



INFINIDEA
CONSULTING

A saját AI

Mire képesek valójában a helyi modellek 2026-ban?

Egy helyi LLM – annyira biztonságos, amennyire csak akarod, és a hardveren túl gyakorlatilag ingyenes – elvégzi a szervezeted AI-munkájának nagy részét.

A mai előadás egy mondatban

Egy helyi LLM – annyira biztonságos, amennyire csak akarod, és a hardveren túl gyakorlatilag ingyenes – elvégzi a szervezeted AI-munkájának nagy részét.



Biztonságos

A modell a te géped futtatja. Bizalmas anyag nem hagyja el a házat.



Szinte ingyenes

A hardver egyszeri beruházás, utána nincs használatarányos díj.



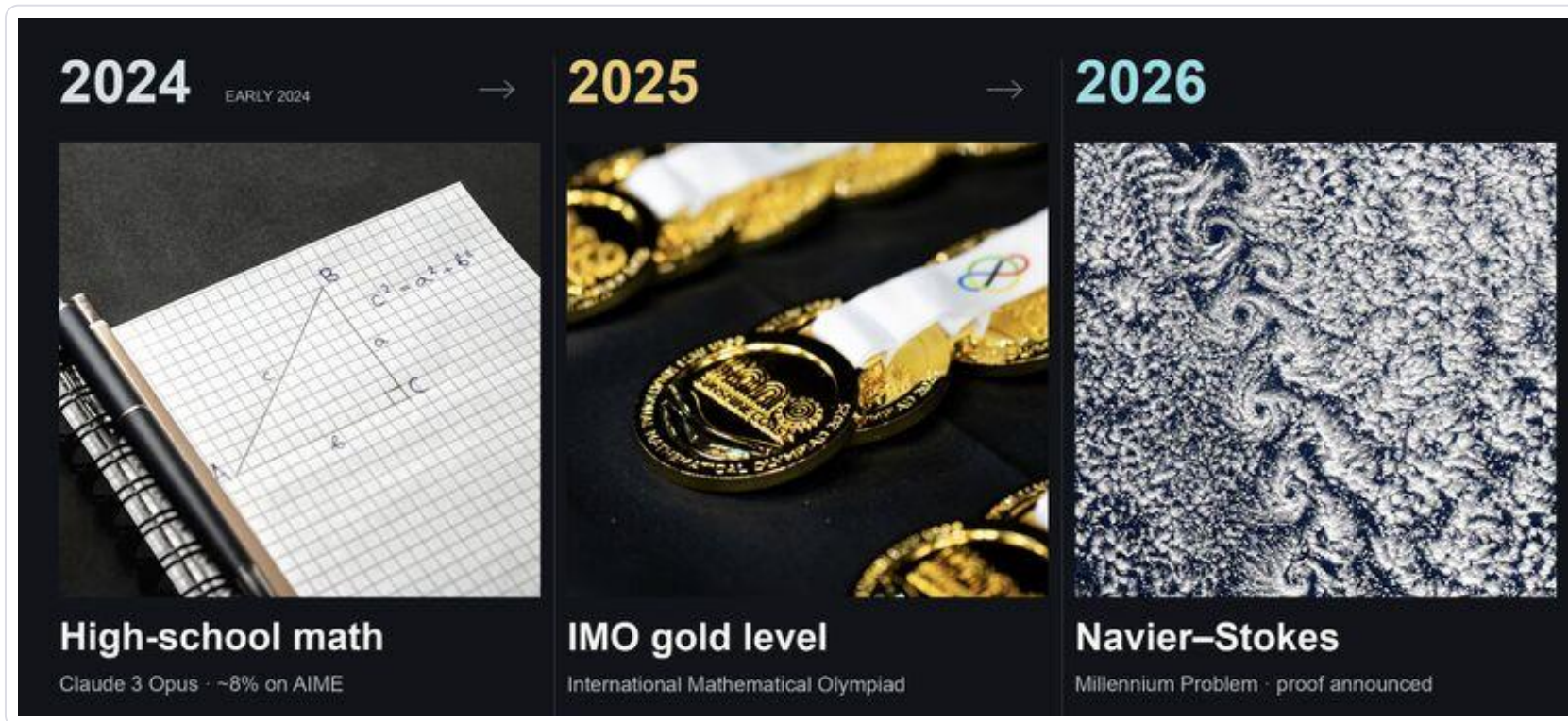
Elég okos

Kevésbé okos, mint a csúcsmodell – de a napi feladatokhoz bőven elég.



Az autokiegészítéstől a Navier–Stokesig

Három év a csúcsmodellek fejlődésében



2024: középiskolás matek · 2025: diákolimpiai arany · 2026: Navier–Stokes

Túl okos az irodai munkához

~80%

az irodai feladatokhoz

nem kell hozzá frontier modell



Szűkülő piac

Egyre kevesebben fizetnek meg egy még nagyobb tudást.



Üzleti plafon

A fejlődés üzleti okokból lassul, nem technikai okokból.



Szabályozás

A frontier laborok maguk sürgetik a szigorúbb kereteket.



A verseny: Egyesült Államok és Kína



Egyesült Államok

- A legjobb képességek
- Hatalmas ráfordítás



Kína

- Egyértelmű második hely
- Nagyságrenddel kisebb ráfordítás
- A legjobb ár-érték arány

A kínai laborok állami háttérét sokan sejtik – hivatalosan nem megerősített.

AI befektetések becsült értéke 1 Trillió dollár

318 878 000

000

000



Az ingyenes ezüstérem

1. hely: nagyon drága. · 2. hely: gyakorlatilag ingyenes.



Nyílt súlyú modellek

Letölthetők, és legálisan futtathatók a saját gépeden – a felhőben is olcsók.



Ingyennel nem lehet versenyezni

Aki a második helyet ingyen adja, kiszorítja a piacról a többi versenyzőt.



Mekkora a lemaradás?

fél–egy év

a csúcsmodellek mögött

a nemzetközi benchmarkok szerint



AI-időben rengeteg

Ez a terület hónapok alatt alakul át.



Üzleti időben semmi

Egy féléves technológiát senki nem hív elavultnak.



Mérni nehéz

Embernél is nehéz – a benchmark a legjobb, amink van.



Einsteinnek is voltak feladatai, amikhez nem kellett zseni

1

Az AI dolga: elvégezni a feladatot

Nem lenyűgözni – a látványos bemutató a csúcsmodellek dolga.

2

Minden munkakörben van rutinfeladat

Szükséges, de nem nehéz. Nálam is bőven akad.

3

Ez a feladatról szól

Nem a dolgozóról. Ezt fontos külön kimondani.

„Egyikünk sem használja az eszét 100%-ban minden egyes feladatnál. És ez teljesen rendben van.”



Illeszd a modellt a feladathoz

Magas – kevés feladat

Közepes – sok feladat

Alacsony – nagyon sok feladat

A feladat igényelte gondolkodás

Miért érdemes kisebbet választani?



Adatvédelem

A bizalmas anyag nem hagyja el a házat.



Költség

Sok felhasználóval gyorsan gyűlik a tokenszámla.



Illeszkedés

A feladat egyszerűen nem igényel többet.



Ezt emberekkel már régen tudjuk

„Egyetlen szervezet sem mondta még: nálunk csak PhD-sok dolgozhatnak.”



Minden szervezetben vegyes a csapat

Vannak kiemelkedő és teljesen átlagos emberek – mint én.



Az AI-nál sincs ez másképp

Nem mindig kell a csúcsmodell, mert az mára túl sokat tud.

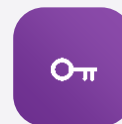


Ez nem újdonság, csak logika



Nagyon okos frontier modellek

Már többet tudnak, mint amennyi a napi munkához kell.



Ingyenes, nyílt súlyú modellek

Fél–egy éves lemaradással, legálisan, ingyen futtathatók.

Ezek a következtetések maguktól adódnak

A legtöbb erős nyílt súlyú modell kínai – de nem mindegyik.



Saját AI, házon belül



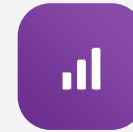
Mi ez?

Letölthető modellfájlok, amiket a saját géped futtat – nincs külső szolgáltató.



Mi kell hozzá?

Erős gamer PC néhány felhasználóig, ERP-szintű szerver az egész szervezetnek.



Mibe kerül?

Egyszeri hardverberuházás, utána gyakorlatilag nulla üzemeltetés.

Rövid ideig kell hozzá szakértő – de ez nem varázslat.



Kontextusablak: mennyi szöveget lát egyszerre?

~1M

token

frontier modellek

128–256K

token

helyi modellek

Ez több száz oldalnyi szöveg. A legtöbb dokumentum 100 oldal alatt van, a hosszabbak pedig értelmes részekre bonthatók – a gyakorlatban ez ritkán valódi korlát.



Amihez nem kell frontier modell

Paradicsomleves-recept haikuban, ló alakba hajtogatva? Az a frontier modell dolga.

- szöveg összefoglalása
- kulcsszavak és nevek kinyerése
- anonimizálás
- átírás más stílusban
- két dokumentum összevetése
- gyors fordítás
- jegyzetek és kérdéslisták

A tudásmunka nagyjából fele ilyen feladatokból áll.



Nem zsenialitás kell, hanem precizitás

„Valaki, aki pontos, és soha nem fárad el.”



Dokumentumok és űrlapok

Ugyanaz a sablon, új tartalommal – ez a szervezetek napi munkája.



API-integrációk

A dokumentáció és a tényleges megvalósítás összevetése.



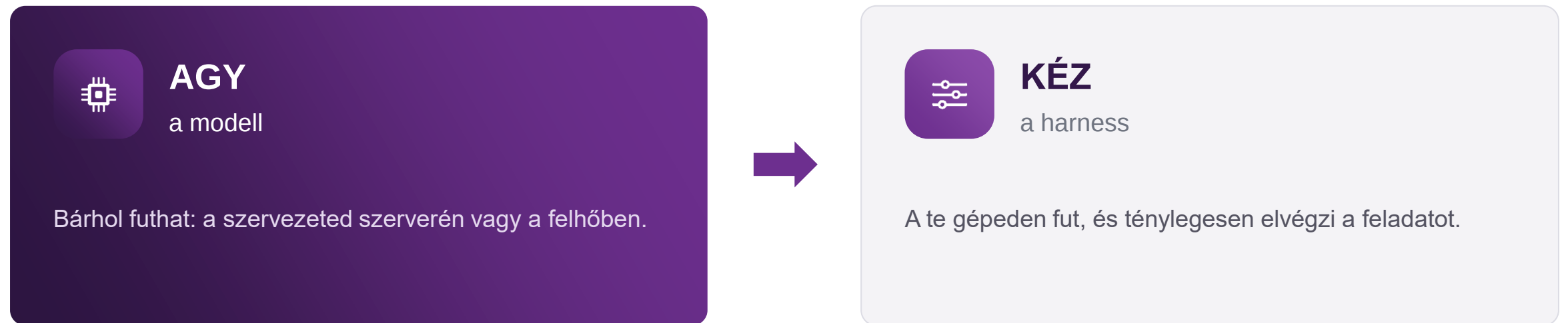
Szabályzat és jogszabály

Mondatról mondatra, akkor is, ha az ellentmondás 20 oldallal később bújik meg.

Gondolj a saját szervezetedre: melyik ismétlődő sablon jut eszedbe?



Agy és kéz: a harness



Claude Code · Codex · DeepSeek agent · saját fejlesztésű harness

A verseny nem a legjobb modellről szól, hanem arról, ki használja jobban. Helyi modellel is működik.



Ne nekem higgy – a benchmarkoknak

Qwen 3.8 27B $\approx$ Claude Opus 4.6

a 2026. februári csúcsmodell szintje



Megfizethető vason fut

Egy olyan szerveren, amit a legtöbb magyar cég meg tud venni.



Ez volt a fordulópont

Az Opus 4.6-tal kezdtünk valóságként beszélni az AI-ügynökökről.



Egy menetben – one-shot

Megmondod, mit szeretnél, és a modell egyetlen nekifutásra megírja a működő programot – közben maga javítja a saját hibáit.

Opus 4.6, 2026. február: egy 80-as évekbeli arcade játék, böngészőben, egy menetben.

Ma már az Opus 5.5-nél tartunk – a februári csúcstudás viszont mostanra helyben is elfut.



CPU, GPU vagy mindkettő



CPU

Kis modellek, egyszerű eszköztár
mindenkinek – lassabb, de
robusztus.



Hibrid

Ugyanaz, nagyobb sebességgel és
több felhasználóval.



Teljes GPU

Az egész szervezet kiszolgálása,
összetettebb feladatokkal.

A CPU-s eszközök tipikus feladatai: rövid fordítás, egyszerű OCR egy-két oldalon, számok kinyerése egy ábrából – nem orvosi képdiagnosztika. A törvény szerint jelezni kell, hogy AI-eszköz.





INFINIDEA
CONSULTING

Tényleg többre van szükséged, mint amit a februári legjobb modell tud?

Néhány feladatnál igen. De legalább a felénél biztosan nem.

Köszönöm a figyelmet!

Kérdés, beszélgetés

Szívesen beszélek a saját szervezeted feladatairól is.



office@infinidea.hu



[+36 30 986 6210](tel:+36309866210)



www.infinidea.hu