ระบบการเรียนรู้ความต้องการเพื่อสั่งซื้อสินค้าที่มีความเกี่ยวข้องกัน เพื่อรองรับระบบลูกค้าสัมพันธ์

Learning System for Customer Requirements to Order Related Products in Support of Customer Relationship Management

กิตติ์ธัญญา เตชานนท์กุลวัฒน์[1], กิติภูมิ หน่อเนื้อ[2], และ ประพจน์ ศรีนุวัตติวงศ์[3]
Kitthanya Teachanontkullawat[1], Kitipum Nornua[2], and Prapoj Srinuwattiwong[3]

[1, 2]นักศึกษาสาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
[3]อาจารย์สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

Corresponding Author: prapoj.sr@kmitl.ac.th
Received: April 9, 2025. **Revised**: April 25, 2025. **Accepted**: April 27, 2025.

บทคัดย่อ

ธุรกิจกระจก และวัสดุก่อสร้างในปัจจุบันประสบปัญหาสำคัญในการเข้าใจความต้องการของลูกค้า ส่งผลให้การเพิ่มยอดขายขาดประสิทธิภาพ งานวิจัยนี้จึงพัฒนาระบบแชทบอทอัจฉริยะผ่านเว็บแอปพลิเคชันที่เชื่อมต่อกับแพลตฟอร์มไลน์ เพื่อเป็นสื่อกลางในการสื่อสาร บันทึกข้อมูลความต้องการ และแนะนำสินค้าให้ตรงกับความต้องการของลูกค้า โดยกลุ่มเป้าหมายหลักคือลูกค้ากลุ่มช่างที่ต้องการซื้อวัสดุอุปกรณ์ก่อสร้างสำหรับงานเฉพาะทาง ระบบแชทบอทนี้ใช้เทคโนโลยีแบบจำลองภาษาขนาดใหญ่ แอลแอลเอ็ม (Large Language Model) ที่ผ่านการปรับแต่ง (Fine-Tuning) ร่วมกับเทคนิคการเพิ่มประสิทธิภาพด้วยการเรียกข้อมูล (Retrieval Augmented Generation: RAG) เพื่อให้เหมาะสมกับข้อมูลความรู้และความสัมพันธ์ระหว่างสินค้า การวิจัยใช้ชุดข้อมูลทดสอบจำนวน 2,732 ชุด และประเมินประสิทธิภาพของแบบจำลองด้วยการเปรียบเทียบความแม่นยำระหว่าง Llama-3.1-8B ที่ผ่านการปรับแต่ง กับแบบจำลองแบบปิดอย่าง GPT-4o และ Claude 3.5 Sonnet

ผลการศึกษาพบว่า Llama-3.1-8B ที่ปรับแต่งแล้วให้ความแม่นยำสูงในการตรวจสอบสินค้า (79%) และการเชื่อมโยงสินค้ากับการใช้งาน (87%) ในขณะที่ GPT-4o มีประสิทธิภาพดีกว่าในด้านการแนะนำสินค้า (86%) นอกจากนี้ ผลการประเมินความพึงพอใจจากผู้ใช้งานยังพบว่า Llama-3.1-8B มีคะแนนความพึงพอใจโดยรวมสูงสุด (4.6 จาก 5.0) จึงได้รับการคัดเลือกเป็นแบบจำลองหลักในการพัฒนาระบบแชทบอทสำหรับการใช้งานจริง

คำสำคัญ: แชทบอท, แบบจำลองภาษาขนาดใหญ่, การปรับแต่งแบบจำลอง, ระบบลูกค้าสัมพันธ์, การเพิ่มประสิทธิภาพด้วยการเรียกข้อมูล

Abstract

Glass and construction material businesses currently face significant challenges in understanding customer needs, resulting in inefficient sales growth. This research developed an intelligent chatbot system through a web application connected to the LINE platform to serve as a medium for communication, recording customer requirements, and recommending products that match customer needs. The primary target group is contractor customers who need to purchase construction materials for specialized work. This chatbot system utilizes Large Language Model (LLM) technology that has been fine-tuned, combined with Retrieval Augmented Generation (RAG) techniques to optimize for product knowledge and inter-product relationships. The research used 2,732 test datasets and evaluated the model performance by comparing the accuracy between fine-tuned Llama-3.1-8B with closed models like GPT-4o and Claude 3.5 Sonnet.

The results showed that fine-tuned Llama-3.1-8B provided high accuracy in product verification (79%) and connecting products with their applications (87%), while GPT-4o performed better in product recommendations (86%). Additionally, user satisfaction evaluations found that Llama-3.1-8B received the highest overall satisfaction score (4.6 out of 5.0), thus being selected as the primary model for developing the chatbot system for real-world implementation.

Keywords: Chatbot, Large Language Model, Fine-Tuning, Customer Relationship, Retrieval Augmented Generation

บทนำ

ในปัจจุบัน ธุรกิจจำหน่ายวัสดุก่อสร้างกำลังเผชิญกับความท้าทายสำคัญในการเข้าใจความต้องการของลูกค้ากลุ่มช่าง ซึ่งส่งผลโดยตรงต่อประสิทธิภาพในการเพิ่มยอดขายและการให้บริการ จากการศึกษาพบว่ากระบวนการสั่งซื้อสินค้าผ่านช่องทางการสื่อสารแบบดั้งเดิม เช่น ไลน์ส่วนตัวหรือโทรศัพท์ มีข้อจำกัดสำคัญคือพนักงานขายไม่สามารถให้คำแนะนำเกี่ยวกับสินค้าที่เกี่ยวข้องได้อย่างรวดเร็วและแม่นยำ เนื่องจากข้อจำกัดด้านความรู้เชิงลึกและเวลาในการตอบสนอง โดยเฉพาะเมื่อลูกค้ากลุ่มช่างมักจะสั่งซื้อสินค้าตามรายการที่ต้องการโดยไม่ได้ระบุวัตถุประสงค์การใช้งาน ทำให้พนักงานไม่สามารถแนะนำสินค้าเพิ่มเติมที่จำเป็นต่อการทำงานได้อย่างครบถ้วน

การใช้เทคโนโลยีแบบจำลองภาษาขนาดใหญ่ แอลแอลเอ็ม (Large Language Model) ได้รับความสนใจอย่างมากในการพัฒนาระบบแชทบอทอัจฉริยะที่สามารถเข้าใจความต้องการและให้คำแนะนำที่เหมาะสมแก่ผู้ใช้ อย่างไรก็ตามการนำแอลแอลเอ็มมาประยุกต์ใช้ในบริบทของธุรกิจจำเพาะยังคงเป็นความท้าทาย เนื่องจากข้อจำกัดในการเข้าถึงข้อมูลเฉพาะทางและการตอบสนองต่อคำถามที่ต้องการความรู้เชิงลึก งานวิจัยของ Neupane et al., (2024) แสดงให้เห็นว่าการใช้เทคนิคการเรียกข้อมูล (Retrieval Augmented Generation: RAG) ร่วมกับการปรับแต่งแบบจำลอง (Fine-Tuning) สามารถเพิ่มประสิทธิภาพของแชทบอทในการให้คำตอบที่แม่นยำในบริบทเฉพาะได้

งานวิจัยนี้จึงพัฒนาระบบแชทบอทอัจฉริยะสำหรับธุรกิจวัสดุก่อสร้าง โดยใช้แบบจำลองแอลแอลเอ็มร่วมกับเทคนิคการปรับแต่งโมเดล และการเรียกข้อมูล เพื่อแก้ไขปัญหาดังกล่าว ระบบที่พัฒนาขึ้นจะช่วยให้การสื่อสารระหว่างลูกค้ากลุ่มช่างและธุรกิจมีประสิทธิภาพมากขึ้น สามารถเข้าใจความต้องการของลูกค้าได้อย่างลึกซึ้ง และแนะนำสินค้าที่เกี่ยวข้องได้อย่างตรงจุด ซึ่งไม่เพียงแต่จะช่วยอำนวยความสะดวกให้กับลูกค้าเท่านั้น แต่ยังช่วยเพิ่มโอกาสในการขายและยกระดับประสิทธิภาพของกระบวนการขายให้ดียิ่งขึ้น

วัตถุประสงค์

1. เพื่อศึกษาและพัฒนาระบบแชทบอทอัจฉริยะที่สามารถเพิ่มยอดขายด้วยการแนะนำสินค้าที่เหมาะสมให้แก่ลูกค้ากลุ่มช่าง โดยใช้เทคโนโลยีแอลแอลเอ็ม (Large Language Model), การปรับแต่ง (Fine-Tuning) และ การเรียกข้อมูล (Retrieval Augmented Generation: RAG)
2. เพื่อวิเคราะห์และประเมินประสิทธิภาพของแบบจำลองภาษาต่างๆ ในการให้คำแนะนำสินค้าที่เกี่ยวข้องกับความต้องการของลูกค้า
3. เพื่อพัฒนาระบบที่สามารถบูรณาการกับแพลตฟอร์มการสื่อสารที่ลูกค้าใช้งานอยู่ ลดความซับซ้อนและข้อผิดพลาดในกระบวนการสั่งซื้อ

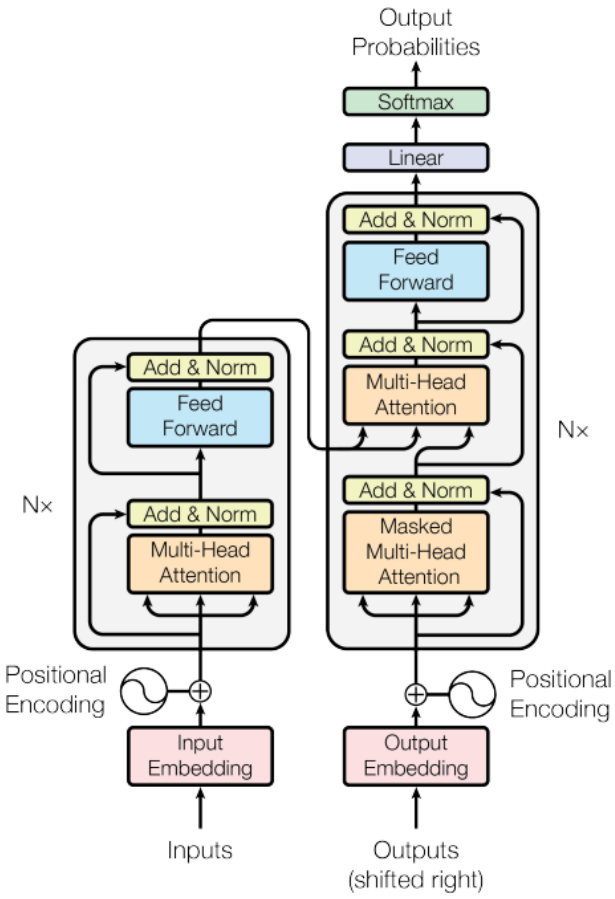
การทบทวนวรรณกรรม

Large Language Model

แอลแอลเอ็ม (Large Language Model) คือ โมเดลประมวลผล พื้นฐานทางภาษาขนาดใหญ่ โดยเป็นเทคโนโลยีหลักสำหรับการทำงานในรูปแบบปัญญาประดิษฐ์ (Generative Artificial Intelligence) ที่มีความสามารถในด้านการทำความเข้าใจ และสื่อสารภาษาได้ใกล้เคียงกับมนุษย์ อีกทั้งยังมีพื้นฐานความรู้ ความ

เข้าใจในความรู้ทั่วไปอย่างกว้างขวางเนื่องจากถูกฝึกฝน ด้วยชุดข้อมูล (Training Dataset) ทางภาษาขนาดใหญ่เป็นจำนวนมาก เช่น หนังสือ เอกสาร และเว็บไซต์ต่างๆ เพื่อให้มีข้อมูลที่ครอบคลุม และหลากหลาย อีกทั้งโครงสร้างของแอลแอลเอ็ม ก็ประกอบไปด้วยพารามิเตอร์ (Parameter) จำนวนมาก ซึ่งเป็นตัวแปรที่ช่วยให้แบบจำลอง (Model) ได้เรียนรู้และปรับตัวในระหว่างการฝึกฝน เพื่อให้สามารถสร้าง ผลลัพธ์ที่แม่นยำและตรงตามบริบท

แอลแอลเอ็มถูกพัฒนาต่อยอดจากสถาปัตยกรรมแบบทรานฟอร์เมอร์ (Transformer) ซึ่งเป็นพื้นฐานสำคัญที่ช่วยให้โมเดลสามารถประมวลผลข้อมูลขนาดใหญ่และซับซ้อนได้อย่างมีประสิทธิภาพ โดยการสร้างข้อความตอบกลับในแอลแอลเอ็มอาศัยการทำนายคำถัดไป (Next Word Prediction) ร่วมกับฐานความรู้ภายในโมเดลเพื่อให้คำตอบที่สร้างขึ้น มีความสอดคล้องกับบริบทของข้อความที่รับเข้ามา โดยสถาปัตยกรรมของทรานฟอร์เมอร์ ประกอบด้วยสองส่วนหลัก ได้แก่ การเข้ารหัส (Encoder) ซึ่งมีหน้าที่ประมวลผลและตีความข้อมูลที่นำเข้า และการถอรหัส (Decoder) ที่รับผิดชอบการสร้างข้อความตอบกลับโดยใช้ข้อมูลที่ได้จากการเข้ารหัส ผสมผสานกับความรู้ที่มีอยู่ในตัวโมเดล กระบวนการนี้ช่วยให้แอลแอลเอ็ม สามารถสร้างคำตอบที่มีความแม่นยำ สอดคล้องกับบริบท และเหมาะสมกับการใช้งานในสถานการณ์ต่างๆ



ภาพที่ 1 สถาปัตยกรรมโมเดลแบบทรานฟอร์เมอร์
ที่มา: Minaee, Mikolov, Nikzad, Chenaghlu, Socher, Amatriain & Gao (2024)

ด้วยศักยภาพในการเลียนแบบการสื่อสารของมนุษย์อย่างเป็นธรรมชาติ ทำให้แอลแอลเอ็มถูกนำมาใช้งานในฐานะแชทบอท (Chatbot) ที่สามารถให้คำแนะนำและตอบคำถามตามข้อมูลที่ผู้ใช้ป้อน (Prompt) อย่างไรก็ตาม แอลแอลเอ็ม มีข้อจำกัดสำคัญในเรื่องการเข้าถึงข้อมูลภายนอก (External Data) หรือข้อมูลเฉพาะในองค์กร ซึ่งส่งผลให้เมื่อเจอคำถามที่ต้องอาศัยข้อมูลดังกล่าว แบบจำลองอาจสร้างคำตอบที่ไม่แม่นยำและผิดเพี้ยนจากความเป็นจริง ซึ่งปัญหานี้เรียกว่า Hallucination ทำให้การใช้งานในบางบริบทอาจขาดความน่าเชื่อถือ และจำเป็นต้องพัฒนาระบบเพิ่มเติมเพื่อแก้ไขข้อจำกัดดังกล่าว

Fine-Tuning

วิธีการปรับแต่ง Fine-Tuning คือการปรับแต่งแอลแอลเอ็มเข้ากับความรู้หรือการทำงานแบบละเอียดและเฉพาะเจาะจง เพื่อให้สามารถกำหนดรูปแบบการตอบกลับเมื่อเผชิญกับข้อมูลที่ผู้ใช้ป้อนที่กำหนด เพื่อให้สามารถรับมือได้อย่างเหมาะสม โดยนำแอลแอลเอ็ม ที่ถูกฝึกฝนมาแล้ว (Pretrained model) มาอัปเดตค่าพารามิเตอร์ให้สอดคล้องกับชุดข้อมูลที่เตรียมไว้ เพื่อให้แบบจำลองนั้นสามารถให้คำตอบ โดยอิงจากชุดข้อมูลดังกล่าวได้ ซึ่งการอัปเดตค่าพารามิเตอร์ ทั้งหมดนั้นใช้เวลาและกำลังในการประมวลผลที่มหาศาล จึงมีการใช้เทคนิค ต่างๆ ดังนี้

เทคนิคการแบ่งนับ (Quantization) คือ การลดข้อมูลตัวเลขทศนิยมของการฝึกฝนแอลแอลเอ็มให้มีความแม่นยำน้อยลงทำให้ขนาดไฟล์ข้อมูล ต่างๆให้เล็กลงได้ จึงสามารถผ่านกระบวนการปรับแต่งแบบจำลองได้รวดเร็ว

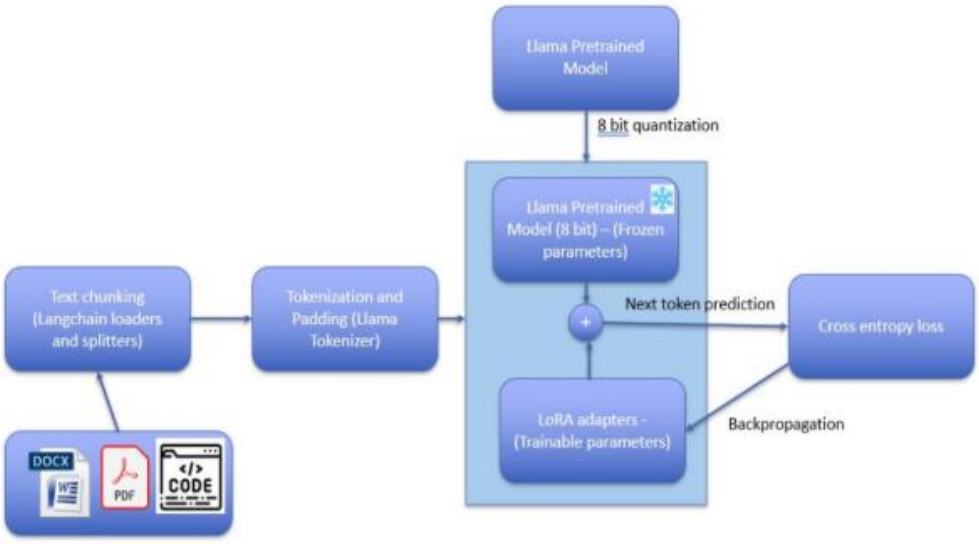
Quantization	GPU memory	What is AION?
Without quantization	28 GB	AION (Artificial Intelligence ON) is a cloud-based platform that enables developers to build, train and deploy machine learning models. It provides an end-to-end solution for data scientists and engineers to create, test, refine, and deploy ML models in production environments.
8 bit quantization	8 GB	AION (Artificial Intelligence ON) is a cloud-based platform that enables developers to build, train and deploy machine learning models. It provides an end-to-end solution for data scientists and engineers to create, test, refine, and deploy predictive modeling solutions in the form of APIs or containerized microservices.

ภาพที่ 2 การแบ่งนับช่วยลดทรัพยากรที่จำเป็น แต่ผลลัพธ์ยังคงมีประสิทธิภาพ
ที่มา: VM, Warrier & Gupta (2024)

จากภาพที่ 2 แสดงให้เห็นว่าการใช้เทคนิคการแบ่งนับ สามารถลดการใช้ทรัพยากรการประมวลผลโดยที่ยังคงคุณภาพของชุดคำตอบไว้ โดยตารางดังกล่าว ได้เทียบคำตอบแบบจำลองจาก คำถาม "What is AION?" จะเห็นได้ว่าแบบจำลองที่ไม่ได้ใช้เทคนิคลดขนาดแบบการแบ่งนับนั้น ใช้หน่วยความจำของการ์ดจอในการประมวลผลถึง 28 GB เพื่อสร้างคำตอบของคำถามดังกล่าว ในขณะที่แบบจำลองที่ผ่านการลดขนาดให้มีรูปแบบข้อมูลในรูปของ INT8 (8 bit integer) จากเทคนิคลดขนาดข้อมูลแบบ 8 bit (8 bit quantization) ดังกล่าวนั้นใช้หน่วยความจำของการ์ดจอในการประมวลผลเพียงแค่ 8 GB และยังคงสามารถสร้างคำตอบที่มี

คุณภาพและใกล้เคียงกับแบบจำลองที่ไม่ได้ใช้เทคนิคดังกล่าวได้ ซึ่งเทคนิคนี้นั้นช่วยให้เครื่องคอมพิวเตอร์ที่มีฮาร์ดแวร์แบบทั่วไป สามารถเข้าถึงและใช้งานเทคโนโลยีแอลแอลเอ็มได้

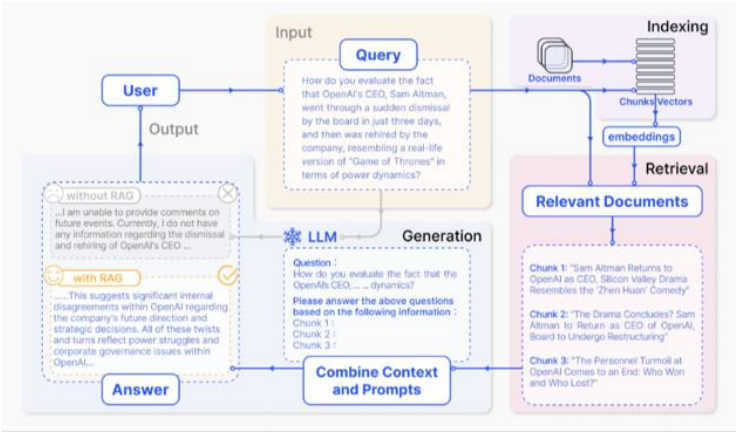
เทคนิค Low Rank Adaptation (LoRA) คือ การอัปเดตเฉพาะพารามิเตอร์ที่มีความจำเป็น เพื่อให้มีการทำงานที่ สอดคล้องกับชุดข้อมูลที่เตรียมไว้ โดยจะเพิกเฉยพารามิเตอร์ (Freeze) ที่ไม่จำเป็น แล้วรวมเข้ากับพารามิเตอร์ที่อัปเดตโดยเทคนิคดังกล่าวนี้ช่วยให้ขนาดของแอลแอลเอ็มที่ผ่านปรับแต่งนั้นมีขนาดเล็กลง



ภาพที่ 3 ตัวอย่างการใช้ Quantization และ LoRA สำหรับการ Fine-Tuning LLM
ที่มา: Gao, Xiong, Gao, Jia, Pan, Bi, & Wang (2023)

Retrieval Augmented Generation

วิธีการเรียกข้อมูล (Retrieval Augmented Generation: RAG) คือ วิธีที่ดึงความรู้ภายนอกมาร่วมประมวลผลกับแอลแอลเอ็ม เพื่อให้สามารถสร้างคำตอบโดยอิงจากข้อมูลดังกล่าวได้ เหมาะกับงานที่อาศัยข้อมูลที่เกิดการเปลี่ยนแปลงได้ตลอดเวลาโดยจะต้องนำข้อมูลดังกล่าวมาแปลงเป็นชุดข้อมูลเวกเตอร์ (Embedding) เก็บไว้ในฐานข้อมูล (Indexing) จากนั้นเมื่อต้องการให้แอลแอลเอ็มตอบคำถามโดยอิงจาก ชุดความรู้นั้นก็สามารถ แปลงข้อมูลที่ผู้ใช้ป้อนให้อยู่ในรูปแบบข้อมูลแบบเวกเตอร์แล้วดึงข้อมูล (Retrieval) ที่มีค่าเวกเตอร์ใกล้เคียงกัน จากฐานข้อมูลมาแล้วส่งให้แอลแอลเอ็มประมวลผลร่วมกัน เพื่อสร้างคำตอบ (Generation) โดยอิงจากชุดข้อมูลที่ได้ดึงมา และมีความเกี่ยวข้องกัน



ภาพที่ 4 ตัวอย่างกระบวนการ RAG
ที่มา: Gupta, Shirgaonkar, Balaguer, Silva, Holstein, Li & Benara (2024)

เริ่มต้นด้วยการปรับแต่งแบบจำลองแอลแอลเอ็มบนชุดข้อมูลที่ต้องการ เพื่อให้โมเดลเข้าใจข้อมูลพื้นฐานและปรับตัวให้เข้ากับบริบทเฉพาะ จากนั้นนำโมเดลที่ผ่านการปรับแต่งเข้าสู่กระบวนการเรียกข้อมูลเพื่อเพิ่มความสามารถในการอ้างอิงข้อมูลล่าสุด ส่งผลให้การตอบมีความแม่นยำและมีประสิทธิภาพมากขึ้น

Model	Fine-tuned	Accuracy	+RAG
Llama-2-chat 13B		76% ±2%	75% ±2%
Vicuna		72% ±2%	79% ±2%
GPT-4		75% ±3%	80% ±4%
Llama2 13B	✓	68% ±3%	77% ±2%
GPT-4	✓	81%±5%	86% ±2%

ภาพที่ 5 เปรียบเทียบประสิทธิภาพความแม่นยำของแอลแอลเอ็มต่างๆ
ที่มา: Gupta, Shirgaonkar, Balaguer, Silva, Holstein, Li & Benara (2024)

วิธีดำเนินการวิจัย

ความสำคัญของการแก้ไขปัญหา (Value Proposition)

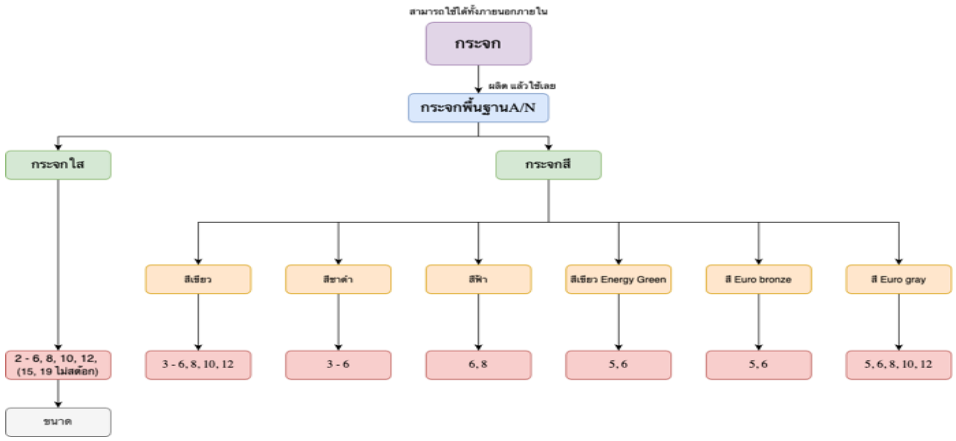
กลุ่มลูกค้าขององค์กรส่วนใหญ่ประกอบด้วยกลุ่มช่างที่ต้องใช้วัสดุก่อสร้างในงานต่างๆ โดยเมื่อมีการสั่งซื้อสินค้า ลูกค้ามักจะสั่งซื้อตามรายการที่ต้องการไม่ได้มีการระบุว่าจะนำสินค้าที่สั่งไปใช้งานในรูปแบบใด ทำให้พนักงานขายจำเป็นต้องมีความรู้เชิงลึก เพื่อสามารถแนะนำสินค้าที่เหมาะสมจากประเภทงานก่อสร้างได้อย่างแม่นยำ

อย่างไรก็ตาม จากการศึกษาพบว่ากระบวนการสั่งซื้อสินค้าผ่านไลน์ มีข้อจำกัดสำคัญในการเพิ่มยอดขายผ่านการแนะนำสินค้าเพิ่มเติม คือ ข้อจำกัดในการสื่อสารระหว่างพนักงานขายและลูกค้า ข้อจำกัด

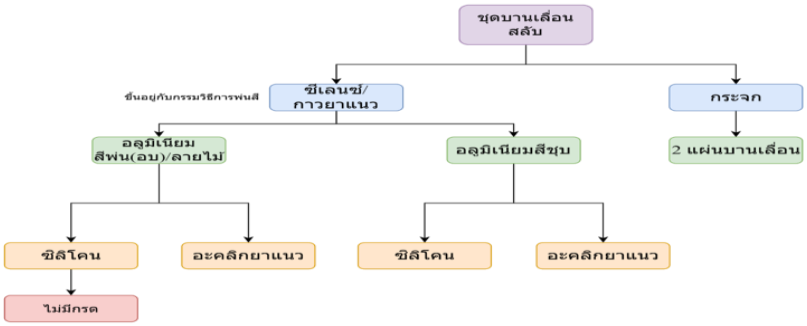
ด้านความรู้เชิงลึกของพนักงานขาย และความเร่งด่วนในการตอบข้อความจากลูกค้า ซึ่งทำให้การสนทนามักมุ่งเน้นไปที่การยืนยันคำสั่งซื้อเพียงอย่างเดียว โดยไม่สามารถแนะนำสินค้าที่เกี่ยวข้องหรือสินค้าอื่นๆ ที่อาจตรงกับความต้องการของลูกค้าได้ ส่งผลให้โอกาสในการเพิ่มยอดขายลดลง

เพื่อแก้ไขปัญหาดังกล่าว จึงได้พัฒนาระบบแชทบอทบนไลน์ออฟฟิศเชียลแอคเคาท์ โดยใช้เทคโนโลยีแอลแอลเอ็มที่สามารถสื่อสารและเข้าใจความต้องการของลูกค้าได้อย่างแม่นยำ โดยการประยุกต์ใช้เทคนิคการปรับแต่งแบบจำลอง และ การเรียกข้อมูล เพื่อเสริมความสามารถในการแนะนำสินค้า โดยการดึงข้อมูลความสัมพันธ์ระหว่างสินค้า

โดยจากการทำงานร่วมกับทีมผู้จัดการผลิตภัณฑ์ (Product Manager) ได้มีการกำหนดแนวทางการเพิ่มยอดขายผ่านการจัดกลุ่มความสัมพันธ์ของสินค้าในรูปแบบ BOM Diagram ซึ่งแบ่ง ความสัมพันธ์ออกเป็น 2 ประเภทหลัก ได้แก่ ความสัมพันธ์ระหว่างสินค้ากับสินค้า และความสัมพันธ์ระหว่างสินค้ากับรูปแบบงานก่อสร้าง ซึ่งจะช่วยให้บอทสามารถแนะนำสินค้าได้อย่างแม่นยำและตรงตามบริบทของลูกค้า



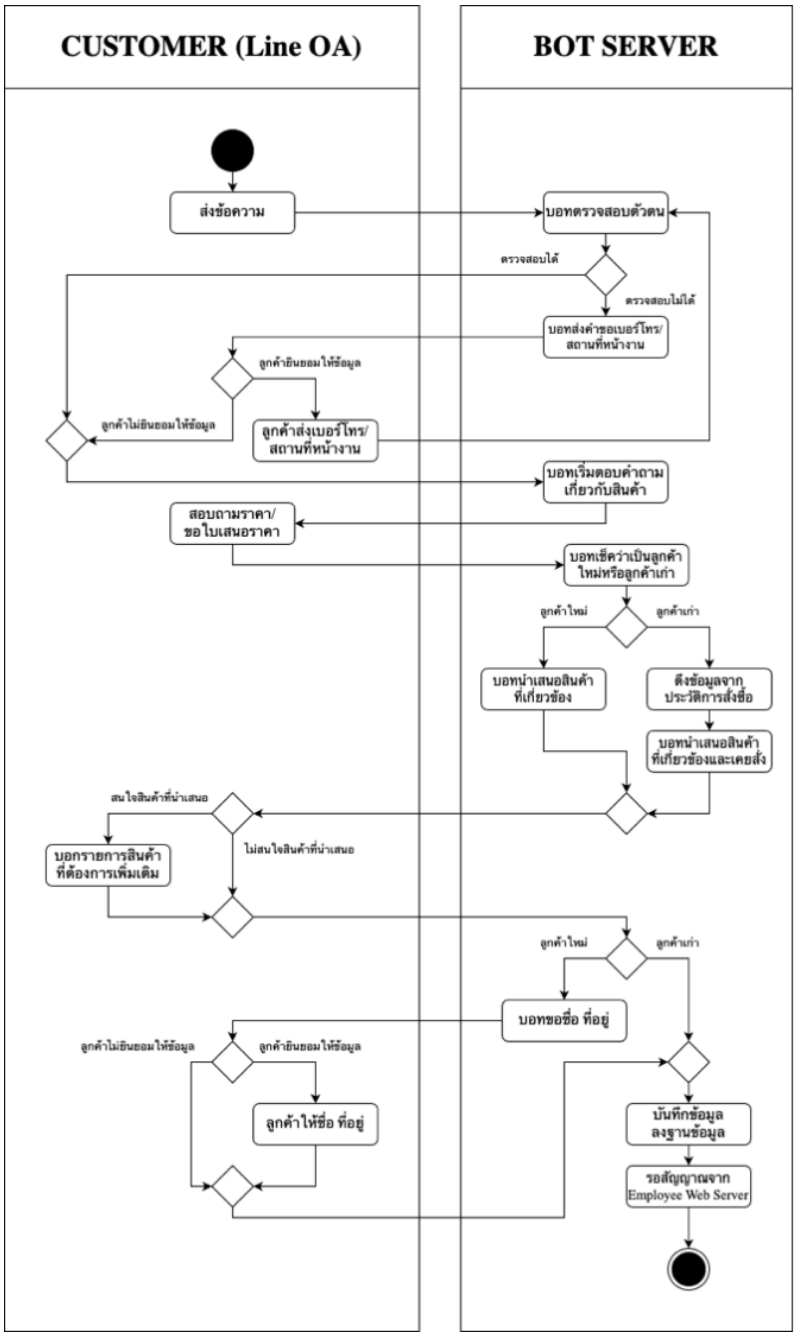
ภาพที่ 6 ตัวอย่างความสัมพันธ์ระหว่างสินค้ากับสินค้าในกลุ่มกระจก
ที่มา: ผู้วิจัย



ภาพที่ 7 ตัวอย่างความสัมพันธ์ระหว่างสินค้ากับรูปแบบงานก่อสร้าง
ที่มา: ผู้วิจัย

การออกแบบขั้นตอนการทำงานของบอท (Activity Diagram)

โครงสร้างลำดับขั้นตอนในการสั่งซื้อสินค้าของลูกค้าและการตอบกลับของบอทที่ส่งผลต่อการประมวลผลคำตอบของแอลแอลเอ็ม ซึ่งสามารถระบุได้ดังนี้



ภาพที่ 8 การออกแบบลำดับขั้นตอนการตอบกลับของแชทบอท
ที่มา: ผู้วิจัย

การออกแบบวิธีการเตรียมข้อมูล (Dataset)

ข้อมูลที่เตรียมจะถอดแบบมาจาก BOM Diagram และจัดเตรียมในรูปแบบชุดข้อมูล Alpaca ที่แต่ละข้อมูลประกอบไปด้วย Instruction, Input โดย 2 สิ่งนี้รวมกันเป็นการป้อนข้อมูลในการส่งให้แอลแอลเอ็มประมวลผล และตอบกลับ

ตาราง 1 ตารางข้อมูลที่จัดเตรียม เพื่อเทรนแอลแอลเอ็ม

Instruction	Input	Output
สินค้าที่มีในระบบ	หมวดหมู่อุปกรณ์	สีทาฝ้า, สกรู, สีย้อมไม้
การใช้งานสินค้า	งานฝ้าเพดาน, ฝ้าเพดาน, อุปกรณ์	สีทาฝ้า, สกรู, ยาแนวอคริลิก
แนะนำสินค้า	สีทาฝ้า, สกรู	สีย้อมไม้, ยาแนวอคริลิค

การออกแบบการประเมินโมเดล

การประเมินประสิทธิภาพของแบบจำลองแอลแอลเอ็ม ใช้การเปรียบเทียบผลลัพธ์กับคำตอบมาตรฐานจากการป้อนข้อมูลที่กำหนด โดยทดสอบกับแบบจำลอง 3 เวอร์ชัน ได้แก่ Llama-3.1-8B ที่ผ่านกระบวนการปรับแต่งแล้ว และ แบบจำลองแบบปิด 2 แบบจำลอง ได้แก่ GPT-4o และ Claude 3.5 Sonnet ผ่านชุดข้อมูล 3 ประเภท ได้แก่ สินค้าที่มีในระบบ ความสัมพันธ์ระหว่างสินค้ากับการใช้งาน และการแนะนำสินค้า ซึ่งแบ่งข้อมูลออกเป็นชุดฝึกอบรม (Training) และชุดทดสอบ (Testing)

การเปรียบเทียบเน้นการวัดค่าความแม่นยำ (Accuracy) ในการนำโมเดลเข้ากระบวนการเรียกข้อมูลเพื่อวิเคราะห์ว่าแบบจำลองใดเหมาะสมที่สุดสำหรับการแนะนำสินค้าและตอบสนองความต้องการลูกค้า ดังแสดงในตารางที่ 2, 3 และ 4

การประยุกต์ใช้แบบจำลอง

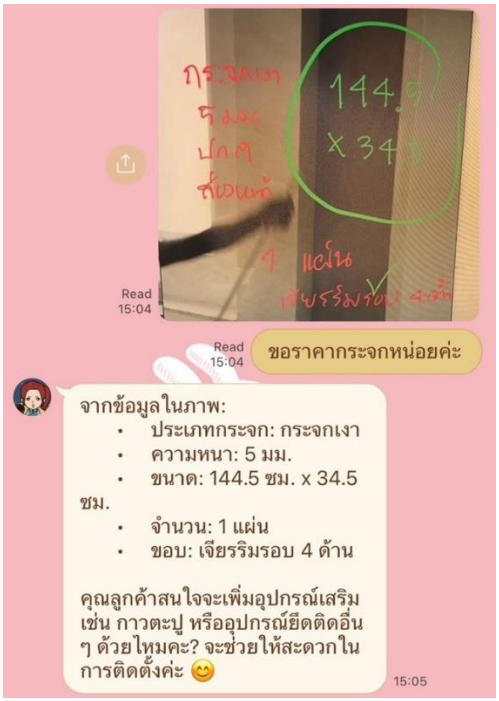
หลังจากการพัฒนาแบบจำลองแอลแอลเอ็ม ด้วยการเทรนจากชุดข้อมูลที่ครอบคลุมสินค้า ความสัมพันธ์ระหว่างสินค้าและการใช้งาน รวมถึงข้อมูลสำหรับการแนะนำสินค้า โมเดลที่ได้จะถูกนำมาใช้งานในระบบแชทบอทที่เชื่อมต่อกับไลน์ออฟฟิศเชียลแอคเคาท์ ด้วย LINE Messaging API ในรูปแบบของ Webhook เพื่อเพิ่มประสิทธิภาพในการให้บริการลูกค้า โดยแบบจำลองจะช่วยตอบคำถามและแนะนำสินค้าที่เหมาะสมสอดคล้องกับความต้องการของลูกค้า

เมื่อแบบจำลองประมวลผลคำถามและวิเคราะห์ความต้องการของลูกค้าได้ หากลูกค้ายืนยันคำสั่งซื้อหรือข้อมูลอื่นที่เกี่ยวข้อง ระบบจะบันทึกคำสั่งเหล่านั้นและส่งต่อไปยังฝ่ายปฏิบัติการ เช่น พนักงานขายหรือฝ่ายการตลาดดิจิทัล เพื่อให้พนักงานสามารถดำเนินการตามขั้นตอนการขาย เช่น การออกใบเสนอราคา การจัดทำเอกสารคำสั่งซื้อ หรือเอกสารอื่นที่เกี่ยวข้อง

นอกจากนี้ ระบบยังจัดเก็บข้อมูลการสนทนาระหว่างลูกค้าและแชทบอท รวมถึงข้อมูลการประสานงานระหว่างลูกค้าและพนักงานที่เกี่ยวข้องผ่านระบบหลังบ้าน พร้อมทั้งติดตามสถานะของคำสั่งซื้อเพื่อวิเคราะห์ผลลัพธ์ โดยข้อมูลจากคำสั่งซื้อที่สำเร็จและคำสั่งซื้อที่ไม่สำเร็จจะถูกนำมาวิเคราะห์เพื่อระบุ

ปัจจัยที่ส่งผลต่อความสำเร็จหรือข้อจำกัดในการทำงาน ซึ่งการวิเคราะห์จะช่วยปรับปรุงกระบวนการแนะนำสินค้า พัฒนากลยุทธ์การขายใหม่ๆ และเพิ่มประสิทธิภาพการให้บริการของแชทบอทเพื่อตอบสนองความต้องการของลูกค้าได้ดียิ่งขึ้น

กระบวนการดังกล่าวทำให้โมเดลที่พัฒนาแล้วสามารถนำมาใช้ได้อย่างครอบคลุมและเป็นระบบ ช่วยเพิ่มความสะดวกและความรวดเร็วในการให้บริการลูกค้า เสริมความคล่องตัวในกระบวนการขาย และสร้างโอกาสในการเพิ่มยอดขายและประสิทธิภาพทางธุรกิจในระยะยาว



ภาพที่ 9 การตอบกลับของแชทบอทที่ใช้แบบจำลองแอลแอลเอ็มบนช่องทางไลน์ออฟฟิศเชียลแอคเคาท์
ที่มา: ผู้วิจัย

การออกแบบการประเมินความพึงพอใจในการใช้งานแชทบอท

เป็นการให้ผู้ใช้ได้ทดลองสนทนากับแชทบอทที่ขับเคลื่อนด้วยแอลแอลเอ็ม และ สำรวจความพึงพอใจในด้านต่างๆ โดยเริ่มจากการประเมินความแม่นยำในการตอบคำถามว่า ผลลัพธ์ที่ได้รับมีความถูกต้องหรือไม่ การประเมินความลื่นไหลของบทสนทนา ซึ่งวัดว่าแอลแอลเอ็ม สามารถตีความคำถามหรือป้อนข้อมูลได้อย่างถูกต้องและสามารถตอบสนองในหัวข้อที่กำลังสนทนาได้อย่างต่อเนื่องและสอดคล้อง

นอกจากนี้ยังมีการประเมินด้านความเร็วในการโต้ตอบ เพื่อวัดว่าระบบสามารถสร้างคำตอบได้รวดเร็วและเหมาะสมกับบริบทที่กำหนดไว้หรือไม่ รวมถึงความเหมาะสมของคำตอบในรูปแบบที่สอดคล้องกับความต้องการของผู้ใช้งาน สุดท้ายคือการประเมินความพึงพอใจโดยรวมเพื่อวิเคราะห์ว่าคำตอบจากระบบสามารถสร้างประสบการณ์ที่น่าพึงพอใจได้หรือไม่ การรวบรวมข้อมูลทั้งหมดนี้ช่วยให้สามารถวิเคราะห์ ตัดสินใจเลือกโมเดลที่เหมาะสมที่สุดและมีประสิทธิภาพสำหรับการใช้งานจริง

ผลการวิจัย

การติดตามสังเกตการณ์และประเมินผลการทำงานมีการแบ่ง การประเมินเป็น 2 รูปแบบดังนี้

การประเมินประสิทธิภาพความแม่นยำของแอลแอลเอ็ม

การประเมินประสิทธิภาพความแม่นยำถูกแบ่งออกตาม แต่ละประเภทหัวข้อของคำถาม หรือการป้อนข้อมูล โดยผลลัพธ์ที่สอดคล้องกับชุดข้อมูลที่เตรียมไว้คือผลลัพธ์ที่ถูกต้อง และผลลัพธ์ที่ไม่สอดคล้อง หรือไม่สามารถตอบได้ตรงประเด็น กับการป้อนข้อมูล ก็จะถูกนับเป็นผลลัพธ์ที่ไม่ถูกต้องหรือ Hallucination ติดตามความพึงพอใจของการนำแอลแอลเอ็มไปใช้งาน

โดยการประเมินผลนี้ได้ใช้แบบจำลอง Llama-3.1-8B ที่ได้ผ่านกระบวนการปรับแต่งแบบจำลองให้เข้ากับชุดข้อมูลความรู้ทั้งหมดที่ได้เตรียมไว้เป็นจำนวน 2,732 ชุดรวมเข้ากับกระบวนการดึงข้อมูลเทียบกับแบบจำลองแบบปิด ทั้ง 2 ที่โด่งดังได้แก่ GPT-4o และ Claude 3.5 Sonnet โดยที่แบบจำลองแบบปิดทั้ง 2 ไม่สามารถปรับแต่ง เพื่อใช้งานแบบส่วนตัวได้ แต่สามารถใช้การดึงข้อมูลร่วมได้

ตาราง 2 ตารางผลการประเมินประสิทธิภาพแบบจำลอง ในการตรวจสอบสินค้าที่มีในระบบ

LLM	# of Embeded Dataset for RAG Pipeline	# of Dataset for Testing	Accuracy (Fine-Tuning + RAG)
Llama-3.1-8B (Fine-Tuned)	12074	100	79%
GPT-4o	12074	100	74%
Claude 3.5 Sonnet	12074	100	69%

ตาราง 3 ตารางผลการประเมินประสิทธิภาพแบบจำลอง ในการเชื่อมโยงสินค้ากับการใช้งาน

LLM	# of Embeded Dataset for RAG Pipeline	# of Dataset for Testing	Accuracy (Fine-Tuning + RAG)
Llama-3.1-8B (Fine-Tuned)	12074	100	87%
GPT-4o	12074	100	82%
Claude 3.5 Sonnet	12074	100	78%

ตาราง 4 ตารางผลการประเมินประสิทธิภาพแบบจำลอง ในด้านการ แนะนำสินค้า

LLM	# of Embeded Dataset for RAG Pipeline	# of Dataset for Testing	Accuracy (Fine-Tuning + RAG)
Llama-3.1-8B (Fine-Tuned)	12074	100	82%
GPT-4o	12074	100	86%
Claude 3.5 Sonnet	12074	100	80%

การประเมินความพึงพอใจในการใช้งานแชทบอท

จากผลสำรวจกลุ่มตัวอย่างผู้ใช้งานจำนวน 112 คน พบว่าการประเมินความพึงพอใจของผู้ใช้งาน (Likert Scale) ในการนำแอลแอลเอ็มที่ผ่านกระบวนการปรับแต่ง และ การดึงข้อมูล แล้วนำไปใช้ในการตอบคำถามตามแต่ละบริบทของการใช้งาน โดยจะวัดค่าเฉลี่ยจาก ระดับความพึงพอใจ 1-5 ในแต่ละหัวข้อประเมิน โดยที่ระดับ 1 คือ พอใจน้อยที่สุด และระดับ 5 คือ พอใจมาก

ตาราง 5 ตารางผลการประเมินค่าเฉลี่ยความพึงพอใจการใช้งาน

หัวข้อประเมิน	ความแม่นยำในการตอบคำถาม	ความลื่นไหลของบทสนทนา	ความเร็วในการตอบสนอง	ความเหมาะสมในบริบท	ความพึงพอใจต่อผลลัพธ์โดยรวม
Llama-3.1-8B (Fine-Tuned)	4.8	4.2	3.9	4.9	4.6
GPT-4o	4.4	4.7	4.8	4.2	4.3
Claude 3.5 Sonnet	4.4	4.4	4.6	4.1	4.1

อภิปรายผลการวิจัย

จากข้อจำกัดในการสื่อสารระหว่างพนักงานขายและลูกค้า การตอบข้อความที่มีความเร่งด่วน รวมถึงการขาด ความรู้เชิงลึกของพนักงานขายในการแนะนำสินค้า ทำให้ไม่สามารถแนะนำสินค้าที่เกี่ยวข้องหรือเพิ่มยอดขายได้อย่างเต็มที่ ดังนั้นการพัฒนา ระบบแชทบอทที่สามารถเข้าใจความต้องการและแนะนำสินค้าอย่างแม่นยำด้วยแอลแอลเอ็ม (Large Language Model) รวมกับเทคนิคการประแต่งแบบจำลอง (Fine-Tuning) และการดึงข้อมูล (Retrieval Augmented Generation:RAG) จึงเป็นแนวทางในการเพิ่มยอดขาย
อย่างมีประสิทธิภาพ

ผลการประเมินประสิทธิภาพและความพึงพอใจของผู้ใช้งาน พบว่าแบบจำลอง Llama-3.1-8B มีความโดดเด่นในด้านความแม่นยำในการตอบคำถาม ความลื่นไหลของการสนทนาและความเหมาะสมกับบริบทของการแนะนำสินค้า แม้ว่าความเร็วในการตอบสนองของโมเดลนี้จะต่ำกว่าแบบจำลองแบบปิดเล็กน้อย แต่ยังคงอยู่ในระดับที่ยอมรับได้สำหรับการใช้งานจริง

ด้วยเหตุนี้แบบจำลอง Llama-3.1-8B จึงได้รับการคัดเลือกเป็นโมเดลหลักในการพัฒนาระบบ แชทบอทสำหรับแพลตฟอร์มไลน์ออฟฟิศเชียลแอคเคาท์ โดยมุ่งเน้นการเพิ่มประสิทธิภาพในการตอบสนองต่อความต้องการของลูกค้า ช่วยให้การแนะนำสินค้ามีความแม่นยำ และสอดคล้องกับความต้องการของลูกค้ามากขึ้น อีกทั้งยังสร้างโอกาสในการเพิ่มยอดขายในระยะยาวได้ดี

นอกจากนี้ แม้ว่าแบบจำลอง Llama-3.1-8B จะให้ผลลัพธ์ที่ดี แต่ด้วยลักษณะการทำงานที่ต้องประมวลผลบนเครื่องคอม ซึ่งแตกต่างจากแบบจำลองแลลปิดอย่าง GPT-4 หรือ Claude 3.5 Sonnet ที่ใช้

การประมวลผลผ่าน API จึงจำเป็นต้องคำนึงถึง ประสิทธิภาพของเซิร์ฟเวอร์ที่ต้องรองรับการทำงานอย่างรวดเร็วและทันต่อความต้องการของผู้สนทนา การพึ่งพาเทคนิคการหาค่าที่เหมาะสมที่สุด (optimization) และแคช (cache) จึงมีความสำคัญในการลดภาระ การทำงานของเซิร์ฟเวอร์

ข้อเสนอแนะ

สำหรับการพัฒนาในอนาคต เมื่อข้อมูลได้รับการอัปเดต มากขึ้น จำเป็นต้องมีการปรับแต่งแบบจำลองข้อมูลใหม่และพิจารณา คัดกรองข้อมูลเก่าที่ล้าสมัยออกจากระบบในขั้นตอนการดึงข้อมูล เนื่องจากการมีข้อมูลชุดเดียวกันปริมาณมากแต่ใช้งานได้จริงน้อย อาจก่อให้เกิด Noise และส่งผลให้แอลแอลเอ็มประมวลผลคำตอบ ผิดพลาดจากการอ้างอิงชุดข้อมูลเก่าร่วมด้วย

1. การปรับปรุงและบำรุงรักษาระบบ เมื่อข้อมูลได้รับการอัปเดตมากขึ้น จำเป็นต้องมีการปรับแต่งแบบจำลองข้อมูลใหม่และพิจารณาคัดกรองข้อมูลเก่าที่ล้าสมัยออกจากระบบในขั้นตอนการดึงข้อมูล ดังนั้นควรพัฒนาระบบการจัดการข้อมูลอัตโนมัติที่สามารถคัดกรองและปรับปรุงชุดข้อมูลให้ทันสมัยอยู่เสมอ

2. การต่อยอดเชิงกลยุทธ์ทางธุรกิจ ควรขยายระบบให้ครอบคลุมกลุ่มลูกค้าอื่นๆ นอกเหนือจากกลุ่มช่าง เช่น เจ้าของโครงการ สถาปนิก หรือผู้รับเหมา ซึ่งแต่ละกลุ่มอาจมีความต้องการและการใช้งานที่แตกต่างกัน ดังนั้นควรพัฒนาระบบวิเคราะห์ข้อมูลเชิงลึก (Analytics) เพื่อติดตามพฤติกรรมการซื้อและช่วยในการวางแผนกลยุทธ์การตลาดและการขายที่เหมาะสม และควรศึกษาโอกาสในการขยายการใช้งานไปยังช่องทางอื่นๆ นอกเหนือจาก LINE เช่น Facebook Messenger หรือเว็บไซต์ของบริษัท

เอกสารอ้างอิง

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., & Wang, H. (2023). *Retrieval-augmented generation for large language models: A survey*. arXiv preprint arXiv:2312.10997. https://doi.org/10.48550/arXiv.2312.10997

Gupta, A., Shirgaonkar, A., Balaguer, A. D. L., Silva, B., Holstein, D., Li, D., & Benara, V. (2024). *RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture*. arXiv preprint arXiv:2401.08406. https://doi.org/10.48550/arXiv.2401.08406

Li, Y., Chen, H., Wei, J., Huang, R., Gu, S., Zha, D., & Williams, A. (2023). Enhancing Enterprise Customer Service with Large Language Models: A Case Study. *In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 952-964). Association for Computational Linguistics.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). *Large language models: A survey. arXiv preprint*. arXiv:2402.06196. https://doi.org/10.48550/arXiv.2402.06196

Neupane, S., Hossain, E., Keith, J., Tripathi, H., Ghiasi, F., Golilarz, N. A., Kaiser, J., Jiang, Y., Zhu, M., & Rahimi, S. (2024). *From Questions to Insightful Answers: Building an Informed Chatbot for University Resources*. arXiv preprint arXiv:2405.08120. https://arxiv.org/abs/2405.08120

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. arXiv preprint arXiv:1706.03762. https://doi.org/10.48550/arXiv.1706.03762

VM, K., Warrier, H., & Gupta, Y. (2024). *Fine Tuning LLM for Enterprise: Practical Guidelines and Recommendations*. arXiv preprint arXiv:2404.10779. https://doi.org/10.48550/arXiv.2404.10779