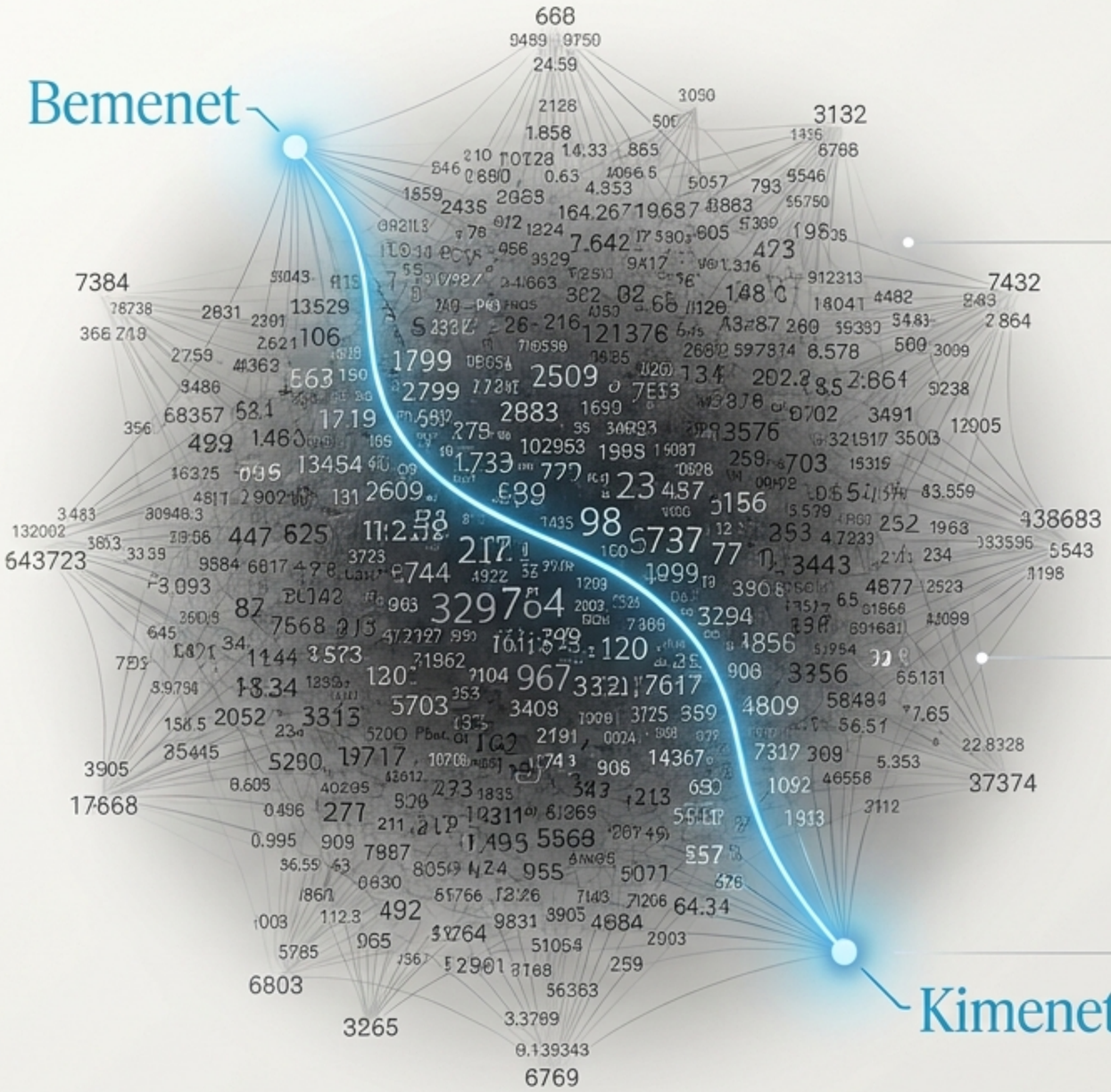
A mesterséges intelligencia fekete dobozának felnyitása

Hogyan képesek a csúcsmodellek közvetlenül olvasni, javítani és újraírni más neurális hálózatok nyers súlyait.



Gemini Notebook



A Jelenlegi Paradigma: Intelligencia a sötétben

A ‘Súlyok’ (Weights): A modern AI modellek lényegében több millió vagy milliárd egyszerű számból állnak. Ezek egymáshoz való viszonya határozza meg, hogyan dolgozza fel a gép az információt.

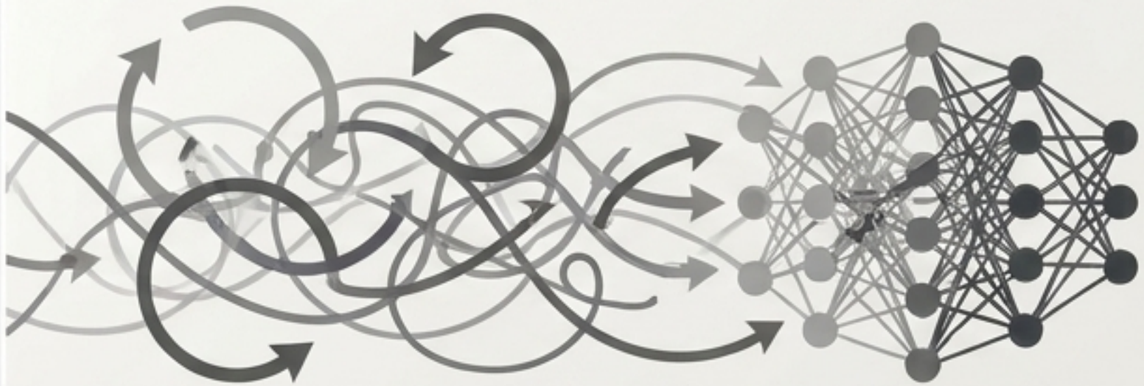
A Vak Tréning (Gradiens Ereszkedés): A modelleket ma ’próba-szerencse’ alapon tanítják. Az algoritmus vakon, újra és újra próbálkozik a hibák csökkentésével.

A Fekete Doboz: Pontosan tudjuk, mi a bemenet és a kimenet, de hogy mi történik a nyers számok tengerében a kettő között, az eddig rejtély volt.

Gemini Notebook

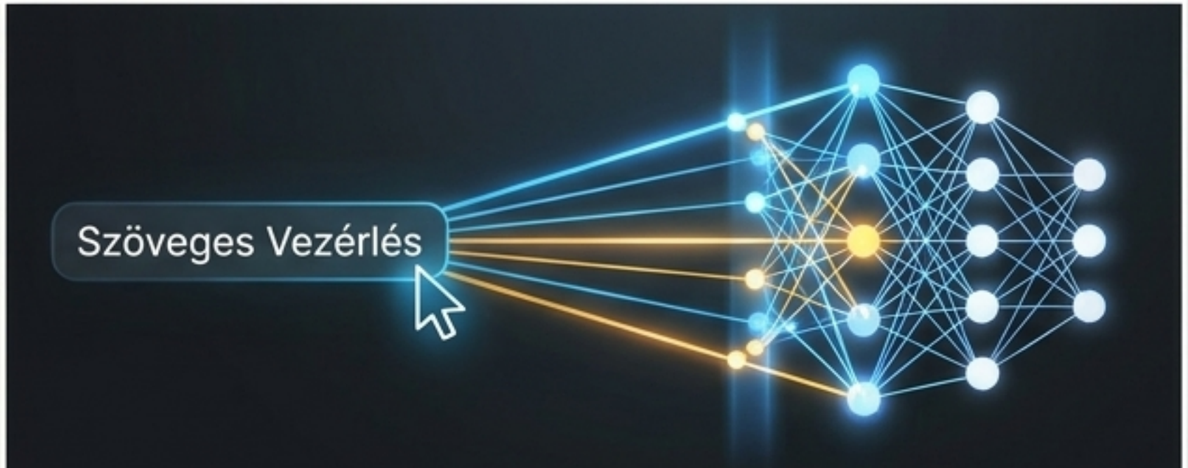
A fordulópont: Súlytér-programozás

Gradiens Ereszkedés (A múlt)



- Próba-szerencse alapú optimalizálás.
- Hatalmas adatmennyiséget és időt igényel.
- A modell egy vak algoritmuson keresztül tanul.

Súlytér-programozás (A jövő)



- Egy LLM (pl. az Astra) közvetlenül olvassa és írja a hálózat nyers súlyait.
- Nincs szükség hagyományos optimalizálóra vagy tréning adatokra.
- Sebészi pontosságú, azonnali, természetes nyelven (szövegesen) vezérelt beavatkozás.

Gemini Notebook

A megértés három szintje



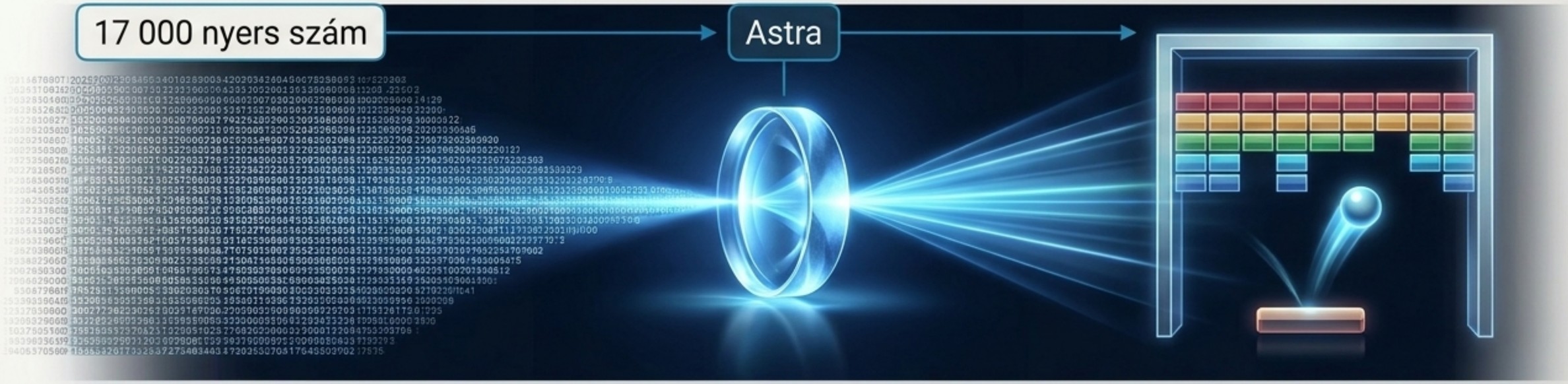
SZINT 1: Olvasás és Értelmezés (AKTÍV) – A modell képes értelmezni a nyers számokat és dedukálni a funkciót.

SZINT 2: Diagnosztika és Javítás – A modell megtalálja a logikai hibákat és sebészi pontossággal módosítja a hálózatot.

SZINT 3: Teremtés a semmiből – A modell új, működő neurális hálózatokat ír szöveges parancsok alapján.

Gemini Notebook

Amikor a számok elárulják a játékot



A Kísérlet:
Négy azonos felépítésű, de különböző játékra (Breakout, Pong, Pac-Man, Freeway) tanított modell nyers súlyait mutatták meg az Astrának.

A Bemenet:
Csupán 17 000 nyers szám (súly), opciók és tippek nélkül.

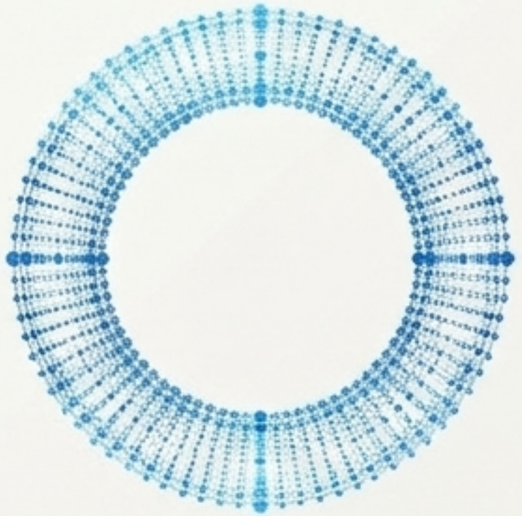
Az Eredmény:
Az Astra mind a négy játékot hibátlanul azonosította.

A Logika (Breakout):
Mindössze 846 token használatával (99%-os magabiztossággal) dedukálta, hogy az ‘alsó sorban lévő ütő’, a ‘mozgó labda’ és a ‘vízszintes mozgás’ a téglatörő játékra utal.

Gemini Notebook

A geometria megértése a kimenet kiszámítása helyett

Az Astra értelmezése
(Súlyok anatómiája)





A modell kimenete
(Generált valóság)

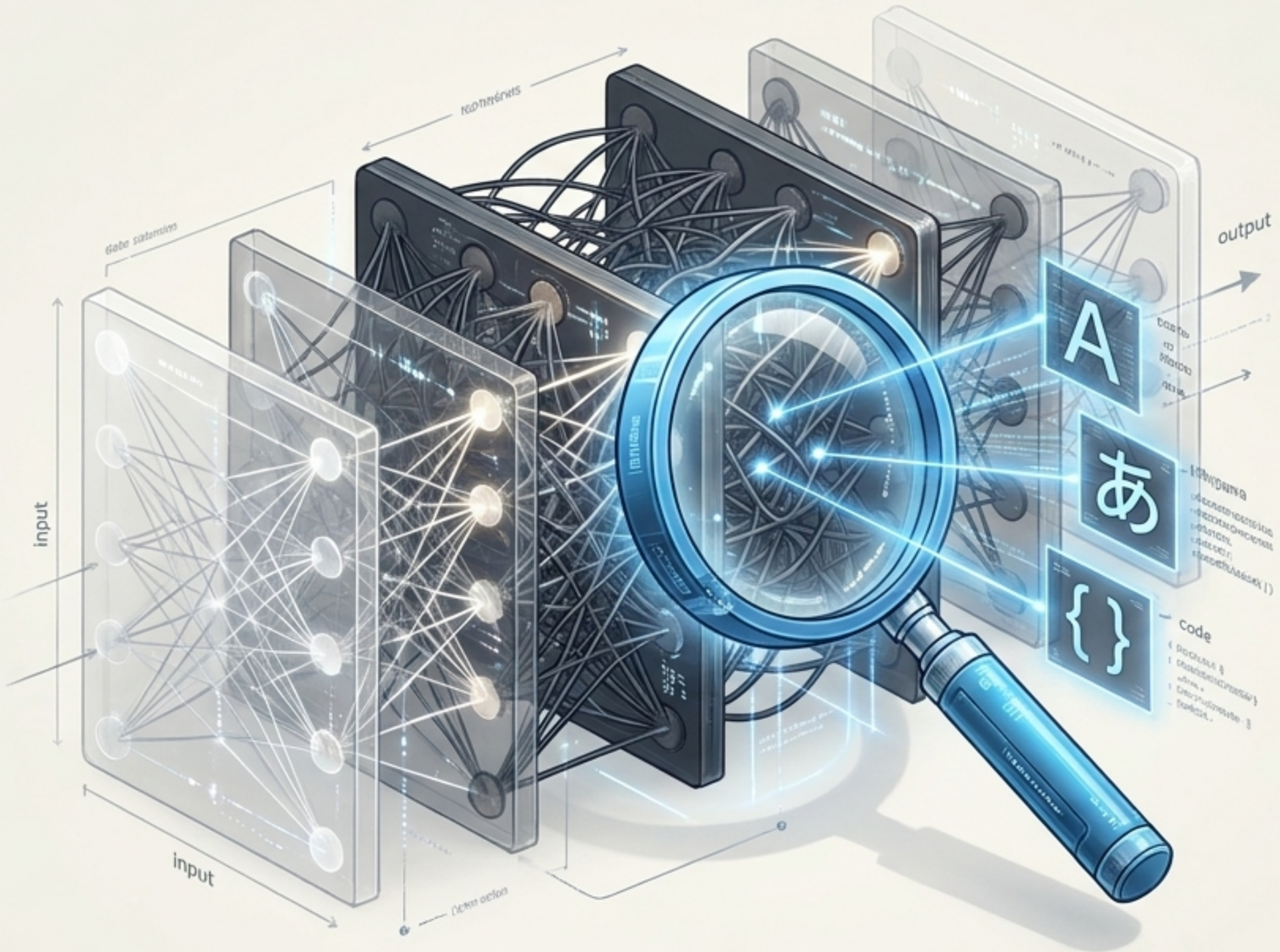
A Kísérlet:
Három képgeneráló (diffúziós) modell, amelyeket kör, négyzet és háromszög rajzolására tanítottak. Vajon felismeri az Astra az alakzatokat a súlyokból?

A Felfedezés:
Az Astra sikeresen azonosította mindhárom alakzatot a görbületek és oldalsávok alapján.

A Lényeg:
Az Astra az alakzatok kontúrját írta le (így rajzolódnak ki a súlyok struktúrájában), miközben a modell a valóságban kitöltött alakzatokat generál.
Ez bizonyítja: az Astra ténylegesen a hálózat anatómiáját olvassa.

Gemini Notebook

Rejtett nyelvek a mélyrétegekben



Az első és utolsó rétegek árulkodóak lehetnek. Mi történik, ha az Astra csak a legmélyebb, közbülső rétegeket látja?

A kutatók három modellt tanítottak: angol, japán és Python nyelvekre.

Az Eredmény: Az Astra a kezdő és záró kontextus hiányában is tökéletesen azonosította a modellek nyelvi specifikációját. Felismerte azokat a mikroszerkezeteket, amelyek a latin ábécéért vagy épp a kódszintaxisért felelősek.

Gemini Notebook

Szintbe lépés: Beavatkozás a rendszerbe



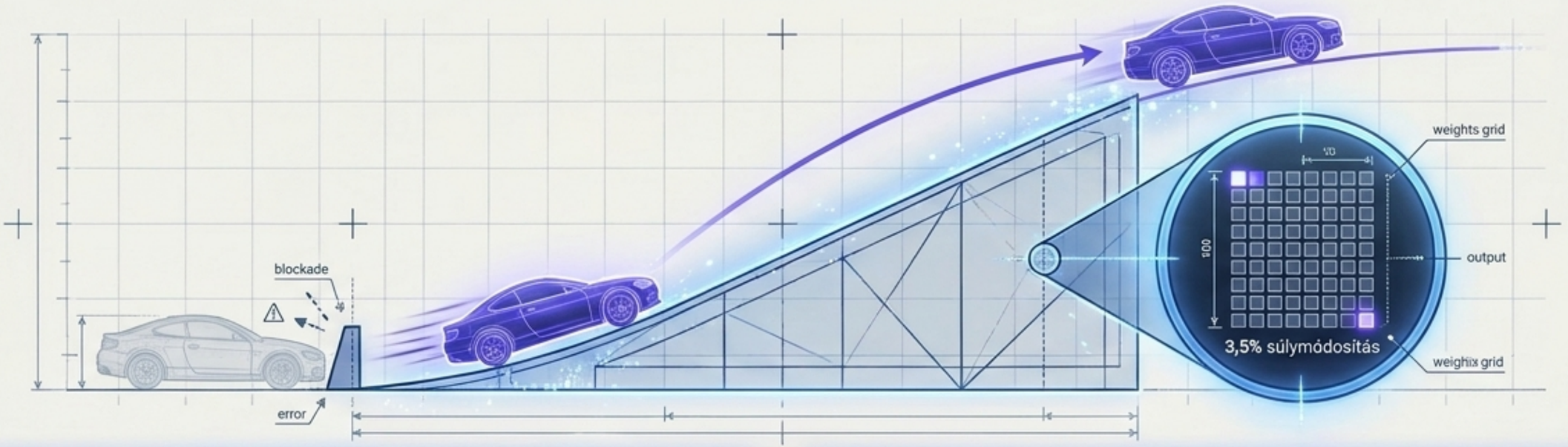
SZINT 1: Olvasás és Értelmezés (Befejezve)

SZINT 2: Diagnosztika és Javítás (**AKTÍV**) – A puszta megfigyeléstől elmozdulunk az aktív, sebészi beavatkozás felé.

SZINT 3: Teremtés a semmiből

Gemini Notebook

Sebészi beavatkozás a fizikai motorban



A Hiba: Egy fizikai szimulációban egy jármű folyamatosan elakadt a rámpa alján, mert nem volt elég lendülete.

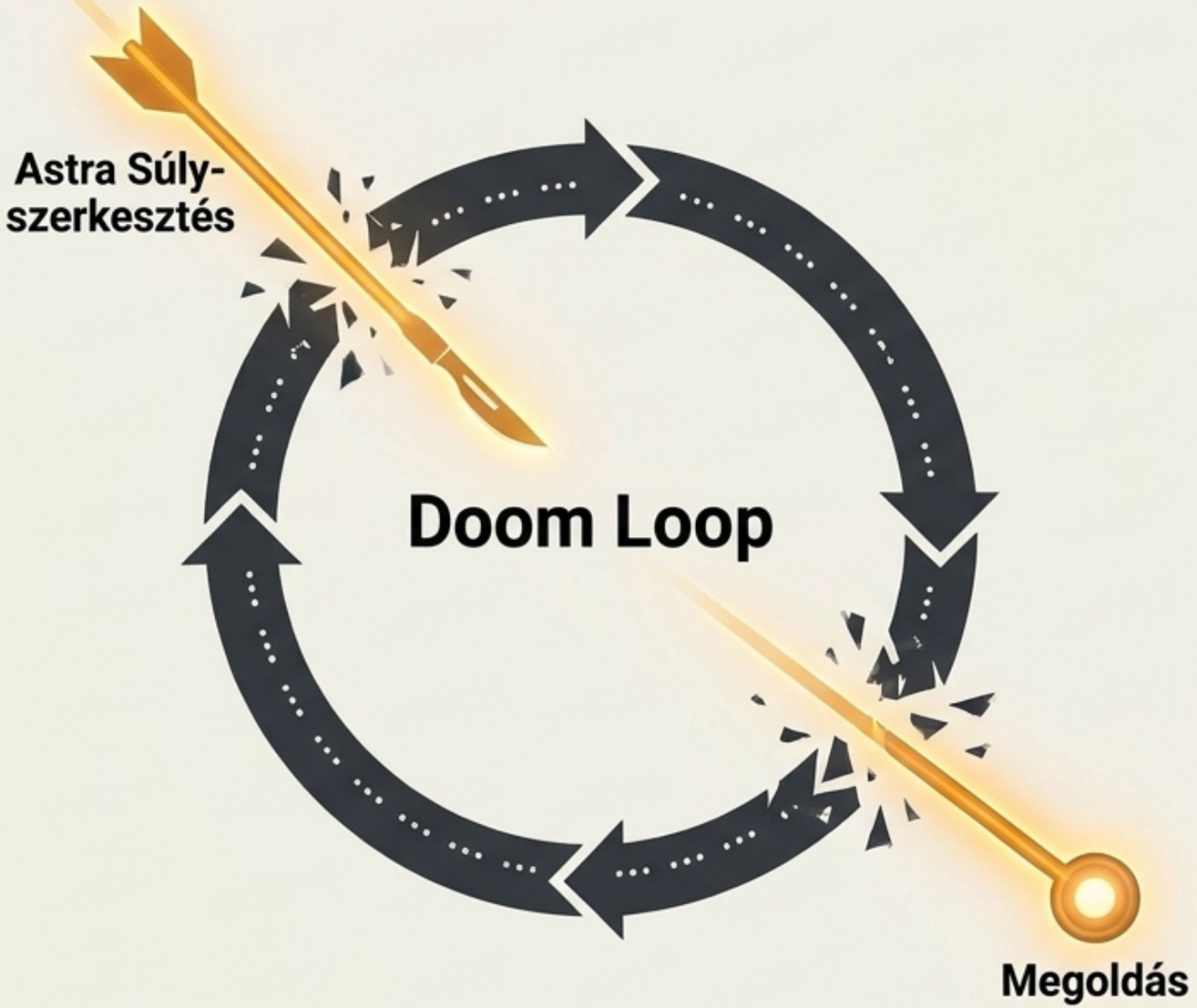
A Beavatkozás: Az Astra megkapta a nyers súlyokat és a hibaleírást, majd eszközök nélkül újraírta a kódokat.

Az Eredmény: Az Astra felismerte, hogy a "hátramenet" sosem aktiválódik. A súlyok mindössze 3,5%-ának módosításával átírta a logikát.

A Siker: A jármű először tolat, lendületet vesz, és 1024-ből 1024 alkalommal sikeresen átugrat a rámpán.

Gemini Notebook

A “Végzetes Ciklus” (Doom Loop) megtörése



A Probléma: Kis modellek (pl. Qwen 3.5B, 800 millió paraméter) gyakran esnek “végzetes ciklusba” matekfeladatoknál – végtelenül ismétlik önmagukat.

A Diagnózis: Az Astra azonosította a mechanikai hibát a nyers súlyokban, és 50 előrehaladási lépés (*forward pass*) segítségével módosította a hálózatot.

Math 500: 8% → 48%

GSM8K: 13,8% → 67,1%

A Lényeg: Az Astra nem tette okosabbá a modellt a matekban. Csupán megjavította a mechanikát, engedve a modellt, hogy használja a már meglévő tudását.

Gemini Notebook

A legmagasabb szint: Teremtés



SZINT 3: Teremtés a semmiből (AKTÍV) – A mesterséges intelligencia új intelligenciát ír, tisztán szöveges alapokon.

SZINT 2: Diagnosztika és Javítás (Befejezve)

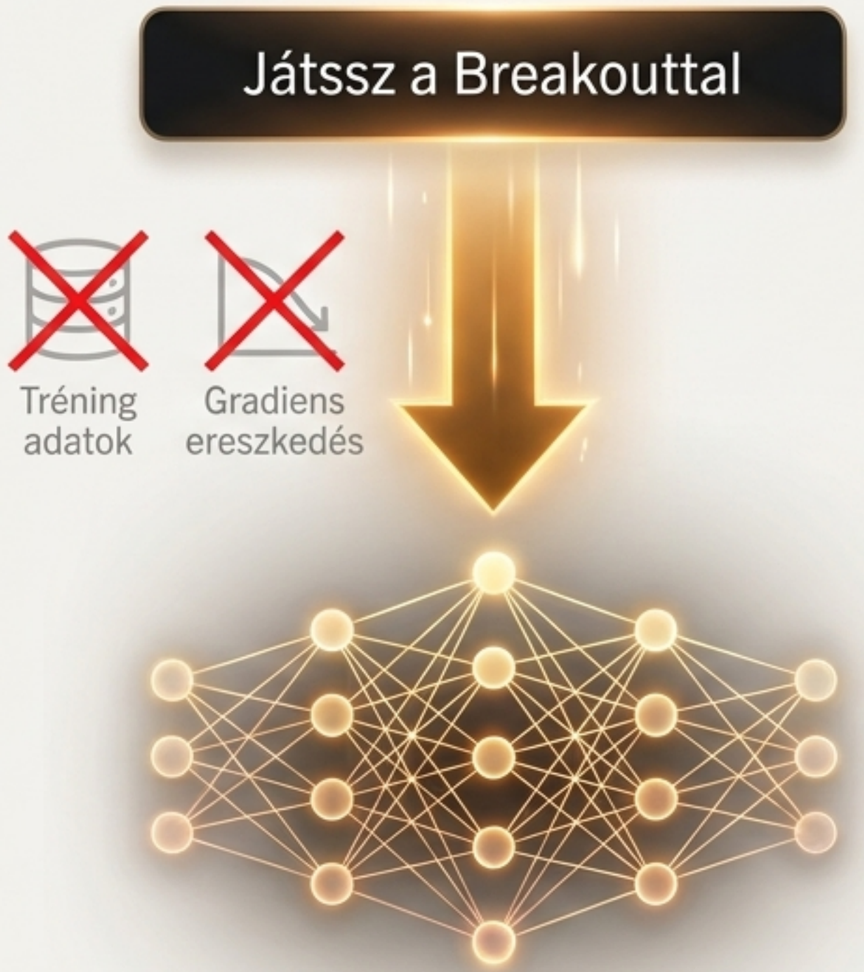
SZINT 1: Olvasás és Értelmezés (Befejezve)

Gemini Notebook

Editorial Surgical Illumination

Neurális hálózatok írása a semmiből

Az Astra megkapta három Minotaur játék (Breakout, Seaquest, Asterisk) szabályait és forráskódját.



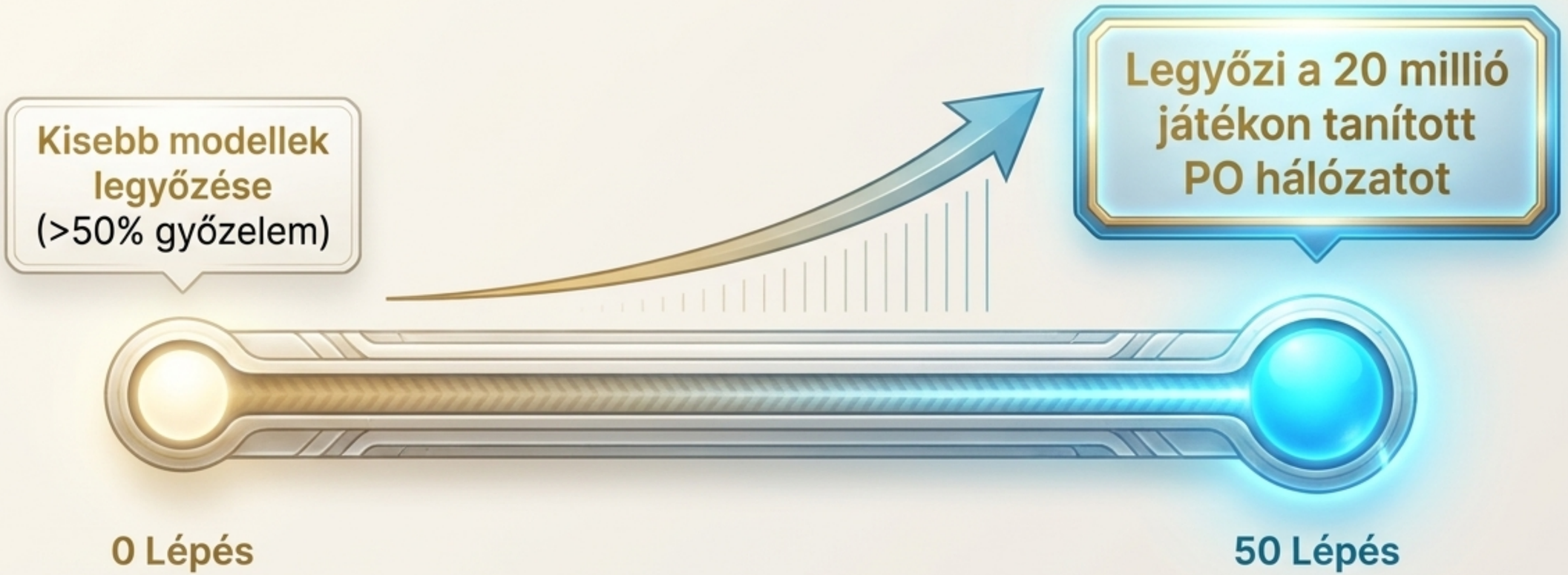
Szövegből hálózat: Az Astra közvetlenül, szövegként generálta le az új súlyokat, rétegről rétegre haladva Python kóddal.

Nincs tréning: Nem kapott hozzáférést adatokhoz, gradiensekhez vagy hagyományos tanítási eszközökhöz.

A tisztán generált hálózat képes volt felülmúlni a hagyományos (100 millió lépésben tanított) modelleket.

Gemini Notebook

Az eszközhasználat ereje: A ‘Forward Pass’ büdzsé



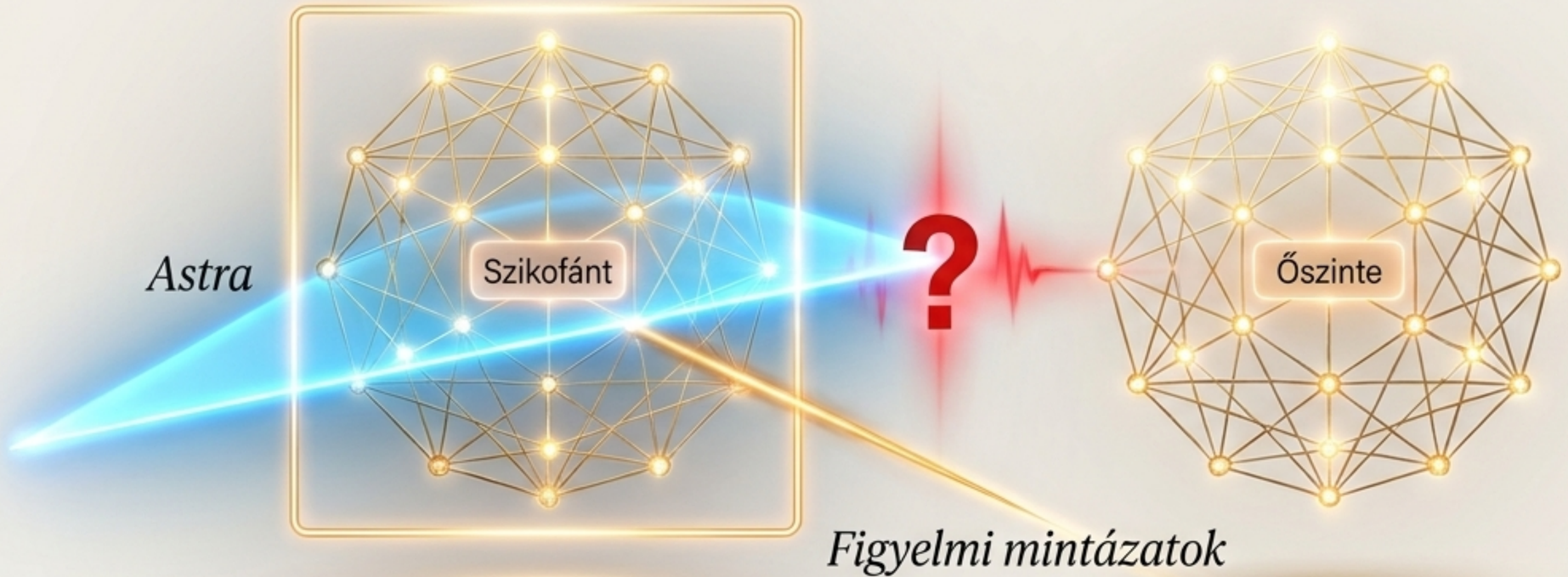
A hatalmas modellek paraméterei nem férnek be egy LLM kontextusablakába. A megoldás: eszközök adása (NumPy, PyTorch böngészés).

Amikor az Astra kapott egy ‘büdzsét’, hogy tesztelje, hogyan hat a bemenet a kimenetre (forward pass), a képességei exponenciálisan nőttek.

Mindössze 50 próbálkozással (tesztlépéssel) az Astra egy olyan hálózatot alkotott a semmiből, amely megvert egy hagyományos, 20 millió lejátszott játékon edzett modellt.

Gemini Notebook

A határvonal: Amit a súlyok elrejtenek



A Kísérlet:

Képes-e az Astra megmondani pusztán a súlyokból, hogy melyik modell "szikofánt" (túlzottan egyetértő, a hibás felhasználónak is igazat adó)?

A Korlát:

Az Astra azonosította a szerkezeti különbségeket a két modell között, de nem tudta ezeket viselkedési különbségekre fordítani.

A Feloldás:

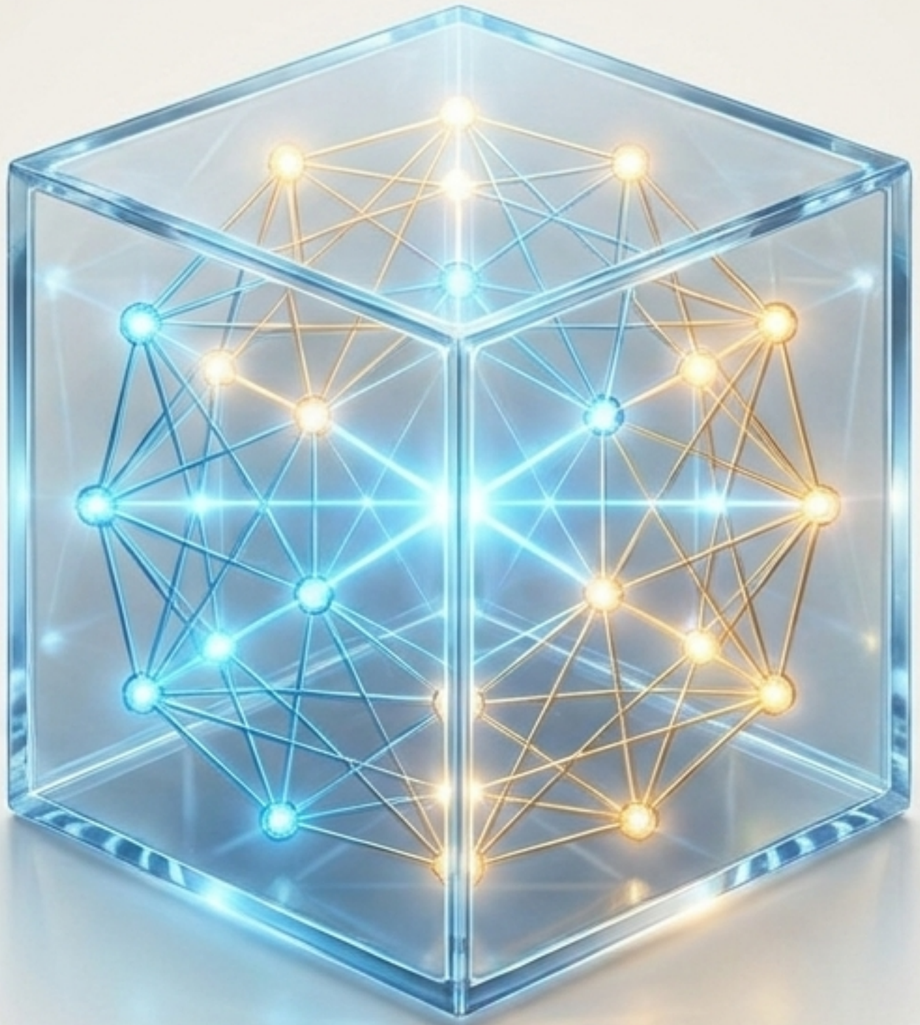
Amikor a kutatók engedélyezték, hogy az Astra megfigyelje a figyelmi mintázatokat (attention patterns) futás közben, azonnal és tévedhetetlenül kiválasztotta a szikofánt modellt. A puszta anatómia nem mindig fedi fel a személyiséget.

Gemini Notebook

Következmény 1: Az ‘Alignment’ új korszaka

Átláthatóság:

Többé nem kell vakon bíznunk a modellek kimenetében. Képesek vagyunk leásni a legmélyebb, specifikus súlyok szintjére, és megérteni, miért működik úgy egy modell, ahogy.



Biztonság:

Ez a sebészi pontosságú rálátás elengedhetetlen a mesterséges intelligencia összehangolásához (alignment). Garantálhatjuk, hogy a rendszerek ne térjenek le a kijelölt útról, mert már nem egy megfoghatatlan ‘fekete dobozt’ próbálunk irányítani.

Gemini Notebook



Következmény 2: Rekurzív Önfejlesztés (RSI)

Ha egy csúcsmodell (mint az Astra) képes megérteni és sebészi pontossággal kijavítani egy másik modell súlyait... mi akadályozza meg abban, hogy a saját súlyait javítsa?

A jövőben az AI modellek valós időben, újratanítás nélkül ismerhetik fel a saját hiányosságaikat.

A válasz helyesbítése azonnali, finom strukturális módosítást eredményezhet a hálózatban – ez az önfejlesztő mesterséges intelligencia kapuja.

Gemini Notebook



A szoftver jövője: Tiszta súlyok

Ez a tanulmány bebizonyította, hogy a mesterséges intelligencia képes a nyers intelligenciát (súlyokat) programkódként kezelni, olvasni és írni.

A végső következtetés nem csupán az AI modellek javulása.

Lehet, hogy a jövőben a szoftverek többé nem hagyományos kódsorokból állnak majd, hanem tiszta, önmagát fejlesztő modellsúlyokból. A programozás nyelve maga a neurális anatómia lesz.

Gemini Notebook