

A Thesis Final Defense Fulfilment of the Requirements for the Degree of  
Master of Engineering in Mechatronic Engineering

Presented by Nutdanai Sritunya st123055



# ENHANCING RETRIEVAL- AUGMENTED GENERATION FOR THAI QUESTION ANSWERING THROUGH COSINE SIMILARITY AND INTERACTIVE FEEDBACK

Examination Committee: Dr. Mongkol Ekpanyapong (Chairperson)  
Dr. Chaklam Silpasuwanchai  
Dr. Ekbordin Winijkul



What do you need help with?

This is the content of my thesis



**1 BACKGROUND &  
PROBLEM STATEMENT**

**2 METHODOLOGY**

**3 EXPERIMENT  
& RESULTS**

**4 DISSCUSION &  
CONCLUSION**

# TABLE OF CONTENTS

# 1. BACKGROUND & PROBLEM STATEMENT



What is the Large Language Models (LLMs)

LARGE LANGUAGE MODELS (LLMs)



## Advancements & Capabilities:

- LLMs have transformed how we interact with AI, excelling in creating text and answering questions with high accuracy.
- They are versatile, enhancing various applications from content creation to customer service.

## Limitations:

- Despite their strengths, LLMs can sometimes create information that seems real but isn't ("hallucination").
- They may not always reflect the most current knowledge, as they're trained on existing datasets.



แหล่งกำเนิดฝุ่นในกรุงเทพมหานครมีอะไรบ้าง



แหล่งกำเนิดฝุ่นในกรุงเทพมหานคร ได้แก่ 1. การเผาไหม้เชื้อเพลิงฟอสซิลจากโรงงานและยานพาหนะ 2. กิจกรรมทางอุตสาหกรรม เช่น การผลิต การก่อสร้าง และการบดอัด 3. การเผาขยะมูลฝอย 4. กิจกรรมทางการเกษตร เช่น การเผาพืชผลและปศุสัตว์ 5. พายุทราย 6. พายุฝุ่น 7. การก่อสร้างถนนและการพัฒนาโครงสร้างพื้นฐาน 8. การไหลบ่าของดินจากภูเขาและพื้นที่สูง 9. กิจกรรมการท่องเที่ยว เช่น การเดินป่าและตั้งแคมป์ 10. การระเบิดของภูเขาไฟ



แหล่งกำเนิดฝุ่นในกรุงเทพมหานครมีอะไรบ้าง



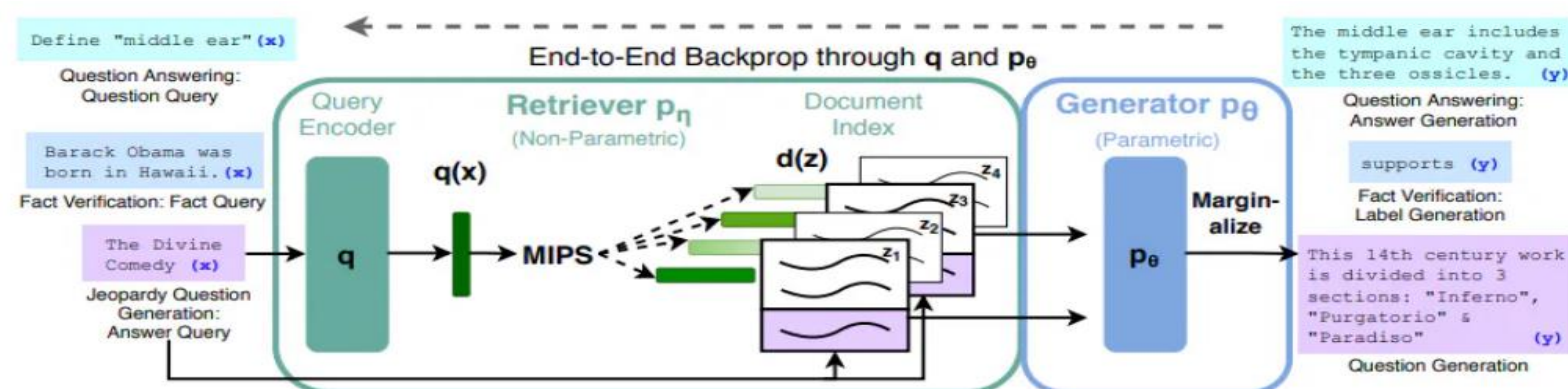
แหล่งกำเนิดหลักของฝุ่น PM 2.5 ในพื้นที่กรุงเทพฯ และปริมณฑล ได้แก่ การจราจร (53%) อุตสาหกรรม (21%) และการเผาในที่โล่ง (13%) ในพื้นที่ภาคเหนือ แหล่งกำเนิดฝุ่นส่วนใหญ่เกิดจากการเผาในที่โล่งและไฟฟ้า ในขณะที่ในพื้นที่ภาคใต้ มาจากมลพิษทางอากาศข้ามแดน เช่น ไฟป่าในประเทศเพื่อนบ้าน



# BACKGROUND

## RETRIEVAL-AUGMENTED GENERATION (RAG)

### WHAT IS RAG?



**Retrieval-Augmented Generation**, or RAG, stands at the cutting edge of artificial intelligence research. It's a sophisticated technique that marries the power of two AI domains: retrieval-based models and generative models. The retrieval component pulls relevant information from vast databases or the internet, akin to finding the best sources in a library. Then, the generative component takes this information to craft responses that are not just accurate but contextually rich and informative. Think of RAG as a highly skilled chef who first sources the freshest ingredients before creating a gourmet dish.

### WHY DO WE NEED RAG?

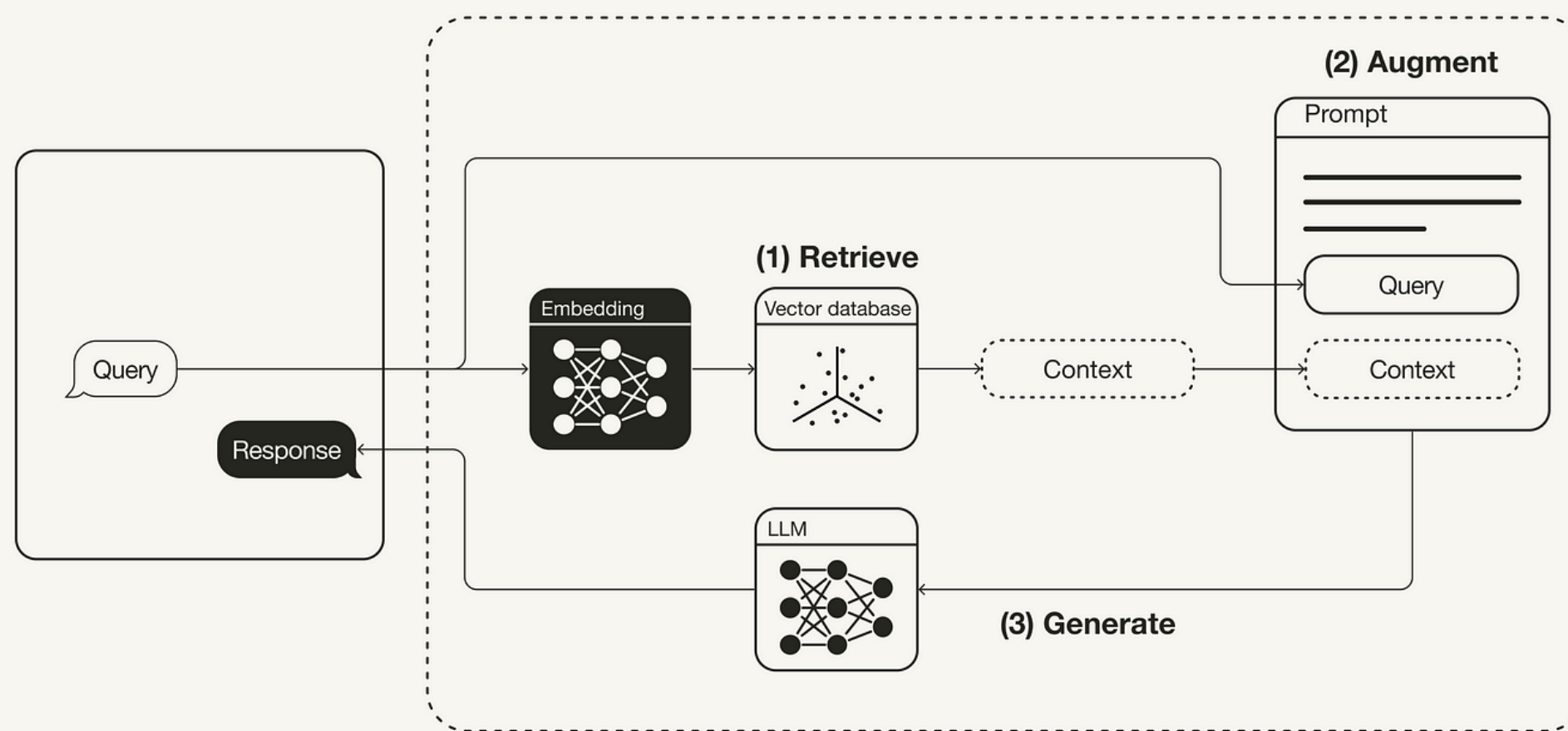


Traditional language models can sometimes produce incorrect but plausible responses, known as "hallucination." RAG (Retrieval-Augmented Generation) addresses this issue by incorporating a retrieval system that gathers up-to-date information from external databases. This ensures accurate and personalized responses backed by factual evidence, minimizing hallucinations and improving relevance for question-answering systems requiring high accuracy.



# PROBLEM STATEMENT

## PROBLEM



- **Relevance Verification Challenge:** RAG systems encounter difficulties in accurately verifying the relevance of retrieved documents, potentially resulting in the inclusion of irrelevant sources.
- **Redundancy Reduction Issue:** The challenge of minimizing redundancy within responses arises due to the incorporation of irrelevant documents during retrieval in RAG systems.
- **Integration of User Feedback:** Incorporating user feedback to validate document relevance presents a challenge in current RAG systems, affecting the quality of responses.

# Enhancing Retrieval-Augmented Generation for Thai Question-Answering through Cosine Similarity and Interactive Feedback

1

2

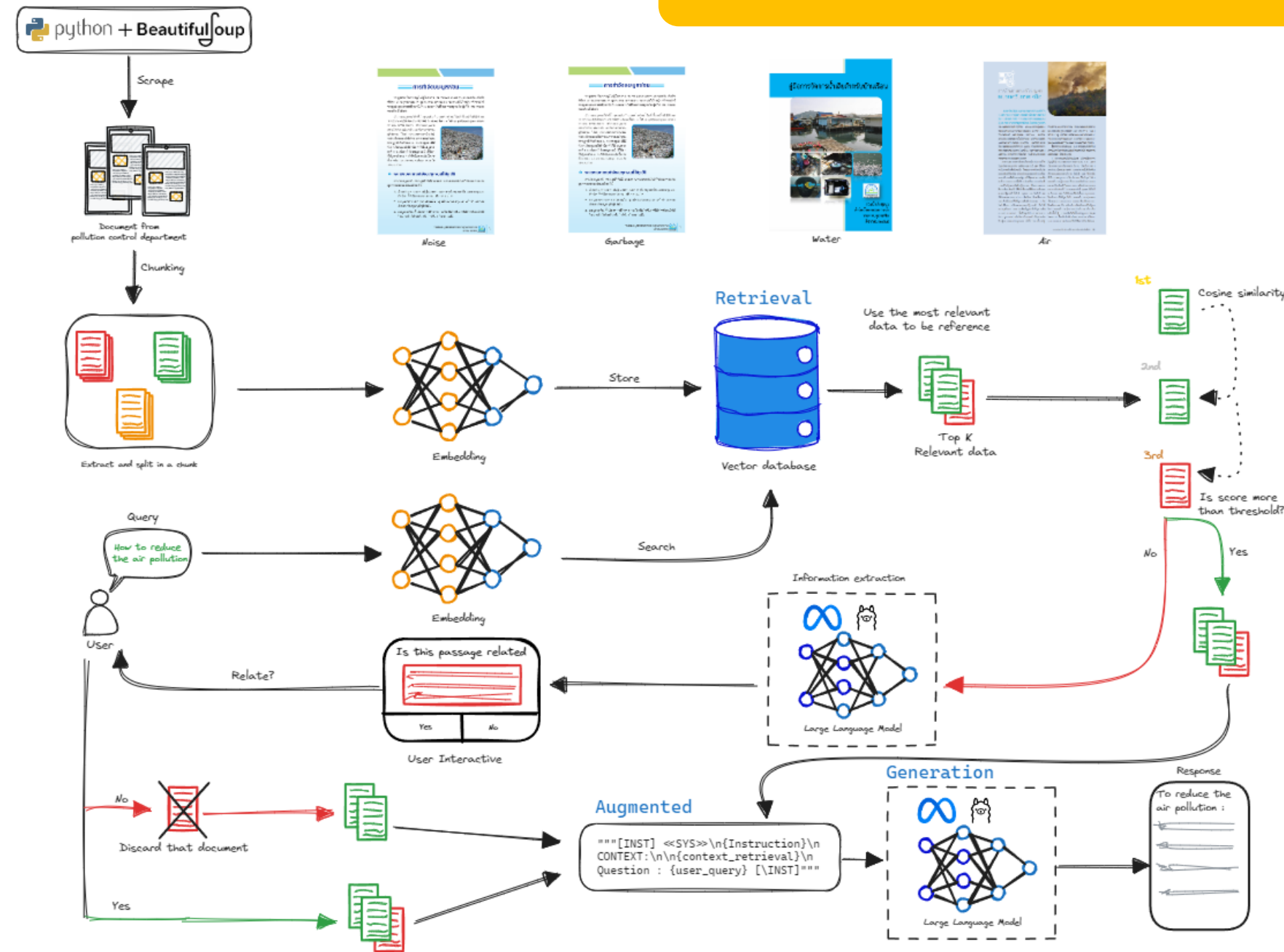
3

4



Can you show overview of your thesis?

This Is an overview for methodology





What you do in the pre processing?

METHODOLOGY

DATA COLLECTION



**Total Documents Retrieved:**  
**663**

- Pollution Statistics: 292 documents
  - Detailed statistical data on various types of pollution.
- Knowledge about Pollution: 371 documents
  - Informative content on pollution causes, effects, and prevention.

DATA COLLECTION



- Solid waste pollutions : 97 document
- Water pollutions : 125 document
- Air pollutions : 126 document
- Noise pollutions : 23 document

หน่วยงาน	ข้าราชการ (ค่าหนังสือ)	ลูกจ้างประจำ (อัตรา)	พนักงานราชการ (อัตรา)	รวม
ส่วนกลาง	5	0	0	5
กลุ่มพัฒนาระบบบริหาร	3	0	2	5
กลุ่มตรวจสอบภายใน	3	0	0	3
สำนักงานเลขาธิการกรม	34	21	28	83
กองกฎหมาย	22	3	9	34
กองจัดการกากของเสียและสารอันตราย	41	2	43	86
กองจัดการคุณภาพน้ำ	45	4	34	83
กองจัดการคุณภาพอากาศและเสียง	44	2	46	92
กองตรวจมลพิษ	39	1	30	70
กองยุทธศาสตร์และแผนงาน	23	3	11	37
ศูนย์เทคโนโลยีสารสนเทศและการสื่อสาร	18	0	8	26
ศูนย์ปฏิบัติการวิเคราะห์มลพิษและสิ่งแวดล้อม	25	3	10	38
สำนักงานสิ่งแวดล้อมและควบคุมมลพิษที่ 1 - 10	284	33	108	425
<b>รวม</b>	<b>586</b>	<b>72</b>	<b>329</b>	<b>987</b>

หมายเหตุ : ข้อมูล ณ วันที่ 9 กันยายน 2555

Due to the complexity and difficulty in extracting data from Pollution Statistics documents filled with tabulated data, this research will primarily focus on documents providing informative content about pollution.

METHODOLOGY



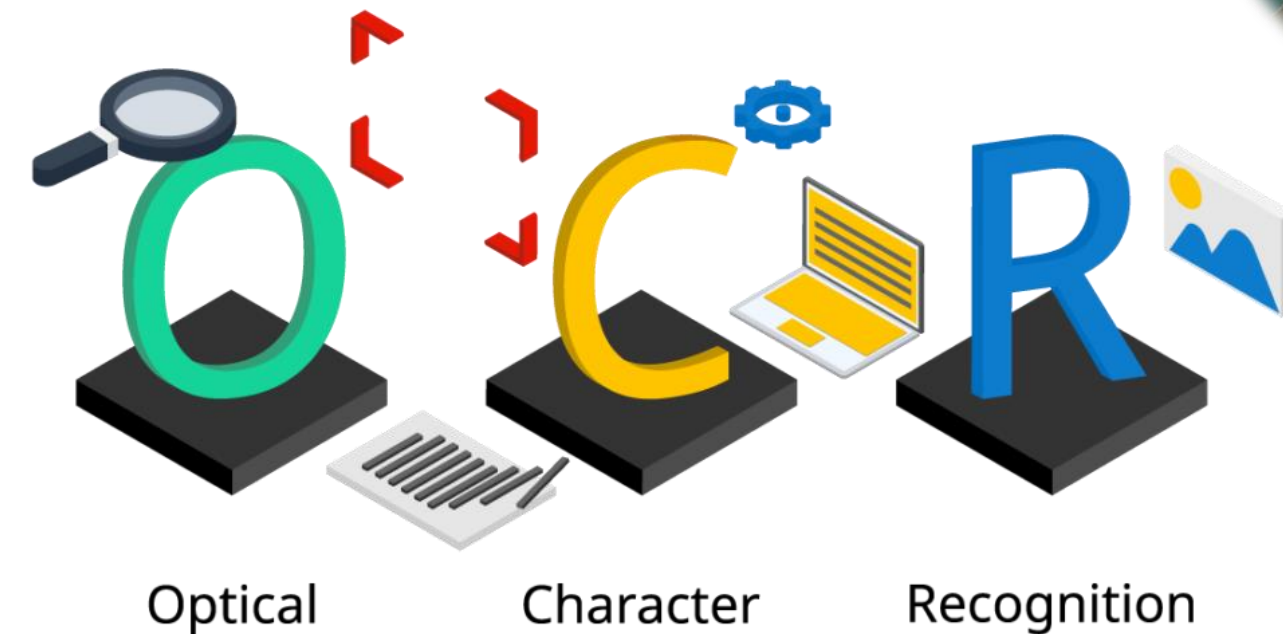
PDF EXTRACTION

Most of the document from Pollution Control Department website is a encrypt document so after I use PY2PDF library to scrape the data from document, The document will look like this

2. vâ«j,,°-πλ-πj“,,™â¿“™π-,™â È”°“vâ-πj“,,™â¿“™π-,™â È” †™àπ ðÿβâ“ μ-°vâ“ †âñ -à“j  
âiðπ,μ°≈âðβæ≈“ μ° †âiπμâπ †âiπ«Π’Σ’È™â«-≈¥âV’j“≥æ≈“ μ°·≈-,øj%â†âiπð-à“β¥’†æv“-¿“™π  
-,™â È”j’ð“-ÿ°“v,,™âβ“πΣ’È-“«π“π “j“vð≈â“βΣ”§«“j -ð“¥%â·≈-%jâ vâ“βç-ÿ≈ðð-†æ’Èjç÷Èπ  
†™àπ †v“§«væ°æ“ðÿβâ“,, à,,π°v-†âÛ“Σ”β“πλvðð,, à%«â,,πvðð-ÿà† jðÈ †jðÈð- ðÈðçðβ°Áπ”¿“™π-,  
™â È”j“,, àçðβ·Σπ°“v,, àðÿβæ≈“ μ° λvðð,,™ââiðπ,μ,, àð“λ“v·Σπ°“v,,™âðÿβæ≈“ μ°·≈-,øj  
ððð·°â«λvðð°v-μ°%â¥â«-†jðÈð%â ðÈð°“·øλvðð†§vðÈðβ¥ðÈj †âiπμâπ¿. À≈’°†≈’È-β°“v,,™âæ≈“  
μ°·≈-,øj ”λv-∫°‘°°v∫j,,π™’«μâv-°”«-π†™àπ π”°v-¥“..j“,,™â†âiπ«- ¥ÿÀ’∫ààð°-π°v-†Σððπ  
”λv-∫ðÿ°°v≥âμà“βÈ π”,,∫μðβ·≈-λ-«°°≈â«-j“,,™â≈≈‘μ°v-Σβ,,π†Σ»°“≈≈ð°v-Σβλvðð≈≈‘μæ«βλv’¥  
†âiπμâπ14



OPTICAL CHARACTERISTICS RECOGNITIONS (OCR)

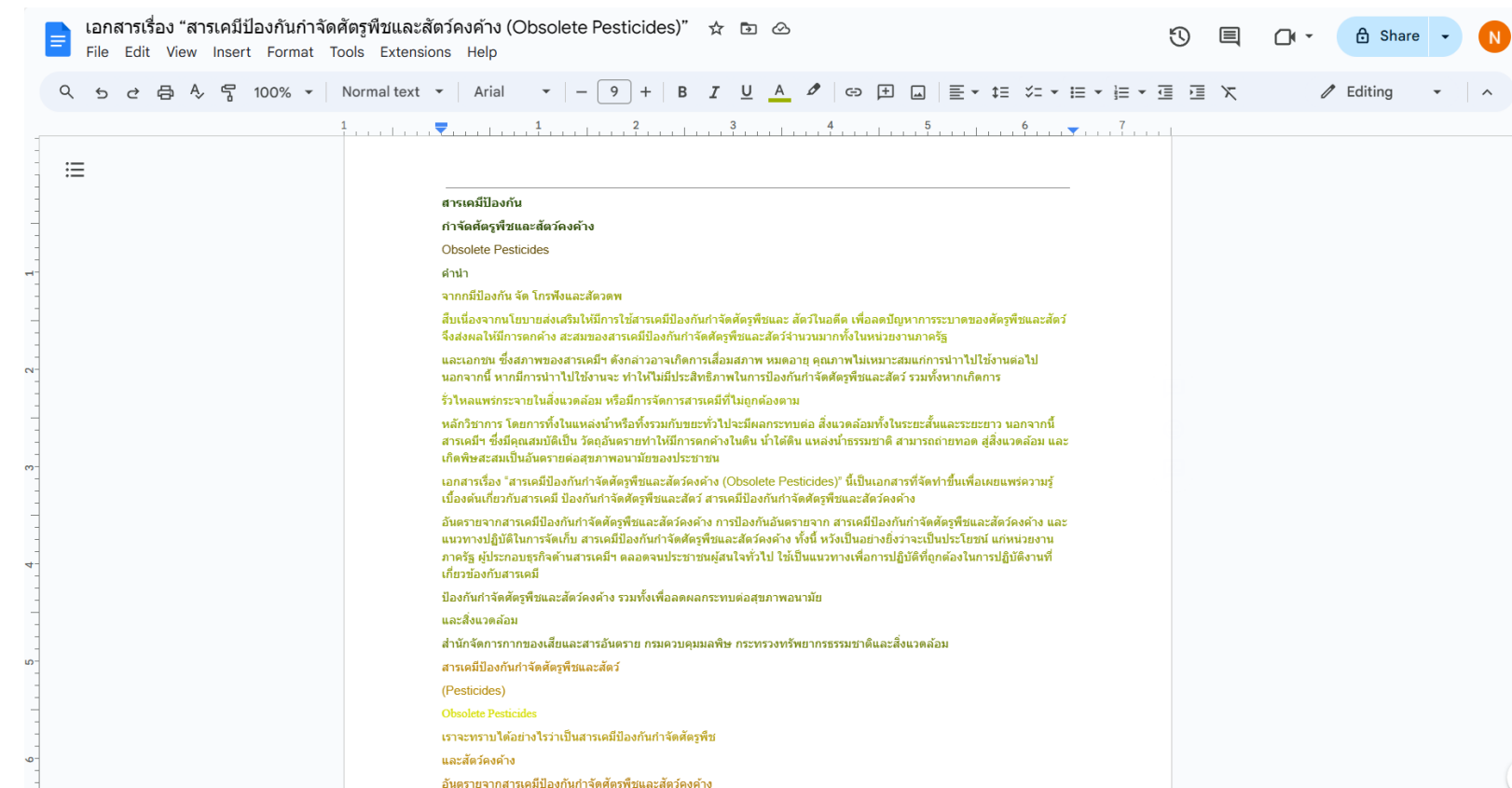
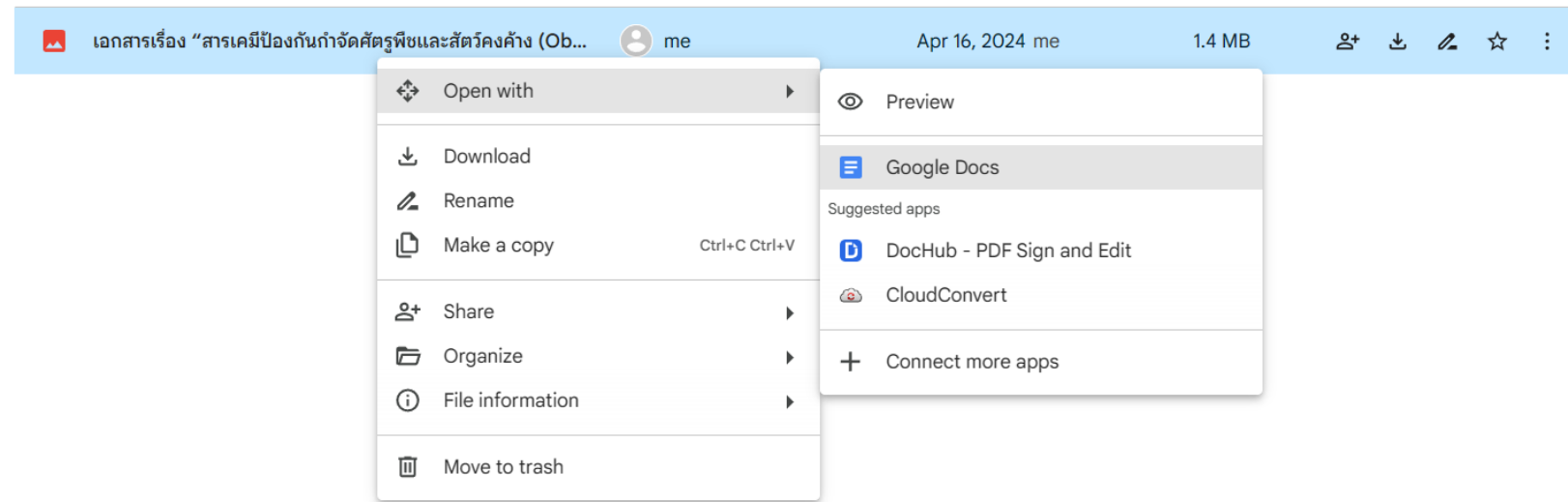


OCR stands for Optical Character Recognition. It is a technology that converts images of text into editable and searchable digital text data.





# OPTICAL CHATERISTICS RECOGNITIONS (OCR)



```
file_metadata = {
    'name': 'Converted Document',
    'mimeType': 'application/vnd.google-apps.document'
}
```

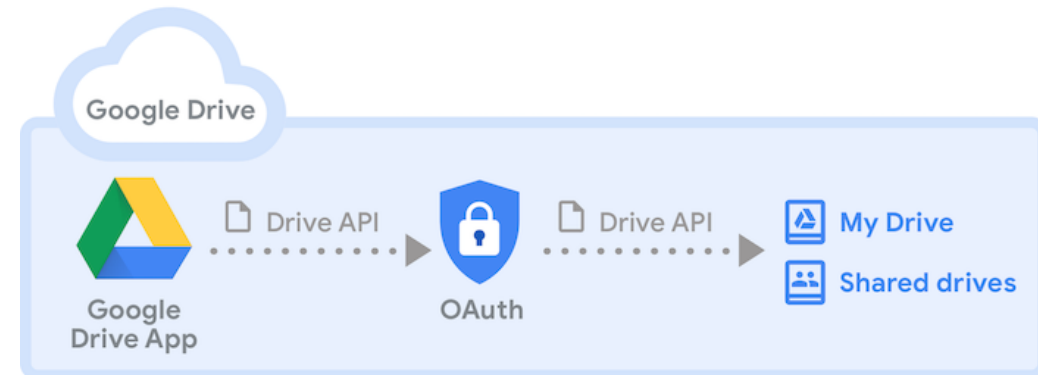
```
media = MediaFileUpload(image_path,
    mimeType='image/jpeg',
    resumable=True)
```

```
file = service.files().create(
    body=file_metadata,
    media_body=media,
    fields='id',
    ocrLanguage='th'
).execute()
```

```
doc_id = file.get('id')
doc = service.files().export(fileId=doc_id,
    mimeType='text/plain').execute()
content = doc.decode('utf-8')
```



# OPTICAL CHATERISTICS RECOGNITIONS (OCR)



### Methods

Method ↑	Requests	Errors	Avg latency	99th percentile latency
google.apps.drive.v3.DriveFiles.Create	13,346	0.04%	2.103 seconds	8.175 seconds
google.apps.drive.v3.DriveFiles.Export	13,332	0.02%	0.268 seconds	1.041 seconds

3. ปัญหาและผลกระทบ

สถานการณ์ของข้อมูล ที่เกิดขึ้นในการจัดเก็บ

3. ปัญหาและผลกระทบด้านการจัดการข้อมูลของ

ที่เก็บข้อมูล

3.1 ปัญหาการจัดการข้อมูลของที่เก็บข้อมูล

การนำปัญหาต่างๆ ที่เกิดขึ้นเกี่ยวกับการจัดการข้อมูลไปเป็นแนวทาง เพื่อการเตรียมตัวกับสถานการณ์ที่เก็บข้อมูลในอนาคต ยังเป็นสิ่งที่มีความสำคัญต่อการจัดการข้อมูลของที่มีประสิทธิภาพ รายละเอียดของสถานการณ์ ดังต่อไปนี้ได้รับจากปัญหาต่างๆ ที่เกิดขึ้นจากการจัดการข้อมูลของที่มีประสิทธิภาพขององค์กรปกครองส่วนท้องถิ่นต่างๆ ซึ่งประกอบด้วย:

3.1.1 ปัญหาการเตรียมข้อมูล

- > เมื่อเจ้าหน้าที่เตรียมข้อมูลเพื่อใช้ในระบบสารสนเทศ การนำปัญหาที่เกิดขึ้นมาวิเคราะห์และ หาสาเหตุของปัญหาที่เกิดขึ้นจากข้อมูลที่เกิดขึ้นในระบบสารสนเทศ
- > ไม่มีการวางแผนหรือพิจารณาการป้องกันสถานการณ์ที่ข้อมูลของระบบสารสนเทศขององค์กรปกครองส่วนท้องถิ่น
- > ปัญหาของระบบสารสนเทศจากภายนอก เช่น ผู้ดูแลระบบ พนักงานที่ทำงาน หรือระบบที่เชื่อมกับระบบภายนอก ซึ่งไม่มีการจัดการป้องกันหรือป้องกันเหตุการณ์ เมื่อเกิดเหตุการณ์ขึ้นทำให้ระบบสารสนเทศส่วนหนึ่งไม่ทำงาน
- > ในกรณีที่ข้อมูลของระบบสารสนเทศมีความผิดปกติประมาณ 3-4 นาที ทำให้เกิดข้อมูลของระบบสารสนเทศ ไม่สามารถดำเนินการกับระบบสารสนเทศขององค์กรปกครองส่วนท้องถิ่นได้เป็นเวลาหลาย ชั่วโมงหรือมากกว่านั้น
- > ขาดการประสานงาน ฉุกเฉิน และบุคลากรที่เกี่ยวข้องเข้ามา เมื่อเกิดเหตุการณ์ของระบบสารสนเทศขึ้น ต้องรีบแก้ไขปัญหาที่เกิดขึ้น โดยที่ข้อมูลของระบบสารสนเทศเป็นจำนวนมาก
- > ขาดการประชาสัมพันธ์ ทำความเข้าใจกับประชาชนในการจัดการข้อมูล และการนำวิธี
- > หน่วยงานอื่น มีการที่ข้อมูลของระบบสารสนเทศ ทำให้ประชาชนสามารถแจ้งข้อผิดพลาดของข้อมูลได้ทันที หรือแจ้งและบุคลากรจัดการข้อมูลสามารถตรวจสอบ
- > ในกรณีฉุกเฉิน องค์กรปกครองส่วนท้องถิ่นสามารถจัดการข้อมูลของระบบสารสนเทศได้ แต่ ป้ายเตือนแจ้งเจ้าหน้าที่ข้อมูลด้านการจัดการข้อมูลของระบบสารสนเทศขององค์กรปกครองส่วนท้องถิ่น ทำให้บุคลากรที่เกี่ยวข้องสามารถแก้ไขปัญหาของระบบสารสนเทศได้

28

3.1.2 ปัญหาการเตรียมข้อมูล

- > องค์กรปกครองส่วนท้องถิ่นสามารถ ขาดอุปกรณ์และ เครื่องมือที่จำเป็นในการจัดการข้อมูลของระบบสารสนเทศ เช่น เครื่องมือระบบคอมพิวเตอร์ ระบบการสื่อสาร การเข้าถึงข้อมูลของระบบสารสนเทศขององค์กรปกครองส่วนท้องถิ่น
- > ข้อมูลของระบบสารสนเทศจากภายนอกของระบบสารสนเทศขององค์กรปกครองส่วนท้องถิ่น

3.1.3 ปัญหาการเตรียมข้อมูล

การเตรียมข้อมูลของระบบสารสนเทศขององค์กรปกครองส่วนท้องถิ่น การนำปัญหาที่เกิดขึ้นมาวิเคราะห์และ หาสาเหตุของปัญหาที่เกิดขึ้นจากข้อมูลที่เกิดขึ้นในระบบสารสนเทศ

3.1.4 ปัญหาการเตรียมข้อมูล

- > ขาดการประสานงาน ฉุกเฉิน และบุคลากรที่เกี่ยวข้องเข้ามา เมื่อเกิดเหตุการณ์ของระบบสารสนเทศขึ้น ต้องรีบแก้ไขปัญหาที่เกิดขึ้น โดยที่ข้อมูลของระบบสารสนเทศเป็นจำนวนมาก
- > ขาดการประชาสัมพันธ์ ทำความเข้าใจกับประชาชนในการจัดการข้อมูล และการนำวิธี
- > หน่วยงานอื่น มีการที่ข้อมูลของระบบสารสนเทศ ทำให้ประชาชนสามารถแจ้งข้อผิดพลาดของข้อมูลได้ทันที หรือแจ้งและบุคลากรจัดการข้อมูลสามารถตรวจสอบ
- > ในกรณีฉุกเฉิน องค์กรปกครองส่วนท้องถิ่นสามารถจัดการข้อมูลของระบบสารสนเทศได้ แต่ ป้ายเตือนแจ้งเจ้าหน้าที่ข้อมูลด้านการจัดการข้อมูลของระบบสารสนเทศขององค์กรปกครองส่วนท้องถิ่น ทำให้บุคลากรที่เกี่ยวข้องสามารถแก้ไขปัญหาของระบบสารสนเทศได้

29

## METHODOLOGY

### OPTICAL CHARACTERISTICS RECOGNITIONS (OCR)



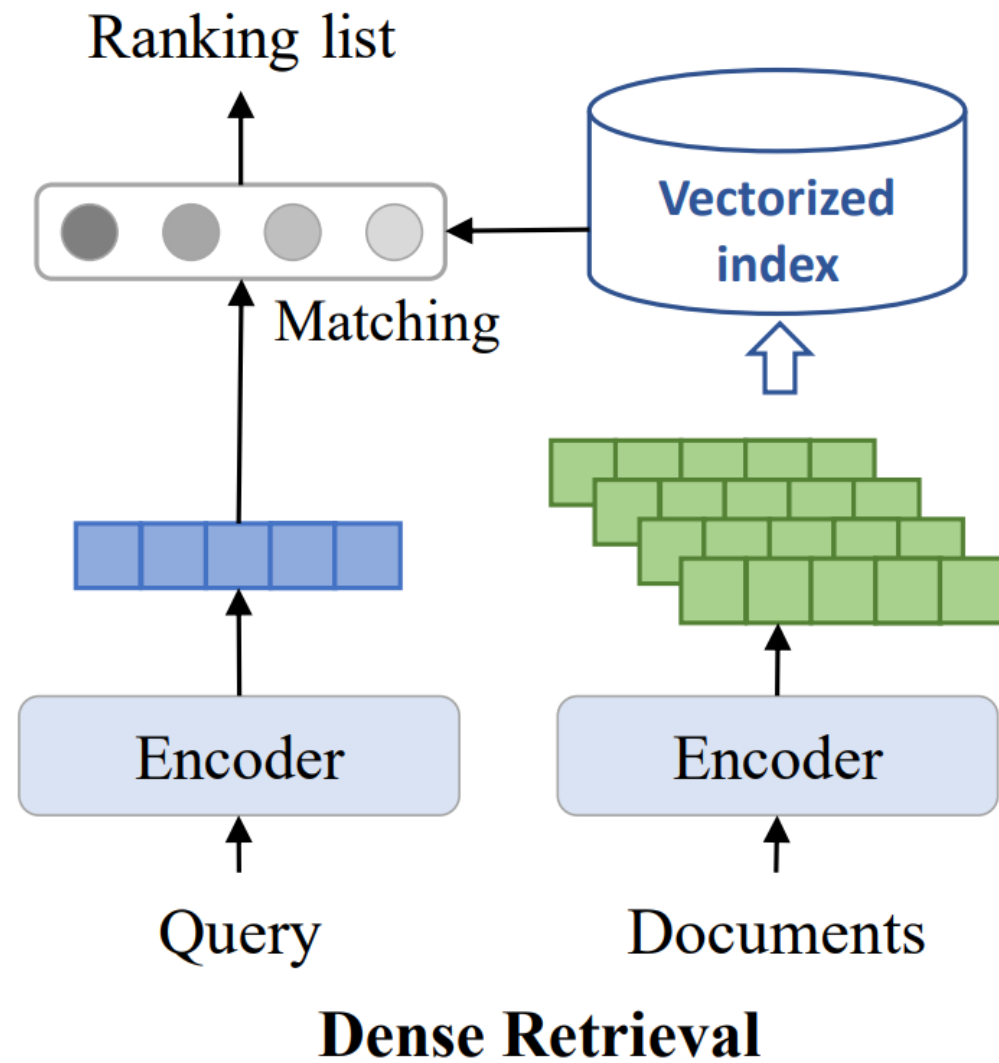
### TEXT CLEANING

- **Delete Punctuation:** In this step, punctuation marks such as \*, /, @, #, \$, %, and & are removed from the extracted text. This process aims to eliminate noise and ensure consistency in the text data.
- **Replace Thai Numbers with Arabic Numbers:** Following punctuation removal, Thai numbers are replaced with Arabic numerals. This standardizes the text format and facilitates easier comprehension and analysis of the data.
- **Delete Separators:** In this additional cleaning step, separators such as commas, periods, and hyphens are removed from the text. This further refines the text data, ensuring its suitability for subsequent analysis and processing.

METHODOLOGY



RETRIEVAL MODEL EXTRACTION



XQuAD

Base Model	R@1	MRR@10	Supervised?
<a href="#">simcse-model-distil-m-bert</a>	18.24	27.19	
<a href="#">simcse-model-m-bert-thai-cased</a>	22.94	30.29	
<a href="#">simcse-model-XLMR</a>	52.02	62.94	
<a href="#">simcse-model-wangchanberta</a>	53.87	65.51	
<a href="#">simcse-model-phayathaibert</a>	73.95	81.67	
<a href="#">SCT-model-XLMR</a>	55.29	65.23	
<a href="#">SCT-model-wangchanberta</a>	66.30	76.14	
<a href="#">SCT-model-phayathaibert</a>	67.56	76.14	
<a href="#">SCT-Distil-model-XLMR</a>	68.91	78.19	
<a href="#">SCT-Distil-model-wangchanberta</a>	62.27	72.53	
<a href="#">SCT-Distil-model-phayathaibert</a>	71.43	80.18	
<a href="#">ConGen-model-XLMR</a>	71.76	80.01	
<a href="#">ConGen-model-wangchanberta</a>	70.92	79.59	
<a href="#">ConGen-model-phayathaibert</a>	71.85	80.33	
<a href="#">ConGen-BGE_M3-model-phayathaibert</a>	85.80	90.48	
<a href="#">distiluse-base-multilingual-cased-v2</a>	49.16	58.19	✓
<a href="#">paraphrase-multilingual-mpnet-base-v2</a>	71.26	79.63	✓
<a href="#">BGE M-3</a>	90.50	94.33	✓
<a href="#">Cohere-embed-multilingual-v2.0</a>	82.52	87.78	✓

MIRACL

Base Model	R@1	MRR@10	Supervised?
<a href="#">simcse-model-distil-m-bert</a>	28.51	37.05	
<a href="#">simcse-model-m-bert-thai-cased</a>	26.19	36.11	
<a href="#">simcse-model-XLMR</a>	34.92	47.51	
<a href="#">simcse-model-wangchanberta</a>	36.29	48.96	
<a href="#">simcse-model-phayathaibert</a>	43.25	57.28	
<a href="#">SCT-model-XLMR</a>	28.51	40.84	
<a href="#">SCT-model-wangchanberta</a>	35.33	48.19	
<a href="#">SCT-model-phayathaibert</a>	37.52	51.02	
<a href="#">SCT-Distil-model-XLMR</a>	40.38	51.68	
<a href="#">SCT-Distil-model-wangchanberta</a>	39.43	50.61	
<a href="#">SCT-Distil-model-phayathaibert</a>	45.16	56.52	
<a href="#">ConGen-model-XLMR</a>	43.11	55.51	
<a href="#">ConGen-model-wangchanberta</a>	41.06	53.31	
<a href="#">ConGen-model-phayathaibert</a>	44.34	55.77	
<a href="#">ConGen-BGE_M3-model-phayathaibert</a>	70.40	79.33	
<a href="#">distiluse-base-multilingual-cased-v2</a>	17.74	27.78	✓
<a href="#">paraphrase-multilingual-mpnet-base-v2</a>	38.20	49.65	✓
<a href="#">BGE M-3</a>	79.67	86.68	✓
<a href="#">Cohere-embed-multilingual-v2.0</a>	66.98	77.58	✓



**DOCUMENT FILTERING AND INTERACTIVE FEEDBACK**

DOCUMENT FILTERING

1st



cosine similarity

$$\frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

2nd



0.73

3rd



0.88

The rang of cosine similarity is from 0-1

We will use the top-ranked document as the reference because I believe that the document most similar to the query is likely to contain more comprehensive information than the query itself. This document should have the most complete answer for that query.

DOCUMENT FILTERING



Query: แหล่งกำเนิดสารมลพิษทางอากาศมาจากอะไร

Document 1:

1.3 แหล่งกำเนิดสารมลพิษทางอากาศ แบ่งออกได้เป็น 2 ประเภทใหญ่ๆ คือ 1.3.1 แหล่งกำเนิดตามธรรมชาติ (Natural Sources) เป็นแหล่งกำเนิด ก่อให้เกิดสารมลพิษทางอากาศตามกระบวนการทางธรรมชาติ หรือไม่มีการกระทำ หรือกิจกรรมของมนุษย์เข้าไปเกี่ยวข้อง เช่น ภูเขาไฟระเบิด แผ่นดินไหว ไฟป่าตามธรรมชาติ การปล่อยก๊าซ จากการย่อยสลายซากพืชซากสัตว์โดยจุลินทรีย์ เป็นต้น 1.3.2 แหล่งกำเนิดที่เกิดจากกิจกรรมของมนุษย์ (Man-Made Sources หรือ Anthropogenic Sources) เป็นแหล่งกำเนิดที่เกิดจากกิจกรรมของมนุษย์

Score: 0.7672

Document 2:

(Emission Sources) อากาศหรือบรรยากาศ (Atmosphere) การแพร่กระจาย (Dispersion) ผู้รับผลกระทบ (Receptors) (1) แหล่งกำเนิดมลพิษทางอากาศ (Emission Sources) เป็นแหล่งที่ก่อให้เกิดมลพิษทางอากาศและระบายออกสู่บรรยากาศ โดยที่ชนิดและปริมาณของสารมลพิษ ที่ถูกระบายออกสู่อากาศขึ้นอยู่กับประเภทของแหล่งกำเนิด ลักษณะการเผาไหม้ ชนิดของเชื้อเพลิงที่ใช้ กิจกรรม และวิธีการควบคุมการระบายสารมลพิษทางอากาศ (2) อากาศหรือบรรยากาศ (Atmosphere) เป็นส่วนของระบบที่รองรับสารมลพิษอากาศที่ถูกระบายออกจากแหล่งกำเนิดต่างๆ และเป็นตัวกลาง

Score: 0.7538

Document 3:

เป็นแหล่งกำเนิดที่เกิดจากกิจกรรมของมนุษย์ ที่ทำให้มีการระบายสารมลพิษทางอากาศ แบ่งได้เป็น 3 ประเภท คือ 1. แหล่งกำเนิดมลพิษทางอากาศที่อยู่กับที่ หรือ Point Sources ได้แก่ การปลดปล่อยมลพิษจาก โรงงานอุตสาหกรรม (Industrial air pollution) ไร่หรือเหมืองจากคลังน้ำมันและสถานีบริการน้ำมัน เป็นต้น 2. แหล่งกำเนิดมลพิษทางอากาศที่เคลื่อนที่ หรือ Mobile Sources ได้แก่ มลพิษจากยานพาหนะ (Automobile air pollution) ประเภทต่าง ๆ เช่น รถยนต์ รถบรรทุก เรือ เครื่องบิน เป็นต้น 3. มลพิษทางอากาศที่ไม่มีแหล่งกำเนิดแน่นอน หรือ

Score: 0.7416

Document 4:

หยดของเหลว หรืออนุภาค ของแข็งก็ได้ สารมลพิษทางอากาศที่สำคัญและมีผลกระทบต่อสุขภาพอนามัย ได้แก่ ฝุ่นละออง สารตะกั่ว ก๊าซคาร์บอนมอนอกไซด์ ก๊าซซัลเฟอร์ไดออกไซด์ ก๊าซออกไซด์ของ ไนโตรเจน ก๊าซโอโซน และสารอินทรีย์ระเหยง่าย เป็นต้น ระบบภาวะมลพิษทางอากาศ มีส่วนประกอบที่มีความสัมพันธ์ กัน 3 ส่วน คือ แหล่งกำเนิดสารมลพิษ (Emission Sources) อากาศหรือ บรรยากาศ (Atmosphere) และผู้รับผลกระทบ (Receptors) แสดงเป็น แผนภูมิความสัมพันธ์ดังรูป สารมลพิษ (Pollutants) แหล่งกำเนิดสารมลพิษ (Emission Sources) อากาศหรือบรรยากาศ

Score: 0.6982

DOCUMENT FILTERING



Query: แหล่งกำเนิดสารมลพิษทางอากาศมาจากอะไร

Document 1:

1.3 แหล่งกำเนิดสารมลพิษทางอากาศ แบ่งออกได้เป็น 2 ประเภทใหญ่ๆ คือ 1.3.1 แหล่งกำเนิดตามธรรมชาติ (Natural Sources) เป็นแหล่งกำเนิด ก่อให้เกิดสารมลพิษทางอากาศตามกระบวนการทางธรรมชาติ หรือไม่มีการกระทำ หรือกิจกรรมของมนุษย์เข้าไปเกี่ยวข้อง เช่น ภูเขาไฟระเบิด แผ่นดินไหว ไฟป่าตามธรรมชาติ การปล่อยก๊าซ จากการย่อยสลายซากพืชซากสัตว์โดยจุลินทรีย์ เป็นต้น 1.3.2 แหล่งกำเนิดที่เกิดจากกิจกรรมของมนุษย์ (Man-Made Sources หรือ Anthropogenic Sources) เป็นแหล่งกำเนิดที่เกิดจากกิจกรรมของมนุษย์

Score: 1.0000

Document 2:

(Emission Sources) อากาศหรือบรรยากาศ (Atmosphere) การแพร่กระจาย (Dispersion) ผู้รับผลกระทบ (Receptors) (1) แหล่งกำเนิดมลพิษทางอากาศ (Emission Sources) เป็นแหล่งที่ก่อให้เกิดมลพิษทางอากาศและระบายออกสู่บรรยากาศ โดยที่ชนิดและปริมาณของสารมลพิษ ที่ถูกระบายออกสู่อากาศขึ้นอยู่กับประเภทของแหล่งกำเนิด ลักษณะการเผาไหม้ ชนิดของเชื้อเพลิงที่ใช้ กิจกรรม และวิธีการควบคุมการระบายสารมลพิษทางอากาศ (2) อากาศหรือบรรยากาศ (Atmosphere) เป็นส่วนหนึ่งของระบบที่รองรับสารมลพิษอากาศที่ถูกระบายออกจากแหล่งกำเนิดต่างๆ และเป็นตัวกลาง

Score: 0.8026

Document 3:

เป็นแหล่งกำเนิดที่เกิดจากกิจกรรมของมนุษย์ ที่ทำให้มีการระบายสารมลพิษทางอากาศ แบ่งได้เป็น 3 ประเภท คือ 1. แหล่งกำเนิดมลพิษทางอากาศที่อยู่กับที่ หรือ Point Sources ได้แก่ การปลดปล่อยมลพิษจาก โรงงานอุตสาหกรรม (Industrial air pollution) ไร่เหี้ยจากคลังน้ำมันและสถานีบริการน้ำมัน เป็นต้น 2. แหล่งกำเนิดมลพิษทางอากาศที่เคลื่อนที่ หรือ Mobile Sources ได้แก่ มลพิษจากยานพาหนะ (Automobile air pollution) ประเภทต่าง ๆ เช่น รถยนต์ รถบรรทุก เรือ เครื่องบิน เป็นต้น 3. มลพิษทางอากาศที่ไม่มีแหล่งกำเนิดแน่นอน หรือ Non-point sources ได้แก่ การเผาในที่โล่ง การลักลอบเผาป่า

Score: 0.9238

Document 4:

หอยดของเหลว หรืออนุภาค ของแข็งก็ได้ สารมลพิษทางอากาศที่สำคัญและมีผลกระทบต่อสุขภาพอนามัย ได้แก่ ฝุ่นละออง สารตะกั่ว ก๊าซคาร์บอนมอนอกไซด์ ก๊าซซัลเฟอร์ไดออกไซด์ ก๊าซออกไซด์ของ ไนโตรเจน ก๊าซโอโซน และสารอินทรีย์ระเหยง่าย เป็นต้น ระบบภาวะมลพิษอากาศ มีส่วนประกอบที่มีความสัมพันธ์ กัน 3 ส่วน คือ แหล่งกำเนิดสารมลพิษ (Emission Sources) อากาศหรือ บรรยากาศ (Atmosphere) และผู้รับผลกระทบ (Receptors) แสดงเป็น แผนภูมิความสัมพันธ์ดังรูป สารมลพิษ (Pollutants) แหล่งกำเนิดสารมลพิษ (Emission Sources) อากาศหรือบรรยากาศ

Score: 0.7409



**DOCUMENT FILTERING AND INTERACTIVE FEEDBACK**

Case1

2nd



0.88

3rd



0.73

IS THIS PAPER RELATED TO YOUR QUERY

Document 3

Relate

Not Relate





**DOCUMENT FILTERING AND INTERACTIVE FEEDBACK**

Case2

2nd



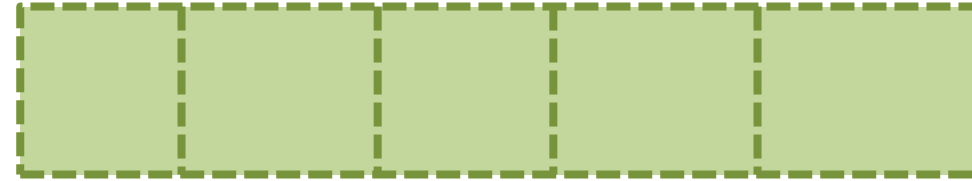
0.89

3rd



0.80

4th

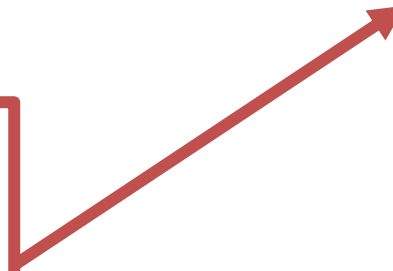


IS THIS PAPER RELATED TO YOUR QUERY

Document 4

Relate


Not Relate





WEBSITE DESIGN (Release Mode)

release mode  
experiment mode

 **AIT**  
Asian Institute of Technology

แอปพลิเคชันนี้เป็นระบบตอบสนองข้อความ  
ที่ใช้ LLMs ร่วมกับเทคโนโลยี RAG โดยใช้  
เอกสารจากกรมควบคุมมลพิษเป็นแหล่ง  
ข้อมูลหลัก

จัดทำโดย นายณัฐดนัย ศรีชัยญา MMI

**เลือกโมเดล**

โมเดลสำหรับการตอบคำถาม

Gemini Pro 1.5

**ประเภทของปัญหา**

เลือกหัวข้อที่สนใจ

- มลพิษทางอากาศ
- มลพิษทางน้ำ
- มลพิษทางเสียง
- มลพิษทางขยะ

ถามได้เลย

ตอบปัญหามลพิษด้วย AI 



## WEBSITE DESIGN (Release Mode)



สาเหตุและผลกระทบของมลพิษทางอากาศ



# สาเหตุและผลกระทบของมลพิษทางอากาศ

## สาเหตุ

มลพิษทางอากาศเกิดจากหลายสาเหตุ แบ่งเป็น 2 ประเภทหลักๆ ดังนี้:

- แหล่งกำเนิดตามธรรมชาติ:
  - ภูเขาไฟระเบิด
  - ไฟป่า
- แหล่งกำเนิดที่เกิดจากกิจกรรมของมนุษย์:
  - โรงงานอุตสาหกรรม: ปล่อยก๊าซต่างๆ เช่น ไนโตรเจนออกไซด์, ซัลเฟอร์ไดออกไซด์, ฝุ่นละออง และสารอินทรีย์ระเหยง่าย (VOCs)
  - ยานพาหนะ: ปล่อยก๊าซเช่น คาร์บอนมอนอกไซด์, ไนโตรเจนออกไซด์, ซัลเฟอร์ไดออกไซด์, คาร์บอน, สารไฮโดรคาร์บอน และฝุ่นละออง
  - กิจกรรมภายในบ้าน: คาร์บอนหรี, สารอินทรีย์ระเหยง่ายจากเครื่องเรือน, ฝุ่น, คาร์บอน, และคาร์บอนจากการทำครัว



## ผลกระทบ

มลพิษทางอากาศมีผลกระทบต่อสุขภาพอนามัยของมนุษย์ สัตว์ พืช และวัสดุต่างๆ ตัวอย่างผลกระทบที่สำคัญ:

- สุขภาพ: ก่อให้เกิดโรคทางเดินหายใจ, โรคหัวใจและหลอดเลือด, มะเร็งปอด และปัญหาระบบประสาท
- สิ่งแวดล้อม: ฝนกรด, ภาวะโลกร้อน, หมอกควัน, และทำลายพืชผล
- เศรษฐกิจ: ค่าใช้จ่ายด้านสุขภาพที่เพิ่มขึ้น, ผลผลิตทางการเกษตรลดลง, และผลกระทบต่อการท่องเที่ยว

ตัวอย่างสารมลพิษทางอากาศที่สำคัญและผลกระทบต่อสุขภาพ:


- ฝุ่นละออง: ระคายเคืองตา จมูก และคอ, โรคหอบหืด, โรคปอด
- สารตะกั่ว: ปัญหาสุขภาพประสาท, พัฒนาการทางสมองของเด็กบกพร่อง
- ก๊าซคาร์บอนมอนอกไซด์: ปวดศีรษะ, เวียนศีรษะ, คลื่นไส้, หมดสติ, เสียชีวิต
- ก๊าซซัลเฟอร์ไดออกไซด์: ระคายเคืองระบบทางเดินหายใจ, โรคหอบหืด



WEBSITE DESIGN (Release Mode)

release mode

experiment mode

 **AIT**  
Asian Institute of Technology

โหมคนี้เป็นส่วนหนึ่งในการทดลองของ  
วิทยานิพนธ์ Enhancing Retrieval-  
Augmented Generation for Thai Question-  
Answering through Cosine Similarity and  
Interactive Feedback

**ข้อมูลโมเดลที่ใช้**


- LLMs : Gemini Pro 1.5
- Temperature : 0.35
- Top P : 0.6
- Top K : 0

**ประเภทของปัญหา**

เลือกหัวข้อที่สนใจ

- มลพิษทางอากาศ
- มลพิษทางน้ำ
- มลพิษทางเสียง
- มลพิษทางขยะ

ถามได้เลย >

ตอบปัญหามลพิษด้วย AI 



WEBSITE DESIGN (Release Mode)

release mode  
experiment mode

AIT  
Asian Institute of Technology

โมเดลนี้เป็นส่วนหนึ่งในการทดลองของ  
วิทยานิพนธ์ Enhancing Retrieval-  
Augmented Generation for Thai Question-  
Answering through Cosine Similarity and  
Interactive Feedback

ข้อมูลโมเดลที่ใช้

- LLMs : Gemini Pro 1.5
- Temperature : 0.35
- Top P : 0.6
- Top K : 0

ประเภทของปัญหา

เลือกหัวข้อที่สนใจ

- มลพิษทางอากาศ
- มลพิษทางน้ำ
- มลพิษทางเสียง
- มลพิษทางขยะ

### เอกสารนี้เกี่ยวข้องกับคำถามของคุณหรือไม่

ความเกี่ยวข้องมากกว่า 0.75 :

1.2 แหล่งกำเนิดและผลกระทบของมลพิษทางอากาศและเสียงจากยานพาหนะ จากการวิเคราะห์ปัญหาและสาเหตุของมลพิษทางอากาศและเสียงที่เกิดขึ้นจากยานพาหนะ พบว่า สาเหตุหลักสองประการคือ ยานพาหนะที่มีจำนวนมากเกินไปและยานพาหนะที่มีมลพิษเกินมาตรฐาน การขาดระบบขนส่งมวลชนและขนส่งสินค้าที่มีประสิทธิภาพ ทำให้มียานพาหนะมากและการจราจรติดขัด ส่งผลให้ยานพาหนะเคลื่อนที่ด้วยความเร็วไม่สม่ำเสมอ มีช่วงเร่งและหยุด (start and stop) มาก ทำให้มีมลพิษทางอากาศและเสียงมากขึ้น

ในส่วนของยานพาหนะที่มีมลพิษเกินมาตรฐานที่มักจะเห็นได้ชัดคือ ยานพาหนะที่ไม่ได้รับการดูแลอย่างเพียงพอ เช่น เครื่องยนต์หลวม ไส้กรองอากาศอุดตัน ท่อไอเสียน้อยดัดแปลงเครื่องยนต์บรรทุกน้ำหนักเกินกำลัง เป็นต้น จากการติดตามตรวจสอบยานพาหนะตามท้องถนนโดยกรมควบคุมมลพิษมักจะพบรถโดยสารประจำทางและรถบรรทุกมีควันดำและเสียงเกินมาตรฐานอยู่เป็นจำนวนมาก

ในการจัดการปัญหามลพิษจากยานพาหนะ ต้องให้ความสนใจในเรื่องดังกล่าวเพื่อให้ผลการดำเนินงานเป็นที่ยอมรับของประชาชน มลภาวะทางอากาศและเสียงที่ก่อให้เกิดผลกระทบต่อคุณภาพสิ่งแวดล้อม เช่น ก๊าซคาร์บอน-มอนอกไซด์ ตะกั่ว ก๊าซซัลเฟอร์ไดออกไซด์ ก๊าซออกไซด์ของไนโตรเจน ก๊าซโอโซน ก๊าซไฮโดรคาร์บอน ฝุ่นรวม และฝุ่นขนาดเล็กที่เกิดจากยานพาหนะ เช่น ควันดำ และควันขาว ตลอดจนเสียงรบกวนที่เกิดจากยานพาหนะ โดยสาเหตุของมลภาวะทางอากาศและเสียงที่เกิดจากยานพาหนะและแหล่งกำเนิดอื่นๆ สรุปได้ดังนี้ 1.2.1


เอกสารนี้เกี่ยวข้องกับคำถาม

เอกสารนี้ไม่เกี่ยวข้องกับคำถาม



What about background in the thesis?

## ตอบปัญหามลพิษด้วย AI

 สาเหตุและผลกระทบของมลพิษทางอากาศ

### สาเหตุและผลกระทบของมลพิษทางอากาศ

มลพิษทางอากาศคือภาวะที่อากาศมีสารเจือปนในปริมาณที่มากเกินไปและเป็นระยะเวลายาวนานพอที่จะส่งผลเสียต่อสุขภาพอนามัยของมนุษย์ สัตว์ พืช และวัสดุต่างๆ

#### สาเหตุของมลพิษทางอากาศ

แหล่งกำเนิดของมลพิษทางอากาศแบ่งออกได้เป็น 2 ประเภทหลัก:

##### 1. แหล่งกำเนิดตามธรรมชาติ:

- ภูเขาไฟระเบิด: ปล่อยก๊าซซัลเฟอร์ไดออกไซด์, ฝุ่นละออง และก๊าซอื่นๆ
- ไฟป่า: ปล่อยก๊าซคาร์บอนมอนอกไซด์, ฝุ่นละออง และสารอินทรีย์ระเหยง่าย (VOCs)

##### 2. แหล่งกำเนิดที่เกิดจากกิจกรรมของมนุษย์:

- โรงงานอุตสาหกรรม: ปล่อยก๊าซไนโตรเจนออกไซด์, ซัลเฟอร์ไดออกไซด์, ฝุ่นละออง และ VOCs
- ยานพาหนะ: ปล่อยก๊าซคาร์บอนมอนอกไซด์, ไนโตรเจนออกไซด์, ซัลเฟอร์ไดออกไซด์, คาร์บอน, สารไฮโดรคาร์บอน และฝุ่นละออง
- กิจกรรมภายในบ้าน: ปล่อยควันบุหรี่, VOCs จากเครื่องเรือน, ฝุ่นละออง และเขม่าควันจากการทำครัว

## LARGE LANGUAGE MODELS (LLMs)



### ผลกระทบของมลพิษทางอากาศ

ผลกระทบต่อสุขภาพ:

- ระบบทางเดินหายใจ: โรคหอบหืด, โรคปอดอุดกั้นเรื้อรัง, มะเร็งปอด
- ระบบหัวใจและหลอดเลือด: โรคหัวใจ, โรคหลอดเลือดสมอง
- ระบบประสาท: ปัญหาพัฒนาการทางสมองในเด็ก, โรคอัลไซเมอร์
- ระบบอื่นๆ: การระคายเคืองตา, ผิวหนัง, และภูมิแพ้

ผลกระทบต่อสิ่งแวดล้อม:

- ฝนกรด: ทำลายพืชผล, ป่าไม้, และระบบนิเวศทางน้ำ
- ภาวะโลกร้อน: ก๊าซเรือนกระจก เช่น คาร์บอนไดออกไซด์ ทำให้เกิดการเปลี่ยนแปลงสภาพภูมิอากาศ
- หมอกควัน: ลดทัศนวิสัย, ส่งผลกระทบต่อการท่องเที่ยว และการบิน

### ตัวอย่างสารมลพิษทางอากาศ

- ก๊าซคาร์บอนมอนอกไซด์ (CO): เกิดจากการเผาไหม้ที่ไม่สมบูรณ์, เป็นพิษต่อระบบประสาท
- ไนโตรเจนออกไซด์ (NOx): เกิดจากการเผาไหม้ที่อุณหภูมิสูง, ก่อให้เกิดฝนกรด และหมอกควัน
- ซัลเฟอร์ไดออกไซด์ (SO2): เกิดจากการเผาไหม้เชื้อเพลิงที่มีกำมะถัน, ก่อให้เกิดฝนกรด
- สารอินทรีย์ระเหยง่าย (VOCs): เกิดจากการระเหยของสารเคมี, ก่อให้เกิดหมอกควัน และเป็นอันตรายต่อสุขภาพ
- ฝุ่นละออง: อนุภาคขนาดเล็ก, สามารถเข้าสู่ปอดและทำให้เกิดโรคทางเดินหายใจ

### การแก้ไขปัญหามลพิษทางอากาศ

- ลดการใช้ยานพาหนะส่วนบุคคล: เดิน, ปั่นจักรยาน, ใช้ระบบขนส่งสาธารณะ
- ใช้พลังงานสะอาด: พลังงานแสงอาทิตย์, พลังงานลม
- ควบคุมการปล่อยมลพิษจากโรงงานอุตสาหกรรม
- ปลูกต้นไม้: ช่วยดูดซับคาร์บอนไดออกไซด์

มลพิษทางอากาศเป็นปัญหาที่ซับซ้อนและมีผลกระทบต่อทุกคน การแก้ไขปัญหาต้องอาศัยความร่วมมือจากทุกภาคส่วน



What about background in the thesis?

## EXPERIMENT & RESULT

### LARGE LANGUAGE MODELS (LLMS)



### 3.1 Response Satisfaction Experiment

The Response Satisfaction Experiment aimed to assess the satisfaction levels of domain experts in the pollution area with the responses generated by two different modes of the system: the Release Mode and the Experiment Mode.

#### Summary of Response Satisfaction Experiment Results

Preferred Mode	Average Satisfaction Rating (1-10)
Release Mode	4.3
Experiment Mode	7.5



What about background in the thesis?

## EXPERIMENT & RESULT

### LARGE LANGUAGE MODELS (LLMs)



### 3.2 User Experience Experiment

The User Experience Experiment aimed to evaluate participants' satisfaction and preferences regarding the system's modes: the Experiment Mode and the Release Mode.

#### Summary of Response Satisfaction Experiment Results

Preferred Mode	Number of Participant	Main Reason for Preference
Release Mode	8	Faster response time, familiarity
Experiment Mode	32	AI responsiveness, accuracy, interactive features





What about background in the thesis?

## DISCUSSION & CONCLUSION

### LARGE LANGUAGE MODELS (LLMs)



Preferred Mode	Average Satisfaction Rating (1-10)			
	Air Pollutions	Water Pollutions	Noise Pollutions	Solid Waste Pollutions
Release Mode	3.9	4.0	4.4	4.7
Experiment Mode	8.6	8.0	4.9	8.3

Datasets Consist of 4 type of pollutions

- Solid waste pollutions : 97 document
- Water pollutions : 125 document
- Air pollutions : 126 document
- Noise pollutions : 23 document

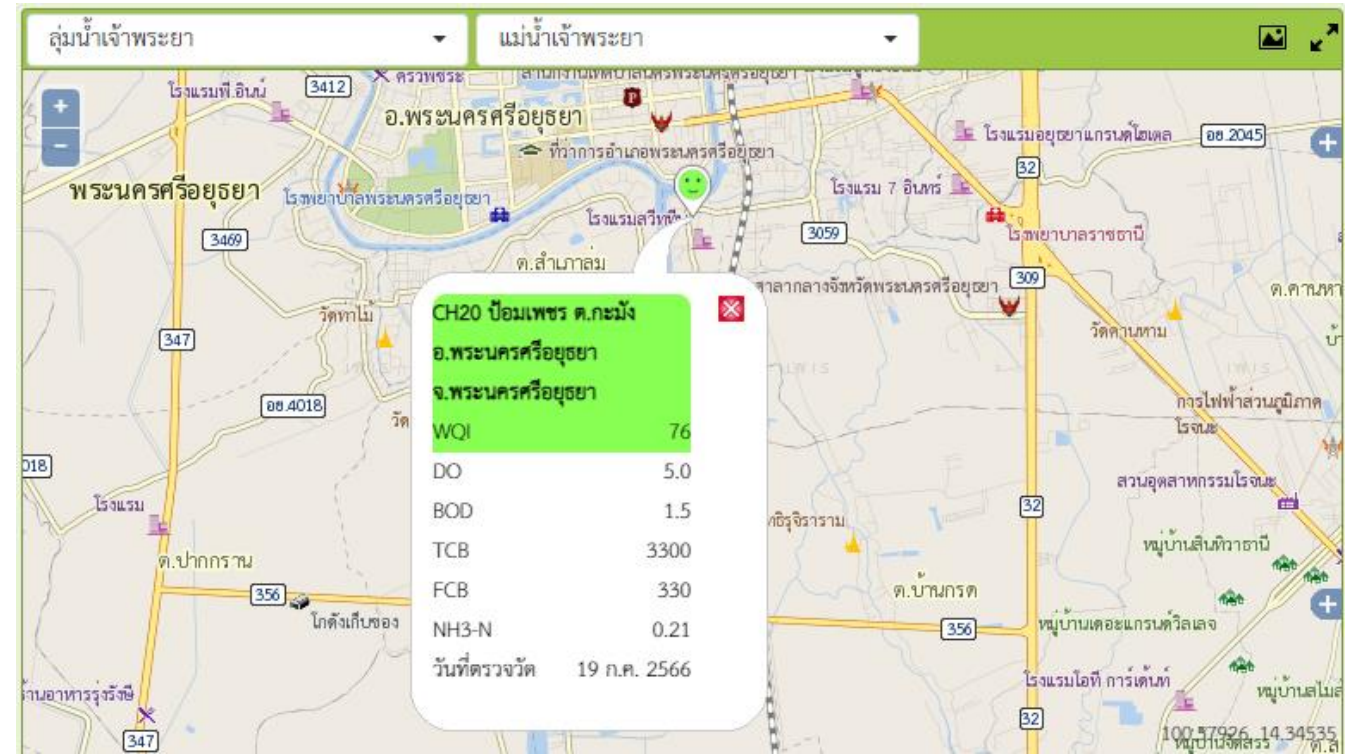




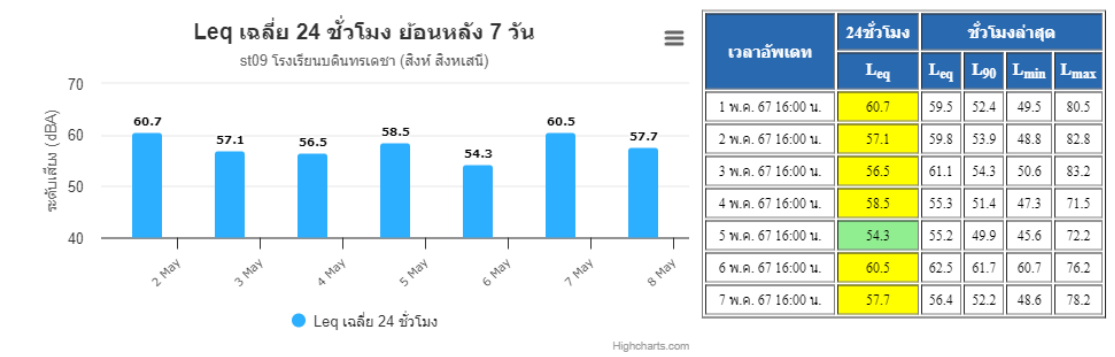
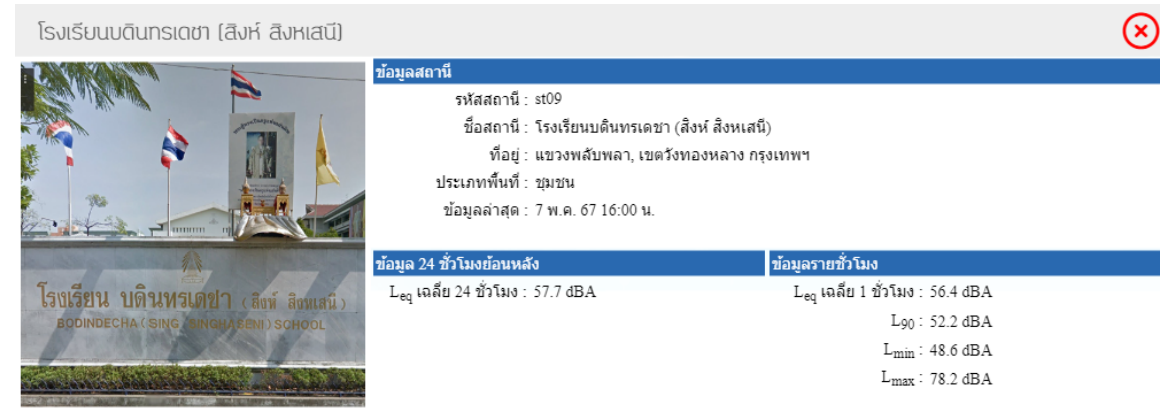
What about background in the thesis?

## DISCUSSION & CONCLUSION

## Future Work



- Expanding Base Knowledge
- Multimodal Interactions
- Real-world Deployment and Evaluation



## CONCLUSION



- Enhanced performance and user experience of Thai question-answering RAG systems in environmental pollution domain.
- Introduced cosine similarity-based filtering and interactive feedback system for better coherence and user input.
- Demonstrated significant improvements over baseline RAG system through experiments and user studies.
- Domain experts rated our system higher for response quality, relevance, and coherence.



**THANK YOU**

