# Minighid AI II


AdrianaS

# Ce este o halucinație?

IA nu minte. Dar poate inventa cu foarte multă încredere. Pare sigur pe el chiar și când spune prostii. Asta se numește „***halucinație***”. Nu e rea intenție (***nu are intenții***). Nu e „IA mincinos”. E doar matematică.
Practic, IA **generează un răspuns plauzibil, dar fals.**
Cum făceai și tu la școală când nu învățasei lecția....

# Exemple clasice

– inventează cărți/autori
– inventează citate
– inventează legi
– inventează rezultate „foarte sigure”
– relații (bine, asta fac și oamenii, dar cine să îi judece....)”

# De ce se întâmplă?

Pentru că un LLM nu știe adevărul. El doar prezice următorul cuvânt.
Dacă datele lipsesc → inventează.
Dacă e neclar promptul → generează ***„ce pare logic”***.
Dacă contextul e prea vag → umple golurile cu ce găsește.

# Convingător...

Pentru că IA este construit să:

- folosească un ton sigur,
- scrie coerent,
- structureze elegant,
- sune *„profesional”*.

**Coerența NU înseamnă adevăr.**

# Cum reduci halucinațiile

**Dă-i context clar.**
*Ambiguitate → imaginație.*
R**estrânge tema.**
*„Explică X în max. 5 rânduri.”*
*„Fără exemple inventate.”*
**Cere-i să îți dea clarificări.**
*„Dacă nu știi, întreabă.”*

# Cum reduci halucinațiile

**Cere surse verificabile.**
*„Dă-mi linkuri oficiale.”*
*„Citează legislație existentă.”*
**Învață-l să recunoască necunoscutul.**
*„Dacă informația nu există, spune «nu știu».”*
**Refă promptul când răspunsul e prea frumos.**
*Ton perfect + zero citate = semnal de alarmă.*

# Red flags

- nume inventate
- date exacte fără surse
- legi sau articole inexistente
- cifre rotunde „prea curate”
- fraze vagi ca „specialiștii spun că...” (așa-i că sună a politician?)

# Exemplu practic

**Prompt prost:** *„Explică-mi legea europeană despre AI.”* → inventează.
**Prompt bun:** *„Pe baza Regulamentului (UE) 2024/1689, rezumă principalele cerințe pentru modele generale. Citează articolele reale.”* → risc minim.