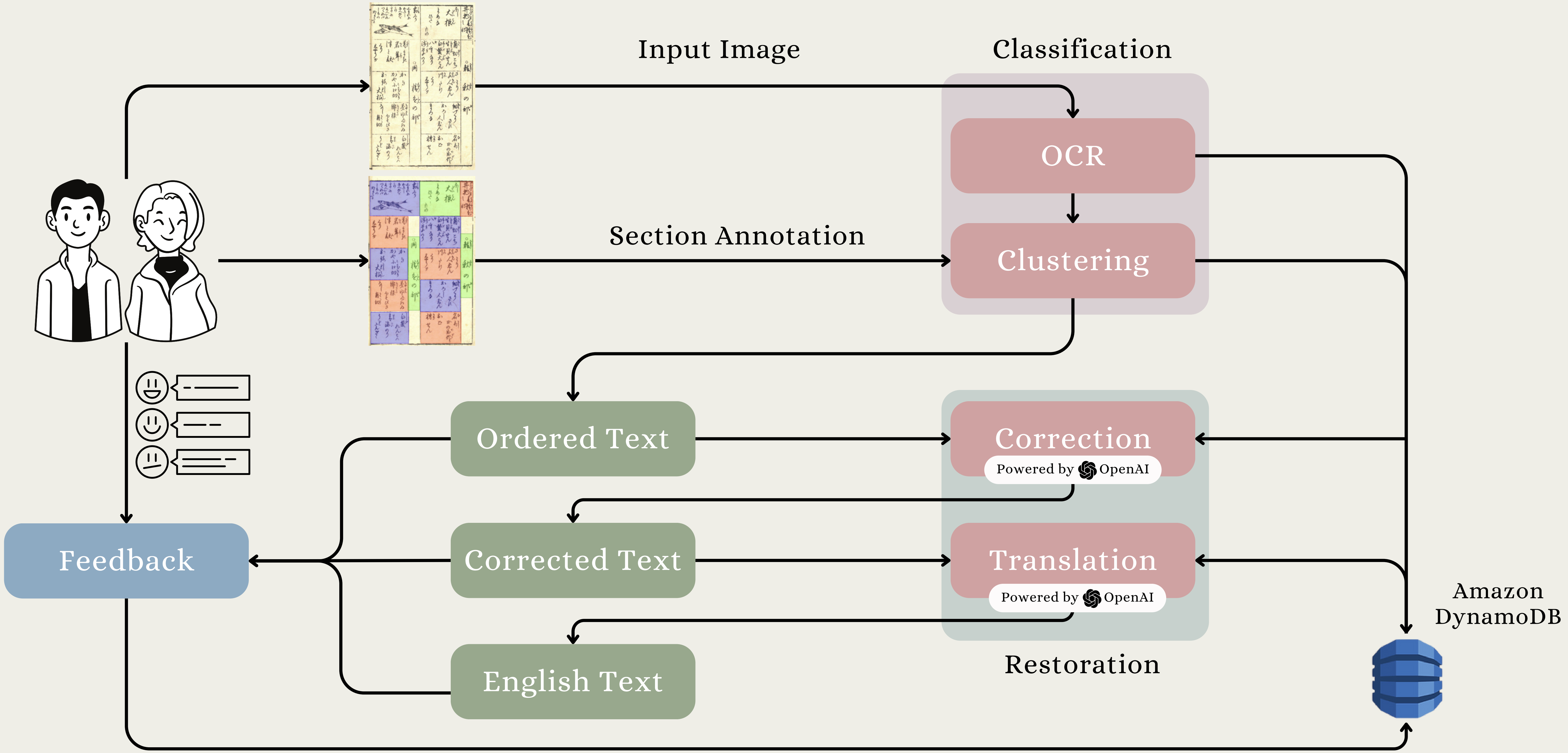
NISE

Full page images from early Japanese texts

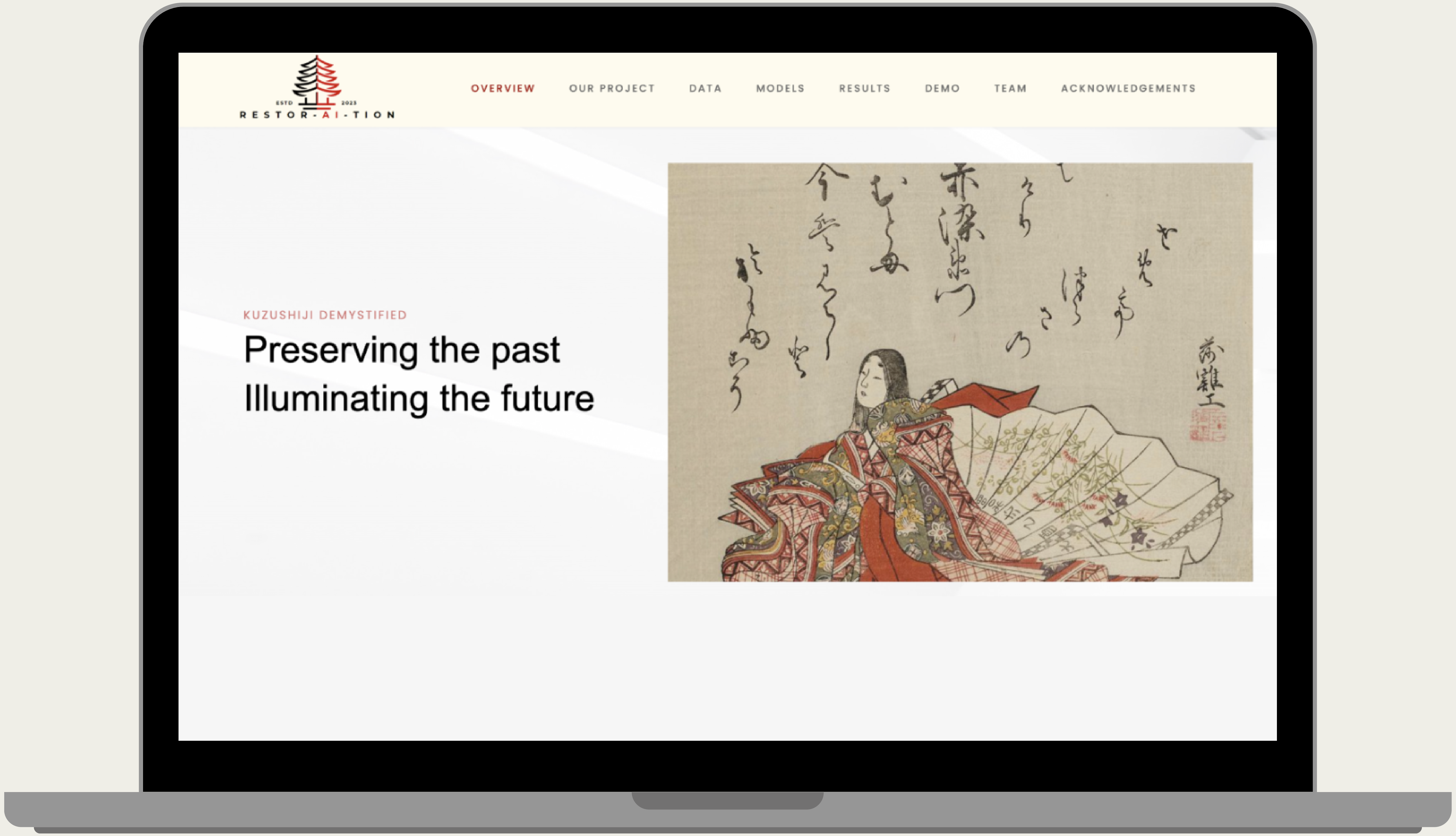
- **44** books
- Over **5** genres
- Published over the span of **200** years from late **1600**'s to **1800**'s
- **1,086,326** total characters

http://codh.rois.ac.jp/

Under the Hood: Architecture



Restor-AI-tion in Action



KUZUSHIJI DEMYSTIFIED

Preserving the past
Illuminating the future

Models: OCR

1

CUSTOM

CNN Model

- CNN based character detection model

75-80% accuracy

2

OPEN SOURCE

Easy OCR

- ResNet + LSTM + CTC model
- Potential for downstream application

Unsuccessful for Kuzushiji

3

OPEN SOURCE

Hanya's OCR

- CenterNet and MobileNetV3
- One of the top solutions in Kaggle competition

> 95% accuracy

4

CLOSED SOURCE

KuroNet

- Residual U-Net architecture
- State of the art solution deployed in miwo app

Unsuccessful in training

5

CUSTOM

Contours

- OpenCV contours and classify
- Non-character contours are problematic

50% accuracy

EasyOCR: https://github.com/JaidedAI/EasyOCR
CTC: https://www.cs.toronto.edu/~graves/icml_2006.pdf
KuroNet: https://arxiv.org/abs/1910.09433
Custom OCR: https://towardsdatascience.com/how-did-i-train-an-ocr-model-using-keras-and-tensorflow-7e10b241c22b
Hanya's OCR: CenterNet: https://paperswithcode.com/method/centernet
MobileNetV3: https://arxiv.org/abs/1905.02244

ESTD 2023

Models: OCR

1

CUSTOM

CNN Model

- CNN based character detection model

75-80% accuracy

2

OPEN SOURCE

Easy OCR

- ResNet + LSTM + CTC model
- Potential for downstream application

Unsuccessful for Kuzushiji

3

OPEN SOURCE

Hanya’s OCR

- CenterNet and MobileNetV3
- One of the top solutions in Kaggle competition

> 90% accuracy

4

CLOSED SOURCE

KuroNet

- Residual U-Net architecture
- State of the art solution deployed in miwo app

Unsuccessful in training

5

CUSTOM

Contours

- OpenCV contours and classify
- Non-character contours are problematic

50% accuracy

ESTD 2023

Performance Metrics: OCR

Performance on total characters in 15 held-out books (2040 pages)

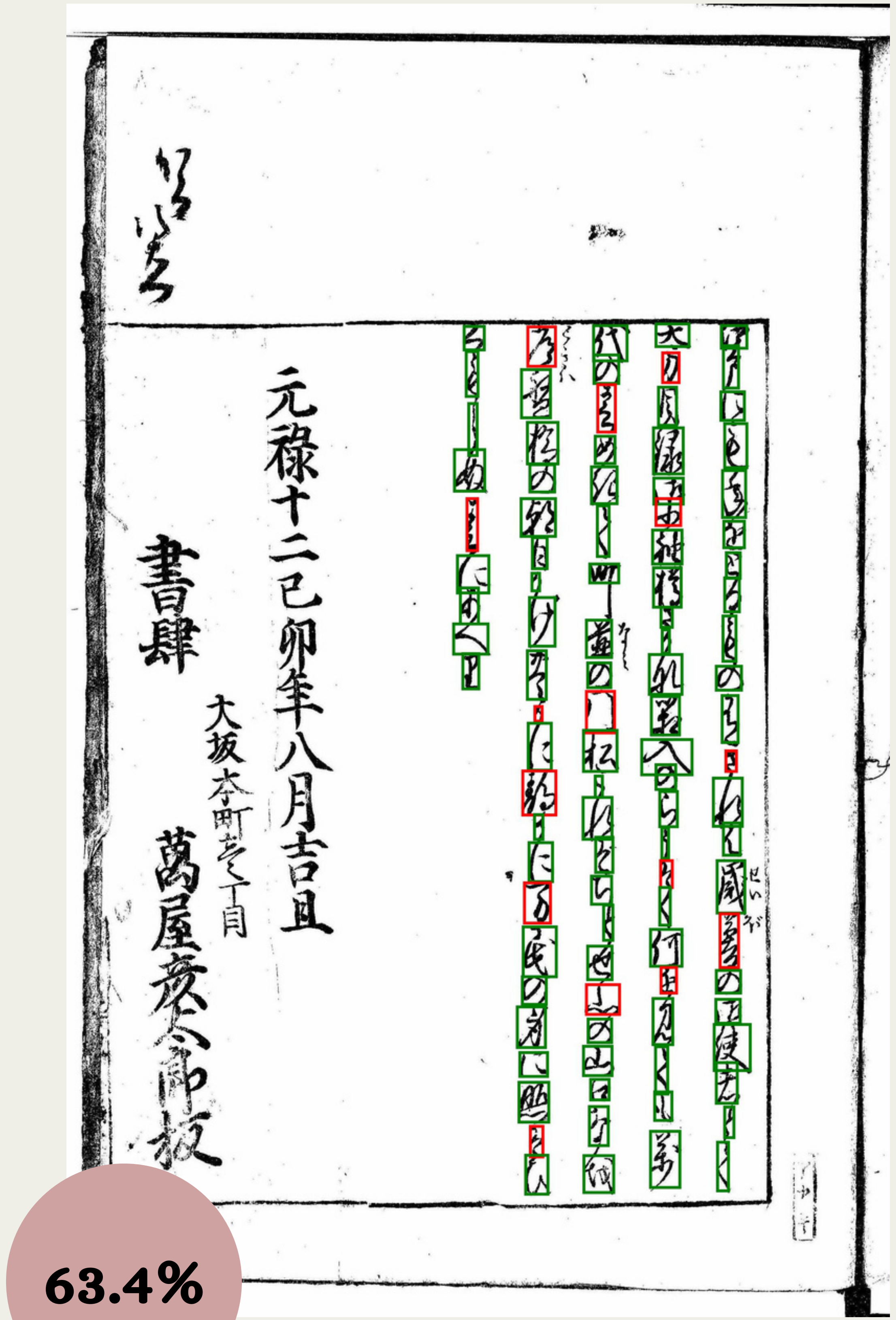
	KuroNet	KuroNet + Reg	Hanya's OCR
Precision	0.7964	0.8889	**0.9101**
Recall	0.7509	**0.9025**	0.8958
F1	0.773	0.8957	**0.9029**

Findings

- Hanya’s OCR is at least as good as KuroNet models
- In majority of the books, recall is better with KuroNet + Reg model, while precision and overall F1 score is better with Hanya’s OCR

KuroNet: https://link.springer.com/article/10.1007/s42979-020-00186-z
Hanya OCR: https://github.com/t-hanya/kuzushiji-recognition

Performance Metrics: Hanya’s OCR



元禄十二己卯年八月吉且

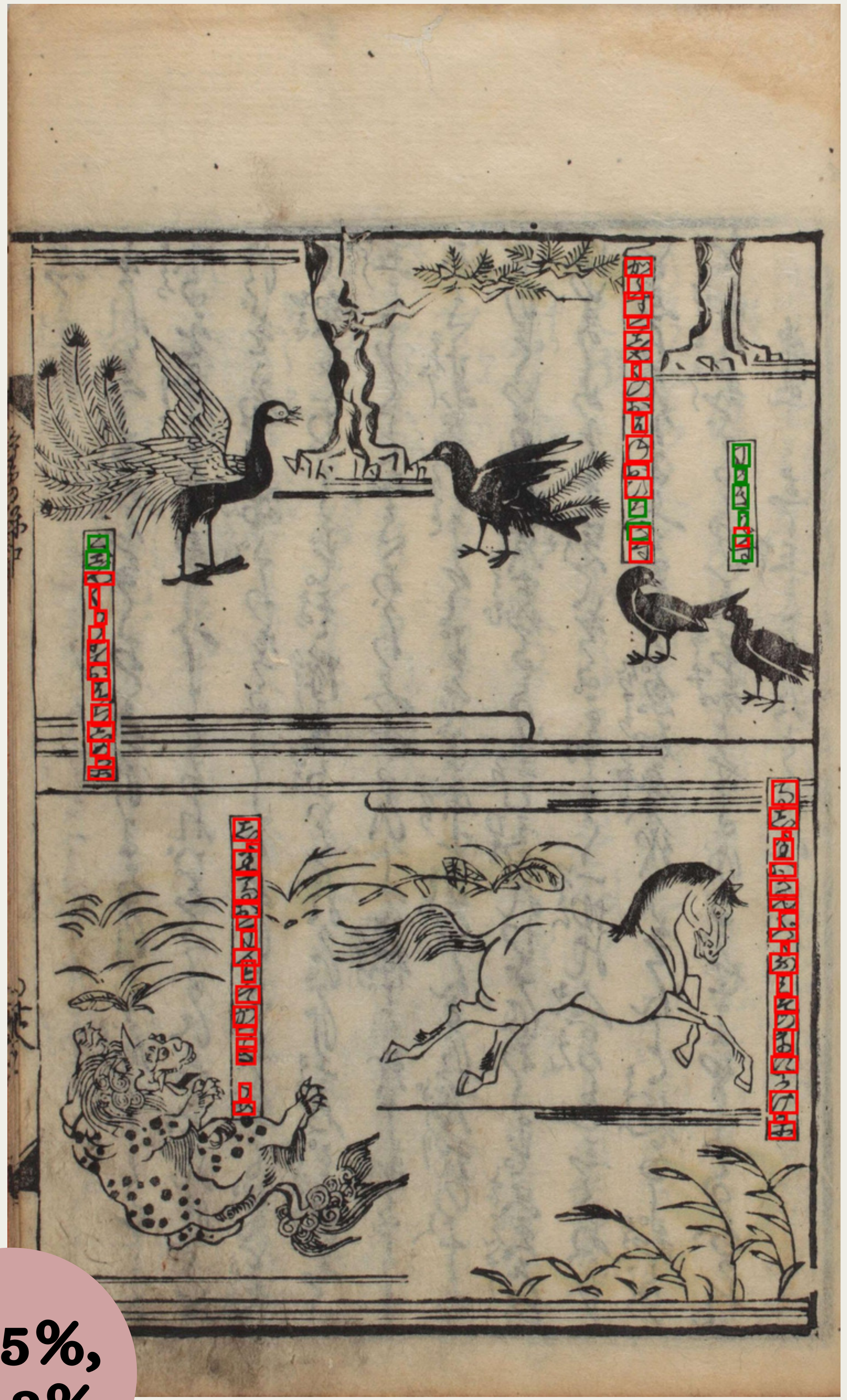
書肆

大坂本町弐丁目

萬屋彦太郎板

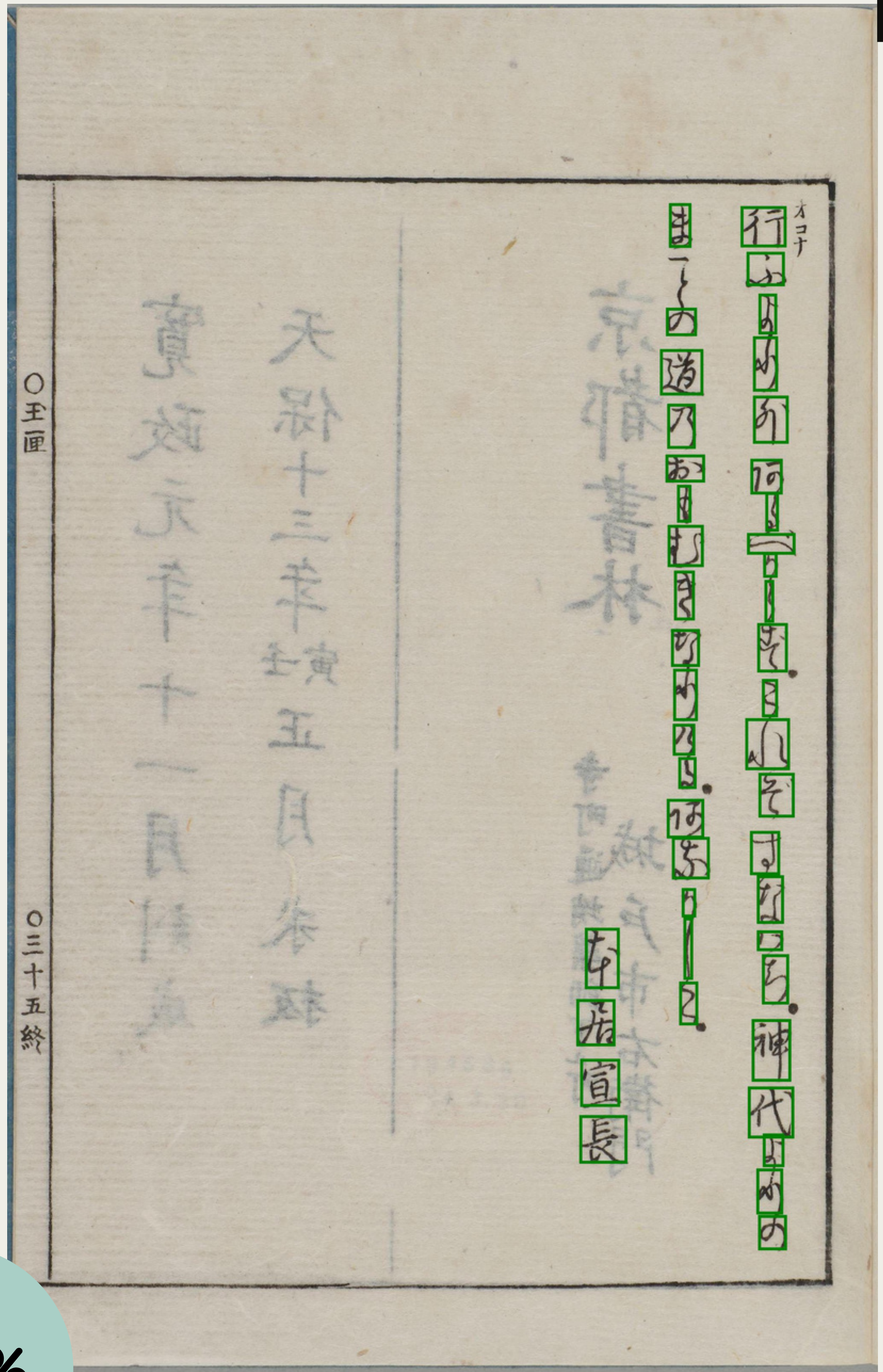
63.4%

Low Precision



12.5%, 22.2%

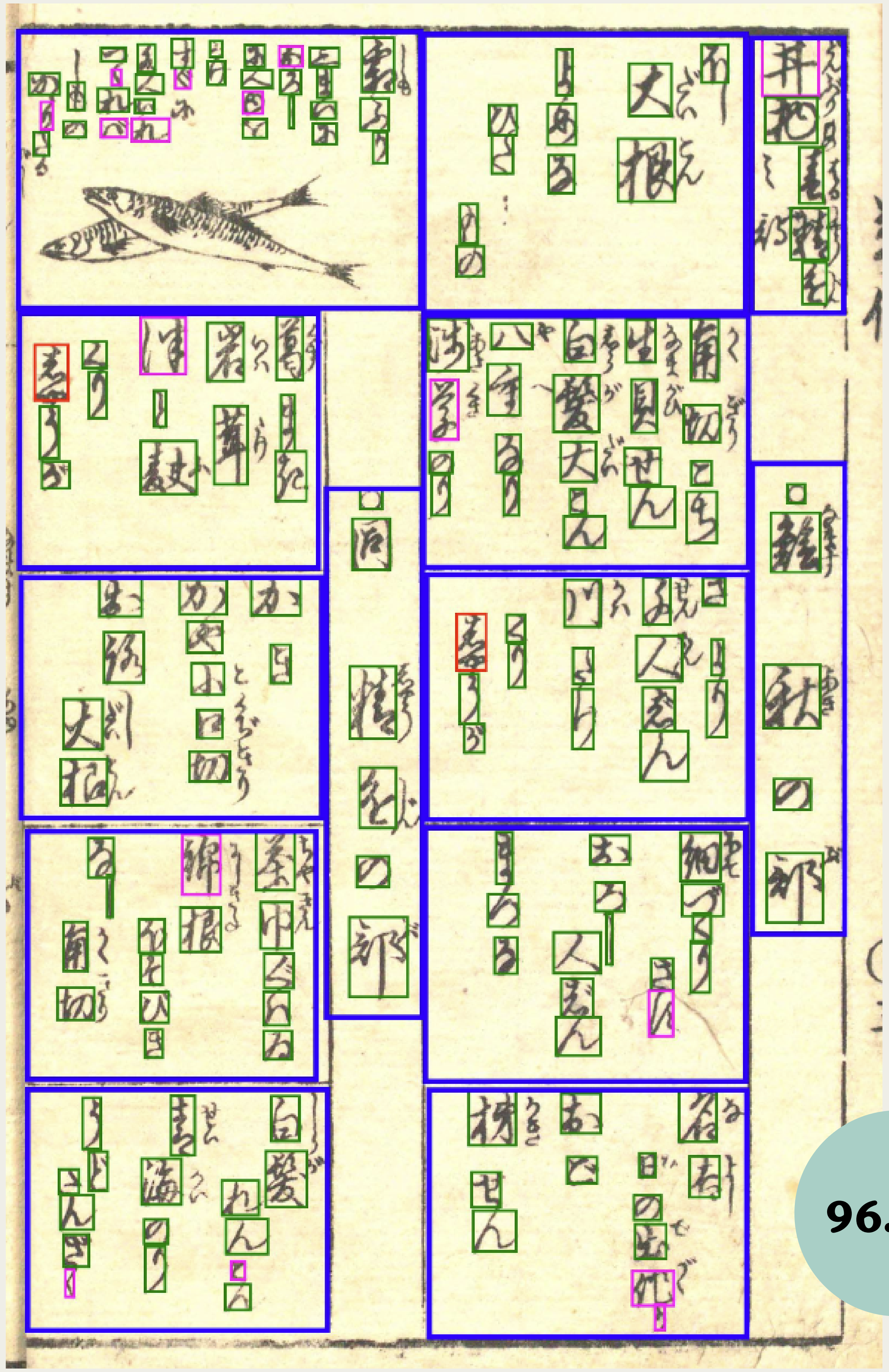
Low Recall & F1



100%

High Precision, Recall & F1

Insights: OCR



96.6%

Visual Feedback for Users: Users receive visual feedback through bounding boxes colored for quick interpretation.

🟩 Probability >= 0.9 – High Confidence

🟪 Probability >= 0.5 and < 0.9 – Moderate Confidence

🟥 Probability < 0.5 – Low Confidence

Confidence Metric: Indicate the overall confidence for the image. A weighted score that penalizes pink and red buckets based on the proportions of characters that fall in those buckets.

Models: Reading Order

1

CLOSED SOURCE

Deep-AR

- Auto-regressive character ordering
- Given a position, predict the character in the next position

2

CUSTOM

Modified K-means

- Only to detect vertical clusters (custom distance metric)
- Logic to collapse overlapping clusters

3

CUSTOM

Transformer

- GPT-Neox Japanese word embeddings
- Positional embeddings from bounding boxes

4

CUSTOM

Fine-tuning T5

- Simple model to rearrange characters into meaningful phrases

5

CUSTOM

LLM Japanese Model

- Large language model with 3.6b parameters developed by LINE, a common messaging app in Asia.

Deep-AR: https://www.arxiv-vanity.com/papers/2106.06786/
K-Means: https://pyclustering.github.io/docs/0.10.1/html/index.html
GPT-NeoX-Japanese: https://huggingface.co/docs/transformers/model_doc/gpt_neox_japanese
T5: https://arxiv.org/abs/1910.10683, LLM: https://huggingface.co/line-corporation/japanese-large-lm-3.6b

Models: Reading Order

1

CLOSED SOURCE

Deep-AR

- Auto-regressive character ordering
- Given a position, predict the character in the next position

2

CUSTOM

Modified K-means

- Only to detect vertical clusters (custom distance metric)
- Logic to collapse overlapping clusters

3

CUSTOM

Transformer

- GPT-Neox Japanese word embeddings
- Positional embeddings from bounding boxes

4

CUSTOM

Fine-tuning T5

- Simple model to rearrange characters into meaningful phrases

5

CUSTOM

LLM Japanese Model

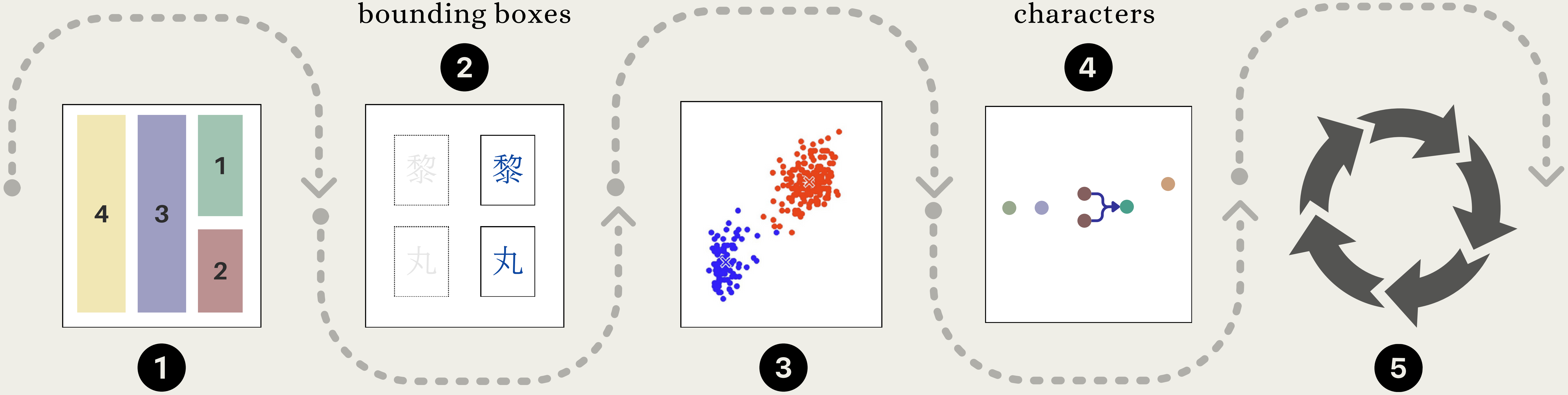
- Large language model with 3.6b parameters developed by LINE, a common messaging app in Asia.

Unable to run

Training accuracy extremely low / Cost prohibitive

ESTD 2023

Models: K-means Clustering



1 User Provided reading order boxes

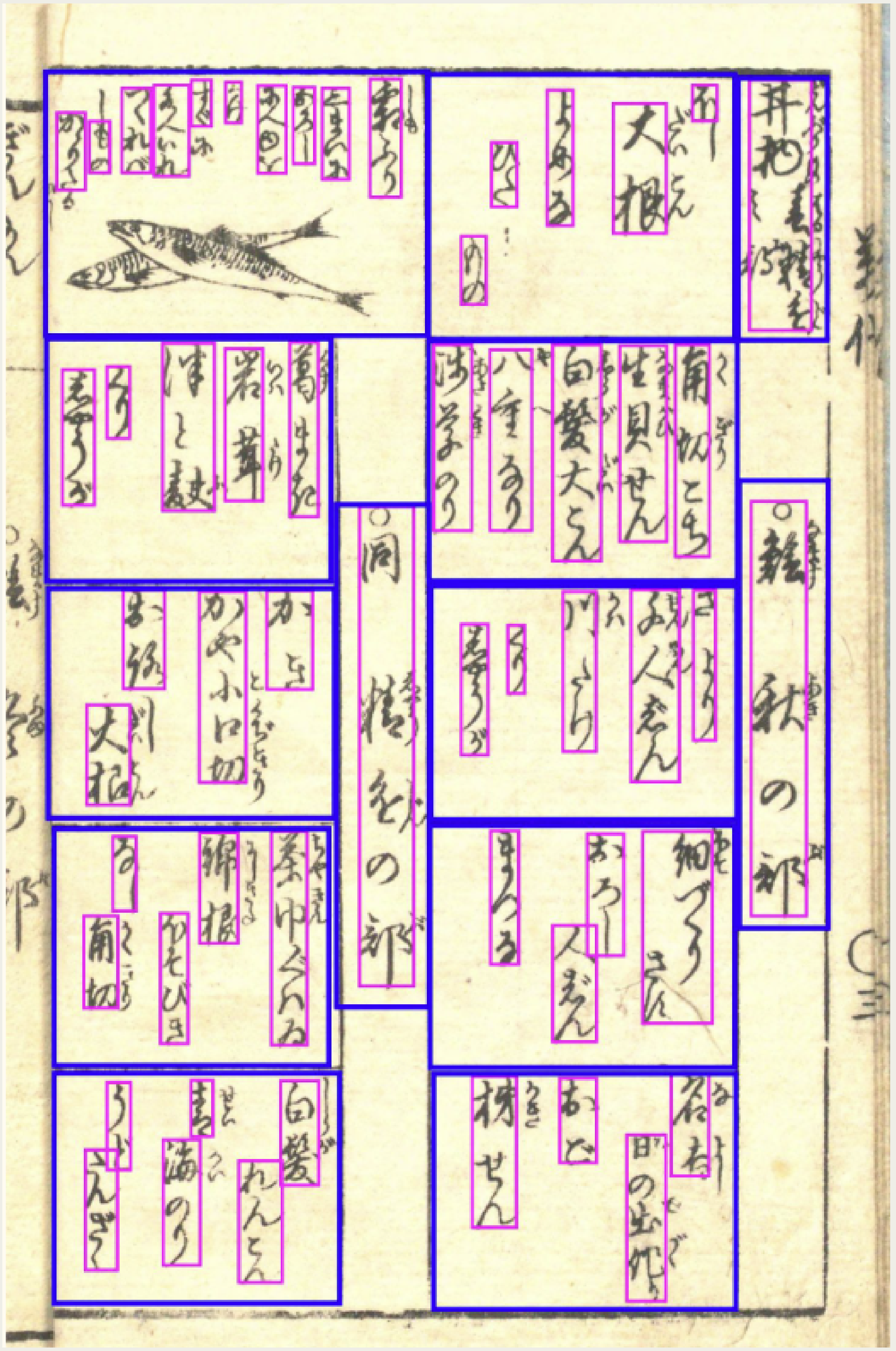
2 Reading order box with OCR recoginized characters with bounding boxes

3 K-means clustering with custom distance metric

4 Merge centers that are within 2 standard deviatons of bounding box of ocr characters

5 Recompute cluster centers and, optionally, repeat until convergence

Performance Metrics: K-means Clustering



Ground Truth: 丼物春精進の部ほし大根よめなひたしもの霜ふり三まいにおろしにへゆをかけすぐに水へいれつくれバしものかりたるべし・ 鱠秋の部角切こち生貝せん白髪大こん八重なり浅草のりさより千人じん川たけくりしやうが細づくりさすおろし人じんまつな名吉日の出作りおご柿せん・同精進の部葛まき岩茸つと麩くりしやうがかきかや小口切おろし大根茶巾ぐハゐ錦根ほそびきなし角切白髪れんこん青海のりう

Full Text: 丼物春精進ほ大根よめなひたもの霜ふり三まいにおろしにへゆをかけすゞ水へいれつゝればものかかた・鱠秋の部角切こち生貝せん白髪大こん八重なり浅草のりさより千人じん川たりくり音うが細づくりさすおろし人じんまつな名吉日の出作りおご柿せん・同精進の部葛まき岩茸つと麩くりしうがかきかや小口切おろ大根茶巾ぐはゐ綿根ほそびきなし角切白髪れんこん青海のりうどたんざく

Bleu Score: 0.7736

Models: Correction & Translation

1

OPEN SOURCE

Easy OCR

- Once OCR is completed, use EasyOCR to read the text
- Difficulties with column-wise texts

2

OPEN SOURCE

Manga OCR

- Once OCR is completed, use MangaOCR to read text
- Difficulties with the reading order

3

OPEN SOURCE

LLM

- Integrate with several LLMs trained with Japanese language
- Has trouble with ancient/archaic format of the text

4

CLOSED, API

GPT 4.0

- Integrate with ChatGPT (GPT 4.0) using APIs
- Easy to integrate and present results

LLM: https://huggingface.co/models
GPT API: https://platform.openai.com/docs/guides/text-generation
EasyOCR: https://github.com/JaidedAI/EasyOCR
MangaOCR: https://github.com/kha-white/manga-ocr

Models: Correction & Translation

1

OPEN SOURCE

Easy OCR

- Once OCR is completed, use EasyOCR to read the text
- Difficulties with column-wise texts

2

OPEN SOURCE

Manga OCR

- Once OCR is completed, use MangaOCR to read text
- Difficulties with the reading order

3

OPEN SOURCE

LLM

- Integrate with several LLMs trained with Japanese language
- Has trouble with ancient/archaic format of the text

4

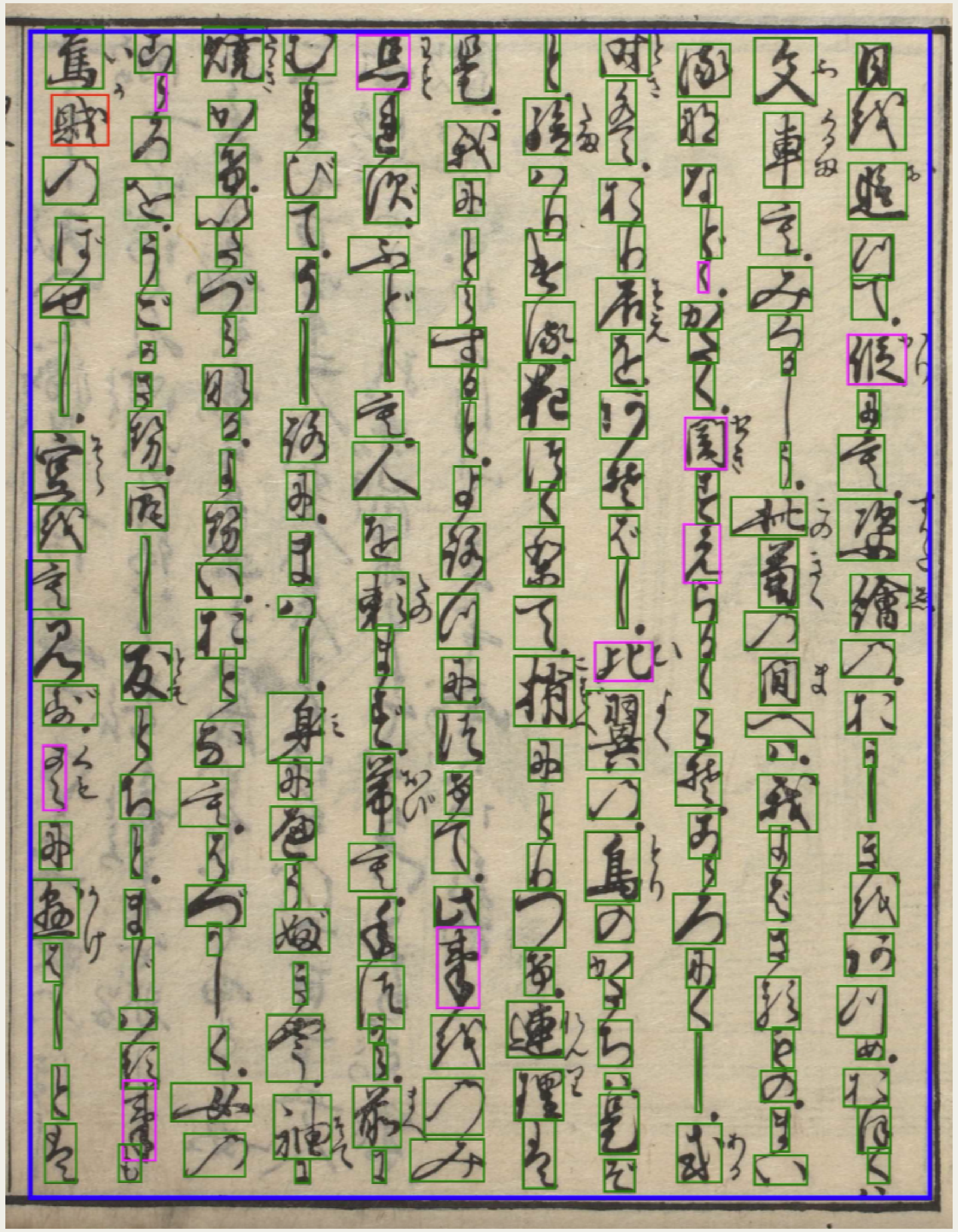
CLOSED, API

GPT 4.0

- Integrate with ChatGPT (GPT 4.0) using APIs
- Easy to integrate and present results

LLM: https://huggingface.co/models
GPT API: https://platform.openai.com/docs/guides/text-generation
EasyOCR: https://github.com/JaidedAI/EasyOCR
MangaOCR: https://github.com/kha-white/manga-ocr

Insights: Correction & Translation



Corrected Text

Confidence Score: 95.00%

此の菊の絵姿もう車にもつかぬ日を待ちて追う文を此方にして仮るにも我が心にかく我にくし絵の関こそ繋がらることを知るほならなくても、ある日は此も東の間にさえられつつある。されども、何ものにとらわれない我が心は、ほんの僅かな希望だけを抱いており、あたかも比翼の鳥のように連理の花を梢につけ、つつじの葉をそばに置いております。あるいは、我が心が動くほどには、こころをくすぐる事や遊びのあらし時と同じく、人への愛もまた帯につかれて前向きに進み、身にまとうぶきや袖に入る風もまた新たな希望をくれ、手元から産まれる喜びは後方に向かう希望となり、つねに身を包み込むなど、さらにどんな厳しい顔色でも女性の笑顔には敵わず、空を見上げて立ち止まることなく進む事に尽きると言え、これが我が生きた証しともなる。

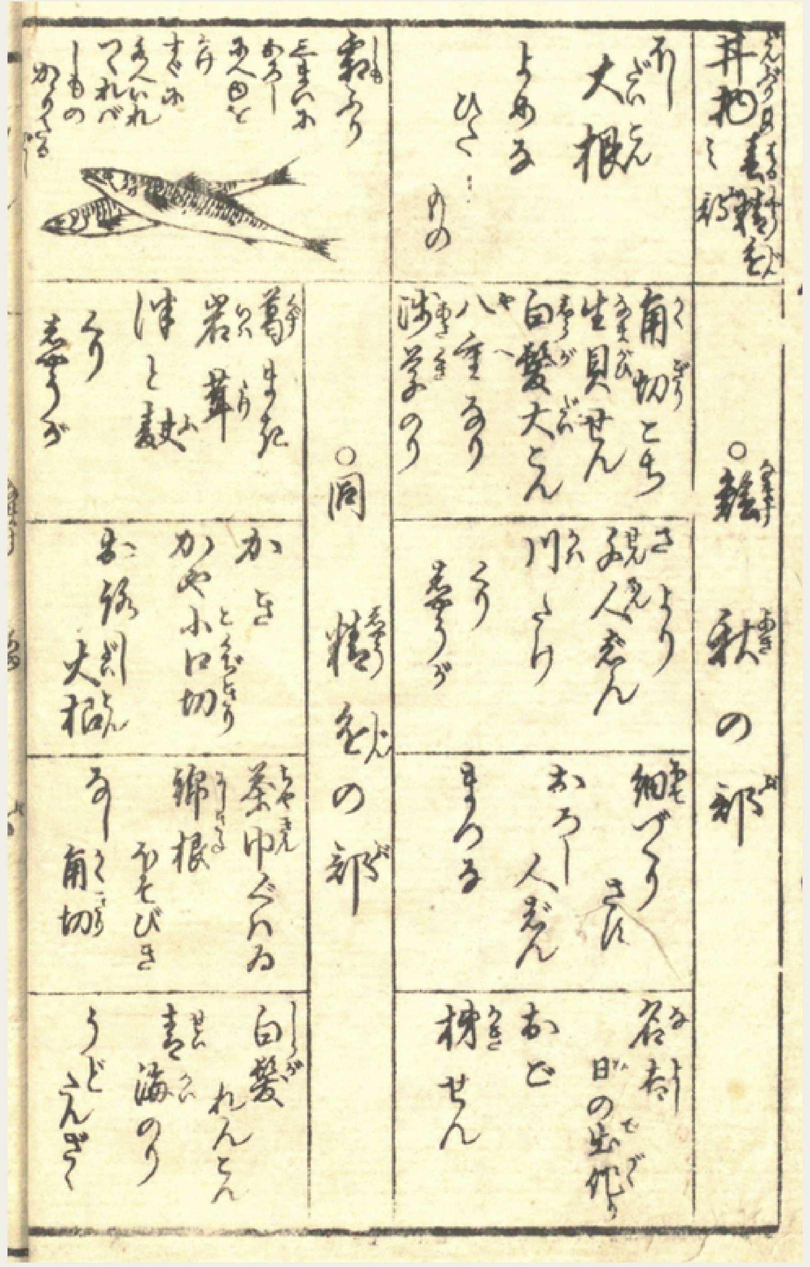
Translated Text

Confidence Score: 97.00%

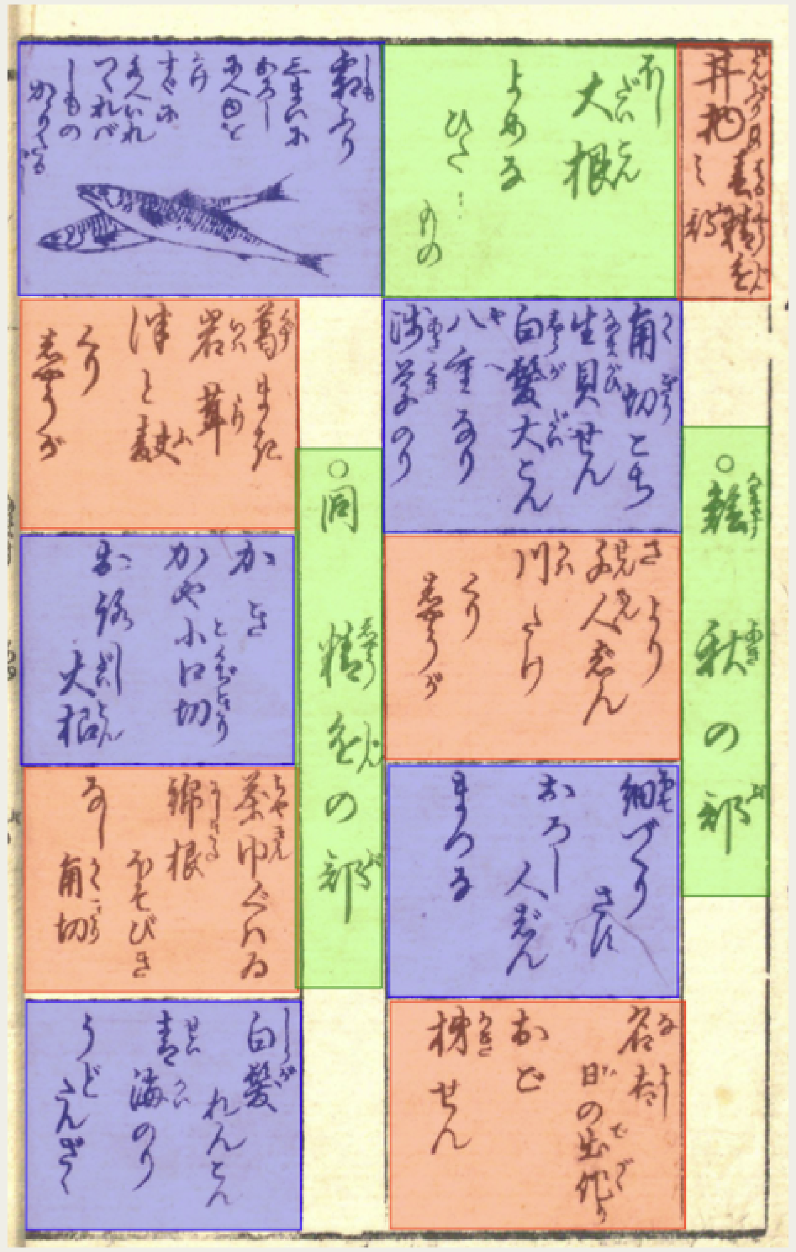
Waiting for the day when this likeness of a chrysanthemum will no longer serve a car, the letter chasing after it is thus provisionally on this side. Even if I do not know clearly that my heart is connected to this painting's barrier, one day this too is being eliminated for a moment. Nevertheless, my heart that is caught by nothing is holding on to only a sliver of hope, just like a bird with conjoined wings attaching a united-flowers to treetops, and placing azalea leaves beside it. Or, my heart is stimulated enough that just like during the storm of ticklish things and games, love for people also moves forward tied to a belt, the spray to wear and the wind entering the sleeves also give new hope, the joy born from hands becomes the hope directed backwards, it always envelops me, no stern expression can compete with a woman's smile, looking up at the sky without stopping and just moving forward is ultimately what matters, and this becomes the proof that I lived.

Models: End to End

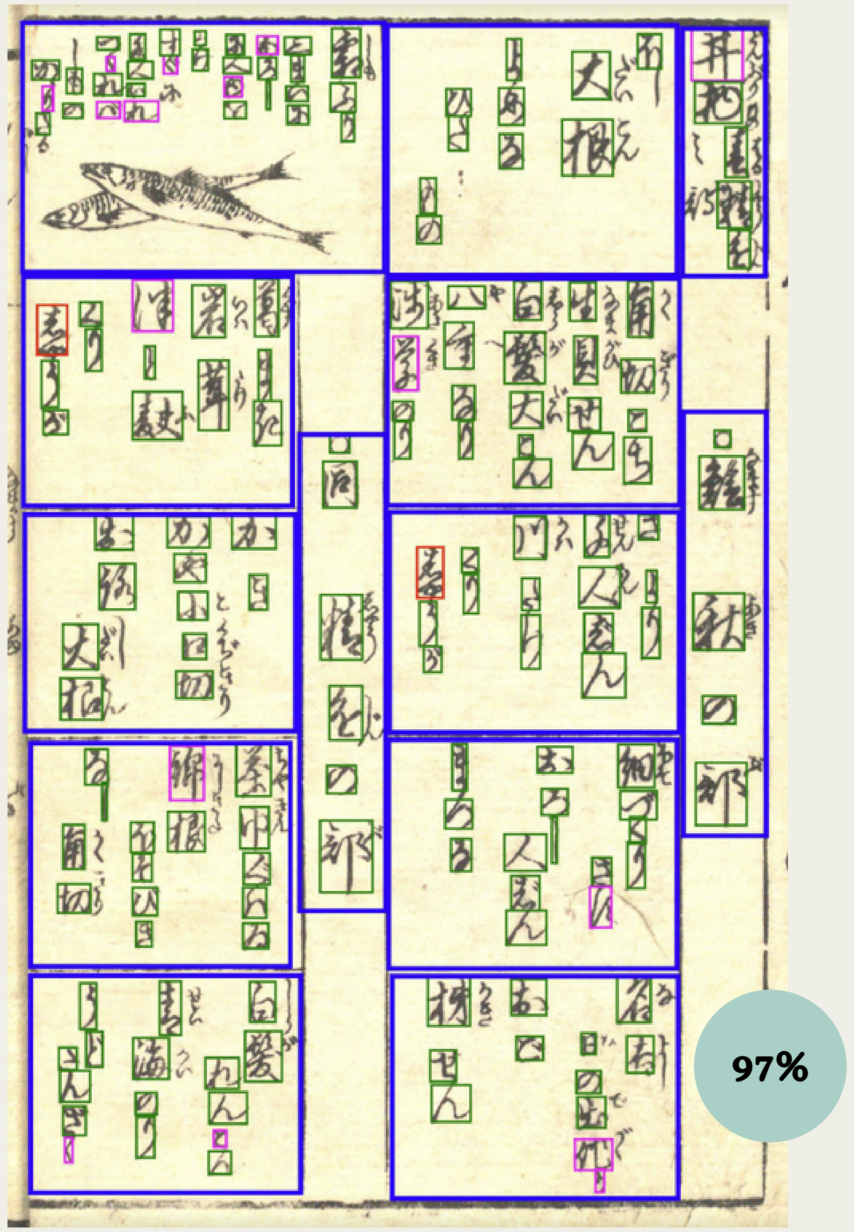
Input Image



Section Annotation



OCR & Clustering



97%

80%

Correction & Translation

Correction: 春の精進物部・大根おろしにひたし三まい、霜ふりほたゆをかけ、すゞを水にいれつゝす。あたかたのもの。秋の鱠部・角切生貝八重あさり、浅草せんなり、大根おろしに千人じんふりまつ。白髪せん名の吉日作り、音せんなり、細作りおりてくくり。おごり柿。同精進物部・葛まきつとくり、岩茸つとくり、くやきつとくり、茶巾ぐわいなしキ。白髪れんこん、青のりたんざく、海大根おろし。綿ぐせんす

Translation: Spring asceticism - soak three grated radishes, sprinkle with frosty mustard sauce, test by putting vinegar in water. A casual dish. In the Fall section - angle cut raw shellfish, eight layers of clams, Asakusa mustard, sprinkle with a grated radish of a thousand people. Good luck made on a lucky day with a white hair mustard, sound mustard, intricately made, tied. Persimmons to boast. The same asceticism - making kuzu, a kind of starch, chestnut, grilling rock mushrooms, grilling chestnuts, without a tea towel. White radish, green seaweed, sea radish grated. Wipe with cotton.

88%

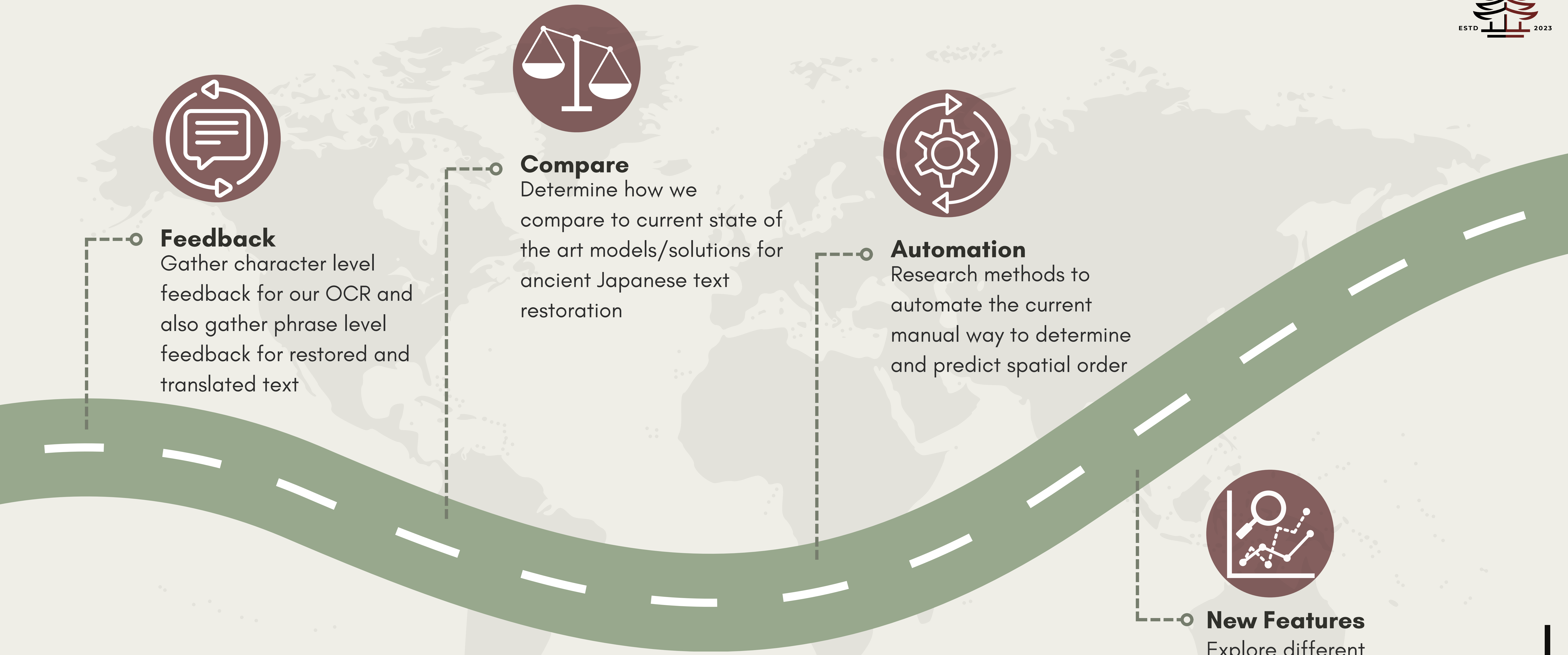
Overcoming Challenges

Reading order determination process

Understanding the nuances of spatial relationships in Kuzushiji characters

Integrating traditional aspects of ancient Japanese texts with cutting-edge technologies

Looking Ahead



Feedback
Gather character level feedback for our OCR and also gather phrase level feedback for restored and translated text

Compare
Determine how we compare to current state of the art models/solutions for ancient Japanese text restoration

Automation
Research methods to automate the current manual way to determine and predict spatial order

New Features
Explore different techniques to find and predict missing/faded characters in ancient texts

Restor-AI-tion: Preserving the past, illuminating the future

ESTD 2023

Contact: team@restor-ai-tion.net