การวิเคราะห์ความผิดพลาดของ OCR ในระบบ Retrieval-Augmented Generation (RAG) สำหรับเอกสารภาษาไทย

Analysis of OCR-Induced Errors in Retrieval-Augmented Generation (RAG) for Thai Documents

อรรถสิทธิ์ พิมพ์อินทร์
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีไทย-ญี่ปุ่น
กรุงเทพมหานคร
attasit@tni.ac.th

พิชิตชัย คำอินทร์
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีไทย-ญี่ปุ่น
กรุงเทพมหานคร
pichitchai@tni.ac.th

ปริทัศน์ ทิพยากุล
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีไทย-ญี่ปุ่น
กรุงเทพมหานคร
paritat.tip@tni.ac.th

บทคัดย่อ — งานวิจัยนี้มุ่งศึกษาผลกระทบของ "สัญญาณรบกวน" จากกระบวนการแปลงไฟล์ภาพเป็นข้อความ (OCR) ในเอกสารภาษาไทยที่ส่งผลโดยตรงต่อประสิทธิภาพของระบบค้นหาและตอบคำถามอัตโนมัติ (RAG) โดยพบว่าแม้เทคโนโลยี OCR จะช่วยเปลี่ยนเอกสารประกาศเก่าให้เป็นข้อมูลดิจิทัลได้ แต่ข้อผิดพลาดเฉพาะตัว เช่น วรรณยุกต์ลอยหรือตัวอักษรที่ผิดเพี้ยน กลับทำให้ความแม่นยำในการจัดอันดับเอกสารลดลง งานวิจัยจึงได้จำแนกความผิดพลาดออกเป็น 5 กลุ่ม (E1–E5) เพื่อให้เห็นปัญหาชัดเจนขึ้น และจากการทดสอบกับเอกสารสถาบัน 197 ฉบับ ด้วยโมเดล BGE-M3 ได้ค้นพบปรากฏการณ์ "ความย้อนแย้งของการค้นคืน (Retrieval Paradox)" กล่าวคือ ข้อมูลจาก OCR สามารถดึงชิ้นส่วนข้อมูลที่เกี่ยวข้องขึ้นมาได้จำนวนมาก (P@10 = 0.418) แต่กลับจัดอันดับความถูกต้องได้ไม่ดีนัก (MRR เพียง 0.738) เมื่อเทียบกับเอกสารต้นฉบับ เพื่อแก้ปัญหาผู้วิจัยจึงใช้กลไกกำกับการอ้างอิง ซึ่งเป็นเทคนิคการสั่งงาน (Prompt) ให้ AI ตรวจสอบเนื้อหาก่อนตอบเสมอเพื่อลดโอกาสการตอบผิด (False Confidence) ผลการศึกษาชี้ให้เห็นว่าระบบ RAG ที่ใช้ข้อมูลจาก OCR สามารถนำไปใช้งานจริงได้ โดยกุญแจสำคัญอยู่ที่การคัดกรองคุณภาพข้อมูล (Chunk Validation) และการบังคับให้ระบบระบุแหล่งที่มาของคำตอบเสมอเพื่อให้ผลลัพธ์มีความน่าเชื่อถือ

คำสำคัญ — *Thai RAG, สัญญาณรบกวนจาก OCR, การประมวลผลเอกสาร, กรอบจำแนกความผิดพลาด, การลดอาการหลอนของโมเดล (Hallucination)*

ABSTRACT — This research aims to investigate the impact of "noise" from Optical Character Recognition (OCR) processes on Thai documents, which directly affects the performance of automated search and question-answering systems (RAG). It was found that although OCR technology enables the digitization of legacy announcements, specific errors—such as floating tone marks or character distortions—degrade document ranking accuracy. The research classifies these errors into 5 groups (E1–E5) to clearly identify the problems. Experiments with 197 institutional documents using the BGE-M3 model reveal a "Retrieval Paradox": OCR data retrieves a high volume of relevant chunks (P@10 = 0.418) but exhibits poor ranking quality (MRR only 0.738) compared to the original documents. To address this, the researchers employ "Citation Guardrails," a prompt engineering technique that requires the AI to validate content before answering to reduce the risk of false confidence. The study suggests that OCR-based RAG systems are viable for practical application, provided that strict Chunk Validation and mandatory source citation (citation enforcement) are implemented to ensure maximum reliability.

Keywords — *Thai RAG, OCR Noise, Document Processing, Error Taxonomy, Hallucination Mitigation.*

I. บทนำ

ระบบ Retrieval-Augmented Generation (RAG) ช่วยเพิ่มให้กับการตอบคำถามของ AI โดยการค้นหาข้อมูลจากเอกสารที่เกี่ยวข้องมาเป็นบริบทก่อนสร้างคำตอบ [1] ในการใช้งานจริงภายในสถาบันการศึกษาของไทย มักพบปัญหาจากเอกสารเก่าที่เป็นไฟล์สแกน ซึ่งต้องแปลงเป็นข้อความดิจิทัลด้วยเทคโนโลยี OCR กระบวนการนี้มักก่อให้เกิดสัญญาณรบกวนจาก OCR ตั้งแต่เรื่องตัวอักษรผิดเพี้ยน การเว้นวรรคผิด ไปจนถึงการสูญเสียโครงสร้างของตารางและย่อหน้า งานวิจัยนี้จึงมุ่งเจาะลึกเพื่อตอบคำถามว่า “ความผิดพลาดเกิดขึ้นเพราะอะไร” เมื่อระบบ RAG ต้องทำงานบนข้อมูลที่มีข้อผิดพลาดจาก OCR โดยทำการวิเคราะห์ความสัมพันธ์ระหว่างคุณภาพการค้นคืน ด้วยค่า MRR และ NDCG กับ ความถูกต้องของคำตอบ ด้วยค่า Accuracy และ Similarity นอกจากนี้ ผู้วิจัยยังได้นำเสนอ การจำแนกประเภทความผิดพลาด (Error Taxonomy) ในการระบุสาเหตุของปัญหาและเสนอแนวทางเพิ่มความน่าเชื่อถือ (Reliability) ให้ระบบ RAG

งานวิจัยนี้เลือกใช้ เอกสารประกาศและคำสั่งจริงของสถาบันการศึกษา [2] เป็นกรณีศึกษา โดยนำมาเปรียบเทียบใน 3 รูปแบบ ได้แก่ ไฟล์ข้อความ (TXT), ไฟล์ข้อความ PDF, และข้อความที่ผ่านกระบวนการ OCR เพื่อจำลองคุณภาพของข้อมูลที่เกิดขึ้น ในการวิเคราะห์ช่องว่างระหว่างการค้นคืน (Retrieval) และการสร้างคำตอบ (Generation แม้ระบบจะสามารถดึงเอกสารที่เกี่ยวข้องขึ้นมาได้ แต่โมเดลภาษากลับให้คำตอบที่ผิด (False Confidence) เนื่องจากข้อมูลสำคัญ เช่น ตัวเลข วันที่ หรือเงื่อนไขกฎระเบียบ ถูกบิดเบือนจาก

กระบวนการ OCR จากนั้นจึงจำแนกความผิดพลาดตามรูปแบบที่พบบ่อย เช่น ตัวสะกดเพี้ยน เว้นวรรคหาย เลข/วันที่ผิด และการเสียรูปของหัวข้อ/ตาราง เพื่อหาแนวทางปรับปรุงที่สามารถนำไปประยุกต์ใช้ได้จริงกับงานเอกสารของสถาบัน

วัตถุประสงค์ของงานวิจัยมีดังนี้

1. เพื่อวิเคราะห์ข้อผิดพลาดจาก OCR ต่อคุณภาพของผลการจัดอันดับการค้นคืนข้อมูล และความถูกต้องของคำตอบคำถามของระบบ RAG สำหรับภาษาไทย

2. เพื่อประเมินแนวทางเพิ่มความน่าเชื่อถือผ่านกลไกการคัดกรองชิ้นส่วนข้อมูลแบบที่คำนึงถึงความผิดพลาด และการอ้างอิงก่อนที่จะสร้างคำตอบ

II. งานวิจัยที่เกี่ยวข้อง

งานวิจัยด้าน Retrieval-Augmented Generation (RAG) ในช่วงหลายปีที่ผ่านมาได้ให้ความสำคัญกับการเพิ่มความถูกต้องและความน่าเชื่อถือของคำตอบ ผ่านการดึงหลักฐานจากเอกสารก่อนการสร้างคำตอบ ข้อมูลการทำเอกสารจำนวนมาก ได้รับการประเมินข้อมูลจากเอกสารภาษาอังกฤษ ทำให้ยังมีช่องว่างในการอธิบายความผิดพลาด เมื่อข้อมูลต้นทางมีข้อผิดพลาดจาก OCR โดยเฉพาะในเอกสารภาษาไทยที่มี “ตัวเลข/วันที่” ซึ่งมีความผิดเพี้ยนระดับอักขระและการสูญเสียโครงสร้างของเอกสาร

ก. รากฐาน RAG และแนวโน้มของการประเมิน งานวิจัยในช่วงหลายปีที่ผ่านมาได้มีการสังเคราะห์องค์ประกอบหลักของระบบ RAG ไว้อย่างครบถ้วนแล้ว ตั้งแต่ขั้นตอนการจัดทำดัชนี (Indexing) การค้นคืน (Retrieval) การจัดอันดับซ้ำ (Reranking) ไปจนถึงการสร้างคำตอบ (Generation) โดยมีแนวคิดสำคัญว่าการประเมินประสิทธิภาพควรพิจารณา “คุณภาพของหลักฐานที่ค้นเจอ” ควบคู่ไปกับ “ความถูกต้องของคำตอบสุดท้าย” สำหรับแนวทางการวัดผลที่ได้รับความนิยม ได้แก่ การใช้ตัวชี้วัดด้านคุณภาพการจัดอันดับ เช่น MRR และ NDCG ร่วมกับตัวชี้วัดด้านความครอบคลุม เช่น Hit@K และ F1@K เพื่อตรวจสอบว่าระบบสามารถดึงข้อมูลที่สำคัญที่สุดขึ้นมาอยู่ในลำดับต้นได้จริงหรือไม่ ซึ่งส่งผลโดยตรงต่อคุณภาพการตอบ นอกจากนี้ยังเน้นย้ำประเด็นเรื่อง ความน่าเชื่อถือ (Reliability) และ อาการหลอนของข้อมูล (Hallucination) ว่าเป็นปัจจัยวิกฤตที่ต้องคำนึงถึงเมื่อนำระบบ RAG ไปใช้งานจริง [3], [4]

ข. Hallucination, False Confidence และ Guardrails ใน RAG ปัญหาการสร้างข้อมูลเท็จ (Hallucination) ในระบบ RAG ไม่ได้เกิดจากตัวโมเดลเพียงอย่างเดียว แต่มักมีสาเหตุหลักมาจาก คุณภาพของบริบทที่ได้รับ (Context Quality) หากบริบทที่ดึงมามีสัญญาณรบกวน (Noise) ข้อมูลเกิดข้อผิดพลาด หรือข้อมูลสำคัญตกหล่นไประหว่างการแบ่งส่วน (Chunking) โมเดลภาษาอาจพยายาม เติมเต็มคำตอบ ช่องว่างเหล่านั้นด้วยการคาดเดา ซึ่งนำไปสู่คำตอบที่ผิดแต่ฟังดูน่าเชื่อถือ หรือที่เรียกว่า ความมั่นใจผิดๆ (False Confidence) แก้ไขปัญหานี้ งานวิจัยในปัจจุบันจึงเสนอแนวทางการสร้าง การป้องกัน (Guardrails) เช่น การกำกับให้โมเดลต้องระบุแหล่งที่มา (Attribution/Citation) ควบคู่กับคำตอบเสมอ และการกำหนดเงื่อนไขอย่างเคร่งครัดว่า “หากไม่พบข้อมูลให้ปฏิเสธการตอบ” วิธีการเหล่านี้ช่วยลดความเสี่ยงของความผิดพลาดของข้อมูล [5], [6]

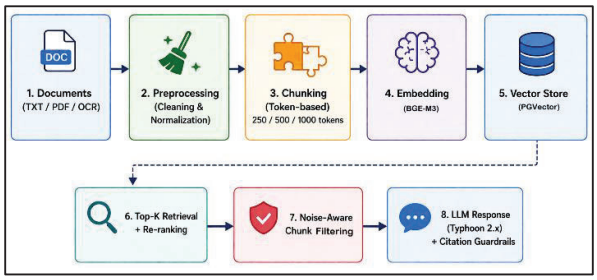
ค. ผลกระทบของ OCR Noise และแนวทางการแก้ไข ในบริบทของเอกสารที่ได้จากการสแกน งานวิจัยด้านระบบค้นคืนและตอบคำถาม (IR/QA) ชี้ให้เห็นว่า "ข้อผิดพลาดจาก OCR" เป็นปัจจัยหลักที่ลดทอนคุณภาพการค้นคืน โดยเฉพาะความผิดพลาดระดับตัวอักษรและตัวเลข รวมถึงการสูญเสียโครงสร้างดั้งเดิมของเอกสาร เช่น ตาราง หรือรายการลำดับหัวข้อ ปัญหานี้ส่งผลให้บริบทข้อมูลเกิดความไม่ต่อเนื่องและกระจัดกระจาย ทำให้ข้อมูลที่ส่งต่อไปยังโมเดลภาษาขาดความสมบูรณ์และลดความถูกต้องลง งานวิจัยในปัจจุบันจึงให้ความสำคัญกับ กระบวนการแก้ไขข้อความหลังจากการทำ OCR (Post-OCR Correction) และ การทำความสะอาดข้อมูล (Data Cleaning) มากขึ้น โดยมีผลการศึกษาที่ยืนยันว่า การคัดกรองชิ้นส่วนข้อมูล (Chunk) ที่มีคุณภาพต่ำออก และการปรับปรุงคุณภาพข้อความก่อนนำไปใช้งาน คือกุญแจสำคัญที่จะช่วยให้ระบบมีความน่าเชื่อถือ และทำงานได้แม้ต้องประมวลผลบนข้อมูลที่มีสัญญาณรบกวน [7]

ง. ส่วนที่งานวิจัยนี้เติมเต็ม แม้งานวิจัยในอดีตจะนำเสนอภาพรวมของสถาปัตยกรรม RAG และแนวทางการลดอาการหลอน (Hallucination) ไว้อย่างครอบคลุม แต่ยังขาดการศึกษาที่เจาะลึกถึงความสัมพันธ์เชิงสาเหตุ ในบริบทเอกสารภาษาไทย โดยเฉพาะประเด็นที่ว่าสัญญาณรบกวนจาก OCR ส่งผลกระทบต่อเนื่องไปยังคุณภาพการจัดอันดับ (MRR/NDCG) และความถูกต้องของคำตอบได้อย่างไร งานวิจัยนี้จึงมุ่งเติมเต็มช่องว่างดังกล่าวด้วยการนำเสนอ กรอบการจำแนกความผิดพลาด (Error Taxonomy: E1–E5) โดยใช้เอกสารจริงของสถาบันในการนำมาประเมิน และทำการวิเคราะห์เชิงลึกเพื่อชี้ให้เห็นว่า ความหนาแน่นของการค้นคืน (Retrieval Density) ที่สูงเพียงอย่างเดียวอาจไม่เพียงพอ หากหลักฐานที่ถูกจัดให้อยู่ในลำดับต้นมีคุณภาพต่ำ ซึ่งสถานการณ์นี้เป็นต้นเหตุสำคัญของการทำให้เกิดความผิดพลาดในการตอบคำถาม และความคลาดเคลื่อนของคำตอบโดยเฉพาะในคำถามที่เกี่ยวข้องกับตัวเลขและวันที่

III. วิธีดำเนินการวิจัย

III.I ภาพรวมสถาปัตยกรรม

งานวิจัยนี้ได้พัฒนากรอบการทำงานของระบบ RAG สำหรับเอกสารภาษาไทยโดยการเพิ่มความน่าเชื่อถือ (Robustness Measures: S1–S3) กับสัญญาณรบกวนจาก OCR เข้ากระบวนการค้นคืนและสร้างคำตอบ เพื่อลดโอกาสเกิดข้อผิดพลาดประเภทต่างๆ (E1–E5) โดยมีภาพรวม ดังแสดงในรูปที่ I



รูปที่ I. ภาพรวมสถาปัตยกรรมและองค์ประกอบต่อสัญญาณรบกวนจาก OCR

III.II ชุดข้อมูลและการเตรียมข้อมูล

งานวิจัยนี้ใช้เอกสารประกาศ/คำสั่งภาษาไทยจำนวน 197 ฉบับ และชุดคำถามทดสอบ 50 ข้อ โดยจัดเตรียมแหล่งข้อมูล 3 ประเภท ได้แก่ (1) TXT ข้อความ (2) PDF-text ข้อความที่สกัดจากไฟล์ PDF และ (3) OCR-text ข้อความที่ได้จากการประมวลผล Tesseract OCR [8] จากนั้นทำการฝังเวกเตอร์โดยใช้ BGE-M3 [9] ในการสร้างเวกเตอร์สโตร์แยกตามแหล่งข้อมูลและขนาดชิ้นส่วนข้อมูล (chunk size) ได้แก่ 250/500/1000 พร้อมการซ้อนทับ (overlap) 50/100/200 เพื่อใช้ในการเปรียบเทียบผลกระทบของคุณภาพแหล่งข้อมูลและการแบ่งส่วนข้อมูลต่อประสิทธิภาพของระบบ

III.III การออกแบบการทดลองและเมตริก

การประเมินประสิทธิภาพแบ่งออกเป็น 2 ส่วน เพื่อให้ครอบคลุมการค้นหาและความถูกต้องของเนื้อหา

1. การประเมินคุณภาพการค้นคืน (Retrieval Quality) ที่ระดับ K=10 โดยใช้ตัวชี้วัด Precision@10 และ Avg. Rel. Chunks@10 เพื่อตรวจสอบความครอบคลุมของข้อมูลที่เกี่ยวข้องร่วมกับค่า MRR และ NDCG@10 เพื่อวัดประสิทธิภาพของการจัดอันดับเอกสาร

2. การประเมินคุณภาพการสร้างคำตอบ (Generation Quality) โดยทดสอบกับโมเดลภาษา Typhoon โดยใช้เกณฑ์การวัด Lax threshold 0.6 และ Strict threshold 0.8 รายงานผ่านค่า Accuracy และ F1 Score นอกจากนี้ ยังมีการวัด ความคล้ายคลึงของคำตอบ (Generation Similarity) ด้วย Normalized Sequence Matching Ratio (ค่า 0.0–1.0) เพื่อตรวจสอบความถูกต้องของการสะกดคำ การเรียบเรียงประโยค และความสอดคล้อง เมื่อเทียบกับคำตอบอ้างอิง

III.IV กรอบการจำแนกความผิดพลาด (Error Taxonomy)

เพื่อระบุสาเหตุที่แท้จริงของปัญหาที่เกิดขึ้นจากการใช้งานข้อมูลข้อความที่ได้จาก OCR ผู้วิจัยได้จำแนกรูปแบบความผิดพลาด (Failure Modes) ที่ส่งผลกระทบโดยตรงต่อประสิทธิภาพของกระบวนการค้นคืนข้อมูล (Retrieval) และการสร้างคำตอบของระบบออกเป็น 5 กลุ่มหลัก (E1–E5) ดังนี้

E1 การอ่านอักขระผิดเพี้ยน (Character Misread) เกิดจากการจำแนกพยัญชนะหรือสระที่มีรูปร่างคล้ายกันผิดพลาด ส่งผลให้คำสำคัญ (Keywords) เปลี่ยนรูปไป ทำให้ระบบจับคู่คำค้นหาได้ยากขึ้น และค่าคะแนนความคล้ายคลึง (Similarity) ลดต่ำลง

E2 การเว้นวรรคคลาดเคลื่อน (Spacing / Word Boundary Errors) ปัญหาการเว้นวรรคผิดตำแหน่งหรือวรรคหายไป ซึ่งส่งผลกระทบโดยตรงต่อการตัดคำ (Tokenization) ของโมเดลภาษา ทำให้กระบวนการแบ่งส่วนข้อมูล (Chunking) และการจับคู่ความหมายผิดเพี้ยนไปจากต้นฉบับ

E3 ความผิดพลาดของตัวเลขและวันที่ (Numeric Distortion) การอ่านตัวเลขหรือวันที่ผิดแม้อักขระเดียว (เช่น '2568' เป็น '2558' หรือ '140' เป็น '1A0') ถือเป็นความผิดพลาดร้ายแรงที่สุดในเอกสารประกาศ เนื่องจากทำให้ข้อเท็จจริงเปลี่ยนไปและนำไปสู่คำตอบที่เป็นเท็จทันที

E4 การสูญเสียโครงสร้างเอกสาร (Structure Loss) การที่โครงสร้างตารางหรือลำดับหัวข้อย่อยถูกแปลงเป็นข้อความเรียงต่อกันเป็นส่วนเดียวกัน ทำให้ความสัมพันธ์ทางความหมายระหว่าง หัวข้อ และ เนื้อหา ขาดหายไป ส่งผลให้ระบบดึงข้อมูลไปใช้ผิดบริบท

E5 ข้อความขยะและส่วนเกิน (Noisy / Short Chunks) ข้อความรบกวนที่ไม่เกี่ยวข้องกับเนื้อหาหลักของเอกสาร เช่น เลขหน้า หัวกระดาษ ท้ายกระดาษ หรืออักขระผิดปกติจาก OCR ซึ่งอาจส่งผลต่อความแม่นยำของกระบวนการค้นคืนข้อมูลและการสร้างคำตอบของระบบ

III.V มาตรการเพิ่มความน่าเชื่อถือของระบบ (Robustness Measures)

เพื่อให้ระบบ RAG สามารถทำงานได้แม้ต้องใช้งานกับเอกสารที่มีสัญญาณรบกวนจาก OCR ไว้ 3 ขั้นตอน ดังนี้

S1 การคัดกรองคุณภาพชิ้นส่วนข้อมูล (Noise-aware Chunk Validation) เป็นการป้องกันเพื่อสกัดกั้น "ข้อมูลขยะ" โดยระบบจะตรวจจับและคัดทิ้งชิ้นส่วนข้อมูล (Chunk) ที่มีความสั้นผิดปกติหรือมีรูปแบบที่ไม่สมบูรณ์ออกไปก่อนเข้าสู่กระบวนการสร้างดัชนี เพื่อป้องกันไม่ให้ข้อมูลลดคุณภาพถูกนำไปใช้ประมวลผล

S2 การทำความสะอาดบริบทข้อมูลภาษาไทย (Thai Contextual Cleaning) กระบวนการปรับปรุงข้อมูลให้เป็นมาตรฐานเดียวกัน (Normalization) ก่อนนำไปใช้งาน ได้มีการดำเนินการทดลองใช้กระบวนการปรับปรุงข้อความหลังทำ OCR (Post-OCR Correction) โดยการจัดการโครงสร้างภาษาไทยที่ส่งผลต่อโมเดล ได้แก่ (1) การแปลงเลขไทย (๐-๙) เป็นเลขอารบิก (0-9) ทั้งหมด และใช้ Regular Expression เพื่อแก้ไขรูปแบบตัวเลขหรือจุดทศนิยมที่อ่านผิดเพี้ยน (2) การจัดระเบียบการเว้นวรรคโดยเพิ่มช่องว่างระหว่างอักขระภาษาไทยและตัวเลขอารบิกหรือภาษาอังกฤษ และ (3) การจำกัดการขึ้นบรรทัดใหม่เพื่อไม่ให้โมเดลตัดคำผิดพลาด การทำความสะอาดข้อมูลในลักษณะนี้ช่วยลดปัญหาอักขระเพี้ยน (E1) และเว้นวรรคคลาดเคลื่อน (E2) ลดความสับสนของโมเดลภาษาในขั้นตอนการจับคู่และสร้างคำตอบได้

S3 กลไกกำกับการอ้างอิง (Citation Guardrails) การควบคุมในขั้นตอนการสร้างคำตอบ โดยใช้เทคนิคการเขียนคำสั่ง (Prompt Engineering) เพื่อบังคับให้โมเดลภาษาต้องตอบคำถามโดยอ้างอิงจากบริบทของข้อมูลที่ค้นมาได้เท่านั้น และกำหนดเงื่อนไขคำตอบเป็น หากไม่พบข้อมูล ให้ตอบ "ไม่มีข้อมูลในเอกสาร" วิธีนี้ช่วยป้องกันการสร้างคำตอบที่ผิดพลาด และช่วยให้แยกแยะได้ว่าที่ระบบไม่ตอบเป็นเพราะ "ไม่มีข้อมูลจริง" หรือเพราะ "ระบบทำงานผิดพลาด"

แนวทางดังกล่าวช่วยลดความเสี่ยงของ Hallucination และช่วยให้สามารถแยกแยะได้ว่า ความผิดพลาดเกิดจากการไม่มีข้อมูลจริงหรือเกิดจากความผิดพลาดของระบบ

IV. ผลการวิจัย

IV.I ผลการค้นคืนข้อมูล (Retrieval)

การทดสอบประสิทธิภาพการค้นคืนข้อมูลผ่านโมเดล BGE-M3 (50 queries) ที่ค่า K=10 แสดงผลลัพธ์ดังตารางที่ I

ตารางที่ I. แสดงผลการค้นคืนที่ K=10 (BGE-M3, 50 queries)

Source	P@10	Avg. Rel. Chunks	MRR	NDCG @10	Hit@10
OCR-text (c250, o50)	0.418	4.180	0.738	0.752	0.880
PDF-text (c500, o100)	0.192	1.920	0.710	0.747	0.900
TXT (c1000, o200)	0.110	1.100	0.837	0.899	0.940

หมายเหตุ: ค่า Avg. Rel. Chunks (Average Number of Relevant Chunks) ค่าเฉลี่ยของจำนวนชิ้นส่วนข้อมูลที่เกี่ยวข้องซึ่งถูกค้นคืนมาได้ในลำดับ Top-10 ต่อหนึ่งคำถาม

จากตารางพบว่าแหล่งข้อมูล OCR-text (c250) มีความหนาแน่นของการค้นคืนสูงสุด โดยมีค่า Precision@10 อยู่ที่ 0.418 (เฉลี่ยพบ 4.180 ชิ้นต่อคำถาม) สะท้อนว่าการดึงข้อมูลที่เกี่ยวข้องเข้าสู่ Top-K ได้มากที่สุด

หากพิจารณาจากคุณภาพการจัดอันดับ พบว่า TXT (c1000) กลับให้ค่า MRR (0.837) และ NDCG@10 (0.899) สูงกว่าอย่างชัดเจน สิ่งนี้เรียกว่า “Retrieval Paradox” แม้ OCR จะครอบคลุมเนื้อหาได้กว้าง แต่ข้อมูลที่ถูกต้องที่สุดมักไม่อยู่ในอันดับต้นๆ ส่งผลให้ LLM อาจนำข้อมูลที่มีคำที่คล้ายกัน แต่รายละเอียดผิดพลาดไปใช้สร้างคำตอบ

เพื่อแก้ปัญหาความคลาดเคลื่อนของข้อมูล ผู้วิจัยได้ประยุกต์ใช้กระบวนการจัดอันดับซ้ำ (Reranking) โดยเมื่อระบบค้นหาชิ้นส่วนข้อมูลเบื้องต้นที่มีความคล้ายคลึงสูงสุด (Initial Retrieval ที่ K=50) ระบบจะใช้โมเดล Cross-Encoder BAAI/bge-reranker-v2-m3 ทำการประเมินความสัมพันธ์เชิงลึกระหว่างคำถามและเนื้อหา เพื่อจัดอันดับใหม่และคัดเลือกเฉพาะหลักฐานที่มีคุณภาพสูงสุด (Top-5) ส่งไปยังโมเดลภาษา กลไกนี้ช่วยคัดกรองสัญญาณรบกวนและชิ้นส่วนข้อมูลขยะ (Garbage Chunks) ออกไป และจัดลำดับเนื้อหาที่สมบูรณ์ที่สุดขึ้นมาไว้ลำดับแรก ควบคู่กับการใช้กลไก Verification เพื่อตรวจสอบความถูกต้องของข้อมูลก่อนส่งต่อให้ LLM ซึ่งสอดคล้องกับผลการประเมินค่า MRR และ NDCG@10 ที่สะท้อนถึงความแม่นยำในการจัดอันดับข้อมูลที่เพิ่มสูงขึ้น

IV.II ผลการทดลองระดับ Generation: Accuracy Gap ภายใต้ Noise

ตารางที่ II แสดงผลการเปรียบเทียบความแม่นยำในการสร้างคำตอบโดยโมเดล Typhoon เมื่อใช้แหล่งข้อมูลที่แตกต่างกัน โดยใช้เกณฑ์ตัดสินความถูกต้องแบบยืดหยุ่น (Lax threshold = 0.6)

ตารางที่ II. ผลการสร้างคำตอบ (Typhoon, threshold=0.6, 50 queries)

Source	Chunk	Accuracy	F1 (Lax)	Avg. Similarity
Clean TXT	1000	0.860	0.925	0.621
PDF-text	500	0.560	0.718	0.477
OCR-text	250	0.700	0.824	0.523

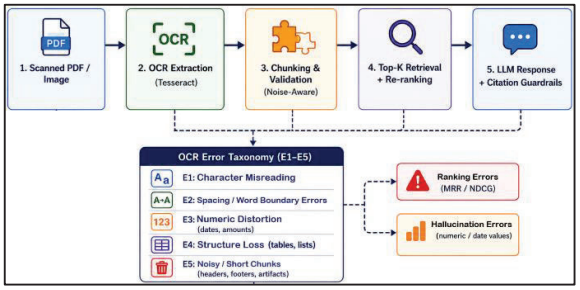
จากผลการทดลองพบว่า Clean TXT (c1000) ให้ค่า Accuracy สูงที่สุดที่ 0.860 ในขณะที่ OCR-text (c250) มีความแม่นยำลดลงเหลือ 0.700 และ PDF-text ต่ำสุดที่ 0.560

นอกจากนี้ เมื่อเปรียบเทียบกลยุทธ์การแบ่งส่วนข้อมูล (Chunk Settings) ที่ขนาด 250, 500 และ 1,000 โทเคน ภายใต้เงื่อนไขเดียวกัน พบว่า แม้ชุดข้อมูล OCR ที่ขนาด Chunk 250 จะให้ประสิทธิผลในขั้นการค้นคืน (Retrieval) สูงที่สุด แต่เมื่อเข้าสู่กระบวนการสร้างคำตอบ (Generation) ชุดข้อมูลที่ใช้ขนาด Chunk 1,000 กลับให้ประสิทธิภาพโดยรวมสูงสุด (F1-Score 0.925 ภายใต้เกณฑ์ Lax threshold 0.6) สาเหตุที่การตั้งค่า Chunk ขนาดใหญ่ (1,000 โทเคน) ให้ผลลัพธ์ดีที่สุดในขั้นสร้างคำตอบ เป็นเพราะชิ้นส่วนข้อมูลขนาดใหญ่ช่วยรักษาความสมบูรณ์ของบริบทแวดล้อม (Context) ไว้ได้มากกว่า ทำให้โมเดลภาษาสามารถชดเชยหรือคาดเดาความหมายของคำที่ถูก OCR อ่านผิดเพี้ยนไปได้ดีกว่าชิ้นส่วนขนาดเล็กที่มักจะถูกตัดขาดจากบริบท

ความแตกต่างของคะแนนนี้ สะท้อนให้เห็นว่า Noise ที่เกิดจากการแปลงไฟล์ของ ตัวอักษรเพี้ยนใน OCR หรือการจัดเรียงประโยคที่ผิดพลาดใน PDF ส่งผลโดยตรงต่อความสามารถของโมเดลภาษาในการจับใจความสำคัญ โดยเฉพาะในคำถามที่ต้องการความแม่นยำอย่าง ตัวเลขและวันที่ ซึ่งความผิดพลาดเพียงเล็กน้อยในข้อมูลอาจทำให้โมเดลสร้างคำตอบที่คลาดเคลื่อนไปทั้งหมด

IV.III การวิเคราะห์ความผิดพลาดเชิงคุณภาพ (Qualitative Error Analysis)

จากการวิเคราะห์ Log การตอบคำถามและการตรวจสอบความผิดพลาด พบว่าความผิดพลาดที่ส่งผลกระทบต่อระบบ RAG สำหรับเอกสารภาษาไทย มีลักษณะการแพร่กระจาย แสดงในรูปที่ II ดังนี้



รูปที่ II. ผังการไหลของความผิดพลาดจาก OCR ในระบบ RAG ภาษาไทย โดยแสดงการแพร่กระจายของประเภทความผิดพลาด (E1–E5)

1. ผลกระทบจากความคลาดเคลื่อนของตัวเลขและอักขระ (E3 & E1) ความผิดพลาดประเภท E3 (Numeric Distortion) ถือเป็นปัญหาที่สร้างผลกระทบสูงสุด เนื่องจากคำถามในเอกสารสถาบันมักเกี่ยวข้องกับ "ตัวเลขและวันที่" ซึ่งความคลาดเคลื่อนเพียง 1 หลักทำให้คำตอบกลายเป็นเท็จทันที โดยพบเป็นกรณีศึกษาได้ดังนี้

กรณีศึกษาที่ 1 (ค่าใช้จ่าย): ในคำถามเรื่อง "เงินช่วยเหลือเดินทาง (OT หลัง 22.00 น.)" ซึ่งคำตอบที่ถูกต้องคือ "140 บาท" แต่ OCR อ่านผิดเป็นอักขระผสมตัวเลข เป็น 1A0 หรืออ่าน "25,000" เป็น

"2S,000" ทำให้โมเดลภาษาไม่สามารถจับคู่ตัวเลขเพื่อนำไปคำนวณหรือตอบคำถามได้

กรณีศึกษาที่ 2 (วันที่มีผลบังคับใช้): การอ่านเลขปี พ.ศ. ผิดเพี้ยน ส่งผลให้คำตอบผิดทันทีเมื่อประเมินด้วยเกณฑ์ Exact Match (Strict)

นอกจากนี้ ความผิดพลาดประเภท E1 (Character Misread) ยังส่งผลโดยตรงต่อการจัดอันดับเอกสาร เช่น คำว่า "สถาบัน" ถูกอ่านเป็น "สกาบัน" ทำให้ค่าคะแนนความคล้ายคลึง (Similarity Score) ลดลงจนเอกสารที่ถูกต้องตกไปอยู่ในลำดับท้าย (Rank 15-20) ซึ่งหลุดจากบริบทที่ส่งให้โมเดลประมวลผล (Top-5) ก่อให้เกิดปรากฏการณ์ Retrieval Paradox ที่ระบบแจ้งว่าไม่พบข้อมูลทั้งที่เอกสารมีอยู่จริง

2. ผลกระทบจากการสูญเสียโครงสร้างเอกสาร (E4) ปัญหา E4 (Structure Loss) ทำให้บริบทสูญเสียความสัมพันธ์ดั้งเดิม เช่น การเว้นบรรทัดผิดตำแหน่ง การสูญเสียหัวข้อย่อย หรือการแยกตารางออกเป็นข้อความต่อเนื่อง ส่งผลให้ข้อมูลกระจัดกระจาย และเกิดสภาวะที่ข้อมูลถูกดึงขึ้นมามีความใกล้เคียงกันเยอะแต่เนื้อหาไม่ถูกต้อง (Context Crowding) ทำให้โมเดลภาษาจับใจความได้ยาก

IV.IV แนวทางการแก้ไขและข้อเสนอแนะ

จากผลการวิเคราะห์ความผิดพลาดข้างต้น ผู้วิจัยเสนอแนวทางปรับปรุงระบบ RAG ดังนี้

1. การจัดการความผิดพลาดเชิงตัวเลข (E3) ควรเพิ่มขั้นตอน การตรวจสอบความถูกต้อง (Verification) ก่อนส่งบริบทให้โมเดลภาษา โดยใช้กฎเกณฑ์ (Rule-based Validation) หรือ Regular Expression (Regex) เพื่อตรวจจับและแก้ไขรูปแบบตัวเลขที่ผิดปกติจาก OCR เช่น แก้ 'S' เป็น '5' หรือ 'O' เป็น '0' ในบริบทตัวเลข รวมถึงการยืนยันรูปแบบวันที่ให้ถูกต้องตามมาตรฐาน

2. การจัดการโครงสร้างเอกสาร (E4) เสนอให้เปลี่ยนจากการใช้ OCR พื้นฐาน เป็น Layout-aware OCR เช่น ThaiTrOCR หรือ LayoutLM ที่มีความสามารถในการรักษาตำแหน่งของตารางและลำดับย่อหน้าได้แม่นยำกว่า หรือใช้เทคนิคการแบ่งส่วนข้อมูลแบบอิงโครงสร้าง (Structure-aware Chunking) เพื่อลดปัญหาการกระจัดกระจายของข้อมูล

V. สรุปและอภิปรายผล

งานวิจัยนี้ได้ข้อสรุปสำคัญว่า ความน่าเชื่อถือของระบบ RAG สำหรับเอกสารภาษาไทยนั้น ขึ้นอยู่กับ “คุณภาพของข้อมูลในการจัดอันดับ (Ranking Quality)” มากกว่า “ปริมาณ” ของบริบทที่ค้นหาได้เพียงอย่างเดียว

จากการทดลองพบว่า แม้ข้อมูลจาก OCR จะช่วยให้ระบบค้นเจอเนื้อหาที่เกี่ยวข้องได้จำนวนมาก (ค่า P@10 สูง) แต่กลับพบว่าข้อมูลที่ถูกดึงขึ้นมาแม้จะมีความคล้ายคลึงกันสูงแต่เนื้อหากลับมีความคลาดเคลื่อน ส่งผลให้ประสิทธิภาพการจัดอันดับความถูกต้องลดลง (ค่า MRR ต่ำ) ทำให้บริบทที่ถูกส่งไปให้โมเดลภาษาประมวลผลเต็มไปด้วยสัญญาณรบกวน (Noise) ซึ่งเป็นสาเหตุหลักของความผิดพลาดในการตอบคำถาม โดยเฉพาะข้อเท็จจริงที่มีความละเอียดอ่อนอย่างตัวเลขและวันที่ (E3)

เพื่อให้ระบบสามารถนำไปใช้งานได้จริงในสถาบันการศึกษา ผู้วิจัยขอเสนอแนวทางในอนาคต ดังนี้

1. ใช้การคัดกรองคุณภาพข้อมูล (Noise-aware Validation) เพื่อตัดข้อมูลที่ผิดปกติออก และใช้ กลไกกำกับการอ้างอิง (Citation Guardrails) เพื่อบังคับให้ AI ตรวจสอบข้อมูลก่อนก่อนตอบเสมอ

2. แนวทางการพัฒนาต่อยอด

(1) การจัดการโครงสร้างโดยการนำเทคโนโลยี Layout-aware OCR มาใช้เพื่อรักษาโครงสร้างตารางและรายการหัวข้อ

(2) การเพิ่มขั้นตอน การจัดอันดับซ้ำ (Reranking) และ การตรวจสอบความถูกต้อง (Verification) เพื่อยืนยันข้อเท็จจริงก่อนตอบ

(3) การประเมินผลระบบภายใต้ระดับสัญญาณรบกวนของ OCR ที่แตกต่างกันในสถานการณ์จริงให้มากขึ้น

บรรณานุกรม

[1] P. Lewis *et al.*, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Apr. 12, 2021, *arXiv*: arXiv:2005.11401. doi: 10.48550/arXiv.2005.11401.

[2] “Announcement Systems.” Accessed: Dec. 03, 2025. [Online]. Available: https://announcement.tni.ac.th/public/homepage

[3] Y. Gao *et al.*, “Retrieval-Augmented Generation for Large Language Models: A Survey,” Mar. 27, 2024, *arXiv*: arXiv:2312.10997. doi: 10.48550/arXiv.2312.10997.

[4] R. Friel, M. Belyi, and A. Sanyal, “RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems,” Jan. 16, 2025, *arXiv*: arXiv:2407.11005. doi: 10.48550/arXiv.2407.11005.

[5] T. A. Chang *et al.*, “Detecting Hallucination and Coverage Errors in Retrieval Augmented Generation for Controversial Topics,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., Torino, Italia: ELRA and ICCL, May 2024, pp. 4729–4743. Accessed: Feb. 14, 2026. [Online]. Available: https://aclanthology.org/2024.lrec-main.423/

[6] S. Ozaki *et al.*, “Understanding the Impact of Confidence in Retrieval Augmented Generation: A Case Study in the Medical Domain,” in *Proceedings of the 24th Workshop on Biomedical Language Processing*, D. Demner-Fushman, S. Ananiadou, M. Miwa, and J. Tsujii, Eds., Viena, Austria: Association for Computational Linguistics, Aug. 2025, pp. 1–17. doi: 10.18653/v1/2025.bionlp-1.1.

[7] J. Kanerva, C. Ledins, S. Käpyaho, and F. Ginter, “OCR Error Post-Correction with LLMs in Historical Documents: No Free Lunches,” in *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, Š. A. Holdt, N. Ilinykh, B. Scalvini, M. Bruton, I. N. Debess, and C. M. Tudor, Eds., Tallinn, Estonia: University of Tartu Library, Estonia, Mar. 2025, pp. 38–47. Accessed: Feb. 14, 2026. [Online]. Available: https://aclanthology.org/2025.resourceful-1.8/

[8] R. Smith, “An Overview of the Tesseract OCR Engine,” in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*, Curitiba, Parana, Brazil: IEEE, Sep. 2007, pp. 629–633. doi: 10.1109/ICDAR.2007.4376991.

[9] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, “BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation,” Jun. 28, 2024, *arXiv*: arXiv:2402.03216. doi: 10.48550/arXiv.2402.03216.

[10] K. Pipatanakul *et al.*, “Typhoon 2: A Family of Open Text and Multimodal Thai Large Language Models,” Dec. 19, 2024, *arXiv*: arXiv:2412.13702. doi: 10.48550/arXiv.2412.13702.