UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the bachelor thesis:

Escudé Macià, Martina. *Disseny i Implementació d'una Arquitectura RAG per la Integració de Models Generatius a Marlex*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Enginyeria de Dades)

This version is available at https://ddd.uab.cat/record/337340

under the terms of the CC BY-SA license.

Disseny i Implementació d'una Arquitectura RAG per la Integració de Models Generatius a Marlex

Martina Escudé Macià

Resum – Aquest Treball de Final de Grau té com a objectiu dissenyar, implementar i avaluar una arquitectura tècnica per integrar models de llenguatge de gran escala (LLMs) en l'entorn empresarial de Marlex, consultora de recursos humans en plena expansió. La incorporació d'IA generativa en entorns corporatius presenta reptes significatius: manca d'accés a dades internes actualitzades, absència de traçabilitat i dificultat per garantir la privacitat i el control del contingut generat.
A partir de l'anàlisi de l'estat de l'art i dels requisits de Marlex, es dissenya un sistema RAG híbrid que combina cerca semàntica (ChromaDB) i cerca lexical (SQL Server), fusionades mitjançant Reciprocal Rank Fusion (RRF), amb processament íntegrament local mitjançant models de codi obert. Els resultats validen empíricament la superioritat de l'enfocament híbrid i estableixen un marc metodològic reutilitzable per a qualsevol organització que vulgui integrar LLMs amb dades corporatives sota requisits estrictes de privacitat i control.

Paraules clau – Intel·ligència Artificial Generativa, Large Language Models (LLM), Retrieval-Augmented Generation (RAG), Embeddings, Vector Store, Reciprocal Rank Fusion (RRF), Prompt Engineering, Governança de Dades.

Abstract – This Bachelor's Thesis aims to design, implement and evaluate a technical architecture for integrating large language models (LLMs) into the enterprise environment of Marlex, a human resources consultancy in full expansion. Deploying generative AI in corporate settings presents significant challenges: lack of access to up-to-date internal data, absence of source traceability, and difficulty ensuring privacy and control over generated content. Based on the state of the art analysis and Marlex's specific requirements, a hybrid RAG system is designed combining semantic search (ChromaDB) and lexical search (SQL Server), fused through Reciprocal Rank Fusion (RRF), with fully local processing using open-source models. Results empirically validate the superiority of the hybrid approach and establish a reusable methodological framework for any organisation seeking to integrate LLMs with corporate data under strict privacy and control requirements.

Keywords – Generative Artificial Intelligence, Large Language Models (LLM), Retrieval-Augmented Generation (RAG), Embeddings, Vector Store, Reciprocal Rank Fusion (RRF), Prompt Engineering, Data Governance.

1 Introducció – Problem Statement

Aquest Treball de Final de Grau (TFG) té com a objectiu principal el disseny d'una arquitectura tècnica que permeti integrar models de llenguatge de gran escala (LLMs) en un entorn empresarial, proporcionant-los context estructurat, governat i alineat amb els sistemes d'informació corporatius. El projecte es planteja des d'una perspectiva externa, assumint el rol de consultora, amb l'objectiu d'analitzar de manera objectiva quina estructura de *backend* (capa de servidor que gestiona la lògica i les dades) i gestió de dades és necessària per desplegar solucions basades en intel·ligència artificial generativa dins d'una organització.

Tot i que el punt de partida del treball va ser la voluntat d'automatitzar la generació de documents de preparació de reunions comercials a partir de les dades del CRM corporatiu (*Salesforce*), el projecte ha evolucionat cap a un enfocament més estructural. L'objectiu ja no és únicament resoldre una necessitat funcional concreta, sinó definir una infraestructura reutilitzable capaç de donar suport a futurs casos d'ús basats en LLM. La generació de documents comercials esdevé un primer cas d'ús específic, parametritzable i mesurable que permet posar a prova l'arquitectura proposada en altres contextos organitzatius.

Per tal d'analitzar quina estratègia de recuperació és més adequada per integrar LLMs amb dades corporatives de Marlex, es dissenya i s'implementa un sistema RAG híbrid basat en un microservei extern amb arquitectura *Retrieval-Augmented Generation*. Aquest sistema combina una cer-

• E-mail: martinaescu10@gmail.com
• Tutor: Ramon Grau Sala
• Curs 2025/26

ca semàntica sobre una base de dades vectorial i una cerca lexical sobre SQL Server, fusionades mitjançant *Reciprocal Rank Fusion* (RRF) [24]. L'avaluació no només permet mesurar el rendiment tècnic de cada modalitat de cerca, sinó també el seu impacte en termes de governança, control de dades, escalabilitat i mantenibilitat.

1.1 Marlex

Marlex, una empresa de consultoria i serveis de recursos humans nascuda a Girona, és actualment una de les empreses líders del seu mercat a Catalunya. La raó de ser de Marlex és ajudar les empreses a desenvolupar els seus projectes mitjançant la selecció del millor talent, a través de tres equips especialitzats: un equip centrat en la cerca de talent directiu i consultoria de lideratge (**Sarah Marlex**), un altre especialitzat en la selecció de perfils de *middle management* (**Paul Marlex**) i un tercer focalitzat en la cerca de perfils operatius, que es poden incorporar directament a l'empresa o a través de treball temporal (TT) (**Marlex People**).



Fig. 1: Logotip de Marlex.

Com a objectiu de negoci, Marlex vol consolidar-se com a referent en tot el territori peninsular i continuar amb l'expansió progressiva a la resta del territori estatal. Per assolir aquest objectiu, Marlex vol reforçar el seu comportament *data-driven*, basant la presa de decisions en dades fiables.

1.2 La transformació de l'entorn analític de Marlex davant d'un nou context empresarial

En els darrers anys, Marlex ha iniciat un procés d'evolució del seu entorn analític amb l'objectiu d'estructurar millor la informació corporativa i consolidar les seves fonts de dades. Aquest procés inclou el desenvolupament d'un *data warehouse* altament governat, escalable i fiable.

En aquest context, la incorporació de models de llenguatge de gran escala (LLMs) requereix una arquitectura capaç de proporcionar context fiable i controlat. Els sistemes generatius no poden operar directament sobre dades disperses o poc governades [27], sinó que necessiten mecanismes de recuperació d'informació estructurats i traçables.

Davant d'aquesta necessitat, la implementació d'un sistema *Retrieval-Augmented Generation* (RAG) es planteja com la solució definitiva al problema de la manca de context empresarial intern dels LLM, ja que permet connectar els models generatius amb dades corporatives de manera controlada i escalable.

2 Objectiu principal

L'objectiu principal d'aquest Treball de Fi de Grau és el disseny, la implementació i l'avaluació d'un sistema de Retrieval-Augmented Generation (RAG) híbrid per a la integració de models de llenguatge de gran escala (LLMs) en entorns empresarials. Mitjançant la combinació de cerca semàntica i cerca lexical sobre fonts corporatives heterogènies, es desacobla la capa de dades de la capa generativa, garantint el control sobre la qualitat, la privacitat i la seguretat del contingut generat sense dependència de serveis externs.

Per assolir aquest propòsit, el treball persegueix un conjunt d'objectius específics que el concreten. En primer lloc, es vol definir un marc d'arquitectura reutilitzable, és a dir, establir una infraestructura de *backend* capaç de donar suport no només al cas d'ús comercial actual, sinó també a futures implementacions d'intel·ligència artificial generativa dins de l'organització. En segon lloc, es pretén analitzar la viabilitat comparativa de la solució, avaluant de forma objectiva l'enfocament desacoblat basat en microserveis i RAG extern i el seu impacte en termes de governança, control de dades, escalabilitat i mantenibilitat en un entorn empresarial real. Igualment, el projecte busca implementar un sistema de governança i traçabilitat que garanteixi que la integració d'LLMs compleixi els estàndards de seguretat de dades, incloent-hi l'auditoria de consultes i l'accés controlat a les fonts d'informació. A més, es vol avaluar l'impacte en l'eficiència comercial, estimant l'estalvi de temps i la qualitat dels documents generats per al cas d'ús de preparació de reunions comercials mitjançant una validació qualitativa amb el responsable de negoci. Finalment, es proposa establir criteris d'escalabilitat i costos, analitzant l'eficiència econòmica de les solucions proposades segons el volum d'execucions i la complexitat de manteniment.

3 Estat de l'art

Els models de llenguatge de gran escala (LLMs) han demostrat una elevada capacitat per generar contingut en llenguatge natural a partir d'instruccions textuals. Tanmateix, quan s'apliquen en entorns empresarials, presenten limitacions significatives: manca d'accés a dades internes actualitzades, absència de traçabilitat sobre les fonts utilitzades i dificultat per garantir coherència amb la informació corporativa.

Aquestes limitacions han impulsat el desenvolupament d'arquitectures que desacoblen la capa generativa de la capa de dades. En aquest context, els sistemes *Retrieval-Augmented Generation* (RAG) introdueixen una capa intermèdia de recuperació d'informació que permet injectar context específic en temps d'execució. Això es materialitza mitjançant mecanismes de cerca sobre bases documentals, sistemes estructurats o magatzems de vectors, els quals proporcionen al model generatiu informació rellevant abans de produir la resposta final.

La integració d'LLMs requereix definir clarament tres components: una capa de dades estructurada i governada per a l'emmagatzematge i ingestió de dades, una capa de recuperació capaç d'identificar el context pertinent i una capa generativa responsable de sintetitzar la informació. La literatura recent destaca que la robustesa del sistema no depèn únicament del model utilitzat, sinó de la qualitat i l'arquitectura del *pipeline* de dades que l'alimenta [1].

En conseqüència, l'estudi de les arquitectures RAG no es limita a l'ús d'un model generatiu, sinó que implica analitzar com s'estructuren, indexen i recuperen les dades corporatives per garantir fiabilitat, escalabilitat i control en entorns empresarials [5, 8, 9].

3.1 Tecnologies punteres al sector

Segons una predicció de Gartner de 2023, més del 80% de les empreses adoptaran o utilitzaran intel·ligència artificial generativa a través d'APIs o aplicacions habilitades amb aquesta tecnologia abans del 2026 [10].

El sector de la consultoria de RRHH mostra una tendència clara cap a la **inversió en plataformes de dades i intel·ligència artificial**. Empreses líders com Randstad i Adecco han apostat per la **centralització de les seves dades comercials sobre CRM al núvol**: Randstad va adoptar Salesforce Sales Cloud com a plataforma de CRM i automatització de vendes [34], i Adecco ha ampliat la seva col·laboració amb Salesforce per integrar **Data Cloud** i capacitats d'IA generativa, unificant les seves instàncies en una visió 360 dels comptes per a 27.000 usuaris [35]. Aquesta tendència confirma que les **bases de dades governades al núvol** s'estan consolidant com el fonament dels projectes d'IA al sector.

3.2 Tendències en TFG i recerca

A partir de l'anàlisi dels TFG recents, les tendències punteres clau són:

- **RAG per extracció d'informació/arquitectures avançades:** Aplicació de *Retrieval-Augmented Generation* per millorar l'extracció de dades de documents intel·ligents [1, 3].
- **Assistents virtuals per suport tècnic:** Desenvolupament d'assistents basats en IA per millorar l'atenció i consulta tècnica interna [2].
- **RAG per consultes clíniques:** Aplicació de RAG per millorar consultes en l'àmbit de la salut, com la immunització pediàtrica [4].
- **Tècniques de llenguatge natural per consultes documentals:** Ús de models de llenguatge per optimitzar les cerques en documents empresarials [5].
- **IA orientada a l'anàlisi de negocis:** Aplicació de IA per desenvolupar funcionalitats orientades al *business intelligence* [6].

3.3 Arquitectures estandarditzades per a sistemes RAG

Les arquitectures més reconegudes per al desenvolupament de sistemes RAG (*Retrieval-Augmented Generation*) inclouen:

- **Arquitectura RAG Tradicional (Arquitectura de Codi Dual):** Aquesta arquitectura es fonamenta en dos processos principals. La *Pipeline de Recuperació* utilitza un model de cerca (per exemple, BM25 o un model *transformer* com BERT [11]) per recuperar documents rellevants d'una base de dades. Un cop recuperada la informació, la *Pipeline de Generació* utilitza un model de llenguatge (LLM), com GPT-3 [12] o BART [13], per generar una resposta contextualitzada a partir de la informació obtinguda durant la cerca. Aquesta arquitectura és àmpliament utilitzada per a tasques com la resposta a preguntes o la generació de respostes a partir de documents de suport.
- **Arquitectura RAG basada en fine-tuning (end-to-end):** En aquesta arquitectura, el sistema es construeix com un model *end-to-end*, entrenant conjuntament les parts de recuperació i generació. El sistema aprofita models com T5 o BART per integrar la recuperació i la generació en un únic model entrenat, millorant l'eficiència i l'adaptabilitat del sistema. Aquesta solució és adequada quan es vol un sistema completament entrenat que integri tots els processos de manera unificada.
- **Arquitectura Híbrida (RAG + models basats en regles):** En una arquitectura híbrida, els models basats en regles utilitzen regles predefinides per generar respostes, en lloc de dependre només de l'aprenentatge automàtic. Aquest enfocament permet un major control sobre la qualitat de les respostes i garanteix que compleixin polítiques internes de seguretat, privacitat o regulacions legals, afegint així un nivell addicional de control al sistema.
- **Arquitectura de Microserveis per a RAG:** Aquesta arquitectura segmenta el sistema en diversos microserveis, cadascun encarregat d'una tasca específica, com la recuperació d'informació, la generació de respostes o el monitoratge de resultats. L'ús de microserveis facilita l'escalabilitat i la mantenibilitat del sistema, ja que cada component pot ser actualitzat o escalat de manera independent sense afectar els altres components.

Aquestes arquitectures no són mútuament excloents: és possible combinar-les, per exemple, adoptant una arquitectura de microserveis on cadascun dels serveis implementi internament una estratègia de recuperació diferent.

A diferència dels enfocaments analitzats, aquest treball adopta una arquitectura de microserveis amb cerca híbrida, ja que és la que millor s'adapta als requisits específics de Marlex: privacitat total de les dades (processament íntegrament local), baix cost operatiu en comparació amb solucions cloud, gràcies a l'ús de models de codi obert i a l'aprofitament de la infraestructura de servidor existent (*on premise*), i heterogeneïtat de fonts (web corporativa, notícies de premsa i CRM relacional). Aquesta elecció es valida empíricament a la secció 7, on es presenten resultats que evidencien la superioritat de l'enfocament híbrid sobre les modalitats individuals.

4 METODOLOGIA

La metodologia del projecte combina una fase inicial de recollida de requisits amb un desenvolupament tècnic guiat per la presa de decisions arquitectòniques. Per a la recollida de requisits es van realitzar reunions amb els responsables de Marlex amb l'objectiu d'entendre les necessitats actuals i identificar quines dades internes i externes intervenen en els seus processos. A partir d'aquesta anàlisi es va determinar quina informació podia ser estructurada i utilitzada com a context per a la integració de models generatius, i es van identificar les limitacions de l'entorn analític existent. Aquesta primera aproximació, recolzada en l'anàlisi dels sistemes ja en ús, especialment el CRM *Salesforce*, que té un paper fonamental en la gestió de dades, va permetre establir els requisits tècnics necessaris per construir una arquitectura capaç de gestionar el flux de dades de manera eficient i escalable, així com definir els criteris d'escalabilitat,

traçabilitat i governança que han regit tota la implementació.

El desenvolupament del projecte es planteja com un procés de presa de decisions arquitectòniques orientades a establir les bases per a la integració d'LLMs en un entorn empresarial. Més enllà de la implementació d'una funcionalitat concreta, l'objectiu és analitzar quina estructura tècnica permet proporcionar context fiable als models generatius, garantint control i escalabilitat. S'adopta una metodologia *Agile* basada en *Scrum*, adaptada a les necessitats del projecte. Aquesta metodologia, centrada en *sprints* iteratius, permet contrastar decisions de disseny i validar progressivament les implicacions tècniques de cada aproximació, tot i ser executada per un únic responsable de la implementació. La flexibilitat d'aquesta metodologia assegura que els canvis i adaptacions en el disseny es puguin incorporar ràpidament durant el procés de desenvolupament.

La validació del sistema s'ha plantejat com un exercici de mesura quantitativa centrat en tres dimensions. La primera és el rendiment de la recuperació, és a dir, la precisió del sistema per retornar els fragments més rellevants per a cada pregunta, mesurada amb la mètrica Precision@k [26]. La segona és la qualitat del context, entesa com la capacitat del sistema d'utilitzar les dades corporatives de manera coherent i de retornar informació fidel als documents originals. La tercera és la millora iterativa, que recull l'evolució dels resultats entre versions del sistema i la correcció dels problemes detectats durant l'avaluació. Finalment, s'han definit els mecanismes de monitoratge necessaris per assegurar el correcte funcionament del sistema i la seva coherència amb els criteris de governança establerts.

5 Disseny d'una Arquitectura RAG Híbrida

5.1 Visió general

El sistema RAG dissenyat per a Marlex transforma la interacció amb la IA en un procés estructurat que garanteix que la resposta no depengui de la memòria interna del model, sinó de dades reals i actualitzades. El flux s'inicia amb el *user prompt*, on l'usuari defineix la seva necessitat. El *retriever* actua com a cercador actiu, consultant la memòria no paramètrica (informació externa: notícies de premsa, pàgines web corporatives i dades del CRM) per extreure el context rellevant. Aquesta informació es fusiona amb la petició original creant un *prompt* enriquit que s'entrega al LLM. Finalment, el model utilitza la seva memòria paramètrica (el coneixement general après durant l'entrenament) per sintetitzar una resposta fonamentada en evidències externes, traçable i amb un risc d'al·lucinacions reduït [27].

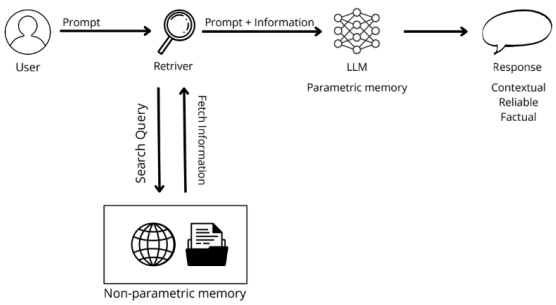


Fig. 2: Arquitectura conceptual del sistema RAG (Retrieval-Augmented Generation) [14]

A nivell estructural, l'arquitectura s'organitza en dos blocs operatius complementaris: l'*Indexing Pipeline*, un procés *offline* que construeix la base de coneixement a partir de les fonts corporatives, i el *Generation Pipeline*, un procés *online* que respon les consultes en temps real. Aquesta separació desacobla la capa de dades de la capa generativa i permet actualitzar, escalar i auditar cada bloc de manera independent.

5.2 Decisions de disseny

El disseny es fonamenta en un conjunt de decisions arquitectòniques orientades a la fiabilitat, el control de les dades i l'escalabilitat. La primera és l'adopció d'una estratègia de recuperació híbrida: en lloc d'una cerca semàntica o lexical pures, es combinen els dos canals en paral·lel i es fusionen mitjançant Reciprocal Rank Fusion (RRF) [24], aprofitant la comprensió del significat dels *embeddings* i la precisió de la coincidència exacta de termes per a dades concretes del CRM (la justificació detallada es recull als Annexos A.5, A.8 i A.9). La segona és un emmagatzematge dual, en què cada fragment es desa simultàniament a ChromaDB per a la cerca semàntica i a SQL Server per a la cerca lexical, reaprofitant la infraestructura ja existent a Marlex (Annexos A.4 i A.7). La tercera és el processament íntegrament local: tots els models (embeddings i generació) s'executen via Ollama als servidors interns de l'empresa, de manera que cap dada de clients surt de la infraestructura corporativa, garantint privacitat, baix cost operatiu i independència de proveïdors externs (Annexos A.6, A.7, A.10 i A.11).

6 Desenvolupament

6.1 Pipeline d'ingesta (*Indexing Pipeline*)

Aquesta fase construeix la memòria no paramètrica de Marlex, transformant dades en brut de fonts heterogènies en una estructura cercable. El procés és modular: cada font té el seu propi mòdul d'ingesta, i tots desemboquen en un *pipeline* comú de fragmentació i indexació.

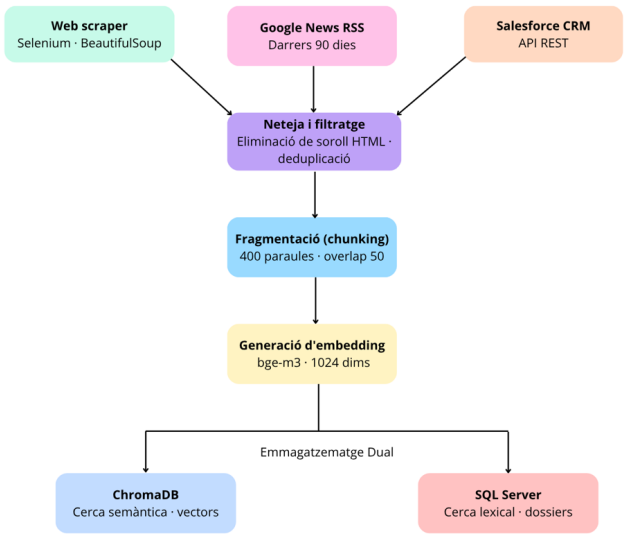


Fig. 3: Diagrama de flux de l'*Indexing Pipeline*: les tres fonts d'informació convergeixen cap al pipeline comú de neteja, fragmentació, generació d'*embeddings* i emmagatzematge dual a ChromaDB i SQL Server.

Càrrega de dades. S'han desenvolupat mòduls Python a mida per a cada font d'informació. Un sistema de *web scraping* (*scraper.py*) basat en Selenium [18] i BeautifulSoup [19] rastreja fins a 20 pàgines de la web corporativa del prospecte, aplicant una fase de neteja que elimina el soroll (codi HTML, menús de navegació, avisos de cookies, peus de pàgina) per conservar únicament el text de valor estratègic, i detecta i salta automàticament les URL de catàlegs de producte o pàgines de cerca. Paral·lelament, un connector a Google News RSS monitoritza les notícies dels darrers 90 dies filtrades per nom d'empresa, navegant les redireccions per arribar a l'article original i aplicant deduplicació per identificador únic. Finalment, el mòdul *salesforce_enrichment.py* extreu les dades relacionals internes de Salesforce via API REST [20]. La decisió tècnica clau d'aquesta fase és l'assignació de metadades estructurades a cada fragment en el moment de la ingesta (empresa, origen i data), que permeten prioritzar la informació interna del CRM per sobre de les fonts externes en cas de conflicte (Annex A.1).

Fragmentació (*chunking*). Un cop les dades estan netes, es divideixen en fragments de 400 paraules amb un solapament de 50 paraules entre fragments consecutius [14]. El solapament actua com a pont contextual: els últims 50 tokens d'un fragment apareixen a l'inici del següent, assegurant que cap idea comercial quedi tallada en el punt de divisió i que el LLM sempre disposi d'un context complet en recuperar qualsevol bloc (Annex A.2).

Generació d'*embeddings*. Cada fragment es converteix en un vector de 1024 dimensions mitjançant el model *bge-m3* [36], executat localment via Ollama [22]. El model és multilingüe i gestiona adequadament la relació semàntica entre el català i el castellà, sense cap cost per token ni dependència de serveis externs (Annexos A.3 i A.11).

Emmagatzematge dual. Cada fragment s'emmagatzema en dos sistemes complementaris: a ChromaDB [23] (base de dades vectorial integrada i local) per a la cerca semàntica, i a SQL Server (*[dbo].[dossiers]* i *[dbo].[noticies]*, Annex A.16) per a la cerca lexical, aprofitant la infraestructura ja existent a Marlex (Annexos A.4 i A.7).

6.2 Pipeline de generació (*Generation Pipeline*)

El *Generation Pipeline* s'activa en temps real quan el Hunter fa una consulta i executa les seves etapes seqüencials.

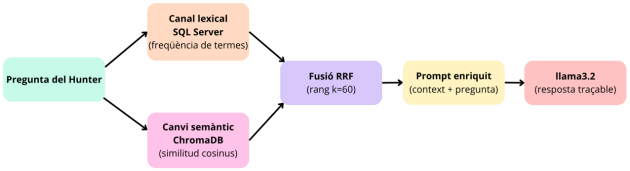


Fig. 4: *Generation Pipeline*: la pregunta del *Hunter* (perfil comercial de Marlex responsable de la prospecció de nous clients) entra als dos canals de recuperació paral·lels (semàntic i lexical), els resultats es fusionen mitjançant RRF i es construeix el *prompt* enriquit que processa llama3.2.

Recuperació híbrida amb RRF. La recuperació opera en dos canals paral·lels. El canal semàntic (ChromaDB) converteix la pregunta en un vector amb *bge-m3* i recupera els fragments amb major similitud de cosinus [26], entenent el significat de la pregunta independentment del vocabulari exacte. El canal lexical (SQL Server) identifica els documents que contenen els termes clau de la pregunta i els puntua per freqüència de termes (TF) [25], essent especialment eficaç per a dades concretes del CRM (noms, xifres, dates) on la coincidència exacta és necessària. Els resultats dels dos canals es fusionen amb Reciprocal Rank Fusion (RRF) [24], que unifica les dues llistes ordenades per rang sense necessitat de normalitzar puntuacions entre sistemes d'escala diferent; un fragment que apareix en ambdós canals puja significativament en el rànquing final (Annexos A.6, A.7, A.8 i A.9).

Augmentació i generació. Es construeix un *prompt* enriquit que combina la pregunta original amb els fragments recuperats. S'aplica Contextual Prompting [29]: el model rep instruccions explícites per respondre estrictament a partir del context proporcionat i per indicar "Informació no trobada" quan la dada no existeix als documents recuperats, reduint substancialment el risc d'al·lucinació [27]. El model llama3.2 (3B de paràmetres) [28], executat localment via Ollama [22], processa el *prompt* final i sintetitza els fragments recuperats en una resposta coherent. Tot el processament té lloc als servidors interns de Marlex: cap dada de clients surt de la infraestructura de l'empresa (Annexos A.7 i A.10).

6.3 Sistema d'avaluació quantitativa

Per validar el sistema de manera objectiva i guiar les millores iteratives, s'ha dissenyat un *framework* d'avaluació propi que mesura tant la qualitat de la recuperació, amb mètriques estàndard d'*Information Retrieval* [26], com la qualitat de la generació de respostes, mitjançant un model de llenguatge que actua com a jutge [30]. El disseny detallat del *framework*, el dataset d'avaluació i la implementació de l'LLM-jutge es descriuen als Annexos A.12 i A.17.

6.4 El cas d'ús: generació del dossier comercial

El cas d'ús que articula el projecte és la generació automàtica d'un dossier de preparació de reunió per a un prospecte. A partir d'una única entrada (el nom de l'empresa i, opcionalment, la seva URL), el pipeline complet (ingesta, recuperació híbrida i generació) produeix un document estructurat en format Markdown (*dossier.md*) llest perquè el consultor comercial (*Hunter*) el consulti abans de la visita. El dossier s'organitza en seccions pensades per a la conversa comercial (vegeu Figura 13 a l'Annex A.19): un indicador de potencial Marlex (puntuació heurística de 0 a 10 segons els senyals detectats al contingut), la informació corporativa (activitat, sector, any de fundació, certificacions), les xifres clau (dimensió operativa i facturació, com a indicador del potencial de contractació), les línies de negoci (per identificar quins perfils de RRHH pot necessitar el prospecte), la cultura i valors (per personalitzar la proposta de Marlex) i un recull de notícies recents dels darrers 90 dies (esdeveniments que poden obrir conversa o anticipar noves necessitats de personal). Per als comptes que ja són clients de Marlex, el dossier incorpora a més la secció "RELACIÓN MARLEX", que integra les dades internes del CRM de Salesforce (estat de la relació, presència a la llista taronja, marge generat, serveis actius i Temps Treball (TT) actives) i fins i tot destaca oportunitats de negoci detectades a partir de l'encreuament de la informació externa i la interna.

A més del document estàtic, el sistema ofereix un mode interactiu: el *Hunter* pot formular preguntes en llenguatge natural sobre qualsevol prospecte i obtenir respostes immediates i traçables ancorades en les fonts (vegeu Figura 12 a l'Annex A.19).

7 PRESENTACIÓ I DISCUSSIÓ DE RESULTATS

7.1 Resultats quantitatius de la recuperació

L'avaluació del sistema s'ha realitzat sobre un dataset de 26 preguntes representatives distribuïdes en tres empreses (Bonàrea, Europastry i Mercadona) i tres tipologies (factual, estratègica i relacional), mesurant la qualitat de la recuperació amb les mètriques Precision@3 i Precision@5 [26]. Els resultats globals es presenten a la Taula 2 i el desglossament per tipologia a la Taula 3 (Annex A.15), i es resumeixen visualment a la Figura 5.

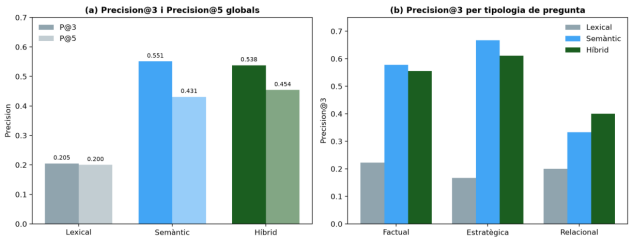


Fig. 5: Resultats de la recuperació: (a) Precision@3 i Precision@5 globals per mode de cerca; (b) Precision@3 desglossat per tipologia de pregunta

El sistema híbrid supera la cerca lexical en un +162% en Precision@3. Respecte a la cerca semàntica, l'híbrid obté un Precision@3 equivalent (0,538 vs 0,551) però la supera en Precision@5 (0,454 vs 0,431) i, sobretot, en les preguntes relacionals (0,400 vs 0,333), les de major valor de negoci, ja que són les úniques que requereixen el canal lexical sobre les dades internes del CRM.

7.2 Interpretació dels resultats

Preguntes factuals (P@3 = 0,556): Són les preguntes amb resposta objectiva i verificable ("Quin any es va fundar Bonàrea?", "Quants treballadors té Mercadona?"). El resultat s'explica perquè aquest tipus de dades sovint apareix en contextos molt breus del dossier i la cerca semàntica pot recuperar fragments relacionats però que no contenen la dada concreta. La cerca lexical complementa adequadament recuperant fragments on apareix l'any o el nombre exacte.

Preguntes estratègiques (P@3 = 0,611): Obtenen el segon millor resultat. La cerca semàntica és especialment eficaç en aquest tipus (P@3 semàntic = 0,667) perquè les preguntes sobre estratègia, sostenibilitat o model de negoci es beneficien de la comprensió del significat per sobre de la coincidència de paraules exactes. El sistema híbrid millora encara més perquè el canal lexical aporta fragments addicionals que contenen els termes clau de la pregunta.

Preguntes relacionals (P@3 = 0,400): Tot i ser les preguntes més difícils, ja que requereixen dades internes del CRM de Salesforce ("Quantes TT actives té Europastry?", "Quin marge ha generat Bonàrea?") que no es troben a cap font pública, són on el sistema híbrid mostra el seu avantatge més clar sobre la cerca semàntica pura (0,400 vs 0,333). La clau és que el canal lexical sobre *[dbo].[dossiers]* (Annex A.16) recupera precisament els fragments de la secció "RELACIÓ MARLEX" del dossier, que contenen aquestes dades. El sistema híbrid els atorga un ponderació de $\times 1{,}5$ i limita el *recall* semàntic per evitar que fragments semànticament propers però relacionalment irrellevants desplacin els *chunks* del CRM. Aquest comportament confirma el valor de la fusió: mentre que en preguntes factuals i estratègiques la cerca semàntica per si sola ja és molt competitiva, en les preguntes relacionals, les de major valor per al perfil Hunter, el canal lexical és imprescindible i és el que justifica l'enfocament híbrid.

7.3 Resultats de la qualitat de la generació

S'han avaluat tres mètriques de generació sobre el mode híbrid. La *Correctness* (0,523) compara la resposta generada amb la resposta de referència (*ground truth*) anotada al dataset, detectant errors factuals que la *Faithfulness* no captura: una resposta pot ser fidel al context recuperat però igualment incorrecta si el context no conté la dada exacta. El valor obtingut reflecteix que aproximadament la meitat de les respostes coincideixen substancialment amb la resposta esperada, un resultat coherent amb les limitacions ja esmentades: model generador i jutge de mida reduïda (3B) i cobertura parcial del corpus per a determinades preguntes factuals.

Més enllà de la recuperació, s'ha avaluat la qualitat de les respostes generades pel model llama3.2 [28] mitjançant un LLM-jutge executat localment via Ollama [22] i [30], aplicat sobre el mode híbrid (el de millor recuperació). Els resultats es presenten a la Taula 6 (Annex A.15).

El sistema obté una Faithfulness de 0,599, cosa que indica que al voltant del 60% de les afirmacions de les respostes es poden atribuir directament als fragments recuperats. Aquest valor confirma que la generació es fonamenta majoritàriament en el context proporcionat i no en el coneixement paramètric del model, complint l'objectiu principal d'una arquitectura RAG: ancorar les respostes en evidència documental i reduir el risc d'al·lucinació [27].

La Answer Relevance se situa en 0,538, reflectint que les respostes són pertinents respecte a la pregunta formulada. Cal tindre en compte la concisió del model llama3.2 (3B de paràmetres) i la presència de preguntes relacionals que requereixen dades del CRM no sempre disponibles al corpus. Aquests valors s'han d'interpretar amb cautela, ja que el model jutge coincideix amb el model generador (discussió de limitacions a l'Annex A.17). Ambdues mètriques són coherents amb la naturalesa d'un model local lleuger, prioritzat per privacitat i cost zero, i estableixen una línia base sòlida per mesurar l'impacte de futures millores com el *re-ranking* o l'ús de models de major mida.

7.4 Comparació amb la literatura i contextos anàlegs

Els valors de Precision@3 obtinguts (0,551) s'emmarquen dins dels rangs reportats a la literatura per a sistemes RAG de domini específic. Estudis com Lewis et al. (2020) [27] i Karpukhin et al. (2020) [32] mostren que sistemes RAG sobre corpus tancats i dominis especialitzats típicament obtenen Precision@5 entre 0,30 i 0,60 per a preguntes de domini, depenent de la qualitat i la cobertura del corpus. El sistema de Marlex assoleix P@3 = 0,538 amb un corpus de tres empreses (relativament petit), cosa que suggereix que el rendiment millorarà a mesura que s'ampliï el corpus amb més prospectes indexats.

La millora del +162% de la cerca híbrida respecte a la lexical és consistent amb els resultats reportats per Luan et al. (2021) [33], en sistemes híbrids sobre dominis corporatius, on la fusió semàntic-lexical assoleix millores d'entre el 60% i el 150% respecte a la modalitat lexical sola, especialment quan el corpus conté documents de fonts heterogènies (web, CRM, notícies) amb vocabularis diferenciats.

Un context anàleg és el dels sistemes d'*enterprise search* desplegats en grans corporacions per als seus equips comercials: solucions com Glean [37] o Guru [38] reporten millores similars en la productivitat del personal de vendes (reducció del 40–60% del temps de preparació de visites) quan s'implementa una cerca semàntica sobre el CRM intern, dada que reforça la viabilitat i el valor de negoci del sistema dissenyat.

7.5 Validació amb la responsable de projecte a Marlex

Per complementar l'avaluació quantitativa amb una perspectiva de negoci, es va presentar el sistema a la Revenue Operations Manager de Marlex, responsable de l'estratègia i l'operativa de l'equip comercial, qui actuà com a stakeholder principal del projecte.

La valoració va ser positiva. La responsable va destacar que el valor del sistema va més enllà de la generació automàtica del dossier comercial previ a la visita: el cas d'ús que considera de major impacte és la possibilitat que el consultor pugui fer preguntes en llenguatge natural sobre un compte concret i obtenir respostes immediates.

En termes d'impacte operatiu, es va estimar que el sistema permetria als Hunters reduir significativament el temps dedicat a la recerca manual d'informació, alliberant capacitat per centrar-se en l'anàlisi estratègica i la interacció directa amb el client. A mode d'estimació orientativa, si considerem que un Hunter dedica aproximadament 30 minuts a preparar manualment la informació d'un compte, i que el sistema genera el dossier en menys de 10 minuts, la ràtio resultant seria d'uns 3x més ràpid. Cal remarcar que aquesta estimació s'hauria de validar empíricament en un desplegament real, ja que el temps manual pot variar segons el compte i el consultor.

7.6 Eficiència econòmica i escalabilitat

A més del rendiment tècnic, s'ha analitzat l'eficiència econòmica de la solució segons el volum d'execucions. L'anàlisi de sensibilitat (Annex A.18) confirma que el cost de la solució local es manté **constant** independentment del volum, mentre que les alternatives de pagament per ús creixen **linealment** fins a esdevenir insostenibles a gran escala. Aquesta característica, sumada a la propietat total de les dades, reforça la viabilitat de l'arquitectura local per a un desplegament corporatiu de gran abast.

En conjunt, els resultats obtinguts confirmen la viabilitat tècnica i el valor de negoci del sistema dissenyat: l'enfocament híbrid supera les modalitats individuals en les preguntes de major valor per al negoci, les respostes generades s'ancoren majoritàriament en evidència documental, i la solució local ofereix el millor perfil cost-escalabilitat sota requisits estrictes de privacitat. Aquests resultats constitueixen la base empírica de les conclusions que es presenten a continuació.

8 CONCLUSIONS

8.1 Conclusió principal

Aquest treball demostra que és possible desplegar un sistema RAG híbrid de qualitat contrastada íntegrament amb tecnologies de codi obert i processament local, sense costos de llicència ni dependència de proveïdors externs, i sota requisits estrictes de privacitat i governança. Més enllà de les xifres concretes assolides, la contribució principal és metodològica: un marc de decisions tècniques, validat empíricament i reproduïble, que redueix la barrera d'entrada perquè una organització integri models generatius sobre les seves pròpies dades de manera fiable i controlada. El cas d'ús de la prospecció comercial de Marlex ha estat el primer banc de proves d'aquesta arquitectura, però el seu valor real rau en la seva capacitat de servir de fonament per a futures aplicacions d'intel·ligència artificial generativa dins de l'empresa.

8.2 Assoliment dels objectius

L'objectiu principal s'ha assolit: s'ha construït un sistema funcional complet, des de la ingesta multimodal de dades fins a la interfície de consulta del consultor, que desacobla

la capa de dades de la capa generativa i ancora les respostes del model en evidència documental recuperada, reduint el risc d'al·lucinació i garantint la traçabilitat de cada afirmació fins a la seva font.

Pel que fa als objectius específics:

- **Marc d'arquitectura reutilitzable:** L'arquitectura modular (*pipeline* d'ingesta, recuperació híbrida amb RRF i generació local) permet substituir cada component de manera independent, constituint una plantilla transferible a futures implementacions d'IA generativa dins de l'organització.
- **Viabilitat comparativa:** El sistema híbrid assoleix un Precision@3 de 0,538, un 162% superior a la cerca lexical, i supera la cerca semàntica pura en les preguntes relacionals, les de major valor per al negoci.
- **Governança i traçabilitat:** Tot el processament s'executa localment, cap dada de clients abandona la infraestructura de Marlex, i cada resposta generada cita les fonts que la sustenten, garantint l'auditabilitat i l'alineació amb la normativa de protecció de dades.
- **Impacte en l'eficiència comercial:** La validació amb la *stakeholder* principal va confirmar qualitativament el valor del sistema, destacant especialment la consulta en llenguatge natural com el cas d'ús de major impacte operatiu, amb una estimació orientativa de fins a 3x més ràpid en la preparació de visites comercials.
- **Escalabilitat i costos:** La solució local manté un cost pràcticament constant independentment del volum, mentre que les alternatives *cloud* creixen linealment fins a esdevenir insostenibles a escala.

8.3 Limitacions

El treball presenta limitacions que cal reconèixer:

- El corpus d'avaluació, format per tres empreses i vint-i-sis preguntes, és reduït i limita la significació estadística dels resultats.
- La qualitat de la generació, mesurada amb un model jutge local de només 3B de paràmetres, s'ha d'interpretar com una línia base orientativa i no com una mesura absoluta.
- Determinades preguntes factuals no poden respondre's perquè la informació no és present a cap de les fonts indexades.

8.4 Línies futures

Les línies de continuació naturals del projecte són:

- **Ampliació del corpus i del dataset d'avaluació** per guanyar robustesa estadística i representativitat dels resultats.
- **Incorporació de noves fonts d'ingesta** (registres mercantils, informes sectorials o perfils professionals) per augmentar la cobertura d'informació disponible.
- **Re-ranking amb models de major capacitat:** els experiments realitzats demostren que amb un model de 3B el *re-ranking* degrada el rendiment; explorar aquesta tècnica amb models més grans podria aportar millores significatives.
- **Desplegament productiu i integració a l'ecosistema Marlex:** el sistema actual constitueix un MVP (*Minimum Viable Product*) funcional i validat. El pla de posada en marxa preveu la seva integració a l'entorn operatiu de Marlex mitjançant una interfície web accessible als consultors, amb un *trigger* automàtic al CRM que executi el *pipeline* quan un Hunter programa una reunió, eliminant qualsevol acció manual prèvia a la visita.
- **Mode proactiu:** evolucionar el sistema d'un mode reactiu a un mode proactiu que detecti automàticament esdeveniments rellevants d'un prospecte (notícies, canvis al CRM, noves oportunitats) i els faci arribar al consultor sense que aquest hagi de fer cap consulta explícita.

AGRAÏMENTS

En primer lloc, a en Ramon Grau, tutor acadèmic, per la seva orientació i suport al llarg de tot el procés. A la Mireia, tutora a Marlex, per molt més que la seva orientació professional: pels seus coneixements, per la seva dedicació genuïna i per haver estat un pilar fonamental en tot el que ha representat aquesta experiència per a mi. A Marlex, per haver-me obert les portes i donat la confiança per desenvolupar aquest projecte en un entorn real. A la Martina, que va aparèixer a la meva vida per quedar-s'hi, per fer que els dies a Marlex fossin molt més que feina i per ser una de les millors coses que m'enduc d'aquesta etapa. A la Júlia i a la Mireia, companyes de la universitat, per acompanyar-me en cada pas d'aquesta etapa. A la meva parella i a tota la meva família, per ser el suport incondicional de sempre. I en especial a la meva mare, la meva guia i el meu referent, per estendre'm la mà sempre que ho he necessitat, i sovint abans fins i tot de demanar-ho.

Gràcies a tots.

REFERÈNCIES

[1] D. Laparra Osca, *Retrieval-Augmented Generation para la Extracción de Información de Documentos Inteligente*, TFG, Universitat Politècnica de València, 2024. Disponible a: https://riunet.upv.es/server/api/core/bitstreams/c775be7d-8724-476b-b0a4-73e055d03d50/content

[2] I. López Montaner, *Asistente virtual para soporte técnico basado en inteligencia artificial*, TFG, Universitat de Saragossa, 2024. Disponible a: https://zaguan.unizar.es/record/134102/files/TAZ-TFG-2024-119.pdf

[3] J. Poza Tamayo, *Estudio e implementación de arquitecturas RAG avanzadas*, TFG, Universitat de Valladolid, 2024. Disponible a: https://uvadoc.uva.es/bitstream/handle/10324/78491/TFG-B.%202490.pdf?sequence=1&isAllowed=y

[4] J. Beltrán Peris, *Arquitectura RAG aplicada a consultas clínicas*, TFG, Universitat Politècnica de València, 2024. Disponible a: https://riunet.upv.es/server/api/core/bitstreams/a40382e5-9e70-432f-8b01-ab041f83d939/content

[5] J. León Martín, *Técnicas de lenguaje natural aplicadas a consultas documentales*, TFG, Universitat d'Alcalá, 2025. Disponible a: https://ebuah.uah.es/dspace/bitstream/handle/10017/67198/TFG_Leon_Martin_2025.pdf

[6] A. García Fábregas, *Desarrollo de funcionalidades mediante IA orientada al análisis de negocios*, TFG, Universitat

Pontifícia Comillas, 2024. Disponible a: https://repositorio.comillas.edu/xmlui/bitstream/handle/11531/94969/TFG%20-%20Garcia%20Fabregas%2c%20Alberto.pdf

[7] TechRadar, *Salesforce says per-user pricing will be new AI norm*, 2024. Disponible a: https://www.techradar.com/pro/salesforce-says-per-user-pricing-will-be-new-ai-norm

[8] InterSystems, *Modelos LLM y aplicaciones RAG paso a paso*, 2024. Disponible a: https://es.community.intersystems.com/post/modelos-llm-y-aplicaciones-rag-paso-paso-parte-i-introduccin

[9] Juice Studio, *Arquitectura RAG empresarial para 2025*, 2024. Disponible a: https://juice-studio.com/arquitectura-rag-empresarial-para-2025/

[10] Gartner, *Generative AI adoption trends*, 2023. Disponible a: https://www.gartner.com/en/newsroom/press-releases/2023-10-11-gartner-says-more-than-80-percent-of-enterprises-will-have-used-generative-ai-apis-or-deployed-generative-ai-enabled-applications-by-2026

[11] J. Devlin, M. Chang, K. Lee i K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2019. Disponible a: https://arxiv.org/abs/1810.04805

[12] T. Brown et al., *Language Models are Few-Shot Learners*, 2020. Disponible a: https://arxiv.org/abs/2005.14165

[13] M. Lewis et al., *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation*, 2019. Disponible a: https://arxiv.org/abs/1910.13461

[14] A. Kimothi, *Simple Guide to Retrieval Augmented Generation*, Medium, 2023.

[15] LangChain, *LangChain Documentation*, 2024. Disponible a: https://python.langchain.com

[16] OpenAI, *Embedding models and API updates*, 2024. Disponible a: https://openai.com/index/new-embedding-models-and-api-updates

[17] Pinecone, *Vector Databases for Scalable AI*, 2024. Disponible a: https://www.pinecone.io/learn/vector-database

[18] Selenium, *Selenium Documentation*, 2024. Disponible a: https://www.selenium.dev/documentation/

[19] L. Richardson, *BeautifulSoup Documentation*, 2024. Disponible a: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[20] Salesforce, *Salesforce REST API Developer Guide*, 2024. Disponible a: https://developer.salesforce.com/docs/atlas.en-us.api_rest.meta/api_rest/intro_what_is_rest_api.htm

[21] Nomic AI, *nomic-embed-text*, 2024. Disponible a: https://ollama.com/library/nomic-embed-text

[22] Ollama, *Ollama Documentation*, 2024. Disponible a: https://ollama.com/

[23] Chroma, *Chroma Documentation*, 2024. Disponible a: https://docs.trychroma.com

[24] G. V. Cormack, C. L. A. Clarke i S. Buettcher, *Reciprocal Rank Fusion outperforms Condorcet and individual Rank Learning Methods*, SIGIR, 2009. Disponible a: https://plg.uwaterloo.ca/~gvcormack/cormacksigir09-rrf.pdf

[25] S. Robertson i H. Zaragoza, *The Probabilistic Relevance Framework: BM25 and Beyond*, 2009. Disponible a: https://www.staff.city.ac.uk/~sbrp622/papers/foundations_bm25_review.pdf

[26] C. Manning, P. Raghavan i H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008. Disponible a: https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

[27] P. Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, NeurIPS, 2020. Disponible a: https://arxiv.org/abs/2005.11401

[28] Meta AI, *Llama 3.2*, 2024. Disponible a: https://ai.meta.com

[29] OpenAI, *Prompt Engineering Guide*, 2024. Disponible a: https://platform.openai.com/docs/guides/prompt-engineering

[30] S. Es, J. James, L. Espinosa-Anke i S. Schockaert, *RAGAS: Automated Evaluation of Retrieval Augmented Generation*, 2023. Disponible a: https://arxiv.org/abs/2309.15217

[31] Parlament Europeu i Consell de la UE, *Reglament General de Protecció de Dades (RGPD)*, 2016. Disponible a: https://eur-lex.europa.eu

[32] V. Karpukhin et al., *Dense Passage Retrieval for Open-Domain Question Answering*, 2020. Disponible a: https://arxiv.org/abs/2004.04906

[33] Y. Luan et al., *Sparse, Dense, and Attentional Representations for Text Retrieval*, 2021. Disponible a: https://arxiv.org/abs/2005.00181

[34] AppsRunTheWorld, *Randstad Group Selects Salesforce Sales Cloud*, 2021. Disponible a: https://www.appsruntheworld.com/customers-database/purchases/view/randstad-group-netherlands-selects-salesforce-sales-cloud-for-sales-automation-crm-sales-engagement

[35] The Adecco Group, *The Adecco Group Accelerates AI Integration with Salesforce Agreement*, 2024. Disponible a: https://www.adeccogroup.com/our-group/media/press-releases/the-adecco-group-accelerates-ai-integration-with-salesforce-agreement

[36] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian i Z. Liu, *BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation*, 2024. Disponible a: https://arxiv.org/abs/2402.03216

[37] Glean, *Glean for Sales Teams*, 2024. Disponible a: https://www.glean.com/solutions/sales

[38] Guru, *Guru for Sales Enablement*, 2024. Disponible a: https://www.getguru.com/solutions/sales

A ANNEXOS

A.1 Justificació de la càrrega i filtratge de dades (data loading)

Per a la ingesta de dades, es va considerar inicialment l'ús de mètodes convencionals com la càrrega manual mitjançant fitxers CSV. No obstant això, aquesta opció es va descartar ràpidament a causa de la falta de sincronització i l'elevat risc d'error humà que comporta la manipulació de dades sensibles. En el seu lloc, es van desenvolupar mòduls Python a mida per a cada font: Selenium [18] i BeautifulSoup [19] per al web scraping de pàgines corporatives, un connector a Google News i RSS per a notícies de premsa, i un mòdul d'enriquiment que consulta Salesforce [20] per extreure les dades relacionals del CRM. Aquesta arquitectura modular permet automatitzar el flux i assignar metadades d'origen a cada fragment en el moment de la ingesta, garantint que el sistema pugui prioritzar la informació interna del CRM per sobre de les fonts externes en cas de conflicte o contradicció.

A.2 Anàlisi de la fragmentació de context (chunking)

En la fase de fragmentació, es va avaluar l'ús del Character Text Splitting (mètode de mida fixa), que divideix el text cada cert nombre de caràcters sense tenir en compte el contingut. Aquesta alternativa es va desestimar perquè sovint tallava frases de manera arbitrària, separant conceptes clau (com el nom d'un client de la seva facturació) i destruint el sentit comercial del text. L'elecció final d'una finestra de 400 paraules amb un *overlap* de 50 paraules respon a la necessitat de respectar l'estructura del llenguatge natural [14]. El solapament actua com un pont contextual: els últims 50 tokens d'un fragment apareixen també a l'inici del següent, assegurant que la informació es mantingui cohesionada i que el LLM sempre disposi d'un context complet en recuperar qualsevol bloc.

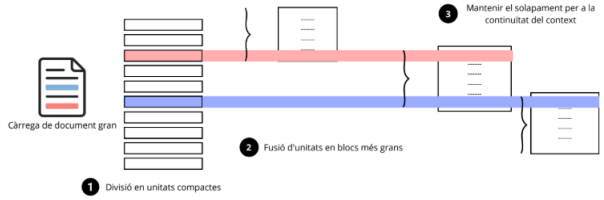


Fig. 6: Procés de fragmentació de dades (*data chunking*)

A.3 Selecció del model d'embeddings

Pel que fa a la conversió vectorial, es van analitzar diverses opcions. Els models d'OpenAI (*text-embedding-3-small*, *text-embedding-ada-002*) [16] ofereixen bona qualitat però requereixen enviar les dades a servidors externs, cosa incompatible amb el requisit de privacitat de Marlex. Els models de codi obert allotjats localment, en canvi, permeten un control total sobre les dades.

S'ha triat el model *bge-m3* [36], executat localment via Ollama [22], per les raons següents: opera completament en local sense cap cost per token ni dependència externa; genera vectors de 1.024 dimensions amb una representació semàntica rica; és un model multilingüe d'última generació que gestiona de manera robusta la relació semàntica entre el català i el castellà; i s'instal·la i executa amb una sola comanda (ollama pull bge-m3), sense configuració addicional de servidors. La tria es va validar empíricament durant la fase de millora iterativa: en substituir el model inicial (*nomic-embed-text*, 768 dimensions) per *bge-m3*, la Precision@3 de la cerca semàntica va passar de 0,256 a 0,551 (vegeu Annex A.11).

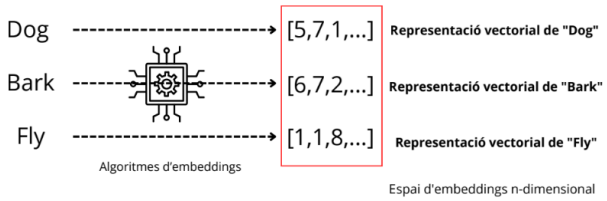


Fig. 7: El procés d'*embedding* converteix el text en vectors i genera un espai vectorial específic per al corpus de dades de Marlex

A.4 Infraestructura d'emmagatzematge (*vector database*)

Per a l'emmagatzematge vectorial, es van avaluar solucions gestionades al núvol com Pinecone [17] o Weaviate, descartades per la dependència de serveis externs, costos per consulta i el risc que les dades de clients surtin de la infraestructura de Marlex. S'ha seleccionat ChromaDB [23] per ser integrada i local (s'executa dins del procés Python), de cost zero (codi obert sense llicència), de desplegament trivial (*pip install chromadb*) i amb integració nativa amb Ollama. Cal diferenciar que SQL Server continua sent el motor principal per a la informació estructurada (taules *[dbo].[noticies]* i *[dbo].[dossiers]*), aprofitant la infraestructura ja existent a Marlex.

A.5 Estratègia de recuperació híbrida (retrieval)

Es va descartar la cerca per paraules clau tradicional perquè és incapaç d'entendre la intenció de l'usuari sense el terme exacte (no relacionaria "problemes de facturació" amb "incidències en els cobraments"). La cerca semàntica pura tampoc és suficient: no garanteix la recuperació de dades concretes del CRM (TT actives, marges) quan la pregunta conté termes que no apareixen en el context semàntic. S'ha implementat una estratègia híbrida que combina:

Cerca semàntica (ChromaDB). Recupera els fragments més propers semànticament, entenent el significat independentment del vocabulari exacte.

Cerca lexical (SQL Server). Identifica documents amb els termes clau i els puntua per freqüència de termes (TF) [25].

Els resultats es fusionen amb RRF [24], que combina les dues llistes sense normalitzar puntuacions. L'avaluació confirma la superioritat: P@3 lexical = 0,205; semàntic = 0,551; híbrid = 0,538 ($+162\%$ respecte a la lexical).

A.6 Selecció del motor d'inferència local davant OpenAI i serveis cloud

Es va avaluar l'ús de l'API d'OpenAI (*gpt-4o* i *text-embedding-3-small*) com a motor d'*embeddings* i genera-

ció, però es va descartar per tres raons:

Privacitat de les dades. Marlex treballa amb informació comercial sensible (marges, estat de la relació, historial de visites). Enviar-la a servidors externs implica un risc incompatible amb les polítiques de l'empresa i el RGPD [31].

Cost operatiu. Tot i que el cost per token és modest, creix linealment amb el volum. Ollama opera a cost marginal zero per consulta. El factor realment decisiu, però, és la privacitat.

Dependència externa. Una interrupció o canvi de preus d'OpenAI paralitzaria el sistema. Amb Ollama [22], el sistema és completament autònom.

Ollama permet executar models LLM localment amb una sola comanda (*ollama run llama3.2*), sense configuració de servidors, i s'ha convertit en l'estàndard *de facto* per a desplegaments LLM privats en entorns corporatius.

A.7 Decisions tecnològiques

Python pur davant LangChain i altres frameworks. Tot i que LangChain és el *framework* més popular per a la construcció de sistemes RAG, s'ha optat per una implementació en Python pur per garantir el control total sobre el codi, eliminar dependències innecessàries (LangChain té més de 200 dependències transitives) i assegurar l'estabilitat a llarg termini davant canvis freqüents d'API entre versions.

Arquitectura local davant solucions SaaS. Es van avaluar solucions com Agentforce de Salesforce, descartades per tres motius principals: absència de suport per a context extern (*web scraping*, notícies), cost de llicència insostenible a escala (Agentforce a 2$/conversa [7]; vegeu l'anàlisi de sensibilitat de costos a l'Annex A.18) i dependència d'un proveïdor extern que limita la governança de les dades.

SQL Server davant Elasticsearch per a la cerca lexical. Marlex ja disposa d'un servidor SQL Server en producció, cosa que elimina qualsevol cost d'implantació addicional. Per al volum actual (centenars de dossiers i milers de notícies), la cerca amb `LIKE` i puntuació TF ofereix temps de resposta inferiors a 200 ms, suficients per al cas d'ús. Elasticsearch aportaria valor diferencial a partir de milions de documents.

A.8 Justificació de l'algorisme RRF davant altres mètodes de fusió

Quan es combinen resultats de dues fonts heterogènies: cerca semàntica (ChromaDB, puntuació de similitud cosinus entre 0 i 1) i cerca lexical (SQL Server, puntuació TF sense rang fix) sorgeix un problema fonamental: les escales de puntuació no són comparables. Un document amb similitud cosinus de 0,85 i un document amb puntuació TF de 12 no es poden sumar directament sense normalitzar, i la normalització introdueix biaixos depenent de la distribució de cada font [26].

Es va avaluar la suma ponderada de puntuacions $\text{score}_{\text{final}} = \alpha \cdot \text{score}_{\text{sem}} + \beta \cdot \text{score}_{\text{lex}}$, però es va descartar perquè requereix calibrar manualment els pesos α i β per a cada tipus de pregunta, i és sensible a valors atípics d'una de les fonts.

La fórmula RRF [24], en canvi, treballa exclusivament amb rangs i no amb puntuacions absolutes:

$$RRFscore(d \in D) = \sum_{r \in R} \frac{1}{k + r(d)}$$

Fig. 8: Fórmula de l'algorisme Reciprocal Rank Fusion (RRF)

on k = 60 és la constant estàndard de la literatura. Aquesta fórmula és robusta per tres motius: és independent de l'escala de puntuació de cada font; penalitza els documents que apareixen en posicions baixes, independentment de la seva puntuació absoluta; i premia els documents que apareixen en ambdues fonts simultàniament, que és exactament el comportament desitjat en un sistema híbrid.

A.9 Justificació de l'arquitectura de cerca híbrida davant modalitats individuals

La decisió d'implementar una cerca híbrida es fonamenta en les limitacions de cada modalitat:

Cerca lexical sola. Precisa amb termes exactes, però falla en preguntes semàntiques ("quina és l'estratègia de creixement?") sense coincidència de paraules clau. **P@3 = 0,205.**

Cerca semàntica sola. Entén el significat independentment del vocabulari, però pot retornar fragments temàticament propers però no factualment rellevants en preguntes relacionals del CRM ("quantes TT actives té aquesta empresa?"). **P@3 = 0,551.**

Cerca híbrida (RRF). Combina la precisió lexical amb la comprensió semàntica dels *embeddings* [26]. Un document que apareix en ambdues llistes puja significativament en el rànquing final. **P@3 = 0,538** ($+162\%$ respecte a la lexical).

L'avaluació sobre 26 preguntes de tres empreses i tres tipologies confirma empíricament la superioritat de l'enfocament híbrid.

A.10 Justificació de la selecció de llama3.2 com a model de generació

Per a la fase de generació de respostes es van avaluar diverses opcions de models locals disponibles a través d'Ollama [22] (vegeu Taula A.15).

S'ha triat llama3.2 [28] (versió de 3B de paràmetres) per l'equilibri òptim entre qualitat i velocitat en el maquinari disponible. El model entén i genera text en català i castellà amb una qualitat suficient per sintetitzar dossiers comercials. La versió de 8B oferiria major qualitat però amb temps de resposta 3 a 4 vegades superiors, cosa que penalitza l'experiència del perfil Hunter. La versió d'1B no assoleix la qualitat mínima per a respostes coherents en contextos multilingües.

A.11 Justificació de la selecció de bge-m3 com a model d'embeddings i migració des de nomic-embed-text

Inicialment es va emprar *nomic-embed-text* (768 dimensions), però es va adoptar *bge-m3* [36] per tres raons: és

un model multilingüe (BAAI) entrenat específicament per a tasques de *retrieval*; les seves 1.024 dimensions ofereixen una representació més rica; i gestiona robustament el bilingüisme català-castellà, clau per a aquest corpus. L'impacte va ser decisiu: P@3 semàntic de 0,256 a 0,551, i híbrid de 0,410 a 0,538. El model *mxbai-embed-large*, tot i la seva qualitat, presentava un temps d'indexació superior sense millora proporcionada.

A.12 Selecció de les mètriques d'avaluació i disseny del dataset

Elecció de la mètrica Es van considerar diverses mètriques d'Information Retrieval [26]: Precision@k (proporció de documents rellevants entre els k primers), Recall@k (menys informativa quan el nombre de rellevants per pregunta varia), MRR (valora la posició del primer resultat rellevant) i NDCG (pondera per posició però requereix gradació manual de rellevància). S'ha triat Precision@3 i Precision@5 per la seva interpretabilitat directa i perquè el Hunter rep un resum basat en els 3 a 5 fragments principals.

Disseny del dataset d'avaluació El dataset (*eval_dataset.json*) conté 26 preguntes distribuïdes en 3 empreses (Bonàrea, Europastry i Mercadona) i 3 tipologies: factual (15), estratègica (6) i relacional (5). Per a cada pregunta es van definir manualment els *chunks* de referència (*ground truth*) [30], mètode estàndard en avaluació RAG quan no es disposa de dades de referència públiques per al domini.

A.13 Evolució iterativa del sistema: de v1 a v4

El desenvolupament ha seguit un cicle iteratiu de quatre versions, cadascuna motivada per l'anàlisi quantitativa de l'anterior.

Versió 1: Baseline híbrid amb *score fusion*. Combinava puntuacions brutes de ChromaDB (cosinus, 0–1) i SQL Server (*hits*, 0–N) mitjançant suma ponderada. Les dues escales no eren comparables, fent les puntuacions difícils d'optimitzar.

Versió 2: RRF, filtre *zero-hits* i dataset ampliat. Es va substituir la suma per RRF [24] i s'hi va afegir un filtre de zero coincidències lexicals. El filtre era correcte per a *[dbo].[noticias]*, però aplicat incorrectament a *[dbo].[dossiers]*: el dossier és en castellà (*lista naranja*) mentre les preguntes relacionals són en català (*llista taronja*), generant 0 *hits* i fent caure les preguntes relacionals a P@3 = 0,0.

Versió 2.1: Correcció del *mismatch* lingüístic. Es va eliminar el filtre *zero-hits* del dossier, s'hi va afegir una puntuació base de $+0{,}5$ i es va introduir un *recall* semàntic adaptatiu ($n \times 1$ per a preguntes relacionals, $n \times 4$ per a la resta).

Versió 3: *upgrade* de l'embedder. Substitució de *nomic-embed-text* (768d) per *bge-m3* (1.024d, multilingüe), elevant P@3 semàntic de 0,256 a 0,551.

Versió 4: correcció del *bug* de fusió. La clau d'identitat dels *chunks* pels primers 80 caràcters fusionava subfragments d'una mateixa secció a l'RRF, inflant fragments irrellevants. Corregit usant un *hash* del text complet, recuperant P@3 híbrid fins a 0,538.

Experiment descartat: *re-ranking* amb LLM. El *re-ranking* posterior a RRF amb llama3.2 (3B) va degradar P@3 de 0,538 a 0,372: un model de 3B actua com avaluador massa sorollós. El *re-ranking* requereix un model més capaç, incompatible amb les restriccions de latència i processament local.

TAULA 1: EVOLUCIÓ DEL RENDIMENT ENTRE VERSIONS

Versió	Canvi principal	P@3 híbrid
v1	Suma ponderada (*score fusion*)	0,417*
v2	RRF + filtre *zero-hits*	0,346
v2.1	Correcció *mismatch* + *recall* adaptatiu	0,423
v3	Embedder *bge-m3* (1.024d)	0,551
v4	Correcció *bug* de fusió (clau de *chunk*)	0,538
v4+RR	*Re-ranking* amb llama3.2 (descartat)	0,372

* No comparable directament: mesurat sobre 20 preguntes i 2 empreses.

A.14 Anàlisi d'errors preguntes amb Precision@3 = 0

L'anàlisi de les preguntes amb pitjor rendiment permet identificar limitacions estructurals i guiar millores futures. Les preguntes s'identifiquen amb la nomenclatura *[empresa]-[número]*, on *ba* correspon a Bonàrea, *ep* a Europastry i *me* a Mercadona.

Preguntes relacionals amb dossier en castellà (resolt a v2.1): Les preguntes *ep-08* (*"Quantes TT actives té Europastry?"*), *ep-11* (*"Quin marge ha generat Europastry l'últim any?"*) i *ba-09* (*"Quina és la relació comercial actual de Bonàrea amb Marlex?"*) obtenien P@3 = 0,0 a v2 a causa d'un *mismatch* lingüístic: els dossiers del CRM estan en castellà mentre les preguntes es formulaven en català, generant zero coincidències lexicals. Corregit a v2.1.

Preguntes factualment específiques no cobertes: Algunes preguntes fan referència a dades que no apareixen als dossiers (expansió geogràfica de Bonàrea, xifres financeres d'Europastry no publicades). Cap estratègia pot retornar el fragment si no existeix al corpus; la solució és ampliar les fonts d'ingesta (informes anuals, registres mercantils, LinkedIn).

Preguntes estratègiques amb respostes disperses: Les preguntes sobre estratègia tenen respostes distribuïdes en múltiples seccions, no concentrades en un sol fragment. Precision@3 penalitza el sistema en aquests casos; una possible millora és implementar *multi-document summarization*.

A.15 Taules de resultats

TAULA 2: RESULTATS GLOBALS DE LA RECUPERACIÓ PER MODE DE CERCA

Mode de cerca	P@3	P@5
Lexical únicament (SQL Server)	0,205	0,200
Semàntic únicament (ChromaDB)	0,551	0,431
Híbrid RRF: sistema actual	0,538	0,454

TAULA 3: RESULTATS DESGLOSSATS PER TIPOLOGIA

Tipologia	Lex. P@3	Sem. P@3	Híb. P@3
Factual	0,222	0,578	0,556
Estratègica	0,167	0,667	0,611
Relacional	0,200	0,333	0,400

TAULA 4: MODELS DE GENERACIÓ LOCALS (OLLAMA)

Model	Mida	Qual.	Vel.	ML
llama3.2:1b	1,3 GB	Baixa	Molt alta	Parc.
llama3.2 (3B)	2,0 GB	Bona	Alta	Sí
llama3.1:8b	4,7 GB	Alta	Moderada	Sí
mistral	4,1 GB	Alta	Moderada	Sí

TAULA 5: COMPARATIVA DE MODELS D'EMBEDDINGS LOCALS (OLLAMA)

Model	Dim.	Qualitat sem.	Velocitat	Multilingüe
all-MiniLM-L6-v2	384	Moderada	Molt alta	Limitada
nomic-embed-text	768	Alta	Alta	Sí
bge-m3 ← sel.	1024	Molt alta	Mitjana	Sí
mxbai-embed-large	1.024	Molt alta	Moderada	Sí
text-emb-3-small (OAI)	1.536	Molt alta	Alta	Sí

TAULA 6: RESULTATS DE LA QUALITAT DE LA GENERACIÓ (MODE HÍBRID, VIA LLM-JUTGE)

Mètrica	Valor	Descripció
Faithfulness	0,599	Proporció d'afirmacions fonamentades en els fragments recuperats
Answer Relevance	0,538	Pertinència de la resposta respecte a la pregunta
Correctness (GT)	0,523	Coincidència amb la resposta de referència (*ground truth*)

A.16 Esquema de les taules SQL Server

El sistema persisteix la informació en tres taules a SQL Server per a la cerca lexical i l'emmagatzematge dels dossiers.

Taula *[dbo].[noticias]*: Notícies de premsa per empresa. Camps principals: *noticia_id* (hash per deduplicació), *empresa_slug* i *empresa_nombre* (claus de cerca), *titulo* (indexat per TF), *fuente*, *fecha_publicacion* (filtre 90 dies), *url* (traçabilitat) i *resumen* (text netejat).

	noticia_id	empresa_slug	empresa_nombre	titulo	fuente	fecha_publicacion	url	resumen
1	073e6f8828784f0f	mercadona	mercadona	Mercadona: Las Coliflores que molaron al súper tradici...	americaeconomia-mailx.com	2026-04-23	https://news.google.com/rss/articles/CBMiekFBVVV5...	Mercadona: Las Coliflores que molaron al súper tradici...
2	0da7fa0f413d3f41	europastry	Europastry	MCH lanza su tercer fondo de agricultura y alimentació...	El Confidencial	2026-05-26	https://news.google.com/rss/articles/CBMiqwFBVVV5...	MCH lanza su tercer fondo de agricultura y alimentació...
3	1433496fdb0f8fb0	esclat	esclat	Abre en Badalona un supermercado Esclat de 1.626 m...	recidenciadigital.cat	2026-03-25	https://news.google.com/rss/articles/CBMiq2wFBVV5...	Abre en Badalona un supermercado Esclat de 1.626 m...
4	150e6715de9f59f5	mercadona	mercadona	Qué supermercados abren en Semana Santa 2026: ho...	Libertad Digital	2026-04-02	https://news.google.com/rss/articles/CBMiiAFBVVV5...	Qué supermercados abren en Semana Santa 2026: ho...
5	1e3ce1c5da724c1	bonarea	bonarea	bonÀrea Corporación factura 2.626 millones de euros e...	Corresponsables	2026-04-14	https://news.google.com/rss/articles/CBMiSkFBVV5...	bonÀrea Corporación factura 2.626 millones de euros e...
6	1f9b778c81e3d9f4	bonarea	bonarea	BonÀrea estrada preparando el relleno en su presidencia...	Segre.com	2026-03-31	https://news.google.com/rss/articles/CBMiaUFBVV5...	BonÀrea estrada preparando el relleno en su presidencia...
7	29771a3017b04901	esclat	esclat	Los clientes del Grupo Bon Preu ahorraron 227 millone...	recidenciadigital.cat	2026-02-27	https://news.google.com/rss/articles/CBMiq2FBVV5...	Los clientes del Grupo Bon Preu ahorraron 227 millone...
8	3b9aa3f9d2014456	bonarea	bonarea	BonÀrea supera los 102 ME de beneficio el 2025, un 15...	Diari AVUI	2026-04-14	https://news.google.com/rss/articles/CBMiqgFBVV5...	BonÀrea supera los 102 ME de beneficio el 2025, un 15...
9	3bc556d5518490fd	bonarea	bonarea	bonÀrea trasteja y amplía su primera tienda en Badal...	Corresponsables	2026-03-26	https://news.google.com/rss/articles/CBMiqFBVV5...	bonÀrea trasteja y amplía su primera tienda en Badal...
10	4038d5154b9de315	mercadona	mercadona	Del retail español a Portugal: Mercadona despega en el...	AnexoAffalle & Retail	2026-03-26	https://news.google.com/rss/articles/CBMiqgFBVV5...	Del retail español a Portugal: Mercadona despega en el...
11	4adefa01afbcb3e1	mercadona	mercadona	Responsabilidad e innovación: Mercadona y Lidl marca...	Directivos y Gerentes	2026-02-12	https://news.google.com/rss/articles/CBMiiwFBVV5...	Responsabilidad e innovación: Mercadona y Lidl marca...
12	52d5c5bcf4d6bf7c	bonarea	bonarea	BonÀrea bate su récord de beneficios con 102,7 millon...	El Món	2026-04-14	https://news.google.com/rss/articles/CBMiiwFBVV5...	BonÀrea bate su récord de beneficios con 102,7 millon...
13	537ab0d547ecaf39	mercadona	mercadona	Mercadona roza el 30% de cuota de mercado en 2025,...	FRS Food Retail & Service	2026-01-27	https://news.google.com/rss/articles/CBMiiwFBVV5...	Mercadona roza el 30% de cuota de mercado en 2025,...
14	54c72b0f0fc19ec9b	bonarea	bonarea	BonÀrea factura 2.626 millones (+5,4%) en 2025 y reba...	FRS Food Retail & Service	2026-04-14	https://news.google.com/rss/articles/CBMiiAFBVV5...	BonÀrea factura 2.626 millones (+5,4%) en 2025 y reba...
15	551bb5072c3e1066	bonarea	bonarea	bonÀrea supera los 100 millones de beneficio en 2025 -...	Sweetpress	2026-04-14	https://news.google.com/rss/articles/CBMiqgFBVV5...	bonÀrea supera los 100 millones de beneficio en 2025 -...

Fig. 9: Taula *[dbo].[noticias]*: notícies indexades per empresa

Taula *[dbo].[dossiers]*: Fragments individuals (*chunks*) del dossier de cada prospecte. Camps principals: *dossier_id* (hash antiduplcats), *empresa_slug/nombre*, *seccio* (tipus de contingut), *contingut* (text per cerca lexical i vectorial), *marlex_score* (0–10), *sector*, *generado_en* i *fecha_carga*.

	dossier_id	empresa_slug	empresa_nombre	seccio	contingut	marlex_score
22	4b22bcc4fe7354c1	consum	consum	RESUM	Dossier Comercial: consum _Generado: 26/04/2026 10:53 | M...	2
23	4cde7c15de6808c1	mango	mango	Certificaciones y Normativas	_No detectadas_	0
24	4d0415c922998ea0	condis	condis	RESUM	Dossier Comercial: condis _Generado: 25/04/2026 18:45 | Ma...	5
25	4ffac8ff3eb5dac7	action	action	Noticias Recientes (últimos 90 días)	### 1. El live-action de One Piece se enfrenta a una prueba d...	0
26	5d8ff1e72f2443f	bonarea	bonarea	RESUM	Dossier Comercial: bonarea _Generado: 27/04/2026 08:39 | ...	0
27	6082ae6c4b3e9c70	tradeinn	tradeinn	CONTACTO	_Persona de referencia para la reunión: cargo, trayectoria y da...	4
28	657a912479f66971	mango	mango	Ficha Rápida	| Campo | Valor | |---|---| | **URL** | https://shop.mango.com/e...	0
29	6ac24e1d85a2ba7b	esclat	esclat	INFO	**Sector.** tecnología **Fundación.** 2026 **Web.** https://w...	2
30	6f57020f5aabbed75	bonarea	bonarea	Productos y Servicios	Comprar Productos innovadores bonÀrea compra online | Dire...	0
31	6f8746d93aa9539f	mercadona	mercadona	INFO EMPRESA	_Descripción corporativa, actividad principal y posicionamiento...	0

Fig. 10: Taula *[dbo].[dossiers]*: fragments per prospecte

Taula *[dbo].[dossier_cache]*: Dossier complet consolidat per empresa en Markdown. Camps principals: *empresa_slug/nombre* (claus primàries), *dossier_md* (document complet per al LLM), *marlex_score*, *sector*, *pages_crawled*, *chunks_indexats*, *generat_en* i *fecha_carga*.

	empresa_slug	empresa_nombre	dossier_md	marlex_score	sector	pages_crawled	chunks_indexats	generat_en	fecha_carga
1	europastry	Europastry	# PREPARACIÓN REUNIÓN: EUROPASTRY _Generado: 20/...	6	alimentacion	20	12	2026-05-20 22:54:39.2501170	2026-05-20 22:54:39.3100000
2	mercadona	mercadona	# PREPARACIÓN REUNIÓN: MERCADONA _Generado: 21/0...	0	merceria	20	7	2026-05-21 11:34:01.7278330	2026-05-21 11:34:01.7833333
3	selectil	selectil	# PREPARACIÓN REUNIÓN: SELVAFIL _Generado: 20/05/20...	6	textil	20	12	2026-05-20 17:04:38.9780260	2026-05-20 17:04:39.0633333

Fig. 11: Taula *[dbo].[dossier_cache]*: dossier consolidat per empresa

A.17 Implementació de l'LLM-jutge per a l'avaluació de la generació

Per avaluar la qualitat de les respostes generades es va considerar el framework RAGAS [30], però en la seva configuració per defecte depèn de l'API d'OpenAI per al model jutge, fet que resulta incompatible amb el principi de privacitat i cost zero que regeix tota l'arquitectura. Per coherència, es va implementar un LLM-jutge propi que utilitza el mateix model local (llama3.2 [28]) via Ollama [22], sense cap mena de dependència externa.

El procediment és el següent: per a cadascuna de les 26 preguntes del dataset d'avaluació, el model jutge rep una sèrie d'instruccions estructurades i retorna una puntuació decimal entre 0,0 i 1,0 per a cada mètrica:

- **Faithfulness:** El jutge rep el context recuperat i la resposta generada, i puntua fins a quin punt la informació està suportada pel context (1,0 = completament suportat; 0,0 = presència d'al·lucinacions).
- **Answer Relevance:** El jutge rep la pregunta i la resposta, i puntua fins a quin punt la resposta respon adequadament la pregunta formulada (1,0 = resposta completa; 0,0 = no respon).

- **Correctness:** El jutge compara la resposta generada amb la resposta de referència (ground truth), detectant errors factuals no capturats pel Faithfulness.

Els valors obtinguts es presenten a la Taula 6: Faithfulness 0,599, Answer Relevance 0,538 i Correctness 0,523.

Cal assenyalar dues limitacions d'aquest enfocament. Primer, el model jutge (llama3.2, 3B) és el mateix que genera les respostes, fet que pot introduir un biaix d'autoavaluació. Segon, un model de només 3B de paràmetres té una capacitat limitada com a avaluador. Per mitigar-ho parcialment, el jutge opera amb instruccions estructurades i verificació contra el ground truth. Tot i així, les puntuacions s'han d'interpretar com una línia base orientativa i no com una mesura absoluta: un model jutge independent i de major mida proporcionaria valors més fiables.

A.18 Anàlisi de sensibilitat de costos segons el volum d'ús

Per avaluar l'eficiència econòmica de les solucions segons el volum d'execucions, es comparen els tres models de desplegament possibles: la solució local implementada (Ollama), una API de generació al núvol (OpenAI gpt-4o-mini) i una solució SaaS integrada al CRM amb cost per conversa (Agentforce [7]).

Supòsits del càlcul: xarxa de 200 consultors; per consulta, ∼3.000 tokens d'entrada (context recuperat + prompt) i ∼300 de sortida; preus gpt-4o-mini de 0,15 $/1M tokens d'entrada i 0,60 $/1M de sortida [16]; Agentforce a 2 $/conversa [7]; cost local estimat com l'electricitat d'un servidor de ∼400 W en funcionament continu (∼40 /mes, infraestructura preexistent a Marlex); conversió 1 $ ≈ 0,92 .

TAULA 7: COST MENSUAL ESTIMAT SEGONS EL VOLUM DE CONSULTES

Consultes/mes	Local (Ollama)	gpt-4o-mini	Agentforce
10.000	∼40 (fix)	∼6	∼18.400
50.000	∼40 (fix)	∼29	∼92.000
100.000	∼40 (fix)	∼58	∼184.000

L'anàlisi mostra dos comportaments clarament diferenciats. El cost de la solució local és **pràcticament pla**: en operar amb maquinari propi, el cost marginal per consulta és nul i només depèn de l'electricitat, de manera que el sistema escala sense increment de cost mentre el maquinari disponible absorbeixi la càrrega. En canvi, les solucions de pagament per ús creixen **linealment amb el volum**: el cas d'Agentforce, amb un cost per conversa, esdevé econòmicament insostenible a escala (desenes o centenars de milers d'euros mensuals per a la xarxa comercial completa).

Cal matisar que el cost per token de l'API de generació (gpt-4o-mini) és modest fins i tot a volums alts; el factor que descarta aquesta opció no és, doncs, el cost, sinó la **privacitat de les dades** i la dependència d'un proveïdor extern (vegeu Annex A.6). La conclusió de l'anàlisi de sensibilitat és que la solució local ofereix el millor perfil cost-escalabilitat *i*, simultàniament, la màxima garantia de governança i propietat de les dades.

A.19 Exemple de dossier comercial generat pel sistema

Les imatges següents mostren un exemple complet de dossier comercial generat pel sistema per a l'empresa fictícia Càrniques del Pirineu, S.A., amb dades fictícies amb finalitat il·lustrativa. A més, s'inclou un exemple del mode interactiu de consulta, on el *Hunter* pot formular preguntes en llenguatge natural i obtenir respostes immediates amb traçabilitat de les fonts.

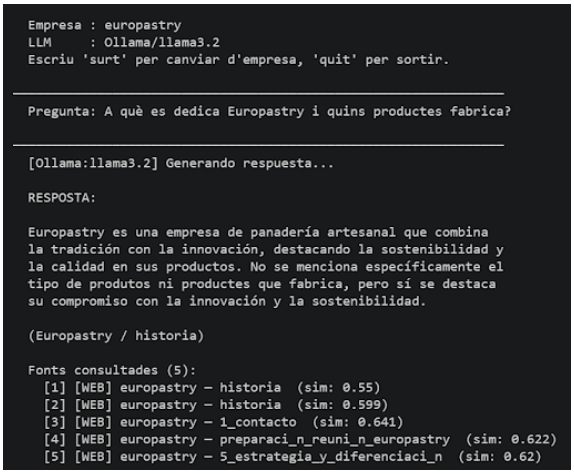
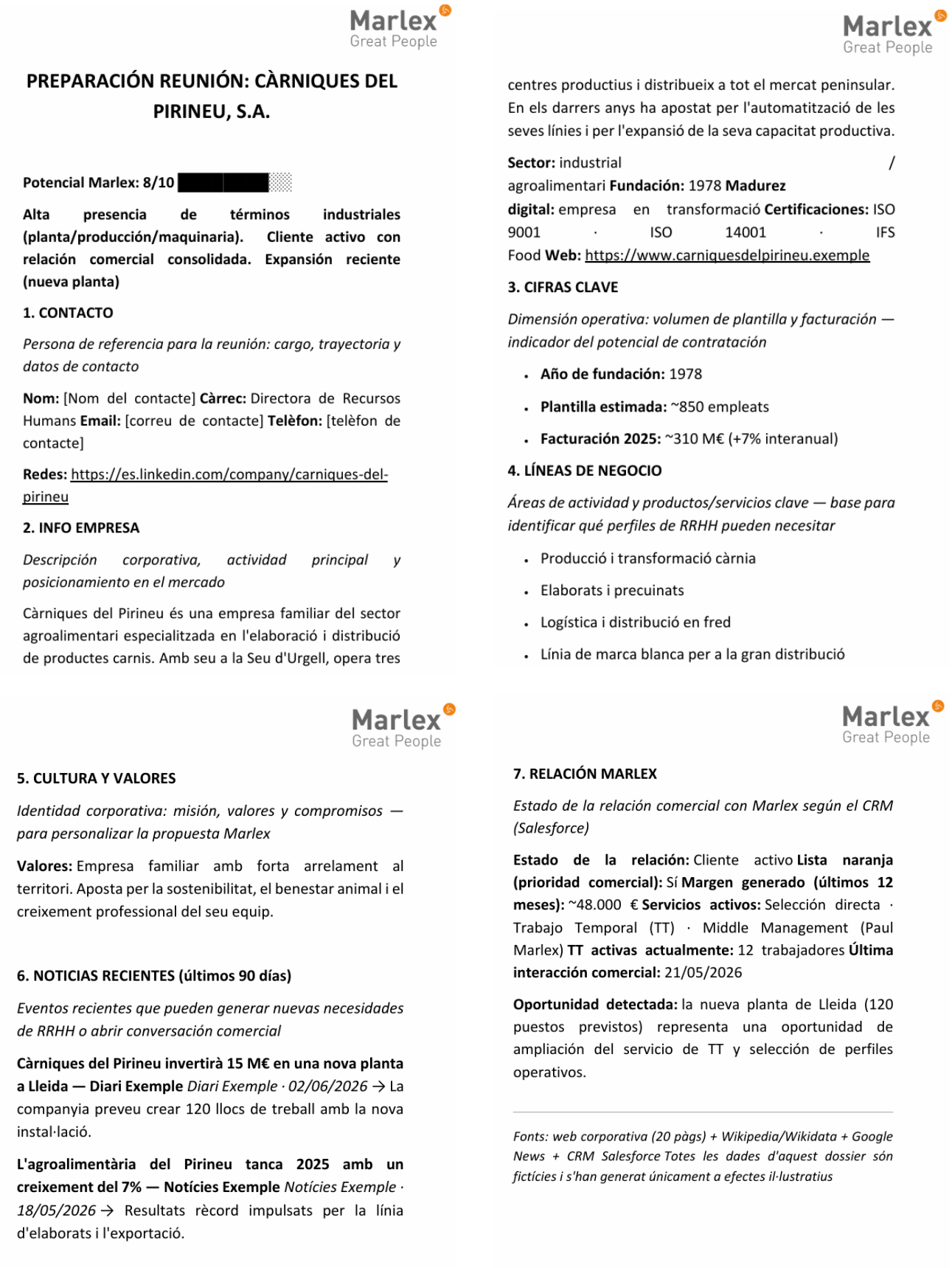


Fig. 12: Exemple de consulta al sistema en mode interactiu. Es mostra la pregunta, la resposta generada per llama3.2 i la traçabilitat de les fonts utilitzades.



Marlex
Great People

PREPARACIÓN REUNIÓN: CÀRNIQUES DEL PIRINEU, S.A.

Potencial Marlex: 8/10 [redacted]

Alta presencia de términos industriales (planta/producción/maquinaria). Cliente activo con relación comercial consolidada. Expansión reciente (nueva planta)

1. CONTACTO

Persona de referencia para la reunión: cargo, trayectoria y datos de contacto

Nom: [Nom del contacte] **Càrrec:** Directora de Recursos Humans **Email:** [correu de contacte] **Telèfon:** [telèfon de contacte]

Redes: https://es.linkedin.com/company/carniques-del-pirineu

2. INFO EMPRESA

Descripción corporativa, actividad principal y posicionamiento en el mercado

Càrniques del Pirineu és una empresa familiar del sector agroalimentari especialitzada en l'elaboració i distribució de productes carnis. Amb seu a la Seu d'Urgell, opera tres centres productius i distribueix a tot el mercat peninsular. En els darrers anys ha apostat per l'automatització de les seves línies i per l'expansió de la seva capacitat productiva.

Sector: industrial / agroalimentari **Fundación:** 1978 **Madurez digital:** empresa en transformació **Certificaciones:** ISO 9001 · ISO 14001 · IFS Food **Web:** https://www.carniquesdelpirineu.exemple

3. CIFRAS CLAVE

Dimensión operativa: volumen de plantilla y facturación — indicador del potencial de contratación

- **Año de fundación:** 1978
- **Plantilla estimada:** ~850 empleats
- **Facturación 2025:** ~310 M€ (+7% interanual)

4. LÍNEAS DE NEGOCIO

Áreas de actividad y productos/servicios clave — base para identificar qué perfiles de RRHH pueden necesitar

- Producció i transformació càrnia
- Elaborats i precuinats
- Logística i distribució en fred
- Línia de marca blanca per a la gran distribució

5. CULTURA Y VALORES

Identidad corporativa: misión, valores y compromisos — para personalizar la propuesta Marlex

Valores: Empresa familiar amb forta arrelament al territori. Aposta per la sostenibilitat, el benestar animal i el creixement professional del seu equip.

6. NOTICIAS RECIENTES (últimos 90 días)

Eventos recientes que pueden generar nuevas necesidades de RRHH o abrir conversación comercial

Càrniques del Pirineu invertirà 15 M€ en una nova planta a Lleida — Diari Exemple *Diari Exemple · 02/06/2026* → La companyia preveu crear 120 llocs de treball amb la nova instal·lació.

L'agroalimentària del Pirineu tanca 2025 amb un creixement del 7% — Notícies Exemple *Notícies Exemple · 18/05/2026* → Resultats rècord impulsats per la línia d'elaborats i l'exportació.

7. RELACIÓN MARLEX

Estado de la relación comercial con Marlex según el CRM (Salesforce)

Estado de la relación: Cliente activo **Lista naranja (prioridad comercial):** Sí **Margen generado (últimos 12 meses):** ~48.000 € **Servicios activos:** Selección directa · Trabajo Temporal (TT) · Middle Management (Paul Marlex) **TT activas actualmente:** 12 trabajadores **Última interacción comercial:** 21/05/2026

Oportunidad detectada: la nueva planta de Lleida (120 puestos previstos) representa una oportunidad de ampliación del servicio de TT y selección de perfiles operativos.

Fonts: web corporativa (20 pàgs) + Wikipedia/Wikidata + Google News + CRM Salesforce Totes les dades d'aquest dossier són fictícies i s'han generat únicament a efectes il·lustratius

Fig. 13: Exemple de dossier comercial generat pel sistema (dades fictícies amb finalitat il·lustrativa)