ISACA MÁSODIK SZERDAI ELŐADÁS

Varga Gábor

ChatGPT és más nyelvi modellek biztonsága az elvek és a gyakorlat szintjén

MI és innovációs szakértő

linkedin.com/in/gaborvar

gaborvar@hotmail.com

2023. május 10.

Az ISACA Magyarországi Egyesület által szervezett Második szerdai előadásokon bármilyen eszközzel történő kép- vagy hangrögzítés tilos. (Idetartozik a mobiltelefonokkal készített kép- vagy hangfelvétel is). A jelen szabály be nem tartása szerzői- és szomszédos jogi jogsértés jogkövetkezményeit vonhatja maga után.

Confidential. For internal use only.

ISACA Budapest Chapter

FELELŐSSÉG KIZÁRÁSA

Az elhangzott prezentációk tartalmát illetően az ISACA Magyarországi Egyesület nem vállal felelősséget az abban nyújtott információk aktualitásáért, helyességéért és teljességéért.

Az itt elhangzott információk nem feltétlenül egyeznek meg az ISACA Magyarországi Egyesület álláspontjával.

Confidential. For internal use only.

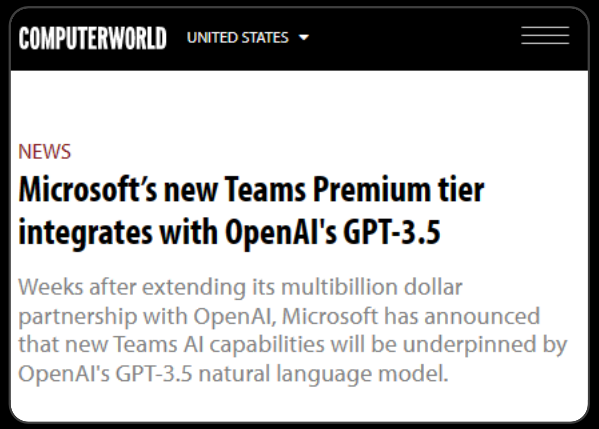
ISACA Budapest Chapter

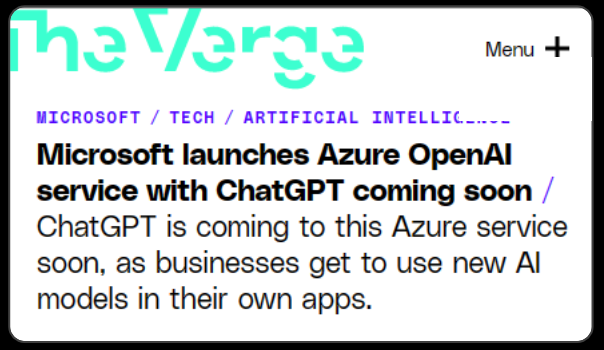
ChatGPT és más nyelvi modellek biztonsága az elvek és a gyakorlat szintjén

Varga Gábor
MI és innovációs szakértő
linkedin.com/in/gaborvar
gaborvar@hotmail.com

Néha zavar minket, s néha gépek zavarják,
Már zsebre vágjuk, s ha rosszul figyelsz, ő is zsebre vág.

A ChatGPT jelenség



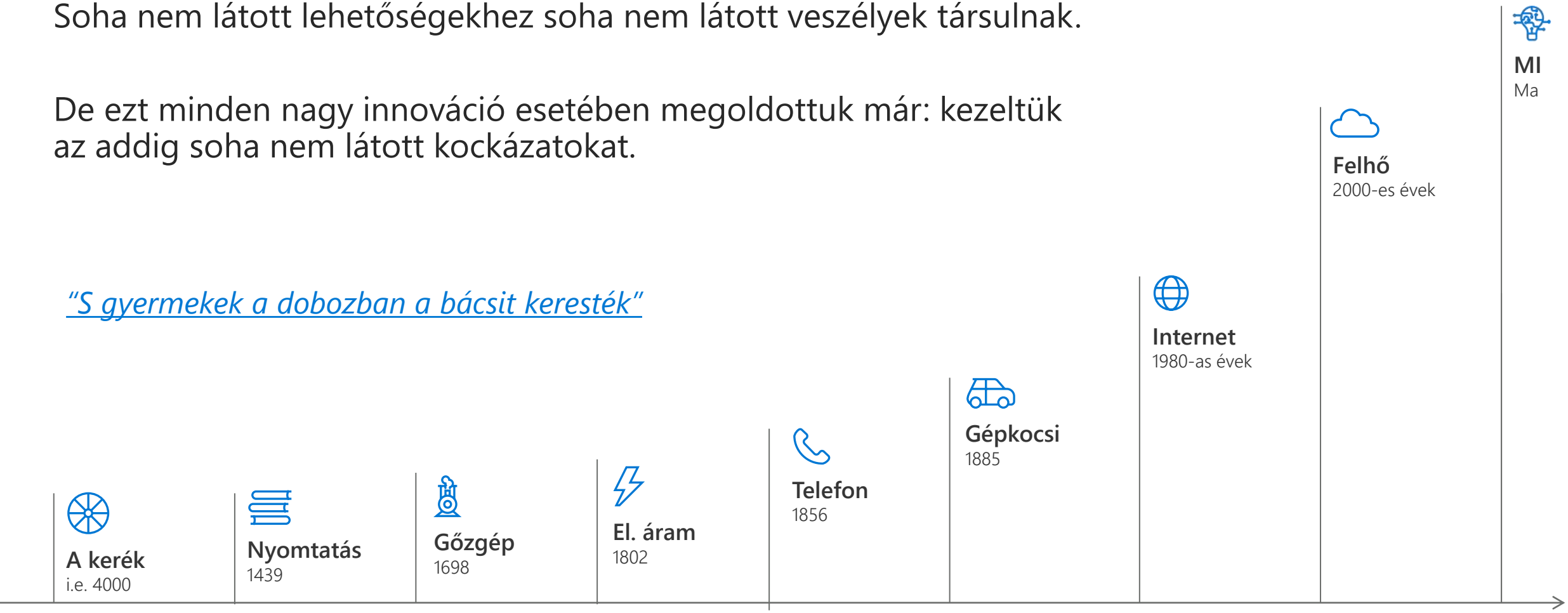


A mesterséges intelligencia a legnagyobb hatású innováció lehet az emberiség történetében

Soha nem látott lehetőségekhez soha nem látott veszélyek társulnak.

De ezt minden nagy innováció esetében megoldottuk már: kezeltük az addig soha nem látott kockázatokat.

"S gyermekek a dobozban a bácsit keresték"



Szabályozási analógia

Az automobil kitermelte a kockázatkezelés eszközeit:

- KRESZ
- oktatás társadalmi méretekben
- ittas vezetés tilalmának elfogadtatása a társadalommal
- töréstesztek és nyilvánosság
- műszaki vizsga, forgalmi engedély
- biztonsági öv – több évtizedes meggyőzés
- stb.

Ellenpélda: piros zászlós jelzőőr autó előtt

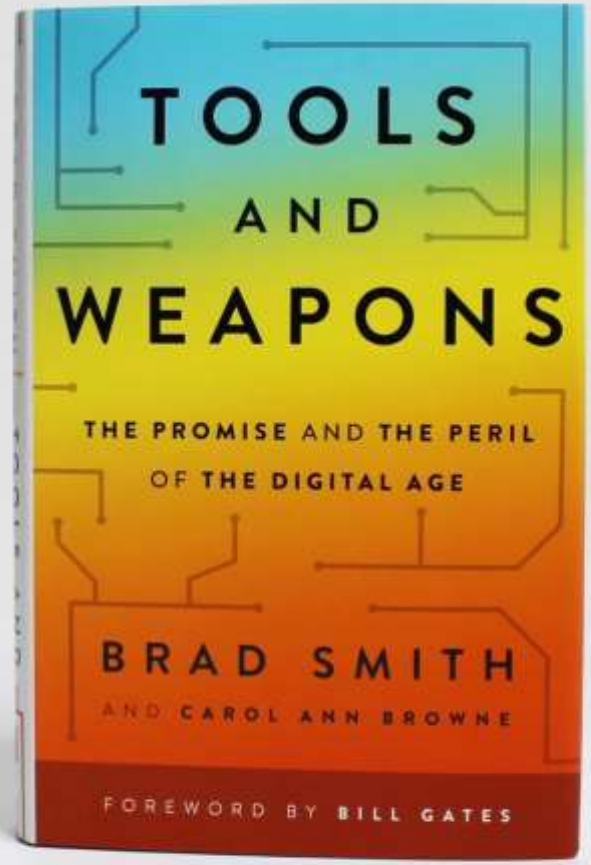
Hogyan alkalmazhatjuk ezeket a tapasztalatokat a mesterséges intelligencia szabályozásában?

Eszköz vagy fegyver?

“Minél hatalmasabb az eszköz, annál nagyobb előnyökkel vagy károkkal járhat az alkalmazása…

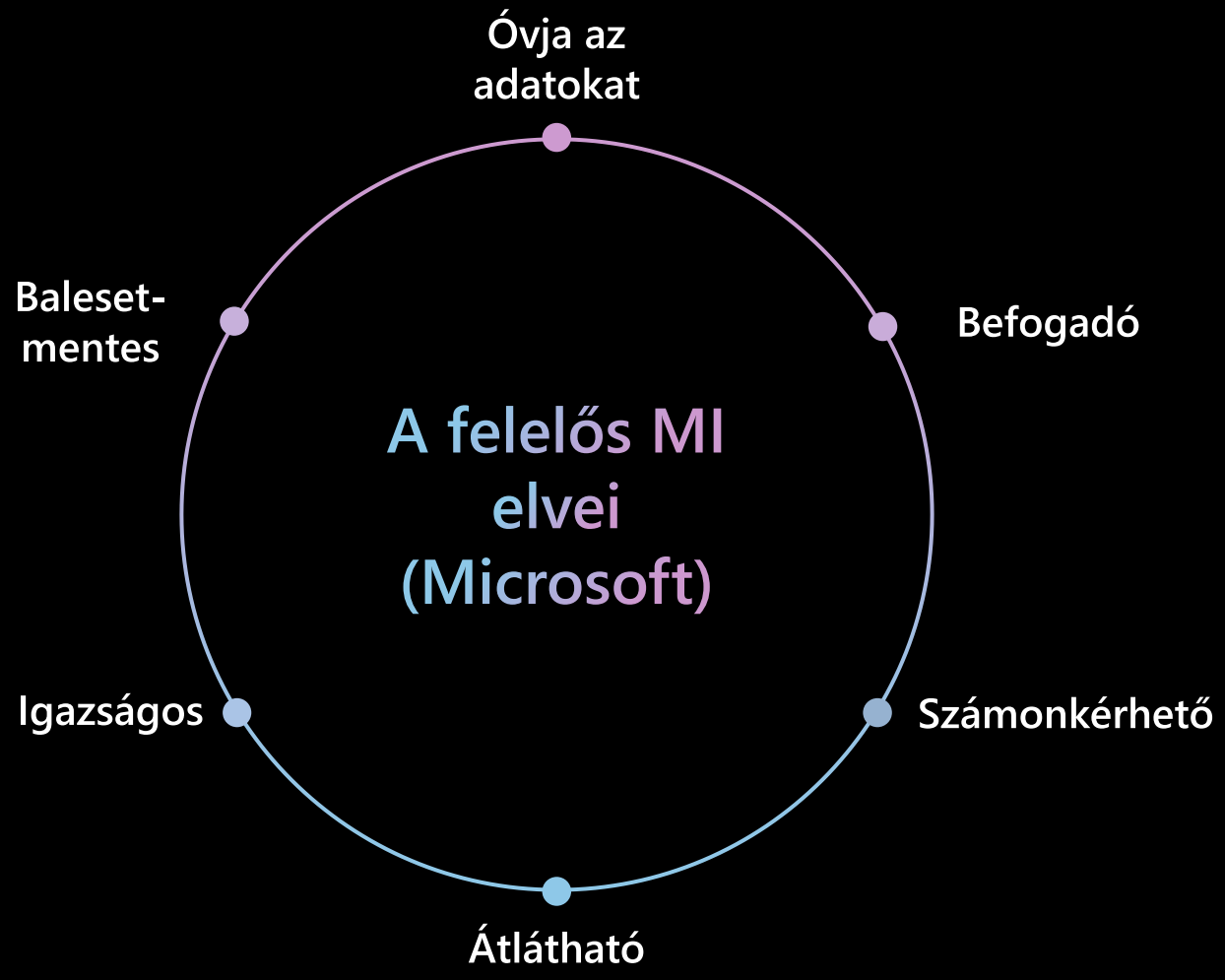
A technológiai fejlődés nem fog lelassulni. Annak a munkának kell felgyorsulnia, amellyel mederben tartható.”

Brad Smith
President and Chief Legal Officer, Microsoft



A mesterséges intelligencia felelős működése

”AI Alignment and AI Capabilities seem orthogonal. Progress in capabilities helps alignment. Alignment helps capability progress.” - Sam Altman, Open AI



Az elvek érvényre jutásához:

- Eszközök és folyamatok
- Oktatás és gyakorlat
- Szabályok, követelmények
- Irányítás (Governance)

Gyakorlatias eszköz: Responsible AI Standard

https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf

Accountability

A1: Impact Assessment
A2: Oversight of significant adverse impacts
A3: Fit for purpose
A4: Data governance and management
A5: Human oversight and control

Transparency

T1: System intelligibility for decision making
T2: Communication to stakeholders
T3: Disclosure of AI interaction

Fairness

F1: Quality of service
F2: Allocation of resources and opportunities
F3: Minimization of stereotyping, demeaning, and erasing outputs

Reliability & Safety

RS1: Reliability and safety guidance
RS2: Failures and remediations
RS3: Ongoing monitoring, feedback, and evaluation

PS1: Privacy Standard compliance
PS2: Security Policy compliance

I1: Accessibility Standards compliance

ACCOUNTABILITY

Példa a Responsible AI Standard alkalmazására

Goal A1: Impact assessment

Tools and Practices

Example Impact Assessment for Hospital Employee & Resource Optimization System (HEROS)

Stakeholders, potential benefits, and potential harms

The stakeholder analysis surfaces stakeholders and allows teams to consider how the system may impact them. Prompts related to the Goals of the Standard provide support for the identification of potential harms.

In the final section of the Impact Assessment, teams explore mitigations to the potential harms that have been identified.

[Microsoft's framework for building AI systems responsibly - Microsoft On the Issues]

Stakeholders, potential benefits, and potential harms

***2.2** Identify the system's stakeholders for this intended use. Then, for each stakeholder, document the potential benefits and potential harms. For more information, including prompts, see the **guidance & activities deck**.*

Stakeholders	Potential system benefits	Potential system harms
Hospital Administrators (end user)	Automated shift scheduling reduces this stakeholder workload that would otherwise have to create schedules manually. This stakeholder benefits from a simplified task that involves making manual changes to the schedule if necessary and approving the automated schedules.	*Reliability & safety*: If this stakeholder finds that they must make a lot of manual changes because of repeated system errors, they may abandon the system all together. *Accountability*: if a lot of nurses have negative feedback this stakeholder may feel disproportionately accountable and spend a long-time reviewing schedules and making manual changes. *Transparency & Accountability* – If this stakeholder does not have enough information about nurses and their personal constraints/needs they may not be able to make informed changes and may be tempted to always approve the automatically generated schedules.
Nurses	Improves the general feeling of fairness. Nurses feel like shifts are allocated systematically and eliminates the fear of favouritism when compared to a human-operated system.	*Fairness*: Unsuitable shifts that do not account for personal circumstances. For example, a parent might not be able to work night shifts, or someone might need to work more hours for financial reasons. *Reliability & safety*: Our expert interviews revealed that nurses are more efficient when they work with the same team. Separation of medical teams could lead to unhappy, inefficient teams.

Mit tehetünk a MI védelmében

Készüljünk fel a támadások fajtáira és természetükre

Failure Modes in Machine Learning (Harvard, Microsoft)

Szándékos - nem szándékos; sértetlenség – bizalmasság – hozzáférhetőség, stb

Egyes támadási fajták kivédhetők a MI előtti módszerekkel

Kockázatelemzés

Hozzáférési jogosultságok

IT biztonsági higiénia – szoftverfrissítés, kétfaktoros autentikáció, stb

Mások MI specifikusak

Modellek, tanítási folyamat stb. támadása

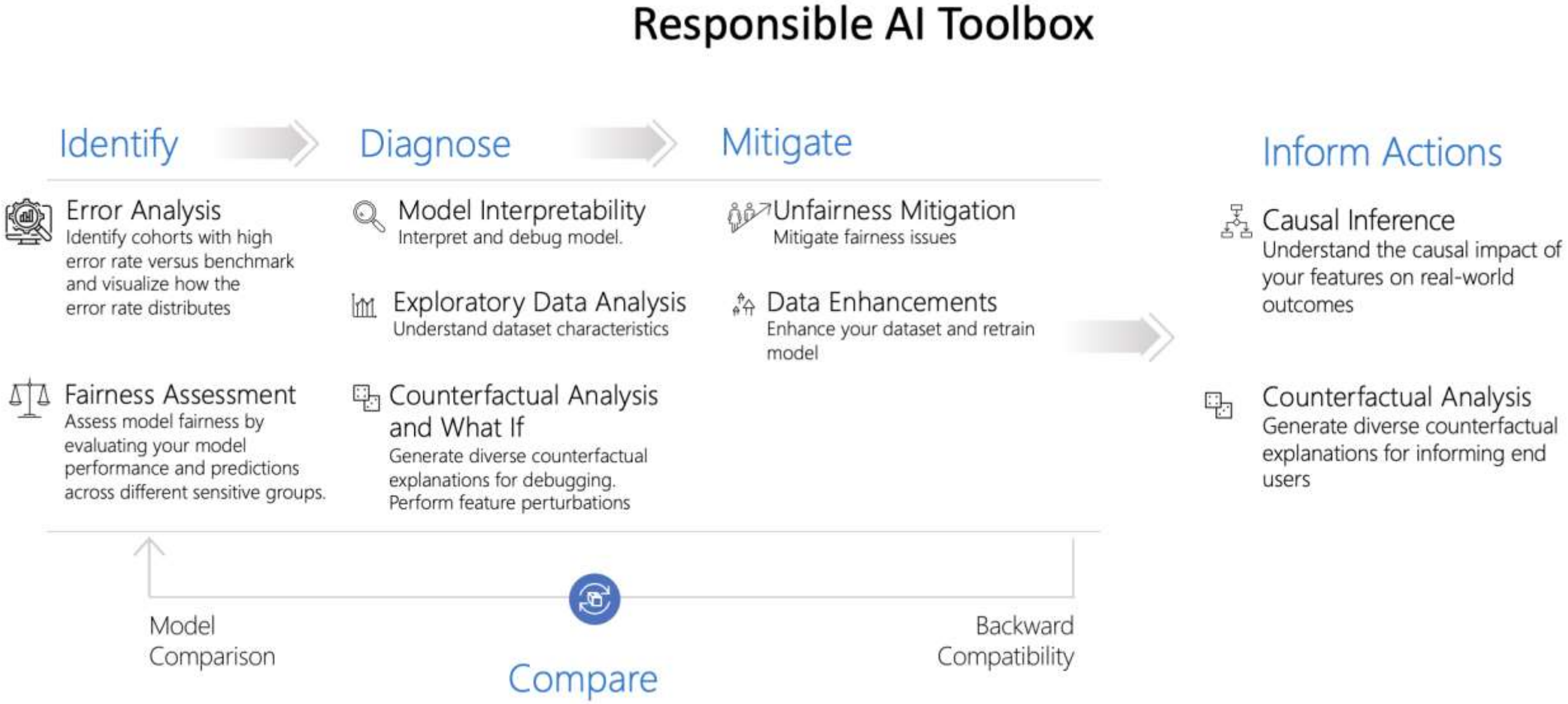
Attack type	Likelihood	Impact	Exploitability
Extraction	High	Low	High
Evasion	High	Medium	High
Inference	Medium	Medium	Medium
Inversion	Medium	High	Medium
Poisoning	Low	High	Low

Best practices for AI security risk management - Microsoft Security Blog

AI-Security-Risk-Assessment/AI_Risk_Assessment_v4.1.4.pdf at main · Azure/AI-Security-Risk-Assessment · GitHub

Eszközök

https://github.com/microsoft/responsible-ai-toolbox



A MI szunnyadó kérdéseket is felszínre hoz

Egyes kockázatok eddig is léteztek, de nem vettük észre őket. Kettős mérce: a gép előtt magasabbra tesszük a lécet, mert nagyobbnak tartjuk az etikátlan döntések veszélyét

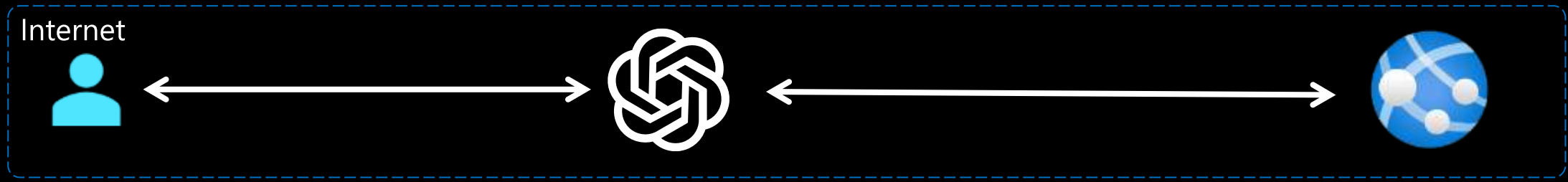
Three measurements of fairness: False positive rate, false negative rate, positive predictive value. Measure these 3 metrics for two subgroups. Each one is desirable. However cannot be satisfied simultaneously unless the base rate is the same. You and AI – the challenges to making machines play fair

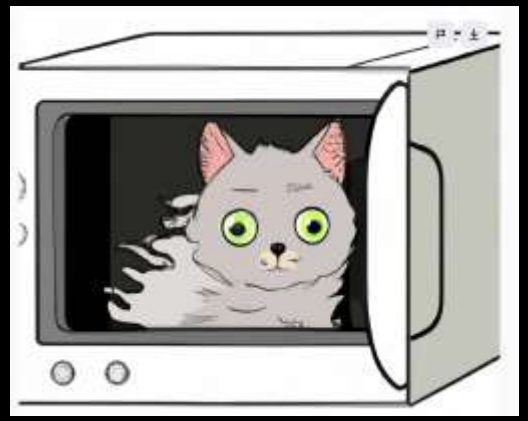
Más kockázatok az új technikai lehetőségek miatt kapnak gyakorlati jelentőséget

ChatGPT is a data privacy nightmare. If you've ever posted online, you ought to be concerned (ampproject.org)
(16) Nissenbaum's Theory of "Contextual Integrity" | LinkedIn

Kockázati térkép – mit vállal és mit nem a szolgáltató?

Példa: OpenAI vs Azure OpenAI vs Azure Cognitive AI





Tudatos felelősségelosztás. Az ügyfélnek kell kezelnie egyes kockázatokat.

Microsoft Azure Cloud

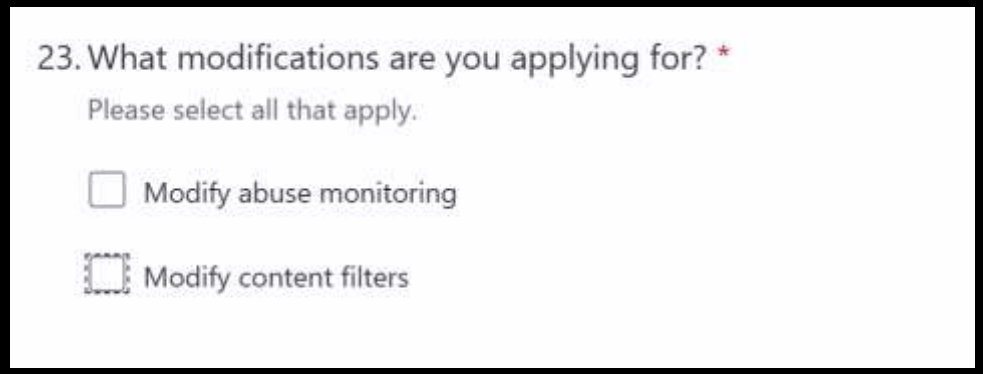
Nagyvállalati/intézményi bizalmi modell

A Te adatod a Tiéd	Az adatokat titkosítva tárolja az Azure előfizetésed
A továbbtanítások (finomhangolások) adatait nem használják modellek (elő)tanításához	Azure OpenAI szolgáltatás az ügyfél Azure előfizetésében létesül A továbbtanított modell az ügyfél Azure előfizetésében marad és soha nem kerül át az alap modellekhez
Az adataidat átfogó nagyvállalati/intézményi biztonsági intézkedések védik, a szabályozói megfelelés is cél	Ügyfél által kezelt kulcsok titkosítják (CMK) Magán VPN, szerepkör alapú hozzáférések (RBAC) SOC2, ISO, HIPPA, CSA STAR Compliant

Tudatos döntések és hatásuk a felelős működésre

A tartalom szűrése bizonyos esetekben kikapcsolható.

Azure OpenAI Limited Access Review:
Modified Content Filters and Abuse Monitoring

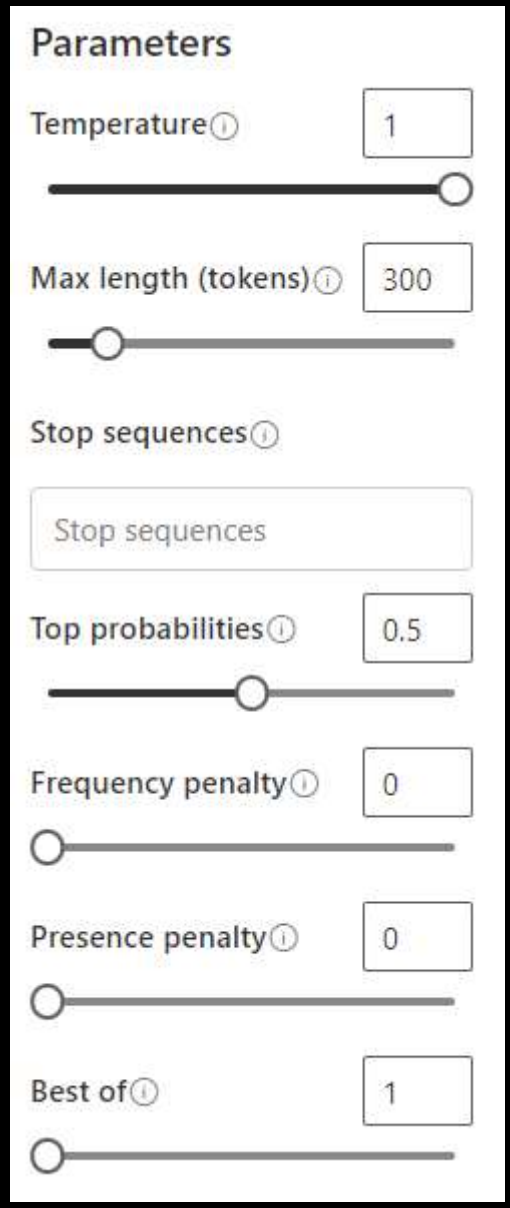


A válaszok “hőmérsékletének” megválasztása 0 és 1 között. Kreatív vagy kockázatkerülő legyen?

A felhasználó döntési lehetőségei és hatásuk a kockázatokra

- Max_length - The maximum number of tokens to generate in the completion. The token count of your prompt plus max_length can't exceed the model's context length.
- Temperature - What sampling temperature to use. Higher values means the model will take more risks. Try 0.9 for more creative applications, and 0 (argmax sampling) for ones with a well-defined answer.
- Top_p - An alternative to sampling with temperature, called nucleus sampling, where the model considers the results of the tokens with top_p probability mass. So 0.1 means only the tokens comprising the top 10% probability mass are considered

https://learn.microsoft.com/en-us/azure/cognitive-services/openai/reference#completions



EU Ethics Guidelines for Trustworthy AI:

"a hátrányos megkülönböztetést el kell kerülni, mert sokszorosan negatív következményei lehetnek, a sérülékeny csoportok marginalizálódásától az előítéletesség és a kizárás erősödéséig."

A magyar nyelvet értő modellekre szükség van a világnyelveken nem beszélők esélyegyenlősége érdekében.

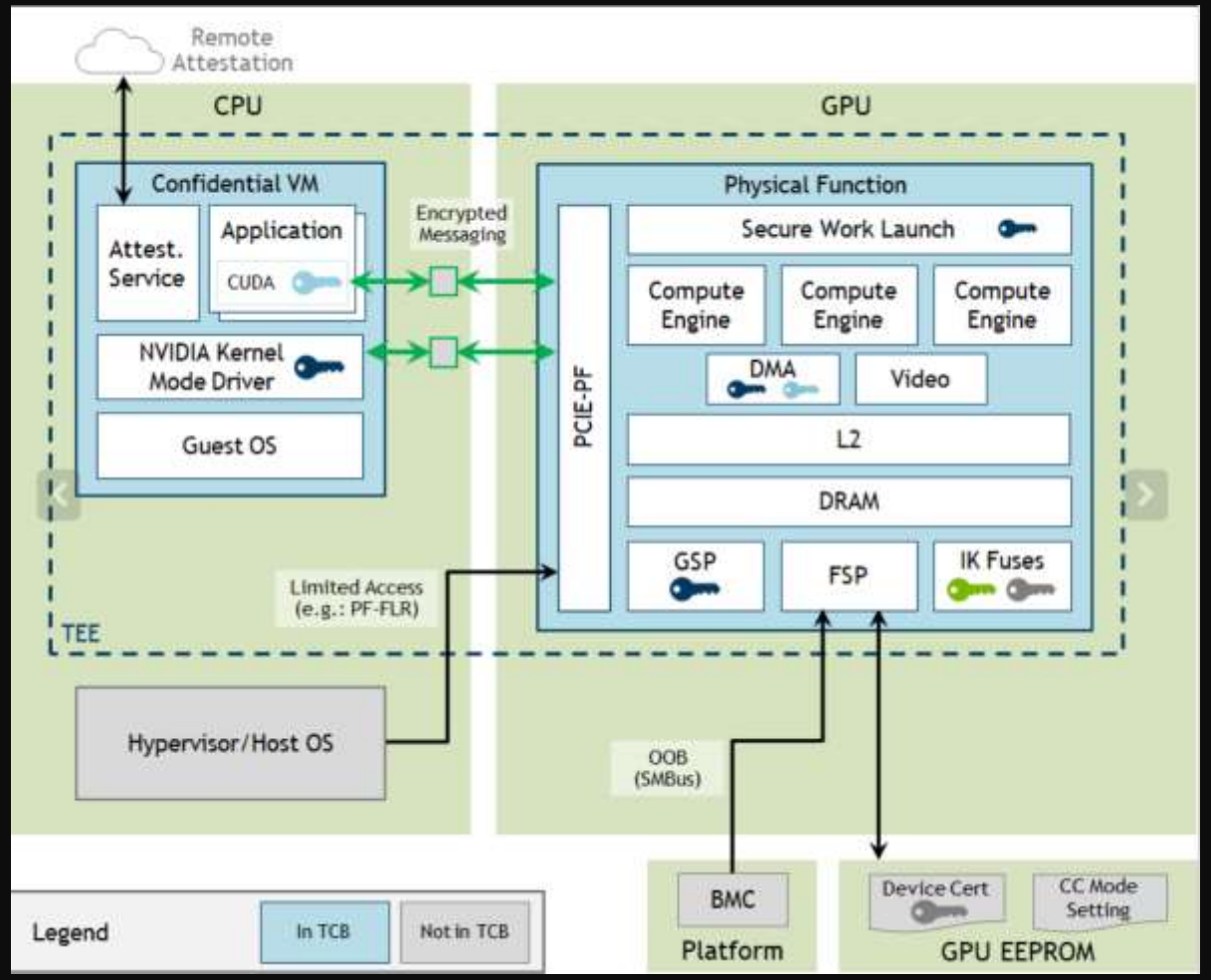
Bizalmi számítástechnika
(Confidential Computing)

Többszereplős adatelemzés

Ha a szereplők nem engednek a többieknek hozzáférést a saját adataikhoz

Védje és tartsa titkosítva az adatokat, amelyek **használatban vannak,** miközben a gép számításokat végez rajtuk.

Nem kell megbíznod a szolgáltatóban. A bizalmi számítástechnika megakadályozza, hogy a rendszergazda hozzáférjen az adatokhoz. Erről kriptográfiai bizonyítékok felmutathatók.



Hogyan vehetünk részt az ökoszisztémában?

A nagy nyelvi modellelk ökoszisztémája

A MI másképp működik, mint a hagyományos IT

Extrém infrastruktúra ☹

Prompt engineering ☺

Továbbtanítás ☺

Optimalizált pipeline ☺

Példa nagy nyelvi modellek többszereplős fejlesztésére

Értéklánc és ökoszisztéma létrehozásának módja

A megkülönböztetés eszközei

Promptok

- “Barátságos, informatív munkatárs vagy”
- “Csak az adatokból válaszolj”
- “Ha nem találsz rá választ, akkor válaszold ezt…”

Saját adatok

- Belső tudásbázisok
- Strukturált és strukturálatlan források
- Operációs és tranzakciós adatok

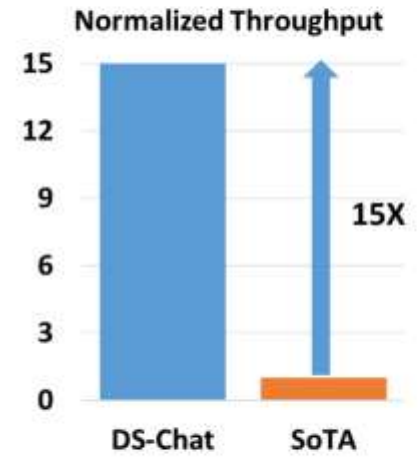
GPT-3.5 | GPT-4 | ChatGPT

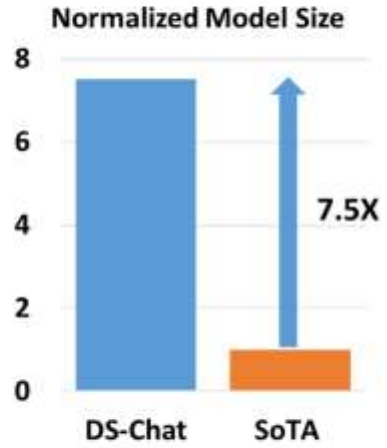
Azure OpenAI Service

DEEPSPEED CHAT

Fast Training with Affordable Cost

Train 15X Faster and Scale to 5x Bigger Models than SOTA RLHFs





Easy-Breezy Training

A complete end-to-end RLHF training experience with a single click

High Performance System

Hybrid Engine achieves 15X training speedup over SOTA RLHF systems with unprecedented cost reduction at all scales

Accessible Large Model Support

Training ChatGPT-Style models with tens to hundreds of billions parameters on a single or multi-GPUs through ZeRO and LoRA

A Universal Acceleration Backend for RLHF

Support InstructGPT pipeline and large-model finetuning for various models and scenarios

Kontextus

Context is King: In-context learning

Emergence

Egy modell, ami “szinte mindenre” képes. Alapjaiban változtatja meg a MI fejlesztésének dinamikáját, szerepköreit.

Azure OpenAI Service Learning Guide

Start Here!

- Create an Azure subscription
- ★ Apply for access to the Azure OpenAI Service: https://aka.ms/oai/access
- Watch the video: An Introduction to Azure OpenAI
- Get Familiarized with the product content: Azure OpenAI Page
- Review how other customers are using Azure AI

Basic

- Understand "What is Azure OpenAI?": compare Azure OpenAI and OpenAI. Review key concepts.
- ★ Work through Introduction to Azure OpenAI Service training module
- Explore key Responsible AI guidelines and principles

Intermediate

- How-to guides: Generate embeddings and Fine-tune
- Complete embeddings and document search tutorial
- How your/your customer data is processed, used and stored in Azure OpenAI: Data, privacy and security
- ★ Review the Enterprise Data with ChatGPT tech blog, and work through the accompanying GitHub repository.
- Become familiar with the APIs

Advanced

- Learn more in the Azure OpenAI Workshop
- ★ See examples in the Open AI Cookbook
- Start experimenting and developing your MVP
- Get support and help

„Ez a forradalom megállíthatatlan.

Ha megpróbálnánk, csak annyit érnénk el, hogy a jövőt átengedjük olyanoknak, akik bátrabban szembe néznek a saját újításaik következményeivel.”

Összegzés

Még kihasználatlan a védelem eszköztára

Az etikus és felelős használatról ne mondjunk le

Társadalmi mélységű erőfeszítés

A hozadékok nagyok, indokolt áldozni rá