การผสานการปรับแต่งโมเดลและการดึงข้อมูล (RAG) เพื่อเพิ่มประสิทธิภาพโมเดลภาษาสำหรับงานสนทนาบนอุปกรณ์เอดจ์

Hybrid Fine-Tuning and Retrieval-Augmented Generation for Efficient Conversational Language Models on Edge Devices

บารมี สุขหงษ์
คณะวิศวกรรมศาสตร์ | วศ.ม. วิศวกรรมมหาบัณฑิต
สถาบันเทคโนโลยีไทย-ญี่ปุ่น
กรุงเทพมหานคร
2461110047@tni.ac.th

อัดนา เซนโต๊ะ
คณะวิศวกรรมศาสตร์ | วศ.บ. วิศวกรรมคอมพิวเตอร์และปัญญาประดิษฐ์ (CE)
สถาบันเทคโนโลยีไทย-ญี่ปุ่น
กรุงเทพมหานคร
*Corresponding author: adna@tni.ac.th

บทคัดย่อ **—ปัจจุบันงานวิจัยเกี่ยวกับการพัฒนาโมเดลภาษาสำหรับใช้งานบนอุปกรณ์ Edge Computing ที่ไม่ต้องอาศัยเครือข่ายภายนอก เช่น Raspberry Pi หรือ Nvidia Jetson เพื่อนำไปใช้งานด้านต่าง ๆ มักประสบข้อจำกัดด้านหน่วยความจำและหน่วยประมวลผล ส่งผลให้เกิดความหน่วงสูงและขาดความราบรื่นในการโต้ตอบอันเป็นอุปสรรคสำคัญต่อประสิทธิภาพการสื่อสาร ด้วยเหตุนี้ งานวิจัยนี้จึงนำเสนอแนวทางการเพิ่มประสิทธิภาพการสื่อสารด้วยโมเดลภาษาขนาดเล็กโดยใช้การปรับแต่ง (Fine-tuning) ให้เหมาะสมกับบริบทของชุดข้อมูลเฉพาะ นอกจากนี้ งานวิจัยนี้ยังใช้เทคนิคการบีบอัดด้วยการควอนไทเซชัน 4 บิต (4-bit Quantization) เพื่อให้สามารถประมวลผลบนสถาปัตยกรรมของอุปกรณ์คอมพิวเตอร์ขนาดเล็ก (Edge Computing) ได้อย่างราบรื่น สำหรับการทดสอบประสิทธิภาพของโมเดลที่นำเสนอนี้ ผู้วิจัยได้นำโมเดลภาษาดังกล่าวไปทดลองบนอุปกรณ์ประมวลผลขนาดเล็กโดยใช้ตัวชี้วัดได้แก่ ค่า BERTScoreF1 ความถูกต้องของเนื้อหา ความเร็วในการสร้างคำตอบ และอัตราการใช้หน่วยความจำเพื่อประเมินประสิทธิภาพของโมเดล จากผลการทดลองแสดงให้เห็นว่า โมเดลนำเสนอนี้สามารถทำงานได้อย่างคล่องตัวและมีประสิทธิภาพสูง โดยยังคงรักษาความแม่นยำของเนื้อหาไว้ในระดับที่ใกล้เคียงกับโมเดลภาษาขนาดใหญ่**

คำสำคัญ — การปรับแต่งโมเดลเฉพาะทาง, การประมวลผลบนอุปกรณ์ขนาดเล็ก, โมเดลภาษาขนาดใหญ่, การประมวลผลภาษาธรรมชาติภาษาไทย, ปัญญาประดิษฐ์เฉพาะโดเมน

ABSTRACT **—Currently, research on developing language models for use on edge computing devices such as Raspberry Pi or Nvidia Jetson, which do not rely on external networks for applications, often face limitations in memory and processing power. This results in high latency and a lack of smooth interaction, significantly hindering communication performance. Therefore, this research proposes an approach to improve communication performance with a small-language model by fine-tuning it to the context of a specific dataset. In addition, this research uses 4-bit quantization to enable smooth processing on edge computing architectures. For performance testing, the proposed model was tested on edge computing devices using BERTScoreF1, content accuracy, solution generation speed, and memory usage rate to evaluate its performance. The experimental results show that the proposed model can operate flexibly and efficiently while maintaining content accuracy at a level comparable to larger language models.**

Keywords — Fine-tuning, Edge Computing, Large Language Model, Thai NLP, Domain-specific AI.

I. บทนำ

Large Language Models (LLM) ได้ปฏิวัติ AI เชิงสนทนา โดยแสดงให้เห็นถึงความสามารถอันโดดเด่นในการสร้างข้อความที่เหมือนมนุษย์ ซึ่งมีการใช้งานอย่างกว้างขวางในด้านการบริการลูกค้า การดูแลสุขภาพ การศึกษา และอื่นๆ อีกมากมาย กระบวนการปรับแต่ง LLM ให้เข้ากับโดเมนเฉพาะ เช่น สภาพแวดล้อมทางวิชาการหรืออุตสาหกรรมเฉพาะ ช่วยให้โมเดลเหล่านี้มีประสิทธิภาพมากขึ้นในการตอบคำถามเฉพาะและให้คำตอบที่แม่นยำ โดยมีตัวอย่างงานวิจัยที่วิจัยที่เกี่ยวกับเรื่องนี้ดังต่อไปนี้ งานวิจัยของ Raj และคณะ ในปี 2024 [1] เป็นงานวิจัยที่เจาะลึกถึงกระบวนการที่ซับซ้อนของการปรับแต่ง (Fine-Tuning) โมเดลภาษาขนาดใหญ่แบบโอเพนซอร์ส (LLMs) เช่น LLaMA สำหรับการใช้งานในองค์กร โดยใช้ชุดข้อมูลที่เป็นกรรมสิทธิ์ ซึ่งประกอบด้วยข้อมูลทั้งแบบข้อความและที่มาจากคลังโค้ด งานนี้นำเสนอเวิร์กโฟลว์แบบครบวงจร ตั้งแต่การเตรียมและการประมวลผลล่วงหน้าของชุดข้อมูลไปจนถึงการทดลองด้วยการตั้งค่าการปรับแต่งที่หลากหลาย เพื่อเป็นแนวทางปฏิบัติสำหรับผู้ที่ต้องการปรับแต่ง LLMs ให้เหมาะสมกับการใช้งานในธุรกิจเฉพาะทาง การศึกษานี้เน้นวิธีการเตรียมชุดข้อมูลที่สำคัญ รวมถึงเทคนิคการแบ่งส่วนข้อมูล (Chunking) สำหรับข้อมูลข้อความ เช่น ข้อมูลดิบ คำสำคัญ หัวข้อ และคำแนะนำที่อิงตามคำค้นหา รวมถึงแนวทางเฉพาะสำหรับชุดข้อมูลโค้ด เช่น สรุปข้อมูลในระดับ

ฟังก์ชันและการแปลงข้อมูลเป็นโทเคน เทคนิคเหล่านี้ช่วยให้ชุดข้อมูลได้รับการปรับให้เหมาะสมกับการฝึกอบรม LLM ในขณะที่ยังคงรักษาบริบทสำคัญและลักษณะเฉพาะของโดเมนไว้ การทดลองแสดงให้เห็นถึงคุณค่าของวิธีการ เช่น การควอนไทเซชัน (Quantization) การสะสมเกรเดียนต์ (Gradient Accumulation) และวิธีการปรับแต่งแบบประหยัดพารามิเตอร์ (PEFT) เช่น LoRA และ QLoRA ซึ่งช่วยลดข้อกำหนดด้านการประมวลผลและหน่วยความจำได้อย่างมีนัยสำคัญ ทำให้สามารถปรับแต่งโมเดลบนฮาร์ดแวร์ที่มีทรัพยากรจำกัดได้โดยไม่ลดทอนประสิทธิภาพ งานวิจัยนี้ยังเน้นถึงประเด็นที่ควรสำรวจเพิ่มเติม เช่น การลดข้อมูลที่ไม่ถูกต้องผ่านเทคนิคการเตรียมชุดข้อมูลขั้นสูงและการปรับปรุงกลยุทธ์การสร้างพรอมต์ ด้วยข้อมูลเชิงลึกและแนวทางปฏิบัติที่ครอบคลุม งานนี้จึงเป็นรากฐานที่แข็งแกร่งสำหรับการใช้ประโยชน์จากโมเดล LLM ที่ได้รับการปรับแต่งเพื่อตอบสนองความต้องการที่เปลี่ยนแปลงไปของแอปพลิเคชัน AI ในองค์กร

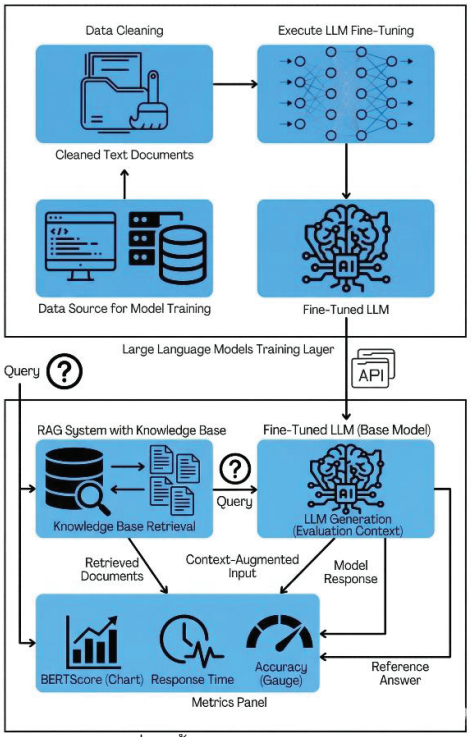
งานวิจัยต่อมาเป็นของ J. M. Bink [2] ซึ่งเป็นการศึกษาเกี่ยวกับศักยภาพที่สำคัญของปัญญาประดิษฐ์เชิงสร้างสรรค์ (Generative AI) โดยเฉพาะโมเดลภาษาขนาดใหญ่ (LLMs) เช่น GPT-4 ในการเปลี่ยนแปลงการสนับสนุนลูกค้าผ่านการปรับแต่งและคุณภาพของการโต้ตอบที่ดีขึ้น ด้วยการผสานเทคนิคขั้นสูง เช่น การสร้างกรอบการประเมินคุณภาพของแชตบอต (Framework for Assessment of Chatbot Quality: FACQ) ซึ่งเป็นวิธีที่มีความแม่นยำในการประเมินและวัดความสามารถในการปรับแต่งของแชตบอตที่ขับเคลื่อนด้วย LLM FACQ อย่างไรก็ตาม การศึกษานี้ยอมรับข้อจำกัดหลายประการ การมุ่งเน้นที่การโต้ตอบแบบเทิร์นเดียวไม่สามารถแก้ไขความซับซ้อนของการมีส่วนร่วมที่ยาวนานหรือแบบหลายรูปแบบได้อย่างสมบูรณ์ และมีงานวิจัยอื่นๆ ที่ได้รับการตีพิมพ์เป็นจำนวนมากเช่น การวิจัยในงานการวิเคราะห์เชิงลึกของห้าโมเดลแชทบอทปัญญาประดิษฐ์ชั้นนำ ได้แก่ ChatGPT, Bard, Llama, Ernie และ Grok โดยประเมินความสามารถ ข้อจำกัด [3] หรือจะเป็นการปรับแต่งแบบประหยัดพารามิเตอร์ (PEFT) และเทคนิค Retrieval-Augmented Generation (RAG) ด้วยการใช้ LangChain เป็นไลบรารีพื้นฐาน [4] หรือจะเป็นงานวิจัยนี้เสนอการตรวจสอบอย่างครอบคลุมถึงผลกระทบทางจริยธรรมที่เกี่ยวข้องกับการพัฒนาและการปรับใช้โมเดลภาษาขนาดใหญ่ (LLMs) ระบบ AI ขั้นสูงเหล่านี้ แม้ว่าจะมีความสามารถในการเปลี่ยนแปลง แต่ก็ก่อให้เกิดความท้าทายเฉพาะตัวที่ต้องการความสนใจและการดำเนินการอย่างเร่งด่วน ซึ่งแตกต่างจาก AI แบบดั้งเดิม LLMs เผชิญกับปัญหา เช่น การหลอนภาพ (hallucination) ความโปร่งใส การเซ็นเซอร์ และความรับผิดชอบ ซึ่งถูกขยายโดยการผสานรวมที่เพิ่มขึ้นในโดเมนต่างๆ ของสังคม [5] และสุดท้ายเป็นงานวิจัยที่เน้นย้ำถึงความจำเป็นอย่างยิ่งสำหรับกลไกความปลอดภัยที่แข็งแกร่งและสามารถปรับตัวได้ในการปรับใช้โมเดลภาษาขนาดใหญ่ (LLMs) ด้วยศักยภาพในการเปลี่ยนแปลงในทุกอุตสาหกรรม LLMs แต่อย่างไรก็ตาม งานวิจัยรับทราบถึงความท้าทาย เช่น การรักษาสมดุลระหว่างการปกป้องความเป็นส่วนตัวกับประสิทธิภาพของระบบ การลดผลบวกปลอม (false positives) และการเพิ่มประสิทธิภาพการคำนวณของกรอบงาน งานวิจัยในอนาคตเสนอให้เพิ่มประสิทธิภาพของโมดูล เพิ่มความสามารถในการปรับตัวแบบไดนามิก และขยายขอบเขตการประยุกต์ใช้ของกรอบงานไปสู่บริบทแบบมัลติโมดัลและข้ามภาษา [6]

II. วัตถุประสงค์ของการวิจัย

1. สร้างชุดข้อมูลเฉพาะของหน้าเว็บของสถาบันเทคโนโลยีไทย-ญี่ปุ่นบนคอมพิวเตอร์ในอยู่ในรูปแบบที่สามารถนำไป Fine-tuning บนโมเดลภาษา
2. พัฒนาโมเดลภาษา AI ด้วยโมเดล LLAMA ให้สามารถเรียนรู้และตอบคำถามเกี่ยวกับข้อมูลเฉพาะของหน้าเว็บของสถาบันเทคโนโลยีไทย-ญี่ปุ่นบนคอมพิวเตอร์ขนาดเล็กได้โดยมีคะแนนมีความแม่นยำและความถูกต้องไม่น้อยกว่าร้อยละ 85%

III. วิธีดำเนินการวิจัย

งานวิจัยนี้มุ่งเน้นค้นหาสถาปัตยกรรมโมเดลภาษาที่เหมาะสมสำหรับใช้งานบนคอมพิวเตอร์ขนาดเล็กสำหรับงานการสนทนา (Chatbot) ด้วยข้อมูลเฉพาะ ตัวอย่างเช่น ในบริบททางการศึกษาโดยเฉพาะภายในสถาบันเทคโนโลยีไทย-ญี่ปุ่น (TNI) ในด้านการสื่อสารของมหาวิทยาลัย ข้อความทางวิชาการ และคำถามที่ถามบ่อยจากนักเรียนและคณาจารย์เพื่อให้ได้โมเดลที่แม่นยำและสามารถตอบสนองต่อผู้ใช้ได้ดี และด้วยเหตุที่ต้องการเพิ่มประสิทธิภาพการสื่อสารในด้านการความเร็ว การเข้าใจบริบทพื้นฐานของกลุ่มข้อมูล และการเข้าใจคำศัพท์เฉพาะด้านงานวิจัยนี้จึงเลือกวิธีการของเทคนิคการปรับแต่ง (Fine-Tunning) เพื่อค้นหาโมเดลภาษาที่เหมาะสม จากนั้นนำโมเดลมาเสริมความถูกต้องด้วยเทคนิคการสืบค้นข้อมูล (Retrieval-Augmented Generation, RAG) โมเดลภาษาขนาดเล็กที่ผ่านการปรับแต่งนี้จะถูกบีบอัดด้วยเทคนิคการควอนไทเซชัน 4 บิตเพื่อลดขนาดของโมเดล ซึ่งขั้นตอนการดำเนินงานการวิจัยที่นำเสนอนี้สามารถแสดงได้ดังรูปที่ 1



รูปที่ I. ขั้นตอนการดำเนินงานการวิจัย

III.I การเก็บรวบรวมข้อมูล

แหล่งข้อมูลที่ใช้ในการฝึกโมเดลจะใช้ข้อมูลจากเว็บไซต์ของสถาบันในส่วนของงานบริการด้านการศึกษา เช่น ข้อมูลเกี่ยวกับหลักสูตร รายชื่อคณาจารย์ ข้อมูลทุนการศึกษา และประกาศสำคัญ ตัวอย่างการสื่อสารภายใน เช่น อีเมล แจ้งข่าวสาร และคำถามที่พบบ่อยจากนักศึกษามาเข้าสู่กระบวนการทำความสะอาดข้อมูล (Data Cleaning and Preprocessing) โดยลบข้อมูลที่ไม่เกี่ยวข้อง กรองข้อมูลให้ตรงกับเนื้อหาที่จำเป็นทำให้ข้อมูลเป็นนิรนาม (Anonymization) เพื่อลดความเสี่ยงด้านความเป็นส่วนตัวของนักศึกษาและบุคลากร แปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสม แยกประเภทข้อมูล เช่น คำถาม-คำตอบ ข้อมูลเชิงโครงสร้าง และข้อมูลที่ใช้สำหรับการฝึกโมเดล โดยจะมีการจัดหมวดหมู่ข้อมูลออกเป็น 3 หมวดหลัก ได้แก่ ข้อมูลทางวิชาการ (เช่น หลักสูตร วิชาเรียน และคู่มือการศึกษา) ข้อมูลการบริหารงาน (เช่น การลงทะเบียน ตารางสอบ และค่าธรรมเนียม) ข้อมูลทั่วไปของสถาบัน (เช่น กฎระเบียบของมหาวิทยาลัย สถานที่ และการติดต่อ) และจะมีการตรวจสอบคุณภาพข้อมูลใช้ Human Review และ Automated Validation เพื่อตรวจสอบว่าข้อมูลที่ใช้มีความถูกต้องและสอดคล้องกับวัตถุประสงค์ของการวิจัย

ชุดข้อมูลที่ใช้ในการ Fine-Tuning มีทั้งหมด 4,441 ตัวอย่าง (rows) ในรูปแบบคู่คำถาม-คำตอบ (Question-Answer pairs) ซึ่งครอบคลุมทั้งสามหมวดหมู่ข้อมูลที่กล่าวถึงข้างต้น ได้แก่ ข้อมูลทางวิชาการ ข้อมูลการบริหารงาน และข้อมูลทั่วไปของสถาบัน สำหรับระบบ RAG นั้น Knowledge Base ถูกสร้างจากเอกสารข้อความที่ผ่านการทำความสะอาดแล้ว โดยใช้เทคนิค Text Chunking แบ่งเนื้อหาออกเป็น chunk ย่อยพร้อม overlap เพื่อรักษาความต่อเนื่องของบริบท จากนั้นแปลงเป็น vector embeddings และจัดเก็บใน vector store สำหรับการสืบค้นแบบ semantic search ในขั้นตอนการ inference โดยระบบจะดึงเอกสารที่เกี่ยวข้องมาเสริมใน context ก่อนส่งให้โมเดลสร้างคำตอบ

III.II วิธีการปรับแต่งโมเดล

เพื่อให้โมเดลสามารถตอบคำถามได้อย่างแม่นยำและสอดคล้องกับบริบทของสถาบัน จึงต้องดำเนินการปรับแต่งโดยใช้เทคนิคที่เหมาะสม ได้แก่

1. การเลือกโมเดลพื้นฐาน (Base Model Selection) กำหนดให้ใช้โมเดล Meta LLaMA เป็นโมเดลพื้นฐานในการปรับแต่ง เนื่องจากมีประสิทธิภาพสูงและสามารถรองรับการประมวลผลข้อความภาษาไทยและอังกฤษ
2. กระบวนการปรับแต่งโมเดล (Fine-Tuning Process) มีการใช้ชุดข้อมูลที่ผ่านการเตรียมไว้เพื่อฝึกโมเดล และใช้เทคนิค Supervised Fine-Tuning โดยมีตัวอย่างคำถามและคำตอบที่ถูกต้อง จากนั้นมีการปรับพารามิเตอร์สำคัญ เช่น Learning Rate, Batch Size และ Training Steps เพื่อให้โมเดลสามารถเรียนรู้ข้อมูลใหม่ได้อย่างมีประสิทธิภาพ
3. การใช้เครื่องมือช่วยปรับแต่ง (Utilizing Optimization Tools) มีการใช้ Parameter-Efficient Fine-Tuning เช่น LoRA (Low-Rank Adaptation) และ QLoRA เพื่อให้สามารถปรับแต่งโมเดลได้โดยใช้ทรัพยากรคอมพิวเตอร์ที่น้อยลง
4. การกำหนดบริบทและน้ำเสียงของโมเดล (Contextual and Tone Adjustments) มีการปรับแต่งให้โมเดลสามารถเข้าใจบริบทของคำถามและให้คำตอบในโทนที่เหมาะสมกับนักศึกษาและบุคลากร และฝึกให้โมเดลสามารถจัดลำดับความสำคัญของข้อมูลที่เกี่ยวข้อง
5. การทดสอบประสิทธิภาพของโมเดล (Testing and Evaluation) มีการทดสอบโมเดลกับคำถามจริงจากนักศึกษาและบุคลากร และใช้ตัวชี้วัด BLEU Score และ F1 Bert Score เพื่อประเมินประสิทธิภาพ
6. การนำผลการทดสอบมาปรับปรุงโมเดล (Iterative Refinement) มีการปรับปรุงโมเดลตามผลการวิเคราะห์ข้อผิดพลาด (Error Analysis) และแก้ไขปัญหาที่เกี่ยวข้องกับการให้คำตอบผิดพลาด (Hallucination) และความลำเอียงของโมเดล (Bias)

โดยมีกระบวนการของการเขียนโค้ดโปรแกรมดังลิตส์ข้างล่างนี้

```
Algorithm: Institutional_Aware_LLM_Training
Input:
    D_train = ชุดข้อมูลคำถาม-คำตอบ (training dataset)
    D_eval  = ชุดข้อมูลทดสอบ (evaluation dataset)
    BaseModel = LLM model
    Hyperparameters = {learning_rate, batch_size, training_steps}
Output:
    FineTunedModel
Begin
1. Model ⟵ Load(BaseModel)
2. Model ⟵ Apply_LoRA(Model)
   Model ⟵ Apply_QLoRA(Model)
3. D_train ⟵ Clean(D_train)
   D_train ⟵ Tokenize(D_train)
   D_train ⟵ Format_QA_Pairs(D_train)
4. For each sample in D_train do
       sample ⟵ Add_Context_Tag(sample, "institutional")
       sample ⟵ Adjust_Tone(sample, "formal-friendly")
   EndFor
5. // Fine-Tuning Process (Supervised Learning)
   For step = 1 to training_steps do
       batch ⟵ Sample(D_train, batch_size)
       predictions ⟵ Model.forward(batch.input)
       loss ⟵ Compute_Loss(predictions, batch.target)
       Model ⟵ Backpropagate(Model, loss, learning_rate)
   EndFor
6. // Evaluation Phase
   results ⟵ {}
   For each sample in D_eval do
       output ⟵ Model.generate(sample.input)
       BLEU Score ⟵ Compute_ BLEU(output)
       F1 Bert Score ⟵ Compute_ F1Bert(output
       results ⟵ Store(results, BLEU Score, F1 Bert Score)
   EndFor
End
```

III.III การวิเคราะห์โมเดล

หลังจากการปรับแต่งโมเดลเสร็จสิ้น จะมีการประเมินผลเพื่อวิเคราะห์ประสิทธิภาพและความถูกต้องของโมเดลภาษา โดยขั้นตอนนี้มีการนำเทคนิคของ RAG มาใช้ด้วย โดยมีเกณฑ์ในการพิจารณาประสิทธิภาพของโมเดลได้แก่ ความแม่นยำของการตอบสนอง (Response Accuracy) ซึ่งจะมีการเปรียบเทียบคำตอบของโมเดลที่ได้รับการปรับแต่งกับคำตอบมาตรฐานจากผู้เชี่ยวชาญ และใช้ตัวชี้วัด เช่น BLEU Score และ F1 Bert Score เพื่อวัดคุณภาพของการตอบกลับ ต่อมาจะวัดความสอดคล้องของคำตอบ (Contextual Coherence) มีการตรวจสอบว่าโมเดลสามารถให้คำตอบที่เหมาะสมกับบริบทของคำถามได้หรือไม่ และวัดความสามารถในการเข้าใจคำถามเชิงลึก (Comprehension and Depth Analysis) มีการทดสอบโมเดลกับคำถามที่ซับซ้อนขึ้น เช่น คำถามแบบ Multi-Turn Conversations หรือคำถามที่ต้องอาศัยการให้เหตุผล และตรวจสอบว่าสามารถเชื่อมโยงข้อมูลจากแหล่งข้อมูลหลายแหล่งได้อย่างมีประสิทธิภาพ พร้อมกันนี้มีการตรวจสอบข้อผิดพลาดของโมเดล (Error Analysis) มีการวิเคราะห์กรณีที่โมเดลให้คำตอบผิดพลาด โดยจำแนกเป็นHallucination (การสร้างข้อมูลที่ไม่มีอยู่จริง) Bias (ความลำเอียงในการให้ข้อมูล) Lack of Specificity (คำตอบที่กว้างเกินไปหรือไม่ชัดเจน) และใช้ข้อมูลจากข้อผิดพลาดเหล่านี้เพื่อปรับแต่งโมเดลเพิ่มเติม และสุดท้ายจะเป็นการเปรียบเทียบกับโมเดลอื่น (Benchmarking Against Other Models) มีการนำโมเดลที่ได้รับการปรับแต่งไปเปรียบเทียบกับโมเดลพื้นฐาน เช่น Llama-3.1-8B, Llama-3.2-3B, และ OpenThaiGPT-7B และประเมินข้อดีข้อเสียของโมเดลที่พัฒนาเทียบกับโมเดลที่มีอยู่

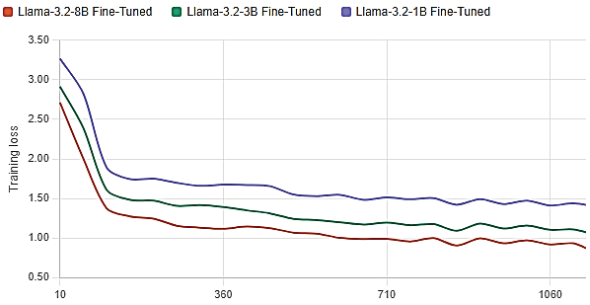
งานวิจัยนี้ใช้ทั้ง BLEU Score และ BERTScore F1 โดยมีบทบาทต่างกัน BLEU ถูกใช้เป็นตัวชี้วัดเสริมเพื่อให้สามารถเปรียบเทียบกับงานวิจัยอื่นได้ แม้จะมีข้อจำกัดในงานสนทนาภาษาไทยเนื่องจากวัดเฉพาะการจับคู่คำแบบตรงตัว (Exact Match) ในขณะที่ BERTScore F1 ถูกเลือกเป็นตัวชี้วัดหลักเพราะวัดความคล้ายคลึงเชิงความหมาย (Semantic Similarity) ด้วย contextual embeddings ทำให้เหมาะสมกับบริบทภาษาไทยมากกว่า ดังนั้นค่า BLEU ที่ต่ำในขณะที่ BERTScore F1 สูงจึงไม่ขัดแย้งกัน แต่สะท้อนว่าโมเดลสร้างคำตอบที่ถูกต้องเชิงความหมายแม้จะใช้คำหรือโครงสร้างประโยคที่แตกต่างจากคำตอบอ้างอิง

IV. ผลการวิจัย

เพื่อประเมินประสิทธิภาพของโมเดลภาษา ทางผู้วิจัยได้เตรียมชุดคำถามโดยเทียบกับเฉลยที่เตรียมเอาไว้จำนวน 10 ตัวอย่าง โดยสุ่มตัวอย่างจากชุดข้อมูลทั้งหมดเพื่อนำไปทดสอบกับโมเดลที่ได้รับการ Fine-Tuning แล้ว และโมเดลภาษาทั่วไปได้แก่ Fine-Tuned Llama-3.1-8B, Llama-3.1-8B, Fine-Tuned Llama-3.2-3B, Llama-3.2-3B, Fine-Tuned Llama-3.2-1B, Llama-3.2-1B, Fine-Tuned OpenthaiGPT7B, และ OpenthaiGPT7B โดยผู้วิจัยได้ดำเนินการทดลองดังนี้ ส่วนแรกเป็นผลการทดลองการเปลี่ยนแปลงของพารามิเตอร์ (Loss) ของการฝึกสอนแสดงดังรูปที่ 2 ต่อมาส่วนของตัวอย่างผลลัพธ์บางส่วนของการโต้ตอบจากโมเดลภาษาแสดงได้ดังรูปที่ 3 และส่วนสุดท้ายจะเป็นส่วนของผลการทดลองเกี่ยวการประเมินประสิทธิภาพของโมเดลภาษาแสดงได้ดังตารางที่ 1 ถึงตารางที่ 4

Instruction
วิชา BUS-211 การเป็นผู้ประกอบการ สอนเรื่อง "Venture Capital" (VC) เบื้องต้นหรือไม่
วิชา ENG-101 ศึกษาเรื่อง "Missing Line/View Problems" เพื่ออะไร
วิชา ENG-102 การเขียนโปรแกรมฯ ศึกษาเรื่อง "Debugging" อย่างไร
Reference_Answer
สอน เพื่อให้วิศวกรเข้าใจแหล่งเงินทุนและแนวทางการระดมทุนสำหรับสตาร์ทอัพ
เพื่อฝึกจินตนาการและการอ่านแบบโดยการหาเส้นหรือมุมมองที่หายไปจากแบบแปลน
ศึกษาขั้นตอนการตรวจสอบและแก้ไขข้อผิดพลาดในโค้ดให้โปรแกรมทำงานได้ตามอัลกอริทึม
Generated_Response
สอน เพื่อให้เข้าใจแหล่งเงินทุนสำหรับธุรกิจสตาร์ทอัพ
เพื่อฝึกการสร้างภาพ 3 มิติจากภาพ 2 มิติในแบบจำลองเครื่องจักร
ศึกษาการตรวจสอบและแก้ไขปัญหาในโค้ดโปรแกรมเพื่อให้ทำงานได้ตามที่ต้องการ

รูปที่ II. ตัวอย่างคำถาม คำตอบอ้างอิง และผลการสร้างคำตอบ



รูปที่ III. กราฟแสดงค่าผิดพลาดระหว่างการฝึกสอน

ตารางที่ 1: ผลการประเมินการทดสอบประสิทธิภาพ

Model	Type	BLEU	BERTScoreF1
Llama-3.1-8B	Base	0.027	0.516
	Fine-Tuned	0.389	0.809
Llama-3.2-3B	Base	0.038	0.610
	Fine-Tuned	0.308	0.778
Llama-3.2-1B	Base	0.039	0.538
	Fine-Tuned	0.235	0.745
OpenthaiGPT7B	Base	0.031	0.609
	Fine-Tuned	0.072	0.689

จากรูปที่ 3 แสดงถึงกราฟของ Loss ในระหว่างการฝึกสอนของโมเดล Llama3.2 จะเห็นว่าค่ารอบประมาณที่ 700 ค่า Loss ถึงจุดที่ไม่เปลี่ยนแปลง กล่าวคือในเหมาะสมของของการฝึกสอนของโมเดลนี้คือ 700 รอบนั้นเอง และสำหรับตารางที่ 1 ผลการประเมินการทดสอบประสิทธิภาพแสดงให้เห็นอย่างชัดเจนว่าการ Fine-Tuning ช่วยพัฒนาประสิทธิภาพของทุกโมเดลอย่างมีนัยสำคัญ โดย Llama-3.1-8B Fine-Tuned มีประสิทธิภาพสูงที่สุดในกลุ่ม มีค่า BLEU เพิ่มขึ้นจาก 0.027 เป็น 0.389 (เพิ่มขึ้น 14 เท่า) และค่า BERTScore F1 เพิ่มขึ้นจาก 0.516 เป็น 0.809 สะท้อนให้เห็นว่าการปรับแต่งเฉพาะ domain ทำให้โมเดลสามารถสร้างคำตอบที่มีความหมายใกล้เคียงกับคำตอบมาตรฐานได้ดีขึ้นอย่างมาก และสำหรับโมเดล Llama-3.2-3B Fine-Tuned มีค่า BLEU เพิ่มจาก 0.038 เป็น 0.308 และ BERTScore F1 จาก 0.610 เป็น 0.778 ในขณะที่ Llama-3.2-1B Fine-Tuned มีค่า

BLEU เพิ่มจาก 0.039 เป็น 0.235 และ BERTScore F1 จาก 0.538 เป็น 0.745 ผลดังกล่าวสะท้อนให้เห็นว่าขนาดของโมเดลมีความสัมพันธ์โดยตรงกับประสิทธิภาพหลัง Fine-Tuning โดย Llama-3.1-8B FT > Llama-3.2-3B FT > Llama-3.2-1B FT ตามลำดับ สำหรับ OpenThaiGPT-7B Fine-Tuned มีค่า BLEU เพิ่มจาก 0.031 เป็น 0.072 และ BERTScore F1 จาก 0.609 เป็น 0.689 ซึ่งแม้จะมีการพัฒนาขึ้น แต่ยังต่ำกว่าโมเดลอื่นในกลุ่ม Fine-Tuned อย่างมีนัยสำคัญ บ่งชี้ว่าสถาปัตยกรรมของ OpenThaiGPT-7B อาจต้องการชุดข้อมูล Fine-Tuning ที่มีขนาดใหญ่กว่าเพื่อให้เกิดการปรับตัวกับ domain เฉพาะได้อย่างเต็มประสิทธิภาพ

ทั้งนี้ ในการตีความค่า BLEU Score ที่ปรากฏในตารางที่ 1 ควรคำนึงว่า BLEU วัดความแม่นยำโดยอาศัยการจับคู่คำแบบตรงตัว (Exact Match) ซึ่งไม่เหมาะกับงานสนทนาภาษาไทยที่สามารถสื่อความหมายเดียวกันได้ด้วยคำพ้องหรือโครงสร้างประโยคที่หลากหลาย ดังนั้นค่า BLEU ที่ต่ำจึงไม่ได้สะท้อนถึงปัญหา Hallucination หรือความล้มเหลวของการ Fine-Tuning แต่อย่างใด โดยหลักฐานที่ชัดเจนกว่าปรากฏจากค่า BERTScore F1 ซึ่งวัดความคล้ายคลึงเชิงความหมาย (Semantic Similarity) โดยโมเดล Fine-Tuned ทุกตัวได้คะแนนสูงกว่า 0.74 และโมเดล Hybrid ได้ถึง 0.9434 จึงควรใช้ BERTScore F1 เป็นตัวชี้วัดหลักในการประเมินโมเดลภาษาสำหรับงานสนทนาภาษาไทย

ตารางที่ 2: การประเมินเชิงคุณภาพ

Model	Type	Contextual Coherence	Comprehension Depth	Multi-turn Capability
Llama-3.1-8B	Base	1.8	1.5	1.4
	Fine-Tuned	4.2	3.9	4.0
Llama-3.2-3B	Base	2.8	2.5	2.2
	Fine-Tuned	3.8	3.5	3.6
Llama-3.2-1B	Base	2.0	1.8	1.6
	Fine-Tuned	3.5	3.2	3.3
Openthai GPT7B	Base	2.3	2.0	1.8
	Fine-Tuned	2.5	2.2	2.0

ตารางที่ 3: การวิเคราะห์ความผิดพลาด

Model	Type	Hallucination (%)	Bias (%)	Lack of Specificity (%)
Llama-3.1-8B	Base	80.0	30.0	90.0
	Fine-Tuned	20.0	10.0	20.0
Llama-3.2-3B	Base	60.0	20.0	70.0
	Fine-Tuned	30.0	10.0	30.0
Llama-3.2-1B	Base	70.0	20.0	80.0
	Fine-Tuned	40.0	10.0	40.0
OpenthaiGPT7B	Base	70.0	20.0	80.0
	Fine-Tuned	60.0	20.0	70.0

ในด้านการประเมินเชิงคุณภาพ (ตารางที่ 2) พบว่า Llama-3.1-8B Fine-Tuned มีคะแนนสูงที่สุดในทุกมิติที่ 4.2, 3.9 และ 4.0 ตามลำดับ สะท้อนให้เห็นว่าการ Fine-Tuning สามารถพัฒนาคุณภาพการตอบได้อย่างมีนัยสำคัญ โดย Llama-3.1-8B Base มีคะแนนต่ำที่สุดที่ 1.8, 1.5 และ 1.4 เนื่องจากพบปัญหา repetition loop ซ้ำในหลายคำถาม Llama-3.2-3B Fine-Tuned และ Llama-3.2-1B Fine-Tuned มีคะแนนในระดับดีที่ 3.5–3.8 ซึ่งสูงกว่า Base อย่างชัดเจน ส่วน OpenThaiGPT-7B Fine-Tuned มีคะแนนต่ำที่สุดในกลุ่ม Fine-Tuned เนื่องจากพบปัญหา token แตก และ response format ที่ผิดปกติ ส่งผลให้คุณภาพการสื่อสารลดลง

จากการวิเคราะห์ข้อผิดพลาด (ตารางที่ 3) พบว่า Llama-3.1-8B Fine-Tuned มีอัตรา Hallucination ต่ำที่สุดในกลุ่มทั้งหมดที่ 20.0% ลดลงจาก Base ถึง 60 percentage points แสดงให้เห็นว่า Fine-Tuning ช่วยควบคุมคุณภาพคำตอบได้อย่างมีประสิทธิภาพ Llama-3.1-8B Base มีอัตรา Hallucination และ Lack of Specificity สูงที่สุดที่ 80.0% และ 90.0% ตามลำดับ OpenThaiGPT-7B Fine-Tuned ยังคงมีอัตรา error สูงที่ 60.0% เนื่องจากปัญหา token fragmentation ที่เกิดจาก tokenizer ที่ไม่เข้ากันกับ prompt format ในภาพรวม (ตารางที่ 4) Llama-3.1-8B Fine-Tuned ถูกกำหนดให้เป็น Proposed Model ในฐานะ baseline ที่มีประสิทธิภาพสูงที่สุด โดยมีค่า BLEU 0.389, BERTScoreF1 0.809 และ Hallucination ต่ำที่สุดที่ 20.0% อย่างไรก็ตาม เนื่องจากโมเดลขนาด 8B มีข้อจำกัดด้านทรัพยากรในการ deploy บนอุปกรณ์ขนาดเล็ก ผู้วิจัยจึงพิจารณา Llama-3.2-3B Fine-Tuned ในฐานะตัวเลือกที่มีศักยภาพสูงสุดสำหรับการนำไปใช้งานบน Edge Device โดยมีค่า Perplexity ต่ำที่สุดในกลุ่มทั้งหมดที่ 11.25 BERTScoreF1 ที่ 0.778 และ BLEU ที่ 0.308 ซึ่งถือว่าใกล้เคียงกับ Proposed Model ในขณะที่ใช้ทรัพยากรน้อยกว่าอย่างมีนัยสำคัญ

ตารางที่ 4: การประเมินก่อนการนำไปใช้งานจริง

Model	Technique	Accuracy	BERT ScoreF1	Final Score	TPS
Llama-3.2-3B	Base	0.4710	0.6943	0.605	15.38
	Fine-Tuning	0.4306	0.7781	0.6391	14.59
	RAG Only	0.8140	0.8358	0.8271	8.49
	Hybrid (FT+RAG)	0.8614	0.9434	0.9106	7.58

หลังจากการประเมินผลเพื่อวิเคราะห์ประสิทธิภาพและความถูกต้องของโมเดลภาษาดังกล่าวที่ได้รับการปรับแต่งโมเดลเสร็จสิ้นแล้ว สามารถสรุปได้ว่าโมเดลที่เหมาะสมที่สุดสำหรับการนำไปใช้งานบนคอมพิวเตอร์ขนาดเล็กคือโมเดล Llama-3.2-3B ด้วยเหตุที่จะต้องนำไปใช้งานจริงจึงได้นำเทคนิคของ RAG มาใช้เพื่อประเมินประสิทธิภาพของโมเดลภาษาโดยวัดความถูกต้องของการสร้างคำตอบ (Accuracy) ความถูกต้องในด้านการเข้าใจบริบท (BertScoreF1) ค่าเฉลี่ยของ Accuracy และ BertScoreF1 เพื่อการตัดสินใจ (Final Score) ค่าดัชนีวัดความเร็วในการสร้างคำตอบ (Token per second, TPS) ซึ่งผลการทดลองแสดงได้ดังตารางที่ 4

จากตารางที่ 4 การประเมินก่อนการนำไปใช้งานจริงบน Edge Device โดยใช้โมเดล Llama-3.2-3B เปรียบเทียบ 4 เทคนิค พบว่า Hybrid (Fine-Tuning + RAG) ให้ผลลัพธ์ดีที่สุดในทุกตัวชี้วัด โดยมีค่า

Accuracy 0.8614 BERTScore 0.9434 และ Final Score 0.9106 ซึ่งสูงกว่า Fine-Tuning เพียงอย่างเดียวที่ Final Score เพียง 0.6391 และ RAG เพียงอย่างเดียวที่ 0.8271 อย่างมีนัยสำคัญ แสดงให้เห็นว่าการผสานทั้งสองเทคนิคเข้าด้วยกันสามารถชดเชยข้อจำกัดของแต่ละวิธีได้อย่างมีประสิทธิภาพ กล่าวคือ Fine-Tuning ช่วยให้โมเดลเข้าใจบริบทและรูปแบบการตอบคำถามของ domain เฉพาะ ในขณะที่ RAG ช่วยดึงข้อมูลที่ถูกต้องและทันสมัยมาเสริมความแม่นยำ ลด Hallucination ที่พบในการใช้ Fine-Tuning เพียงอย่างเดียว อย่างไรก็ตาม โมเดล Hybrid มีค่า Token per Second (TPS) อยู่ที่ 7.58 ซึ่งเป็นค่าต่ำที่สุดในขณะที่โมเดลอื่นๆ ได้แก่ โมเดล RAG Only, โมเดล Fine-Tuning, โมเดล Base เพียงอย่างเดียวมีค่า TPS อยู่ที่ 8.49, 14.59, 15.38 ตามลำดับ ซึ่งผลการทดลองนี้บ่งชี้ว่าโมเดลทั้ง 3 นี้ (RAG Only, Fine-Tuning และ Base) มีต้นทุนด้านเวลาในการสืบค้นข้อมูลที่ดีกว่า อย่างไรก็ตาม ค่า TPS ของ Hybrid (Fine-Tuning + RAG) ยังอยู่ในระดับที่ยอมรับได้สำหรับการใช้งานบน Edge Device โดยไม่ส่งผลกระทบต่อประสบการณ์การสนทนาของผู้ใช้ จึงสามารถสรุปได้ว่าสถาปัตยกรรม Hybrid Fine-Tuning และ RAG บน Llama-3.2-3B เป็นแนวทางที่เหมาะสมที่สุดสำหรับการนำไปใช้งานระบบสนทนาบน Edge Device ในบริบทเฉพาะ โดยสามารถรักษาความแม่นยำในระดับสูงควบคู่กับ Token per second ที่เหมาะสมสำหรับการใช้งานจริง

V. สรุปและอภิปรายผล

งานวิจัยนี้นำเสนอแนวทางการเพิ่มประสิทธิภาพการสื่อสารด้วยโมเดลภาษาขนาดเล็กโดยใช้การปรับแต่ง (Fine-tuning) ให้เหมาะสมกับบริบทของชุดข้อมูลเฉพาะ ซึ่งจากผลการทดลองแสดงให้เห็นว่าโมเดลภาษา Llama-3.2-3B Fine-Tuned เหมาะสมสำหรับการใช้งานบนอุปกรณ์ Edge Computing โดยพิจารณาจากความสมดุลระหว่างประสิทธิภาพและทรัพยากรที่ต้องการ แม้ Llama-3.1-8B Fine-Tuned จะมีประสิทธิภาพสูงที่สุดในเชิงตัวชี้วัด แต่ขนาด 8,000 ล้านตัวแปรทำให้มีข้อจำกัดในการนำไปใช้งานบนอุปกรณ์ที่มีหน่วยความจำที่จำกัด ในขณะที่ Llama-3.2-3B Fine-Tuned มีขนาดเล็กกว่าถึง 2.5 เท่า แต่ยังคงให้ประสิทธิภาพที่ดี นอกจากนี้ งานวิจัยนี้ยังใช้ด้วยเทคนิคการบีบอัดด้วยการควอนไทเซชัน 4 บิต (4-bit Quantization) เพื่อขนาดเล็กลงและยังคงรรักษาความถูกต้องที่ใกล้เคียงกับโมเดลเดิมทำให้สามารถประมวลผลบนสถาปัตยกรรมของอุปกรณ์คอมพิวเตอร์ขนาดเล็ก (Edge Computing) ได้อย่างราบรื่น อย่างไรก็ตาม การเปรียบเทียบระหว่าง Llama-3.1-8B Fine-Tuned และ Llama-3.2-3B Fine-Tuned บน Edge Device ชี้ให้เห็นถึงข้อดีข้อเสียที่ชัดเจน กล่าวคือ Llama-3.1-8B Fine-Tuned มีความแม่นยำและมีค่าดัชนีวัดความเร็วในการสร้างคำตอบสูงกว่าอย่างมีนัยสำคัญ แต่ต้องการทรัพยากรสูงกว่าเมื่อนำไปใช้งานบนอุปกรณ์ขนาดเล็ก ดังนั้นการเลือกใช้โมเดลที่เหมาะสมจึงขึ้นอยู่กับลักษณะการใช้งาน หากต้องการความแม่นยำสูงสุดและมีทรัพยากรเพียงพอควรเลือก Llama-3.1-8B Fine-Tuned แต่หากต้องการนำไปใช้งานบนอุปกรณ์ขนาดเล็กที่มีทรัพยากรที่จำกัด Llama-3.2-3B Fine-Tuned เป็นตัวเลือกที่เหมาะสมกว่า และสำหรับข้อจำกัดด้านค่าของอัตรา Hallucination ของ Llama-3.2-3B Fine-Tuned ที่ยังสูงก็สามารถลดด้วยการผสมผสานเทคนิค Retrieval-Augmented Generation (RAG) เข้ากับโมเดลดังกล่าวที่ผ่านการ Fine-Tuning บน Edge Device เพื่อเพิ่มความแม่นยำของคำตอบในบริบทเฉพาะทางโดยไม่เพิ่มขนาดโมเดล ซึ่งจากผลการทดลองแสดงให้เห็นอย่างชัดเจนว่าสถาปัตยกรรม Hybrid (Fine-Tuning + RAG) ให้ผลลัพธ์ดีที่สุดในทุกมิติสะท้อนให้เห็นว่าการผสานทั้งสองเทคนิคสามารถชดเชยข้อจำกัดของแต่ละวิธีได้ กล่าวคือ Fine-Tuning ช่วยให้โมเดลเข้าใจบริบทและรูปแบบการตอบคำถามของ domain เฉพาะ ขณะที่ RAG ช่วยดึงข้อมูลที่ถูกต้องมาเสริมความแม่นยำและลด Hallucination ที่เป็นข้อจำกัดของ Fine-Tuning เพียงอย่างเดียวงานวิจัยในอนาคตควรทำการประเมินคุณภาพของระบบ Retrieval อย่างเป็นเชิงปริมาณเพิ่มเติม เช่น การวัดค่า Retrieval Precision และ Recall ของ Knowledge Base แยกออกจากการประเมินโมเดล เพื่อให้เข้าใจได้ว่าขั้นตอนใดเป็น bottleneck หลักของระบบ Hybrid และนำไปสู่การปรับปรุงที่ตรงจุดยิ่งขึ้น นอกจากนี้ควรขยายขนาดและความหลากหลายของชุดข้อมูล Fine-Tuning เพื่อเพิ่มความสามารถในการ Generalize ของโมเดลในบริบทการศึกษาภาษาไทย

บรรณานุกรม

[1] Raj, M., Hcltech Bengaluru, J., Warrier, H., Bengaluru, H., & Gupta, Y. (2024). "*Fine Tuning LLM for Enterprise: Practical Guidelines and Recommendations*". https://arxiv.org/abs/2404.10779v1

[2] J. M. Bink, "Personalized Response with Generative AI: Improving Customer Interaction with Zero-Shot Learning LLM Chatbots," Master's thesis, Dept. Math. Comput. Sci., Eindhoven Univ. Technol., Eindhoven, Netherlands, 2023. [Online]. Available: https://research.tue.nl/en/studentTheses/personalized-response-with-generative-ai/

[3] Wangsa, K.; Karim, S.; Gide, E.; Elkhodr, M. "*A Systematic Review and Comprehensive Analysis of Pioneering AI Chatbot Models from Education to Healthcare: ChatGPT, Bard, Llama, Ernie and Grok*". https://doi.org/10.3390/fi16070219

[4] Vidivelli, S., Ramachandran, M., & Dharunbalaji, A. (n.d.). *"Efficiency-Driven Custom Chatbot Development: Unleashing LangChain, RAG, and Performance-Optimized LLM Fusion".* https://doi.org/10.32604/cmc.2024.054360

[5] Jiao, Junfeng, et al. *"Navigating llm ethics: Advancements, challenges, and future directions." arXiv preprint arXiv:2406.18841* (2024).

[6] Biswas, Anjanava, and Wrick Talukdar. *"Guardrails for trust, safety, and ethical development and deployment of Large Language Models (LLM)."* Journal of Science & Technology 4.6 (2023): 55-82.