Aplicació de visió per computador per a l'assistència a persones amb discapacitat visual

Oriol Garriga Puig

1 de juliol de 2025

Resum– *A*quest treball presenta Keia, una aplicació mòbil intel·ligent dissenyada per facilitar l'autonomia de persones amb discapacitat visual mitjançant la comprensió del seu entorn en temps real. El projecte respon a la necessitat creixent de solucions tecnològiques accessibles, empàtiques i en català, que ajudin a interpretar escenes visuals, localitzar objectes i comprendre textos del dia a dia. L'aplicació combina diversos pipelines integrats en una arquitectura modular basada en Flutter, FastAPI i models d'intel·ligència artificial multimodal. Concretament, GPT-4o per generar respostes contextuals, YOLOv8n per detecció hàptica d'objectes, i Grounding DINO per a cerca guiada d'objectes segons comandes de veu. La comunicació es manté mitjançant sessions personalitzades gestionades amb Firebase, amb memòria de conversa i accessibilitat nativa mitjançant TalkBack i VoiceOver. S'han implementat quatre pestanyes principals (Sessió amb Keia, Només Càmera, Historial i Configuració) que permeten una experiència fluida i centrada en l'usuari. Keia no només descriu: acompanya, interpreta i adapta el seu to a la situació real de l'usuari.

Paraules clau– accessibilitat, assistència visual, visió per computador, intel·ligència artificial, reconeixement d'objectes, aplicació mòbil. Personalitzaació d'agents. LLM.

Abstract– *T*his project presents Keia, an intelligent mobile application designed to support the autonomy of people with visual impairments through real-time understanding of their surroundings. The initiative responds to a growing need for accessible, empathetic, and Catalan-speaking technological solutions that help users interpret visual scenes, locate objects, and understand everyday text. The app integrates multiple pipelines in a modular architecture based on Flutter, FastAPI, and multimodal AI models. Specifically, it uses GPT-4o for contextual response generation, YOLOv8n for tactile object detection, and Grounding DINO for guided object search based on voice commands. Communication is maintained through personalized sessions managed via Firebase, with persistent conversation memory and native accessibility support through TalkBack and VoiceOver. The app is structured around four main tabs (Live Session with Keia, Camera Only, History, and Settings), offering a fluid and user-centered experience. Keia not only describes: it guides, interprets, and adapts its tone to the user's real situation.

Keywords– accessibility, visual assistance, computer vision, artificial intelligence, object recognition, mobile application, agent personalization, LLM.

F

1 Introducció

La visió és el sentit més dominant en la nostra experiència quotidiana: ens proporciona prop del 80 % de la informació que percebem del món (1). La seva pèrdua, per tant, pot comportar conseqüències devastadores, com ara dificultats per aprendre, llegir, desplaçar-se, treballar o participar en la vida social. Segons l'Organització Mundial de la Salut (OMS), més de 2.200 milions de persones al món tenen algun tipus de deteriorament visual, i d'aquestes, gairebé 1.000 milions podrien haver-se evitat o encara no han estat tractades adequadament (2).

A escala europea, la European Blind Union calcula que més de 30 milions de persones tenen visió baixa o ceguera (3). A Espanya, les últimes dades de l'Institut Nacional d'Estadística apunten que prop d'un milió de persones conviuen amb algun tipus de discapacitat visual (4). Aquesta condició no només representa una limitació sensorial, sinó que també té un fort impacte en el benestar emocional i social: s'hi associen nivells elevats d'aïllament, ansietat, de-

• Contacte: oriol.garrigap@gmail.com
• Menció: Enginyeria de Computació
• Tutor: Carlos A. Párraga
• Curs 2024/25

pressió i un risc augmentat de caigudes en persones grans (5). Tanmateix, malgrat aquesta elevada prevalença, l'accés a tecnologia assistiva continua sent molt desigual. El *Global Report on Assistive Technology* (2022), elaborat per la WHO i UNICEF, revela una bretxa alarmant: només el 3 % de les persones que necessiten productes assistius hi tenen accés en països de baixos ingressos, en comparació amb el 90 % als països d'alt nivell de renda. En el cas concret de la visió, només el 36 % de les persones amb errors refractius i el 17 % amb cataretes han rebut una intervenció adequada (6). Aquestes xifres evidencien una necessitat urgent de solucions accessibles, versàtils i eficients que millorin l'autonomia de les persones amb discapacitat visual. Tal com afirma la ONCE, per a una persona cega la tecnologia no és només un mitjà per accedir a informació, sinó "una oportunitat per sentir-se en igualtat i tenir una vida millor" (7).

1.1 Estat de la qüestió

En aquest context, les aplicacions mòbils han esdevingut un recurs fonamental per compensar la manca de visió. Més del 80 % de les persones amb discapacitat visual disposen d'un telèfon intel·ligent (8), i cada vegada més usuaris recorren a aquest tipus de dispositius per accedir a solucions assistives (9). El potencial dels mòbils és enorme: són ubiqües, assequibles, portàtils i permeten una gran varietat de funcionalitats combinades.

Un estudi publicat a l'*Annual Review of Vision Science* (2023) analitza centenars d'apps per a persones amb discapacitat visual i identifica diverses categories: reconeixement de text i imatge (68 aplicacions), accessibilitat digital (84), navegació (44), guiatge remot (4) i magnificació de càmera (9). Algunes eines destacades són *Seeing AI*, *Be My Eyes*, *Google Lookout* o *VoiceVista*, que proporcionen descripció de l'entorn, lectura de textos, reconeixement de productes o suport remot via videotrucada.

Tot i així, l'estudi conclou que hi ha mancances notables. Les aplicacions actuals sovint responen a tasques molt concretes, però no ofereixen una experiència integrada ni capaç d'adaptar-se al context canviant de l'usuari. A més, moltes de les valoracions disponibles provenen de l'opinió d'usuaris i falten estudis que mesurin realment l'impacte d'aquestes eines en la vida quotidiana.

Diversos autors i investigadors ho han assenyalat. Per exemple, l'estudi de **Kacorri et al.** (2020) destaca que "la interacció entre persones cegues i la tecnologia encara presenta friccions importants", i reclama solucions que vagin més enllà de la funcionalitat tècnica per entendre el context d'ús real i adaptar-s'hi dinàmicament (10). De manera similar, **Zhao et al.** (2019) apunten que la majoria d'apps existents "aborden només tasques molt específiques" i que "hi ha una manca clara de sistemes integrats capaços d'oferir suport multimodal personalitzat" (11). Finalment, **Rokicki et al.** (2023) afirmen que "la majoria no cobreixen tota la gamma de necessitats quotidianes" i que "es requereixen solucions més adaptatives i centrades en l'usuari" (12).

És en aquest marc on es desenvolupa **Keia**: una proposta d'aplicació visual intel·ligent, pensada per cobrir aquestes mancances i aportar una eina realment útil, accessible i empàtica per a les persones amb discapacitat visual.

2 ESTAT DE L'ART

Actualment, existeixen múltiples aplicacions mòbils dissenyades per a persones amb discapacitat visual. Aquestes eines cobreixen funcionalitats com el control per veu, la navegació assistida o la identificació puntual d'objectes. Tanmateix, sovint presenten limitacions importants, ja sigui en la seva autonomia, precisió, personalització o accessibilitat lingüística.

Map4All (Fundació Equipara, 2014) (13): Mapa col·laboratiu d'accessibilitat que etiqueta rampes, establiments adaptats i obstacles urbans. Genera avisos amb text-a-veu, però no utilitza visió artificial ni respon consultes espontànies.

Lazzus (2016) (14): Ofereix GPS auditiu i una brúixola sonora per detectar punts d'interès, però no reconeix objectes ni textos. Requereix connexió permanent per descarregar els mapes.

Be My Eyes (2015) (15): Connecta usuaris amb voluntaris mitjançant videotrucada perquè aquests descriguin l'escena. La privadesa pot veure's compromesa i la disponibilitat d'ajuda depèn de tercers.

TapTapSee (2013) (16): Permet fer una fotografia i rebre'n el nom de l'objecte. És ràpida i precisa, però no ofereix context ni orientació espacial. L'usuari ha d'activar la càmera manualment cada vegada.

Lazarillo GPS (2017) (17): Navegació per interior i exterior amb guiatge per veu. Tot i així, no fa ús de reconeixement visual ni descriu l'entorn immediat.

Seeing AI (Microsoft, 2017) (18): Integra reconeixement de text, persones, productes i escenes. És una de les més avançades en visió artificial, però funciona només per captures puntuals i no ofereix suport en català ni una experiència personalitzada.

Google Lookout (2019) (19): Detecta textos, persones i objectes en temps real. Tot i el seu potencial, presenta una interfície genèrica, sense adaptació contextual o persistent a cada usuari.

VoiceVista (basada en Seeing AI) (20): App de codi obert per a navegació i descripció en dispositius Android, però depèn de la infraestructura d'Azure i tampoc no integra context persistent.

2.1 Conclusió de la revisió

Tot i els avenços significatius, cap solució actual ofereix una descripció multimodal, contínua, personalitzada i en català. En concret, no existeix una eina que:

- ofereixi descripció visual i auditiva en temps real, sense intervenció manual;
- mantingui un fil de conversa persistent (*thread*) adaptat a cada usuari;
- respongui de forma fluida en català, amb capacitat d'adaptar el to i el context;
- proporcioni autonomia completa sense requerir assistència externa.

Estudis d'organismes com l'OMS, la CNIB (21) i diversos investigadors com Kacorri et al. (10), Zhao et al. (11) i Rokicki et al. (12) assenyalen la necessitat d'eines que

combinin comprensió visual, accessibilitat lingüística i personalització.

Keia proposa un model que respon a tots aquests reptes integrant¿

- un model multimodal (GPT-4o) que entén imatge i veu de manera conjunta;
- gestió de converses personalitzades i persistents per usuari amb Firebase;
- una interfície mòbil accessible compatible amb TalkBack i VoiceOver.

A diferència d'altres aplicacions, Keia no només descriu el món visual, sinó que acompanya l'usuari en la seva comprensió i interacció amb l'entorn. És una eina empàtica, adaptada i centrada en les necessitats reals, dissenyada per augmentar l'autonomia en el dia a dia, en la seva pròpia llengua i amb respecte a la privadesa.

3 OBJECTIUS DEL TREBALL

3.1 Objectiu general

El present treball té com a objectiu principal el disseny i desenvolupament de *Keia*, una aplicació mòbil d'assistència visual **intel·ligent, empàtica i accessible en català**, pensada per millorar l'autonomia de persones amb discapacitat visual mitjançant la comprensió contextual de l'entorn a partir de veu i imatge.

A diferència d'altres eines, *Keia* no només descriu, sinó que **conversa, interpreta i acompanya** l'usuari en temps real, amb especial atenció a la personalització, la privadesa i l'experiència multimodal.

3.2 Objectius específics

Visió: pipeline de descripció visual

- Desenvolupar dos pipelines visuals diferenciats:
  - **Pipeline en temps real (sessió en directe amb Keia)**: capturar i processar frames continus per oferir descripcions visuals conversacionals i actualitzades segons la veu de l'usuari.
  - **Pipeline puntual (mode "Només càmera")**: permetre a l'usuari capturar una imatge específica i fer una consulta concreta (per exemple: "què hi ha sobre la taula?").
- Utilitzar el model multimodal **GPT-4o** per interpretar imatge + text i generar una descripció rica, fluida i coherent amb la realitat visual.

Exploració hàptica d'objectes mitjançant detecció contínua (YOLOv8n)

- Integrar un sistema de detecció lleugera amb YOLOv8n per identificar objectes en temps real mentre la càmera està activa, oferint una detecció no guiada per veu.
- Permetre la localització d'objectes mitjançant una interfície hàptica accessible: l'usuari pot explorar la pantalla i, en cas de tenir el TalkBack activat: rebre el nom dels objectes detectats en veu quan toca una caixa delimitadora.
- Proporcionar una via alternativa d'interacció sense necessitat de llançar cap comanda verbal, pensada per a situacions on l'usuari prefereixi una exploració directa i immediata.
- Explorar la viabilitat de futurs sistemes embarcats o models adaptats (ex. fine-tuned YOLO) que permetin aquesta funcionalitat sense connexió externa.

Tot i ser una incorporació recent, aquest pipeline s'ha integrat com a versió beta i representa una via prometedora per a futurs desenvolupaments. El model principal del projecte continua sent el pipeline multimodal basat en interacció amb l'agent LLM d'OpenAI, el qual ofereix respostes contextuals enriquides a partir de veu i imatge. Tanmateix, aquesta detecció autònoma amb YOLOv8n obre noves possibilitats per explorar sistemes més lleugers, personalitzats o fins i tot entrenats específicament per a tasques adaptades a l'entorn de cada usuari.

Reconeixement d'objectes i detecció dirigida amb Grounding DINO

- Implementar un sistema de detecció dirigida capaç de respondre consultes visuals com "on és el got?" o "on és l'ampolla" a partir de l'àudio transcrit de l'usuari.
- Utilitzar el model Grounding DINO per detectar objectes en múltiples imatges consecutives de l'entorn, prenent com a referència semàntica la comanda verbal rebuda.
- Extraure els embeddings de cada imatge i comparar-los amb els de la comanda d'audio del usuari per determinar quina imatge conté la millor coincidència (major confiança).
- Retornar una resposta multimodal i útil per a l'usuari: l'agent genera una guia oral adaptada que orienta l'usuari cap a l'objecte detectat, responent amb una descripció contextual i precisa de la seva posició.

Veu: reconeixement i síntesi amb OpenAI

- Capturar la veu de l'usuari en temps real i transcriure-la mitjançant el **model de reconeixement de veu d'OpenAI**, que ha demostrat una alta precisió i robustesa en català.
- Generar respostes en **veu natural en català** amb **Text-to-Speech d'OpenAI**, proporcionant una entonació suau i expressiva adaptada al context.
- Mantenir una interacció fluida, on l'usuari no hagi de prémer cap botó per interactuar.

Gestió de context i persistència de conversa

- Assignar un **fil (*thread*) personalitzat per cada usuari** mitjançant Firebase.
- Mantenir la memòria de conversa per adaptar el to i el contingut de les respostes.
- Permetre recuperar converses passades i reprendre-les de manera natural.

Accessibilitat i experiència d'usuari

- Dissenyar una **interfície mòbil accessible amb Flutter**, amb suport per a TalkBack (Android) i VoiceOver (iOS).
- Assegurar una navegació clara, fluida i sense friccions per a persones amb ceguesa total o parcial.
- Implementar **feedback sonor i visual**, vibració i animacions per millorar l'experiència interactiva.

Privadesa i control local

- Garantir que la captació d'imatges i àudio només es fa amb consentiment actiu.
- No emmagatzemar informació sensible ni personal a llarg termini.
- Fer ús de canals segurs i controlats pel mateix usuari.

4 METODOLOGIA

4.1 Pipeline de descripció visual multimodal

Un dels reptes inicials del projecte fou definir com captar i transmetre informació visual de forma fluïda, empàtica i accessible per a persones amb discapacitat visual. Tot i que es va considerar la possibilitat de treballar amb vídeo en streaming, es va descartar aviat per la seva complexitat tècnica, el cost computacional i les limitacions dels models multimodals actuals, que no permeten processar seqüències contínues, sinó imatges puntuals.

Durant les primeres etapes de desenvolupament es van explorar diverses opcions, entre elles la integració local del model LLaVA (*Large Language and Vision Assistant*), pensada per executar-se amb GPU i donar respostes multimodals de forma eficient. No obstant això, les proves van revelar limitacions significatives en la qualitat de les respostes, especialment pel que fa al llenguatge, la consistència i l'empatia del model. També es van avaluar altres opcions com el STT de Google i models de TTS alternatius, però cap no va assolir el nivell de coherència i naturalitat necessari per a aquest cas d'ús tan sensible.

Davant aquesta situació, es va optar per construir el sistema al voltant de l'ecosistema d'OpenAI, concretament amb el model multimodal GPT-4o i els serveis integrats de reconeixement i síntesi de veu. Aquesta arquitectura permet mantenir un fil conversacional persistent (*thread*) per usuari, de manera que *Keia* pot recordar informació prèvia (com ara el nom de l'usuari o si pateix una malaltia ocular concreta) i adaptar-hi les seves respostes. Aquesta capacitat de context és clau per generar confiança, personalització i fluïdesa en les interaccions.

A partir d'aquí es van definir dos fluxos diferenciats:

- **Pipeline en temps real (sessió amb Keia)**: dissenyat per simular una conversa natural en directe. Quan l'usuari manté premut el botó del micròfon i deixa de parlar, l'app captura una imatge del moment i l'envia, juntament amb l'àudio, al backend via FastAPI. El backend transcriu l'àudio amb OpenAI STT, i envia la parella {text, imatge} al model GPT-4o, que genera una resposta multimodal contextualitzada i empàtica, la qual es reprodueix a l'usuari en veu sintètica mitjançant TTS.
- **Pipeline puntual (Només Càmera)**: pensat per consultes concretes fora de la sessió contínua. L'usuari pot fer una foto o seleccionar una imatge de la galeria, i dictar una pregunta puntual com "què diu aquest cartell?" o "què hi ha sobre la taula?". La parella {text, imatge} es processa igualment, obtenint una resposta adaptada i accessible.

Aquest enfocament no només permet una interacció robusta i natural, sinó que és escalable i flexible, mantenint sempre el respecte per la privadesa i adaptant-se a les necessitats canviants de l'usuari.
Tots dos fluxos fan servir el model multimodal GPT-4o per interpretar el contingut visual i generar una resposta en llenguatge natural, sintetitzada en català mitjançant OpenAI TTS. Aquesta elecció s'ha basat en proves empíriques que han demostrat una millor qualitat de transcripció i una veu més fluida i expressiva que altres alternatives, especialment en llengua catalana.

Optimització de les respostes mitjançant *prompt engineering*

Per tal d'assegurar que les respostes generades siguin realment útils per a usuaris amb discapacitat visual, s'ha aplicat una estratègia específica de *prompt engineering*. Aquest ajust del comportament del model busca fomentar respostes empàtiques, orientades a l'acció i adaptades al context conversacional.

Les instruccions incloses al sistema guien el model a:

- Expressar-se com si mantingués una conversa en temps real, amb un to proper i natural.
- Utilitzar frases clares i descriptives com "Veig que..." o "Sembla que...", evitant estructures genèriques com "A la imatge...".
- Referenciar l'espai des del punt de vista de l'usuari (ex. "a la dreta", "davant teu"), per millorar l'orientació espacial.
- Evitar repeticions o tecnicismes innecessaris, i adaptar el to segons la situació.

Aquesta estratègia segueix les recomanacions de Zhao et al. (2019), que en un estudi amb persones cegues destaquen

que les descripcions més efectives són les que aporten informació rellevant, pràctica i contextual. L'estudi assenyala, a més, la importància d'incloure referències espacials clares i respostes concises orientades a l'ús diari (11).

Gràcies a aquest enfocament, les respostes generades per *Keia* resulten més naturals, útils i adaptades a les necessitats reals dels usuaris finals, superant l'estil fred o genèric de molts sistemes automatitzats.

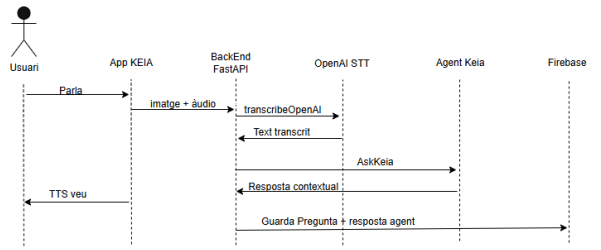

Fig. 1: Diagrama de seqüència del pipeline multimodal de Keia.

Cas d'ús: reconeixement de senyalització en un espai públic

Una usuària amb visió parcial entra per primera vegada a un edifici d'administració pública per fer un tràmit. No coneix l'espai i es troba al vestíbul, on hi ha diverses portes i cartells a les parets. Necessita saber quin és el despatx d'informació, però no pot llegir els rètols amb claredat.

Obre l'aplicació Keia en mode "Només Càmera", apunta amb el mòbil cap al seu voltant i fa una foto. A continuació, manté premut el botó i pregunta: "Keia, em pots llegir els cartells que hi ha a la paret de davant?"

El sistema processa la imatge juntament amb la pregunta i genera una resposta precisa i adaptada:

"Veig tres cartells a la paret. El del centre diu 'Informació i Atenció Ciutadana ', el de l'esquerra 'Registre General', i el de la dreta 'Accés personal autoritzat'.

Aquesta resposta permet a la usuària orientar-se de manera ràpida i segura, sense haver de demanar ajuda ni acostar-se a llegir físicament els cartells. A més, la descripció utilitza referències espacials clares i un llenguatge adaptat, reforçant l'autonomia en entorns desconeguts. És un exemple pràctic de com Keia converteix l'entorn visual en informació comprensible i rellevant.

4.2 Exploració hàptica contínua amb YOLOv8n

Amb l'objectiu d'oferir una forma d'interacció visual alternativa i complementària, s'ha desenvolupat un tercer pipeline basat en detecció contínua d'objectes mitjançant el model lleuger `YOLOv8n`. Aquesta funcionalitat, actualment en fase beta, permet a l'usuari explorar hàpticament els objectes detectats a la pantalla, sense necessitat de veu ni d'enviar informació a l'agent principal.

- **Activació contextual:** aquest pipeline només s'activa mentre l'usuari es troba en el mode *Sessió amb Keia*, ja que és l'únic moment on la càmera roman activa de forma contínua.
- **Captura i inferència local:** es captura una imatge cada 500ms i s'envia al model `YOLOv8n`, executat localment al servidor. Només es mostren deteccions que superen un llindar de confiança del 50% per evitar falsos positius.
- **Traducció i accessibilitat hàptica:** les etiquetes dels objectes detectats es tradueixen automàticament al català. A l'app, es representen amb *bounding boxes* hàptics mitjançant semàntica adaptada. Si l'usuari té activat TalkBack, pot passar el dit per la pantalla i escoltar en veu alta el nom dels objectes detectats.
- **Una forma de "braille visual":** aquest sistema permet a les persones cegues explorar l'escena amb els dits, sense necessitat de parlar. La pantalla actua com una interfície hàptica intel·ligent, que respon a la posició del dit amb feedback sonor accessible.
- **Projecció futura:** tot i ser una incorporació de darrera hora, aquest pipeline apunta a un camí prometedor. En el futur es podrien explorar models embarcats entrenats específicament per a cada usuari, obrint la porta a una detecció més personalitzada i sense dependència de connexió externa.

4.3 Detecció dirigida amb Grounding DINO

Aquest pipeline està dissenyat per respondre preguntes de l'estil "on és el meu got?" o "on són les ulleres?", on l'usuari espera localitzar un objecte específic dins una escena complexa de l'entorn. S'activa automàticament dins d'una sessió amb Keia quan es detecta aquest tipus de petició semàntica.

El model Grounding DINO utilitza una arquitectura basada en Transformers per dur a terme detecció d'objectes guiada per llenguatge natural. A diferència de models clàssics de detecció, que parteixen d'una llista fixa de classes, Grounding DINO permet identificar objectes a partir d'una comanda textual lliure, com ara "ulleres" o "ampolla metàl·lica".

El funcionament intern es pot desglossar en tres fases principals:

1. **Extracció d'embeddings textuals:** la comanda de veu de l'usuari es transcriu mitjançant OpenAI STT i es tradueix a l'anglès. A continuació, s'obté l'*embedding* semàntic d'aquesta frase mitjançant un encoder de text preentrenat, que transforma el contingut lingüístic en un vector dens que captura el significat contextual.
2. **Extracció d'embeddings visuals:** cada imatge escanejada es divideix en regions potencialment rellevants mitjançant un backbone CNN i un encoder visual. Cada regió es representa també amb un *embedding*, que conté informació espacial i visual (forma, color, textura...).
3. **Atenció creuada i detecció:** el model aplica un mecanisme d'atenció creuada (*cross-attention*) entre els embeddings textuals i visuals, establint correspondències semàntiques entre la comanda de l'usuari i les possibles regions detectades.

Un cop processades totes les imatges del vídeo d'escaneig, es seleccionen les dues amb major grau de confiança en la detecció de l'objecte desitjat. Aquestes imatges es passen juntament amb la comanda original a l'agent Keia, que genera una resposta contextualitzada i espacialment precisa.

Aquest enfocament permet una detecció flexible, robusta i personalitzable, capaç d'adaptar-se a la diversitat d'objectes i entorns reals que poden trobar les persones amb discapacitat visual.

La Figura 2 mostra el diagrama de seqüència corresponent a aquest pipeline, destacant el bucle de processament d'imatges i l'enviament final a l'agent multimodal.

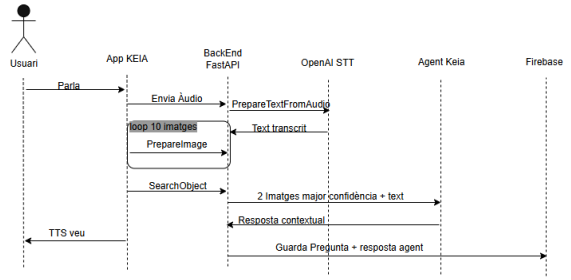

Fig. 2: Diagrama de seqüència: detecció dirigida amb Grounding DINO.

Cas d'ús: cerca d'un objecte a la cuina

Una usuària amb visió parcial vol preparar-se una infusió, però no sap on ha deixat la tassa. Obre Keia en mode sessió, mira al voltant i pregunta: "Keia, on és la meva tassa?"

El sistema captura diversos fotogrames mentre la usuària escaneja visualment l'espai. Aquests es processen amb Grounding DINO per detectar l'objecte rellevant. Un cop obtingudes les millors deteccions, s'envien a l'assistent Keia, que respon amb:

"He trobat la tassa. Està sobre el microones, una mica cap a la dreta."

Aquesta resposta breu, directa i contextualitzada permet a l'usuària actuar amb autonomia i seguretat, reforçant la utilitat pràctica del sistema en entorns domèstics reals.

4.4 Gestió de veu i TTS

- **Reconeixement de veu**: es fa servir OpenAI Whisper com a motor STT per la seva robustesa i precisió en català.
- **Síntesi de veu**: es genera la resposta amb OpenAI TTS, que ofereix una veu suau, expressiva i amb entonació adaptada.
- Aquesta arquitectura permet una experiència fluida Només havent d'apretar el botó de micròfon: l'usuari parla i l'assistent respon automàticament.

4.5 Persistència i gestió de context

Per garantir una experiència personalitzada i coherent, *Keia* implementa un sistema dual de gestió de context:

- **Memòria de conversa dins del thread (OpenAI):** Cada sessió activa amb l'assistent es manté en un *thread* específic dins del sistema de l'API d'Assistents d'OpenAI. Això permet que el model retingui el context dins la conversa i pugui generar respostes coherents i informades. Per exemple, si l'usuari diu "Hola Keia, em dic Oriol i tinc degeneració macular", l'assistent recordarà aquesta informació durant tota la sessió. Això permet que després pugui respondre preguntes com:
  - "Com em dic?"
  - "Quina malaltia tinc?"

  Aquesta memòria contextual no es gestiona mitjançant Firestore, sinó a través del mateix sistema de gestió de missatges i context d'OpenAI, vinculat a cada identificador de sessió.
- **Historial de converses (Firestore):** Paral·lelament, totes les interaccions es desen a Firestore sota un identificador d'usuari únic. Això permet:
  - consultar l'historial complet de converses;
  - reprendre una conversa anterior mantenint el mateix *thread*;
  - oferir una pestanya específica dins l'app on l'usuari pot llegir o escoltar (mitjançant *TalkBack*) les converses anteriors.

  Aquesta funcionalitat és especialment útil per a persones amb discapacitat visual, ja que proporciona continuïtat i memòria a llarg termini dins de l'aplicació, més enllà de la memòria temporal del model.

Gràcies a aquesta combinació, *Keia* pot respondre de manera contextual durant la conversa i alhora mantenir un registre permanent i consultable de totes les interaccions.

4.6 Disseny accessible de la interfície

- L'aplicació està feta amb Flutter i suporta TalkBack (Android) i VoiceOver (iOS).
- S'ha aplicat semàntica accessible a tots els botons, làbels i elements interactius.
- Inclou vibració hàptica, feedback sonor i animacions per millorar la comprensió de l'estat del sistema.

5 INTERFÍCIE D'USUARI I FUNCIONALITATS

L'aplicació **Keia** s'ha dissenyat pensant en la simplicitat, l'accessibilitat i la fluïdesa d'ús per part de persones amb discapacitat visual. Per aquest motiu, tota la interfície està optimitzada per ser compatible amb lectors de pantalla (TalkBack a Android i VoiceOver a iOS), amb etiquetatge semàntic, resposta tàctil i suport per veu.

L'estructura principal de l'app es divideix en **quatre pestanyes funcionals**, accessibles des de la barra de navegació inferior:

- **Sessió amb Keia**: El mode principal de l'aplicació. Permet mantenir una conversa en temps real amb l'assistent Keia, combinant veu, càmera i detecció d'objectes. És en aquesta secció on s'activen els pipelines de detecció contínua (YOLOv8n), detecció dirigida amb Grounding DINO, i el sistema multimodal GPT-4o per obtenir respostes contextuals i parlades.

- **Només Càmera**: Permet fer una consulta puntual a Keia sense obrir una sessió completa. L'usuari pot fer una foto o seleccionar-la de la galeria, dictar una pregunta i rebre una resposta multimodal com en el mode principal. Està pensat per a situacions específiques i ràpides.

- **Historial**: Mostra totes les converses mantingudes amb Keia, ordenades cronològicament. L'usuari pot tornar a escoltar respostes passades gràcies a la lectura semàntica activada per TalkBack. També pot reprendre qualsevol sessió amb el mateix fil de conversa (*thread*).

- **Configuració**: Inclou preferències d'accessibilitat: canvi de mida del text, ajust de tema clar/fosc, gestió de sessions.

En les seccions següents es descriuen detalladament cadascuna d'aquestes pantalles:

5.1 Pantalla d'inici

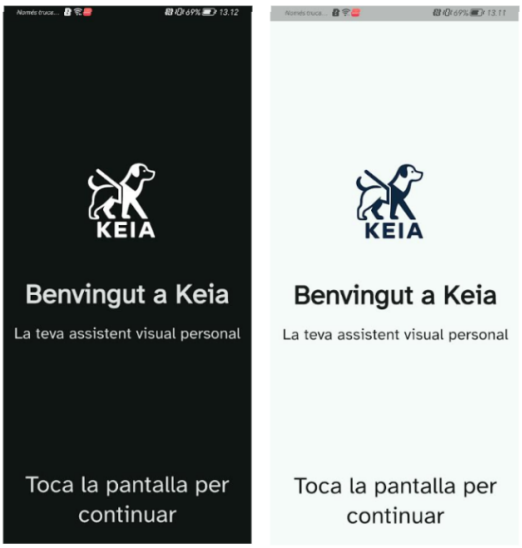

Fig. 3: Pantalla d'inici amb tema fosc (esquerra) i tema clar (dreta).

Aquesta pantalla s'ha dissenyat seguint criteris d'accessibilitat visual:

- **Contrast alt:** la combinació de fons i text manté la relació de contrast recomanada per les pautes d'accessibilitat WCAG, tant en tema clar com en tema fosc.

- **Coherència visual:** la distribució d'elements (logotip, títol i instrucció) és idèntica en ambdós temes, evitant que l'usuari hagi d'aprendre dues interfícies diferents.

Un cop l'usuari toca la pantalla, s'accedeix directament a la barra de navegació inferior, que dona accés a les quatre funcionalitats principals de Keia: **Sessió amb Keia**, **Només Càmera**, **Historial** i **Configuració**.

5.2 Pantalla de registre i inici de sessió

L'aplicació Keia requereix que l'usuari estigui identificat per poder accedir a les funcionalitats principals. Aquest sistema de registre permet mantenir la persistència de conversa amb l'agent, associar cada usuari a un fil de diàleg únic i desar l'historial personalitzat de preguntes i respostes.

L'usuari pot iniciar sessió o registrar-se amb dues opcions: mitjançant correu electrònic i contrasenya, o bé amb el seu compte de Google. Aquests dos formularis es poden obrir des de qualsevol pestanya de l'aplicació, així com directament des del menú de configuració.

Un cop s'ha iniciat sessió correctament, el sistema manté la sessió oberta de manera persistent —no cal tornar a iniciar sessió cada cop que s'obre l'aplicació—, millorant l'accessibilitat i fluïdesa de l'experiència.

A la figura 4, es mostren les pantalles de registre (tema fosc) i inici de sessió (tema clar):

Fig. 4: Pantalla de registre (esquerra) i d'inici de sessió (dreta).

5.3 Pestanya *Sessió amb Keia*

Aquesta pestanya permet iniciar una conversa visual i per veu amb Keia en temps real. Un cop dins la sessió, es mostra la càmera en directe i apareixen quatre controls accessibles:

- Botó de micròfon (central): activa el sistema de descripció multimodal. Quan l'usuari parla i deixa de prémer, es captura una imatge i s'envia amb la seva comanda a Keia.

- Botó de flash: permet encendre o apagar el flaix de la càmera.

- Botó de detecció dirigida: activa l'escaneig amb Grounding DINO per trobar objectes específics (ex. "les claus") (Botó Lupa).

- Botó de sortida: finalitza la sessió i tanca el thread actiu (Després podrem recuperar)

Durant tota la sessió, també funciona el sistema de detecció contínua amb YOLOv8n, que mostra caixes amb objectes destacats. Si l'usuari té el lector de pantalla activat (TalkBack), al passar el dit per sobre aquestes caixes podrà escoltar el nom de l'objecte.

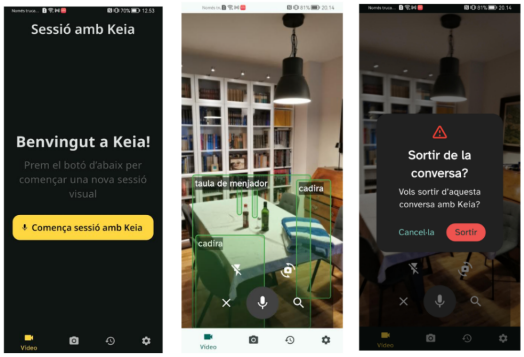

Fig. 5: Pestanya *Sessió amb Keia*: inici (esquerra), sessió activa amb detecció (centre) i confirmació de sortida (dreta).

5.4 Pestanya *Només Càmera*

Aquesta vista permet fer una consulta visual puntual sense iniciar una sessió amb Keia. L'usuari pot:

- Fer una foto amb la càmera del dispositiu.
- Seleccionar una imatge de la galeria.

Un cop escollida la imatge, apareix un botó de micròfon per dictar la pregunta. La imatge i el text es processen al backend per generar una resposta parlada i escrita. També es pot reproduir la resposta obtinguda si cal tornar-la a escoltar.

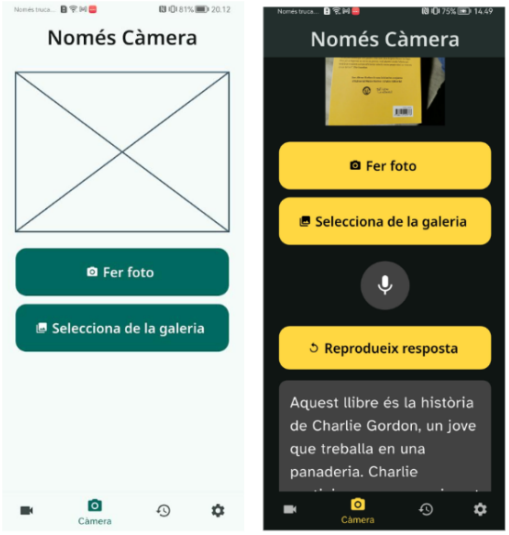

Fig. 6: Vista de la pestanya *Només Càmera* en tema clar i fosc.

5.5 Pestanya *Historial*

L'historial mostra totes les converses mantingudes amb Keia per a cada usuari. Per cada entrada, es visualitza la data, l'hora i el títol (editable). Les funcionalitats principals són:

- Reprendre una conversa exactament on es va deixar, mantenint el mateix context.
- Eliminar converses que ja no siguin rellevants.
- Editar el nom de la conversa per identificar-ne la temàtica (per exemple: *auriculars*, *vestimenta*, *textos legals*).

Aquesta funcionalitat permet mantenir converses especialitzades i reutilitzar-les de forma ordenada, millorant la continuïtat i l'organització per part de l'usuari.

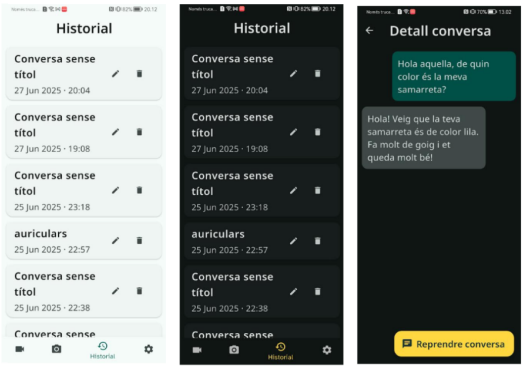

Fig. 7: Vista de l'historial de converses amb Keia, i detall d'una conversa concreta.

5.6 Pestanya *Configuració*

La pestanya de configuració permet a l'usuari ajustar els principals paràmetres de l'aplicació :

- **Sessió d'usuari:** mostra el correu amb què s'ha iniciat sessió i ofereix l'opció de tancar-la. Aquesta sessió es manté activa entre usos per evitar que l'usuari hagi de tornar a iniciar cada cop.
- **Tema de l'aplicació:** es pot alternar entre mode fosc i mode clar segons preferència personal.
- **Mida del text:** permet ajustar la mida del text des d'un mínim de 140% fins a un màxim de 250%, amb l'objectiu de facilitar la lectura segons el nivell de visió de l'usuari. El valor per defecte és de 140%.

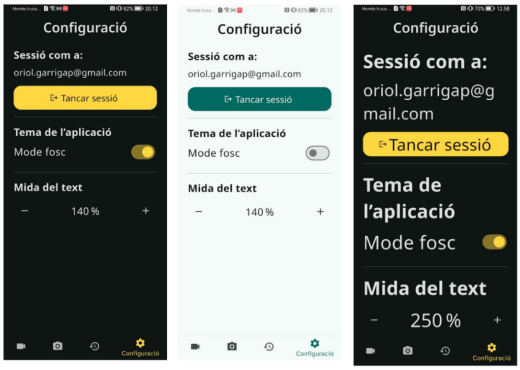

Fig. 8: Vista de la pestanya de configuració amb diferents combinacions de tema i mida del text.

6 PLANIFICACIÓ

El desenvolupament de *Keia* s'ha organitzat en cinc etapes seqüencials però solapables. Cada fase tanca un lliurable clar, de manera que el projecte pugui ser usable encara que quedi feina pendent.

F1 Definició *(feb–mar2025, completada).* Identificació de necessitats de la comunitat amb visió reduïda, revisió d'aplicacions existents i selecció preliminar de tecnologies (visió, veu i frameworks mòbils).

F2 Prototip inicial amb *LiveKit* *(mar–abr2025, completada).* Es va construir un demostrador CLI que enviava flux continu d'àudio i vídeo al servidor mitjançant LiveKit i rebia respostes d'un GPT-4o genèric. Va servir per provar la viabilitat tècnica però es va descartar per culpa de la gran quantitat de consum de dades i la impossibilitat de poder crear un agent personalitzat. Donava respostes genèriques del model GPT-4o.

F3 Creació de una app flutter per a poder organitzar les sessions dels usuaris a Firebase amb *(abr–juny2025, en curs).* S'ha creat una app multiplataforma (Android/iOS) amb quatre pestanyes —Sessió en directe, Només Càmera, Historial i Configuració— i amb mesures d'accessibilitat integrades (tema clar/fosc, text fins al 250 %, vibració i control per gest).

F4 Migració de LiveKit a arquitectura diferent *(maig–juny2025, completada).* – Es canvia l'arquitectura per un flux *push-to-talk*: es grava l'àudio, es captura un sol fotograma quan l'usuari deixa de parlar i s'envien tots dos al nou *Assistant* assistent: Keia, agent personalitzat, preparat per a rebre consultes. – Cada usuari disposa del seu propi *thread* emmagatzemat a Firebase Firestore, de manera que les converses queden contextualitzades i persistents.

F5 Integració de detecció visual (Grounding DINO i YOLOv8n) *(juny2025, completada).* – Es va implementar un sistema de detecció dirigida amb Grounding DINO per permetre a l'usuari localitzar objectes concrets mitjançant una ordre verbal com ara "on són les claus?", escanejant l'espai amb diverses imatges i retornant una resposta guiada. – Com a millora de darrera hora, es va afegir un pipeline lleuger amb YOLOv8n per detectar objectes en temps real mentre la càmera està activa. Aquest sistema permet una exploració hàptica de l'escena: si l'usuari passa el dit per damunt d'una bounding box, se li llegeix l'etiqueta de l'objecte detectat. Es tracta d'una primera versió experimental que obre la porta a futures millores com un "Braille visual" accessible des de la pantalla.

F6 Avaluació i refinament *(juny2025, completada).* Proves amb usuaris per mesurar latència i satisfacció; filtratge d'al·lucinacions, ajust del to empàtic i millora del pipeline TTS/STT. Un cop validades les metriques, es publicarà la versió candidata.

7 TESTS I AVALUACIÓ

Durant el desenvolupament de *Keia*, s'han realitzat diverses proves per validar-ne el funcionament correcte, la coherència de les respostes generades, i la seva adequació a les necessitats d'usuaris amb discapacitat visual.

Proves tècniques de backend

- **Test del servidor FastAPI**: s'han realitzat proves unitàries i d'integració per assegurar que els endpoints responen correctament amb els formats requerits, especialment en el pipeline d'àudio, d'imatge i d'embeddings.
- **Validació del sistema de fils (threads)**: s'ha verificat que cada usuari manté una conversa persistent, i que el backend relaciona correctament cada pregunta amb el seu context anterior.
- **Test de detecció amb Grounding DINO i YOLOv8n**: s'ha comprovat que les deteccions retornen bounding boxes amb confiança adequada i que la selecció de les imatges òptimes en escaneig es realitza segons la millor coincidència semàntica.

Validació funcional i de resposta

- **Respostes generades per Keia**: s'ha testejat que les respostes multimodals generades per l'agent compleixin els criteris d'utilitat, empatia i claredat. Es va fer servir un conjunt de preguntes típiques per simular escenaris d'ús (per exemple: "Què hi ha sobre la taula?", "On són les meves ulleres?").
- **Coherència espacial**: s'ha revisat que les respostes incloguin referències relatives (com "a la dreta", "davant teu") i no meres descripcions abstractes.
- **Validació del *prompt engineering***: es va ajustar iterativament fins obtenir un estil de resposta natural, que evités tecnicismes i estructures artificials, seguint les directrius de Zhao et al. (2019).

Proves d'accessibilitat i usabilitat

- **Compatibilitat amb TalkBack**: s'ha verificat que totes les pestanyes i botons siguin accessibles amb TalkBack activat, i que els elements es llegeixin amb l'etiquetatge correcte.

- **Proves de mides de text i contrast**: s'ha comprovat que l'aplicació es manté usable amb qualsevol mida de lletra entre 140% i 250%, i amb els dos temes (clar/fosc) segons la configuració de l'usuari.
- **Temps de resposta**: les latències d'inferència s'han mantingut per sota dels 5-7s en la majoria de casos, depenent de la qualitat de la connexió i del tipus de petició.

8 Conclusions

Aquest treball ha donat lloc al desenvolupament d'una aplicació funcional i accessible que apropa la intel·ligència artificial a les persones amb discapacitat visual. *Keia* neix amb l'objectiu de ser una eina empàtica, contextual i útil, i les diferents funcionalitats implementades —descripció visual, detecció d'objectes, reconeixement de veu, historial persistent i interfície adaptada— han estat pensades amb aquest propòsit.

La integració de models com GPT-4o i Grounding DINO ha permès oferir respostes multimodals contextuals, millorant la comprensió de l'entorn per part de l'usuari. A més, la inclusió recent del pipeline amb YOLOv8n ha obert la porta a una forma alternativa i contínua d'exploració hàptica, amb potencial per a futurs sistemes embarcats.

Un dels aspectes més rellevants ha estat l'estratègia de *prompt engineering*, decisiva per adaptar el comportament del model a les necessitats específiques del col·lectiu. L'aplicació ha estat validada amb èxit tant tècnicament com des del punt de vista de l'accessibilitat, mantenint una latència acceptable i una usabilitat adequada en totes les seves pestanyes.

Finalment, aquest projecte no només ha representat un repte tècnic i de disseny, sinó també una aproximació real i empàtica a les necessitats d'un col·lectiu sovint oblidat pel desenvolupament tecnològic. Les proves realitzades demostren que és possible construir interfícies conversacionals multimodals amb aplicabilitat directa i valor social tangible.

9 Possibles millores i línies futures

Tot i que l'aplicació actual ofereix una experiència completa i funcional per a l'assistència visual, s'han identificat algunes millores que podrien ampliar-ne les prestacions en futures versions:

- **Detecció hàptica més avançada:** el sistema de detecció contínua amb YOLOv8n es podria millorar amb models més precisos o entrenats amb dades reals de cada usuari, per fer-lo més fiable en entorns domèstics.
- **Execució parcialment offline:** implementar funcionalitats bàsiques com la detecció d'objectes o lectura de text sense connexió, mitjançant models lleugers embarcats, podria millorar-ne la disponibilitat en contextos sense cobertura.
- **Avaluació amb usuaris finals:** una prova pilot amb persones amb visió reduïda permetria validar l'efectivitat del sistema en situacions reals i detectar ajustos de to o funcionalitat que podrien millorar encara més l'experiència d'ús.

Aquestes línies d'evolució no responen a mancances crítiques, sinó que apunten a un desenvolupament progressiu i enfocat a l'adaptació personalitzada i la robustesa de l'assistent.

Agraïments

Vull agrair en primer lloc al meu tutor, per totes aquelles sessions del principi en què cap dels dos ho vèiem del tot clar, ja que jo tenia masses idees i no sabia exactament per on començar. Per donar-me pau i ajudar-me a ordenar el meu cap ple d'idees inacabables.

En segon lloc, als meus pares i a la meva germana, per ajudar-me en tot aquest llarg procés i fer-me veure que, al final, tot acabaria sent possible.

Als meus amics, per tots els dies de biblioteca en què els explicava que no sabia si això seria útil, i ells em motivaven a fer-ho possible.

En particular a l'Aitor, per ajudar-me a trobar el nom de l'aplicació mentre dinàvem a la gespa de la UAB, i al Pere, que només pel fet de voler sortir als agraïments s'ha passat tots aquests mesos donant-me suport.

I, per últim, i el més important: al meu avi. Gràcies per tot el que has fet per mi aquests anys. M'hauria fet molta il·lusió provar l'aplicació amb tu i ensenyar-te allò que m'has motivat a fer.

Referències

[1] W. H. Organization, "World report on vision," 2019. [Online]. Available: https://www.who.int/

[2] ——, "Vision impairment and blindness," 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment

[3] E. B. Union, "Statistics on blind and partially sighted people in europe," 2023. [Online]. Available: https://www.euroblind.org/

[4] I. N. de Estadística, "Discapacidad en españa," 2022. [Online]. Available: https://www.diariosanitario.com

[5] W. H. Organization, "Blindness and vision impairment," 2022. [Online]. Available: https://www.who.int

[6] WHO and UNICEF, "Global report on assistive technology," 2022. [Online]. Available: https://www.who.int

[7] ONCE, "Tecnología y discapacidad visual." [Online]. Available: https://www.once.es

[8] L. Guild, "Smartphone use by people with vision loss," 2021. [Online]. Available: https://www.lhpb.org

[9] J. Rokicki and et al., "Mobile applications for blind and low vision users," *Annual Review of Vision Science*, vol. 9, 2023.

[10] H. Kacorri and et al., “Enabling people with visual impairments to interact with technology,” *Communications of the ACM*, vol. 63, no. 6, pp. 70–77, 2020.

[11] S. A. Zhao and et al., “Designing multimodal apps for people with visual impairments,” in *Proceedings of ASSETS ’19*, 2019.

[12] J. Rokicki and et al., “Mobile applications for blind and low vision users,” *Annual Review of Vision Science*, 2023.

[13] F. Equipara, “Map4all - mapa col·laboratiu d’accessibilitat,” 2014, accessed: 2025-06-25. [Online]. Available: https://www.map4all.com/

[14] Neosentec, “Lazzus - personal assistant for the visually impaired,” 2016, accessed: 2025-06-25. [Online]. Available: https://lazzus.com/en/

[15] B. M. Eyes, “Be my eyes - lend your eyes to the blind,” 2015, accessed: 2025-06-25. [Online]. Available: https://www.bemyeyes.com/

[16] TapTapSee, “Taptapsee - mobile camera application designed for the blind and visually impaired,” 2013, accessed: 2025-06-25. [Online]. Available: https://www.taptapseeapp.com/

[17] LazarilloApp, “Lazarillo gps for the blind,” 2017, accessed: 2025-06-25. [Online]. Available: https://www.lazarillo.app/

[18] M. Research, “Seeing ai - talking camera app for the blind community,” 2017, accessed: 2025-06-25. [Online]. Available: https://www.microsoft.com/en-us/ai/seeing-ai

[19] G. LLC, “Lookout - assisted vision by google,” 2021, accessed: 2025-06-25. [Online]. Available: https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.reveal

[20] V. App, “Voicevista - gps guidance for the blind,” 2023, accessed: 2025-06-25. [Online]. Available: https://apps.apple.com/us/app/voicevista/id6445925255

[21] Canadian National Institute for the Blind, “Removing barriers to accessibility: Our 2022 annual report,” 2022, accessed: 2025-06-25. [Online]. Available: https://www.cnib.ca/en/about-cnib/reports/annual-report-2022

RECURSOS WEB CONSULTATS

- `https://developer.android.com/guide/topics/ui/accessibility` — Guia oficial d’accessibilitat d’Android.
- `https://material.io/design` — Recomanacions de disseny Material Design.
- `https://www.w3.org/WAI/standards-guidelines/wcag` — Directrius WCAG sobre contingut web accessible.
- `https://docs.flutter.dev` — Documentació oficial de Flutter.
- `https://platform.openai.com/docs` — API d’OpenAI per a models multimodals.
- `https://platform.openai.com/docs/assistants/deep-dive#managing-threads-and-messages` — Gestió de *threads* amb Assistants.
- Hackernoon – tutorial de creació d’un agent OpenAI — Guia pas a pas en Python.
- `https://www.youtube.com/watch?v=7xTGNNLPyMI` — Vídeo ≪Deep Dive into LLMs like ChatGPT≫.
- `https://bbycroft.net/llm` — Visualització interactiva d’una arquitectura LLM.
- `https://arxiv.org/abs/2203.02155` — Article sobre RLHF per ajustar models amb feedback humà.