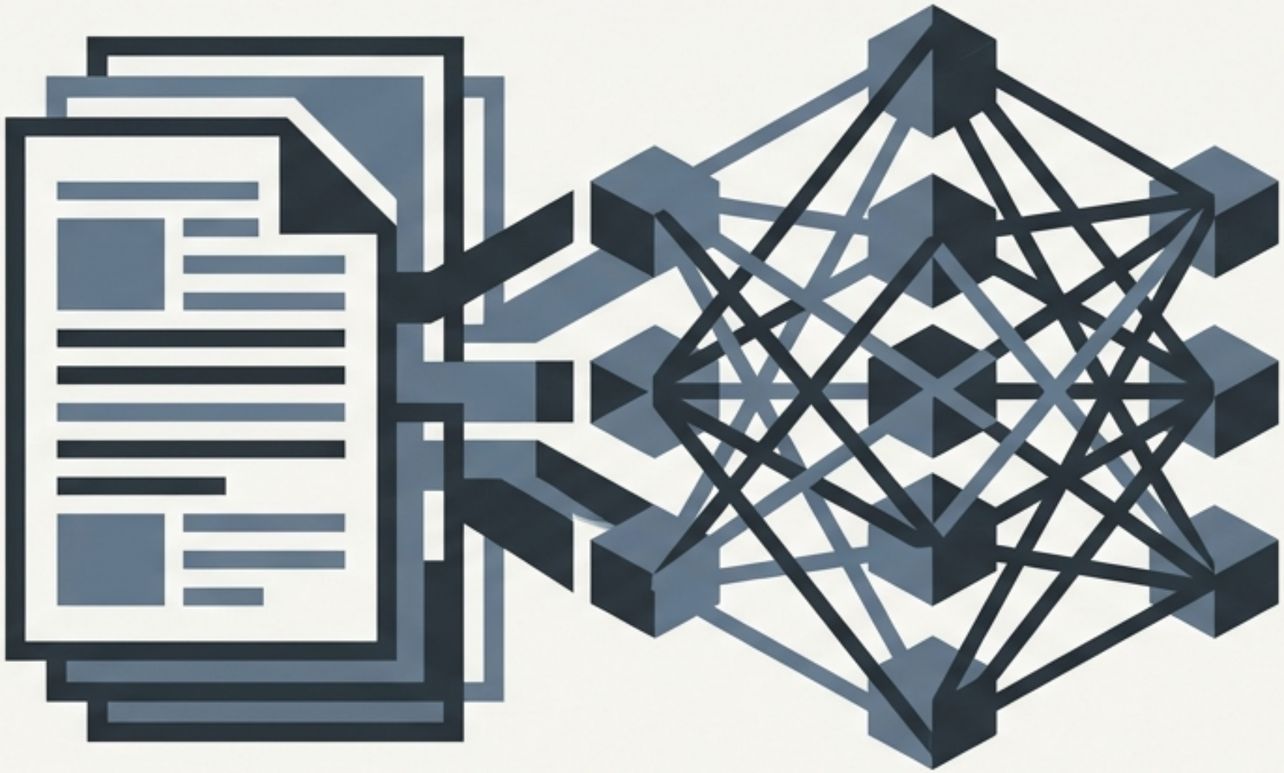


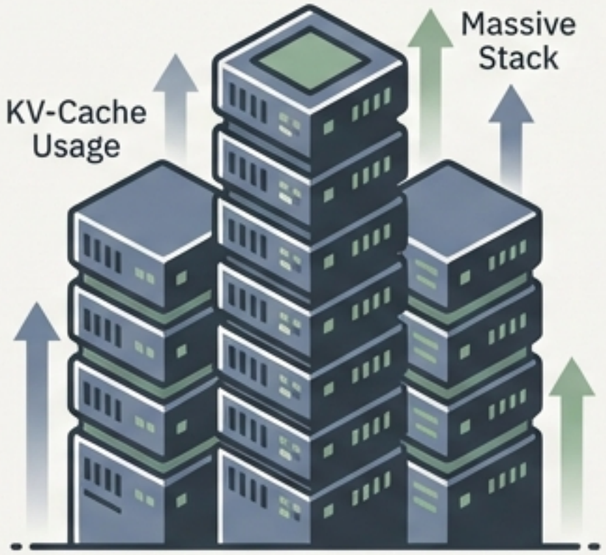
การผนวกความรู้เข้าสู่ LLMs ในพริบตาด้วย Doc-to-LoRA

ก้าวข้ามข้อจำกัดของ Context Window ด้วย Zero-Shot Context Distillation ภายใน Forward Pass เดียว

อ้างอิงจากงานวิจัย: "Doc-to-LoRA: Learning to Instantly Internalize Contexts"

NotebookLM

ข้อจำกัดของการปรับแต่ง LLMs ในปัจจุบัน



In-Context Learning (ICL):

- สะดวก แต่ใช้งานชั่วคราวและกินทรัพยากรมหาศาล (Transient & Memory-Intensive)
- Prompt ที่ยาวทำให้เกิดความหน่วงสูง และใช้หน่วยความจำ KV-cache เพิ่มขึ้นแบบทวีคูณ (Quadratic Attention)

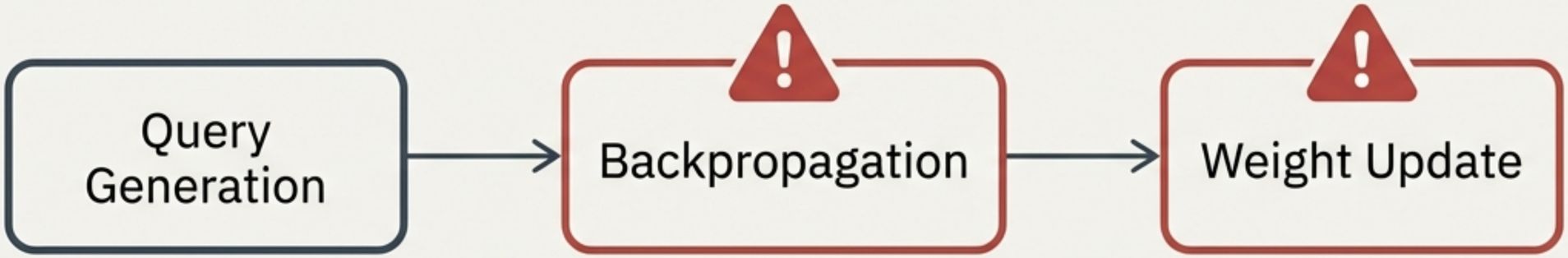
Supervised Finetuning (SFT):

- ถาวร แต่ช้าและมีค่าใช้จ่ายสูง (Slow & Expensive)
- ต้องสร้าง Dataset เฉพาะทาง เสี่ยงต่อการเกิด Overfitting และไม่สามารถปรับเปลี่ยนแบบ Real-time ตามความต้องการของผู้ใช้ได้

NotebookLM

ความหวังลวง: Context Distillation (CD) มีประสิทธิภาพสูง แต่ใช้งานจริงแบบ Real-time ไม่ได้

แนวคิดของ CD: นำข้อมูลบริบท (Context) แปลงเป็น Parameter ของโมเดล เพื่อลดต้นทุนขณะ Inference



ปัญหาที่พบ (The Bottleneck):

- **ช้ามาก** (Agonizingly Slow): ต้องใช้เวลาหลายนาทีหรือหลายชั่วโมงในการ Train ต่อหนึ่ง Prompt
- **ใช้หน่วยความจำมหาศาล (Memory-Heavy):** การทำ Backpropagation บน Queries จำนวนมาก กินทรัพยากร VRAM มหาศาล
- **สรุป:** ไม่สามารถนำไปใช้กับแอปพลิเคชั่นที่ต้องการโต้ตอบแบบทันที (Interactive) หรือทำงานบนอุปกรณ์ขนาดเล็ก (On-device) ได้

NotebookLM

ทางออก: Doc-to-LoRA (D2L) เรียนรู้กระบวนการ Distillation ในพริบตา

D2L คือ Hypernetwork ขนาดเล็กที่ทำ Meta-learning กระบวนการ CD ให้เสร็จสิ้นภายใน Forward Pass เดียว ไม่ต้องทำ Backpropagation ที่กินเวลา

< 1 วินาที

Sub-second Update Latency

< 50 MB

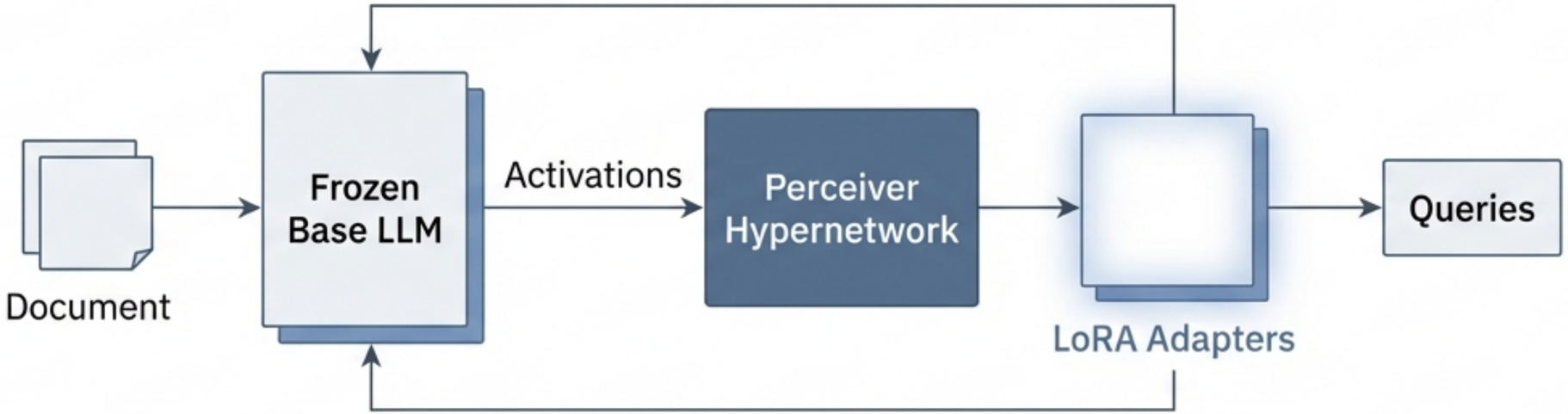
Additional Inference Memory

4x

ขยายขีดจำกัด Context Window ได้ถึง 4 เท่าโดยไ ไม่เสียความแม่นยำ

NotebookLM

สถาปัตยกรรม: การสร้าง LoRA Adapter แบบ Real-time ด้วย Perceiver

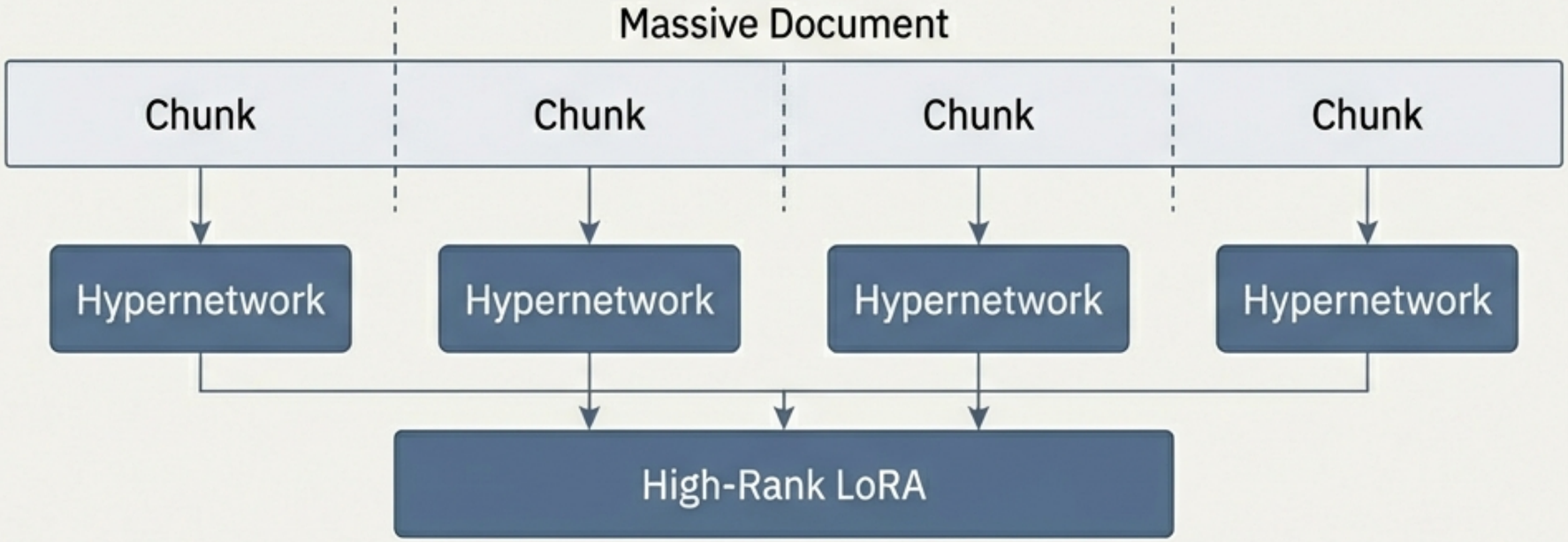


1. **Extract:** ดึง Token Activations จาก Frozen Base LLM โดยไม่ต้องแก้ไขโมเดลหลัก
2. **Encode:** ส่งผ่าน Perceiver-style Cross-Attention Module เพื่อรองรับข้อมูลที่มีความยาวไม่จำกัด
3. **Adapt:** แปลงข้อมูลให้กลายเป็น Low-Rank LoRA Matrices ส่งตรงไปยังแต่ละ Layer ของ LLM ทันที ทำให้โมเดลตอบคำถามได้โดยไม่ต้องอ่าน Context ซ้ำ

NotebookLM

กลไก Chunking: ทลายขีดจำกัดความยาวของ Context Window

LLM พื้นฐานมีข้อจำกัดด้านความยาว (เช่น 8K Tokens) แต่ D2L ข้ามข้อจำกัดนี้ด้วย Dynamic Chunking

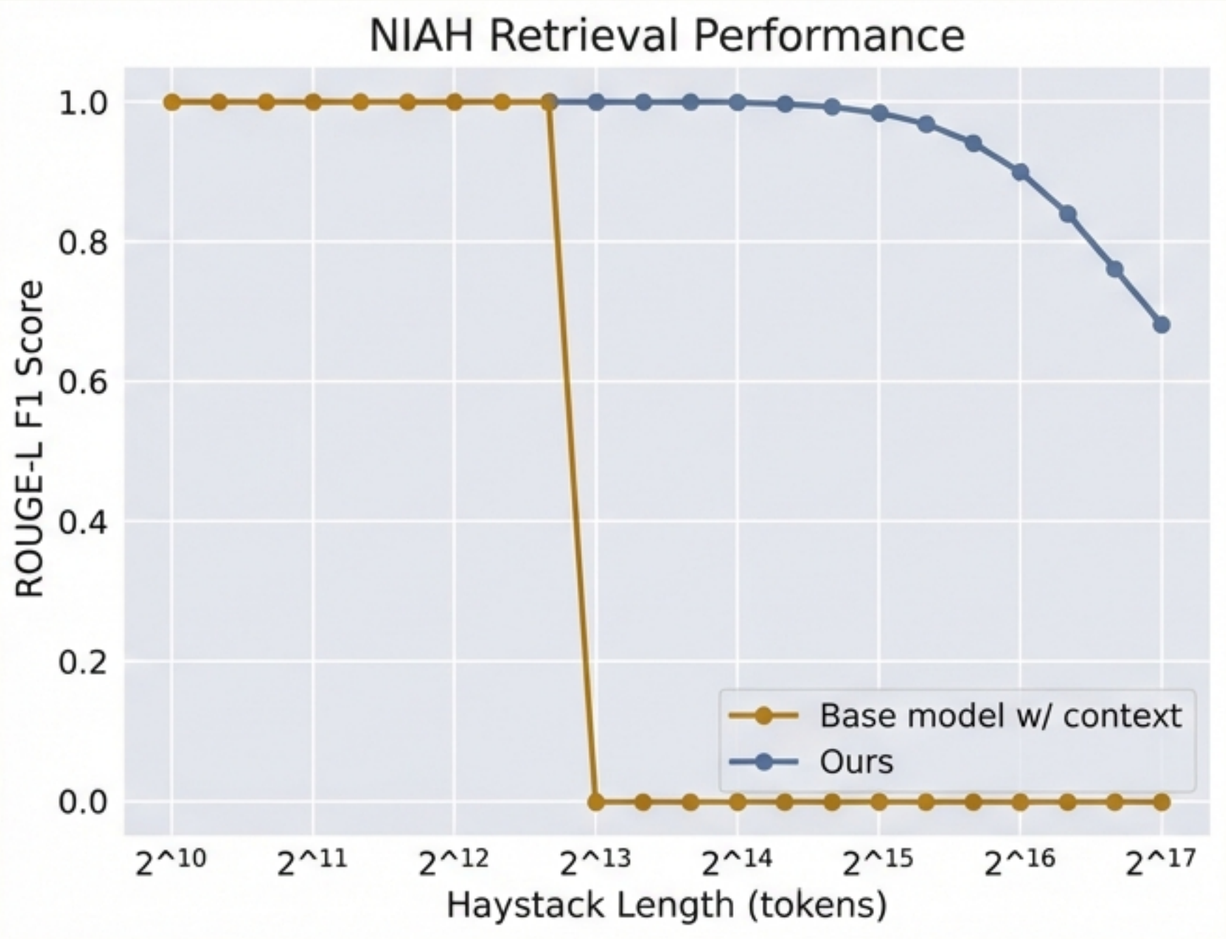


- **วิธีทำงาน**: แบ่งเอกสารขนาดยาวออกเป็นส่วนๆ (Chunks) ประมวลผลแต่ละส่วนแยกกันผ่าน Hypernetwork
- **ผลลัพธ์**: นำ Adapter ของแต่ละ Chunk มาเรียงต่อกัน (Concatenate) กลายเป็น LoRA ที่มี Rank สูงขึ้น สามารถเก็บข้อมูลเอกสารขนาดใหญ่ได้โดยไม่ต้องเปลี่ยนรูปร่าง (Output Shape) ของ Hypernetwork

NotebookLM

บทพิสูจน์ที่ 1: ดึงข้อมูลแม่นยำ 100% แม้ความยาวเกินข้อจำกัดโมเดลถึง 4 เท่า

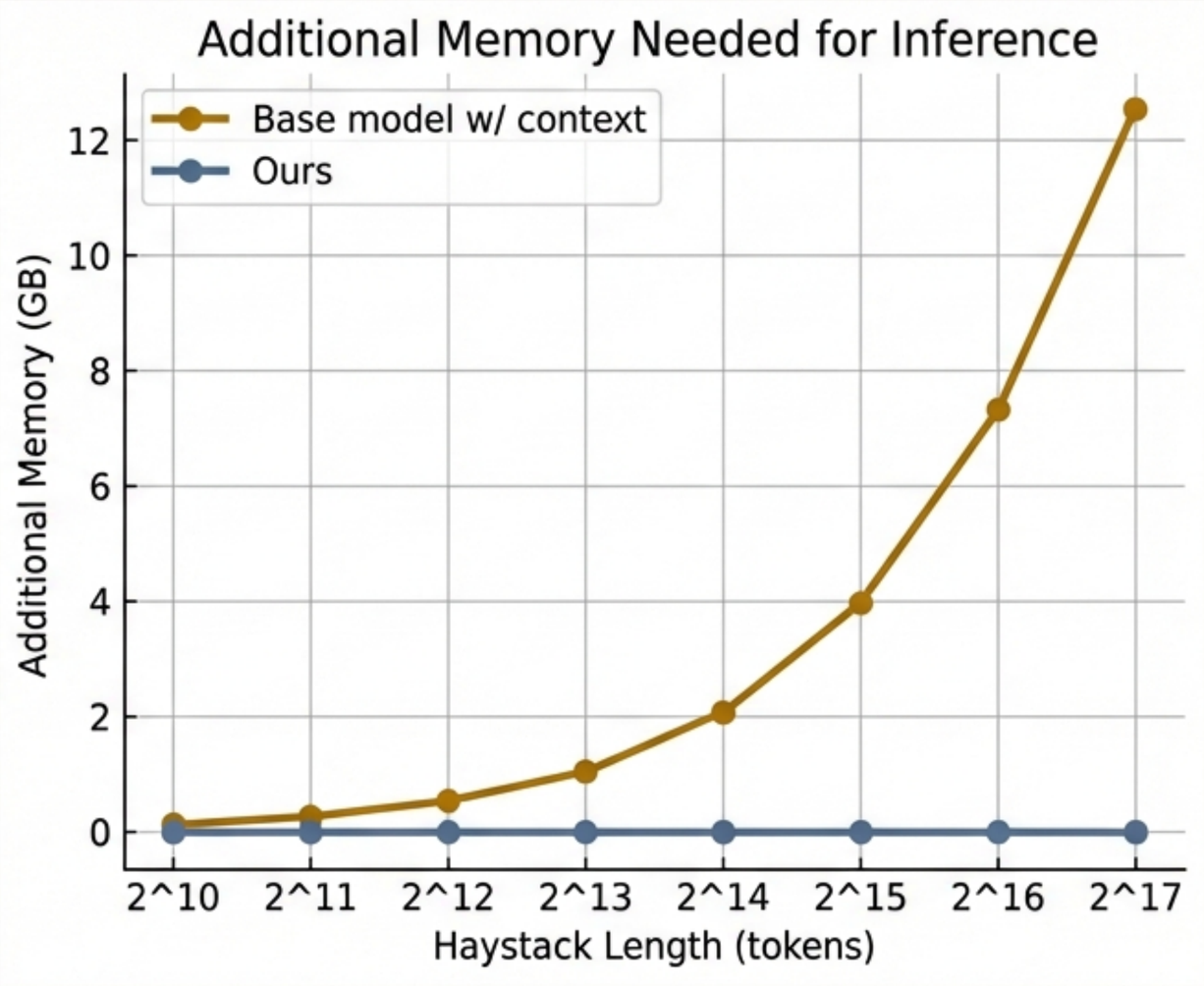
- การทดสอบ Needle-in-a-Haystack (NIAH): ซ่อนข้อความสั้นๆ ในเอกสารขนาดยักษ์
- Base Model (Gemma-2-2b-it) **สูญเสียควา**มสามารถทันทีเมื่อความยาวเกิน 8K Tokens
- Doc-to-LoRA: รักษาความแม่นยำในระดับสมบูรณ์แบบได้จนถึง 40,000 Tokens (40 Chunks) แม้ตอน Train จะเคยเห็นความยาวสูงสุดแค่ 256 Tokens เท่านั้น (Zero-shot Generalization)



NotebookLM

บทพิสูจน์ที่ 2: ปาฏิหาริย์แห่งการลดหน่วยความจำ (The Memory Miracle)

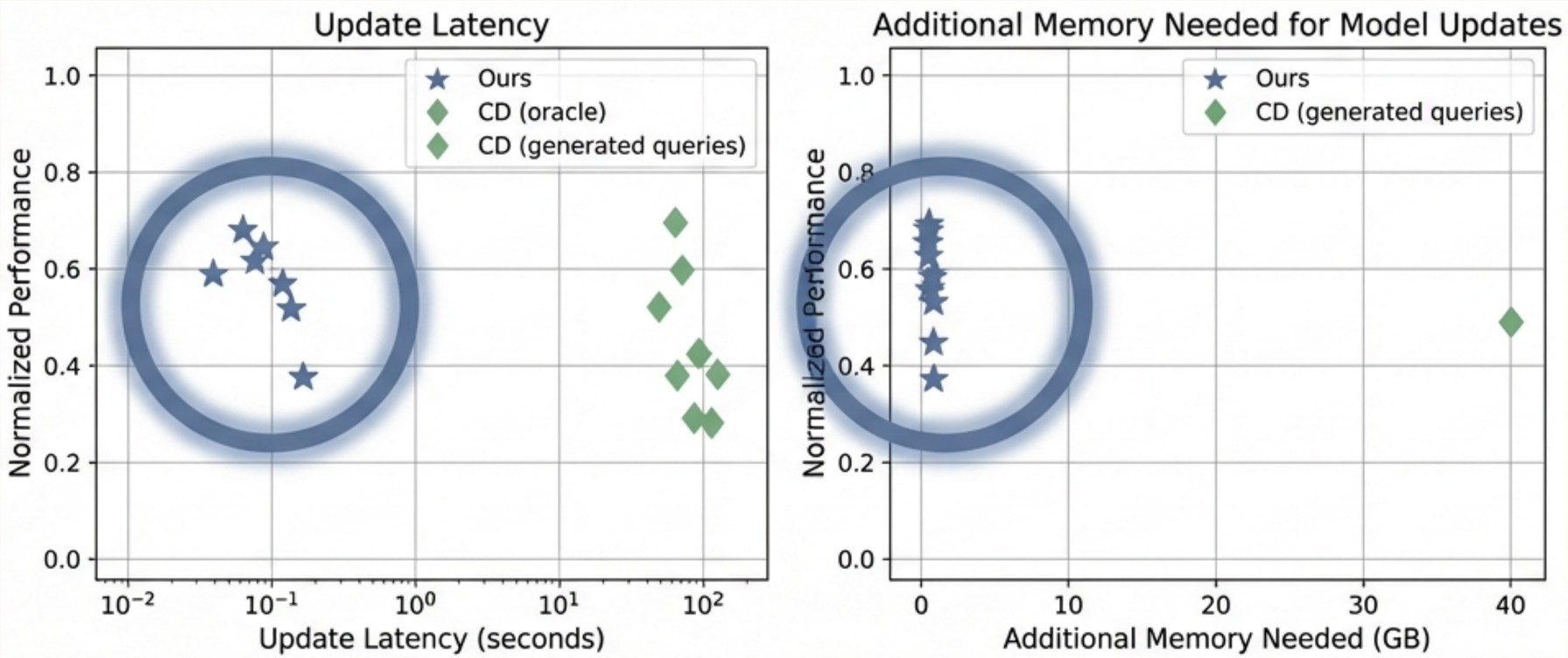
- ปัญหาหลักของ ICL คือการใช้ KV-cache ที่กินหน่วยความจำมหาศาลเมื่อเจอกับเอกสารยาว
- **Base Model**: ใช้ VRAM เพิ่มเติมมากกว่า 12 GB สำหรับการสร้างคำตอบจากเอกสาร 128K Tokens
- **Doc-to-LoRA**: ใช้ VRAM เพิ่มเติม **น้อยกว่า 50 MB** คงที่เสมอ ไม่ว่าเอกสารจะยาวแค่ไหน เพราะข้อมูลถูกฝังลงใน Weight ของ LoRA เรียบร้อยแล้ว



NotebookLM

บทพิสูจน์ที่ 3: อัปเดตความรู้จริงเร็วกว่า CD แบบดั้งเดิมเป็นร้อยเท่า

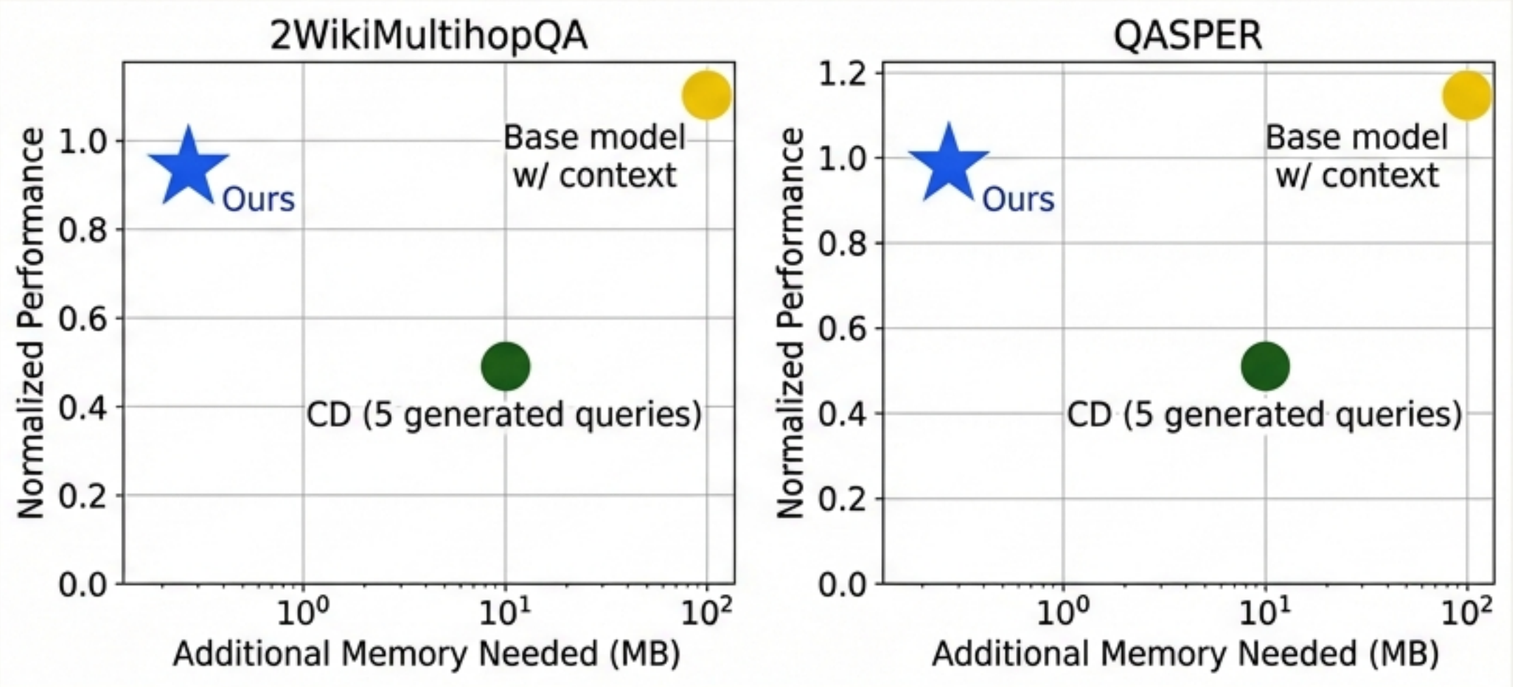
- **Update Latency:** D2L ใช้เวลาประมวลผลและสร้าง LoRA ไม่ถึง 1 วินาที (Sub-second) ในขณะที่ CD แบบดั้งเดิมใช้เวลาตั้งแต่ 40 วินาที ไปจนถึงเกิน 100 วินาทีต่อ 1 Prompt
- **Memory for Updates:** D2L ใช้ VRAM ระหว่างอัปเดตเพียงเล็กน้อย (< 2GB) เทียบกับ CD ที่ต้องใช้ VRAM ถึง 40GB+ สำหรับการรัน Backpropagation บนชุดคำถาม



NotebookLM

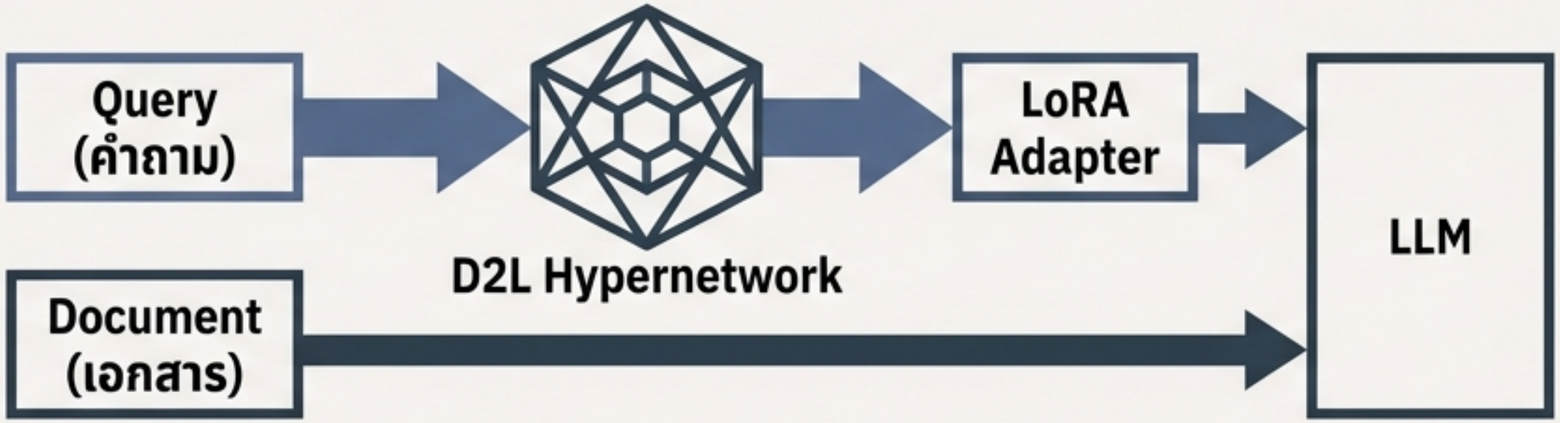
บทพิสูจน์ที่ 4: ประสิทธิภาพสูงบนชุดข้อมูลจริงขนาดใหญ่

- ทดสอบบนเอกสารจริงความยาวสูงสุดถึง 32K Tokens (2WikiMultihopQA, MultiFieldQA, QASPER)
- Zero-Shot Generalization: D2L สามารถเข้าใจและทำคะแนนได้สูเว่า Baseline ของ CD ที่ต้องสร้างคำถามขึ้นมาเอง (Generated-query CD)
- ทำได้โดยใช้พลังการประมวลผลและหน่วยความจำน้อยกว่าหลายเท่าตัว



NotebookLM

ขีดสุดของการ Generalize: การฝัง "คำถาม" แทน "เอกสาร"



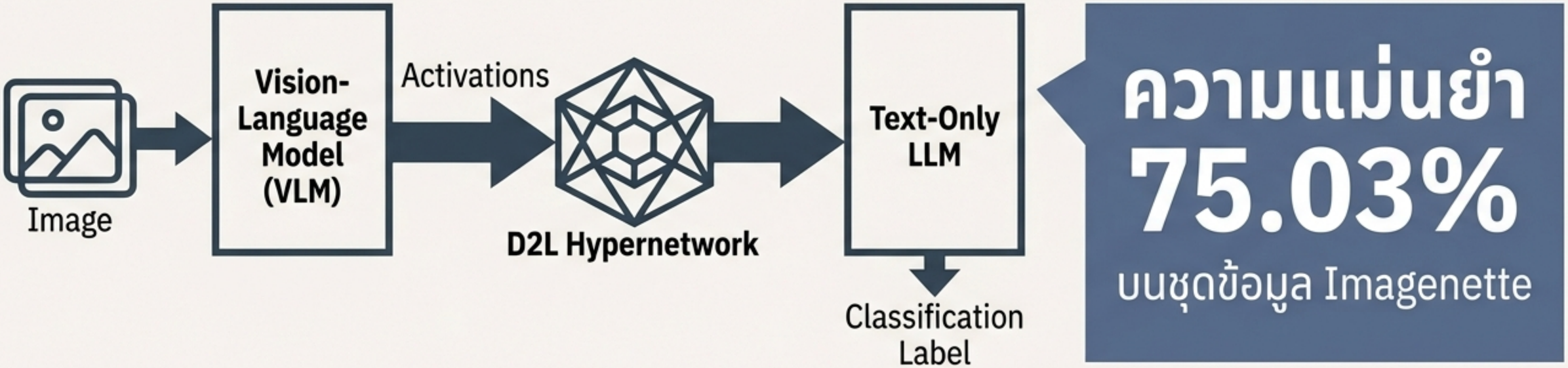
Recall: 0.587 vs 0.185

(D2L vs โมเดลปกติที่ไม่มี Context)

- The Swapped Test: จะเกิดอะไรขึ้นถ้าเราใช้ D2L เพื่อแปลง "คำถาม" (Query) ให้กลายเป็น LoRA แทนที่จะเป็นเอกสาร?
- แม้ D2L จะถูก Train มาเพื่อจัดการกับข้อมูลความรู้ (Documents) แต่ผลลัพธ์พบว่าโมเดลได้เรียนรู้กระบวนการผนวกข้อมูลแบบครอบจักรวาล
- ผลลัพธ์: D2L สามารถดึงข้อมูลได้ด้วยค่า Recall 0.587 (เทียบกับ 0.185 ของโมเดลปกติที่ไม่มี Context) พิสูจน์ความยืดหยุ่นขั้นสุดของสถาปัตยกรรมนี้

NotebookLM

ความสามารถเหนือความคาดหมาย: Zero-Shot Vision Transfer

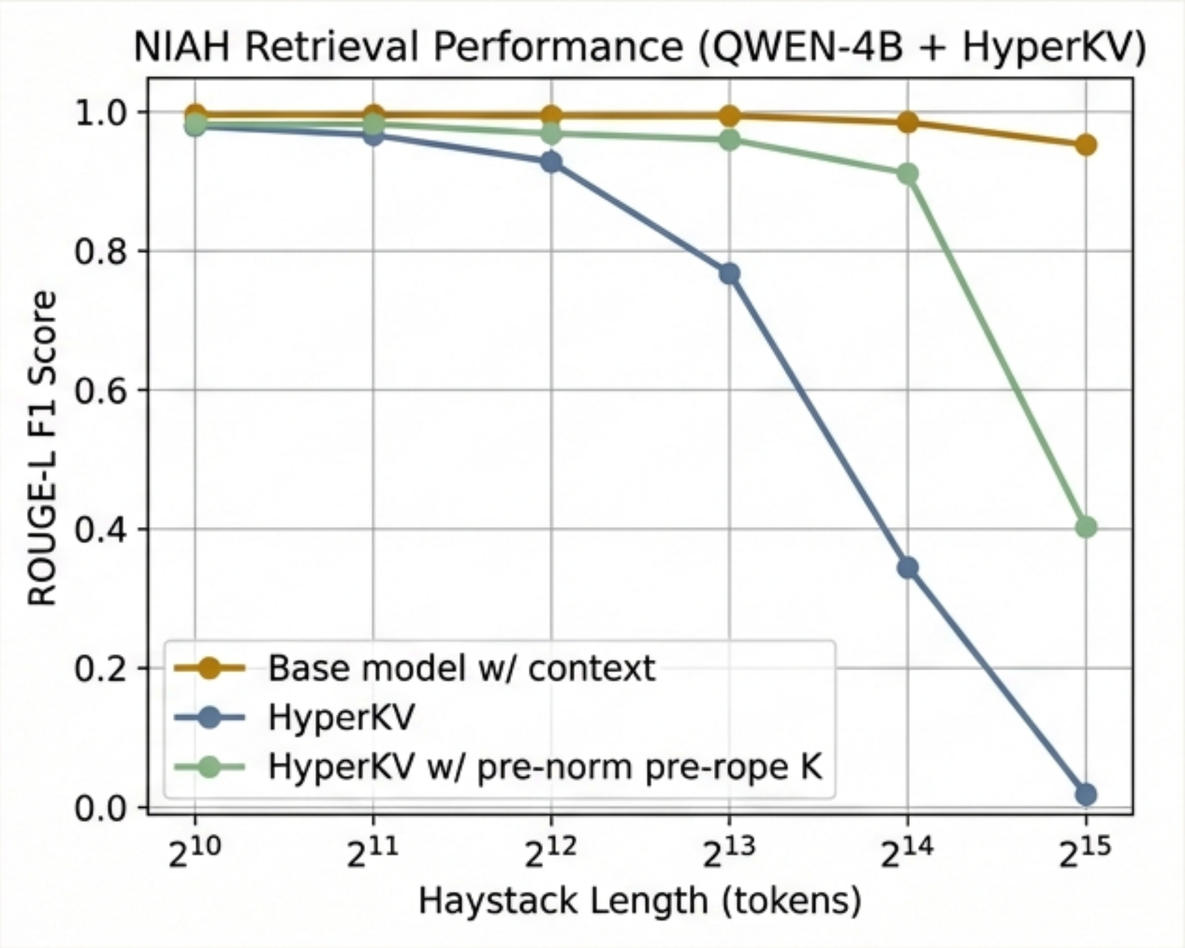


- เชื่อมต่อ Vision-Language Model (VLM - Gemma-3-4b) เข้ากับโมเดลที่อ่าน Text ได้อย่างเดียว (Gemma-2-2b) ผ่าน D2L
- D2L แปลง Activations ของภาพจาก VLM ให้กลายเป็น LoRA ของโมเดล Text
- ผลลัพธ์ระดับทลายกรอบ: โมเดล Text สามารถแยกแยะภาพ (Image Classification บนชุดข้อมูล Imagenette) ได้ความแม่นยำถึง 75.03% โดยที่ D2L ไม่เคยเห็นรูปภาพใดๆ ในขั้นตอนการ Train เลย

NotebookLM

ความยืดหยุ่นของโครงสร้าง: สร้าง KV-Cache ได้ทันที

- โครงสร้าง Perceiver ของ D2L ไม่ได้ถูกจำกัดแค่การสร้าง LoRA
- การประยุกต์ใช้: สามารถตั้งค่าให้ Hypernetwork สร้าง Compressed KV-cache (Prefix-tuning) แบบทันทีได้
- จากการทดสอบกับ Qwen3-4B โมเดลยังคงรักษาความสามารถในการดึงข้อมูลแบบ Needle-in-a-Haystack ได้อย่างยอดเยี่ยม เป็นเครื่องพิสูจน์ความอเนกประสงค์ของวิธีนี้

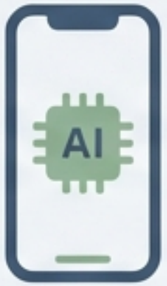


NotebookLM

ปลดล็อกขีดความสามารถ AI ในโลกความเป็นจริง

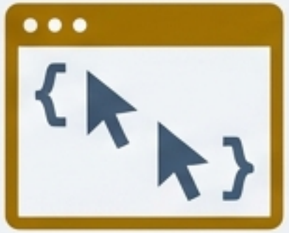
Personalized Chat Behavior

สามารถอัปเดตความชอบและข้อมูลส่วนตัวของผู้ใช้ (User Preferences) ลงในโมเดลได้ทันที โดยไม่ต้องกินพื้นที่ KV-cache ทำให้บทสนทนายาวขึ้นและลื่นไหลขึ้น



Edge-Device LLMs

ตอบโจทย์อุปกรณ์ที่มี VRAM จำกัด (มือถือ/IoT) เพราะสามารถซึมซับเอกสารยาวๆ ได้ด้วยหน่วยความจำชั่วคราวเพียงไม่ถึง 50 MB



Instant Knowledge Updates

อัปเดตข้อมูลแบบ Real-time เหมาะสำหรับระบบ Agent หรือ Active Codebases ที่ข้อมูลเปลี่ยนแปลงทุกวินาที

NotebookLM

ก้าวใหม่ของวิธีการประมวลผล Context ใน Large Language Models

Doc-to-LoRA ผสมผสานความสะดวกของ In-Context Learning
เข้ากับประสิทธิภาพของ Context Distillation

>4x Context Extension	<1 วินาที Update Latency	<50MB Inference Memory

รันโค้ดและทดสอบด้วยตัวคุณเองที่:
github.com/SakanaAI/doc-to-lora

NotebookLM