ระบบปัญญาประดิษฐ์แบบ Retrieval-Augmented Generation สำหรับการให้คำแนะนำทางกฎหมายคุ้มครองผู้บริโภคในประเทศไทย

กริสนา ทรัพย์สินชัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
บัณฑิตศึกษา สาขาเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีไทย-ญี่ปุ่น
ปีการศึกษา 2568

A RETRIEVAL-AUGMENTED GENERATION SYSTEM FOR A LEGAL ADVISORY SYSTEM IN THAI CONSUMER PROTECTION

Kritsana Sapsinchai

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Information Technology
Graduate Studies
Thai-Nichi Institute of Technology
Academic Year 2025

หัวข้อวิทยานิพนธ์	ระบบปัญญาประดิษฐ์แบบ Retrieval-Augmented Generation สำหรับการให้คำแนะนำทางกฎหมายคุ้มครองผู้บริโภคในประเทศไทย
โดย	กริสนา ทรัพย์สินชัย
สาขาวิชา	เทคโนโลยีสารสนเทศ
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ดร.ศรายุทธ นนท์ศิริ

บัณฑิตศึกษา สถาบันเทคโนโลยีไทย-ญี่ปุ่น อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....รองอธิการบดีฝ่ายวิชาการ
(รองศาสตราจารย์ ดร.วรากร ศรีเชวงทรัพย์)
วันที่.....เดือน.....พ.ศ.....

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(รองศาสตราจารย์ ดร.อรรณพ หมั่นสกุล)

.....กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ฐิติพร เลิศรัตน์เดชากุล)

.....กรรมการ
(ดร.ณปภัช วิชัยดิษฐ์)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์
(ดร.ศรายุทธ นนท์ศิริ)

กริสนา ทรัพย์สินชัย : ระบบปัญญาประดิษฐ์แบบ Retrieval-Augmented Generation สำหรับการให้คำแนะนำทางกฎหมายคุ้มครองผู้บริโภคในประเทศไทย. อาจารย์ที่ปรึกษา : ดร.ศรายุทธ นนท์ศิริ, 46 หน้า.

เทคโนโลยีที่พัฒนาอย่างก้าวกระโดดได้เปลี่ยนพฤติกรรมของผู้บริโภคให้หันมาซื้อขายสินค้าและบริการผ่านช่องทางออนไลน์มากขึ้น ทำให้ตลาดอีคอมเมิร์ซเติบโตอย่างต่อเนื่อง แต่ในขณะเดียวกันมีผู้บริโภคจำนวนไม่น้อย ที่ประสบปัญหาจากการซื้อสินค้าและบริการผ่านช่องทางออนไลน์ โดยเฉพาะอย่างยิ่งผู้บริโภคที่ขาดความรู้ความเข้าใจในเรื่องกฎหมายที่เกี่ยวข้อง เพื่อการปกป้องสิทธิของตนเองตามกฎหมาย

งานวิจัยนี้มุ่งเน้นที่การนำเทคโนโลยีปัญญาประดิษฐ์เชิงสร้างสรรค์ ที่ใช้เทคนิค Retrieve Augmented Generation (RAG) ร่วมกับ Sematic Legal Preprocessing (SLP) ในการดึงข้อมูลกฎหมายที่เกี่ยวข้องกับคำถามจากผู้ใช้งาน นำมาประมวลผลร่วมกับโมเดลภาษาขนาดเล็ก Small Language Model (SLM) สร้างเป็นระบบแชทบอทเพื่อให้คำแนะนำทางด้านกฎหมายคุ้มครองผู้บริโภคของประเทศไทย ที่มีความแม่นยำและใช้ภาษาที่เข้าใจได้ง่ายให้แก่ผู้ใช้งาน โดยในงานวิจัยนี้ให้ผลลัพธ์ค่า Mean Reciprocal Rank (MRR) อยู่ที่ 0.54 ในขณะที่ Hit-Rate อยู่ที่ 0.71 และ Precision อยู่ที่ 0.71 สำหรับ Embedding Model และ Faithfulness อยู่ที่ 0.1249

บัณฑิตศึกษา ลายมือชื่อนักศึกษา .....

สาขาวิชา เทคโนโลยีสารสนเทศ ลายมือชื่ออาจารย์ที่ปรึกษา .....

ปีการศึกษา 2568

KRITSANA SAPSINCHAI : A RETRIEVAL-AUGMENTED GENERATION SYSTEM FOR A LEGAL ADVISORY SYSTEM IN THAI CONSUMER PROTECTION. ADVISOR : DR.SARAYUT NONSIRI, 46 PP.

Rapid technological advancements have significantly shifted consumer behavior toward online transactions, driving continuous growth in the e-commerce market. However, a significant number of consumers encounter issues when purchasing goods and services online, particularly those lacking the legal knowledge necessary to protect their statutory rights.

This research focuses on the implementation of Generative Artificial Intelligence, utilizing Retrieval-Augmented Generation (RAG) combined with Semantic Legal Preprocessing (SLP) to retrieve legal information relevant to user queries. By processing this data in conjunction with a Small Language Model (SLM), a chatbot system was developed to provide accurate and accessible advice on Thai consumer protection laws. The experimental results demonstrated a Mean Reciprocal Rank (MRR) of 0.54, a Hit-Rate of 0.71, and a Precision of 0.71 for the Embedding Model, with a Faithfulness score of 0.1249.

Graduate Studies Student's Signature.....

Field of Study Information Technology Advisor's Signature.....

Academic Year 2025

กิตติกรรมประกาศ

งานวิจัยฉบับนี้สำเร็จลุล่วงได้ด้วยดี เนื่องจากได้รับความกรุณาอย่างสูงจาก ดร.ศรายุทธ นนท์ศิริ อาจารย์ที่ปรึกษา ที่กรุณาให้คำแนะนำ ปรึกษา ตรวจสอบแก้ไขข้อบกพร่องต่างๆ ด้วยความเอาใจใส่อย่างดียิ่ง ผู้วิจัยตระหนักถึงความตั้งใจจริงและความทุ่มเทของอาจารย์ และขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ ที่นี้

ขอขอบพระคุณบิดามารดา ที่ช่วยเหลือสนับสนุนทั้งด้านกำลังใจและกำลังทรัพย์ด้วยดีตลอดมา จึงขอกราบขอบพระคุณ ณ โอกาสนี้ด้วย

ผู้วิจัยหวังเป็นอย่างยิ่งว่า วิทยานิพนธ์ฉบับนี้จะเกิดประโยชน์ต่อผู้ที่มีความสนใจ สามารถนำไปต่อยอดองค์ความรู้เพื่อประยุกต์ใช้กับงานด้านต่างๆ ที่เกี่ยวข้อง เพื่อให้เกิดคุณค่าและสามารถใช้งานได้จริง

งานวิจัยนี้ได้รับการสนับสนุนเงินทุนวิจัยจากกองทุนวิจัย Dr.Noriaki Kano

กริสนา ทรัพย์สินชัย

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง
บทคัดย่อภาษาอังกฤษ ..... จ
กิตติกรรมประกาศ ..... ฉ
สารบัญ ..... ช
สารบัญตาราง ..... ฌ
สารบัญรูป ..... ญ

บทที่

1 บทนำ 1
1.1 ความเป็นมาและความสำคัญ 1
1.2 วัตถุประสงค์ของงานวิจัย 2
1.3 ขอบเขตการวิจัย 2
1.4 ประโยชน์ที่จะได้รับ 3

2 ทฤษฎี หลักการ และงานวิจัยที่เกี่ยวข้อง 4
2.1 ปัญญาประดิษฐ์ (Artificial Intelligence - AI) 4
2.2 Machine Learning (ML) 4
2.3 Deep Learning และ Neural Network 5
2.4 Natural Language Processing (NLP) 6
2.5 Tokenization 7
2.6 Byte-Pair Encoding (BPE) 10
2.7 Transformer Architecture 12
2.8 Language Model (LM) 13
2.9 Chunk of text 14
2.10 Embedding model 14
2.11 Retrieval Augmented Generation (RAG) 15
2.12 Evaluation 16

สารบัญ (ต่อ)

บทที่		หน้า
3	วิธีดำเนินงานวิทยานิพนธ์	20
	3.1 การสร้างชุดข้อมูล (Dataset)	21
	3.2 Semantic Legal Preprocessing (SLP)	25
	3.3 Embedding	29
	3.4 ฐานข้อมูล เวกเตอร์ (Vector Database)	30
	3.5 Small Language Model (SLM)	31
	3.6 การประเมินผล (Evaluation)	32
	3.7 Hardware, Software and Service	35
4	ผลการดำเนินงาน การวิเคราะห์และสรุปผลต่างๆ	36
	4.1 ผลการเปรียบเทียบประสิทธิภาพของ Embedding Model	36
	4.2 ผลการเปรียบเทียบประสิทธิภาพของ Based Model	37
	4.3 การวิเคราะห์ข้อมูล	37
5	บทสรุป อภิปราย และข้อเสนอแนะ	40
	5.1 สรุปผลการวิจัย	40
	5.2 ปัญหาและอุปสรรคในการทำวิจัย	40
	5.3 ข้อเสนอแนะงานวิจัย	41
บรรณานุกรม		42
ประวัติย่อผู้วิจัย		46

สารบัญตาราง

ตาราง		หน้า
1.1	แผนการดำเนินงานวิจัย	3
2.1	สรุปงานวิจัยที่เกี่ยวข้อง	17
3.1	รายการข้อมูลดิบสำหรับสร้างชุดข้อมูล (Dataset)	22
3.2	รายการ Embedding Model	29
3.3	ตัวอย่างข้อมูลใน Vector Database	31
3.4	รายการ Small Language Model (SLM)	31
4.1	คะแนนประเมิน Embedding Model	36
4.2	คะแนนประเมิน Generation Model	37

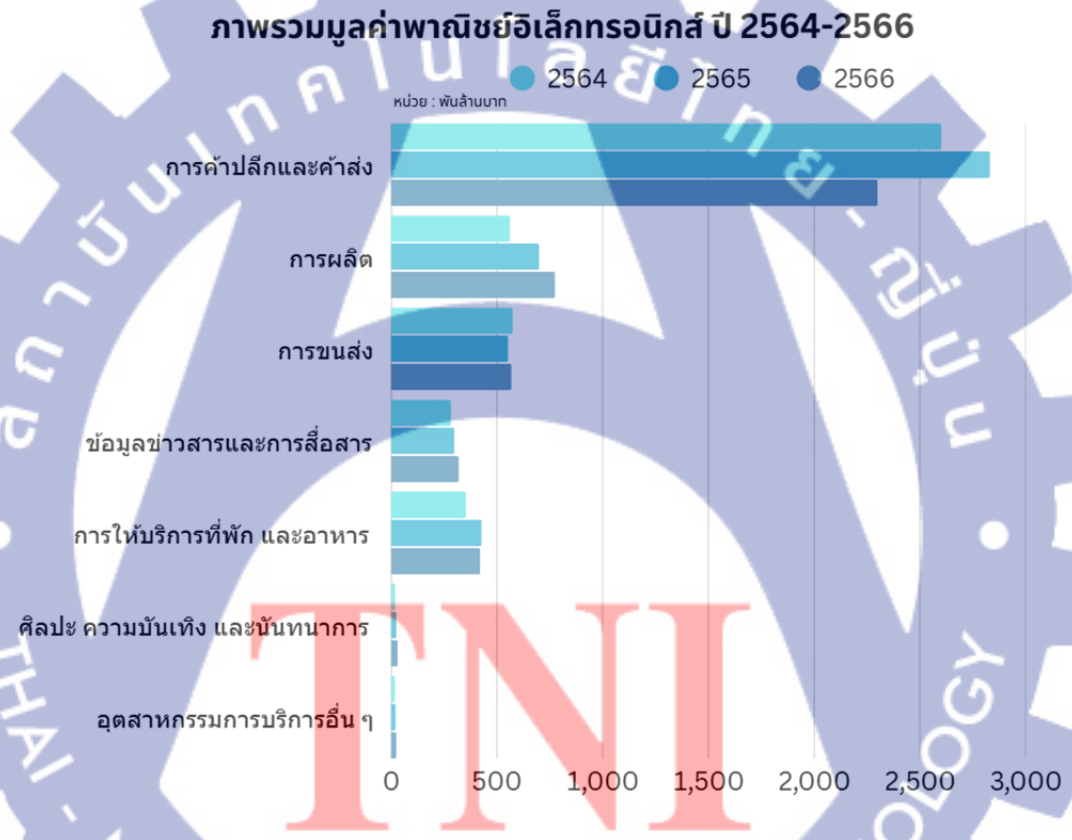
สารบัญรูป

รูป		หน้า
1.1	ภาพรวมมูลค่าพาณิชย์อิเล็กทรอนิกส์ ปี 2564-2566	1
2.1	Neural Network	5
2.2	วิวัฒนาการของ Natural Language Processing (NLP)	7
2.3	Word Tokenization	8
2.4	Character Tokenization	9
2.5	Sub-word Tokenization	10
2.6	Byte-Pair Encoding (BPE)	11
2.7	Byte-Pair Encoding (BPE)	11
2.8	Byte-Pair Encoding (BPE)	11
2.9	The Transformer-Model Architecture	13
2.10	ภาพจำลองการทำงานของ Embedding Model	15
3.1	ขั้นตอนการดำเนินงานวิจัย	20
3.2	Retrieval Augmented Generation (RAG)	21
3.3	ต้นฉบับเอกสารพระราชบัญญัติ	23
3.4	เอกสาร PDF หลังลบลายน้ำออกและบันทึกเป็นเอกสาร Microsoft Word	23
3.5	หน้าเว็บเพจของ สภาองค์กรของผู้บริโภค	24
3.6	Node SLP ในระบบ RAG	25
3.7	การทำงานภายใน SLP	28
3.8	ไฟล์ชุดข้อมูลในรูปแบบตาราง (Tabular Data)	28
3.9	Retrieval Evaluation	32
3.10	Generation Evaluation	34
4.1	ตัวอย่างการทำงานของโปรแกรม	38

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญ

เนื่องจากการเติบโตอย่างรวดเร็วของธุรกิจอีคอมเมิร์ซทั่วโลก การซื้อสินค้าและบริการผ่านช่องทางออนไลน์เป็นที่นิยมอย่างกว้างขวางในปัจจุบัน ในประเทศไทยโดยเฉพาะอุตสาหกรรมการค้าปลีกและค้าส่ง จากสถิติแสดงให้เห็นว่ามูลค่าของอุตสาหกรรมการค้าปลีกและค้าส่งในธุรกิจ อีคอมเมิร์ซของประเทศไทย สูงกว่าอุตสาหกรรมอื่น และเติบโตขึ้นอย่างต่อเนื่องดังแสดงในรูปที่ 1.1



รูปที่ 1.1 ภาพรวมมูลค่าพาณิชย์อิเล็กทรอนิกส์ ปี 2564-2566 [1]

อย่างไรก็ตาม มีผู้บริโภคจำนวนไม่น้อยที่ประสบปัญหาในการซื้อสินค้าและบริการผ่านช่องทางออนไลน์ ตัวอย่างเช่น การได้รับสินค้าไม่ตรงตามที่แสดงรายละเอียดไว้ สินค้าที่ได้รับชำรุดเสียหาย ได้รับสินค้าล่าช้า การบริการหลังการขายไม่ดี ถูกหลอกลวง หรือได้รับบริการที่ไม่เป็นไปตามข้อตกลงหรือไม่เป็นธรรม การถูกเอาเปรียบจากผู้ให้บริการ เป็นต้น

ซึ่งหากเกิดปัญหาเหล่าขึ้น กฎหมายคุ้มครองผู้บริโภคจึงถือเป็นเครื่องมือสำคัญในการช่วยเหลือผู้บริโภคให้ได้รับความเป็นธรรม อย่างไรก็ตามกฎหมายเป็นเรื่องที่เข้าใจยาก ต้องอาศัยผู้เชี่ยวชาญในกฎหมายเฉพาะด้านในการให้คำปรึกษา ซึ่งมีค่าใช้จ่ายที่สูง ขั้นตอนยุ่งยากและใช้เวลานาน ก่อให้เกิดความล่าช้า ทำให้ผู้บริโภคหลายคนรู้สึกท้อแท้ หรือไม่รู้จะรักษาสิทธิของตัวเองอย่างไร

เนื่องด้วยปัญหาดังกล่าวมีผลกระทบอย่างมากต่อเศรษฐกิจและสังคม จึงได้นำเสนอระบบให้คำปรึกษากฎหมาย โดยใช้ปัญญาประดิษฐ์เชิงสร้างสรรค์บูรณาการกับชุดข้อมูลทางกฎหมายที่เกี่ยวข้องกับผู้บริโภค เพื่อใช้ในการให้คำปรึกษากฎหมายแก่ผู้บริโภคได้อย่างถูกต้อง รวดเร็ว และเป็นภาษาที่เข้าใจได้ง่าย เพื่อให้ผู้บริโภคทราบถึงสิทธิของตนซึ่งกฎหมายคุ้มครอง

1.2 วัตถุประสงค์ของงานวิจัย

1.2.1 เพื่อสร้างระบบแชทบอทสำหรับให้คำปรึกษาปัญหากฎหมายคุ้มครองผู้บริโภคแก่ประชาชน

1.2.2 เพื่อเปรียบเทียบประสิทธิภาพของ Embedding Model และ Based Model ในภาษาไทยสำหรับกฎหมายไทย

1.3 ขอบเขตการวิจัย

1.3.1 ข้อมูลที่นำมาใช้ในระบบแชทบอท ได้แก่ พระราชบัญญัติคุ้มครองผู้บริโภค พ.ศ.2522 และกฎหมายที่เกี่ยวข้องกับการคุ้มครองผู้บริโภคของประเทศไทย

1.3.2 Embedding Model ที่นำมาเปรียบเทียบ ได้แก่

1.3.2.1 BGE-M3

1.3.2.2 WangchanX-Legal-ThaiCCL-Retriever

1.3.2.3 Snowflake-arctic-embed-m-v1.5

1.3.2.4 Multilingual-e5-base

1.3.2.5 All-MiniLM-L6-v2

1.3.2.6 Static-retrieval-mrl-en-v1

1.3.3 Base Model ที่นำมาเปรียบเทียบ ได้แก่

1.3.3.1 Open thaiGPT-1.5

1.3.3.2 Typhoon2.1-gemma3

1.3.3.3 Llama3.2

1.4 ประโยชน์ที่จะได้รับ

ได้ระบบสำหรับให้คำปรึกษากฎหมายคุ้มครองผู้บริโภคแก่ประชาชนทั่วไปในธุรกิจอีคอมเมิร์ซของประเทศไทย ให้มีความสะดวก และรวดเร็ว

ตารางที่ 1.1 แผนการดำเนินงานวิจัย

การดำเนินงาน	ระยะเวลา																	
	ปี 2567					ปี 2568												ปี 2569
	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1
ขั้นตอนเบื้องต้น																		
การหาหัวข้อ																		
ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง																		
ร่างเค้าโครงวิทยานิพนธ์																		
นำเสนอเค้าโครง																		
ขั้นตอนการดำเนินงาน																		
เก็บรวบรวมและเตรียมข้อมูลจากเว็บและเอกสารกฎหมาย																		
การประมวลผลและทำความสะอาดข้อมูล																		
จัดการข้อมูลด้วย SLP และสร้างระบบ RAG																		
ประเมินและเปรียบเทียบประสิทธิภาพโมเดล																		
เขียนรายงานผลการวิจัย																		
จัดทำและส่งตีพิมพ์/นำเสนอผลงาน																		
เขียนวิทยานิพนธ์เล่มสมบูรณ์																		
สอบป้องกันวิทยานิพนธ์																		

บทที่ 2
ทฤษฎี หลักการ และงานวิจัยที่เกี่ยวข้อง

ในงานวิจัยนี้ ได้มีการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง และได้นำเสนอตามหัวข้อต่อไปนี้

2.1 ปัญญาประดิษฐ์ (Artificial Intelligence - AI)

สาขาวิชาทางวิทยาศาสตร์คอมพิวเตอร์ที่มุ่งเน้นการพัฒนาโปรแกรมหรือเครื่องจักรให้มีความสามารถในการคิด แก้ปัญหา และตัดสินใจได้เหมือนกับมนุษย์ โดยทั่วไปแล้ว AI สามารถแบ่งได้เป็น 3 ประเภท ดังนี้

1) AI แบบแคบ (Narrow AI หรือ Weak AI) เป็น AI ที่ถูกออกแบบมาเพื่อทำงานเฉพาะอย่างใดอย่างหนึ่งเท่านั้น เช่น AI ที่ใช้เล่นหมากรุก หรือผู้ช่วยอัจฉริยะเช่น Siri และ Alexa AI ซึ่งมีให้เห็นกันอยู่ในปัจจุบัน

2) AI แบบทั่วไป (General AI หรือ Strong AI) เป็น AI ที่มีความสามารถในการเรียนรู้และแก้ปัญหาได้ทุกประเภทเหมือนกับมนุษย์ ซึ่งในปัจจุบันยังเป็นเพียงแนวคิดทางทฤษฎีและยังไม่มี AI ประเภทนี้ที่ใช้งานได้จริง

3) AI แบบขั้นสูง (Artificial Superintelligence) เป็น AI ที่มีความสามารถทางสติปัญญาสูงกว่ามนุษย์ในทุกๆ ด้าน ทั้งในด้านการคิดวิเคราะห์ ความคิดสร้างสรรค์ การแก้ปัญหา และการตัดสินใจที่ซับซ้อน ซึ่งในปัจจุบันยังเป็นเพียงแนวคิดเท่านั้น

2.2 Machine Learning (ML)

สาขาย่อยหนึ่งของ ปัญญาประดิษฐ์ (AI) ที่มุ่งเน้นการสร้างระบบที่สามารถ เรียนรู้ได้เองจากข้อมูล โดยไม่ต้องถูกตั้งโปรแกรมอย่างชัดเจนสำหรับแต่ละงาน แนวคิดหลักคือการปรับเปลี่ยนรูปแบบจากการกำหนดชุดคำสั่งทีละขั้นตอน (Imperative Programming) มาเป็นการประยุกต์ใช้ข้อมูลร่วมกับอัลกอริทึมเพื่อให้ระบบเรียนรู้และประมวลผลด้วยตนเอง แล้วให้คอมพิวเตอร์เรียนรู้จากข้อมูลเหล่านั้นเพื่อหา “รูปแบบ” (Patterns) หรือความสัมพันธ์ที่ซ่อนอยู่ ทำให้เมื่อคอมพิวเตอร์ได้รับข้อมูลใหม่ที่ไม่เคยเห็นก็จะสามารถเข้าใจและทำงานกับข้อมูลนั้นได้

การฝึกฝนโมเดล Machine Learning แบ่งได้เป็น 3 ประเภท

1) Supervised Learning คือการฝึกโมเดลโดยนำข้อมูลที่มีการติดฉลากข้อมูล (Data Labeling) ซึ่งเป็นป้ายกำกับข้อมูลดิบ เช่น รูปภาพ ข้อความ และเสียง เช่น ติดฉลาก “แมว” บนภาพ เพื่อที่จะได้ส่งเข้าไปให้โมเดลเรียนรู้ว่าภาพแมวมีคุณสมบัติอย่างไร ซึ่งโมเดลที่ถูกฝึกด้วยข้อมูลประเภทนี้เหมาะกับงานประเภท ทำนาย (Predictive) เช่น ทำนายสภาพอากาศ ทำนายราคาบ้าน เป็นต้น

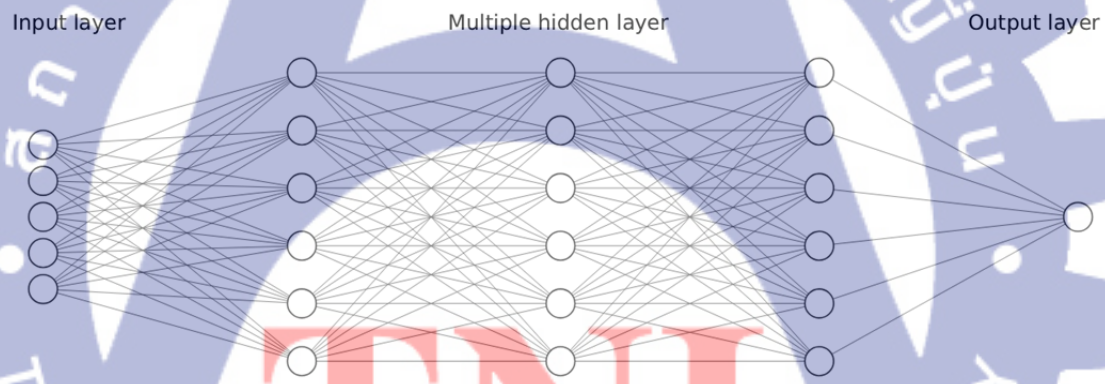
2) Unsupervised Learning คือการฝึกโมเดลโดยไม่การติดฉลากข้อมูล หรือไม่มีคำตอบของข้อมูลที่นำเข้า ซึ่งโมเดลที่ถูกฝึกด้วยข้อมูลประเภทนี้จะตัดสินใจแบ่งกลุ่มข้อมูลเองโดยพิจารณาจาก “รูปแบบ” (Patterns) ซึ่งโมเดลที่ฝึกด้วยวิธีนี้จะเหมาะกับงานประเภท แบ่งกลุ่ม (Segmentation) เช่น แบ่งกลุ่มผู้ซื้อสินค้า เป็นต้น

3) Reinforcement Learning คือการฝึกโมเดลโดยการให้คะแนนหรือรางวัล (Reward) ซึ่งข้อมูลดิบโดยไม่มีการติดฉลากข้อมูลจะถูกนำเข้าสู่โมเดล และโมเดลจะทำการตัดสินใจเองโดยการลองผิดลองถูก โดยมีเป้าหมายเพื่อเพิ่มคะแนนหรือ “รางวัล” (Reward) ที่ได้รับให้สูงสุด

การฝึกฝน Machine Learning จะทำให้โมเดลมีความรู้ความเข้าใจ เมื่อพบข้อมูลใหม่จะสามารถตัดสินใจได้อย่างถูกต้อง กระบวนการนี้ถือเป็นรากฐานสำคัญของปัญญาประดิษฐ์แขนงต่างๆ

2.3 Deep Learning และ Neural Network

สาขาย่อยหนึ่งของ Machine Learning ที่ใช้โครงข่ายประสาทเทียม (Neural Network) ที่มีหลายชั้นในการประมวลผลข้อมูลเพื่อให้ได้ Output ที่ต้องการ ดังแสดงในรูปที่ 2.1



รูปที่ 2.1 Neural Network

หลักการทำงานของ Neural Network คือการเรียนรู้เชิงลึก (Deep Learning) ผ่านชั้นข้อมูล (Layers) และโหนด (Nodes) ที่เปรียบเสมือนโครงข่ายประสาทในสมองมนุษย์ แต่ละโหนดจะรับค่าจากชั้นก่อนหน้าและประมวลผลด้วยหลักการทางคณิตศาสตร์

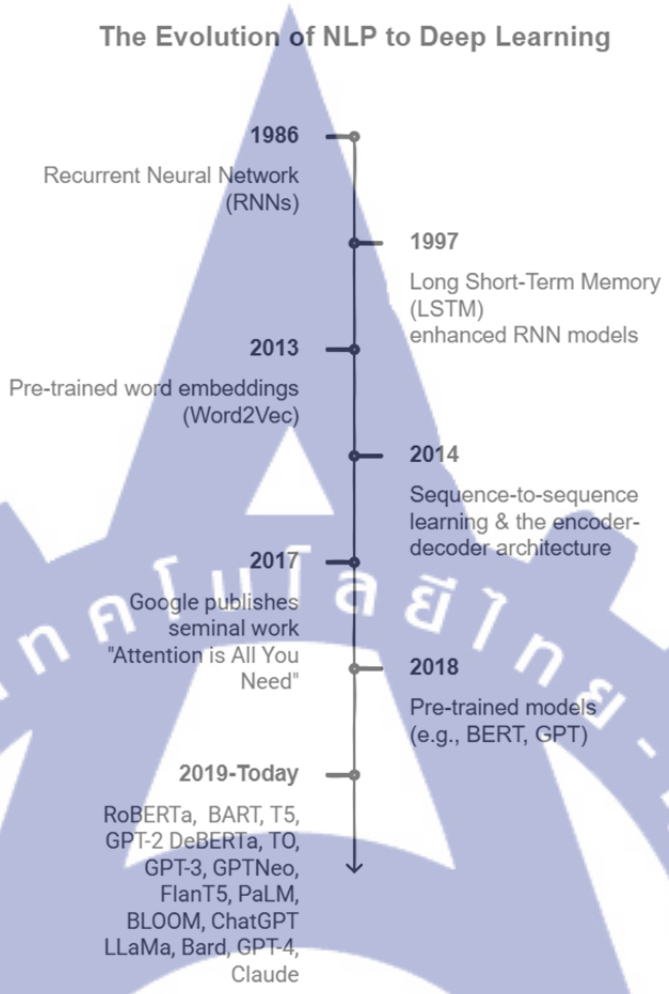
โมเดลจะเรียนรู้โดยการเปรียบเทียบผลลัพธ์ที่ได้กับข้อมูลจริง (Ground Truth) หากผลลัพธ์ยังไม่ตรงกัน ระบบจะใช้เทคนิคต่างๆ เช่น Gradient Descent เพื่อย้อนกลับไปปรับค่าน้ำหนัก (Weights) และไบแอส (Biases) ของแต่ละโหนด เพื่อให้สามารถหาค่าที่ถูกต้องและแม่นยำที่สุดในระยะเวลาที่เหมาะสม

การเรียนรู้แบบนี้ทำให้ Deep Learning แตกต่างจาก Machine Learning แบบดั้งเดิมอย่างชัดเจน ซึ่ง Deep Learning สามารถเรียนรู้และสกัดคุณสมบัติ (features) ที่สำคัญของข้อมูลได้ด้วยตัวเองโดยอัตโนมัติ โดยไม่ต้องอาศัยการกำหนดคุณสมบัติจากมนุษย์

2.4 Natural Language Processing (NLP)

Natural Language Processing (NLP) คือการประมวลผลภาษาธรรมชาติหรือภาษาที่ใช้สื่อสารกันระหว่างมนุษย์ เช่น ภาษาอังกฤษ ภาษาไทย ภาษาจีน เป็นต้น เป้าหมายของการประมวลผลภาษาธรรมชาติคือการทำให้มนุษย์สามารถสื่อสารกับคอมพิวเตอร์ได้ด้วยภาษาธรรมชาติ

จากรูปที่ 2.2 ในช่วงทศวรรษ 1950-1960 Natural Language Processing (NLP) เริ่มต้นจากแนวคิดในการใช้คอมพิวเตอร์แปลภาษา Machine Translation โดยใช้วิธีการถอดรหัสเพื่อแปลจากภาษาหนึ่งเป็นอีกภาษาหนึ่งแบบคำต่อคำโดยผ่านตัวกลาง อย่างไรก็ตาม เนื่องจากภาษามนุษย์เป็นภาษาที่มีทั้งไวยากรณ์และความหมาย การถอดรหัสเพียงอย่างเดียวจึงไม่เพียงพอต่อการประมวลผลภาษาธรรมชาติ ในช่วงทศวรรษ 1960-1970 มีการพัฒนาด้าน Natural Language Processing (NLP) หลายประการ ได้แก่ การแปลด้วยเครื่องแบบอิงตามกฎ (Rule-Based Machine Translation - RBMT) เป็นวิธีการเริ่มต้นที่อาศัยกฎทางภาษาที่กำหนดไว้ล่วงหน้า เช่น กฎไวยากรณ์และพจนานุกรม อย่างไรก็ตามวิธีนี้มีข้อจำกัดในการจัดการกับความกำกวมทางภาษาและความซับซ้อนของภาษาธรรมชาติ จึงได้มีการพัฒนาเป็นรูปแบบการแปลเชิงสถิติ (Statistical Machine Translation - SMT) โดยในช่วงทศวรรษที่ 1990 มีการใช้งาน Natural Language Processing (NLP) ที่กว้างขวางขึ้น และขนาดข้อมูลที่ใหญ่ขึ้น มีการรวบรวมข้อมูลจำนวนมากสร้างเป็นคลังข้อมูล (Corpus) เมื่อโครงสร้างของข้อมูลมีขนาดใหญ่ขึ้นจึงต้องการการประมวลผลด้วยคอมพิวเตอร์ที่เร็วขึ้น ปัจจัยต่างๆ เหล่านี้ ทำให้เกิดการพัฒนาอย่างรวดเร็วจนมีความสามารถในการดึงเอาความหมายของภาษาได้โดยอัตโนมัติโดยอาศัยหลักการทางคณิตศาสตร์และสถิติ ทำให้การพัฒนาด้าน Natural Language Processing (NLP) มีความก้าวหน้าอย่างรวดเร็ว



รูปที่ 2.2 วิวัฒนาการของ Natural Language Processing (NLP)

Natural Language Processing (NLP) มีขอบเขตกว้างขวางซึ่งครอบคลุมงานหลากหลายประเภทที่เกี่ยวกับภาษาธรรมชาติ เช่น การวิเคราะห์ความรู้สึก การจำแนกข้อความ และการสกัดข้อมูล งานวิจัยนี้จึงได้นำความสามารถดังกล่าวมาใช้ในด้านกฎหมายเพื่อใช้อธิบายหรือตีความภาษากฎหมายได้ แต่ยังคงมีความท้าทายในการฝึกให้เครื่องจักรเข้าใจภาษากฎหมายมีความซับซ้อน โดยเฉพาะอย่างยิ่งในภาษาไทย

2.5 Tokenization

Tokenization [2], [3] คือ กระบวนการแบ่งข้อความอินพุตออกเป็นส่วนย่อยๆ เรียกว่า โทเค็น (Tokens) โทเค็นเหล่านี้มีหลายระดับ ได้แก่ คำ ส่วนของคำ เครื่องหมายวรรคตอน หรือแม้แต่ตัวอักษรเดียวก็ได้ วัตถุประสงค์ของการแบ่ง โทเค็น (Tokens) เพื่อให้คอมพิวเตอร์สามารถประมวลผลได้อย่างมี

ประสิทธิภาพ เนื่องจากข้อความที่ไม่มีโครงสร้างได้ถูกแปลงให้อยู่ในรูปแบบที่มีโครงสร้างที่โมเดลแมชชีนเลิร์นนิงเข้าใจได้ การแบ่งโทเค็น (Tokens) ในระดับคำ จะช่วยในการจัดการคลังคำศัพท์ที่แตกต่างกันได้เป็นอย่างดี และยังช่วยในการประมวลผลคำศัพท์นอกคลัง (Out-of-Vocabulary - OOV) หรือคำศัพท์ที่ไม่คุ้นเคยด้วยการแบ่งเป็นคำศัพท์ย่อย Sub-word Tokenization ที่โมเดลแมชชีนเลิร์นนิงคุ้นเคยและเข้าใจคำศัพท์ได้ดียิ่งขึ้น ในการทำ Tokenization แบ่งออกได้เป็น 4 ประเภท ได้แก่

การทำโทเค็นระดับคำ (Word Tokenization) เป็นวิธีการพื้นฐานใน Natural Language Processing (NLP) ที่แบ่งข้อความออกเป็นคำเดี่ยว ๆ เหมาะกับคำที่มีขอบเขตชัดเจน เช่น ภาษาอังกฤษ อย่างไรก็ตาม การทำโทเค็นระดับคำ (Word Tokenization) นั้นยังคงมีปัญหาในเรื่องการขยายตัวของคำศัพท์ โดยวิธีนี้จะสร้างโทเค็นที่ไม่ซ้ำกันสำหรับคำแต่ละรูปแบบ แสดงในรูปที่ 2.3

“Let us Learn Tokenization.”

การทำโทเค็นระดับคำ (Word Tokenization)

“Let”, “us”, “Learn”, “Tokenization.”

รูปที่ 2.3 Word Tokenization

โดยจะแบ่งตามช่องว่างระหว่างคำแต่ละคำในประโยค อย่างไรก็ตามยังคงมีปัญหาสำหรับการแบ่งโทเค็นระดับคำ เช่น “run”, “running”, “ran” และ “runner” จะถูกนับเป็นโทเค็นแยกกัน แม้จะมีความหมายที่เกี่ยวข้องกันก็ตาม ซึ่งทำให้เกิดการใช้ทรัพยากรในการประมวลผลสูงและมีค่าใช้จ่ายมาก และยังพบปัญหาคำศัพท์นอกคลัง (Out-of-Vocabulary - OOV) หมายถึงคำศัพท์ที่หายากหรือไม่เคยปรากฏในข้อมูลการฝึกมาก่อนจะไม่สามารถประมวลผลได้ ทำให้โมเดลไม่รู้จักคำเหล่านั้น การแบ่งโทเค็นระดับคำไม่เหมาะกับภาษาที่มีโครงสร้างคำซับซ้อนหรือไม่ชัดเจน เช่น ภาษาจีน ญี่ปุ่น อาหรับ และภาษาไทย ที่ไม่มีการเว้นวรรคที่ชัดเจนระหว่างคำ ทำให้การแบ่งคำทำได้ยากและซับซ้อน อีกทั้งยังจัดการคำที่ประกอบขึ้นจากหน่วยย่อยหลายคำได้ยาก เช่น ที่อยู่อีเมล, URL หรือสัญลักษณ์พิเศษต่างๆ

การทำโทเค็นระดับตัวอักษร (Character Tokenization) [4] คือกระบวนการแบ่งข้อความออกเป็นหน่วยที่เล็กที่สุด นั่นคือตัวอักษรแต่ละตัว วิธีนี้จะช่วยให้โมเดลสามารถประมวลผลข้อความได้ในระดับที่ละเอียดมาก ดังตัวอย่างในรูปที่ 2.4

“Let us Learn Tokenization.”

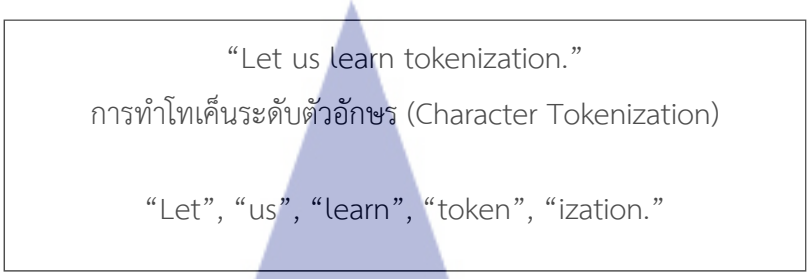
การทำโทเค็นระดับตัวอักษร (Character Tokenization)

“L”, “e”, “t”, “u”, “s”, “l”, “e”, “a”, “r”, “n”, “t”, “o”,
“k”, “e”, “n”, “i”, “z”, “a”, “t”, “i”, “o”, “n”, “.”

รูปที่ 2.4 Character Tokenization

การทำโทเคนประเภทนี้เหมาะสำหรับงานที่ต้องการการวิเคราะห์อย่างละเอียด เช่น การแก้ไขการสะกดคำ (Spelling Correction) มีประโยชน์อย่างยิ่งสำหรับภาษาที่ไม่มีการแบ่งคำที่ชัดเจน เช่น ภาษาจีน ญี่ปุ่น อาหรับ และภาษาไทย ซึ่งมักไม่มีช่องว่างที่ชัดเจนระหว่างคำ ช่วยแก้ปัญหาคำที่ไม่มีอยู่พจนานุกรม (Out-of-Vocabulary - OOV) เนื่องจากคำใดๆ ก็สามารถแยกย่อยเป็นชุดตัวอักษรที่รู้จักได้ ทำให้ไม่มีคำ “นอกพจนานุกรม” ที่ไม่เคยพบเจอ อย่างไรก็ตามวิธีการนี้ไม่สามารถจับความหมายเชิงอรรถ (Semantic Meaning) ของประโยคได้โดยตรง ซึ่งอาจนำไปสู่ความกำกวมได้ เนื่องจากเป็นการแบ่งข้อความออกเป็นหน่วยที่เล็กเกินไป ทำให้ความหมายของคำทั้งคำหรือวลีถูกแยกออกจากกัน อีกทั้งยังขาดความเข้าใจในด้านไวยากรณ์หรือโครงสร้างคำ เนื่องจากโมเดลจะพิจารณาเฉพาะความถี่ของรูปแบบตัวอักษรเท่านั้น ไม่ได้มีความรู้ด้านภาษาศาสตร์ในระดับที่สูงขึ้น และยังเพิ่มภาระการคำนวณ (Computational Burden) สำหรับลำดับข้อความที่ยาวขึ้น เพราะการแปลงทุกคำให้เป็นตัวอักษรแต่ละตัวจะส่งผลให้มีจำนวนโทเค็นที่ต้องประมวลผลเพิ่มขึ้นอย่างมาก ซึ่งอาจทำให้การฝึกโมเดลช้าลงและยากขึ้นในการจับความสัมพันธ์ระยะยาวในข้อความ

การทำโทเค็นแบบซับเวิร์ด (Sub-word Tokenization) [4] เป็นเทคนิคที่อยู่กึ่งกลางระหว่างการทำโทเค็นระดับคำ (Word Tokenization) และระดับตัวอักษร (Character Tokenization) โดยจะแบ่งคำออกเป็นหน่วยย่อยที่มีความหมาย ที่ใหญ่กว่าตัวอักษรหนึ่งตัวแต่เล็กกว่าคำเต็มๆ ซึ่งสามารถเป็นส่วนหนึ่งของคำหรือเป็นคำทั้งคำได้ วิธีนี้ช่วยให้โมเดลสามารถเข้าใจทั้งความหมายและโครงสร้างของข้อความได้อย่างมีประสิทธิภาพยิ่งขึ้น ดังตัวอย่างที่แสดงในรูปที่ 2.5



รูปที่ 2.5 Sub-word Tokenization

โดยรูปแบบการแบ่งคำขึ้นอยู่กับ อัลกอริทึม (Algorithm) และ คำศัพท์ หรือ “running”, “ran” และ “runner” ซึ่งจะไม่แยกเป็นคนละโทเค็น แต่จะแยกเป็นรากศัพท์ “run” และส่วนต่อท้าย “ning”, “an”, “ner “ ซึ่งช่วยลดขนาดคลังคำศัพท์ให้จัดการได้ง่ายขึ้น การจัดการคำที่ไม่เคยพบในคลังคำศัพท์ (Out-of-Vocabulary - OOV) เช่น “rerun” ซึ่งมีความคล้ายกัน และ โมเดลเรียนรู้คำศัพท์ “re”, “run” ก็ยังสามารถเข้าใจคำใหม่นี้ได้ โดยการรวมโทเค็นซับเวิร์ดเหล่านี้เข้าด้วยกัน และยังช่วยให้โมเดลเข้าใจความสัมพันธ์ทางความหมายและโครงสร้างของคำ ตัวอย่างเช่น โมเดลสามารถเชื่อมโยงความหมายของ "running", "ran" และ "runner" เข้ากับรากศัพท์ "run" ได้ ซึ่งช่วยให้การแสดงความหมายของคำ (Semantic Representation) มีประสิทธิภาพมากขึ้น

การทำโทเค็นแบบซับเวิร์ดเป็นเทคนิคสำคัญที่ใช้ในโมเดลภาษาขนาดใหญ่หลายตัว เช่น BERT, GPT และ T5 และเป็นแนวทางที่ช่วยให้โมเดลเหล่านี้สามารถประมวลผลภาษาธรรมชาติได้อย่างยืดหยุ่นและมีประสิทธิภาพ เหมาะสมอย่างยิ่งสำหรับงานที่ต้องใช้ความถูกต้องของภาษาเป็นสำคัญ อย่างเช่น งานด้านกฎหมาย

2.6 Byte-Pair Encoding (BPE)

อัลกอริทึม (Algorithm) ในการสร้างโทเค็นแบบ Sub-word (Sub-word Tokenization) ที่ใช้กันอย่างแพร่หลายในการประมวลผลภาษาธรรมชาติ (NLP) ได้แก่ Byte-pair Encoding [4] โดยเฉพาะอย่างยิ่งในโมเดลภาษาขนาดใหญ่ (Large Language Models : LLMs) และโมเดลภาษาขนาดเล็ก (Small Language Models : SLMs) เทคนิคนี้มีพื้นฐานมาจากอัลกอริทึมการบีบอัดข้อมูล จุดประสงค์คือการแยกข้อความออกเป็นหน่วยย่อยที่เล็กกว่าคำ แต่ใหญ่กว่าอักขระเดี่ยวๆ ซึ่งช่วยให้สามารถจัดการกับคำศัพท์ที่ไม่เคยเห็น (Out-of-Vocabulary - OOV) โดยไม่เพิ่มขนาดโมเดลได้อย่างมีประสิทธิภาพ

การทำงานของ Byte-Pair Encoding (BPE) เริ่มต้นด้วยการนำชุดข้อมูลที่เป็นข้อความมานับความถี่ของแต่ละคำที่ปรากฏ และแยกตัวอักษรแต่ละตัวของแต่ละคำเพื่อสร้างคลังคำศัพท์

ข้อความ

low low low low low lower lower newest newest widest widest widest

low: 5

lower: 2

newest: 2

widest: 3

คลังคำศัพท์เริ่มต้น รวมสัญลักษณ์สิ้นสุดของคำ </w>

l, o, w, e, r, n, s, t, i, d, </w>

รูปที่ 2.6 Byte-Pair Encoding (BPE)

จากรูปที่ 2.6 สังเกตว่าในคลังคำศัพท์ ตัวอักษรที่ซ้ำกันจะนับแค่ครั้งเดียว จากนั้นจึงเริ่มจับคู่ไบต์ที่พบบ่อยที่สุด จากตัวอย่างคือ o w (มาจาก low, lower) มีความถี่ที่พบถึง 7 ครั้ง จากนั้นรวมคู่ไบต์ o w เป็น ow และอัปเดตคลังคำศัพท์ โดยที่คำว่า low และ lower จะถูกนับ ow เป็น 1 ไบต์

l, o, w, e, r, n, s, t, i, d, </w>, ow

รูปที่ 2.7 Byte-Pair Encoding (BPE)

ทำซ้ำไปเรื่อยๆ จนไม่มีคู่ไบต์ที่สามารถรวมกันได้อีกแล้ว สุดท้ายคลังคำศัพท์จะประกอบไปด้วย

ตัวอักษรเดี่ยวๆ : l, o, w, e, r, n, s, t, d, i

Subwords ที่ถูกสร้างขึ้น: ow, es, est, lo, low, lowe, lower, new, newest, wi, wid, widest

สัญลักษณ์พิเศษ: </w>

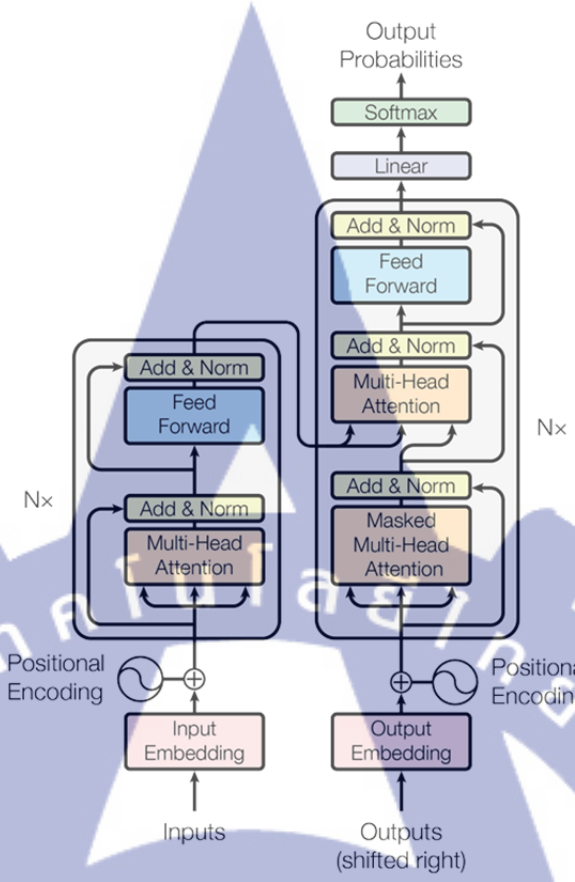
รูปที่ 2.8 Byte-Pair Encoding (BPE)

นี่คือหลักการที่ Byte-Pair Encoding (BPE) ใช้ในการสร้าง คลังคำศัพท์ (Vocabulary) ที่มีขนาดเหมาะสมและสามารถรองรับคำศัพท์ใหม่ๆ ได้อย่างมีประสิทธิภาพ การทำ Tokenization เป็นขั้นตอนการประมวลผลล่วงหน้า (Pre-Processing Step) ที่สำคัญสำหรับการฝังคำ (Embedding) ซึ่งเป็นกระบวนการที่แปลงโทเค็นเหล่านี้ให้เป็นเวกเตอร์ตัวเลขในพื้นที่ที่มีมิติสูง โดยที่โทเค็นที่มีความหมายคล้ายกันจะอยู่ใกล้กันในพื้นที่เวกเตอร์ (Vector Space) ซึ่ง Transformer Model จะเรียกใช้การฝังคำจากเมทริกซ์การฝัง เพื่อให้เข้าใจบริบทและความหมายของข้อความตามเวกเตอร์เหล่านั้น

2.7 Transformer Architecture

สถาปัตยกรรมเครือข่ายประสาทเทียม ซึ่งถูกออกแบบมาเพื่อจัดการกับข้อมูลประเภทลำดับ เช่น ข้อความ แต่มีจุดเด่นสำคัญคือสามารถประมวลผลข้อมูลได้พร้อมกันทั้งหมด ซึ่งช่วยให้สามารถจับความสัมพันธ์ระหว่างคำที่อยู่ห่างกันในประโยคได้ดี

สถาปัตยกรรม Transformer [5] ประกอบด้วยสองส่วนหลัก คือ Encoder และ Decoder แต่ละส่วนประกอบด้วยเลเยอร์ย่อยหลาย ๆ เลเยอร์ที่เรียงซ้อนกัน ดังที่แสดงในรูปที่ 2.9 โดยกลไกที่สำคัญในสถาปัตยกรรม Transformer คือ Self-Attention ซึ่งเป็นกระบวนการให้ความสำคัญในแต่ละคำในประโยคเพื่อดูว่าคำไหนเกี่ยวข้องกันมากที่สุด ทำให้เข้าใจความหมายของประโยคทั้งหมดได้ โดยไม่ต้องให้ความสำคัญกับทุกคำในประโยค ยกตัวอย่างเช่น เมื่ออ่านประโยค จะไม่ให้ความสำคัญกับทุกคำอย่างเท่าเทียมกัน แต่จะเน้นที่คำที่มีความหมายและเกี่ยวข้องกันจริงๆ ในบริบทนั้น ทำให้สามารถเข้าใจความสัมพันธ์ของคำและเข้าใจประโยคได้ ตัวอย่างโมเดลภาษาที่โดดเด่นและใช้กันอยู่ในปัจจุบัน เช่น GPT-5 from OpenAI [6], LLaMA from Meta [7], and Gemini from Google [8]. ล้วนถูกพัฒนาขึ้นโดยใช้ สถาปัตยกรรมTransformer ทั้งสิ้น



รูปที่ 2.9 The Transformer-Model Architecture

2.8 Language Model (LM)

โมเดลภาษาที่ถูกฝึกฝนด้วยชุดข้อมูลข้อความขนาดใหญ่และมีขนาดพารามิเตอร์ (Parameters) ตั้งแต่หลายพันล้านไปจนถึงหลายแสนล้านตัว ทำให้โมเดลเหล่านี้มีความสามารถในการเข้าใจภาษาธรรมชาติของมนุษย์ในแบบของเครื่องจักร โดยถูกสร้างจากสถาปัตยกรรม Transformer ดังแสดงในรูปที่ 2.9 โดยโมเดลภาษาจะแบ่งออกเป็น Large Language Model และ Small Language Model โดยมีลักษณะที่แตกต่างกัน ดังนี้

Large Language Model (LLM) คือ โมเดลภาษาขนาดใหญ่ที่ถูกฝึกด้วยข้อมูลจำนวนมาก โดยโมเดลจะมีขนาดตั้งแต่ 100 พันล้าน (100B) ตัว ถึงไปจนถึงหลาย Trillion (ล้านล้าน) พารามิเตอร์ ทำให้โมเดลสามารถตอบคำถามที่ซับซ้อนได้ดี โมเดลที่มีชื่อเสียง เช่น Gemini, ChatGPT, Sonnet เป็นต้น ซึ่งในการประมวลผลโมเดลภาษาขนาดใหญ่นี้ต้องให้ทรัพยากรจำนวนมากในการสร้างคำตอบออกมา

Small Language Model (SLM) คือ โมเดลภาษาที่มีขนาดเล็กกว่าโมเดลภาษาขนาดใหญ่ (Large Language Mode) โดยทั่วไปแล้ว SLM จะมีจำนวนพารามิเตอร์ (Parameters) น้อยกว่า 10 พันล้าน (10B) ตัว หรือบางครั้งอาจมีเพียงไม่กี่ร้อยล้านตัว เช่น Meta-Llama-3-8B เป็นต้น ทำให้สามารถประมวลผลโดยใช้ทรัพยากรที่น้อยกว่าโมเดลภาษาขนาดใหญ่มาก เช่น คอมพิวเตอร์ส่วนบุคคล อย่างไรก็ตาม ผลลัพธ์ที่ได้ไม่อาจเทียบเท่ากับโมเดลภาษาขนาดใหญ่ เนื่องจากโมเดลถูกฝึกด้วยข้อมูลจำนวนน้อยกว่า

2.9 Chunk of text

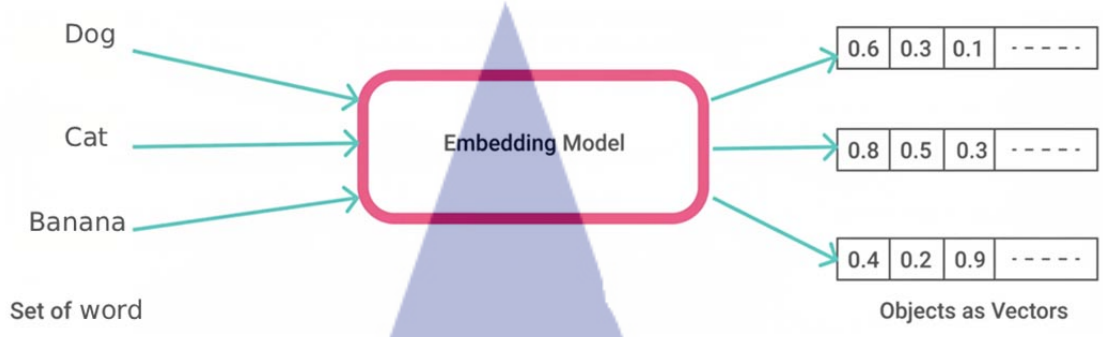
วิธีการแบ่งข้อความขนาดใหญ่ออกเป็นส่วนย่อยๆ เพื่อให้มีขนาดเล็กลงและสามารถจัดการได้ง่ายขึ้น เป้าหมายหลักของการแบ่งส่วนนี้คือ

- ลดขนาดข้อมูล ทำให้ข้อมูลมีขนาดเล็กพอที่จะส่งไปยังโมเดลภาษาได้
- เพิ่มประสิทธิภาพการค้นหา ช่วยให้การค้นหาข้อมูลที่เกี่ยวข้องกับคำถามทำได้แม่นยำและรวดเร็วยิ่งขึ้น

การแบ่งข้อมูลทำโดยการแบ่งเอกสารออกเป็นส่วนๆ โดยกำหนดจำนวนอักขระ (Characters) ที่แน่นอน เช่น ทุกๆ 500 อักขระ ซึ่ง Chunk of Text นั้นมีความสำคัญอย่างยิ่งในกระบวนการ Retrieval Augmented Generation (RAG) เพื่อใช้ในการเตรียมข้อมูล อย่างไรก็ตามในบริบททางกฎหมาย ที่ต้องการเหตุผลและความหมาย วิธีนี้ไม่อาจรักษาบริบททางกฎหมายไว้ได้ รวมถึงข้อความที่แบ่งได้อาจเกิดความไม่สอดคล้องกับโครงสร้างทางกฎหมาย และยังส่งผลให้ได้ผลลัพธ์ที่ผิดพลาดได้

2.10 Embedding model

เป็นแบบจำลองทางคณิตศาสตร์ที่ทำหน้าที่แปลงข้อความให้อยู่ในรูปของ Vector (เวกเตอร์) หรือชุดตัวเลขที่แสดงถึงความหมายและบริบทของข้อความนั้นๆ ตัวเลขเหล่านี้จะถูกจัดเรียงในช่องว่างแบบหลายมิติ โดยที่เวกเตอร์ของข้อมูลที่มีความหมายคล้ายคลึงกันหรืออยู่ในบริบทเดียวกันจะอยู่ใกล้กันในช่องว่างนั้นดังแสดงในรูปที่ 2.10 ซึ่งมีความสำคัญต่อการทำให้คอมพิวเตอร์เข้าใจภาษาของมนุษย์ ซึ่งมีประโยชน์อย่างมากในการประยุกต์ใช้ร่วมกับโมเดลภาษา โดยเฉพาะอย่างยิ่งในงานที่ต้องการความเข้าใจภาษามนุษย์อย่างมีเหตุมีผล



รูปที่ 2.10 ภาพจำลองการทำงานของ Embedding Model

ในงานวิจัยนี้นำโมเดล BGE-M3, WangchanX-Legal-ThaiCCL-Retriever, Snowflake-Arctic-Embed-m-v1.5, Multilingual-e5-Base, all-MiniLM-L6-v2, Static-Retrieval-mrl-en-v1 เพื่อใช้ในการแปลงข้อความภาษาไทยให้เป็นตัวเลขเวกเตอร์เพื่อใช้ในการประมวลผล

2.11 Retrieval Augmented Generation (RAG)

เทคนิคที่ช่วยให้โมเดลภาษาสามารถสร้างคำตอบที่ถูกต้องและเป็นปัจจุบันมากขึ้น โดยการดึงข้อมูลจากแหล่งความรู้ภายนอกมาประกอบการสร้างคำตอบ แทนที่จะอาศัยเพียงแค่ข้อมูลที่ถูกฝึกมาเท่านั้นซึ่งหากข้อมูลมีการเปลี่ยนแปลงแก้ไข โมเดลภาษาจะไม่มีข้อมูลทำให้ไม่สามารถใช้งานได้อย่างถูกต้อง อาจให้ข้อมูลที่ไม่เป็นจริง (hallucination) โดยขั้นตอนการทำงานของ RAG แบ่งออกเป็น 2 ขั้นตอนหลัก ได้แก่ Retrieval (การดึงข้อมูล) และ Augmentation and Generation (การเพิ่มบริบทและการสร้างคำตอบ)

Retrieval (การดึงข้อมูล) ขั้นตอนนี้คือการค้นหาและดึงข้อมูลที่เกี่ยวข้องจากฐานความรู้ภายนอกโดยการรวบรวมเอกสาร เช่น ประมวลกฎหมาย และแปลงเอกสารเหล่านี้ให้เป็น ตัวเลขเวกเตอร์ จากนั้นจัดเก็บในฐานข้อมูลเวกเตอร์ เมื่อผู้ใช้ถามคำถาม ระบบจะแปลงคำถามนั้นเป็น ในตัวเวกเตอร์เช่นกัน แล้วนำไปค้นหาเอกสารในฐานข้อมูลเวกเตอร์ที่มีความคล้ายคลึงทางความหมายโดยการเปรียบเทียบตัวเลขเวกเตอร์ เอกสารที่ถูกค้นพบว่าเกี่ยวข้องมากที่สุดจะถูกดึงออกมา เพื่อใช้เป็นข้อมูลอ้างอิง

Augmentation and Generation (การเพิ่มบริบทและการสร้างคำตอบ) ขั้นตอนนี้คือการนำข้อมูลที่ดึงมาได้ไปใช้ประกอบการสร้างคำตอบ โดยระบบจะนำคำถามของผู้ใช้และข้อมูลที่ดึงมาได้จากขั้นตอน Retrieval มาสร้างเป็น Prompt หรือชุดคำสั่งใหม่ที่สมบูรณ์ โดย Prompt ใหม่นี้จะถูกส่งไปยังโมเดลภาษาเพื่อใช้ในการปรับแต่งข้อมูลให้มีภาษาที่เหมาะสม มีความถูกต้องและมีเหตุผล

2.12 Evaluation

ในการสร้างระบบค้นหาข้อมูลหรือระบบตอบคำถามจำเป็นต้องมีการประเมินประสิทธิภาพของระบบ การเลือกเมตริกที่ใช้ประเมินประสิทธิภาพของระบบนั้นมีความสำคัญอย่างมาก หากใช้เมตริกที่ไม่เหมาะสม จะทำให้การประเมินประสิทธิภาพของระบบคลาดเคลื่อน ส่งผลให้เกิดปัญหาเมื่อนำระบบไปใช้

เมตริกที่เหมาะสมสำหรับระบบค้นหาข้อมูลหรือระบบตอบคำถาม สำหรับงานวิจัยนี้ แบ่งเป็น 2 ส่วน ได้แก่ ส่วนการดึงข้อมูล Retrieval ซึ่งใช้ 3 เมตริก ได้แก่ ได้แก่ Hit-Rate@k, Mean Reciprocal Rank (MRR), Precision@k และส่วนสร้างคำตอบ Generation ซึ่งใช้ 1 เมตริก ได้แก่ Faithfulness โดยมีหลักการทำงานดังนี้

1) Mean Reciprocal Rank (MRR) คือเมตริกที่ใช้ประเมินประสิทธิภาพของระบบการดึงข้อมูล โดยจะให้ความสำคัญกับ ตำแหน่งของข้อมูลที่เกี่ยวข้องที่อยู่อันดับแรก ที่ระบบค้นเจอ โดย MRR จะให้คะแนนสูงขึ้น หากข้อมูลที่ถูกต้องรายการแรกนั้นอยู่ใกล้ตำแหน่งบนสุดของรายการผลลัพธ์ โดยมีสมการดังนี้

$$MRR = \frac{1}{N}\sum_{i-1}^{N}\frac{1}{rank\_i} \quad (2.1)$$

2) Precision@k คือเมตริกที่ใช้ประเมินประสิทธิภาพของระบบการการดึงข้อมูล โดยจะวัดจำนวนของรายการทั้งหมดที่อยู่ใน Top-k ซึ่ง k หมายถึงตัวเลขที่กำหนดขอบเขตของการพิจารณา เช่น k = 5 หมายถึง 5 ลำดับแรกของผลลัพธ์ที่นำมาพิจารณา เมตริกนี้จะให้คะแนนสูงขึ้นหากระบบสามารถดึงข้อมูลที่ถูกต้องมาไว้ในกลุ่มลำดับต้นๆ ได้เป็นจำนวนมาก โดยมีสมการดังนี้

$$Precision@k = \frac{Relevant\ Items\ in\ Top\ k}{k} \quad (2.2)$$

3) Hit-Rate@k คือเมตริกที่ใช้ประเมินประสิทธิภาพของระบบการการดึงข้อมูล โดยจะวัดว่าในคำค้นหานั้นระบบสามารถพบข้อมูลที่เกี่ยวข้อง “อย่างน้อยหนึ่งรายการ” ในผลลัพธ์ Top-k หรือไม่ ซึ่ง k หมายถึงตัวเลขที่กำหนดขอบเขตของการพิจารณา เช่น k = 5 หมายถึง 5 ลำดับแรกของผลลัพธ์ที่นำมาพิจารณา โดยคะแนนที่สูงขึ้นหมายถึงระบบมีคำตอบที่เกี่ยวข้องให้กับคำถามมากขึ้น โดยมีสมการดังนี้

$$\text{Hit Rate@k} = \frac{\text{Queries} + \text{Relevant in Top k}}{\text{All Queries}} \quad (2.3)$$

4) Faithfulness คือหนึ่งในเมตริกที่สำคัญที่สุดในการประเมินระบบ RAG (Retrieval-Augmented Generation) ซึ่งใช้ในการประเมินประสิทธิภาพของส่วนการสร้างคำตอบให้แก่ผู้ใช้ Generation ซึ่งจะเปรียบเทียบว่าผลลัพธ์คำตอบที่โมเดลสร้างมาเกี่ยวข้องกับข้อมูลที่ดึงมาจากฐานข้อมูลเวกเตอร์หรือไม่ ซึ่งเมตริกนี้จะแสดงให้เห็นถึงประสิทธิภาพของโมเดลภาษา โดยมีสมการดังนี้

$$F = \frac{|\text{Sverified}|}{|\text{Stotal}|} \quad (2.4)$$

โดยสรุป การเลือกใช้เมตริกสำหรับการวัดผลนั่นมีความสำคัญอย่างยิ่งในการพัฒนาระบบที่ทำงานร่วมกับปัญญาประดิษฐ์ เพราะจะแสดงให้เห็นได้ถึงความก้าวหน้าในการพัฒนาและแสดงถึงจุดอ่อนในระบบเพื่อใช้สำหรับการปรับปรุงระบบให้มีประสิทธิภาพยิ่งขึ้น

ตารางที่ 2.1 สรุปงานวิจัยที่เกี่ยวข้อง

ลำดับที่	ปี	วัตถุประสงค์	เทคนิค	ชุดข้อมูล	ผลลัพธ์
1	[9], 2025	เพื่อเพิ่มประสิทธิภาพของ RAG โดยการแก้ไขปัญหาความไม่สอดคล้องทางด้านรูปแบบ ระหว่างคำถามของผู้ใช้และเนื้อหาในเอกสาร	Hypothetical Prompt Embeddings (HyPE)	MS MARCO, Ragas-WikiQA, RAG-dataset-12000, MultiHopRAG	เพิ่มความแม่นยำของบริบทในการดึงข้อมูลได้สูงสุดถึง 42% และเพิ่มความครบถ้วนของการอ้างอิงข้อมูลได้สูงสุดถึง 45% เมื่อเทียบกับวิธีการมาตรฐานทั่วไป
2	[10], 2025	เพื่อนำเสนอภาพรวมที่ครอบคลุมของวิธีการ ความท้าทาย และการประยุกต์ใช้ RAG ในด้านกฎหมาย	RAG Stages, The Core Processes Include Information Retrieval (IR), Augmentation, and Generation	BorderLegal-QA, JEC-QA, CJRC, CAIL2020, CAIL2021	เพื่อเพิ่มความแม่นยำของ โมเดลภาษาขนาดใหญ่ และลดการสร้างข้อมูลที่ผิดพลาด ในสาขาที่มีความสำคัญสูง อย่างเช่นกฎหมาย

ตารางที่ 2.1 สรุปงานวิจัยที่เกี่ยวข้อง (ต่อ)

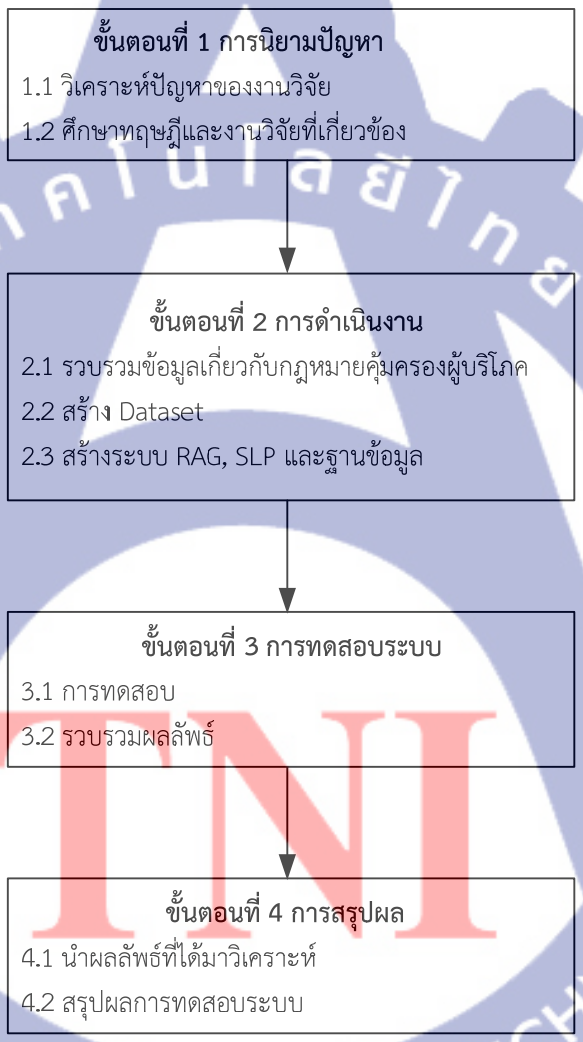
ลำดับที่	ปี	วัตถุประสงค์	เทคนิค	ชุดข้อมูล	ผลลัพธ์
3	[11], 2025	เพื่อนำเสนอ “ผู้ช่วยตรวจปล่อยสินค้าอัจฉริยะ” สำหรับงานศุลกากร โดยใช้เทคโนโลยี RAG ในการจัดการข้อมูลที่หลากหลายรูปแบบ และแก้ไขปัญหาคำถามที่มีความคลุมเครือ	Multimodal Parsing, Processes PDFs, Images, and Tables Using Tools Like PyMuPDF and Tesseract OCR.	Dataset of Chinese Customs Materials, including Tariff Chapters, Customs Laws, and inspection and Quarantine Ordinances	ความถูกต้องของคำตอบ เพิ่มขึ้น 20.1%, ความเกี่ยวข้องของเนื้อหาเพิ่มขึ้น 15.3% และความซื่อตรงของข้อมูล เพิ่มขึ้น 18.7%
4	[12], 2025	เพื่อนำเสนอโครงสร้างระบบ ใหม่ที่ใช้การตอบกลับแบบเรียกซ้ำ เพื่อเพิ่มความโปร่งใส รวมถึงป้องกันการสร้างข้อมูลที่ผิดพลาด (Hallucinations) และอคติ (Bias) ในการประยุกต์ใช้กับงานด้านกฎหมาย	Hybrid Fine-Tuned Generative LLM (HFM)	Dsynthetic, Domain-specific Q&A	ประสิทธิภาพโดยรวมเพิ่มขึ้น 24% เมื่อเทียบกับ LLM ทั่วไป โดยแบ่งเป็นประสิทธิภาพที่เพิ่มขึ้น 9% ในงานทั่วไป และเพิ่มขึ้นถึง 38% ในงานเฉพาะทางด้านกฎหมาย
5	[13], 2025	เพื่อเพิ่มประสิทธิภาพการสรุปเนื้อหาทางกฎหมาย โดยการรวมระบบ RAG แบบไดนามิก เข้ากับการปรับแต่งให้เหมาะสมกับบริบทเฉพาะทางด้านกฎหมาย	Dynamic Legal RAG System, Legal (NER)	Indian Penal Code, Civil Procedure Code, Criminal Procedure Code	BERTScore: 0.89, ROUGE-L:0.8894, Cosine Similarity: 0.94
6	[14], 2023	เพื่อนำเสนอวิธีการกำหนดคำสั่งแบบ Zero-shot (Zero-Shot Prompting) ที่เรียบง่าย เพื่อสอนให้ LLM สามารถคาดการณ์คำพิพากษาได้	Legal Syllogism Prompting (LoT)	CAIL2018, Criminal Case	Logic-of-Thought (LoT) สามารถทำความแม่นยำ (Accuracy) ได้ที่ 0.6850

งานวิจัยเหล่านี้นำเสนอแนวทางที่แตกต่างกันเพื่อเพิ่มประสิทธิภาพของระบบ Retrieval Augmented Generation (RAG) [9] สำหรับสาขาเฉพาะทาง เช่น กฎหมาย [10], [11], [12], [13], [14] และระเบียบศุลกากร โดยแนวทางเหล่านี้มีเป้าหมายเพื่อแก้ไขข้อจำกัดที่สำคัญของ Large Language Models (LLMs) เช่น การสร้างข้อมูลที่ไม่เป็นจริง (Hallucinations) และความรู้ที่ล้าสมัย ด้วยการอ้างอิงข้อมูลภายนอกที่เจาะจงเฉพาะสาขา ทำให้ระบบเหล่านี้ทำงานได้ดีและแสดงให้เห็นถึงการปรับปรุงที่สำคัญในเรื่องความแม่นยำของคำตอบและความเข้าใจในบริบท แสดงในตารางที่ 2.1

อย่างไรก็ตาม ยังคงมีช่องว่างหลายประการโดยงานวิจัยนี้มุ่งเน้นเป็นพิเศษในสาขากฎหมายที่มีความละเอียดอ่อนด้านกฎหมายคุ้มครองผู้บริโภคและรองรับภาษาไทย ซึ่งเป็นบริบทที่มักถูกมองข้ามในงานวิจัยที่มีอยู่ ชุดข้อมูลของงานวิจัยนี้ถูกคัดสรรมาอย่างพิถีพิถันโดยใช้เทคนิค Prompt Engineering เพื่อสร้างคำถามและคำตอบที่มีโครงสร้าง และเพื่อให้แน่ใจว่าการประเมินผลมีความน่าเชื่อถือ และยังใช้เมตริกที่เหมาะสมสำหรับ RAG เช่น MRR, Precision และ Hit-Rate ซึ่งให้การประเมินประสิทธิภาพของ Embedding Model ในส่วนของการดึงข้อมูล (Retrieve) และ Faithfulness ซึ่งเมตริกที่เหมาะสมสำหรับการประเมินประสิทธิภาพในส่วนของการสร้างข้อมูลคำตอบ (Generation) ที่ครอบคลุมมากกว่าวิธีการแบบดั้งเดิม

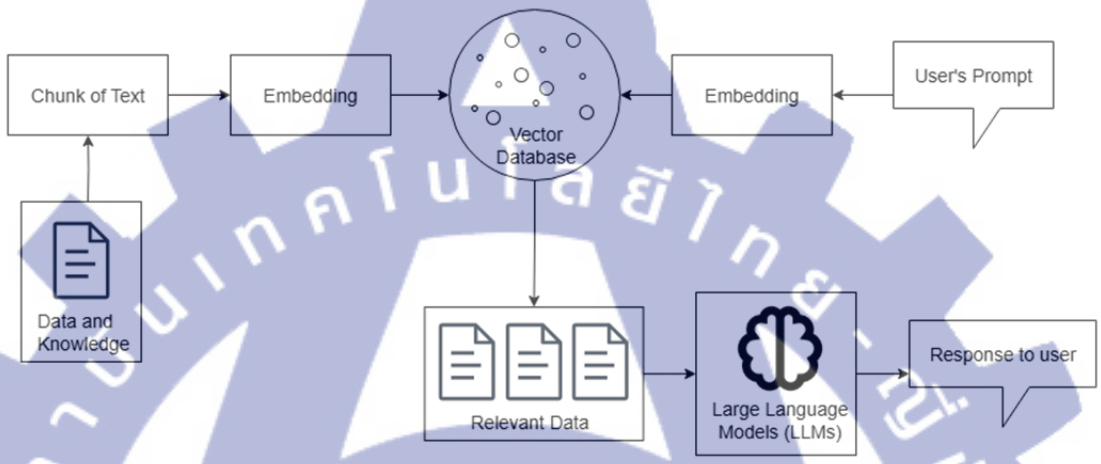
บทที่ 3
วิธีดำเนินงานวิทยานิพนธ์

การวิจัยเรื่อง การให้คำปรึกษากฎหมายคุ้มครองผู้บริโภคโดยใช้ปัญญาประดิษฐ์เชิงสร้างสรรค์สำหรับอีคอมเมิร์ซของประเทศไทย เป็นการสร้างระบบที่บูรณาการองค์ความรู้ด้านกฎหมายเข้ากับเทคโนโลยีปัญญาประดิษฐ์ โดยมีขั้นตอนการดำเนินงานดังแสดงในรูปที่ 3.1**ผิดพลาด! ไม่พบแหล่งการอ้างอิง**



รูปที่ 3.1 ขั้นตอนการดำเนินงานวิจัย

ในงานวิจัยนี้ เริ่มด้วยการหาหัวข้องานวิจัย วิเคราะห์ปัญหา โดยปัญหาของงานวิจัยนี้คือการให้คำปรึกษาด้านกฎหมายคุ้มครองผู้บริโภคโดยใช้เทคโนโลยีปัญญาประดิษฐ์เชิงสร้างสรรค์ (Generative AI) เพื่อให้ได้ผลลัพธ์ที่แม่นยำในข้อกฎหมาย และเป็นภาษาที่ผู้ใช้เข้าใจได้ง่าย และสามารถประมวลผลได้บนเครื่องคอมพิวเตอร์ส่วนบุคคล โดยได้นำเทคนิค Retrieval Augmented Generation (RAG) มาทำงานร่วมกับโมเดลภาษาขนาดเล็กช่วยทำให้ระบบสามารถตอบคำถามได้อย่างแม่นยำและสามารถทำงานบนเครื่องคอมพิวเตอร์ส่วนบุคคลได้ โดยมีรูปแบบการทำงานดังแสดงในรูปที่ 3.2



รูปที่ 3.2 Retrieval Augmented Generation (RAG)

ด้วยระบบ RAG แบบดั้งเดิมพบว่ายังคงมีปัญหาเรื่องของการจัดการข้อมูลก่อนส่งข้อมูลเข้าสู่ขั้นตอนการฝังเวกเตอร์ ซึ่งโดยทั่วไปจะใช้วิธีการ แบ่งข้อความ Chunks of Text ซึ่งจะแบ่งข้อความยาวๆ ออกเป็นส่วนๆ ตามจำนวนความยาวที่กำหนดไว้ ผลที่ได้คือข้อความยาวๆ ที่ถูกตัดเป็นส่วนเท่าๆ กัน โดยไม่สนใจบริบทในข้อความ

3.1 การสร้างชุดข้อมูล (Dataset)

การรวบรวมข้อมูลที่เกี่ยวข้องกับกฎหมายคุ้มครองผู้บริโภคนั้นได้มาจาก 2 แหล่ง ได้แก่

1) สำนักงานคณะกรรมการคุ้มครองผู้บริโภค (สคบ.) โดยจะเป็นข้อมูลดิบในส่วนของพระราชบัญญัติจำนวน 11 ฉบับ ซึ่งเป็นกฎหมายที่บังคับใช้โดยทั่วกัน และเป็นเอกสารที่เผยแพร่เป็นสาธารณะ โดยประกอบไปด้วยกฎหมายที่เกี่ยวข้องกับ สินค้าและบริการ เช่น พระราชบัญญัติเครื่องสำอาง พ.ศ. 2558, พระราชบัญญัติคุ้มครองผู้ประสบภัยจากรถ พ.ศ. 2535 เป็นต้น โดยพระราชบัญญัติเหล่านี้จัดเกี่ยวข้องโดยตรงกับผู้บริโภค ซึ่งได้บัญญัติไว้เพื่อคุ้มครองผู้บริโภคในด้านต่างๆ ทั้งที่เกี่ยวข้องกับสินค้าและบริการ

รวมถึงที่ได้บัญญัติไว้สำหรับกระบวนพิจารณาในชั้นศาล พระราชบัญญัติเหล่านี้จึงถือเป็นข้อมูลหลักที่นำมาใช้ในงานวิจัยนี้

2) สภาองค์กรของผู้บริโภค [15] ซึ่งเกิดจากการรวมตัวขององค์กรผู้บริโภค โดยเป็นข้อมูลในส่วนของบทความ ข่าวสาร และกรณีที่เกิดขึ้นจริงซึ่งเกี่ยวข้องกับการคุ้มครองผู้บริโภคที่เผยแพร่เป็นสาธารณะ จำนวน 47 บทความ ข้อมูลส่วนนี้ถือว่ามีความสำคัญและเป็นประโยชน์ไม่น้อยในเรื่องของการแสดงตัวอย่าง หรือสถานการณ์จริงที่เคยเกิดขึ้น ทำให้ระบบเข้าใจบริบทและให้ผลลัพธ์ได้อย่างมีประสิทธิภาพ

ตารางที่ 3.1 รายการข้อมูลดิบสำหรับสร้างชุดข้อมูล (Dataset)

สำนักงานคณะกรรมการคุ้มครองผู้บริโภค (สคบ.)		
ชื่อพระราชบัญญัติ (ปรับปรุงล่าสุด)	**ปี**	**จำนวนมาตรา**
พระราชบัญญัติยา	2510	129
พระราชบัญญัติคุ้มครองผู้บริโภค	2522	63
พระราชบัญญัติอาหาร	2522	78
พระราชบัญญัติวัตถุอันตราย	2535	93
พระราชบัญญัติคุ้มครองผู้ประสบภัยจากรถ	2535	47
พระราชบัญญัติขายตรงและตลาดแบบตรง	2545	56
พระราชบัญญัติวิธีพิจารณาคดีผู้บริโภค	2551	66
พระราชบัญญัติความรับผิดต่อความเสียหายที่เกิดขึ้นจากสินค้าที่ไม่ปลอดภัย	2551	16
พระราชบัญญัติเครื่องสำอาง	2558	94
พระราชบัญญัติการไกล่เกลี่ยข้อพิพาท	2562	96
พระราชบัญญัติคุ้มครองข้อมูลส่วนบุคคล	2562	72

สภาองค์กรของผู้บริโภค	
ประเภทเนื้อหา	**จำนวนเนื้อหา**
บทความ	20
ข่าว	12
กรณีศึกษา	15

ในการเก็บรวบรวมข้อมูลดังที่ได้แสดงในตารางที่ 3.1 ในส่วนเอกสารของสำนักงานคณะกรรมการคุ้มครองผู้บริโภค (สคบ.) ซึ่งเป็นพระราชบัญญัติ นั้นเป็นเอกสารประเภท PDF มีโครงสร้างที่ไม่เอื้อต่อการแยกส่วนข้อมูล ดังแสดงในรูปที่ 3.3

รูปที่ 3.3

มาตรา ๓ ในพระราชบัญญัตินี้

"ซื้อ" หมายความรวมถึง เช่า เช่าซื้อ หรือได้มาไม่ว่าด้วยประการใด ๆ โดยให้ค่าตอบแทนเป็นเงินหรือผลประโยชน์อย่างอื่น

"ขาย" หมายความรวมถึง ให้เช่า ให้เช่าซื้อ หรือจัดหาให้ไม่ว่าด้วยประการใด ๆ โดยเรียกค่าตอบแทนเป็นเงินหรือผลประโยชน์อย่างอื่น ตลอดจนการเสนอหรือการชักชวนเพื่อการดังกล่าวด้วย

"สินค้า" หมายความว่า สิ่งของที่ผลิตหรือมีไว้เพื่อขาย

"บริการ" หมายความว่า การรับจัดทำการงาน การให้สิทธิใด ๆ หรือการให้ใช้หรือให้ประโยชน์ในทรัพย์สินหรือกิจการใด ๆ โดยเรียกค่าตอบแทนเป็นเงินหรือผลประโยชน์อื่นแต่ไม่รวมถึงการจ้างแรงงานตามกฎหมายแรงงาน

"ผลิต" หมายความว่า ทำ ผสม ปรุง ประกอบ ประดิษฐ์ หรือแปรสภาพและหมายความรวมถึงการเปลี่ยนรูป การดัดแปลง การคัดเลือก หรือการแบ่งบรรจุ

รูปที่ 3.3 ต้นฉบับเอกสารพระราชบัญญัติ

เอกสาร PDF นี้จะมีลายน้ำ "สำนักงานคณะกรรมการกฤษฎีกา" ตลอดทั้งฉบับซึ่งไม่สามารถนำไปใช้กับระบบ SLP เพื่อสร้างชุดข้อมูลสำหรับใช้ในระบบได้ จึงต้องมีการทำความสะอาดข้อมูล โดยผู้วิจัยเขียนโค้ดด้วยภาษา Python ให้ดึงข้อความทั้งหมดในเอกสาร PDF ออกมาและลบคำว่า "สำนักงานคณะกรรมการกฤษฎีกา" ซึ่งเป็นลายน้ำออกทั้งหมด จึงเหลือเพียงข้อความที่เป็นเนื้อหาที่ต้องการเท่านั้น และบันทึกเป็นเอกสาร Microsoft Word ดังแสดงในรูปที่ 3.4

มำตรำ ๓ ในพระรำชบัญญัตินี้ "ซื้อ" หมำยควำมรวมถึง เช่ำ เช่ำซื้อ หรือ ได้มำไม่ว่ำด้วยประกำรใด ๆ โดยให้ ค่ำตอบแทนเป็นเงินหรือผลประโยชน์อย่ำงอื่น "ขำย" หมำยควำมรวมถึง ให้เช่ำ ให้เช่ำซื้อ หรือจัดหำให้ไม่ว่ำด้วยประกำรใด ๆ โดยเรียกค่ำตอบแทนเป็นเงินหรือผลประโยชน์อย่ำงอื่น ตลอดจนกำรเสนอหรือกำรชักชวนเพื่อกำร ดังกล่ำวด้วย "สินค้ำ" หมำยควำมว่ำ สิ่งของที่ผลิตหรือมีไว้เพื่อขำย "บริกำร" หมำยควำมว่ำ กำรรับจัดท ำกำรงำน กำรให้สิทธิใด ๆ หรือกำรให้ใช้หรือ ให้ประโยชน์ในทรัพย์สินหรือกิจกำรใด ๆ โดยเรียกค่ำตอบแทนเป็นเงินหรือผลประโยชน์อื่นแต่ไม่ รวมถึงกำรจ้ำงแรงงำนตำมกฎหมำยแรงงำน "ผลิต" หมำยควำมว่ำ ท ำ ผสม ปรุง ประกอบ ประดิษฐ์ หรือแปรสภำพและ หมำยควำมรวมถึงกำรเปลี่ยนรูป กำรดัดแปลง กำรคัดเลือก หรือกำรแบ่งบรรจ

"ผู้บริโภค" ๒ หมำยควำมว่ำ ผู้ซื้อหรือผู้ได้รับบริกำรจำกผู้ประกอบธุรกิจหรือผู้ซึ่ง ได้รับกำรเสนอหรือกำรชักชวนจำกผู้ประกอบธุรกิจเพื่อให้ซื้อสินค้ำหรือรับบริกำรและหมำยควำม รวมถึงผู้ใช้สินค้ำหรือผู้ได้รับบริกำรจำกผู้ประกอบธุรกิจโดยชอบแม้มิได้เป็นผู้เสียค่ำตอบแทนก็ดำม "ผู้ประกอบธุรกิจ" หมำยควำมว่ำ ผู้ขำย ผู้ผลิตเพื่อขำย ผู้สั่งหรือน ำเข้ำมำใน รำชอำณำจักรเพื่อขำยหรือผู้ซื้อเพื่อขำยต่อซึ่งสินค้ำ หรือผู้ให้บริกำร และหมำยควำมรวมถึงผู้ ประกอบกิจกำร โฆษณำด้วย

รูปที่ 3.4 เอกสาร PDF หลังลบลายน้ำออกและบันทึกเป็นเอกสาร Microsoft Word

แม้จะได้ตัดเอาลายน้ำออกแล้ว จะเห็นว่าสระในภาษาไทยนั้นไม่ถูกต้อง เนื่องจากปัญหารูปแบบ “การเข้ารหัสฟอนต์” (Font Encoding) ที่ไม่ตรงกันระหว่างต้นทาง (PDF) และปลายทาง (Word) อย่างไรก็ตามปัญหานี้จะถูกแก้ไขเมื่อเข้าสู่กระบวนการ SLP

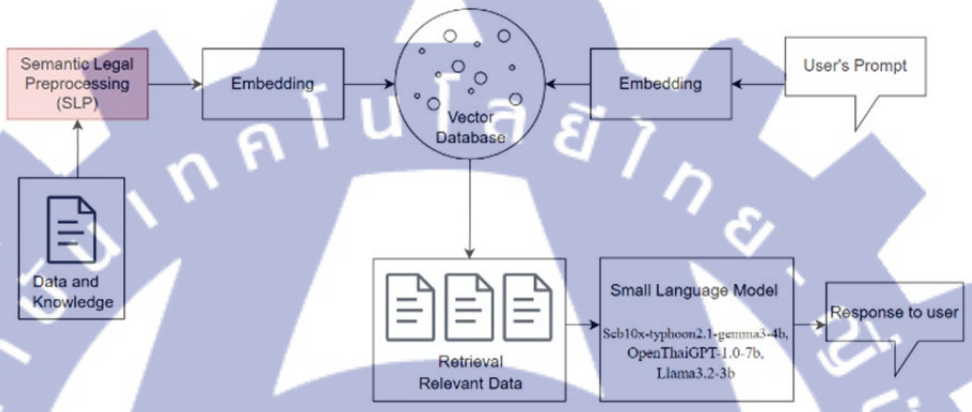
สำหรับข้อมูลที่ได้จาก สภาองค์กรของผู้บริโภค ซึ่งจะเป็นบทความ ข่าวสาร และกรณีที่เกิดขึ้นจริงซึ่งเกี่ยวข้องกับการคุ้มครองผู้บริโภค เป็นข้อมูลซึ่งอยู่ในหน้าเว็บเพจของ สภาองค์กรของผู้บริโภค ที่เผยแพร่เป็นสาธารณะ ดังแสดงในรูปที่ 3.5 ผู้วิจัยได้เขียนโค้ดด้วยภาษา Python ให้ดึงข้อความจากหน้าเว็บเพจ แล้วจึงบันทึกเป็นเอกสาร Microsoft Word

รูปที่ 3.5 หน้าเว็บเพจของ สภาองค์กรของผู้บริโภค [15]

หลังจากแหล่งข้อมูลและบันทึกลงเอกสาร Microsoft Word แล้ว ขั้นตอนต่อไปจะเป็นการนำข้อมูลดิบนี้ส่งเข้าไปในระบบ Semantic Legal Preprocessing (SLP) เพื่อประมวลผลข้อมูลล่วงหน้าสำหรับระบบ RAG ต่อไป

3.2 Semantic Legal Preprocessing (SLP)

เป็นกระบวนการจัดการข้อมูลให้มีประสิทธิภาพเพื่อส่งเข้าสู่ขั้นตอนการแปลง ขั้นตอนนี้ถือเป็นหัวใจสำคัญของงานวิจัยนี้ เนื่องจาก ระบบ RAG ดั้งเดิมใช้วิธีการจัดการข้อมูลโดยการแบ่งข้อมูลแบบ Chunk of Text ซึ่งแต่ละส่วนจะมีขนาดความยาวของตัวอักษรคงที่ โดยการด้วยวิธี Chuck of Text จะตัดแบ่งข้อความโดยไม่สนใจว่าข้อความนั้นมีบริบทยาวเพียงใด จึงเป็นไปได้ยากที่จะได้ข้อความที่ครบถ้วนและมีบริบท งานวิจัยนี้จึงเสนอ ระบบ Semantic Legal Preprocessing (SLP) เพื่อแก้ปัญหาดังกล่าว ดังแสดงในรูปที่ 3.6



รูปที่ 3.6 Node SLP ในระบบ RAG

จากไดอะแกรม การทำงานจะเริ่มจากการนำข้อมูลดิบ (Raw Data) ทางกฎหมายที่เตรียมไว้เป็นเอกสาร Microsoft Word สำหรับพระราชบัญญัติ จะนำมาแบ่งเป็นเรื่องๆ โดยผู้วิจัยเอง และในส่วนของบทความ ได้ถูกบันทึกเป็นเอกสาร Microsoft Word เป็นเรื่องๆ อยู่แล้ว เมื่อแบ่งได้แล้วจึงนำแต่ละเรื่องมาจัดรูปแบบที่เรียกว่า Few-Shot Learning เทคนิค Prompt Engineering แบบหนึ่ง ซึ่งเป็นการให้ตัวอย่างผลลัพธ์ที่ต้องการกับ ChatGPT 5 ซึ่งเป็น Large Language Model (LLM) หรือโมเดลภาษาที่ใช้สร้างข้อความ เพื่อเรียนรู้รูปแบบและขอบเขต เพื่อสร้างผลลัพธ์ตามต้องการ โดยผู้เขียนให้บทบาทกับ ChatGPT 5 ตัวอย่างเช่น

“คุณคือผู้เชี่ยวชาญด้านกฎหมายที่เกี่ยวข้องกับการคุ้มครองผู้บริโภคของประเทศไทย”

เหตุที่ต้องให้บทบาทกับ LLM เนื่องจากจะทำให้ LLM เข้าในบริบทในภาพรวมว่าต้องสร้างผลลัพธ์ในเชิงของนักกฎหมายที่มีความรู้ในกฎหมายเกี่ยวกับการคุ้มครองผู้บริโภค เพื่อไม่ให้ LLM สร้างข้อความที่นอกเหนือจากบทบาทที่ให้ เป็นการลดอาการหลอน (Hallucination)

“จงสร้าง คำถาม, คำตอบ, คำอธิบาย, ตัวอย่าง ด้วยภาษาที่เข้าใจง่าย
จำนวน 10 รายการ จากข้อความที่ให้ไป โดยให้มีรูปแบบดังนี้
Q = คำถาม
A = คำตอบ
E = อธิบาย
S = ตัวอย่าง
ให้ถือว่าเป็น 1 ชุด และเว้นบรรทัด”

ใน Prompt ส่วนนี้เป็นการบอก LLM ว่าต้องการให้สร้างข้อความแบบใดออกมา ในรูปแบบใด รวมถึงกำหนดขอบเขต รูปแบบการแสดงผล จำนวนข้อมูลที่ต้องการ

“ตัวอย่างดังนี้
Q : ผู้บริโภคมีสิทธิพื้นฐานอะไรบ้าง?
A : ได้รับข่าวสารที่ชัดเจน เลือกซื้ออย่างเสรี ได้รับความปลอดภัย ร้องเรียนได้ และได้รับความเป็นธรรม
E : เป็นสิทธิพื้นฐานที่กฎหมายคุ้มครองผู้บริโภคกำหนดไว้
S : ซื้อของแล้วแต่ได้ของไม่ครบ ถือว่าละเมิดสิทธิ”

ใน Prompt ส่วนนี้เป็นการบอก LLM ว่าตัวอย่างผลลัพธ์ที่ต้องการเป็นอย่างไร สิ่งสำคัญของเทคนิค Prompt Engineering แบบ Few-Shot Learning คือการใช้ตัวอย่างเพื่อให้ LLM เรียนรู้ เพื่อให้สามารถสร้างผลลัพธ์ตามตัวอย่างได้

“หลังจากได้ข้อมูลเรียบร้อยแล้ว ให้นำข้อความ A, E มาต่อกัน ตามด้วยคำว่า “ตัวอย่างเช่น” และต่อด้วยข้อความ S เป็นแถวเดียวกัน และครอบข้อความทั้งหมดด้วยเครื่องหมายคำพูด ดังรูปแบบนี้

“ได้รับข่าวสารที่ชัดเจน เลือกซื้ออย่างเสรี ได้รับความปลอดภัย ร้องเรียนได้ และได้รับความเป็นธรรม เป็นสิทธิพื้นฐานที่กฎหมายคุ้มครองผู้บริโภคกำหนดไว้ ตัวอย่างเช่น ซื้อของแล้วไม่ได้ข้อมูลครบ ถือว่าละเมิดสิทธิ”

แล้วนำทั้งหมดมาเรียงต่อกันโดยแต่ละชุดคั่นด้วยเครื่องหมาย “,”

ใน Prompt ส่วนนี้ คือการสั่ง LLM ให้จัดรูปแบบข้อความ (Format) ตามที่ผู้วิจัยต้องการ สำหรับงานวิจัยนี้ ต้องให้มีข้อมูลที่เป็น คำถาม Q, คำตอบ A, คำอธิบาย E, ตัวอย่าง S แยกกันก่อน ดังตัวอย่าง

“Q : ถ้าผู้ขายไม่คืนเงินตามกำหนดจะเกิดอะไรขึ้น?
A : ต้องจ่ายเบี้ยปรับเพิ่มเติม
E : เป็นบทลงโทษเพื่อให้ผู้ขายปฏิบัติตามกฎหมาย
S : ร้านไม่คืนเงินตามเวลา โดนปรับเพิ่มให้ผู้บริโภค”

จากนั้นจึงนำข้อความ A , E มาต่อกัน ตามด้วยคำว่า “ตัวอย่างเช่น” และต่อด้วยข้อความ S เป็นแถวเดียวกัน และครอบข้อความทั้งหมดด้วยเครื่องหมายคำพูด จะได้ผลลัพธ์ดังนี้

“ต้องจ่ายเบี้ยปรับเพิ่มเติม เป็นบทลงโทษเพื่อให้ผู้ขายปฏิบัติตามกฎหมาย ตัวอย่างเช่น ร้านไม่คืนเงินตามเวลา โดนปรับเพิ่มให้ผู้บริโภค”

จากนั้นสั่งให้นำข้อความทั้งหมดทุกชุดมาเรียงต่อกันโดยแต่ละชุดคั่นด้วยเครื่องหมายจุลภาค (,) ทำให้ได้ผลลัพธ์ดังตัวอย่าง

“ต้องมีชื่อผู้ซื้อ-ผู้ขาย วันที่ซื้อขาย วันที่ส่งมอบ และสิทธิเลิกสัญญา เพื่อให้ผู้บริโภคได้รับข้อมูลครบถ้วนและเข้าใจง่ายตามกฎหมาย ตัวอย่างเช่น ซื้อของจากขายตรงแล้วได้ใบเสร็จที่มีข้อมูลครบตามที่กำหนด”

“ผู้บริโภคมีสิทธิเลิกสัญญาภายใน 7 วันหลังได้รับสินค้า เป็นการคุ้มครองผู้บริโภคจากการตัดสินใจที่อาจถูกเร่งรัด ตัวอย่างเช่น สั่งของจากขายตรงแล้วเปลี่ยนใจ สามารถส่งหนังสือขอยกเลิกได้ใน 7 วัน”

“การซื้อขายจะไม่ผูกพันผู้บริโภค เพื่อป้องกันไม่ให้ผู้บริโภคเสียเปรียบจากการซื้อขายที่ไม่ชอบด้วยกฎหมาย ตัวอย่างเช่น ได้รับของแต่ไม่มีเอกสารตามที่กฎหมายกำหนด ถือว่าไม่ต้องรับผิดชอบต่อสัญญานั้น”

“ต้องคืนสินค้าให้ผู้ขาย หรือเก็บรักษาไว้ 21 วัน เพื่อให้ผู้ขายมีโอกาสมารับสินค้าคืนอย่างเป็นธรรม ตัวอย่างเช่น ใช้สิทธิเลิกสัญญาแล้ววางของไว้รอให้ตัวแทนมารับภายใน 21 วัน”

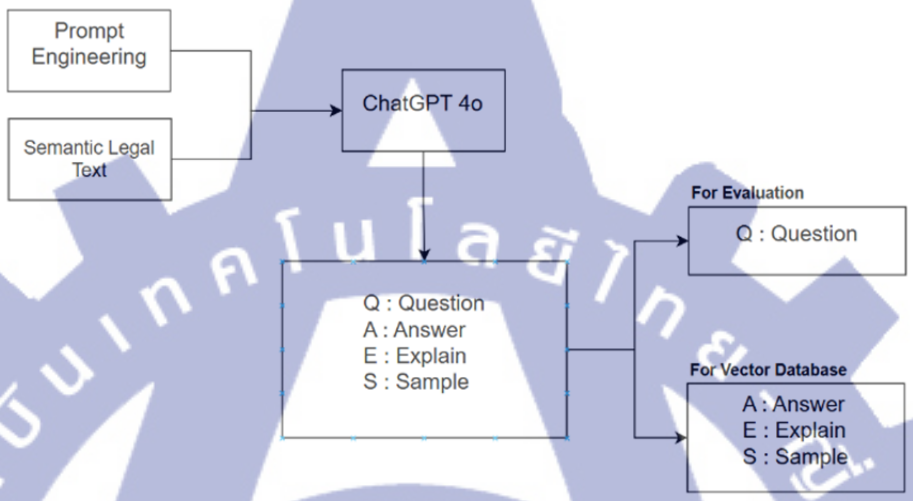
“ผู้บริโภคต้องชดใช้ค่าเสียหาย ยกเว้นความเสียหายจากการใช้งานตามปกติ เพื่อความเป็นธรรมทั้งสองฝ่ายในกรณีที่สินค้ามีปัญหา ตัวอย่างเช่น สินค้าแตกเพราะทำตกเอง ต้องจ่ายค่าเสียหาย”

“ยึดไว้ได้จนกว่าจะได้รับเงินคืน เป็นการคุ้มครองสิทธิของผู้บริโภคไม่ให้เสียเปรียบ ตัวอย่างเช่น ขอคืนของแล้วเงินยังไม่มา ผู้บริโภคเก็บสินค้าไว้ก่อนได้”

“ต้องคืนเงินเต็มจำนวนภายใน 15 วัน เพื่อไม่ให้ผู้บริโภครอเงินนานเกินไป ตัวอย่างเช่น แจ้งยกเลิกภายใน 7 วัน แล้วได้เงินคืนภายในอีก 15 วัน”

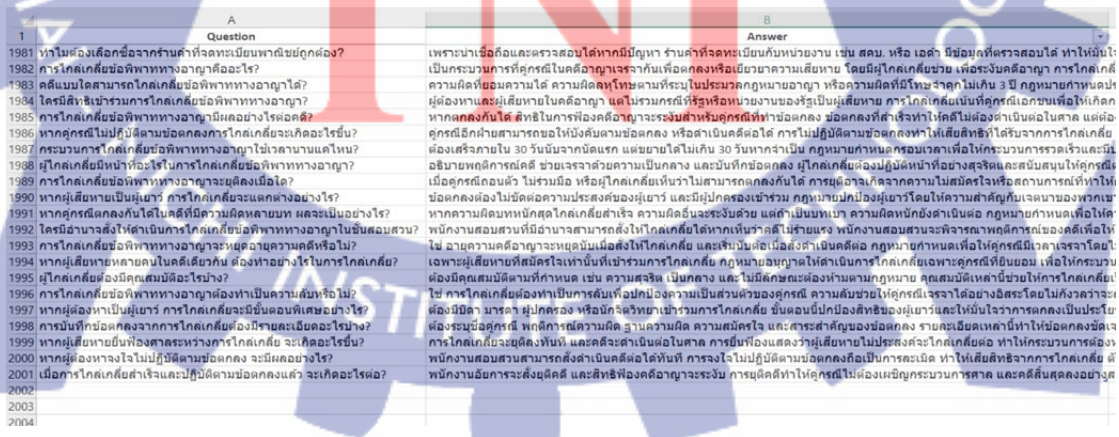
“ต้องจ่ายเบี้ยปรับเพิ่มเติม เป็นบทลงโทษเพื่อให้ผู้ขายปฏิบัติตามกฎหมาย ตัวอย่างเช่น ร้านไม่คืนเงินตามเวลา โดนปรับเพิ่มให้ผู้บริโภค”

โดยกระบวนการ Semantic Legal Preprocessing (SLP) ได้แสดงไว้ในรูปที่ 3.7



รูปที่ 3.7 การทำงานภายใน SLP

ชุดข้อความทั้งหมดจะถูกบันทึกเป็นไฟล์รูปแบบตาราง (Tabular Data) โดยแบ่งเป็น 2 คอลัมน์ ได้แก่ คำถาม Q และ ข้อความรวมของ คำตอบ A, คำอธิบาย E, และ ตัวอย่าง S ซึ่งถูกรวมเป็นข้อความเดียวกัน ดังแสดงในรูปที่ 3.8



	A	B
1	Question	Answer
1981	ทำไมต้องเลือกชื่อจากร้านค้าที่จดทะเบียนพาณิชย์ถูกต้อง?	เพราะนำเชื่อถือและตรวจสอบได้หากมีปัญหา ร้านค้าที่จดทะเบียนกับหน่วยงาน เช่น สคบ. หรือ เอค่า มีข้อมูลที่ตรวจสอบได้ ทำให้มั่นใ
1982	การโกลเกลี่ยข้อพิพาททางอาญาคืออะไร?	เป็นกระบวนการที่คู่กรณีในคดีอาญาเจรจากันเพื่อตกลงหรือเยียวยาความเสียหาย โดยมีผู้ไกล่เกลี่ยช่วย เพื่อระงับคดีอาญา การไกล่เกลี่
1983	คดีแบบใดสามารถไกล่เกลี่ยข้อพิพาททางอาญาได้?	ความผิดที่ยอมความได้ ความผิดลหุโทษตามที่ระบุในประมวลกฎหมายอาญา หรือความผิดที่มีโทษจำคุกไม่เกิน 3 ปี กฎหมายกำหนดป
1984	ใครมีสิทธิเข้าร่วมการไกล่เกลี่ยข้อพิพาททางอาญา?	ผู้ต้องหาและผู้เสียหายในคดีอาญา แต่ไม่รวมกรณีที่รัฐหรือหน่วยงานของรัฐเป็นผู้เสียหาย การไกล่เกลี่ยเน้นที่คู่กรณีเองเพื่อให้เกิดก
1985	การไกล่เกลี่ยข้อพิพาททางอาญามีผลอย่างไรต่อคดี?	หากตกลงกันได้ สิทธิในการฟ้องคดีอาญาจะระงับสำหรับกรณีที่ทำข้อตกลง ข้อตกลงที่สำเร็จทำให้คดีไม่ต้องดำเนินต่อในศาล แต่ต่อ
1986	หากคู่กรณีไม่ปฏิบัติตามข้อตกลงการไกล่เกลี่ยจะเกิดอะไรขึ้น?	คู่กรณีอีกฝ่ายสามารถขอให้บังคับตามข้อตกลง หรือดำเนินคดีต่อได้ การไม่ปฏิบัติตามข้อตกลงทำให้เสียสิทธิที่ได้รับจากการไกล่เกลี่ย
1987	กระบวนการไกล่เกลี่ยข้อพิพาททางอาญาใช้เวลานานแค่ไหน?	ต้องเสร็จภายใน 30 วันนับจากนัดแรก แต่ขยายได้ไม่เกิน 30 วันหากจำเป็น กฎหมายกำหนดกรอบเวลาเพื่อให้กระบวนการรวดเร็วและมีป
1988	ผู้ไกล่เกลี่ยมีหน้าที่อะไรในการไกล่เกลี่ยข้อพิพาททางอาญา?	อธิบายพฤติการณ์คดี ช่วยเจรจาด้วยความเป็นกลาง และบันทึกข้อตกลง ผู้ไกล่เกลี่ยต้องปฏิบัติหน้าที่อย่างสุจริตและสนับสนุนให้คู่กรณี
1989	การไกล่เกลี่ยข้อพิพาททางอาญาจะยุติลงเมื่อใด?	เมื่อคู่กรณีถอนตัว ไม่ร่วมมือ หรือผู้ไกล่เกลี่ยเห็นว่าไม่สามารถตกลงกันได้ การยุติอาจเกิดจากความไม่สมัครใจหรือสถานการณ์ที่ทำให้
1990	หากผู้เสียหายเป็นผู้เยาว์ การไกล่เกลี่ยจะแตกต่างอย่างไร?	ข้อตกลงต้องไม่ขัดต่อความประสงค์ของผู้เยาว์ และมีผู้ปกครองเข้าร่วม กฎหมายคุ้มครองเข้าร่วม กฎหมายคุ้มครองผู้เยาว์โดยให้ความสำคัญกับเจตนาของพวกเข
1991	หากคู่กรณีตกลงกันได้ในคดีที่มีความผิดหลายบท ผลจะเป็นอย่างไร?	หากความผิดบทหนักสุดไกล่เกลี่ยสำเร็จ ความผิดอื่นจะระงับด้วย แต่ถ้าเป็นบทเบา ความผิดหนักยังดำเนินต่อ กฎหมายกำหนดเพื่อให้ค
1992	ใครมีอำนาจสั่งให้ดำเนินการไกล่เกลี่ยข้อพิพาททางอาญาในชั้นสอบสวน?	พนักงานสอบสวนที่มีอำนาจสามารถสั่งให้ไกล่เกลี่ยได้หากเห็นว่าคดีไม่ร้ายแรง พนักงานสอบสวนจะพิจารณาพฤติการณ์ของคดีเพื่อให้
1993	การไกล่เกลี่ยข้อพิพาททางอาญาจะหยุดอายุความคดีหรือไม่?	ไม่ อายุความคดีอาญาจะหยุดนับเมื่อสั่งให้ไกล่เกลี่ย และเริ่มนับต่อเมื่อสั่งดำเนินคดีต่อ กฎหมายกำหนดเพื่อให้คู่กรณีมีเวลาเจรจาโดยไ
1994	หากผู้เสียหายหลายคนในคดีเดียวกัน ต้องทำอย่างไรในการไกล่เกลี่ย?	เฉพาะผู้เสียหายที่สมัครใจเท่านั้นที่เข้าร่วมการไกล่เกลี่ย กฎหมายอนุญาตให้ดำเนินการไกล่เกลี่ยเฉพาะคู่กรณีที่ยินยอม เพื่อให้กระบวน
1995	ผู้ไกล่เกลี่ยต้องมีคุณสมบัติอะไรบ้าง?	ต้องมีคุณสมบัติตามที่กำหนด เช่น ความสุจริต เป็นกลาง และไม่มีลักษณะต้องห้ามตามกฎหมาย คุณสมบัติเหล่านี้ช่วยให้การไกล่เกลี่ยเ
1996	การไกล่เกลี่ยข้อพิพาททางอาญาต้องทำเป็นความลับหรือไม่?	ใช่ การไกล่เกลี่ยต้องทำเป็นการลับเพื่อปกป้องความเป็นส่วนตัวของคู่กรณี ความลับช่วยให้คู่กรณีเจรจาได้อย่างอิสระโดยไม่กังวลว่าจะ
1997	หากผู้ต้องหาเป็นผู้เยาว์ การไกล่เกลี่ยจะมีขั้นตอนพิเศษอย่างไร?	ต้องมีบิดา มารดา ผู้ปกครอง หรือนักจิตวิทยาเข้าร่วมการไกล่เกลี่ย ขั้นตอนนี้ปกป้องสิทธิของผู้เยาว์และให้มั่นใจว่าการตกลงเป็นประโย
1998	การนัดที่ตกลงจากการไกล่เกลี่ยต้องมีรายละเอียดอะไรบ้าง?	ต้องระบุข้อคู่กรณี พฤติการณ์ความผิด ฐานความผิด ความสมัครใจ และสาระสำคัญของข้อตกลง รายละเอียดเหล่านี้ทำให้ข้อตกลงชัดเจ
1999	หากผู้เสียหายยืนฟ้องศาลระหว่างการไกล่เกลี่ย จะเกิดอะไรขึ้น?	การไกล่เกลี่ยจะยุติลงทันที และคดีจะดำเนินต่อในศาล การยื่นฟ้องแสดงว่าผู้เสียหายไม่ประสงค์จะไกล่เกลี่ยต่อ ทำให้กระบวนการต้องง
2000	หากผู้ต้องหาจงใจไม่ปฏิบัติตามข้อตกลง จะมีผลอย่างไร?	พนักงานสอบสวนสามารถสั่งดำเนินคดีต่อได้ทันที การจงใจไม่ปฏิบัติตามข้อตกลงถือเป็นการละเมิด ทำให้เสียสิทธิจากการไกล่เกลี่ย ดั
2001	เมื่อการไกล่เกลี่ยสำเร็จและปฏิบัติตามข้อตกลงแล้ว จะเกิดอะไรต่อ?	พนักงานอัยการจะสั่งยุติคดี และสิทธิฟ้องคดีอาญาจะระงับ การยุติคดีทำให้คู่กรณีไม่ต้องเผชิญกระบวนการศาล และคดีสิ้นสุดลงอย่างส
2002		
2003		
2004		

รูปที่ 3.8 ไฟล์ชุดข้อมูลในรูปแบบตาราง (Tabular Data)

3.3 Embedding

เมื่อได้ข้อมูลและรูปแบบตามที่ต้องการแล้ว ผู้วิจัยได้เขียนโค้ดด้วยภาษา Python ให้ดึงข้อมูลที่อยู่ในรูปแบบตาราง (Tabular Data) ไปแปลงเป็นเวกเตอร์ตัวเลขด้วย Embedding Model และบันทึกไปยังฐานข้อมูลเวกเตอร์ (Vector Database)

กระบวนการแปลงข้อมูลที่อยู่ในรูปข้อความให้กลายเป็นตัวเลขเวกเตอร์ที่มีความหมาย เพื่อให้คอมพิวเตอร์สามารถเข้าใจและประมวลผลได้ โดยใช้ Embedding Model โดยในงานวิจัยนี้ ได้เลือก Model จำนวน 6 ตัว ในการทดสอบเพื่อเปรียบเทียบประสิทธิภาพ ดังที่แสดงในตารางที่ 3.2

ตารางที่ 3.2 รายการ Embedding Model

Embedding Model	ความยาวสูงสุดของข้อความที่รองรับ
BGE-M3 [16]	8,192 Tokens
WangchanX-Legal-ThaiCCL-Retriever [17]	8,192 Tokens
Snowflake-Arctic-Embed-m-v1.5 [18]	512 Tokens
Multilingual-e5-Base [19]	512 Tokens
All-MiniLM-L6-v2 [20]	256 Tokens
Static-Retrieval-mrl-en-v1 [21]	Inf Tokens

รายละเอียดของแต่ละโมเดลมีดังต่อไปนี้

• BGE-M3 (BAAI General Embedding - M3): เป็น Embedding Model พัฒนามาจาก XLM-RoBERTa ซึ่งเป็น Based Model และรองรับการใช้งานมากกว่า 100 ภาษาทั่วโลก ซึ่งรวมถึงภาษาไทยด้วย รองรับความยาว Sequence หรือความยาวของข้อความ สูงสุด 8192 tokens

• WangchanX-Legal-ThaiCCL-Retriever: เป็น Embedding Model ที่นำโมเดล BGE-M3 มาฝึกกับชุดข้อมูลที่ครอบคลุมประมวลกฎหมายแพ่งและพาณิชย์ และกฎหมายธุรกิจที่เกี่ยวข้อง นอกจากนี้ยังสืบทอดคุณสมบัติจาก BGE-M3 โดยรองรับความยาว Sequence หรือความยาวของข้อความสูงสุด 8192 tokens เช่นกัน เหมาะสำหรับการประมวลผลข้อมูลเกี่ยวกับกฎหมายไทยซึ่งมีลักษณะเป็นข้อความยาว

• Snowflake Arctic Embed M v1.5: ที่เป็น Embedding Model ที่พัฒนามาจาก BERT-Based Model โดยรองรับความยาว Sequence หรือความยาวของข้อความ สูงสุด 512 Tokens มีจุดเด่นที่ขนาดที่เล็กลงแต่ยังคงความแม่นยำได้ เหมาะสำหรับงานที่มีข้อจำกัดด้านพื้นที่จัดเก็บข้อมูล

• Multilingual-E5-Base: เป็น Embedding Model ที่พัฒนามาจาก XLM-RoBERTa-base ถูกออกแบบมาและรองรับการใช้งานมากกว่า 100 ภาษาทั่วโลก ซึ่งรวมถึงภาษาไทยด้วย โดยรองรับความยาว Sequence หรือความยาวของข้อความ สูงสุด 512 tokens

• all-MiniLM-L6-v2: เป็น Embedding Model ที่พัฒนามาจาก Microsoft MiniLM ซึ่งเป็น Based Model ที่ปรับจูนให้จับความหมายเชิงลึก (Semantic) ได้ดี เหมาะสำหรับงาน Clustering และ Semantic Search

• Static-Retrieval-mrl-en-v1: เป็น Embedding Model แบบ Bag-of-Embeddings ซึ่งทำให้มีความเร็วในการประมวลผลที่สูง และใช้ทรัพยากรในการประมวลผลน้อย และเนื่องจากโมเดลนี้ไม่ได้ใช้สถาปัตยกรรม Transformer แบบปกติ ที่มี Self-Attention Mechanism ซึ่งกินทรัพยากรมากตามความยาวข้อความ แต่ใช้สถาปัตยกรรมแบบ Static Embedding (Bag-of-Embeddings)

ในงานวิจัยนี้กระบวนการแปลงข้อความเป็นเวกเตอร์ตัวเลขโดยใช้ Embedding Model นั้นจะถูกใช้ทั้งในส่วนของการแปลงชุดข้อมูลเพื่อเก็บเข้าในฐานข้อมูลเวกเตอร์ และยังใช้ในส่วนของการรับคำถามจากผู้ใช้อีกด้วย เนื่องจาก ต้องแปลงข้อความทั้ง 2 ส่วน เป็น เวกเตอร์ตัวเลขเพื่อนำมาคำนวณหาระยะห่างเพื่อดึงข้อความ ดังนั้นในส่วนของการประเมินประสิทธิภาพของ Embedding Model จึงถือว่าเป็นขั้นตอนที่มีความสำคัญอย่างยิ่งในระบบ RAG

3.4 ฐานข้อมูล เวกเตอร์ (Vector Database)

หลังจากกระบวนการ Embedding ที่แปลงข้อความเป็นเวกเตอร์ตัวเลขแล้ว จึงได้จัดเก็บข้อมูลไปยังฐานข้อมูลเวกเตอร์ (Vector Database) ผู้วิจัยได้เขียนโค้ดด้วยภาษา Python ในการบันทึกข้อมูลไปยังฐานข้อมูล ในงานวิจัยนี้ ใช้บริการฐานข้อมูลบนระบบคลาวด์ TiDB ในการจัดเก็บข้อมูลแบบคลาวด์ (Cloud) ที่มีข้อดีในการลดความเสี่ยงด้านอุปกรณ์จัดเก็บข้อมูลในสถานที่ เนื่องจากการจัดเก็บข้อมูลแบบคลาวด์นี้อยู่ที่สถานที่ของผู้ให้บริการด้านฐานข้อมูลโดยเฉพาะซึ่งมีความเสถียรและปลอดภัยกว่า และยังมีระบบสำรองข้อมูลที่ปลอดภัย ซึ่งฐานข้อมูลดังกล่าวประกอบด้วย 2 คอลัมน์ ดังนี้

1) คอลัมน์ Document สำหรับจัดเก็บข้อความต้นฉบับที่ได้จากคอลัมน์ Answer + Explain + Sample

2) คอลัมน์ Embedding สำหรับจัดเก็บเวกเตอร์ตัวเลขที่ได้จากขั้นตอน Embedding

การจัดเก็บข้อมูลในลักษณะนี้ช่วยให้สามารถค้นหาข้อความที่ใกล้เคียงกันได้อย่างรวดเร็วและแม่นยำ โดยอาศัยการคำนวณระยะห่างของเวกเตอร์ที่จัดเก็บไว้ในฐานข้อมูล ดังที่แสดงเป็นตัวอย่างข้อมูลบางส่วนในตารางที่ 3.3

ตารางที่ 3.3 ตัวอย่างข้อมูลใน Vector Database

Document	Embedding
เพราะมีผู้เสียหายจากการซื้อของออนไลน์ จำนวนมาก จากสถิติพบ...	[-0.0005730653,0.025092807, -0.02163328,0.013231898, -0.02...
คดีหลอกลวงซื้อขายสินค้าออนไลน์ คดีนี้คิดเป็นสัดส่วนมากถึง 42...	[-0.03236147,0.03022134, -0.03018961, -0.004339998, -0.02742...
ประมาณ 4,000 ล้านบาท แม้ไม่สูงสุด แต่มีจำนวนคดีมากที่สุด สะ...	[0.00082940236,0.04540198, -0.034941282, -0.0084000295, -...

3.5 Small Language Model (SLM)

เมื่อผู้ใช้ถามคำถาม ระบบจะดึงข้อมูลจากฐานข้อมูลซึ่งเราได้จัดเตรียมไว้ อย่างไรก็ตาม ข้อมูลที่ถูกดึงมายังไม่อาจนำไปเป็นคำตอบได้เนื่องจากเป็นข้อมูลดิบที่ลักษณะของภาษายังไม่เป็นธรรมชาติ ดังนั้นจึงต้องนำข้อมูลที่ดึงมาพร้อมคำถามของผู้ใช้เข้าสู่กระบวนการแปลงให้เป็นภาษาธรรมชาติที่ผู้ใช้อ่านและเข้าใจได้ง่าย ซึ่งถือเป็นวัตถุประสงค์ของงานวิจัยนี้ ซึ่งในกระบวนการแปลงนี้ต้องอาศัยโมเดลภาษาในการแปลงจากข้อมูลดิบให้เป็นภาษาธรรมชาติ โดยงานวิจัยนี้เลือกใช้ Small Language Model (SLM) ซึ่งเป็นโมเดลภาษาพื้นฐาน Based Model ขนาดเล็ก ใช้ในขั้นตอนการ Generation เนื่องจากสามารถประมวลผลได้โดยใช้ทรัพยากรน้อย เหมาะกับการใช้งานจริง เช่น ในคอมพิวเตอร์บุคคล โดยในงานวิจัยนี้ได้เลือกโมเดล 3 ตัว สำหรับการสร้างข้อความดังตารางที่ 3.4

ตารางที่ 3.4 รายการ Small Language Model (SLM)

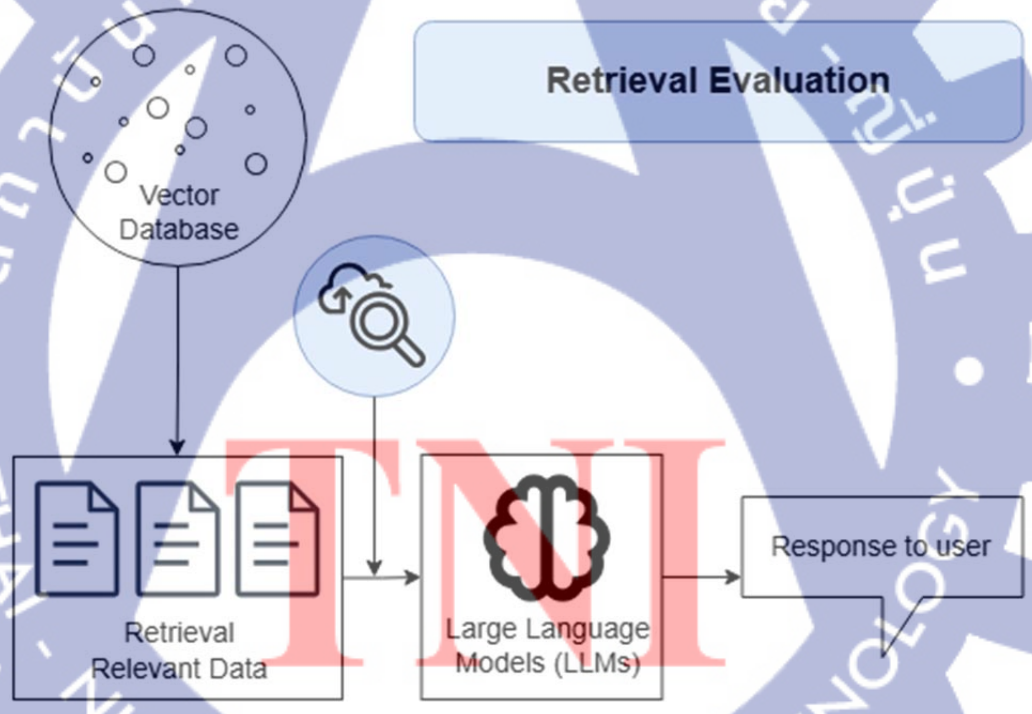
โมเดลภาษาขนาดเล็ก	**ขนาดของโมเดล**
Open thaiGPT-1.5 [22]	7B
Typhoon2.1-gemma3 [23]	4B
Llama3.2 [24]	3B

โมเดลภาษาขนาดเล็กโดยส่วนใหญ่มีขนาดไม่เกิน 10B หรือ หนึ่งหมื่นล้านพารามิเตอร์ ซึ่งพารามิเตอร์นี้คือตัวเลขที่เป็นค่าน้ำหนักของโมเดลที่ได้จากการฝึกฝน จำนวนพารามิเตอร์ที่สูงสะท้อนถึงขีดความสามารถของโมเดลในการเรียนรู้และทำความเข้าใจโครงสร้างทางภาษาที่ซับซ้อนได้อย่างมีประสิทธิภาพ อย่างไรก็ตามในการคำนวณจะใช้ทรัพยากรมากและใช้พลังงานมหาศาลในการทำงาน สำหรับงานวิจัยนี้เน้นให้คำแนะนำในกฎหมายคุ้มครองผู้บริโภคซึ่งเป็นข้อมูลที่จำเพาะเจาะจง จึงเหมาะสม

อย่างยิ่งในการนำโมเดลภาษาขนาดเล็กมาใช้ เนื่องจากเป็นโมเดลที่ถูกฝึกด้วยข้อมูลพื้นฐานขนาดเล็ก ทำให้ตัวโมเดลมีความสามารถให้การเข้าใจเรื่องทั่วไปในภาษาธรรมชาติ ซึ่งถือว่าเพียงพอแล้วสำหรับการใช้งานกับระบบ RAG ในการเสริมความรู้เฉพาะทางให้แก่โมเดลภาษาขนาดเล็กทำให้มีความสามารถเฉพาะทางได้

3.6 การประเมินผล (Evaluation)

ในการวิจัยนี้ ได้ประเมินประสิทธิภาพของระบบโดยใช้เมตริกซึ่งแบ่งออกเป็น 2 ส่วน ได้แก่ Retrieval Evaluation ซึ่งจะประเมินความสามารถของระบบการดึงข้อมูล ดังแสดงในรูปที่ 3.9 ด้วยวิธีการวัดความใกล้เคียงกันของเวกเตอร์ตัวเลขที่อยู่ในฐานข้อมูลเวกเตอร์ ซึ่งใช้เมตริก 3 ตัว ได้แก่ Mean Reciprocal Rank (MRR), Precision@k และ Hit-Rate@k โดยใช้ชุดคำถามแบบสุ่มจำนวน 400 ข้อ ซึ่งคิดเป็น 25% ของชุดข้อมูลคำถาม-คำตอบด้านกฎหมายทั้งหมด 2,000 แถว



รูปที่ 3.9 Retrieval Evaluation

MRR (Mean Reciprocal Rank) ใช้เพื่อวัดความสามารถของระบบในการจัดอันดับคำตอบที่ถูกต้องให้อยู่ในตำแหน่งที่สูงที่สุด โดยจะให้ความสำคัญกับอันดับของคำตอบที่ถูกต้องอันดับแรกที่ระบบค้นเจอ โดย MRR จะให้คะแนนสูงเมื่อคำตอบที่เกี่ยวข้องอยู่ใกล้ลำดับต้นๆ แสดงให้เห็นว่าระบบมีประสิทธิภาพ

ในการจัดอันดับที่แม่นยำ ซึ่งเป็นประโยชน์อย่างยิ่งสำหรับผู้ใช้ที่ต้องการคำตอบที่ถูกต้องอย่างรวดเร็ว โดยกระบวนการทำงานเริ่มจากเตรียมคำถาม 1 คำถามซึ่งคำถามนี้มาจากชุดข้อมูลที่เราใช้อยู่ที่เราเตรียมไว้จากขั้นตอน SLP และระบบ RAG ไม่เคยเห็นคำถามนี้มาก่อน เช่น

"กระบวนการไกล่เกลี่ยข้อพิพาททางอาญาใช้เวลานานแค่ไหน?"

จากนั้น ผู้วิจัยได้เขียนโค้ดด้วยภาษา Python เพื่อส่งคำถามเข้าไปเพื่อให้ระบบ RAG ดึงข้อมูล (Retrieve) ออกมาแสดงจำนวน 5 รายการ โดยเรียงลำดับจากข้อความที่มีตัวเลขเวกเตอร์ใกล้เคียงกับตัวเลขเวกเตอร์ของคำถามมากที่สุดไปจนถึงน้อยที่สุด ดังตัวอย่าง

"แล้วแต่ความซับซ้อนของเรื่อง บางเรื่องใช้เวลาแค่ไม่กี่ชั่วโมง เรื่องง่ายอาจตกลงกันได้เร็ว เรื่องยากอาจต้องใช้หลายครั้ง ตัวอย่างเช่น คู่กรณีตกลงกันได้ภายในครึ่งวันหลังจากเริ่มไกล่เกลี่ย", 0.3937480342720995

"ไม่เกิน 30 วัน นับจากวันนัดครั้งแรก (ขยายได้อีก 30 วันหากจำเป็น) เพื่อให้กระบวนการมีประสิทธิภาพและไม่ล่าช้า ตัวอย่างเช่น กรณีคดีลหุโทษอาจไกล่เกลี่ยเสร็จภายใน 2 สัปดาห์", 0.40097815320380414

"ต้องเสร็จภายใน 30 วันนับจากนัดแรก แต่ขยายได้ไม่เกิน 30 วันหากจำเป็น กฎหมายกำหนดกรอบเวลาเพื่อให้กระบวนการรวดเร็วและมีประสิทธิภาพ ตัวอย่างเช่น คดีลักทรัพย์ใช้เวลาไกล่เกลี่ย 20 วันจนได้ข้อตกลง", 0.4056141087555718

"เป็นกระบวนการที่คู่กรณีในคดีอาญาเจรจากันเพื่อตกลงหรือเยียวยาความเสียหาย โดยมีผู้ไกล่เกลี่ยช่วย เพื่อระงับคดีอาญา การไกล่เกลี่ยช่วยลดความขัดแย้งและป้องกันการฟ้องคดีในศาล โดยคู่กรณีต้องสมัครใจเข้าร่วม ตัวอย่างเช่น ผู้ต้องหาทำร้ายร่างกายผู้อื่นเล็กน้อยและตกลงชดใช้ค่าเสียหายผ่านการไกล่เกลี่ย ทำให้คดีไม่ต้องขึ้นศาล", 0.40773292157437924

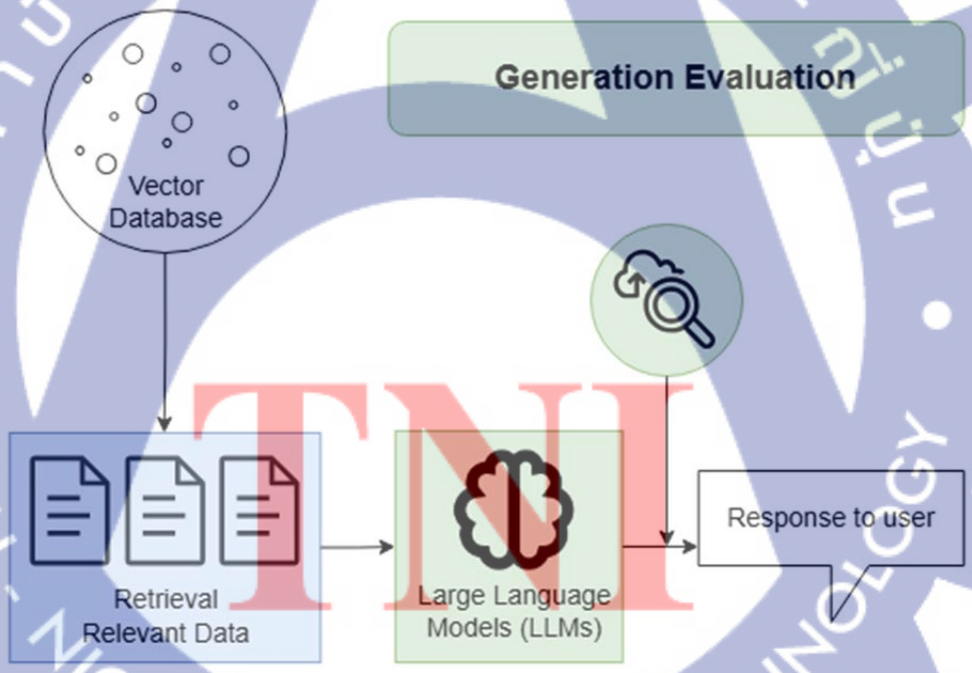
"คือการพูดคุยตกลงกันระหว่างคู่กรณีเพื่อยุติข้อพิพาทโดยไม่ต้องขึ้นศาล เป็นกระบวนการที่ช่วยให้ทั้งสองฝ่ายหาทางออกอย่างสันติ โดยไม่ต้องให้ศาลตัดสิน ตัวอย่างเช่น สองฝ่ายมีปัญหาเรื่องที่ดิน จึงมาไกล่เกลี่ยกันก่อนขึ้นศาล", 0.4150314963406827

จากตัวอย่างแสดงให้เห็นถึงคำตอบที่ระบบ RAG ดึงมาจากฐานข้อมูลเวกเตอร์นั้นมีตัวเลขเวกเตอร์กำกับอยู่ในทุกข้อความ ซึ่งเป็นค่าที่แสดงระยะห่างระหว่างคำถามและคำตอบ โดยข้อความที่มีค่าเวกเตอร์ใกล้ 1 ที่สุดหมายถึงมีความหมายใกล้กับคำตอบมากที่สุด จากนั้นผู้วิจัยได้เขียนโค้ดภาษา Python ตรวจสอบว่าข้อความที่ถูกต้องตรงกันกับคำตอบที่แท้จริง (Ground Truth) ซึ่งอยู่ในชุดข้อมูลนั้นอยู่ในลำดับที่เท่าไรของข้อมูลที่ถูกดึงมาทั้ง 5 รายการ แล้วจึงนำไปคำนวณเพื่อให้คะแนน โดยยิ่งคะแนนมีค่าสูง หมายถึง Embedding Model และระบบดึงข้อมูลยิ่งมีความแม่นยำมาก

Precision@k ใช้เพื่อวัดความแม่นยำของผลลัพธ์ที่อยู่ใน Top-k ซึ่งในการศึกษานี้คือกำหนดค่า k=5 อันดับ โดยผู้วิจัยได้เขียนโค้ดด้วยภาษา Python เพื่อส่งคำถามเข้าไปเพื่อให้ระบบ RAG ดึงข้อมูล (Retrieve) ออกมาแสดงจำนวน 5 รายการ และนับว่ามีกี่รายการที่เกี่ยวข้อง เพื่อนำไปคำนวณและให้คะแนน โดยคะแนนของ Precision@k นี้ที่มีค่าสูง หมายถึง Embedding Model และระบบดึงข้อมูลยิ่งมีความแม่นยำมาก

Hit-Rate@k ใช้เพื่อวัดความสามารถโดยรวมของระบบในการหาคำตอบที่ถูกต้องอย่างน้อยหนึ่งรายการในผลลัพธ์ Top-k ในการศึกษานี้คือ 5 อันดับแรก โดยไม่พิจารณาว่าคำตอบนั้นจะอยู่ในลำดับที่เท่าไร โดยคะแนนที่มีค่าสูง หมายถึง Embedding Model และ ระบบดึงข้อมูลยิ่งมีความแม่นยำมาก

Generation Evaluation ซึ่งจะประเมินความสามารถโมเดลภาษาขนาดเล็ก ในการสร้างข้อความที่เป็นธรรมชาติ ดังแสดงในรูปที่ 3.10 โดยใช้เมตริก Faithfulness ซึ่งจะประเมินความถูกต้องและความซื่อสัตย์ของส่วนการสร้างคำตอบ (Generation) ในระบบ RAG โดยจะตรวจสอบว่าคำตอบที่โมเดลสร้างขึ้น (Generated Answer) มีเนื้อหาที่อ้างอิงและสอดคล้องกับข้อมูลที่ดึงมา (Retrieved Context) หรือไม่



รูปที่ 3.10 Generation Evaluation

เมตริกนี้มีจุดประสงค์หลักเพื่อตรวจจับอาการหลอนของโมเดล (Hallucination) โดยการให้ให้ ChatGPT สกัดใจความย่อยออกมาจากตำตอบที่ระบบสร้างขึ้น จากนั้นนำไปตรวจสอบยืนยันกับบริบท (Context) ที่ระบบดึงมาจากฐานข้อมูลเวกเตอร์ในขั้นตอนการดึงข้อมูล (Retrieval) ว่ามีความ

เกี่ยวข้องกันหรือไม่ เพื่อเป็นข้อพิสูจน์ว่าระบบสร้างคำตอบ (Generation) ไม่ได้สร้างคำตอบขึ้นมาเอง โดยใช้ชุดคำถามสุ่ม 100 คำถามในการวัดผล โดยคะแนนของเมตริก Faithfulness ที่สูงแสดงให้เห็นว่าข้อความที่ถูกสร้างด้วยโมเดลภาษานั้นมีความเกี่ยวข้องกับข้อมูลที่ดึงมาจากฐานข้อมูลเวกเตอร์

กล่าวโดยสรุป การใช้ MRR, Precision@k และ Hit-Rate@k ประเมินระบบในส่วนของ Retrieval และการใช้ Faithfulness ประเมินในส่วนของ Generation เพื่อใช้ในการวัดผลว่าระบบมีประสิทธิภาพตรงตามวัตถุประสงค์เพียงใด

3.7 Hardware, Software and Service

ในงานวิจัยนี้เน้นที่การใช้งานโดยทั่วไป เช่น การใช้งานในคอมพิวเตอร์ส่วนบุคคล ดังนั้นจึงใช้ Hardware Software และ Service ดังต่อไปนี้

Hardware ที่ใช้สำหรับงานวิจัย ได้แก่

- Intel Core i5-12500H (2.50 GHz) Processor (3.33 GHz)
- 16.00 GB RAM
- Nvidia GeForce RTX 3050 (4 GB) Graphics Card

Software ที่ใช้สำหรับงานวิจัย ได้แก่

- Python Programming Version 3.12
- Visual Studio Code (VS Code) สำหรับการเขียนโค้ดโปรแกรม
- TiDB ผู้ให้บริการ Cloud สำหรับ Vector Database
- Gradio Library สำหรับการสร้าง User Interface
- Ollama ใช้ประมวลผล Small Language Model ในเครื่องคอมพิวเตอร์ส่วนบุคคล

Service ที่ใช้สำหรับงานวิจัย ได้แก่

- Colab Pro+ สำหรับประมวลผลการสร้างชุดข้อมูล
- ChatGPT 5 สำหรับการสร้างชุดข้อมูล
- Google Gemini สำหรับการประเมิน Generation Model

บทที่ 4
ผลการดำเนินงาน การวิเคราะห์และสรุปผลต่างๆ

จากการศึกษาแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้องเพื่อเป็นความรู้สำหรับการศึกษาในการวิจัยครั้งนี้ ผู้วิจัยได้ทำการทดสอบระบบเพื่อวัตถุประสงค์ในการให้คำแนะนำเกี่ยวกับกฎหมายคุ้มครองผู้บริโภค และเพื่อเปรียบเทียบประสิทธิภาพของโมเดลภาษาพื้นฐาน โดยได้นำผลการทดลองมาวิเคราะห์และสรุปผลลัพธ์การวิจัย โดยมีรายละเอียดดังนี้

4.1 ผลการเปรียบเทียบประสิทธิภาพของ Embedding Model

ในการเปรียบเทียบประสิทธิภาพของ Embedding Model ทั้ง 6 โมเดลที่นำมาใช้ในงานวิจัยนี้สำหรับขั้นตอนของการแปลงข้อความให้เป็นตัวเลขเวกเตอร์ที่มีความหมาย และบันทึกในฐานข้อมูลเวกเตอร์สำหรับเรียกใช้งาน ได้มีการใช้เมตริกความแม่นยำในการดึงข้อมูล Mean Reciprocal Rank (MRR), Precision@k และ Hit-Rate@k ในการประเมินประสิทธิภาพ โดยใช้ชุดคำถามแบบสุ่มจำนวน 400 ข้อ ซึ่งคิดเป็น 25% ของชุดข้อมูลคำถาม-คำตอบด้านกฎหมายทั้งหมด 2,000 แถว

ตารางที่ 4.1 คะแนนประเมิน Embedding Model

Embedding Model	Evaluation Metric		
	Mean Reciprocal Rank (MRR)	Hit Rate@k	Precision@k
BGE-M3	**0.54**	0.70	0.70
Thai CCL	0.52	**0.71**	**0.71**
ARCTIC	0.49	0.64	0. 64
E5 Base	0.47	0.66	0.66
MiniLMV2	0.28	0.42	0.42
MRL	0.01	0.00	0.01

จากผลการประเมินความแม่นยำในการดึงข้อมูลของ Embedding Model แสดงให้เห็นว่าค่า Mean Reciprocal Rank (MRR) Metric ของโมเดล BGE-M3 ได้คะแนน 0.54 เมื่อเทียบกับโมเดลอื่น อย่างไรก็ตาม ค่า Hit Rate@k และค่า Precision@k ของโมเดล WangchanX-Legal-ThaiCCL-Retriever ได้คะแนน 0.71 และ 0.71 ตามลำดับ ซึ่งมากกว่า BGE-M3 และโมเดลตัวอื่นเช่นกัน ดังที่แสดงในตารางที่ 4.1

4.2 ผลการเปรียบเทียบประสิทธิภาพของ Based Model

ในการเปรียบเทียบประสิทธิภาพของ Based Model ทั้ง 3 โมเดลที่นำมาใช้ในงานวิจัยนี้ สำหรับขั้นตอนของการสร้างข้อความที่ได้จากการดึงข้อมูลจากฐานข้อมูลเวกเตอร์ เพื่อสร้างคำตอบให้แก่ผู้ใช้ ได้มีการใช้เมตริก Faithfulness เพื่อความแม่นยำของ RAG โดยใช้ชุดคำถามแบบสุ่มจำนวน 100

ตารางที่ 4.2 คะแนนประเมิน Generation Model

Generation Models	Faithfulness
Llama3.2-3b	0.1128
Scb10x-typhoon2.1-gemma3-4b	**0.1249**
OpenThaiGPT-1.5-7b	0.1179

จากผลการประเมินความแม่นยำในการสร้างข้อความของ Generation Model แสดงให้เห็นว่าค่า Faithfulness ของโมเดล Scb10x-typhoon2.1-gemma3-4b ได้คะแนน 0.1249 เมื่อเทียบกับโมเดลอื่น ดังที่แสดงในตารางที่ 4.2

4.3 การวิเคราะห์ข้อมูล

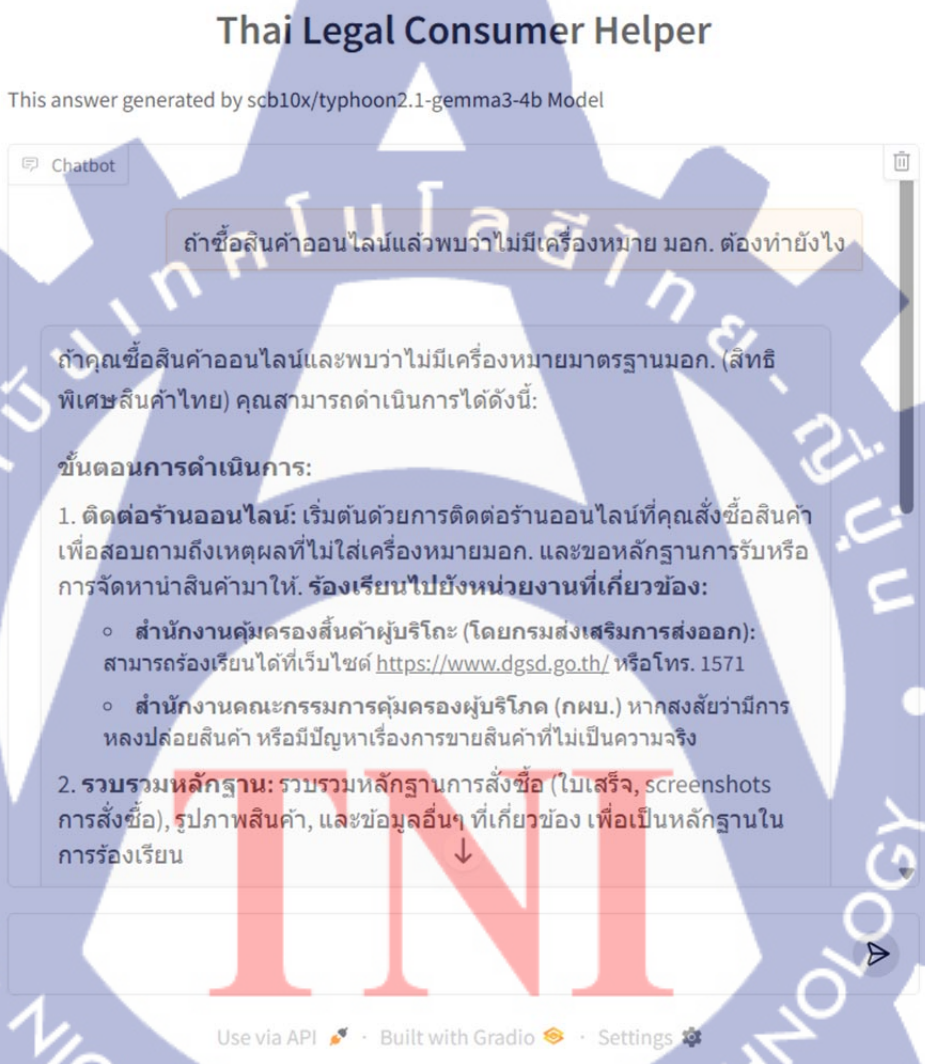
จากผลการทดลองทั้งในส่วนของ Retrieval และ Generation แสดงให้เห็นว่าระบบ Retrieval Augmented Generation (RAG) ที่เสริมด้วยระบบจัดเตรียมข้อมูลกฎหมายล่วงหน้า Semantic Legal Preprocessing (SLP) ที่ได้นำเสนอในงานวิจัยนี้ให้ผลลัพธ์ที่น่าสนใจ โดยแบ่งเป็น 3 ส่วน ได้แก่

ส่วนของการเตรียมข้อมูลกฎหมายล่วงหน้า Semantic Legal Preprocessing (SLP) ซึ่งเป็นวิธีที่นำเสนอในงานวิจัยนี้ซึ่งเป็นส่วนสำคัญในการเตรียม

ส่วนของดึงข้อมูล Retrieval เมื่อใช้งาน Embedding Model WangchanX-Legal-ThaiCCL-Retriever ที่ให้ผลลัพธ์ในเมตริก Hit Rate 0.71 และ Precision 0.71 ซึ่งสูงกว่า Embedding Model อื่นที่เลือกมาทดสอบในงานวิจัยนี้ เนื่องจาก WangchanX-Legal-ThaiCCL-Retriever เป็น Embedding Model ที่ถูกฝึกด้วยข้อมูลกฎหมายโดยเพราะกฎหมายธุรกิจ ทำให้โมเดลมีความเข้าใจภาษาและคำศัพท์ทางกฎหมายที่เป็นความเชี่ยวชาญเฉพาะด้านได้เป็นอย่างดี ส่งผลให้การดึงข้อมูล (Retrieval) ได้ผลลัพธ์ซึ่งเป็นข้อความที่เกี่ยวข้องกับคำถามของผู้ใช้

ส่วนของการสร้างข้อมูล Generation เมื่อใช้งาน Based Model Scb10x-typhoon2.1-gemma3-4b ที่ให้ผลลัพธ์ที่ให้ผลลัพธ์ในเมตริก Faithfulness 0.1249 ซึ่งสูงกว่า Based Model ตัวอื่นที่เลือกมาทดสอบในงานวิจัยนี้

สรุปผลการวิจัยในการสร้างระบบแชทบอทเพื่อให้คำแนะนำกฎหมายคุ้มครองผู้บริโภคด้วยระบบปัญญาประดิษฐ์แบบ Retrieval-Augmented Generation ให้ผลลัพธ์ที่ดีเมื่อใช้ กระบวนการประมวลผลข้อมูลกฎหมายล่วงหน้า Semantic Legal Preprocessing (SLP) ร่วมกับ Embedding Model: WangchanX-Legal-ThaiCCL-Retriever และ Based Model: Scb10x-typhoon2.1-gemma3-4b ดังแสดงในรูปที่ 4.1 และ รูปที่ 4.2



รูปที่ 4.1 ตัวอย่างการทำงานของโปรแกรม

รูปที่ 4.1 ตัวอย่างการทำงานของโปรแกรม (ต่อ)

บทที่ 5
บทสรุป อภิปราย และข้อเสนอแนะ

การศึกษาวิจัยเรื่องระบบปัญญาประดิษฐ์แบบ Retrieval-Augmented Generation สำหรับการให้คำแนะนำทางกฎหมายคุ้มครองผู้บริโภคในประเทศไทย ซึ่งการวิจัยนี้มีวัตถุประสงค์ตามบทที่ 1 ได้นำเสนอ ข้อสรุปตามลำดับดังนี้

5.1 สรุปผลการวิจัย

5.2 ปัญหาและอุปสรรคในการทำวิจัย

5.3 ข้อเสนอแนะงานวิจัย

5.1 สรุปผลการวิจัย

ผลลัพธ์ของงานวิจัยในการสร้างระบบแชทบอทเพื่อให้คำปรึกษาปัญหากฎหมายคุ้มครองผู้บริโภคแก่ประชาชน และการเปรียบเทียบประสิทธิภาพของ Embedding Model และ Based Model ในภาษาไทยสำหรับกฎหมายไทย พบว่า ระบบสามารถให้คำตอบที่เป็นโครงสร้างและมีการใช่ภาษาที่เข้าใจง่าย มีความแม่นยำ โดยพบว่าระบบ RAG จะมีความแม่นยำมากน้อยเพียงใด ขึ้นอยู่กับปัจจัยต่างๆ ได้แก่ จำนวนข้อมูล การเตรียมข้อมูล คุณภาพของข้อมูล การเลือก Embedding Model การเลือก Generation Model ซึ่งคุณสมบัติของปัจจัยเหล่านี้ นำมาซึ่งคำตอบที่แม่นยำและตรงตามวัตถุประสงค์ของงานวิจัย โดยเฉพาะอย่างยิ่ง ขั้นตอนการเตรียมข้อมูล สอดคล้องกับแนวคิด “Garbage In, Garbage Out” (GIGO) หมายถึง การนำเข้าข้อมูลคุณภาพต่ำนำมาซึ่งผลลัพธ์ที่คุณภาพต่ำ แนวคิดนี้ถือเป็นหัวใจสำคัญในงานวิจัยสาขาวิทยาการคอมพิวเตอร์ การวิเคราะห์ข้อมูล และสาขาอื่นๆ อีกมากมาย

5.2 ปัญหาและอุปสรรคในการทำวิจัย

ปัญหาที่พบในงานวิจัยนี้ ได้แก่ จำนวนข้อมูลที่ใช้สำหรับระบบ RAG มีจำกัด เนื่องจากตัวบทกฎหมายและพระราชบัญญัติที่เกี่ยวข้องโดยตรงกับการคุ้มครองผู้บริโภคมีจำนวนจำกัด ผู้วิจัยจึงจำเป็นต้องรวบรวมข้อมูลเพิ่มเติมจากแหล่งอื่น เช่น บทความหรือกรณีศึกษาต่างๆ ซึ่งข้อมูลที่เป็นสาธารณะที่เป็นทางการและน่าเชื่อถือนั้นยังมีจำนวนน้อย รวมถึงในการรวบรวมข้อมูลที่เกี่ยวข้องกับกฎหมายต้องอาศัยเวลามากเนื่องจากต้องคัดกรองอย่างละเอียดถี่ถ้วนเพื่อไม่ให้เกิดความผิดพลาด และยังคงต้องการปรับปรุงข้อมูลสำหรับรองรับกรณีใหม่ที่เกิดขึ้นอย่างเป็นปัจจุบัน

5.3 ข้อเสนอแนะงานวิจัย

เนื่องจากชุดข้อมูลที่ใช้สำหรับระบบ RAG ในงานวิจัยนี้มีจำกัด ทำให้ระบบอาจไม่ครอบคลุมกรณีต่างๆ อย่างทั่วถึง เช่น กรณีคดีความใหม่ๆ หรือปัญหาที่เพิ่งเคยเกิดขึ้นเป็นครั้งแรกในสังคมไทย ผู้วิจัยจึงแนะนำให้นำเข้าข้อมูลจากแหล่งอื่น เช่น จากวิดีโอ ข่าวสาร ที่คัดกรองแล้ว รวมถึงสื่อที่ให้ความรู้ด้านกฎหมายคุ้มครองผู้บริโภคจากแหล่งข้อมูลที่เป็นทางการและน่าเชื่อถือ นำมาปรับปรุงชุดข้อมูลด้วยกระบวนการเตรียมข้อมูลล่วงหน้า SLP ที่นำเสนอใจงานวิจัยนี้เพื่อให้ได้ข้อมูลที่ใหม่และมีจำนวนข้อมูลที่ครอบคลุมกรณีต่างๆ มากยิ่งขึ้น เพื่อให้ได้คำตอบสำหรับคำถามที่ครอบคลุมมากขึ้น


สถาบันเทคโนโลยีไทย-ญี่ปุ่น
THAI - NICHI INSTITUTE OF TECHNOLOGY
TNI

บรรณานุกรม

บรรณานุกรม

[1] Electronic Transactions Development Agency, *"Slide ผลการสำรวจมูลค่าพาณิชย์อิเล็กทรอนิกส์ ปี 2566,"* [Online]. Available : https://www.etda.or.th/th/Useful-Resource/documents-for-download.aspx?group=1&page=2. [Accessed : 26 กรกฎาคม 2568].

[2] J. Ling and M. Afzaal, "Automatic question-answer pairs generation using pre-trained large language models in higher education," *Computers and Education : Artificial Intelligence,* vol. 6, no. 3, pp. 252, June 2024.

[3] M. Raheem and Y. C. Chong, "E-commerce fake reviews detection using LSTM with word2vec embedding," *Journal of Computing and Information Technology,* vol. 32, no. 2, pp. 65-80, September 2024.

[4] C. Si et al., "Sub-character tokenization for chinese pretrained language models," *Transactions of the Association for Computational Linguistics,* vol. 11, no. 5, pp. 469-487, May 2023.

[5] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems,* NIPS 2017, Long Beach, California, USA, December 4-9, 2017, pp. 5,998-6,008.

[6] M. Rahman et al., "ChatGPT and academic research : A review and recommendations based on practical examples," *Journal of Education, Management and Development Studies,* vol. 3, no. 1, pp. 1-12, March 2023.

[7] G. Lotfian, et al., "Performance review of meta LLaMa 3.1 in thoracic imaging and diagnostics," *iRADIOLOGY,* vol 3, no. 4, pp. 279-288, August 2025.

[8] A. Pande et al., "Comprehensive study of google gemini and text generating models : Understanding capabilities and performance," *Advances in Computing Control and Telecommunication Technologies,* ACT 2024, Hyderabad, India, June 21-22, 2024, pp. 1,482-1,484.

[9] D. Vake et al., "Bridging the question-answer gap in retrieval-augmented generation : Hypothetical prompt embeddings," *Institute of Electrical and Electronics Engineers,* vol. 13, no. 8, pp. 129,952-129,961, July 2025.

[10] M. Hindi et al., "Enhancing the precision and interpretability of retrieval-augmented generation (RAG) in legal technology : A survey," *Institute of Electrical and Electronics Engineers,* vol. 13, no. 3, pp. 46,171-46,189, March 2025.

[11] R. Hu et al., "ICCA-RAG : Intelligent customs clearance assistant using retrieval-augmented generation (RAG)," *Institute of Electrical and Electronics Engineers,* vol. 13, no. 5, pp. 39,711-39,726, February 2025.

[12] R. S. M. Wahidur et al., "Legal query rag," *Institute of Electrical and Electronics Engineers,* vol. 13, no. 2, pp. 36,978-36,994, February 2025.

[13] S. Ajay Mukund and K. S. Easwarakumar, "Optimizing legal text summarization through dynamic retrieval-augmented generation and domain-specific adaptation," *Symmetry,* vol. 17, no. 5, p. 633, April 2025.

[14] C. Jiang and X. Yang, "Legal syllogism prompting : Teaching large language models for legal judgment prediction," *International Conference on Artificial Intelligence and Law,* ICAIL 2023, University of Minho in Braga, Portugal, June 19-23, 2023, pp. 223-225.

[15] สภาองค์กรของผู้บริโภค, *"บทเรียนผู้บริโภค,"* [Online]. Available : https://www.tcc.or.th/document-category/summary-of-important-judgments. [Accessed : 26 กรกฎาคม 2568].

[16] J. Chen et al., "M3-embedding : Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," *Findings of the Association for Computational Linguistics,* ACL 2024, Bangkok, Thailand, August 11-16, 2024, pp. 875-890.

[17] VISTEC-depa AI Research Institute of Thailand, *"WangchanX-Legal-ThaiCCL-Retriever,"* [Online]. Available : https://huggingface.co/airesearch/WangchanX-Legal-ThaiCCL-Retriever. [Accessed : March 15, 2025].

[18] Snowflake, *"Snowflake-Arctic-Embed-l-v2.0,"* [Online]. Available : https://huggingface.co/Snowflake/snowflake-arctic-embed-l-v2.0. [Accessed : May 19, 2025].

[19] J. Chen et al., *"M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation,"* [Online]. Available : https://huggingface.co/papers/2402.05672. [Accessed : April 28, 2025].

[20] Sentence Transformers, *"All-MiniLM-L6-v2,"* [Online]. Available : https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. [Accessed : June 12, 2025].

[21] Sentence Transformers, *"Static-Retrieval-mrl-en-v1,"* [Online]. Available : https://huggingface.co/sentence-transformers/static-retrieval-mrl-en-v1. [Accessed : July 5, 2025].

[22] OpenThaiGPT, *"OpenThaiGPT 1.5 : A Thai-Centric Open Source Large Language Model,"* [Online]. Available : https://openthaigpt.aieat.or.th/previous-versions-and-resources/openthaigpt-1.5. [Accessed : November 11, 2024].

[23] Typhoon AI, *"Typhoon2.1-Gemma3-4b,"* [Online]. Available : https://huggingface.co/typhoon-ai/typhoon2.1-gemma3-4b. [Accessed : November 18, 2026].

[24] Meta Llama, *"Llama-3.2-3B,"* [Online]. Available : https://huggingface.co/meta-llama/Llama-3.2-3B. [Accessed : November 28, 2024].