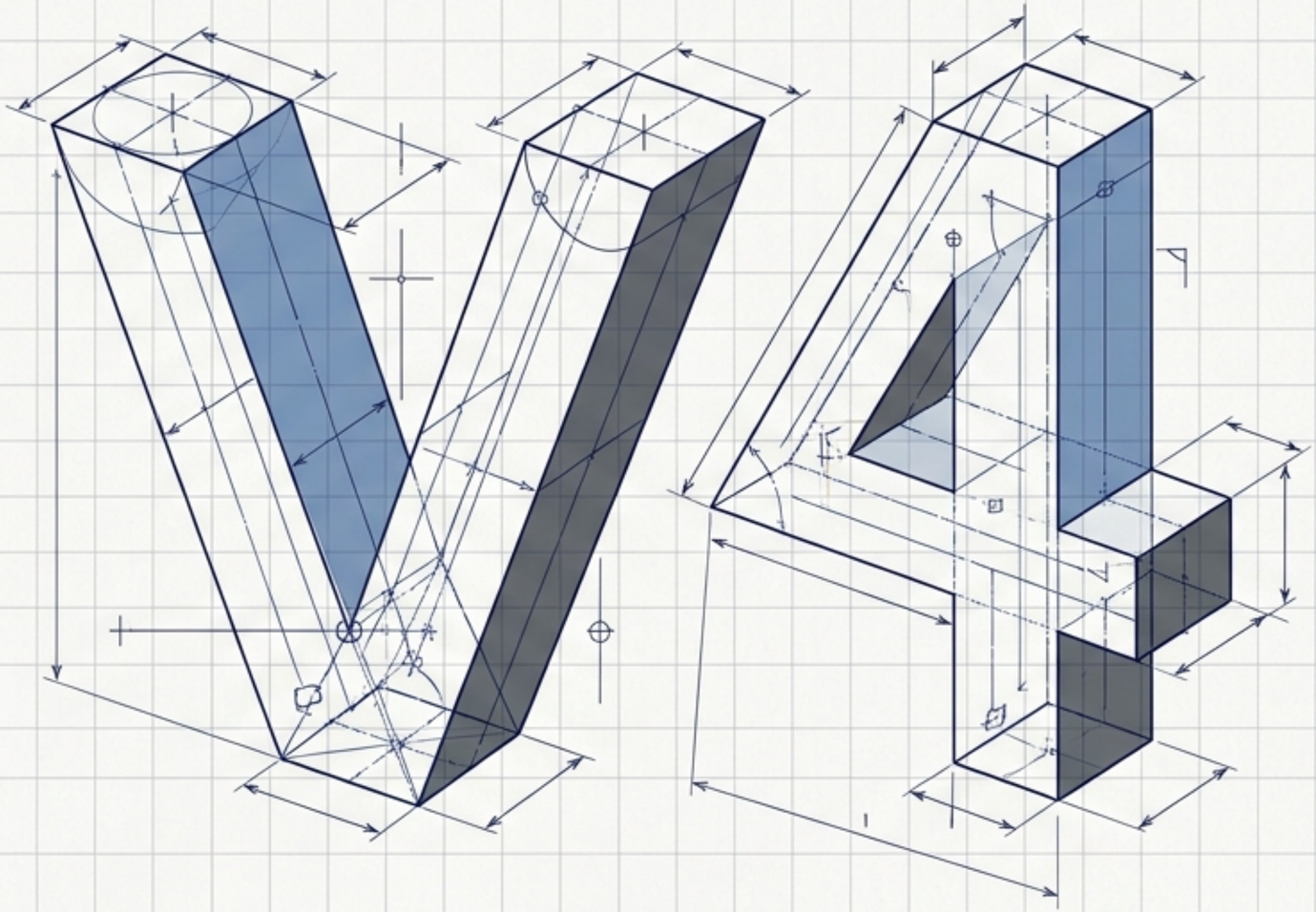
ถอดรหัสสถาปัตยกรรมแห่งอนาคตของ DeepSeek V4

กำหนดการเปิดตัว: ช่วงเทศกาลตรุษจีนปี 2026 (กลางเดือนกุมภาพันธ์)

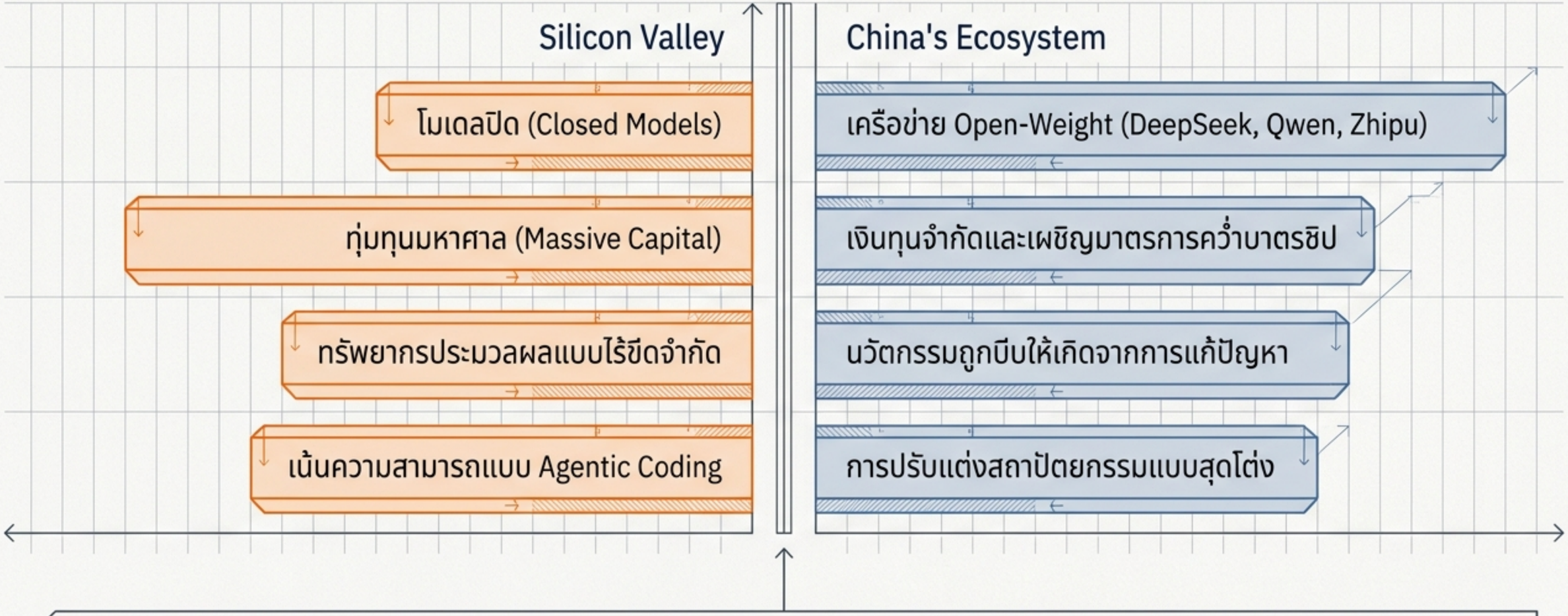
โปรเจกต์ลับภายใต้รหัส: MODEL1



นี่ไม่ใช่เพียงการอัปเดตโมเดล แต่คือการรื้อโครงสร้างสถาปัตยกรรมใหม่ทั้งหมดที่จะกำหนดทิศทางของ Open-Source AI ทั่วโลกในปี 2026 และ 2027

NotebookLM

ขั้วตรงข้ามของการพัฒนา AI โลก



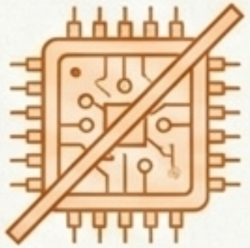
ห้องปฏิบัติการ AI ในจีนกำลังสร้างนวัตกรรมในระดับแนวหน้าด้วยทรัพยากรทางการเงินเพียงเศษเสี้ยวของคู่แข่งในสหรัฐฯ

NotebookLM

ข้อจำกัดทางฮาร์ดแวร์คือตัวเร่งปฏิกิริยาแห่งนวัตกรรม

The Bottleneck

ความพยายามฝึกสอนโมเดล R2 บนชิป Huawei ล้มเหลวจากปัญหาความเสถียรและความเร็วที่ช้า

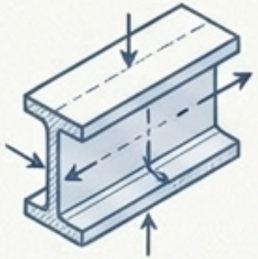


The Pivot

ต้องถอยกลับมาใช้ฮาร์ดแวร์ Nvidia สำหรับ Training และลดบทบาท Huawei เหลือเพียง Inference

The Forced Innovation

ละทิ้งการพึ่งพาพลังประมวลผลดิบ และหันมาโฟกัสที่ประสิทธิภาพของสถาปัตยกรรม (Architectural Efficiency)

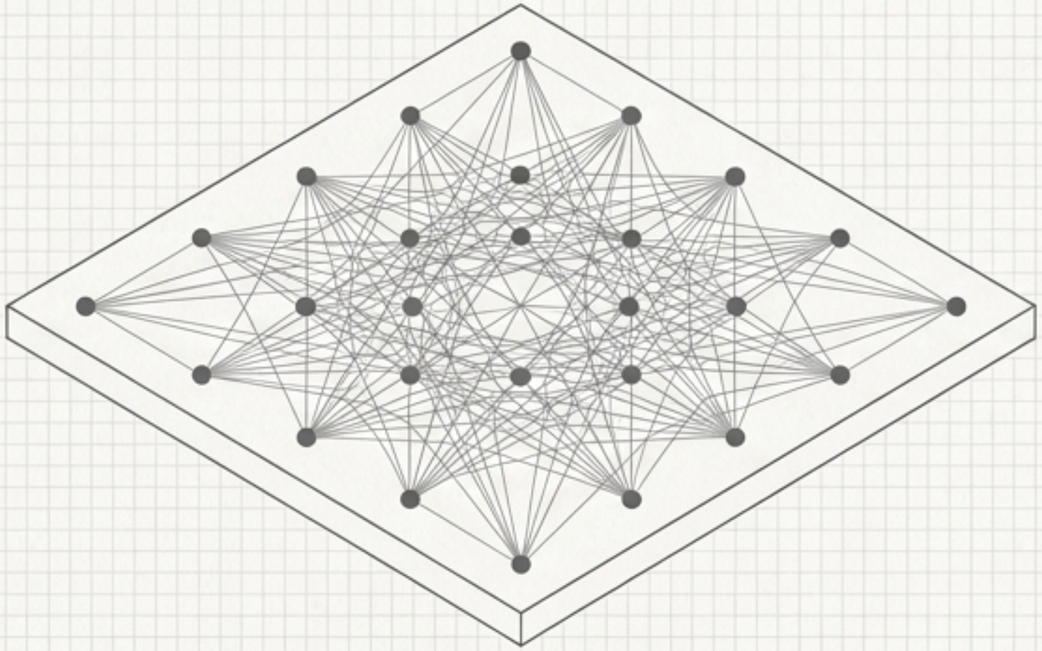


เมื่อไม่สามารถเอาชนะด้วยจำนวนชิป ทางรอดเดียวคือการเปลี่ยนวิธีที่โมเดลใช้คำนวณข้อมูล

NotebookLM

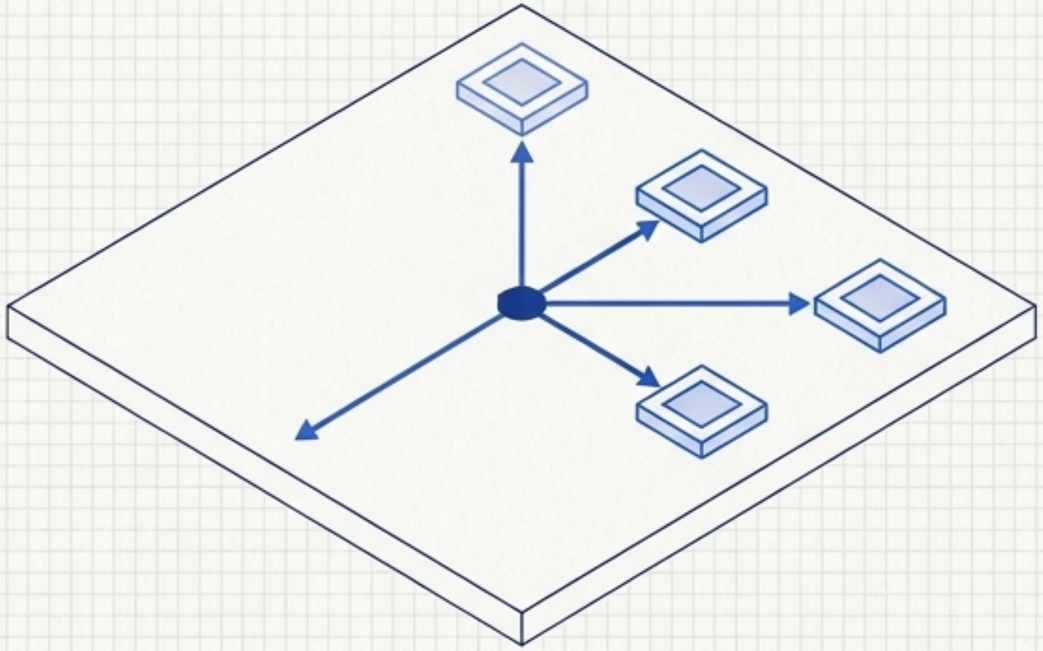
ปรัชญาแกนหลักของ DeepSeek: Sparsity

Dense Network (ขีดจำกัดเดิม)



ระบบส่ง Token ข้อมูลผ่านเครือข่ายทั้งหมด
ทำให้สูญเสียพลังงานและเวลาไปกับการคำนวณที่ไม่จำเป็น

Sparsity (การจัดการอย่างชาญฉลาด)



ระบบเลือกประมวลผลโดยส่งข้อมูลไปยังกลุ่มพารามิเตอร์หรือ
ผู้เชี่ยวชาญ (Experts) ที่เกี่ยวข้องที่สุดเท่านั้น

The Bottom Line: การขยายความฉลาดไม่ได้มาจากการเพิ่มฮาร์ดแวร์ แต่มาจากการรู้วิธี 'เพิกเฉย' ต่อข้อมูลที่ไม่จำเป็น

NotebookLM

ลำดับขั้นการแก้ปัญหาคอขวด (The Evolution of Sparsity)

Network Sparsity (V1 - V3.0)	Attention Sparsity (V2 - V3.2)	The Frontier (V4 - Q1 2026)
กลไก: การจัดการเครือข่ายผ่านสถาปัตยกรรม MoE (Mixture of Experts) **การพัฒนา:** ขยายจาก 64 Experts (V1) สู่ 256 Experts (V3) **สถานะ:** ปรับจูนสำเร็จแล้ว (Optimized)	**กลไก:** การบีบอัดและลดภาระการประมวลผล Context Window **การพัฒนา:** เริ่มจาก MLA (บีบอัดข้อมูล) สู่ NSA (Native Sparse Attention) **สถานะ:** ปรับจูนสำเร็จแล้ว (Optimized)	**กลไก:** การจัดการหน่วยความจำขั้นสูงและการเชื่อมต่อชั้นลึก **เทคโนโลยีหลัก:** Engram (แยกความจำ), DSA (ดึงเฉพาะจุด), mHC (เสถียรภาพ) **สถานะ:** กำลังจะเกิดขึ้น (Pending Release)

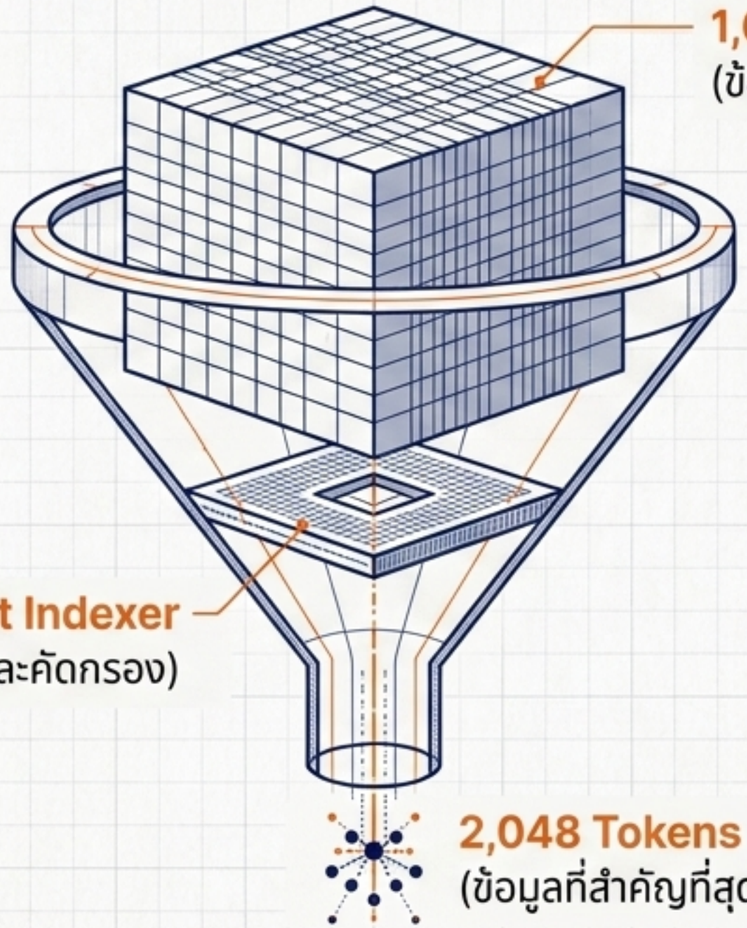
การแก้ปัญหาอย่างเป็นระบบ (Systematic Progression of Solving Bottlenecks)

NotebookLM

นวัตกรรมที่ 1: การตระหนักรู้ข้อมูลระดับล้านด้วย DeepSeek Sparse Attention (DSA)

The Mechanism:

สถาปัตยกรรม DSA ข้ามข้อจำกัดการอ่านข้อมูลแบบเดิมด้วยการใช้ Indexer ขนาดเล็กที่จำลองรูปแบบความสนใจที่จำลองรูปแบบความสนใจเพื่อคัดกรองข้อมูล



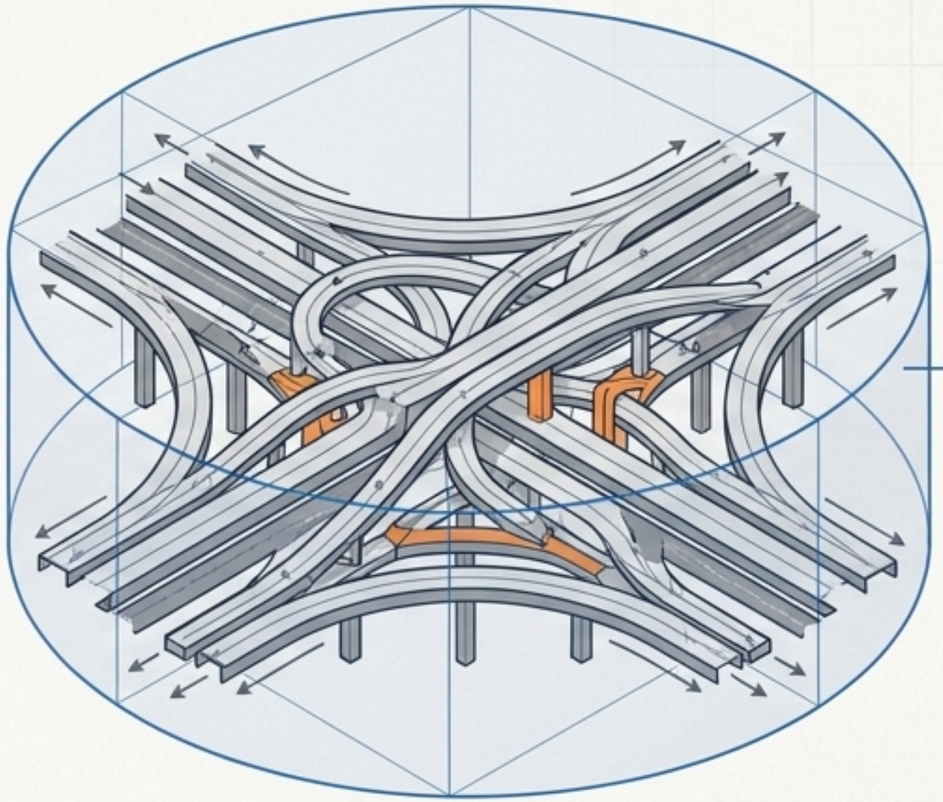
The Business Value:

เปิดประตูสู่ความสามารถในการวิเคราะห์โค้ดคอมพิวเตอร์แบบข้ามไฟล์ (Multi-file debugging) ระดับหลายพันไฟล์ได้ในคราวเดียว โดยไม่กินพลังประมวลผลจนเกินขีดจำกัด

NotebookLM

นวัตกรรมที่ 2: เสถียรภาพระดับโครงสร้างชั้นลึกด้วย mHC

- The Concept: การสร้างเส้นทางด่วนพิเศษข้ามชั้นเลเยอร์ของโมเดล (Hyper-connections) เพื่อให้ข้อมูลไหลเวียนได้อย่างอิสระ
- The Problem: ป้องกันสภาวะ Gradient Explosion (ข้อมูลล้น) หรือ Vanishing Gradients (ข้อมูลสูญหาย) ในโครงข่ายแบบลึก



How it Works: บังคับใช้สมการคณิตศาสตร์แบบ Manifold เพื่อรักษาสมดุลของกระแสข้อมูลแม้การเชื่อมต่อจะซับซ้อน ช่วยลดการใช้พลังงานและรักษาสภาพการฝึกสอน (Training) ให้เสถียรสูงสุด

NotebookLM

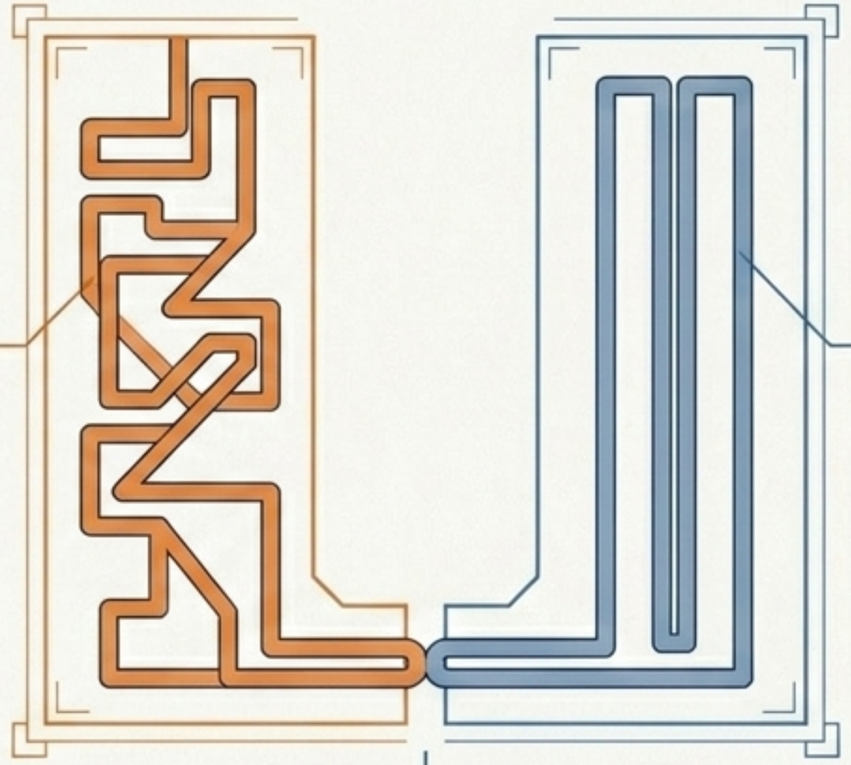
นวัตกรรมที่ 3: ระบบแยกส่วนการรู้คิดด้วย Engram (The Cognitive Split)

สถาปัตยกรรมเดิมบังคับให้ GPU ต้องประมวลผลใหม่ทุกครั้ง แม้จะเป็นข้อเท็จจริงตายตัว

Dynamic Neural Computation

ชิป GPU

สงวนไว้สำหรับการให้เหตุผลที่ซับซ้อน, การสังเคราะห์สิ่งใหม่, และงานที่ต้องการบริบทเฉพาะ



Static Memory Lookup

CPU Host Memory

ดึงข้อเท็จจริงตายตัวระดับ 100 Billion+ Parameters ด้วยกลไก O(1) Constant Time โดยกินทรัพยากรส่วนเพิ่มต่ำกว่า 3%

The Breakthrough: รักษาโรค ‘ความจำเสื่อม’ ของ LLM เมื่อต้องรับบริบทอยาวๆ และลดการกินพื้นที่ GPU Memory ลงได้อย่างมหาศาล

NotebookLM

ร่องรอยสถาปัตยกรรมลับจากโค้ดรหัส MODEL1

```
[dim_config: 512, previous: 576]
```

Evidence 1: การเปลี่ยนผ่านสู่ 512-Dim

ออกแบบมาเพื่อดึงประสิทธิภาพสูงสุดจากสถาปัตยกรรม NVIDIA Blackwell (SM100)

```
[module: FlashMLA,
feature: VVPA_enable]
```

Evidence 2: กลไก VVPA (Value Vector Position Awareness)

รักษารายละเอียดความสัมพันธ์เชิงพื้นที่ (Spatial info) ป้องกันไม่ให้โมเดลสูญเสียรายละเอียดเมื่อเจอกับ Context ที่ยาวระดับอนันต์

```
[cache_storage: tiered,
dtype: FP8 + bfloat16]
```

Evidence 3: สถาปัตยกรรมหน่วยความจำแบบแบ่งชั้น (Tiered Storage)

ลดการใช้พื้นที่ GPU Memory ลงถึง 40% ทำให้มีต้นทุนการรันโมเดลที่ถูกกว่าแบบก้าวกระโดด

NotebookLM

ข้อมูลทางเทคนิคที่คาดการณ์ของสถาปัตยกรรม DeepSeek V4

1,000,000+

Tokens Context Window

รองรับการอ่านและวิเคราะห์ฐานข้อมูลโปรแกรมมิ่ง (Codebase) ทั้งระบบได้ในรวดเดียว

O(1)

Conditional Memory (Engram)

เรียกข้อมูลข้อเท็จจริงตายตัวได้ทันทีด้วยความเร็วคงที่ โดยไม่เปลืองรอบคำนวณของ GPU

-40%

Operational Cost Efficiency

ประหยัดพื้นที่หน่วยความจำ GPU จากการออกแบบสถาปัตยกรรม Tiered Storage

SM100

Hardware Optimization

รีดประสิทธิภาพขีดสุดจากชิป NVIDIA Blackwell ควบคู่กับการใช้ CPU เพื่อลดคอขวด

Key Takeaway: V4 ไม่ใช่แค่การอัปเดตเวอร์ชั่น แต่คือสถาปัตยกรรมใหม่หมดจด ที่ออกแบบมาเพื่อการเขียนโค้ดและการวิเคราะห์บริบทขนาดยักษ์โดยเฉพาะ

NotebookLM

รัศมีแรงกระแทกจากปรากฏการณ์ DeepSeek Effect

The Global Convergence

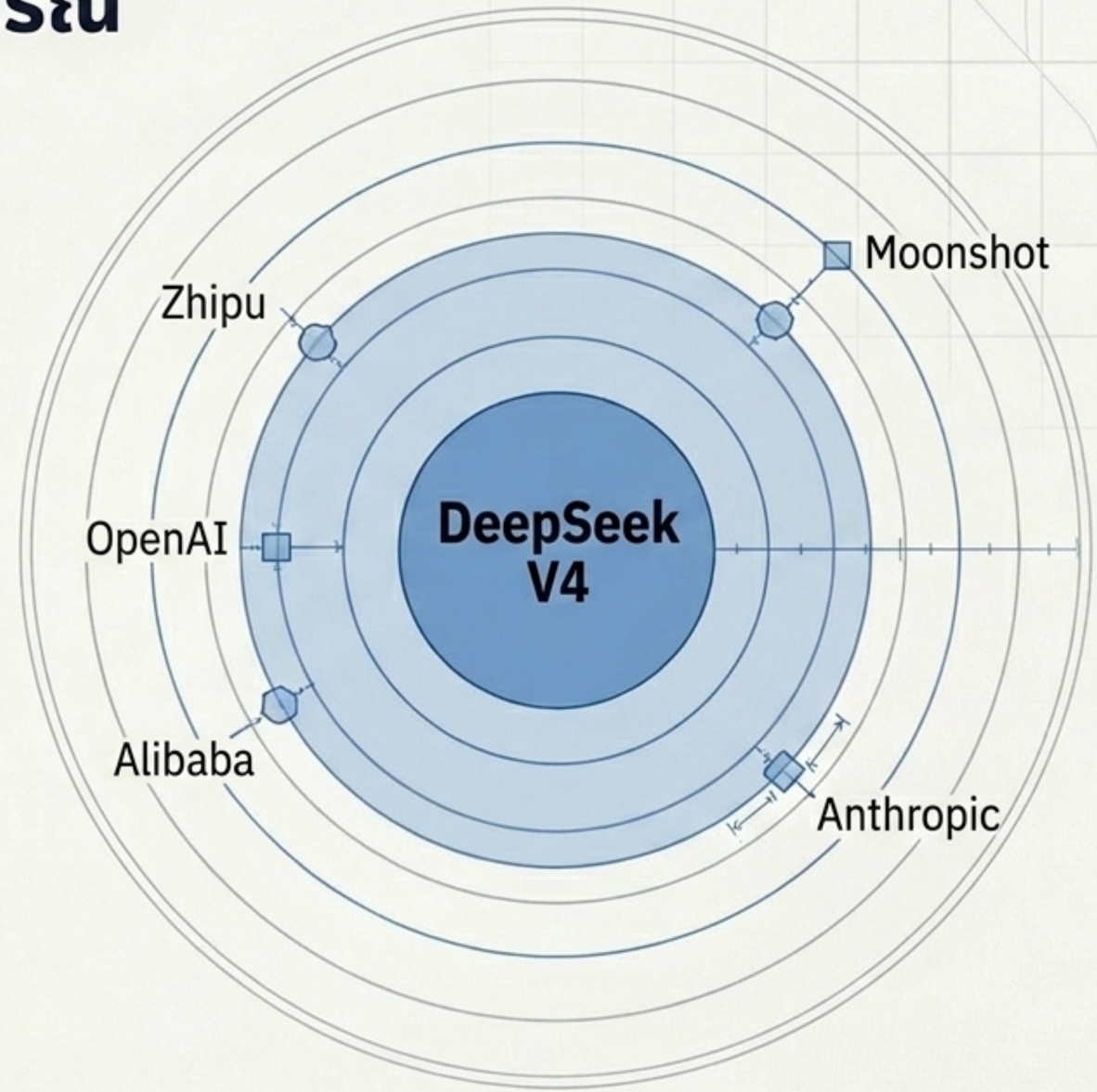
ช่องว่างความสามารถระหว่างโมเดล Open-source ของจีนและ Proprietary Models จากสหรัฐฯ กำลังถูกบีบให้แคบลงอย่างรวดเร็ว

Ecosystem Activation

ปรากฏการณ์นี้กระตุ้นให้บริษัท AI ในจีนเร่งพัฒนาระบบนิเวศ LLM ของตนเองในระดับแนวหน้าอย่างดุเดือด

The 2026 Battlefield

ปี 2026 จะไม่ใช่แค่สงครามของการทุ่มเม็ดเงินซื้อชิปประมวลผล แต่จะเป็นสมรภูมิของการปรับแต่งสถาปัตยกรรม (Extreme Optimization) ปะทะกับอำนาจทุน (Capital)



NotebookLM

“ข้อจำกัดทางทรัพยากรไม่ได้หยุดยั้งนวัตกรรม แต่มันบังคับให้ผู้พัฒนาต้องสร้างสถาปัตยกรรมแห่งความชาญฉลาดที่แท้จริง”

โมเมนตัมของ Open-Source AI จากจีน ขับเคลื่อนด้วยแนวคิด Sparsity
ภายใต้แรงกดดันจากข้อจำกัดทางฮาร์ดแวร์ กำลังเขียนกฎเกณฑ์และมาตรฐานใหม่ให้กับโลกวิศวกรรม AI ทั้งระบบ

NotebookLM