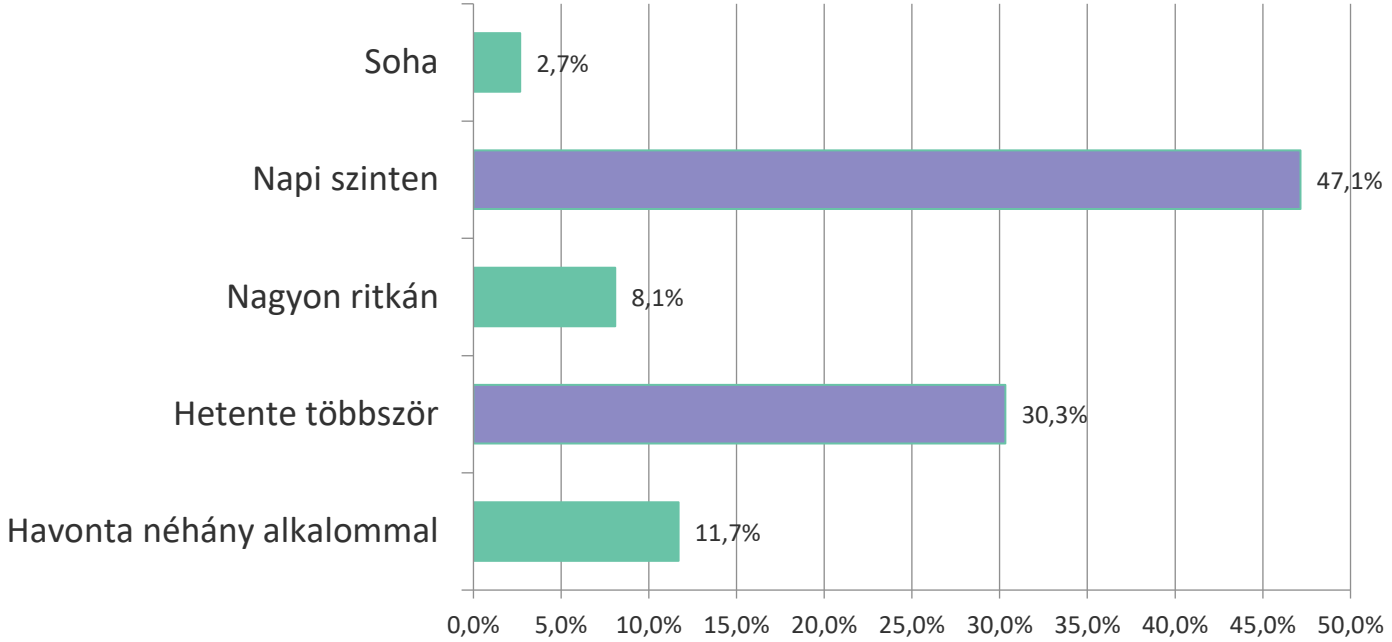
Az AI lufi lassan leereszt

hogyan számoljuk fel a Shadow AI-t és vezessünk be valóban használható vállalati AI képességeket

2026.04.25

Jagusztin László, Alerant Zrt.

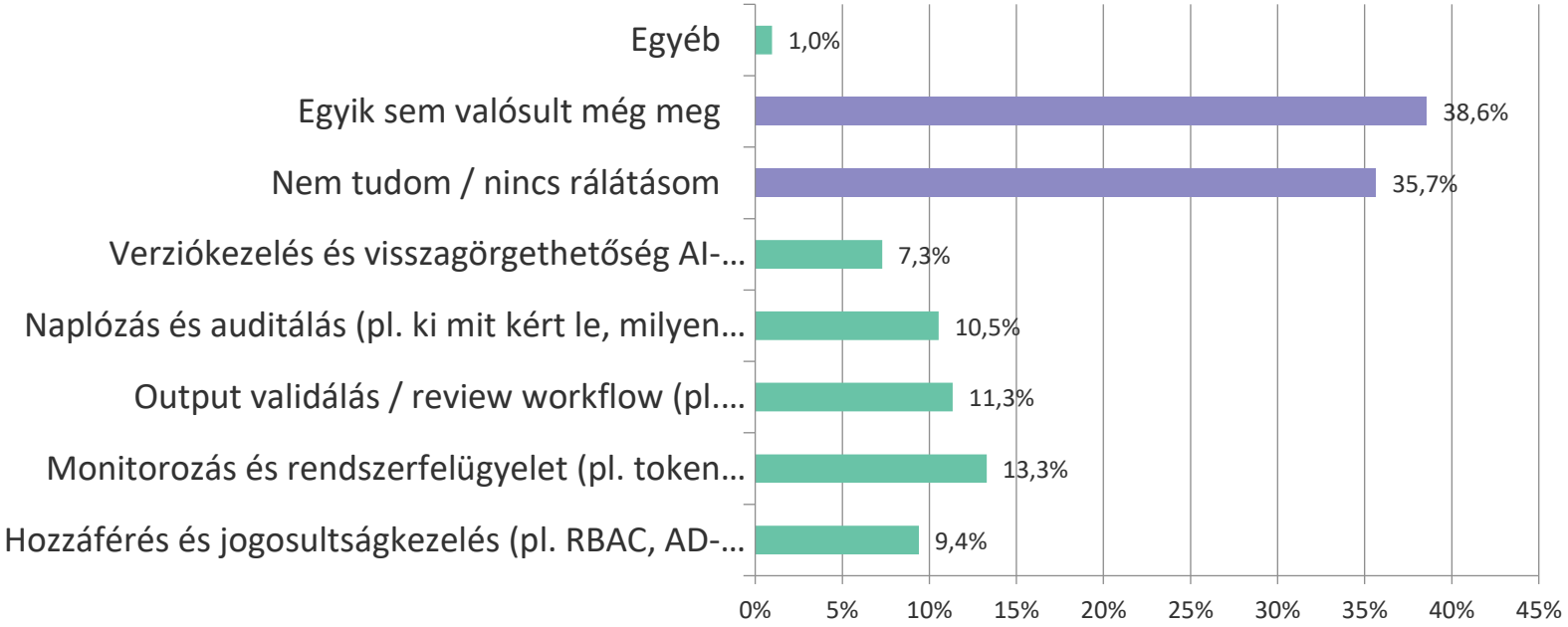
Milyen gyakran használod a munkádhoz az AI-t?



ALERANT

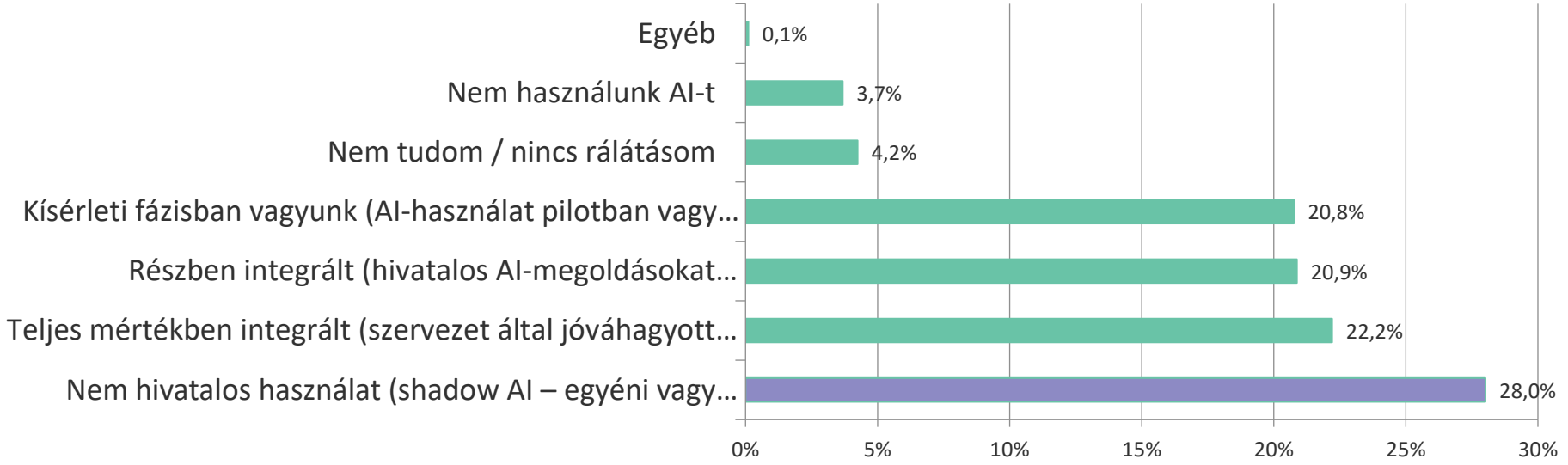
itkorkep2025.alerant.hu, 1000+ magyar IT szakember válaszai alapján

Milyen vállalati funkciókra használjátok?



itkorkep2025.alerant.hu, 1000+ magyar IT szakember válaszai alapján

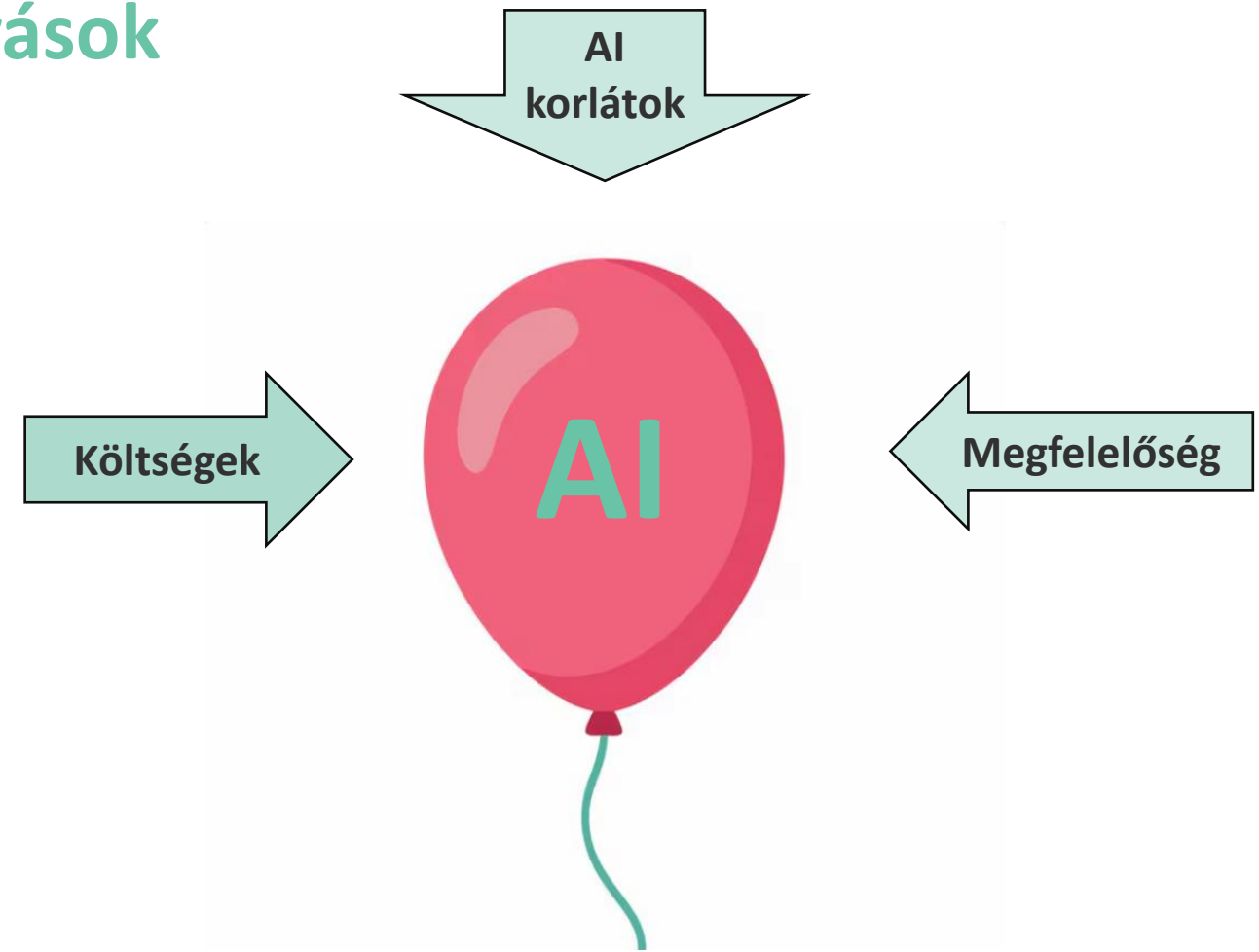
Milyen mértékben van integálva a működésbe az AI?



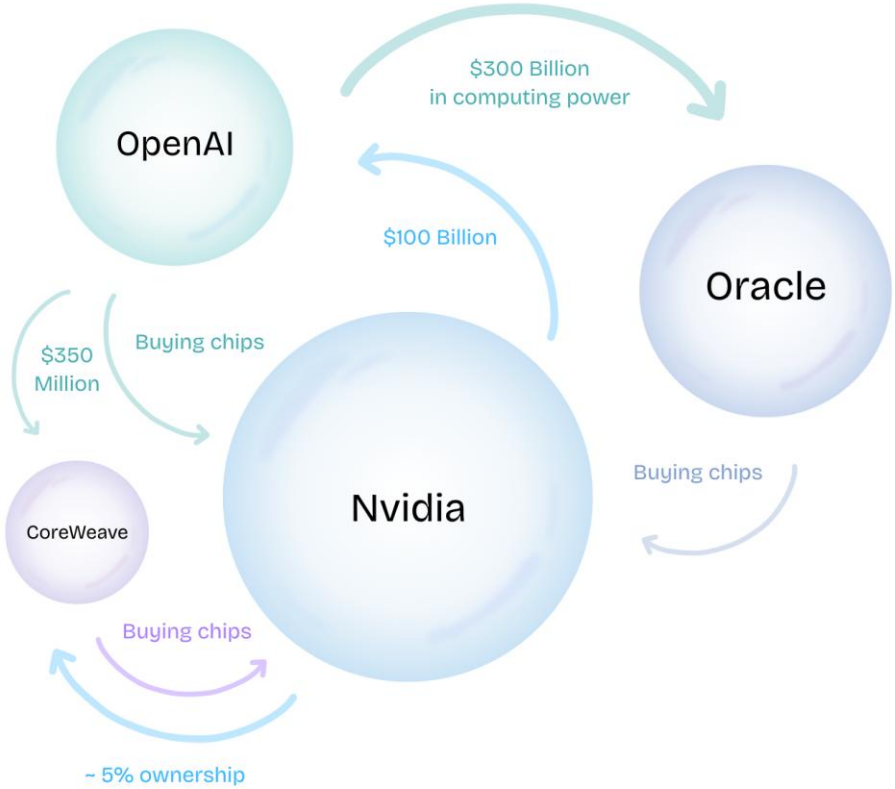
ALERANT

itkorkep2025.alerant.hu, 1000+ magyar IT szakember válaszai alapján

AI kihívások

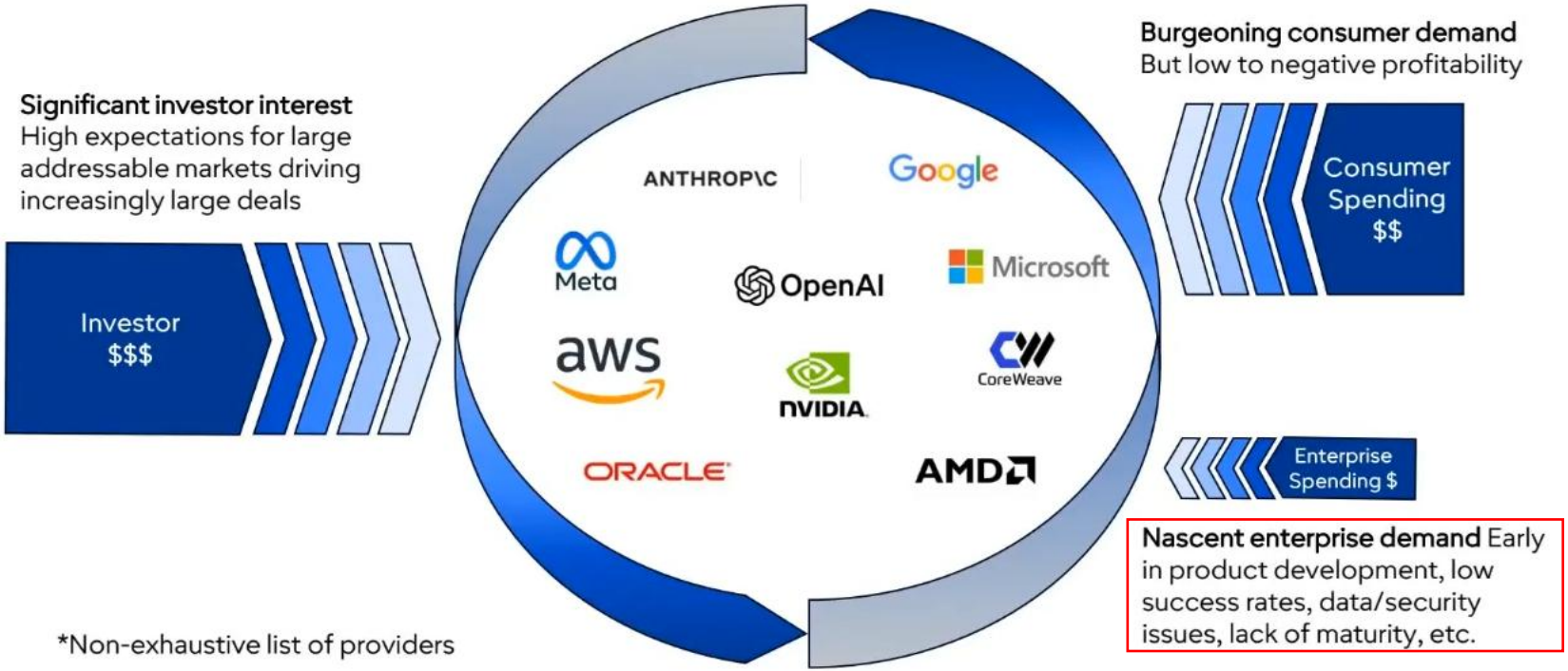


AI Lufi



https://en.wikipedia.org/wiki/AI_bubble

AI Circular Investment Cycle



Source: Assessing the Risks in AI's Circular Investment Cycle

Gartner

Mit hoz a jövő?

Sam Altman, San Francisco

2026 Január 27.

- **2027: Az év, amikor az intelligencia „túl olcsó lesz ahhoz, hogy mérni kelljen”**
- 2027 végére az intelligencia költsége jelentősen csökkenni fog — akár 100x mértékben
- Ugyanakkor egy új kompromisszum körvonalazódik: **sebesség vs. költség**.
- **Piaci kettéválás:** lesz „lassú, olcsó gondolkodás” a háttérfeladatokra, és „gyors, prémium gondolkodás” a valós idejű felhasználói interakciókhoz.

ALERANT

2026 Január óta a változások

- A szolgáltatási verseny eltolódott az agent-platformok felé:
  - Minden szolgáltatónál megjelent Batch API 50% költségcsökkentésse
  - Gemini 3.1 Flash-Lite: ultra-olcsó modell az API-felhasználóknak (0,25$/1M token)
  - A Claude Pro csomagból áprilisban kikerül a Claude Code – majd visszakerül..
- Új verseny a token gazdaságosság:
  - A Gemini 3 Pro esetében az árazás már a kontextus méretétől függ: a 200 ezer token feletti az input költség megduplázódik, ösztönözve a felhasználókat a hatékonyabb adatkezelésre.
  - A ChatGPT Codex díjazása üzenetalapúról tokenalapúra váltott; április 23-tól ezt a meglévő Enterprise/Edu/Health/Gov/Teachers csomagokra is kiterjesztették.
  - A MS leállítja a GitHub Copilot Pro, Pro+ új előfizetéseket, bevezeti a token alapú árazást, új limiteket vezet be az alacsonyabb csomagkora és kivezeti a drágább modelleket

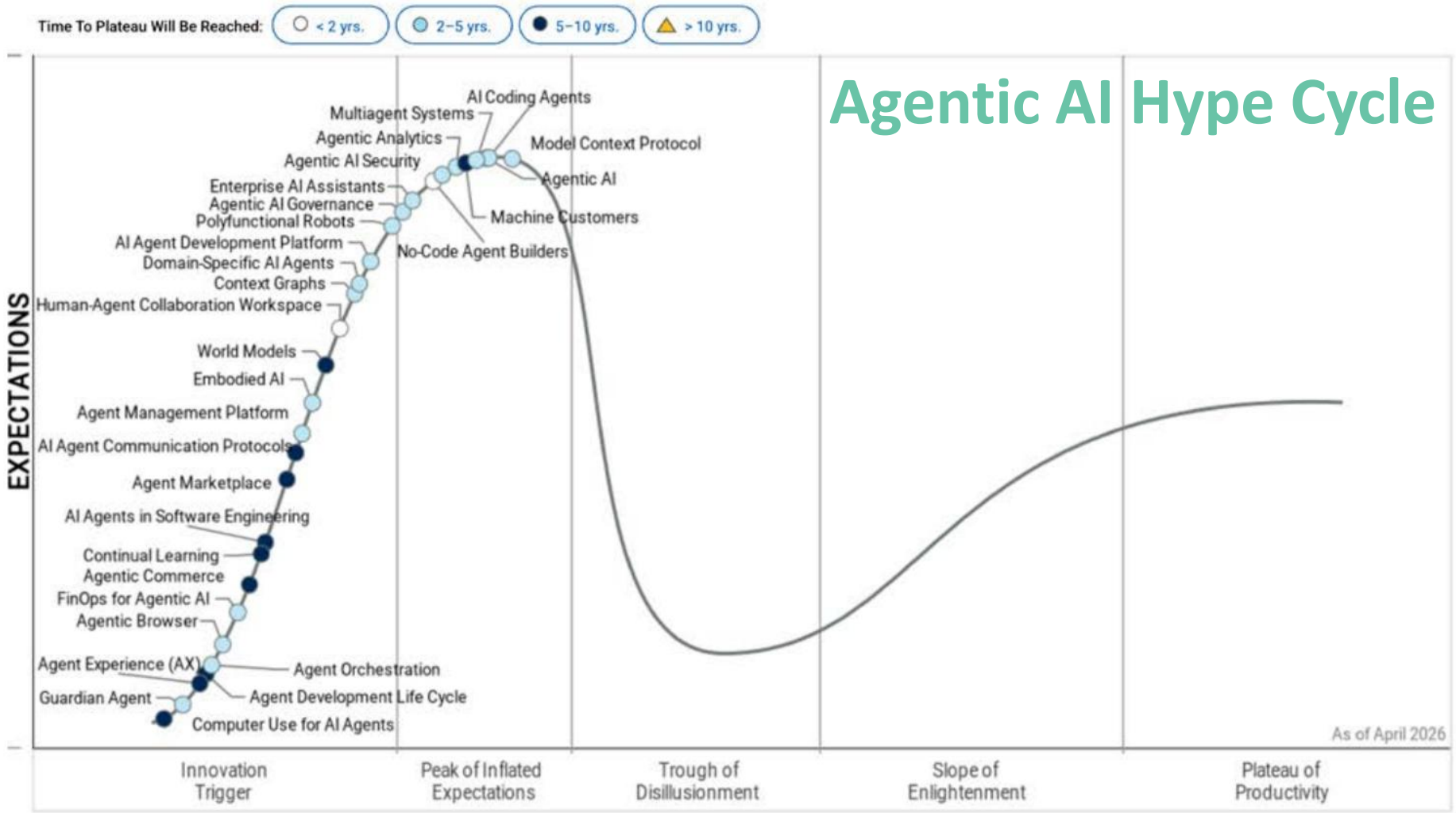
MS GitHub Copilot magyarázat

https://github.blog/news-insights/company-news/changes-to-github-copilot-individual-plans/

- „Azt tapasztaljuk, hogy a használat *minden* felhasználónál intenzívebbé vált, ahogy felismerik az agent-ek értékét”
- „Ezek a hosszan futó, párhuzamosított munkafolyamatok … komoly kihívás elé állították az infrastruktúránkat és az árazási struktúránkat”

- Népszerű agentek: OpenClaw, OpenCode, Hermes, NemoClaw.. mind a tömeges felhasználást erősítik

Agentic AI Hype Cycle



Source: Gartner

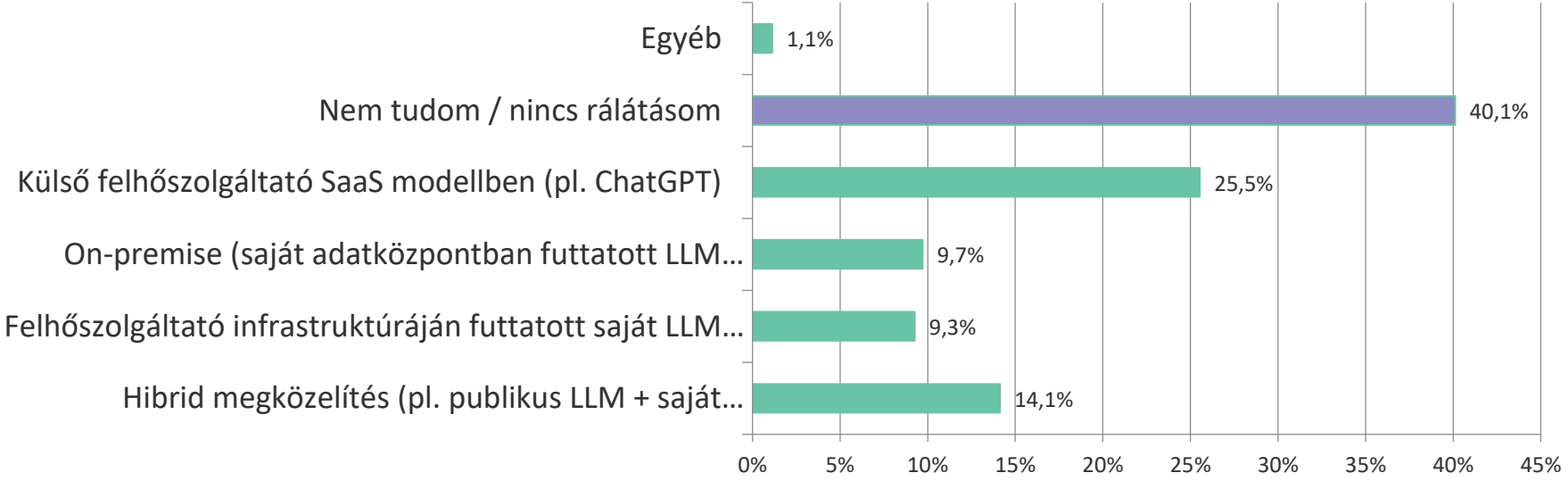
AI FinOPS

$$K_{\text{cloud}} = \frac{Q \cdot D \cdot T_{in}}{1\,000\,000} \cdot C_{in} + \frac{Q \cdot D \cdot T_{out}}{1\,000\,000} \cdot C_{out}$$

Egy egyszerű számítás:

- A GPT-5.2 futtatása 1,75 USD-be kerül 1millió bemeneti tokenenként, és 14 USD-be 1millió kimeneti tokenenként.
  Egy napi 10 000 lekérdezést kezelő ügyfélszolgálati alkalmazás esetében
  (500 tokenes bemeneti és 1000 kimeneti kontextussal számolva) ez **4462** USD/hó
- Egy helyben futtatott 7B modell egy 2 000 USD értékű szerveren:
  Körülbelül 50 USD havonta villamos energiára.
  A szerver két év alatti amortizációját figyelembe véve a teljes havi költség **133** USD.
- **Ez 97% költség csökkentés!**

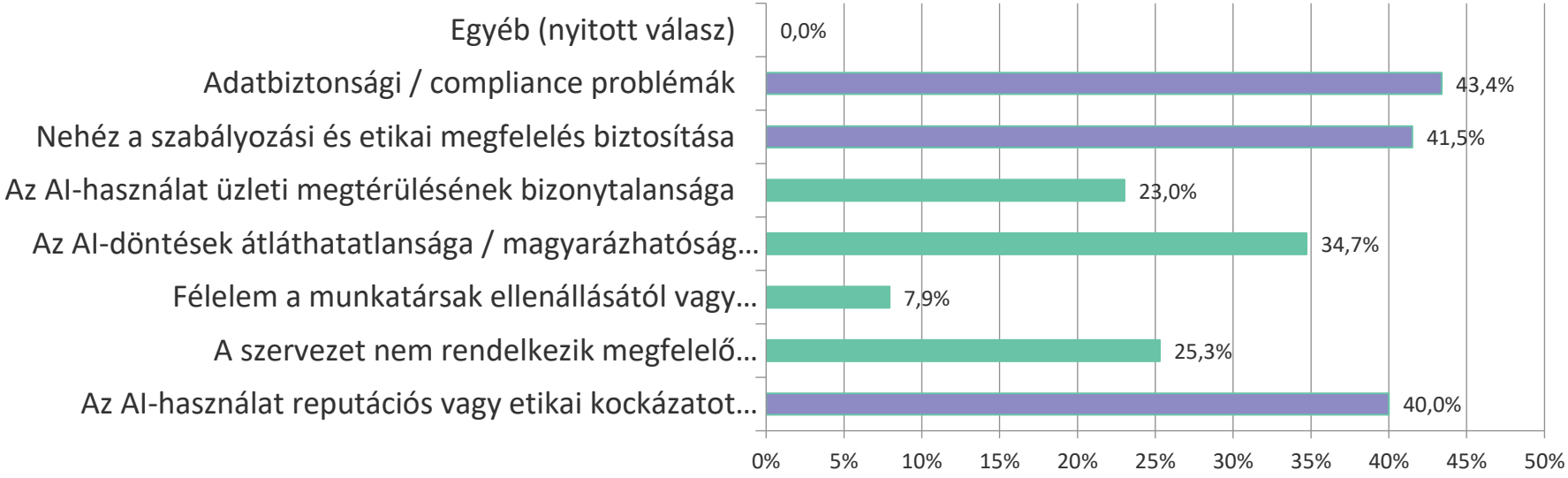
Milyen szolgáltatási modellben van használva ?



ALERANT

itkorkep2025.alerant.hu, 1000+ magyar IT szakember válaszai alapján

Milyen kihívások aggasztanak AI használat kapcsán?



itkorkep2025.alerant.hu, 1000+ magyar IT szakember válaszai alapján

AI szolgáltatási modellek

SaaS

On-premise

Private Cloud

AI megoldások – publikus szolgáltatás

Public service

- ChatGPT, Gemini, Deepseek…
  - Megfelelőségi és biztonsági problémák
  - Használat alapú költség modell – könnyen elszaladnak a költségek
  - Az API nem stabil, példa: GPT-4o -> GPT-5 váltás problémák
  - Model tanításra használják az adatainkat - és minden másra is..

Adatelemzés: Harvard AI study, 700M user

Public service

Occupation Group	Documenting/ Recording Information	Making Decisions And Solving Problems	Thinking Creatively	Working With Computers	Interpreting The Meaning Of Information For Others	Getting Information	Providing Consultation And Advice To Others
Management	2	1	3	6	4	5	8
Business	2	1	3	6	4	5	7
Computer/Math	4	2	5	1	3	6	7
Engineering	3	1	5	2	4	6	7
Science	2	1	4	3	6	5	7
Social Service	2	1	3	X	5	4	X
Legal	1	X	X	X	X	X	X
Education	1	2	3	4	6	5	7
Arts/Design/Media	2	1	3	5	4	6	7
Health Professionals	1	2	3	X	5	4	6
Food Service	1	X	X	X	X	X	X
Personal Service	1	2	3	X	4	5	X
Sales	2	1	3	6	4	5	7
Administrative	2	1	3	7	4	5	8
Transportation	2	1	3	X	X	4	X
Military	2	1	X	X	X	X	X

AI megoldások – privát szolgáltatás

- Azure OpenAI, AWS Bedrock, GCP Vertex AI, ChatGPT Team/Enterprise
  - Megfelelőségi és biztonsági problémák
    - Megfelelés magasabb szinten: SOC 2 Type II, GDPR stb.
    - Az adataink nem kerülnek ki EU-n kívülre (EU Data Boundary)
    - Adatainkat nem használják fel modell tanításra
    - Biztonságos hálózat kezelés (titkosított/private link)
- Továbbra is probléma:
  - Használat alapú költség modell – könnyen elszaladnak a költségek
  - Az API nem stabil: GPT-4o vs GPT-5
- Adatkezelés:
  - Visszaélés elemzési célra tárolva vannak a chat adataink is, időbeni korlát nélkül
  - Bírósági megkeresésre chat adataink kiadhatóak
  - CMK/BYOK támogatás van, de nem teljeskörű, származtatott adatokra nincs (pl. logok)
  - Model tanításra nem használják az adatainkat – de emberek olvassák őket

AI szolgáltatások adatkezelése

Emberek olvassák a chat tartalmakat:

- **Azure OpenAI**: „preventing abuse, ...prompts and completions are stored for human review. …automated review means including by AI models such as LLMs by default, with additional reviews by human reviewers as necessary. The human reviewers are authorized Microsoft employees..”
  https://learn.microsoft.com/en-us/azure/ai-foundry/responsible-ai/openai/data-privacy?tabs=azure-portal
- **Google Gemini**: Egyes mentett csevegéseket emberek vizsgálnak felül a Google AI továbbfejlesztése érdekében. …ne adjon meg olyan információkat, amelyeket nem szeretné, hogy átnézzünk vagy felhasználjunk.
- A **GitHub** bejelentette, hogy 2026. április 24-től alapértelmezetten felhasználja a Free, Pro és Pro+ felhasználók adatait a modellek fejlesztésére. Ez azt jelenti, hogy a beírt promptok, a generált válaszok és a kódrészletek is bekerülhetnek a tanításba, ha ezt nem kapcsolod ki.
  https://github.blog/news-insights/company-news/updates-to-github-copilot-interaction-data-usage-policy/
- Bírósági megkeresésre meg kall tartani őket és ki kell adni:
  ChatGPT: „court ordered OpenAI to preserve all output log data that would otherwise be deleted, even if a user requests the deletion of a chat or if privacy laws require OpenAI to delete data.”
  https://www.theverge.com/news/681280/openai-storing-deleted-chats-nyt-lawsuit?utm_source=chatgpt.com

AI szolgáltatói adatkezelés

	Claude Anthropic	DeepSeek DeepSeek	Gemini Google	Copilot Microsoft	Le Chat Mistral	ChatGPT OpenAI
PROVIDER BENEFITS						
Trains on user data	✓	✓	~	✓	✓	✓
Opt-out available (not opt-in)	✓	✓	✓	✓	✓	✓
Data still used after opt-out	✓	✗	✗	✗	✗	✗
Shares with third parties	✓	✓	✓	✓	✓	✓
Used for advertising	✗	✗	~	✓	~	~
User prohibited from training on outputs	✓	✗	✓	✓	✓	✓

ACCM Conference, Terms of (Ab)Use: An Analysis of GenAI Services, https://arxiv.org/abs/2603.18964v2

AI szolgáltatók EU jogszabályi megfelelése

	Claude Anthropic	DeepSeek DeepSeek	Gemini Google	Copilot Microsoft	Le Chat Mistral	ChatGPT OpenAI
LEGAL & JURISDICTION						
Acknowledges EU jurisdiction	✓	✗	✓	✓	✓	✓
Specifies applicable laws clearly	✗	✗	✗	✗	✓	✗
Local consumer laws acknowledged	✓	✗	✓	✓	✓	✓
Mentions AI Act	✗	✗	✗	✗	✓	✗

ACCM Conference, Terms of (Ab)Use: An Analysis of GenAI Services, https://arxiv.org/abs/2603.18964v2

Privát AI platform – a hiányzó képesség

Private AI

- Miért használjunk helyi AI megoldásokat?
  - A biztonsági és megfelelőségi problémákat megoldja – éveket nyerünk vele
  - Adatok nem kerülnek ki harmadik félhez
  - Kiszámítható, korlátlan, nem használat alapú költségek – jóval olcsóbb
  - Nem ügyfél függő a használata (mit lehet használni és mit nem)
  - Nem függ külső szolgáltatótól a működése (stabil válaszok)
  - Testre szabható nyelvi modellek, megoldások, komponensek
  - OpenAI kompatibilis API használata integrációra

LLM vs SLM

https://www.gartner.com/en/newsroom/press-releases/2025-04-09-gartner-predicts-by-2027-organizations-will-use-small-task-specific-ai-models-three-times-more-than-general-purpose-large-language-models

- **A Gartner előrejelzése szerint 2027-re a szervezetek háromszor gyakrabban fognak kis, feladatspecifikus modelleket használni, mint általános célú nagy nyelvi modelleket.**
- **Az NVIDIA Research 2025 június (Belcak és mtsai):**
  „A kis, nem pedig a nagy nyelvi modellek jelentik az agentikus MI jövőjét.”

A megbízható és költséghatékony megoldások

Gyorsválaszokat adnak, alacsony késleltetéssel

Kevesebb számítási kapacitást igényelnek

LLM válaszok pontossága rosszabb a speciális szakterületek esetén

Egyetlen modell valós feladatok esetén sokszor nem elég

A finomhangoláshoz az LLM nem megfelelő, költséges

Saját tulajdonú modelljeik hasznosítása a vállalatoknak új bevételi forrás

Megfelelőség

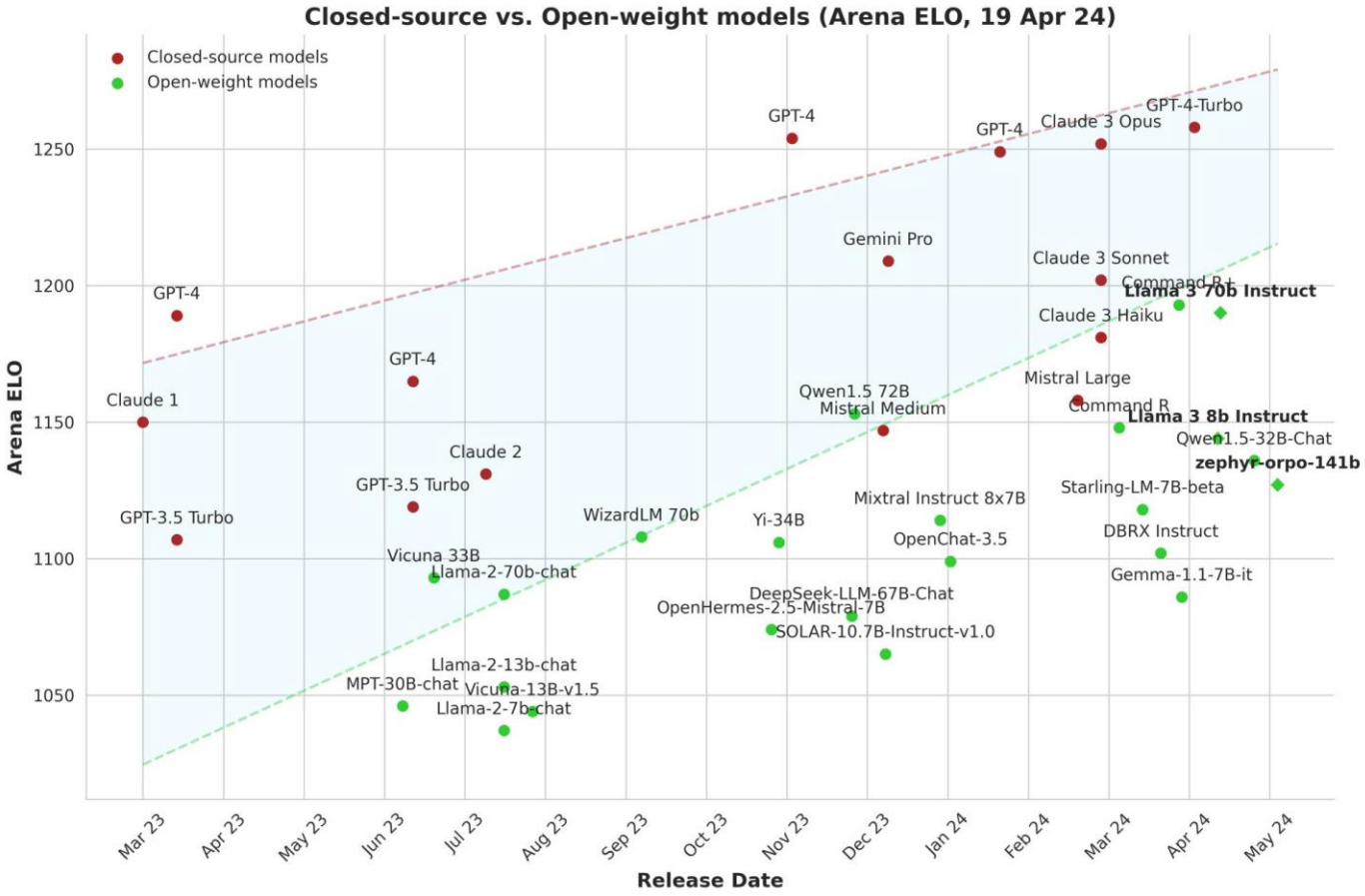
A konkrét üzleti problémákra a kisebb, finomhangolt modellek jobb eredményeket adnak nagyobb hatékonysággal — különösen a szabályozott iparágakban.

GDPR: egyszerűbb megfelelés (helyi adatok)	HIPAA: natív megfelelés az egészségügyben	EU AI Act: teljes kontroll a modell felett	Pénzügy: a szenzitív adatok kitettségének kockázata megszűnik	Digitális szuverenitás: garancia

- Rami Luisto, a Digital Workforce:
  „Amikor a magyarázhatóság és a bizalom kulcsfontosságú, egy SLM auditálása sokkal egyszerűbb, mint feltárni egy LLM viselkedésének okait.”
- Egy 3B paraméteres SLM vizsgálható, megérthető, auditálható. Egy 400 milliárdos LLM? Sok szerencsét..

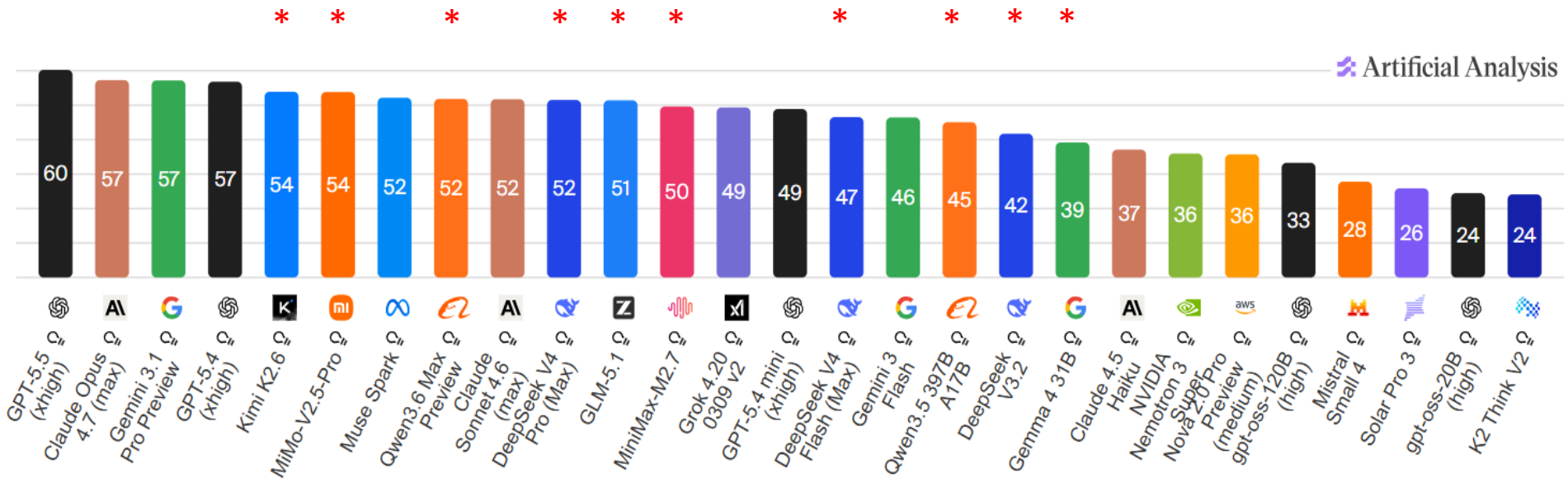
ALERANT

Zárt és nyílt modellek közti különbség



ALERANT

AI Index 2026 Április



* Nyílt, helyi modell

ALERANT

A Gartner jóslata szerint 2027-re az országok 35%-a régió-specifikus AI platformokat fog használni

Digitális szuverenitás és helyi modellek

- **Hazai fejlesztések:** A digitális szuverenitási célokat követő országok növelik befektetéseiket a hazai AI infrastruktúrákba, alternatívát keresve a zárt amerikai modellekkel szemben.
- **Kulturális illeszkedés:** A döntéshozók ma már előnyben részesítik a helyi értékekhez, jogi szabályozáshoz és felhasználói elvárásokhoz igazodó AI platformokat a legnagyobb modellekkel szemben.
- **Regionális előny: A** regionális nyelvi modellek (LLM-ek) jobban teljesítenek az oktatás, a jogi megfelelőség és a közszolgáltatások terén, különösen a nem angol nyelvű környezetekben.

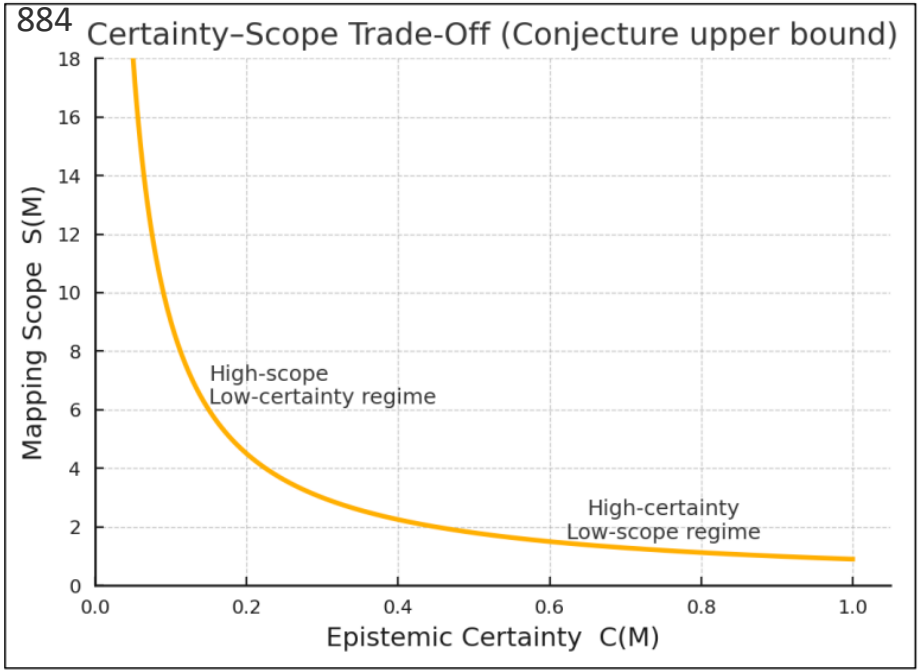
ALERANT

https://www.gartner.com/en/newsroom/press-releases/2026-01-29-gartner-predicts-35-percent-of-countries-will-be-locked-into-region-specific-ai-platforms-by-2027

LLM növekedési korlát

- **Floridi sejtés:** kimondja, hogy egy AI abszolút nem rendelkezhet egyszerre nagy mennyiségű információval és nagyfokú bizonyossággal.
- Másképpen megfogalmazva: ahogy egy AI modell egyre általánosabbá válik, elkerülhetetlenül **csökken a pontossága**, és rendellenes válaszokat produkál (melyeket „hallucinációknak” neveznek).

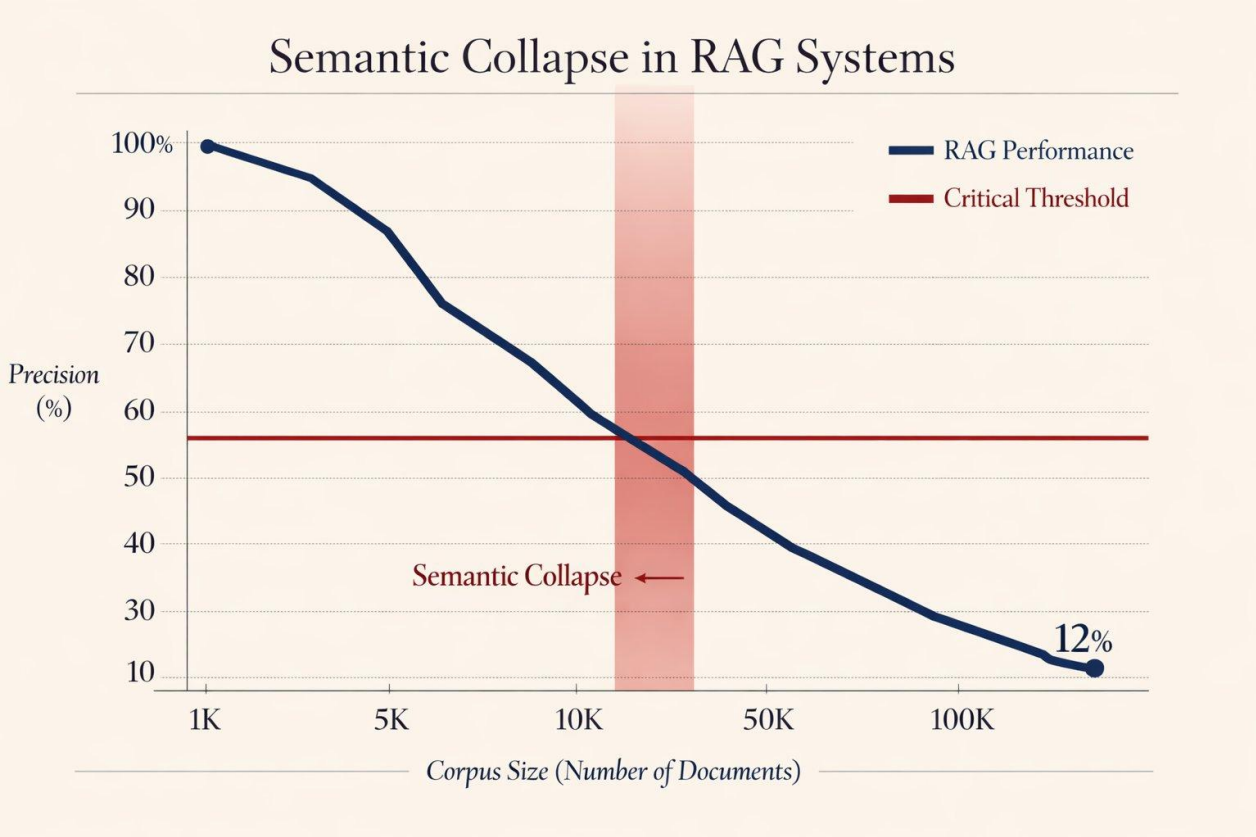
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5289



Ez a tétel jól megmagyarázza, miért látunk elmozdulást a vállalati szférában a nagy, általános modellek felől a kisebb, precízebb, specifikus megoldások irányába.

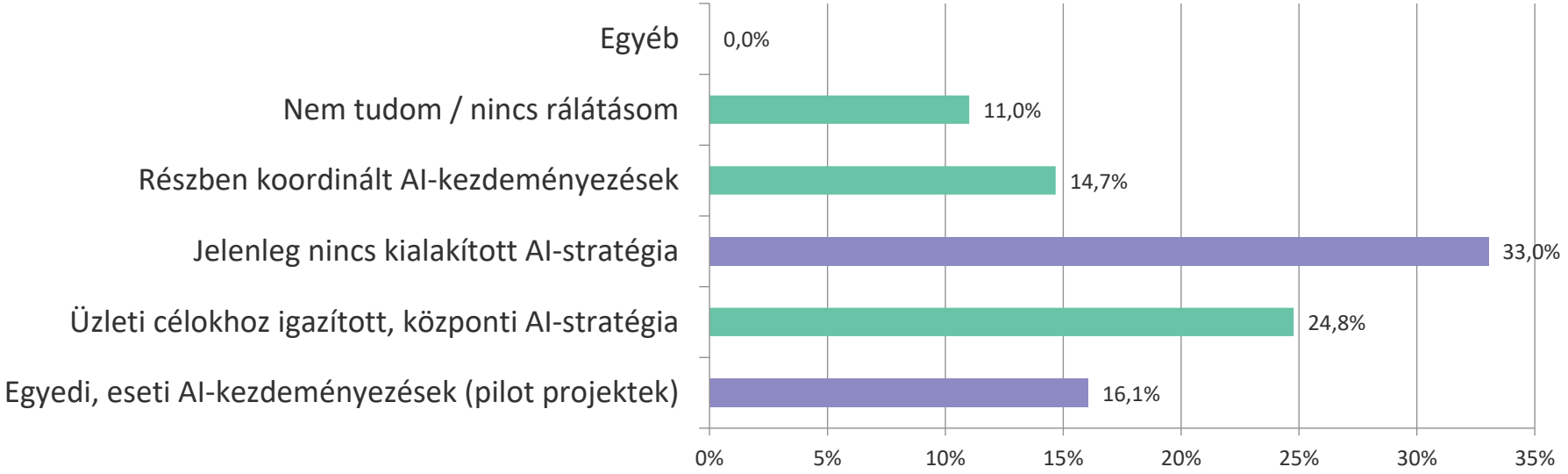
ALERANT

Stanford research: A RAG méretezhetősége korlátos



ALERANT

Van-e AI stratégiája a szervezetnek?



ALERANT

itkorkep2025.alerant.hu, 1000+ magyar IT szakember válaszai alapján

AI Stratégiai irányok 2026-ra

- Megfelelőség és költség hatékonyság biztosítása
  - Szolgáltatási modellek tudatos használata: Privát AI, PAAS és SAAS
  - SLM és LLM modellek vegyes használata, célhoz kötötten
- Az AI korlátok figyelembe vétele a megoldások tervezésénél
  - Hallucináció, kontextus méret, számítás igény, költség, adatmennyiség
  - Agent alapú megoldások a korlátok kompenzációjára (workflow, kontroll)
- A Shadow AI felszámolása valóban használható vállalati AI platformokkal
  - A Platform Engineering szerepe nőni fog ezen a területen is

KÖSZÖNÖM A FIGYELMET!

Jagusztin László
Alerant Zrt.