Real or Fake? - A Mobile Game for AI Content Detection

Oriol Oliva Riera

Resum—En els últims anys, cada vegada són més populars arreu del món les eines de generació de contingut amb intel·ligència artificial, fins al punt de formar part del día a día de molta gent. Aquesta tecnologia, si bé no és nova, ha anat evolucionant durant els últims anys, i especialment, durant l'última dècada gràcies a l'increment de potència computacional. El projecte abarca una aplicació mòbil d'un joc que ensenya imatges a l'usuari i li proposa una pregunta: És aquesta imatge IA? L'usuari respondrà el que consideri i, en cas d'encertar, s'incrementarà la seva puntuació, mentre que en cas de fallar s'acabarà la partida. L'objectiu amb aquesta aplicació és observar el comportament i la capacitat d'encert d'una diversitat d'usuaris davant diferents imatges, i comparar la taxa d'encert i error de diferents models de generació d'imatges, a més concienciar a la gent de l'existència d'aquestes imatges i educar en com identificar-les.

Paraules clau— IA, Intel·ligència Artificial, Dataset, GenAI, LLM, Deepfake, React, React Native, GenAI-Bench, Dall·E 3, Midjourney, DeepFloyd, StableDiffusion

Abstract—For the last few years, tools for content generation through artificial intelligence are getting more popular, to the point of becoming part of our daily basis. This type of technology, if not brand new, has seen a lot of growth these past few years, and especially, during the last ten years with the increase in computational power. This project is based on making a phone app of a game showing images to the user, and asking them: Is this image AI? The user then has to answer according to what they consider, and in case they are correct, the game will increase their score and they are allowed to keep playing, however, if they fail it is game over. The aim with the app is to test and observe the capabilities of different users to try and identify images, and compare their success rate and error rate over different image generation models, as well as raise awareness about the existence of these images, and to educate on how to identify them.

Index Terms— AI, Artifficial Intelligence, Dataset, GenAI, LLM, Deepfake, React, React Native, GenAI-Bench, Dall·E 3, Midjourney, DeepFloyd, StableDiffusion

1 Introducció - Context del treball

Quan parlem d'eines de generació de contingut amb intel·ligència artificial, habitualment estem parlant d'eines de GenAI, o Intel·ligència Artificial Generativa. Aquesta, com bé diu el nom, és capaç d'entendre i analitzar uns valors d'entrada i generar una resposta que correspongui, sigui un text de resposta o potser, una imatge.

Dins d'aquest tipus de IA, tenim doncs eines de LLM (Large Language Model), que són models gegants entrenats a partir d'infinitat de text per a poder entendre, interpretar i expressar-se en un llenguatge humà. Amb aquesta tecnologia podem crear els chatbots que han revolucionat el món durant l'últim parell d'anys, com Chat GPT [5] i Gemini [6], que són un exemple de LLM que interpreta un missatge enviat per l'usuari, i a partir del que entén i el que troba, genera una resposta que espera satisfer a l'usuari.

Una altra tecnologia que també ha avançat en els últims 5 anys són els models de Text-to-Image [7]. Aquests, que són els models que tenen una relació directa amb l'objectiu d'aquest treball, funcionen a partir d'un input: un text que l'usuari envia al model amb les instruccions del que vol que el model representi, que serà enviat i interpretat per aquest model, i que generarà una imatge, output, acord a l'input de l'usuari. Això es possible ja que el model Text-to-Image, a partir de l'ús de Machine Learning, ha après a relacionar una imatge amb un prompt associat, de manera que al finalitzar el periode d'aprenentatge pot fer aquesta relació de manera inversa i crear una imatge a partir d'un prompt proposat per l'usuari.

Estem vivint un moment únic i revolucionari a la història,

- *E-mail de contacte: Oriol.OlivaR@autonoma.cat*
- *Menció realitzada: Enginyeria del Software*
- *Treball tutoritzat per: Alexandra Gomez Villa (Departament de Ciències de la Computació)*
- *Curs 2025/26*

del que molt possiblement se'n parli al futur gràcies a tots els avenços tecnològics, però no tot és positiu en relació amb aquest tema. El fet de que qualsevol persona disposi d'accés a aquestes eines, i les pugui utilitzar per a generar articles, informació, imatges i vídeos que semblen reals, però que no ho són és una realitat amb la que ens hem d'afrontar i no estem suficientment educats per a diferenciar entre contingut real o generat amb completa certesa.
Per altra banda, els deepfakes [10] son imatges, vídeos o audio que ha estat modificat o generat des de zero amb intel·ligència artificial i que mostres o bé persones reals, o personatges ficticis.
Aquests tenen cada vegada més presència a les xarxes socials, i si bé n'hi ha creats amb fins humorístics i que clarament són identificables com a tal, n'hi ha d'altres on es pot veure una figura pública declarant comentaris que mai han sortir de la seva boca, i en casos pot ser molt difícil distingir si això que s'està visualitzant és o no real [11]. També hi ha casos on es viola la privacitat de les persones amb aquestes eines, amb desenes de notícies de deepfakes a escoles generant imatges no desitjades sobre companyes de classe.
Altrament, també s'ha de considerar l'àmbit artístic, doncs aquests models de generació de contingut, específicament d'imatges, poden generar una imatge a partir d'un prompt, paraules claus introduïdes per a guiar al model, amb l'estil 'robat' d'un artista que durant anys ha practicat i que en molts casos, no ha donat el seu permís.
En aquest tema, cada vegada podem trobar més casos de reclamacions per drets d'autor a aquests models, que com exposen els reclamants, utilitzen milers d'imatges sense permís per a entrenar aquests models. Casos d'aquest estil importants en aquest àmbit poden ser el cas 'Getty Images v Stability-AI'[12] o bé, un cas centrat en artistes contra aquestes grans empreses d'intel·ligència artificial: Andersen v. Stability AI Ltd.[13], que reclama que les empreses Stability AI, DeviantArt i Midjourney han estat utilitzant art de diferents artistes sense cap mena de permís per a entrenar els seus models.
La intenció que abarca aquest projecte és crear consciència de que cada vegada aquests models de generació d'imatges són més avançats, i que cada cop és més complicat diferenciar entre si una imatge és real o no, i intentar educar en com poder identificar una imatge generada a partir de dades i entrenament vía un joc d'identificació d'imatges.
Aquest joc serà una aplicació mòbil que presentarà a l'usuari una imatge i li donarà dues opcions a escollir: Aquesta imatge és real, o la ha fet una IA?
En cas d'encertar, l'usuari guanyarà punts, mentre que si no encerta, haurà perdut la partida.

2 Estat de l'art

2.1 Estudi de joc

Si bé no existeix cap aplicació mòbil similar, si que s'han trobat dues pàgines web amb un joc molt semblant. El primer a observar és: AIGuessGame [14].

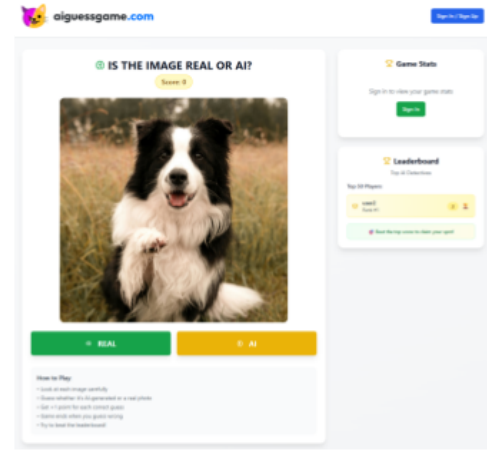


Figura 1. Interfície de aiguessgame.com

En aquest cas, podem veure que el layout de la pantalla és molt similar a la idea d'aquest projecte, però en el nostre cas li donarem un enfoc afegit, amb el que guardarem estadístiques sobre el percentatge d'encert dels usuaris, per a poder analitzar diferents perfils d'usuari i com de bé ho fan. A més, al ser la proposta d'aquest projecte un joc d'aplicació mòbil, un major nombre de persones hi podrà tenir accés en comparació amb aquesta pàgina web.

El segon exemple es: Fakeout.dev [15].

En aquest segon exemple, podem veure la mateixa dinàmica de joc, amb dos modes diferents.
El primer i el que es mostra només entrar a la pàgina és el mode video, on es pregunta a l'usuari: Quin d'aquests dos vídeos és generat per IA? A continuació, mostra la categoria del prompt i les dues imatges mostrades, per a guiar lleugerament als usuaris, i al centre de la pantalla tenim els dos videos en bucle. Per a continuar, l'usuari ha de clicar al vídeo que cregui que es generat, i en cas d'encertar es sumarà un punt a la puntuació. A més, al respondre es mostra per pantalla estadístiques d'encert dels usuaris que ja han participat a la mateixa pregunta, i fonts d'ambdós vídeos, per al video generat per IA en podem veure el prompt, i per al video real en podem veure els crèdits. Per a continuar jugant, caldrà clicar al botó de 'Següent ronda', amb el que carregarà la següent parella de videos, fins un total de 5 rondes que es reinicien un cop per setmana.

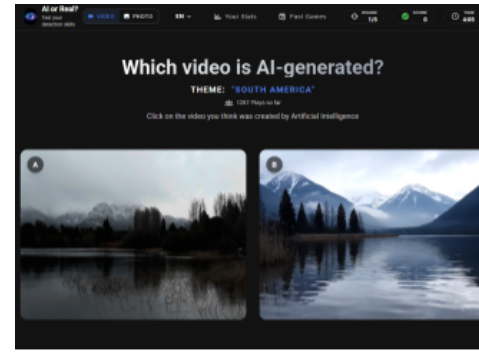


Figura 2. Interfície de fakeout.dev

Al acabar la partida, la pàgina mostra la puntuació de l'usuari i el temps que ha tardat, i ens convida a provar l'altre mode de la pàgina.

Aquest altre és el mode d'imatges, on, semblant a la pàgina anterior, es mostra dues imatges i s'ha d'identificar quina és la generada per IA. Igual que amb el mode anterior, al completar la ronda mostrarà estadístiques dels usuaris que ja han participat i la puntuació final.
Cal tenir en compte que aquesta segona pàgina explica motius pels quals s'ha creat el joc: convertir aquestes imatges en una activitat interactiva que serveixi com a eina per a desenvolupar pensament crític envers el contingut generat per intel·ligència artificial, i poder navegar per aquest món amb confiança.
La principal diferència entre aquests dos jocs és el disseny, els modes de joc (imatges o vídeos) i la duració de partides. El primer deixa jugar infinitament, mentre que el segon només deixa jugar un nombre de rondes limitades a la setmana, una aproximació més semblant a jocs que han guanyat popularitat en els últims anys, com Wordle [16]. Tothom té la mateixa ronda al joc, i aquesta s'actualitza al mateix temps per a tothom. Així, els usuaris poden competir directament entre ells i compartir les puntuacions a les xarxes.

2.2 Estudi d'investigació

Per altra banda, ja s'han fet diversos estudis sobre la capacitat humana d'identificar imatges generades per IA, i de distingir-les d'imatges reals.

2.2.1 Interpretation of AI-Generated vs. Human-Made Images [17]

A l'estudi "**Interpretation of AI-Generated vs. Human-Made Images**", observem com es realitza un estudi semblant al que es vol fer per aquest projecte. Es disposa de 32 imatges, amb diferents classificacions de les quals 24 són generades pels 3 models de generació d'imatges per Intel·ligència Artificial següents: Midjourney, Dalle_3 i Firefly. Per altra banda, tenen 8 imatges reals, és a dir, fetes per humans. Totes aquestes imatges s'agrupen en 4 categories diferents, per a classificar les dades dels resultats que s'aconseguiran al finalitzar l'estudi: retrat humà, paisatge, escena quotidiana i objectes detallats.
Al distribuir les imatges a usuaris i aconseguir un total de 5152 respostes, s'obtenen els següents resultats:
61.08% d'encert sobre les imatges generades per IA, mentre que les imatges reals obtenen un 78.26% d'encert.

A més, també s'entra en detall per a analitzar el percentatge d'encert a les diferents categories, començant per la primera, retrat humà. A aquesta categoria, els resultats de l'experiment sorprenentment obtenen un percentatge d'encert del 54.35% per les imatges generades, mentre que el percentatge d'encert sobre les imatges humanes és del 79.81%.
A la categoria de paisatges, els resultats milloren en relació als retrats, doncs s'aconsegueix un 65.74% de percentatge d'encert a les imatges generades, mentre que s'obté un percentatge similar per a les imatges reals, d'un 68.32%.
A la següent categoria, escenes quotidianes, aconsegueixen uns resultats similars a la primera categoria, amb un percentatge del 62.94% d'encert a les imatges generades, mentre que aconsegueixen un 79.81% d'encert a les imatges humanes.
A l'última categoria, objectes detallats, aconsegueixen un encert del 61.28% a les imatges generades i un 84.47% d'encert a les imatges humanes, classificant com a primera categoria en aquest àmbit.
Amb aquests resultats, poden arribar a la conclusió de que la majoria d'usuaris té certa dificultat a l'hora d'identificar si una imatge és generada per un model, resaltant la alta capacitat de realisme que aquests models poden aconseguir, tot i l'estat en el que es troben, sent un camp de desenvolupament molt nou i amb una ràpida evolució.
També resalten la falta de balanç en el nombre d'imatges reals en base a les generades, que pot afectar a les respostes dels usuaris, però es realitza per a representar l'estat de la xarxa a internet, on la predominancia de les imatges generades és cada día més real.
A aquest projecte, es pot observar com l'objectiu és aconseguir uns resultats qualitatius, més que quantitatius. Disposen d'un nombre limitat d'imatges, models i usuaris, escollits específicament per a fer-ne aquest estudi, i així poder veure en els pocs exemples què fa que als usuaris els resulti més o menys complicat identificar una imatge correctament.

2.2.2 Seeing is not always believing: Benchmarking Human and Model Perception of AI-Generated Images [18]

Aquest estudi es centra també en observar la percepció dels humans davant imatges generades, però també enfrontant els resultats dels humans amb models d'identificació d'imatges generades, que són i seran cada vegada més prevalents en línea, per a verificar l'origen d'una imatge.
Disposen d'un dataset format per imatges reals d'internet i imatges generades pels següents models: StableDiffusion [19], StyleGAN3 [20] i DeepFloyd IF [21].
Amb aquest dataset, entrenen un model d'identificació d'imatges generades, i fan l'experiment amb els usuaris, aconseguint els següents resultats:
Els participants identifiquen correctament un 61.3% de les imatges, on el participant amb millor puntuació va ser capaç d'identificar un 73% de les imatges correctament. A més, comparant amb la seva expectativa de que els participants serien capaços d'identificar el 100% d'imatges reals, troben un resultat del 66.9% d'encert a imatges reals, a l'hora de fer-ne la classificació.
En aquest cas, igual que a l'anterior estudi, tenen també diferents exemples de classificació d'imatge, comprenent les següents categories: multiperson,a paisatge, home, dona, record, planta, animal i objecte.

Amb els participants humans, a les diferents categories aconsegueixen els següents resultats:

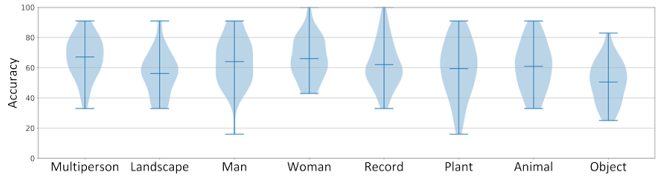


Figura 3. Distribució de la puntuació d'avaluació humana en vuit categories utilitzant totes les dades dels participants. Seeing is not always believing: Benchmarking Human and Model Perception of AI-Generated Images [18]

On podem observar que totes les categories tenen al voltant d'un 60% d'encert medià, i mentre que les categories amb figures humanes presents són les que millors resultats han obtingut, les categories amb objectes i paisatges són les pitjor classificades.

Per últim, i com a tancament del projecte, posen a prova el model d'identificació d'imatges generades que han entrenat amb el seu propi dataset d'imatges. Posant a prova el millor model, a un experiment que inclou 50 imatges reals i 50 imatges generades, el model pot aconseguir un 87% de percentatge d'encert.

A aquest projecte, l'enfocament principal es aconseguir resultats d'identificació d'imatges en base a diferents categories d'imatges, i comparar els resultats humans obtinguts amb els resultats d'un model d'identificació d'imatges per a veure quin dels dos grups obté millors resultats, és a dir, un percentatge d'encert a l'hora de classificar una imatge correctament més elevat.

2.2.3 How good are humans at detecting ai-generated images? Learnings from an experiment [22]

Per últim, a aquest experiment també es presenta una base de dades amb imatges generades per intel·ligència artificial i imatges fetes per humans.

D'un total de 287.269 imatges, els participants de l'experiment en classifiquen correctament el 62%, mentre que si ens fixem únicament en les imatges generades per IA observem un percentatge d'encert del 63%, obtenint uns resultats en que els usuaris són lleugerament millors que si es respon a l'experiment a partir de tirar una moneda.

A més, aquest estudi a diferència dels altres, presenta un gràfic de densitat d'encert, on es pot observar que la gran majoria d'imatges tendeixen a ser encertades entre un 60 i un 80% de les vegades, i que hi ha molt poques imatges que siguin consistentment mal classificades, doncs la població de la distribució a partir del 40% de l'encert cau en picat.

Per altra banda, també analitzen si la qualitat d'una imatge pot afectar a la capacitat d'identificació dels participants de l'experiment, que utilitzen al seu anàlisis qualitatiu, observant les úniques 3 imatges reals amb un encert menor al 20%, i comparant-les amb les imatges generades per ia amb el menor percentatge d'encert de l'experiment.

Igual que a l'anterior (2.2.2), es compara la capacitat d'identificar i classificar correctament imatges entre els humans participants i models de detecció d'imatges generades per ia. En aquest cas, trobem una comparació entre el 62% d'encert dels usuaris de l'experiment, contra un 95% d'encert de les eines de detecció.

De nou i per a tancar, ressalten que el 62% d'encert és poc millor que respondre a l'experiment tirant una moneda, i fan una crida a la transparència del contingut generat a la xarxa, considerant l'alarmant increment de la quantitat d'aquest tipus de contingut i que, com es pot observar és i serà cada vegada més difícil d'identificar i distinguir del contingut real.

A aquest projecte, en contrast del primer analitzat (2.2.1), es dona un enfoc més quantitatiu que qualitatiu a l'estudi i experiment. Al disposar d'un major nombre d'imatges, es pot aconseguir uns resultats més similars a un cas real amb el que es trobaria un usuari, per exemple, navegant per les xarxes socials. A aquestes xarxes, hi podem trobar milers d'imatges de tota mena, des de reals, a imatges generades amb models molt bàsics i per tant, senzilles d'identificar, o bé tot el contrari, imatges generades amb models d'última generació.

3 Objectius

- Crear un joc funcional
- Crear un sistema de puntuació al joc per a incentivar a l'usuari
- Mostrar una taula de puntuacions
- Emmagatzemar dades sobre encert dels usuaris
- Generar estadístiques a partir de les dades d'encert dels usuaris

4 Metodologia

4.1 Eines de desenvolupament

Cal concretar una eina a la que desenvolupar el projecte, en el meu cas utilitzaré Visual Studio Code [23] ja que considero que es un programa molt versàtil amb el que podré utilitzar diferents llenguatges de programació i tots els plugins que consideri necessaris durant el desenvolupament.

En relació amb els llenguatges de programació:

- Pel Frontend utilitzaré React Native [24], un llenguatge molt versàtil per a poder crear una aplicació multiplataforma adaptable
  A més, amb el llenguatge de React Native, utilitzaré el framework Expo [25]: un framework senzill que em facilitarà la creació de l'aplicació i la transició a generar una apk d'Android per a realitzar les proves i l'experiment
- Pel Backend: utilitzaré Python, un dels llenguatge més utilitzats al món de la intel·ligència artificial
- Per a la gestió de bases de dades, utilitzaré la eina Supabase, amb PostgreSQL, què em servirà per a emmagatzemar les dades de partida dels usuaris de prova, per a l'ús analític posterior

4.1 Datasets

El cor de l'aplicació són les imatges que es presentaran a l'usuari per a que pugui seleccionar si considera que són reals o no, de manera que aquestes imatges les treuré de les bases de dades següents:

- GenAI-Bench [26]: Base de dades d'imatges generades, de 6 models diferents
- relaion2b-natural (Roth and Hebart) [27]: Base de dades d'imatges naturals

Inicialment, vaig utilitzar el dataset DiffusionDB [28] per a extreure les imatges generades, però en fases inicials del projecte es canvia a GenAI-Bench al ser una base de dades més moderna i completa.

5 Desenvolupament

5.1 Inici i layout

El desenvolupament de l'aplicació ha tingut els següents passos:
Primer de tot, la creació de l'aplicació mòbil en si mateixa. Disposem d'un disseny simple, amb una disposició de 3 pantalles:

- una pantalla Menú que ens servirà com a índex per a entrar a les altres dues
- una pantalla de Joc, on mostrarem les imatges i l'usuari respondrà
- i una pantalla amb la Taula de resultats

Aquesta aplicació, com ja s'ha comentat es va desenvolupar en React Native, amb l'ajuda d'Expo com a framework base de React Native, que facilita el desenvolupament oferint-nos un model i eines base per a començar el desenvolupament.
Al disposar de l'esquelet de l'aplicació, ofert pel projecte base de Expo, cal muntar el layout que tindrà l'aplicació, que com ja s'ha comentat començarà pel menú, des del qual podem anar a la pantalla de joc o a la pantalla de puntuacions, mentre que des d'aquestes dues podrem tornar a la pantalla de menú en cas d'anar enrere.

5.2 Funcionalitat de joc

El següent pas es omplir-ne les funcionalitats. La pantalla de joc ha de ser senzilla i fàcil d'entendre, de manera que disposarà d'un títol: "És aquesta imatge IA?". Això, ho acompanyarem del número que representarà la puntuació de la partida actual, la imatge sobre la qual s'ha de respondre, i un botó afirmatiu i un botó negatiu. En cas de que l'usuari encerti, la puntuació s'incrementarà en un punt i la pantalla mostrarà la següent imatge.
Quan l'usuari no premi el botó correcte, el joc finalitzarà i un menú apareixerà per pantalla preguntant a l'usuari si vol guardar la puntuació i associar-la a un nom de jugador.
A l'apèndix 1, es mostra imatges de l'estat final de l'aplicació desenvolupada, mostrant 4 figures per a les 4 possibles pantalles del joc final.

5.3 Càrrega d'imatges

La següent part del desenvolupament suposa la càrrega d'imatges per a que el joc pugui funcionar. La base de dades GenAI-bench, disposa de 6 models diferents de IA generativa: DALLE_3, DeepFloyd_I_XL_v1, Midjourney_6, SDXL_2_1, SDXL_Base i SDXL_Turbo.
A aquesta base de dades, i específicament a la utilitzada per a aquest projecte, GenAI-bench1600, disposem de 1600 prompts per a generació d'imatges. Cada un d'aquests prompts s'utilitzen als 6 diferents models per a generar 6 imatges diferents en funció del mateix prompt. Utilitzant aquest coneixement com a base, per a simplificar les dades de l'aplicació, decideixo utiltizar els 100 primers prompts, per a un total de 600 imatges. Per a igualar aquest nombre d'imatges doncs, també caldrà utilitzar 600 imatges 'reals' del dataset relaion2b-natural, sobre les quals decideixo escollir-ne les primeres 600 amb una puntuació natural de 0.7 o superior, per recomanació dels creadors del dataset.
Així doncs, per a les proves amb usuaris reals disposarem de 1200 imatges, suficients per a que sigui poc probable que es repeteixin imatges entre partides d'un mateix usuari.
Al tenir un nombre reduït d'imatges, la solució més senzilla per a utilitzar-les a la aplicació és penjar-les a una base de dades o un repositori en línea. En aquest cas, vaig utilitzar el mateix respositori del projecte per a pujar les 600 imatges categoritzades com a ai, i les 600 imatges categoritzades com a real. A més, cada imatge té unes metadades associades, id i link, però per a les imatges IA també disposem del model que les ha generat i el prompt utilitzat.
Amb aquestes metadades, podrem realitzar la següent part del projecte, que es preparar l'enviament de dades de les partides al backend, la base de dades del projecte a Supabase. De nou, a aquesta base de dades recollim les següents. .

5.4 Test

Per últim, s'han realitzat les següents proves per a verificar el funcionament correcte de l'aplicació finalitzada. Al ser una aplicació simple, si bé podem fer proves automatitzades vía el framework d'Expo per a verificar-ne el correcte renderitzat de les diferents pantalles, i que la connexió entre aquestes sigui correcta, al projecte m'he centrat més en les proves d'usuari.

Al necessitar proves amb usuari per a la recol·lecció de dades també es poden utilitzar les proves d'usuari per a comprovar el correcte funcionament del joc, si els usuaris troben intuitiva la interfície i les funcionalitats sense prèvia explicació, si l'enviament de dades a la base de dades és correcte durant les partides que fan els usuaris, i si les dades es corresponen amb els resultats que es mostren per pantalla.

6 Resultats

6.1 Resultats generals

Per a obtenir resultats, s'ha provat amb diferents usuaris el joc, i les estadístiques de partides dels usuaris són els següents:
D'un total de 35 usuaris de prova, i unes 2011 imatges mostrades al llarg de totes les partides, tenim els següents resultats: 1637 encerts i 374 errors
Això implica, com podem veure a la figura 4, un **81.4% d'encert** per part dels usuaris a l'hora d'identificar les diferents imatges, sense distingir entre imatges reals o d'IA.

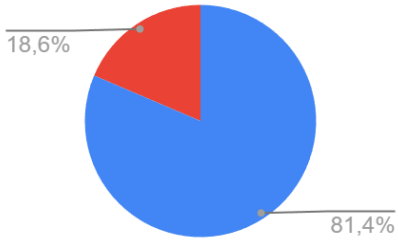


Figura 4. Estadístiques d'encert totals

Per altra banda, del total de 2011 imatges, 1025 eren imatges generades per algun dels 6 models d'intel·ligència artificial generativa comentats anteriorment.

Sobre aquestes imatges, tal i com es mostra a la figura 5, tenim els següents resultats:
825 encerts i 200 errors.

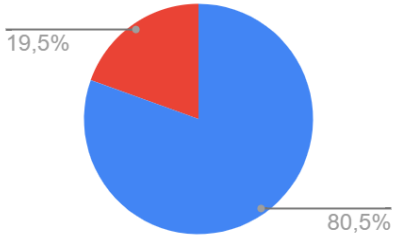


Figura 5. Estadístiques d'encert a les imatges generades

Amb això, podem observar que el percentatge d'encert disminueix lleugerament, però pràcticament es manté, quedant en un **80.5% d'encert**.

Curiosament, si observem els resultats de les imatges reals, també obtenim un percentatge d'encert similar, aconseguint un **82.4% d'encert**, amb 812 encerts i 174 errors

6.2 Resultats per model

Si ens fixem en els 200 errors obtinguts a les imatges generades per IA, aconseguim la següent distribució de resultats:

- DALLE_3: 9 errors d'un total de 177 imatges d'aquest model mostrades
- SDXL_Base: 22 errors d'un total de 159 imatges d'aquest model mostrades
- Midjourney: 28 errors d'un total de 171 imatges d'aquest model mostrades
- SDXL_Turbo: 37 errors d'un total de 169 images d'aquest model mostrades
- SDXL_2: 43 errors d'un total de 193 imatges d'aquest model mostrades
- DeepFloyd: 51 errors d'un total de 154 imatges d'aquest model mostrades

Amb això, vol dir que tenim la següent taxa d'encert per model, com es pot observar a la figura 6:

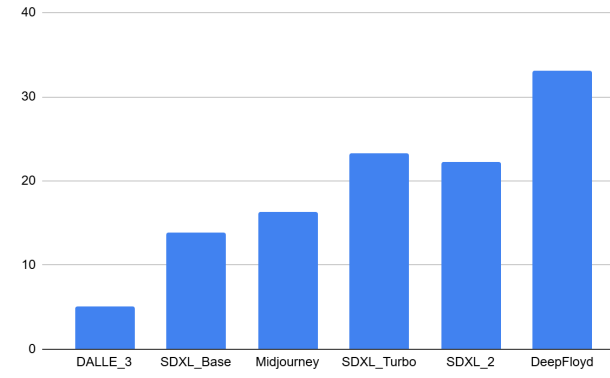


Figura 6. Percentatge d'errors per model d'IA generativa

Amb els resultats obtinguts durant l'experiment i les proves del joc, podem observar com el model que més ha enganyat als usuaris de prova ha estat **DeepFloyd**, amb un percentatge de 33% d'error, és a dir, 1 de cada 3 imatges d'aquest model enganyava als usuaris, fent que marquessin la imatge com a real.

6.3 Resultats per edat

Dins dels usuaris de l'experiment, les distribucions de puntuació mitjana en base a l'edat han estat les següents:

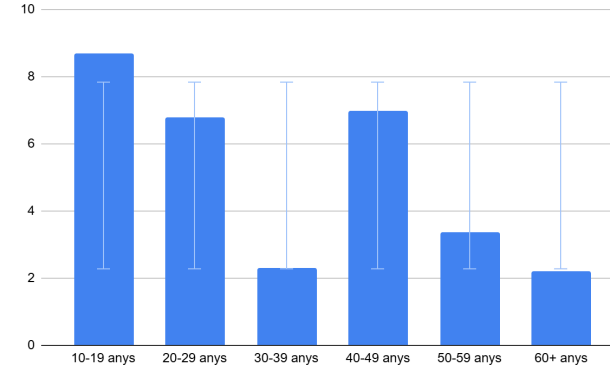


Figura 7. Puntuació mitjana per franja d'edat

6.4 Observacions

Cal tenir en compte, que durant l'experiment s'ha pogut observar que els usuaris tendeixen a la millora, segons el número de partides jugades va augmentant.
La majoria de persones obtenen un resultat de entre 2 i 4 punts a la seva primera partida, sent la mediana de puntuació total de l'experiment de **3**.
Això pot ser degut a diversos factors, però principalment s'atribueix a la pràctica i la soltesa que pot tenir un usuari en reconèixer diferents patrons que comprenen les imatges generades menys avançades.

També, cal tenir en compte que les dades estan limitades per una distribució no equilibrada dels usuaris a les diferents franges, però és una de les limitacions del projecte.
21 de 35 persones estaven a la franja de 20-29 anys, i les mitjanes de puntuació de 10-19 i 40-49 són més elevades degut a que tenien 2 i 1 persones, respectivament.

Per altra banda, el fet de que tot i tenir 1200 imatges, com hem pogut veure hi ha 600 imatges d'IA de les quals 6 tenen el mateix prompt. En cas de que un usuari es trobi amb una d'aquestes imatges amb prompt repetit, pot considerar que ja ha vist la mateixa imatge, 'repetida', encara que no s'hagi repetit la mateixa imatge.

Per últim, cal analitzar quines imatges o quin tipus d'imatge feia confondre més als usuaris de l'experiment, i per què.
Primerament, si ens basem en els resultats d'imatges generades per IA, observem que les imatges amb una composició simple, i sobretot que contenien paisatges, són les que més sovint han enganyat als usuaris.

D'aquest tipus d'imatge, en podem observar els següents exemples:

Aquesta primera imatge apareix 3 vegades a les proves, i d'aquestes 3 vegades a totes la resposta dels diferents usuaris ha estat que aquesta imatge era 'real'.

Figura 8. Una bicicleta vermella davant d'una paret blava

Prompt: "A red bicycle against a blue wall."
Model: DeepFloyd

La següent imatge, de nou, presenta una escena semblant.

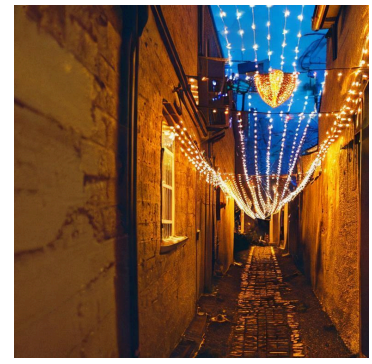


Figura 9. Un estret carreró il·luminat per cordes de llums decoratives.

Prompt: "A narrow alleyway illuminated by strings of fairy lights."
Model: SDXL_2
Novament, aquesta imatge es va mostrar 3 vegades durant les proves, i a les 3 va ser classificada incorrectament pels usuaris.

Per últim dins aquest mateix tipus d'imatge, tenim la següent:

Figura 10. Una fila de cases colorides a un dia solejat.

Prompt: "A row of colorful townhouses on a sunny street."
Model: DeepFloyd
De nou, 3 errors a 3 mostres de la imatge.

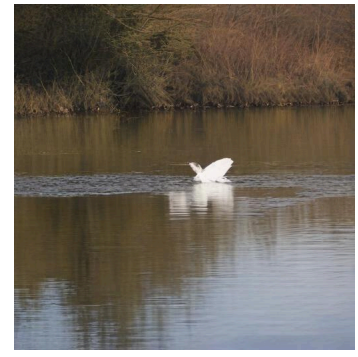


Figura 11. Un estany petit amb un cigne sol que es mou cap a l'esquerra.

Prompt: "A small pond with a single swan gliding towards the left."
Model: SDXL_Turbo
3 errors a 3 vegades que s'ha mostrar la imatge

Més exemples d'imatges generades amb alta taxa d'error a l'Annexe 2.

També hi ha hagut errors per part dels usuaris en la identificació de les imatges reals, que a vegades enganyaven els usuaris. Aquestes imatges, com podem veure a continuació, també són majoritariament paisatges, amb algún element no ordinari que generava dubtes.

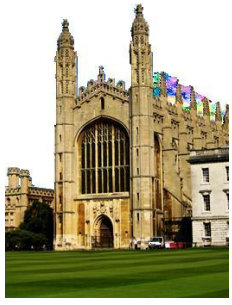


Figura 12. Imatge real d'una església.

Una església que de 3 vegades que es va mostrar als usuaris, 3 vegades que aquests van interpretar que la imatge era generada.

Figura 13. Imatge real d'una cabanya de fusta.

Igual que l'anterior, aquesta cabanya de 3 vegades que va ser mostrada, 3 vegades que va ser classificada com a imatge generada per IA.
De nou, més exemples d'imatges reals amb alta taxa d'error a l'apartat d'Annexes 3.

I per últim, també cal veure exemples de les imatges que no han enganyat als usuaris.
La majoria d'aquestes imatges estan composades per figures humanes, normalment amb algun tipus d'error de generació que es pot diferenciar fàcil, o bé són imatges amb conceptes tant extravagants que no poden ser reals, i tampoc enganyen als usuaris per a semblar un dibuix.
Cal tenir en compte, també, que moltes d'aquestes imatges tenen una estètica similar, que poc a poc es va popularitzant a les imatges generades que podem trobar a les xarxes socials, i que pot facilitar-ne la identificació.

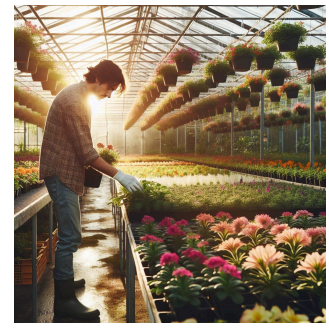


Figura 14. Un jardiner cuidant flors en un hivernacle ple de llum.

Imatge amb una figura humana, i l'estètica 'd'imatge generada' habitual a les xarxes. Demostra que aquestes imatges són comuns, i per tant, fàcilment identificables pels usuaris de l'experiment.
Prompt: "A gardener tending to flowers in a greenhouse filled with sunlight."
Model: Dalle_3
De 4 vegades que es va mostrar, té una taxa del 100% d'encert per part dels usuaris.

Altres exemples d'imatges amb alta taxa d'encert, com ja hem comentat presenten un estil fantàstic, que no convenç als usuaris, com la imatge següent:

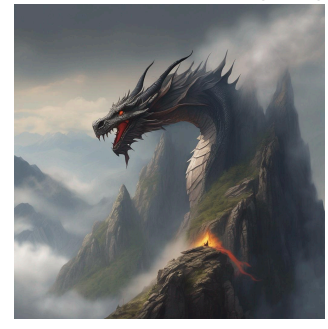


Figura 15. Un drac col·locat majestuosament a una muntanya rocosa i coberta de fum.

Prompt: "A dragon perched majestically on a craggy, smoke-wreathed mountain."
Model: SDXL_Base
3 encerts per part dels usuaris, amb una taxa del 100%.

Per últim, un altre tipus d'imatge fàcilment identificable és aquella que conté alguna mena de text, de cartell o de qualsevol tipus de lletra, que habitualment si forma part de la composició d'un paisatge, contindrà lletres sense sentit o malformades. Un exemple d'una imatge així és la següent:

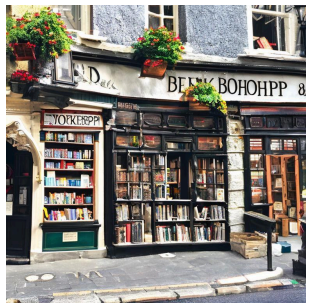


Figura 16. Una botigua de llibres peculiar.

Prompt: "A quaint bookshop."
Model: SDXL_2_1
4 encerts de 5 vegades que es va mostrar la imatge.

6 Limitacions i pròxims passos

Al projecte hi ha hagut certes limitacions, ja sigui per falta de temps, coneixements o infraestructura, a continuació llistaré les possibles pròximes millores que es poden dur a terme al projecte:

- Gestió d'usuaris: Actualment, a l'aplicació no es fa cap mena de gestió del perfil d'un usuari. Es guarda el resultat de la partida amb el sobrenom que l'usuari escull, si decideix guardar-ne els resultats a la taula de puntuació, i d'aquesta manera es podria emmagatzemar més informació personalitzada sobre un mateix usuari, millor puntuació visible de manera senzilla, registre i personalització del compte, entre d'altres opcions

- Escalabilitat: Si bé una distribució de 1200 imatges és un nombre suficient per a fer proves amb un nombre limitat d'usuaris sense que les partides tendeixin a mostrar les mateixes imatges, és un nombre que no ofereix una infinitat de partides. Per això, un pas futur seria el d'ampliar la base de dades a un nombre major d'imatges, amb una infraestructura de backend que ho pugui soportar.

- Diversitat a les dades: En les proves, s'ha tractat de diversificar al màxim dins de les possibilitats el tipus d'usuari de proves, però els recursos són limitats. En cas d'ampliar els recursos i publicar l'aplicació, la diversitat d'usuaris es veurà incrementada, poden arribar així a obtenir uns resultats més correctes.

- Dataset modern: Cal resaltar la dificultat de disposar d'un dataset actualitzat per a fer l'experiment, doncs en qüestió de setmanes els datasets es poden quedar obsolets. En cas de disposar d'un dataset amb imatges generades més modernes i avançades, és probable que els resultats de l'experiment empitjorin.

7 Conclusió

Si ve els resultats són optimistes, cal tenir en compte les limitacions de les proves i la ràpida evolució de les tecnologies d'intel·ligència artificial generativa. No és cap misteri que fa uns anys dur a terme aquest mateix experiment hagués obtingut un resultat encara més elevat a favor dels encerts humans degut a la pobre qualitat de les imatges generades pels models de la època, que no enganyaven a ningú.
Si tenim en compte els resultats, observem un 80% d'encert general davant les imatges d'IA, però cal considerar que els models utilitzats per a aquest projecte, i les imatges utilitzades, no estan actualitzades a les últimes tecnologies de dia d'avui, al venir un dataset generat a l'any 2024, amb models ja existents abans. L'avenç d'aquestes tecnologies és exponencial, cada vegada les xarxes socials estan més plenes de contingut generat, i cada vegada és més difícil per a l'usuari distingir què és real i què no.
Observem també, que setmanes després de l'experiment, els participants comenten ser més conscients a l'hora de navegar per les xarxes, i tenir cert pensament crític i dubtar a l'hora de veure certes imatges i vídeos que segueixen els patrons mostrats per imatges generades a l'experiment.
De nou, els resultats obtinguts a l'investigació són optimistes, per tant encara estem a temps de preparar a l'usuari per al dia de demà, doncs l'avenç és inevitable i el futur és la IA generativa, però també cal recordar d'on s'originen aquestes imatges, i que és real i que no.

Ús de la IA Generativa

Per a ajudar amb el desenvolupament del projecte, s'ha utilitzat l'ús de la IA generativa per a revisar el codi i format desenvolupat de l'aplicació, i per a facilitar el desenvolupament de la càrrega de datasets a la base de dades.

Agraïments

Vull agrair primerament a la meva família i amics, que han ajudat tant en el procés de proves d'usuari com en la part experimental com a subjectes de prova, i a la meva parella que a més, també m'ha ajudat a dissenyar el logo de l'aplicació i l'estil. També vull agrair a la meva tutora tota la informació que m'ha facilitat, ajudant molt a aconseguir bones fonts per a verificar l'estat de l'art i les bases de dades utilitzades al projecte.

Bibliografia

[1] IBM Technology, 2024, Aug 5. AI, Machine Learning, Deep Learning and Generative AI Explained. [Video]. Youtube.AI, Machine Learning, Deep Learning and Generative AI Explained
[2] Vox, 2022, Jun 1. AI, Machine Learning, Deep Learning and Generative AI Explained. [Video]. Youtube.AI art, explained
[3] TED, 2025, Jul 18. How to Spot Fake AI Photos | Hany Farid | TED. [Video]. Youtube.How to Spot Fake AI Photos | Hany Farid | TED
[4] IBM Technology, 2025, Jan 30. Diffusion Models for AI Image Generation. [Video]. Youtube.Diffusion Models for AI Image Generation

[5] ChatGPT: Radford, A., & Narasimhan, K., 2018. Improving Language Understanding by Generative Pre-Training.
[6] Gemini: Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J. B., Yu, J., ... & Hauth, A. (2024). Gemini: A family of highly capable multimodal models, 2024. arXiv preprint arXiv:2312.11805, 10.
[7] Text-to-image Ramadevi, Y., Vanapalli, A., & Kashyap, K. V. H. (2025, March). A Systematic Review of Text-to-Image and Image-to-Image Synthesis Models. In 2025 6th International Conference on Recent Advances in Information Technology (RAIT) (pp. 1-6). IEEE.
[8] What is Text-to-Image? - Hugging Face- https://huggingface.co/tasks/text-to-image
[9] Google Research, 2023, Jan 19. Text-to-image generation explained. [Video]. Youtube.Text-to-image generation explained
[10] Deepfake-Wikipedia- https://en.wikipedia.org/wiki/Deepfake
[11] The Deepfake detection Challenge: Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Canton Ferrer, C. (2019). *T*he Deepfake Detection Challenge (DFDC) preview dataset. arXiv. https://arxiv.org/abs/1910.08854
[12] Getty Images v Stability AI: https://www.judiciary.uk/wp-content/uploads/2025/11/Getty-Images-v-Stability-AI.pdf
[13] https://www.loeb.com/en/insights/publications/2024/08/andersen-v-stability
[14] AIGuessGame: An interactive game where you guess if an image is AI Generated or not. https://aiguessgame.com/
[15] Fakeout.dev: AI or Real Game. Spot AI Generated Videos & Images. Fakeout. https://fakeout.dev/
[16] Wordle- https://www.nytimes.com/games/wordle/index.html
[17] Roca, T., Cintron Roman, A., Torres Vega, J., Duarte, M., Wang, P., White, K., Misra, A., & Lavista Ferres, J. (2025). How good are humans at detecting AI-generated images? Learnings from an experiment. arXiv. https://arxiv.org/abs/2507.18640
[18] Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X., & Ouyang, W. (2023). Seeing is not always believing: Benchmarking human and model perception of AI-generated images. arXiv. https://arxiv.org/abs/2304.13023
[19] Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., ... & Rombach, R. (2024, July). Scaling rectified flow transformers for high-resolution image synthesis. In Forty-first international conference on machine learning.
[20] Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-free generative adversarial networks. Advances in neural information processing systems, 34, 852-863.
[21] Deepfloyd IF: https://github.com/deep-floyd/IF
[22] Velásquez-Salamanca, D., Martín-Pascual, M. Á., & Andreu-Sánchez, C. (2025). Interpretation of AI-Generated vs. Human-Made Images. Journal of imaging, 11(7), 227. https://doi.org/10.3390/jimaging11070227
[23] Visual Studio Code: https://code.visualstudio.com/
[24] ReactNative.dev: https://reactnative.dev/docs/getting-started Pàgina web oficial de desenvolupament de React Native. Disposa de tota la informació necessària per a iniciar un desenvolupament amb el framework.
[25] Expo: https://docs.expo.dev/ Pàgina web oficial de documentació sobre el framework
[26] GenAI-Bench: Li, Baiqi and Lin, Zhiqiu and Pathak, Deepak and Li, Jiayao and Fei, Yixin and Wu, Kewen and Ling, Tiffany and Xia, Xide and Zhang, Pengchuan and Neubig, Graham and others. "GenAI-Bench: Evaluating and Improving Compositional Text-to-Visual". 2024. arXiv, https://arxiv.org/abs/2406.13743.
[27] Relaion2b-natural: Roth, Johannes, and Martin N. Hebart. "How to Sample the World for Understanding the Visual System." 8th Annual Conference on Cognitive Computational Neuroscience, 2025, https://openreview.net/forum?id=T9k6KkZoca.
[28] DiffusionDB: Wang, Zijie J., Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A Large-Scale Prompt Gallery Dataset for Text-to-Image Generative Models. 2022. arXiv, https://arxiv.org/abs/2210.14896.

APÈNDIX

A1. Imatges de la interfície de l'aplicació

Figura 1. Interfície de la pantalla Menú de l'aplicació:.

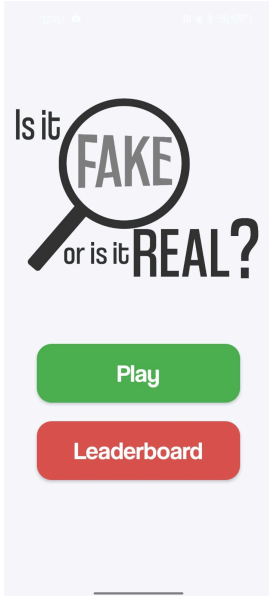


Figura 2. Interfície de la pantalla de Joc de l'aplicació:

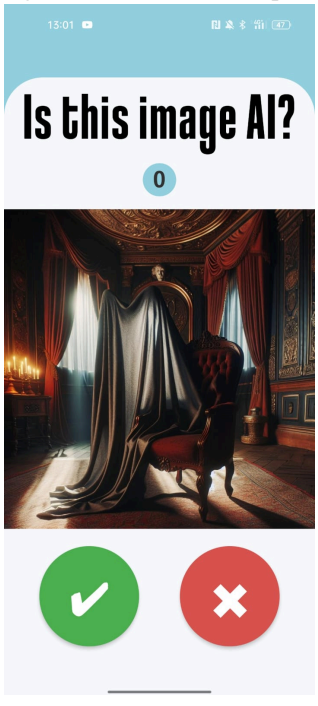


Figura 3.Interfície de la pantalla de fi de joc de l'aplicació:

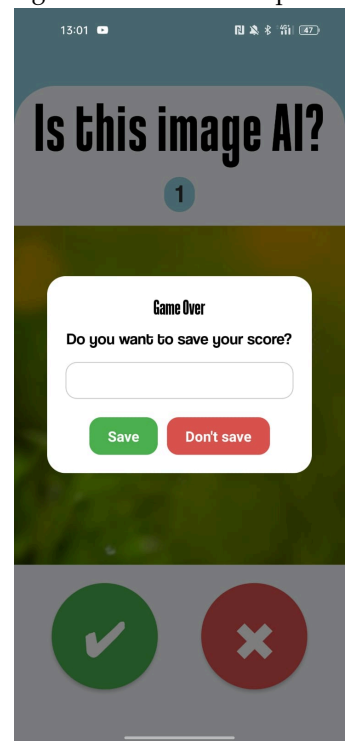
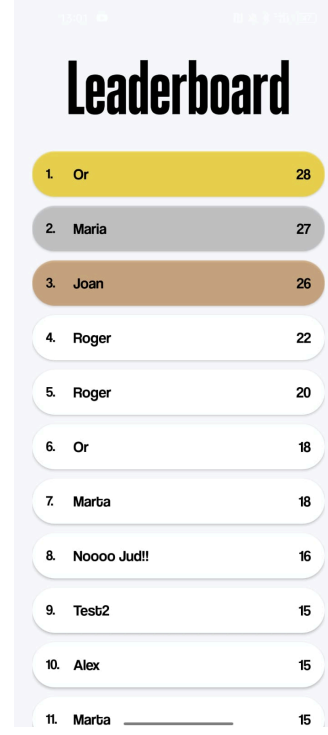
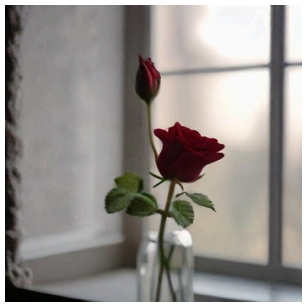


Figura 4. Interfície de la pantalla de puntuació de l'aplicació:



A2. Imatges amb elevada taxa d'error d'IA



Prompt: "A single red rose in a vase on the right side of a windowsill."
Model: SDXL_Turbo
3 de 3 errors

Prompt: "A butterfly perched on a wildflower in a meadow."
Model: SDXL_Base
3 de 3 errors

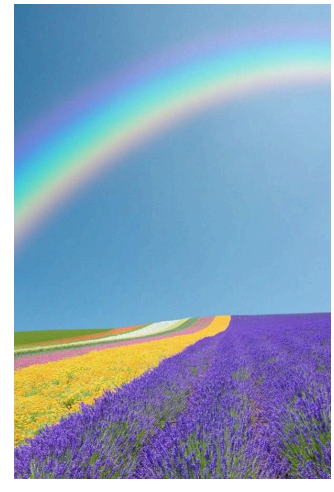
Prompt: "A lone lighthouse standing guard on a rocky coastline."
Model: DeepFloyd
3 de 4 errors

Prompt: "A garden path lined with glowing stones under a twilight sky."
Model: DeepFloyd
3 de 4 errors

A3. Imatges amb elevada taxa d'error Reals

Imatges reals que semblen altament artificials:

4 vegades que s'ha mostrat, 4 vegades que s'ha classificat com a imatge generada



5 vegades que s'ha mostrat, 3 vegades que s'ha classificat com a imatge generada

3 vegades que s'ha mostrat, 3 vegades que s'ha classificat com a imatge generada

A4. Imatges senzilles d'identificar

Més exemples d'imatges generades que els usuaris no han tingut problema en identificar durant les proves

Prompt: "A fairy dancing lightly atop a blooming, moonlit flower."
Model: DeepFloyd
3 de 3 encerts

Prompt: "A phoenix soaring above a city, aglow with golden flames."
Model: SDXL_Base
3 de 3 encerts

Prompt: "A sorcerer's hat casting shadows over a cluttered, enchanted desk."
Model: Midjourney_6
3 de 3 encerts

Prompt: "A dog tunes a violin."
model: Midjourney_6
4 de 4 encerts

Prompt": "A mermaid singing softly near a coral throne undersea."
Model: SDXL_Base
5 de 5 encerts