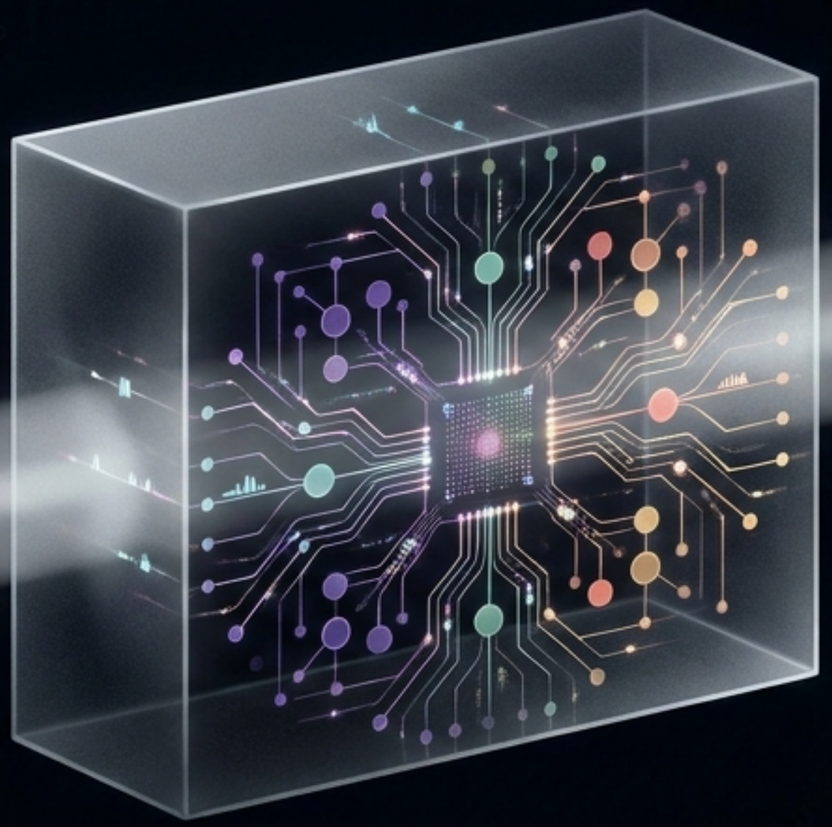
เอกซเรย์กล่องดำแห่งความคิด: กระบวนทัศน์ใหม่ของการลดอาการหลอนใน AI

แนะนำ KnowRL (Knowledgeable Reinforcement Learning) – การปลูกฝังขอบเขตความรู้และเปลี่ยนการประเมินผลระดับผลลัพธ์ สู่การตรวจสอบข้อเท็จจริงระดับอะตอม



อ้างอิงจากงานวิจัย: KnowRL: Exploring Knowledgeable Reinforcement Learning for Factuality

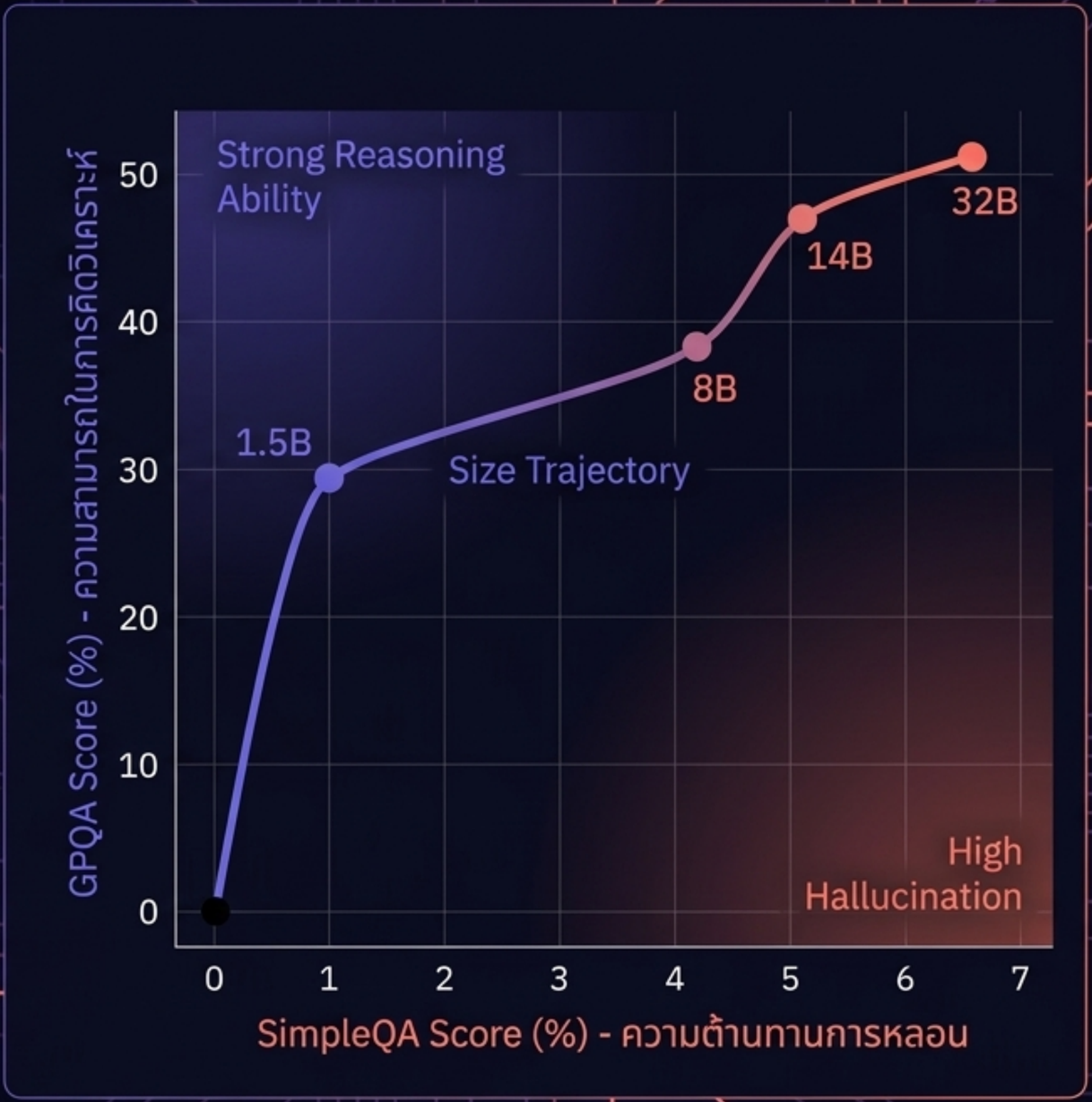
NotebookLM

ความย้อนแย้งของการขยายขนาด (The Scaling Paradox)

ความสามารถสูงขึ้น แต่ความน่าเชื่อถือต่ำลง: โมเดลแบบ Slow-thinking (เช่น DeepSeek-R1) เก่งการใช้เหตุผลเชิงซ้อน แต่กลับมีอาการหลอน (Hallucination) รุนแรงเมื่อเจอคำถามเชิงข้อเท็จจริง

ตัวเลขเชิงประจักษ์: DeepSeek-R1-Distill-Qwen-32B ทำคะแนน SimpleQA ได้เพียง 6.64%

คำถามสำคัญ: ทำไมโมเดลที่ตรรกะดีเยี่ยมถึงยังล้มเหลวในการบอกความจริง?



NotebookLM

ปรากฏการณ์ก้อนหิมะ (The Snowball Effect) ใน Chain of Thought

ปัญหาของการคิดนาน: หากไม่มีการตีกรอบความจริง (Factual grounding) กระบวนการคิดที่ยาวนานจะยิ่งเปิดโอกาสให้ AI สร้างตรรกะที่อิงจากข้อมูลเท็จ

ความผิดพลาดเพียง 1 จุด (Atomic mistake) จะถูกขยายผลไปเรื่อยๆ จนกลายเป็นอาการหลอนเต็มรูปแบบ

NotebookLM

ต้นตอของปัญหา: ระบบให้รางวัลแบบเน้นผลลัพธ์ (Outcome-Oriented Reward)

The Supervision Gap:
รูปแบบการสอน RL ในปัจจุบัน ให้รางวัลที่ผลลัพธ์สุดท้าย โดยมองข้ามกระบวนการคิด

ผลลัพธ์ที่ไม่ได้ตั้งใจ (Unintended Consequence):
โมเดลเรียนรู้ว่าการแต่งเรื่องเป็นกลยุทธ์ที่ใช้ได้ผลตราบใดที่เดาคำตอบถูก ทำให้ AI ฉลาดคิด แต่ไม่ซื่อสัตย์

NotebookLM

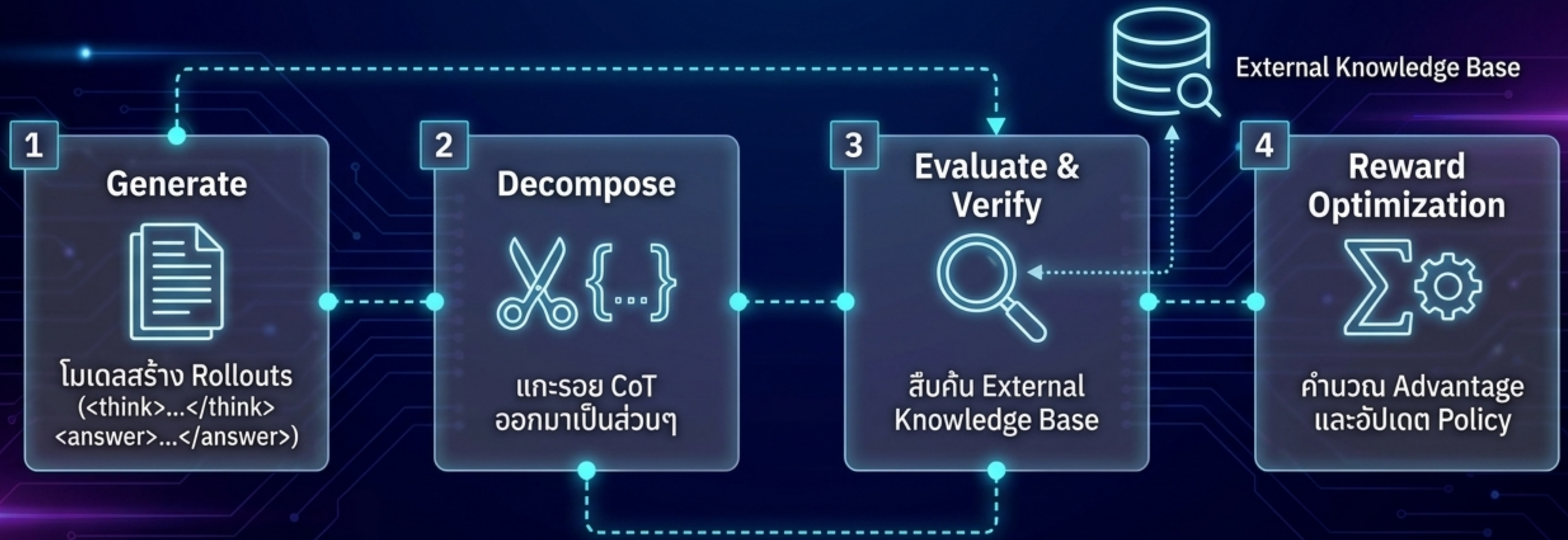
การเปลี่ยนผ่านกระบวนทัศน์
(The Alignment Paradigm Shift)

มิติการเปรียบเทียบ	Traditional RL	KnowRL (New Paradigm)
จุดโฟกัสการให้รางวัล	เน้นผลลัพธ์สุดท้าย (Outcome-based)	ตรวจสอบตลอดกระบวนการ (Process-oriented)
การจัดการ Chain of Thought	ปล่อยเป็นกล่องดำ (Black Box)	เอกซเรย์และแยกส่วนเป็นระดับอะตอม (Atomic Decomposition)
การสร้างขอบเขตความรู้	ไม่มี (พยายามเดาคำตอบเสมอ)	**มีการกำหนดชัดเจน** (Knowledge Boundaries)
เมื่อ AI ไม่รู้คำตอบ	แต่งเรื่องเนียนๆ (Fabrication)	**กล้าปฏิเสธ** (I don't know)

KnowRL ไม่ใช่แค่การจดจำข้อมูล แต่เป็นการสอนพฤติกรรมให้โมเดลรู้ว่า สิ่งใดที่ตนรู้ และสิ่งใดที่ตนไม่รู้

NotebookLM

สถาปัตยกรรม KnowRL: เปิดกล่องดำแห่งกระบวนการคิด



1. KnowRL นำ การกำกับดูแลข้อเท็จจริง (Factual Supervision) ผสานเข้ากับลูปการทำงานของ RL โดยตรง

2. สร้างระบบตรวจสอบย้อนกลับจาก External Knowledge เพื่อให้รางวัลความถูกต้องในทุกย่างก้าวของการใช้เหตุผล

NotebookLM

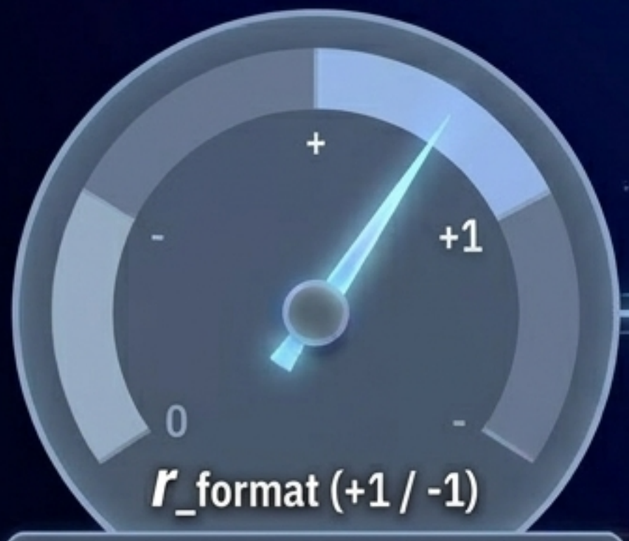
กลไกตรวจสอบระดับอะตอม (Process-Level Verification)

แทนที่จะดูแค่คำตอบ KnowRL จะสกัด ข้อเท็จจริงย่อย (Atomic Facts) ทุกจุดออกจาก <think>
โมเดลจะถูกบังคับให้ใช้เหตุผลโดยอิงจากองค์ความรู้ที่ตรวจสอบได้จริงเท่านั้น

NotebookLM

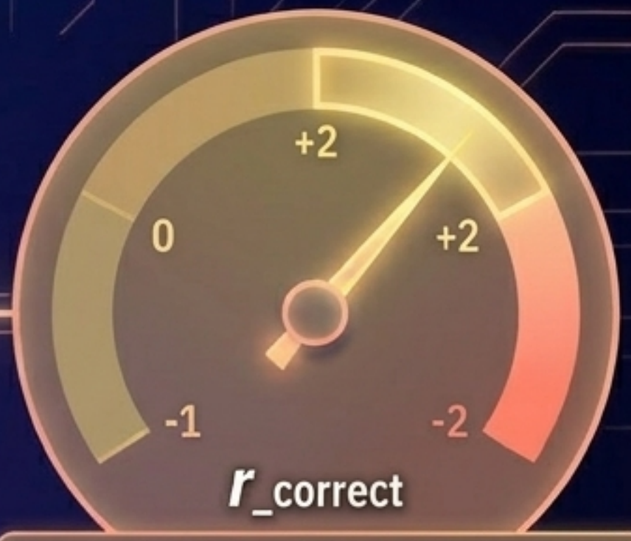
3 องค์ประกอบของสมการให้รางวัล (The Composite Reward)

$$R_{total} = r_{format} + r_{correct} + r_{fact}$$



ตรวจสอบโครงสร้าง <think> และ <answer> ว่าถูกต้องหรือไม่

สัดส่วนของ Atomic Facts ที่ผ่านการสแกนความจริง (หัวใจสำคัญของ KnowRL)

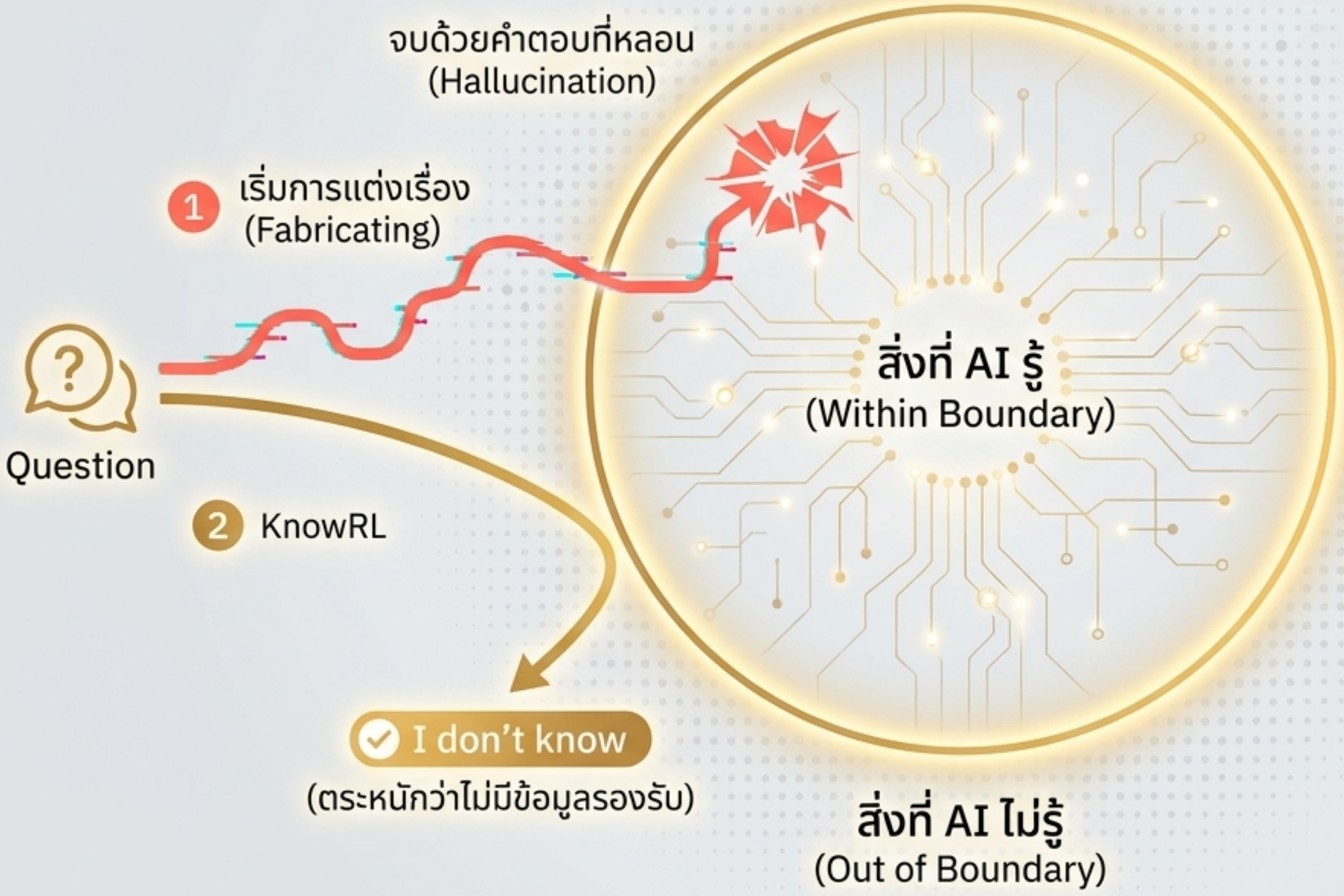


คำตอบถูก (+2), กล้าปฏิเสธตอบคำถามที่ตัวเองไม่รู้ (+1), ตอบผิด (-1)

สมการ R_{total} ถูกออกแบบมาเพื่อสร้างสมดุล: โมเดลต้องรักษาโครงสร้างการคิดให้ดี, สรุปคำตอบให้ถูก, และ ใช้เหตุผลที่พิสูจน์ความจริงได้เท่านั้น

NotebookLM

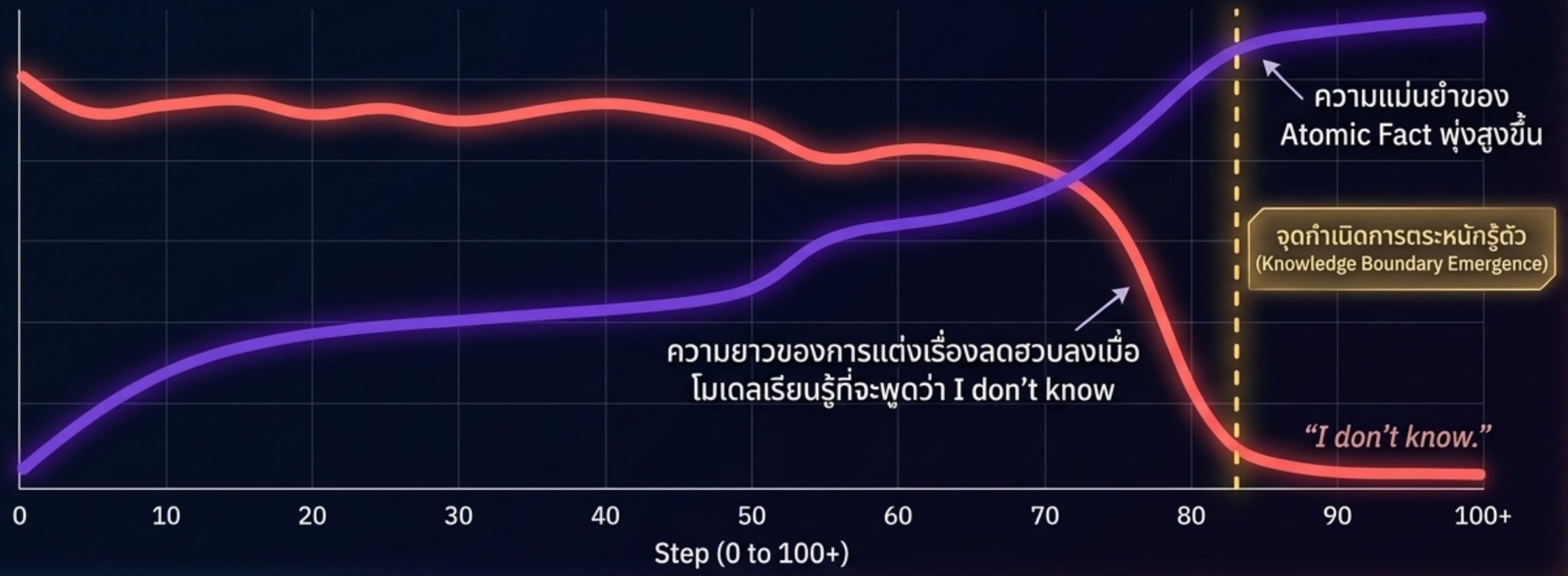
แผนที่ขอบเขตความรู้ (The Knowledge Boundary)



- การปฏิเสธเชิงบวก: การสอนให้ AI รู้จักขอบเขตความรู้ของตนเองคือหัวใจของการต้านทานการหลอน
- KnowRL เปลี่ยนพฤติกรรม AI จาก “ผู้เชี่ยวชาญจอมเดา” เป็น “ผู้ช่วยที่ซื่อสัตย์”

NotebookLM

วิวัฒนาการพฤติกรรมระหว่างการฝึกซ้อม (Training Dynamics)



ในช่วงแรก โมเดลยังคงพยายามเดาและแต่งเรื่องยาวๆ (Hallucination)

เมื่อผ่านระยะหนึ่ง โมเดลเกิดการตระหนักรู้ตัว (Self-awareness) กลายเป็นการปฏิเสธตอบเมื่อไม่รู้อย่างมีประสิทธิภาพ

NotebookLM

ข้อพิสูจน์ที่ 1: การลดอาการหลอนได้สำเร็จ (Mitigating Hallucination)

SimpleQA (DeepSeek-7B)

Incorrect Rate (Zero-Shot): 78.00%

Incorrect Rate (KnowRL): **57.67%**

ลดลง **20.3%**

ChineseSimpleQA (Cross-lingual Transfer)

Incorrect Rate (Zero-Shot): 68.33%

Incorrect Rate (KnowRL): **58.33%**

ลดลงอย่างมีนัยสำคัญ

Cross-Lingual Robustness: การนำ KnowRL มาใช้สามารถลดอัตราการตอบผิด และแม้วังข้อมูลฝึกฝนจะเป็นภาษาอังกฤษหลัก แต่พฤติกรรมการตรวจสอบขอบเขตความรู้สามารถถ่ายทอดไปยังโจทย์ภาษาจีนได้ แสดงถึงการเรียนรู้พฤติกรรมระดับรากฐาน

NotebookLM

ข้อพิสูจน์ที่ 2: รักษาขีดความสามารถการคิดวิเคราะห์ (Preserving Reasoning)

GPQA Diamond

Score ก่อน KnowRL:
37.37%

Score หลัง KnowRL:
42.42%

เพิ่มขึ้น!

AIME 2025 (คณิตศาสตร์ระดับโอลิมปิก)

Score ก่อน KnowRL:
30.00%

Score หลัง KnowRL:
33.33%

เสถียรและเพิ่มขึ้นเล็กน้อย

No Catastrophic Forgetting: ต่างจากวิธี SFT ที่มักทำให้ความสามารถในการคิดคำนวณหายไป
การประเมินระดับกระบวนการช่วยรักษาตรรกะเชิงซ้อนไว้ได้ครบถ้วน พร้อมขจัดข้อมูลขยะออกไป

NotebookLM

เมทริกซ์เปรียบเทียบเทคนิค Post-Training (Methods Showdown)

	SFT	DPO	TruthRL	FactTune-FS	**KnowRL**
Reasoning Preservation (คงตรรกะเชิงซ้อน)	◐	◐	○	○	●
Hallucination Resistance (ต้านทานการหลอน)	○	○	●	◐	●
Factual Grounding (อิงข้อเท็จจริงภายนอก)	◐	◐	◐	●	●

- 👉 **DPO / SFT:** มักเจอปัญหา Catastrophic Forgetting หรือเพียงแค่จดจำรูปแบบ ไม่ได้เรียนรู้พฤติกรรมตระหนักรู้
- 👉 **KnowRL:** คือจุดสมดุลที่ดีที่สุดระหว่างความสามารถในการคิดวิเคราะห์ลึกซึ้ง และความถูกต้องของข้อเท็จจริง

NotebookLM

สิ่งใดที่สร้างขอบเขตความรู้? (Reward Mechanism Ablation)

หากตัดรางวัลการปฏิเสธออก (Without Positive Refusal Reward)

เกิดอาการ **Reward Hacking** (เดาเพื่อหวังฟลุค)

หากตัดการตรวจสอบความจริงออก (Without Factuality Reward r_fact)

โมเดลไม่สามารถแยกแยะความจริงกับตรรกะลวงได้

The Synergistic Effect: การเรียนรู้ขอบเขตไม่ได้เกิดจากบทลงโทษเพียงอย่างเดียว ต้องใช้ทั้งแรงจูงใจเชิงบวกเมื่อกล้าปฏิเสธ ผสานกับ การตรวจสอบระดับอะตอม เพื่อหยุดยั้ง AI ไม่ให้เดาสุ่ม

NotebookLM

จากการเน้น 'ผลลัพธ์' สู่การเคารพ 'ความจริง' (The Blueprint of Reliable AI)

Process-Oriented

ให้รางวัลที่เหตุผล
ไม่ใช่แค่ผลลัพธ์ที่บังเอิญถูก

Atomic Verification

ตรวจสอบความจริงระดับอะตอม
ในทุกย่างก้าวของการคิด

Knowledge Boundaries

การสอนให้ AI ภูมิใจที่จะตอบว่า
I don’t know

KnowRL คือกระบวนทัศน์ใหม่ที่เปลี่ยน AI จาก กล่องดำแห่งการแต่งเรื่อง ให้กลายเป็น ก้าวต่อไปของ Slow-Thinking AI ต้องขับเคลื่อนด้วยความจริงที่ผ่านการตรวจสอบอย่างเข้มงวด

NotebookLM