IA generativa i ús responsable

Pregunta a un xatbot qui va fundar el club de futbol del teu poble, i et respon en un segon: un nom, un any, una frase bonica sobre els fundadors. Sona segur. És cert?

De vegades sí. De vegades el nom, l'any i fins i tot la "font" que cita s'han inventat en aquell moment, escrits amb la mateixa fluïdesa que la veritat. Les empreses de les Illes Balears ja fan servir aquestes eines per respondre els hostes, escriure descripcions i fer esborranys de codi. El xatbot d'un hotel que s'inventa una política de cancel·lació no és un error divertit: és una reclamació, potser una demanda. Avui aprendràs com funcionen aquests models, com enxampar-los quan s'equivoquen i quines normes els ha posat Europa.

Què aprendràs

- Explicar què és la **IA generativa** i com escriu un **gran model de llenguatge**, sense màgia
- Enxampar una **al·lucinació** contrastant les respostes d'un model amb fonts reals
- Redactar normes per a l'**ús responsable** de la IA en una empresa
- Classificar sistemes d'IA segons els quatre nivells de risc del **Reglament d'IA**, i saber quines parts ja s'apliquen
- Evitar la trampa del dia: un model que escriu bé no és un model que sap

Models que creen

> La **IA generativa** és la IA que **produeix contingut nou** (text, codi, imatges, àudio, vídeo) a partir d'una instrucció, en lloc de només classificar o predir un valor.

El teu model de Teachable Machine de la lliçó anterior **classificava**: correcta o defectuosa. Un model generatiu **crea**: un paràgraf, una imatge, una funció.

> Un **gran model de llenguatge (LLM)** és una xarxa neuronal molt gran entrenada amb quantitats enormes de text per **predir el tros de text següent**. Repetint aquesta predicció una vegada i una altra, escriu respostes senceres.

Aquest és tot el truc. No cerca la resposta en una base de dades; escriu la continuació més probable del teu text, tros a tros. Al joc de construir ciutats, és un assessor que parla amb molta seguretat, fins i tot quan no ha mirat el mapa.

Terme	Què vol dir
Prompt	La instrucció o el text d'entrada que dones al model. Més context i més precisió donen millors respostes
Token	Un tros de text (una paraula o part d'una paraula) amb què treballa el model. Els preus i els límits es compten en tokens
Al·lucinació	Una resposta falsa o inventada, ben escrita i convincent
Data de tall del coneixement	El model només "sap" el que hi havia a les seves dades d'entrenament; el que és posterior no ho sap si no l'hi dones tu o si no cerca a la web

Què fa bé	Els seus límits
Escriure, resumir, traduir i reformular	Tot sol no comprova els fets: pot inventar dades, cites i fonts
Explicar i generar codi, proves i documentació	Pot repetir els **biaixos** de les seves dades d'entrenament
Treure informació de textos desordenats	La mateixa pregunta pot donar respostes diferents
Donar-te un primer esborrany ràpid	El que escrius pot sortir de l'empresa (privadesa); consumeix molta energia

> **Compte: "sona segur, per tant és correcte"**
> Un LLM escriu una resposta correcta i una d'equivocada exactament amb el mateix estil segur. El problema no és que falli: totes les eines fallen. El problema és que **falla sense avisar**. Per això tot el que produeix és un esborrany fins que una persona el revisa.

A classe: caçadors d'errors

Aquesta és l'**única activitat de la unitat en què la IA generativa està autoritzada**, i només en aquestes condicions:

> **Què està autoritzat i què no**
> **Autoritzat:** fer servir l'assistent d'IA que indiqui el professor, **a classe i només durant aquesta activitat**, per fer les preguntes de més avall i analitzar-ne les respostes.
>
> **No autoritzat:** redactar les teves conclusions amb IA, qualsevol altra activitat de classe i qualsevol part del Lliurament 4. No escriguis mai a l'assistent dades personals (teves o de ningú) ni dades internes de l'empresa.

Per parelles. Cada resposta s'ha de contrastar amb una **font real**: un web oficial, la premsa, la documentació del llenguatge, o executant el codi.

Pas	Pregunta a l'assistent	Comprova-ho amb
1	Tres preguntes sobre coses que coneixes molt bé: el teu poble o barri, el teu institut, un esport o un joc que domines	El que tu saps, confirmat amb un web
2	Dos fets sobre l'empresa del vostre projecte que **ja heu citat** al dossier (any de fundació, nombre d'hotels o de treballadors…)	Comparant-los amb la font del dossier
3	Una funció curta en el llenguatge que fas servir a Programació	Executant-la, amb un cas normal i un cas límit
4	Les fonts d'una de les seves respostes	Obrint els enllaços: existeixen? Diuen el que afirma el model?

Per a cada resposta, marca **correcta**, **equivocada** o **no es pot comprovar**, i enganxa l'enllaç que has fet servir per comprovar-la. Després posa nom al **tipus** d'error: fet inventat, font inexistent, codi que no funciona, resposta desfasada, biaix.

Normes per a un ús responsable

Dels errors que enxampis en surten cinc normes. La majoria d'empreses que fan servir IA escriuen una cosa semblant.

Norma	Què vol dir
Revisa-ho	Tot el que produeix un model és un esborrany fins que algú ho revisa
Protegeix les dades	Res de dades personals, ni codi o documents confidencials, en eines que l'empresa no ha aprovat
Sigues transparent	Digues quan has fet servir IA i per a què
El responsable ets tu	Qui signa la feina o publica el codi en respon, no l'eina
Respecta l'autoria	Respecta l'autoria i les llicències del contingut i del codi

El teu grup les convertirà en normes per a la **vostra** empresa, amb com a mínim una norma pròpia del seu sector. En un hotel, per exemple: "el xatbot no confirma mai una reserva ni un preu sense el sistema de reserves".

No és només una qüestió de bones maneres. Des del **2 de febrer de 2025**, el Reglament d'IA obliga les empreses que fan servir IA a assegurar-se que el seu personal en té prou coneixements, el que s'anomena **alfabetització en IA**.

El Reglament d'IA: normes segons el risc

> El **Reglament d'IA** és el **Reglament (UE) 2024/1689**, la llei europea sobre intel·ligència artificial. Va entrar en vigor l'**1 d'agost de 2024** i s'aplica per etapes.

Afecta sobretot els **proveïdors** (qui desenvolupa o ven un sistema d'IA) i els **responsables del desplegament** (qui en fa servir un de manera professional) a la Unió Europea. La seva idea és un **enfocament basat en el risc**: com més mal pot fer un sistema a les persones, més obligacions té. La Comissió Europea ho resumeix en quatre nivells.

Nivell	Què vol dir	Exemples
Inacceptable	Pràctiques prohibides	Puntuació social dels ciutadans; manipulació que aprofita vulnerabilitats; reconeixement d'emocions a la feina o a l'escola (llevat de motius mèdics o de seguretat)
Alt	Permès amb obligacions estrictes: gestió del risc, qualitat de les dades, documentació, registre, supervisió humana	Ordenar currículums per contractar persones, avaluar alumnes, decidir sobre préstecs, infraestructures crítiques
Limitat (transparència)	Obligacions de transparència	Els xatbots han de dir que parles amb una màquina; el contingut generat o manipulat (*deepfakes*) s'ha d'etiquetar
Mínim	Sense obligacions específiques	Filtres de correu brossa, IA als videojocs

Els grans models de llenguatge també tenen obligacions pròpies, com a **models d'IA d'ús general**: documentació, transparència i respecte dels drets d'autor, amb més obligacions per als més grans.

Què s'aplica, i des de quan

Data	Què s'hi comença a aplicar
1 d'agost de 2024	Entrada en vigor
2 de febrer de 2025	Pràctiques prohibides i alfabetització en IA
2 d'agost de 2025	Governança i obligacions dels models d'IA d'ús general

Data	Què s'hi comença a aplicar
2 d'agost de 2026	Aplicació general, incloses les obligacions de transparència
2 de desembre de 2027	Sistemes d'alt risc com la contractació, el crèdit o l'educació (ajornat per l'*Òmnibus d'IA*)
2 d'agost de 2028	Sistemes d'alt risc integrats en productes regulats

El calendari ja ha canviat una vegada

Les dates de l'alt risc es van fixar primer per al 2026 i el 2027. Una modificació, l'**Òmnibus d'IA**, va entrar en vigor el **27 de juliol de 2026** i les va ajornar. Quan escriguis sobre el Reglament d'IA, comprova les dates vigents a la pàgina de la Comissió Europea i cita-la: les lleis es mouen. I no el confonguis amb el **RGPD** (Reglament (UE) 2016/679), la llei de protecció de dades que veuràs a la Unitat 5.

A classe: quin nivell de risc?

Amb el grup del projecte, classifica cada cas i justifica-ho en una frase. Després classifica les dues idees d'IA que vas apuntar per a la vostra empresa a la lliçó anterior.

Cas	El teu nivell
Un xatbot que respon les preguntes dels clients al web d'un hotel	
Un programa que ordena currículums i descarta candidats	
El filtre de correu brossa d'una empresa de transport	
Un sistema d'una administració pública que puntua els ciutadans segons el seu comportament	
Una IA que corregeix exàmens i decideix qui entra en un curs	
Un vídeo promocional amb la veu d'una persona generada per IA	
Càmeres que detecten l'estat d'ànim dels treballadors d'una oficina	
Un model que decideix si un banc concedeix un préstec	

Lliura a Classroom (per parelles, "S16 · Caçadors d'errors i nivells de risc", fins a les 23:59 del mateix dia): la teva taula d'errors amb l'enllaç de cada comprovació, les cinc normes adaptades a la vostra empresa i la classificació dels vuit casos i de les vostres dues idees. Recorda: l'autorització de la IA s'ha acabat amb l'activitat de caçar errors.

Per al teu projecte

Aquesta lliçó alimenta la **part de grup** del Lliurament 4 · Pla d'IA:

La feina d'avui	Es converteix en	Criteri
Les vostres cinc normes, adaptades a l'empresa	Part de grup: normes d'ús responsable	c
El nivell de risc de les vostres idees	Part de grup: el nivell de risc segons el Reglament d'IA de cada proposta	c
Què pot fer i què no pot fer la IA generativa	Secció C: present i futur de la IA al sector	c, f

Idees clau

- La **IA generativa** crea contingut; un **LLM** escriu predint el tros de text següent, una vegada i una altra.
- Escriu igual de bé les respostes correctes i les equivocades: les **al·lucinacions** arriben sense avisar. Contrasta-ho sempre amb fonts.
- Ús responsable: **revisa-ho**, **protegeix les dades**, **sigues transparent**, **el responsable ets tu**, **respecta l'autoria**.
- El **Reglament d'IA** (Reglament (UE) 2024/1689) segueix un **enfocament basat en el risc**: inacceptable, alt, limitat (transparència), mínim.
- Les prohibicions i l'alfabetització en IA s'apliquen des del 2 de febrer de 2025; la transparència, des del 2 d'agost de 2026; les obligacions d'alt risc s'han ajornat al desembre de 2027 i a l'agost de 2028.
- En aquest mòdul, la IA **només** està permesa quan el professor l'autoritza per a una activitat concreta.

Comprovació ràpida

1 Què fa bàsicament un gran model de llenguatge quan respon?

A Cerca la resposta en una base de dades de fets verificats
B Prediu, tros a tros, la continuació més probable del text
C Cerca a internet i copia la millor pàgina
D Executa regles escrites per experts per a cada pregunta

2 Un xatbot et dona una cita d'un informe, amb el títol i l'any. Cerques l'informe i no existeix. Com se'n diu, d'això?

A Un token
B Un prompt
C Una al·lucinació
D Una data de tall del coneixement

3 Quines dues accions segueixen les normes d'ús responsable?

A Enganxar una llista de clients amb els seus telèfons en un xatbot públic per escriure correus més de pressa
B Dir al teu equip que un primer esborrany s'ha escrit amb una eina d'IA aprovada per l'empresa
C Provar una funció generada per un assistent abans d'afegir-la al projecte
D Publicar una resposta de la IA tal com és, perquè l'eina en és la responsable

4 Segons el Reglament d'IA, de quin nivell és un programa que ordena currículums i descarta candidats?

A Inacceptable: està prohibit
B Alt risc: permès amb obligacions estrictes com la supervisió humana
C Limitat: només ha de dir que és una IA
D Mínim: sense obligacions

5 Un hotel posa un xatbot al seu web. Què és el mínim que exigeix el Reglament d'IA?

- **A** Res: els xatbots són de risc mínim
- **B** Una llicència de la Comissió Europea
- **C** Dir als usuaris que parlen amb una IA
- **D** Apagar-lo de nit

6 Quina d'aquestes és una pràctica prohibida segons el Reglament d'IA?

- **A** Un filtre de correu brossa
- **B** Un videojoc amb enemics controlats per IA
- **C** Càmeres que reconeixen les emocions dels treballadors d'una oficina
- **D** Un recomanador en una botiga en línia

7 El teu grup proposa una IA que prediu quines màquines d'una fleca s'espatllaran la setmana que ve, a partir de les dades dels seus sensors. Quin nivell de risc hi encaixa millor, i què has de fer encara?

- **A** Inacceptable; has d'abandonar la idea
- **B** Probablement mínim: no decideix res sobre persones, però una persona n'hauria de revisar igualment els avisos
- **C** Alt, perquè fa servir sensors
- **D** Limitat, i per tant ha d'anunciar a les màquines que és una IA

Ja saps com escriu un model generatiu, per què l'has de revisar i què li demana Europa. Ara canvia de costat: ets l'empresa, i vols saber si la IA surt a compte. Següent: Automatització i rendibilitat.

Solucions

1. B · Un **LLM** és un predictor de text. Per això pot escriure una resposta convincent que és falsa: res de la predicció no comprova els fets.

2. C · Una resposta inventada però convincent és una **al·lucinació**. Les fonts inventades en són una de les formes més habituals, i per això sempre obres els enllaços.

3. B, C · **Sigues transparent** i **revisa-ho**. Les dades personals no van mai a eines no aprovades, i qui publica és el responsable, no l'eina.

4. B · Les decisions de contractació afecten la vida de les persones, i per això el Reglament d'IA les classifica com d'**alt risc**: gestió del risc, qualitat de les dades, documentació i **supervisió humana**.

5. C · Els xatbots són al nivell **limitat (transparència)**: les persones han de saber que parlen amb una màquina. Aquestes obligacions de transparència s'apliquen des del **2 d'agost de 2026**.

6. C · El reconeixement d'emocions **a la feina o a l'escola** és una pràctica de risc **inacceptable** (llevat de motius mèdics o de seguretat), prohibida des del 2 de febrer de 2025.

7. B · Una predicció de manteniment no decideix sobre els drets de les persones, i per això encaixa en el risc **mínim**. L'ús responsable continua valent: algú ho **revisa** abans d'actuar, i les dades han de ser bones.