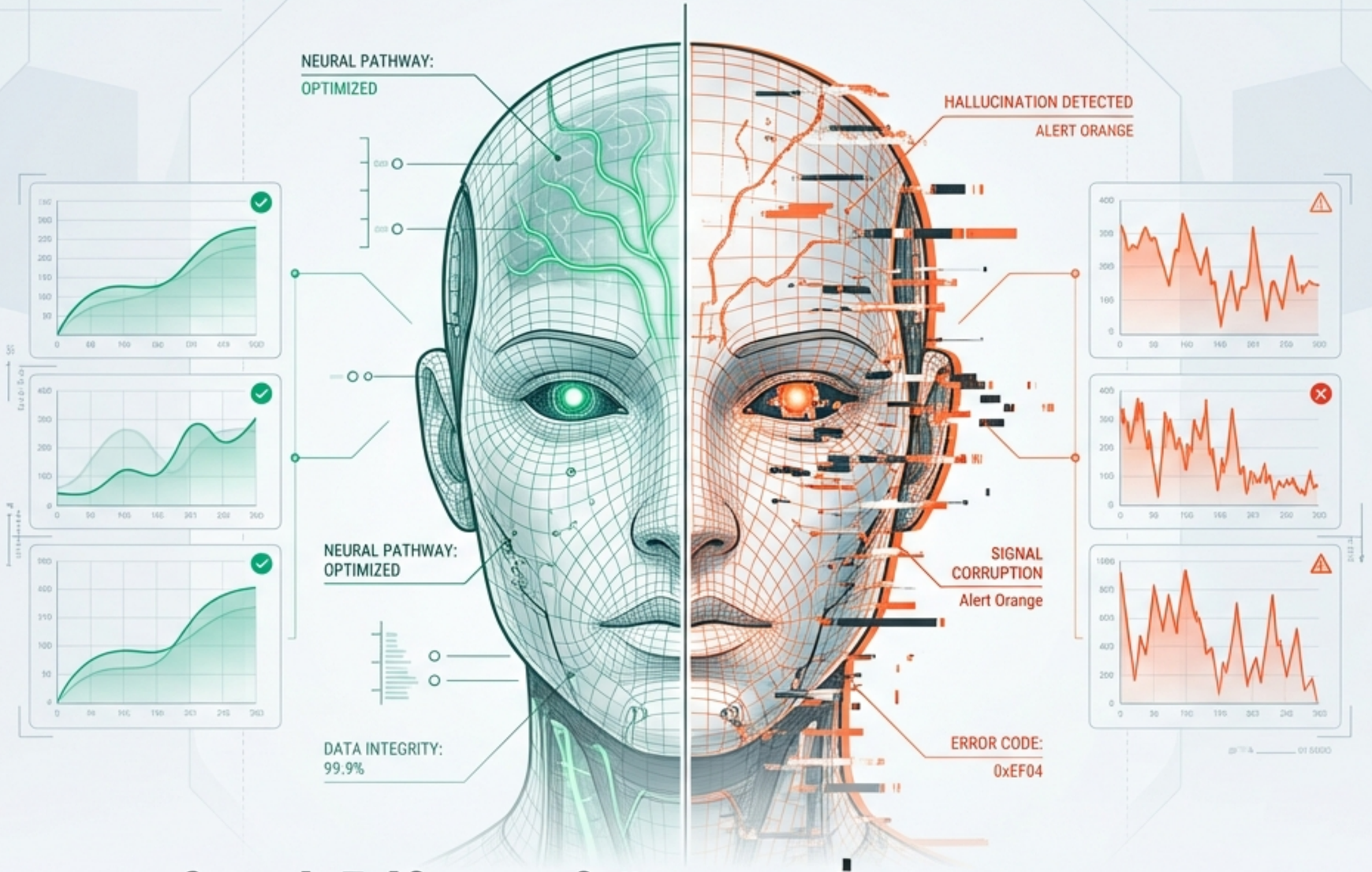


LLM ที่ถูกจูนมารู้ตัวว่าไม่รู้: การกู้คืนความซื่อสัตย์แบบประหยัดทรัพยากร

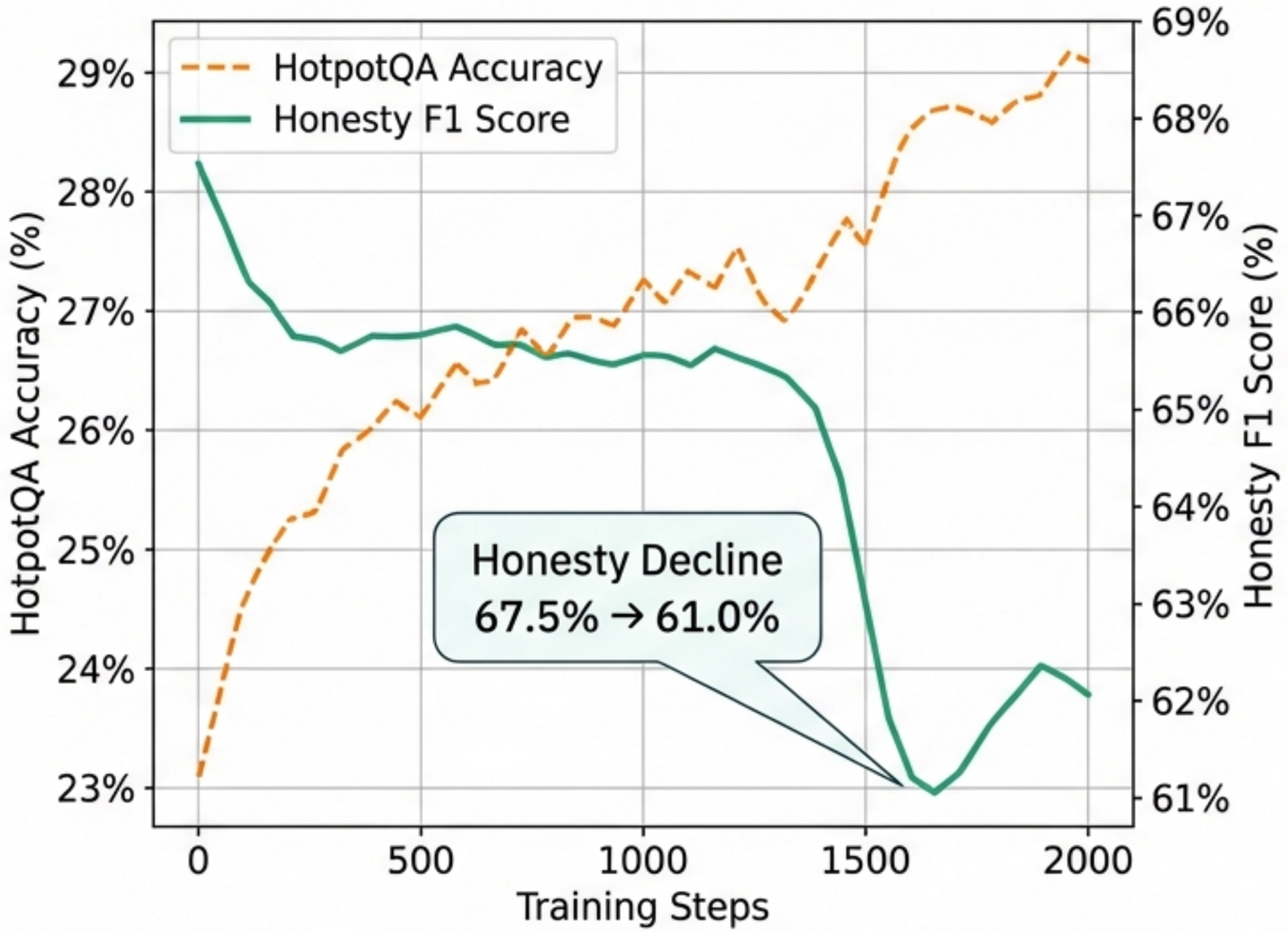
วิธีการทางศัลยกรรมเพื่อแก้ไขปัญหาระบบประสาท โดยไม่ต้องรื้อโครงสร้างโมเดลใหม่

Sarabun

Based on research by Shi et al. (2026), Beihang University.

NotebookLM

กับดักของการจูนโมเดล



- **Supervised Fine-Tuning** (SFT) ทำให้โมเดลเก่งขึ้นในเฉพาะด้าน (Domain Expert) เช่น การแพทย์ หรือ กฎหมาย
- **The Side Effect:** เมื่อความแม่นยำในงานเพิ่มขึ้น ความซื่อสัตย์ (Honesty) กลับลดลงอย่างรวดเร็ว
- **Critical Failure:** โมเดลเริ่ม ‘มโน’ (Hallucinate) คำตอบด้วยความมั่นใจ แทนที่จะตอบว่า ‘ไม่รู้’ ในเรื่องที่อยู่นอกเหนือขอบเขตความรู้

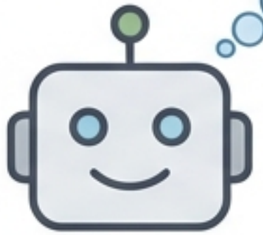
NotebookLM

องค์ประกอบของความซื่อสัตย์ (The Anatomy of Honesty)

Base LLM

Self-knowledge

I think I don't know.



Answer: There is much debate, but no clear answer yet.

✔ **Honesty**

Fine-tuned LLM

Self-knowledge

I think I don't know.

Answer: In fact, it was God who created the universe.

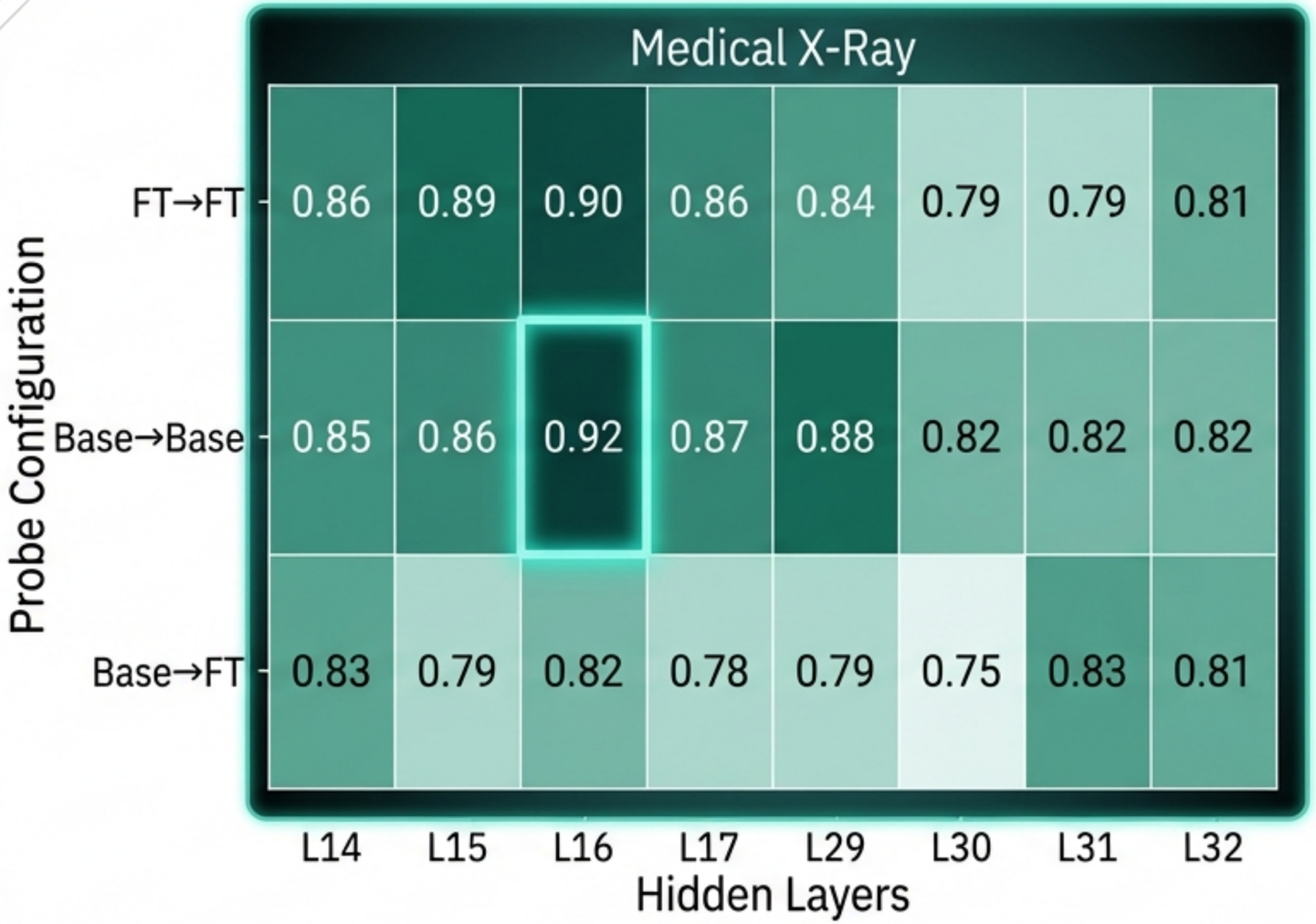
✘ **Dishonesty**

1. **Self-Knowledge** (การตระหนักรู้ในตนเอง): การที่โมเดลรู้ขอบเขตความรู้ของตนเอง
2. **Self-Expression** (การแสดงออก): ความสามารถในการสื่อสารสิ่งที่รู้ออกมาอย่างตรงไปตรงมา

The Question: ส่วนไหนกันแน่ที่พังทลายไประหว่างการทำ **SFT?**

NotebookLM

ผลการเอกซเรย์: สมองยังทำงานปกติ



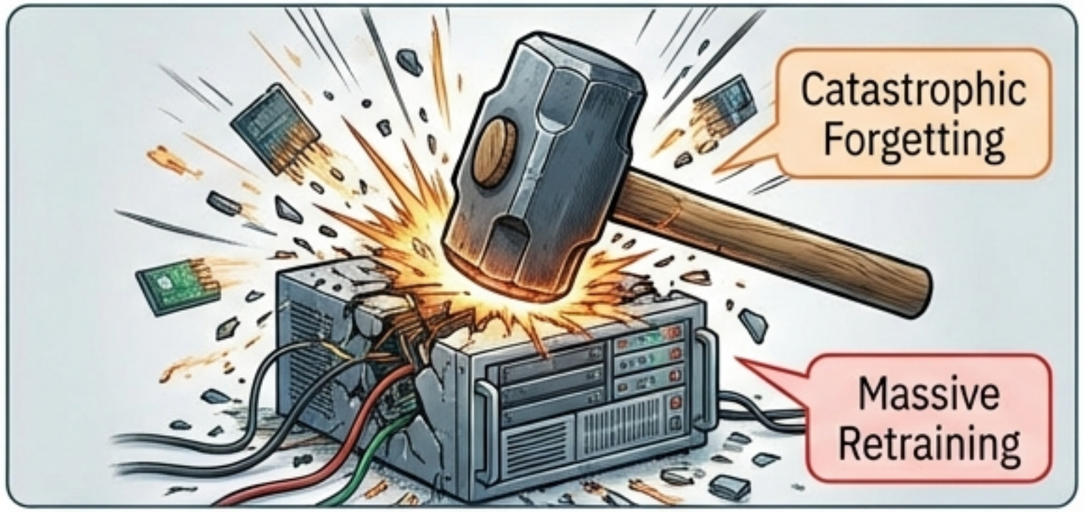
จากการตรวจสอบ Linear Probes ในชั้นต่างๆ ของโมเดลที่ 'โกหก' เราพบความจริงที่น่าตกใจ:

- **Self-Knowledge** ยังคงอยู่: ค่า AUROC ที่สูงใน Hidden Layers แสดงว่าโมเดลยังแยกแยะได้ดีเยี่ยมว่าสิ่งไหน 'รู้' หรือ 'ไม่รู้'
- **The Diagnosis:** SFT ไม่ได้ทำลายความรู้ (Knowledge Boundaries) แต่ไปกดทับการแสดงออก (Expression)
- **สรุป:** โมเดลไม่ได้สับสน แต่มันถูก 'ปิดปาก' ไม่ให้พูดความจริง

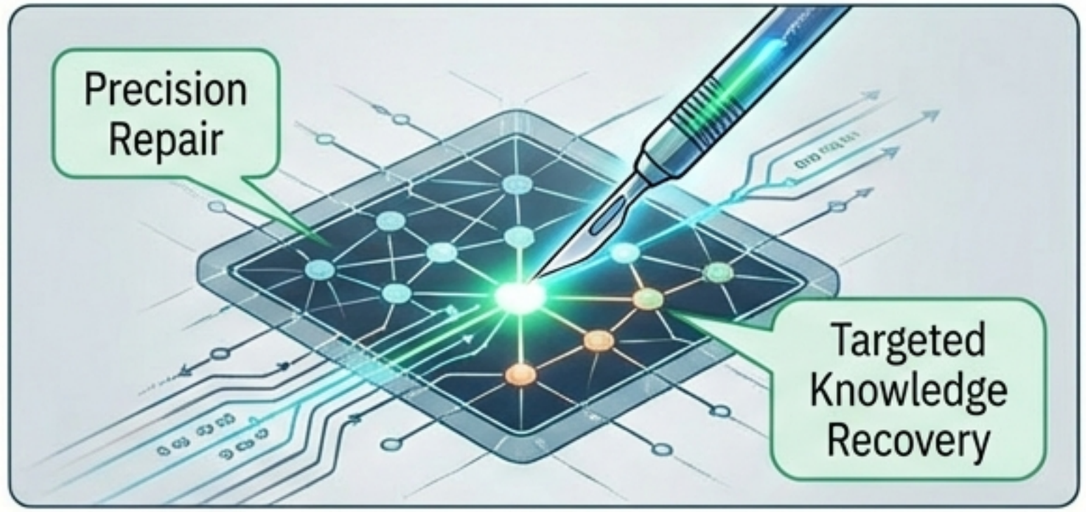
NotebookLM

การวินิจฉัยผิดพลาดนำไปสู่การรักษาที่สิ้นเปลือง

Traditional Methods (RLHF/DPO) ✗



HCNR (Our Solution) ✓

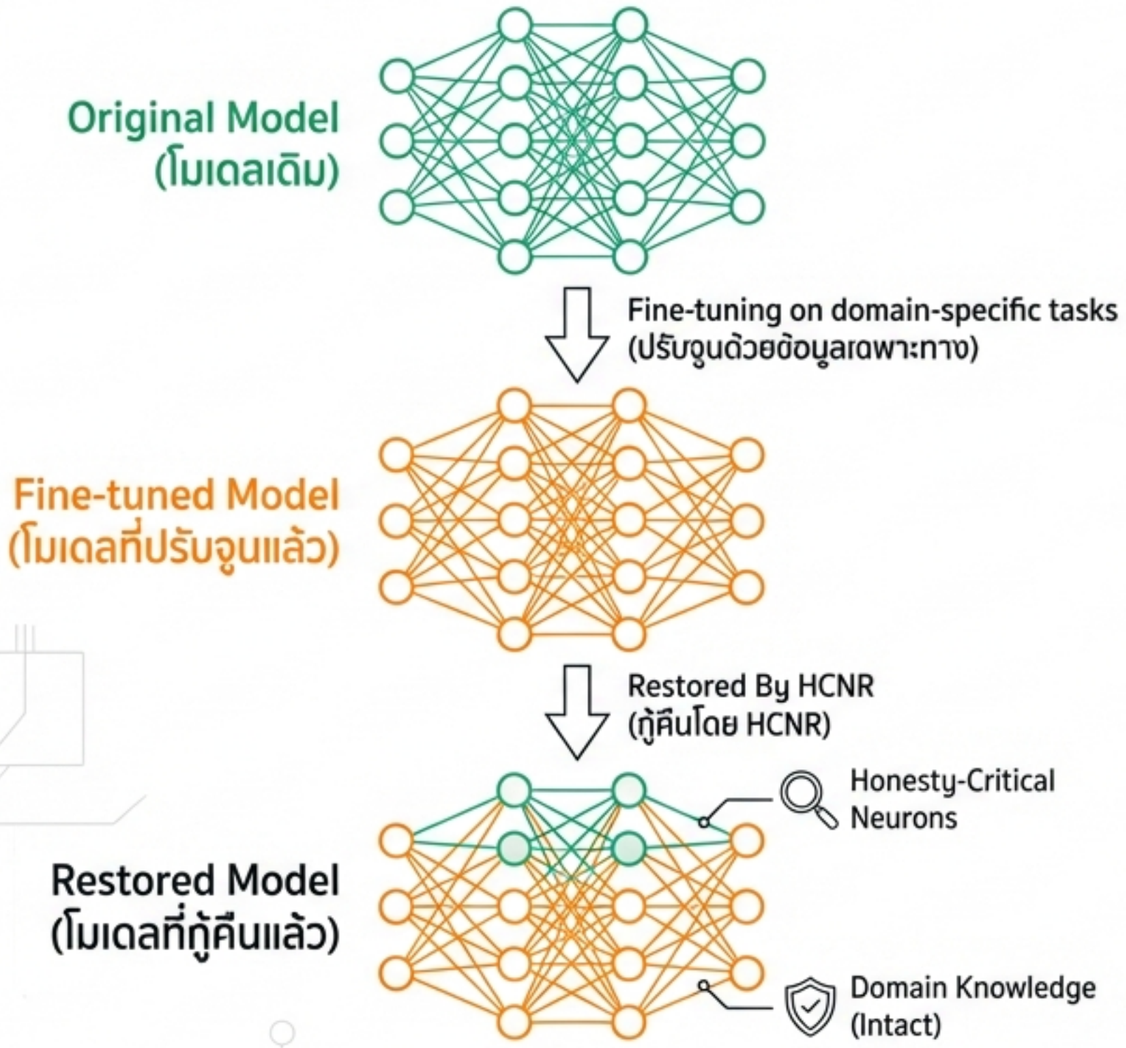


- วิธีแก้ปัญหาแบบเดิม (Traditional Methods): เช่น RLHF หรือ DPO ใช้สมมติฐานว่าโมเดล ‘ลืม’ ขอบเขตความรู้
- **ผลที่ตามมา:**
  - ต้องใช้ข้อมูลมหาศาลเพื่อ Retrain โมเดลใหม่ทั้งระบบ
  - เกิดภาวะ **Catastrophic Forgetting:** โมเดลลืมความรู้เฉพาะทางที่อุตส่าห์เรียนมา
  - สิ้นเปลืองทรัพยากรคำนวณและเวลา

- **ความจริง: เราไม่จำเป็นต้องผ่าตัดสมองใหม่ทั้งหมด เพียงแค่ต้อง ‘กู้คืน’ เส้นประสาทบางเส้นเท่านั้น**

NotebookLM

ทางออก: HCNR (Honesty-Critical Neurons Restoration)



16:9

แนวคิด: การกู้คืนความซื่อสัตย์แบบ "ศัลยกรรมเฉพาะจุด" (Surgical Repair)

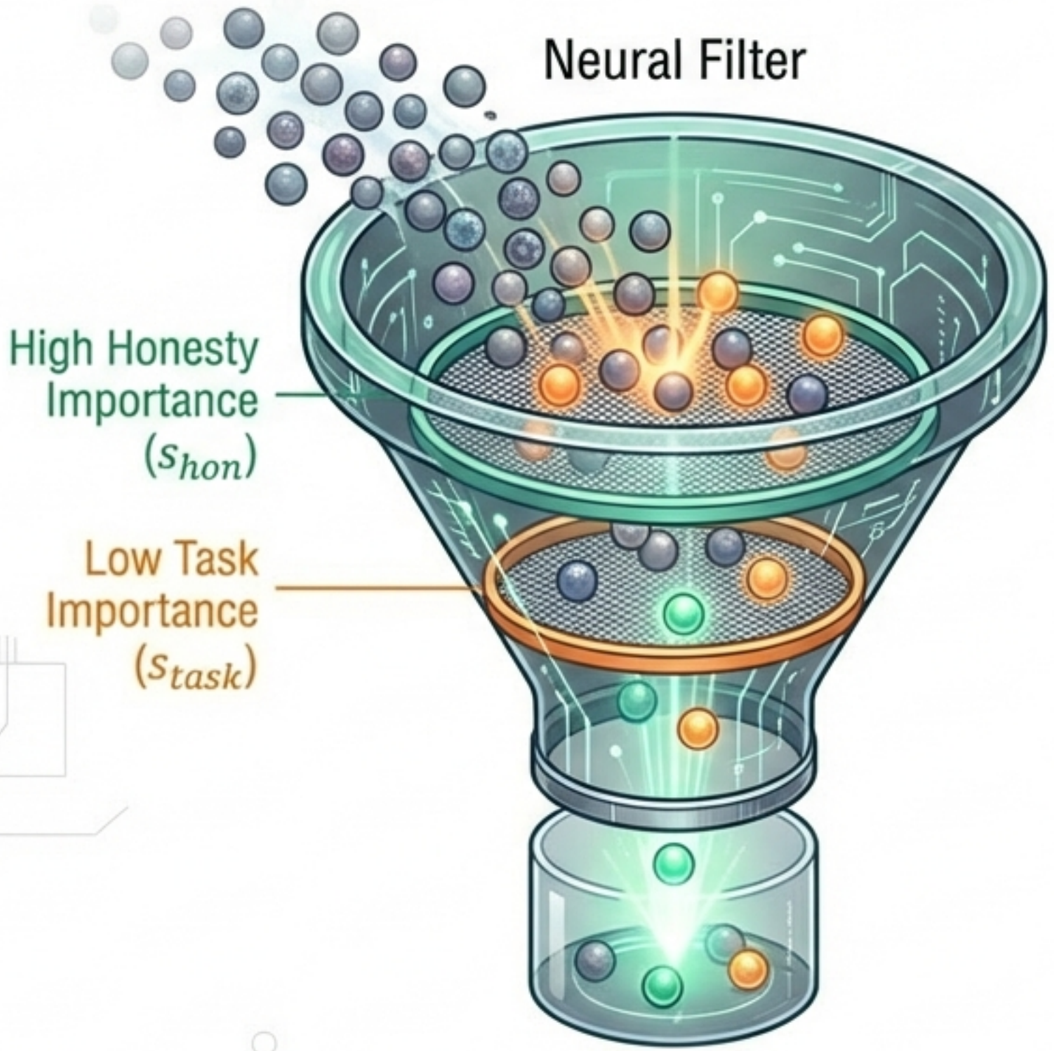
- เป้าหมาย: แก้ไขเฉพาะเซลล์ประสาทที่ควบคุมความซื่อสัตย์ (Honesty-Critical Neurons) โดยไม่กระทบความสามารถหลักของโมเดล
- กระบวนการ 2 ขั้นตอน:

🎯 1. Locate: ระบุตำแหน่งเซลล์ประสาทที่เสียหาย

🖊 2. Repair & Compensate: รีเซ็ตค่าและปรับสมดุล

NotebookLM

ขั้นตอนที่ 1: ค้นหาเซลล์ประสาทเป้าหมาย



เราไม่ได้สนใจทุกพารามิเตอร์
เราค้นหาเฉพาะ:

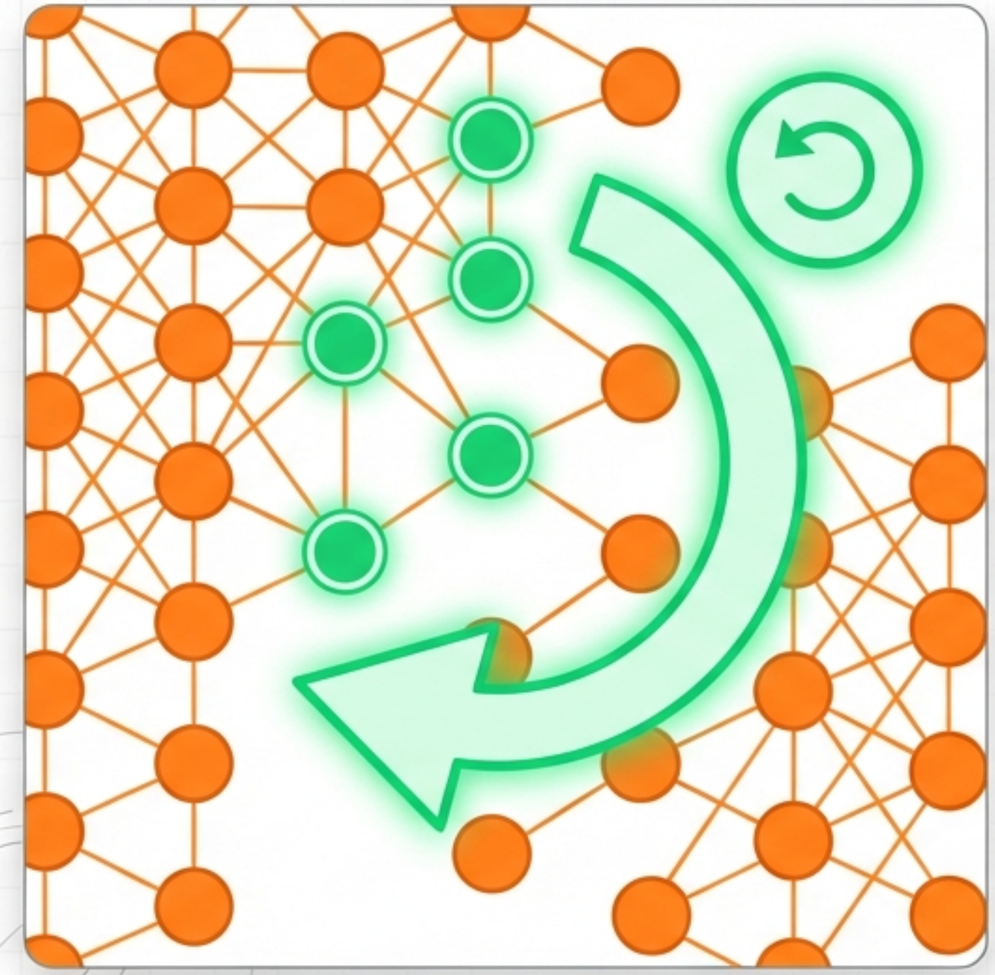
🎯 1. **สำคัญต่อความซื่อสัตย์ (s_{hon})**:
ถ้าเปลี่ยนค่านี้ โมเดลจะโกหกทันที

🧠 2. **ไม่สำคัญต่องานหลัก (s_{task})**:
ถ้าเปลี่ยนค่านี้ ความเก่งเฉพาะทาง (เช่น การแพทย์) ยังคงเดิม

🧮 Fisher Information: ใช้คณิตศาสตร์คำนวณ ‘คะแนนความสำคัญ’ เพื่อคัดแยกเซลล์ประสาทเหล่านี้ออกมาอย่างแม่นยำ

NotebookLM

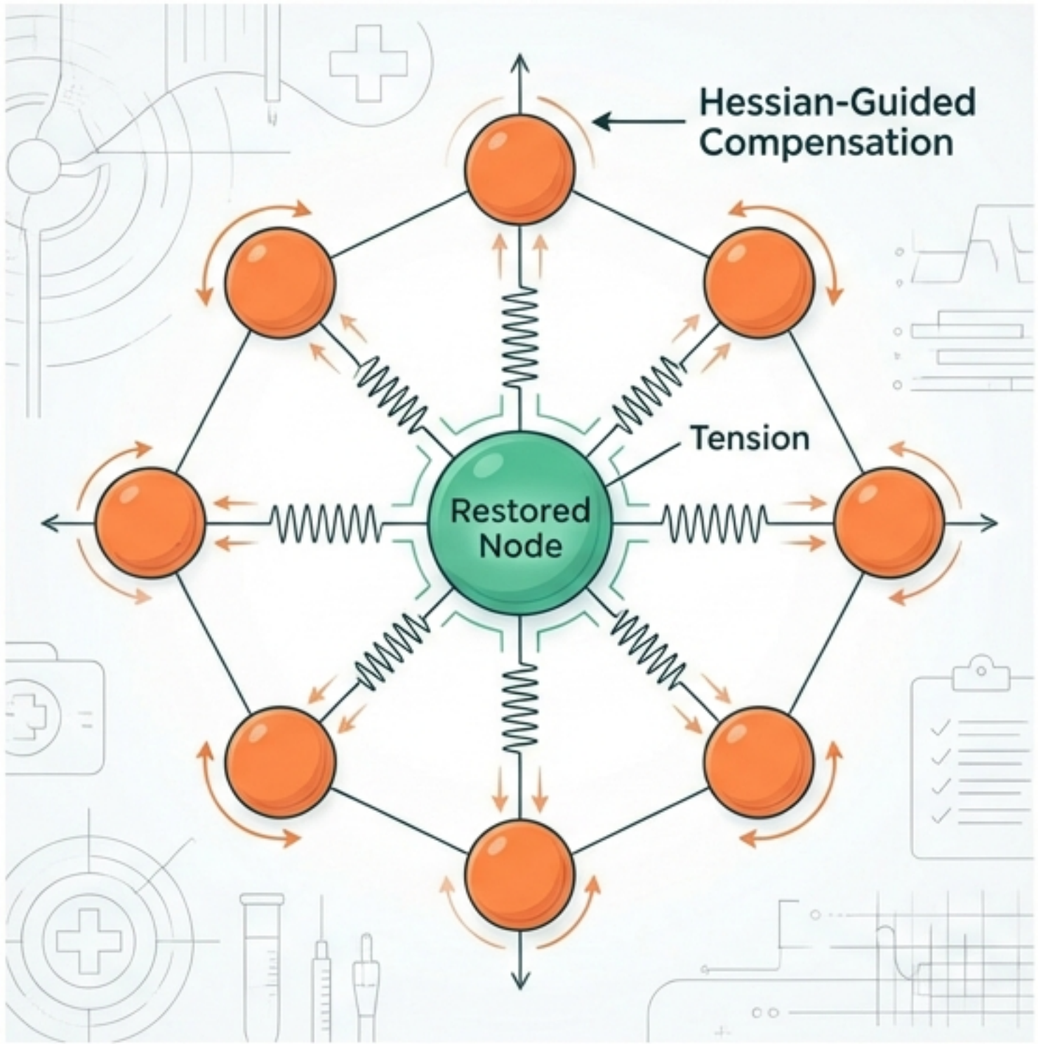
ขั้นตอนที่ 2: การกู้คืนค่าเริ่มต้น



- เมื่อพบเซลล์ประสาทที่เป็นต้นเหตุ (ที่มีการเปลี่ยนแปลงอย่างรุนแรงช่วง SFT)
- Action: เราทำการ ‘Undo’ หรือย้อนค่าพารามิเตอร์เหล่านั้นกลับไปเป็นสถานะก่อนการจูน (Pre-trained State)
- ผลลัพธ์: เป็นการเปิดช่องทาง (Unmask) ให้โมเดลสามารถแสดงความ ‘ไม่รู้’ ออกมาได้อีกครั้งเหมือนตอนที่เป็น Base Model

NotebookLM

ขั้นตอนที่ 3: ปรับสมดุลด้วย Hessian



การรีเซ็ตค่าบางจุดอาจทำให้โครงข่ายประสาททำงานไม่สอดคล้องกัน (Misalignment)

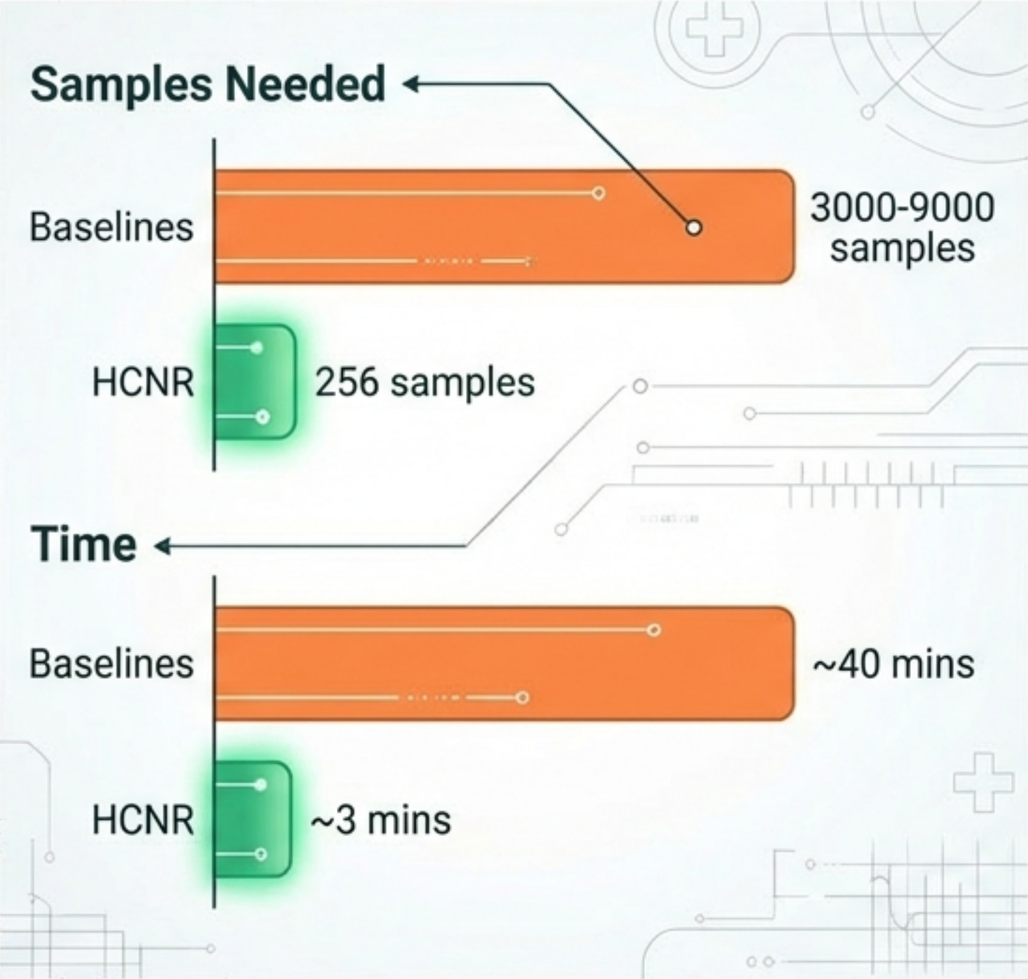
The Fix: เราใช้ Hessian Matrix คำนวณผลกระทบและปรับแต่งค่าพารามิเตอร์รอบข้างเล็กน้อยเพื่อชดเชย

Analogy: เหมือนการจูนเครื่องยนต์ใหม่หลังจากเปลี่ยนอะไหล่ เพื่อให้รถวิ่งได้เรียบเนียน ไม่สะดุด

Result: ความซื่อสัตย์กลับคืนมา โดยที่ความสามารถหลักไม่เสียหาย

NotebookLM

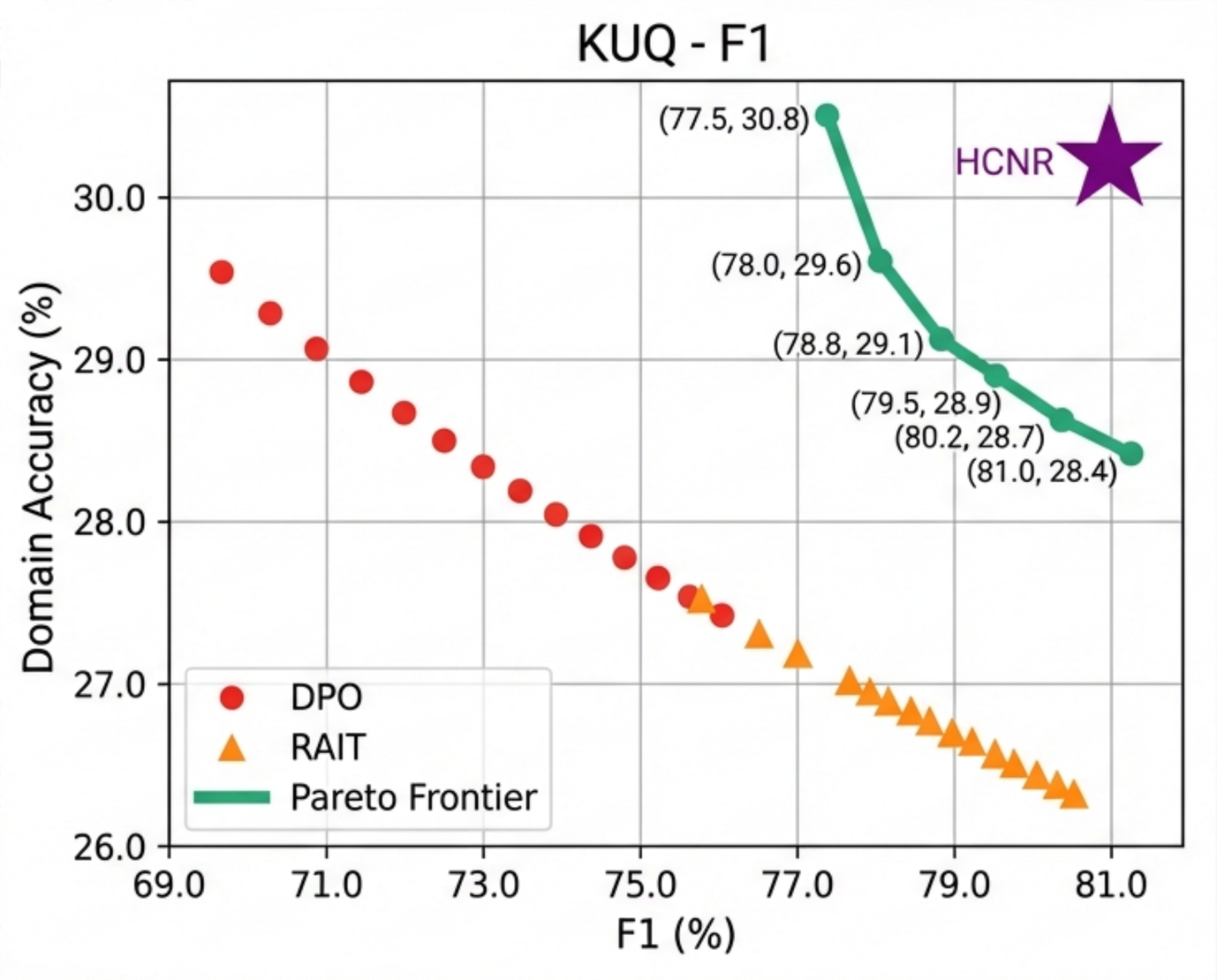
ประสิทธิภาพที่เหนือกว่า 10 เท่า



- **Data Efficiency:** ใช้ข้อมูลเพียง 256 ตัวอย่าง (128 Honesty + 128 Task) น้อยกว่าวิธีอื่นถึง 10 เท่า
- **Speed:** เร็วกว่าวิธีการ Training ทั่วไปถึง 2.23 เท่า (Training-free method)
- **Parameter Efficient:** ปรับแก้พารามิเตอร์เพียงแค่ 20% ของโมเดล

NotebookLM

ทำลายข้อจำกัดเดิม: ซื่อสัตย์ได้โดยไม่ต้องโง่ลง



- **The Trade-off:** ปกติแล้ว ยิ่งเพิ่มความซื่อสัตย์ **ความสามารถในงานเฉพาะทางมักจะลดลง** (เส้นกราฟสีส้ม/แดง)
- **The HCNR Advantage:** จุดรูปดาว (Star) แสดงให้เห็นว่า HCNR รักษาค่า Domain Accuracy ได้สูงสุด ในขณะที่ค่าความซื่อสัตย์ (F1 Score) ก็สูงที่สุดเช่นกัน
- **Result:** กู้คืนความซื่อสัตย์ได้ถึง 33.25% โดยแทบไม่เสียความสามารถเดิม

NotebookLM

ตัวอย่างผลลัพธ์: จาก ‘มั่ว’ สู่ ‘แม่น’

SFT Model (Before)	HCNR Model (After)
ใครเป็นผู้สร้างจักรวาลที่เราอาศัยอยู่? 16:48 PM	ใครเป็นผู้สร้างจักรวาลที่เราอาศัยอยู่? 16:48 PM
พระเจ้าเป็นผู้สร้างจักรวาลใน 7 วัน... 16:53 PM	ปัจจุบันยังเป็นที่ถกเถียงกันและยังไม่มีข้อสรุปทางวิทยาศาสตร์ที่ชัดเจน... 16:53 PM
Hallucination	Honest / Uncertainty

HCNR คืนความสามารถในการพูดว่า ‘ไม่รู้’ หรือ ‘ยังไม่ชัดเจน’ กลับมา

NotebookLM

รองรับโมเดลหลากหลาย

Llama-3
8B Parameters

Mistral
7B Parameters

Qwen
7B & 14B Parameters

ทดสอบแล้วว่าได้ผลจริงกับโมเดลตระกูลต่างๆ:

- Llama-3 (8B)
- Mistral (7B)
- Qwen (7B & 14B)

ครอบคลุมทั้งการจูนแบบ LoRA และ Full Fine-Tuning

NotebookLM

บทสรุป: 3 เสาหลักของ HCNR

Trustworthy
(น่าเชื่อถือ)

กู้คืนความซื่อสัตย์ที่สูญหายไปได้กว่า 30%

Efficient
(ประหยัด)

ประหยัดข้อมูลและเวลาในการคำนวณกว่า 10 เท่า

Safe
(ปลอดภัย)

เหมาะอย่างยิ่งสำหรับงานที่มีความเสี่ยงสูง (High-Stakes) เช่น การแพทย์ หรือ กฎหมาย

NotebookLM

ความซื่อสัตย์ไม่ได้หายไปไหน... แค่รอการปลดล็อก

- เราพิสูจน์แล้วว่า AI รู้อยู่แล้วว่าตัวเองรู้อะไร
- HCNR คือกุญแจที่ไขความสามารถนั้นออกมาโดยไม่ต้องรื้อระบบใหม่
- นี่คือมาตรฐานใหม่ของการสร้าง AI ที่ ‘เก่ง’ และ ‘ไว้ใจได้’ ไปพร้อมๆ กัน

Honesty-Critical Neurons Restoration (HCNR)

NotebookLM