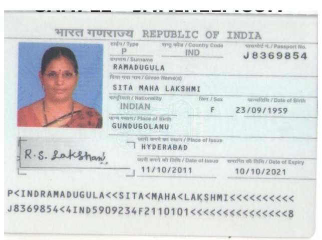
How do we build an In-house Government Document Understanding Service

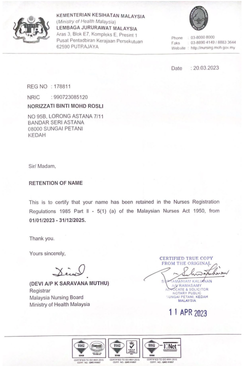
The technical development journey

GOVTECH
SINGAPORE

Direct Motivation – *PRS for MOH*

- ***Foreign Nurse Registration*** involves the submission and review of multiple documents (e.g., passports, nursing education transcripts, certificate verifications)
- All the documents are being processed ***manually***
- This leads to ***high error rates*** for mismatched inputs during submission, and ***low efficiency*** during review
- The MOH system is ***sensitive-high***
- PRS uses READ to automate this processing





How we got started – *initial architecture*

Scanned Documents → Off-the-shelf Document Analyzer API (e.g., AWS Textract, LLMs) → Structured Outputs

Textract

First name
LINDA

Document number
S7788888H

Expiration date
23 AUG 2010

Date of birth
27 JUN 1977

Date of issue
23 AUG 2005

ID type
PASSPORT

Place of birth
SINGAPORE

Mrz code
P<GPNG<<LINDA<ZEE<ZEE<<<<<<<<<<<<<<<<<<<<<<<
X2000444N4SGP7706275F1008232S7788888H<<<<<96

Okay! Mission Accomplished! But…

We need an ***in-house*** solution to process sensitive-high documents, that can function with no internet access

So, we cannot use 3rd-party Commercial Off-The-Shelf (COTS) solutions; even if we can, most of them cannot provide satisfactory performance.

Overall, our in-house solution is better than COTS!

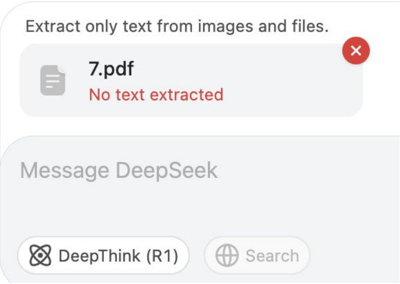
- >100 real document submissions from PRS were collected for comparison
- 3 different document types were tested
- Our in-house solution performs much better than COTS consistently across all types

Document Type	**AWS Textract (Commercial Off-The-Shelf)**	**Our Solution (In-house & On-prem)**
Passport	54/78 (69.2%)	76/78 (97.4%)
Transcripts of Nursing Education	8/16 (50.0%)	13/16 (81.3%)
Practising Certificate	11/18 (61.1%)	17/18 (94.4%)

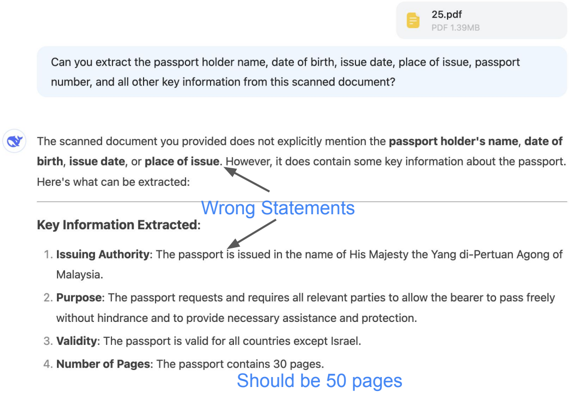
* The reported accuracy above is measured at the document level. It checks if all the text for target entities can be correctly OCRed/extracted for each document.

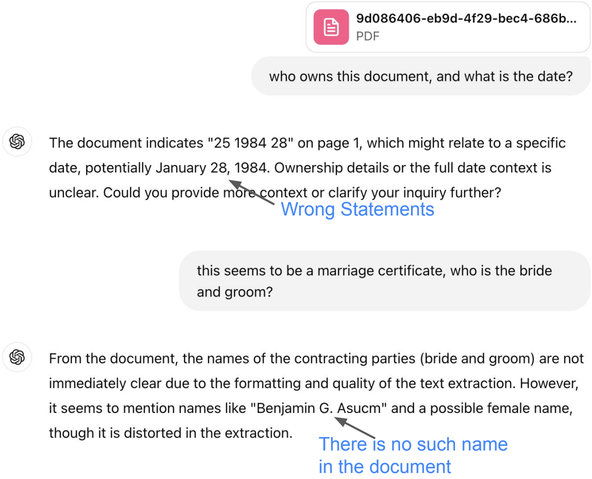
Many are also questioning: why not just use LLMs?

- Majority (>60%) docs incurred errors when uploading, preventing further questions



- The rest were even worse, with the LLM confidently returning incorrect results





* These screenshots were tested as of 2-Feb-2025 in DeepSeek-R1 and GPT-o1

Building an end-to-end solution took longer than expected

The implementation wasn’t as straightforward as we initially thought though. Here is the start…

We start from comparing locally-hosted OCR engines

Scanned Documents → Text Extraction with Local OCR Engine → Structured Outputs

Tesseract OCR

飞桨 PaddleOCR

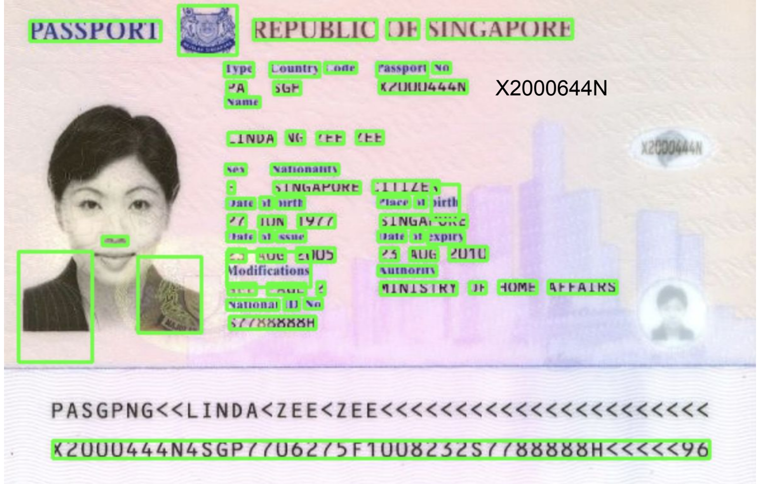
Easy OCR

* All above are under Apache, the local deployment would be suitable for even commercial usage

Try before making the choice – *local OCR comparison*

Tesseract OCR *

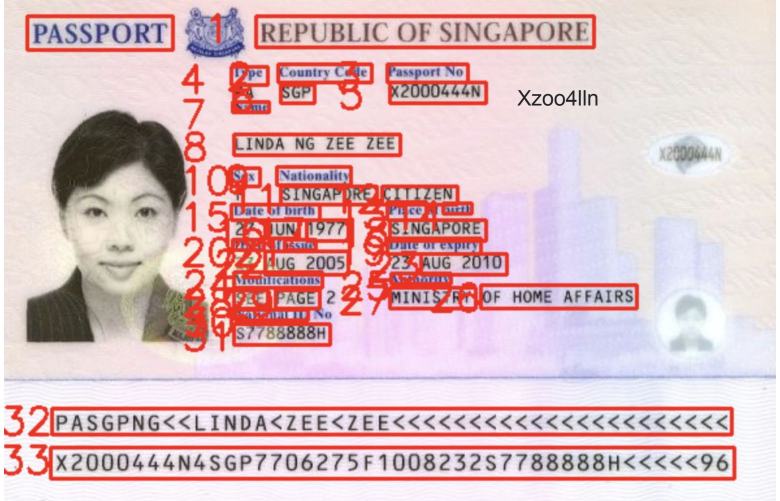
- One of the earliest open-source OCR engines funded in 1980s by HP; handed over to Google in 2006.
- LSTM-based algorithm for [line/word detection and classifier-based character recognition].
- Deep Learning framework was not popular that time, thus it wasn't used much.



- High text recognition error rate
- High chance of missing words
- Wrongly separate words within the same field

Try before making the choice – *local OCR comparison*

- Text detection is implemented using [CRAFT algorithm] with VGG-16 as the backbone in PyTorch.
- Text recognition is backboned by [CRNN], consisting of [ResNet] as feature extraction, [LSTM] as sequence encoder, and [CTC] as the decoder.
- The overall framework is modernized, but still doesn't benefit from the latest transformer-based architecture.

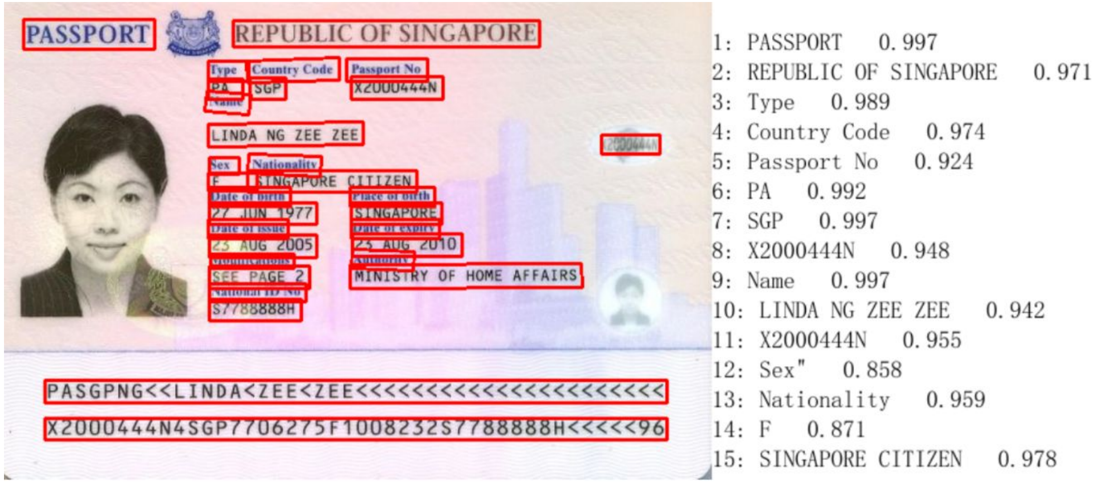


- Better at word detection
- Worse recognition error rate
- Better at finding long-field text, but still not satisfactory

Try before making the choice – *local OCR comparison*

飞桨 PaddleOCR

- Text detection by default uses DB++, a ResNet-50 backboned model with Differentiable Binarization and Adaptive Scale Fusion.
- Text recognition by default uses SVTR, a Vision Transformer-like backboned model, with faster inference and higher accuracy.
- An additional text direction classifier to deal with deskewed inputs.



- Almost perfect at text detection and recognition
- Runnable on CPUs with 2-3s per page, but much faster on GPU at <10ms per page

Larger-scale testing on bigger dataset

- We tried extracting the text fields from [EdisonTD dataset]*, using the three different local OCR engines
  - >200 passport images from various countries

- Paddle OCR did consistently well, while the performance from Tesseract and easyOCR were consistent too (on the negative side) :(

* EdisonTD is an open dataset covers containing information about travel and travel-related documents from almost every country on the globe

Now, we are ready to go!??

There is always a gap between expectation and reality

Expectation

ph.jpg

Reality

ph.pdf

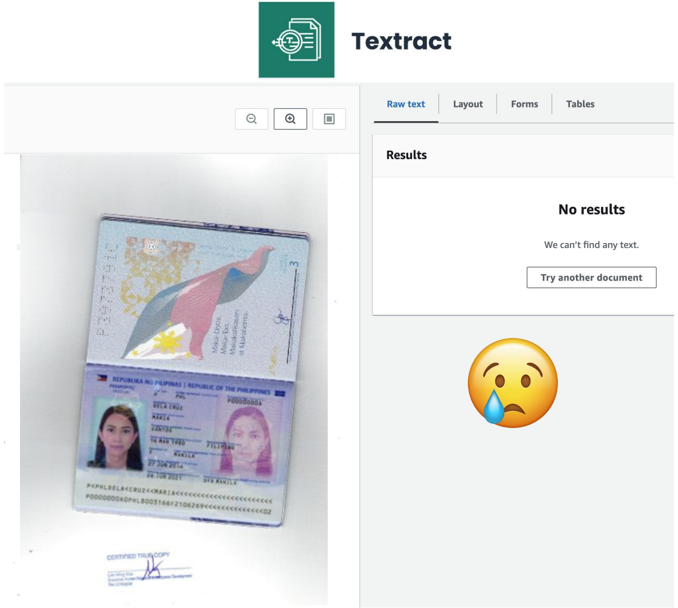
- Always in png or jpg format
- Perfectly cropped with a central view
- Highest possible scanning quality
- All samples are normalized to the same size
- No rotation or significant image skewing

- Most are in pdf format
- Actual image can be anywhere, with noisy background
- Scanning quality varies significantly
- Image files come in different sizes
- Rotations are very common with random angles

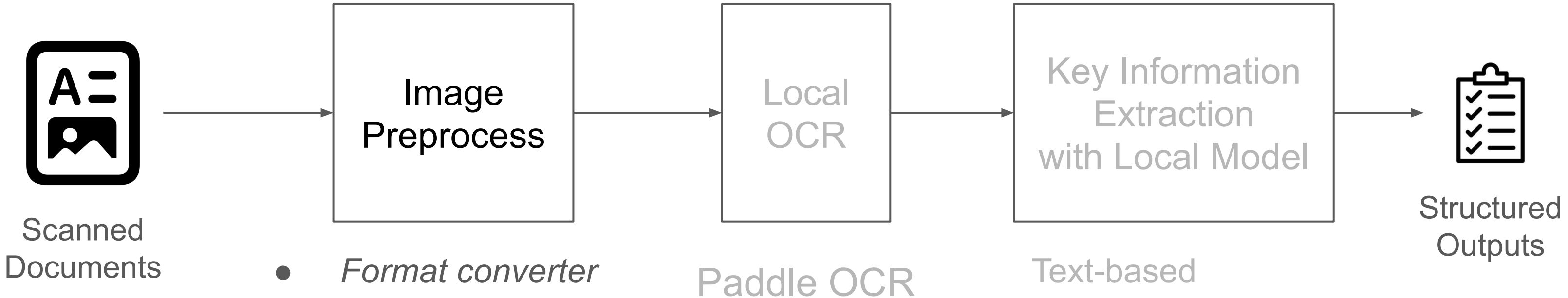
But technically, what do they mean?

- Straightforward ones
  - Inputs are in pdf format — Convert pdfs into jpg/png formats
  - Images are rotated — Deskew the image based on the text direction
  - Images come in different sizes — Normalize the size

- Less intuitive ones
  - Scanned images are low-quality, even mature commercial OCR engines cannot do the job
    - Computer vision tricks to enhance the quality



Add the Preprocess module into the pipeline



- *Format converter*
- *Size normalization*
- *Image cropping*
- *Re-normalization*
- *Image deskew*
- *Text enhancement*

Paddle OCR

Text-based
- BERT-Ner
- Spacy-Ner

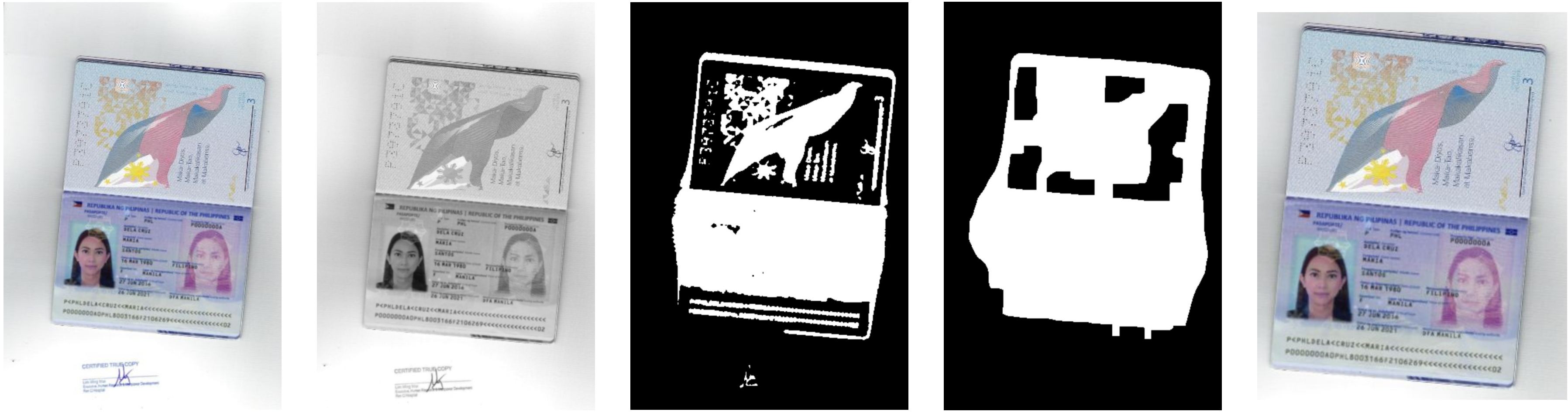
Multimodal-based
- LayoutLM-series

LLM-based
- Llama 3
- Phi-3

Zoom into image pre-processing

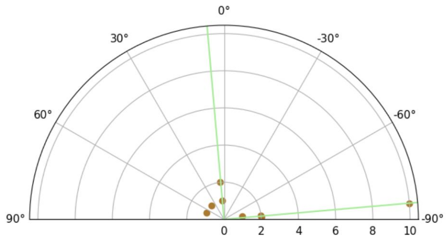
PDF → Image Loader / Converter → Normalize Image → Image Cropping → Image deskew → Re-normalize Image → Text Enhance

Zoom into image pre-processing – *image cropping*



Zoom into image pre-processing – *image deskew*

Cropped Image → Hough Transform → Calculate Skewed Angle → CNN-based 0° vs 180° Classifier



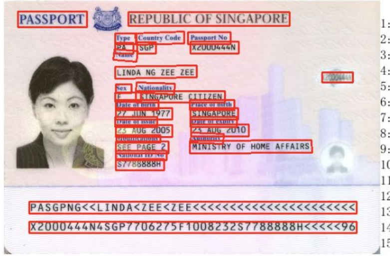
Zoom into image pre-processing – *text enhancement*

De-skewed Image → [Erode]() → [Dilate]() → [Blackhat]()

Now are we ready for production usage yet?

Yes, we get good OCR results, but what’s more?

- OCR
  - Detect text bounding boxes
  - Recognize text in each box

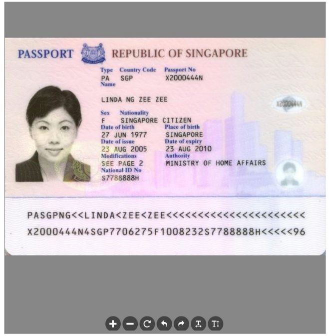


1: PASSPORT 0.997
2: REPUBLIC OF SINGAPORE 0.971
3: Type 0.989
4: Country Code 0.974
5: Passport No 0.924
6: PA 0.992
7: SGP 0.997
8: X2000444N 0.948
9: Name 0.997
10: LINDA NG ZEE ZEE 0.942
11: X2000444N 0.955
12: Sex" 0.858
13: Nationality 0.959
14: F 0.871
15: SINGAPORE CITIZEN 0.978
16: Date of birth 0.959
17: Place of birth 0.974
18: 27JUN 1977 0.953
19: SINGAPORE 0.997
20: Date of issue 0.984
21: Date of expiry 0.924
22: 23 AUG 2005 0.915
23: 23AUG 2010 0.945
24: Modifications 0.986
25: Authority 0.993
26: SEE PAGE 2 0.927
27: MINISTRY OF HOME AFFAIRS 0.946
28: National ID No 0.953
29: S7788888H 0.990
30: PASGPNG<<LINDA<ZEE<ZEE<<<<

- KIE
  - Accurately map relevant text fields into preloaded taxonomy
  - Discard irrelevant text fields from OCR

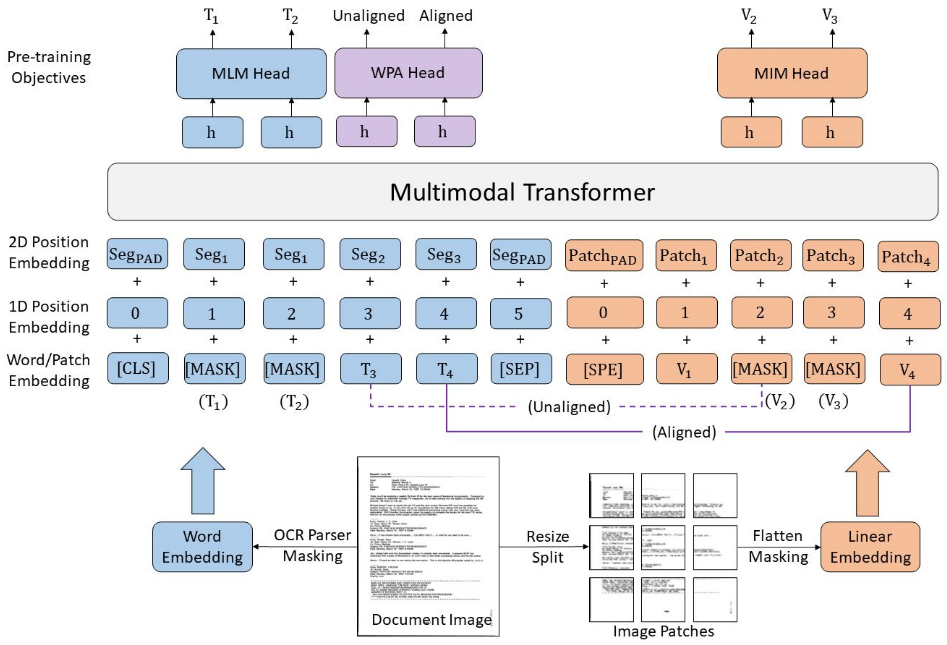
Document: PASSPORT

country	SGP
surname	NG
firstname	LINDA ZEE ZEE
sex	F
passport_number	X2000444N
birth_date	27 Jun 1977
expiration_date	23 Aug 2010
date_of_issue	23 Aug 2005
place_of_birth	SINGAPORE
passport_photo	true
personal_number	S7788888H
authority	MINISTRY OF HOME AFFAIRS



KIE using *LayoutLM* as a downstream task after OCR

- Pre-trained transformer for
  - Masked language model (MLM)
  - Masked image model (MIM)
  - Word patch alignment (WPA)
- A classification head for KIE in downstreaming usage
- In layman's terms:
  - The model uses multimodal info,
    - word meanings and positions
    - image content and positions
    - text-image alignment



* Reference: Huang Yupan, et al. "*Layoutlmv3: Pre-training for document ai with unified text and image masking*." ACM International Conference on Multimedia, MM 2022 [paper] [code]

Oh no! we are lacking training samples — *data augmentation*

Original Image

Border Crop

Add Noise

Rotation

Padding

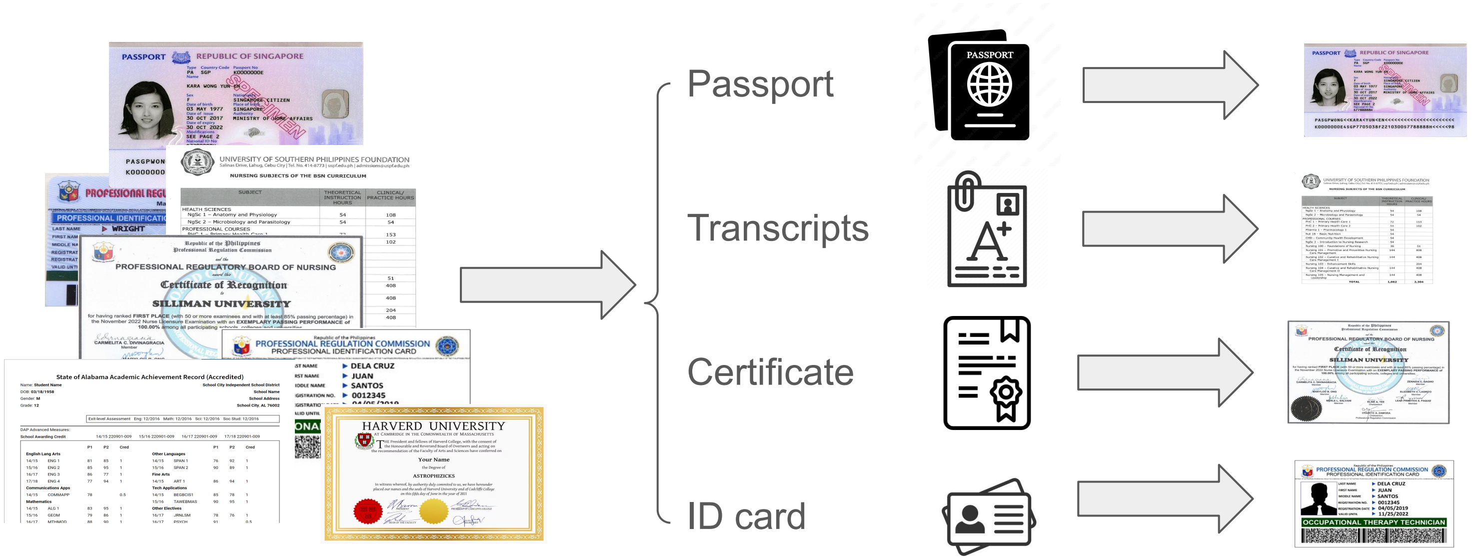
Perspective

Color Jitter

Posterized

And random combinations of them

We also need the right page from the right document

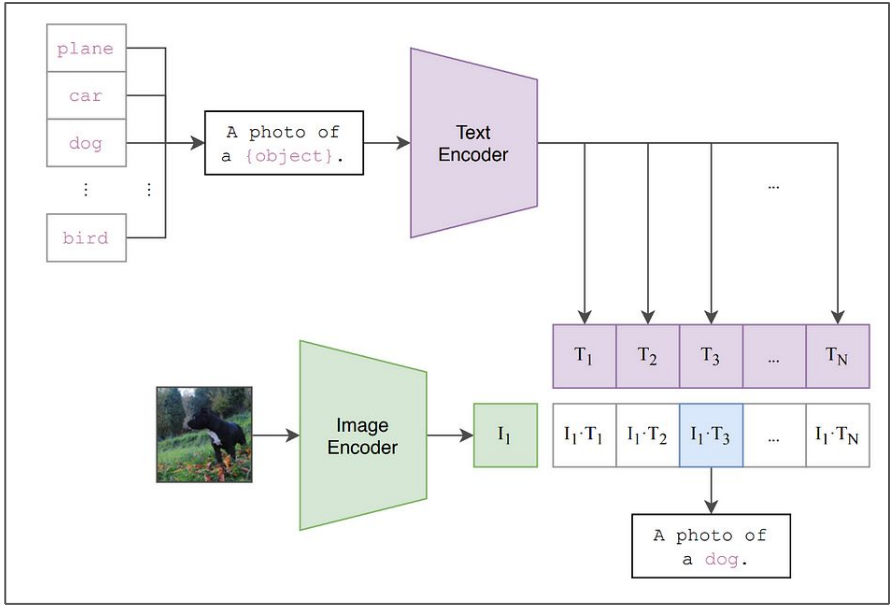


Input documents

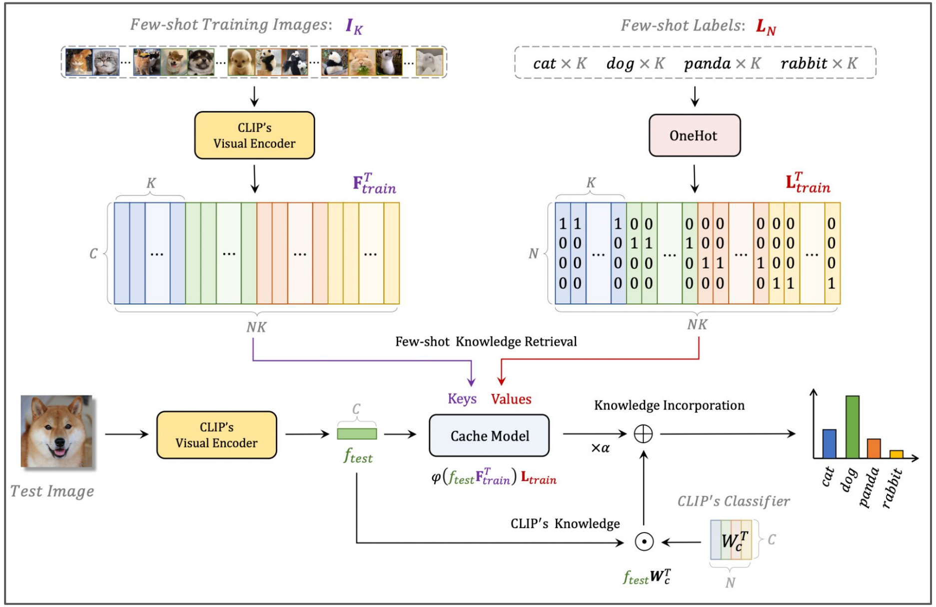
Stage 1. Document-level Classifier

Stage 2. Page-level Classifier

Document classifier – *vision features*



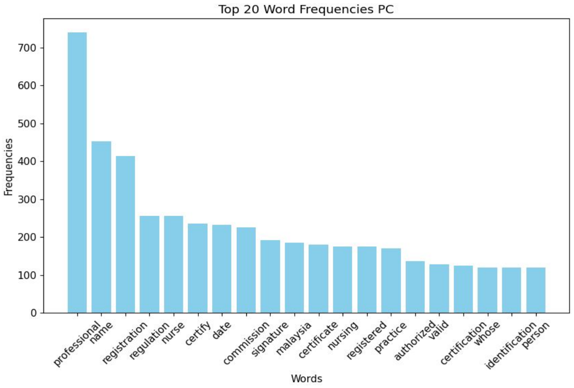
CLIP (Contrastive Image-Language Pretraining): Trained to match similarity between image and text caption pairs. Most similar caption used as prediction.

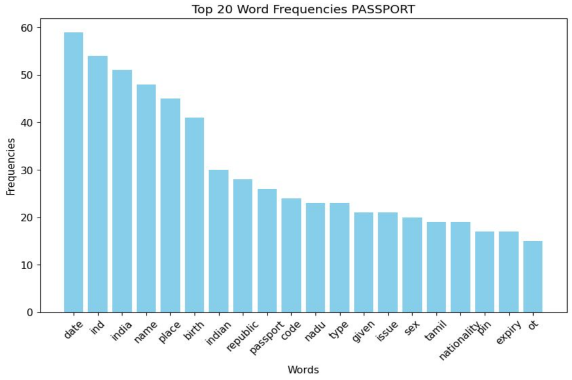


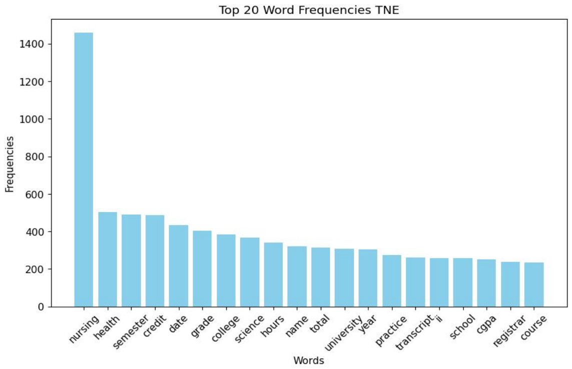
Tip-Adapter: Cache image features of a few training examples and their one-hot encoded labels. Take weighted sum of CLIP logits and Tip logits before making prediction.

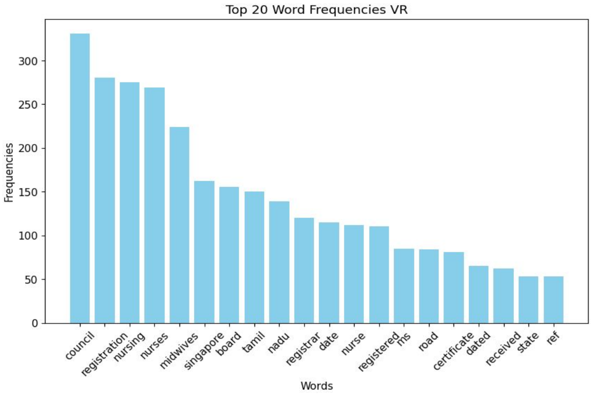
Document classifier – *text features and ensembles*

- TF-IDF
  - A statistical way to assign each word a score (in terms of frequency) to each class
- Ensemble
  - BaggingClassifier to combine vision and text classifier
  - Achieves ***97%*** F1 for in-house PRS documents

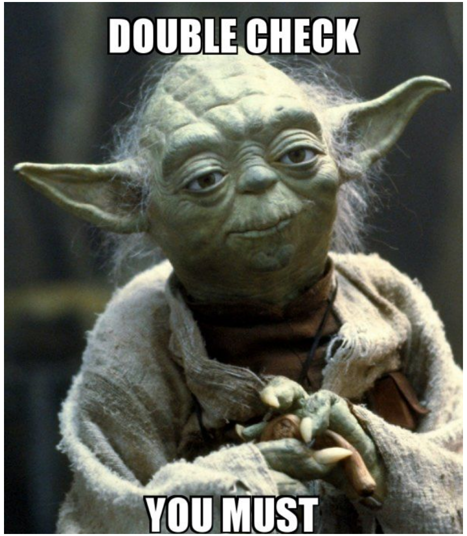




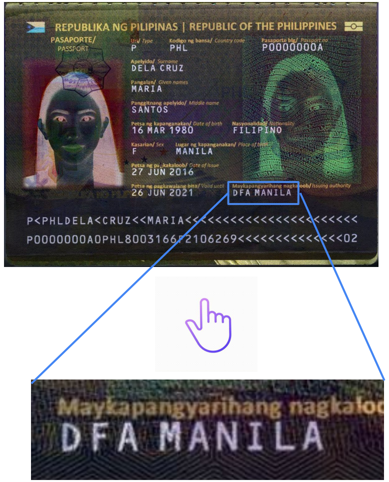




Zoom-in to double check low-confidence entities



- Double-confirm entities
  - *Some extractions could still have low confidence scores, especially for blurry entities*
  - *If confidence score is low, we zoom in and re-do a local OCR*



Further post-processing with domain knowledges

24/12/1990	1990-12-24
01/07/1980	1980-01-07
2000-12-24	2000-12-24
24 Nov 2005	2005-11-24
September 24, 2010	2010-09-24
Jul 24, 2016	2016-07-24
12/24/21	2021-12-24

- Date format normalization
  - *Different document type naturally has different date format*
  - *Even for the same document type, different country comes with different date format*
  - *Rules are added to normalize date according to document context*

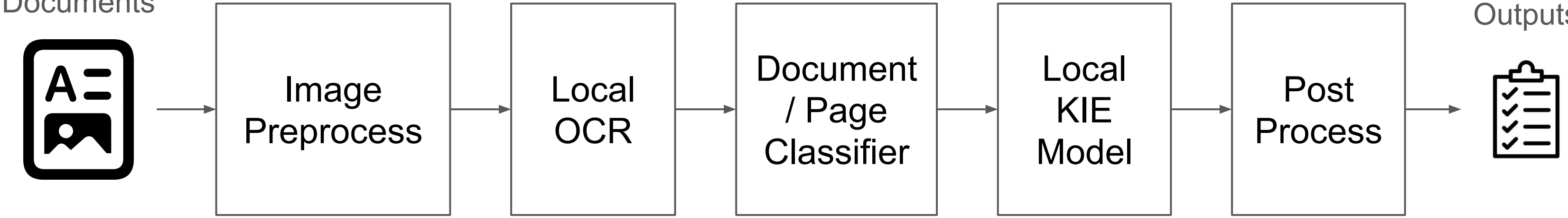
- Typo auto-correction & Field validation
  - *Quite often, OCR mis-recognizes one or two letters among long sentences*
  - *Maintain common words set and common OCR errors*
  - *Check Levenshtein distance and candidate frequency to auto-correct typos*

Now, the completed pipeline looks like this!

- Paddle OCR
- LayoutLMv3

Scanned Documents

Structured Outputs



- *Format converter*
- *Size normalization*
- *Image cropping*
- *Re-normalization*
- *Image deskew*
- *Text enhancement*

- *CLIP vision model*
- *TF-IDF text model*

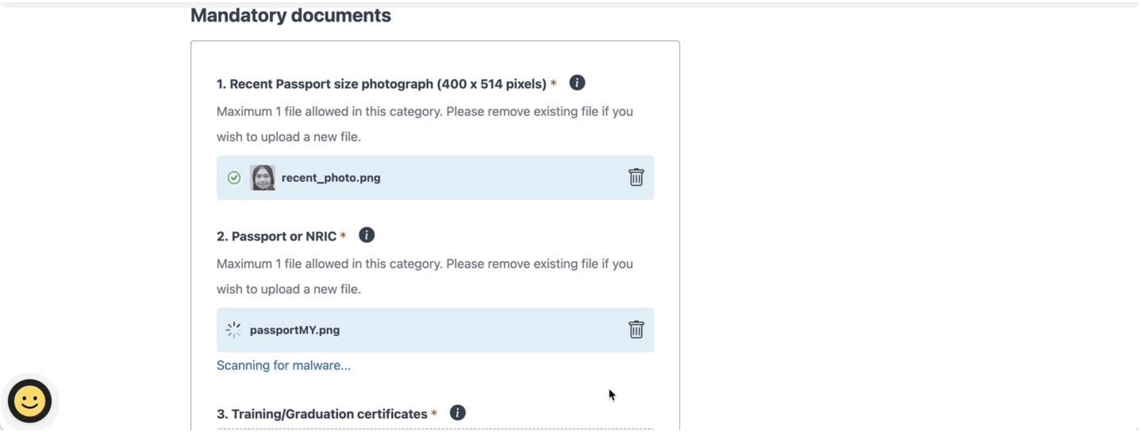
- *Zoom-in for double-confirm*
- *Date normalization*
- *Typo correction*
- *Field validation*

And here is the real production impact on PRS

- Integrated with PRS online submission system, providing ***real-time*** document processing with 1-2s latency
- Launched publicly **in June 2024**
- **18%** of applications were **prevented from being routed back** due to submitted information mismatches

SINGAPORE NURSING BOARD

MINISTRY OF HEALTH
SINGAPORE

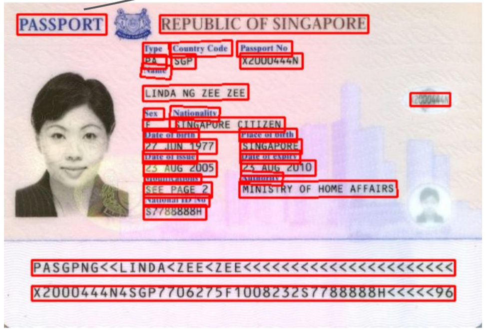


Can we move it further!

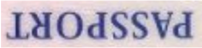
Let’s turn OCR into a white box – *optimize from inside out*

OCR Engine

Text Detection → Text Orientation Classifier → Text Recognition



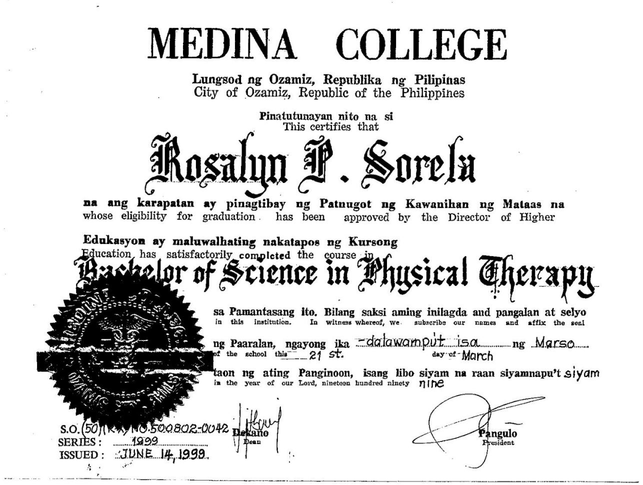
or

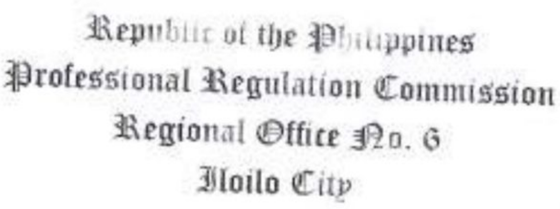


PASSPORT

Why fine-tune OCR models?

- Special and unique fonts are not uncommon
    - *Educational certificates / transcripts*
    - *Professional verification certificates*
    - *Marriage certificates*
- These cannot be reliably detected, classified, or recognized



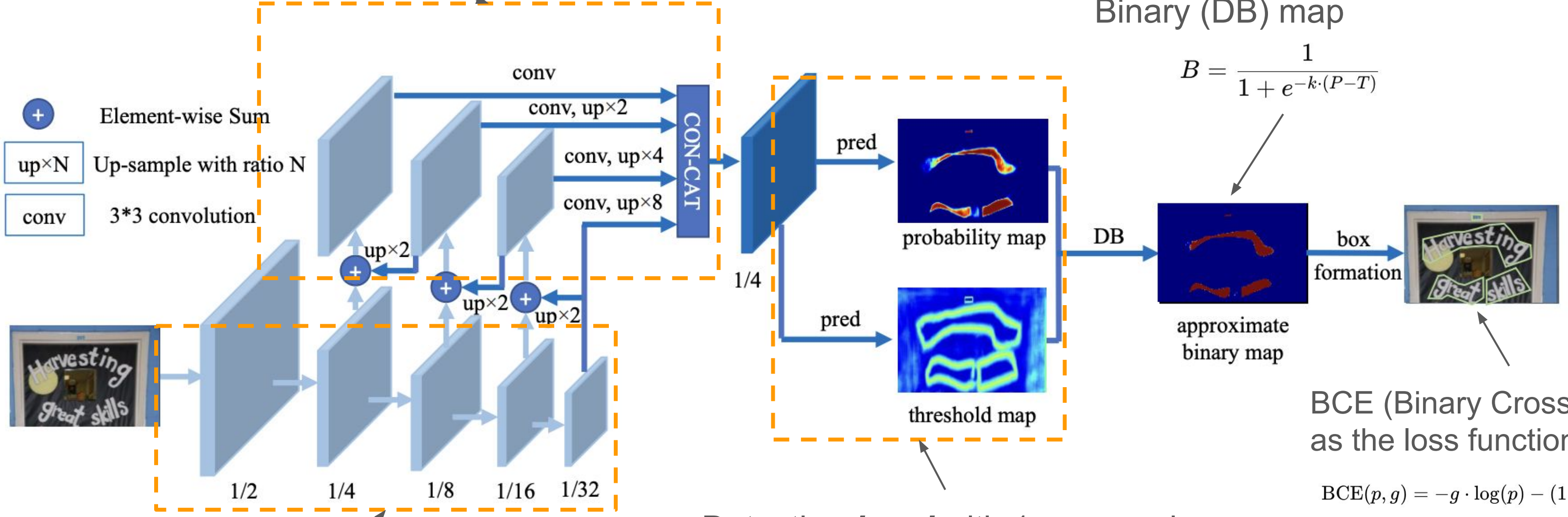


Fine-tune text detection model – *DB algorithm as an example*

Pre-trained ResNet-FPN (Feature Pyramid Network) as the ***neck*** for image feature enhancement

Calculate Differentiable Binary (DB) map

$$B = \frac{1}{1 + e^{-k \cdot (P - T)}}$$



BCE (Binary Cross-Entropy) as the loss function

$$\mathrm{BCE}(p, g) = -g \cdot \log(p) - (1 - g) \cdot \log(1 - p)$$

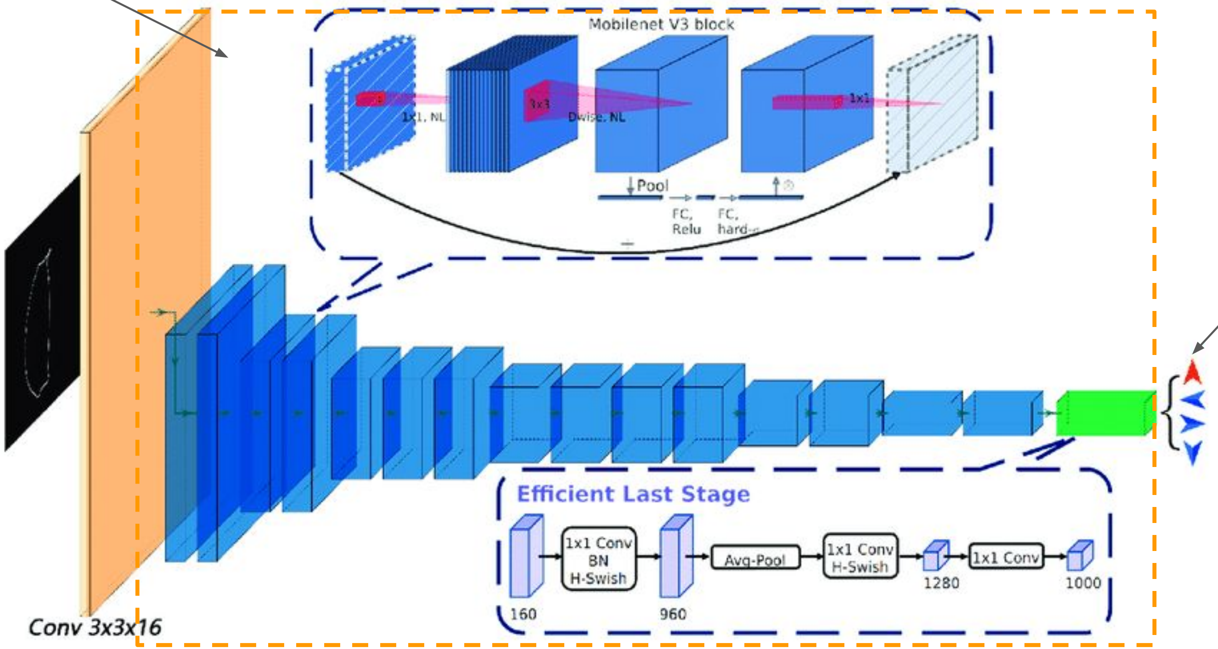
Pre-trained ResNet50 as the ***backbone*** for image feature extraction

Detection ***head*** with 1 conv and 2 de-conv layers each for probability and threshold map

* Reference: Minghui Liao, et al. "*Real-time Scene Text Detection with Differentiable Binarization.*" AAAI, 2020 [paper]

Fine-tune text direction classification model

Pre-trained light-weighted MobileNet-v3 as the ***backbone*** for image feature extraction



1 FC (Fully Connected) layer, followed by softmax as the classification ***head***, with Cross-Entropy as the loss

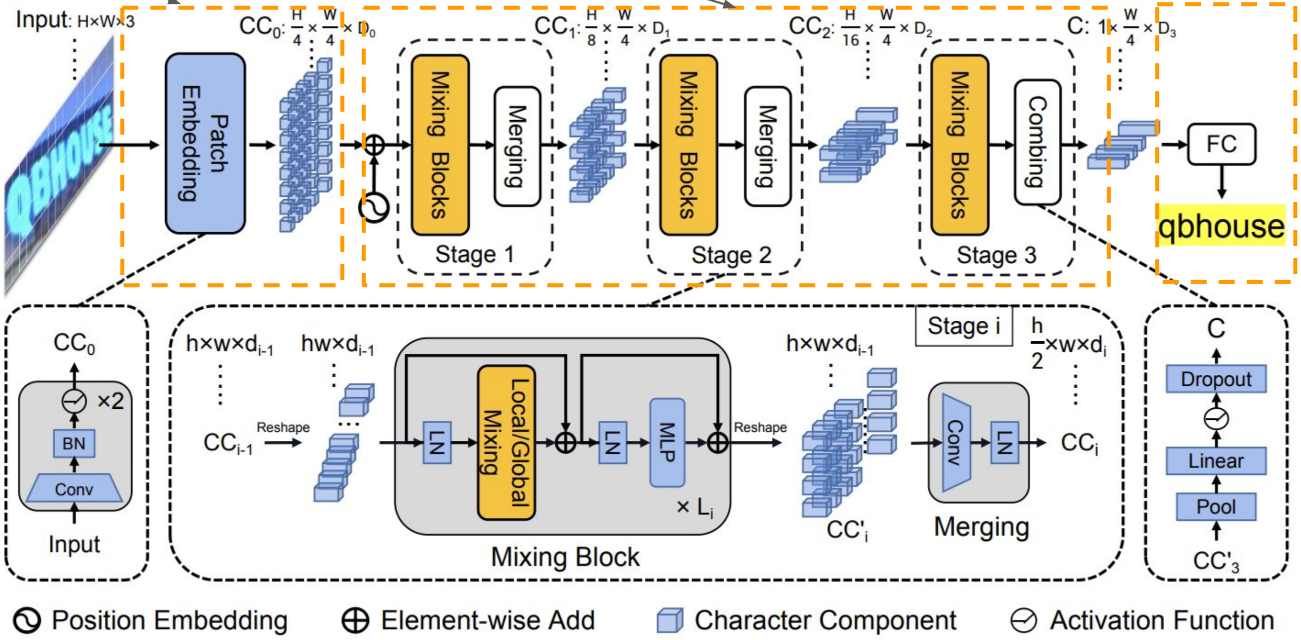
* Reference: Andrew Howard, et al. "*Searching for MobileNetV3*." ICCV, 2019 [paper]

Fine-tune text recognition model – *SVTR as an example*

ViT (Vision Transformer) based patch embedding as the ***backbone***

Height progressively decreased network as the ***neck*** to aggregate spatial features into sequential ones

A fully connected layer using CTC (Connectionist Temporal Classification) loss as the ***head***



* Reference: Yongkun Du, et al. "*SVTR: Scene Text Recognition with a Single Visual Model.*" IJCAI, 2022 [paper]

Blurry detection – *ask for immediate reupload if blurry*

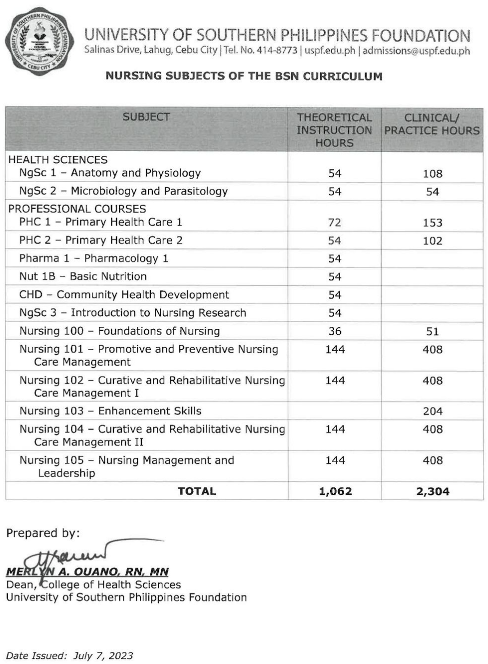
- The Laplacian operator is applied to an image by convolving the operator with each pixel

- The result of the convolution is a new image that highlights the edges in the original image

- ***A Laplacian variance*** can be used as a focus measure to differentiate blurry vs. in-focus images

Table detection & extraction – *dealing with tables*

- A two-step processing for tables
  - Object detection using [yolo] to locate tables, if any
  - Re-construct tables using table structure and cell locations, using [SLAnet]



UNIVERSITY OF SOUTHERN PHILIPPINES FOUNDATION
Salinas Drive, Lahug, Cebu City | Tel. No. 414-8773 | uspf.edu.ph | admissions@uspf.edu.ph

NURSING SUBJECTS OF THE BSN CURRICULUM

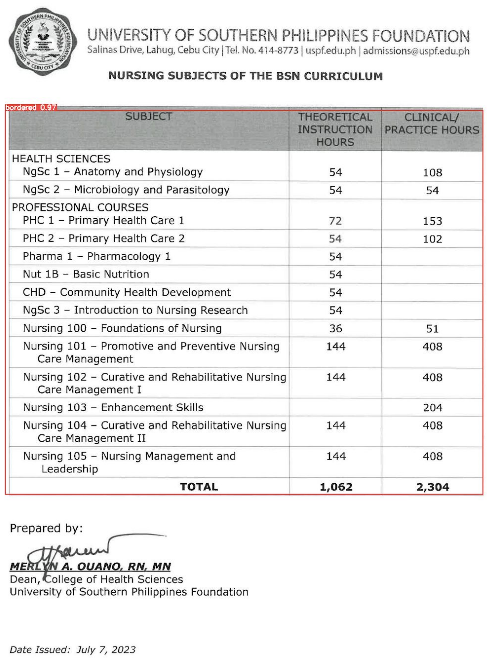
SUBJECT	THEORETICAL INSTRUCTION HOURS	CLINICAL/ PRACTICE HOURS
HEALTH SCIENCES NgSc 1 – Anatomy and Physiology	54	108
NgSc 2 – Microbiology and Parasitology	54	54
PROFESSIONAL COURSES PHC 1 – Primary Health Care 1	72	153
PHC 2 – Primary Health Care 2	54	102
Pharma 1 – Pharmacology 1	54	
Nut 1B – Basic Nutrition	54	
CHD – Community Health Development	54	
NgSc 3 – Introduction to Nursing Research	54	
Nursing 100 – Foundations of Nursing	36	51
Nursing 101 – Promotive and Preventive Nursing Care Management	144	408
Nursing 102 – Curative and Rehabilitative Nursing Care Management I	144	408
Nursing 103 – Enhancement Skills		204
Nursing 104 – Curative and Rehabilitative Nursing Care Management II	144	408
Nursing 105 – Nursing Management and Leadership	144	408
TOTAL	**1,062**	**2,304**

Prepared by:

MERLYN A. OUANO, RN, MN
Dean, College of Health Sciences
University of Southern Philippines Foundation

Date Issued: July 7, 2023

Table detection



UNIVERSITY OF SOUTHERN PHILIPPINES FOUNDATION
Salinas Drive, Lahug, Cebu City | Tel. No. 414-8773 | uspf.edu.ph | admissions@uspf.edu.ph

NURSING SUBJECTS OF THE BSN CURRICULUM

SUBJECT	THEORETICAL INSTRUCTION HOURS	CLINICAL/ PRACTICE HOURS
HEALTH SCIENCES NgSc 1 – Anatomy and Physiology	54	108
NgSc 2 – Microbiology and Parasitology	54	54
PROFESSIONAL COURSES PHC 1 – Primary Health Care 1	72	153
PHC 2 – Primary Health Care 2	54	102
Pharma 1 – Pharmacology 1	54	
Nut 1B – Basic Nutrition	54	
CHD – Community Health Development	54	
NgSc 3 – Introduction to Nursing Research	54	
Nursing 100 – Foundations of Nursing	36	51
Nursing 101 – Promotive and Preventive Nursing Care Management	144	408
Nursing 102 – Curative and Rehabilitative Nursing Care Management I	144	408
Nursing 103 – Enhancement Skills		204
Nursing 104 – Curative and Rehabilitative Nursing Care Management II	144	408
Nursing 105 – Nursing Management and Leadership	144	408
TOTAL	**1,062**	**2,304**

Prepared by:

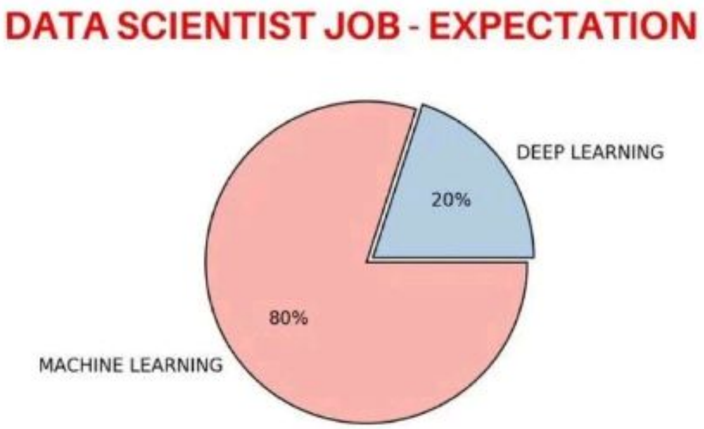
MERLYN A. OUANO, RN, MN
Dean, College of Health Sciences
University of Southern Philippines Foundation

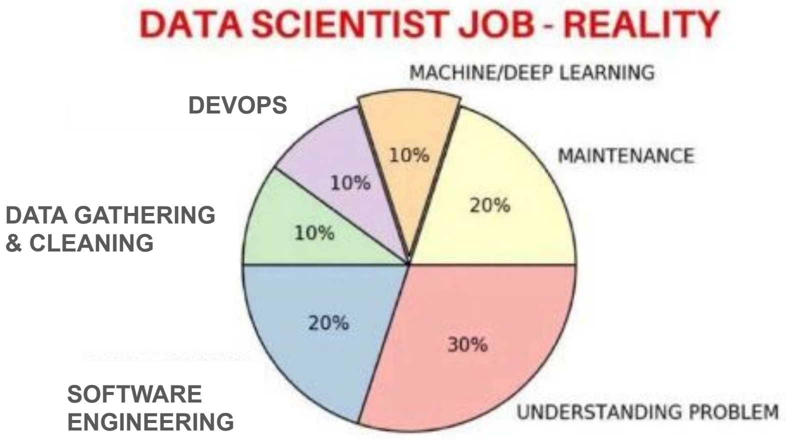
Date Issued: July 7, 2023

Reconstruct tables

SUBJECT	THEORETICAL INSTRUCTION HOURS	CLINICAL/ PRACTICE HOURS
HEALTH SCIENCES NgSc 1 - A...	54.0	108.0
NgSc 2 - Microbiology and ...	54.0	54.0
PROFESSIONAL COURSES	NaN	NaN
PHC 1 - Primary Health Care 1	72.0	153.0
PHC2-Primary Health Care 2	54.0	102.0
Pharma 1 - Pharmacology 1	54.0	NaN
Nut 1B-Basic Nutrition	54.0	NaN
CHD Community Health Devel...	54.0	NaN
NgSc 3 - Introduction to N...	54.0	NaN
Nursing 100 - Foundations ...	36.0	51.0
Nursing 101 - Promotive an...	144.0	408.0
Nursing 102 - Curative and...	144.0	408.0
Nursing 103-Enhancement Sk...	NaN	204.0
Nursing 104 - Curative and...	144.0	408.0
Nursing 105 - Nursing Mana...	144.0	408.0
TOTAL	1062.0	2304.0

End-to-End delivery with a team of 3 data scientists





Tech Stack

Language	Python, Typescript, Shell
Backend	FastAPI, Uvicorn, PyTorch, Transformers, OpenCV, PaddleOCR, OpenClip, Spacy, XGBoost, GliNER PyMuPDF, Ultralytics, PyTest
DevOps	GitLab CI/CD, Kubernetes, Docker, CloudWatch, Parameter Store
Cloud	AWS S3, ECR, ECS (CPU Fargate / GPU), ALB+ASG, SageMaker, NAT Gateway, Lambda, EventBridge

Engineering practices – *quality assurance with enough tests*

Unit Test

Ensures each individual unit works as expected.

Integration Test

Ensures the end-to-end outputs align with our expectations.

Regression Test

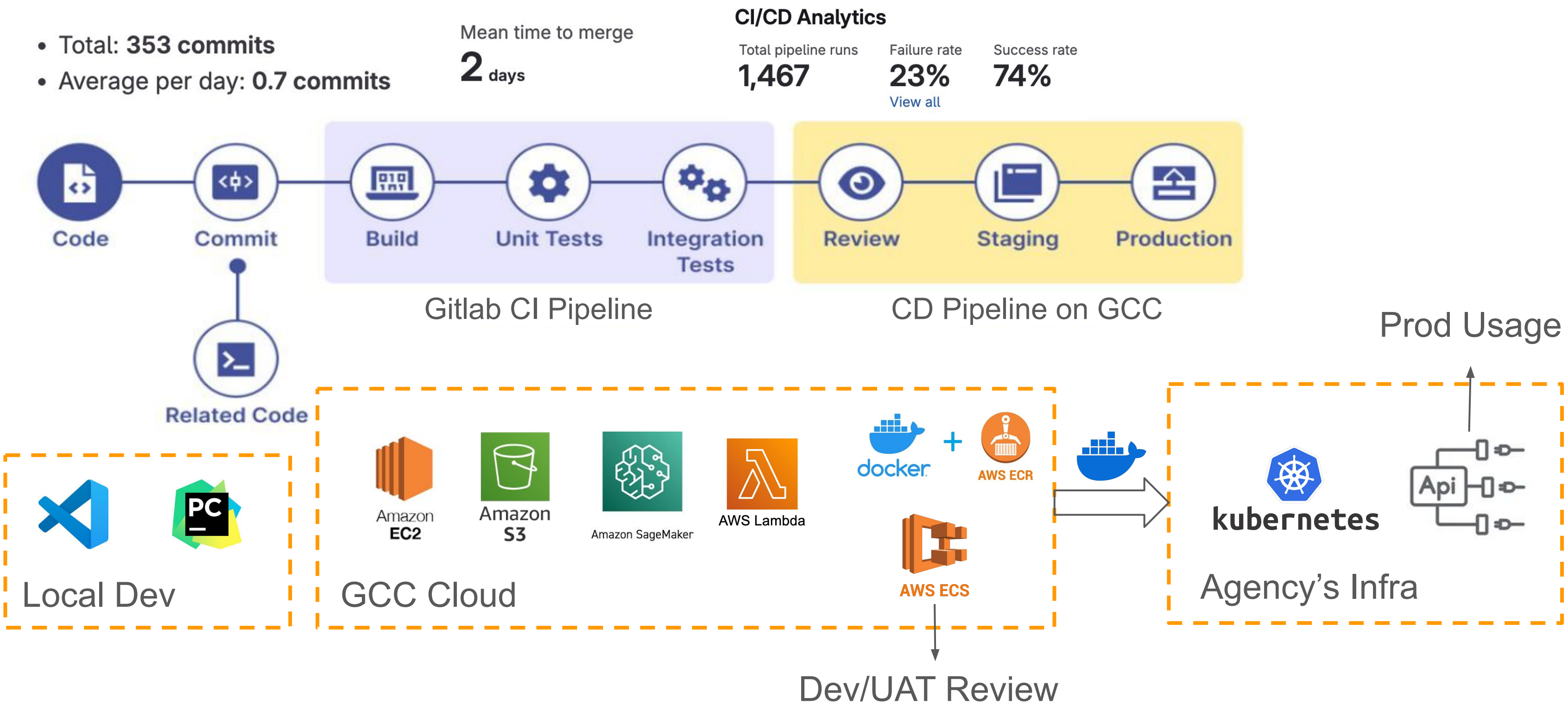
Ensures any new logic, models, or features won't cause a decline in the functionality of existing systems.

Performance Test

Ensures the latency and scalability under certain workload.

```
============================================================== test session starts ==============================================================
platform darwin -- Python 3.11.4, pytest-8.2.0, pluggy-1.5.0 -- /Users/yichao/Documents/document-understanding-service/.venv/bin/python
cachedir: .pytest_cache
rootdir: /Users/yichao/Documents/document-understanding-service
configfile: pytest.ini
plugins: anyio-3.7.1
collected 2083 items
```

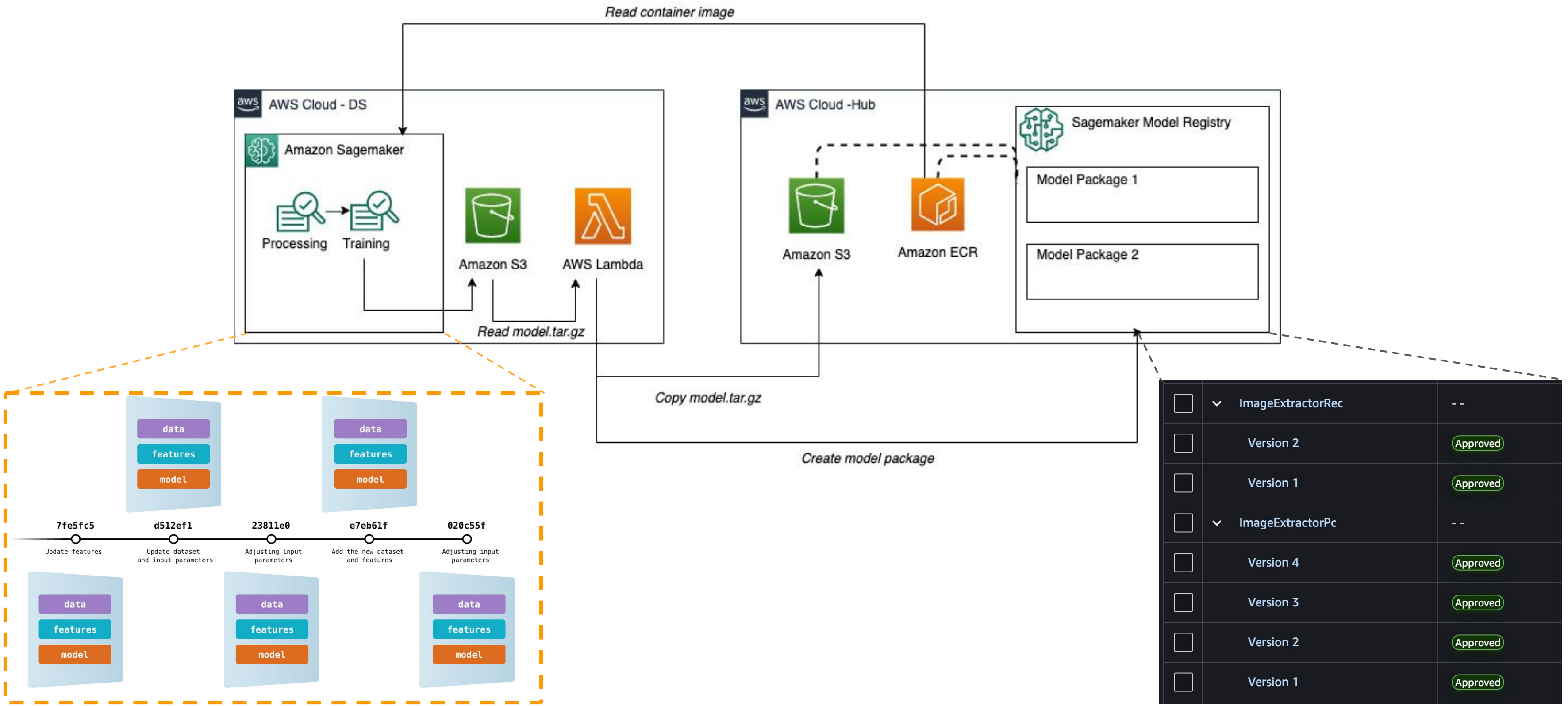
Engineering practices – *CI/CD workflow for in-house services*



Engineering practices – *real-time metrics monitor dashboard*

- One-single place to know your deployed service
  - Overall product metrics
  - Technical metrics
  - Detailed views with breakdown from different document types
  - Real-time alarming if anything goes wrong

Engineering practices – *model iteration and versioning*



Business highlights – *unique advantages of READ*

- Security Compliance → Able to serve **sensitive-high** documents (e.g. medical reports, PII, etc)
- Vision Extractions → Able to serve **both text and visual data** (e.g. photo, signature, stamp, medical plot, etc)
- Cost Efficiency → Fixed infra cost means lower unit price (much lower than COTS) with big volume
- Accuracy & speed → Able to **achieve >95% accuracy** with **1-2s end-to-end delay***

* the accuracy and latency is reported based on the documents from our SNB/MOH use case, which is significantly better than COTS

What are our next steps?

Product expansion – *potential use cases beyond PRS*

- Agencies handling complex documents/images/video preferably in sensitive-high systems with large volume

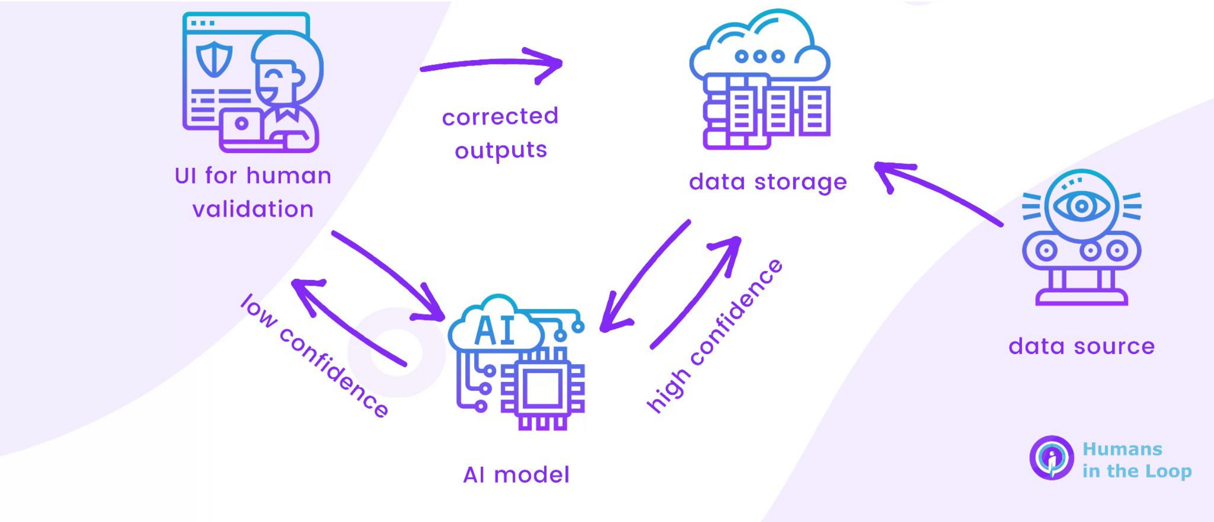
MINISTRY OF HEALTH SINGAPORE
ICA Securing Our Borders, Safeguarding Our Home
DSTA Defence Science & Technology Agency
MINDEF SINGAPORE
CENTRAL PROVIDENT FUND BOARD SINGAPORE

- Enable complex document/image/video understanding capabilities for other GovTech platform services

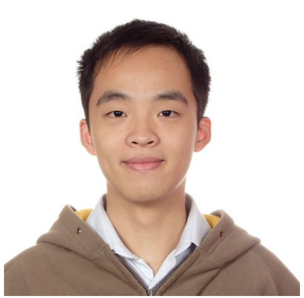
VAS
CLOAK The Central Privacy Toolkit
AIBots

Technical advance – *an AI agent to learn from the mistakes*

- Complete the loop to automate model iteration with human-corrected outputs
  - *So that the same mistakes won't repeat again and again*
- All agencies can benefit from model iteration, without actually sharing the data



Interested to know more? – *the team is ready for questions*



JIN Yichao
Data Scientist

CHONG Zi Kang
Data Scientist

SHEN Lin
Data Scientist

jin.yichao@gt.tech.gov.sg

chong.zi.kang@gt.tech.gov.sg

shen.lin@gt.tech.gov.sg

GOVTECH
SINGAPORE