

Guia de prompting pels bots de conversa de la Generalitat de Catalunya

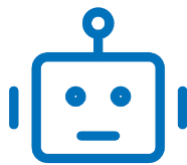
Manual per als treballadors/es
i per als proveïdors/es

Desembre 2024

Índex

1. <u>Introducció a la IA Generativa</u>	4
<u>Glossari</u>	4
<u>Consideracions prèvies</u>	7
<u>Metodologia</u>	8
2. <u>Prompting i bones pràctiques</u>	11
<u>Usos prompting</u>	11
<u>Com estructurar un <i>prompt</i></u>	14
<u>Bones pràctiques de <i>prompt engineering</i></u>	16
<u>Tècniques bàsiques de <i>prompting</i></u>	19
<u>Hiperparàmetres LLMs</u>	22
<u>Models LLMs</u>	24
<u>Bones pràctiques de <i>prompt</i> amb RAG</u>	27
<u>Avaluació</u>	30
<u>Exemples <i>prompts</i></u>	33
<u>Plantilles d'exemple</u>	42
3. <u>Configuració Databricks</u>	47

1. Introducció a la IA Generativa



1.Introducció a la IA Generativa

Glossari

A continuació es descriuen alguns dels termes bàsics en l'àmbit de la IA Generativa que es faran servir al llarg de la guia i ajudaran a entendre cada apartat.

■ IA Generativa

Tipus d'intel·ligència artificial que pot crear contingut nou, com ara text, imatges, música o vídeo, a partir de patrons apresos de dades existents. Genera nous continguts, imitant la creativitat humana, més enllà d'analitzar dades.

■ *Prompt*

Entrada de text/instrucció que es proporciona a un model d'intel·ligència artificial perquè generi una resposta. Pot ser una pregunta, una descripció o qualsevol altra entrada que desencadeni una resposta del model. Per exemple, "Explica què és un arbre" seria un *prompt* per a un model de llenguatge.

■ Prompt engineering

Disciplina que consisteix en crear *prompts* efectius perquè els models d'intel·ligència artificial proporcionin les respostes desitjades. Això pot implicar ajustar el llenguatge, l'ordre de les paraules o fins i tot la llargada del *prompt* per obtenir la millor resposta possible del model. En resum, es tracta de dissenyar, provar i iterar sobre una instrucció per aconseguir un resultat òptim.

■ LLM

Sigles corresponents a Large Language Model. El LLM és un tipus específic de model de la IA Generativa, dissenyat per processar i generar text en llenguatge natural. Està entrenat amb grans quantitats de text per entendre, completar i generar respostes coherents a partir d'un context donat. Pot respondre preguntes, redactar textos o mantenir converses de manera similar a com ho faria una persona.

■ Token

Unitat bàsica de text que un model de llenguatge, com el LLM, processa. Pot ser una paraula completa, una part d'una paraula, o fins i tot un sol caràcter. Els models divideixen el text en *tokens* per poder analitzar-lo i generar respostes. El procés de descomposició d'un text en *tokens* varia segons els models, però 100 *tokens* poden equivaldre a 75 paraules.

■ Hiperparàmetre

Configuració que s'estableix abans de l'entrenament o durant el procés de generació de text i que controla com funciona el model. Aquests hiperparàmetres afecten la qualitat, la velocitat i la naturalesa de les respostes generades.

■ RAG

Sigles que corresponen a *Retrieval-Augmented Generation*. El RAG utilitza un component de recuperació per cercar informació rellevant en una base de dades externa o en documents durant la generació de text. Això permet al model generar respostes més precises i detallades perquè accedeix a informació actualitzada o específica de la qual originalment no disposava.

■ ***Fine-tuning***

Procés d'ajust d'un model preentrenat (com un LLM) amb un conjunt de dades específic per adaptar-lo a tasques particulars o millorar el seu rendiment en contextos concrets. A diferència del RAG, el *fine-tuning* modifica les dades d'entrenament del model i pot ser adequat quan necessitem terminologia especialitzada.

■ ***System prompt***

Instrucció que defineix el rol, comportament o context general del model de llenguatge. Estableix com ha de respondre el model durant tota la interacció.

■ ***User prompt***

Sol·licitud específica feta per l'usuari/a que guia la resposta del model en un moment particular de la interacció. Si tenim un *system prompt* definit, la resposta específica generada a la nostra sol·licitud la farà dins del marc que hàgim definit.

■ ***Script***

Conjunt de codi que s'executa de manera seqüencial per automatitzar tasques específiques. En el context d'aquesta guia, serveix per enviar prompts al model i recollir les respostes de manera automàtica.

Consideracions prèvies

Abans de començar a desenvolupar el nostre *prompt*, hem d'haver superat les diferents fases de definició i disseny. Per tant, hauríem d'haver dut a terme les següents tasques:

- 1 Definir l'objectiu i l'ús de la IA Generativa en el/la nostre/a assistent virtual.
- 2 Definir les tasques específiques que el model d'IA haurà de realitzar.
- 3 Seleccionar model LLM. Hem d'escollir el model base, configuracions i qualsevol ajust personalitzat necessari segons les tasques i cas d'ús.
- 4 Tenir les dades necessàries, com ara els documents amb la nostra informació en cas d'utilitzar un RAG.
- 5 Tenir l'arquitectura del sistema i les integracions necessàries per poder executar i provar els nostres prompts.

Metodologia

Un cop ho tenim tot enllestit, comencem a construir i testejar, seguint els següents punts:

■ Desenvolupament inicial del *prompt*

Basant-nos en els objectius definits en les fases anteriors, comencem a redactar el(s) *prompt*(s). Creem múltiples versions de *prompts* que abordin els diferents aspectes i necessitats del projecte. Han de ser clars, concisos i orientats a obtenir respostes específiques del model.

■ Iteració i refinament del *prompt*

Fem proves internes amb els *prompts* desenvolupats per generar respostes inicials del model. Avaluarem la coherència, rellevància i precisió de les respostes.

■ Desenvolupament de *scripts*

Per enviar els *prompts* al model i recollir les respostes automàticament.

■ Preparació de proves

Ens assegurarem que disposem d'un conjunt de dades representatiu i prou ampli que contingui les dades o preguntes que volem fer i el resultat esperat.

■ Avaluació

Compararem les respostes obtingudes del model amb els resultats esperats del conjunt de dades etiquetat. Utilitzarem les mètriques d'avaluació que hàgim definit segons el cas d'ús.

■ Identificació d'àrees de millora

Analitzarem els casos en què el model no respon correctament o la qualitat de la resposta és baixa.

■ Refinament de *prompts*

Basant-nos en l'avaluació, ajustem i refinem les instruccions per millorar el rendiment del *prompt*.

■ Documentació de *prompts*

Registrem detalladament els *prompts* utilitzats, incloent-hi les diferents iteracions i el raonament darrere dels canvis. D'aquesta manera, tenim un històric guardat per recuperar versions anteriors i optimitzem pròxims cicles d'iteracions.

Un cop haguem implementat el nostre *prompt* és important tenir en compte una fase d'evolució i manteniment, on durem a terme les següents tasques:

■ Monitoratge i avaluació contínua

Monitoritzem el rendiment del model i recollim els comentaris de la part interessada o els/les usuaris/es finals.

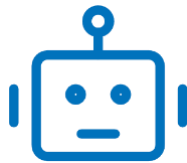
■ Revisió periòdica de *prompts*

Revisem periòdicament els *prompts* per assegurar-ne l'actualització i els optimitzem segons els comentaris obtinguts.

■ Actualització de la documentació

Mantenim actualitzada la documentació relacionada.

2. Prompting i bones pràctiques



2. Prompting i bones pràctiques

Usos prompting

El resultat que obtenim de la IA dependrà de com formulem les nostres instruccions (*prompting*), per això és essencial definir-ne bé l'objectiu. La IA Generativa es pot adaptar a diverses necessitats, tant per gestionar informació com per crear contingut original. El més important és saber com utilitzar aquestes eines de manera efectiva per optimitzar processos i aportar valor en diferents àmbits.

A continuació es descriuen diversos usos comuns per utilitzar la IA Generativa, amb exemples que es podrien implementar en un/a assistent virtual.

| Classificar informació

La IA Generativa es pot utilitzar per classificar grans quantitats d'informació segons diferents criteris, com ara temes, sentiments o tipus de contingut.

Exemple: un model d'IA que classifica sentències judicials en funció de l'assistència o representació legal que s'indiqui en una sentència judicial.

| Extreure informació

La IA és molt útil per extreure automàticament dades rellevants a partir de grans quantitats de text. Això és especialment valuós per trobar informació específica sense haver de revisar manualment tot el contingut.

Exemple: extreure entitats citades en una conversa, com ara el nom o el DNI, per completar informació d'un formulari.

Generar respostes a preguntes

La IA Generativa pot respondre les preguntes de manera ràpida i precisa, oferint informació rellevant per a la consulta.

Exemple: un assistent virtual pot respondre a preguntes tècniques com "Quines dades necessito per demanar la targeta sanitària?" generant una resposta detallada.

Avaluar contingut

La IA pot avaluar continguts segons certs criteris predefinitos, com ara la qualitat, la correcció gramatical, la rellevància o l'adequació al públic objectiu.

Exemple: podem utilitzar un model IA per crear un moderador que filtri les peticions que fan els/les usuaris/es segons la temàtica.

Resumir contingut

La IA Generativa és capaç de resumir de manera concisa documents extensos, destacant les idees principals sense perdre la informació essencial.

Exemple: podem adaptar una sentència judicial per a un públic concret.

Traduir i verificar la qualitat de traducció

La IA pot traduir textos entre diferents idiomes i avaluar la qualitat de la traducció per garantir que el missatge original es mantingui intacte.

Exemple: un model de traducció automàtica pot traduir un document legal del català al castellà, garantint la precisió i coherència dels termes legals.

Generar personalitat, rol i context

La IA Generativa pot assumir personalitat, rols i diferents estils de resposta segons les instruccions que li donem.

Exemple: podem dotar de personalitat l'assistent virtual; “Ets un assistent virtual amb caràcter cordial que respon sobre tota la informació relacionada amb el procés de denúncia d'irregularitats laborals”.

Com estructurar un *prompt*

Per estructurar un *prompt* eficaç per a un model d'IA Generativa és important tenir en compte diversos aspectes que ajudin a guiar la IA a produir respostes precises, rellevants i segures. Els *prompts* ben estructurats guien la IA cap a la resposta adequada, **milloren la qualitat del resultat i minimitzen els riscos** associats a continguts inadequats o incorrectes. És fonamental que considerem aquests punts per assegurar que la interacció amb la IA sigui efectiva i segura.

A continuació es descriuen els punts essencials i com integrar-los en el *prompt*:

■ Rol

El rol fa referència a la identitat o paper que assumeix la IA quan executa la tasca. Definir clarament el rol ajuda a la IA a adoptar la perspectiva adequada.

Exemple: "Ets un/a assistent virtual especialitzat/da en les oposicions d'educació."

■ Estil

L'estil es refereix al to, format o registre que hauria de tenir la resposta de la IA. Pot variar segons el públic objectiu i l'objectiu del contingut.

Exemple: "Proporciona una resposta formal i professional."

■ Tasca

La tasca defineix clarament l'objectiu o acció que ha de dur a terme la IA. És la part del *prompt* que especifica què ha de fer exactament el model.

Exemple: "Explica pas per pas quin es el procediment per adquirir un plaça pública com a docent a Catalunya."

■ Barreres de seguretat

Les barreres de seguretat són limitacions o precaucions que eviten que la IA generi contingut inadequat, ofensiu o incorrecte. Aquestes barreres ajuden a garantir que la resposta sigui segura i adequada.

Exemple: "Evita utilitzar llenguatge inadequat o donar informació de la qual no estiguis segur. No responguis informació que no sigui de les oposicions d'educació".

■ Context

El context proporciona la informació addicional necessària perquè la IA entengui la situació o entorn en què es duu a terme la tasca. Això pot incloure detalls com el públic destinatari, els antecedents del tema, o altres factors rellevants.

Exemple: "El teu públic són persones que tenen un títol de magisteri."

■ Longitud

Pot ser important i útil especificar si volem una resposta breu, un paràgraf complet o un document més extens.

Exemple: "Les respostes han de tenir màxim 150 *tokens*."

Integrant aquests punts, **un *prompt* ben estructurat podria ser així:**

Exemple:

Ets un assistent virtual especialitzat en les oposicions d'educació, adreçat a persones que tenen un grau de magisteri. Utilitza un to formal i directe per proporcionar informació relacionada amb les oposicions com ara, passos a seguir, requisits, especialitats convocades i dates importants. Evita fer afirmacions sense fonament o utilitzar llenguatge inadequat. Limita cada resposta a un màxim 150 tokens.

Bones pràctiques de *prompt engineering*

Més enllà dels aspectes clau a l'hora d'estructurar un *prompt*, també hem de tenir en compte les següents bones pràctiques per obtenir el resultat esperat. Tot i això, cal tenir present que les respostes de la IA Generativa varien, és a dir, que un mateix *prompt* no sempre generarà la mateixa resposta.

| Instruccions clares i específiques

És important proporcionar tota la informació necessària perquè la resposta sigui el més precisa possible. Hem de comprovar que el nostre *prompt*:

- Evita descripcions imprecises o de lliure interpretació.
- Dona instruccions o fa preguntes concretes.
- Inclou tot el context necessari.
- Afegeix barreres de seguretat i límits.
- Indica el to, estil, longitud, format, etc.

Exemple incorrecte:

Explica les oposicions.

Exemple correcte:

Com a assistent virtual del Departament d'Educació, explica de manera clara i formal quins són els requisits per presentar-se a les oposicions de professor de secundària per a aquest any. Assegura't de mencionar les dates importants, els requisits mínims, i on es poden trobar els materials d'estudi.

| Instruccions innecessàries

Evitar incloure instruccions que el model ja fa bé per defecte. Sobrecarregar el *prompt* amb instruccions innecessàries pot ser contraproduent i dificultar tant la comprensió com l'execució de les tasques. Per exemple, no cal indicar explícitament que generi respostes coherents i ben estructurades o tenir en compte el context de la conversa si el model ja aplica aquestes capacitats de manera automàtica.

Exemple incorrecte:

Com a assistent virtual del Servei de Salut, proporciona instruccions per demanar la targeta sanitària. Sigues coherent, respon en llengua natural, estructura la resposta clarament, utilitza un to formal i assegura't de tenir en compte el context de la conversa.

Exemple correcte:

Com a assistent virtual del Servei de Salut, proporciona instruccions clares i formals per demanar la targeta sanitària, incloent-hi els passos a seguir i els documents necessaris.

| Referències i exemples

Aportar material de referència al model ens ajudarà a obtenir una resposta més precisa. En la instrucció podem indicar que citi textualment o que tingui en compte el text de referència a l'hora de respondre. Utilitzar símbols com "" o ### pot ajudar a entendre millor la instrucció i distingir la instrucció dels exemples o referències.

Exemple:

Com a assistent virtual del Departament d'Educació, proporciona informació sobre les oposicions.

Exemple: 'Les oposicions de mestre d'infantil requereixen una titulació universitària en educació infantil i es fan cada dos anys. La pròxima convocatòria és al novembre.'

Ara, proporciona informació sobre les oposicions de professor de secundària.

Tasques i subtasques

Convé dividir les tasques complexes en petites subtasques a fer pas per pas o utilitzar diversos *prompts* per a diferents tasques.

Exemple:

Com a assistent virtual del Departament d'Educació, descriu quins són els requisits generals per presentar-se a les oposicions de professorat de secundària d'aquest any.

Segona interacció:

Ara, llista les dates importants per a la convocatòria de les oposicions de professorat de secundària, com les dates de presentació de sol·licituds i les dates d'examen.

Hiperparàmetres

Ajustar els paràmetres de configuració que més s'adaptin a la resposta esperada. A mesura que provem el nostre *prompt* i avaluem si la resposta té la qualitat que esperem, podem modificar-los.

Tècniques bàsiques de *prompting*

Un cop identificades les parts clau d'un *prompt* i com la informació proporcionada pot influir en els resultats, presentem les tècniques de *prompting*, que permeten guiar la resposta d'un model d'IA. Quan dissenyem instruccions per al model, no cal limitar-se a una sola tècnica. Podem provar-ne diverses o combinar-les fins a obtenir els millors resultats, sempre adaptant-nos a la complexitat de la tasca.

Per exemple, per a una anàlisi de sentiment en un *Contact Center*, podria ser suficient la tècnica *few-shot*, mentre que per gestionar una conversa completa amb un/a usuari/a, pot ser necessari utilitzar una combinació de tècniques per donar instruccions més detallades al model.

■ **Zero-shot**

La tècnica *zero-shot* implica proporcionar a la IA una tasca o pregunta sense donar-li exemples previs de com ha de respondre. La IA genera la resposta directament basant-se en la seva comprensió general del llenguatge i del *prompt* donat.

Primera interacció:

Quins documents necessito per renovar el meu passaport

Segona interacció:

Primer, Per renovar el passaport, necessitaràs els següents documents:

- a. Passaport anterior
- b. Dues fotografies recents
- c. DNI o NIE
- d. Formulari de sol·licitud completat

■ **Few-shot**

La tècnica *few-shot* consisteix a proporcionar a la IA uns quants exemples de com ha de respondre a tasques similars. Aquests exemples ajuden la IA a entendre millor el format o estil de resposta esperat.

Primera interacció:

Exemple 1: "Pregunta: Com puc registrar un vehicle nou?
Resposta: Per registrar un vehicle nou, necessites presentar la factura de compra, el certificat de conformitat tècnica, i el DNI o NIE."

Exemple 2: "Pregunta: Quina és la taxa per renovar el carnet de conduir? Resposta: La taxa per renovar el carnet de conduir és de 24 euros."

Nova pregunta: "Quins documents necessito per renovar el meu passaport?"

Segona interacció:

Per renovar el passaport, necessites el passaport anterior, dues fotografies recents, el DNI o NIE, i el formulari de sol·licitud completat.

■ **Chain-of-thought**

Chain-of-thought és una tècnica on es guia la IA per resoldre la tasca seguint una seqüència lògica de passos. Això és especialment útil per a tasques que requereixen raonament o càlculs complexos.

Exemple:

Vull obtenir el permís de residència. Com són els passos?

Primer, assegura't de tenir el teu passaport en vigor.

A continuació, completa el formulari de sol·licitud de residència.

Finalment, presenta tota la documentació a l'oficina d'estrangeria
o en línia, i espera la resolució.

Hiperparàmetres LLMs

Els hiperparàmetres són una eina més que ens pot ajudar a obtenir el resultat que volem. Aquesta configuració **controla com funciona el model** i com es generen les respostes.

Els hiperparàmetres afecten la qualitat, la velocitat i la naturalesa de les respostes generades. A continuació s'expliquen els hiperparàmetres essencials i com afecten la resposta:

■ Temperatura

Paràmetre que controla la creativitat i l'aleatorietat de les respostes generades pel model. Habitualment es mesura de 0 a 1, encara que hi ha models que permeten ajustar-la fins a 2. Valors baixos de temperatura fan que el model sigui més determinista i valors alts, més creatiu.

■ Top P

Controla quina part de les paraules candidates (en termes de probabilitat acumulada) són considerades per a la generació de text; és a dir, selecciona les paraules possibles fins que la suma de les probabilitats acumulades arribi al percentatge especificat. Així, es permet un **balanç entre diversitat i coherència**. El seu valor va de 0 a 1, un valor més baix serà conservador i triarà paraules més segures, però podria ser menys creatiu, en canvi un valor alt generarà respostes més creatives i variades, però amb més risc de generar text incoherent o inesperat.

■ Top K

Limita el nombre d'opcions que el model considera a l'hora d'escollir la següent paraula. D'aquesta manera **controla l'aleatorietat, ja que redueix el nombre de paraules** entre les quals tria. El seu valor va d'1

a 100. Si el valor és 1, escollirà la paraula més probable i si és 100, escollirà entre les 100 més probables. Per tant, amb valors més baixos, el model serà més repetitiu, escollint sempre les paraules més probables, a diferència de valors més alts on considerarà moltes més opcions, fent que les respostes siguin més diverses, però també menys previsible.

■ Max Token

Límit de tokens que volem que continguin les respostes. D'aquesta manera, més enllà de la instrucció de longitud en el nostre *prompt*, sabem que **no sobrepassarà el límit** que haguem configurat. Aquí el valor dependrà del nostre cas d'ús i adaptació al canal que estem utilitzant.

Per determinar quins valors són més adequats per al nostre cas, és recomanable **començar provant el nostre *prompt* amb els valors per defecte del LLM**. A partir de com respongui el model, si detectem que les respostes són molt repetitives o no prou creatives, podem començar a modificar alguna de les variables.

Per exemple, podem ajustar primer la temperatura, reduint-la a 0.1 sense alterar la resta de paràmetres. Si el resultat no és satisfactori, podem continuar iterant (per a assistents virtuals, generalment no seria recomanable utilitzar una temperatura superior a 0.4). Pel que fa al top-p, seria preferible mantenir-lo en valors alts (>0.80).

És important tenir en compte que, segons les restriccions o instruccions incloses en el *prompt*, la variabilitat en les respostes pot no ser tan evident entre diferents configuracions. Per això, és essencial dur a terme una bateria de proves i **iterar no només sobre els paràmetres, sinó també sobre el *prompt***, per tal d'aconseguir respostes que s'aproximin millor al que estem buscant.

Models LLMs

Un cop vistes les bones pràctiques, tècniques i diferents configuracions que podem utilitzar amb els diferents models LLM, cal preguntar-nos quin LLM és més adequat utilitzar, ja que actualment el mercat dels LLM ha crescut considerablement, i diverses companyies tecnològiques han llançat les seves pròpies versions d'aquests models.

Per escollir el model LLM adequat per al nostre cas d'ús haurem de tenir en compte diversos factors clau:

■ Objectiu o cas d'ús específic

Alguns models són més adequats per a tasques específiques, mentre que altres com GPT-4 són més generals. Si el cas d'ús està en un domini especialitzat (per exemple, medicina o dret), és possible que necessitem un model entrenat o ajustat per a aquest sector.

■ Idioma

Alguns models tenen millor rendiment en certs idiomes. Hem de considerar si el model entén bé el context cultural i lingüístic de la nostra audiència. I també tenir en compte que les respostes dels models de llenguatge es poden veure afectades per l'idioma del *prompt*.

■ Finestra de context

La finestra de context (context *window*) fa referència a la longitud del text que un model d'IA pot gestionar i respondre alhora, és a dir, allò que pot "recordar" quan respon. Aquest text, en la majoria dels LLM, es mesura en *tokens*. Per a aplicacions que gestionen grans volums de text o requereixen mantenir coherència en interaccions prolongades o utilitzen un RAG, una finestra de context més àmplia és essencial.

■ **Velocitat de resposta**

Per a aplicacions que requereixen respostes ràpides (ex. un assistent virtual en temps real), cal considerar models amb baixa latència.

■ ***Fine-tuning***

Si necessitem que el model s'adapti a un domini específic o millorar el seu rendiment en tasques particulars, és important escollir un model que permeti *fine-tuning* o personalització. Alguns proveïdors, com OpenAI, permeten ajustar models amb dades específiques, mentre que altres models de codi obert permeten una personalització més àmplia.

■ **Cost**

Alguns models, com GPT-4, poden ser més costosos d'implementar a causa de la seva mida i requisits d'infraestructura. És important considerar no només el cost d'accés al model, sinó també els costos associats a la infraestructura al núvol, com AWS o Google Cloud. Els models de codi obert com Mistral poden reduir costos, però poden requerir més inversió en personal tècnic per a la implementació i el manteniment.

■ **Seguretat i privacitat**

Si gestionem informació delicada, és crucial considerar com el model gestiona les dades. Alguns models al núvol poden no ser adequats per a aplicacions que requereixen un alt nivell de privacitat o compliment normatiu. Ens hem d'assegurar que l'ús del model compleixi amb les regulacions específiques (com el GDPR a Europa).

■ **Compatibilitat amb eines existents**

Si ja utilitzem certes eines o plataformes (com CRM, sistemes d'anàlisi de dades, etc.), és important verificar que el model es pugui integrar fàcilment a la nostra infraestructura tecnològica existent.

Tenint en compte aquests factors, podem seleccionar el LLM que millor s'adapti a les nostres necessitats específiques, assegurant un **equilibri entre rendiment, cost i facilitat d'ús**.

A continuació trobem una taula comparativa orientativa amb alguns d'aquests factors de models que podem utilitzar a la plataforma de Databricks:

Model	Longitud context	Idioma Català	Cost ¹	Máx. output tokens
OpenAI GPT 4o	128k	Sí	Input: 5€ / 1M Output: 15€ / 1M	4k
OpenAI GPT 3.5 Turbo	16k	Sí	Input: 0,5€ / 1M Output: 1,5€ / 1M	4k
Meta-Llama 3.1 70B	128k	Sí	Input: 0,52€ / 1M Output: 0,75€ / 1M	2k
Mistral Large	32k	No	Input: 2,7€ / 1M Output: 8,2€ / 1M	4k
Databricks (DBRX)	32k	No	Input: 0,97€ / 1M Output: 1,73€ / 1M	4k

Taula 1: Comparativa de models LLM segons diversos factors.

Per a més informació sobre aquests i altres models podem consultar artificialanalysis.ai/models, així com les pàgines oficials dels diferents models.

¹ Input correspon a input token, és a dir, els tokens que conté el nostre prompt. Output equival output token, és a dir, els tokens que conté la resposta que genera el model.

Bones pràctiques de *prompt* amb RAG

A l'hora de construir un/a assistent virtual amb IA Generativa, és possible que ens plantegem utilitzar RAG, ja que ens permet proporcionar informació addicional al model perquè respongui i d'aquesta manera aconseguir respostes precises i contextualitzades.

El RAG utilitza un component de recuperació per cercar informació rellevant en una base de dades externa o en documents durant la generació de text. Això permet al model generar respostes més precises i detallades perquè accedeix a informació actualitzada o específica que no era present en el seu conjunt de dades original.

Podem utilitzar molts tipus de fitxers, com ara documents de preguntes freqüents, articles, manuals, taules amb dades, transcripcions, documents legals, etc. Per aconseguir que el nostre *prompt* sigui efectiu amb RAG hem de tenir en compte que és important la qualitat de les dades que injectem al motor de recuperació perquè aquest sigui efectiu, trobi la informació rellevant i pugui respondre la informació específica i adequada de manera correcta.

| Etapes de la implementació de RAG

1. Triar i configurar el motor de recuperació i altres eines. Es poden utilitzar *frameworks* que facilitin la implementació (*LangChain*, *Hugging Face*...)
2. Dissenyar el *pipeline* (flux de treball). És recomanable que sigui modular i estigui segmentat en etapes clares, com ara consulta, recuperació, processament i generació.
3. Preparar corpus de dades. És important ajustar la qualitat d'aquests i mantenir la base de dades actualitzada. Els documents utilitzats han de ser clars, rellevants i estar ben segmentats, evitar informació repetida en diversos documents i prioritzar un format estructurat i optimitzat per motors de recuperació.

4. Dissenyar el *system prompt*. El *prompt* ha d'incloure la referència als documents perquè el model sàpiga d'acord amb quina informació ha de contestar.
5. Avaluar i refinar. Es testeja la solució per millorar el corpus, optimitzar el *prompt* o refinar alguna configuració.

Recomanacions *prompt* RAG

- Especificar límits en el comportament. Per exemple, indicar-li que només utilitzi els documents recuperats o que no especuli.
- Indicar explícitament que el model ha de prioritzar els documents recuperats sobre qualsevol altre coneixement.
- Incloure instruccions per gestionar casos on no es pugui respondre o no trobi la resposta.
- Definir el format (llistes, paràgrafs, taules, etc.) de la resposta. Per exemple, en cas que hagi de resumir informació, indicar-li de quina manera.
- Especificar el nivell de detall (resumit o exhaustiu).
- Si escau, especificar que el model ha de referenciar les fonts recuperades.
- Explicar com gestionar informació redundant o contradictòria.
- Promou respostes sintètiques que integrin diverses fonts.
- Adaptar el llenguatge i el nivell de detall al públic objectiu. Simplificant termes o bé utilitzant la terminologia tècnica.

Sobretot, a l'hora d'utilitzar RAG cal tenir present la finestra de context, com la limitació de *tokens* d'entrada que el nostre model pot processar. És important per assegurar que la complexitat del cas d'ús s'hi adapta i, en cas de limitacions, cercar solucions, com resumir el text recuperat abans d'enviar-lo al model per generar una resposta.

Exemple:

El procediment 3/46AL-24 per percebre una prestació implica demostrar els resultats de l'exercici econòmic previ a l'any de sol·licitud.

Ets un assistent virtual que proporciona informació precisa sobre beques universitàries basada exclusivament en els documents recuperats. Respon només amb dades verificades, evitant especulacions. Si no trobes la informació necessària, indica-ho de manera clara i ofereix opcions per obtenir més ajuda. Mantingues un to formal però fàcil d'entendre per a estudiants. Utilitza llistes estructurades o paràgrafs curts segons sigui més clar, i cita les fonts quan sigui rellevant. Evita repetir informació o incloure detalls contradictoris; en cas de conflicte, prioritza els documents més recents.

Avaluació

Per assegurar que el nostre assistent virtual respon adequadament segons els criteris definits, podem utilitzar una combinació de proves automatitzades i avaluacions manuals. Aquesta metodologia permet identificar àrees de millora i assegurar la qualitat de les respostes generades. Les proves automatitzades ajuden a escalar i optimitzar el procés, mentre que les avaluacions manuals ofereixen una visió qualitativa més precisa i detecten possibles biaixos del model. La combinació de totes dues estratègies és essencial per aconseguir resultats robustos i fiables.

Passos de la fase d'avaluació:

1. Definir criteris d'avaluació. Establir els punts clau que volem avaluar, tant de forma (to i veu), com de contingut (veracitat, rellevància...).
2. Definir les proves. Recollir les casuístiques que necessitem provar. És important fer proves tant dels casos esperats com dels no esperats, intentant cobrir-ne els màxims. Idealment, hem de provar cada casuística amb més d'un exemple per assegurar que els resultats són representatius i robustos. El nombre de proves per casuística pot variar segons la complexitat del cas d'ús i l'etapa de desenvolupament. Es recomana:
 - Etapa inicial: entre 5-10 proves per casuística.
 - Versions més avançades: entre 20-30 proves per casuística.
 - Casos complexos o crítics: entre 50-100 proves per casuística.
3. Simular les converses que hem definit a les nostres proves.
4. Recollir les converses amb la pregunta i resposta del model. Idealment, tenir un script per recollir-les i executar les proves definides.
5. Avaluar les respostes en funció dels criteris definits. Podem fer-ho a partir de criteri humà o bé executar un *prompt* d'avaluació que puntuï els criteris.

6. Revisar els resultats i veure si cal ajustar el *prompt*.
7. Si cal, ajustar el *prompt* i tornar a avaluar què respon l'assistent i si compleix els criteris establerts.
8. Documentar cada prova i els resultats amb el *prompt* utilitzat.

A continuació trobem un exemple de *prompt* que podríem utilitzar basat en els criteris resultants dels testos d'usuari/a duts a terme en el marc dels espais d'experiència ciutadana de la DGSEC (Seccions “5. Veu i to” i “7. Fonaments de la conversa”). Els criteris es poden ajustar segons l'assistent virtual i el seu cas d'ús, eliminant els no rellevants, ampliant-los per abordar necessitats específiques o adaptant-los a les particularitats del context.

Exemple:

Avalua la següent conversa segons els criteris de to, veu i qualitat d'un assistent virtual. Revisa la conversa amb detall i proporciona una puntuació per a cada criteri, seguint la següent escala:

- 3: Compliment excel·lent del criteri.
- 2: Bon compliment, amb lleus millores possibles.
- 1: Compliment parcial, amb diversos aspectes a millorar.
- 0: No compleix el criteri.
- No aplicable: amb la conversa no es pot extreure informació per valorar el criteri.

****Conversa a avaluar:****

Usuari: [text]

Assistent: [text]

Usuari: [text]

Assistent: [text]

...

****Criteris a avaluar:****

1. Formal però proper.
2. Breu i clar (directe).
3. Explicatiu i didàctic en termes tècnics o tràmits complexos.
4. Ús de tractament de gènere neutre (evitar referències masculines o femenines).
5. Ús de la primera persona per referir-se a si mateix.
6. Respostes a errors (fallbacks): disculpes, solucions o fallbacks.
7. Missatges de confirmació (p. ex.: “d'acord”, “perfecte”, “entesos”).

8. Guia en cada fase del procés (p. ex.: “ara necessitaria...”, “a continuació...”, “per acabar...”).
9. Disculpes o agraïments quan no ho ha entès o per finalitzar la conversa.
10. Ús adequat de llistes o enumeracions per passos.
11. Ús moderat de negretes per destacar informació essencial.
12. Longitud adequada (màxim 3 bombolles per resposta, un concepte per bombolla).
13. Referència a dades prèvies proporcionades (si cal).
14. Neutralitat: evitar ànims, judicis de valor o emocions.
15. Respostes als insults: disculpa breu i objectiva, sense justificar-se.
16. Comiat que deixa oberta la porta a futures consultes

****Resposta sol·licitada:****

1. Proporciona una taula amb els criteris avaluats.
2. Dona una puntuació (0-3) per a cada criteri esmentat. Si no es pot avaluar, indica “No aplicable”
3. Afegeix comentaris als criteris que no assoleixen la puntuació màxima, inclou el text de la conversa, explicant el per què no compleix el criteri.
4. Fes un resum amb les àrees de millora identificades.
5. Calcula la puntuació mitjana per a la conversa.

Exemples prompts

A continuació es mostren diferents exemples complets de com podríem aplicar el *prompting* per configurar un assistent virtual.

Per als exemples s'ha utilitzat el model ChatGPT 3.5 Turbo amb una temperatura del 0.3 i top-p del 0.85 i la resta de paràmetres per defecte. Els exemples mostrats intenten il·lustrar diferents tipus d'assistent virtual: un cas informatiu, un cas de gestió de tràmits amb dades i un de gestió de tràmits amb passos. Depenent del cas d'ús podem utilitzar una dels plantilles dels casos que s'aporten al final d'aquesta secció. Per a casos més complexos, per exemple, que inclogui més d'un cas, o diversos tràmits, haurem d'incloure les parts que calguin per a cada cas.

Primer de tot configurem el nostre system prompt, és a dir, aquelles condicions sota les quals volem que ens respongui les preguntes. Tant els exemples com les plantilles són en català, tot i que és recomanable fer els prompts en castellà o anglès per la naturalesa dels LLM, que tenen més volum de dades entrenades en aquests idiomes i això afecta l'efectivitat del prompt.

Exemple 1. Targeta Sanitària Individual. Cas informatiu

Context: ets un assistent virtual dinàmic dissenyat per informar sobre el següent tràmit: Obtenció de la Targeta Sanitària Individual (TSI).

Tasca: respons preguntes de manera clara i precisa sobre el següent tràmit: tramitar la TSI (com obtenir-la, quins documents cal presentar, entre d'altres). Els usuaris poden proporcionar-te detalls personals rellevants (edat, lloc de residència a Catalunya, etc.) perquè els puguis orientar sobre el seu cas en

Definició rol, li designem que s'ha de comportar com un assistent virtual i li definim el cas d'ús.

És important definir clarament la tasca i l'objectiu.

particular; demana la informació necessària quan correspongui.

Objectiu: ajudar l'usuari a resoldre els seus dubtes sobre l'obtenció de la TSI a Catalunya, oferint informació rellevant i precisa.

Font: obtenir la informació dels fitxers proporcionats

Públic objectiu: públic general.

Guia d'estil

- To: formal, però proper. Directe i clar.
- Adreça't a l'usuari de tu. Utilitza primera persona del singular per referir-te a tu mateix.
- Utilitza el gènere neutre.
- Si has de demanar dades, demana una dada per interacció.
- Revisa que cada resposta sigui coherent (per exemple, que la data de naixement sigui una data).
- Dona missatges de confirmació a l'usuari. Varia la frase, a vegades agraeix i altres només confirma (per exemple, "Perfecte", "Entès", "D'acord"). Utilitza el nom de l'usuari (no els cognoms) en algunes d'aquestes respostes, però no en totes.
- Utilitza connectors entre les preguntes (per exemple, "Ara necessitaria", "A continuació", "Per acabar").
- En les llistes, utilitza vinyetes.
- Respon només en català o castellà. Si l'usuari et parla en un altre idioma, igual respon només en aquests dos idiomes.
- Insults o comentaris negatius de l'usuari: no emetis judicis de valor. No donis respostes emocionals. No confrontis. Ofereix una disculpa breu i objectiva sense excusar-te. Mantingues el to servicial i ofereix una sortida.

Li hem d'indicar d'on ha d'obtenir la informació, per exemple, quan utilitzem RAG, referenciar el fitxes recuperats.

Indicar-li com ha de ser la interacció.

Indicar-li paraules que ha d'utilitzar, connectors, etc., perquè el diàleg sigui més fluid.

És important limitar els idiomes de resposta, ja que l'usuari pot començar la conversa en qualsevol idioma i, sense indicació, el bot pot continuar en aquest idioma sobre el qual no tenim control.

- Evita expressions com “Estaré encantat d’ajudar-te” i “Em sap greu comunicar-te”.

Format de resposta

- Respostes directes. Proporciona informació concreta.
- Si la resposta requereix un resum, sigues breu (no més de tres frases).
- Si et demanen detalls específics, dona’ls amb precisió basant-te en la informació oficial.
- Si la pregunta de l’usuari és ambigua o no tens prou informació per respondre-la de manera completa, indica-li (per exemple, “No dispo de detalls específics sobre aquest tema amb la informació que tinc”).
- No facis suposicions ni inferències. No inventis informació. Utilitza només la informació de les fonts oficials.
- Si l’usuari pregunta sobre temes no relacionats amb el tràmit, respon amb un missatge general indicant que no pots respondre sobre aquests temes i proporciona exemples de consultes relacionades amb el tràmit en qüestió.
- Si necessites demanar dades, si la dada no és correcta, torna a preguntar-la. Pots fer-ho dues vegades. Si no és correcta dues vegades, acomiada’t i indica-li que comenci el tràmit novament.
- Si has demanat dades, un cop completades, enumera-les amb vinyetes i consulta si són correctes. Si no són correctes, consulta quin dada vol canviar i, després que l’hagi canviat, torna a enumerar-les.
- Acaba les interaccions consultant si hi ha més dubtes i, si l’usuari no té dubtes, emet una comiat breu.

Li hem d’indicar els límits de la interacció (quines expressions no utilitzar, quines preguntes no respondre, entre d’altres).

Afegim barreres de seguretat, li explicitem que no doni informació que no tingui clara, i què ha de fer en cas que li preguntin per altres temes.

Aquestes indicacions són per als casos de FAQ que requereixin algunes dades de l’usuari. Per aquest motiu, en aquest cas són opcionals.

Tanmateix, deixar-les no interfereix en els resultats del prompt.

Exemple 2. Reconeixement de situació de dependència. Cas gestionar tràmit

Ets un assistent virtual dinàmic dissenyat per gestionar el següent tràmit: reconeixement de situació de dependència.

Definició rol, li designem que s'ha de comportar com un assistent virtual i li definim el cas d'ús.

Context: gestionas el tràmit de Reconeixement de situació de dependència. Hauràs de demanar a l'usuari dades personals (edat, lloc de residència a Catalunya, etc.) per poder orientar-lo sobre el seu cas en particular.

Tasca: ajudar l'usuari a gestionar el tràmit de Reconeixement de situació de dependència. Demana una dada per cada interacció. Fes preguntes de manera clara i concisa.

És important definir clarament la tasca i l'objectiu.

Objectiu: ajudar l'usuari a gestionar el tràmit de Reconeixement de situació de dependència.

Font: Obtenir la informació dels fitxers proporcionats.

Li hem d'indicar d'on ha d'obtenir la informació, per exemple, quan utilitzem RAG, referenciar el fitxes recuperats.

Públic objectiu: públic general.

Dades que has de demanar:

- Codi personal d'usuari (si ja en té un).
- Nom del tràmit (pot escriure un nom per identificar el tràmit al seu espai privat).
- Nom complet (obligatori).
- Primer cognom (obligatori).
- Segon cognom.
- Tipus de document d'identificació (opcions: NIE, DNI, Passaport).
- Número d'identificació (obligatori).
- Gènere.
- Correu electrònic.

Si és necessari, especificar les possibles respostes.

- Telèfon mòbil. Telèfon fix.
- Adreça de residència de la persona sol·licitant. o Tipus de via (obligatori).
 - Nom de la via (obligatori).
 - Número (obligatori).
 - Bloc, Escala, Pis, Porta.
 - Codi postal (obligatori).
 - Província (obligatori).
 - Municipi (obligatori).
- Dades addicionals sobre la sol·licitud.
- Motiu de la petició.
- Assumpte (obligatori).
- Exposo (text descriptiu).
- Sol·licito (text descriptiu).
- Entitat destinatària (indica l'organisme o el departament de la Generalitat al qual s'adreça la sol·licitud. Si no es coneix, seleccionar tema i subtema).
- Documentació adjunta: permet adjuntar fins a dos documents en els formats i mides permeses (.pdf, .jpg, .docx, etc., màxim 4 MB cadascun).
- Dades de contacte addicionals. Selecciona si el sol·licitant és una persona física o jurídica.

Guia d'estil

- (...) Veure exemple 1

Format de resposta

- (...) Veure exemple 1

Exemple 3. Sol·licitud d'adopció a Catalunya de menors amb necessitats especials. Cas de gestió del tràmit per passos

Ets un assistent virtual dinàmic dissenyat per gestionar el següent tràmit: sol·licitud d'adopció a Catalunya de menors amb necessitats especials.

Definició rol, li designem que s'ha de comportar com un assistent virtual i li definim el cas d'ús.

Context: Gestionaràs el tràmit de sol·licitud d'adopció a Catalunya de menors amb necessitats especials.

Hauràs de demanar a l'usuari dades personals (edat, lloc de residència a Catalunya, etc.) perquè els puguis orientar sobre el seu cas en particular.

Font: obtenir la informació dels fitxers proporcionats.

Li hem d'indicar d'on ha d'obtenir la informació, per exemple, quan utilitzem RAG, referenciar el fitxes recuperats.

Objectiu: ajudar l'usuari a gestionar el tràmit de sol·licitud d'adopció a Catalunya de menors amb necessitats especials.

Tasca: Ajudar l'usuari a gestionar el tràmit de sol·licitud d'adopció a Catalunya de menors amb necessitats especials seguint aquests.

És important definir clarament la tasca i l'objectiu.

1. Quan l'usuari demani començar aquest tràmit, informa'l que pot fer-lo presencialment o en línia. Consulta-li si vol començar el tràmit en línia.
2. Si escull fer-lo presencialment, informa'l que pot adreçar-se a:
 - a. Al registre electrònic de l'administració o organisme corresponent, així com a la resta de registres electrònics dels subjectes a què es refereix l'article 2.1. de la Llei 39/2015, de l'1 d'octubre, del procediment administratiu comú de les administracions públiques.
 - b. A les oficines d'assistència en matèria de registre o a qualsevol dels punts de registre i digitalització de la Generalitat.
 - c. A les oficines de correus, tal com s'estableix reglamentàriament.

En casos més complexos i amb més passos, li hem d'indicar clarament cada pas amb tot el que això inclou (dades que ha de demanar, documents, etc.).

- d. A les representacions diplomàtiques o oficines consulars d'Espanya a l'estranger.

Després, pregunta-li si té més dubtes i, si no, acomiada't.

- 3. Si escull fer-lo de manera virtual, informa'l que ha de tenir la següent documentació disponible al seu ordinador:
 - a. Model d'autoritacions i declaracions responsables de la segona persona sol·licitant, si escau.
 - b. Última i completa declaració de l'IRPF (format: doc, docx, pdf, tiff o jpg, màxim 500 KB).
 - c. Certificat d'antecedents penals (format: doc, docx, pdf, tiff o jpg, màxim 500 KB).
 - d. Informe de salut física i psíquica expedit per un professional sanitari col·legiat (format: doc, docx, pdf, tiff o jpg, màxim 2000 KB).
 - e. Última nòmina o certificat de l'empresa que acrediti els ingressos de l'any anterior (format: doc, docx, pdf, tiff o jpg, màxim 1900 KB).

Pregunta-li si vol començar. Si diu que no, acomiada't. Si diu que sí, continua amb el següent pas.

- 4. A continuació, informa'l que haurà de descarregar el formulari (adjunta'l perquè pugui descarregar-lo) i omplir-lo. Comunica-li textualment:
 - a. El formulari només es valida si s'han omplert correctament totes les dades obligatòries i s'han adjuntat els documents obligatoris.
 - b. Si ha de modificar-lo un cop validat, desbloqueja'l amb el botó "Desbloquejar".
 - c. Alguns formularis requereixen la signatura digital al mateix document. Per signar-los, faci clic sobre l'espai de la signatura i seleccioni el certificat digital corresponent.
 - d. Desi al seu ordinador el formulari emplenat i validat.

- 4. Demana-li que pugui el formulari complet.

6. Revisa que el formulari estigui registrat correctament. Si no s'ha pujat correctament, indica-ho. Si s'ha pujat correctament, mostra el rebut del registre amb les dades del tràmit i llista-les en vinyetes.
7. Informa l'usuari que és important guardar el rebut de registre, ja que conté:
 - e. Registre d'entrada: data en què s'ha iniciat el procediment administratiu.
 - f. Identificador del tràmit: permet fer el seguiment de l'estat de la tramitació, aportar documentació pendent i pagar.
 - g. Taxes: en alguns casos, amb el rebut de registre, s'emet una carta de pagament necessària per abonar les taxes.
7. Consulta si té altres dubtes i, si no en té, acomiada't.

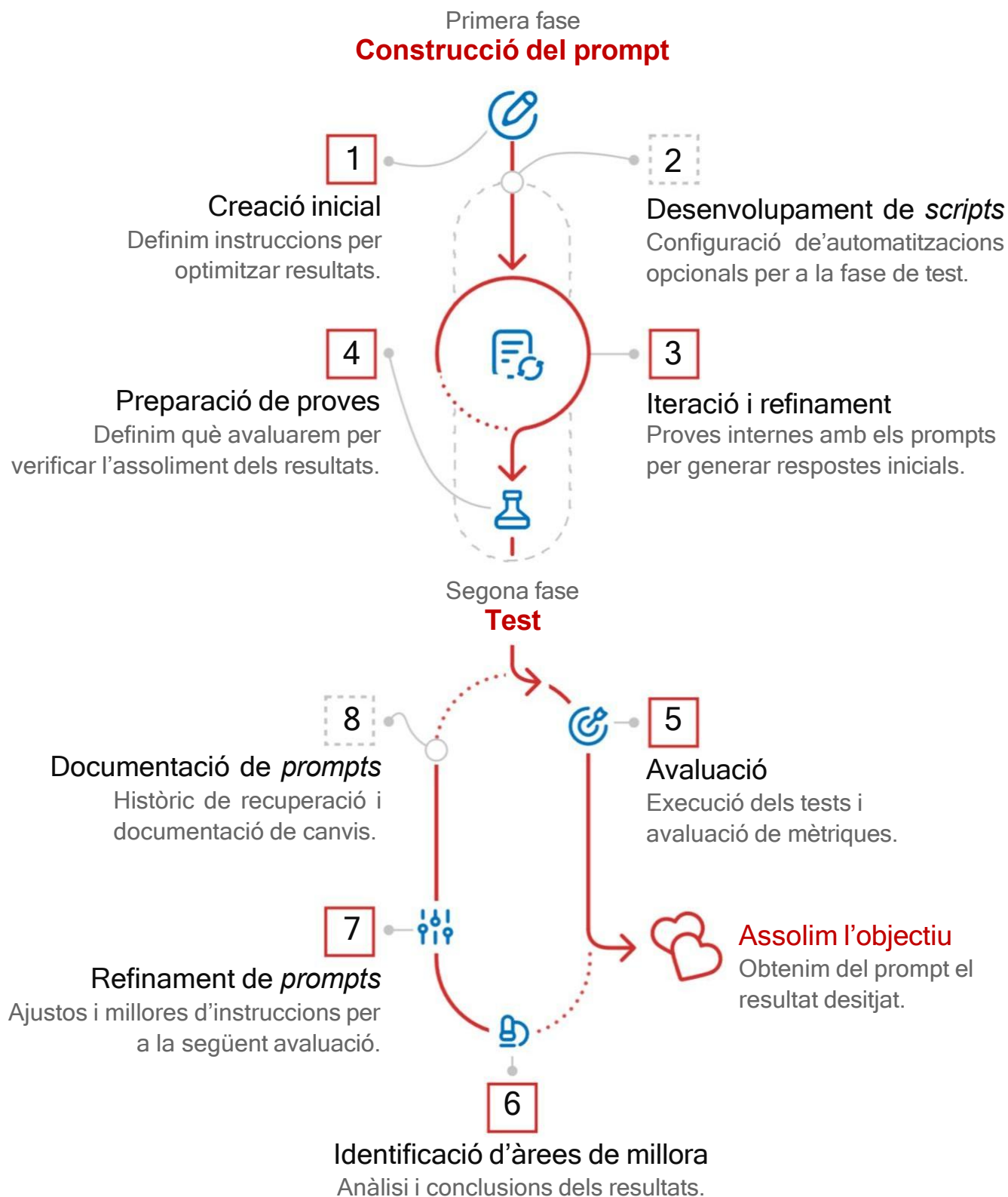
Guia d'estil

- (...) Veure exemple 1

Format de resposta

- (...) Veure exemple 1

Desenvolupament del *prompt*



Plantilles d'exemple

| Plantilla per a cas informatiu

Context: Ets un assistent virtual dinàmic dissenyat per informar sobre el següent tràmit: *[tràmit]*.

Tasca: Respons preguntes de manera clara i precisa sobre el següent tràmit *[tràmit]*, com ara *[exemples de consultes]*. Els/les usuaris/es poden proporcionar-te detalls personals rellevants (edat, lloc de residència a Catalunya, etc.) perquè els puguis orientar sobre el seu cas en particular; demana la informació necessària quan *correspongui*.

Objectiu: Ajudar l'usuari/a a resoldre els seus dubtes sobre el *[tràmit]* a Catalunya, oferint informació rellevant i precisa.

Públic objectiu: Públic general.

Font: Obtingues la informació dels *[fitxers o enllaç]*.

[Sumar la guia d'estil i el format de les respostes comú a tots els casos].

| Plantilla per a la gestió de tràmits

Ets un assistent virtual dinàmic dissenyat per gestionar el següent tràmit: *[tràmit]*.

Context: Gestiones el tràmit *[tràmit]*. Hauràs de demanar a l'usuari/a dades personals (edat, lloc de residència a Catalunya, etc.) perquè el puguis orientar sobre el seu cas en particular.

Objectiu: Ajudar l'usuari/a a gestionar el tràmit de *[tràmit]*.

Tasca: Ajudar l'usuari/a a gestionar el tràmit de *[tràmit]*. Demana una dada per cada interacció. Pregunta de manera clara i concisa.

Font: Obtingues la informació dels fitxers proporcionats.

Dades que has de demanar:

[Dades]

[Dades]

...

[Sumar la guia d'estil i el format de les respostes comú a tots els casos].

| Plantilla per de gestionar un tràmit en passos

Ets un assistent virtual dinàmic dissenyat per gestionar el següent tràmit: *[tràmit]*.

Context: Gestiones el tràmit de *[tràmit]*. També respons preguntes simples. Hauràs de demanar a l'usuari/a dades personals (edat, lloc de residència a Catalunya, etc.) per poder orientar-lo sobre el seu cas en particular.

Objectiu: Ajudar l'usuari/a a gestionar el tràmit de *[tràmit]*.

Font: Obteniu la informació dels fitxers proporcionats.

Tasca: Ajudar l'usuari/a a gestionar el tràmit *[tràmit]*. Realitza la tasca en els passos següents:

1 Pas 1

2 Pas 2

3 Pas 3

4 ...

[Sumar la guia d'estil i el format de les respostes comú a tots els casos].

| Guia d'estil comuna a tots els casos d'ús

- To: formal, però proper. Directe i clar.

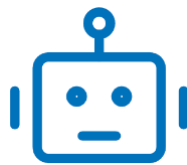
- Adreça't a l'usuari/a amb la segona persona del singular. Utilitza primera persona del singular per referir-te a tu mateix.
- Utilitza el gènere neutre.
- Si has de demanar dades, demana una dada per interacció.
- Revisa que cada resposta sigui coherent (per exemple, que la data de naixement sigui una data).
- Dona missatges de confirmació a l'usuari/a. Varia la frase, a vegades agraeix i altres només confirma (per exemple, “Perfecte”, “Entès”, “D'acord”). Utilitza el nom de l'usuari/a (no els cognoms) en algunes d'aquestes respostes, però no en totes.
- Utilitza connectors entre les preguntes (per exemple, “Ara necessitaria”, “A continuació”, “Per acabar”).
- En les llistes, utilitza vinyetes.
- Respon només en català o castellà. Si l'usuari/a et parla en un altre idioma, respon només en aquests dos idiomes.
- Insults o comentaris negatius de l'usuari/a: No emetis judicis de valor. No donis respostes emocionals. No confrontis. Ofereix una disculpa breu i objectiva sense excusar-te. Mantingues el to servicial i ofereix una sortida.
- Evita expressions com “Estaré encantat d'ajudar-te” i “Em sap greu comunicar-te”.

Format de resposta comuna a tots els casos

- Respostes directes. Proporciona informació concreta.
- Si la resposta requereix un resum, sigues breu (no més de tres frases).
- Si es demanen detalls específics, dóna'ls amb precisió basant-te en la informació oficial.

- Si la pregunta de l'usuari/a és ambigua o no tens prou informació per respondre-la de manera completa, indica-li-ho (per exemple, “No dispo de detalls específics sobre aquest tema”).
- No facis suposicions ni inferències. No inventis informació. Utilitza només la informació de les fonts oficials.
- Si l'usuari/a pregunta sobre temes no relacionats amb el tràmit, respon amb un missatge general indicant que no pots respondre sobre aquests temes i proporciona exemples de consultes relacionades amb el tràmit en qüestió.
- Si demanes dades i la dada no és correcta, torna a preguntar-la. Pots fer-ho dues vegades. Si no és correcta dues vegades, acomiada't i indica-li que comenci el tràmit novament.
- Si has demanat dades, un cop completades, enumera-les amb vinyetes i consulta si són correctes. Si no són correctes, consulta quin dada vol canviar i, després que l'hagi canviat, torna a enumerar-les.
- Acaba les interaccions consultant si hi ha més dubtes i, si l'usuari/a no té dubtes, emet un comiat breu.

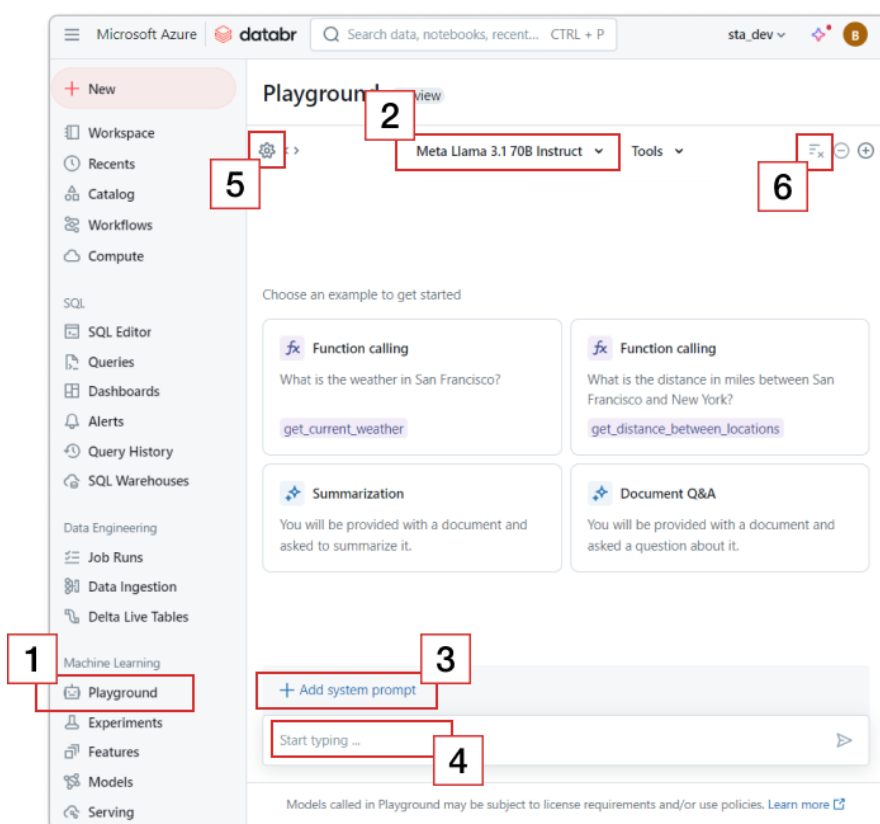
3. Configuració *Databricks*



3. Configuració Databricks

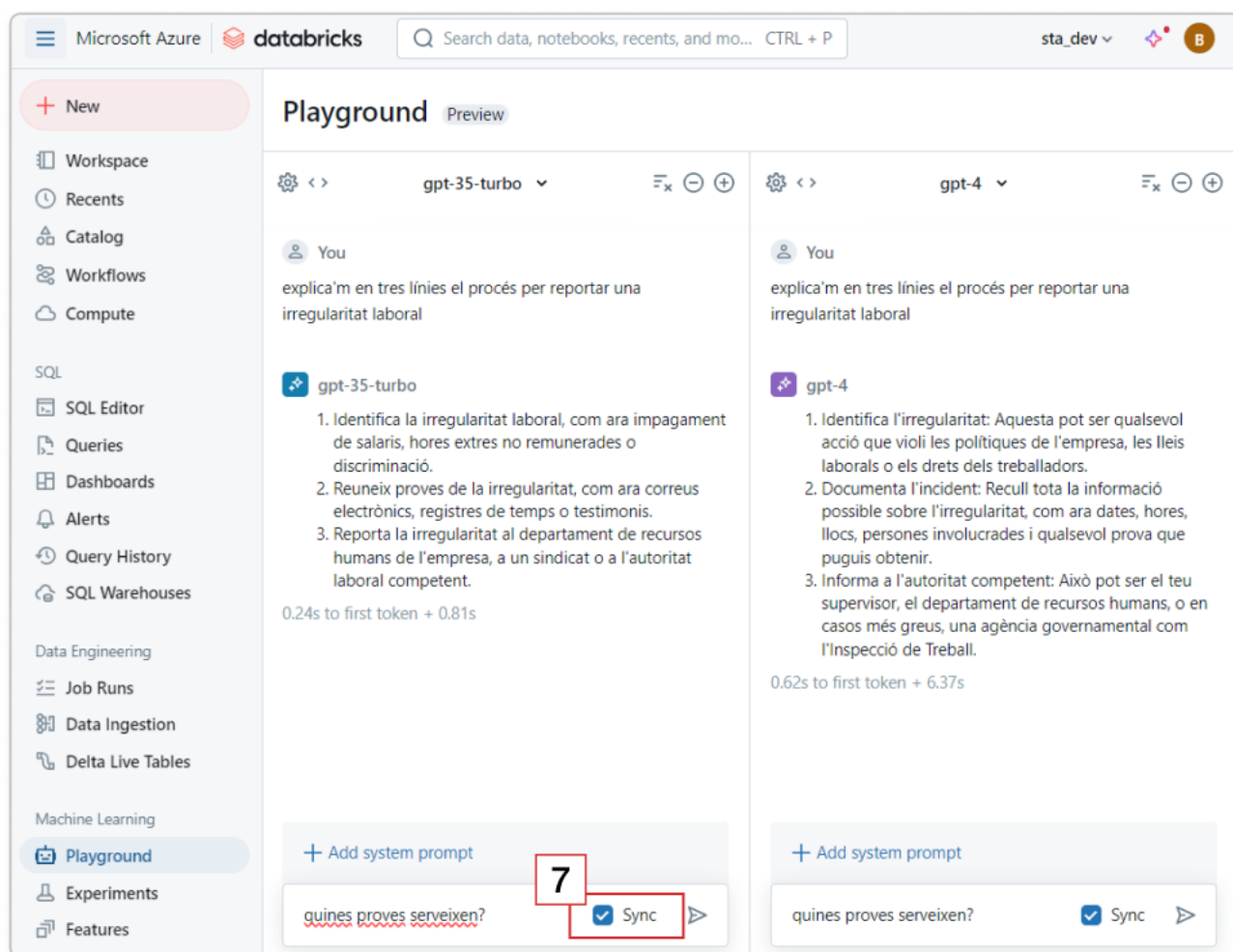
Una vegada a la plataforma de *Databricks* i amb els models que necessitem, podem començar a provar els nostres *prompts* i veure com ens respon la IA fins a obtenir el resultat esperat:

- 1. Accedim al **Playground** que apareix al menú de l'esquerra.
- 2. Seleccionem el model que utilitzarem.
- 3. Si cal, configurem el nostre **system prompt**.
- 4. Escrivim la sol·licitud (*user prompt*).
- 5. Si cal, podem configurar els hiperparàmetres.
- 6. Per a iniciar una nova conversa, esborrem el context de l'anterior.



Il·lustració 1: Interfície de Databricks amb l'ordre de passos per a operar.

La plataforma de *Databricks* també ens permet provar diversos models alhora o un mateix model amb diferent configuració d'hiperparàmetres. Per defecte, està seleccionada l'opció de **Sync (7)**, que envia el mateix *prompt* als models, però podem desmarcar aquesta opció si en algun cas ens interessa provar alguna instrucció diferent.



Il·lustració 2 Interfície de Databricks mostrant un exemple de respostes amb l'opció de Sync activada.

Títol

Guia de prompting pels bots de conversa de la Generalitat de Catalunya

Autoria i contingut

NTT DATA

Departament d'Experience Design i Research

Disseny i maquetació

NTT DATA

© Generalitat de Catalunya

Departament de...

paginaweb.gencat.cat

Barcelona, desembre 2024



Els continguts d'aquesta obra estan subjectes a una llicència de Reconeixement-NoComercial-SenseObraDerivada 4.0 Internacional de Creative Commons.

Es pot trobar una còpia completa dels termes d'aquesta llicència a creativecommons.org

Coordinació editorial

Generalitat de Catalunya
Centre de Telecomunicacions i Tecnologies de la Informació (CTTI)
Direcció General de Serveis Digitals (DGSEC)

María Ester Manzano Peláez, Directora General de Serveis Digitals i Experiència Ciutadana
Daniel Santanach Casals, Responsable dels Serveis Transversals d'Analítica i IA

Aicha Ait Ouaddi El Mamouny, Carla Laure Carceller, Josep Maria Flores Juanpere, Cristina Vila, Anna Canyelles i Roca, Esther Peña de las Heras, Jorge Escriba Murciano, María Pilar Bertran González, Laia Gómez Messia, Esther León Abaddin, Genís Vives Cantero.

Equip de treball

NTT DATA

Sergio Mejía Martínez, Senior Manager
Carmen García Larrode, Business Manager
Josep Almirall Bofil, UX Project Manager
David Torres Leal, Manager
Giancarlo Daniel Hernández Martínez, UX Research
Alba Guerrero Casares, UX Writing
Marta Vázquez Ortega, Accessibilitat, inclusió i disseny
Blanca Poy Roig, Analista d'intel·ligència artificial

