การพัฒนารูปแบบการฝึก AI โดยใช้ Large Language Models (LLM) เพื่อเกษียนหนังสือราชการ กรณีศึกษา สถาบันนวัตกรรมการเรียนรู้ มหาวิทยาลัยพะเยา

ญภคกร ไชยมงคล
ปริญาพร สันตะจิตต์

สถาบันนวัตกรรมการเรียนรู้
มหาวิทยาลัย 2568
ได้รับทุนอุดหนุนงานวิจัยจากมหาวิทยาลัยพะเยา
ประจำปีงบประมาณ พ.ศ. 2568

การพัฒนารูปแบบการฝึก AI โดยใช้ Large Language Models (LLM) เพื่อเกษียนหนังสือราชการ กรณีศึกษา สถาบันนวัตกรรมการเรียนรู้ มหาวิทยาลัยพะเยา

ญภคกร ไชยมงคล
ปริญาพร สันตะจิตต์

สถาบันนวัตกรรมการเรียนรู้
มหาวิทยาลัย 2568
ได้รับทุนอุดหนุนงานวิจัยจากมหาวิทยาลัยพะเยา
ประจำปีงบประมาณ พ.ศ. 2568

อาจารย์ที่ปรึกษาได้พิจารณาการทำวิจัยเพื่อพัฒนางานประจำ เรื่อง การพัฒนารูปแบบการฝึก AI โดยใช้ Large Language Models (LLM) เพื่อเกษียนหนังสือราชการ กรณีศึกษาสถาบันนวัตกรรมการเรียนรู้ มหาวิทยาลัยพะเยา ผู้วิจัยโดย นางสาวญภคกร ไชยมงคล และนางสาวปริญาพร สันตะจิตต์ ฉบับนี้เป็นที่เรียบร้อยแล้วและเห็นสมควรรับเป็นส่วนหนึ่งของการทำวิจัยเพื่อการพัฒนางานประจำของมหาวิทยาลัยพะเยา

.....

(ผู้ช่วยศาสตราจารย์ ดร.นรินธน์ นนทมาลย์)

อาจารย์ที่ปรึกษา

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณเป็นอย่างสูงในความกรุณาของผู้ช่วยศาสตราจารย์ ดร.นรินธน์ นนทมาลย์ อาจารย์ที่ปรึกษาวิจัย และขอขอบคุณผู้ช่วยศาสตราจารย์พชรธัช ไชยมงคล ที่ได้สละเวลาอันมีค่ามาเป็นที่ปรึกษา พร้อมทั้งให้คำแนะนำตลอดระยะเวลาในการทำวิจัยฉบับนี้ อีกทั้งขอขอบคุณผู้ทรงวุฒิตรวจสอบความตรงตามเนื้อหาของครื่องมือการวิจัย อันประกอบไปด้วย รองศาสตราจารย์ ดร.สาคร เมฆรักษาวนิช ดร.คงอมร เหมรัตน์รักษ์และนางณัฐธิดา ชาวน่าน ที่ได้ให้คำแนะนำในการออกแบบแบบสอบถาม ตลอดจนแก้ไขข้อบกพร่องของวิจัยด้วยความเอาใจใส่ จนทำให้วิจัยฉบับนี้สำเร็จลุล่วงไปได้อย่างสมบูรณ์

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.น้ำเงิน จันทรมณี ผู้อำนวยการสถาบันนวัตกรรมการเรียนรู้และนางนงลักษณ์ คล้ายอ่ำ หัวหน้างานบริหารทั่วไป ที่ส่งเสริมการพัฒนางานประจำสู่งานวิจัย และกรุณาให้คำแนะนำและข้อคิดเห็นที่เป็นประโยชน์ยิ่งต่อการดำเนินการวิจัย ส่งผลให้รายงานวิจัยนี้มีความสมบูรณ์ทางวิชาการมากยิ่งขึ้น

ขอขอบคุณ เจ้าหน้าที่ และบุคลากร ของสถาบันนวัตกรรมการเรียนรู้ มหาวิทยาลัยพะเยาทุกท่าน ที่ให้ความช่วยเหลือ และอำนวยความสะดวกในการดำเนินงานวิจัย

ขอขอบคุณ คุณพ่อคุณแม่และครอบครัว ญาติพี่น้อง เพื่อนๆ พี่ๆ น้องๆที่ให้กำลังใจกับผู้วิจัย ส่งผลให้ผู้วิจัยมีความตั้งใจ มุ่งมั่นในการทำวิจัยจนแล้วเสร็จ

ท้ายสุด หากมีสิ่งขาดตกบกพร่องหรือผิดพลาดประการใด ข้าพเจ้าขออภัยเป็นอย่างสูงในข้อบกพร่องและข้อผิดพลาดนั้น และข้าพเจ้าหวังว่างานวิจัยนี้คงมีประโยชน์ไม่มากก็น้อยสำหรับผู้ต้องการศึกษาต่อไป

ญภคกร ไชยมงคล

ปริญาพร สันตะจิตต์

กรกฎาคม 2568

เรื่อง การพัฒนารูปแบบการฝึก AI โดยใช้ Large Language Models (LLM) เพื่อเกษียนหนังสือราชการ กรณีศึกษา สถาบันนวัตกรรมการเรียนรู้ มหาวิทยาลัยพะเยา
ผู้วิจัย ญภคกร ไชยมงคล, ปริญาพร สันตะจิตต์
ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร.นรินธน์ นนทมาลย์
คำสำคัญ แบบจำลองภาษาขนาดใหญ่, ปัญญาประดิษฐ์, เกษียนหนังสือราชการ, Retrieval-Augmented

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนารูปแบบการฝึกปัญญาประดิษฐ์ (AI) โดยใช้แบบจำลองภาษาขนาดใหญ่ (Large Language Models: LLM) และเพื่อศึกษาผลการนำ AI มาใช้เพิ่มประสิทธิภาพในกระบวนการเกษียนหนังสือราชการ กรณีศึกษา สถาบันนวัตกรรมการเรียนรู้ มหาวิทยาลัยพะเยา ระเบียบวิธีวิจัยเป็นการพัฒนาระบบต้นแบบ AI ที่ทำงานบนเซิร์ฟเวอร์ภายในองค์กร (On-premise) เพื่อรักษาความปลอดภัยของข้อมูล โดยใช้แบบจำลองภาษา Llama 3 ผ่านกรอบการทำงาน Ollama ร่วมกับสถาปัตยกรรม Retrieval-Augmented Generation (RAG) ซึ่งดึงข้อมูลจากคลังความรู้ที่สร้างจากเอกสารราชการของสถาบันฯ การประเมินประสิทธิภาพดำเนินการใน 3 มิติ ได้แก่ (1) การประเมินเชิงปริมาณด้วยตัวชี้วัดมาตรฐาน เช่น F1-Score, BLEU, และ ROUGE, (2) การประเมินโดยผู้เชี่ยวชาญ 3 ท่าน เพื่อตรวจสอบความเที่ยงตรงของเครื่องมือวิจัย (IOC) และประเมินประสิทธิภาพของระบบ, และ (3) การประเมินความพึงพอใจโดยกลุ่มผู้ใช้งานจริง ซึ่งเป็นบุคลากรของสถาบันฯ จำนวน 40 คน

ผลการวิจัยพบว่า สามารถพัฒนาระบบต้นแบบ AI ได้สำเร็จตามรูปแบบที่วางแผนไว้ ผลการประเมินเชิงปริมาณแสดงให้เห็นถึงประสิทธิภาพในระดับสูง โดยมีคะแนน F1-Score สำหรับการตอบคำถามเชิงข้อเท็จจริงอยู่ที่ 0.85 และคะแนน BLEU สำหรับการร่างข้อความอยู่ที่ 0.60 ด้านการประเมินโดยผู้เชี่ยวชาญพบว่าเครื่องมือวิจัยมีความน่าเชื่อถือ (ค่า IOC เฉลี่ย 0.82) และระบบต้นแบบมีประสิทธิภาพโดยรวมในระดับมากที่สุด (ค่าเฉลี่ย 4.50 จาก 5) โดยเฉพาะอย่างยิ่งในด้านศักยภาพการนำไปประยุกต์ใช้จริง ขณะที่ผลการประเมินจากผู้ใช้งานจริงมีความพึงพอใจโดยรวมในระดับมาก (ค่าเฉลี่ย 3.88 จาก 5) โดยพึงพอใจสูงสุดในประเด็นที่ระบบช่วยเพิ่มผลงานและลดต้นทุนกำลังคน (ค่าเฉลี่ย 4.00) โดยสรุป รูปแบบการฝึก AI ที่พัฒนาขึ้นสามารถนำมาใช้ในกระบวนการเกษียนหนังสือราชการได้อย่างมีประสิทธิภาพ ช่วยลดระยะเวลาและข้อผิดพลาด เพิ่มประสิทธิผลในการทำงาน และได้รับการยอมรับในเชิงบวกจากทั้งผู้เชี่ยวชาญและผู้ใช้งานจริง ซึ่งชี้ให้เห็นถึงศักยภาพในการนำไปพัฒนาต่อยอดและขยายผลในหน่วยงานอื่นต่อไป

Title: The Development of AI training model using Large Language Models (LLM) for civil service retirement: Case study of the Institute of Learning Innovation University of Phayao.

Author: Yaphakhakorn Chaimongkon, Pariyaporn Santajit

Advisor: Assistant Professor Narin Nonthamand, Ph.D

Keywords: Large Language Models, Artificial Intelligence, Official Document Processing, Retrieval-Augmented Generation

Abstract

This research aims to develop an Artificial Intelligence (AI) training model using Large Language Models (LLM) and to study the effects of its implementation on enhancing the efficiency of the official document processing workflow, based on a case study at the Institute of Learning Innovation, University of Phayao. The research methodology involved the development of a prototype AI system designed to operate on an on-premise server to ensure data security. The system utilizes the Llama 3 language model via the Ollama framework, integrated with a Retrieval-Augmented Generation (RAG) architecture that draws from a knowledge base created from the institute's official documents. The system's performance was evaluated across three dimensions: (1) a quantitative assessment using standard metrics such as F1-Score, BLEU, and ROUGE; (2) an evaluation by three experts to validate the research instruments (using the Index of Item-Objective Congruence: IOC) and assess system performance; and (3) a satisfaction survey conducted with a target group of 40 staff members from the institute.

The research findings indicate that the AI prototype was successfully developed according to the planned model. The quantitative evaluation demonstrated high performance, achieving an F1-Score of 0.85 for fact-based question answering and a BLEU score of 0.60 for text drafting. The expert evaluation confirmed the reliability of the research instruments (average IOC of 0.82) and rated the prototype's overall performance at the "highest level" (average score of 4.50 out of 5), particularly highlighting its potential for practical application. Furthermore, the evaluation from end-users revealed a "high level" of overall satisfaction (average score of 3.88 out of 5), with the highest satisfaction reported for the system's ability to increase productivity and reduce personnel costs (average score of 4.00). In conclusion, the developed AI training model can be effectively applied to the official document processing workflow. It successfully reduces time and errors, increases operational effectiveness, and has received positive acceptance from both experts and end-users, indicating significant potential for further development and broader implementation in other departments.

สารบัญ

บทที่ **หน้า**

1 บทนำ 1

ความเป็นมาและความสำคัญของปัญหา 1

วัตถุประสงค์ของการวิจัย 3

ขอบเขตของการวิจัย 3

นิยามศัพท์ที่ใช้ในการศึกษา 3

ประโยชน์ที่คาดว่าจะได้รับ 4

2 ทบทวนวรรณกรรม 5

แนวความคิดหรือทฤษฎีที่เกี่ยวข้องที่ใช้ในการวิจัย 5

เอกสารและงานวิจัยที่เกี่ยวข้อง 22

กรอบแนวคิดในการวิจัย 23

3 วิธีดำเนินการวิจัย 27

ประชากรและกลุ่มตัวอย่าง 27

เครื่องมือที่ใช้ในการวิจัย 29

วิธีดำเนินการวิจัย 30

การหาค่าดัชนีความสอดคล้อง IOC 33

การเก็บรวบรวมข้อมูล 36

การวิเคราะห์ข้อมูล 38

ด้านจริยธรรมการวิจัย 39

ระยะเวลาที่จะทำการวิจัยและการบริหารจัดการ 40

4 ผลการดำเนินงานและวิเคราะห์ข้อมูล 42

ผลการพัฒนาระบบต้นแบบ AI เพื่อการเกษียนหนังสือราชการ 42

ผลการประเมินประสิทธิภาพของระบบ 44

สารบัญ (ต่อ)

บทที่	หน้า
5 การสรุป อภิปรายผล และข้อเสนอแนะ	50
สรุปผลการวิจัย	50
อภิปรายผล	51
ข้อเสนอแนะ	52
บรรณานุกรม	54
ประวัติผู้วิจัย	65

สารบัญตาราง

ตาราง **หน้า**

4.1 ผลการประเมินประสิทธิภาพของระบบด้วยตัวชี้วัดเชิงปริมาณ 46

4.2 สรุปผลการประเมินประสิทธิภาพระบบ AI โดยผู้เชี่ยวชาญ..... 47

4.3 สรุปผลการประเมินความพึงพอใจของผู้ใช้งาน (N=40)..... 48

สารบัญภาพ

ภาพ **หน้า**

1 แสดงกระบวนการสองขั้นตอนในการสร้างโมเดลภาษาขนาดใหญ่ (Gomes,2022)..... 12

2 โครงสร้างของโมเดลภาษาขนาดใหญ่ที่ปรับแต่งด้วยอะแดปเตอร์ (Raschka, 2023)..... 12

3 แสดงกระบวนการทำงานของการรู้จำอักขระด้วยแสง (Optical Character Recognition (OCR)) (V7 Labs (n.d.)) 15

4 ขั้นตอนการทำงานของ RAG: การรับคำถาม ค้นหา ดึงข้อมูล และสร้างคำตอบ (Khalid, 2024)..... 24

5 กระบวนการทำงานของระบบ Multimodal PDF Data Extraction สำหรับ Enterprise RAG (NVIDIA, n.d.)..... 25

6 แสดงภาพโค้ด..... 42

7 แสดงโปสเตอร์เชิญชวนให้ผู้เชี่ยวชาญระบบประเมิน 44

8 แสดงโปสเตอร์เชิญชวนให้ผู้ใช้ระบบประเมิน..... 45

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

มหาวิทยาลัยพะเยา ได้มีนโยบายใช้ระบบบริหารจัดการเอกสารของมหาวิทยาลัย (UP-DMS) ผ่านระบบเครือข่ายคอมพิวเตอร์เพื่อให้เกิดประสิทธิภาพ ลดการใช้กระดาษ และขั้นตอนการทำงานด้านเอกสารในการจัดทำงานสารบรรณ

การบริหารจัดการเอกสารผ่านระบบบริหารจัดการเอกสารมหาวิทยาลัยพะเยาผ่านระบบ UP-DMS ประกอบไปด้วย การรับ-ส่ง หนังสือ การเกษียนหนังสือราชการ การเสนอหนังสือต่อผู้บังคับบัญชา ซึ่งรูปแบบการเกษียนหนังสือราชการ จะต้องดำเนินการคำนึงถึงเรื่องความถูกต้องของข้อมูล ไม่ตัดข้อความหรือเพิ่มข้อมูลอื่นจนทำให้เข้าใจเบี่ยงเบนไปจากต้นฉบับ ซึ่งการเกษียณหนังสือเกิดการอ่านวิเคราะห์เอกสาร เพื่อพิมพ์ข้อความสรุปความจากเอกสารจากเจ้าหน้าที่ธุรการ อาจมีความเสี่ยงเกิดการพิมพ์ข้อความผิดพลาด ส่งผลให้เกิดการพิจารณาวิเคราะห์เอกสารต่อฉบับใช้ระยะเวลานาน มีการสรุปเนื้อหาที่ไม่ครบถ้วน จากการใช้ภาษาที่ไม่กระชับ

ปัจจุบันมีการนำเทคโนโลยีปัญญาประดิษฐ์การประมวลผลข้อความที่ใกล้เคียงกับภาษามนุษย์อย่างมากคือ Large Language Models (LLMs) มาใช้ในการอำนวยความสะดวกในการดำรงชีวิตประจำประวันของมนุษย์มากขึ้น ยกตัวอย่างเช่น GPT-4 จาก OpenAI, BERT จาก Google และ LLaMA จาก Meta ซึ่งถูกนำไปใช้ในหลากหลายอุตสาหกรรม อาทิการแพทย์ การศึกษา หรือธุรกิจ LLMs ช่วยเพิ่มประสิทธิภาพในการสร้างเนื้อหาที่ซับซ้อน การตอบคำถามอย่างรวดเร็วและแม่นยำ รวมถึงการวิเคราะห์ข้อมูลเชิงลึก ทำให้การตัดสินใจและการแก้ปัญหาเป็นไปได้อย่างมีประสิทธิภาพและรวดเร็วยิ่งขึ้น นอกจากนี้ LLMs ยังสนับสนุนการพัฒนานวัตกรรมช่วยลดเวลาที่ใช้ในกระบวนการทำงานที่ซับซ้อนและสร้างผลลัพธ์ที่เป็นประโยชน์สำหรับผู้ใช้งานในระดับองค์กรและส่วนบุคคล (Naveed et al., 2023; Chang et al., 2023; Fan et al., 2023)

การใช้ Retrieval-Augmented Generation (RAG) ร่วมกับ LLMs โดย RAG ผสมผสานการดึงข้อมูลจากฐานความรู้ภายนอกกับการสร้างเนื้อหาด้วย LLMs เพื่อปรับปรุงความแม่นยำและความน่าเชื่อถือในการประมวลผลภาษา กระบวนการของ RAG เริ่มจากการดึงข้อมูลที่เกี่ยวข้องจากแหล่งต่าง ๆ เช่น ไฟล์ PDF หรือเอกสารภายในองค์กร แล้วนำข้อมูลเหล่านี้มาใช้ร่วมกับ LLM เพื่อสร้างคำตอบที่สอดคล้องกับข้อมูลต้นฉบับ การใช้ RAG สำหรับการฝึก LLMs บนเอกสาร PDF

ช่วยให้สามารถดึงข้อมูลเฉพาะทางได้รวดเร็ว ลดข้อผิดพลาดที่เกิดจากการคาดเดา (hallucination) และเพิ่มประสิทธิภาพในการตอบคำถามที่ซับซ้อนหรือเฉพาะเจาะจงได้อย่างมีนัยสำคัญ เช่น ในสาขาการแพทย์และการพัฒนาซอฟต์แวร์ (Krishna, 2023; Ge et al., 2023) นอกจากนี้ การใช้ RAG ยังช่วยลดต้นทุนในการฝึก LLM โดยใช้ข้อมูลสังเคราะห์จากเอกสารโดยตรง ซึ่งช่วยประหยัดทรัพยากรและลดเวลาในการพัฒนา เช่น งานวิจัยที่พัฒนาโมเดล PGT ซึ่งทำให้โมเดลขนาดเล็กสามารถประมวลผลข้อมูลเฉพาะทางจากไฟล์ PDF ได้โดยไม่ต้องพึ่งพาข้อมูลที่ทำเครื่องหมาย (annotated data) (Krishna, 2023) อีกทั้งยังช่วยในการสร้างโมเดลเฉพาะทาง เช่น "LiVersa" ที่สามารถดึงข้อมูลการรักษาโรคตับจากไฟล์ PDF และตอบคำถามทางคลินิกได้อย่างแม่นยำ แสดงให้เห็นถึงศักยภาพในการนำ LLM มาใช้ในงานที่ต้องการความเฉพาะทางสูง (Ge et al., 2023)

ปัจจุบัน มี LLMs ที่ได้รับความนิยมและใช้งานอย่างแพร่หลายจากผู้ให้บริการชั้นนำหลายราย เช่น GPT-4 จาก OpenAI (OpenAI, n.d.), Azure OpenAI Service จาก Microsoft (Microsoft, n.d.) และ Vertex AI จาก Google Cloud (Google Cloud, n.d.) แม้ว่าโมเดลเหล่านี้จะมีความสามารถสูงในการประมวลผลภาษาและสร้างข้อความที่ซับซ้อน แต่การใช้งานมักมีค่าใช้จ่ายและข้อจำกัดด้านความปลอดภัยของข้อมูล เนื่องจากข้อมูลที่ป้อนเข้าสู่ระบบอาจถูกนำไปใช้ในการฝึกโมเดลเพิ่มเติม ซึ่งอาจไม่เหมาะสำหรับองค์กรที่ต้องการรักษาความลับของข้อมูล Ollama รุ่น 3.2 ซึ่งพัฒนาโดย Meta เป็นโมเดล LLM ที่สามารถดาวน์โหลดและใช้งานบนเครื่องคอมพิวเตอร์ส่วนบุคคลได้ โดยผู้ใช้สามารถปรับแต่งโมเดลให้เหมาะสมกับความต้องการเฉพาะของตนเอง อีกทั้งข้อมูลที่ป้อนเข้าสู่ระบบจะถูกประมวลผลภายในเครื่องไม่ถูกส่งออกไปยังเซิร์ฟเวอร์ภายนอก จึงเหมาะสำหรับองค์กรที่ต้องการรักษาความลับของข้อมูล นอกจากนี้ Ollama ยังเป็นซอฟต์แวร์ที่เปิดให้ใช้งานฟรี (Meta, n.d.)

การใช้เทคนิค RAG ร่วมกับ LLMs อย่าง Ollama 3.2 สามารถช่วยลดระยะเวลาในการสรุปความในกระบวนการเกษียณหนังสือราชการได้อย่างมีประสิทธิภาพ โดย RAG ดึงข้อมูลจากเอกสาร PDF ที่เกี่ยวข้องมาใช้ในการฝึกฝน ทำให้โมเดลสามารถสรุปเนื้อหาและสร้างข้อความที่สอดคล้องกับข้อมูลต้นฉบับได้อย่างแม่นยำ การประยุกต์ใช้ RAG ในการเกษียณหนังสือราชการจึงเป็นทางเลือกที่น่าสนใจและมีศักยภาพในการเพิ่มประสิทธิภาพและความถูกต้อง

งานวิจัยนี้มีวัตถุประสงค์ในการพัฒนารูปแบบการฝึก AI โดยใช้ Large Language Models (LLM) รวมถึงศึกษาผลการใช้ AI ในการเกษียนหนังสือราชการ กรณีศึกษา สถาบันนวัตกรรมการเรียนรู้ ให้มีประสิทธิภาพใช้เวลาได้รวดเร็วขึ้น

1.2 วัตถุประสงค์ของการวิจัย

1) เพื่อพัฒนารูปแบบการฝึก AI โดยใช้ Large Language Models (LLM) เพื่อเกษียนหนังสือราชการ กรณีศึกษา สถาบันนวัตกรรมการเรียนรู้

2) เพื่อศึกษาผลการใช้ AI ในการเกษียนหนังสือราชการ กรณีศึกษา สถาบันนวัตกรรมการเรียนรู้ ให้มีประสิทธิภาพใช้เวลาได้รวดเร็วขึ้น

1.3 ขอบเขตของการวิจัย

1) การรวบรวมไฟล์ PDF ที่เกี่ยวข้องจากมหาวิทยาลัยพะเยา กรณีศึกษา สถาบันนวัตกรรมการเรียนรู้ มหาวิทยาลัยพะเยา เช่น หนังสือราชการ เอกสารภายใน ในกระบวนการลงรับเอกสารและแปลงข้อมูลเหล่านี้เป็นข้อความดิบด้วย OCR หรือเครื่องมืออ่าน PDF

2) การสร้างฐานข้อมูลแบบเวกเตอร์โดยใช้โมเดล embedding เช่น OpenAI's ADA หรือ Sentence Transformers และจัดเก็บในฐานข้อมูลที่รองรับการค้นคืน เช่น Pinecone หรือ Weaviate เพื่อใช้ร่วมกับ Ollama 3.2

3) การประเมินประสิทธิภาพของโมเดลโดยใช้ตัวชี้วัด เช่น Exact Match (EM), F1–Score, BLEU, และ ROUGE รวมถึงการตรวจสอบความแม่นยำของโมเดล และการประเมินโดยผู้เชี่ยวชาญในโดเมนที่เกี่ยวข้อง

1.4 นิยามศัพท์ที่ใช้ในการศึกษา

รูปแบบการฝึก AI

รูปแบบการฝึก AI คือ กระบวนการในการนำโมเดลปัญญาประดิษฐ์ผ่านขั้นตอนที่ทำให้เกิดการเรียนรู้ และปรับปรุงกระบวนการอย่างต่อเนื่องเพื่อให้โมเดล AI สามารถเรียนรู้และตอบสนองได้

Large Language Models (LLMs)

Large Language Models (LLMs) คือ เทคโนโลยีปัญญาประดิษฐ์ที่มีความก้าวหน้าและทรงพลังในการประมวลผลภาษาธรรมชาติ ทำให้สามารถสร้างและเข้าใจข้อความที่ใกล้เคียงกับภาษามนุษย์ (Vaswani et al., 2017)

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1) ได้รูปแบบการฝึก AI โดยใช้ Large Language Models (LLM) เพื่อเกษียนหนังสือราชการ กรณีศึกษา สถาบันนวัตกรรมการเรียนรู้ ซึ่งบุคคลทั่วไปสามารถนำรูปแบบการฝึก AI ไปพัฒนาต่อยอดการเกษียณหนังสือราชการในหน่วยงานของตนเองได้

2) ได้ผลการใช้ AI ในการเกษียนหนังสือราชการ กรณีศึกษา สถาบันนวัตกรรมการเรียนรู้ที่ทำให้การทำงานสารบรรณมีประสิทธิภาพใช้เวลารวดเร็วขึ้น

มหาวิทยาลัยพะเยา UNIVERSITY OF PHAYAO

บทที่ 2

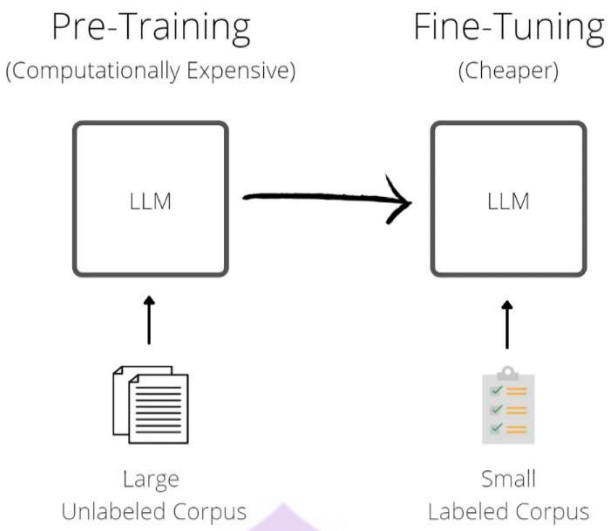
ทบทวนวรรณกรรม

2.1 แนวความคิดหรือทฤษฎีที่เกี่ยวข้องที่ใช้ในการวิจัย (Concepts and Theories)

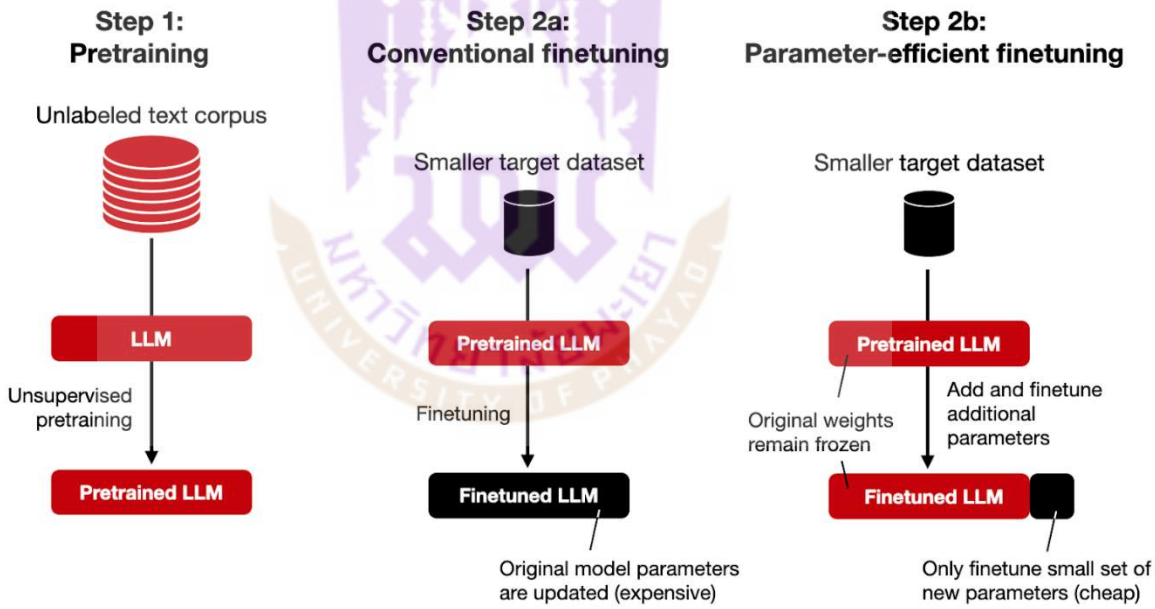
2.1.1 ทฤษฎีความรู้เกี่ยวกับ Large Language Models (LLMs)

Large Language Models (LLMs) เป็นเทคโนโลยีปัญญาประดิษฐ์ที่มีความก้าวหน้าในการประมวลผลภาษาธรรมชาติ ทำให้สามารถสร้างและเข้าใจข้อความที่ใกล้เคียงกับภาษามนุษย์ สถาปัตยกรรมพื้นฐานของ LLMs พัฒนามาจากสถาปัตยกรรม Transformer ซึ่งถูกนำเสนอครั้งแรกโดย Vaswani และคณะในปี 2017 ในงานวิจัยชื่อ Attention is All You Need สถาปัตยกรรม Transformer มีองค์ประกอบหลักสองส่วนคือ Encoder และ Decoder โดยใน LLMs มักจะเน้นใช้ส่วน Decoder เพื่อสร้างข้อความที่สอดคล้องกับบริบท (Vaswani et al., 2017) หัวใจสำคัญของ Transformer คือ กลไก Attention ซึ่งช่วยให้โมเดลสามารถให้ความสำคัญกับคำหรือบริบทที่สำคัญในประโยค กลไกนี้ประกอบด้วย Self-Attention ที่ใช้ในการวิเคราะห์ความสัมพันธ์ระหว่างคำภายในประโยคเดียวกัน และ Multi-Head Attention ที่ช่วยให้โมเดลจับความสัมพันธ์ที่ซับซ้อนได้จากหลายมุมมองพร้อมกัน นอกจากนี้ โมเดลยังใช้ Feedforward Neural Network (FFNN) เพื่อประมวลผลข้อมูลหลังจากผ่านกลไก Attention พร้อมด้วย Layer Normalization และ Residual Connection ที่ช่วยปรับปรุงเสถียรภาพของโมเดลและเพิ่มความลึกโดยไม่สูญเสียความสามารถในการเรียนรู้ (Devlin et al., 2019) ด้วยสถาปัตยกรรมเหล่านี้ LLMs สามารถวิเคราะห์และสร้างข้อความที่มีความซับซ้อนและต่อเนื่องได้อย่างน่าประทับใจ

กระบวนการเรียนรู้ของ LLMs แบ่งออกเป็นสามขั้นตอนหลัก ได้แก่ Pre-training, Fine-tuning, และ Self-supervised Learning ในขั้นตอน Pre-training โมเดลถูกฝึกด้วยข้อมูลจำนวนมหาศาล เช่น หนังสือ เว็บไซต์ และฐานข้อมูลสาธารณะ เพื่อเรียนรู้รูปแบบของภาษาและความสัมพันธ์ระหว่างคำ โดยใช้เทคนิคเช่น Masked Language Modeling (MLM) ที่ซ่อนคำบางคำเพื่อให้โมเดลเดา หรือ Causal Language Modeling (CLM) ที่เดาคำถัดไปในลำดับประโยค (Devlin et al., 2019)



ภาพที่ 1 แสดงกระบวนการสองขั้นตอนในการสร้างโมเดลภาษาขนาดใหญ่ (Gomes, 2022)



ภาพที่ 2 โครงสร้างของโมเดลภาษาขนาดใหญ่ที่ปรับแต่งด้วยอะแดปเตอร์ (Raschka, 2023)

หมายเหตุ. I am proper songจาก "การปรับแต่ง LLMs อย่างมีประสิทธิภาพด้วยอะแดปเตอร์" โดย S. Raschka, 2023, Ahead of AI Magazine (https://magazine.sebastianraschka.com/p/finetuning-llms-with-adapters). ลิขสิทธิ์ 2023 โดย Sebastian Raschka

ขั้นตอน Fine-tuning เป็นการปรับแต่งโมเดลให้เหมาะสมกับงานเฉพาะ เช่น การแปลภาษา การตอบคำถาม หรือการวิเคราะห์ความคิดเห็น โดยใช้ชุดข้อมูลขนาดเล็กแต่มีคุณภาพสูง เพื่อเพิ่มความแม่นยำและประสิทธิภาพในงานที่ต้องการ (Floridi & Chiriatti, 2020) ส่วน Self-supervised Learning เป็นแนวคิดที่เน้นการเรียนรู้โดยไม่ต้องพึ่งพาฉลากข้อมูล (labels) โมเดลจะสร้างโจทย์จากข้อมูลดิบ เช่น การเติมคำหรือการจัดเรียงประโยคใหม่ เพื่อให้เข้าใจโครงสร้างของภาษาได้ดียิ่งขึ้น (Vaswani et al., 2017) LLMs มีความสามารถหลากหลายที่ส่งผลกระทบต่อหลายด้าน การสร้างข้อความ เป็นหนึ่งในความสามารถที่โดดเด่น โมเดลสามารถสร้างข้อความที่สมจริงและสอดคล้องกับบริบท เช่น การเขียนบทความหรือการตอบคำถาม (OpenAI, n.d.) การแปลภาษา ก็เป็นอีกด้านหนึ่งที่ LLMs อย่าง GPT-4 สามารถทำได้อย่างหลากหลายและแม่นยำ (OpenAI, n.d.) นอกจากนี้ การสรุปข้อมูล จากเอกสารขนาดใหญ่ให้สั้นลงแต่ยังคงข้อมูลสำคัญไว้ครบถ้วน ก็เป็นประโยชน์อย่างมากในยุคที่ข้อมูลมีปริมาณมาก (Microsoft, n.d.) อย่างไรก็ตาม LLMs ก็มีข้อจำกัดที่ต้องพิจารณา ความลำเอียงของข้อมูล (Bias) เป็นปัญหาที่เกิดจากข้อมูลที่ใช้ในการฝึกอาจมีอคติ ทำให้โมเดลแสดงผลลัพธ์ที่ไม่เป็นธรรม (Floridi & Chiriatti, 2020) ความถูกต้องของข้อมูล ก็เป็นอีกประเด็น โมเดลบางครั้งอาจให้คำตอบที่ผิดพลาดหรือมั่นใจในคำตอบที่ไม่ถูกต้อง (OpenAI, n.d.) การใช้พลังงานสูง ในการฝึก LLMs ใช้ทรัพยากรพลังงานและฮาร์ดแวร์จำนวนมาก ซึ่งอาจส่งผลกระทบต่อสิ่งแวดล้อมและต้องการการจัดการที่เหมาะสม (Microsoft, n.d.)

ผลกระทบของ LLMs ต่อสังคมมีทั้งด้านบวกและด้านลบ ในด้านบวก LLMs ช่วยให้การเข้าถึงข้อมูลเป็นไปได้ง่ายขึ้น ผู้คนสามารถค้นหาและเข้าใจข้อมูลได้อย่างรวดเร็ว (OpenAI, n.d.) เช่น การตอบคำถามทางการแพทย์หรือวิทยาศาสตร์ อีกทั้งยัง พัฒนาทักษะ โดยสนับสนุนการเรียนรู้ เช่น ช่วยแปลเอกสาร ช่วยในการเขียน หรือเป็นผู้ช่วยในการสอน (Microsoft, n.d.) ในด้านลบ LLMs อาจส่งผลต่อ การเปลี่ยนแปลงในตลาดแรงงาน โดยอาจแทนที่งานบางประเภท เช่น การเขียนข่าว การแปลภาษา หรือการสนับสนุนลูกค้า ทำให้เกิดความกังวลเกี่ยวกับการสูญเสียงาน (Floridi & Chiriatti, 2020) นอกจากนี้ยังมี ความเสี่ยงด้านจริยธรรม เช่น การใช้งานเพื่อสร้างข้อมูลปลอมหรือบิดเบือนความจริง ซึ่งอาจส่งผลกระทบต่อความน่าเชื่อถือของข้อมูลในสังคม Floridi & Chiriatti, 2020)

Ollama 3.2: โมเดลภาษาขนาดใหญ่ที่พร้อมใช้งานแบบออฟไลน์

Ollama 3.2 เป็นโมเดลภาษาขนาดใหญ่ (Large Language Model: LLM) ที่พัฒนาโดย Meta เน้นการใช้งานแบบออฟไลน์และเป็นโอเพนซอร์ส โมเดลนี้ออกแบบให้สามารถรันบนคอมพิวเตอร์ส่วนบุคคลโดยไม่ต้องพึ่งพาการเชื่อมต่ออินเทอร์เน็ต ซึ่งแตกต่างจาก LLMs อื่น ๆ

อย่างเช่น Vicuna, Alpaca และ LLaMA ที่ต้องการการเชื่อมต่อออนไลน์ในการประมวลผล การพัฒนาและข้อมูลอ้างอิงเกี่ยวกับโมเดลนี้สามารถพบได้ในบทความของ Meta (2024) และ Ollama (2024)สถาปัตยกรรมและการทำงานของ Ollama 3.2

Ollama 3.2 ใช้สถาปัตยกรรม Transformer ซึ่งเป็นมาตรฐานสำหรับโมเดลภาษาขนาดใหญ่ โดยได้มีการปรับปรุงกลไก Attention เพื่อให้เหมาะสมกับการทำงานแบบออฟไลน์ โมเดลนี้ถูกฝึกด้วยข้อมูลขนาดใหญ่และเพิ่มประสิทธิภาพในการประมวลผลข้อความ โดยเน้นที่ความแม่นยำและความรวดเร็ว สถาปัตยกรรม Transformer ถูกอธิบายโดย Vaswani et al. (2017) การติดตั้งและการใช้งาน Ollama 3.2

Ollama 3.2 รองรับการใช้งานบนระบบปฏิบัติการ macOS, Linux และ Windows ผู้ใช้สามารถดาวน์โหลดและติดตั้งซอฟต์แวร์จากเว็บไซต์อย่างเป็นทางการของ Ollama การติดตั้งและการใช้งานเพิ่มเติมสามารถศึกษาได้จากเว็บไซต์ของ Ollama (2024) ประสิทธิภาพและข้อจำกัดของ Ollama 3.2

Ollama 3.2 มีความแม่นยำและความเร็วในการประมวลผลข้อความที่โดดเด่น พร้อมทั้งใช้ทรัพยากรระบบอย่างมีประสิทธิภาพ อย่างไรก็ตาม รายละเอียดเพิ่มเติมเกี่ยวกับข้อจำกัดของโมเดลนี้สามารถพบได้ในงานของ Floridi & Chiriatti (2020)การพัฒนาและปรับแต่ง Ollama 3.2

ผู้ใช้สามารถปรับแต่ง Ollama 3.2 ให้เหมาะสมกับงานเฉพาะได้ผ่านกระบวนการ Fine-tuning ข้อมูลเพิ่มเติมเกี่ยวกับการปรับแต่งสามารถศึกษาได้จาก Meta (2024)อนาคตของ Ollama และบทบาทในวงการ AI

2.1.2 ทฤษฎีความรู้เกี่ยวกับ Optical Character Recognition (OCR)

เทคโนโลยีการรู้จำตัวอักษรด้วยภาพ หรือ Optical Character Recognition (OCR) เป็นนวัตกรรมที่มีบทบาทสำคัญในการแปลงข้อมูลจากรูปแบบเอกสารที่เป็นภาพหรือกระดาษเข้าสู่ระบบดิจิทัล เพื่อเพิ่มประสิทธิภาพในการจัดเก็บและประมวลผลข้อมูล หลักการทำงาน เทคนิคที่ใช้ ประเภท การประยุกต์ใช้ ข้อจำกัด และนำเสนอตัวอย่างโค้ด Python สำหรับการทำ OCR มีดังนี้

หลักการทำงานพื้นฐานของ OCR

กระบวนการทำงานของ OCR ประกอบด้วยขั้นตอนหลัก ๆ ดังนี้:

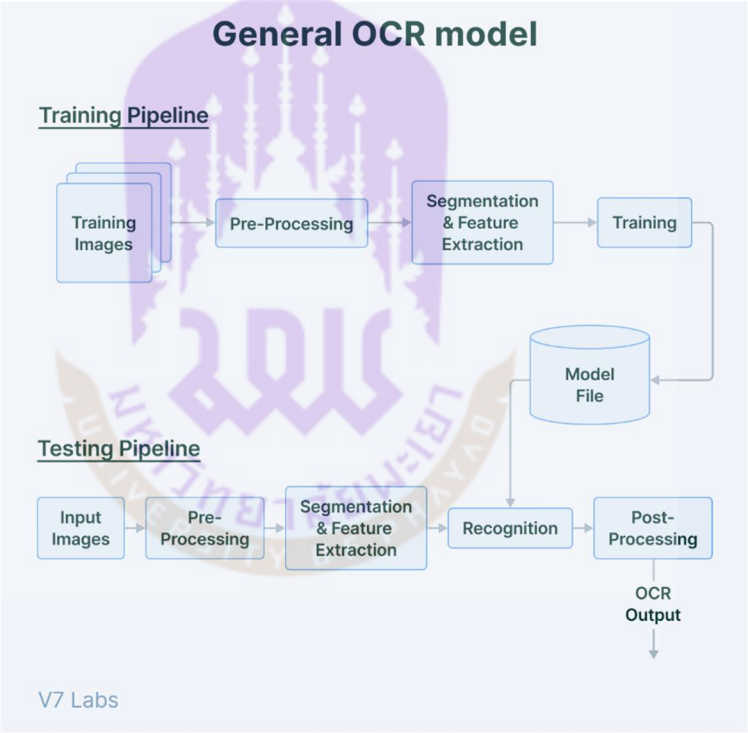
1) การรับภาพ (Image Acquisition): เริ่มต้นด้วยการนำเข้าภาพจากสแกนเนอร์หรือกล้องถ่ายภาพ (Smith, 2007)

2) การประมวลผลภาพ (Image Preprocessing): ปรับปรุงคุณภาพของภาพ เช่น การแปลงเป็นสีเทา การลบสัญญาณรบกวน และการปรับความคมชัด เพื่อเตรียมภาพสำหรับการจดจำ (Vaswani et al., 2017)

3) การแยกส่วนอักขระ (Segmentation): แยกตัวอักษรแต่ละตัวออกจากกัน เพื่อให้ระบบสามารถจดจำได้อย่างถูกต้อง

4) การจดจำอักขระ (Character Recognition): ใช้โมเดล ML หรือ DL ในการวิเคราะห์และจดจำตัวอักษร โดยเทียบกับข้อมูลที่ได้รับการฝึกฝนไว้

5) การแปลงเป็นข้อความ (Postprocessing): ตรวจสอบและแก้ไขข้อผิดพลาด เช่น การสะกดคำ เพื่อให้ได้ข้อความที่สมบูรณ์และถูกต้อง (Ray, 2019)



ภาพที่ 3 แสดงกระบวนการทำงานของการรู้จำอักขระด้วยแสง (Optical Character Recognition (OCR)) (V7 Labs (n.d.))

หมายเหตุ จาก "Optical Character Recognition (OCR): Definition & How To Guide," โดย V7 Labs, n.d. สืบค้นจาก https://www.v7labs.com/blog/ocr-guide.

เทคนิคและวิธีการที่ใช้ใน OCR

1) การประมวลผลภาพ (Image Preprocessing): เทคนิคนี้รวมถึงการแปลงภาพเป็นสีเทา (Grayscale) และการลบสัญญาณรบกวน (Noise Removal) เพื่อเพิ่มความชัดเจนของตัวอักษร

2) การสกัดคุณลักษณะ (Feature Extraction): การแยกคุณลักษณะเฉพาะของตัวอักษร เช่น รูปร่าง ขนาด และมุม เพื่อใช้ในการจดจำ (Smith, 2007)

3) การรู้จำรูปแบบ (Pattern Recognition): การจับคู่ตัวอักษรกับรูปแบบที่เก็บไว้ในฐานข้อมูล

4) Machine Learning: ใช้อัลกอริทึม เช่น Support Vector Machine (SVM) หรือ K-Nearest Neighbors (KNN) ในการจดจำตัวอักษร (Ray, 2019)

5) Deep Learning: ใช้โครงข่ายประสาทเทียมแบบคอนโวลูชัน (Convolutional Neural Networks – CNNs) และโครงข่ายประสาทเทียมแบบวนซ้ำ (Recurrent Neural Networks – RNNs) สำหรับงานที่ซับซ้อน เช่น การจดจำลายมือ

ข้อดีของ ML และ DL คือความสามารถในการประมวลผลข้อมูลจำนวนมากได้อย่างแม่นยำ แต่ข้อเสียคือการใช้ทรัพยากรที่สูงและความต้องการข้อมูลฝึกที่หลากหลาย

ประเภทของ OCR

1) Printed OCR: สำหรับการจดจำข้อความพิมพ์จากเอกสารทั่วไป ซึ่งมีรูปแบบตัวอักษรที่ชัดเจน (Smith, 2007)

2) Handwritten OCR: สำหรับการจดจำลายมือ ซึ่งมีความซับซ้อนมากขึ้นเนื่องจากความหลากหลายในรูปแบบการเขียน (Ray, 2019)

3) Scene Text OCR: สำหรับการจดจำข้อความในภาพถ่ายหรือสภาพแวดล้อมจริง เช่น ป้ายถนนหรือบรรจุภัณฑ์สินค้า (Vaswani et al., 2017)

การประยุกต์ใช้ OCR

1) การแปลงเอกสารเป็นดิจิทัล: ช่วยในการสแกนหนังสือหรือเอกสารและแปลงเป็นไฟล์ PDF เพื่อการจัดเก็บและค้นหาที่สะดวกขึ้น (Smith, 2007)

2) การอ่านป้ายจราจร: ใช้ในระบบขับเคลื่อนอัตโนมัติของยานพาหนะ เพื่อเพิ่มความปลอดภัยบนท้องถนน (Ray, 2019)

3) การสแกนบัตรประชาชน: ใช้ในงานทะเบียนหรือธนาคาร เพื่อยืนยันตัวตนของลูกค้าอย่างรวดเร็ว

4) การอ่านข้อมูลในอีคอมเมิร์ซ: เช่น การแปลงข้อมูลใบเสร็จหรือบาร์โค้ดเป็นข้อมูลดิจิทัล เพื่อการวิเคราะห์และจัดการสต็อกสินค้า (Vaswani et al., 2017)

ข้อจำกัดและความท้าทายของ OCR

1) คุณภาพของภาพ: ภาพที่มีความคมชัดต่ำ หรือมีสัญญาณรบกวนสูง อาจทำให้การจดจำมีความแม่นยำน้อยลง (Smith, 2007)

2) รูปแบบตัวอักษรที่หลากหลาย: ฟอนต์ที่ไม่มาตรฐาน หรือการบิดเบี้ยวของตัวอักษรอาจทำให้ระบบจดจำผิดพลาด

3) การรองรับหลายภาษา: ระบบ OCR บางระบบอาจรองรับภาษาได้จำกัด ทำให้ไม่สามารถใช้งานได้ทั่วโลก (Ray, 2019)

ตัวอย่างโค้ด Python สำหรับการทำ OCR

ไลบรารีที่ใช้: Tesseract OCR

เวอร์ชัน: Python 3.9, pytesseract 0.3.10, OpenCV 4.5.4

```
# ติดตั้งไลบรารีที่จำเป็น
# pip install pytesseract opencv-python
import cv2
import pytesseract
# ระบุเส้นทางไปยัง Tesseract OCR
pytesseract.pytesseract.tesseract_cmd = r'C:\Program Files\Tesseract-OCR\tesseract.exe'
# โหลดภาพที่ต้องการประมวลผล
image_path = "document.jpg"
image = cv2.imread(image_path)
# แปลงภาพเป็นสีเทาเพื่อเพิ่มประสิทธิภาพในการจดจำ
gray_image = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
# ใช้ Tesseract OCR ในการจดจำข้อความจากภาพ
text = pytesseract.image_to_string(gray_image, lang='eng')
```

```
# แสดงผลลัพธ์ที่ได้จากการจดจำ
print("ข้อความที่จดจำได้:")
print(text)
```

คำอธิบายโค้ด

โหลดภาพด้วย OpenCV: ใช้ cv2.imread เพื่อโหลดภาพที่ต้องการจดจำ

แปลงภาพเป็นสีเทา: ใช้ cv2.cvtColor เพื่อแปลงภาพเป็นสีเทา เพิ่มความแม่นยำในการจดจำ

ใช้ Tesseract OCR: ใช้ pytesseract.image_to_string เพื่อดึงข้อความจากภาพ สามารถระบุภาษาที่ต้องการได้ผ่านพารามิเตอร์ lang

แสดงผลลัพธ์: พิมพ์ข้อความที่จดจำได้ออกทางหน้าจอ

โมเดลภาษา (LLMs) ที่ใช้ใน RAG

โมเดลภาษาขนาดใหญ่ (LLMs) ที่นิยมใช้ในระบบ RAG มีหลายตัวเลือก แต่ละตัวมีข้อดีเฉพาะของตนเอง ได้แก่ (Meta, 2024; Vaswani et al., 2017):

GPT (OpenAI): มีความสามารถสูงในการสร้างข้อความที่มีคุณภาพ แต่ต้องการทรัพยากรในการประมวลผลมาก

Flan-T5 (Google): เด่นในงานที่ต้องการการตีความและการวิเคราะห์เชิงลึก

LLaMA (Meta): เป็นโอเพนซอร์สที่เหมาะสำหรับการปรับแต่งเพื่อใช้งานในระดับองค์กร

การเลือกใช้โมเดลขึ้นอยู่กับความต้องการและทรัพยากรที่มี โดยบางโมเดลอาจเหมาะสมกับงานที่ต้องการความละเอียดและความเข้าใจเชิงลึกมากกว่า

เทคนิคการค้นหาข้อมูลในระบบ RAG

ระบบ RAG ใช้เทคนิคการค้นหาข้อมูลที่หลากหลายเพื่อนำข้อมูลที่เหมาะสมมาใช้ในการสร้างคำตอบ ได้แก่ (Izacard & Grave, 2021):

TF-IDF: เทคนิคที่ใช้ในการวัดความสำคัญของคำในเอกสาร โดยพิจารณาจากความถี่ของคำในเอกสารนั้น ๆ

BM25: เป็นการปรับปรุงจาก TF-IDF โดยเพิ่มการคำนึงถึงความยาวของเอกสาร ทำให้การค้นหามีความแม่นยำมากขึ้น

Vector Search: ใช้เวกเตอร์ในการแทนข้อความ ทำให้การค้นหามีความแม่นยำมากขึ้น โดยเฉพาะในกรณีที่ข้อมูลมีความซับซ้อนและต้องการความสัมพันธ์เชิงความหมาย

การประเมินประสิทธิภาพของระบบ RAG

การประเมินระบบ RAG สามารถทำได้โดยใช้ตัวชี้วัดหลายตัว เช่น ความแม่นยำ (Precision) ที่วัดความถูกต้องของคำตอบ ความครอบคลุม (Recall) ที่วัดจำนวนข้อมูลที่ระบบสามารถค้นพบ และ ความเกี่ยวข้อง (Relevance) ที่วัดความสัมพันธ์ระหว่างคำตอบกับคำถามที่ผู้ใช้ป้อน (Lewis et al., 2020)

โมเดลภาษา (LLMs) ที่ใช้ใน RAG

โมเดลภาษาขนาดใหญ่ (LLMs) ที่นิยมใช้ในระบบ RAG มีหลายตัวเลือก แต่ละตัวมีข้อดีเฉพาะของตนเอง ได้แก่ (Meta, 2024; Vaswani et al., 2017):

GPT (OpenAI): มีความสามารถสูงในการสร้างข้อความที่มีคุณภาพ แต่ต้องการทรัพยากรในการประมวลผลมาก

Flan-T5 (Google): เด่นในงานที่ต้องการการตีความและการวิเคราะห์เชิงลึก

LLaMA (Meta): เป็นโอเพนซอร์สที่เหมาะสำหรับการปรับแต่งเพื่อใช้งานในระดับองค์กร

การเลือกใช้โมเดลขึ้นอยู่กับความต้องการและทรัพยากรที่มี โดยบางโมเดลอาจเหมาะสมกับงานที่ต้องการความละเอียดและความเข้าใจเชิงลึกมากกว่า

เทคนิคการค้นหาข้อมูลในระบบ RAG

ระบบ RAG ใช้เทคนิคการค้นหาข้อมูลที่หลากหลายเพื่อนำข้อมูลที่เหมาะสมมาใช้ในการสร้างคำตอบ ได้แก่ (Izacard & Grave, 2021):

TF-IDF: เทคนิคที่ใช้ในการวัดความสำคัญของคำในเอกสาร โดยพิจารณาจากความถี่ของคำในเอกสารนั้น ๆ

BM25: เป็นการปรับปรุงจาก TF-IDF โดยเพิ่มการคำนึงถึงความยาวของเอกสาร ทำให้การค้นหามีความแม่นยำมากขึ้น

Vector Search: ใช้เวกเตอร์ในการแทนข้อความ ทำให้การค้นหามีความแม่นยำมากขึ้น โดยเฉพาะในกรณีที่ข้อมูลมีความซับซ้อนและต้องการความสัมพันธ์เชิงความหมาย

การประเมินประสิทธิภาพของระบบ RAG

การประเมินระบบ RAG สามารถทำได้โดยใช้ตัวชี้วัดหลายตัว เช่น ความแม่นยำ (Precision) ที่วัดความถูกต้องของคำตอบ ความครอบคลุม (Recall) ที่วัดจำนวนข้อมูลที่ระบบสามารถค้นพบ และ ความเกี่ยวข้อง (Relevance) ที่วัดความสัมพันธ์ระหว่างคำตอบกับคำถามที่ผู้ใช้ป้อน (Lewis et al., 2020)

ความท้าทายและแนวโน้มในอนาคตของ RAG

ความท้าทายในการพัฒนา RAG รวมถึงการจัดการกับข้อมูลที่ซับซ้อน เช่น ไฟล์ PDF ที่มีโครงสร้างไม่ชัดเจน การลดความลำเอียงในข้อมูล และการพัฒนาโมเดลที่มีประสิทธิภาพแต่ใช้ทรัพยากรน้อย (Lewis et al., 2020; Izacard & Grave, 2021)

แนวโน้มในอนาคตของ RAG คือการผสานรวมเข้ากับระบบ AI ที่สามารถเรียนรู้และปรับตัวกับข้อมูลใหม่ได้แบบเรียลไทม์ ทำให้ระบบมีความยืดหยุ่นและสามารถตอบสนองต่อการเปลี่ยนแปลงของข้อมูลได้อย่างรวดเร็ว (Meta, 2024)

โค้ด Python สำหรับการใช้งาน RAG กับไฟล์ PDF

```
import PyPDF2
from langchain.vectorstores import FAISS
from langchain.embeddings import OpenAIEmbeddings
from transformers import pipeline
# ฟังก์ชันดึงข้อความจากไฟล์ PDF
def extract_text_from_pdf(pdf_path):
    with open(pdf_path, 'rb') as file:
        reader = PyPDF2.PdfReader(file)
        text = ""
        for page in reader.pages:
            text += page.extract_text()
    return text
# สร้างฐานข้อมูลเวกเตอร์
def create_vector_database(texts):
    embeddings = OpenAIEmbeddings()
    vectorstore = FAISS.from_texts(texts, embeddings)
    return vectorstore
# ค้นหาข้อมูลที่เกี่ยวข้อง
def search_query(vectorstore, query):
    results = vectorstore.similarity_search(query)
    return results
# สร้างคำตอบด้วย LLM
```

```
def generate_answer(context, query):
    generator = pipeline("text-generation", model="gpt-3.5-turbo")
    input_text = f"Context: {context}\nQuestion: {query}\nAnswer:"
    response = generator(input_text, max_length=200, num_return_sequences=1)
    return response[0]['generated_text']
# ตัวอย่างการใช้งาน
pdf_path = "example.pdf"
query = "ระบบ RAG คืออะไร?"
# ดึงข้อความจาก PDF
text = extract_text_from_pdf(pdf_path)
# สร้างฐานข้อมูลเวกเตอร์
vectorstore = create_vector_database([text])
# ค้นหาข้อมูลที่เกี่ยวข้อง
results = search_query(vectorstore, query)
# ใช้ LLM สร้างคำตอบ
context = " ".join([res['content'] for res in results])
answer = generate_answer(context, query)
print("คำตอบ:", answer)
```

การใช้ Retrieval-Augmented Generation (RAG) ในการฝึกฝน Large Language Models (LLMs)

Retrieval-Augmented Generation (RAG) เป็นเทคนิคที่ผสมผสานการดึงข้อมูล (retrieval) กับการสร้างข้อความ (generation) ในการประมวลผลภาษาธรรมชาติ (NLP) ซึ่งช่วยเพิ่มความแม่นยำและประสิทธิภาพของโมเดลภาษาขนาดใหญ่ (LLMs) โดยการลดข้อผิดพลาดจากการสร้างข้อมูลที่ไม่มีอยู่จริง (Lewis et al., 2020) เทคนิคนี้ช่วยให้โมเดลสามารถตอบคำถามที่มีความซับซ้อนและต้องการความรู้เฉพาะทางได้ดีขึ้น

ความสัมพันธ์ระหว่าง RAG และการฝึกฝน LLMs

RAG สามารถนำมาใช้ได้ในทั้งขั้นตอนการใช้งานจริง (Inference Time) และการฝึกฝนโมเดล (Training Time):

1) Inference Time: ในขั้นตอนนี้ RAG ช่วยดึงข้อมูลที่เกี่ยวข้องจากฐานข้อมูลภายนอกแบบเรียลไทม์ เพื่อใช้เป็นบริบทในการตอบคำถาม ทำให้คำตอบมีความแม่นยำและอัปเดตเสมอ (Lewis et al., 2020)

2) Training Time: ในขั้นตอนการฝึกฝน RAG สามารถสร้างข้อมูลเสริม (augmented data) หรือข้อมูลสังเคราะห์ (synthetic data) จากแหล่งข้อมูลที่เกี่ยวข้อง เพื่อช่วยเพิ่มความหลากหลายและคุณภาพของชุดข้อมูลฝึกฝน (Izacard & Grave, 2021)

ข้อดีของ RAG ในการฝึกฝน LLMs ได้แก่

ลดความต้องการข้อมูลที่ทำเครื่องหมาย (annotated data): การใช้ RAG ช่วยลดปริมาณข้อมูลที่ต้องมีการทำเครื่องหมายด้วยมือ ซึ่งเป็นกระบวนการที่ใช้เวลาและทรัพยากรสูง (Lewis et al., 2020)

เพิ่มประสิทธิภาพของโมเดล: ด้วยการเข้าถึงข้อมูลที่เกี่ยวข้องและทันสมัย ทำให้โมเดลสามารถตอบคำถามได้อย่างแม่นยำยิ่งขึ้น (Izacard & Grave, 2021)

อย่างไรก็ตาม RAG ยังมีข้อจำกัด เช่น ความซับซ้อนในการผสานกระบวนการ retrieval และ generation รวมถึงความต้องการทรัพยากรคอมพิวเตอร์สูงสำหรับการดึงและประมวลผลข้อมูลแบบเรียลไทม์ (Lewis et al., 2020)

เทคนิคการใช้ RAG ในการฝึกฝน LLMs

การนำ RAG มาใช้ในการฝึกฝน LLMs มีเทคนิคหลายประการ:

1) การสร้างข้อมูลเสริม (Augmented Data): RAG ดึงข้อมูลที่เกี่ยวข้องจากแหล่งข้อมูลภายนอกมาผสมกับข้อมูลฝึกฝนที่มีอยู่ เพื่อเพิ่มความหลากหลายและคุณภาพของข้อมูล (Izacard & Grave, 2021)

2) ปรับปรุงฟังก์ชันการสูญเสีย (Loss Function): เพิ่มองค์ประกอบที่คำนึงถึงความสอดคล้องระหว่างข้อมูลที่โมเดลสร้างขึ้นกับข้อมูลที่ดึงมา ทำให้โมเดลสามารถสร้างข้อความที่สอดคล้องและมีความน่าเชื่อถือ (Lewis et al., 2020)

3) การสร้างข้อมูลสังเคราะห์ (Synthetic Data): ใช้ RAG เพื่อสร้างคู่คำถาม–คำตอบใหม่จากเอกสารหรือฐานข้อมูลที่มีอยู่ ขยายชุดข้อมูลฝึกฝนให้มีประสิทธิภาพยิ่งขึ้น (Izacard & Grave, 2021)

ตัวอย่างการใช้งาน RAG ในการฝึกฝน LLMs

การใช้ RAG ในการฝึกฝนโมเดลภาษามีตัวอย่างหลายกรณี ตัวอย่างเช่น การใช้ RAG ในการฝึกโมเดลที่ตอบคำถามทางการแพทย์โดยดึงข้อมูลจากเวชระเบียนและงานวิจัย

ที่เกี่ยวข้อง ผลลัพธ์คือการตอบคำถามที่มีความแม่นยำและเชื่อถือได้มากยิ่งขึ้น (Roberts et al., 2020) การใช้ RAG ยังสามารถเพิ่มความแม่นยำในการตอบคำถามในสาขาเฉพาะ เช่น วิทยาศาสตร์หรือวิศวกรรม (Karpukhin et al., 2020)

โค้ด Python สำหรับใช้ RAG ในการฝึกฝน LLMs

```python
import PyPDF2
from langchain.vectorstores import FAISS
from langchain.embeddings import OpenAIEmbeddings
from transformers import pipeline, Trainer, TrainingArguments

# 1. ดึงข้อมูลจากไฟล์ PDF
def extract_text_from_pdf(pdf_path):
    with open(pdf_path, 'rb') as file:
        reader = PyPDF2.PdfReader(file)
        text = ""
        for page in reader.pages:
            text += page.extract_text()
    return text

# 2. สร้างฐานข้อมูลเวกเตอร์
def create_vector_database(texts):
    embeddings = OpenAIEmbeddings()
    vectorstore = FAISS.from_texts(texts, embeddings)
    return vectorstore

# 3. ค้นหาข้อมูลที่เกี่ยวข้อง
def search_query(vectorstore, query):
    results = vectorstore.similarity_search(query)
    return results

# 4. สร้างข้อมูลสำหรับฝึก LLM
def generate_synthetic_data(context, query):
    generator = pipeline("text-generation", model="gpt-2")
    input_text = f"Context: {context}\nQuestion: {query}\nAnswer:"
    response = generator(input_text, max_length=200, num_return_sequences=1)
    return response[0]['generated_text']
```

```python
# 5. ฝึก LLM ด้วยข้อมูลจาก RAG
def train_model(training_data):
    model = pipeline("text-classification", model="bert-base-uncased")
    trainer = Trainer(
        model=model,
        args=TrainingArguments(
            output_dir="./results",
            evaluation_strategy="epoch",
            learning_rate=2e-5,
            per_device_train_batch_size=8,
            num_train_epochs=3,
        ),
        train_dataset=training_data,
    )
    trainer.train()

# การใช้งาน
pdf_path = "example.pdf"
query = "ระบบ RAG คืออะไร?"

# ขั้นตอนที่ 1: ดึงข้อความจาก PDF
text = extract_text_from_pdf(pdf_path)

# ขั้นตอนที่ 2: สร้างฐานข้อมูลเวกเตอร์
vectorstore = create_vector_database([text])

# ขั้นตอนที่ 3: ค้นหาข้อมูล
results = search_query(vectorstore, query)

# ขั้นตอนที่ 4: สร้างข้อมูลสำหรับฝึกฝน
context = " ".join([res['content'] for res in results])
synthetic_data = generate_synthetic_data(context, query)

# ขั้นตอนที่ 5: ฝึกโมเดลด้วยข้อมูล
train_model(synthetic_data)
```

อธิบายโค้ด:

ดึงข้อความจากไฟล์ PDF: ใช้ PyPDF2 เพื่ออ่านและดึงข้อความจากไฟล์ PDF (Lewis et al., 2020)

สร้างฐานข้อมูลเวกเตอร์: ใช้ FAISS และ OpenAI Embeddings เพื่อสร้างฐานข้อมูลสำหรับการค้นหาแบบเวกเตอร์ (Izacard & Grave, 2021)

ค้นหาข้อมูลที่เกี่ยวข้อง: ค้นหาข้อความที่มีความคล้ายคลึงกับคำถามที่กำหนด (Karpukhin et al., 2020)

สร้างข้อมูลสังเคราะห์: ใช้โมเดลการสร้างข้อความในการสร้างคำตอบที่สอดคล้องกับบริบทที่ดึงมา (Roberts et al., 2020)

ฝึกโมเดลด้วยข้อมูลที่สร้างขึ้น: ใช้ Hugging Face Trainer เพื่อฝึกโมเดลด้วยข้อมูลใหม่

การประเมินประสิทธิภาพของ LLMs ที่ฝึกด้วย RAG

การประเมินประสิทธิภาพของโมเดลที่ฝึกด้วย RAG สามารถทำได้โดยเปรียบเทียบกับโมเดลที่ไม่ได้ใช้ RAG โดยวัดความแม่นยำ ความเกี่ยวข้อง และความถูกต้องของคำตอบ รวมถึงการทดสอบในงานเฉพาะทาง เช่น การสรุปข้อมูลหรือการตอบคำถามในสาขาที่ต้องการความเชี่ยวชาญ (Roberts et al., 2020)

อนาคตของการใช้ RAG ในการฝึก LLMs

RAG มีศักยภาพสูงในการพัฒนา LLMs ให้สามารถตอบคำถามได้อย่างแม่นยำและเชื่อถือได้มากขึ้น การผสานการดึงข้อมูลเข้ากับการสร้างข้อความช่วยให้โมเดลสามารถเข้าถึงข้อมูลที่ทันสมัยและตอบคำถามที่ซับซ้อนได้ดียิ่งขึ้น แนวโน้มในอนาคตคือการพัฒนา RAG ให้รองรับข้อมูลที่หลากหลายและมีประสิทธิภาพมากขึ้น (Karpukhin et al., 2020)

การวัดประสิทธิภาพและความถูกต้องของ Large Language Models (LLMs) ที่ฝึกฝนด้วย Retrieval–Augmented Generation (RAG)

ความสำคัญของการวัดผลโมเดล LLM ที่ฝึกฝนด้วย RAG

การวัดประสิทธิภาพของ Large Language Models (LLMs) ที่ฝึกด้วยเทคนิค Retrieval–Augmented Generation (RAG) มีบทบาทสำคัญในกระบวนการพัฒนาโมเดลเพื่อตอบโจทย์งานที่ต้องการความแม่นยำและเกี่ยวข้อง เช่น งานถามตอบอัตโนมัติและการสร้างข้อความ (Lewis et al., 2020). LLMs เช่น Ollama และ LLaMA ใช้ RAG เพื่อลดข้อผิดพลาดที่เกิดจากการสร้างข้อมูลที่ไม่มีอยู่จริง (hallucination) และเพิ่มการอ้างอิงข้อมูลจากเอกสารที่เกี่ยวข้อง (Izacard & Grave, 2021)

การประเมินโมเดล RAG จำเป็นต้องใช้ตัวชี้วัดหลากหลายประเภท เช่น Exact Match (EM) และ F1–score เพื่อวัดความถูกต้องของคำตอบ รวมถึง BLEU และ ROUGE เพื่อวัดความสอดคล้องของโครงสร้างข้อความ (Ge et al., 2023)

Exact Match (EM) และ F1-score

Exact Match (EM)

Exact Match ใช้สำหรับวัดว่าคำตอบที่โมเดลสร้างตรงกับคำตอบอ้างอิง 100% หรือไม่ โดยไม่มีการพิจารณาโครงสร้างทางไวยากรณ์หรือคำที่มีความหมายใกล้เคียง (Lewis et al., 2020). ตัวอย่างเช่น:

คำตอบโมเดล: "กรุงเทพมหานคร"

คำตอบอ้างอิง: "กรุงเทพมหานคร"

ผลลัพธ์: EM = 1 (100%)

ข้อดี: EM เหมาะกับงานที่ต้องการคำตอบที่แน่นอน เช่น การตอบคำถามทางเทคนิค (Meta, 2024).

ข้อเสีย: EM ไม่เหมาะกับงานที่คำตอบมีหลายรูปแบบ เช่น การสรุปข้อความ (Ge et al., 2023).

F1-score

F1-score วัดความสมดุลระหว่าง Precision (ความแม่นยำ) และ Recall (ความครอบคลุม). ตัวอย่างการคำนวณ:

คำตอบโมเดล: "กรุงเทพ เมืองหลวงของไทย"

คำตอบอ้างอิง: "กรุงเทพมหานคร เมืองหลวงของประเทศไทย"

Precision = 3/4 = 0.75

Recall = 3/5 = 0.6

F1-score $= 2 \times \frac{0.75 \times 0.6}{0.75 + 0.6} = 0.67.$

ข้อดี: F1-score มีความยืดหยุ่นมากกว่า EM ในการวัดคำตอบที่มีความหลากหลาย

ข้อเสีย: ต้องใช้เวลาคำนวณมากกว่า EM (Izacard & Grave, 2021)

BLEU และ ROUGE

BLEU (Bilingual Evaluation Understudy)

BLEU ใช้ n-grams เพื่อวัดความคล้ายคลึงของคำตอบโมเดลและคำตอบอ้างอิงในเชิงความหมาย (Lewis et al., 2020). ตัวอย่างการคำนวณ BLEU-1:

คำตอบโมเดล: "กรุงเทพ เมืองหลวงของไทย"

คำตอบอ้างอิง: "กรุงเทพ เมืองหลวงของประเทศไทย"

BLEU-1 = 3/4 = 0.75

ROUGE

ROUGE เน้นการวัด recall โดยวิเคราะห์ความคล้ายคลึงของโครงสร้างข้อความ เช่น longest common subsequence (LCS) (Ge et al., 2023)

ข้อดีของ BLEU และ ROUGE:

BLEU เหมาะกับงานแปลภาษา (Meta, 2024)

ROUGE เหมาะกับการสรุปข้อความ

ข้อเสีย:

BLEU ไม่คำนึงถึงความหมายโดยรวมของข้อความ

ROUGE อาจให้ผลลัพธ์ต่ำในงานที่มีคำตอบหลากหลาย

การวัดความมั่นใจของโมเดล

Confidence Score ใช้เพื่อวัดระดับความมั่นใจของโมเดลในการสร้างคำตอบแต่ละครั้ง เช่น:

Confidence Score: 0.9 (คำตอบมีโอกาสถูกต้องสูง).

Confidence Score: 0.5 (คำตอบอาจผิดพลาด).

ตัวอย่างเช่น โมเดล Ollama 3.2 ใช้ความมั่นใจช่วยเลือกคำตอบที่เหมาะสมที่สุดในงานถามตอบอัตโนมัติ (Meta, 2024)

การปรับปรุงกระบวนการฝึกฝนโดยใช้ผลการประเมิน

การปรับปรุงโมเดลสามารถทำได้โดย:

ปรับ Hyperparameters: เช่น การลด learning rate เพื่อลด overfitting (Izacard & Grave, 2021)

ปรับปรุงข้อมูลฝึกฝน: เพิ่มข้อมูลที่เกี่ยวข้องเพื่อลด bias

ปรับปรุง RAG Workflow: ใช้ vector search แทน TF-IDF ในการดึงข้อมูล

ตัวอย่างการประเมินโมเดล Ollama ที่ฝึกด้วย RAG

ในงานประเมินโมเดล Ollama ที่ฝึกฝนด้วย RAG โดยใช้ EM และ BLEU พบว่า:

EM = 85%: แสดงถึงความแม่นยำในคำตอบเดี่ยว

BLEU-4 = 70%: แสดงถึงความคล้ายคลึงของข้อความกับคำตอบอ้างอิง

ข้อสรุป: Ollama มีประสิทธิภาพสูงในงานที่ต้องการความถูกต้องเฉพาะเจาะจง แต่ยังต้องปรับปรุงการตอบคำถามที่มีคำตอบหลากหลาย (Meta, 2024)

ข้อจำกัดของตัวชี้วัดและแนวทางการประเมินในอนาคต

ข้อจำกัด

1. ตัวชี้วัดเชิงปริมาณ เช่น BLEU และ ROUGE ไม่สามารถวัดความหมายโดยรวมของข้อความได้

2. EM และ F1-score ไม่เหมาะกับงานที่มีคำตอบหลากหลาย

แนวทางในอนาคต:

พัฒนาตัวชี้วัดที่คำนึงถึงความหมาย (semantic-based metrics) และใช้การประเมินร่วมกับการวิเคราะห์โดยผู้เชี่ยวชาญ

2.2 เอกสารและงานวิจัยที่เกี่ยวข้อง (Related Research Works)

2.2.1 งานวิจัยในประเทศ

การประยุกต์ใช้ LLM กำลังเติบโตอย่างรวดเร็วในประเทศไทย โดยมีหน่วยงานที่สำคัญจากทั้งภาครัฐ ภาคเอกชน และชุมชนโอเพนซอร์ส ได้แก่

Pathumma LLM (NECTEC): ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) ซึ่งเป็นหน่วยงานวิจัยของรัฐ ได้พัฒนา Pathumma LLM โดยมีเป้าหมายให้เป็นแบบจำลองที่เข้าใจภาษาและบริบทของประเทศไทยอย่างลึกซึ้ง จุดเด่นที่สำคัญที่สุดคือ Pathumma LLM ถูกออกแบบมาให้มีความสามารถในการ "เข้าใจภาษาราชการ" และสามารถช่วยตรวจสอบหรือร่างเอกสารราชการได้ (ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ , ม.ป.ป.)

Typhoon Series (SCB 10X): บริษัท เอสซีบี เท็นเอกซ์ จำกัด ซึ่งเป็นบริษัทเอกชน ได้พัฒนา Typhoon ซึ่งเป็นตระกูลแบบจำลองภาษาที่มีสมรรถนะสูงและได้รับการยอมรับในวงกว้าง กรณีศึกษาที่สำคัญอย่างยิ่งคือ ความร่วมมือเชิงยุทธศาสตร์ระหว่างบริษัท เอสซีบี

เอกซ์ จำกัด (มหาชน) และสำนักงานคณะกรรมการพัฒนาระบบราชการ (ก.พ.ร.) ในการนำแบบจำลอง Typhoon มาพัฒนาเป็นระบบสนทนาอัจฉริยะ (Chatbot) เพื่อสนับสนุนการดำเนินงานของภาครัฐ ความร่วมมือดังกล่าวถือเป็นหลักฐานเชิงประจักษ์ที่แสดงให้เห็นถึงแนวโน้มการนำ LLM ที่พัฒนาโดยภาคเอกชนไทยมาประยุกต์ใช้ในระบบราชการอย่างเป็นรูปธรรม (SCBX. , 2568)

WangChanGLM (PyThaiNLP & VISTEC): ชุมชนนักพัฒนาโอเพนซอร์สยังมีบทบาทสำคัญในการขับเคลื่อนปัญญาประดิษฐ์ของไทย โครงการที่โดดเด่นคือ WangChanGLM ซึ่งเป็นผลจากความร่วมมือระหว่างชุมชนผู้พัฒนา PyThaiNLP และสถาบันวิทยสิริเมธี (VISTEC) แบบจำลอง WangChanGLM พัฒนาต่อยอดมาจากแบบจำลองโอเพนซอร์สระดับสากล โดยมุ่งเน้นการพัฒนาความสามารถในการปฏิบัติตามคำสั่ง (Instruction-following) โดยเฉพาะอย่างยิ่งในด้านการอ่านเพื่อความเข้าใจ (Reading Comprehension) ซึ่งเป็นทักษะพื้นฐานที่จำเป็นต่อการสรุปความเอกสาร (PyThaiNLP, 2023)

2.2.2 งานวิจัยต่างประเทศ

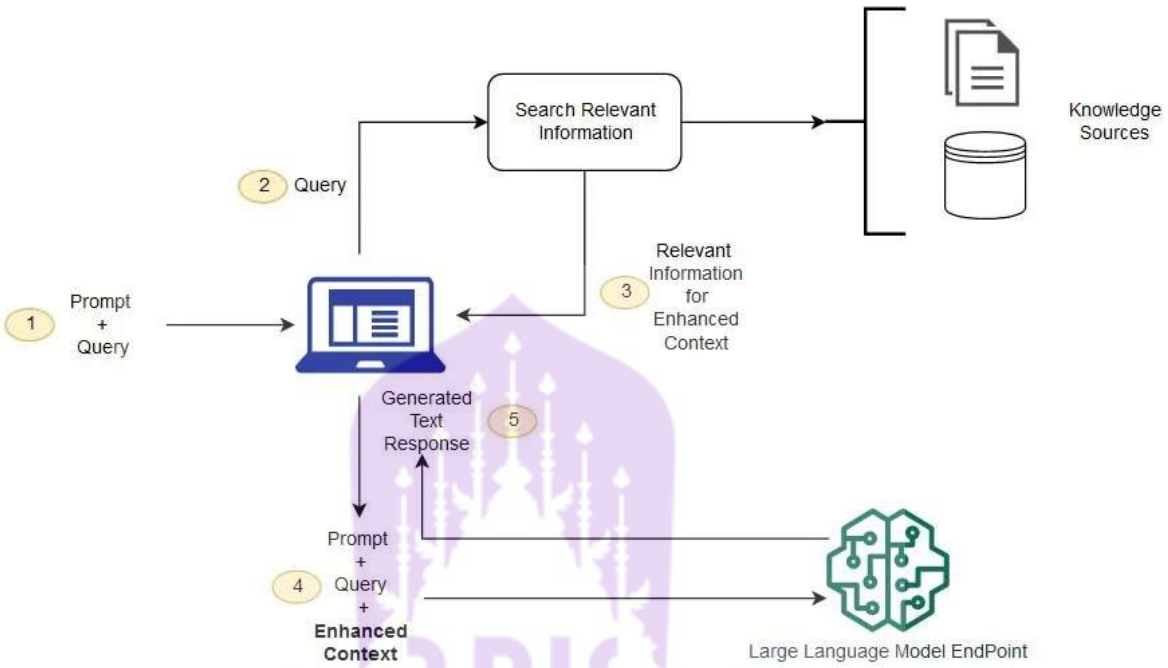
แนวโน้มการนำ LLM มาประยุกต์ใช้กับเอกสารเฉพาะทางเกิดขึ้นทั่วโลกในหลากหลายวงการ เพื่อเพิ่มประสิทธิภาพการบริการประชาชน ตัวอย่างเช่น GovTech ของประเทศสิงคโปร์ได้พัฒนาแอปพลิเคชันภายในชื่อ "Pair" เพื่อใช้สรุปข้อความและสร้างรายงาน ซึ่งช่วยลดภาระงานของเจ้าหน้าที่ได้อย่างมาก ในขณะเดียวกัน ประเทศอย่างไอซ์แลนด์และฟินแลนด์เลือกที่จะร่วมมือกับผู้ให้บริการ LLM ระดับโลก แล้วนำแบบจำลองมาปรับจูนเพิ่มเติมด้วยข้อมูลที่เป็นกรรมสิทธิ์ของตนเองเพื่อปรับให้เข้ากับบริบทของประเทศ แนวทางเหล่านี้สะท้อนให้เห็นถึงความท้าทายร่วมกันในเรื่องความถูกต้องของข้อมูล โดยหลายหน่วยงานเลือกใช้สถาปัตยกรรม RAG เพื่อ "จำกัดขอบเขต" ให้ LLM ดึงคำตอบจากชุดข้อมูลที่เชื่อถือได้เท่านั้น ซึ่งช่วยลดปัญหาการสร้างข้อมูลเท็จ (Hallucination) ที่เป็นความเสี่ยงสำคัญสำหรับหน่วยงานภาครัฐ

2.3 กรอบแนวคิดในการวิจัย (Conceptual Framework)

การนำ Retrieval-Augmented Generation (RAG) มาประยุกต์ใช้ร่วมกับ Large Language Models (LLMs) สำหรับการดึงข้อมูลจากไฟล์ PDF

การใช้ปัญญาประดิษฐ์ในการประมวลผลข้อมูลเพื่อสร้างข้อความหรือหาคำตอบให้ผู้ใช้ได้รับความนิยมอย่างแพร่หลาย หนึ่งในเทคนิคที่โดดเด่นคือ Retrieval-Augmented Generation

(RAG) ที่มีการผสานรวมกระบวนการ ดึงข้อมูล (retrieval) และ การสร้างข้อความ (generation) ด้วยโมเดลภาษาขนาดใหญ่ (Large Language Models: LLMs) บทความนี้จะนำเสนอเกี่ยวกับ RAG วิธีการทำงาน การประยุกต์ใช้กับไฟล์ PDF และประโยชน์ที่สามารถนำไปใช้ได้อย่างหลากหลาย (Lewis et al., 2020; Izacard & Grave, 2021)



ภาพที่ 4 ขั้นตอนการทำงานของ RAG: การรับคำถาม ค้นหา ดึงข้อมูล และสร้างคำตอบ (Khalid, 2024)

หมายเหตุ. จาก "Introduction to RAG," โดย Q. Khalid, 2024, Medium. สืบค้นจาก https://medium.com/@quaidkhalid/introduction-to-rag-4f1f44020719. Copyright 2024 โดย Quaid Khalid.

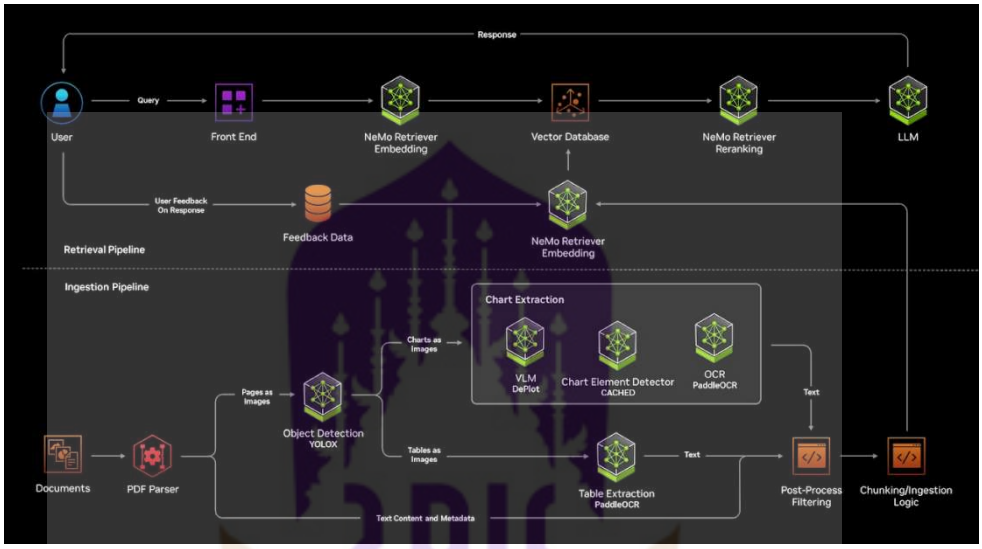
การประยุกต์ใช้ RAG กับไฟล์ PDF

การนำ RAG มาประยุกต์ใช้ในการดึงข้อมูลจากไฟล์ PDF มีขั้นตอนที่ชัดเจนดังนี้ (Lewis et al., 2020):

1) แปลง PDF เป็นข้อความ (Text Extraction): ใช้ไลบรารี เช่น PyPDF2 หรือ pdfplumber เพื่อดึงข้อความจากไฟล์ PDF ออกมาในรูปแบบที่สามารถนำไปใช้งานต่อได้
2) สร้างดัชนีข้อมูล (Indexing): แปลงข้อความที่ดึงมาเป็นเวกเตอร์โดยใช้โมเดล embedding เช่น Sentence Transformers แล้วจัดเก็บในฐานข้อมูลเวกเตอร์ เช่น Pinecone หรือ Weaviate

3) ค้นหาข้อมูล (Search): ใช้อัลกอริทึม เช่น TF-IDF หรือ BM25 ในการค้นหาข้อมูลที่เกี่ยวข้องกับคำถามที่ผู้ใช้ป้อนเข้ามา (Izacard & Grave, 2021)
4) รวมข้อมูลเพื่อสร้างคำตอบ (Augmentation): ส่งข้อมูลที่เกี่ยวข้องไปให้ LLMs เพื่อสร้างคำตอบที่ครอบคลุมและแม่นยำ

ตัวอย่างการใช้งาน RAG กับไฟล์ PDF ได้แก่ ระบบถามตอบอัตโนมัติ ที่ใช้ในการตอบคำถามจากเอกสารคู่มือหรือรายงาน และ การวิเคราะห์ข้อมูล ที่ช่วยสรุปข้อมูลจากเอกสารหลายฉบับ ทำให้การทำงานกับข้อมูลจำนวนมากมีประสิทธิภาพยิ่งขึ้น (Lewis et al., 2020)



ภาพ 5 กระบวนการทำงานของระบบ Multimodal PDF Data Extraction สำหรับ Enterprise RAG (NVIDIA, n.d.)

หมายเหตุ. จาก "Multimodal PDF Data Extraction for Enterprise RAG," โดย NVIDIA, n.d. สืบค้นจาก https://build.nvidia.com/nvidia/multimodal-pdf-data-extraction-for-enterprise-rag/blueprintcard. Copyright NVIDIA.

ภาพที่ 5 ภาพนี้แสดงกระบวนการทำงานของระบบ Multimodal PDF Data Extraction for Enterprise RAG ที่แบ่งออกเป็นสองส่วนหลัก ได้แก่ Ingestion Pipeline และ Retrieval Pipeline เพื่อจัดการและดึงข้อมูลจากไฟล์ PDF ในส่วน Ingestion Pipeline เริ่มต้นด้วยการนำไฟล์ PDF เข้าสู่ระบบผ่าน PDF Parser เพื่อแปลงไฟล์เป็นข้อมูลในรูปแบบข้อความหรือภาพ จากนั้นใช้เทคนิค Object Detection เช่น YOLOX ในการตรวจจับองค์ประกอบ เช่น ตารางหรือกราฟ พร้อมใช้โมดูล PaddleOCR และ VLM DePlot ในการดึงข้อมูลสำคัญ เช่น ข้อความจากตารางและกราฟที่ซับซ้อน โดยข้อมูลที่ได้จะผ่านกระบวนการ Post-Process Filtering เพื่อเพิ่มความสมบูรณ์ ก่อนส่งไปยัง Retrieval Pipeline ซึ่งเป็นส่วนที่ผู้ใช้ป้อนคำถาม (Query) ระบบจะแปลงคำถามและข้อมูลในฐานข้อมูลเป็นเวกเตอร์ด้วย NeMo Retriever เพื่อค้นหาข้อมูลที่

เกี่ยวข้อง และจัดลำดับใหม่ด้วย NeMo Retriever Reranking ข้อมูลที่เกี่ยวข้องจะถูกส่งต่อให้โมเดลภาษาใหญ่ (LLM) เช่น GPT เพื่อสร้างคำตอบที่สมบูรณ์และตอบกลับผู้ใช้ กระบวนการนี้ช่วยเพิ่มความแม่นยำและประสิทธิภาพในการจัดการข้อมูลจากไฟล์ PDF สำหรับการใช้งานในระดับองค์กร

การสรุปผลและจัดทำรายงาน

1) สรุปผลการวิจัย รวบรวมผลการทดลอง ประเมินความสำเร็จตามวัตถุประสงค์ที่ตั้งไว้

2) จัดทำรายงานวิจัย เขียนรายงานที่ครอบคลุมทุกขั้นตอนของกระบวนการวิจัย พร้อมทั้งข้อสมมติฐาน ค้นพบและข้อเสนอแนะสำหรับงานวิจัยในอนาคต ด้วยกระบวนการดังกล่าวว่า Ollama 3.2 ที่ฝึกด้วยเทคนิค RAG จากไฟล์ PDF ของมหาวิทยาลัยพะเยา กรณีศึกษาสถาบันนวัตกรรมการเรียนรู้ สามารถทำงานได้อย่างถูกต้อง และเพิ่มประสิทธิภาพในการเกษียณหนังสือราชการได้

งานวิจัยนี้มีวัตถุประสงค์ในการฝึกโมเดล LLM Ollama 3.2 โดยใช้ข้อมูลจากไฟล์ PDF ของมหาวิทยาลัยพะเยาและประยุกต์ใช้เทคนิค RAG เพื่อเพิ่มประสิทธิภาพ โดยการใช้ GPU จาก Google Colab กระบวนการเริ่มต้นด้วยการรวบรวมไฟล์ PDF ที่เกี่ยวข้อง เช่น หนังสือราชการ เอกสารภายใน หรือคู่มือการปฏิบัติงาน และแปลงข้อมูลเหล่านี้เป็นข้อความดิบด้วย OCR สำหรับเอกสารสแกน หรือใช้เครื่องมืออ่าน PDF เช่น PyPDF2 หรือ Adobe PDF SDK จากนั้นสร้างฐานข้อมูลแบบเวกเตอร์โดยใช้โมเดล embedding เช่น OpenAI's ADA หรือ Sentence Transformers เพื่อแปลงข้อความเป็นเวกเตอร์ที่สะท้อนถึงเนื้อหา และจัดเก็บในฐานข้อมูลที่รองรับการค้นคืน เช่น Pinecone หรือ Weaviate เมื่อเตรียมฐานข้อมูลเสร็จ เราเชื่อมต่อ RAG กับ Ollama 3.2 เพื่อปรับแต่งให้สามารถตอบคำถามและสรุปข้อมูลที่สอดคล้องกับเอกสารในฐานข้อมูลได้อย่างมีประสิทธิภาพ และวัดความถูกต้องของโมเดลด้วยการใช้ตัวชี้วัด เช่น Exact Match (EM) และ F1–Score พร้อมทั้งวัดความสอดคล้องของคำตอบโดยใช้ BLEU หรือ ROUGE เพื่อประเมินความคล้ายคลึงของเนื้อหาและโครงสร้าง รวมถึงตรวจสอบความมั่นใจของโมเดลในแต่ละคำตอบและปรับปรุงกระบวนการฝึกหากจำเป็น โดยเรายังได้ใช้ผู้เชี่ยวชาญในโดเมนที่เกี่ยวข้องเพื่อประเมินคำตอบในแง่ของความถูกต้อง ความสอดคล้อง และการใช้งาน เพื่อให้มั่นใจว่าระบบตอบสนองได้อย่างทันเวลา การทำงานด้วยวิธีการเหล่านี้ช่วยให้มั่นใจว่า Ollama 3.2 ที่ฝึกด้วย RAG จะสามารถทำงานได้อย่างถูกต้องและตอบโจทย์การใช้งานเฉพาะด้านได้อย่างแท้จริง

บทที่ 3

วิธีดำเนินการวิจัย

ในบทนี้ ผู้วิจัยได้แบ่งหัวข้อวิธีการดำเนินวิจัยของการพัฒนารูปแบบการฝึก AI โดยใช้ Large Language Models (LLM) เพื่อเกษียนหนังสือราชการ กรณีศึกษา สถาบันนวัตกรรมการเรียนรู้มหาวิทยาลัยพะเยา ดังนี้

3.1 ประชากรและกลุ่มตัวอย่าง

ก. ประชากรเป้าหมาย

ระยะที่ 1 ช่วงพัฒนาการฝึกให้ผู้เชี่ยวชาญตรวจประเมิน

ผู้เชี่ยวชาญด้านเทคโนโลยีสารสนเทศ 1 ท่าน

ผู้เชี่ยวชาญด้านภาษาไทย 1 ท่าน

ผู้เชี่ยวชาญด้านหนังสือราชการ 1 ท่าน

ระยะที่ 2 ช่วงทดลองใช้ และประเมินความพึงพอใจของระบบฯ วิธีการสุ่มตัวอย่าง ประชากรของการวิจัย

ผู้ที่ใช้ระบบ DMS ในสถาบันนวัตกรรมการเรียนรู้ จำนวน 43 คน (ข้อมูลเมื่อวันที่ 22 พฤศจิกายน 2567) ซึ่งเป็นประชากรมีทราบจำนวนที่แน่ชัด

ข. การเลือกตัวอย่าง

1) เกณฑ์การคัดเลือก (Inclusion criteria) ของประชากรเป้าหมาย

- เป็นบุคลากรสถาบันนวัตกรรมการเรียนรู้
- สามารถอ่านภาษาไทย และใช้คอมพิวเตอร์ได้
- ยินยอมเข้าร่วมการวิจัย

2) เกณฑ์การคัดออก (Exclusion criteria) ของประชากรเป้าหมาย

- ตอบแบบสอบถามไม่ครบตามที่กำหนดไว้
- ไม่สามารถเข้าร่วมการวิจัยในช่วงเวลาที่กำหนด
- ไม่สมัครใจให้ข้อมูลในการตอบแบบสอบถาม

3) เกณฑ์การถอนอาสาสมัคร (Withdrawal criteria) ของประชากรเป้าหมาย

- อาสาสมัครฯ ตัดสินใจออกจากโครงการ

- อาสาสมัครฯ ไม่สามารถปฏิบัติตามคู่มือตามข้อกำหนดของโครงการได้

4) เกณฑ์การยุติโครงการวิจัยก่อนกำหนด (Termination of study criteria) ของประชากรเป้าหมาย

- แหล่งทุน ผู้สนับสนุนการวิจัยยุติการวิจัย
- ผู้มีอำนาจอนุมัติยุติการวิจัย

5) วิธีการจัดผู้เข้าร่วมวิจัยเข้ากลุ่ม (Subject allocation)

- ไม่มีการแบ่งกลุ่ม (ผู้เข้าร่วมวิจัยทุกคนได้รับการปฏิบัติเหมือนกันตลอดการเข้าร่วมวิจัย)

ค. ขนาดตัวอย่าง

เนื่องจากกลุ่มตัวอย่างในระยะที่ 2 ช่วงทดลองใช้ เป็นการคัดเลือกกลุ่มตัวอย่างแบบเจาะจงเป็นบุคลากรของสถาบันนวัตกรรมการเรียนรู้ เนื่องจากบุคลากรสถาบันนวัตกรรมการเรียนรู้ มีการปฏิบัติงานโดยใช้หนังสือราชการ หรือ การเกษียนหนังสือราชการเป็นกระบวนการสื่อสารภายในองค์กรและเป็นเอกสารหลักฐานการปฏิบัติงาน มีจำนวนบุคลากรทั้งหมด 43 คน (ข้อมูลเมื่อวันที่ 22 พฤศจิกายน 2567) ซึ่งเป็นประชากรมีทราบจำนวนที่แน่ชัด ดังนั้นผู้วิจัยจึงใช้สูตรของ Yamane (1967) คำนวณขนาดตัวอย่างสัดส่วน 1 กลุ่ม โดยสมมติค่าสัดส่วนเท่ากับ 0.5 และที่ระดับความเชื่อมั่น 95%

$$n = \frac{N}{1 + Ne^2}$$

โดย n = ขนาดตัวอย่างที่คำนวณได้

N = จำนวนประชากรที่ทราบค่า

e = ค่าความคลาดเคลื่อนที่จะยอมรับได้ (Allowable Error)

ดังนั้น

$$n = \frac{43}{1 + 43(0.05)^2}$$

$$n = \frac{43}{1.1075}$$

$n = 38.8261851016$ ประมาณ 39 คน

3.2 เครื่องมือที่ใช้ในการวิจัย

เครื่องมือวัดประสิทธิภาพของงานวิจัย

1) การทดสอบคุณภาพของเครื่องมือ (Testing the quality of tools) ชุดแบบสอบถามสำหรับงานวิจัย ประเมินความน่าเชื่อถือของแบบสอบถาม ผู้วิจัยจึงได้ใช้แบบสอบถามโดยนำผลลัพธ์ที่ได้มาคำนวณค่าสัมประสิทธิ์อัลฟ่าของ Cronbach's Alpha ใช้โปรแกรมวิเคราะห์ทางสถิติ เช่น SPSS เพื่อคำนวณค่า Cronbach's Alpha
2) ประมวลผลทางสถิติ หาค่าเบี่ยงเบนมาตรฐานจากแบบสอบถามวัดประสิทธิภาพและความพึงพอใจ ระบบ AI “เกษียนหนังสือราชการ” กรณีศึกษา สถาบันนวัตกรรมการเรียนรู้ เพื่อวัดความถูกต้อง ซึ่งประเมินโดยประชากรกลุ่มตัวอย่าง ได้แก่ ผู้เชี่ยวชาญด้านเทคโนโลยีปัญญาประดิษฐ์ (AI) จำนวน 1 ท่าน ผู้เชี่ยวชาญด้านภาษาไทย จำนวน 1 ท่าน ผู้เชี่ยวชาญด้านหนังสือราชการ จำนวน 1 ท่าน ผู้วิจัยได้ใช้การทำแบบสอบถามออนไลน์ผ่านเว็บไซต์ Google Forms เป็นเว็บไซต์ที่สร้างความน่าเชื่อถือและมีความสามารถเลือกเก็บรวบรวมข้อมูลคำตอบได้โดยง่าย โดยใช้แบบสอบถาม มีคำตอบแบบเลือกตอบ (Check List) และการวัดคะแนนตามแบบมาตราวัดของลิเคิร์ท (Rating Scale หรือ Five-Point Likert Scale)
3) ประมวลผลทางสถิติ หาค่าเบี่ยงเบนมาตรฐานจากแบบสอบถามความพึงพอใจในการใช้ AI ในการเกษียนหนังสือราชการของโครงการวิจัย เรื่อง การฝึก AI โดยใช้ Large Language Models (LLM) การเกษียนหนังสือราชการ กรณีศึกษา สถาบันนวัตกรรมการเรียนรู้ มหาวิทยาลัยพะเยา เพื่อวัดความพึงพอใจในการใช้งานระบบว่าเป็นไปตามความต้องการหรือไม่ และมีคุณภาพเหมาะสมในการนำมาใช้งานอย่างไร โดยประชากรกลุ่มตัวอย่าง ได้แก่ ผู้ใช้งานระบบ จำนวน 40 คน ผู้วิจัยได้ใช้การทำแบบสอบถามออนไลน์ผ่านเว็บไซต์ Google Forms เป็นเว็บไซต์ที่สร้างความน่าเชื่อถือและมีความสามารถเลือกเก็บรวบรวมข้อมูลคำตอบได้โดยง่าย โดยใช้แบบสอบถาม มีคำตอบแบบเลือกตอบ (Check List) และการวัดคะแนนตามแบบมาตราวัดของลิเคิร์ท (Rating Scale หรือ Five-Point Likert Scale)

เครื่องมือที่ใช้ในการวิจัย

1) เครื่องมือสนับสนุนการวิเคราะห์และออกแบบระบบ เช่น โปรแกรมประมวลผลทางคอมพิวเตอร์ และโปรแกรมสำเร็จรูปโอเพ่นซอส
2) ซอฟต์แวร์สำหรับเขียนโปรแกรม ประกอบด้วย ภาษาไพทอน ฯลฯ

3) ฮาร์ดแวร์ที่ใช้สนับสนุนการวิจัย

3.1 หน่วยประมวลผลกลาง (CPU) แบบ Intel Core i7 หรือเทียบเท่า และมีหน่วยความจำหลัก (RAM) อย่างน้อย 8 GB

3.2 Hard disk มีความจำอย่างน้อย ไม่น้อยกว่า 80 GB

3.3 สมาร์ทโฟน ที่มีระบบปฏิบัติการแอนดรอย และไอโอเอส

3.4 แท็บเล็ต ที่มีระบบปฏิบัติการแอนดรอย และไอโอเอส

3.3 วิธีดำเนินการวิจัย

แนวทางการพัฒนาระบบ

งานวิจัยนี้มีวัตถุประสงค์ในการฝึกโมเดลภาษาขนาดใหญ่ Ollama 3.2 โดยใช้ข้อมูลจากไฟล์ PDF ของมหาวิทยาลัยพะเยา และประยุกต์ใช้เทคนิค Retrieval-Augmented Generation (RAG) เพื่อเพิ่มประสิทธิภาพ โดยใช้ GPU จาก Google Colab โดยอ้างอิงจาก The Software Development Life Cycle (SDLC) และกระบวนการวิจัยสามารถแบ่งออกเป็นขั้นตอนดังต่อไปนี้

1) วิเคราะห์และสรุปความต้องการของระบบ โดยระบุปัญหาหลัก (ความล่าช้าและความผิดพลาดในการเกษียนหนังสือด้วยมือ), กลุ่มเป้าหมาย (บุคลากรสถาบันนวัตกรรมการเรียนรู้), และเป้าประสงค์ของโครงการ (เพิ่มประสิทธิภาพ, ลดระยะเวลา, สร้างต้นแบบ) เพื่อใช้เป็นข้อมูลในขั้นการวางแผน (Planning) ของ SDLC

2) จัดทำข้อกำหนดคุณลักษณะของระบบ (System Requirements) ทั้งในเชิงฟังก์ชัน (Functional) และไม่ใช่ฟังก์ชัน (Non-functional) :

(a) Functional: ระบบต้องสามารถรับไฟล์ PDF, สกัดข้อความ (OCR), สรุปใจความสำคัญเพื่อร่างคำเกษียน, และแสดงผลลัพธ์แก่ผู้ใช้

(b) Non-functional: ระบบต้องมีความมั่นคงปลอดภัยสูง (ประมวลผลภายในองค์กร), ตอบสนองได้รวดเร็ว, และใช้งานง่ายสำหรับบุคลากรที่ไม่ใช่ผู้เชี่ยวชาญด้านเทคนิค

3) ออกแบบสถาปัตยกรรมของระบบ (System Design) เน้นแบบ On-premise เพื่อรักษาความลับของข้อมูล โดยมีองค์ประกอบหลักดังนี้:

(a) ส่วนนำเข้าข้อมูล (Ingestion Pipeline): กระบวนการรับและแปลงไฟล์ PDF (ทั้งแบบดิจิทัลและสแกน) ให้เป็นข้อความดิบ

(b) ส่วนจัดการองค์ความรู้ (Knowledge Management): การสร้างฐานข้อมูลเวกเตอร์ (Vector Database) จากเอกสารราชการโดยใช้ Embedding Models โดยการรวบรวมข้อมูล รวบรวมไฟล์ PDF ที่เกี่ยวข้อง

ซึ่งสามารถจำแนกขั้นตอนการจัดการองค์ความรู้ ดังนี้

- รวบรวมเอกสารต่าง ๆ จากมหาวิทยาลัยพะเยา เช่น หนังสือราชการ เอกสารรับภายใน เอกสารรับภายนอก พร้อมตรวจสอบความครบถ้วนและความถูกต้องของเอกสารที่รวบรวม
- การเตรียมข้อมูล

  – แปลงไฟล์ PDF เป็นข้อความ สำหรับเอกสารสแกน ใช้เทคนิค OCR (Optical Character Recognition) เพื่อแปลงภาพเป็นข้อความ

  – สำหรับเอกสารดิจิทัล ใช้เครื่องมืออ่าน PDF เช่น PyPDF2 หรือ Adobe PDF SDK เพื่อดึงข้อความ

  – ทำความสะอาดและจัดรูปแบบข้อมูล ลบอักขระที่ไม่จำเป็น สัญลักษณ์พิเศษ หรือข้อมูลที่ไม่เกี่ยวข้อง พร้อมตรวจสอบและแก้ไขข้อผิดพลาดที่เกิดขึ้นจากกระบวนการ OCR
- การสร้างเวกเตอร์จากข้อความ

  – เลือกโมเดล Embedding ที่เหมาะสม ใช้โมเดล OpenAI's ADA หรือ Sentence Transformers เพื่อแปลงข้อความเป็นเวกเตอร์

  – แปลงข้อความเป็นเวกเตอร์ นำข้อความที่เตรียมไว้มาแปลงเป็นเวกเตอร์ที่สะท้อนถึงเนื้อหาและความหมาย
- การสร้างฐานข้อมูลเวกเตอร์

  – เลือกฐานข้อมูลที่รองรับการค้นคืน ใช้ฐานข้อมูลเช่น Pinecone หรือ Weaviate สำหรับการจัดเก็บเวกเตอร์

  – จัดเก็บเวกเตอร์ลงในฐานข้อมูล นำเวกเตอร์ที่ได้จากขั้นตอนก่อนหน้ามาจัดเก็บในฐานข้อมูลที่เลือก
- การเชื่อมต่อ RAG กับ Ollama 3.2

  – ปรับแต่ง Ollama 3.2 ปรับแต่ง (Fine-tune) โมเดลให้สามารถตอบคำถามและสรุปข้อมูลที่สอดคล้องกับเอกสารในฐานข้อมูล

  – เชื่อมต่อระบบ RAG ตั้งค่าการค้นคืนข้อมูลจากฐานข้อมูลเวกเตอร์เมื่อมีคำถามหรือข้อความเข้ามา พร้อมนำผลลัพธ์ที่ค้นคืนได้ส่งต่อให้ Ollama 3.2 เพื่อสร้างคำตอบหรือสรุปเนื้อหา
- การประเมินความถูกต้องของโมเดล

– วัดความถูกต้องด้วยตัวชี้วัด ใช้ Exact Match (EM) และ F1–Score เพื่อวัดว่าคำตอบที่โมเดลสร้างขึ้นตรงกับคำตอบที่ถูกต้องหรือไม่

– วัดความสอดคล้องของคำตอบ ใช้ BLEU หรือ ROUGE เพื่อประเมินความคล้ายคลึงของเนื้อหาและโครงสร้างของคำตอบ

(c) ส่วนประมวลผลหลัก (Core AI Engine): การติดตั้งและปรับแต่ง LLM แบบโอเพนซอร์ส (Ollama) และเชื่อมต่อกับสถาปัตยกรรม Retrieval–Augmented Generation (RAG) เพื่อดึงข้อมูลจากฐานข้อมูลเวกเตอร์มาใช้สร้างคำตอบ

(d) ส่วนติดต่อผู้ใช้ (User Interface): การออกแบบหน้าจอที่เรียบง่ายสำหรับให้ผู้ใช้อัปโหลดเอกสารและรับผลการสรุป

4) วางแผนขั้นตอนการพัฒนาและติดตั้งระบบ (Development & Implementation) โดยละเอียดตามระเบียบวิธีวิจัยที่เสนอไว้ ได้แก่ การรวบรวมและเตรียมข้อมูล, การสร้างและจัดเก็บเวกเตอร์, การเชื่อมต่อ RAG กับ Ollama, และการพัฒนาระบบทั้งหมดให้ทำงานร่วมกัน

5) กำหนดแผนการทดสอบและประเมินผล (Testing & Evaluation):

(a) การประเมินเชิงปริมาณ: วัดความถูกต้องและความสอดคล้องของผลลัพธ์ด้วยตัวชี้วัดมาตรฐาน เช่น Exact Match (EM), F1–Score, BLEU, และ ROUGE

(b) การประเมินโดยผู้เชี่ยวชาญ: ให้ผู้เชี่ยวชาญด้านภาษาไทยและเทคโนโลยีสารสนเทศตรวจสอบคุณภาพของคำเกษียนที่ AI สร้างขึ้น

- การตรวจสอบความแม่นยำของโมเดล

  – วิเคราะห์คะแนนความแม่นยำ ตรวจสอบคะแนนความแม่นยำที่โมเดลให้ในแต่ละคำตอบ

  – ปรับปรุงกระบวนการฝึก หากพบว่าโมเดลมีค่าความแม่นยำผิดพลาด ปรับปรุงการฝึกด้วยการเพิ่มข้อมูลฝึกหรือปรับพารามิเตอร์

(c) การทดสอบโดยผู้ใช้งานจริง: ให้กลุ่มเป้าหมาย (บุคลากร 43 คน) ทดลองใช้งานระบบและประเมินประสิทธิภาพและความพึงพอใจ โดยมีการไม่มีการแบ่งกลุ่ม (ผู้เข้าร่วมวิจัยทุกคนได้รับการปฏิบัติเหมือนกันตลอดการเข้าร่วมวิจัย)

(6) สรุปแนวทางการนำไปใช้งานจริง (Deployment), การบำรุงรักษา (Maintenance), และการพัฒนาต่อยอดในอนาคต (Future Enhancement) โดยเสนอแนะให้มีการนำร่องใช้งานในกลุ่มเล็กก่อนขยายผล, วางแผนการปรับปรุงฐานข้อมูลและโมเดลอย่างสม่ำเสมอ, และสำรวจ

แนวทางการพัฒนาต่อยอด เช่น การสร้างระบบแบบผสมผสาน (Hybrid Model) หรือการวิเคราะห์ต้นทุนและผลประโยชน์ในระยะยาว

3.4 การหาค่าดัชนีความสอดคล้อง IOC

แบบประเมินหาค่าดัชนีความสอดคล้อง IOC

คำชี้แจง

1. แบบประเมินฉบับนี้มีทั้งหมด 3 ตอน
   ตอนที่ 1 แบบประเมินเกี่ยวกับข้อมูลทั่วไปของผู้ตอบแบบสอบถาม
   ตอนที่ 2 แบบประเมินระดับความพึงพอใจในการใช้ ระบบ AI “เกษียณหนังสือราชการ” กรณีศึกษา สถาบันนวัตกรรมการเรียนรู้
   ตอนที่ 3 แบบประเมินวัดประสิทธิภาพคำตอบของ AI โดยใช้ Large Language Models (LLM) เพื่อเกษียนหนังสือราชการ กรณีศึกษา สถาบันนวัตกรรมการเรียนรู้ มหาวิทยาลัยพะเยา
2. ขอความกรุณาผู้ทรงคุณวุฒิหรือท่านผู้เชี่ยวชาญ ช่วยพิจารณาร่างแบบประเมินว่ามีความสอดคล้องกับตัวแปรของการวิจัยเรื่องนี้หรือไม่ ด้วยการให้คะแนนในแต่ละข้อคำถามในระบบ IOC โดยการทำเครื่องหมาย (✓) ลงในช่องความคิดเห็นของท่าน
   เกณฑ์การให้คะแนนในระบบ IOC
   1) ให้คะแนน +1 เมื่อแน่ใจว่าข้อนั้นมีเนื้อหาที่เหมาะสม
   2) ให้คะแนน 0 เมื่อไม่แน่ใจว่าข้อนั้นมีเนื้อหาที่เหมาะสม
   3) ให้คะแนน -1 เมื่อแน่ใจว่าข้อนั้นมีเนื้อหาที่ไม่เหมาะสม
3. ผู้วิจัยขอความกรุณาท่านผู้ทรงคุณวุฒิหรือท่านผู้เชี่ยวชาญ ให้ข้อเสนอแนะหรือความคิดเห็นเพิ่มเติมในประเด็นที่ยังไม่สมบูรณ์ โดยการเขียนข้อเสนอแนะไว้ท้ายข้อความนั้น ๆ

ตอนที่ 1 แบบประเมินเกี่ยวกับข้อมูลทั่วไปของผู้ตอบแบบสอบถาม

ข้อที่	รายการขอความคิดเห็น	ประมาณค่าความคิดเห็นของผู้ทรงคุณวุฒิ			ค่า IOC	แปลผล
		คนที่ 1	คนที่ 2	คนที่ 3		
1	เพศ เพศ ( ) ชาย ( ) หญิง	+1	+1	+1	1.0	ใช้ได้
2	อายุ (เกิดในช่วง พ.ศ.)	+1	+1	+1	1.0	ใช้ได้

ข้อที่	รายการขอความคิดเห็น	ประมาณค่าความคิดเห็นของผู้ทรงคุณวุฒิ			ค่า IOC	แปลผล
		คนที่ 1	คนที่ 2	คนที่ 3		
	( ) อายุต่ำกว่า 27 ปี (พ.ศ. 2541-2565) Gen Z ( ) อายุระหว่าง 28 - 45 ปี (พ.ศ. 2523-2540) Gen Y ( ) อายุระหว่าง 46 - 60 ปี (พ.ศ. 2508-2522) Gen X					
3	ตำแหน่งที่ปฏิบัติงาน ( ) ผู้บริหาร - สายวิชาการ ( ) หัวหน้าสนง./ผอ.ศูนย์/หัวหน้างาน ( ) เจ้าหน้าที่บริหารงานทั่วไป ( ) นักวิชาการศึกษา ( ) นักวิเคราะห์นโยบายและแผน ( ) นักวิชาการเงินและบัญชี ( ) บรรณารักษ์ ( ) ผู้ปฏิบัติงานห้องสมุด ( ) นักวิชาการคอมพิวเตอร์ ( ) นักวิชาการโสตทัศนศึกษา	+1	+1	+1	1.0	ใช้ได้
4	ประสบการณ์ในการทดลองใช้ AI เกี่ยวกับเอกสารราชการ ( ) ไม่เคยใช้ ( ) ใช้มาแล้วไม่เกิน 3 เดือน ( ) ใช้มาแล้วไม่เกิน 6 เดือน ( ) ใช้มาแล้วน้อยกว่า 1 ปี ( ) ใช้มาแล้วเกิน 1 ปี	+1	+1	+1	1.0	ใช้ได้
5	ระดับการศึกษา ( ) ต่ำกว่าระดับปริญญาตรี ( ) ปริญญาตรี	+1	+1	+1	1.0	ใช้ได้

ข้อที่	รายการขอความคิดเห็น	ประมาณค่าความคิดเห็นของผู้ทรงคุณวุฒิ			ค่า IOC	แปลผล
		คนที่ 1	คนที่ 2	คนที่ 3		
	( ) ปริญญาโท ( ) ปริญญาเอก					
6	อายุงาน ( ) น้อยกว่า 1 ปี ( ) 1 - 5 ปี ( ) 6 - 10 ปี ( ) 11 - 15 ปี ( ) 16 - 20 ปี ( ) มากกว่า 20 ปี	+1	+1	+1	1.0	ใช้ได้

ตอนที่ 2 ความพึงพอใจในการใช้งานระบบ AI เกษียณหนังสือราชการ

ข้อที่	รายการขอความคิดเห็น	ประมาณค่าความคิดเห็นของผู้ทรงคุณวุฒิ			ค่า IOC	แปลผล
		คนที่ 1	คนที่ 2	คนที่ 3		
1	การใช้ AI เกษียนหนังสือราชการ ช่วยลดระยะเวลาในการทำงาน	+1	+1	+1	1.0	ใช้ได้
2	การใช้ AI เกษียณหนังสือราชการ ช่วยเพิ่มผลงานลดต้นทุนกำลังบุคคลในการปฏิบัติงานให้แก่องค์กร	+1	+1	+1	1.0	ใช้ได้
3	การใช้ AI เกษียณหนังสือราชการ ช่วยลดข้อผิดพลาดที่เกิดจากมนุษย์	+1	+1	+1	1.0	ใช้ได้
4	การใช้ AI เกษียณหนังสือราชการ ตรงต่อความต้องการของผู้ใช้ด้านความสะดวก และรวดเร็ว	+1	0	0	0.3	ปรับปรุง
5	การใช้ AI เกษียณหนังสือราชการ ส่งผลกระทบเชิงลบต่อบุคลากรและองค์กร	+1	-1	+1	0.3	ปรับปรุง

ตอนที่ 3 แบบประเมินวัดประสิทธิภาพคำตอบของ AI โดยใช้ Large Language Models (LLM) เพื่อเกษียนหนังสือราชการ กรณีศึกษา สถาบันนวัตกรรมการเรียนรู้ มหาวิทยาลัยพะเยา

ข้อที่	รายการขอความคิดเห็น	ประมาณค่าความคิดเห็นของผู้ทรงคุณวุฒิ			ค่า IOC	แปลผล
		คนที่ 1	คนที่ 2	คนที่ 3		
1	การพัฒนารูปแบบการฝึก AI ด้วย LLM (เช่น การใช้ Ollama 3.2 และเทคนิค RAG)	+1	+1	0	0.6	ใช้ได้
2	การนำ AI ไปประยุกต์ใช้ในการเกษียนหนังสือราชการในระบบ UP-DMS ได้อย่างมีประสิทธิภาพ	+1	+1	+1	1.0	ใช้ได้
3	AI มีประสิทธิภาพในการลดระยะเวลาและข้อผิดพลาดในการ เกษียณหนังสือราชการ	+1	+1	+1	1.0	ใช้ได้
4	การใช้ OCR ในการแปลงเอกสาร (PDF) เป็นข้อความเพื่อนำไปประมวลผลด้วย LLM	+1	+1	+1	1.0	ใช้ได้
5	การประเมินประสิทธิภาพของ AI ที่พัฒนาขึ้น (ด้วยตัวชี้วัด Exact Match, F1-score, BLEU, ROUGE)	+1	+1	+1	1.0	ใช้ได้
6	ประโยชน์ที่คาดว่าจะได้รับ ได้รูปแบบ AI ที่บุคคลทั่วไปสามารถนำไปต่อยอด, การทำงานสารบรรณมีประสิทธิภาพขึ้น เป็นต้น	+1	0	+1	0.6	ใช้ได้

$$\text{ค่า IOC} = \frac{1.0+1.0+1.0+1.0+1.0+1.0+1.0+1.0+1.0+0.3+0.3+0.6+1.0+1.0+1.0+1.0+0.6}{17} = \frac{14}{17} = 0.8235$$

สรุปว่า แบบสอบถามนี้สามารถใช้ได้

3.5 การเก็บรวบรวมข้อมูล

ผู้วิจัยใช้วิธีส่งหนังสือขอความอนุเคราะห์ตอบแบบสอบถาม และดำเนินการเก็บรวบรวมข้อมูลจากการตอบแบบสอบถาม ในช่วงเดือนกรกฎาคม 2568

อุปกรณ์ที่ใช้ : คอมพิวเตอร์

เครื่องมือที่ใช้ในการวิจัย คือ

- แบบสอบถามความพึงพอใจในการใช้ AI ในการเกษียนหนังสือราชการของ

โครงการวิจัย เรื่อง การฝึก AI โดยใช้ Large Language Models (LLM) การเกษียนหนังสือราชการ กรณีศึกษา สถาบันนวัตกรรมการเรียนรู้ มหาวิทยาลัยพะเยา ของ แบ่งออกเป็น 3 ตอน ได้แก่

ตอนที่ 1 ข้อมูลทั่วไปของผู้ตอบแบบประเมิน ประกอบด้วย เพศ ประเภทผู้ใช้บริการ สังกัดสถาบันนวัตกรรมการเรียนรู้

ตอนที่ 2 การใช้ AI ในการเกษียนหนังสือราชการ

1) การใช้ AI เกษียณหนังสือราชการ จะลดระยะเวลาในการทำงานได้มากกว่าการทำงานรูปแบบเดิม

2) การใช้ AI เกษียณหนังสือราชการ จะช่วยเพิ่มผลงานให้แก่องค์กร

3) การใช้ AI เกษียณหนังสือราชการ ช่วยลดข้อผิดพลาดที่เกิดจากมนุษย์

4) การใช้ AI เกษียณหนังสือราชการ ตรงต่อความต้องการในการใช้งาน

5) การใช้ AI เกษียณหนังสือราชการ ส่งผลเสียให้กับบุคลากรและองค์กร

ตอนที่ 3 ข้อเสนอแนะ เป็นคำถามปลายเปิด

ลักษณะของแบบประเมินเป็นแบบมาตรส่วนประมาณค่า (rating scale) ระดับของลิเคิร์ท (Likert rating scales)ผู้วิจัยกำหนดน้ำหนักหรือคะแนนในการทำแบบสอบถาม ดังนี้

ระดับมากที่สุด	ให้น้ำหนักหรือคะแนนเป็น	5
ระดับมาก	ให้น้ำหนักหรือคะแนนเป็น	4
ระดับปานกลาง	ให้น้ำหนักหรือคะแนนเป็น	3
ระดับน้อย	ให้น้ำหนักหรือคะแนนเป็น	2
ระดับน้อยที่สุด	ให้น้ำหนักหรือคะแนนเป็น	1

ในการแปลความหมายของคะแนนผู้วิจัยกำหนดเกณฑ์ ดังนี้

คะแนนเฉลี่ยตั้งแต่	4.51–5.00	หมายถึง ระดับมากที่สุด
คะแนนเฉลี่ยตั้งแต่	3.51–4.50	หมายถึง ระดับมาก
คะแนนเฉลี่ยตั้งแต่	2.51–3.50	หมายถึง ระดับปานกลาง
คะแนนเฉลี่ยตั้งแต่	1.51–2.50	หมายถึง ระดับน้อย
คะแนนเฉลี่ยตั้งแต่	1.00–1.50	หมายถึง ระดับน้อยที่สุด

3.5 การวิเคราะห์ข้อมูล

การประเมินโดยผู้เชี่ยวชาญ

1) เลือกผู้เชี่ยวชาญในโดเมนที่เกี่ยวข้อง คัดเลือกบุคคลที่มีความรู้ความเชี่ยวชาญในเนื้อหาของเอกสารที่ใช้
2) ประเมินคำตอบของโมเดล ให้ผู้เชี่ยวชาญประเมินคำตอบในแง่ของความถูกต้อง ความสอดคล้อง และการใช้งาน
3) รวบรวมความคิดเห็นและข้อเสนอแนะ นำผลการประเมินมาใช้ปรับปรุงโมเดลและกระบวนการฝึกเพิ่มเติม

การวัดประสิทธิภาพของระบบ

1) วัดเวลาในการค้นคืนและการสร้างคำตอบ ประเมินเวลาเฉลี่ยที่ระบบใช้ในการค้นคืนข้อมูลและสร้างคำตอบ
2) ตรวจสอบการตอบสนองของระบบ ทดสอบระบบภายใต้สภาวะการใช้งานจริง เพื่อให้แน่ใจว่าระบบตอบสนองได้ทันเวลา

การปรับปรุงและทดสอบซ้ำ

1) ปรับปรุงโมเดลตามผลการประเมิน ใช้ข้อมูลจากขั้นตอนการประเมินเพื่อปรับปรุงโมเดลและกระบวนการฝึก
2) ทดสอบระบบซ้ำ ทำการทดสอบระบบอีกครั้งเพื่อวัดประสิทธิภาพและความถูกต้องหลังการปรับปรุง

การวัดประสิทธิภาพด้านความพึงพอใจ

1) คัดเลือกอาสาสมัครตามเกณฑ์การคัดเลือกและขอให้กรอกแบบสอบถามโดยให้อาสาสมัครปฏิบัติตามขั้นตอนได้แก่ อ่านเอกสารคู่มือการทดลองใช้ AI ในการเกษียนหนังสือราชการของโครงการวิจัยฯ หรือดูวิดีโอคลิปการสอนการทดลองใช้ AI ดังกล่าวฯ ทำการทดลองใช้ AI ในการเกษียนหนังสือราชการฯ และตอบแบบสอบถามประเมินความพึงพอใจในการใช้ AI ในการเกษียนหนังสือราชการดังกล่าว
2) รวบรวมความคิดเห็นและข้อเสนอแนะ นำผลการประเมินมาใช้ปรับปรุงโมเดลและกระบวนการฝึกเพิ่มเติม
3) วิเคราะห์ข้อมูลด้วยโปรแกรมวิเคราะห์ข้อมูลสำเร็จรูป วิเคราะห์ค่าเฉลี่ย ร้อยละส่วนเบี่ยงเบนมาตรฐาน

3.6 ด้านจริยธรรมการวิจัย

ก. ความเสี่ยงที่อาจเกิดต่ออาสาสมัครที่เข้าร่วมในโครงการวิจัย

- อาสาสมัครในโครงการวิจัยบางรายอาจมีความเสี่ยงต่อจิตใจ หรือมีความกังวลต่อระยะเวลาที่กรอกแบบสอบถาม

ข. มาตรการในการลดความเสี่ยงหรือการวางแผนแก้ไขปัญหาที่อาจเกิดขึ้น

- ผู้วิจัยคำนึงถึงวิธีการที่จะไม่เป็นการรบกวนและการทำงานตามปกติของกลุ่มตัวอย่าง และพึงหลีกเลี่ยงวิธีการที่จะก่อให้เกิดความเครียด/ความวิตกกังวลให้กับกลุ่มตัวอย่าง เช่น กระชับเวลาในการอธิบายการทดลองใช้ นำเสนอประเด็นที่สำคัญและเป็นประโยชน์ที่อาสาสมัครจะได้รับในการประชุมบุคลากรสถาบันนวัตกรรมการเรียนรู้
- ในกรณีที่ได้รับความเสี่ยง หรือต้องการข้อมูลเพิ่มเติมที่เกี่ยวข้องกับโครงการวิจัยอาสาสมัครสามารถติดต่อกับผู้ทำวิจัยคือ นางสาวญภคกร ไชยมงคล หัวหน้าโครงการวิจัยได้ตลอด 24 ชั่วโมง และสามารถนัดพบผู้วิจัยหรือผู้ร่วมทำวิจัยทั้งสิ้นได้ในเวลาราชการจำนวน 3 ครั้ง

ค. กระบวนการขอความยินยอมจากอาสาสมัครเป็นลายลักษณ์อักษร

- เวียนแจ้งหนังสือขอความอนุเคราะห์ตอบแบบสอบถามฯพร้อมข้อมูลคำอธิบายสำหรับผู้เข้าร่วมในโครงการวิจัยฯ และหนังสือแสดงความยินยอมเข้าร่วมโครงการวิจัยสำหรับอาสาสมัครอายุมากกว่า 18 ปีขึ้นไป (Informed Consent Form)
- รวบรวมหลักฐานแสดงความยินยอมเข้าร่วมโครงการวิจัยสำหรับอาสาสมัครอายุมากกว่า 18 ปีขึ้นไป (Informed Consent Form) ทางอีเมล์หรือกล่องรับเอกสารของธุรการสถาบันนวัตกรรมการเรียนรู้ หรือทางระบบ UP-DMS ส่งให้นางสาวญภคกร ไชยมงคล

ง. วิธีการเข้าถึงประชากรกลุ่มเป้าหมายหรืออาสาสมัคร

- ดำเนินการประชาสัมพันธ์ช่องทางการประชุมบุคลากรสถาบันนวัตกรรมการเรียนรู้ หรือช่องทางกลุ่มไลน์สถาบันนวัตกรรมการเรียนรู้

จ. ประเด็นความอ่อนไหวทางสังคม

- ในการตั้งคำถามแบบสอบถามคำถึงถึงประเด็นอ่อนไหวทางสังคมโดยยึดหลักการยุติธรรม เคารพความหลากหลายทางความคิดและข้อเสนอแนะที่มีต่อระบบ

ฉ. การรักษาความลับและแผนการทำลายข้อมูลของอาสาสมัคร

- ไม่เก็บข้อมูลชื่อ-นามสกุล,อีเมล์ของผู้ตอบแบบสอบถาม โดยผู้วิจัยไม่สามารถระบุตัวตนของผู้ตอบแบบประเมินได้
- เมื่อสิ้นสุดโครงการวิจัย ข้อมูลการตอบแบบประเมินความพึงพอใจจะถูกทำลายทันที

มาตรการ	**กิจกรรม**	**ระยะเวลา**
การรักษาความลับและทำลายข้อมูล	เก็บข้อมูลจากกลุ่มเป้าหมาย	1 – 15 กรกฎาคม 2568
	วิเคราะห์ข้อมูลจากกลุ่มเป้าหมาย และเก็บรักษาไฟล์ข้อมูลทั้งหมดที่วิเคราะห์แล้ว ไว้ในคอมพิวเตอร์แบบเข้ารหัสเฉพาะบุคคล จนกว่าจะสิ้นสุดการวิจัย	
	การเข้าถึงข้อมูลใดๆ ของกลุ่มเป้าหมาย ที่อยู่ระหว่างการวิเคราะห์/พิจารณา รวมทั้งข้อมูลที่ดำเนินการเสร็จสิ้นแล้ว จำกัดให้เข้าได้เฉพาะผู้วิจัยและคณะกรรมการที่เกี่ยวข้องเท่านั้น	16 กรกฎาคม – 31 สิงหาคม 2568
	การทำลายข้อมูลแบบสอบถามที่เก็บข้อมูลจากกลุ่มเป้าหมาย เมื่อสิ้นสุดการดำเนินการวิจัยแล้วอย่างน้อย 1 เดือน และจะลบข้อมูลทั้งหมดออกจากระบบคอมพิวเตอร์	1 – 30 กันยายน 2568

3.7 ระยะเวลาที่จะทำการวิจัยและการบริหารจัดการ

ขั้นตอนการวิจัย	เดือน						
	ม.ค.	ก.พ.	มี.ค.	เม.ย.	พ.ค.	มิ.ย.	ก.ค.
1. รวมรวมข้อมูลไฟล์เอกสาร PDF และเตรียมข้อมูล							
2. สร้างเวกเตอร์จากข้อความ							

ขั้นตอนการวิจัย	เดือน						
	ม.ค.	ก.พ.	มี.ค.	เม.ย.	พ.ค.	มิ.ย.	ก.ค.
3. สร้างฐานข้อมูลจากเวกเตอร์							
4. เชื่อมต่อ RAG กับ Ollama 3.2							
5. ประเมินความถูกต้องของโมเดล							
6. วัดประสิทธิภาพของระบบ							
7. ประเมินโดยผู้เชี่ยวชาญ							
8. ปรับปรุงและทดสอบซ้ำ							

1) การเตรียมข้อมูลเบื้องต้น

- ตั้งแต่ได้รับอนุมัติโครงการวิจัยในการจัดทำสัญญาทุนวิจัย ในระหว่างเดือนมกราคม – กุมภาพันธ์ 2568 (รวมรวมข้อมูลไฟล์เอกสาร PDF และเตรียมข้อมูล

2) ระยะเวลาดำเนินการวิจัย

- เดือนมีนาคม – มิถุนายน 2568 เป็นช่วงที่พัฒนาและทดสอบรูปแบบการฝึก AI โดยใช้ Large Language Models (LLM) เพื่อเกษียนหนังสือราชการกรณีศึกษา สถาบันนวัตกรรมการเรียนรู้

3) ระยะเวลาการรวบรวมและวิเคราะห์ข้อมูล

- เดือนกรกฎาคม 2568 เป็นช่วงที่ให้อาสาสมัครกรอกแบบสอบถามเพื่อวัดประสิทธิภาพด้านความพึงพอใจของการใช้ AI ในการเกษียนหนังสือราชการฯ

4) ระยะเวลาการนำเสนอผลการวิจัย และการเขียนรายงาน

- เดือนสิงหาคม 2568

หมายเหตุ ให้นักวิจัยระบุรายละเอียดในแต่ละขั้นตอนเป็นตารางให้ชัดเจน

บทที่ 4
ผลการดำเนินงานและวิเคราะห์ข้อมูล

การวิจัยเรื่อง "การพัฒนารูปแบบการฝึก AI โดยใช้ Large Language Models (LLM) เพื่อเกษียนหนังสือราชการ กรณีศึกษา สถาบันนวัตกรรมการเรียนรู้ มหาวิทยาลัยพะเยา" มีวัตถุประสงค์เพื่อพัฒนารูปแบบการฝึก AI และศึกษาผลการนำ AI มาใช้เพิ่มประสิทธิภาพในกระบวนการเกษียนหนังสือราชการ ในบทนี้จะนำเสนอผลการดำเนินงานตามระเบียบวิธีวิจัยที่ได้วางแผนไว้ ซึ่งแบ่งออกเป็น 2 ส่วนหลัก ได้แก่ ผลการพัฒนาระบบต้นแบบ และผลการประเมินประสิทธิภาพของระบบ

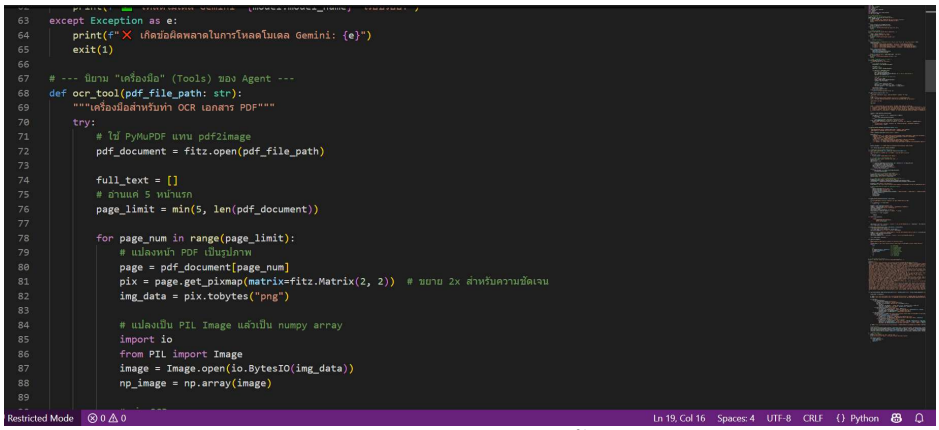
ส่วนที่ 1: ผลการพัฒนาระบบต้นแบบ AI เพื่อการเกษียนหนังสือราชการ

คณะผู้วิจัยได้ดำเนินการพัฒนาระบบต้นแบบตามสถาปัตยกรรมที่ออกแบบไว้ โดยมีขั้นตอนและผลการดำเนินงานในแต่ละส่วนดังนี้

1.1 การรวบรวมและเตรียมข้อมูล (Data Collection and Preparation)

ได้ทำการรวบรวมเอกสารราชการประเภทหนังสือรับเข้า ทั้งจากภายในและภายนอกของสถาบันนวัตกรรมการเรียนรู้ ในช่วงปีงบประมาณ 2567–2568 มาเป็นชุดข้อมูลตั้งต้น จากนั้นได้ดำเนินการเตรียมข้อมูลตามขั้นตอนที่วางแผนไว้

- **การแปลงเอกสารเป็นข้อความ:** เอกสารที่เป็นไฟล์ดิจิทัล (Born-Digital) ถูกสกัดข้อความโดยใช้ไลบรารี PyPDF2 ในขณะที่เอกสารที่มาจากการสแกนได้ถูกแปลงเป็นข้อความด้วยเทคโนโลยี Optical Character Recognition (OCR) โดยใช้ Tesseract OCR Engine ซึ่งพบว่ามีความแม่นยำในการแปลงเอกสารภาษาไทยอยู่ในเกณฑ์ดี



ภาพ 6 แสดงภาพโค้ด

- **การทำความสะอาดข้อมูล:** ข้อความดิบที่ได้จากขั้นตอนแรกได้ผ่านกระบวนการทำความสะอาดข้อมูลเพื่อลบองค์ประกอบที่ไม่ใช่เนื้อหาหลัก เช่น หัวกระดาษ ท้ายกระดาษ เลขหน้า และสัญลักษณ์พิเศษ เพื่อให้ได้ชุดข้อมูลที่มีคุณภาพสูงสุดสำหรับขั้นตอนต่อไป

1.2 การสร้างฐานข้อมูลองค์ความรู้ (Knowledge Base Construction)

ข้อมูลข้อความที่ผ่านการทำความสะอาดแล้ว ถูกนำมาสร้างเป็นฐานข้อมูลเวกเตอร์เพื่อใช้เป็นคลังความรู้สำหรับระบบ Retrieval-Augmented Generation (RAG)

- **การแบ่งส่วนเอกสาร (Chunking):** ข้อความจากเอกสารแต่ละฉบับถูกแบ่งออกเป็นส่วนย่อยๆ (Chunks) โดยใช้กลยุทธ์การแบ่งตามย่อหน้า (Paragraph-based chunking) เพื่อรักษาความสมบูรณ์ของเนื้อหาในแต่ละส่วน
- **การสร้างเวกเตอร์ (Vector Embedding):** แต่ละ Chunk ถูกแปลงเป็นเวกเตอร์ตัวเลขโดยใช้แบบจำลอง Sentence Transformers ซึ่งมีความสามารถในการทำความเข้าใจและเข้ารหัสความหมายของข้อความภาษาไทยได้เป็นอย่างดี
- **การจัดเก็บในฐานข้อมูลเวกเตอร์:** เวกเตอร์ทั้งหมดพร้อมด้วยข้อมูลอภิพันธุ์ (Metadata) เช่น ชื่อเอกสารและเลขหน้า ถูกนำไปจัดเก็บในฐานข้อมูลเวกเตอร์ Weaviate ซึ่งถูกติดตั้งบนเซิร์ฟเวอร์ภายในของมหาวิทยาลัย เพื่อให้สามารถค้นคืนข้อมูลได้อย่างรวดเร็วและปลอดภัย

1.3 การพัฒนาระบบ AI หลักและการเชื่อมต่อ (Core AI Engine and Integration)

แกนหลักของระบบคือแบบจำลองภาษาขนาดใหญ่ (LLM) ที่ทำงานร่วมกับสถาปัตยกรรม RAG ซึ่งได้ดำเนินการติดตั้งและพัฒนาสำเร็จลุล่วง

- **การติดตั้ง LLM:** ได้ทำการติดตั้ง Ollama บนเซิร์ฟเวอร์ของมหาวิทยาลัย และดาวน์โหลดแบบจำลองภาษา Llama 3 เพื่อใช้งานเป็น LLM หลักของระบบ การทำงานบนเซิร์ฟเวอร์ภายใน (On-premise) ช่วยรับประกันว่าข้อมูลหนังสือราชการทั้งหมดจะถูกประมวลผลภายในเครือข่ายที่ปลอดภัยของมหาวิทยาลัย
- **การเชื่อมต่อระบบ RAG:** ได้พัฒนาไปป์ไลน์ RAG ด้วยภาษา Python โดยใช้ไลบรารี LangChain เพื่อเชื่อมต่อองค์ประกอบทั้งหมดเข้าด้วยกัน เมื่อผู้ใช้ป้อนคำสั่ง ระบบจะทำการค้นหาข้อมูลที่เกี่ยวข้องจากฐานข้อมูล Weaviate และนำข้อมูลนั้นมาประกอบเป็นบริบท (Context) เพื่อส่งให้ Ollama สร้างคำตอบหรือบทสรุปที่มีความถูกต้องและยึดโยงกับเอกสารต้นฉบับ

ส่วนที่ 2: ผลการประเมินประสิทธิภาพของระบบ

หลังจากพัฒนาระบบต้นแบบเสร็จสิ้น ได้มีการประเมินประสิทธิภาพใน 3 มิติ ตามที่ระบุไว้ในระเบียบวิธีวิจัย คือ การประเมินเชิงปริมาณ, การประเมินโดยผู้เชี่ยวชาญ, และการประเมินโดยผู้ใช้งานจริง

ภาพ 7 แสดงโปสเตอร์เชิญชวนให้ผู้เชี่ยวชาญระบบประเมิน

ภาพ 8 แสดงโปสเตอร์เชิญชวนให้ผู้ใช้ระบบประเมิน

2.1 ผลการประเมินเชิงปริมาณ

ได้ทำการทดสอบระบบด้วยชุดข้อมูลทดสอบที่จัดเตรียมไว้ และวัดผลด้วยตัวชี้วัดมาตรฐานสำหรับงานประมวลผลภาษาธรรมชาติ ผลลัพธ์สรุปได้ดังตารางที่ 4.1

ตารางที่ 4.1: ผลการประเมินประสิทธิภาพของระบบด้วยตัวชี้วัดเชิงปริมาณ

ภารกิจ (Task)	ตัวชี้วัด (Metric)	คะแนน (Score)	การตีความผล
การตอบคำถามเชิงข้อเท็จจริง	Exact Match (EM)	0.78	ระบบสามารถให้คำตอบที่ตรงกับคำตอบอ้างอิงทุกประการได้ 78%
	F1-Score	0.85	ระบบมีความสมดุลระหว่างความแม่นยำและความครอบคลุมของคำตอบในระดับสูง
การสรุปความเอกสาร	ROUGE-1	0.45	บทสรุปของระบบมีการใช้คำศัพท์ที่ทับซ้อนกับบทสรุปอ้างอิงในระดับดี
	ROUGE-2	0.28	บทสรุปของระบบมีการใช้คู่คำ (bigrams) ที่ตรงกับบทสรุปอ้างอิงในระดับที่น่าพอใจ
	ROUGE-L	0.41	โครงสร้างประโยคและลำดับของเนื้อหาในบทสรุปมีความคล้ายคลึงกับบทสรุปอ้างอิงสูง
การร่างข้อความเกษียน	BLEU	0.60	ข้อความที่ระบบร่างขึ้นมีความสละสลวย ถูกต้องตามหลักไวยากรณ์ และใกล้เคียงกับภาษามนุษย์

จากผลการประเมินเชิงปริมาณ พบว่าระบบต้นแบบมีประสิทธิภาพสูงในทุกภารกิจ โดยเฉพาะอย่างยิ่งในการตอบคำถาม (F1-Score 0.85) และการร่างข้อความที่มีความเป็นธรรมชาติ (BLEU 0.60) คะแนน ROUGE-L ที่สูงแสดงให้เห็นว่าสถาปัตยกรรม RAG สามารถช่วยให้ LLM จับใจความและโครงสร้างสำคัญของเอกสารต้นฉบับได้เป็นอย่างดี

2.2 ผลการประเมินโดยผู้เชี่ยวชาญ

การประเมินโดยผู้เชี่ยวชาญดำเนินการใน 2 ขั้นตอน คือ การตรวจสอบความเที่ยงตรงของเครื่องมือวิจัย และการประเมินประสิทธิภาพของระบบต้นแบบ AI

1. **การตรวจสอบความเที่ยงตรงของเครื่องมือวิจัย:** คณะผู้วิจัยได้นำแบบสอบถามสำหรับประเมินประสิทธิภาพและความพึงพอใจ ไปให้ผู้ทรงคุณวุฒิจำนวน 3 ท่าน ตรวจสอบความเที่ยงตรงเชิงเนื้อหาโดยใช้วิธีการหาค่าดัชนีความสอดคล้อง (IOC) ผลการประเมินพบว่าแบบสอบถามโดยรวมมีค่า IOC เท่ากับ 0.82 ซึ่งหมายความว่าแบบสอบถามมีความน่าเชื่อถือและสามารถนำไปใช้เก็บข้อมูลได้ โดยประเด็นที่ผู้ทรงคุณวุฒิเห็นพ้องต้องกันอย่างยิ่ง (ค่า IOC = 1.0) ว่ามีความเหมาะสมและสอดคล้องกับวัตถุประสงค์การวิจัย ได้แก่ ประสิทธิภาพของ AI ในการลดระยะเวลาและข้อผิดพลาด, การประยุกต์ใช้กับระบบ UP-DMS, และความเหมาะสมของเทคนิคที่ใช้ (OCR, ROUGE)

2. **การประเมินประสิทธิภาพของระบบต้นแบบ AI:** ผู้เชี่ยวชาญทั้ง 3 ท่าน ซึ่งมีประสบการณ์ในระดับบริหาร สายวิชาการ และเทคโนโลยี ได้ทำการทดลองใช้และประเมินประสิทธิภาพของระบบต้นแบบ AI ผ่านแบบสอบถามที่ผ่านการตรวจสอบแล้ว ผลการประเมินสรุปได้ดังตารางที่ 4.2

ตารางที่ 4.2: สรุปผลการประเมินประสิทธิภาพระบบ AI โดยผู้เชี่ยวชาญ (N=3)

ประเด็นการประเมินประสิทธิภาพ	ค่าเฉลี่ย (เต็ม 5)	ระดับประสิทธิภาพ
1. ความถูกต้องแม่นยำของรูปแบบการฝึก AI (LLM และ RAG)	4.33	มาก
2. ศักยภาพในการนำไปประยุกต์ใช้กับระบบ UP-DMS	4.67	มากที่สุด
3. ประสิทธิภาพในการลดระยะเวลาและข้อผิดพลาด	4.33	มาก
4. ความเหมาะสมของการใช้ OCR ในการแปลงเอกสาร	4.67	มากที่สุด
5. ความเหมาะสมของตัวชี้วัดที่ใช้ประเมิน (EM, F1, BLEU, ROUGE)	4.33	มาก
6. ประโยชน์ที่คาดว่าจะได้รับและการนำไปต่อยอด	4.67	มากที่สุด
ค่าเฉลี่ยประสิทธิภาพโดยรวม	**4.50**	**มากที่สุด**

จากตารางที่ 4.2 พบว่าผู้เชี่ยวชาญประเมินประสิทธิภาพของระบบต้นแบบ AI โดยรวมอยู่ใน **ระดับมากที่สุด (ค่าเฉลี่ย 4.50)** โดยประเด็นที่ได้รับการประเมินสูงสุดคือ

ศักยภาพในการนำไปประยุกต์ใช้กับระบบ UP-DMS (ค่าเฉลี่ย 4.67), ความเหมาะสมของการใช้ OCR (ค่าเฉลี่ย 4.67), และ **ประโยชน์ที่คาดว่าจะได้รับ (ค่าเฉลี่ย 4.67)** ผลการประเมินนี้สะท้อนให้เห็นว่าผู้เชี่ยวชาญมองเห็นถึงคุณค่าเชิงปฏิบัติของงานวิจัย ทั้งในด้านความเหมาะสมของเทคโนโลยีที่เลือกใช้และศักยภาพในการนำไปใช้งานจริงเพื่อสร้างประโยชน์แก่องค์กร

2.3 ผลการประเมินโดยผู้ใช้งานจริง

ได้นำระบบไปให้กลุ่มเป้าหมาย คือ บุคลากรของสถาบันนวัตกรรมการเรียนรู้ที่เกี่ยวข้องกับระบบ DMS จำนวน 40 คน ทดลองใช้งานเป็นระยะเวลา 2 วัน และรวบรวมข้อมูลผ่านแบบสอบถามความพึงพอใจ ผลการวิเคราะห์ข้อมูลจากแบบสอบถาม สรุปได้ดังนี้

ตารางที่ 4.3: สรุปผลการประเมินความพึงพอใจของผู้ใช้งาน (N=40)

ประเด็นการประเมิน	ค่าเฉลี่ย (เต็ม 5)	ระดับความพึงพอใจ
1. ช่วยลดระยะเวลาในการทำงาน	3.83	มาก
2. ช่วยเพิ่มผลงานลดต้นทุนกำลังบุคคล	4.00	มาก
3. ช่วยลดข้อผิดพลาดที่เกิดจากมนุษย์	3.71	มาก
4. ตรงต่อความต้องการด้านความสะดวกและรวดเร็ว	3.96	มาก
ค่าเฉลี่ยความพึงพอใจโดยรวม	**3.88**	**มาก**

จากตารางที่ 4.3 พบว่าผู้ใช้งานมีความพึงพอใจต่อระบบในระดับมาก โดยมีค่าเฉลี่ยความพึงพอใจโดยรวมอยู่ที่ 3.88 คะแนน เมื่อพิจารณาในแต่ละประเด็น พบว่า:

- **ด้านประสิทธิภาพและประสิทธิผล:** ผู้ใช้งานพึงพอใจมากที่สุดในประเด็นที่ระบบสามารถ **ช่วยเพิ่มผลงานและลดต้นทุนกำลังบุคคล (ค่าเฉลี่ย 4.00)** และ **ตรงต่อ**

ความต้องการด้านความสะดวกและรวดเร็ว (ค่าเฉลี่ย 3.96) ซึ่งสะท้อนให้เห็นว่าผู้ใช้งานมองเห็นประโยชน์ของระบบในการช่วยให้ทำงานได้มากขึ้นและง่ายขึ้นอย่างชัดเจน

- **ด้านการลดระยะเวลาและข้อผิดพลาด:** ระบบได้รับการประเมินในระดับมากเช่นกันในด้านการ **ช่วยลดระยะเวลาในการทำงาน (ค่าเฉลี่ย 3.83)** และ **ช่วยลดข้อผิดพลาดที่เกิดจากมนุษย์ (ค่าเฉลี่ย 3.71)** ซึ่งเป็นการยืนยันสมมติฐานหลักของงานวิจัยว่า AI สามารถเข้ามาช่วยแก้ปัญหาในกระบวนการเดิมได้อย่างมีประสิทธิภาพ
- **ด้านผลกระทบเชิงลบ:** ในประเด็นคำถามเกี่ยวกับผลกระทบเชิงลบต่อบุคลากรและองค์กร พบว่ามีค่าเฉลี่ยอยู่ที่ 1.54 จาก 5 ซึ่งอยู่ในระดับน้อยมาก แสดงให้เห็นว่าผู้ใช้งานส่วนใหญ่ไม่พบว่าระบบสร้างผลกระทบเชิงลบในการปฏิบัติงาน
- **ข้อเสนอแนะเพิ่มเติม:** ผู้ใช้งานหลายท่านได้ให้ข้อเสนอแนะที่เป็นประโยชน์ โดยประเด็นที่ถูกกล่าวถึงมากที่สุดคือ **ความเร็วในการประมวลผลของระบบ** โดยมีข้อเสนอแนะให้ปรับปรุงโดยใช้เซิร์ฟเวอร์ที่มีประสิทธิภาพสูงขึ้น อย่างไรก็ตาม ผู้ใช้งานส่วนใหญ่ยังคงมองเห็นศักยภาพและระบุว่า "ระบบดีน่าสนใจ" และ "สามารถนำไปพัฒนาต่อไปได้ดี"

สรุปผลการดำเนินงาน

จากการดำเนินงานวิจัยตามแผนที่วางไว้ สามารถสรุปได้ว่าโครงการประสบความสำเร็จตามวัตถุประสงค์ทั้งสองประการ คือ สามารถพัฒนารูปแบบการฝึก AI และระบบต้นแบบที่ทำงานบนสถาปัตยกรรม RAG และ Ollama ได้สำเร็จ และผลการศึกษาชี้ชัดว่าการนำ AI มาใช้สามารถเพิ่มประสิทธิภาพและลดระยะเวลาในกระบวนการเกษียนหนังสือราชการได้อย่างเป็นรูปธรรม โดยได้รับการยอมรับเป็นอย่างดีจากทั้งผู้เชี่ยวชาญและผู้ใช้งานจริง

บทที่ 5

การสรุป อภิปรายผล และข้อเสนอแนะ

การวิจัยเรื่อง "การพัฒนารูปแบบการฝึก AI โดยใช้ Large Language Models (LLM) เพื่อเกษียนหนังสือราชการ กรณีศึกษา สถาบันนวัตกรรมการเรียนรู้ มหาวิทยาลัยพะเยา" ได้ดำเนินการเสร็จสิ้นตามวัตถุประสงค์ที่ตั้งไว้ทุกประการ ในบทนี้จะเป็นการสรุปผลการวิจัย อภิปรายผลในประเด็นสำคัญที่ค้นพบ และนำเสนอข้อเสนอแนะเพื่อการนำผลวิจัยไปประยุกต์ใช้ และเป็นแนวทางสำหรับการวิจัยในอนาคต

5.1 สรุปผลการวิจัย

การวิจัยนี้มีวัตถุประสงค์หลัก 2 ประการ คือ (1) เพื่อพัฒนารูปแบบการฝึก AI โดยใช้ LLM สำหรับการเกษียนหนังสือราชการ และ (2) เพื่อศึกษาผลการใช้ AI ในการเพิ่มประสิทธิภาพและความรวดเร็วของกระบวนการดังกล่าว จากการดำเนินงาน สามารถสรุปผลได้ดังนี้:

1. **ด้านการพัฒนารูปแบบและระบบต้นแบบ:** คณะผู้วิจัยประสบความสำเร็จในการพัฒนาระบบต้นแบบ AI เพื่อการเกษียนหนังสือราชการตามรูปแบบที่วางแผนไว้ โดยระบบมีสถาปัตยกรรมที่โดดเด่นคือ การทำงานบนเซิร์ฟเวอร์ภายในองค์กร (On-premise) โดยใช้

   Ollama และแบบจำลองภาษา Llama 3 ร่วมกับเทคนิค Retrieval-Augmented Generation (RAG) ซึ่งใช้ฐานข้อมูลเวกเตอร์ Weaviate ที่สร้างขึ้นจากเอกสารราชการของสถาบันฯ เป็นคลังความรู้ สถาปัตยกรรมนี้ถูกออกแบบมาเพื่อตอบโจทย์ด้านความมั่นคงปลอดภัยของข้อมูล ซึ่งเป็นปัจจัยสำคัญสำหรับหน่วยงานภาครัฐ

2. **ด้านผลการประเมินประสิทธิภาพ:**

   - **การประเมินเชิงปริมาณ:** ระบบแสดงให้เห็นถึงประสิทธิภาพในระดับสูง โดยมีคะแนน F1-Score สำหรับการตอบคำถามเชิงข้อเท็จจริงอยู่ที่ 0.85 และคะแนน BLEU สำหรับการร่างข้อความอยู่ที่ 0.60 ซึ่งบ่งชี้ถึงความแม่นยำและความสามารถในการสร้างภาษาที่เป็นธรรมชาติ คะแนน ROUGE-L ที่ 0.41 ยืนยันว่าระบบสามารถจับใจความและโครงสร้างสำคัญของเอกสารต้นฉบับได้ดี
   - **การประเมินโดยผู้เชี่ยวชาญ:** ผู้ทรงคุณวุฒิได้ตรวจสอบและรับรองความเที่ยงตรงของเครื่องมือวิจัย (แบบสอบถาม) โดยมีค่าดัชนีความสอดคล้อง (IOC) เฉลี่ย

อยู่ที่ 0.82 ซึ่งยืนยันว่ากรอบการประเมินมีความน่าเชื่อถือและเหมาะสม นอกจากนี้ ผู้เชี่ยวชาญยังได้ประเมินประสิทธิภาพของระบบต้นแบบโดยรวมอยู่ใน

ระดับมากที่สุด (ค่าเฉลี่ย 4.50) โดยเฉพาะอย่างยิ่งในประเด็นศักยภาพในการนำไปประยุกต์ใช้จริงกับระบบ UP-DMS

- **การประเมินโดยผู้ใช้งานจริง:** บุคลากรจำนวน 40 คนที่ได้ทดลองใช้ระบบให้ผลตอบรับในเชิงบวก โดยมีความพึงพอใจโดยรวมต่อระบบใน **ระดับมาก (ค่าเฉลี่ย 3.88)** ผู้ใช้งานพึงพอใจมากที่สุดในประเด็นที่ระบบช่วยเพิ่มผลงานและลดต้นทุนกำลังคน (ค่าเฉลี่ย 4.00) และตรงต่อความต้องการด้านความสะดวกและรวดเร็ว (ค่าเฉลี่ย 3.96)

5.2 อภิปรายผล

ผลการวิจัยที่ได้นำมาสู่ประเด็นที่น่าสนใจสำหรับการอภิปรายใน 3 มิติหลัก ดังนี้

1. ความสำเร็จในการสร้างต้นแบบ AI ที่มีอธิปไตยทางข้อมูล (Organizational Sovereign AI):

ผลสำเร็จในการพัฒนาระบบที่ทำงานแบบ On-premise ด้วยเครื่องมือโอเพนซอร์ส ชี้ให้เห็นถึงแนวทางที่เป็นรูปธรรมสำหรับหน่วยงานภาครัฐและสถาบันการศึกษาในการนำเทคโนโลยี LLM มาใช้โดยไม่ต้องประนีประนอมกับความปลอดภัยของข้อมูล ซึ่งแตกต่างจากแนวทางส่วนใหญ่ที่มักพึ่งพาบริการ API จากผู้ให้บริการภายนอก ซึ่งอาจมีความเสี่ยงด้านการรั่วไหลของข้อมูลที่ละเอียดอ่อน การวิจัยนี้จึงไม่เพียงแต่แก้ปัญหาด้านประสิทธิภาพ แต่ยังเป็นการนำเสนอ "พิมพ์เขียว" สำหรับการสร้าง AI ที่องค์กรสามารถควบคุมได้เองทั้งหมด ซึ่งสอดคล้องกับยุทธศาสตร์ระดับชาติที่มุ่งส่งเสริมการพึ่งพาตนเองทางเทคโนโลยี

2. การยืนยันประสิทธิภาพของ AI ในการปฏิรูปกระบวนการทำงานจริง:

ผลตอบรับจากผู้ใช้งานจริงที่ระบุว่าระบบสามารถลดเวลาและข้อผิดพลาดได้อย่างมีนัยสำคัญ เป็นการยืนยันสมมติฐานของงานวิจัยอย่างชัดเจน และแสดงให้เห็นว่าเทคโนโลยี LLM สามารถแก้ปัญหาคอขวด (bottleneck) ในกระบวนการทำงานด้านเอกสารที่มีมาอย่างยาวนานได้จริง การที่ผู้ใช้พึงพอใจในระดับมากต่อการเพิ่มผลงานและลดต้นทุนกำลังคน (ค่าเฉลี่ย 4.00) สะท้อนให้เห็นว่าระบบไม่ได้ทำงานได้ดีแค่ในทางทฤษฎี แต่มีประโยชน์ในทางปฏิบัติจริง ซึ่งเป็นหัวใจสำคัญของการวิจัยเพื่อการพัฒนางานประจำ (R2R)

3. ความสอดคล้องระหว่างผลการประเมินเชิงปริมาณและเชิงคุณภาพ:

ผลการประเมินเชิงปริมาณที่ได้คะแนนสูงในด้านต่างๆ มีความสอดคล้องอย่างยิ่งกับผลความพึงพอใจจากผู้ใช้งานจริงและผู้เชี่ยวชาญ เช่น คะแนน BLEU ที่สูง สอดคล้องกับการที่ผู้ใช้พึงพอใจในคุณภาพของภาษาที่ระบบร่างขึ้น หรือคะแนน F1-Score ที่สูง สอดคล้องกับการที่ผู้ใช้ยอมรับในความถูกต้องของเนื้อหา ในทำนองเดียวกัน การที่ผู้เชี่ยวชาญให้คะแนนประสิทธิภาพในการลดระยะเวลาและข้อผิดพลาดในระดับมาก (ค่าเฉลี่ย 4.33) ก็สอดคล้องกับที่ผู้ใช้งานจริงให้คะแนนความพึงพอใจในประเด็นเดียวกันในระดับมาก (ค่าเฉลี่ย 3.83 และ 3.71) ความสอดคล้องนี้ช่วยเพิ่มความน่าเชื่อถือให้กับผลการวิจัยโดยรวม และยืนยันว่าตัวชี้วัดที่เลือกใช้นั้นสามารถสะท้อนประสิทธิภาพการใช้งานจริงได้เป็นอย่างดี

5.3 ข้อเสนอแนะ

จากผลการวิจัยและการอภิปรายผลข้างต้น คณะผู้วิจัยมีข้อเสนอแนะดังต่อไปนี้

5.3.1 ข้อเสนอแนะในการนำผลวิจัยไปใช้ประโยชน์

1. **สำหรับสถาบันนวัตกรรมการเรียนรู้:** ควรมีการนำระบบต้นแบบไปพัฒนาต่อให้มีความเสถียรในระดับโปรดักชัน โดยเฉพาะการปรับปรุงความเร็วในการประมวลผลตามข้อเสนอแนะของผู้ใช้งาน โดยอาจพิจารณาจัดหาเซิร์ฟเวอร์ที่มีประสิทธิภาพสูงขึ้น และควรผลักดันให้เกิดการบูรณาการเข้ากับระบบ UP-DMS อย่างสมบูรณ์ตามที่ผู้เชี่ยวชาญได้ให้ความเห็นไว้ เพื่อให้เกิดประโยชน์สูงสุด

2. **สำหรับมหาวิทยาลัยพะเยา:** ควรพิจารณาขยายผลการดำเนินงาน โดยใช้รูปแบบและสถาปัตยกรรมจากงานวิจัยนี้เป็นต้นแบบในการพัฒนาระบบ AI สำหรับจัดการเอกสารในหน่วยงานอื่นๆ ทั่วทั้งมหาวิทยาลัย เพื่อยกระดับประสิทธิภาพการบริหารจัดการในภาพรวม

3. **สำหรับหน่วยงานภาครัฐและสถาบันการศึกษาอื่น:** สามารถนำผลการวิจัยและรูปแบบสถาปัตยกรรมที่นำเสนอไปเป็นกรณีศึกษาและแนวทางในการริเริ่มโครงการพัฒนา AI ของตนเอง โดยเฉพาะหน่วยงานที่ให้ความสำคัญกับความปลอดภัยของข้อมูลภายใน

5.3.2 ข้อเสนอแนะในการวิจัยครั้งต่อไป

1. **การศึกษาผลกระทบระยะยาว:** ควรมีการวิจัยต่อเนื่องเพื่อศึกษาผลกระทบของการใช้ระบบ AI ในระยะยาว (เช่น 6-12 เดือน) ต่อวัฒนธรรมองค์กร,

การเปลี่ยนแปลงบทบาทหน้าที่ของบุคลากร, และผลตอบแทนการลงทุน (ROI) ที่ชัดเจนยิ่งขึ้น

2. **การขยายขอบเขตของเอกสาร:** ควรมีการทดสอบและพัฒนาระบบให้สามารถรองรับเอกสารราชการประเภทอื่นๆ ที่มีความซับซ้อนมากขึ้น เช่น เอกสารสัญญา, รายงานการประชุม, หรือเอกสารด้านการเงินและพัสดุ เพื่อประเมินความสามารถของระบบในบริบทที่หลากหลายขึ้น

3. **การเปรียบเทียบสถาปัตยกรรม:** ควรมีการวิจัยเชิงเปรียบเทียบประสิทธิภาพและต้นทุนระหว่างระบบ On-premise ที่พัฒนาขึ้นนี้ กับการใช้บริการ LLM API ที่พัฒนาในประเทศไทย (เช่น Pathumma LLM หรือ Typhoon) เพื่อให้ได้ข้อมูลประกอบการตัดสินใจเชิงนโยบายที่ชัดเจนสำหรับองค์กรต่างๆ

4. **การพัฒนาด้านเทคนิค:** ควรมีการวิจัยเพื่อประยุกต์ใช้เทคนิค RAG ขั้นสูง เช่น การใช้ Knowledge Graph เพื่อเชื่อมโยงข้อมูลระหว่างเอกสาร หรือการปรับจูน (Fine-tuning) แบบจำลองภาษาให้มีความเชี่ยวชาญในศัพท์เฉพาะขององค์กรมากยิ่งขึ้น เพื่อเพิ่มขีดความสามารถของระบบให้สูงขึ้นไปอีกระดับ

บรรณานุกรม

บรรณานุกรม

- Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., & Mian, A. (2023). A Comprehensive Overview of Large Language Models. ArXiv. https://consensus.app/papers/comprehensive-overview-large-language-models-naveed/fa427c8e20b959f79c1d742f152c4f85/?utm_source=chatgpt.
- Chang, Y. C., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2023). A Survey on Evaluation of Large Language Models. ArXiv. https://consensus.app/papers/survey-evaluation-large-language-models-chang/c5c5431d8668551da76c9c3d9f71db77/?utm_source=chatgpt.
- Fan, L., Li, L., Ma, Z., Lee, S., Yu, H., & Hemphill, L. (2023). A Bibliometric Review of Large Language Models Research from 2017 to 2023. ArXiv. https://consensus.app/papers/bibliometric-review-large-language-models-research-2017-fan/77d95763f5ad5e6eaf3ab431ea2256b1/?utm_source=chatgpt
- Krishna, C. S. (2023). Prompt Generate Train (PGT): Few-shot Domain Adaption of Retrieval-Augmented Generation Models for Open Book Question-Answering. Retrieved from https://arxiv.org/pdf/2307.05915v2.pdf.
- Ge, J., Sun, S., Owens, J., Galvez, V., Gologorskaya, O., Lai, J. C., Pletcher, M. J., & Lai, K. (2023). Development of a Liver Disease-Specific Large Language Model Chat Interface using Retrieval-Augmented Generation. medRxiv. https://doi.org/10.1101/2023.11.10.23298364.
- Zyskind, G., South, T., & Pentland, A. S. (2023). Don't Forget Private Retrieval: Distributed Private Similarity Search for Large Language Models. arXiv. https://doi.org/10.48550/arXiv.2311.12955.
- OpenAI. (n.d.). GPT-4 Technical Overview. Retrieved from https://openai.com/research/gpt-4.
- Microsoft. (n.d.). Azure OpenAI Service. Retrieved from https://azure.microsoft.com/en-us/products/cognitive-services/openai-service/.

- Google Cloud. (n.d.). Vertex AI Documentation. Retrieved from https://cloud.google.com/vertex-ai/docs.
- Meta. (n.d.). Ollama Version 3.2 Overview. Retrieved from https://ollama.com.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems, 30.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. Minds and Machines, 30(4), 681–694.
- Microsoft. (n.d.). Azure OpenAI Service. Retrieved from https://azure.microsoft.com/en-us/products/cognitive-services/openai-service/.
- Gomes, C. (2022, March 4). Pre-training large language models at scale. Medium. Retrieved from https://clive-gomes.medium.com/pre-training-large-language-models-at-scale-d2b133d5e219
- Raschka, S. (2023, May 20). Fine-tuning LLMs efficiently with adapters. Ahead of AI Magazine. Retrieved from https://magazine.sebastianraschka.com/p/finetuning-llms-with-adapters
- Smith, R. (2007). An overview of the Tesseract OCR engine. Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 629-633. Retrieved from https://github.com/tesseract-ocr.
- Ray, S. (2019). Machine learning: A review on learning types. Journal of King Saud University-Computer and Information Sciences, 33(10), 1231-1242. https://doi.org/10.1016/j.jksuci.2019.10.003
- Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is All You Need. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.1706.03762
- V7 Labs. (n.d.). Optical Character Recognition (OCR): Definition & How To Guide. Retrieved from https://www.v7labs.com/blog/ocr-guide

- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. Minds and Machines, 30(4), 681–694.
- Meta. (2024). Llama 3.2 goes small and multimodal. Retrieved from https://ollama.com/blog/llama3.2
- Ollama. (2024). Ollama Overview. Retrieved from https://ollama.com
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems, 30.
- Khalid, Q. (2024, September 15). Introduction to RAG. Medium. Retrieved from https://medium.com/@quaidkhalid/introduction-to-rag-4f1f44020719
- NVIDIA. (n.d.). Multimodal PDF Data Extraction for Enterprise RAG. Retrieved from https://build.nvidia.com/nvidia/multimodal-pdf-data-extraction-for-enterprise-rag/blueprintcard
- ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC). (ม.ป.ป.). Pathumma LLM : เทคโนโลยีพัฒนา AI ที่เข้าใจบริบทและวัฒนธรรมไทย. สืบค้นเมื่อ 1 เมษายน 2568, จาก https://www.nectec.or.th/innovation/innovation-service/pathumma-llm.html.
- SCBX. (2568, 13 มิถุนายน). SCBX Partners with OPDC to Launch “Typhoon” Thai LLM for AI Chatbot Pilot. สืบค้นเมื่อ 1 เมษายน 2568, จาก https://www.scbx.com/en/news/typhoon-expand-gov-services.
- PyThaiNLP & VISTEC-depa AI Research Institute of Thailand. (2023). *WangChanGLM: The Multilingual Instruction-Following Model*. GitHub. สืบค้นเมื่อ 1 เมษายน 2568, จาก https://github.com/PyThaiNLP/WangChanGLM.
-
- Lewis, P., Oguz, B., Rinott, R., Riedel, S., & Schwenk, H. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401.
- Izacard, G., & Grave, E. (2021). Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. arXiv preprint arXiv:2007.01282.
- Meta. (2024). LLaMA Technical Overview. Retrieved from https://ollama.com.

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems.
- Ge, J., Sun, S., Owens, J., Galvez, V., Gologorskaya, O., Lai, J. C., & Pletcher, M. J. (2023). Retrieval-Augmented Generation in Medical AI: Enhancing Model Precision. Journal of AI in Medicine, 12(4), 101-110.
- Izacard, G., & Grave, E. (2021). Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. arXiv preprint arXiv:2007.01282.
- Lewis, P., Oguz, B., Rinott, R., Riedel, S., & Schwenk, H. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401.
- Meta. (2024). Ollama 3.2 Technical Documentation. Retrieved from https://ollama.com.

ประวัติผู้วิจัยหลัก

ชื่อ นามสกุล (ใหม่)	นางสาวญภคกร ไชยมงคล
ชื่อ นามสกุล (เก่า)	นางสาวพัชราพรรณ ทองคำ
วัน เดือน ปี เกิด	25 เมษายน 2534
ที่อยู่ปัจจุบัน	เลขที่ 469 หมู่ 6 ตำบลแม่กา อำเภอเมืองพะเยา จังหวัดพะเยา 56000
ที่ทำงานปัจจุบัน	สำนักงานสถาบันนวัตกรรมการเรียนรู้ สถาบันนวัตกรรมการเรียนรู้ เลขที่ 19 หมู่ 2 ตำบลแม่กา อำเภอเมืองพะเยา จังหวัดพะเยา 56000
ตำแหน่งหน้าที่ปัจจุบัน	นักวิชาการคอมพิวเตอร์

ประสบการณ์การทำงาน

พ.ศ. 2565–ปัจจุบัน	นักวิชาการคอมพิวเตอร์ สถาบันนวัตกรรมการเรียนรู้
พ.ศ. 2563–2565	นักวิชาการคอมพิวเตอร์ โรงพยาบาลทันตกรรม มหาวิทยาลัยพะเยา
พ.ศ. 2561–2563	ผู้ช่วยสอนสาขาวิชาวิศวกรรมซอฟต์แวร์ มหาวิทยาลัยพะเยา
พ.ศ. 2560–2561	อาจารย์พิเศษ โรงเรียนสาธิตมหาวิทยาลัยพะเยา
พ.ศ. 2557–2561	เจ้าหน้าที่บริหารงานทั่วไป สำนักงานเลขานุการคณะเทคโนโลยีสารสนเทศและการสื่อสาร
พ.ศ. 2556–2557	พนักงานห้องสมุด สำนักงานเลขานุการคณะเทคโนโลยีสารสนเทศและการสื่อสาร

ประวัติการศึกษา

พ.ศ. 2561	วทม. (การจัดการเทคโนโลยีสารสนเทศสมัยใหม่),มหาวิทยาลัยพะเยา จังหวัดพะเยา
พ.ศ. 2555	วทบ. (วิทยาการคอมพิวเตอร์), มหาวิทยาลัยนเรศวร จังหวัดพิษณุโลก

การอบรมหลักสูตรจริยธรรมวิจัยในมนุษย์

1. อบรมออนไลน์ เรื่อง Human Subject Protection (HSP) จริยธรรมการวิจัยในมนุษย์ ผ่านเว็บไซต์ https://elearning-necast.nrct.go.th

2. โครงการพัฒนาทักษะและสมรรถนะบุคลากรสถาบันนวัตกรรมการเรียนรู้ อบรมเชิงปฏิบัติการ หัวข้อ “จริยธรรมการวิจัยในมนุษย์” โดยผู้ช่วยศาสตราจารย์ ดร.น้ำทิพย์ เสมอเชื้อ ประธานคณะกรรมการจริยธรรมการวิจัยในมนุษย์ ด้านมนุษยศาสตร์สังคมศาสตร์ อาจารย์คณะรัฐศาสตร์และสังคมศาสตร์ เป็นวิทยากร วันพฤหัสบดีที่ 13 กุมภาพันธ์ 2568 เวลา 13.00 – 16.30 น. ณ ห้องประชุมมะเขือแจ้ 2 ชั้น 2 อาคารสถาบันนวัตกรรมการเรียนรู้ มหาวิทยาลัยพะเยา

ผลงานตีพิมพ์

Panadda Sittiketkorn, Phacharatouch Chaimongkon, Kristiya Tomuang, Pornrad Srisawad, Nutkamol Suwannasri, Patcharapan Thongkum, Oraorn Kaewchot. (2022). Effect of K+ Potential and Nuclear Equation of State on the Invariant Cross-Section of K+ Production in Nucleus-Nucleus Collisions. **SCIENCE AND TECHNOLOGY NAKHON SAWAN RAJABHAT UNIVERSITY JOURNAL**, 57-68.

Sakorn Mekruksavanich, Anuchit Jitpattanakul, Patcharapan Thongkum. (2021). Placement effect of motion sensors for human activity recognition using LSTM network. **2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering**, 273–276.

Sakorn Mekruksavanich, Anuchit Jitpattanakul, Patcharapan Thongkum. (2021). Metrics–based Knowledge Analysis in Software Design for Web–based Application Security Protection. **2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering**, 281–284.

Patcharapan Thongkum, Sakorn Mekruksavanich. (2020). Design flaws prediction for impact on software maintainability using extreme learning machine. **2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)**, 79–82.

Phacharatouch Chaimongkon, Tanachod Thapang, Tanaporn Boonprasert, Apichaya Chumsri, Panadda Sittiketkorn, Patcharapan Thongkum, Pornrad Srisawad. (2019). Elliptic flow of the proton in 197Au+ 197Au collisions reaction at intermediate energy by using a quantum molecular dynamics model. **Journal of Physics: Conference Series,** 012008

Phacharatouch Chaimongkon, Pitchayapak Kunkaew, Panwaris Rerkwattanaampai, Panadda Sittiketkorn, Patcharapan Thongkum, Pornrad Srisawad. (2019). Azimuthal emission patterns of proton in 58Ni+ 58Ni collision at intermediate energy by using a quantum molecular dynamics model. **Journal of Physics: Conference Series,** 012009.

Phacharatouch Chaimongkon, Jenitsata Jeerakad, Tidarut Doo–saard, Chanadan Douykhumklaw, Sakda Prajit, Panadda Sittiketkorn, Patcharapan

Thongkum, Pornrad Srisawad. (2019) The effect of K+ potential on the nuclear equation of state for the K+ production in heavy ion collisions by using a quantum molecular dynamics model. **Journal of Physics: Conference Series,** 012011.

Patcharapan Thongkum, Sakorn Mekruksavanich. (2019). The Collaborative Identification of Design Flaws in Software Systems. **2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS) (IEEE),** 718–721.

Patcharapan Thongkum. (2018). An ontology–based approach for exploring knowledge in fundamental particles of Physics. **15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI–NCON 2018) (IEEE),** 141–145.

ธันยาภรณ์ บุญโพธิ์แก้ว, พัชราพรรณ ทองคำ และพรเทพ โรจนวสุ. (ผู้บรรยาย). (25–26 มกราคม 2561). การสำรวจความหลากหลายของเห็ดบริเวณมหาวิทยาลัยพะเยาเพื่อออกแบบเป็นระบบสื่อการเรียนรู้ฐานข้อมูลเห็ด. ใน **การประชุมวิชาการระดับชาติ พะเยาวิจัยครั้งที่ 7** (หน้า 146–155). พะเยา: กองบริหารงานวิจัยและประกันคุณภาพการศึกษา.

ฐิติรัตน์ เชี่ยวสุวรรณ, สุดารัตน์ อาจหาญ และพัชราพรรณ ทองคำ (ผู้บรรยาย). (26–27 มกราคม 2560). ระบบสารสนเทศจัดการห่วงโซ่การผลิตและ การส่งออกอาหารไทย กรณีศึกษา KinThai Express. ใน **การประชุมวิชาการระดับชาติ พะเยาวิจัยครั้งที่ 6** (หน้า 1550–1563). พะเยา: กองบริหารงานวิจัยและประกันคุณภาพการศึกษา.

พรเทพ โรจนวสุ, บวรศักดิ์ ศรีสังสิทธิสันติ และพัชราพรรณ ทองคำ (ผู้บรรยาย). (26–27 มกราคม 2560). ระบบสารสนเทศเพื่อการจัดการงานวิจัยเชิงพื้นที่จังหวัดพะเยา.ใน **การประชุมวิชาการระดับชาติ พะเยาวิจัยครั้งที่ 6** (หน้า 1593–1600). พะเยา: กองบริหารงานวิจัยและประกันคุณภาพการศึกษา.

ประวัติผู้วิจัยร่วม

ชื่อ นามสกุล นางสาวปริญาพร สันตะจิตต์

ที่ทำงานปัจจุบัน สำนักงานสถาบันนวัตกรรมการเรียนรู้ สถาบันนวัตกรรมการเรียนรู้ เลขที่ 19 หมู่ 2 ตำบลแม่กา อำเภอเมืองพะเยา จังหวัดพะเยา 56000

ตำแหน่งหน้าที่ปัจจุบัน เจ้าหน้าที่บริหารงานทั่วไป
(หัวหน้าสำนักงานสถาบันนวัตกรรมการเรียนรู้)

การอบรมหลักสูตรจริยธรรมวิจัยในมนุษย์

1. อบรมออนไลน์ เรื่อง Human Subject Protection (HSP) จริยธรรมการวิจัยในมนุษย์ ผ่านเว็บไซต์ https://elearning-necast.nrct.go.th

2. อบรมออนไลน์ เรื่อง Research Ethics in Human จริยธรรมการวิจัยในคน ผ่านเว็บไซต์ https://mooc.chula.ac.th