Optical Character Recognition: **Converting** text in an image into a machine-readable format

motivation

✅ OCR widely studied for high-resource languages like English and Chinese

❌ **Limited understanding** on effectiveness and limitations of OCR on **Southeast Asian (SEA) languages**

- Low-resource: Less training data
- SEA population: 700 million people

Current status of NLP in South East Asia with insights from multilingualism and language diversity (Aji et al., 2023)

motivation

Studying OCR for SEA languages is **valuable**.

- Enhanced **digital accessibility** for low-resource language users
- **Resource creation** for downstream Natural Language Processing tasks
  - E.g., machine translation, named-entity recognition

OCR improves machine translation for low-resource languages (Ignat et al., 2022)

Benchmarking and Improving OCR Systems for Southeast Asian Languages

Project ID: H0792230
Supervisor: A/P Min-Yen Kan
Advisor: Tongyao Zhu
Student Name: Qiu Jiasheng, Jason

RQ1: How do popular OCR tools **perform** on SEA scripts?

RQ2: What **script-related challenges** affect OCR accuracy on SEA languages?

RQ3: How can fine-tuning **enhance** OCR accuracy on SEA languages?

roadmap

01 Motivation

02 Related Work

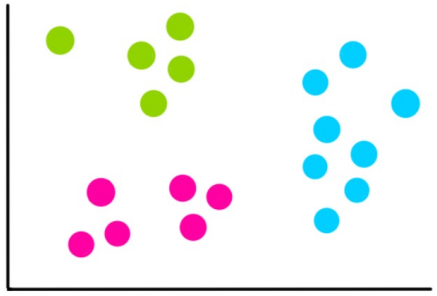
03 Experiment 1: Benchmarking on Real-world Wikipedia **Screenshots**

04 Experiment 2: Benchmarking on **Synthetic** Data

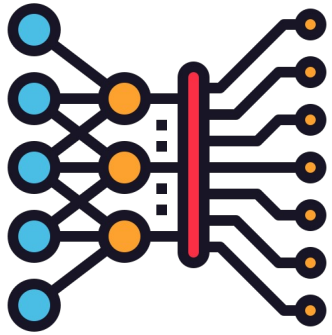
05 Experiment 3: **Fine-tuning**

related work

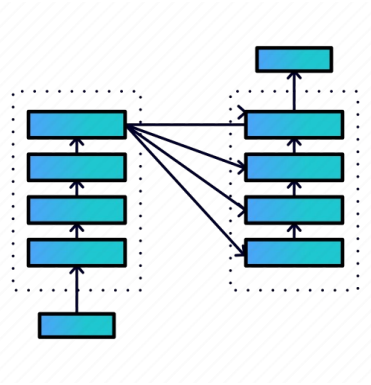
OCR model architectures are **evolving**.



Traditional machine learning (KNN, SVM) on manual features[1]



Deep learning (CNN, RNN)[1]



Transformer-based architecture[2]

Our research: Compares performance of **different OCR models** on SEA languages

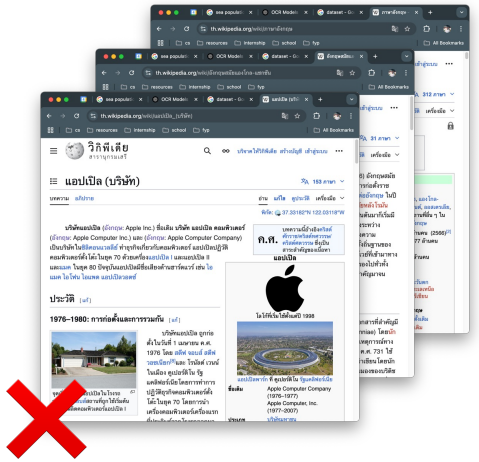
[1] A survey of deep learning approaches for OCR and document understanding (Subramani et al., 2020)
[2] Fine-tuning vision encoder-decoder transformers for handwriting text recognition on historical documents (Parres & Paredes, 2023)

related work

OCR on SEA languages is **challenging**.

ทุกคนมีสิทธิที่จะออกจากประเทศใด ๆ ไป
รวมทั้งประเทศของตนเองด้วย และที่จะ
กลับยังประเทศตน

SEA scripts are more
complex



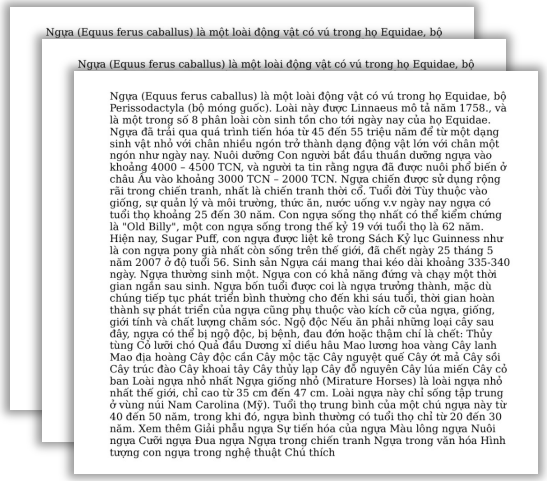
Lack of quality datasets

Our research: Develops a reusable **pipeline** for **collecting** real-world data and **generating** synthetic data

A concise survey of OCR for low-resource languages (Agarwal & Anastasopoulos, 2024)

related work

Synthetic data with noise helps with data scarcity.



Easy to generate large-scale, quality data for low-resource languages

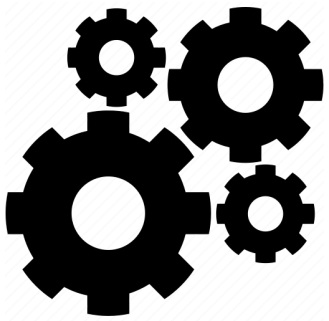
Adding **noise** better simulates real-world conditions

Our research: Adopts **similar** techniques by generating and benchmarking on synthetic data with noise

Lights, camera, action! A framework to improve NLP accuracy over OCR documents (Gupte et al., 2021)

related work

Fine-tuning can adapt OCR models to new languages.



Fine-tuning: **Further training** pre-trained models on task-specific dataset

Retains prior knowledge + Adapts to new dataset

↓

Limited data can achieve **good** results

Our research: **Fine-tunes** existing model on SEA languages

Fine-tuning vision encoder-decoder transformers for handwriting text recognition on historical documents (Parres & Paredes, 2023)

roadmap

01 Motivation

02 Related Work

03 Experiment 1: Benchmarking on Real-world Wikipedia **Screenshots**

04 Experiment 2: Benchmarking on **Synthetic** Data

05 Experiment 3: **Fine-tuning**

Experiment **setup**

Experiment 1: Benchmarking on Real-world Wikipedia **Screenshots**

Experiment 2: Benchmarking on **Synthetic** Data

Experiment 3: **Fine-tuning**

1. Methodology
2. Results
3. Discussion

experiment setup

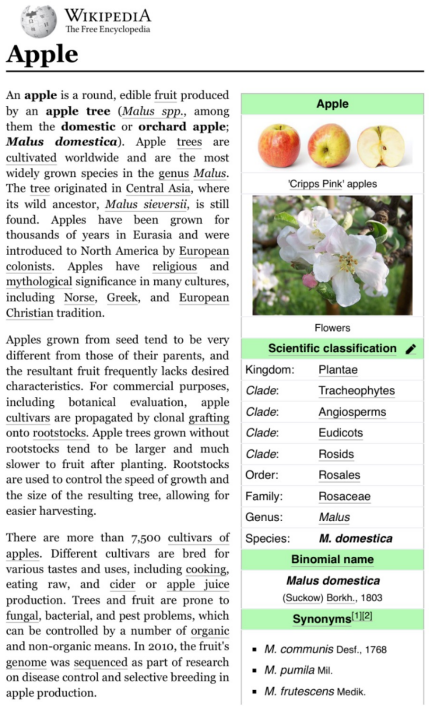
4 languages: English, Indonesian, Vietnamese, Thai

	Speaker population	Script Type	Example
English	1.5 billion	Latin	Good morning
Indonesian	252 million	**Latin**	Selamat pagi
Vietnamese	97 million	Latin with **diacritics**	Chào buổi sáng
Thai	71 million	**Brahmic**	สวัสดีตอนเช้า

Why?

1. Capture diverse range of **script types**
2. Wide usage makes it feasible to obtain **data**

Data source: **Wikipedia**



WIKIPEDIA
The Free Encyclopedia

Apple

An **apple** is a round, edible fruit produced by an **apple tree** (*Malus spp.*, among them the **domestic** or **orchard apple**; ***Malus domestica***). Apple trees are cultivated worldwide and are the most widely grown species in the genus *Malus*. The tree originated in Central Asia, where its wild ancestor, *Malus sieversii*, is still found. Apples have been grown for thousands of years in Eurasia and were introduced to North America by European colonists. Apples have religious and mythological significance in many cultures, including Norse, Greek, and European Christian tradition.

Apples grown from seed tend to be very different from those of their parents, and the resultant fruit frequently lacks desired characteristics. For commercial purposes, including botanical evaluation, apple cultivars are propagated by clonal grafting onto rootstocks. Apple trees grown without rootstocks tend to be larger and much slower to fruit after planting. Rootstocks are used to control the speed of growth and the size of the resulting tree, allowing for easier harvesting.

There are more than 7,500 cultivars of apples. Different cultivars are bred for various tastes and uses, including cooking, eating raw, and cider or apple juice production. Trees and fruit are prone to fungal, bacterial, and pest problems, which can be controlled by a number of organic and non-organic means. In 2010, the fruit's genome was sequenced as part of research on disease control and selective breeding in apple production.

Apple

'Cripps Pink' apples

Flowers

Scientific classification

Kingdom:	Plantae
Clade:	Tracheophytes
Clade:	Angiosperms
Clade:	Eudicots
Clade:	Rosids
Order:	Rosales
Family:	Rosaceae
Genus:	*Malus*
Species:	***M. domestica***

Binomial name

Malus domestica
(Suckow) Borkh., 1803

Synonyms[1][2]

- *M. communis* Desf., 1768
- *M. pumila* Mil.
- *M. frutescens* Medik.

วิกิพีเดีย
สารานุกรมเสรี

แอปเปิล

แอปเปิล (อังกฤษ: apple; ชื่อวิทยาศาสตร์: *Malus domestica*) เป็นต้นไม้ผลัดใบในวงศ์กุหลาบ มีผลรสหวานเรียกว่า **ผลแอปเปิล** แอปเปิลมีปลูกอยู่ทั่วโลกในลักษณะของไม้ผล และสายพันธุ์ที่ถูกปลูกมากที่สุดคือสกุล *Malus* ต้นแอปเปิลมีต้นกำเนิดในเอเชียกลาง ซึ่งบรรพบุรุษคือ *Malus sieversii* ยังคงพบได้ในปัจจุบัน แอปเปิลมีปลูกเป็นเวลาหลายพันปีในเอเชียและยุโรป และกลุ่มอาณานิคมชาวยุโรปนำมาปลูกที่อเมริกาเหนือ แอปเปิลมีความสำคัญทางศาสนาและเทพปกรณัมในหลายวัฒนธรรม รวมถึงนอร์ส กรีก และประเพณีต่าง ๆ ของคริสต์ศาสนิกชนของชาวยุโรป

ต้นแอปเปิลจะมีขนาดใหญ่หากเติบโตจากเมล็ด แต่จะมีขนาดเล็กถ้าถูกตัดต่อเนื้อเยื่อเข้ากับราก ปัจจุบันมีแอปเปิลที่พันธุ์ปลูกมากกว่า 7,500 ชนิด ทำให้แอปเปิลมีลักษณะพิเศษหลากหลาย พันธุ์ปลูกแต่ละพันธุ์จะมีรสชาติแตกต่างกัน และการนำไปใช้ต่างกันด้วย เช่น นำไปประกอบอาหาร กินดิบ ๆ หรือนำไปผลิตไซเดอร์ ปกติแอปเปิลจะแพร่พันธุ์ด้วยการตัดต่อเนื้อเยื่อ แต่แอปเปิลป่าจะเติบโตได้เองจากเมล็ด ต้นแอปเปิลและผลแอปเปิลอาจประสบปัญหาจากเห็ดรา แบคทีเรีย และศัตรูพืชต่าง ๆ ซึ่งอาจควบคุมได้ด้วยวิธีการทางเกษตรอินทรีย์และอนินทรีย์หลายวิธี ใน ค.ศ. 2010 มีการถอดรหัสจีโนมของแอปเปิล เป็นส่วนหนึ่งของงานวิจัยควบคุมโรคและการคัดเลือกผสมพันธุ์ในการผลิตแอปเปิล

ใน ค.ศ. 2013 มีการปลูกแอปเปิลประมาณ 80 ล้านตันขึ้นทั่วโลก ประเทศจีนผลิตได้จำนวนเกือบครึ่งหนึ่งของจำนวนดังกล่าว[3] สหรัฐอเมริกาเป็นผู้ผลิตแอปเปิลมากที่เป็นอันดับที่สอง ด้วยการผลิตมากกว่า 6% ประเทศตุรกีเป็นที่สาม ตามด้วยประเทศอิตาลี อินเดีย และโปแลนด์ แอปเปิลมักนิยมกินดิบ แต่สามารถพบได้ในอาหารที่เตรียมขึ้น (โดยเฉพาะของหวาน) และเครื่องดื่ม มีความคิดว่า

แอปเปิล

แอปเปิลพันธุ์ 'คริปส์พิงก์'

ดอก

การจำแนกชั้นทางวิทยาศาสตร์

โดเมน:	ยูแคริโอต Eukaryota
อาณาจักร:	พืช Plantae
เคลด:	พืชมีท่อลำเลียง Tracheophytes
เคลด:	พืชดอก Angiosperms
เคลด:	พืชใบเลี้ยงคู่แท้ Eudicots
เคลด:	โรสิด Rosids
อันดับ:	กุหลาบ Rosales
วงศ์:	กุหลาบ Rosaceae

Why?

1. Can be **easily** converted into images via screenshots + **Accessible** API for article text
2. Large **availability** of data in SEA languages

experiment setup

OCR systems: **EasyOCR**, **Tesseract**, **GOT**

Criteria:

1. **Different** underlying architectures
2. **Open-source** with wide support for SEA languages

EasyOCR

- Text detection: **CRAFT** algorithm (CNN)
- Text recognition: **Convolutional Recurrent Neural Network**
- Supports **83** languages

Tesseract

- One of the most **well-known** open-source OCR engines
- Long Short-Term Memory (**LSTM**) RNN
- Supports **116** languages

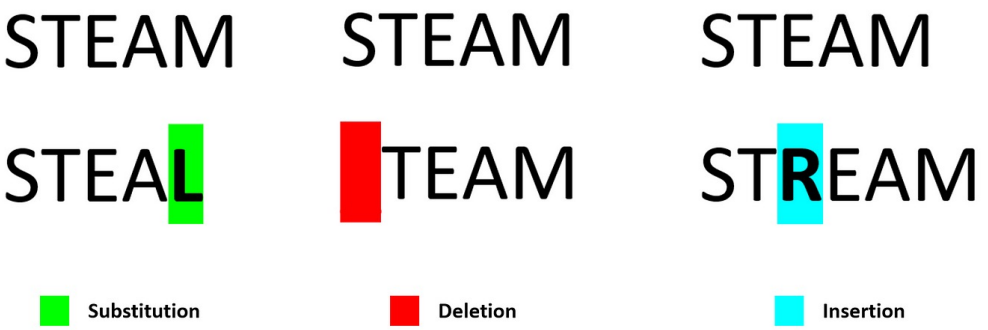
GOT
General OCR Theory

- Designed to recognize **beyond** traditional text (e.g., sheet music)
- **Vision Encoder Decoder** architecture
- **580 million** parameters
- Supports only English and Simplified Chinese

Evaluation metrics: **CER** and **WER**

Character Error Rate (CER):

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



- N: Number of characters in ground truth
- ⬇️ Lower value → **Better**
- **Word Error Rate** (WER): Word-equivalent of CER
  - Commonly used when preserving entire word is important

How good can it get? Analyzing and improving OCR accuracy in large scale historic newspaper digitization programs (Holley, 2009)

roadmap

01 Motivation

02 Related Work

03 Experiment 1: Benchmarking on Real-world Wikipedia **Screenshots**

04 Experiment 2: Benchmarking on Synthetic Data

05 Experiment 3: Fine-tuning

experiment 1: benchmarking on wiki. screenshots

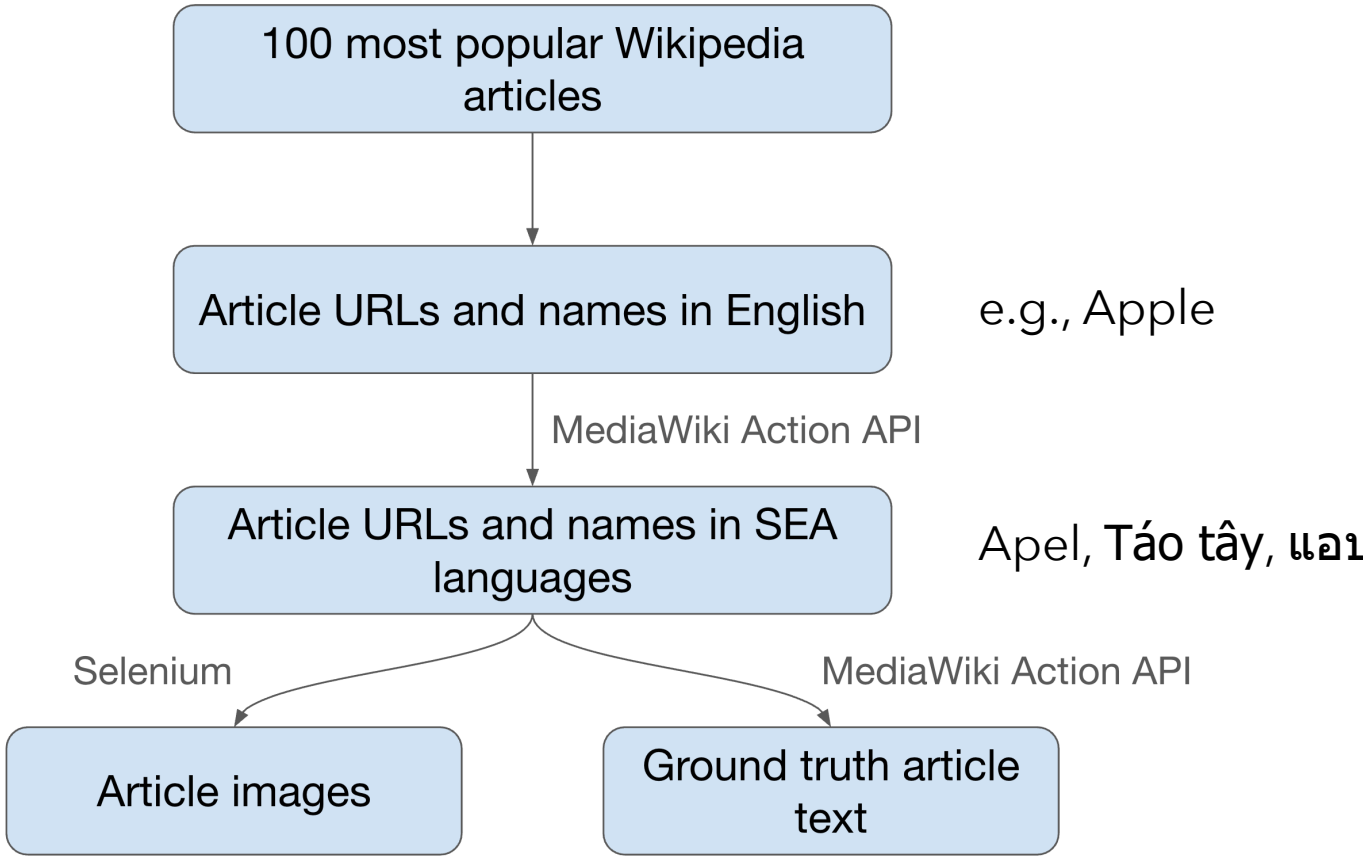
Goal: Explore performance of OCR tools on SEA scripts (RQ1)

Steps:

1. **Collect** article screenshots and texts from Wikipedia
2. **Evaluate** OCR tools

experiment 1: benchmarking on wiki. screenshots

1. **Collect** article screenshots and texts from Wikipedia



วิกิพีเดีย
สารานุกรมเสรี

แอปเปิล

แอปเปิล (อังกฤษ: apple; ชื่อวิทยาศาสตร์: *Malus domestica*) เป็นต้นไม้ผลัดใบในวงศ์กุหลาบ มีผลรสหวานเรียกว่า **ผลแอปเปิล** แอปเปิลมีปลูกอยู่ทั่วโลกในลักษณะของไม้ผล และสายพันธุ์ที่ถูกปลูกมากที่สุดคือสกุล *Malus* ต้นแอปเปิลมีต้นกำเนิดในเอเชียกลาง ซึ่งบรรพบุรุษคือ *Malus sieversii* ยังคงพบได้ในปัจจุบัน แอปเปิลมีปลูกเป็นเวลาหลายพันปีในเอเชียและยุโรป และกลุ่มอาณานิคมชาวยุโรปนำมาปลูกที่อเมริกาเหนือ แอปเปิลมีความสำคัญทางศาสนาและเทพปกรณัมในหลายวัฒนธรรม รวมถึงนอร์ส กรีก และประเพณีต่าง ๆ ของคริสต์ศาสนิกชนของชาวยุโรป

ต้นแอปเปิลจะมีขนาดใหญ่หากเติบโตจากเมล็ด แต่จะมีขนาดเล็กถ้าถูกตัดต่อเนื้อเยื่อเข้ากับราก ปัจจุบันมีแอปเปิลที่พันธุ์ปลูกมากกว่า 7,500 ชนิด ทำให้แอปเปิลมีลักษณะพิเศษหลากหลาย พันธุ์ปลูกแต่ละพันธุ์จะมีรสชาติแตกต่างกัน และการนำไปใช้ต่างกันด้วย เช่น นำไปประกอบอาหาร กินดิบ ๆ หรือนำไปผลิตไซเดอร์ ปกติแอปเปิลจะแพร่พันธุ์ด้วยการตัดต่อเนื้อเยื่อ แต่แอปเปิลป่าจะเติบโตได้เองจากเมล็ด ต้นแอปเปิลและผลแอปเปิลอาจประสบปัญหาจากจากเห็ดรา แบคทีเรีย และศัตรูพืชต่าง ๆ ซึ่งอาจควบคุมได้ด้วยวิธีการทางเกษตรอินทรีย์และอนินทรีย์หลายวิธี ใน ค.ศ. 2010 มีการถอดรหัสจีโนมของแอปเปิล เป็นส่วนหนึ่งของงานวิจัยควบคุมโรคและการคัดเลือกผสมพันธุ์ในการผลิตแอปเปิล

ใน ค.ศ. 2013 มีการปลูกแอปเปิลประมาณ 80 ล้านตันขึ้นทั่วโลก ประเทศจีนผลิตได้จำนวนเกือบครึ่งหนึ่งของจำนวนดังกล่าว[3] สหรัฐอเมริกาเป็นผู้ผลิตแอปเปิลมากที่เป็นอันดับที่สอง ด้วยการผลิตมากกว่า 6% ประเทศตุรกีเป็นที่สาม ตามด้วยประเทศอิตาลี อินเดีย และโปแลนด์ แอปเปิลมักนิยมกินดิบ แต่สามารถพบได้ในอาหารที่เตรียมขึ้น (โดยเฉพาะของหวาน) และเครื่องดื่ม มีความคิดว่า

แอปเปิล

แอปเปิลพันธุ์ 'คริปส์พิงก์'

ดอก

การจำแนกชั้นทางวิทยาศาสตร์

โดเมน:	ยูแคริโอต Eukaryota
อาณาจักร:	พืช Plantae
เคลด:	พืชมีท่อลำเลียง Tracheophytes
เคลด:	พืชดอก Angiosperms
เคลด:	พืชใบเลี้ยงคู่แท้ Eudicots
เคลด:	โรสิด Rosids
อันดับ:	กุหลาบ Rosales
วงศ์:	กุหลาบ Rosaceae

ได้ มาจากต้นแอปเปิล (Malus spp.) รวมถึง แอปเปิลบ้าน และแอปเปิลสวนผลไม้ (Malus domestica) ต้นแอปเปิลได้รับการเพาะปลูกทั่วโลก และเป็นสายพันธุ์ที่ปลูกมากที่สุดในสกุล Malus มีถิ่นกำเนิดในเอเชียกลาง โดยมี Malus sieversii เป็นบรรพบุรุษป่า ซึ่งยังคงพบอยู่ในปัจจุบัน แอปเปิลถูกเพาะปลูกในยูเรเซีย มานานนับพันปี ก่อนถูกนำเข้าสู่ อเมริกาเหนือ โดยชาวอาณานิคมยุโรป นอกจากนี้ แอปเปิลยังมีความสำคัญทางศาสนาและตำนานในหลายวัฒนธรรม เช่น นอร์ส กรีก และคริสเตียนยุโรป

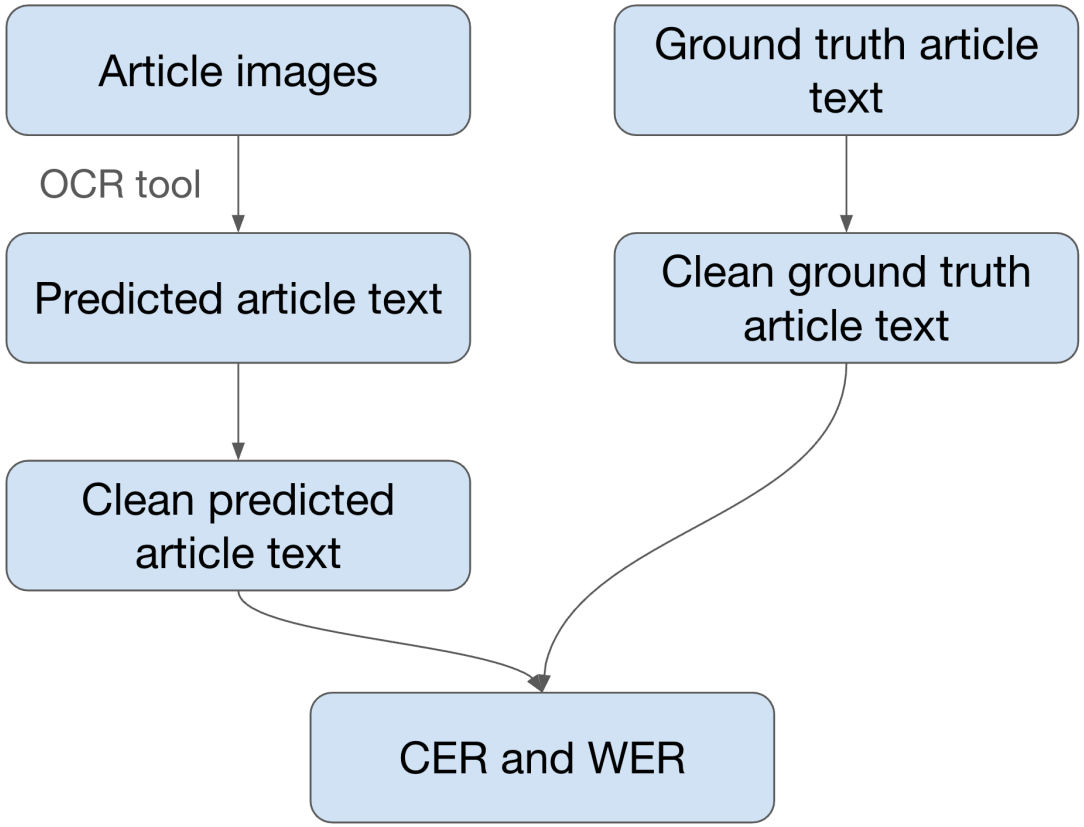
Result: **7,976** Wikipedia screenshots + Ground truth text

experiment 1: benchmarking on wiki. screenshots

2. **Evaluate** OCR tools

ได้ มาจากต้นแอปเปิล (Malus spp.) รวมถึง แอปเปิลบ้าน และแอปเปิลสวน ผลไม้ (Malus domestica) ต้นแอปเปิล ได้รับการเพาะปลูกทั่วโลก และเป็นสาย พันธุ์ที่ปลูกมากที่สุดในสกุล Malus มีถิ่น กำเนิดในเอเชียกลาง โดยมี Malus sieversii เป็นบรรพบุรุษป่า ซึ่งยังคงพบ อยู่ในปัจจุบัน แอปเปิลถูกเพาะปลูกใน ยูเรเซีย มานานนับพันปี ก่อนถูกนำเข้า สู่ อเมริกาเหนือ โดยชาวอาณานิคม ยุโรป นอกจากนี้ แอปเปิลยังมี ความสำคัญทางศาสนาและตำนานใน หลายวัฒนธรรม เช่น นอร์ส กรีก และค ริสเตียนยุโรป

Predicted text



CER: 0.06
WER: 0.10

ได้ มาจากต้นแอปเปิล (Malus spp.) รวมถึง แอปเปิลบ้าน และแอปเปิลสวน ผลไม้ (Malus domestica) ต้นแอปเปิล ได้รับการเพาะปลูกทั่วโลก และเป็นสาย พันธุ์ที่ปลูกมากที่สุดในสกุล Malus มีถิ่น กำเนิดในเอเชียกลาง โดยมี Malus sieversii เป็นบรรพบุรุษป่า ซึ่งยังคงพบ อยู่ในปัจจุบัน แอปเปิลถูกเพาะปลูกใน ยูเรเซีย มานานนับพันปี ก่อนถูกนำเข้า สู่ อเมริกาเหนือ โดยชาวอาณานิคม ยุโรป นอกจากนี้ แอปเปิลยังมี ความสำคัญทางศาสนาและตำนานใน หลายวัฒนธรรม เช่น นอร์ส กรีก และค ริสเตียนยุโรป

Ground truth text

experiment 1: benchmarking on wiki. screenshots

OCR performance on Wikipedia screenshots

	Character Error Rate			Word Error Rate		
	EasyOCR	Tesseract	GOT	EasyOCR	Tesseract	GOT
English	0.21	0.21	0.67	0.27	0.29	0.67
Indonesian	0.28	0.28	0.61	0.36	0.42	0.71
Vietnamese	0.47	0.38	-	0.45	0.39	-
Thai	0.35	0.55	-	1.45	1.45	-

* Lower → Better

experiment 1: benchmarking on wiki. screenshots

OCR performance on Wikipedia screenshots

	Character Error Rate			Word Error Rate		
	EasyOCR	Tesseract	GOT	EasyOCR	Tesseract	GOT
English	0.21	0.21	0.67	0.27	0.29	0.67
Indonesian	0.28	0.28	0.61	0.36	0.42	0.71
Vietnamese	0.47	0.38	-	0.45	0.39	-
Thai	0.35	0.55	-	1.45	1.45	-

* Lower → Better

- Error rate: **Latin** script

experiment 1: benchmarking on wiki. screenshots

OCR performance on Wikipedia screenshots

	Character Error Rate			Word Error Rate		
	EasyOCR	Tesseract	GOT	EasyOCR	Tesseract	GOT
English	0.21	0.21	0.67	0.27	0.29	0.67
Indonesian	0.28	0.28	0.61	0.36	0.42	0.71
Vietnamese	0.47	0.38	-	0.45	0.39	-
Thai	0.35	0.55	-	1.45	1.45	-

* Lower → Better

- Error rate: **Latin** script < Latin script with **diacritics**

experiment 1: benchmarking on wiki. screenshots

OCR performance on Wikipedia screenshots

	Character Error Rate			Word Error Rate		
	EasyOCR	Tesseract	GOT	EasyOCR	Tesseract	GOT
English	0.21	0.21	0.67	0.27	0.29	0.67
Indonesian	0.28	0.28	0.61	0.36	0.42	0.71
Vietnamese	0.47	0.38	-	0.45	0.39	-
Thai	0.35	0.55	-	1.45	1.45	-

* Lower → Better

- Error rate: **Latin** script < Latin script with **diacritics** < **Brahmic** script

experiment 1: benchmarking on wiki. screenshots

OCR performance on Wikipedia screenshots

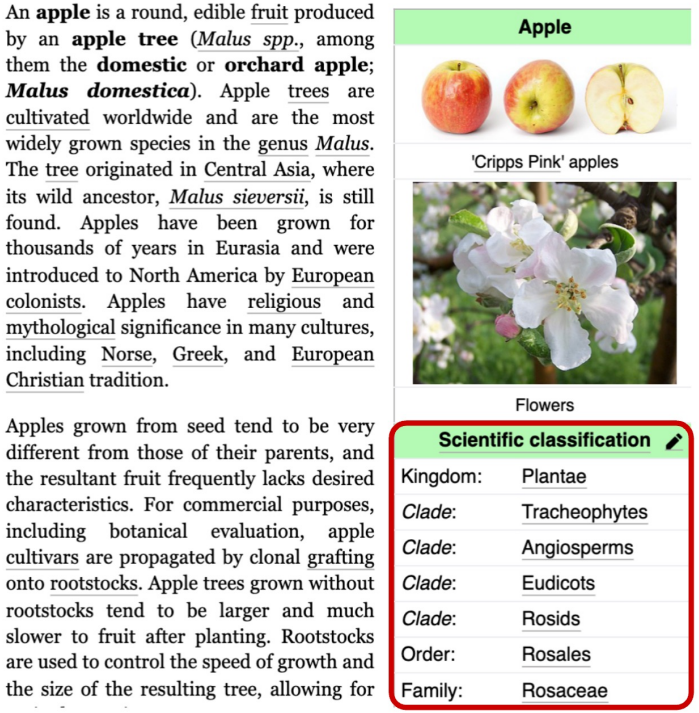
	Character Error Rate			Word Error Rate		
	EasyOCR	Tesseract	GOT	EasyOCR	Tesseract	GOT
English	0.21	0.21	0.67	0.27	0.29	0.67
Indonesian	0.28	0.28	0.61	0.36	0.42	0.71
Vietnamese	0.47	0.38	-	0.45	0.39	-
Thai	0.35	0.55	-	1.45	1.45	-
Average	**0.33**	0.35	0.64			

* Lower → Better

- Error rate: Latin script < Latin script with diacritics < Brahmic script
- **EasyOCR** achieved the best overall performance

experiment 1: benchmarking on wiki. screenshots

Limitation: Complex layouts and multi-modal elements **interfere** with OCR results.



An **apple** is a round, edible fruit produced by an **apple tree** (*Malus spp.*, among them the **domestic** or **orchard apple**; ***Malus domestica***). Apple trees are cultivated worldwide and are the most widely grown species in the genus *Malus*. The tree originated in Central Asia, where its wild ancestor, *Malus sieversii*, is still found. Apples have been grown for thousands of years in Eurasia and were introduced to North America by European colonists. Apples have religious and mythological significance in many cultures, including Norse, Greek, and European Christian tradition.

Apples grown from seed tend to be very different from those of their parents, and the resultant fruit frequently lacks desired characteristics. For commercial purposes, including botanical evaluation, apple cultivars are propagated by clonal grafting onto rootstocks. Apple trees grown without rootstocks tend to be larger and much slower to fruit after planting. Rootstocks are used to control the speed of growth and the size of the resulting tree, allowing for

Apple

'Cripps Pink' apples

Flowers

Scientific classification

Kingdom:	Plantae
Clade:	Tracheophytes
Clade:	Angiosperms
Clade:	Eudicots
Clade:	Rosids
Order:	Rosales
Family:	Rosaceae

```
Apples grown from seed tend to be very
Scientific classification different from
those of their parents, and the resultant
fruit frequently lacks desired Kingdom:
Plantae characteristics. For commercial
purposes, Clade: Tracheophytes including
botanical evaluation, apple cultivars are
propagated by clonal grafting Clade:
Angiosperms onto rootstocks Apple trees
grown without Clade: Eudicots rootstocks
tend to be larger and much Clade: Rosids
slower to fruit after planting: Rootstocks
```

→

Inflated CERs and WERs for Wikipedia screenshots

Need for **layout analysis** to process digital docs.

roadmap

01 Motivation

02 Related Work

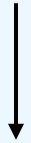
03 Experiment 1: Benchmarking on Real-world Wikipedia Screenshots

04 Experiment 2: Benchmarking on **Synthetic** Data

05 Experiment 3: Fine-tuning

experiment 2: benchmarking on synthetic data

Synthetic data allows **controlled distortions** to dataset and **minimizes** annotation errors



Goals:

1. Analyze OCR robustness against **noise** on SEA languages (RQ1)
2. **Isolate** script-related errors (RQ2)

experiment 2: benchmarking on synthetic data

1. Analyze OCR robustness against **noise**

Types of noise
Bold
Italic
Link
Heading
No noise

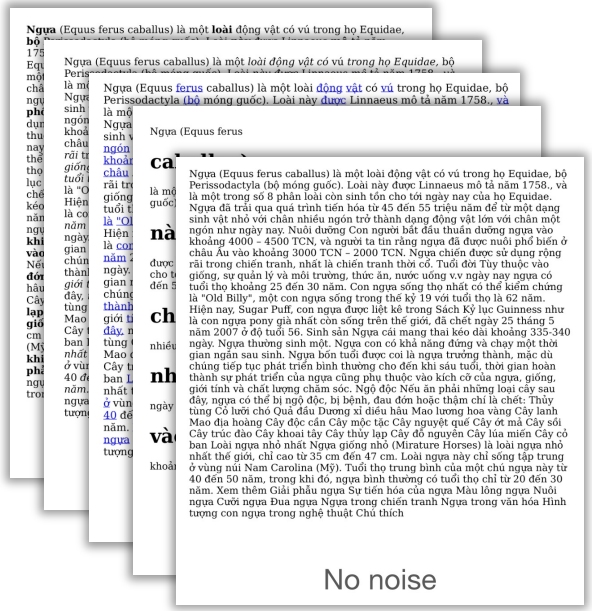
Why?

1. **Commonly** found in digital text and web-based content
2. Each noise type presents a **unique challenge** for OCR systems

experiment 2: benchmarking on synthetic data

1. Analyze OCR robustness against **noise**

Types of noise
Bold
Italic
Link
Heading
No noise



Generate document images with noise using WeasyPrint

Run and **evaluate** OCR tools on each dataset

experiment 2: benchmarking on synthetic data

OCR performance on noise-free synthetic data

	Character Error Rate		
	EasyOCR	Tesseract	GOT
English	0.03	0.02	0.01
Indonesian	0.02	0.02	0.01
Vietnamese	0.10	0.03	-
Thai	0.07	0.09	-

* Lower → Better

experiment 2: benchmarking on synthetic data

Similar trend as benchmark on Wiki. screenshots

OCR performance on noise-free synthetic data

	Character Error Rate			Average
	EasyOCR	Tesseract	GOT	
English	0.03	0.02	0.01	**0.02**
Indonesian	0.02	0.02	0.01	**0.02**
Vietnamese	0.10	0.03	-	**0.06**
Thai	0.07	0.09	-	**0.08**

* Lower → Better

- Error rate: **Latin** script < Latin script with **diacritics** < **Brahmic** script

experiment 2: benchmarking on synthetic data

OCR performance on noise-free synthetic data

	Character Error Rate		
	EasyOCR	Tesseract	GOT
English	0.03 ↓	0.02 ↓	0.01 ↓
Indonesian	0.02 ↓	0.02 ↓	0.01 ↓
Vietnamese	0.10 ↓	0.03 ↓	-
Thai	0.07 ↓	0.09 ↓	-

* Lower → Better

Avg. decrease of **0.34**

- Error rate: Latin script < Latin script with diacritics < Brahmic script
- OCR tools performed **better** on synthetic data than on Wiki. screenshots

experiment 2: benchmarking on synthetic data

OCR performance on noise-free synthetic data

	Character Error Rate		
	EasyOCR	Tesseract	GOT
English	0.03	0.02	0.01
Indonesian	0.02	0.02	0.01
Vietnamese	0.10	0.03	-
Thai	0.07	0.09	-
Average for synthetic data	0.05	> 0.04	> 0.01
Average for Wiki. screenshots	0.33	< 0.35	< 0.64

Tesseract and GOT potentially **less robust to noise**

- Error rate: Latin script < Latin script with diacritics < Brahmic script
- OCR tools performed better on synthetic data than on Wiki. screenshots
- Tesseract and GOT **performed well** on synthetic data, but poorly on Wikipedia screenshots

experiment 2: benchmarking on synthetic data

Impact of noise on OCR accuracy

EasyOCR

	No Noise CER	Bold % Change	Italic % Change	Link % Change	Heading % Change
English	0.03	-7.1%	0.0%	3.6%	-7.1%
Indonesian	0.02	-5.6%	0.0%	22.2%	-5.6%
Vietnamese	0.10	-1.0%	-1.9%	-6.7%	-11.5%
Thai	0.07	0.0%	0.0%	7.2%	0.0%

Tesseract

	No Noise CER	Bold % Change	Italic % Change	Link % Change	Heading % Change
English	0.02	-4.2%	0.0%	0.0%	8.3%
Indonesian	0.02	0.0%	0.0%	0.0%	21.7%
Vietnamese	0.03	0.0%	3.8%	3.8%	26.9%
Thai	0.09	3.2%	0.0%	5.4%	4.3%

GOT

	No Noise CER	Bold % Change	Italic % Change	Link % Change	Heading % Change
English	0.01	27.3%	36.4%	-18.2%	218.2%
Indonesian	0.01	9.1%	18.2%	36.4%	63.6%

experiment 2: benchmarking on synthetic data

Impact of noise on OCR accuracy

EasyOCR

	No Noise CER	Bold % Change	Italic % Change	Link % Change	Heading % Change
English	0.03	-7.1%	0.0%	3.6%	-7.1%
Indonesian	0.02	-5.6%	0.0%	22.2%	-5.6%
Vietnamese	0.10	-1.0%	-1.9%	-6.7%	-11.5%
Thai	0.07	0.0%	0.0%	7.2%	0.0%

Tesseract

	No Noise CER	Bold % Change	Italic % Change	Link % Change	Heading % Change
English	0.02	-4.2%	0.0%	0.0%	8.3%
Indonesian	0.02	0.0%	0.0%	0.0%	21.7%
Vietnamese	0.03	0.0%	3.8%	3.8%	26.9%
Thai	0.09	3.2%	0.0%	5.4%	4.3%

GOT

	No Noise CER	Bold % Change	Italic % Change	Link % Change	Heading % Change
English	0.01	27.3%	36.4%	-18.2%	218.2%
Indonesian	0.01	9.1%	18.2%	36.4%	63.6%

- Tesseract and GOT **sensitive to noise**

experiment 2: benchmarking on synthetic data

Impact of noise on OCR accuracy

EasyOCR

	No Noise CER	Bold % Change	Italic % Change	Link % Change	Heading % Change
English	0.03	-7.1%	0.0%	3.6%	-7.1%
Indonesian	0.02	-5.6%	0.0%	22.2%	-5.6%
Vietnamese	0.10	-1.0%	-1.9%	-6.7%	-11.5%
Thai	0.07	0.0%	0.0%	7.2%	0.0%

Tesseract

	No Noise CER	Bold % Change	Italic % Change	Link % Change	Heading % Change
English	0.02	-4.2%	0.0%	0.0%	8.3%
Indonesian	0.02	0.0%	0.0%	0.0%	21.7%
Vietnamese	0.03	0.0%	3.8%	3.8%	26.9%
Thai	0.09	3.2%	0.0%	5.4%	4.3%

GOT

	No Noise CER	Bold % Change	Italic % Change	Link % Change	Heading % Change
English	0.01	27.3%	36.4%	-18.2%	218.2%
Indonesian	0.01	9.1%	18.2%	36.4%	63.6%
	Average	2.2%	5.6%	5.4%	**31.9%**

- Tesseract and GOT **sensitive to noise**
- **Heading** noise presents greatest challenge for OCR tools

2. **Classify** errors by character type

	Included Characters
Arabic digit	0-9
Latin letter	a-z, A-Z
Latin letter with diacritic	à-ỹ, À-Ỹ
Vietnamese special letter	đ, Đ
Thai letter	ก-ฮ
Special symbol	.,!?;:()-”–’–$%/&+-=[]
Whitespace	␣
Other	

OCR results from synthetic dataset **without noise**

↓

Identify and **classify** misrecognized characters using RegEx

experiment 2: benchmarking on synthetic data

Error classification in English and Indonesian articles

English

	Count	EasyOCR % Missed	Tesseract % Missed	GOT % Missed
Arabic digit	38,324	0.7%	1.9%	0.3%
Latin letter	1,546,964	1.3%	1.8%	0.4%
Latin letter with diacritic	452	100.0%	56.0%	15.9%
Special symbol	55,110	29.5%	3.7%	4.3%
Whitespace	317,587	4.9%	4.3%	3.6%
Other	1,562	103.6%	92.6%	127.9%

Indonesian

	Count	EasyOCR % Missed	Tesseract % Missed	GOT % Missed
Arabic digit	24,947	0.4%	1.8%	0.2%
Latin letter	1,208,707	0.5%	1.8%	0.4%
Latin letter with diacritic	276	7.2%	100.0%	16.7%
Special symbol	38,980	22.9%	4.9%	1.4%
Whitespace	207,556	4.8%	5.1%	4.1%
Other	1,262	94.4%	99.6%	64.8%

Articles with Latin scripts:

experiment 2: benchmarking on synthetic data

Error classification in English and Indonesian articles

English

	Count	EasyOCR % Missed	Tesseract % Missed	GOT % Missed
Arabic digit	38,324	0.7%	1.9%	0.3%
Latin letter	1,546,964	1.3%	1.8%	0.4%
Latin letter with diacritic	452	100.0%	56.0%	15.9%
Special symbol	55,110	29.5%	3.7%	4.3%
Whitespace	317,587	4.9%	4.3%	3.6%
Other	1,562	103.6%	92.6%	127.9%

Indonesian

	Count	EasyOCR % Missed	Tesseract % Missed	GOT % Missed
Arabic digit	24,947	0.4%	1.8%	0.2%
Latin letter	1,208,707	0.5%	1.8%	0.4%
Latin letter with diacritic	276	7.2%	100.0%	16.7%
Special symbol	38,980	22.9%	4.9%	1.4%
Whitespace	207,556	4.8%	5.1%	4.1%
Other	1,262	94.4%	99.6%	64.8%

Articles with Latin scripts:

- High accuracies on **Latin letters** and **Arabic digits**

experiment 2: benchmarking on synthetic data

Error classification in English and Indonesian articles

English

	Count	EasyOCR % Missed	Tesseract % Missed	GOT % Missed
Arabic digit	38,324	0.7%	1.9%	0.3%
Latin letter	1,546,964	1.3%	1.8%	0.4%
Latin letter with diacritic	452	100.0%	56.0%	15.9%
Special symbol	55,110	29.5%	3.7%	4.3%
Whitespace	317,587	4.9%	4.3%	3.6%
Other	1,562	103.6%	92.6%	127.9%

Indonesian

	Count	EasyOCR % Missed	Tesseract % Missed	GOT % Missed
Arabic digit	24,947	0.4%	1.8%	0.2%
Latin letter	1,208,707	0.5%	1.8%	0.4%
Latin letter with diacritic	276	7.2%	100.0%	16.7%
Special symbol	38,980	22.9%	4.9%	1.4%
Whitespace	207,556	4.8%	5.1%	4.1%
Other	1,262	94.4%	99.6%	64.8%

Articles with Latin scripts:

- High accuracies on **Latin letters** and **Arabic digits**
- Lower accuracies for Latin letters with **diacritics**

experiment 2: benchmarking on synthetic data

Error classification in English and Indonesian articles

English

	Count	EasyOCR % Missed	Tesseract % Missed	GOT % Missed
Arabic digit	38,324	0.7%	1.9%	0.3%
Latin letter	1,546,964	1.3%	1.8%	0.4%
Latin letter with diacritic	452	100.0%	56.0%	15.9%
Special symbol	55,110	29.5%	3.7%	4.3%
Whitespace	317,587	4.9%	4.3%	3.6%
Other	1,562	103.6%	92.6%	127.9%

Indonesian

	Count	EasyOCR % Missed	Tesseract % Missed	GOT % Missed
Arabic digit	24,947	0.4%	1.8%	0.2%
Latin letter	1,208,707	0.5%	1.8%	0.4%
Latin letter with diacritic	276	7.2%	100.0%	16.7%
Special symbol	38,980	22.9%	4.9%	1.4%
Whitespace	207,556	4.8%	5.1%	4.1%
Other	1,262	94.4%	99.6%	64.8%

Articles with Latin scripts:

- High accuracies on **Latin letters** and **Arabic digits**
- Lower accuracies for Latin letters with **diacritics**
- Most **uncategorized characters** misclassified
  - E.g., Greek, Cyrillic, Arabic, Chinese characters

experiment 2: benchmarking on synthetic data

Error classification in Vietnamese and Thai articles

Vietnamese

	Count	EasyOCR % Missed	Tesseract % Missed
Arabic digit	31,473	1.1%	2.2%
Latin letter	916,667	8.5%	1.8%
Latin letter with diacritic	294,406	14.9%	1.9%
Vietnamese special letter	30,903	4.2%	1.8%
Special symbol	41,655	25.4%	3.1%
Whitespace	367,936	10.9%	5.3%
Other	1,909	105.4%	88.6%

Thai

	Count	EasyOCR % Missed	Tesseract % Missed
Arabic digit	22,580	0.9%	6.7%
Latin letter	36,174	100.0%	100.0%
Latin letter with diacritic	101	100.0%	100.0%
Thai letter	1,014,395	1.2%	4.2%
Special symbol	14,246	8.0%	10.6%
Whitespace	58,164	37.5%	37.2%
Other	890	105.2%	110.4%

Articles with more complex scripts:

experiment 2: benchmarking on synthetic data

Error classification in Vietnamese and Thai articles

Vietnamese

	Count	EasyOCR % Missed	Tesseract % Missed
Arabic digit	31,473	1.1%	2.2%
Latin letter	916,667	8.5%	1.8%
Latin letter with diacritic	294,406	14.9%	1.9%
Vietnamese special letter	30,903	4.2%	1.8%
Special symbol	41,655	25.4%	3.1%
Whitespace	367,936	10.9%	5.3%
Other	1,909	105.4%	88.6%

Thai

	Count	EasyOCR % Missed	Tesseract % Missed
Arabic digit	22,580	0.9%	6.7%
Latin letter	36,174	100.0%	100.0%
Latin letter with diacritic	101	100.0%	100.0%
Thai letter	1,014,395	1.2%	4.2%
Special symbol	14,246	8.0%	10.6%
Whitespace	58,164	37.5%	37.2%
Other	890	105.2%	110.4%

Articles with more complex scripts:

- Tesseract achieved high accuracies on **Vietnamese letters**

experiment 2: benchmarking on synthetic data

Error classification in Vietnamese and Thai articles

Vietnamese

	Count	EasyOCR % Missed	Tesseract % Missed
Arabic digit	31,473	1.1%	2.2%
Latin letter	916,667	8.5%	1.8%
Latin letter with diacritic	294,406	14.9%	1.9%
Vietnamese special letter	30,903	4.2%	1.8%
Special symbol	41,655	25.4%	3.1%
Whitespace	367,936	10.9%	5.3%
Other	1,909	105.4%	88.6%

Thai

	Count	EasyOCR % Missed	Tesseract % Missed
Arabic digit	22,580	0.9%	6.7%
Latin letter	36,174	100.0%	100.0%
Latin letter with diacritic	101	100.0%	100.0%
Thai letter	1,014,395	1.2%	4.2%
Special symbol	14,246	8.0%	10.6%
Whitespace	58,164	37.5%	37.2%
Other	890	105.2%	110.4%

Articles with more complex scripts:

- Tesseract achieved high accuracies on **Vietnamese letters**
- High accuracies on **Thai letters**

experiment 2: benchmarking on synthetic data

Error classification in Vietnamese and Thai articles

Vietnamese

	Count	EasyOCR % Missed	Tesseract % Missed
Arabic digit	31,473	1.1%	2.2%
Latin letter	916,667	8.5%	1.8%
Latin letter with diacritic	294,406	14.9%	1.9%
Vietnamese special letter	30,903	4.2%	1.8%
Special symbol	41,655	25.4%	3.1%
Whitespace	367,936	10.9%	5.3%
Other	1,909	105.4%	88.6%

Thai

	Count	EasyOCR % Missed	Tesseract % Missed
Arabic digit	22,580	0.9%	6.7%
Latin letter	36,174	100.0%	100.0%
Latin letter with diacritic	101	100.0%	100.0%
Thai letter	1,014,395	1.2%	4.2%
Special symbol	14,246	8.0%	10.6%
Whitespace	58,164	37.5%	37.2%
Other	890	105.2%	110.4%

Articles with more complex scripts:

- Tesseract achieved high accuracies on **Vietnamese letters**
- High accuracies on **Thai letters**
- Thai OCR models did not recognize **Latin letters with or without diacritics**

experiment 2: benchmarking on synthetic data

Implication 1: OCR tools recognize **Viet.** and **Thai scripts** well but struggle with **non-native characters**.

- Related work:
  - OCR tools perform **poorly** on complex scripts[1]
  - Latin scripts perform **better** than SEA scripts[2]
- Our results: **Latin** script < Latin script with **diacritics** < **Brahmic** script

- High accuracies on **Vietnamese letters**
- High accuracies on **Thai letters**
- Thai OCR models did not recognize **Latin letters with or without diacritics**

[1] A concise survey of OCR for low-resource languages (Agarwal & Anastasopoulos, 2024)
[2] OCR improves machine translation for low-resource languages (Ignat et al., 2022)

experiment 2: benchmarking on synthetic data

Implication 1: OCR tools recognize **Viet.** and **Thai scripts** well but struggle with **non-native characters**.

- Related work:
  - OCR tools perform **poorly** on complex scripts[1]
  - Latin scripts perform **better** than SEA scripts[2]
- Our results: **Latin** script < Latin script with **diacritics** < **Brahmic** script

- High accuracies on **Vietnamese letters**
- High accuracies on **Thai letters**
- Thai OCR models did not recognize **Latin letters with or without diacritics**

↓

Challenges common belief

[1] A concise survey of OCR for low-resource languages (Agarwal & Anastasopoulos, 2024)
[2] OCR improves machine translation for low-resource languages (Ignat et al., 2022)

experiment 2: benchmarking on synthetic data

Implication 1: OCR tools recognize **Viet.** and **Thai scripts** well but struggle with **non-native characters**.

- High accuracies on **Vietnamese letters**
- High accuracies on **Thai letters**
- Thai OCR models did not recognize **Latin letters with or without diacritics**

←

Misclassification of characters **outside of target language** inflates overall CER

↓

OCR tools struggle with **multilingual content**

↓

Challenges common belief

Need for OCR systems to better handle **mixed-script content**

experiment 3:
fine-tuning

Goal: Explore how fine-tuning can enhance OCR accuracy on SEA languages (RQ3)

↓

Fine-tune GOT on Vietnamese and Thai

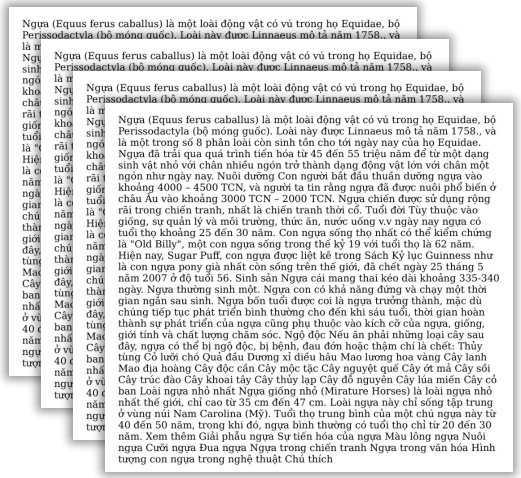
Why?

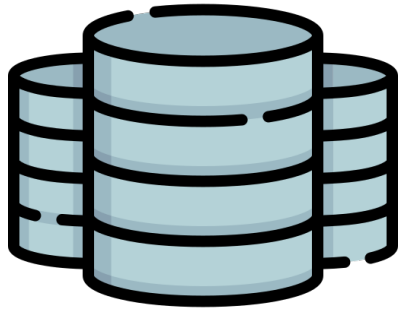
1. GOT does **not** officially support Vietnamese and Thai
2. Related work: Fine-tuning pre-trained **transformer-based** models is efficient in adapting new languages
3. GOT already achieved state-of-the-art results on **Indonesian**

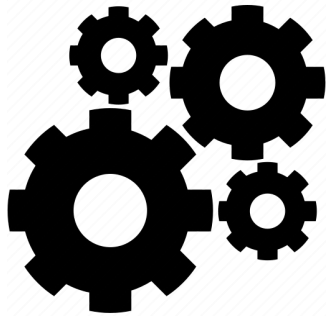
Fine-tuning vision encoder-decoder transformers for handwriting text recognition on historical documents (Parres & Paredes, 2023)

experiment 3:
fine-tuning

Fine-tune GOT on Vietnamese and Thai







Generate more noise-free synthetic data

960 training images
+ **50** test images

Split dataset in sizes of increasing scale

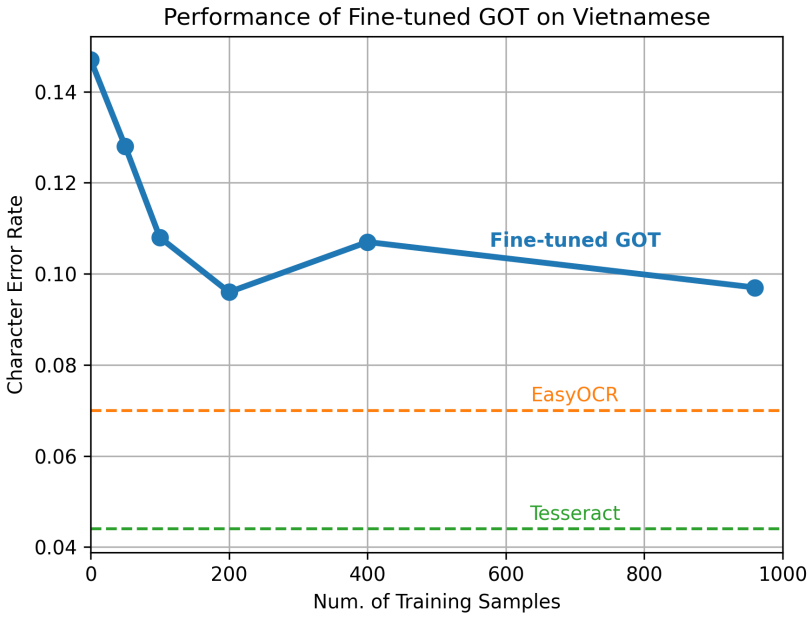
to approximate diminishing returns

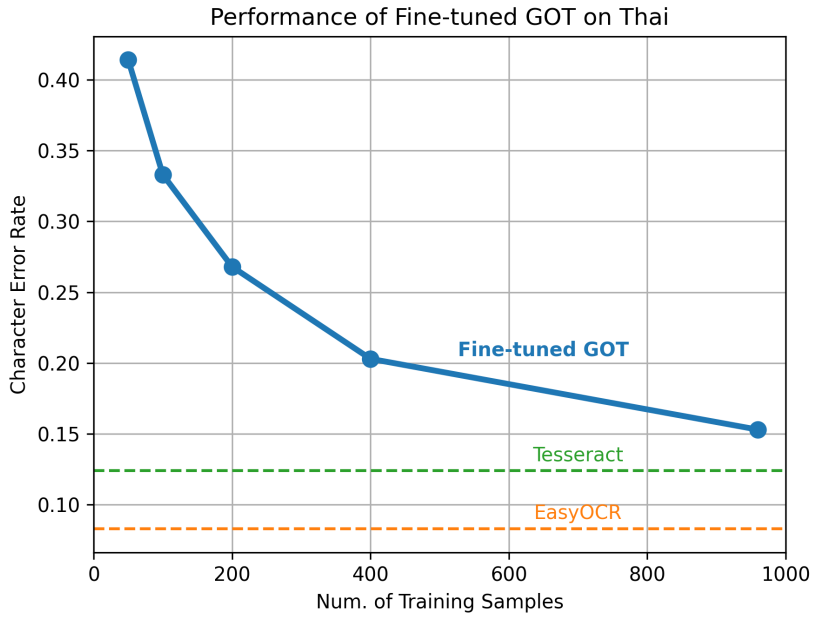
Datasets with 50, 100, 200, 400, 960 images

Fine-tune GOT using each dataset

Fine-tuned GOTs

experiment 3: fine-tuning





experiment 3:
fine-tuning

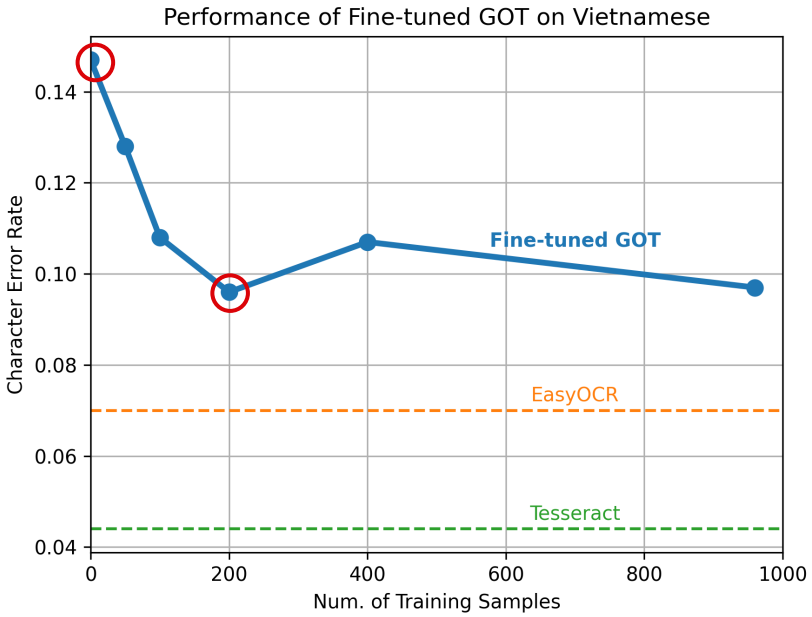
No fine-tuning:

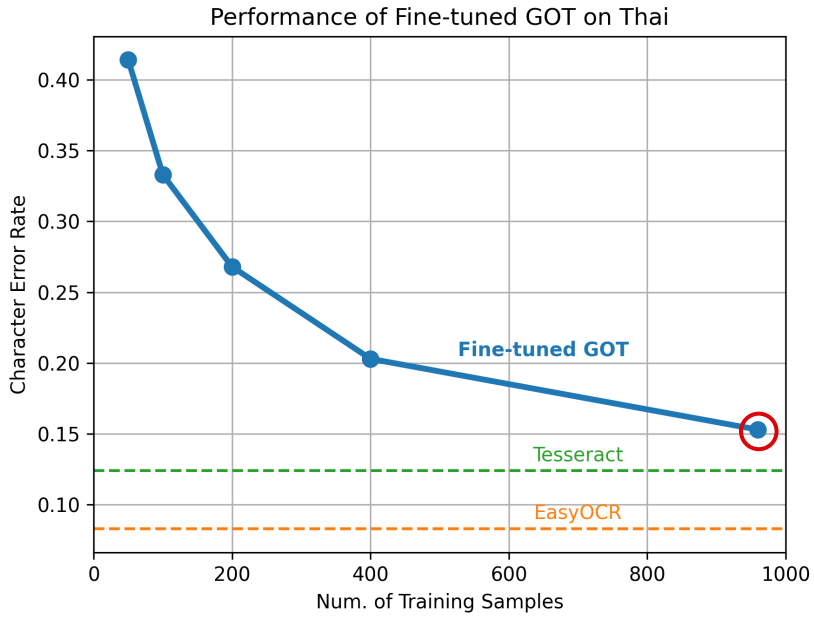
0.15

-33%

200 samples:

0.10





No fine-tuning:

Cannot recognize

960 samples:

0.15

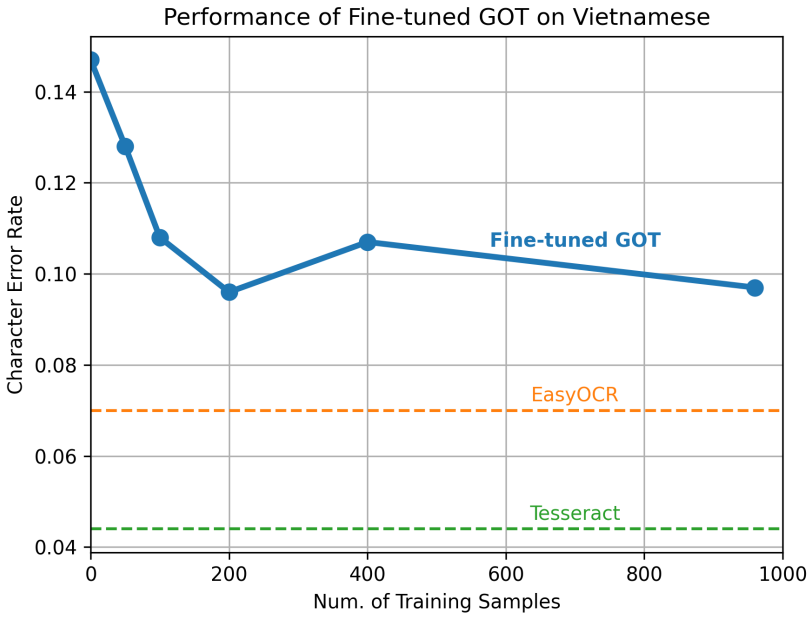
- Fine-tuning **improved** performance with relatively little data

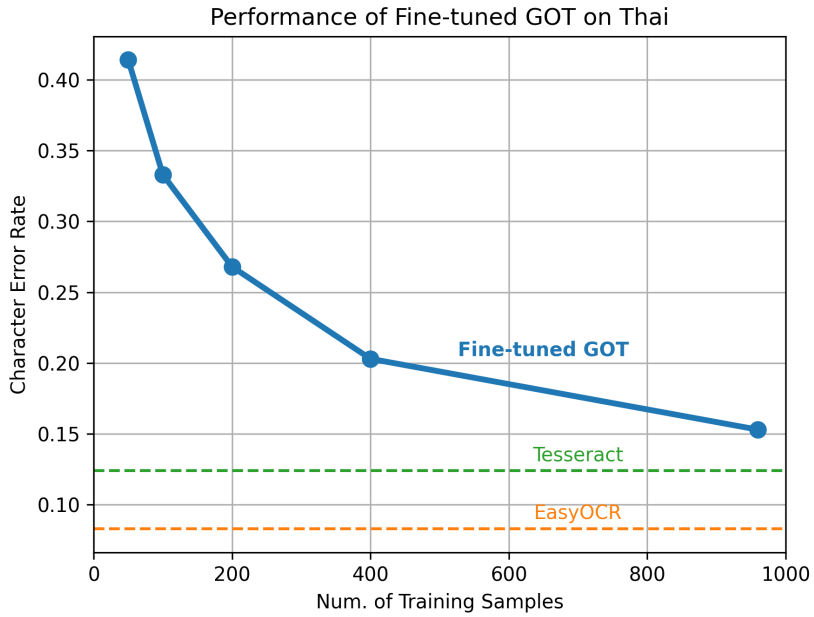
experiment 3: fine-tuning

Ours: **0.10**

EasyOCR: 0.07

Tesseract: 0.04





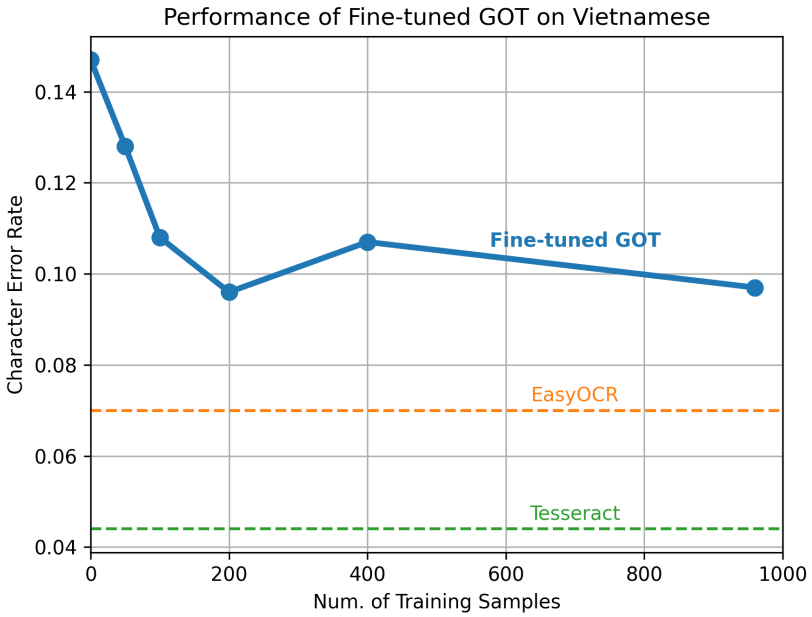
Ours: **0.15**

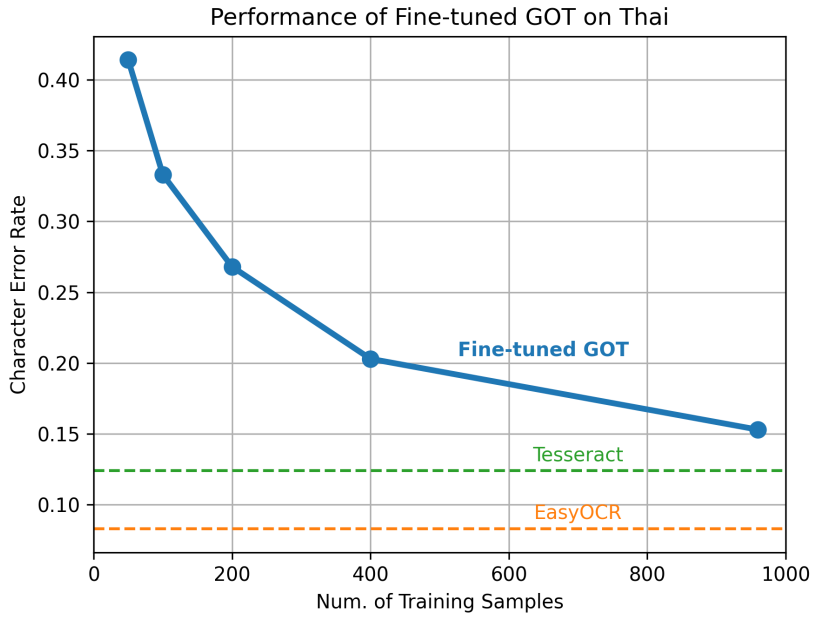
Tesseract: 0.12

EasyOCR: 0.08

- Fine-tuning **improved** performance with relatively little data
- Fine-tuned GOT **underperformed** relative to EasyOCR and Tesseract

experiment 3: fine-tuning





- Fine-tuning **improved** performance with relatively little data
- Fine-tuned GOT **underperformed** relative to EasyOCR and Tesseract
- Model performance did **not** consistently improve with more data

experiment 3: fine-tuning

Implication 2: Fine-tuning is **effective** in adapting OCR systems to new SEA languages.

Fine-tuning **improved** performance with relatively little data

↓

Practical approach for adapting to new **low-resource** SEA languages

Surpassing state-of-the-art OCR tools may require exploring **alternative models** and optimizing **fine-tuning strategies**

🔎 Goal: **Benchmark** and **improve** OCR for SEA languages

What I did:

1. **Benchmarked** EasyOCR, Tesseract, and GOT on English, Indo., Viet., and Thai
2. Created **reusable pipeline** for collecting Wiki. data and benchmarking OCR tools
3. **Fine-tuned** GOT on Vietnamese and Thai

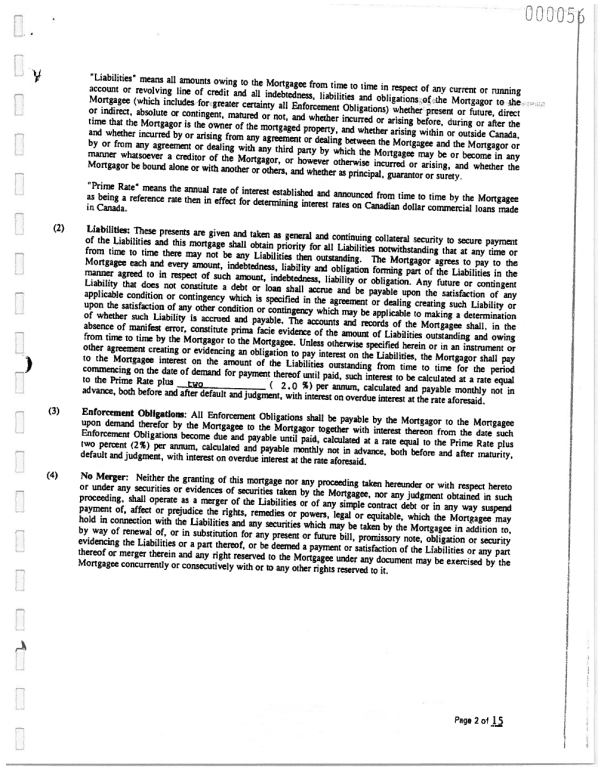
Key findings & implications:

✨ OCR tools recognize **Vietnamese** and **Thai scripts** well, but struggle with non-native characters → Need for **mixed-script OCR**

✨ Fine-tuning is **effective** in adapting OCR systems to new SEA languages

experiment 1: benchmarking on wiki. screenshots

Limitation: Wikipedia articles **cannot fully replace** scanned physical documents.



000056

"Liabilities" means all amounts owing to the Mortgagee from time to time in respect of any current or running account or revolving line of credit and all indebtedness, liabilities and obligations of the Mortgagor to the Mortgagee (which includes for greater certainty all Enforcement Obligations) whether present or future, direct or indirect, absolute or contingent, matured or not, and whether incurred or arising before, during or after the time that the Mortgagor is the owner of the mortgaged property, and whether arising within or outside Canada, by or from any agreement or dealing with any third party by which the Mortgagee may be or become in any manner whatsoever a creditor of the Mortgagor, or however otherwise incurred or arising, and whether the Mortgagor be bound alone or with another or others, and whether as principal, guarantor or surety.

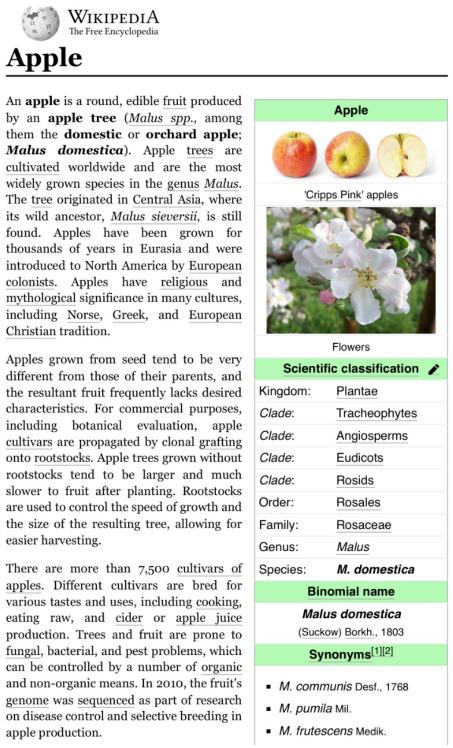
"Prime Rate" means the annual rate of interest established and announced from time to time by the Mortgagee as being a reference rate then in effect for determining interest rates on Canadian dollar commercial loans made in Canada.

(2) **Liabilities:** These presents are given and taken as general and continuing collateral security to secure payment of the Liabilities and this mortgage shall obtain priority for all Liabilities notwithstanding that at any time or from time to time there may not be any Liabilities then outstanding. The Mortgagor agrees to pay to the Mortgagee each and every amount, indebtedness, liability and obligation forming part of the Liabilities in the manner agreed to in respect of such amount, indebtedness, liability or obligation. Any future or contingent Liability that does not constitute a debt or loan shall accrue and be payable upon the satisfaction of any applicable condition or contingency which is specified in the agreement or dealing creating such Liability or upon the satisfaction of any other condition or contingency which may be applicable to making a determination of whether such Liability is accrued and payable. The accounts and records of the Mortgagee shall, in the absence of manifest error, constitute prima facie evidence of the amount of Liabilities outstanding and owing from time to time by the Mortgagor to the Mortgagee. Unless otherwise specified herein or in an instrument or other agreement creating or evidencing an obligation to pay interest on the Liabilities, the Mortgagor shall pay to the Mortgagee interest on the amount of the Liabilities outstanding from time to time for the period commencing on the date of demand for payment thereof until paid, such interest to be calculated at a rate equal to the Prime Rate plus two (2.0 %) per annum, calculated and payable monthly not in advance, both before and after default and judgment, with interest on overdue interest at the rate aforesaid.

(3) **Enforcement Obligations:** All Enforcement Obligations shall be payable by the Mortgagor to the Mortgagee upon demand therefor by the Mortgagee to the Mortgagor together with interest thereon from the date such Enforcement Obligations become due and payable until paid, calculated at a rate equal to the Prime Rate plus two percent (2%) per annum, calculated and payable monthly not in advance, both before and after maturity, default and judgment, with interest on overdue interest at the rate aforesaid.

(4) **No Merger:** Neither the granting of this mortgage nor any proceeding taken hereunder or with respect hereto or under any securities or evidences of securities taken by the Mortgagee, nor any judgment obtained in such proceeding, shall operate as a merger of the Liabilities or of any simple contract debt or in any way suspend payment of, affect or prejudice the rights, remedies or powers, legal or equitable, which the Mortgagee may hold in connection with the Liabilities and any securities which may be taken by the Mortgagee in addition to, by way of renewal of, or in substitution for any present or future bill, promissory note, obligation or security evidencing the Liabilities or a part thereof, or be deemed a payment or satisfaction of the Liabilities or any part thereof or merger therein and any right reserved to the Mortgagee under any document may be exercised by the Mortgagee concurrently or consecutively with or to any other rights reserved to it.

Page 2 of 15



WIKIPEDIA
The Free Encyclopedia

Apple

An **apple** is a round, edible fruit produced by an **apple tree** (*Malus spp.*, among them the **domestic** or **orchard apple**; ***Malus domestica***). Apple trees are cultivated worldwide and are the most widely grown species in the genus *Malus*. The tree originated in Central Asia, where its wild ancestor, *Malus sieversii*, is still found. Apples have been grown for thousands of years in Eurasia and were introduced to North America by European colonists. Apples have religious and mythological significance in many cultures, including Norse, Greek, and European Christian tradition.

Apples grown from seed tend to be very different from those of their parents, and the resultant fruit frequently lacks desired characteristics. For commercial purposes, including botanical evaluation, apple cultivars are propagated by clonal grafting onto rootstocks. Apple trees grown without rootstocks tend to be larger and much slower to fruit after planting. Rootstocks are used to control the speed of growth and the size of the resulting tree, allowing for easier harvesting.

There are more than 7,500 cultivars of apples. Different cultivars are bred for various tastes and uses, including cooking, eating raw, and cider or apple juice production. Trees and fruit are prone to fungal, bacterial, and pest problems, which can be controlled by a number of organic and non-organic means. In 2010, the fruit's genome was sequenced as part of research on disease control and selective breeding in apple production.

Apple

'Cripps Pink' apples

Flowers

Scientific classification

Kingdom:	Plantae
Clade:	Tracheophytes
Clade:	Angiosperms
Clade:	Eudicots
Clade:	Rosids
Order:	Rosales
Family:	Rosaceae
Genus:	*Malus*
Species:	***M. domestica***

Binomial name

Malus domestica
(Suckow) Borkh., 1803

Synonyms[1][2]

- *M. communis* Desf., 1768
- *M. pumila* Mil.
- *M. frutescens* Medik.

Physical documents have **more diverse noise** types and **practical use cases**

Why still use Wikipedia?

1. Similar **linguistic** features
2. **More accessible** to obtain lots of data
3. Growing use cases of OCR on **digital screenshots**