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

Parareda Casablanças, Oriol; Ortega Gil, Marc, dir. Desenvolupament d'un xat bot basat en IA generativa amb integració en un lloc web. 2024. (Enginyeria Informàtica)

This version is available at <https://ddd.uab.cat/record/298937>

under the terms of the  license

Desenvolupament d'un xatbot basat en IA generativa amb integració en un lloc web

Oriol Parareda Casablanacas

1 de juliol de 2024

Resum– Aquest treball desenvolupa un API REST per integrar un xat dinàmic amb IA de text generatiu en pàgines web. L'API permet entrenar la IA amb contingut personalitzat o recursos especialitzats, adaptant-se a necessitats específiques. Inclou capacitat multilingüe per a una aplicació global. El backend es desenvolupa en Convex, una nova eina BaaS que s'encarrega de facilitar l'accés a les dades a temps real, mentre que el frontend està fet en React. Es prioritza la seguretat amb mesures per protegir dades i garantir comunicacions segures. El projecte també ofereix una pàgina de mostra i documentació detallada, facilitant la integració de l'API. Aquest treball demostra el potencial de les IA generatives en la millora de la interacció digital.

Paraules clau– Intel·ligència artificial (IA), API REST, Peticions asíncrones, Gestió d'usuaris, Xatbot de text generatiu, Scrumban, ChatGPT, React, Convex, Backend as a Service(BaaS).

Abstract– This project's aim is to develop a REST API to integrate a dynamic chat with generative text AI into web pages. The API allows training the AI with customized content or specialized resources, adapting to specific needs. It includes multilingual capabilities for global application. The backend is developed over Convex, a cutting edge BaaS that allows an easy access to the stored data in real time, the frontend will be developed in React. Security is prioritized with measures to protect data and ensure secure communications. The project also provides a sample web page and detailed documentation to facilitate API integration. This work demonstrates the potential of generative AI in enhancing digital interaction.

Keywords– Artificial Intelligence (AI), API REST, Asynchronous Requests, User Management, Generative Text Chatbot, Scrumban, ChatGPT, React, Convex, Backend as a Service(BaaS).



1 INTRODUCCIÓ - CONTEXT DEL TREBALL

AQUESTA secció presenta el context i les premisses del treball juntament amb observacions i propòsits. Actualment l'ús de les Intel·ligències Artificials (IA) és present en molts camps del nostre dia a dia, des de programes de text generatiu com seria ChatGPT fins a estris de cuina que aprenen punts de cocció per facilitar la cuina. Tenint en compte aquestes eines i pensant en aplicar-les a més àmbits, l'objectiu d'aquest treball serà la creació d'un API REST que permeti incorporar un xat dinàmic amb una IA de text generatiu en qualsevol pàgina web.

La idea principal serà, doncs, la creació d'aquesta API REST que ens permeti incorporar el xat. A més, ja que no totes les webs necessitaran el mateix nivell d'entrenament de la IA, s'afegirà la possibilitat d'entrenar la IA generati-

va a partir de contingut propi o de portals web especialitzats en entrenament d'IA. Aquesta flexibilitat permetrà que empreses i individus puguin personalitzar l'experiència del xat segons les seves necessitats específiques, assegurant una major eficàcia i rellevància en les respostes generades per la IA.[2]

Un problema detectat és l'idioma del xat. Tot i això, un dels propòsits del treball és poder aplicar aquest xat de text generatiu en qualsevol idioma en el que pugui ser entrenat. Això implicarà la integració de capacitats multilingües a la IA, utilitzant bases de dades de text i recursos lingüístics per a garantir que la IA pugui operar eficaçment en diversos idiomes, proporcionant així una solució veritablement global.

La proposta, doncs, consistirà en la creació d'una part de backend, un frontend, una pàgina web de mostra i una documentació a mode de tutorial, encara per determinar de quin estil. El backend es desenvoluparà en Convex. La part visual del programa es desenvoluparà en React ja que és un framework dels més utilitzats i que permet millor accessibilitat desde Convex.

El sistema ha de ser segur i fiable, per tant serà necessari implementar mètodes de seguretat en la connexió entre el

- E-mail de contacte: 1423671@autonoma.es
- Menció realitzada: Tecnologies de la Informació
- Treball tutoritzat per: Marc Ortega Gil (DEIC)
- Curs 2023/24

backend i el frontend assegurant que no hi ha cap filtració de dades al ser una API REST d'ús públic. Això inclourà l'ús de protocols segurs com HTTPS, autenticació d'usuaris i altres mesures de seguretat per protegir la privadesa i la integritat de les dades dels usuaris. A més, es realitzaran proves exhaustives per assegurar que el sistema compleix amb els estàndards de seguretat i funcionalitat necessaris per a una aplicació d'aquesta naturalesa.[3]

2 OBJECTIUS

En aquest apartat s'especificaran els objectius del projecte. Els objectius específics són els següents:

A Desenvolupar una API REST:

- Crear una API REST robusta que serveixi com a interfície per a la interacció amb el xat bot de text generatiu guiat per una IA.
- L'API ha de ser capaç de gestionar sol·licituds d'entrada de text dels usuaris i retornar respostes generades per la IA.
- Garantir que l'API sigui segura, escalable i fàcil d'integrar amb altres sistemes i aplicacions.

B Crear una base de dades d'usuaris:

- Desenvolupar una base de dades per emmagatzemar informació dels usuaris, incloent dades de registre, historial de converses i preferències.
- Implementar funcionalitats de registre i autenticació d'usuaris, assegurant la protecció i privacitat de les dades personals.
- Permetre als usuaris registrar-se, iniciar sessió i gestionar el seu perfil de manera eficient i segura.

C Desenvolupar una interfície web per al xat:

- Crear una interfície web intuïtiva i atractiva per presentar el xat bot en forma de bombolla de diàleg, facilitant la interacció dels usuaris amb el bot.
- Assegurar que la interfície sigui responsive, de manera que sigui accessible i funcional des de diferents dispositius, incloent ordinadors de sobretaula, tauletes i telèfons mòbils.
- Incloure elements visuals i d'UX/UI que millorin l'experiència de l'usuari durant les converses amb el xat bot.

D Desenvolupar una interfície web senzilla per a demostracions:

- Crear una interfície web addicional, senzilla però efectiva, per mostrar com funciona el sistema de xat bot en un cas real.
- Aquesta interfície ha de servir com a eina de demostració i prova, permetent als usuaris veure exemples de com la IA respon a diferents tipus de preguntes i situacions.
- Incloure una secció de documentació i ajuda per guiar els usuaris en l'ús de la interfície i el xat bot.

Aquests objectius proporcionaran una estructura clara i definida per al desenvolupament del projecte, assegurant que cada component es dissenya i implementa de manera coherent i alineada amb les necessitats dels usuaris.

2.1 Requeriments

2.1.1 Requeriments funcionals

R1: Gestió d'usuaris

Aquest requeriment es centra en el control dels usuaris. Per fer que sigui utilitzable per usuaris necessitem implementar:

R1.1: Creació d'usuaris

Aquest requeriment es centra en el control dels usuaris del sistema, assegurant que aquests poden gestionar els seus comptes de manera efectiva.

R1.2: Autenticació

Garantir que els usuaris puguin iniciar sessió de manera segura. Això implica implementar mecanismes d'autenticació, com ara inici de sessió amb correu electrònic i contrasenya.

R1.3: Gestió dels perfils

Permetre als usuaris modificar la seva informació de perfil. Això inclou opcions per canviar la contrasenya, actualitzar dades personals (com nom i correu electrònic), i altres preferències del compte, assegurant que no hi hagi duplicació de comptes.

R2: Gestió de converses

Aquest requeriment assegura que les converses amb la IA siguin accessibles, organitzades i manejables.

R2.1: Crear converses

Permetre als usuaris iniciar noves converses amb la IA. Això implica proporcionar una interfície on els usuaris poden començar una interacció amb la IA generativa.

R2.2: Mantenir converses

Assegurar que les converses puguin continuar de manera coherent. Això vol dir que les interaccions anteriors es guarden i es poden consultar per mantenir la continuïtat del diàleg.

R2.3: Gestionar converses

Proporcionar funcionalitats per iniciar, pausar, aturar i reprendre converses segons les necessitats dels usuaris. Això inclou opcions per guardar l'estat de la conversa, arxivar converses antigues i eliminar les que no es necessitin.

R3: Interfície de comunicació

Crear una interfície d'usuari clara i intuïtiva per a la comunicació amb la IA. Aquesta interfície ha de ser fàcil d'utilitzar, permetent als usuaris interactuar amb la IA de manera natural i eficient, incloent elements visuals com finestres de xat, botons per enviar missatges, i indicadors de resposta.

R4: Integració amb la IA

Assegurar una integració fluida amb el sistema d'IA

generativa. Això implica connectar el front-end (interfície d'usuari) amb el back-end (sistema d'IA) de manera que les sol·licituds d'usuari es processen correctament i les respostes de la IA es mostren en temps real. Aquesta integració ha de ser robusta per garantir que les respostes de la IA siguin ràpides i rellevants.

R5: Gestió d'entrenament de la IA

Aquest requeriment se centra en assegurar que la IA generativa pot ser entrenada i millorada de manera contínua, eficient i segura. La gestió de l'entrenament és crucial per mantenir la qualitat i la rellevància de les respostes de la IA.

R6: Personalització i configuració

Aquest requeriment es centra en proporcionar als usuaris i administradors la capacitat de personalitzar i configurar el comportament del xat amb la IA generativa segons les seves necessitats i preferències.

2.1.2 Requeriments no funcionals

RN1: Rendiment

El rendiment es refereix a la capacitat del sistema per respondre de manera eficient i ràpida a les sol·licituds dels usuaris.

RN2: Disponibilitat

La disponibilitat es refereix a la capacitat del sistema per estar operatiu i accessible als usuaris en tot moment.

RN3: Fiabilitat

La fiabilitat implica que el sistema funcioni de manera consistent i correcte durant un període prolongat.

RN4: Seguretat

La seguretat és fonamental per protegir la informació dels usuaris i prevenir accessos no autoritzats. Això inclou l'ús de mecanismes d'autenticació, xifratge de dades tant en trànsit com en repòs, i mesures per prevenir atacs com ara el phishing, el malware i altres tipus de ciberamenaces.

RN5: Escalabilitat

L'escalabilitat es refereix a la capacitat del sistema per gestionar un augment en la càrrega de treball sense perdre rendiment.

RN6: Compatibilitat

La compatibilitat fa referència a la capacitat del sistema per funcionar correctament en diversos entorns i amb diferents dispositius i sistemes operatius. Això implica que el xat amb IA ha de ser accessible tant des de dispositius mòbils com des d'ordinadors de sobretaula, i ha de suportar múltiples navegadors web i aplicacions.

RN7: Accessibilitat

L'accessibilitat es refereix a la capacitat del sistema per ser utilitzat per tothom, incloses persones amb discapacitats. Això implica complir amb estàndards d'accessibilitat i proporcionar suport per a tecnologies assistives com lectors de pantalla.

RN8: Monitoratge i registre

El monitoratge i el registre són importants per mantenir el sistema i detectar problemes. Això implica tenir eines per supervisar el rendiment, l'ús i la seguretat del sistema en temps real, així com per registrar les activitats per a anàlisi posterior i compliment normatiu.

RN9: Mantenibilitat

La mantenibilitat es refereix a la facilitat amb què el sistema pot ser modificat per corregir errors, millorar el rendiment o adaptar-se a noves necessitats. Això implica tenir una arquitectura modular, un codi ben documentat i processos eficaços per a la gestió de canvis.

RN10: Compliment normatiu

El compliment normatiu implica assegurar-se que el sistema compleix totes les lleis i regulacions aplicables, com les de protecció de dades i altres requisits legals rellevants per a la indústria o el context geogràfic on opera. Això inclou la protecció de la privadesa dels usuaris i la transparència en el processament de dades.

3 ESTAT DE L'ART

Actualment existeixen IA que es poden aplicar en qualsevol àmbit del nostre dia a dia. En aquest treball ens centrarem en les IA de text generatiu. Al estar en un moment en el que a nivell publicitari tenir el teu propi bot de text et dona un estatus millor, les grans empreses, sobretot, han creat les seves pròpies IA. A més, se'ls ha demanat a les diferents IAs que creïn una presentació pròpia amb el següent prompt: "Pots presentarte al públic en unes 30 paraules?":

- Google: Gemini, abans anomenat Bard permet generar text, traduir idiomes, escriure contingut creatiu i respondre preguntes de manera informativa. [7]

– Arquitectura

LaMDA està dissenyat específicament per a diàlegs i converses, capturant millor el context i les nuances del llenguatge conversacional.

– Aplicacions:

Gemini s'utilitza principalment en aplicacions de cerca i assistència, integrat amb altres serveis de Google per oferir respostes més precises i contextualment rellevants.

– Autopresentació:

Sóc en Gemini, un model de llenguatge gran creat per Google. Encara estic en desenvolupament, però puc generar text, traduir idiomes, escriure diferents tipus de contingut creatiu i respondre les teves preguntes d'una manera informativa. Estic entrenat en una gran quantitat de dades de text i puc accedir i processar informació del món real a través de la Cerca de Google. El meu objectiu és ajudar els usuaris amb les seves preguntes i sol·licituds al màxim de les meves capacitats, utilitzant els coneixements i la informació en què he estat entrenat.

- Amazon: Amazon Lex és un servei de xatbots ofert per AWS que utilitza les mateixes tecnologies de reconeixement de veu i llenguatge natural que Amazon Alexa. [8]

- Arquitectura

Utilitza models de llenguatge profundament entrenats per gestionar text i veu, proporcionant respostes conversacionals coherents.

- Aplicacions:

Utilitzat principalment en assistència al client, interfícies de veu per a aplicacions, i sistemes d'automatització de serveis.

- IBM Watson Assistant: Watson Assistant és una plataforma de xatbots desenvolupada per IBM, que ofereix eines per crear assistents virtuals personalitzats. [9]

- Arquitectura

Utilitza models de llenguatge basats en l'aprenentatge automàtic i l'aprenentatge profund, amb capacitats avançades d'entrenament i personalització.

- Aplicacions:

Utilitzat en sectors com la salut, els serveis financers, i el comerç al detall per millorar l'experiència del client i l'eficiència operativa.

- Microsoft: aquesta gran empresa del sector de la informàtica presenta dues IA, Azure AI, una IA centrada en empreses amb una idea similar a Amazon Q i Copilot, una IA més centrada en la creació, ja sigui de contingut creatiu. [10]

- Arquitectura

Utilitza models de llenguatge desenvolupats per Microsoft, integrant-se amb el framework de serveis cognitius d'Azure per proporcionar capacitats avançades de processament del llenguatge natural.

- Aplicacions:

S'utilitza en aplicacions empresarials, assistència al client, i serveis de productivitat. Zo és un exemple d'un xatbot desenvolupat amb aquesta plataforma.

- Autopresentació:

Certament! Sóc un assistent d'intel·ligència artificial dissenyat per ajudar-te. Pots preguntar-me sobre qualsevol cosa i faré el possible per proporcionar-te una resposta útil.

- Open.ai: ChatGPT, possiblement una de les més conegudes a escala mediàtica, és capaç de crear contingut des de text a música i ajudar en tasques analítiques o comunicatives. [11]

- Arquitectura

Utilitza l'arquitectura GPT-4, basada en transformers, amb milions de paràmetres per processar i generar text natural de manera coherent i contextual.

- Aplicacions:

Utilitzat en una àmplia varietat d'aplicacions, incloent assistència al client, creació de continguts, tutorització, i més.

- Autopresentació:

Sóc ChatGPT, un assistent virtual desenvolupat per OpenAI. Utilitzo intel·ligència artificial per respondre preguntes, generar text, i ajudar en tasques variades. Estic aquí per oferir suport i informació.

- Meta: tot i que han fet intents de presentar algun assistent a escala dels mencionats anteriorment el que realment té presència és el seu llenguatge de programació específic per fer IA.

4 METODOLOGIA

En aquest apartat es presenten diferents metodologies i s'argumenta el perquè de la metodologia triada. [1]

Scrum: Scrum és la metodologia àgil per excel·lència en el desenvolupament de programari des de principis dels anys noranta. Aquesta metodologia proporciona un marc teòric flexible i adaptable, que permet la implementació de diferents tècniques. La seva principal característica és la definició d'etapes incrementals i flexibles per adaptar-se als canvis que puguin sorgir. Aquestes etapes, anomenades sprint o iteració, tenen una durada fixa d'entre una i quatre setmanes, i al final de cada etapa, el producte ha d'estar llest per a ser testejat.

Kanban: Kanban va aparèixer per primera vegada en els sistemes de producció de Toyota i es basa en un sistema de planificació que permet controlar l'avanç de la producció. Kanban es caracteritza per una taula visual en què els diferents apartats del procés de producció es presenten en una de les següents categories: "Per fer", "En procés", "En espera" i "Finalitzat". Aquesta estructura facilita el control de la situació actual i permet fer una estimació de la situació futura de manera senzilla.

Scrumban: Scrumban és una metodologia que combina les característiques principals de Scrum i Kanban, utilitzant iteracions amb un control visual de les tasques pendents del projecte. Aquesta serà la metodologia emprada en aquest projecte.

La implementació es realitzarà mitjançant ClickUp, una eina especialitzada en taulells Kanban que permet la implementació d'un flux de treball clar i eficient.

5 PLANIFICACIO

En aquesta secció s'explica la planificació de desenvolupament de l'API REST. Tal i com s'ha presentat en l'apartat anterior, la metodologia emprada es basa en iteracions.

- Plantejament:

- anàlisi d'objectius

Definir clarament els objectius del projecte. Identificar les funcions principals que l'API ha de proporcionar.

- anàlisi de requeriments i necessitats

TAULA 1: TAULA DE PLANIFICACIÓ

Fase	Data d'inici	Data de finalització
Plantejament	18/02/2024	10/03/2024
Iteració 1	11/03/2024	21/04/2024
Iteració 2	22/04/2024	26/05/2024
Iteració 3	27/05/2024	16/06/2024
Iteració 4	17/06/2024	29/06/2024
LLiurament	01/07/2024	07/07/2024

Recopilar els requisits funcionals i no funcionals del sistema. Determinar les necessitats tècniques i d'infraestructura per suportar l'API.

- estudi de l'estat de l'art

Investigació de les tecnologies existents i les millors pràctiques en el desenvolupament d'APIs REST. Avaluar les solucions existents per integrar la IA i gestionar la base de dades d'usuaris.

- iteració 1:

creació del backend i funcionament del servidor amb Convex

- obertura del servidor

Configurar i desplegar el servidor backend utilitzant Convex. Assegurar la connexió segura i estable del servidor.

- inclusió de la IA

Integrar el model de IA al servidor per gestionar les sol·licituds de text generatiu. Configurar la IA per treballar amb l'API REST.

- estudi del funcionament

Avaluar el rendiment del servidor i la IA. Identificar possibles millores en la configuració inicial.

- testeig d'entrenament IA

Realitzar proves d'entrenament per ajustar i optimitzar el model de IA. Verificar la precisió i eficàcia de les respostes generades.

- iteració 2:

creació d'una pàgina web on poder aconseguir la api.

- obertura d'un portal web senzill

Desenvolupar una pàgina web bàsica per proporcionar informació sobre l'API. Assegurar una interfície d'usuari intuïtiva i fàcil d'utilitzar.

- inclusió de gestió d'usuaris

Implementar funcions de registre, inici de sessió i gestió de perfils d'usuari. Assegurar la seguretat i privacitat de les dades dels usuaris.

- implementació d'un frontend senzill

Crear un frontend senzill que permeti als usuaris interactuar amb l'API. Assegurar la compatibilitat amb diversos dispositius i navegadors.

- iteració 3:

Implementació de les API REST

- inclusió d'un sistema per crear API importables

Desenvolupar un sistema que permeti la creació i la gestió d'APIs importables. Assegurar que les APIs siguin fàcilment integrables en altres aplicacions.

- inclusió del xat amb la ia en la web senzilla

Integrar el sistema de xat amb IA dins la pàgina web. Proporcionar una interfície per interactuar amb el xat bot.

- sistema per descarregar l'API des del portal web

Implementar un sistema que permeti als usuaris descarregar l'API directament des del portal web. Assegurar que la documentació de l'API estigui disponible i sigui comprensible.

- iteració 4:

Aquesta iteració es guardarà exclusivament per ampliar el temps per resoldre problemes que s'hagin detectat durant la realització del treball, creant un marge de temps per la seva resolució.

6 CONVEX

6.1 Introducció

Convex [4] és un Backend as a Service (BaaS) que proporciona serveis de backend a través del núvol, permetent als desenvolupadors centrar-se més en el frontend i la lògica de l'aplicació, mentre que el backend s'encarrega automàticament de la gestió de les bases de dades, l'autenticació d'usuaris, l'emmagatzematge, i altres funcionalitats habituals de les aplicacions.

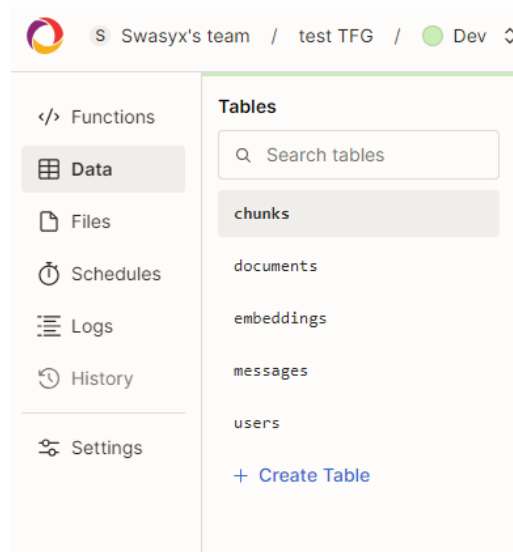


Fig. 1: Menú principal del projecte en Convex

6.2 Backend as a Service (BaaS)

Backend as a Service és una solució que proporciona serveis de backend a través del núvol, permetent als desenvolupadors centrar-se més en el frontend i la lògica de l'aplicació, mentre el backend s'encarrega automàticament de la gestió de bases de dades, l'autenticació d'usuaris, l'emmagatzematge d'arxius, i altres funcionalitats habituals d'una aplicació.



id	createdAt	updatedAt	text	createdAt
1	2023-07-10T10:00:00.000Z	2023-07-10T10:00:00.000Z	"Hi"	2023-07-10T10:00:00.000Z
2	2023-07-10T10:01:00.000Z	2023-07-10T10:01:00.000Z	"Hola"	2023-07-10T10:01:00.000Z
3	2023-07-10T10:02:00.000Z	2023-07-10T10:02:00.000Z	"Bona nit"	2023-07-10T10:02:00.000Z

Fig. 2: Visió d'una taula amb accés en temps real desdel dashboard de Convex

6.3 Estat de l'art

Hi ha diversos BaaS disponibles en el mercat, el més conegut és Firebase, el BaaS de Google, sent el més utilitzat gràcies a la capacitat d'oferir i modificar les dades a temps real i a la gestió d'usuaris automàtica de google. Amazon Web Services (AWS) també disposa de les seves BaaS, tot i que en aquest cas, no es considera un sol programa, sinó que està format per diferents mòduls. Back4app i Parse són dues plataformes BaaS que van ser adquirides per Facebook (ara Meta) i que actualment estan disponibles en codi obert.

Convex encaixa en aquest context omplint buits de mercat com serien la fàcil implementació, el ràpid accés a les dades, ja sigui des del client o des del portal personal de convex. la seva senzillesa, el fet que sigui un programa de codi obert i les possibilitats que ofereix fan que sigui un clar competidor de Firebase i no es pot descartar la possibilitat que sigui el BaaS per excel·lència en els proxims anys.

6.4 Funcionament

Convex genera de manera automàtica l'estructura del backend al iniciar el projecte. La base de dades que es crea no és ni relacional ni no-relacional, depenent del que necessita el projecte permet fer un ús selectiu de les dues, fins i tot barrejant les dues en el mateix projecte. El seu centre d'accions es basa en "schema" "mutation". Schema consisteix en l'estructura de la base de dades, la definició de les taules, la forma en la que estan relacionades i els valors que s'esperen en els seus camps. "Mutation" consisteix en les crides que modifiquen i actuen sobre aquests "schema". Les mutacions, funcions que es programen en la part del servidor, preparen aquest servidor per poder modificar les dades (per fer-ho s'utilitza la part de la llibreria de convex basada en mutacions). Per altre banda, per poder treballar amb les dades des del client, fem ús de "useMutation", una part de la llibreria de Convex centrada en enviar el context en el que estem treballant (schema i la resta del backend) i permet accedir a les mutacions creades en el client desde la api del backend. Fer-ho així afegeix un grau de seguretat extra, ja que les dades es treballen abans de fer cap consulta ni crida a la base de dades. A més, consta de crides a accions, que acostumen a ser internes, en les que pots declarar les funcions per ser accedides des de l'api interna. D'aquesta

manera no cal importar arxius extres i les accions queden amagades en el programa. [5]

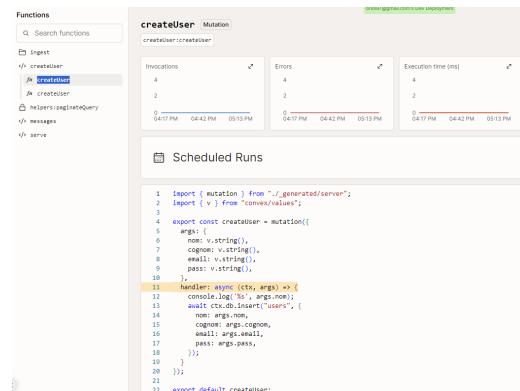


Fig. 3: Ús d'una mutació vista desde la dashboard del projecte en Convex

6.5 Perquè Convex per aquest projecte?

Per acostar-me a les eines que estan fent servir actualment les aplicacions com ChatGPT volia intentar treballar amb algun dels seus programes. Al començar a investigar sobre Convex vaig veure que m'anava molt bé per fer el projecte. Tot i no ser un programa conegut i que per mi era del tot nou, ha suposat un repte ja que no hi ha tanta informació i suport com en la resta de programes similars. LA part positiva és que Convex ja està preparat, de forma automàtica al iniciar-lo, per treballar amb una IA generativa; la forma en la que la "base de dades" funciona permet creat Embeddings de forma ràpida i senzilla, facilitant la feina a l'hora de preparar, entrenat i utilitzar el bot generatiu.

7 RESULTATS

L'apartat de resultats estarà dividit en tres: el frontend i els seus apartats, el backend i finalment la gestió d'usuaris.

7.1 Frontend

Tot i ser un portal senzill consta de 4 apartats (Login, registre, home, perfil). L'apartat del Login permet a un usuari registrat accedir a entrenar el seu chatbot. Per poder fer-ho, l'usuari s'ha de poder registrar, cosa que es fa en l'apartat de registre. En aquest l'usuari entrarà les seves dades (nom, cognom i email) juntament amb una contrassenya. Dins l'apartat del perfil, de moment, nomès hi ha integrada una entrada de text en la que, al entrar-hi una url a un portal web, aquest sera enviat al backend per fer l'entrenament. Finalment, en la pàgina principal hi ha una mostra de com seria el botó que obra el xat, amb el chatbot generatiu funcional a mode de demostració i entrenable gràcies a l'apartat anteriorment mencionat.

7.2 Backend

El Backend està basat en Convex. Per fer-ne un resum dels apartats que em semblen destacables em centraré en els embeddings i en l'ús d'aquests per entrenar la IA. Els embeddings són uns vectors en hexadecimal que contenen l'informació que la IA ha anat aprenent. Tot i ser vectors que no



Fig. 4: Vista de la pàgina d'entrada amb el botó en negre i el xat obert amb la salutació del bot

podem entendre a simple vista, aquests li permeten al xatbot generatiu entendre, aprendre i respondre en els diferents idiomes que hagi estat entrenat. Cada cop que es demana al bot d'entrenar en una web nova, aquest converteix tot el seu contingut en vectors i els organitza dins dels vectors ja existents d'aquest, actualitzant el contingut i creant nous vincles de paraules, estructures i respostes. Això permet, entre altres, entrenar el bot exclusivament en un tipus concret de contingut (posem, per exemple, que volem fer un codi en python. Es molt possible que el nostre bot agafi l'entrenament que tenen la resta i ens farà un codi concret. Si l'hem entrenat en la forma que volem que el faci, el bot sabrà fer el codi tal i com ens agradaria que fos).

```
export const embedList = internalAction({
  args: {
    documentIds: v.array(v.id("documents")),
  },
  handler: async (ctx, { documentIds }) => {
    const chunks = (
      await map(documentIds, (documentId) => {
        ctx.runQuery(internal.ingest.embed.chunksNeedingEmbedding, {
          documentId,
        });
      })
    ).flat();

    const embeddings = await embedTexts(chunks.map((chunk) => chunk.text));
    await map(embeddings, async (embedding, i) => {
      const { id: chunkId } = chunks[i];
      await ctx.runMutation(internal.ingest.embed.addEmbedding, {
        chunkId,
        embedding,
      });
    });
  },
});
```

Fig. 5: Snapshot del codi que permet ens embeddings

7.3 Gestió d'usuaris

Tot i ser un apartat que hauria de ser dins del backend m'agradaria comentar que he investigat formes de fer que els usuaris fossin segurs. Convex conté llibreries i programes secundaris que permeten la gestió d'aquests usuaris sense necessitat, gairabé, de programar. Tot i això, he volgut fer jo el meu propi sistema, per poder veure com era en els inicis i explorar en el seu funcionament. Com a punt central en aquesta gestió he afegit un xifratge amb sha256, evitant guardar les dades en obert. [6]

8 SEGUIMENT

• 10/03/2024

L'objectiu principal del TFG s'ha definit, juntament amb la planificació i la metodologia que s'adequa més per realitzar-lo.

• 21/04/2024

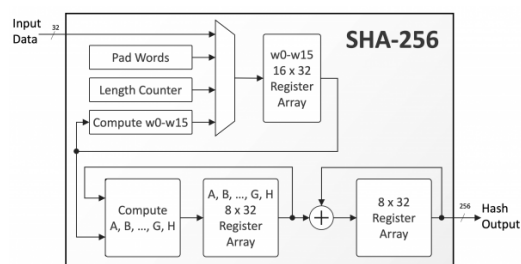


Fig. 6: Mapa explicatori de SHA256 []

Descoberta de Convex. Utilitzant aquesta eina de backend la IA és funcional i permet la creació dels xats diferents. Es segueix amb la planificació esperada per la segona iteració.

• 26/05/2024

Diseny i creació del portal web des del que crear l'usuari, entrenar el chatbot i controlar el seu ús. Es segueix la planificació, la taula kaban comença a alliberar-se i l'esperat en els sprints segueix el seu procediment correcte.

• 16/06/2024

Actualment el xat amb el bot funciona correctament però per temes de gestió d'usuaris estic plantejant la possibilitat d'un canvi d'eines. Al ser un plantejament encara no tinc clara la seva utilitat. Al acabar de decidir ampliaré el treball amb l'incorporació d'una descripció detallada de Convex, si segueix sent l'eina emprada.

9 CONCLUSIONS

9.1 Eficàcia del Portal Web Desenvolupat amb React i Convex

El desenvolupament del portal web utilitzant React i Convex ha demostrat ser una elecció tècnica sòlida i efectiva. React ha proporcionat una interfície d'usuari dinàmica i reactiva, permetent una experiència d'usuari fluida i interactiva. Convex, per la seva banda, ha facilitat la gestió eficient de l'estat de l'aplicació i la manipulació de dades en temps real, millorant així el rendiment general del portal.

9.2 Integració del Xat amb ChatGPT

La integració d'un xat amb ChatGPT ha afegit una dimensió interactiva significativa al portal web. Aquesta funcionalitat permet als usuaris obtenir respostes instantànies a les seves consultes, millorant l'experiència d'usuari i oferint suport continuat.

9.3 Millora de l'Experiència d'Usuari

L'ús combinat de React i Convex, juntament amb la integració del xat, ha resultat en una experiència d'usuari notablement millorada. Els usuaris poden navegar pel portal de manera intuïtiva, interactuar amb el contingut de manera dinàmica i rebre assistència en temps real. Aquest conjunt

de funcionalitats ha fet que el portal sigui més atractiu i útil per als usuaris finals.

9.4 Reptes i Limitacions

Tot i els avantatges esmentats, s'han identificat alguns reptes i limitacions durant el desenvolupament i la implementació del portal. La complexitat inicial de la configuració de Convex i la necessitat de formació específica per a l'ús de ChatGPT són alguns dels obstacles que s'han hagut de superar.

AGRAÏMENTS

En primer lloc, vull expressar la meua sincera gratitud al meu tutor, Marc Ortega, per la seva inestimable guia, suport continu i paciència durant tot el desenvolupament d'aquest projecte. La seva experiència i els seus comentaris perspicaces han estat crucials per a l'èxit d'aquest treball.

També vull agrair a la meua família i amics el seu incondicional suport i comprensió durant els moments en què estava profundament immers en la meua recerca. El seu suport moral m'ha mantingut motivat i enfocat.

A més, vull expressar el meu agraïment als meus companys de classe, que han proporcionat crítiques constructives i suggeriments que han millorat notablement la qualitat d'aquest projecte.

Finalment, vull reconèixer els recursos i el suport proporcionats per l'Universitat Autònoma de Barcelona, que han estat instrumentals en el desenvolupament i la finalització d'aquest treball. Gràcies a tots per fer d'aquest viatge una experiència enriquidora.

REFERÈNCIES

- [1] J. F. González, "Metodologies àgils."
- [2] P. F. Delgado, "Disseny i desenvolupament d'un xatbot de suport a l'usuari basat en models generatius," 2023.
- [3] S. B. Casado, "Desenvolupament de sistema de documentació automàtica aplicant IA per APIs REST."
- [4] "Convex Docs | Convex Developer Hub." [Online]. Available: <https://docs.convex.dev/home>
- [5] "Build AI Chat with Convex Vector Search." [Online]. Available: <https://stack.convex.dev/ai-chat-with-convex-vector-search>
- [6] "SHA-256 | 256-bit SHA Secure Hash Crypto Engine IP Core." [Online]. Available: <https://www.cast-inc.com/security/encryption-primitives/sha-256>
- [7] "LaMDA: our breakthrough conversation technology," May 2021. [Online]. Available: <https://blog.google/technology/ai/lamda/>
- [8] "Build Chatbots using Serverless Bot Framework with Salesforce Integration | AWS Architecture Blog," Mar. 2021, section: Amazon CloudFront. [Online]. Available:

<https://aws.amazon.com/blogs/architecture/build-chatbots-using-serverless-bot-framework-with-salesforce-integration/>

- [9] "IBM Cloud Docs." [Online]. Available: <https://cloud.ibm.com/docs/watson-assistant?topic=watson-assistant-architecture-data>
- [10] RobBagby, "Artificial intelligence (AI) architecture - Azure Architecture Center," Aug. 2023. [Online]. Available: <https://learn.microsoft.com/en-us/azure/architecture/ai-ml/>
- [11] A. Wagh, "Open AI — Understand Foundational Concepts of ChatGPT and cool stuff you can explore!" Jul. 2023. [Online]. Available: <https://medium.com/@amol-wagh/open-ai-understand-foundational-concepts-of-chatgpt-and-cool-stuff-you-can-explore-a7a77baf0ee3>