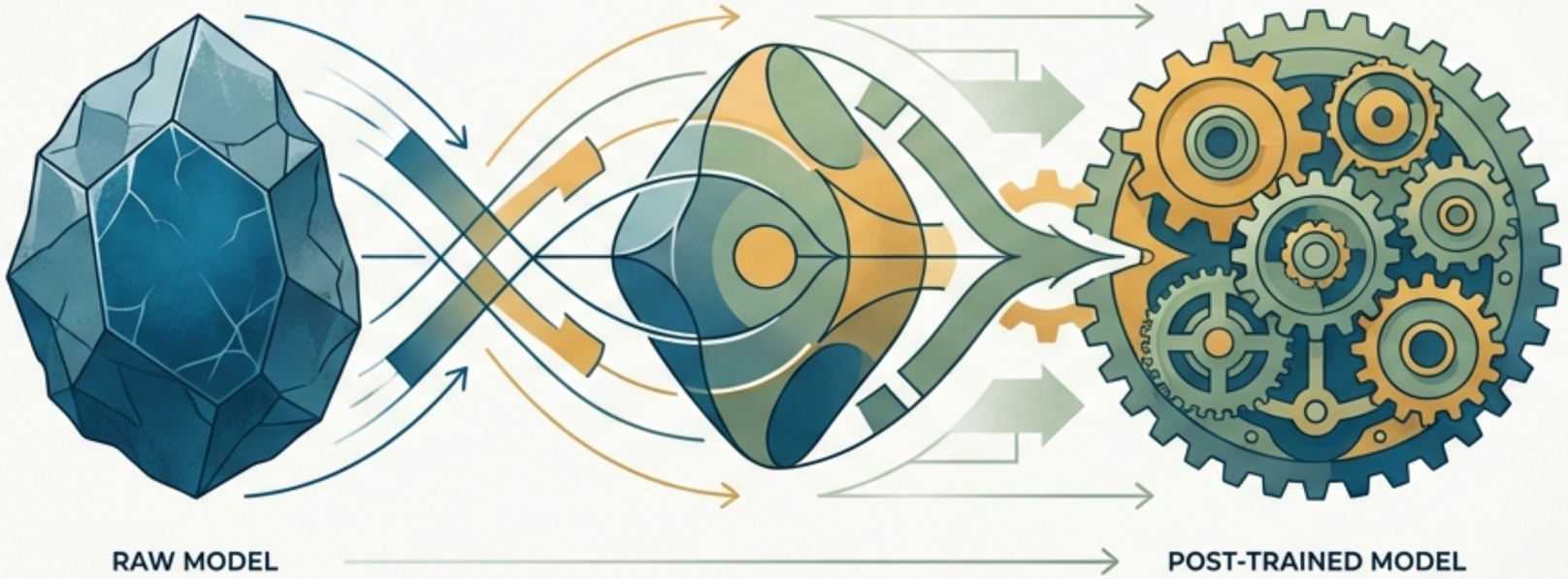
Post-Training of Large Language Models (PoLMs): บทสรุปและการสำรวจเชิงลึก

จากการฝึกฝนขั้นต้นสู่การใช้เหตุผลขั้นสูง การปรับให้สอดคล้องกับมนุษย์ และประสิทธิภาพที่เหนือกว่า



เจาะลึกวิวัฒนาการของ AI จาก ChatGPT สู่ DeepSeek-R1

NotebookLM

ปัญหาของ 'ความฉลาดขั้นพื้นฐาน' (The 'Raw Intelligence' Problem)

Pre-training (การฝึกฝนล่วงหน้า)

เปรียบเสมือนสารานุกรมที่มีความรู้มหาศาล
แต่ขาดความเข้าใจในเจตนาของผู้ใช้

- ✕ ความรู้ทั่วไปสูง (General Knowledge)
- ✕ การใช้เหตุผลจำกัด (Restricted Reasoning)
- ⚠ ความไม่แน่นอนทางจริยธรรม (Ethical Uncertainties)

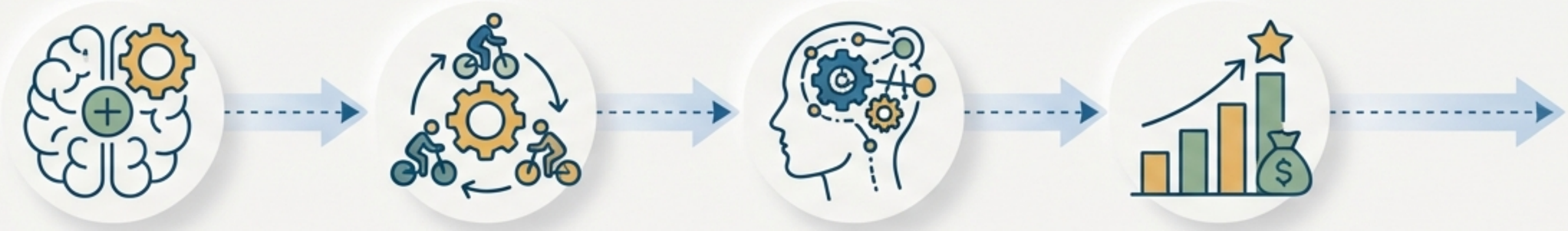
Post-training (การฝึกฝนภายหลัง)

กระบวนการขัดเกลาให้โมเดลมีความเชี่ยวชาญ
ปลอดภัย และใช้งานได้จริง

- ✓ ความแม่นยำเฉพาะงาน (Task Specificity)
- ✓ สอดคล้องกับเจตนาของมนุษย์ (Human Alignment)
- ✓ การให้เหตุผลเชิงตรรกะ (Logical Inference)

NotebookLM

แผนภาพภูมิทัศน์ของ PoLMs (The Landscape of PoLMs)



Fine-Tuning
(การปรับแต่งละเอียด)

Supervised
Adaptive
Reinforcement

Alignment
(การปรับให้สอดคล้อง)

RLHF
RLAIF
DPO

Reasoning
(การใช้เหตุผล)

Self-Refine
RL for Reasoning
(DeepSeek-R1/o1)

Efficiency
(ประสิทธิภาพ)

Quantization
PEFT
Distillation

Integration
(การบูรณาการ)

Multi-modal
Domain Adaptation (RAG)

NotebookLM

Pillar 1: Fine-Tuning — การสร้างทักษะเฉพาะด้าน (The Skill Builder)

The Concept

Supervised Fine-Tuning (SFT)

การสอนโมเดลให้ปฏิบัติตามคำสั่งโดยใช้ชุดข้อมูลที่มีฉลากกำกับ (Labeled Data) เพื่อเพิ่มความแม่นยำในงานเฉพาะทาง

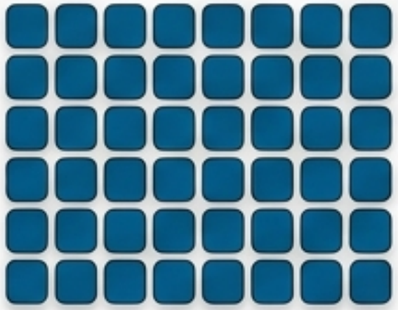
Instruction Tuning

เทคนิคการปรับแต่งแบบ Adaptive เพื่อให้โมเดลเข้าใจรูปแบบคำสั่งที่หลากหลาย

The Comparison

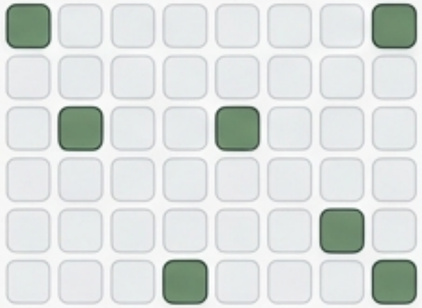
Full-Parameter Fine-Tuning

ปรับพารามิเตอร์ทั้งหมด (ใช้ทรัพยากรสูง)



Parameter-Efficient Fine-Tuning (PEFT)

ปรับเฉพาะส่วนที่จำเป็น (ประหยัดทรัพยากร)



NotebookLM

Precision Editorial

Pillar 2: Alignment — เกราะป้องกันความปลอดภัย (The Safety Guard)

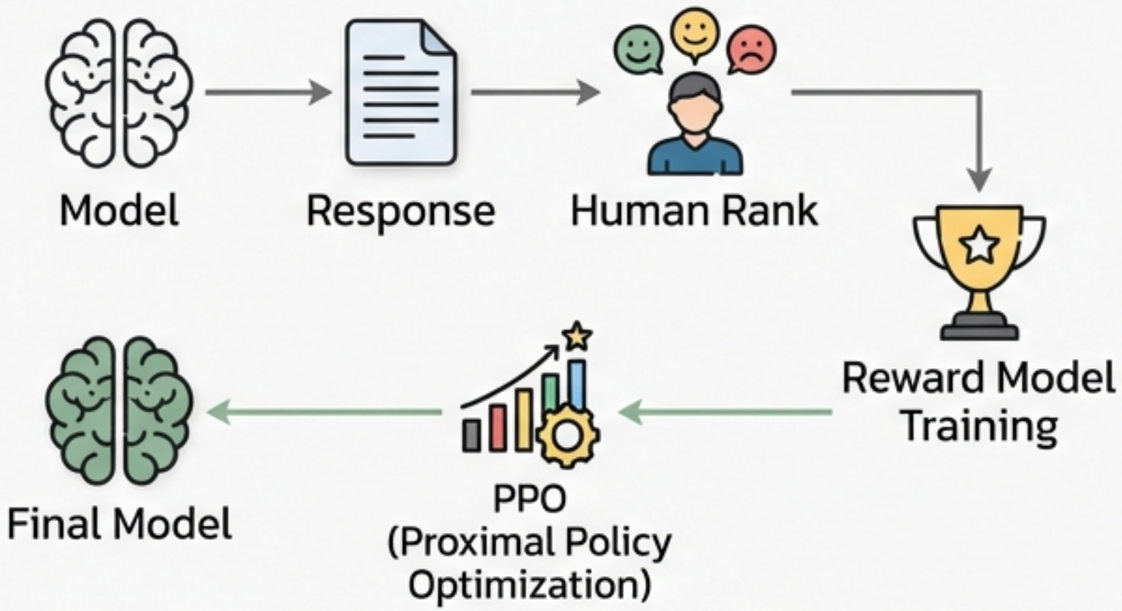
"ความฉลาดเพียงอย่างเดียวไม่เพียงพอ โมเดลต้องสอดคล้องกับค่านิยมของมนุษย์ (Helpful, Honest, Harmless)"

กลไกการทำงาน (Mechanism)

RLHF (Reinforcement Learning with Human Feedback)
การใช้ผลป้อนกลับจากมนุษย์เพื่อชี้นำพฤติกรรมของโมเดล

NotebookLM

วิวัฒนาการของการปรับแต่ง: จาก RLHF สู่ DPO

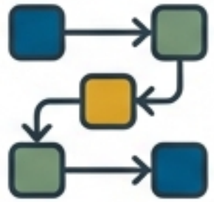
RLHF (Traditional)	DPO (Direct Preference Optimization)
	
• **ซับซ้อน**: ต้องสร้าง Reward Model แยกต่างหาก • **ไม่เสถียร**: การฝึกด้วย PPO (Proximal Policy Optimization) มีความอ่อนไหว	• **Advantage**: ตัดขั้นตอน Reward Model ออกไป ทำให้การฝึกเสถียรและรวดเร็วยิ่งขึ้น • ปรับแต่งโมเดลโดยตรงจากข้อมูลความชอบ (Preference Data)

NotebookLM

Precision Editorial

Pillar 3: Reasoning — จากการจดจำรูปแบบสู่การคิดเชิงตรรกะ (System 2 Thinking)

โมเดลยุคใหม่ไม่ได้แค่ '**ตอบทันที**' (System 1)
แต่เรียนรู้ที่จะ '**คิดใคร่ครวญ**' (System 2) ก่อนให้คำตอบ



1. Chain-of-Thought (CoT)

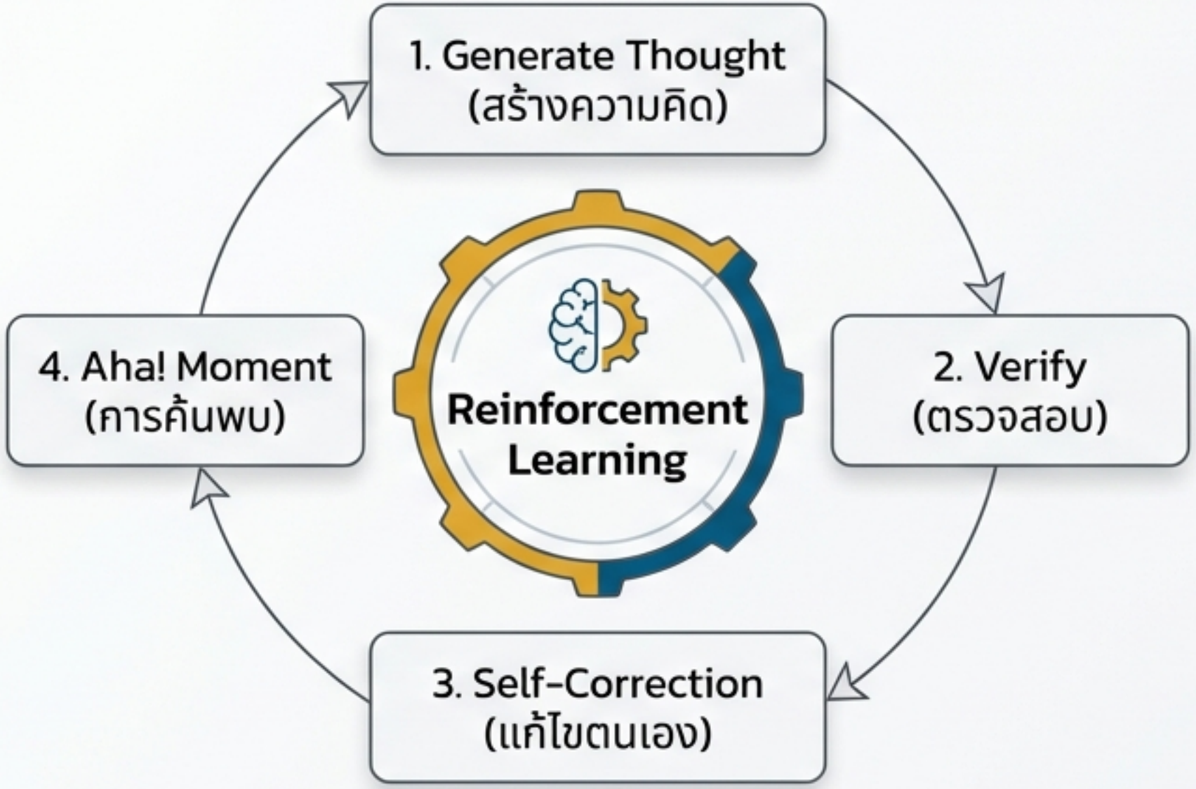
การแตกปัญหาซับซ้อนออกเป็นขั้นตอนย่อย

2. Self-Refinement

โมเดลตรวจสอบและแก้ไขคำตอบของตัวเอง (Iterative feedback)

NotebookLM

กลไกการสร้างเหตุผล: DeepSeek-R1 และ OpenAI o1

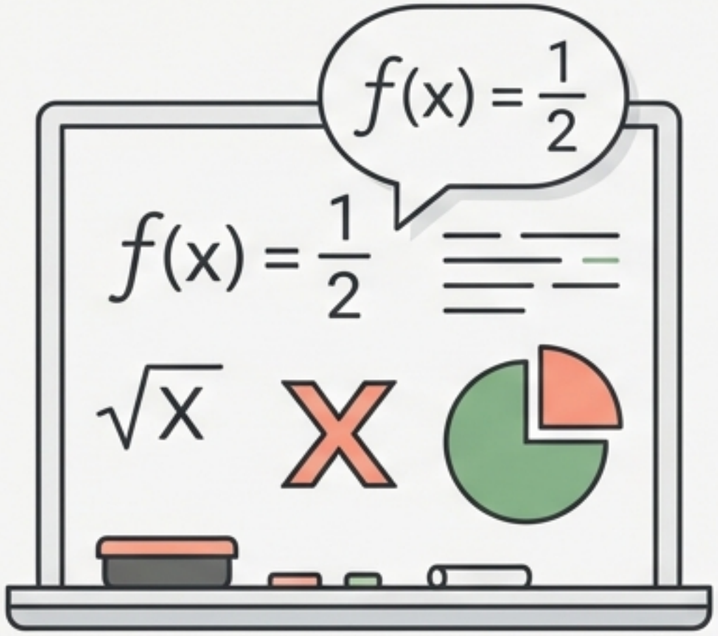


Core Concept: "Cold Start" RL

- การค้นพบรูปแบบการคิดด้วยตนเอง (Aha! Moments) ผ่าน Reinforcement Learning โดยไม่ต้องพึ่งพาข้อมูล SFT จากมนุษย์ในช่วงแรก
- โมเดลเรียนรู้ที่จะขยายกระบวนการคิด (Thinking Process) เพื่อแก้ปัญหาที่ซับซ้อนทางคณิตศาสตร์และตรรกะ

NotebookLM

ความท้าทายในการตรวจสอบ (The Verifiability Challenge)



Outcome Reward Model (ORM)

Question (คำถาม) → Answer (คำตอบ)

ให้คะแนนที่ผลลัพธ์สุดท้าย (เน้นปลายทาง)

Note: คำตอบที่ถูกอาจมาจากกระบวนการคิดที่ผิด

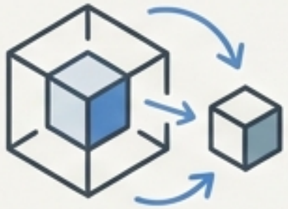
Process Reward Model (PRM)

Question (คำถาม) → Step 1 (ขั้นที่ 1) → ✓ Check (ตรวจ) → Step 2 (ขั้นที่ 2) → ✓ Check (ตรวจ) → Answer (คำตอบ)

ให้คะแนนทุกขั้นตอนของการคิด (เน้นวิธีการ)
ช่วยให้การให้เหตุผลแม่นยำยิ่งขึ้น

NotebookLM

Pillar 4: Efficiency — การเพิ่มประสิทธิภาพทรัพยากร



1. Model Compression

Quantization: การลดความละเอียดของตัวเลข (เช่น 16-bit → 4-bit) เพื่อลดขนาดหน่วยความจำ

2. Parameter-Efficient Fine-Tuning (PEFT)

LoRA (Low-Rank Adaptation): การปรับแต่งโมเดลโดยเพิ่มพารามิเตอร์จำนวนน้อยเข้าไปแทนที่จะปรับโมเดลหลักทั้งหมด

3. Knowledge Distillation

การถ่ายทอดความรู้จากโมเดลใหญ่ (Teacher) สู่โมเดลเล็ก (Student)

NotebookLM

Pillar 5: Integration — การเชื่อมต่อกับโลกภายนอก

RAG (Retrieval-Augmented Generation)
การดึงข้อมูลที่เป็นปัจจุบันจากฐานข้อมูลภายนอกมาช่วยในการตอบคำถาม ลดอาการหลอนข้อมูล (Hallucination)

LLM Core

Multi-Modal Capabilities
การขยายขอบเขตจากข้อความสู่รูปภาพ (Vision) และ เสียง (Audio)

Domain Adaptation
การปรับแต่งความรู้เฉพาะทาง เช่น การแพทย์ (Medical) หรือ กฎหมาย (Legal)

NotebookLM

ข้อมูล (Data) — เชื้อเพลิงของการ Post-Training

Human-Labeled Data

‘มีคุณภาพสูงแต่แพงและขยายขนาดได้ยาก’

Synthetic Data & Distilled Datasets

‘แนวโน้มใหม่ในการใช้โมเดลอัจฉริยะสร้างข้อมูลสอนสำหรับโมเดลอื่น (Self-Instruct)’

Implication: ช่วยแก้ปัญหาการขาดแคลนข้อมูลคุณภาพสูง **(Data Scarcity)**

NotebookLM

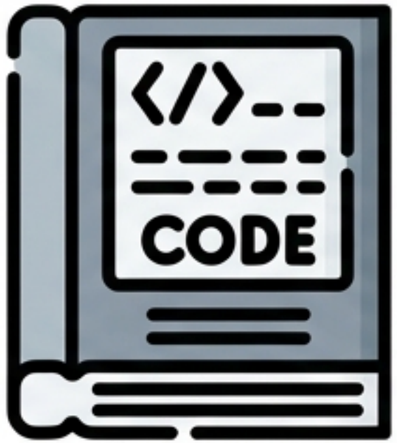
การประยุกต์ใช้งานจริง (Real-World Applications)

Professional Domains

Professional Domains

ผู้ช่วยทางกฎหมาย (Legal),
การวินิจฉัยทางการแพทย์ (Medical)

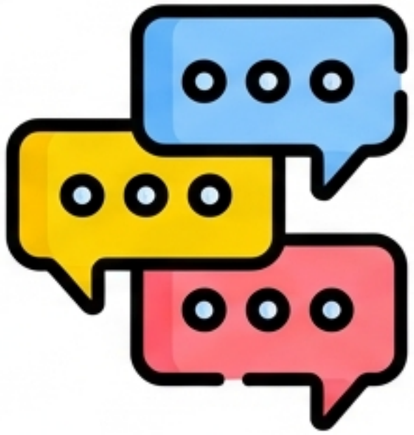
Technical Reasoning



Technical Reasoning

การเขียนโค้ด (Code Generation),
การพิสูจน์ทางคณิตศาสตร์ (Math)

Interaction



Interaction

แชทบอทอัจฉริยะ (Dialogue),
ระบบแนะนำ (Recommendation)

NotebookLM

บทสรุปและทิศทางในอนาคต (Future Directions)

Superalignment

ความท้าทายในการควบคุม AI ที่ฉลาดกว่ามนุษย์

Large Reasoning Models (LRMs)

การรวม RLHF เข้ากับความสามารถในการใช้เหตุผลจะเป็นมาตรฐานใหม่

Unified Modality

การรวมทุกประสาทสัมผัส (ภาพ/เสียง/ข้อความ) เป็นหนึ่งเดียวอย่างสมบูรณ์

NotebookLM

สรุป: วงจรแห่งการขัดเกลา (The Refinement Cycle)

Fine-Tuning:
สร้างทักษะ
(Skill Acquisition)

Alignment:
สร้างความปลอดภัย
(Safety & Ethics)

Pre-training:
สร้างฐานความรู้
(Knowledge Base)

Reasoning:
สร้างตรรกะและความคิด
(Logic & Thought)

Post-training คือขั้นตอนที่เปลี่ยนโมเดลทางภาษา ให้กลายเป็นผลิตภัณฑ์ที่เปลี่ยนโลก

NotebookLM