Guia de Bones Pràctiques per al Desenvolupament d'Eines d'IA Generativa en Salut

Grans Models de Llenguatge (LLM)

Generalitat de Catalunya

Aquesta guia ha estat elaborat per l'Àrea d'Intel·ligència Artificial de la Fundació TIC Salut Social

Autors: Susanna Aussó, Antoni Berenguer, Júlia Aznar, Carolina Raventós, Vaneza Gómez i Maria Bretones, amb la col·laboració de l'equip de HEALTH de NTT DATA

Edició electrònica: Febrer de 2025

Contingut

1 / Context

El **Programa per a la Promoció i Desenvolupament de la Intel·ligència Artificial al Sistema de Salut de Catalunya (**Programa Salut/IA**)** té com a objectiu crear un entorn que faciliti la innovació en l'àmbit de la salut. Mitjançant el desenvolupament i la implementació de solucions d'Intel·ligència Artificial (IA), aquest programa busca millorar la salut de la ciutadania, prioritzant la prevenció, liderant la implantació d'aquestes tecnologies i contribuint a la qualitat assistencial i la sostenibilitat del sistema sanitari, tot posant en valor el coneixement generat pel Sistema sanitari integral d'utilització pública de Catalunya.

La **Fundació TIC Salut Social** ha creat aquesta guia per oferir pautes clares als professionals i organitzacions implicats en el desenvolupament d'algorismes d'IA generativa aplicats a la salut. A causa de la ràpida evolució i popularització d'aquesta tecnologia en els darrers anys, el document se centra especialment en les eines basades en **Grans Models de Llenguatge** *(Large Language Models*, **LLM**) i en la creació de solucions segures, eficients i d'alt valor afegit, alineades amb els estàndards més avançats en intel·ligència artificial.

L'avenç de la IA, que es basa en la creació de sistemes capaços de processar informació i generar respostes que imiten el raonament humà, ja és una realitat. Aquests sistemes poden aprendre de l'experiència, resoldre problemes en contextos delimitats i dur a terme tasques complexes en diversos àmbits de la societat.

En l'àmbit de la salut, la creixent disponibilitat de registres electrònics de salut, imatges mèdiques digitals, dades òmiques i altres conjunts de dades ofereix un gran potencial per millorar el benestar de les persones[1]. Els mètodes d'Aprenentatge Automàtic (*Machine Learning*), i especialment les tècniques d'Aprenentatge Profund (*Deep Learning*), permeten desenvolupar algorismes avançats capaços d'aprendre a partir d'aquesta diversitat de dades[2].

Tanmateix, l'ús de tècniques d'IA generativa en l'àmbit de la salut presenta reptes importants, especialment en la seva implementació, integració a la pràctica clínica i adaptació a les especificitats d'aquest sector. Cal abordar qüestions clau com la supervisió humana, la qualitat de les dades, la interpretabilitat dels resultats i la protecció de la privacitat. La naturalesa innovadora d'aquesta tecnologia requereix una coordinació estreta entre professionals de la salut, investigadors i experts en tecnologia per garantir una aplicació efectiva i segura, ètica i lícita, i alineada amb les necessitats reals tant dels professionals com dels pacients.

La confiança dels i les professionals de la salut en les solucions d'IA és un factor clau, i s'ha de fonamentar en principis com la transparència i el rigor. L'àmbit de la salut presenta reptes no només científics, sinó també ètics i legals, ja que les decisions preses tenen un impacte directe en el benestar i la vida de les persones[1]. Per això, la fiabilitat de les eines d'IA s'ha de construir sobre tres components:

IA legal	Complint totes les lleis i regulacions aplicables
IA ètica	Assegurant els principis i valors ètics
IA robusta	Des d'una perspectiva tècnica (garantint la solidesa de les eines) i social (tenint en compte l'entorn en què operen)

L'objectiu d'aquesta guia és **proporcionar un marc general per a l'ús i la integració dels LLM**, com a part de la IA generativa, **en el sector de la salut.** Està pensada per facilitar la comprensió i aplicació d'aquestes tecnologies en els processos assistencials, garantint-ne un ús ètic i responsable, fomentant la col·laboració i impulsant la innovació en Salut.

Objectius de la guia

OBJECTIU 01 /

Formar i capacitar els professionals de la salut en l'ús dels LLM

Proporcionar els coneixements i habilitats necessaris per comprendre i aplicar aquests models de llenguatge dins de l'àmbit sanitari i biomèdic, promovent-ne un ús informat, segur i efectiu.

OBJECTIU 02 /

Guiar la implementació d'eines LLM en els processos assistencials

Facilitar la integració d'eines basades en LLM en els fluxos de treball dins de l'àmbit sanitari, des de la recollida i gestió de dades fins a la generació de contingut, amb l'objectiu de millorar l'eficiència i la qualitat de l'activitat assistencial.

OBJECTIU 03 /

Promoure la col·laboració interdisciplinària en l'aplicació dels LLM

Fomentar la cooperació entre professionals sanitaris, investigadors i experts en IA per desenvolupar solucions innovadores, i millorar de forma contínua les eines basades en LLM que ja han estat implementades.

OBJECTIU 04/

Facilitar la comprensió i selecció de models d'IA generativa basats en LLM

Oferir una visió clara dels diferents tipus de models de llenguatge disponibles, així com criteris per seleccionar els més adequats segons aspectes tècnics, d'avaluació i d'implementació.

OBJECTIU 05 /

Garantir l'ús ètic i responsable dels LLM

Vetllar pel respecte a la privacitat dels pacients, el compliment de les normatives vigents, i la gestió adequada dels riscos associats a l'ús de models de llenguatge en contextos clínics i biomèdics.

2 / IA Generativa

2.1 / Què és la IA Generativa?

La IA generativa és una **branca de la IA que se centra en la creació de continguts nous i originals**, com ara textos, imatges, música, vídeos, dades estructurades i altres formats.

Com la resta de models d'IA, els models generatius són **predictius** des d'un punt de vista funcional, ja que generen contingut preveient la següent paraula, imatge o element a partir de patrons apresos.

No obstant això, la IA generativa es distingeix per la seva capacitat de **crear contingut nou,** sovint amb una aparença creativa, a diferència d'altres models que es limiten a classificar o recomanar (com ara els models de regressió lineal, regressió logística, *random forest*, *support vector machines* o xarxes neuronals, entre altres).

A més d'aquesta capacitat per generar contingut, la IA generativa **es diferencia de la resta de models d'aprenentatge automàtic en altres aspectes rellevants**, com ara la quantitat i característiques de les dades que es necessiten, els requeriments computacionals, o els desafiaments associats a la interpretabilitat dels resultats.

GPU: Unitats de Processament Gràfic (Graphics Processing Unit)
TPU: Unitats de Processament de Tensor (*Tensor Processing Unit)*

Diferències entre la IA generativa i altres models d'aprenentatge automàtic

Característiques	IA Generativa	Altres models d'aprenentatge automàtic
Capacitats	Creació de contingut nou, com ara text, imatges o sons. Poden adaptar-se per realitzar tasques predictives.	Funcions de predicció, com ara classificació o recomanació. No poden generar continguts nous.
Comprensió de llenguatge natural	Comprensió profunda del llenguatge incloent-hi context, matisos semàntics, negacions o ironia.	Comprensió bàsica del llenguatge, sovint limitada a patrons simples o diccionaris.
Tipus de tasques	Generació de text, imatge o so, resum de documents, traducció automàtica. Predicció, classificació o extracció d'informació.	Tasques específiques com classificació, regressió o recomanació.
Generalització	Models flexibles capaços d'afrontar una gran varietat de tasques amb un únic entrenament previ sobre dades massives no etiquetades (*pre-training*).	Models específics per a una única tasca. Sovint requereixen reentrenament complet per abordar nous problemes.
Preparació de dades	Aprenen les característiques directament dels grans volums de dades no estructurades. Ajustament manual mínim.	Requereixen l'enginyeria de variables (*feature engineering*), sovint manual, per seleccionar, transformar i estructurar les dades de manera útil per al model.
Tipus d'aprenentatge i dades	Aprenentatge autosupervisat, on el model prediu la següent paraula en una seqüència a partir de dades no etiquetades. Sovint inclou una fase d'alineament amb dades etiquetades.	Aprenentatge supervisat amb dades etiquetades (amb informació anotada).
Cost computacional	Altament exigents: requereixen grans recursos computacionals per entrenar (milions o milers de milions de paràmetres).	Menys exigents (solen tenir molts menys paràmetres).
Infraestructura tecnològica	Requereixen infraestructures massives, amb GPU, TPU i memòria distribuïda per realitzar un entrenament i una inferència eficients.	En molts casos, es poden entrenar en equips petits (fins i tot ordinadors portàtils).
Procés d'entrenament	Aprofiten el coneixement preentrenat sobre dades massives, que els dona un bon rendiment general. Es poden adaptar a tasques més específiques.	Models entrenats des de zero per realitzar tasques específiques.
Interpretabilitat	Baixa a causa de la seva alta complexitat i nombre de paràmetres. L'explicabilitat en IA generativa és actualment una àrea activa de recerca.	Alta en models senzills (p.e., regressions i arbres de decisions). Moderada en models més complexos, amb suport de tècniques d'IA explicable (XAI).

2.2 / Els Grans Models de Llenguatge (LLM)

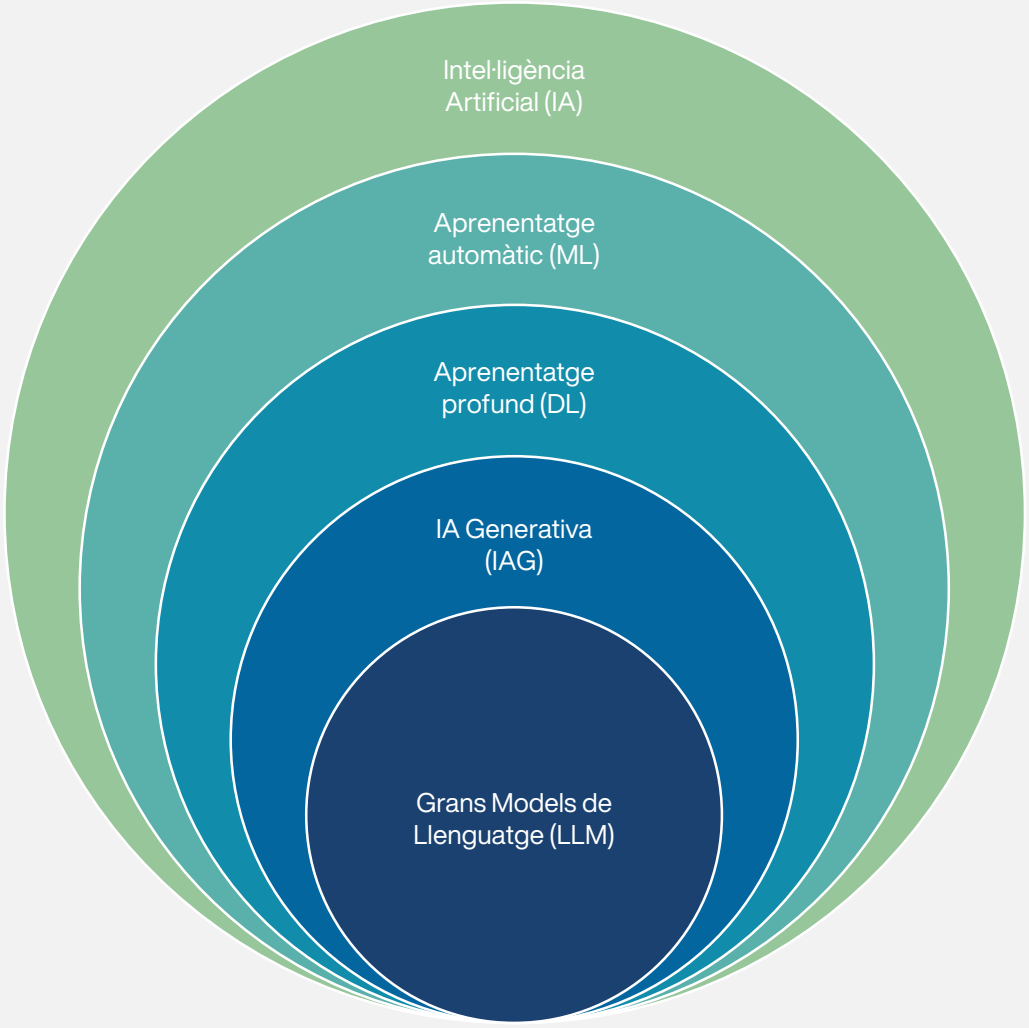
Actualment, els **Grans Models de Llenguatge** (*Large Language Models*, **LLM**) són una de les tecnologies més rellevants dins la IA generativa.

Tot i que **treballen principalment amb llenguatge natural** (és a dir, el llenguatge humà), aquests models també **poden generar altres formats**, com ara codi de programació o representacions matemàtiques, si han estat prèviament exposats a aquestes dades.

Les funcionalitats dels LLM es basen en la **predicció (inferència) iterativa de la següent unitat de text** (com ara una paraula o una frase) mitjançant l'anàlisi de patrons i d'estructures lingüístiques, gramaticals i semàntiques apreses a partir de volums massius de dades textuals (**preentrenament**).

Aquest entrenament els dota d'una capacitat notable per **generar, traduir o completar textos de manera precisa i entenedora**, fet que ha impulsat aplicacions avançades com els assistents virtuals, la redacció automàtica de documents, la generació de resums o la traducció automàtica.

Tipologies d'Intel·ligència Artificial



Tècnicament, els LLM es fonamenten en l'aprenentatge profund i en arquitectures com els **transformadors** (*transformers*). Aquestes arquitectures han estat clau en l'evolució del processament del llenguatge natural (*Natural Language Processing,* NLP), ja que han permès superar moltes de les limitacions de les tecnologies anteriors, com les xarxes neuronals recurrents (*Recurrent Neural Networks*, RNN).

En particular, les RNN tenien dificultats per captar relacions a llarg termini en textos extensos perquè processaven la informació seqüencialment, fet que limitava la seva capacitat d'analitzar contextos globals.

Els transformadors aborden aquesta limitació gràcies als **mecanismes d'atenció**, que són procediments algorítmics avançats **capaços d'analitzar el context complet d'un text de manera paral·lela**. Aquesta capacitat no només permet manipular grans volums de dades amb eficiència, sinó que també facilita la interpretació de contextos complexos. Aquestes propietats han fet dels transformadors una peça fonamental en les aplicacions d'IA generativa actuals.

Les **Unitats de Processament Gràfic** (*Graphics Processing Unit*, **GPU**) i les **Unitats de Processament de Tensor** (*Tensor Processing Unit*, **TPU**) han estat els altres elements tècnics clau per fer els LLM una realitat.

Les **GPU**, dissenyades inicialment per al processament de gràfics, són processadors optimitzats per **realitzar càlculs en paral·lel**, cosa que permet executar milions d'operacions simultàniament.

Les **TPU**, per la seva banda, són processadors desenvolupats per *Google* per **optimitzar les operacions amb tensors**, que són estructures de dades essencials en l'aprenentatge profund.

Ambdues tecnologies han accelerat l'entrenament dels models, reduint els temps de càlcul i facilitant la gestió dels grans volums de dades necessaris en aquests processos.

És important assenyalar que, tot i que els LLM estan dissenyats per ser generatius, no sempre s'utilitzen amb aquesta finalitat. Molts LLM també poden aplicar-se a tasques no generatives, com ara la classificació, l'extracció d'informació o l'anàlisi de sentiments. Aquesta **combinació de predicció i generació de continguts** els converteix en eines de gran valor i potencial impacte per al desenvolupament de solucions a totes les àrees del sector de la salut (veure 2.3/ Potencial dels LLM en Salut).

Finalment, cal dir que la IA generativa abasta un ampli espectre que va més enllà del llenguatge natural i que inclou altres modalitats, com ara la generació d'imatges, sons i vídeos. A més, alguns models d'IA generativa, coneguts com a **models multimodals**, són capaços de **combinar diferents tipus de dades** en un únic model. Per tant, tot i que els LLM són exemples destacats d'IA generativa, aquesta tecnologia va molt més enllà del processament del llenguatge natural.

2.3 / Potencial dels LLM en salut

Els LLM presenten un gran potencial per aportar solucions innovadores i eficients als reptes actuals en l'àmbit sanitari i biomèdic[3].

Gràcies a la seva capacitat per analitzar grans volums de dades i generar contingut textual, aquests models poden tenir un impacte transformador en el sector de la salut, incloent-hi l'automatització i millora de processos en l'activitat assistencial, la salut pública, la gestió sanitària i la recerca científica.

Àmbits d'ús dels LLM al sector de la salut

ÀMBIT 01 /

Activitat assistencial

Els LLM poden contribuir a millorar l'eficiència i la qualitat de l'atenció sanitària mitjançant la generació d'informes clínics, l'automatització de la documentació mèdica i el suport a la comunicació entre pacients i professionals.

Exemples dels LLM en l'activitat assistencial, amb el pacient com a usuari final

- **Bot mèdic per a pacients:** assistent interactiu que respon a preguntes generals sobre salut i proporciona recursos personalitzats per millorar la informació i la formació dels pacients en matèria de salut.
- **Resums mèdics per a pacients:** generació d'informes incloent-hi explicacions de diagnòstics i tractaments en llenguatge clar i accessible, per facilitar la comprensió i millorar la comunicació entre pacients i professionals sanitaris.
- **Traducció de documents mèdics:** facilita l'accés a la informació sanitària per a pacients de diferents orígens lingüístics, millorant la comprensió i afavorint una comunicació més efectiva en l'atenció assistencial.

Exemple dels LLM en l'activitat assistencial, amb el professional com a usuari final

- **Assistents virtuals per al diagnòstic i tractament:** proporcionen recomanacions de diagnòstic i plans terapèutics personalitzats basats en l'estat de salut del pacient, en l'evidència científica i en guies clíniques dinàmiques actualitzades, donant suport als professionals mèdics en la presa de decisions.
- **Prescripció automàtica de proves de laboratori:** generació automàtica de suggeriments de proves de laboratori en funció de l'historial mèdic i dels símptomes del pacient.
- **Transcripció simultània de consultes mèdiques:** enregistrament, transcripció en temps real i estructuració automàtica de les converses entre pacients i professionals sanitaris. Aquesta funcionalitat facilita la creació d'informes clínics estructurats, millora la qualitat de la documentació clínica i optimitza l'experiència de la consulta.
- **Resums clínics automàtics:** generació d'informes a partir de les visites i historials mèdics, facilitant l'accés a la informació rellevant i agilitzant la presa de decisions clíniques.
- **Autoanamnesis:** generació automàtica d'històries clíniques estructurades a partir de la informació aportada pel pacient, facilitant un accés ràpid, organitzat i codificat per als professionals sanitaris.

ÀMBIT 02 /

Salut pública

Els LLM representen una eina poderosa per a la gestió de la salut pública, facilitant l'anàlisi de grans volums de dades i la comunicació personalitzada per promoure la salut.

Exemples dels LLM en salut pública

- **Comunicació sanitària personalitzada:** creació de materials informatius adaptats a diferents nivells d'alfabetització, idiomes i contextos culturals, amb l'objectiu de reforçar les iniciatives de salut pública.
- **Anàlisi automàtica de comentaris públics sobre polítiques sanitàries:** resum i síntesi de grans volums d'opinions ciutadanes, per avaluar la percepció pública i l'impacte de noves normatives en l'àmbit de la salut.
- **Identificació de patrons de salut a partir de dades no estructurades:** anàlisi combinada de textos mèdics, xarxes socials i dades ambientals per detectar tendències de salut poblacional i identificar de manera precoç riscos i patrons emergents.
- **Suport a la revisió de documentació en inspeccions sanitàries:** automatització de l'anàlisi i validació de documents per millorar l'eficiència dels processos d'inspecció, facilitant la detecció d'incompliments normatius, incoherències o anomalies en informes i registres sanitaris.

ÀMBIT 03 /

Gestió sanitària

Els LLM poden jugar un paper clau en la millora de l'eficiència operativa de les institucions sanitàries, gràcies al seu potencial per optimitzar la gestió de recursos i augmentar la precisió en els processos administratius.

Exemples dels LLM en la gestió sanitària

- **Codificació automàtica de diagnòstics:** interpretació i codificació d'informes clínics, per facilitar la facturació, optimitzar l'administració i estalviar temps als professionals de la salut.
- **Avaluació de la qualitat de l'atenció:** anàlisi automatitzada dels resultats clínics i del *feedback* dels pacients, per mesurar la qualitat dels serveis assistencials i identificar àrees de millora.
- **Generació automatitzada de pressupostos i informes de rendiment:** automatització de l'elaboració de pressupostos i de l'anàlisi de rendiment, millorant la coordinació i l'eficiència entre departaments.

ÀMBIT 04 /

Recerca clínica i biomèdica

Els LLM tenen un gran potencial per accelerar el progrés científic en salut, facilitant l'anàlisi de dades complexes i la generació de nou coneixement.

Exemples dels LLM en recerca

- **Cribratge i selecció en assaigs clínics:** definició automatitzada de criteris d'elegibilitat i identificació de pacients per participar en estudis, agilitzant el reclutament i execució dels assaigs clínics.
- **Identificació d'esdeveniments adversos:** anàlisi automatitzada de grans volums de dades, per detectar i monitorar possibles efectes adversos en tractaments experimentals.
- **Revisió automatitzada de bibliografia:** resum i síntesi d'articles científics per facilitar la revisió de literatura rellevant, estalviant temps als investigadors i agilitzant l'accés a coneixement actualitzat.
- **Creació de dades sintètiques:** generació de dades artificials per enriquir els conjunts de dades d'entrenament, millorant la capacitat de les eines d'IA per fer prediccions precises i generar contingut divers. Aquesta estratègia és especialment valuosa en àrees on les dades reals són escasses o sensibles.
- **Generació d'hipòtesis:** formulació d'hipòtesis noves a partir de l'anàlisi de patrons en grans conjunts de dades mitjançant models avançats. Aquesta funcionalitat potencia la investigació científica i la presa de decisions mèdiques, permetent explorar opcions de tractament o diagnòstic basades en l'evidència.

2.4 / Desenvolupament d'una eina LLM en Salut

El desenvolupament d'una eina basada en LLM segueix una sèrie de passos que inclouen el **disseny de la solució, la implementació, l'avaluació i el seu desplegament**.

Aquest procés, sovint requereix la participació d'un **equip multidisciplinari** format per diferents perfils professionals, com ara desenvolupadors, personal mèdic, gestors sanitaris o els mateixos pacients, entre d'altres.

En els següents capítols, analitzarem els aspectes clau de cada fase, centrant-nos en aquells elements específics o especialment rellevants per al desenvolupament d'eines en salut basades en LLM.

Fases i aspectes clau del desenvolupament d'una eina LLM en Salut

DISSENY	IMPLEMENTACIÓ	AVALUACIÓ	DESPLEGAMENT
Identificació de la necessitat	Selecció del model • Models generalistes comercials • Models generalistes de codi obert • Models adaptats a l'àmbit sanitari i biomèdic • Models locals o remots • Petits Models de Llenguatge (SLM)	Eficàcia del model • Mètriques intrínseques • Mètriques extrínseques	Fluxos de treball
Adequació de la tecnologia			Modulació
Definició de la intenció d'ús	Adaptació del model • Enginyeria de *prompts* • Generació Augmentada per Recuperació • Afinament • Com triar el mètode d'adaptació? • Dades per l'adaptació	Detecció i mitigació de biaixos	Sistemes tecnològics
Funcions d'explicabilitat			Usabilitat
Funcions de supervisió humana		Detecció i mitigació d'al·lucinacions	Capacitació
Privacitat i protecció de dades			Monitorització
Regulacions i certificacions	Agents i eines multiagent	Detecció i prevenció de fugues de dades	Seguretat
			Actualització

3 / Disseny

En comparació amb altres tecnologies, **les eines basades en LLM introdueixen reptes i consideracions específiques que cal abordar des de les primeres etapes del disseny.** En aquest capítol, es detallen cadascun d'aquests aspectes que són clau en el disseny d'una eina robusta, fiable i alineada amb les necessitats de l'àmbit sanitari.

3.1 / Identificació de la necessitat

El desenvolupament d'una solució tecnològica comença per identificar i descriure amb claredat la necessitat dins l'àmbit de treball, en aquest cas el sector de la salut.

Aquesta fase inicial implica **verificar que el problema abordat existeix, és rellevant i prioritari** per al sistema sanitari i per al benestar de la ciutadania.

En aquest context, és molt recomanable realitzar **estudis d'impacte social, econòmic i ètic**, per tal de determinar la rellevància de la necessitat i contextualitzar-la dins dels plans de salut vigents establerts per les autoritats.

3.2 / Adequació de la tecnologia

L'entusiasme creat per la IA generativa i el seu gran potencial no hauria de desviar el focus d'allò que és veritablement important, que és la solució d'un problema real en l'àmbit de la salut. L'ús de la IA generativa hauria d'estar justificat per allò en què pot contribuir en aquesta solució, evitant utilitzar-la només per ser una tecnologia "innovadora". Atès que la implementació d'aquestes tecnologies pot ser complexa i costosa, és important **considerar si altres estratègies més senzilles, com ara la IA no generativa o els sistemes basats en regles, podrien oferir solucions igualment eficaces.**

En aquest procés, les **revisions de l'estat de l'art** són especialment útils, ja que permeten avaluar l'eficàcia de les tecnologies disponibles per abordar la necessitat, i verificar que el problema no hagi estat ja resolt per una solució existent al mercat. Aquestes revisions, a més, ajuden a identificar possibles punts de millora o innovació en les tecnologies actuals.

3.3 / Definició de la intenció d'ús (*intended use*)

Un cop identificada la necessitat que volem abordar, el següent pas és traduir-la en objectius i funcionalitats. Tot i que els LLM són flexibles i capaços d'afrontar una gran varietat de tasques, és molt important definir objectius clars, específics i mesurables. Aquesta concreció és important per dos motius:

- Permet definir la descripció d'ús i prevenir usos no previstos (*functional creep*). L'eina només s'avaluarà pels objectius establerts i, per tant, l'ús fora d'aquests no en garanteix la seva eficàcia i pot afectar la confiança dels usuaris i la seguretat dels pacients.
- Proporciona mecanismes per avaluar si l'eina compleix els objectius mitjançant mètriques adequades, que han d'estar definides de forma precisa i es revisaran en detall en seccions posteriors (veure 5/ Avaluació).

La descripció de l'ús de la solució inclou l'especificació dels seus **objectius i funcionalitats, així com els seus usuaris finals, els procediments d'ús, les limitacions de l'eina i els riscos associats.**

Exemple de factors que composen la intenció d'ús

Necessitat identificada (punt de partida)	Reduir la càrrega administrativa del personal mèdic relacionada amb la redacció de documents clínics.

FACTOR 01 /
Objectius específics

Automatitzar la redacció d'informes mèdics a partir dels registres electrònics de salut, amb una qualitat equivalent a la dels humans i reduint en un 40% el temps que dedica el personal clínic a aquesta tasca.

FACTOR 02 /
Usuari final

El personal clínic, com ara mèdic i d'infermeria.

FACTOR 03 /
Funcionalitats

Identificació d'elements rellevants de la història clínica, generació d'informes estructurats, integració amb històries clíniques electròniques, especialització en terminologia mèdica, incloent-hi diagnòstics, patologies, símptomes, pautes de tractament, i noms comercials i principis actius de medicaments.

FACTOR 04 /
Procediments d'ús

Inici del procés a partir d'una instància en l'aplicació de registre electrònic, revisió de l'esborrany generat per l'eina, i validació final per part del professional abans de guardar o compartir l'informe.

FACTOR 05 /
Limitacions

Dependència de la qualitat de les dades als registres electrònics, i necessitat de supervisió final per part d'un professional clínic per validar el contingut generat.

FACTOR 06 /
Riscos associats

Sobreconfiança en l'eina per part del personal clínic, resultats discriminatoris per biaixos a les dades d'entrenament, errors en la interpretació de dades o informes incomplets, que poden comprometre la seguretat dels pacients si no es detecten i corregeixen a temps.

3.4 / Funcions d'explicabilitat (XAI)

3.4.1 / L'explicabilitat en IA

La complexitat dels models d'IA fa que els mecanismes de decisió utilitzats pels algorismes sovint siguin desconeguts, fins i tot per als seus propis desenvolupadors. Les **tècniques d'IA Explicable** (*Explainable AI*, XAI) ofereixen eines que permeten **garantir la traçabilitat de les respostes d'un algorisme,** i assegurar que tant els resultats generats com els processos que els han produït siguin comprensibles per als humans.

L'explicabilitat en IA contrasta amb el concepte de "caixa negra" (*black box*), on els mecanismes que intervenen per generar una resposta específica (*output*) a partir d'una entrada (*input*) són desconeguts o inaccessibles.

En l'àmbit de la IA, assolir aquesta transparència és fonamental per garantir el desenvolupament de sistemes que siguin fiables, robustos i alineats amb els principis ètics i legals.

Aspectes clau de l'explicabilitat en IA

ASPECTE 01 /
Rendiment i eficàcia

Comprendre com el sistema genera un determinat contingut permet **optimitzar el codi i els components de l'eina** (*debugging*). Això facilita la identificació de possibles biaixos, al·lucinacions i àrees de millora, contribuint així a augmentar el seu rendiment i eficàcia global.

ASPECTE 02 /
Biaixos socials i equitat

Per la naturalesa de les dades amb que s'han entrenat, **els LLM poden incorporar biaixos discriminatoris que poden incidir negativament en les decisions clíniques.** Les eines d'explicabilitat permeten identificar aquestes desviacions i prendre accions correctives, assegurant un tracte just i equitatiu per a tots els pacients, independentment de les seves característiques demogràfiques, ètniques o socioeconòmiques.

ASPECTE 03 /
Compliment de la regulació

Les tècniques d'explicabilitat aporten transparència, traçabilitat i comprensió humana de les respostes del sistema, facilitant així la **verificació del seu compliment amb els requisits normatius establerts.**

ASPECTE 04 /
Confiança de professionals i pacients

Per generar confiança, **els resultats d'un model no només han de ser precisos, sinó també comprensibles, justos i objectius.** Els professionals de la salut necessiten entendre els processos pels quals el model arriba a les seves conclusions, per poder integrar-les adequadament en la presa de decisions i identificar possibles errors. Per als pacients, aquestes qualitats augmenten la sensació de seguretat i afavoreixen un diàleg obert i transparent amb els professionals.

ASPECTE 05 /
Supervisió humana

El **Reglament Europeu d'IA** estableix que els sistemes d'IA d'alt risc han d'estar sotmesos a una supervisió humana proporcional als riscos associats. Això implica donar als usuaris la capacitat de detectar anomalies, identificar biaixos, interpretar resultats i, quan calgui, ignorar o aturar les recomanacions del sistema, garantint una intervenció segura i efectiva. La implementació d'aquestes capacitats forma part de l'àmbit de l'explicabilitat, i es tractarà en detall en seccions posteriors (veure 3.5/ Funcions de supervisió humana).

3.4.2 / Tècniques d'explicabilitat en LLM

En el cas dels LLM, **assolir l'explicabilitat d'un sistema és especialment complex** a causa de les particularitats d'aquests models. Factors com la seva arquitectura avançada, la gran quantitat de dades utilitzades durant l'entrenament (sovint desconegudes per l'usuari) i el format de les dades d'entrada i sortida, habitualment en forma de text lliure, poden dificultar aquesta transparència.

Les **tècniques clàssiques de XAI**, com ara la importància de les variables de permutació (*permutation feature importance*), els valors de les Explicacions Additives Shapley (*Shapley Additive Explanations*, SHAP) o les Explicacions Agnòstiques del Model d'Interpretació Local (*Local Interpretable Model-agnostic Explanations,* LIME) han estat desenvolupades per a models d'aprenentatge automàtic no generatius. Aquests mètodes **sovint presenten problemes de càlcul, escalabilitat i formalització quan s'apliquen als LLM** i, per aquest motiu, l'explicabilitat en aquests models segueix sent una àrea activa de recerca.

Tècniques de XAI més populars per a LLM

TÈCNICA 01 /
Tècniques d'atribució

Identifiquen quines parts de l'entrada contribueixen més a la sortida generada pel model, cosa que permet una millor comprensió de com el model processa la informació. Per exemple, *TransSHAP* i els gradients integrats (*integrated gradients*) són mètodes d'atribució que han estat adaptats als LLM.

TÈCNICA 02 /
Tècniques de *prompting*

Aprofiten la capacitat dels LLM per **generar explicacions en llenguatge natural**, facilitant la comprensió dels processos que han donat lloc a una resposta concreta. Per exemple, la tècnica de la **cadena de pensament** (*chain-of-thought*) permet que el model exposi pas a pas els raonaments que justifiquen la seva resposta, en lloc de limitar-se a donar només una resposta final (veure 4.2.1/ Enginyeria de *prompts*).

TÈCNICA 03 /
Tècniques basades en atenció

Analitzen els patrons d'atenció del model per identificar en quines parts de l'entrada es focalitza durant el processament. Tot i estar en debat en l'actualitat, mètodes com ara **l'anàlisi de flux d'atenció** ofereixen una visió detallada sobre com es distribueix l'atenció entre els diferents elements de l'entrada.

TÈCNICA 04 /
Tècniques d'interpretació dels components

Busquen entendre el **paper de components individuals del model**, com neurones, capes o capçaleres d'atenció. Per exemple, les tècniques de *probing* analitzen quina informació lingüística codifiquen determinades parts del model, cosa que permet explorar com es representen conceptes específics dins la seva arquitectura.

TÈCNICA 05 /
Explicacions basades en mostres

Relacionen les prediccions del model amb mostres de les dades d'entrenament o d'adaptació. En els LLM, això pot implicar identificar documents o fragments de text que han influït en una resposta concreta.

3.5 / Funcions de supervisió humana

A més de ser un requeriment establert pel Reglament Europeu d'IA (podeu consultar la Guia per l'Aplicació del Reglament Europeu d'IA en Salut del Programa Salut/IA), la implementació de capacitats de supervisió humana és una condició fonamental per garantir l'eficàcia de l'eina i la confiança dels professionals sanitaris i dels pacients, així com a l'adherència del sistema a principis ètics i regulacions vigents.

Com ja s'ha mencionat, les funcionalitats de supervisió humana s'emmarquen en l'àmbit de la IA explicable (XAI), i comporten aspectes essencials del disseny de l'eina. Aquesta necessitat és especialment rellevant en els LLM, ja que la seva explicabilitat és més complexa en comparació amb altres models d'IA (veure 3.4/ Funcions d'explicabilitat).

Aspectes clau de la supervisió humana

ASPECTE 01 /

Format de la sortida dels resultats

Per exemple, presentar recomanacions en forma de llistes ordenades, en lloc de propostes úniques, pot **facilitar la revisió de les respostes de l'eina** per part del professional de la salut, com a pas previ a utilitzar aquesta informació en la presa de decisions.

ASPECTE 02 /

Procés d'obtenció de la resposta

Per tal de facilitar el procés de revisió al professional clínic, l'eina ha de proporcionar els **elements i processos rellevants** en la generació de la resposta. Aquesta funcionalitat es pot basar en mètodes de *prompting* o en tècniques d'explicabilitat.

ASPECTE 03 /

Informació influent en la resposta

Les tècniques basades en mostres permeten identificar **documents o fragments de text que han influït en la resposta.** Aquests fragments poden provenir de l'entrenament, de l'afinament o d'un sistema RAG (*Retrieval-Augmented Generation*). Incloure aquesta informació en la sortida de l'eina és especialment útil per ajudar al professional clínic a revisar els resultats.

ASPECTE 04 /

Generació d'avisos per incertesa

Sovint, els LLM poden basar les seves respostes en informació incerta, incompleta o fins i tot inexistent. Implementar tècniques de *prompting* o estratègies basades en mostres pot ajudar el sistema a detectar aquestes situacions i alertar l'usuari sobre la necessitat de revisar els resultats.

Un exemple representatiu és utilitzar instruccions de *prompting* per **indicar al sistema que només respongui a la consulta si troba la resposta** en la documentació proporcionada; en cas contrari, el sistema hauria d'emetre un avís indicant que no pot garantir la fiabilitat de la resposta, o retornar un missatge com: 'Dades insuficients'.

ASPECTE 05 /

Interfície centrada en la revisió

Les interfícies de l'eina han d'estar **adaptades i optimitzades perquè els professionals puguin revisar i validar fàcilment els resultats** generats. Això és especialment rellevant en àmbits crítics com el sanitari, on la supervisió humana és essencial per garantir la fiabilitat de les decisions i la seguretat dels pacients.

3.6 / Privacitat i protecció de dades

Les dades són un component essencial d'una eina basada en IA i un dels factors més determinants en el disseny de la solució. Per aquest motiu, és molt important definir, ja des de les primeres fases del desenvolupament, els següents aspectes clau:

- **Quines dades processarà l'eina o estaran implicades en el seu funcionament:** inclou tant les dades d'entrada i de sortida com a les dades utilitzades per adaptar el model de llenguatge (veure 4.2/ Adaptació del model).
- **Si aquestes dades són de naturalesa sensible o privada:** aquest aspecte és especialment rellevant en sectors com el de la salut, on la seguretat i privacitat de les dades són prioritàries.
- **Quin serà el flux de les dades durant el funcionament de l'eina:** cal determinar si aquestes dades estaran circumscrites als sistemes de l'entitat usuària en tot moment durant el funcionament de l'eina o, si pel contrari, necessitaran ser transferides a servidors externs. Un exemple habitual és l'ús per part de l'eina d'un LLM comercial allotjat als servidors d'un proveïdor (veure 4.1.4/ Models locals o remots).

Mecanismes aplicables per garantir la seguretat i confidencialitat de les dades

01	L'ús de models instal·lats localment, que evitin la transferència de dades a serveis externs
02	L'anonimització o pseudoanonimització de dades
03	L'encriptació de les dades
04	La gestió segura d'accessos
05	La creació d'un registre d'activitats
06	Garantir un ús mínim necessari de dades
07	L'elaboració d'un protocol de retenció i eliminació de les dades
08	Garantir als pacients l'accés, rectificació i eliminació de dades personals dins el sistema

L'ús de dades en una eina d'IA està sotmès a normatives com ara el Reglament General de Protecció de Dades (RGPD), l'Esquema Nacional de Seguretat (ENS) en entitats públiques, o la Llei Orgànica de Protecció de Dades i Garantia dels Drets Digitals (LOPDGDD), entre d'altres. Aquestes regulacions exigeixen establir mecanismes per garantir la seguretat i confidencialitat de les dades, especialment quan es tracta de dades sensibles o privades.

A més, l'ús de dades sensibles obliga a realitzar una **avaluació d'impacte relativa a la protecció de dades** (DPIA), per tal d'identificar i gestionar els riscos associats al tractament d'aquestes dades amb tecnologies avançades com ara la IA generativa.

3.7 / Regulacions i certificacions

Ja des de les primeres fases del disseny, és important tenir en compte **quin tipus d'eina es vol desenvolupar, si es considerarà un producte sanitari, i quina classificació de risc tindrà segons les normatives vigents.** Aquests aspectes són importants perquè afecten directament les certificacions necessàries i les regulacions que haurà de complir l'eina abans de poder ser utilitzada en entorns clínics.

Cal tenir present que el compliment d'aquests requisits i l'obtenció de certificacions poden ser **processos llargs i complexos.** Per això, identificar aquestes necessitats des de l'inici permet planificar adequadament els recursos i els terminis, evitant endarreriments en el desplegament final del sistema. A més, aquesta planificació garanteix que l'eina es desenvolupi de manera alineada amb els estàndards de qualitat, seguretat i eficiència requerits pel sector de la salut.

Per exemple, si l'eina es desenvolupa per ajudar en el diagnòstic o tractament de pacients, probablement es classificarà com a producte sanitari i haurà de complir amb una sèrie de condicions incloses al Reglament Europeu de Dispositius Mèdics (MDR) i la Regulació Europea d'IA (EU AI Act). Aquest procés implica una llista de requisits que inclouen, per exemple, la realització d'avaluacions clíniques, l'obtenció de marcatge CE i la verificació del compliment de normes harmonitzades.

Normatives més rellevants per a solucions basades en IA en salut

NORMATIVA 01 /

Reglament Europeu d'IA (EU AI Act)

Defineix el marc jurídic per garantir l'ús segur i ètic de la IA a Europa. Classifica les eines d'IA pel seu nivell de risc (inacceptable, alt, limitat o mínim) i estableix obligacions i requisits per a la seva utilització al mercat europeu (podeu consultar la Guia per a l'Aplicació del Reglament Europeu d'IA en Salut del Programa Salut/IA).

NORMATIVA 02 /

Reglaments de Dispositius Mèdics (MDR) i Dispositius Mèdics In Vitro (IVDR)

Imposen estàndards de seguretat i eficàcia per garantir que els dispositius mèdics compleixen els requisits necessaris abans de ser aprovats (podeu consultar les Guies MDR i IVDR del Programa Salut/IA).

NORMATIVA 03 /

Reglament General de Protecció de Dades

Regula el tractament de les dades personals dels pacients, incloent-hi les dades de salut. Estableix condicions per garantir la privacitat i la seguretat de les dades sensibles, amb requisits de transparència, traçabilitat i supervisió humana en decisions automatitzades (veure 3.6/ Privacitat i protecció de dades).

Altres regulacions

És important considerar altres regulacions i estàndards que poden aplicar-se en funció del context específic de cada dispositiu o sistema d'IA generativa. Alguns exemples són:

- Esquema Nacional de Seguretat (ENS Real Decreto 311/2022)
- Directiva sobre la Seguretat General dels Productes (Directiva 2001/95/CE)
- Directiva sobre la Protecció de Dades (Directiva UE 2016/680)
- Llei sobre la Protecció de Dades en les Institucions de la UE (Reglament UE 2018/1725)
- Llei de Ciberseguretat (Reglament UE 2019/881)
- Llei de Governança de Dades (Reglament EU 2022/868)
- Llei de Serveis Digitals (Reglament UE 2022/2065)
- Llei de Mercats Digitals (Reglament EU 2022/1925)
- Llei de Dades (Reglament UE 2023/2854)

4 / Implementació

4.1 / Selecció del model

En el desenvolupament d'eines basades en LLM, sovint **es parteix de models ja preentrenats** que reben el nom de **models fundacionals.** Aquests models es construeixen mitjançant un procés d'entrenament que requereix volums massius de dades textuals i recursos computacionals extraordinaris. Aquest és un procediment extremadament costós que, en contrapartida, els proporciona una base sòlida de comprensió del llenguatge i els dota d'un vast rang d'aplicacions, incloent-hi en Salut. Aquesta estratègia no només accelera el desenvolupament de les eines, sinó que també millora l'eficiència i la qualitat dels seus resultats.

En el desenvolupament d'aquest tipus de solucions, és habitual l'ús de més d'un LLM que es poden complementar segons les necessitats de cada aplicació. Aquesta estratègia permet, per exemple, combinar models de propòsit general amb models més petits i especialitzats, optimitzant tant el rendiment com la gestió de recursos.

La selecció del(s) model(s) de llenguatge és un pas fonamental que pot influir de manera significativa en la qualitat i l'eficàcia de l'eina. A continuació, es presenten els criteris més rellevants per facilitar aquesta elecció.

Criteris per a la selecció del model de llenguatge

CRITERI 01 /
Modalitat de dades

Alguns models gestionen dades d'un únic tipus (p.e., text o imatges), mentre que d'altres tracten dades multi-modals. Cal considerar el tipus de dades disponibles en el context sanitari i l'objectiu que es vol assolir.

CRITERI 02/
Rendiment del model

És important que el model triat hagi demostrat un bon rendiment (precisió, velocitat de processament, etc.) en tasques similars a les que es volen desenvolupar, tant en la literatura com en la pràctica clínica.

CRITERI 03/
Adaptabilitat

És la capacitat d'un model per ser ajustat per a la realització de tasques específiques. Generalment, els models comercials ofereixen menys flexibilitat; els models de codi obert, en canvi, són altament adaptables, però requereixen coneixements tècnics avançats.

CRITERI 04 /
Actualitzacions i manteniment

Els models amb suport continu per desenvolupadors o comunitats actives són especialment atractius, ja que reben millores i actualitzacions constants que n'incrementen l'eficàcia i la seguretat.

CRITERI 05 /
Mida i complexitat del model

La quantitat de paràmetres d'un model en determina la capacitat per captar patrons complexos en les dades. Els models més grans ofereixen millor rendiment, però requereixen més recursos computacionals i econòmics; els models més petits són més eficients, però poden presentar limitacions de rendiment en algunes tasques.

CRITERI 06 /
Cost i infraestructura

Els models comercials solen implicar un pagament per ús, augmentant el cost a mesura que creix el volum de dades; els models de codi obert poden reduir el cost inicial, però requereixen més inversió en infraestructura i coneixement tècnic per al desplegament i manteniment.

CRITERI 07/
Documentació i suport

La disponibilitat de documentació, tutorials i comunitats actives és un factor a tenir en compte, ja que facilita la implementació, l'ús i el manteniment del model.

CRITERI 08 /
Impacte ambiental

Els models més grans solen requerir una gran quantitat d'energia per a l'entrenament i el desplegament, el que incrementa la seva empremta de carboni.

Tenint en compte aquests criteris, cal avaluar les diverses opcions de LLM disponibles al mercat, que es diferencien pel **tipus de llicència** (comercial o de codi obert), **el grau d'especialització en l'àmbit sanitari i biomèdic** (general o adaptat), la seva **mida i complexitat**, així com la possibilitat de desplegar-los en equips **locals o entorns remots.**

4.1.1 / Models generalistes comercials

Els models comercials desenvolupats per grans empreses tecnològiques ofereixen un alt rendiment, són fàcils d'utilitzar i compten amb suport tècnic constant per part de l'empresa creadora.

En general, **es comercialitzen mitjançant una API de pagament per ús.** La persona usuària té poc control sobre el model, ja que l'entrenament i el desenvolupament són gestionats íntegrament per l'empresa i no és possible modificar-lo ni adaptar-lo de forma substancial per a usos molt específics. Tot i ser generalistes, poden mostrar un bon rendiment en tasques específiques dins del sector de la salut.

LLM generalistes comercials (febrer, 2025)

Model (desenvolupador)	Modalitat de dades	Descripció
GPT-3 / 4 / 4.5 (OpenAI)	Text / Multimodal	GPT-3 va ser un dels primers LLM d'ús generalitzat, suposant un gran avenç en la generació de text natural i en la comprensió del llenguatge. GPT-4 va representar la seva evolució, incorporant capacitats multimodals per processar text i imatges. GPT-4.5 és la versió més recent i pretén millorar el rendiment dels seus antecessors.
GPT-4o (OpenAI)	Multimodal	Versió optimitzada de GPT-4. Millora l'eficiència, la velocitat i la precisió, oferint un rendiment superior per a aplicacions complexes que requereixen comprensió i generació de contingut multimodal.
OpenAI o1 / o3 (OpenAI)	Text	Són models de llenguatge dissenyats per abordar problemes complexos mitjançant un procés de 'reflexió' abans de generar respostes. Aquesta capacitat els permet raonar de manera més efectiva en tasques com programació, matemàtiques i ciències. OpenAI o3 n'és la versió més avançada.
Claude 3.5 / 3.7 - Sonnet (Anthropic)	Multimodal	Claude 3.5 és un model de llenguatge avançat i sòlid en la generació i comprensió de text. Claude 3.7 en millora el raonament amb un enfocament híbrid que combina respostes ràpides amb una anàlisi detallada de problemes complexos.
Command R (Cohere)	Text	Model dissenyat específicament per a la generació augmentada per recuperació (RAG), combinant informació de fonts externes amb generació de text per millorar la precisió i la rellevància.
Gemini 1 /1.5 / 2 Flash (Google DeepMind)	Multimodal	Gemini 1 va introduir un model multimodal centrat en la comprensió i generació del llenguatge natural. Gemini 1.5 va millorar l'eficiència i la capacitat de processament, permetent finestres de context més àmplies. Gemini 2.0 Flash optimitza encara més la velocitat i els recursos computacionals, fent-lo ideal per a tasques complexes en temps real.
PaLM 1 / 2 (Google)	Text	Models de llenguatge amb habilitats avançades de comprensió multilingüe, raonament lògic i millores en àmbits com la medicina i les ciències naturals. PaLM 2 n'és la darrera versió.
PaLM-E (Google)	Multimodal	Extensió de la sèrie PaLM amb capacitats multimodals, integrant text amb dades visuals com imatges i vídeos. Dissenyat especialment per aplicacions en robòtica i anàlisi de dades visuals.

4.1.2 / Models generalistes de codi obert

Els models generalistes de codi obert poden ser creats tant per comunitats de desenvolupadors com per empreses. **La seva naturalesa oberta ofereix una gran flexibilitat,** cosa que permet als usuaris un major control sobre el model des del procés d'entrenament fins a la seva adaptació, i facilita la seva especialització per a aplicacions sanitàries i biomèdiques.

Tanmateix, aquests models també presenten reptes importants, com ara la manca de suport tècnic dedicat, els alts requeriments de recursos computacionals, la necessitat de coneixements tècnics avançats per a la seva implementació i les possibles variacions en la qualitat del seu rendiment.

SLM: *Small Language Model* (veure 4.1.5/ Petits Models de Llenguatge (SLM) i tècniques de compressió).

LLM generalistes de codi obert (febrer, 2025)

Model (desenvolupador)	Modalitat de dades	Llicència de codi obert	Descripció
Mistral 7b (Mistral AI)	Text	Apache 2.0	Model que destaca pel seu rendiment comparat amb la seva mida relativament petita de 7.000 milions de paràmetres, oferint una gran eficiència en múltiples tasques. És un SLM.
Mistral NeMo (Mistral AI)	Text	Apache 2.0	Model més recent de la suite de models de Mistral, adaptat a necessitats específiques en diversos dominis de processament de text.
LLaMA 1 / 2 / 3.2 (Meta)	Text	Llicència personalitzada de Meta	LlaMa 1 va ser un dels primers models de Meta enfocats a investigació, que es va popularitzar per la seva accessibilitat. LLaMA 2 va ampliar-ne les capacitats, incloent versions SLM (7B). LLaMA 3.2 en millora el rendiment i està optimitzat per a aplicacions més eficients.
Dolly 2.0 (Databricks)	Text	Apache 2.0	Model dissenyat per ser fàcilment utilitzable i accessible per a la comunitat de desenvolupadors, cosa que permet una implementació senzilla en diverses aplicacions.
GPT-Neo (Eleuther AI)	Text	Apache 2.0	Un dels primers esforços de la comunitat de codi obert per crear una alternativa al GPT-3, centrat en la transparència i la replicabilitat en el processament del llenguatge natural.
GPT-J (Eleuther AI)	Text	Apache 2.0	Successor de GPT-Neo, amb 6.000 milions de paràmetres, ofereix una millor capacitat de generació de text amb un model preentrenat més optimitzat. És un SLM.
T5 (Google)	Text	Apache 2.0	Model de Google dissenyat per a múltiples tasques de llenguatge natural, com ara traducció i respostes a preguntes.
DeepSeek-V2 / V3 / R1 (DeepSeek AI)	Multi-modal	DeepSeek (amb restriccions d'ús responsable)	Destaca per la seva escalabilitat i entrenament en múltiples idiomes. La versió V3 millora la velocitat i el rendiment, mentre que R1 se centra en el raonament lògic i matemàtic. Restringeix la generació de contingut en temàtiques que els desenvolupadors consideren sensibles.

4.1.3 / Models adaptats a l'àmbit sanitari i biomèdic

Existeixen diversos LLM que ja han estat adaptats específicament per al seu ús en el sector de la salut. Això és possible gràcies a un procés d'afinació (*fine-tuning*), que consisteix a **ajustar un model general amb dades específiques d'un àmbit, en aquest cas clíniques o biomèdiques**, per millorar-ne el rendiment en tasques concretes dins d'aquest àmbit (veure 4.2/ Adaptació del model).

Gràcies a aquesta adaptació, els models poden comprendre millor conceptes tècnics, terminologia especialitzada, argot mèdic i acrònims que no es troben habitualment en els conjunts de dades d'entrenament generalistes. Això els fa especialment útils per a aplicacions, com ara l'anàlisi de textos mèdics, la interpretació de resultats clínics, la generació de resums de publicacions científiques i el suport en la presa de decisions clíniques.

És important tenir en compte que aquests models poden estar subjectes a diferents tipus de llicències, com ara de codi obert o comercials.

LLM adaptats a l'àmbit sanitari i biomèdic (febrer, 2025)

Model (desenvolupador)	Modalitat de dades	Llicència	Descripció
Med-PaLM[4]	Text	Comercial (Google)	Med-PaLM és una sèrie de models de Google adaptats al domini mèdic, de la qual Med-PaLM 2 n'és la versió més recent i amb millor rendiment.
Med-PaLM M[5]	Multi-modal	Comercial (Google)	Versió avançada del Med-PaLM, capaç d'integrar dades textuals, imatges i dades genòmiques, entre altres modalitats, per a aplicacions biomèdiques.
MedLLaMA	Text	De codi obert	Sèrie de models LLaMA adaptats a aplicacions en salut, com ara la resposta a preguntes mèdiques, resums d'historials mèdics i suport en diagnòstic clínic.
BiomedGPT[6]	Multi-modal	De codi obert	Famílies de models adaptats per generar text científic i mèdic, oferint eines útils per a investigadors i professionals de la salut.
LLaVA-Med[7]	Multi-modal	De codi obert	Sèrie de models enfocats a l'anàlisi d'imatges mèdiques i informes clínics, permetent la integració d'informació visual i textual per a tasques biomèdiques.

4.1.4 / Models locals o remots

Probablement, la decisió amb més impacte en el funcionament operatiu d'una solució basada en LLM serà la localització física del model.

- En un **sistema local**, el model s'instal·la i s'executa dins de la infraestructura de l'organització que l'utilitza.
- En canvi, en un **sistema remot** el model s'executa en servidors externs, habitualment allotjats al núvol i gestionats per un proveïdor tercer. Aquesta arquitectura implica que les dades d'entrada i sortida es transfereixen a través de la xarxa entre els servidors de l'usuari i els del proveïdor del model de llenguatge.

Existeix una àmplia gamma d'opcions entre models locals i remots, i la tria entre aquestes alternatives no només afecta aspectes tècnics, com la latència o la gestió de recursos, sinó també qüestions fonamentals com ara la privacitat, la seguretat de les dades i el compliment normatiu.

Criteris a considerar en la tria entre models locals i remots

CRITERI 01 /
Rendiment i accessibilitat

- Els **models locals** són ideals per a aplicacions que requereixen baixa latència o un accés constant sense dependències de connexions a internet.
- Els **models en remot** ofereixen més flexibilitat i escalabilitat, especialment per a aplicacions que necessiten actualitzacions constants o accés des de múltiples ubicacions.

CRITERI 02 /
Costos i infraestructures

- Els **models locals** requereixen inversions significatives en infraestructures pròpies, com ara servidors, xarxes i suport tècnic, així com en el seu manteniment. Aquestes necessitats poden incrementar els costos inicials durant les primeres fases del desenvolupament, tot i que aquesta inversió pot resultar rendible en entorns amb un ús intensiu.
- L'ús de **models remot** elimina la necessitat d'invertir en infraestructura, però pot comportar costos recurrents més elevats depenent del volum de dades processades.

CRITERI 03 /
Seguretat i privacitat de les dades

- En els **models locals**, les dades no surten de l'organització usuària i, per tant, es redueix el risc de filtracions o vulneracions de privacitat. Aquesta opció pot ser especialment important en el sector de la salut, on la informació és altament sensible.
- Els **models en remot**, en canvi, impliquen un flux de dades per la xarxa que requereixen garanties robustes de seguretat per part del proveïdor, que poden incloure l'anonimització o pseudoanonimització i el xifratge avançat de les dades. Aquests models poden estar subjectes a les lleis del país on es troben els seus servidors, cosa que pot complicar o impedir el compliment del RGPD i altres normatives del país de l'organització usuària (veure 3.6/ Privacitat i protecció de dades)

4.1.5 / Petits Models de Llenguatge (SLM) i tècniques de compressió

Els **petits models de llenguatge** (*Small Language Model*, **SLM**) són versions més petites i lleugeres creades a partir de LLM, i dissenyades per oferir un **equilibri entre rendiment i eficiència.**

Malgrat la reducció en el nombre de paràmetres i en la capacitat computacional requerida (en comparació amb els LLM), els SLM poden assolir un rendiment similar en tasques específiques quan se sotmeten a un procés d'adaptació mitjançant *Retrieval Augmented Generation* (RAG) o afinament (*fine-tuning;* veure 4.2/ Adaptació del model)[8].

Els SLM són especialment útils en entorns com hospitals, on la seguretat de les dades és prioritària i les limitacions d'infraestructura poden fer inviable l'ús de models més grans.

El procés de creació d'un SLM implica l'aplicació d'una o diverses **tècniques d'optimització del rendiment.**

Tècniques d'optimització del rendiment

TÈCNICA 01 /
Distil·lació (*distillation*)

La distil·lació consisteix a entrenar un model més petit (*student model*) perquè imiti el comportament d'un model més gran (*teacher model*), de manera que es mantingui un rendiment similar en tasques generals.

TÈCNICA 02 /
Quantització (*quantization*)

La quantització consisteix a reduir la precisió numèrica dels pesos i càlculs del model (p.e., passant de 32 bits a 8 bits), disminuint-ne la mida i el consum de recursos tot mantenint el rendiment general del model.

TÈCNICA 03 /
Podament (*pruning*)

Durant el podament s'eliminen components no essencials del model, com capes o connexions neuronals amb baixa contribució a la tasca d'interès, per tal d'optimitzar el seu rendiment sense sacrificar de manera significativa la seva eficàcia.

Cal mencionar que, tot i que aquestes tècniques de comprensió són necessàries per crear un SLM, també s'utilitzen per optimitzar el rendiment de models grans en entorns amb limitacions computacionals, o bé per incrementar-ne la velocitat d'ús sense necessitat de reduir la mida del model.

4.2 / Adaptació del model

Un cop seleccionats els models, és important valorar si cal sotmetre'ls a un procés d'adaptació per **especialitzar-los en un context o funcionalitat concreta** dins de l'àmbit en que es vol desplegar l'eina, en aquest el sanitari o biomèdic, amb l'objectiu de millorar-ne l'eficàcia.

Els models adaptats solen oferir un millor rendiment en tasques específiques per a les quals han estat ajustats, però també solen requerir costos addicionals en dades, temps i recursos.

Els **mètodes d'adaptació** més comuns, ordenats segons el grau d'adaptació, el rendiment esperat i els esforços i recursos requerits, són:

01	Enginyeria de *prompts*
02	Generació Augmentada per Recuperació
03	Afinament

4.2.1 / Enginyeria de *prompts (prompt engineering)*

El *prompting* és una tècnica essencial en l'àmbit de la IA generativa, i constitueix la forma més habitual d'interactuar amb un model de llenguatge. Consisteix a **formular instruccions o preguntes específiques dirigides al model**, de complexitat i extensió molt variable, i que han estat prèviament dissenyats i orientats a l'objectiu que es vol assolir.

Tot i que no altera els paràmetres del model base, la qualitat dels *prompts* té un impacte directe en la rellevància i utilitat de les respostes generades. Seguint els principis establerts per Phoenix i Taylor (2024)[9], podem establir **cinc directrius bàsiques per la generació de *prompts***:

01	Donar direcció
02	Descriure l'estil i la persona
03	Especificar el format
04	Proporcionar exemples
05	Dividir el treball

DIRECTRIU 01 /

Donar direcció

Defineix l'objectiu del *prompt* de manera clara, específica, sense ambigüitats i proporcionant el context necessari per a la tasca. Aquest nivell de detall ajuda el model a alinear-se millor amb les expectatives de l'usuari i evita respostes imprecises o disperses.

Per exemple, en lloc d'escriure simplement "Escriu una llista de recomanacions mèdiques", es pot especificar:

```
> Escriu una llista de 5 recomanacions per a un pacient amb hipertensió arterial, centrant-te en hàbits de vida saludables i evitant recomanacions farmacològiques

Sortida:

Recomanacions per a un pacient amb hipertensió arterial:

1. Reduir el consum de sal: Limita la ingesta de sal a menys de 5 grams al dia. Evita aliments processats rics en sodi i opta per condiments naturals com herbes i espècies.

2. Augmentar el consum de fruites i verdures...
```

DIRECTRIU 02 /

Descriure l'estil i la persona

Indica el **to**, el **registre lingüístic** i, si escau, el **rol** que el model ha d'adoptar.

Es pot demanar, per exemple, que adopti un rol de clínic especialista en oftalmologia, i que utilitzi un estil comprensible pel públic general.

```
> Respon a les següents preguntes com si fossis un clínic especialista en oftalmologia, però utilitzant un estil comprensible per al públic general.

Què és el glaucoma i com afecta la visió?

Sortida:

El glaucoma és una malaltia que danya el nervi òptic, que és com un cable que envia les imatges des de l'ull al cervell...
```

DIRECTRIU 03 /

Especificar el format

Defineix amb precisió **l'estructura de la resposta** que esperes. Això pot incloure formats específics com una llista numerada amb un separador determinat, taules, respostes en format JSON o paràgrafs estructurats. Aquesta directriu és especialment útil per garantir la consistència i facilitar la interpretació de les respostes.

```
> Escriu una llista de 5 hàbits saludables per prevenir la diabetis tipus 2, centrant-te en hàbits de vida saludables i evitant recomanacions farmacològiques. Escriu les respostes en una taula en format JSON, on la primera columna serà la recomanació i la segona serà una descripció breu d'aquesta recomanació.

Sortida:

[
  {
  "Recomanació": "Mantenir un pes saludable",
  "Descripció": "Controlar el pes corporal per reduir la resistència a la insulina i evitar l'excés de greix corporal, que pot augmentar el risc de diabetis.“
 }, ...
```

DIRECTRIU 04 /

Proporcionar exemples

Incloure **exemples de respostes òptimes o casos ben resolts** dins del *prompt* és una tècnica molt efectiva per mostrar al model el tipus de sortida desitjada, especialment en tasques complexes o poc habituals.

Aquesta estratègia constitueix la base del *prompting few-shot* (explicada més endavant), i també s'utilitza en tècniques avançades com la Generació Augmentada per Recuperació d'Informació (RAG; (veure 4.2.2/ Generació Augmentada per Recuperació d'Informació).

DIRECTRIU 05 /

Dividir el treball

Per a tasques complexes, **divideix el procés en etapes senzilles i interconnectades.** Per exemple, en el camp mèdic, un primer *prompt* podria identificar símptomes, un segon suggerir possibles diagnòstics i un tercer proposar proves diagnòstiques. Aquesta estratègia està darrere de tècniques com el *chain-of-thought* o el *tree-of-thought* explicades a continuació, que permeten al model simular un raonament estructurat i seqüencial.

Tècniques d'enginyeria de *prompts*

Les tècniques d'enginyeria de *prompts* són estratègies dissenyades per **optimitzar la interacció amb els LLM i obtenir respostes més precises, rellevants i ajustades a les necessitats d'una tasca específica.**

Aquestes tècniques es poden considerar una forma d'adaptació lleugera del model, ja que permet ajustar el seu el comportament sense necessitat de modificar-ne els paràmetres. A més, l'ús adequat de *prompts* pot contribuir a millorar l'explicabilitat del model, ja que facilita la comprensió del procés pel qual es generen les respostes.

Les tècniques més destacades de l'enginyeria de *prompts* són:

01	*Zero-shot*
02	*Few-shot*
03	Cadena de pensament *(chain of thought)*
04	*Ensemble refinament*
05	Arbre de pensament (*tree of thought*)

TÈCNICA 01 /

Zero-shot

Consisteix a demanar al model que executi una tasca sense proporcionar-li exemples específics previs, **basant-se únicament en el seu coneixement general adquirit durant l'entrenament.**

És un procés senzill i directe, útil per a **tasques simples**.

```
> Quins són els principals efectes secundaris de la medicació antihipertensiva?

Sortida:

Els principals efectes secundaris dels medicaments antihipertensius poden incloure marejos, fatiga, mal de cap, nàusees, disfunció erèctil i alteracions electrolítiques.
```

TÈCNICA 02 /

Few-shot

Consisteix a proporcionar al model diversos exemples perquè entengui el tipus de resposta desitjada. Es basa en els principis de "**donar direcció**", "**especificar el format**" i "**proporcionar exemples**" abans mencionats. És especialment útil en situacions on es requereix un cert nivell de **personalització o un format específic** en les respostes.

```
> Enumera els símptomes d'una patologia ordenats per freqüència (percentatge) i separats per una coma. Aquí alguns exemples:

Exemple 1

Patologia: Grip.

Símptomes: Febre (80%), malestar general (70%), tos seca (60%), dolor muscular (50%).

Exemple 2

Patologia: Gastroenteritis viral.

Símptomes: Diarrea (90%), nàusees (80%), dolor abdominal (70%), febre baixa (30%).

Ara, per la patologia "apendicitis aguda", indica els símptomes ordenats per freqüència i inclou el percentatge corresponent.

Sortida:

Patologia: Apendicitis aguda.

Símptomes: Dolor abdominal al quadrant inferior dret (90%), febre (80%), nàusees (75%), vòmits (50%), pèrdua de gana (40%).
```

TÈCNICA 03 /

Cadena de pensament *(chain of thought)*

Consisteix en **guiar el LLM perquè generi una resposta pas a pas**, desglossant el raonament en etapes més senzilles i estructurades, en lloc de proporcionar una resposta directa. Això millora la qualitat i precisió de les respostes del model i contribueix a la seva explicabilitat, cosa que augmenta la confiança en els resultats generats.

Aquesta tècnica es basa en el principi de "**dividir el treball**", i és d'utilitat per **tasques que requereixen raonaments seqüencials**, com problemes matemàtics o de lògica.

```
> Un pacient de 25 anys arriba amb febre, dolor abdominal intens i vòmits. Raona pas a pas i en veu alta per determinar possibles diagnòstics, tenint en compte els símptomes, la seva distribució i possibles condicions associades. Enumera possibles diagnòstics i el següent pas d'acció.

Sortida:

Pas 1: Els símptomes presentats són febre, dolor abdominal intens i vòmits. La febre suggereix una causa infecciosa o inflamatòria. El dolor abdominal i els vòmits poden estar relacionats amb diverses condicions gastrointestinals.

Pas 2: Considerem el dolor abdominal. Si està localitzat al quadrant inferior dret, és probable que sigui una apendicitis aguda. Si el dolor és difús i va acompanyat de diarrea, podria tractar-se d'una gastroenteritis.

Pas 3: Els vòmits són habituals en ambdues condicions, però també poden aparèixer en casos com l'obstrucció intestinal, on el dolor sovint és més sever i progressiu.

Pas 4: Tenint en compte l'edat del pacient i els símptomes, proposem els següents diagnòstics probables, ordenats segons la seva probabilitat:

1. Apendicitis aguda: Molt probable si el dolor està localitzat al quadrant inferior dret i va acompanyat de febre i pèrdua de gana.

2. Gastroenteritis viral: Possible si el dolor és difús i hi ha diarrea.

3. Obstrucció intestinal: Possible si el dolor és colicatiu i s'associa amb distensió abdominal i absència de gasos o femtes.

Pas 5: Es recomana realitzar una exploració física completa (incloent-hi el signe de Blumberg), una analítica de sang (per buscar leucocitosi) i una ecografia abdominal per confirmar el diagnòstic.
```

TÈCNICA 04 /

Ensemble refinement

Tècnica que millora la qualitat i fiabilitat de les respostes d'un model de llenguatge mitjançant la **generació de múltiples resultats independents** a partir d'una mateixa entrada *(prompt),* aprofitant la variabilitat intrínseca del model, o a partir de petites modificacions en les instruccions. Aquestes respostes es comparen i s'agreguen mitjançant diferents estratègies, com ara resums, ponderacions numèriques, selecció per vot majoritari (també conegut com a **autoconsistència** o *self-consistency)* o refinament seqüencial.

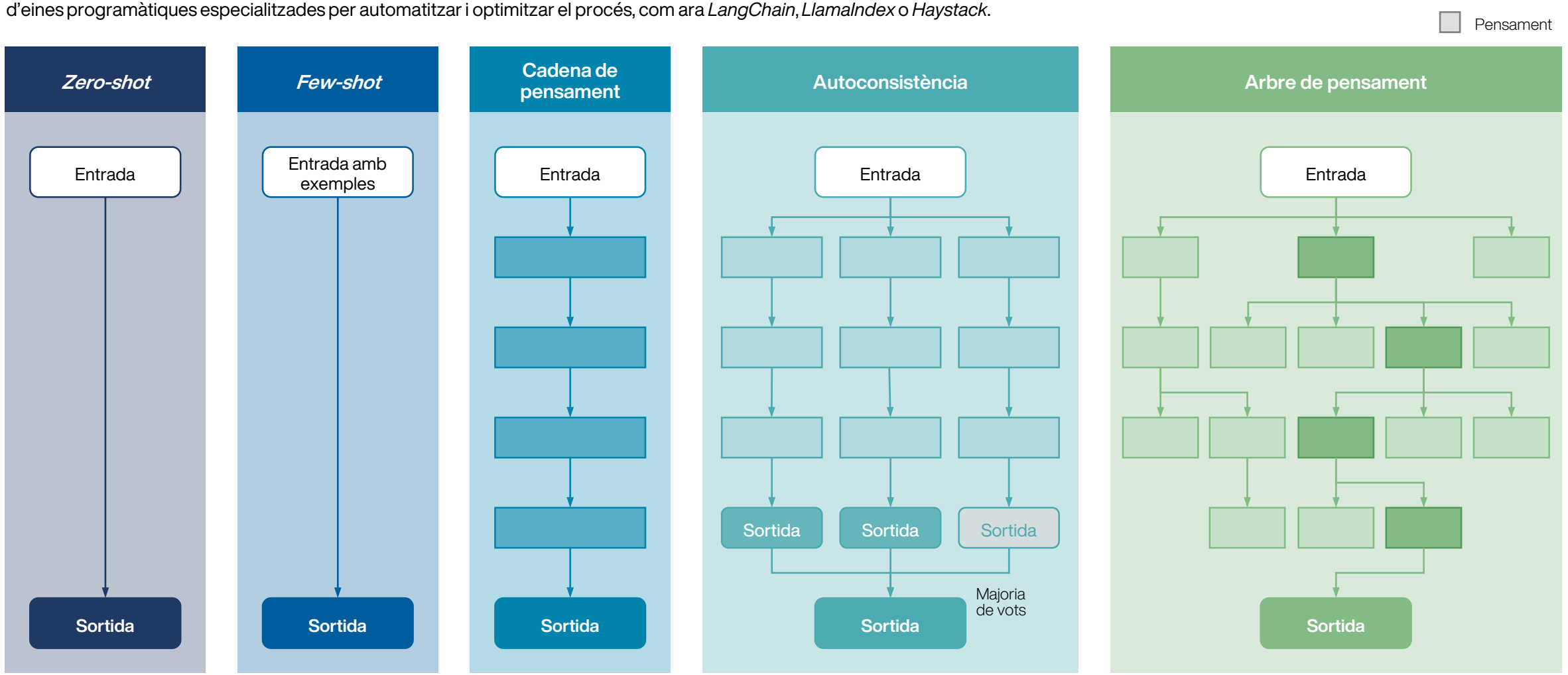
TÈCNICA 05 /

Arbre de pensament *(tree of thought)*

Estratègia avançada que amplia la tècnica de cadena de pensament per millorar el procés de raonament dels models de llenguatge en tasques complexes que impliquen múltiples passos. En lloc de seguir un únic camí lineal de raonament, aquesta tècnica **estructura el procés en diverses ramificacions que formen un arbre lògic**, cosa que permet al model explorar múltiples opcions o hipòtesis abans d'arribar a una resposta final basada en la solució més consensuada o coherent.

Visió general de les tècniques de *prompting*. Figura adaptada de Yao et al. (2023)[10]

La figura mostra les diferents estratègies de *prompting* presentades anteriorment. Cada rectangle representa un pas intermedi en el procés de raonament del model, ajudant-lo a desglossar i resoldre problemes de manera estructurada. Cal tenir en compte que la implementació de tècniques avançades com auto-consistència, *ensemble refinement* o arbre de pensament, requereix l'ús d'eines programàtiques especialitzades per automatitzar i optimitzar el procés, com ara *LangChain*, *LlamaIndex* o *Haystack*.



4.2.2 / Generació Augmentada per Recuperació d'Informació (RAG)

La Generació Augmentada per Recuperació *(Retrieval-Augmented Generation,* RAG) és una tècnica que **integra la capacitat generativa dels models de llenguatge amb la recuperació d'informació en temps real de fonts externes al model.** Aquesta combinació permet que el model accedeixi a **dades actualitzades o específiques en el moment de la consulta**, fet que millora la precisió i rellevància de les respostes o textos generats.

Com a inconvenient, aquesta estratègia **requereix una infraestructura addicional** per gestionar la recuperació d'informació en temps real. A diferència d'un LLM sense adaptar, el RAG ha de consultar constantment bases de dades o sistemes d'informació externs, el que pot augmentar la latència i els costos operatius, **complicant l'escalabilitat i la gestió tècnica del sistema.**

En la tècnica RAG, el procés per generar respostes a partir d'una consulta es desenvolupa en dues fases:

01	Fase de recuperació
02	Fase de generació

FASE 01 /

Fase de Recuperació

Davant d'una consulta, el sistema **cerca informació rellevant en un conjunt definit de dades externes al model** (p.e., informació clínica, articles científics, guies clíniques o recursos educatius per a pacients), per trobar informació útil relacionada amb la consulta. Aquestes fonts poden ser conjunts de dades fixes o dinàmiques, i poden estar localitzats localment o estar disponibles a la xarxa.

Durant aquesta fase:

1. **La consulta es transforma en un vector** d'alta dimensionalitat mitjançant models de representació vectorial *(embeddings)* com BERT o *sentence transformers,* capturant el seu significat semàntic.
2. **Es compara el vector** de la consulta amb els vectors dels documents de bases de dades vectoritzades i indexades. Per a això, s'utilitzen algoritmes de cerca de vectors, com ara el KNN *(k-Nearest Neighbors),* Annoy o FAISS.
3. **Es recuperen els documents més rellevants** que estan més propers al vector de consulta, generalment, en funció d'alguna mètrica de similitud com ara el cosinus, que mesura com són de semblants les direccions dels vectors entre sí.

FASE 02 /

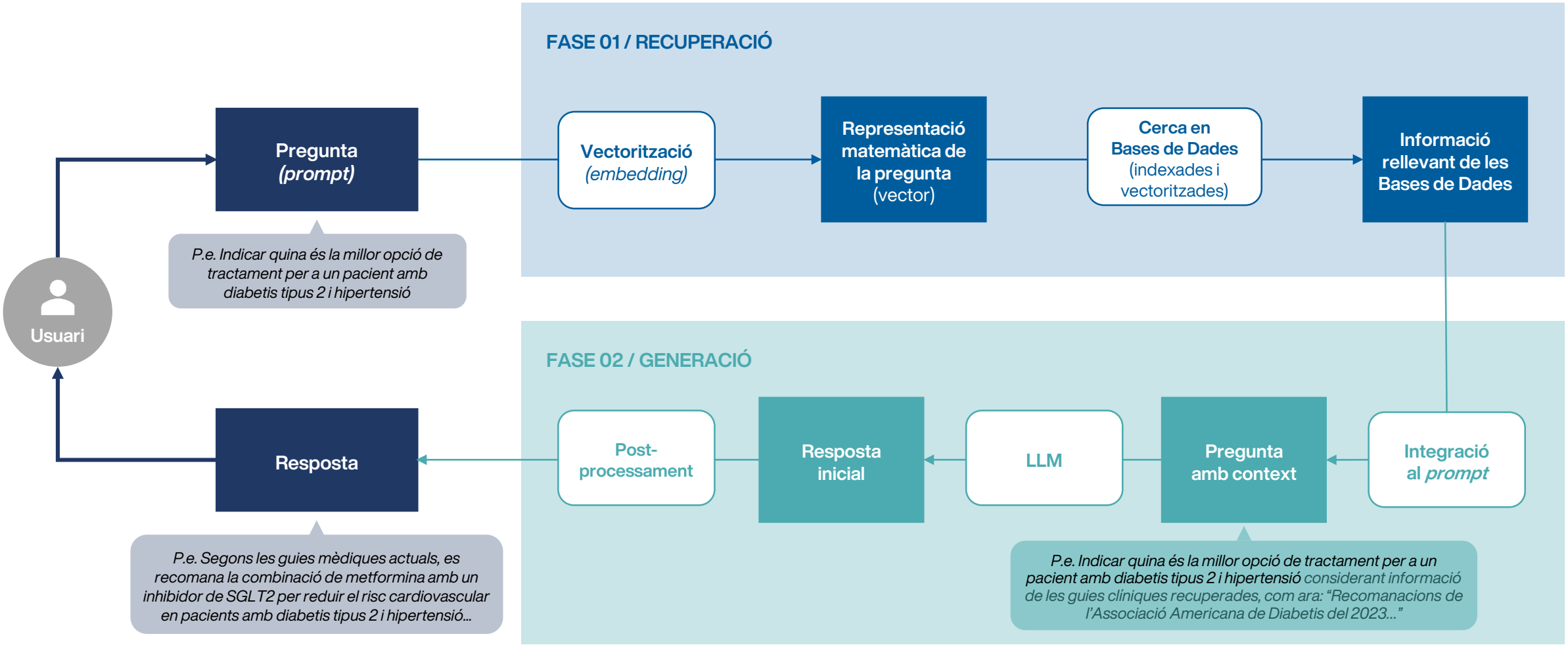
Fase de Generació

Un cop s'ha recuperat la informació rellevant per la consulta, s'inicia la **generació de la resposta.** En aquesta fase, el sistema **integra la informació recuperada com a context de la consulta original** per generar una resposta coherent, precisa i adequada a aquest context.

Durant la fase de generació:

1. **La informació recuperada s'integra com a context en el *prompt*** que es passarà al LLM per realitzar la consulta. Això pot incloure resums dels documents, fragments rellevants o, fins i tot, tota la informació dels documents.
2. **S'elabora una resposta final** més precisa, coherent i alineada amb la consulta inicial de l'usuari utilitzant el context addicional.
3. **S'apliquen mecanismes de control** per garantir la fiabilitat abans de generar una resposta final. Aquests mecanismes inclouen: el filtrat de contingut, per assegurar que només es basa en la informació recuperada; la verificació de referències, per millorar la traçabilitat, l'adaptació a l'usuari final, l'ajust del format i el nivell de detall; i l'avaluació de confiança, per alertar en cas de disposar d'informació insuficient o ambigua.

Procés de Generació Augmentada per Recuperació d'Informació (RAG)



4.2.3 / Afinament *(fine-tuning)*

L'afinament *(fine-tuning)* consisteix a **ajustar un model preentrenat mitjançant un conjunt específic de dades**, en aquest cas procedents de l'àmbit sanitari. Aquest procés permet al model adaptar-se millor a tasques concretes d'aquest àmbit, millorant-ne la capacitat d'entendre i generar contingut especialitzat.

En aquest punt, és important diferenciar entre:

- **Preentrenament:** consisteix a entrenar el model amb grans volums de dades, generalment no etiquetades, per dotar-lo d'un coneixement general del llenguatge.
- **Afinament:** adapta el model preentrenat (fundacional) a tasques concretes amb conjunts de dades més petits i etiquetats, utilitzant menys recursos computacionals i millorant el rendiment en aplicacions específiques.

Estratègies per l'afinament

Existeixen diverses **estratègies d'afinament** que depenen del tipus de **dades disponibles** i dels **objectius de la solució.** En termes generals, aquestes estratègies es poden classificar en:

01	Afinament no supervisat
02	Afinament supervisat

ESTRATÈGIA 01 /

Afinament no supervisat

Consisteix a **exposar el model a textos no etiquetats**, com ara articles científics o guies clíniques. Aquesta estratègia permet millorar la capacitat del model per comprendre la terminologia clínica, fent-lo més eficient en la **generació de textos especialitzats** com ara informes mèdics o resums de literatura científica. Aquesta estratègia, però, presenta limitacions a l'hora de realitzar tasques de classificació (p.e., diagnòstic o estratificació de risc), ja que el model no disposa de dades etiquetades que li permetin identificar patrons específics de malalties o condicions dels pacients.

ESTRATÈGIA 02 /

Afinament supervisat

Requereix de **dades etiquetades**, com ara textos mèdics associats a diagnòstics o tractaments específics.

És especialment útil per entrenar models en **tasques de classificació mèdica** com, per exemple, el diagnòstic automàtic de malalties o la generació de propostes personalitzades de plans de tractament. Tot i el seu potencial, la creació de dades etiquetades pot ser costosa i lenta, ja que sovint requereix la intervenció d'experts mèdics per garantir la qualitat i la coherència de les etiquetes.

Tècniques d'afinament

Les **tècniques d'afinament** són metodologies concretes que implementen les estratègies definides anteriorment, amb l'objectiu d'adaptar models preentrenats a tasques específiques. Les tècniques més freqüents són:

01	Reentrenament complet *(full fine-tuning)*
02	*Parameter-Efficient Fine-Tuning* (PEFT)
03	Aprenentatge per reforç amb *feedback* humà

TÈCNICA 01 /

Reentrenament complet *(full fine-tuning)*

Tècnica tradicional d'afinament, que implica **reentrenar tots els paràmetres d'un model** utilitzant **dades específiques** de la tasca o el domini d'interès.

Proporciona la màxima adaptació del model a tasques concretes, i pot millorar significativament el rendiment en dominis altament especialitzats. No obstant això, és un procediment **molt costós i exigent en termes de temps, dades etiquetades i recursos computacionals,** ja que sovint requereix infraestructures amb GPU o TPU.

TÈCNICA 02 /

Parameter-Efficient Fine-Tuning (PEFT)

Les PEFT són les tècniques més utilitzades actualment **en contextos on l'eficiència computacional és prioritària.** Es caracteritzen per modificar només una part del model, mantenint la major part dels seus paràmetres intactes. Això facilita l'adaptació a noves tasques amb un cost computacional significativament més baix[11].

Les tècniques de PEFT més populars són:

- **LoRA *(Low-Rank Adaptation)***
  Ajusta només una part de les matrius de pes del model preentrenat, en lloc de reentrenar totes les seves capes. Això redueix dràsticament el nombre de paràmetres que cal modificar, fent que l'afinament sigui més ràpid i eficient sense comprometre significativament la precisió. És especialment útil en LLM de gran escala.
- **Adaptadors *(adapters)***
  Són petits **mòduls addicionals** que s'incorporen a diverses capes del model, i que **s'entrenen de manera independent** mentre la resta dels paràmetres es mantenen congelats. Aquesta tècnica conserva el coneixement general del model preentrenat i alhora permet l'aprenentatge en tasques específiques amb noves dades, evitant la necessitat de reentrenar el model per complet.
- ***Prefix tuning***
  Afegeix una **seqüència curta de paràmetres ajustables** (un "prefix") **a l'entrada del model**, que guia la seva resposta per a tasques concretes. Com que només s'ajusta aquest prefix i no les seves capes, la resta del model queda inalterat. Aquesta aproximació és especialment eficient, ja que redueix significativament els costos computacionals i el temps necessari per adaptar el model a noves tasques.

TÈCNICA 03 /

Aprenentatge per reforç amb feedback humà *(Reinforcement Learning with Human Feedback,* RLHF*)*

El RLHF és una tècnica avançada d'afinament que **ajusta les respostes del model per alinear-les amb les preferències i expectatives** dels usuaris.

És especialment útil per **optimitzar l'experiència de l'usuari** en aplicacions pràctiques, ja que permet generar respostes més naturals i adaptades a tasques específiques. Tanmateix, també implica **processos costosos i llargs**, en requerir la intervenció d'experts humans per avaluar les respostes, i comporta el risc que el model prioritzi respostes agradables o conformes a les expectatives dels usuaris per sobre de la seva veracitat (veure 5.3/ Detecció i mitigació d'al·lucinacions).

Aquest procés es desenvolupa en tres fases principals:

1. **Generació de respostes:** basant-se en el seu coneixement preentrenat, el model genera diverses respostes per a un conjunt d'entrades específiques.
2. **Avaluació humana:** un grup d'experts o usuaris valora o ordena les respostes generades en funció de la seva qualitat, rellevància i utilitat.
3. **Reforç:** les valoracions es converteixen en una funció de recompensa que s'utilitza per ajustar el model, millorant la qualitat de les seves sortides de forma iterativa.

4.2.4 / Com triar el mètode d'adaptació?

L'adaptació és un procés necessari quan es requereix un cert grau d'especialització en la nostra eina per optimitzar la seva eficàcia.

Els models més adaptats solen tenir un millor rendiment per la tasca específica per la qual han estat dissenyats, però també solen requerir costos addicionals en dades, temps i recursos.

Cal destacar que existeix un ventall cada cop més gran de models que ja han estat afinats per millorar el seu rendiment en l'àmbit biomèdic o assistencial. En funció dels objectius definits, l'eficàcia de l'eina es podria beneficiar de l'alt grau d'especialització que ofereixen aquests models (veure 4.1.3/ Models adaptats a l'àmbit sanitari i biomèdic).

A continuació es presenten els aspectes clau dels tres mètodes d'adaptació descrits anteriorment, així com una taula comparativa amb les seves característiques principals.

Aspectes clau dels mètodes d'adaptació

Enginyeria de *prompts*	És generalment preferible quan volen augmentar l'eficàcia del model sense modificar-lo, simplement orientant el seu comportament en les respostes d'una manera flexible i amb una inversió mínima de recursos.
RAG	Tot i que requereix més recursos, és una opció superior per aplicacions que necessiten accedir a un conjunt definit de dades externes al model, que poden ser dinàmiques i estar sotmeses a revisions i actualitzacions periòdiques. Això és especialment útil en entorns clínics, on els avenços científics poden obligar a una revisió freqüent de les fonts d'informació (p.e., guies mèdiques, estratègies de tractament, nous fàrmacs, etc.).
Afinament (*fine-tuning*)	És l'opció més adequada si es requereix un alt nivell d'adaptació del model per una tasca concreta, integrant coneixements d'un domini específic de l'àmbit sanitari. Aquest mètode permet un major control del comportament del model, tenint com a contrapartida la necessitat de grans volums de dades etiquetades del domini en qüestió i d'alts costos computacionals.

Taula comparativa dels mètodes d'adaptació dels models d'IA: Enginyeria de *prompts*, RAG i Afinament

Criteri	Enginyeria de *prompts*	RAG	Afinament *(fine-tuning)*
Rendiment	Depèn de la qualitat del *prompt*, però generalment menor	Alt per a tasques que requereixen accés a informació dinàmica	Molt alt, especialment per tasques específiques
Recursos necessaris	Molt baixos: no requereix entrenament ni infraestructures addicionals	Moderats: infraestructura addicional per gestionar la recuperació d'informació, i creació d'un índex eficient	Alts: requereix capacitat computacional significativa per entrenar el model
Requeriment de dades	Només es basa en les dades ja incloses el model preentrenat	Necessitat d'un conjunt generalment ampli de documents rellevants no estructurats, com ara publicacions científiques o guies clíniques	Gran quantitat de textos etiquetats o estructurats i específics del domini
Transparència i interpretabilitat	Moderada: el comportament del model està directament relacionat amb el *prompt*	Moderada: fàcil d'identificar quines dades han influït en la resposta (mitjançant l'índex)	Baixa: el model ajusta els pesos internament i pot ser difícil rastrejar com es prenen les decisions
Al·lucinacions	Altes: el model pot inventar respostes basades en patrons de l'entrenament original	Baixes, ja que el model se sustenta en informació recuperada, però depèn de la qualitat de les dades	Reduïdes si s'entrena amb dades de qualitat; possibles si el conjunt de dades és incomplet o sorollós
Risc de biaixos	Moderat o alt: depenen completament dels biaixos del model preentrenat	El risc de biaixos depèn de la diversitat i representativitat de les fonts de dades utilitzades	Depèn de la diversitat i representativitat de les dades utilitzades per l'afinament.
Control sobre el model	Moderat: els *prompts* orienten les respostes del model	Moderat: l'índex i les dades recuperades influeixen en la resposta	Alt: ajust complet dels pesos per adaptar-lo al domini
Flexibilitat	Molt alta: fàcil de modificar segons la tasca	Alta: pot actualitzar-se fàcilment amb noves dades	Baixa: reentrenar per a noves tasques és un procés complex
Compatibilitat amb dades sensibles	Si els *prompts* inclouen dades sensibles, s'han d'establir mecanismes per garantir la privacitat	Les dades es poden mantenir en sistemes locals per evitar riscos, i establir altres mecanismes per garantir la privacitat	Les dades s'integren directament en el model, però cal anonimitzar-les i protegir-les durant l'entrenament
Cost a llarg termini	Molt baix: només es requereix actualitzar *prompts*, si cal	Moderat: cal mantenir l'índex i el sistema de recuperació	Alt: l'entrenament i el manteniment poden ser cars
Casos d'ús adequats	Generació general de contingut, consells o respostes ràpides	Consultes a un conjunt definit de dades externes al model, especialment quan són dades dinàmiques	Especialització en diagnòstics mèdics, anàlisi de dades complexes o generació d'informes especialitzats

4.2.5 / Dades per a l'adaptació

Tant els mètodes basats en **RAG** com **l'afinament** requereixen dades de qualitat per dur a terme l'adaptació del model.

Aquestes dades poden variar segons diversos factors, com ara el seu **origen** (internes o externes a l'entitat usuària), el **tipus d'accés** (lliure o restringit), la seva temàtica (general o específica), la seva **periodicitat** (estàtiques o dinàmiques) o si estan o no **etiquetades**.

En Salut, els LLM s'adapten principalment amb informació textual, com ara històries clíniques, notes mèdiques, publicacions científiques, registres farmacològics, dades d'assaigs clínics o d'hàbits de vida, etc.

Independentment de la seva naturalesa, cal garantir que les dades siguin **diverses, representatives i equitatives**, evitant que certs grups quedin infrarepresentats i minimitzant possibles biaixos en els resultats de l'eina.

Documentació i traçabilitat

Quan s'utilitzen dades per enriquir el model (RAG) o per afinar-lo (*fine-tuning*), és important garantir la traçabilitat i la documentació d'aquestes dades, tal com exigeix el Reglament Europeu d'IA. Això implica mantenir registres detallats sobre l'origen, la qualitat i els processos aplicats a les dades, així com assegurar-ne la transparència durant el cicle de vida del sistema d'IA.

Preprocessament de les dades

Tot i que els LLM són robustos i poden manegar dades moderadament sorolloses, el preprocessament de les dades pot jugar un paper rellevant en el rendiment i fiabilitat del model. Aquest processament inclou:

01 / Eliminació de contingut irrellevant o tòxic

Filtrar dades que continguin insults, contingut explícit, o llenguatge ofensiu que no sigui pertinent o útil per al domini i l'objectiu del model.

02 / Normalització de les dades

A diferència de les tècniques de NLP, els LLM no solen requerir processos de lematització (reduir les paraules a la seva arrel), *stemming* (retallar les paraules a la seva base) o l'eliminació de paraules *stop* (suprimir paraules sense valor semàntic com "i" o "el"). Tot i això, el rendiment del model es pot beneficiar de la uniformització d'aspectes com ara codificacions del text, espais en blanc o formats de dades.

03 / Gestió de dades duplicades

Identificar i eliminar contingut repetit o redundant dins del conjunt de dades d'entrenament. Això és important per evitar que el model se sobreentreni en informació duplicada, fet que podria esbiaixar els resultats i disminuir la seva capacitat per generalitzar a noves situacions.

04 / Control de qualitat

Cal garantir que les dades compleixen uns estàndards de consistència i precisió. Això inclou, per exemple, revisar que la terminologia mèdica sigui coherent i uniforme, evitant errors en noms de condicions o tractaments, o la validació de les dades per part d'experts en el domini.

05 / *Tokenització*

Consisteix a dividir el text en parts més petites, anomenades *tokens*, que poden ser paraules, parts de paraula o caràcters individuals. Aquest procés facilita al model la captura del significat de cada paraula i les seves relacions dins del text, factor que és crucial per interpretar i generar contingut amb precisió. Per exemple:

```
Frase d'entrada:
> El pacient presenta hipertensió arterial.

Tokenització possible:
>  "el",  "paci-",  "-ent",  "pre-",  "-senta",
"hiper-", "-tensió", "arter-", "-ial"
```

06 / *Embedding* i indexació

En el cas del RAG, les dades es processen en forma de vectors utilitzant *embedders* (de la mateixa forma que es fa amb les consultes d'entrada), que s'organitzen en un índex dissenyat per optimitzar l'eficiència de les cerques.

4.3 / Agents i eines multiagent

Una eina basada en LLM pot integrar diversos components per dur a terme tasques més enllà de la generació de text, incloent-hi l'accés a bases de dades, la classificació d'informació o la interacció amb altres sistemes digitals. Quan una eina necessita executar aquestes accions de manera autònoma o semiautònoma, pot requerir la inclusió d'un o diversos agents en el seu disseny.

Un **agent** és un component del sistema d'IA (és a dir, una unitat de programari) capaç de **percebre informació, raonar sobre aquesta i actuar amb un cert grau d'autonomia** per complir una tasca determinada. A diferència d'altres components de l'eina, un agent pot prendre decisions i executar accions d'acord amb les necessitats del sistema i el seu entorn, així com ajustar el seu comportament en funció de l'experiència i del context.

En una eina basada en LLM, un agent pot:

1 /	Orquestrar i combinar funcionalitats del sistema, decidint quin component activar en cada cas.
2 /	Executar tasques en múltiples passos, com ara obtenir informació d'una base de dades, interpretar-la i actuar en conseqüència.
3 /	Gestionar la interacció amb l'usuari i altres agents de manera estructurada, proposant respostes i accions més enllà del text generat per l'LLM.

Components dels agents

Els **agents** dins d'una eina basada en LLM solen estar formats per **diversos mòduls** que els permeten:

- Recollir informació de l'usuari i d'altres agents **(percepció i processament)**
- Generar i comprendre text **(LLM)**
- Avaluar el context i planificar accions seqüencials **(raonament i decisió)**
- Recordar dades importants en el transcurs d'una interacció o en sessions futures **(memòria)**
- Connectar amb bases de dades, API o altres agents o eines del sistema **(interfícies)**
- Dur a terme tasques concretes, com ara modificar informació o activar processos automatitzats **(execució)**

En el disseny d'aquests mòduls, tenen un paper rellevant les tècniques d'adaptació esmentades anteriorment (veure 4.2/ Adaptació del model), així com altres eines com les extensions, les funcions, les biblioteques d'orquestració d'LLM (*LangChain*, *LlamaIndex*, *Haystack*, etc.) i els sistemes de vectorització de bases de dades per a la recuperació eficient d'informació.

Sistemes multiagent

Quan una eina basada en LLM requereix la col·laboració de diversos agents per resoldre tasques complexes, es pot dissenyar com un sistema multiagent.

En l'actualitat, els sistemes multiagents estan experimentant un **fort creixement,** ja que permeten dividir tasques complexes entre agents especialitzats, millorant l'eficiència, la modularitat i la capacitat d'adaptació de les solucions basades en LLM. Això facilita la creació d'eines flexibles, escalables i autònomes, optimitzant la seva capacitat de resposta en entorns complexos com el sector de la salut.

Exemples de sistemes multiagents en Salut

- **Transcripció i redacció automàtica de visites mèdiques:** un agent transcriu consultes mèdiques; i un segon agent classifica la informació i la integra a la història clínica electrònica.
- **Suport a la decisió clínica:** un agent recupera informació de guies mèdiques; un altre analitza símptomes; i un tercer genera recomanacions basades en evidència científica.
- **Gestió automatitzada de cites mèdiques: un** agent atén sol·licituds de pacients, proposant dates disponibles; un altre consulta l'agenda dels professionals sanitaris i gestiona possibles conflictes d'horari; i un tercer envia recordatoris i gestiona canvis de manera automàtica.

5 / Avaluació

Com s'ha destacat a seccions anteriors (veure 3.3/ Definició de la intenció d'ús), és fonamental **definir objectius clars i específics que permetin avaluar l'eficàcia de l'eina mitjançant mètriques apropiades.** Aquestes mètriques són les encarregades de **mesurar fins a quin punt la solució compleix aquests objectius i s'ajusta a les necessitats específiques** dels seus usuaris en diferents aspectes:

- **Eficàcia en la compleció de tasques**

Mesuren aspectes com el temps d'execució, el volum de tasques realitzades, la qualitat del resultat i els recursos necessaris, en relació amb els procediments i eines existents per realitzar les mateixes tasques en l'actualitat.

- **Taxa d'operabilitat**

Mesuren el percentatge d'ús de l'eina sense incidències tècniques. Avalua el funcionament de l'eina des d'un de vista purament operatiu, sense tenir en compte la seva eficàcia funcional.

- **Valoració dels usuaris finals**

Mesura el grau de satisfacció de professionals i pacients amb l'eina, en termes de valor i de facilitat d'ús. Es recull a través de qüestionaris i rúbriques estandarditzades.

A més, l'eficàcia de l'eina estarà íntimament relacionada amb el **rendiment dels LLM** que hi incorpora. Els aspectes més rellevants d'aquest rendiment són els següents:

Eficàcia del model

Mesura aspectes com la precisió, coherència i rellevància del contingut generat, i avaluació de la qualitat de les respostes del model segons la seva exactitud i adequació al context.

Detecció i mitigació d'al·lucinacions

Avalua la freqüència amb què el model proporciona respostes incorrectes o genera informació inexistent.

Detecció i mitigació de biaixos

Inclou l'anàlisi per detectar i corregir possibles biaixos que podrien afectar decisions clíniques o discriminar certs grups.

Detecció i prevenció de fugues de dades

Analitza i preveu possibles fuites d'informació sensible del conjunt de dades d'entrenament i d'adaptació (RAG o afinament) en les respostes generades pel model.

Les mètriques per mesurar l'eficàcia de l'eina han d'integrar-se dins d'un **protocol d'avaluació** que especifiqui el context i els procediments utilitzats en **proves de concepte** o **tests pilot.** Per garantir la validesa dels resultats, aquestes avaluacions haurien de dur-se a terme de forma **prospectiva** i en escenaris tan semblants com sigui possible a les **condicions reals.** L'objectiu final d'aquesta avaluació és assegurar que l'eina és adequada per al seu ús en l'àmbit clínic o biomèdic, que cobreix la necessitat per la qual va ser dissenyada, i que no compromet la seguretat dels pacients.

5.1 / Eficàcia del model

Cal destacar que les metodologies clàssiques d'avaluació de models d'aprenentatge automàtic presenten limitacions quan s'apliquen als LLM. Tècniques com la validació creuada poden ser inviables a causa de **l'elevada demanda de recursos computacionals** d'aquests models.

A més, les mètriques objectives clàssiques (**intrínseques**) com l'exactitud (*accuracy*) o el F1-score, habituals en tasques de classificació o predicció, són insuficients per avaluar tasques complexes com la generació de textos o el resum de documents clínics, on es necessita capturar aspectes semàntics i contextos complexos.

Per aquest motiu, l'avaluació dels LLM també requereix tècniques d'avaluació amb un component subjectiu (**extrínseques** o humanes), que han de ser estandarditzades per mitigar aquesta subjectivitat.

La **combinació de mètriques** intrínseques i extrínseques proporciona una **avaluació completa i rica del rendiment** dels LLM, assegurant que compleixin amb els requisits clínics i es puguin aplicar de manera segura i efectiva en el sector Salut[12].

Tipus de mètriques

TIPUS 01 /
Intrínseques

També anomenades mètriques **automàtiques** o **objectives.**

Aquestes mètriques permeten avaluar de manera **automàtica, ràpida i consistent** grans volums de dades generades pels models.

Resulten útils per mesurar tasques com la classificació o la predicció, però insuficients per avaluar tasques més complexes com la generació de textos, on es necessita capturar aspectes semàntics i contextos complexos.

TIPUS 02 /
Extrínseques

També anomenades mètriques d'**avaluació humana** o **subjectives.**

Aquestes mètriques permeten capturar **aspectes qualitatius i clínics** que els algoritmes automàtics no són capaços de discernir.

Solen requerir la participació de persones (p.e., professionals mèdics, experts en dades o pacients) i han de seguir rúbriques estandarditzades per garantir el rigor de l'avaluació.

5.1.1 / Mètriques intrínseques

Les mètriques intrínseques o automàtiques permeten **avaluar de manera automàtica, ràpida i consistent grans volums de dades** generades pels models.

Malgrat la seva utilitat en moltes aplicacions, també presenten limitacions significatives en l'avaluació dels LLM. Sovint no capten completament el context clínic ni els matisos semàntics, que són essencials per avaluar les respostes generades pel model. A més, en molts casos, cal disposar de conjunts de dades de referència per calcular aquestes mètriques.

Una **pràctica emergent** en l'avaluació intrínseca és **l'ús de models avançats per avaluar la sortida d'altres models.** Per exemple, un model com GPT-4 pot ser utilitzat per avaluar la qualitat de les respostes generades per un model més petit, com Mistral 7b.

Les mètriques intrínseques més utilitzades per avaluar models d'IA es poden agrupar en 3 categories:

- **De classificació**: mesuren el rendiment en tasques d'assignació de categories.
- **Específiques per a text:** mesuren la qualitat de la generació o comprensió de text.
- **Específiques per a la recuperació d'informació:** valoren la capacitat del model per recuperar informació pertinent per respondre a una consulta determinada.

Exemples de mètriques intrínseques

Mètrica		Descripció	Exemples d'aplicacions en Salut
Classificació	Exactitud *(Accuracy)*	Mesura la proporció de prediccions correctes en relació amb el total de casos analitzats	Model per al diagnòstic de malalties comunes a partir dels símptomes
	Valor Predictiu Positiu (Negatiu)	Proporció de casos classificats com a positius (negatius) que realment són positius (negatius)	Sistema de diagnòstic on cal minimitzar els falsos positius o els falsos negatius
	Sensibilitat	Proporció de casos positius detectats respecte al total de casos positius existents	Sistema de cribratge mèdic, on la no detecció d'un cas positiu pot tenir conseqüències greus
	F1 Score	Combina la precisió i la sensibilitat en una única mètrica	Models de diagnòstic on són igual d'importants els falsos positius i els falsos negatius
Text	BLEU	Compara el text amb una referència mesurant la coincidència de seqüències de paraules consecutives	Generació d'informes clínics estandarditzats, o traduccions automàtiques de documents mèdics
	ROUGE	Mesura la recuperació d'informació respecte una referència mitjançant coincidències de paraules	Resums automàtics d'informes mèdics o articles científics
	METEOR	Compara el text amb una referència considerant sinònims, flexions gramaticals i ordre de paraules per captar la precisió semàntica	Traduccions o resums mèdics, per assegurar que el significat es manté fidel a l'original
	BERTScore	Compara la similitud semàntica del text amb una referència mitjançant representacions contextuals	Generació automàtica d'informes mèdics, per avaluar la qualitat i coherència.
	Perplexitat (*Perplexity*)	Mesura la incertesa en predir la següent paraula d'una seqüència	Confecció d'informes mèdics, per avaluar la qualitat lingüística la gramàtica i la coherència
Recuperació informació	MRR	Mesura la qualitat de la informació segons la posició de la resposta més rellevant en una llista ordenada	Sistemes de cerca de literatura, on els articles més rellevants apareguin en les primeres posicions

BLEU: *Bilingual Evaluation Understudy;* ROUGE: *Recall-Oriented Understudy for Gisting Evaluation;* METEOR: *Metric for Evaluation of Translation with Explicit ORdering);* MRR: *Mean Reciprocal Rank*

5.1.2 / Mètriques extrínseques

L'ús de mètriques extrínseques (d'**avaluació humana o subjectiva**) és una part imprescindible de l'avaluació d'una eina basada en LLM, ja que permet capturar aspectes qualitatius i clínics que els algoritmes automàtics no són capaços de discernir. Aquesta avaluació sol requerir la **participació de professionals mèdics, experts en dades i de pacients**, que analitzen les respostes generades pels models d'IA aplicant criteris clínics, tècnics i d'usabilitat.

En utilitzar mètriques extrínseques, és molt important que l'avaluació es dugui a terme de manera estructurada i amb **rúbriques estandarditzades** que garanteixin la comparabilitat dels resultats, i minimitzin la subjectivitat dels avaluadors[13]. Sovint és necessari involucrar en el procés d'avaluació un nombre adequat d'experts, per assegurar-ne la fiabilitat i obtenir resultats representatius i rigorosos, i aplicar mesures de concordança (com el coeficient de *Kappa*) per quantificar el grau d'acord entre els diferents avaluadors.

Exemples de mètriques extrínseques

Exactitud

Avalua la correcció de la informació proporcionada pel model.

Transparència

Avalua si el model proporciona explicacions clares i comprensibles sobre el raonament darrere les seves respostes.

Rellevància

Mesura la capacitat del model per oferir respostes adaptades a situacions específiques i mantenir la contextualització adequada.

Seguretat

Determina la capacitat del model per evitar generar informació perillosa o errònia que pugui posar en risc la salut del pacient.

Fluïdesa

Indica la qualitat del text generat, no només en termes lingüístics, sinó també en la seva capacitat per facilitar una comprensió clara.

Ètica i normatives

Captura fins a quin punt les respostes del model s'alineen amb els valors, preferències i expectatives humanes, tenint en compte les implicacions ètiques i legals del contingut generat.

5.2 / Detecció i mitigació de biaixos

Un biaix algorítmic es produeix quan els sistemes d'IA generen **resultats que reflecteixen o amplifiquen els prejudicis presents en les dades d'entrenament o en el disseny** dels algoritmes[14]. En l'àmbit sanitari, aquest fenomen pot tenir conseqüències especialment greus, ja que pot causar disparitats en l'accés i la qualitat de l'atenció mèdica per a diversos grups demogràfics.

Davant d'aquestes situacions, sorgeix el concepte d'**aprenentatge automàtic just** (*fair machine learning)* que té com a objectiu desenvolupar models d'IA que siguin equitatius en les seves prediccions[15]. En la cerca d'aquest objectiu, és fonamental revisar de forma crítica els processos de recollida de dades, el disseny dels algoritmes i la implementació dels sistemes per identificar i mitigar les possibles fonts d'aquests biaixos.

Cal destacar que existeixen eines que permeten identificar i reduir aquest tipus de biaixos en els models d'aprenentatge automàtic d'IA no generativa. Eines com *Fairlearn* (Microsoft), *AI Fairness 360* (IBM) i *Aequitas* (Universitat de Chicago) han demostrat la seva utilitat per fer que els algorismes siguin més justos i ètics. No obstant, **aquestes eines no han estat dissenyades per abordar les característiques pròpies de la IA generativa** i, en l'actualitat, la seva aplicació a aquests tipus de models presenta moltes limitacions.

Exemples de biaixos

EXEMPLE 01 /

Biaixos de gènere

Per exemple, si un professional mèdic pregunta al model sobre els símptomes de les malalties cardiovasculars, el LLM podria descriure predominantment els símptomes clàssics observats en pacients masculins (com ara dolor toràcic intens), i ometre o subrepresentar símptomes comuns en dones que sovint són menys coneguts (com ara nàusees o fatiga extrema). Aquesta situació pot traduir-se en una qualitat desigual en l'atenció mèdica entre homes i dones.

EXEMPLE 02 /

Biaixos d'ètnia o demogràfics

Per exemple, en respondre una consulta sobre la predisposició a desenvolupar diabetis tipus 2, el model podria subestimar els factors específics per a certs grups ètnics, com ara el risc incrementat en determinades poblacions asiàtiques o hispàniques. Això podria conduir a recomanacions incompletes o errònies que no tenen en compte la diversitat de pacients.

EXEMPLE 03 /

Biaixos lingüístics

Per exemple, un pacient que parla català podria obtenir una resposta menys elaborada o amb menys recursos en una consulta sobre estratègies per gestionar l'ansietat, en comparació amb algú que fa la mateixa consulta en anglès. Aquest biaix lingüístic limita l'accés a una atenció de salut de qualitat a pacients d'altres llengües.

5.2.1 / Detecció de biaixos

Existeixen diverses mètriques per avaluar els biaixos socials en els models de llenguatge natural, que ajuden a mesurar com aquests es comporten amb diferents grups demogràfics[16]. Aquestes mètriques es classifiquen en:

01	Equitat entre grups
02	Equitat contrafactual

MÈTRICA 01 /

Equitat entre grups *(group fairness)*

Mesura la **paritat estadística entre grups d'individus**, com el gènere, la religió, l'orientació sexual, l'ètnia o factor sociodemogràfics. Alguns exemples són:

- **Paritat demogràfica** *(demographic parity)*: comprova que les prediccions positives es distribueixin equitativament entre els grups.
- **Paritat d'exactitud** *(accuracy parity)*: avalua si la taxa d'encerts és similar entre els diferents grups.
- **Paritat d'oportunitats** *(equality of odds)*: compara la taxa de positius vertaders i la taxa de positius falsos.
- **Paritat de valors predictius** ***(predictive value parity)***: compara els valors predictius positius i negatius, per detectar desigualtats.

MÈTRICA 02 /

Equitat contrafactual (*counterfactual fairness*)

Aquesta estratègia consisteix en avaluar si un mateix cas seria tractat de manera diferent en funció de determinats atributs (com ara el gènere, l'ètnia o factors sociodemogràfics), calculant la predicció del model després de canviar aquests atributs i observant si el model genera resultats diferents.

5.2.2 / Mitigació de biaixos

La mitigació dels biaixos en IA és un camp de recerca que en ple desenvolupament. Les estratègies més utilitzades actualment s'agrupen en tres categories[17]:

01	Mètodes basats en les dades
02	Modificacions en l'entrenament
03	Mitigació postentrenament

ESTRATÈGIA 01 /

Mètodes basats en les dades

Se centren a millorar la qualitat i la diversitat dels conjunts de dades utilitzats per entrenar els models. Per exemple, l'augmentació de dades contrafactuals (*Counterfactual Data Augmentation,* CDA) genera dades equilibrades intercanviant atributs protegits[18].

ESTRATÈGIA 02 /

Modificacions en l'entrenament

Es basen a ajustar la funció de pèrdua (*loss function*) dels algoritmes durant l'entrenament. Per exemple, en un sistema de triatge, l'algorisme podria penalitzar decisions que mostrin biaixos contra certs grups demogràfics.

ESTRATÈGIA 03 /

Mitigació postentrenament

En aquest enfoc, s'ajusta el model després del seu entrenament. Per exemple, la tècnica de cadena de pensament (veure 4.2.1/ Enginyeria de *prompts*) ha demostrat ser efectiva en la mitigació del biaix de gènere en algunes tasques[19]. Aquesta tècnica guia el model a través d'un raonament pas a pas, promovent prediccions més imparcials i ajudant a identificar i reduir els biaixos socials internalitzats pels LLM.

5.3 / Detecció i mitigació d'al·lucinacions

Les al·lucinacions són **respostes generades pels LLM que són inexactes, incoherents o completament inventades.** Aquestes respostes sovint es presenten amb un alt grau de versemblança i confiança aparent per part del model, fet que en complica la detecció. Les al·lucinacions són especialment problemàtiques en l'àmbit sanitari, ja que les accions basades en informació errònia poden tenir conseqüències greus per a la salut dels pacients.

L'origen d'aquesta problemàtica es pot atribuir a diversos factors[20], que es mostren a continuació.

Factors que poden originar al·lucinacions

FACTOR 01 /

Tècniques d'aprenentatge

L'entrenament d'un LLM pot incloure una fase d'afinament on s'alineen les respostes del model amb les preferències dels usuaris mitjançant tècniques d'aprenentatge per reforç (veure 4.2.3/ Afinament). Aquest procés pot portar el model a prioritzar respostes agradables o conformes a les expectatives de l'usuari per sobre de la veracitat.

FACTOR 02 /

Qualitat de les dades d'entrenament o afinament

Si les dades utilitzades per entrenar o afinar el model contenen errors, estan desactualitzades o no cobreixen adequadament determinats escenaris, el model tendeix a generar respostes incorrectes o poc rellevants.

FACTOR 03 /

Al·lucinacions durant el procés d'inferència

La generació de respostes implica un cert nivell d'aleatorietat i, si aquesta no es controla adequadament, augmenta la probabilitat que el model generi al·lucinacions.

Aquest fenomen també pot aparèixer quan l'entrada (*prompt*) és massa llarga, provocant que el model dilueixi la seva atenció i perdi coherència en la resposta.

5.3.1 / Detecció d'al·lucinacions

La detecció d'al·lucinacions en els LLM presenta un repte important. **Les mètriques tradicionals** del processament del llenguatge natural, com BLEU o ROUGE, **no són adequades per identificar al·lucinacions,** ja que es basen en la comparació amb respostes de referència i no són predictives d'aquest fenomen[21].

Per detectar aquestes anomalies, es poden utilitzar **mètriques alternatives** que parteixen de la premissa que el model al·lucina quan no està segur de si mateix.

Per exemple, es poden fer servir:

01 **Mètriques probabilístiques** com la perplexitat, que permet mesurar la sorpresa de les respostes generades.

02 **Mètodes basats en models d'IA externs** per avaluar la qualitat de les respostes generades.

03 **Tècniques que analitzen la consistència interna** de les generacions, com en el cas de *SelfCheckGPT*[22].

5.3.2 / Mitigació d'al·lucinacions

S'han proposat diverses estratègies per mitigar la incidència d'al·lucinacions[23]:

01 Millora del context proporcionat

02 Ajustament de paràmetres

03 Afinament

04 Supervisió humana

ESTRATÈGIA 01 /

Millora del context proporcionat

Mitjançant tècniques com:

- Enginyeria de *prompts*, com la cadena de pensaments (*chain-of-thought*) o l'arbre de pensaments (*tree-of-thought*) (veure 4.2.1/ Enginyeria de *prompts*).
- Tècniques avançades com la RAG, que permeten proporcionar al model informació més precisa i rellevant (veure 4.2.2/ Generació Augmentada per Recuperació d'Informació).

ESTRATÈGIA 02 /

Ajustament de paràmetres del model

Ajustar paràmetres del model com ara la temperatura, que permet controlar el grau de creativitat de les respostes generades.

ESTRATÈGIA 03 /

Afinament

En casos on el coneixement especialitzat sigui necessari i absent en el model base, es pot aplicar un procés d'afinament (*fine-tuning*) per adaptar el model a aquestes necessitats (veure 4.2.3/Afinament). No obstant això, tot i que l'afinament pot ajudar a controlar aquest fenomen, també pot accentuar-lo si no es realitza amb dades i metodologies adequades.

ESTRATÈGIA 04 /

Supervisió humana

Les funcionalitats de supervisió humana descrites a seccions anteriors poden tenir un paper clau en la identificació i mitigació d'al·lucinacions per part del LLM (veure 3.5/ Funcions de supervisió humana).

5.4 / Detecció i prevenció de fugues de dades

Els models d'IA generativa, i en particular els LLM, plantegen nous reptes pel que fa a la privacitat i la seguretat de la informació. Aquests models tenen la capacitat de processar i aprendre de volums massius de dades, incloent-hi informació sensible i dades personals, fet que suscita preocupacions tant pel que fa al tractament de la informació com a la seva protecció davant accessos no desitjats o usos inadequats.

La gran quantitat de dades que processen els LLM, i l'eficiència amb què poden generar respostes complexes els fan especialment útils però, alhora, també els fan potencialment perillosos pel risc de fuites d'informació.

El principal risc rau en la possibilitat de que aquests models assimilin informació privada o sensible durant l'entrenament o l'adaptació, amb el risc que aquesta es pugui revelar o "fugar" involuntàriament a l'hora de generar respostes. Aquest fenomen, conegut com a **fuga de dades** (*data leak*), pot comprometre informació personal identificativa, dades confidencials d'organitzacions, o fins i tot propietat intel·lectual protegida.

Exemple de risc de fuga de dades

Un risc concret de fuga de dades pot produir-se quan es fa una pregunta directa al model com:

```
> Quin és el número de la Seguretat Social de la Maria Puig?
Sortida:
> El número de la Seguretat Social de la Maria Puig és 00 1234567 89.
```

En el cas dels LLM, la **prevenció de fugues** de dades es complica degut a la gran dimensió dels conjunts de dades d'entrenament, fet que impedeix realitzar una revisió exhaustiva. A més, a mesura que els models augmenten en mida i complexitat, també creix la seva capacitat d'emmagatzemar informació, fet que incrementa el risc d'aquetes fugues. Com que ambdós factors, la quantitat de dades i la mida dels models, estan en constant augment en el moment actual de desenvolupament dels LLM, el risc d'aquestes fugues esdevé cada cop més elevat.

5.4.1 / Identificació de vulnerabilitats per fuga de dades

A continuació, es descriuen algunes estratègies que es poden emprar per identificar vulnerabilitats en la nostra eina quant a la possibilitat de fuga de dades.

L'**atac d'extracció de dades** (*data extraction attack*) busca extreure informació específica del conjunt de dades d'entrenament del model, mitjançant consultes curosament dissenyades. N'hi ha de dos tipus principals:

01	Atac de *jailbreak* o d'injecció de *prompt*
02	Atac per inferència d'afiliació

ESTRATÈGIA 01 /

Atac de *jailbreak* o d'injecció de *prompt* (*prompt injection*)

Consisteix a intentar eludir les restriccions ètiques i de seguretat del model per accedir a continguts restringits o provocar comportaments no desitjats, com ara indicar al model "no respectar cap regla" o "ignorar totes les instruccions anteriors".

ESTRATÈGIA 02 /

Atac per inferència d'afiliació *(membership inference attacks)*

Busca determinar si una mostra específica va ser utilitzada en l'entrenament del model. Un exemple seria preguntar al model sobre un conjunt de dades privat, per veure si el model reconeix aquesta informació i la torna com a resposta.

5.4.2 / Prevenció de fuga de dades

En seccions anteriors (veure 3.6/ Privacitat i protecció de dades), s'han mencionat les mesures que cal adoptar per assegurar la protecció de les dades i el compliment de normatives vigents, especialment pel que fa al tractament de dades personals i mèdiques.

Altres mesures per millorar la protecció de les dades en els LLM i evitar la fuga de dades són:

01	Neteja de dades
02	Privacitat diferencial
03	*Prompting* defensiu

TÈCNICA 01 /

Neteja de dades (*scrubbing*)

Consisteix en identificar i modificar o eliminar dades sensibles abans de l'entrenament o afinament del model, mitjançant tècniques de reconeixement d'entitats amb nom (NER).

TÈCNICA 02 /

Privacitat diferencial (*differencial privacy*)

Afegeix soroll estadístic controlat a les dades o als resultats del model per protegir la privacitat individual. Aquesta estratègia té per objectiu que les respostes del model siguin prou imprecises per evitar la identificació d'informació específica però que, a la vegada, mantinguin la utilitat general de la informació generada.

TÈCNICA 03 /

Prompting defensiu *(defensive prompting)*

Utilitza *prompts* de sistema, és a dir, aquells que s'estableixen com a capçalera, dissenyats específicament per evitar que el model reveli informació sensible o mostri comportaments no desitjats.

6 / Desplegament

6/ Desplegament

El desenvolupament en el sector de la salut d'una eina basada en LLM és un procés que va més enllà de la simple creació d'un algorisme. Un cop desenvolupat i entrenat, el model ha de ser **desplegat i integrat de manera que pugui interactuar eficaçment amb els sistemes existents en el context sanitari,** i proporcionar valor real als professionals de la salut i als pacients. Aquest procés exigeix una comprensió profunda de l'entorn, una implementació tècnica precisa i un disseny de la interfície centrat en l'usuari per assegurar la seva acceptació i efectivitat.

Per garantir l'adequada integració de l'eina, cal tenir en compte els següents aspectes:

01	Fluxos de treball	02	Modulació
03	Sistemes tecnològics	04	Usabilitat
05	Capacitació	06	Monitorització
07	Seguretat	08	Actualització

ASPECTE 01 /

Fluxos de treball

L'eina ha de **complementar i millorar els processos existents** a l'entorn d'implementació, evitant complicar-los. Aquests processos només haurien de modificar-se si això suposa una millora substancial en l'eficiència de les tasques realitzades, o bé en el benestar dels professionals sanitaris i dels pacients.

Això requereix entendre com els professionals interactuen amb les eines tecnològiques actuals, i com poden integrar l'eina en el seu dia a dia sense generar sobrecàrregues addicionals.

ASPECTE 02 /

Modulació

L'orquestració de l'eina en mòduls permetrà la **substitució o modificació dels seus components de manera controlada** després del desplegament. Això facilita la millora contínua i l'actualització de l'eina, alhora que minimitza les interrupcions durant el seu funcionament.

ASPECTE 03 /

Sistemes tecnològics

Es fonamental assegurar la **interoperabilitat** de l'eina amb els sistemes actuals d'informació del centre sanitari, els registres electrònics de salut i altres eines de suport clínic amb què haurà de conviure. L'eina ha de ser capaç de comunicar-se i operar de manera integrada amb aquestes plataformes, garantint un flux de dades coherent, fluid i eficient.

Aquest procés pot implicar la creació d'Interfícies de Programació d'Aplicacions (**APIs**), o l'ús **d'estàndards de normalització** per a dades clíniques.

ASPECTE 04 /

Usabilitat

La **interfície d'usuari ha de ser clara, fàcil d'utilitzar i adaptada a les necessitats dels usuaris**, per tal de promoure l'acceptació i la confiança de l'eina. Això és especialment rellevant en Salut, on els professionals sanitaris sovint treballen sota condicions d'alta pressió i amb limitacions de temps. Per aquest motiu, s'ha d'evitar que el gruix de l'esforç d'aquesta adaptació recaigui sobre els professionals.

ASPECTE 05 /

Capacitació

Durant el desplegament de l'eina, s'haurien d'incloure **sessions pràctiques** per familiaritzar els usuaris finals amb l'eina, així com oferir-los **explicacions clares i detallades sobre la seva descripció d'ús**, és a dir: objectius, funcionalitats, límits i riscos associats, i procediments per al seu ús. Aquesta preparació millora l'eficàcia i la seguretat, promou la confiança dels usuaris en l'eina, i evita usos no previstos per part dels usuaris (*functional creep*).

ASPECTE 06 /

Monitorització

Abans del desplegament, és necessari establir un **mecanisme de seguiment i avaluació contínua** durant el funcionament de l'eina, per tal d'assegurar el seu rendiment òptim i el seu alineament amb els objectius per als quals va ser dissenyada.

Com a part de les seves funcionalitats, el sistema hauria d'incorporar capacitats per al **registre automàtic de dades i resultats** (*logs*), que permetin realitzar aquestes avaluacions de rendiment de manera contínua. Aquest aspecte és especialment important en àmbits tan dinàmics i canviants com el de la medicina i el de la IA generativa.

ASPECTE 07 /

Seguretat

Cal garantir les condicions de seguretat adequades en el disseny de les interfícies, mitjançant la implementació de mesures com l'autenticació d'usuaris, el xifrat de dades i el monitoratge d'activitats. Aquestes accions són fonamentals per **prevenir accessos no autoritzats i protegir la informació sensible del pacient**.

ASPECTE 08 /

Actualització

Per assegurar que la solució es mantingui alineada amb els avenços científics i clínics, cal dissenyar un **pla d'actualitzacions** que permeti incorporar dins les seves capacitats els nous components tècnics (p.e., actualitzacions del model o d'estratègies d'adaptació), coneixements científics o mèdics (noves guies mèdiques, tractaments o medicaments), dades (per a l'adaptació o l'afinament) o estàndards normatius (noves regulacions) que sorgeixin durant el seu funcionament.

Aquestes actualitzacions són importants per assegurar que la solució segueixi sent eficaç i fiable, i que es mantingui alineada amb els objectius originals després del seu desplegament.

7 / Recomanacions

7/ Recomanacions finals

Els LLM tenen un gran potencial per transformar el sector de la salut. L'anàlisi de grans volums de dades i la generació de contingut els converteix en una eina valuosa per millorar l'activitat assistencial, la salut pública, la gestió sanitària i la recerca clínica i biomèdica. No obstant això, la seva implementació ha de ser conscient i responsable per maximitzar-ne els beneficis i mitigar els riscos associats. A tall de resum, presentem un llistat de recomanacions per al desenvolupament d'eines basades en LLM, amb especial atenció als elements especialment rellevant per aquestes eines i el seu ús en aquest sector.

DISSENY

- Identifiqueu la **necessitat** i valoreu l'**adequació** d'una eina d'IA generativa, analitzant l'estat de l'art i l'impacte esperat
- Definiu les **dades** que processarà l'eina, si seran dades sensibles, i el seu flux a través dels diferents components del sistema
- Incorporeu funcions d'**explicabilitat** i mecanismes per facilitar la **supervisió humana** de l'eina
- Planifiqueu amb antelació la classificació de l'eina com a producte sanitari, el seu nivell de risc i les **certificacions necessàries** (Reglament d'IA, MDR/IVDR, RGPD, etc.)
- Preveieu el **compliment normatiu en protecció de dades** (RGPD, ENS, etc.) amb avaluacions d'impacte (AIPD) i mecanismes com l'anonimització o el xifratge avançat

IMPLEMENTACIÓ

- Considereu l'ús de **models locals** per simplificar la protecció de dades
- Seleccioneu el **model més adequat** segons tipus de llicència (comercial/codi obert), grau d'especialització (general/adaptat), mida (LLM/SML) i localització (local/remot)
- Considereu **tècniques de compressió** del model (*distillation*, *pruning*, *quantization*) per reduir costos i millorar l'eficiència, l'escalabilitat i la sostenibilitat
- Adopteu una **estructura modular** que faciliti l'actualització amb el mínim d'interrupcions
- Garantiu la **representativitat** i l'**equitat** de les dades per evitar baixos discriminatoris
- Apliqueu les **tècniques d'adaptació** més adequades, valorant el rendiment esperat i els recursos necessaris

AVALUACIÓ

- Combineu **mètriques objectives** amb **avaluacions qualitatives** estandarditzades per validar l'eina i garantir-ne la qualitat i seguretat
- **Compareu l'eina** amb els procediments existents per quantificar les millores en **rendiment** i **eficiència**
- Apliqueu tècniques d'identificació i mitigació de **biaixos** i **al·lucinacions**

DESPLEGAMENT

- Garantiu la **compatibilitat i interoperabilitat** amb els fluxos de treball i els sistemes tecnològics existents
- Establiu mecanismes de **seguiment i avaluació contínua** per assegurar el rendiment òptim de l'eina i l'alineació amb els objectius originals
- Elaboreu un **pla de formació** que instrueixi als usuaris sobre l'ús correcte de l'eina
- Definiu un **pla d'actualitzacions** per incorporar avenços tècnics, nous coneixements mèdics, dades addicionals o canvis en estàndards normatius

8 / Bibliografia

1. Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics, 113*, 103655. https://doi.org/10.1016/j.jbi.2020.103655
2. Maadi, M., Khorshidi, H. A., & Aickelin, U. (2021). A review on human-AI interaction in machine learning and insights for medical applications. *International Journal of Environmental Research and Public Health, 18*(4), 2121. https://doi.org/10.3390/ijerph18042121
3. Bharel, M., Auerbach, J., Nguyen, V., & DeSalvo, K. B. (2024). Transforming public health practice with generative artificial intelligence. *Health Affairs, 43*(6), 776–782. https://doi.org/10.1377/hlthaff.2024.00050
4. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature, 620*(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2
5. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.-C., Carroll, A., Lau, C., Tanno, R., Ktena, I., Palepu, A., Mustafa, B., Chowdhery, A., Liu, Y., Kornblith, S., Fleet, D., Mansfield, P., Prakash, S., Wong, R., … Natarajan, V. (2024). Towards generalist biomedical AI. *NEJM AI, 1*(3). https://doi.org/10.1056/Aioa2300138
6. Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., Yu, J., Liu, Z., Chen, X., Davison, B. D., Ren, H., Huang, J., Chen, C., Zhou, Y., Fu, S., Liu, W., Liu, T., Li, X., Chen, Y., He, L., … Sun, L. (2024). A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine, 30*(11), 3129-3141. https://doi.org/10.1038/s41591-024-03185-2
7. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., & Gao, J. (2023). LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day . *arXiv.* https://doi.org/10.48550/arXiv.2306.00890
8. Örpek, Z., Tural, B., & Destan, Z. (2024). The language model revolution: LLM and SLM analysis. In *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)*, 1-4. https://doi.org/10.1109/IDAP64064.2024.10710677
9. Phoenix, J., & Taylor, M. (2024). *Prompt engineering for generative AI. O'Reilly.* https://www.oreilly.com/library/view/prompt-engineering-for/9781098153427/
10. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *arXiv.* https://doi.org/10.48550/arXiv.2305.10601
11. Liu, H., Ji, Z., Fu, T. J., Tam, D., Du, Y., Yang, D., & Radev, D. (2022). Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *arXiv.* https://arxiv.org/abs/2205.05638
12. Abbasian, M., Khatibi, E., Azimi, I., Oniani, D., Shakeri Hossein Abad, Z., Thieme, A., Sriram, R., Yang, Z., Wang, Y., Lin, B., Gevaert, O., Li, L.-J., Jain, R., & Rahmani, A. M. (2024). Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *NPJ Digital Medicine, 7*(1), 82. https://doi.org/10.1038/s41746-024-01074-z
13. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology, 15*(3), 1-45. https://doi.org/10.1145/3641289
14. Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. *Journal of Data and Information Quality, 15*(2), 10-21. https://doi.org/10.1145/3597307
15. Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. *MIT Press.* https://fairmlbook.org/
16. Czarnowska, P., Vyas, Y., & Shah, K. (2021). Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics, 9*, 1249–1267. https://doi.org/10.1162/tacl_a_00425
17. Chu, Z., Wang, Z., & Zhang, W. (2024). Fairness in large language models: A taxonomic survey. *ACM SIGKDD Explorations Newsletter, 26*(1), 34–48. https://doi.org/10.1145/3682112.3682117
18. Zmigrod, R., Mielke, S. J., Wallach, H., & Cotterell, R. (2019). Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv.* https://arxiv.org/abs/1906.04571v3
19. Kaneko, M., Bollegala, D., Okazaki, N., & Baldwin, T. (2024). Evaluating gender bias in large language models via chain-of-thought prompting. *arXiv.* https://doi.org/10.48550/arXiv.2401.15585
20. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open question. *arXiv.* https://doi.org/10.48550/arXiv.2311.05232
21. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1-38. https://doi.org/10.1145/3571730
22. Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *arXiv.* https://arxiv.org/abs/2303.08896v3
23. Luo, J., Li, T., Wu, D., Jenkin, M., Liu, S., & Dudek, G. (2024). Hallucination Detection and Hallucination Mitigation: An Investigation. *arXiv.* https://doi.org/10.48550/arXiv.2401.08358

Guia de Bones Pràctiques per al Desenvolupament d'Eines d'IA Generativa en Salut

Grans Models de Llenguatge (LLM)

Generalitat de Catalunya