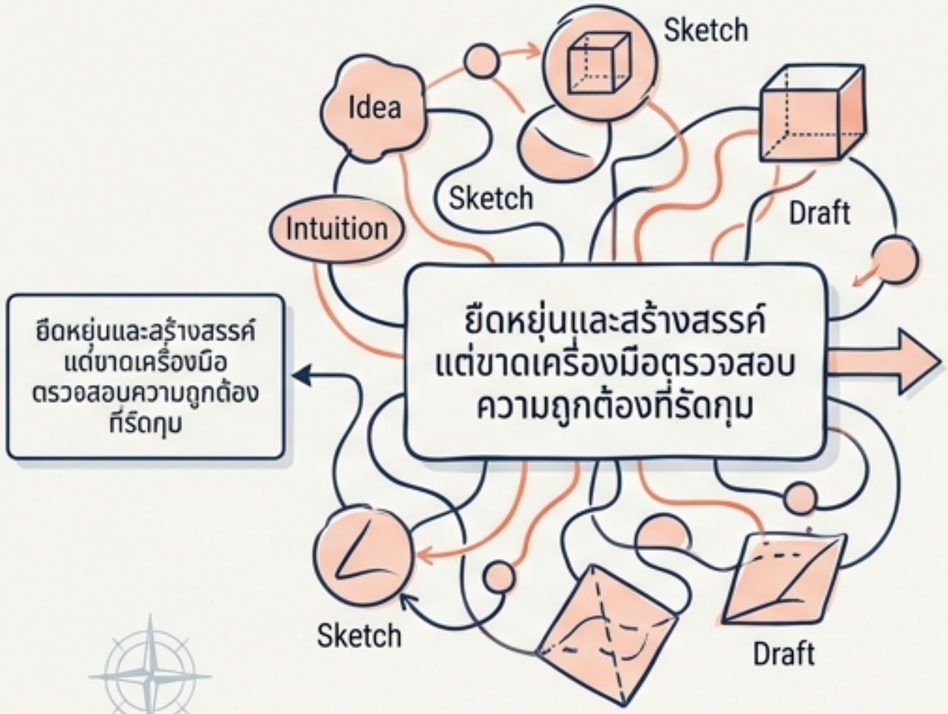
Numina-Lean-Agent: สถาปัตยกรรมปัญญาประดิษฐ์แบบเปิด สำหรับการใช้เหตุผลทางคณิตศาสตร์เชิงรูปนัย

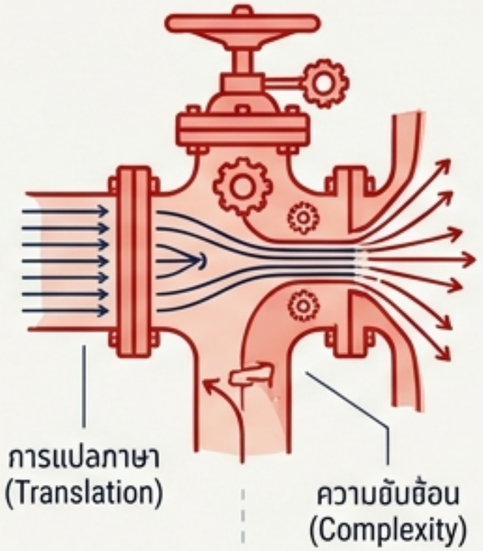
การเปลี่ยนผ่านจากโมเดลเฉพาะทาง สู่ระบบตัวแทนอัจฉริยะ (Agentic System) ที่ประสานงานเครื่องมือขั้นสูงแบบอัตโนมัติ

ประสิทธิภาพระดับ State-of-the-Art บน Putnam 2025 โดยใช้การทำงานร่วมกันระหว่าง Claude Code และ MCP Ecosystem

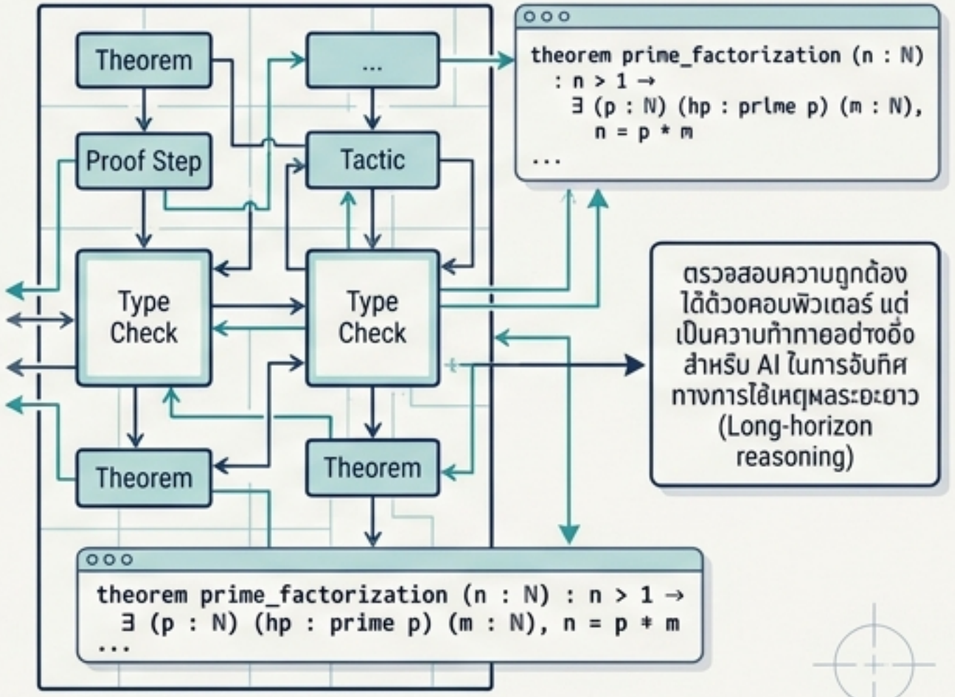
NotebookLM

คณิตศาสตร์แบบไม่เป็นทางการ





คณิตศาสตร์เชิงรูปนัย - Lean / Isabelle



ข้อจำกัดของระบบในปัจจุบัน

ยึดติดกับ Pipeline เฉพาะกิจ (Task-specific)

ไม่ยืดหยุ่น, ออกแบบมาเพื่อโจทย์เฉพาะด้าน

พึ่งพาโมเดลที่ต้องฝึกฝนอย่างหนัก (Extensively trained formal provers)

ต้องใช้ทรัพยากรการคำนวณและข้อมูลมหาศาล

เป็นระบบปิด (Closed-source) ขยายผลและทำซ้ำได้ยาก

เข้าถึงยาก, จำกัดการพัฒนาและตรวจสอบร่วมกัน

NotebookLM

จุดเปลี่ยนผ่านสำคัญ (The Paradigm Shift)

ทำไม General Coding Agent + MCP จึงเหนือกว่า Specialized Math Model?

Native Interface

Coding agents มีอินเทอร์เฟซที่เป็นธรรมชาติสำหรับจัดการงานวิศวกรรมการพิสูจน์ (Proof Engineering) ที่หลากหลายเกินกว่าแค่การแก้โจทย์

Zero-Training Upgrades

สามารถอัปเกรดขีดความสามารถได้ทันทีเพียงแค่เปลี่ยน Base Model (เช่น ไปใช้ Claude Opus 4.5) โดยไม่ต้องเสียเวลาเทรนใหม่

Plug-and-Play Extensibility

การใช้มาตรฐาน MCP (Model Context Protocol) ช่วยให้ระบบสามารถเลือกเรียกใช้เครื่องมือเฉพาะทางได้อย่างยืดหยุ่นและอัตโนมัติตามบริบท

NotebookLM

สถาปัตยกรรมของระบบ: ผู้ควบคุมการก่อสร้าง (The Master Builder)

Informal Prover
(การร่างบทพิสูจน์แบบไม่เป็นทางการ)

Discussion Partner
(การปรึกษาโมเดลภายนอก)

Claude Code
(อ้างอิงพลังการคิดจาก Claude Opus 4.5)

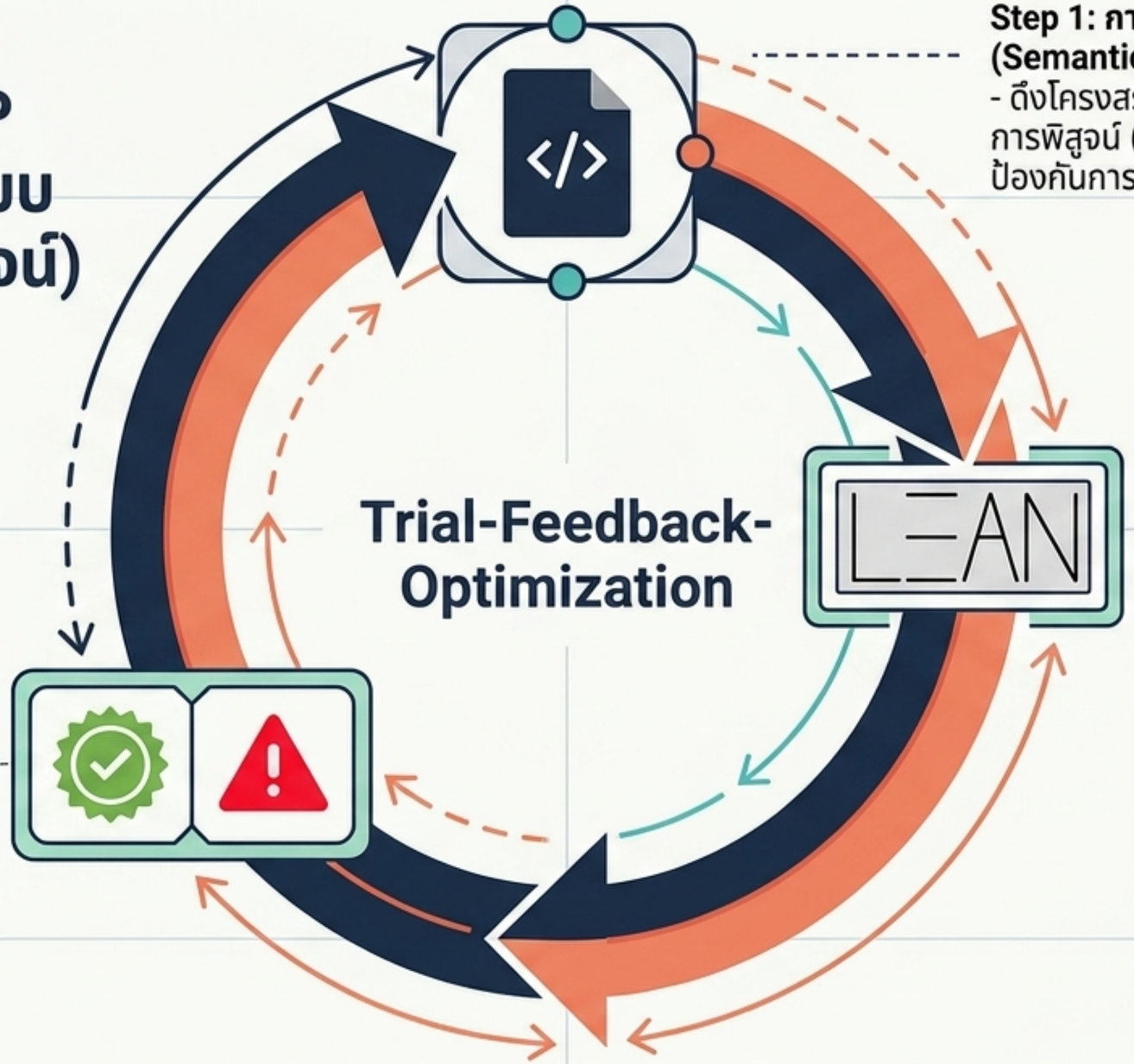
Lean-LSP-MCP
(การโต้ตอบกับคอมไพเลอร์)

LeanDex
(การดึงข้อมูลและค้นหาทฤษฎีบท)

โมเดลตัวกลางทำหน้าที่ประเมินสถานการณ์และเรียกใช้ 4 เครื่องมือนี้อย่างอิสระเพื่อแก้ปัญหาย่อยแต่ละจุด

NotebookLM

เครื่องมือที่ 1: Lean-LSP-MCP (สะพานเชื่อมสู่ระบบการจัดเตรียมพิสูจน์)



Step 1: การรับรู้และโต้ตอบ (Semantic Awareness)
- ดึงโครงสร้างไฟล์และเช็คเป้าหมายการพิสูจน์ (Proof State) ป้องกันการเดาแบบไร้ทิศทาง

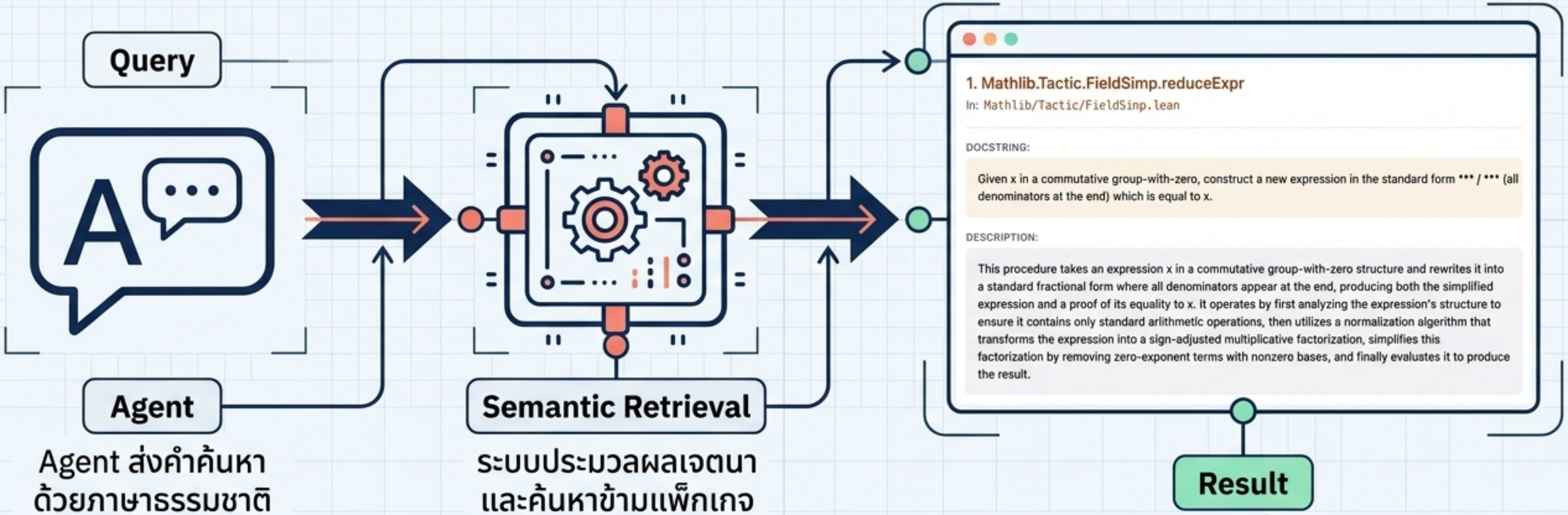
Step 2: การทำงานคู่ขนาน (Strategy Exploration)
- คอมไพล์โค้ดทดสอบแบบแยกส่วนผ่าน `lean_run_code` และใช้ `lean_multi_attempt` ประเมินกลยุทธ์หลายทางพร้อมกัน

Step 3: การแก้ไข (Closed-Loop Optimization)
- นำ Diagnostic Messages (ข้อผิดพลาด) จาก Lean มาปรับแต่งและวนลูปพิสูจน์ใหม่ทันที

NotebookLM

เครื่องมือที่ 2: LeanDex (ระบบฐานข้อมูลความหมายเชิงรุก)

Key Concept: ก้าวข้ามข้อจำกัดของการค้นหาแบบเดิม (เช่น Loogle หรือ Local Search) ที่บังคับรูปแบบไวยากรณ์ตายตัว



Agent ส่งคำค้นหา
ด้วยภาษาธรรมชาติ
(Natural Language)

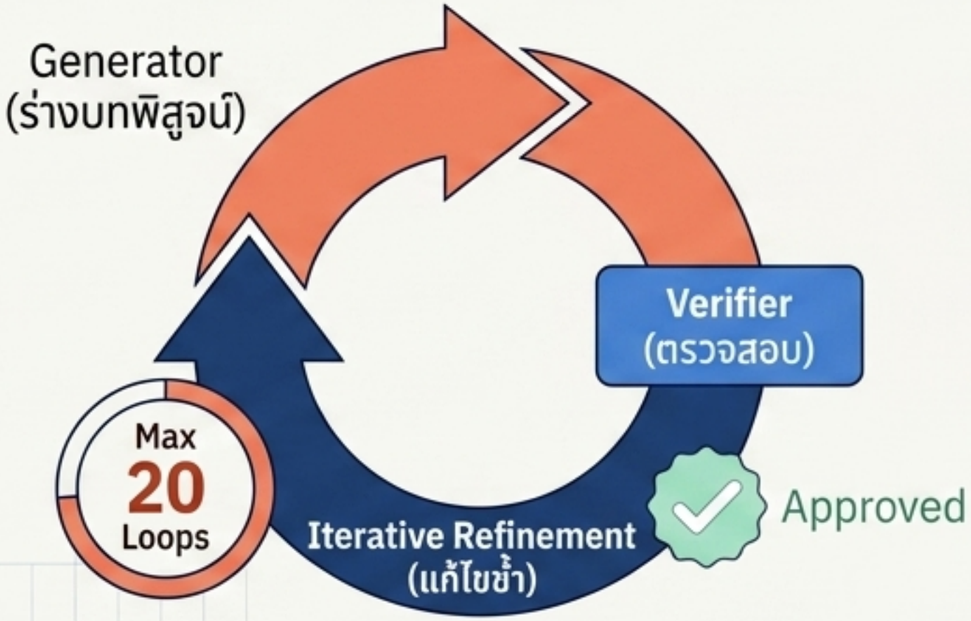
ระบบประมวลผลเจตนา
และค้นหาข้ามแพ็กเกจ
(เช่น Mathlib, FLT)
ใน Lean v4.26.0

ดึงนิยามและทฤษฎีบทที่ถูกต้องแม่นยำมาใช้
ปิดประตูการเกิดภาพหลอน (Hallucination)
ของ AI ในการอ้างอิงสูตรคณิตศาสตร์

NotebookLM

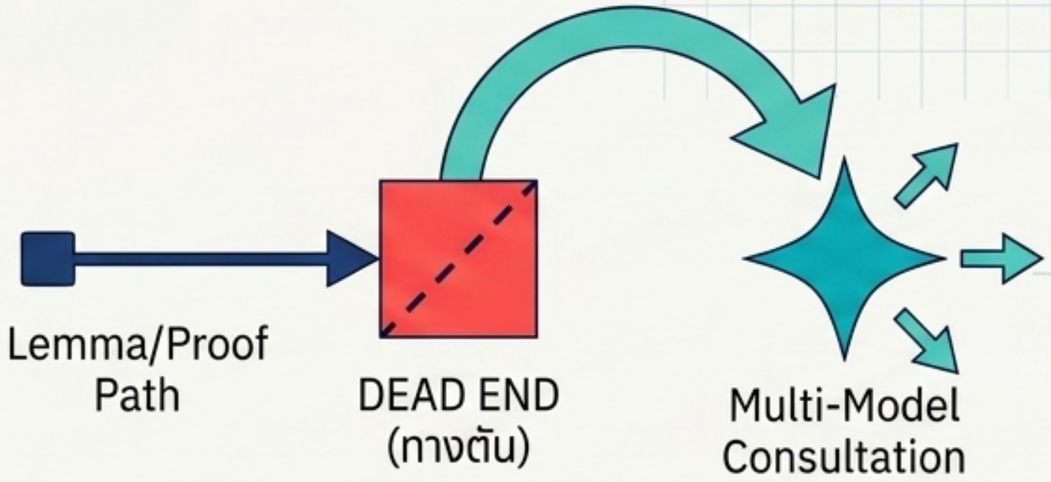
เครื่องมือที่ 3 & 4: การร่างโครงสร้างและการระดมสมอง (Informal Prover & Discussion Partner)

Informal Prover (Gemini IMO Agent)



- ระบบ Generator ร่างบทพิสูจน์ และส่งให้ Verifier ตรวจสอบ
- แก้ไขซ้ำ (Iterative Refinement) ตามคำติชม ทำซ้ำสูงสุด 20 รอบ
- ต้องผ่านการตรวจสอบอิสระ 3 ครั้งจึงจะได้รับการยอมรับ

Discussion Partner



- เมื่อเจอ ‘ทางตัน’ หรือความคลุมเครือใน Lemma ย่อย
- Claude Code สามารถเรียก AI ตัวอื่นมา ‘ปรึกษาหารือ’ (Multi-model collaboration)
- เพื่อหาทิศทางการแก้ปัญหาใหม่ ช่วยหลุดพ้นจากมุมมองเดิมๆ

NotebookLM

การทดสอบ Putnam 2025: ผลลัพธ์ระดับ State-of-the-Art (12/12)

Benchmark Dashboard

Numina-Lean-Agent	✓	(เทียบเท่าโมเดลระบบปิดที่ดีที่สุด)
AxiomProver	■	(Closed-source ensemble)
Seed-Prover 1.5	■	
Aristotle (Harmonic)	■	

ความเข้มงวดในการทดสอบ

ไม่มีสิทธิ์เข้าถึงอินเทอร์เน็ต (Zero online search)
และทำงานแบบลำดับขั้น (Strictly sequential)
ไม่มีการประมวลผลแบบขนาน (No parallelization)

Insight

ชนะกลยุทธ์ Independent Sampling ⇒ ✓
แบบขาดลอยด้วยกลยุทธ์ Iterative Refinement

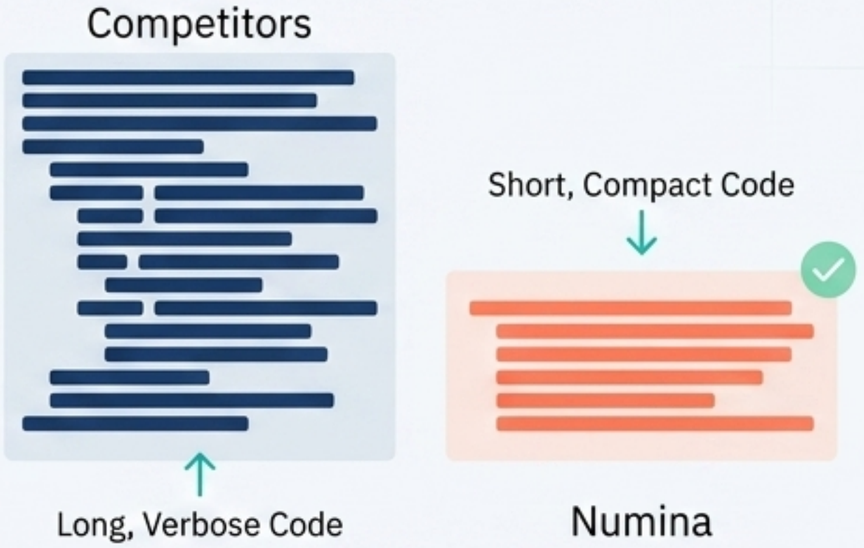
NotebookLM

ประสิทธิภาพและความสง่างาม (Efficiency & Elegance)

Time Efficiency (เวลา)

แม้ไม่มีการประมวลผลแบบขนาน (Parallelization) แต่ Numina ทำเวลาได้ดีกว่าคู่แข่งในหลายโจทย์

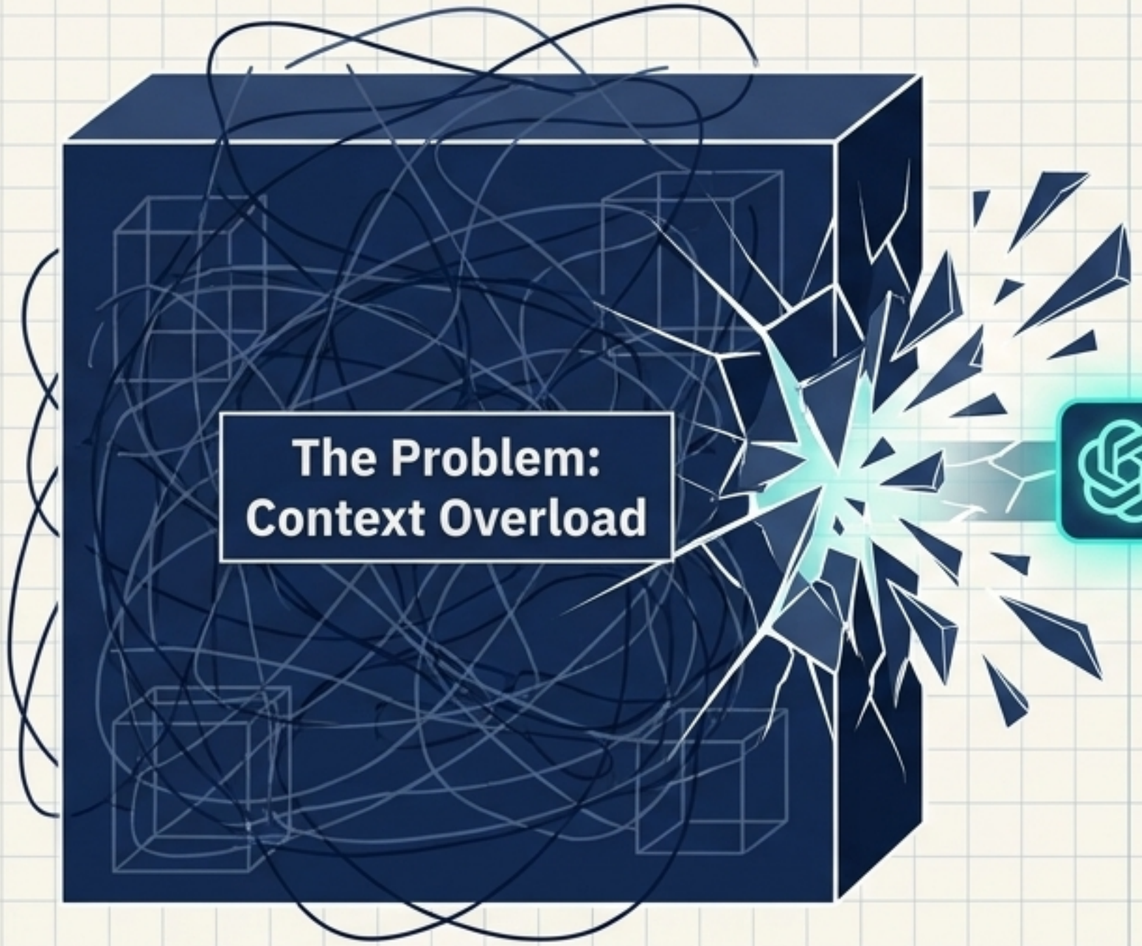
Code Elegance (ความยาวของโค้ด)



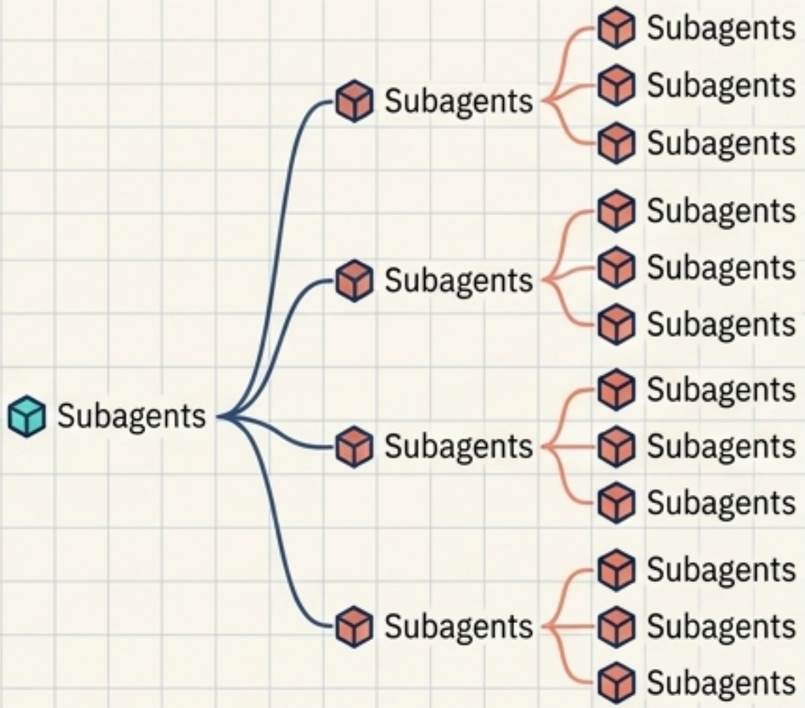
- Numina สังเคราะห์โค้ดที่ ‘สั้นและกระชับกว่า’ อย่างมีนัยสำคัญเมื่อเทียบกับ Axiom และ Seed-Prover 1.5 (เห็นได้ชัดในโจทย์ A3, B1 และ B5)
- โค้ดที่สั้นกว่าหมายถึงบทพิสูจน์ที่ตรงประเด็นและลดความซับซ้อนที่ไม่จำเป็น (Compact and efficient formalizations)

NotebookLM

การจัดการความซับซ้อนขั้นสุด: กรณีศึกษา Problem A5 (Subagents)



เมื่อเนื้อหายาวเกินไป โมเดลจะหลุดโฟกัสและละเลยคำสั่งสำคัญ
ส่งผลให้ติดแหง็กใน Lemma หลัก



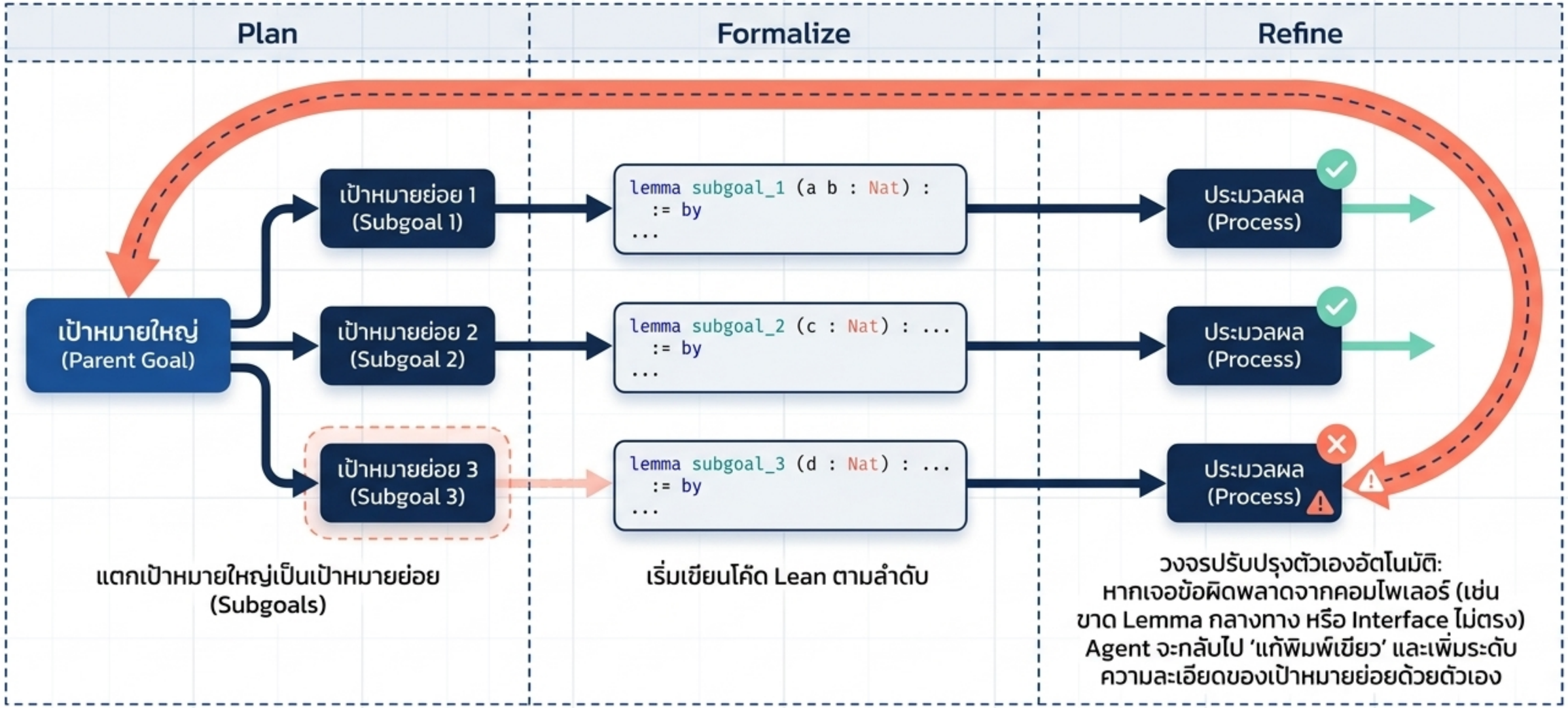
The Breakthrough: ระบบ Subagent Mechanism

- แยก Lemma สำคัญออกมาจากเป้าหมายหลัก
- เรียกใช้ GPT-5.2 เพื่อสร้างคำใบ้ (Informal Hint) เฉพาะเจาะจง
- ให้ Subagent ดำเนินการพิสูจน์ย่อยแบบแยกขาดอิสระ จนกว่าจะสำเร็จ
- นำผลลัพธ์กลับมารวมกัน ป้องกันอาการหลงทางในบริบทที่ลึกเกินไป

NotebookLM

กลยุทธ์พิมพ์เขียว (The Blueprint Strategy: Plan–Formalize–Refine)

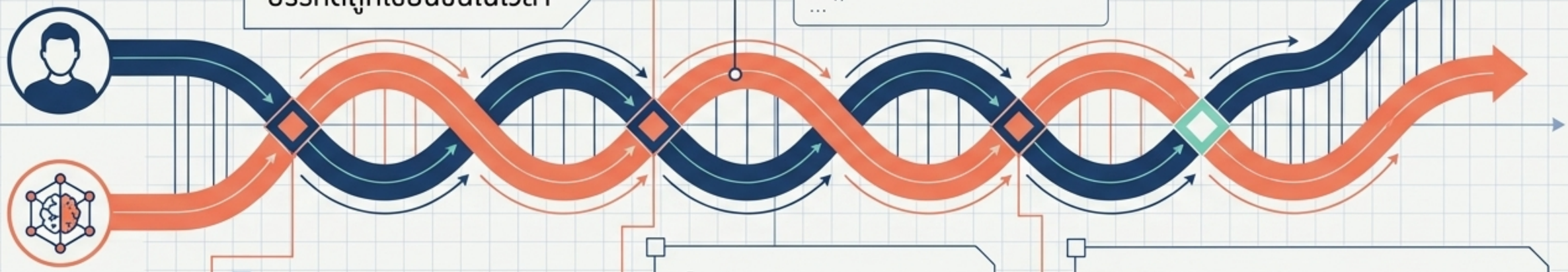
The Concept: การพิสูจน์ระยะยาวไม่สามารถทำแบบรวดเดียวจบ (One-shot) ต้องใช้แผนผัง (DAG) ที่กำหนดลำดับชั้นของ Lemma



NotebookLM

การวิจัยร่วมระหว่างมนุษย์และ AI (Human-AI Vibe Proving)

Case Study: The Brascamp-Lieb Theorem (ทำงานร่วมกับนักคณิตศาสตร์ตัวจริง)



The Achievement
โค้ด Lean มากกว่า **8,000** บรรทัดถูกเขียนขึ้นในเวลา

```
definition BrascampLieb (A : Matrix R) ... {
    podlate Brascamplieb (A
    set A = R.ienvz,
    formal definition = {
        A:atrtx: 8,
        cn-ix.mots-$i/bii.Ionom) 7)
        cow BrascampLieb[[: Matrix 8]),
    };
...
```

The Achievement
โค้ด Lean มากกว่า **8,000** บรรทัดถูกเขียนขึ้นในเวลาไม่ถึง **2 สัปดาห์**

Autonomy
Agent สร้างนิยามและทฤษฎีบทใหม่กว่า **70 รายการ**ด้วยตัวเองด้วยตัวเอง

Self-Correction
เมื่อพบว่า Formal Statement เริ่มต้นตั้งมาผิด Agent สามารถตรวจจับจุดผิดและ '**เสนอการแก้ไขสมมติฐาน (Statement)**' ได้เอง ซึ่งไม่เคยมีมาก่อนในโมเดลเฉพาะทาง ✅

NotebookLM

ข้อจำกัดในปัจจุบัน และรุ่งอรุณแห่งคณิตศาสตร์แบบเปิด

ข้อจำกัดที่ต้องพัฒนาต่อ (Limitations)

- **Verbosity**: โค้ดที่สร้างในระดับโครงสร้างใหญ่บางครั้งยังเยิ่นเย้อ ขาดความเป็น Idiomatic Mathlib (ความสวยงามเชิงโครงสร้างแบบมนุษย์)
- **Type-Level Constraints**: ความท้าทายในการแปลง Type (เช่น Real vs NNReal) ซึ่ง AI มักมองข้ามเนื่องจากความคุ้นเคยในคณิตศาสตร์แบบไม่เป็นทางการ

อนาคต (The Future)

Numina-Lean-Agent พิสูจน์แล้วว่าการเชื่อมต่อโมเดลทั่วไปเข้ากับเครื่องมือเฉพาะทาง (General Agent + MCP) คือก้าวต่อไปที่ทรงพลัง และที่สำคัญที่สุดคือ เป็นระบบเปิด (Open-source) ที่พร้อมให้คอมมูนิตี้นำไปต่อยอดได้อย่างไร้ขีดจำกัด

NotebookLM