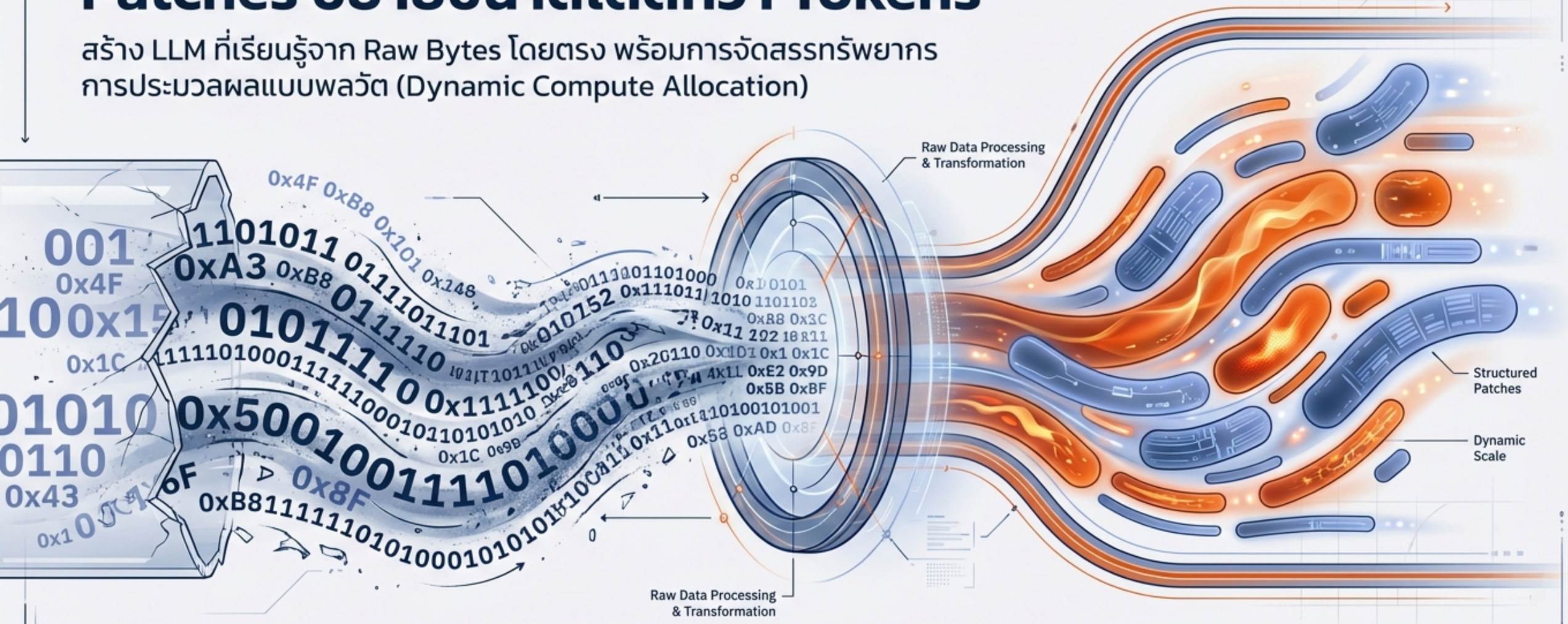
Byte Latent Transformer (BLT): Patches ขยายขนาดได้ดีกว่า Tokens

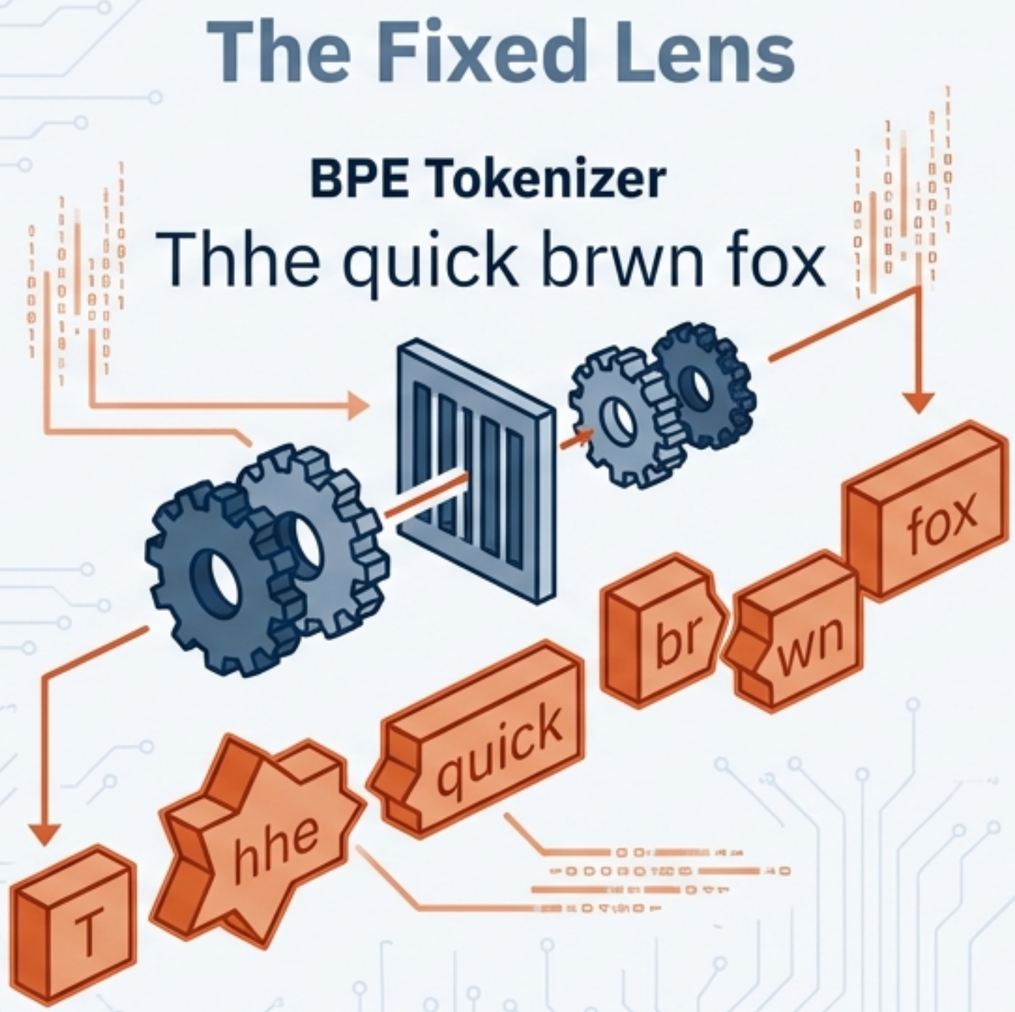
สร้าง LLM ที่เรียนรู้จาก Raw Bytes โดยตรง พร้อมการจัดสรรทรัพยากรการประมวลผลแบบพลวัต (Dynamic Compute Allocation)



Based on: Byte Latent Transformer: Patches Scale Better Than Tokens (Meta FAIR, 2024)

NotebookLM

คอขวดของ Tokenization: ปัญหาของ "Static Compute"



Fixed Lens (มุมมองที่ตายตัว):
Tokenization คือการประมวลผลล่วงหน้าแบบ Heuristic ไม่ใช่การเรียนรู้ที่แท้จริง ทำให้โมเดลมีอคติ

Brittleness (ความเปราะบาง):
อ่อนไหวต่อสัญญาณรบกวน (Noise) และคำผิด (Typos)

Inequity (ความไม่เท่าเทียม):
ประสิทธิภาพลดลงในภาษาที่ 'Out of Vocabulary' หรือภาษาที่มีทรัพยากรน้อย

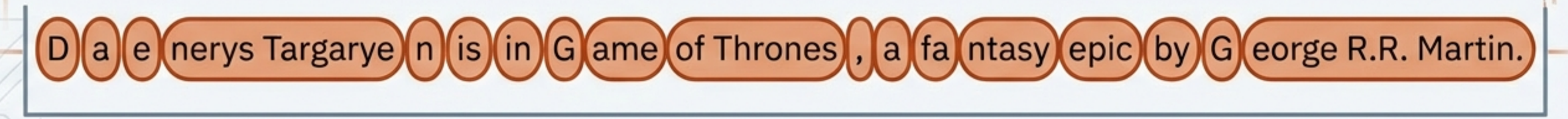
Wasted Compute:
ใช้พลังในการประมวลผลเท่ากันสำหรับคำง่ายๆ (เช่น 'the') และคำยากๆ

NotebookLM

การเปลี่ยนผ่านสู่ยุคใหม่: จาก Static Tokens สู่ Dynamic Patches

Old Paradigm: Fixed Compute (BPE)

New Paradigm: Dynamic Compute (BLT)

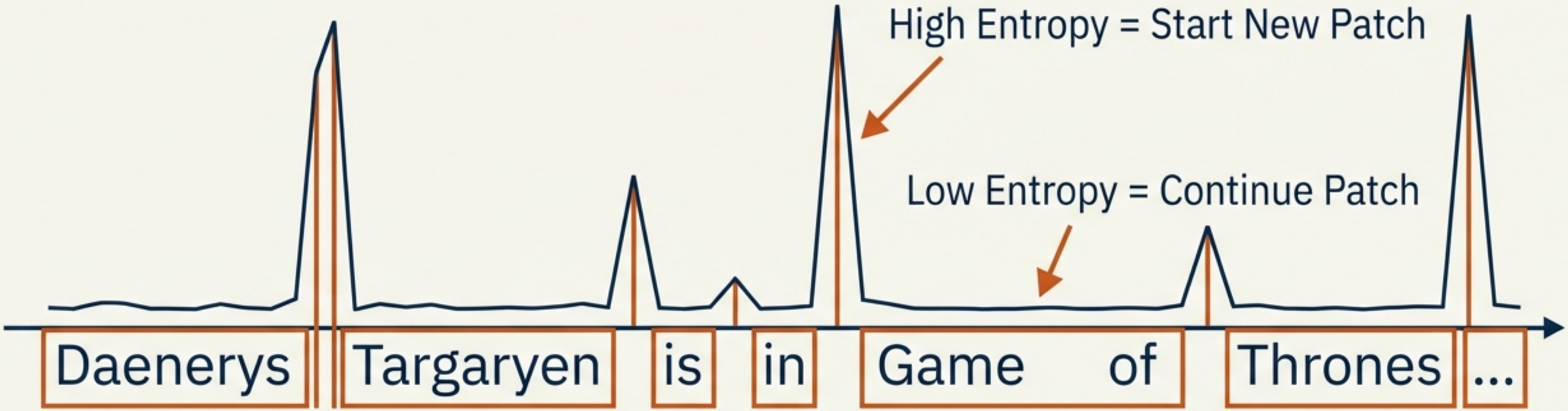


BLT Concept:
โมเดลจะตัดสินใจเองว่าจะจัดกลุ่มข้อมูลอย่างไรตามความซับซ้อนของเนื้อหา

Benefit:
'Compute where it counts'
(ประมวลผลเน้นหนักเฉพาะจุดที่สำคัญ)

NotebookLM

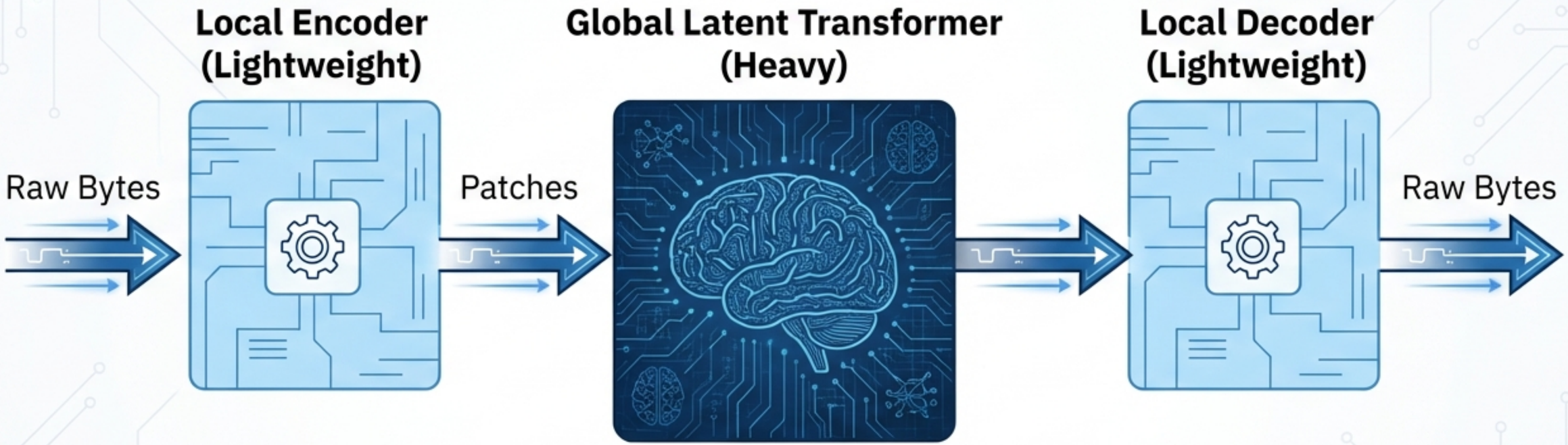
กลไกหลัก: การสร้าง Patch ด้วยค่า Entropy



- **Entropy ($H(x_i)$):** มาตรวัดความ ‘ประหลาดใจ’ ของข้อมูลถัดไป
- **Mechanism:** เมื่อโมเดลเจอสิ่งที่คาดเดายาก (Entropy สูง) มันจะสร้าง Patch ใหม่ทันที เพื่อจัดสรรทรัพยากรมาวิเคราะห์จุดนั้น

NotebookLM

สถาปัตยกรรม ‘The Sandwich Model’

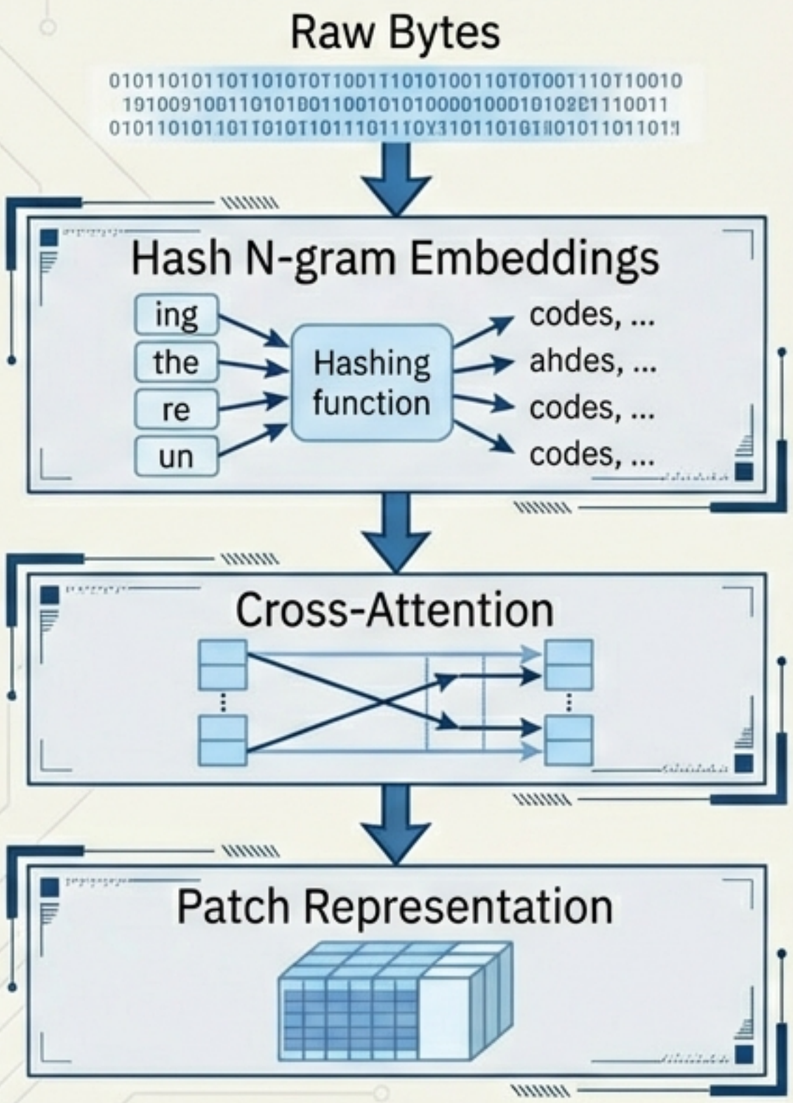


‘สมอง’ หลักที่ประมวลผล Patches
(ทำงานหนักที่สุด แต่รันจำนวนครั้งน้อยกว่า)

Insight: โมเดล Global Transformer ขนาดใหญ่ไม่ต้องรันทุก Byte แต่รันเฉพาะระดับ Patch ทำให้ประหยัดพลังงานมหาศาล

NotebookLM

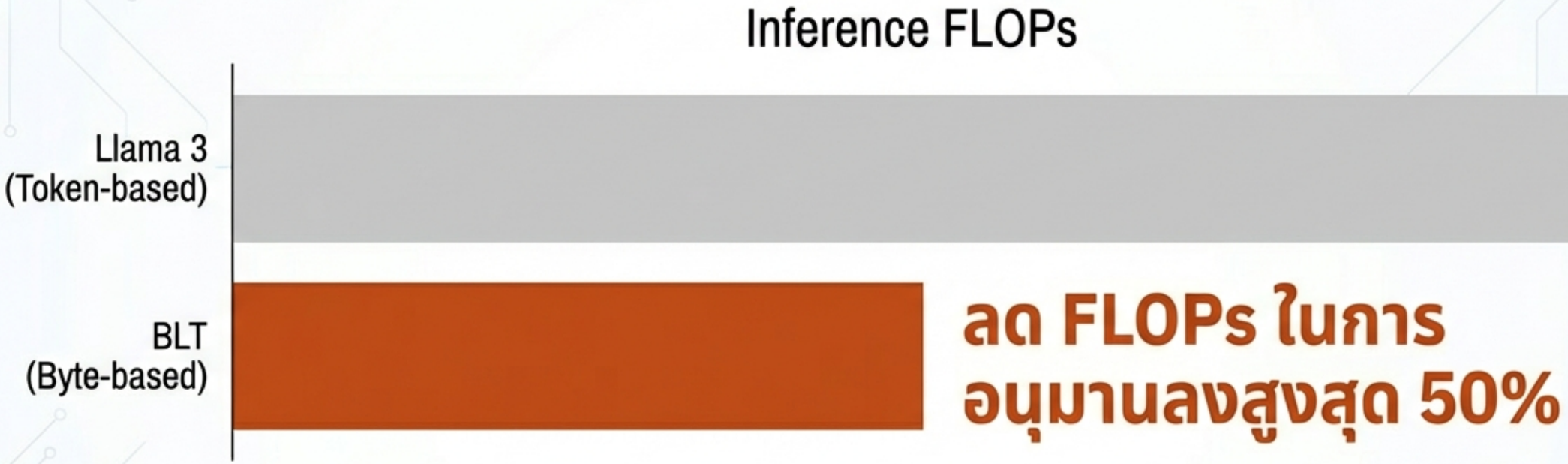
Zoom In: Local Encoder (จาก Bytes สู่ Meaning)



- **No Vocabulary:**
  ไม่มีการเก็บคลังคำศัพท์ตายตัว
- **Hash N-gram Embeddings:**
  เทคนิคการจับบริบทระยะใกล้ (เช่น ‘ing’, ‘the’) โดยไม่ต้องใช้ Tokenizer
- **Result:**
  สร้าง Representation ที่แข็งแกร่งและยืดหยุ่นกว่า Token

NotebookLM

ประสิทธิภาพ: ทำได้มากกว่า ในทรัพยากรที่น้อยกว่า

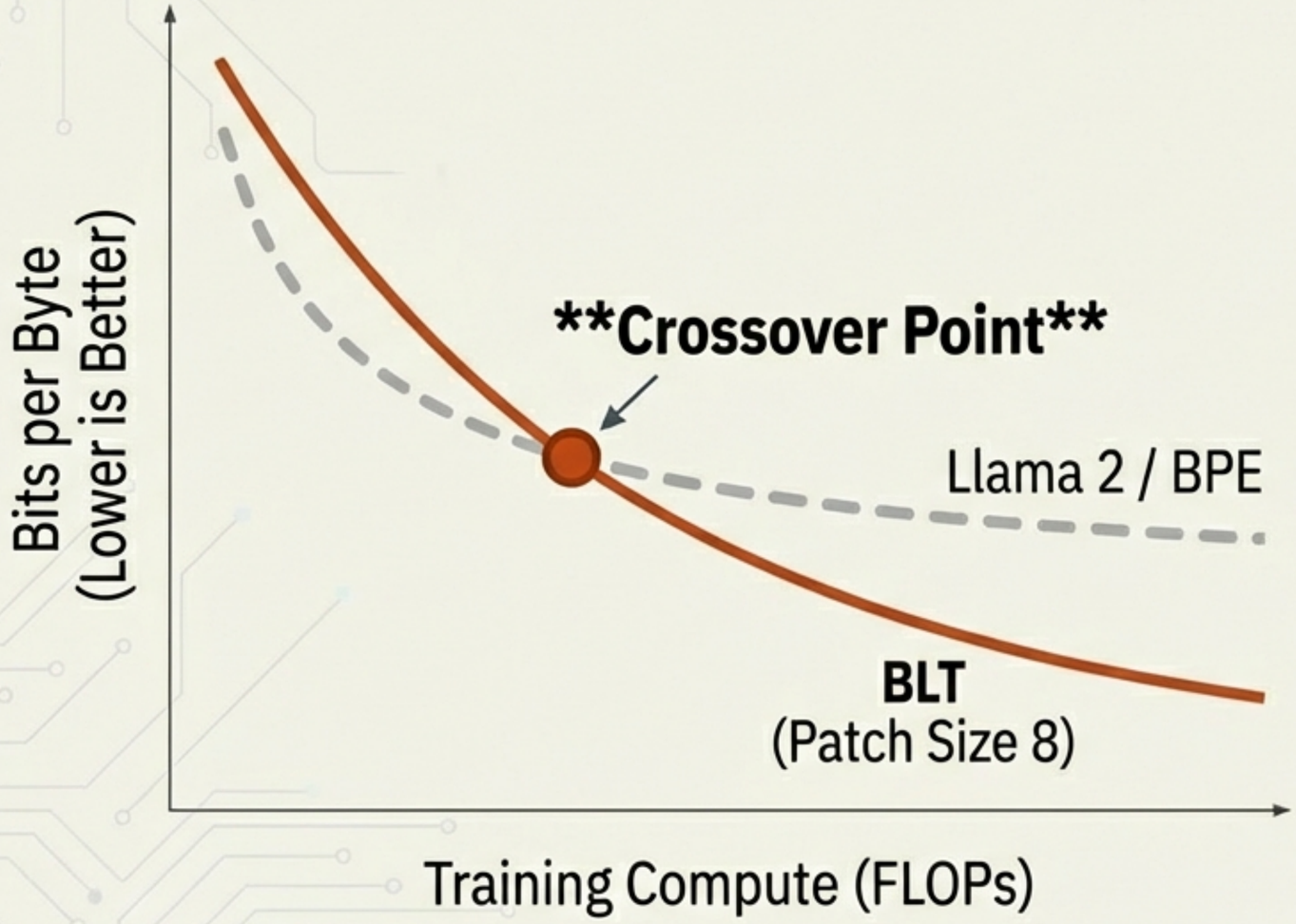


เนื่องจาก BLT สร้าง Patch ขนาดยาวสำหรับข้อความที่ง่าย (Predictable)
ทำให้ลดจำนวนครั้งที่ Global Transformer ต้องทำงาน

Key Takeaway: การประหยัดทรัพยากรนี้ทำให้เราสามารถ**ขยาย**
ขนาดโมเดลให้ใหญ่ขึ้น หรือทำงานได้เร็วขึ้นในงบประมาณเดิม

NotebookLM

กฎการขยายขนาด (Scaling Laws): เมื่อ Patch ชนะ Token



- **BLT** ให้ประสิทธิภาพเทียบเท่า Llama 3 ที่ระดับ 8B parameters
- **New Scaling Axis:** เราสามารถเพิ่มขนาด Patch (Patch Size) เพื่อลดต้นทุนโดยไม่เสียประสิทธิภาพ ซึ่งเป็นสิ่งที่ Token-based models ทำไม่ได้

NotebookLM

ความทนทาน (Robustness): จุดแข็งของระดับ Byte

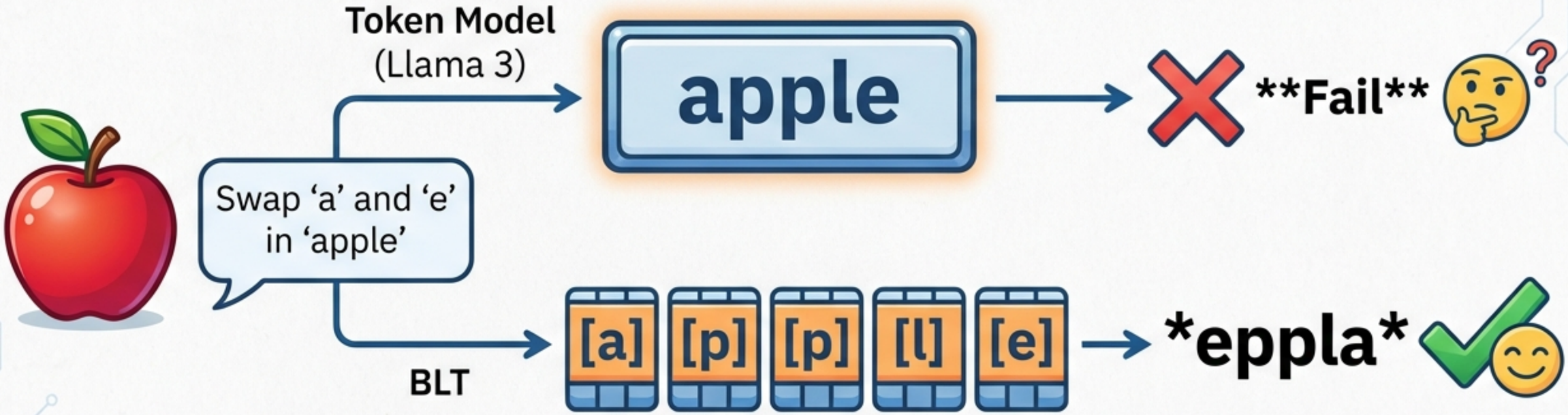
Stress Test

Test Type	Input Example	Performance Impact
Normal	The capital of France is Paris.	Llama 3: High ✅ | BLT: High ✅
UPPERCASE Noise	THE CAPITAL OF FRANCE IS PARIS.	Llama 3: **Crash (Low Accuracy)** ❌ | BLT: **Stable** ✅
Random Chars	T.h.e c.a.p.i.t.a.l o.f...	Llama 3: **Crash** ❌ | BLT: **Stable** ✅
Typos	The capitel of Franse is...	Llama 3: **Confused** ❌ | BLT: **Robust** ✅

✅ BLT ทำคะแนนนำ Llama 3 ถึง 8 จุดในแบบทดสอบที่มี Noise

NotebookLM

ความเข้าใจระดับ Sub-Word (ผลลัพธ์จาก CUTE Benchmark)



Tokenizers จำคำเป็นก้อนๆ แต่ BLT เข้าใจการสะกดคำ

Result: ชนะขาดลอยในงานด้าน Orthography และ Character Manipulation

NotebookLM

ไร้กำแพงภาษา: รองรับ Multilingual และ Long-Tail Data

Generalization: BLT ทำงานได้ดีกว่าในภาษาที่ Tokenizer ไม่ได้ถูกปรับจูนมาโดยเฉพาะ (เช่น ภาษาไทย)

NotebookLM

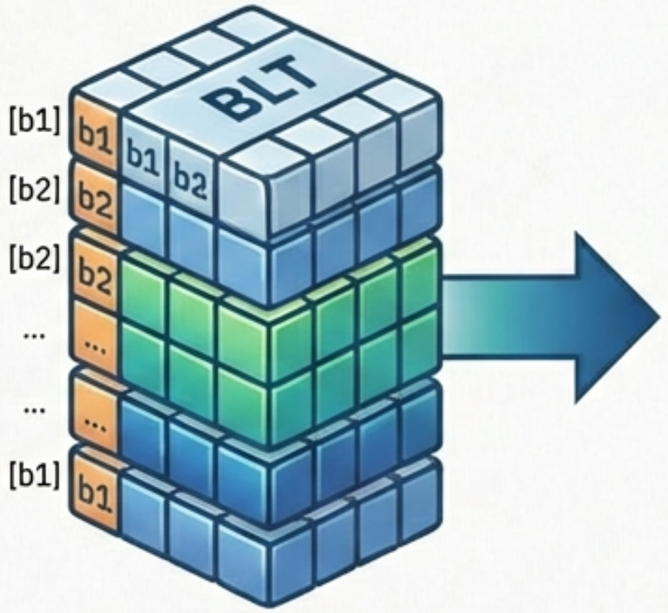
“Byte-ifying”: การอัปเกรด Llama 3 ให้เป็น BLT

Llama 3 Weights

Transfer Learning

การเรียนรู้แบบถ่ายโอน

BLT Global Transformer



Performance Boost

ประสิทธิภาพเพิ่มขึ้น

เราไม่จำเป็นต้องเทรนใหม่จากศูนย์ (Scratch) เสมอไป

การนำ Weight ของ **Llama 3** มาใช้ใน BLT ช่วยเร่งความเร็วในการเทรนและเพิ่มประสิทธิภาพได้ทันที

NotebookLM

ตัวอย่างการใช้งานจริง: MMLU Reasoning

The following are multiple choice questions (with answers) about college physics.

Single Patch / 1 Compute Step

A refracting t telescope consists of two c o n n converging lenses separated by 10 0 cm. The eyepiece lens has a focal length of 20 cm. The angular magnification of the telescope is
A. 4
B. 5
D. 2 0
Answer: A
...
The muon d e cays with a c characteristic l i fetime of about 1 0^ - 6 s econd into an e lectron, a m uon ne nutrino, and an electron a n ti neutrino. The muon is forbidden from decaying into an electron and just a s ingle neutrino by the la w of co onservation of
A. c harge
B. m ass
C. energy a nd momentum
D. l epton number
Answer: D
The quantum e efficiency of a p photon detector is 0. 1 . If 1 00 p hotons are s ent into the detector, one a fter the other, the detector will detect photons
A. a n a verage of 1 0 times, with an rms de viation of about 4
B. an average of 10 times, with an rms deviation of about 3
C. an average of 10 times, with an rms deviation of about 1
D. an average of 10 times, with an rms deviation of about 0. 1
Answer: A

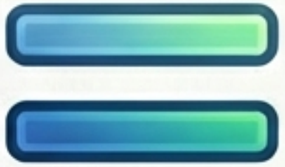
Many Patches / High Compute

Smart Skipping: โมเดลเรียนรู้ที่จะ 'อ่านข้าม' ส่วนที่ซ้ำซ้อนในคำถามปรนัยโดยอัตโนมัติ ทำให้ประหยัดพลังงานสูงสุด

NotebookLM

บทสรุป: ทำไม BLT ถึงคืออนาคต

Efficiency
ลด Inference FLOPs ลง 50%



Performance
ประสิทธิภาพเทียบเท่า Llama 3 ที่ระดับ Scale เดียวกัน

Robustness
ทนทานต่อ Noise และเข้าใจโครงสร้างคำ (Spelling)

Simplicity
ตัดความซับซ้อนของการจัดการ Vocabulary ทิ้งไป

BLT ไม่ใช่แค่ทางเลือก แต่คือวิวัฒนาการถัดไปของสถาปัตยกรรม LLM

NotebookLM

อนาคตที่ไร้ Token (Token-Free Future)

ยุคสมัยของ Tokenizer กำลังจะสิ้นสุดลง ยุคของ Byte เริ่มต้นขึ้นแล้ว

Code & Paper: https://github.com/facebookresearch/blt

Based on 'Byte Latent Transformer: Patches Scale Better Than Tokens' (2024).

NotebookLM