

# Eines Intel·ligents pel Resum i Generació de Contingut Audiovisual en Notícies



## CIDAI-PAI 2024-01

Drets reservats. Aquest treball està disponible sota la llicència Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Segons els termes d'aquesta llicència, podeu copiar, redistribuir i adaptar l'obra amb fins no comercials, sempre que l'obra sigui citada adequadament, tal com s'indica a continuació.

En qualsevol ús d'aquest treball, no s'ha de suggerir que el CIDAI doni suport a cap organització, producte o servei específic. No es permet l'ús del logotip CIDAI.

Si adapteu l'obra, heu de llicenciar-la amb la mateixa llicència Creative Commons o equivalent.

Si creeu una traducció d'aquest treball, heu d'afegir la següent exempció de responsabilitat juntament amb la cita suggerida: "Aquesta traducció no la va crear el Centre of Innovation for Data tech and Artificial Intelligence (CIDAI). CIDAI no es fa responsable del contingut ni de l'exactitud d'aquesta traducció. L'edició original en català serà l'edició autèntica i vinculant".

Qualsevol mediació relacionada amb disputes derivades de la llicència es durà a terme d'acord amb les normes de mediació de la World Intellectual Property Organization.

Cita suggerida. CIDAI-PAI 2024-01// Eines Intel·ligents pel Resum i Generació de Contingut Audiovisual en Notícies per Rafael Redondo Tejedo – Eurecat, Bogdan Raducanu – CVC i José Adrián Rodríguez Fonollosa – IDEAI UPC, CIDAI, 2024. Llicència: CC BY-NC-SA 4.0

---

<b>Títol:</b>	Eines intel·ligents pel resum i generació de contingut audiovisual en notícies
<b>Responsables tècnics:</b>	Rafael Redondo Tejedor <a href="mailto:rafael.redondo@eurecat.org">rafael.redondo@eurecat.org</a>  Bogdan Raducanu <a href="mailto:bogdan@cvc.uab.cat">bogdan@cvc.uab.cat</a>  José Adrián Rodríguez Fonollosa <a href="mailto:jose.fonollosa@upc.edu">jose.fonollosa@upc.edu</a>
<b>Data:</b>	01-03-2024
<b>Prescriptor:</b>	Corporació Catalana de Mitjans Audiovisuals (CCMA)
<b>Receptor:</b>	Rafael Bermúdez Product Owner equip de Recerca i Serveis Transversals Corporació Catalana de Mitjans Audiovisuals (CCMA) <a href="mailto:rbermudez.h@3cat.cat">rbermudez.h@3cat.cat</a>
<b>Adreça:</b>	C. d'Oriol Martorell, 1 08970 Sant Joan Despí, Barcelona <a href="http://www.ccma.cat">www.ccma.cat</a>
<b>Contextualització:</b>	Aquest es un projecte que s'ha fet en el marc dels projecte d'Alt Impacte del CIDAI a partir del reptes proposat per la CCMA i ha estat desenvolupat per Eurecat, UPC i CVC.

---

# Índex

1. INTRODUCCIÓ .....	5
1.1. Models Grans de Llenguatge (LLMs).....	6
1.2. Generació d'Imatges amb Models de Difusió .....	7
1.3. Reconeixement de la Parla .....	8
1.4. Funcionalitats i Interacció.....	8
2. RESUM AUTOMÀTIC DE NOTÍCIES.....	9
2.1. Transcripció de la Parla .....	11
2.2. Extracció Automàtica de Frases Més Significatives.....	15
2.3. Definició de Funcionalitats.....	22
2.4. Detalls de la Implementació Final .....	23
2.5. Interfície d'Usuari UX .....	24
2.6. Avaluació Final d'Usuaris PT1 .....	25
2.7. Conclusions PT1 .....	27
3. NARRACIÓ GRÀFICA GENERATIVA DE NOTÍCIES .....	28
3.1. Resum Abstractiu de Notícies .....	30
3.2. Generació Automàtica d'Il·lustracions .....	31
3.3. Conjunt de Notícies i Àudios .....	33
3.4. Selecció i Experimentació amb Models de Llenguatge i Imatge.....	34
3.5. Interfície d'Usuari i Control d'Estil.....	39
3.6. Avaluació Final d'Usuaris PT2 .....	41
3.7. Conclusions PT2 .....	42

# Introducció



En una època on la informació és abundant i el consum de notícies s'ha convertit en plataformes multimèdia, la necessitat d'eines de resum eficients s'ha fet més pronunciada que mai. El nostre projecte s'embarca en un viatge transformador a la intersecció de la tecnologia i el periodisme, introduint eines automàtiques d'avantguarda dissenyades per a la síntesi integral del contingut de vídeo i la il·lustració dinàmica, tot impulsat pel poder dels mètodes d'aprenentatge profund.

En l'àmbit de la difusió de notícies, els vídeos s'han convertit en un mitjà potent, que ofereix narracions amb una barreja d'elements visuals i auditius. Tanmateix, el gran volum de contingut de vídeo suposa un repte important tant per al públic com per als creadors de contingut. Aquest repte és on intervé el nostre projecte, aprofitant les capacitats de l'aprenentatge profund per destil·lar automàticament l'essència de les notícies de vídeo en resums concisos i informatius.

El nostre esforç va més enllà del simple resum, ja que reconeixem el paper fonamental de la il·lustració per millorar la comprensió. Mitjançant la utilització de tècniques avançades d'aprenentatge profund, pretenem generar de manera dinàmica representacions visuals que encapsulin els elements clau de les notícies. Això no només serveix per enriquir l'experiència del consumidor, sinó que també permet als creadors de contingut transmetre informació complexa amb claredat i impacte.

A mesura que ens endinsem en l'àmbit de les eines automàtiques per al resum de notícies, ens embarquem en la missió de redefinir com el públic interacciona amb el contingut de notícies de vídeo. En combinar perfectament els àmbits de la intel·ligència artificial i el periodisme, el nostre projecte aspira a contribuir a una experiència de consum de notícies més eficient i visualment atractiva. En definitiva, desenvolupem eines innovadores per crear noves formes de narració, revolucionant la manera com interactuem amb les notícies a l'era digital.

## 1.1. Models Grans de Llenguatge (LLMs)

Els grans models de llenguatge (LLM) són potents models d'intel·ligència artificial que s'han entrenat amb quantitats massives de dades de text per entendre i generar un llenguatge semblant a l'ésser humà. Una de les capacitats clau dels LLM és la seva capacitat per realitzar resums de text, un procés que consisteix a destil·lar les idees principals o la informació clau d'un determinat fragment de text.

El resum extractiu consisteix a seleccionar i extreure les frases o frases més importants del text original per crear un resum. Els LLM utilitzen representacions del llenguatge prèviament entrenats per entendre el context i el significat del text d'entrada. El model assigna puntuacions d'importància a cada frase o frase del text en funció de factors com la rellevància, la informació i la coherència. Les frases amb puntuacions d'importància més altes es seleccionen per formar el resum extractiu. Els LLM poden utilitzar mecanismes d'atenció per centrar-se en parts clau del text d'entrada durant el procés de resum.

El resum abstractiu implica generar un resum concís d'una manera que pot implicar parafrasejar i reformular el contingut original, utilitzant potencialment frases o expressions noves. Els LLM aprofiten la seva comprensió del llenguatge per generar noves frases que capturen el significat essencial del text d'entrada. El model pot utilitzar incrustacions de paraules conscients del context i mecanismes d'atenció per produir resums coherents i contextualment rellevants. El resum abstractiu permet una major flexibilitat a l'hora d'expressar el resum, però requereix una comprensió més profunda del contingut.

Creat amb ChatGPT4.



Il·lustració de resum Extractiu vs. Abstractiu dels models de Processament de Llenguatge Natural.



Font <https://www.abstractivehealth.com/extractive-vs-abstractive-summarization-in-healthcare>.

Comprensió contextual: els LLM són excel·lents en la captura dels matisos contextuals del llenguatge, cosa que els permet crear resums que van més enllà de la mera extracció. Els LLM es poden ajustar per a dominis específics, fent-los adaptables a diferents tipus de contingut, inclosos articles de notícies, articles de recerca i molt més. Els LLM poden manejar una àmplia gamma d'idiomes i adaptar-se a diversos estils d'escriptura.

D'altra banda, els LLM poden mostrar biaixos presents a les dades d'entrenament, i la qualitat dels resums es pot veure influenciada per la diversitat i la representativitat del corpus de formació. Tot i que els LLM han fet avenços significatius, garantir que els resums generats siguin alhora coherents i informatius segueix sent un repte.

## 1.2. Generació d'Imatges amb Models de Difusió

Els models de difusió són una classe de models generatius utilitzats per a la síntesi d'imatges generant mostres mitjançant una sèrie de transformacions iteratives. Un cop entrenat, el model de difusió pot generar noves mostres aplicant les transformacions apreses en ordre invers en el procés d'iteració.

Els models de difusió poden generar imatges i seqüències diverses i realistes capturant distribucions de dades complexes a les dades d'entrenament. El procés de difusió permet controlar les mostres generades fins a cert punt, permetent la creació de vídeos amb diferents nivells de soroll o incertesa. No obstant, el control de la coherència temporal, persistències de cert elements i control de la distribució espacial dels elements encara és un repte considerable.

Creat amb ChatGPT4.



### 1.3. Reconeixement de la Parla

Els mètodes de transformació de veu a text (STT), també coneguts com a reconeixement automàtic de veu (ASR), impliquen la conversió del llenguatge parlat en text escrit. Aquesta tecnologia té nombroses aplicacions, des de serveis de transcripció fins a assistents virtuals activats per veu.

Les arquitectures de transformadors, dissenyades originalment per al processament del llenguatge natural, s'han aplicat amb èxit a ASR. Permetent una formació d'extrem a extrem, eliminant la necessitat de components separats com ara models de llenguatge i acústics. Aquests models utilitzen un mecanisme d'atenció per centrar-se en diferents parts de la seqüència d'entrada, alineant-la amb la seqüència de sortida corresponent.

Els sistemes ASR han de gestionar les variacions d'accent, ritme de parla, soroll de fons i diferents idiomes. Per millorar l'eficiència de aquests models s'utilitzen tècniques com la perturbació de la velocitat, l'addició de soroll de fons i la reverberació per augmentar les dades d'entrenament i millorar la robustesa del model. També, els models ASR es poden afinar o adaptar a dominis específics per millorar el rendiment en contextos especialitzats.

Els mètodes de veu a text han estat testimonis d'avanços significatius, especialment amb l'augment de les tècniques d'aprenentatge profund. La millora contínua de les arquitectures de models, les estratègies d'entrenament i els mètodes de pre-processament de dades, contribueix a l'evolució de sistemes de reconeixement de veu cada vegada més precisos i robusts.

### 1.4. Funcionalitats i Interacció

En el panorama dinàmic de les experiències digitals, les funcionalitats d'experiència d'usuari (UX) són la pedra angular per crear interaccions significatives entre els usuaris i la tecnologia. Les funcionalitats UX engloben un conjunt divers de funcions i capacitats meticulosament dissenyades per millorar la satisfacció dels usuaris, la usabilitat i el plaer general en l'ús de productes i serveis digitals. A mesura que la tecnologia evoluciona i les expectatives dels usuaris augmenten, la comprensió i la implementació de funcionalitats efectives d'UX s'han convertit en primordials per als dissenyadors, desenvolupadors i empreses per igual.

En la seva essència, les funcionalitats UX fan referència a la varietat d'eines, elements i interaccions integrades en una interfície digital per optimitzar el viatge de l'usuari. Aquestes funcionalitats s'estenen més enllà de la mera estètica, i s'endinsen en els àmbits de l'accessibilitat, la navegació, la capacitat de resposta i la interactivitat. Un conjunt ben dissenyat de funcionalitats UX garanteix que els usuaris no només compleixin les seves tasques de manera eficient, sinó que també ho facin amb una sensació de compromís i gaudiment.

Creat amb ChatGPT4.





# Resum Automàtic de Notícies



El resum automàtic de vídeo és el procés de condensar un vídeo llarg en una versió més curta alhora que conserva el seu contingut més important i rellevant. Les tècniques de resum de vídeo poden variar en complexitat i enfocament, i es poden classificar en dos tipus principals:

- **Extracció de fotogrames clau:** seleccionar fotogrames representatius del vídeo que capturen escenes o moments importants.
- **Detecció de salts de escena:** la detecció dels límits dels plans ajuda a identificar els canvis en l'escena o els angles de la càmera, cosa que pot ajudar a resumir vídeos seleccionant plans representatius.
- **Reconeixement d'objectes/acció:** el reconeixement d'objectes o accions poden identificar i seleccionar fotogrames que contenen objectes o accions importants, proporcionant un resum més basat en contingut semàntic.
- **Extracció de funcions:** es basa en extreure funcions dels fotogrames de vídeo, l'àudio i les metadades del text. Aquestes funcions poden incloure histogrames de color, vectors de moviment, energia d'àudio i molt més.

Aquestes tècniques requereixen de un sistema de classificació posterior de manera que els fotogrames s'agrupen en grups en funció de la seva similitud de característiques.

Alternativament, l'enfocament per aprenentatge profund inclou Xarxes neuronals recurrents (RNN) que s'utilitzen per capturar dependències temporals en dades de vídeo i generar resums seleccionant fotogrames seqüencialment. També, Xarxes neuronals convolucionals (CNN) per extreure característiques visuals, que es poden utilitzar per identificar fotogrames o plans importants per a un resum.

En general, molts mètodes de resum de vídeo combinen diverses tècniques per millorar la qualitat del resum. Per exemple, combinar l'extracció de fotogrames clau amb la classificació basada en contingut pot produir resums ben arrodonits.

El resum automàtic de vídeo és una àrea de recerca activa amb desenvolupaments constants tant en mètodes tradicionals com en mètodes basats en l'aprenentatge profund. L'elecció de la tècnica depèn de factors com el tipus de contingut de

vídeo, la durada desitjada del resum i els recursos computacionals disponibles.

La dificultat d'aquest repte i la seva resolució mitjançant tècniques de visió per ordinador, no obstant radica en que el contingut visual sovint no té tota la informació semàntica necessària per entendre el contingut audiovisual. En el cas particular dels diàlegs és clar que una gran part significació radica en l'àudio i en particular en l'anàlisi de la parla.

D'altra banda, els grans models de llenguatge (LLM) són una classe de models d'intel·ligència artificial dissenyats per entendre i generar llenguatge semblant a l'ésser humà. Aquests models aprofiten quantitats massives de dades de text per aprendre patrons, context i semàntica del llenguatge. Els LLM han demostrat capacitats notables en diverses tasques de processament del llenguatge natural (NLP), com ara traducció d'idiomes, generació de text, resposta a preguntes, anàlisi de sentiments i molt més.

És per tant, que en aquests casos d'ús concrets, la generació de resums audiovisuals basat en la transcripció de la notícia combinat amb l'anàlisi de text de LLMs pot oferir una solució efectiva.

#### Proposta de Valor:

- El resum de vídeo és un procés per crear un resum concís i coherent d'un vídeo més llarg seleccionant moments, escenes o contingut clau. En el cas particular de mitjans de comunicació el resum de vídeo és valuós per crear versions més curtes i atractives de notícies, entrevistes o esdeveniments més llargs.
- Aquest projecte ha desenvolupat tecnologia innovadora basada en grans models de llenguatge (LLMs) per l'assistència a la generació automàtica de resums audiovisuals.
- Les eines desenvolupades han sigut integrades en una interfície (UX) amigable de fàcil instal·lació i ús que permet la interacció intuïtiva i personalització del contingut segons les necessitats de l'usuari.

En aquest paquet de treball ens hem enfocat en la generació de resums de notícies mitjançant la combinació de models intel·ligents de transcripció (parla a text) i les capacitats de resum de models intel·ligents de processament de llenguatge natural NLP per inferir quines parts son informativament més rellevants.

Flux de treball pel resum audiovisual de notícies: reconeixement de la parla (speech to text), resum extractiu mitjançant LLM, personalització de l'usuari, localització temporal del resum i composició del video.



## 2.1. Transcripció de la Parla

En aquesta tasca es fan servir models de transcripció de la parla a text STT que avui en dia ofereixen una tecnologia molt madura. En aquest pas el registre de les metadades extrems amb els codis de temps dona informació del moment en que cadascuna de les paraules ha sigut pronunciada en la pista d'àudio.

### 2.1.1. Model Whisper

Actualment el programa Whisper<sup>1</sup> desenvolupat per la companyia americana OpenAI és el sistema de referència en transcripció de la parla. Aquest model és compatible amb més de 50 idiomes diferents, inclosos els principals idiomes mundials com l'anglès, el francès, l'alemany, l'espanyol, el mandarí, l'àrab, el japonès, el rus i el català. Whisper tradueix directament l'idioma demanat.

El codi i els models de transcripció i traducció de veu Whisper desenvolupat per OpenAI està disponible de forma gratuïta sota la llicència MIT. La llicència MIT és una llicència permissiva que en permet l'ús comercial sense cap cost ni restricció.

<sup>1</sup> <https://openai.com/research/whisper>

Des que es va publicar el codi del sistema Whisper, diferents investigadors i empreses han desenvolupat diverses variants de Whisper per intentar millorar alguns aspectes com ara el cost computacional (*faster-whisper*) o la qualitat de la segmentació en paraules (*whisperX*).

Tot i això, la mateixa empresa OpenAI també ha anat traient noves versions de Whisper per millorar aquests aspectes. És per això que la nostra recomanació és utilitzar el programari mantingut per OpenAI si el temps de càlcul no és important.

Si la velocitat de processament és important o es vol obtenir una diarització, és a dir, una assignació automàtica de cada frase als diferents locutors que intervien en l'enregistrament, aleshores el programa WhisperX és la millor opció. WhisperX combina una versió ràpida de Whisper, una segmentació basada en detecció de silencis, un alineat de les paraules més precís, i un sistema de diarització de locutors.

### 2.1.2. Avaluació en Conjunt de Notícies

En utilitzar Whisper, l'usuari pot seleccionar diferents models en funció de la qualitat desitjada. Els models més grans proporcionen més qualitat, però són també els que tenen més cost computacional i transcriuen amb menys velocitat. Per executar Whisper amb models més grans és aconsellable disposar una GPU amb un mínim de 11GB RAM.

Per tal d'avaluar el model Whisper i triar la mida més òptima, s'han seleccionat un conjunt de notícies com referència. El script de descàrrega de TV3 inclou subtítols en català (editat) i de vegades resum.

Pel conjunt de notícies seleccionades, es calcula la taxa d'error de paraula i temps total d'execució de cada model en la transcripció de la base de dades d'avaluació. La duració total dels vídeos és de 570 segons. El temps d'execució s'ha obtingut utilitzant una GPU NVIDIA GeForce RTX 2080 Ti. L'opció 'WordTimestamps' proporciona el temps inicial i final de cada paraula.

#### Vídeos d'avaluació dels resultats de transcripció de la parla

Tòpic	Duració	URL
Violència Masclista	115s	<a href="https://www.ccma.cat/3cat/unitat-daccio-entre-ministeris-contra-la-violencia-masclista/video/6255130/">https://www.ccma.cat/3cat/unitat-daccio-entre-ministeris-contra-la-violencia-masclista/video/6255130/</a>
Política	134s	<a href="https://www.ccma.cat/3cat/sumar-i-podem-cronica-dun-divorci-anunciat/video/6255298/">https://www.ccma.cat/3cat/sumar-i-podem-cronica-dun-divorci-anunciat/video/6255298/</a>
Biometria	110s	<a href="https://www.ccma.cat/3cat/els-estats-de-la-ue-podran-recorrer-al-reconeixement-biometric-en-casos-concrets-sota-ordre-judicial/video/6255299/">https://www.ccma.cat/3cat/els-estats-de-la-ue-podran-recorrer-al-reconeixement-biometric-en-casos-concrets-sota-ordre-judicial/video/6255299/</a>
Sequera	95s	<a href="https://www.ccma.cat/3cat/la-generalitat-es-prepara-per-portar-aigua-en-vaixells/video/6253672/">https://www.ccma.cat/3cat/la-generalitat-es-prepara-per-portar-aigua-en-vaixells/video/6253672/</a>
Subsidi d'atur	114s	<a href="https://www.ccma.cat/3cat/batalla-politica-per-reformar-el-subsidi-datur-treball-proposa-ampliar-ledat-i-economia-endurir-condicions/video/6253664/">https://www.ccma.cat/3cat/batalla-politica-per-reformar-el-subsidi-datur-treball-proposa-ampliar-ledat-i-economia-endurir-condicions/video/6253664/</a>
<b>TOTAL</b>	<b>568s</b>	

#### Informació disponible de cada vídeo

KEY	VALUE
title	Unitat d'acció entre ministeris contra la violència masclista
description	El ministre de l'Interior, Fernando Grande-Marlaska, i la ministra d'Igualtat, Ana Redondo, han volgut mostrar unitat d'acció contra la violència masclista en un any que fins a l'1 de desembre ha registrat 16 víctimes més que en tot el 2022. Els recursos s'han incrementat en els darrers anys i tot i que la tendència històrica és a la baixa en el nombre de víctimes, hi ha repunts preocupants. Moltes vegades es posa tot el pes en la víctima a l'hora de denunciar, però avui la ministra d'Igualtat ha fet una crida a l'entorn de la dona maltractada. N'hi ha qui pensa que es pot parlar de violència institucional en casos en què hi ha prou indicis de violència masclista, la víctima retira la denúncia o nega el maltracte, i l'administració no investiga ni persegueix el delictes d'ofici."
imgsrc	<a href="https://img.ccma.cat/multimedia/jpg/3/1/1701983196013.jpg">https://img.ccma.cat/multimedia/jpg/3/1/1701983196013.jpg</a>
videos	[{"format": "MP4", "quality": "480p", "url": "https://mp4-down-medium-int.ccma.cat/3/0/1701983168403.mp4"}, {"format": "MP4", "quality": "720p", "url": "https://mp4-down-high-int.ccma.cat/8/8/1701983237488.mp4"}]
audio-DescriptionVideos	[]
subtitles	[{"text": "Català", "iso": "ca", "url": "https://statics.ccma.cat/multimedia/vtt/2/8/1701986479882.vtt", "format": "vtt"}]

Rendiment de models Whisper. Medium és un bon compromís entre precisió i velocitat

Model	Word Timestamps			
	WER	segons	WER	segons
tiny	55.54%	40	65.73%	40
base	39.21%	39	46.51%	42
small	25.18%	46	25.56%	38
medium	17.49%	92	18.64%	72
large-v1	16.72%	125	16.21%	120
large-v2	18.19%	128	23.64%	155
large-v3	15.63%	123	15.50%	119

Per calcular la taxa d'error s'han utilitzat com a referència els subtítols proporcionats per TV3. Però aquests subtítols no són sempre una transcripció literal de la notícia, i en força ocasions poden incloure a l'inici el text corresponent a una notícia anterior, o deixar d'incloure la transcripció de les darreres locucions (per exemple el nom del locutor). Per tant, la taxa d'error mostrada a la taula és útil únicament a efectes de comparar els diversos models, però el valor real de la taxa d'error és molt més baix que el calculat automàticament per a aquesta comparativa.

El python notebook STT<sup>2</sup> disponible al repositori github del projecte mostra la utilització del sistema Whisper des d'un programa en Python, així com les opcions disponibles de major interès com la selecció del model, els paràmetres del descodificador i el generador de subtítols.

<sup>2</sup> <https://github.com/multimedia-eurecat/eines-intelligents/blob/UPC/PT1/STT.ipynb>

### 2.1.3. Comparativa Whisper versus WhisperX

En aquesta apartat es compara la velocitat i la qualitat de la transcripció dels sistemes Whisper i WhisperX. Tots dos utilitzen els mateixos models entrenats per OpenAI, però WhisperX inclou codi optimitzat per fer la transcripció en paral·lel i aprofita millor la GPU.

La taula següent mostra que el sistema WhisperX proporciona una qualitat similar però amb una velocitat molt més gran, és especial per a enregistraments de llarga durada. WhisperX és capaç d'obtenir la transcripció i l'alineat de les paraules del fitxer 'original\_3' d'uns 57 minuts de durada en menys de 2 minuts, davant dels 14 minuts que necessita Whisper.

Comparativa de rendiment de models Whisper y WhisperX amb timestamps de paraules.

Model	Word Timestamps			
	WER	segons	WER	segons
tiny	55.54%	40	53.24%	31
base	39.21%	39	49.14%	32
small	25.18%	46	25.75%	33
medium	17.49%	92	21.14%	39
large-v1	16.72%	125	24.02%	47
large-v2	18.19%	128	25.50%	47
large-v3	15.63%	123	14.35%	49
large-v3 (original_3) (3440 segons)	-	840	-	120

Tot i que la diarització o assignació de locutors queda fora dels objectius del projecte, hem realitzat algunes primeres proves del funcionament de WhisperX en aquesta tasca, atès l'interès mostrat per alguns usuaris durant l'avaluació dels prototips.

A la taula següent es mostra la diarització d'una notícia on apareixen 6 locutors diferents, 4 parlant en castellà, i dos en català. Els resultats mostren la tònica general dels resultats obtinguts en aquesta primera avaluació informal d'aquest sistema. El sistema integrat a WhisperX identifica 4 veus diferents ajuntant dues veus en castellà, i també ajunta les dues veus en català (tot i que en aquest cas els dos locutors en català són un home i una dona).

Locutor	Inici	Frase
SPEAKER_03	0.01	"El Ministeri d'Igualtat també està ara a mans dels socialistes i aquest matí la seva titular i el ministre d'Interior han volgut mostrar unitat d'acció."
SPEAKER_03	7.83	"L'objectiu, analitzar el repunt de casos de violència masclista d'aquest 2023."
SPEAKER_00	12.79	"No podemos evitar un sentimiento de profunda frustración cada vez que se produce un asesinato."
SPEAKER_00	19.11	"Es cierto que el modelo no es infalible."
SPEAKER_03	22.12	"I el model encara pot fer menys si no hi ha una denúncia."
SPEAKER_03	24.83	"De les 42 dones assassinades que no havien denunciat, en 22 casos l'entorn coneixia els maltractaments."
SPEAKER_02	31.23	"Además de vivir una situación terrible, de terror diario, tengan que denunciar."
SPEAKER_02	37.23	"Muchas veces no pueden, pero los entornos lo conocen, los entornos pueden ayudar, los entornos son esenciales."
SPEAKER_03	43.04	"Per això, Igualtat posarà en marxa una campanya de sensibilització"
SPEAKER_03	47.5	"dels entorns de les víctimes en la línia del que ha engegat el govern català."
SPEAKER_03	51.54	"I malgrat els recursos i la revisió constant del sistema, a 1 de desembre hi ha hagut 16 víctimes més que en tot l'any anterior."
SPEAKER_03	59.17	"Com a exemple, un dels últims casos d'assassinat masclista a Madrid."
SPEAKER_03	62.95	"Durant el judici, la víctima va dir que no recordava els maltractaments i el jutge va absoldre el seu maltractador."
SPEAKER_03	68.81	"N'hi ha que parlen de negligència per part de l'administració en casos com aquest."
SPEAKER_01	72.83	"Sería un caso de violencia institucional."
SPEAKER_01	75.9	"por falta de protección a esa víctima, tanto por parte del juez como también por parte del Ministerio Fiscal."
SPEAKER_01	83.44	"Evidentemente, si el Estado no actúa con todos esos mecanismos para proteger a esas mujeres, el Estado es responsable."
SPEAKER_01	89.83	"Es un delito público y perseguible de oficio."
SPEAKER_01	92.87	"Lo que ella diga a la víctima no puede recaer sobre la víctima todo el peso de la resolución del asunto."
SPEAKER_03	99.04	"Igualtat participarà a partir d'ara en la taula de seguiment mensual que Interior fa amb les policies de l'Estat, autonòmiques i municipals."
SPEAKER_03	107.62	"Redondo vol engrèixar la maquinària per tal que la informació flueixi entre els ministeris."
SPEAKER_03	112.6	"Manuel Onso Calvo, TV3, Madrid."

## 2.2. Extracció Automàtica de Frases Més Significatives

En aquesta tasca el models LLMs donaran automàticament aquelles frases més significatives des de el punt de vista informatiu del text de la notícia transcrita. Amb aquestes frases i els codis de temps registrats per Whisper es pot compondre un vídeo resum.

La majoria dels LLMs més grans, tipus ChatGPT, només són accessibles mitjançant API, limitant l'accés al codi font i als paràmetres del model. Actualment, altres models, com LLaMA<sup>1</sup> o Mixtral<sup>3</sup>, destaquen per ser d'alt rendiment i d'accés obert i molt més eficients en recursos computacionals.

En les següents seccions descriurem una sèrie d'estratègies i experiments preliminars per la prospecció dels mètodes més efectius. En aquesta selecció s'ha tingut en compte també la disponibilitat del català, de l'accés al codi i la llicència d'ús de cara al seu desplegament final.

### 2.2.1. Models de Llenguatge

MODEL	Variant	Arena Elo Ratings	MEMORIA (Gb)	Temps Per Token	Altres Limitacions	Context (tokens)	Llicència	URL
GPT	GPT4, GPT3-5, Turbo	1100-1200	-	-	-	-	Propietari	
Claude	1,2, Instant	1100-1200	-	-	-	-	Propietari	
Bard			-	-	-	-	Propietari	
Wizard	70b v1.0	1093	-	-		-	Llama 2 Community	
Vicuna	33B	1090	-	-		-	Non-commercial	
Vicuna	13B	1169	9.9 VRAM	13.39 tokens/s	-	-	Non-commercial	<a href="https://lmsys.org/blog/2023-03-30-vicuna/">https://lmsys.org/blog/2023-03-30-vicuna/</a>
Alpaca	13B	1008	9.4 VRAM	15 tokens/s	Tendeix a resumir en angles	-	Non-commercial	<a href="https://crfm.stanford.edu/2023/03/13/alpaca.html">https://crfm.stanford.edu/2023/03/13/alpaca.html</a>
Koala	13B		9.2 VRAM	16 tokens/s	Tendeix a resumir en angles	-	Llama 2 Community	<a href="https://bair.berkeley.edu/blog/2023/04/03/koala/">https://bair.berkeley.edu/blog/2023/04/03/koala/</a>
Stable Beluga	7B (16b)		8.5 VRAM	4.3 Tokens/s	Tendeix a resumir en angles	-	Non Commercial Llama 2 Community	<a href="https://huggingface.co/stabilityai/StableBeluga-7B">https://huggingface.co/stabilityai/StableBeluga-7B</a>
OpenChat	3.5	1070				-	Apache 2.0	
Llama 2	70B chat	1065				-	Llama 2 Community	
Wizard	13B v1.2	1047	9.9 VRAM	17 Tokens/s	Lleugera tendència a resumir en castellà	-	Llama 2 Community	
Zephyr	7B	1042	-	-	-	-	Non-Commercial & Licenced	<a href="https://stability.ai/news/stablelm-zephyr-3b-stability-llm">https://stability.ai/news/stablelm-zephyr-3b-stability-llm</a>
MPT	30B chat	1031	-	-	-	-	CC-BY-NC-SA-4.0	
Mixtral-8x7b	46.7B	1109	-	-	-	32k	Open weights. Licensed under Apache 2.0	<a href="https://mistral.ai/news/mixtral-of-experts/">https://mistral.ai/news/mixtral-of-experts/</a>

<sup>3</sup> <https://mistral.ai/news/mixtral-of-experts/>

## 2.2.2. Conjunt de Vídeos i Notícies

Títol	Format	Paraules Duració	Tòpic	URL	Idioma	Resultat
Aràbia Saudita	Vídeo	1:31	Política	<a href="https://www.ccma.cat/3cat/per-que-larabia-saudita-es-tan-important-i-per-que-te-tanta-por/video/5481833/">https://www.ccma.cat/3cat/per-que-larabia-saudita-es-tan-important-i-per-que-te-tanta-por/video/5481833/</a>	Català	<a href="https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/tv3-arabia_saudita_summarized.mp4">https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/tv3-arabia_saudita_summarized.mp4</a>
Obre col.legis	Vídeo	0:48	Política	<a href="https://www.ccma.cat/3cat/obren-els-col-legis-amb-aplaudiments/video/5335511/">https://www.ccma.cat/3cat/obren-els-col-legis-amb-aplaudiments/video/5335511/</a>	Català	<a href="https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/tv3-collegis_aplaudiments_summarized.mp4">https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/tv3-collegis_aplaudiments_summarized.mp4</a>
Tovallotes	Vídeo	1:26	Medi Ambient	<a href="https://www.ccma.cat/3cat/tones-de-tovallotes-que-recullen-al-mar/video/5594828/">https://www.ccma.cat/3cat/tones-de-tovallotes-que-recullen-al-mar/video/5594828/</a>	Català	<a href="https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/tv3-tovallotes_mar_summarized.mp4">https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/tv3-tovallotes_mar_summarized.mp4</a>
Aigua Mar	Vídeo	1:14	Medi Ambient	<a href="https://www.ccma.cat/3cat/aigua-de-mar-embotellada/video/4120813/">https://www.ccma.cat/3cat/aigua-de-mar-embotellada/video/4120813/</a>	Català	<a href="https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/tv3-Aigua_embotellada_summarized.mp4">https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/tv3-Aigua_embotellada_summarized.mp4</a>
Testimoni preeclàmpsia	Vídeo	10:00	Salud	<a href="https://www.ccma.cat/3cat/va-perdre-un-nado-als-6-mesos-dembaras-per-preclamsia/video/6256778/">https://www.ccma.cat/3cat/va-perdre-un-nado-als-6-mesos-dembaras-per-preclamsia/video/6256778/</a>	Català	<a href="https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/original_1_720p_summarized.mp4">https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/original_1_720p_summarized.mp4</a>
Pannetone	Vídeo	20:28	Social	<a href="https://www.ccma.cat/3cat/els-millors-panettones-son-a-sant-esteve-de-palautordera/video/6257824/">https://www.ccma.cat/3cat/els-millors-panettones-son-a-sant-esteve-de-palautordera/video/6257824/</a>	Català	<a href="https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/original_2_720p_summarized.mp4">https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/original_2_720p_summarized.mp4</a>
Fiv postmortem	Vídeo	57:19	Salud	<a href="https://www.ccma.cat/3cat/ser-pare-despres-de-morir/video/6254977/">https://www.ccma.cat/3cat/ser-pare-despres-de-morir/video/6254977/</a>	Català	<a href="https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/original_3_720p_summarized.mp4">https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/original_3_720p_summarized.mp4</a>
Saurí alforja	Vídeo	8:38 i 5:12	Agricultura	<a href="https://www.ccma.cat/3cat/alforja-aconsegueix-aigua-gracies-a-un-sauri-misterios/video/6254763/">https://www.ccma.cat/3cat/alforja-aconsegueix-aigua-gracies-a-un-sauri-misterios/video/6254763/</a>	Català	<a href="https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/original_4_1_720p_summarized.mp4">https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/original_4_1_720p_summarized.mp4</a> <a href="https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/original_4_2_720p_summarized.mp4">https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/original_4_2_720p_summarized.mp4</a>
Teràpia per a adolescents salut mental	Vídeo	59:29	Salud	<a href="https://www.ccma.cat/3cat/falten-centres-de-llarga-estada-per-a-la-salut-mental-infantojuvenil/video/6255151/">https://www.ccma.cat/3cat/falten-centres-de-llarga-estada-per-a-la-salut-mental-infantojuvenil/video/6255151/</a>	Català (Inici Castellà)	<a href="https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/original_5_720p_summarized.mp4">https://d8k4k5rse7nb7.cloudfront.net/projectes/2024-CIDAI-PAI/original_5_720p_summarized.mp4</a>



### 2.2.3. Inferència de Frases Més Significatives Mitjançant Ponderació

Una opció habitual per obtenir les frases més significatives es basa en donar una ponderació a cada una de les frases, per a continuació ordenar-les en funció d'aquest valor. El resultat final és una llista ordenada de la qual podem extreure el nombre de frases que desitgem fins a complir certa condició: nombre de frases del resum, longitud total en paraules o caràcters, tant per cent de frases, etc.

Per obtenir aquesta ponderació de les frases es pot seleccionar un sistema basat en paraules clau de forma que es podran utilitzar tant les paraules claus extretes de forma automàtica mitjançant models de llenguatge, com KeyBERT, juntament amb paraules clau proporcionades per l'usuari.

El codi proporcionat al quadern ExtractiveSummary<sup>4</sup> es basa en el paquet de python summarizer, però reemplaça, opcionalment, la part d'extracció de paraules clau pel model de llenguatge KeyBERT i permet, a més, la incorporació de paraules clau de l'usuari.

KeyBERT és una tècnica d'extracció de paraules clau basada en el conegut model de llenguatge BERT i que aprofita els embeddings d'aquest model per crear les paraules clau més representatives d'un document. KeyBERT necessita pocs recursos computacionals i és fàcil d'utilitzar. KeyBERT té també una llicència MIT, una llicència permissiva que en permet l'ús comercial sense cap cost ni restricció.

Transcripció proporcionada per Whisper:

*El Ministeri d'Igualtat també està ara en mans dels socialistes i aquest matí la seva titular i el ministre d'Interior han volgut mostrar unitat d'acció. L'objectiu, analitzar el repunt de casos de violència masclista d'aquest 2023. No podemos evitar un sentimiento de profunda frustración cada vez que se produce un asesinato. Es cierto que el modelo no es infalible. I el model encara pot fer menys si no hi ha una denúncia de les 42 dones assassinades que no havien denunciat en 22 casos l'entorn coneixia els maltractaments. Además de vivir una situación terrible, de terror diario, tengan que denunciar. Muchas veces no pueden, pero los entornos lo conocen, los entornos pueden ayudar, los entornos son esenciales.*

<sup>4</sup> <https://github.com/multimedia-eurecat/eines-intelligents/blob/UPC/PT1/ExtractiveSummary.ipynb>

*Per això, Igualtat posarà en marxa una campanya de sensibilització dels entorns de les víctimes en la línia del que ha engegat el govern català. I malgrat els recursos i la revisió constant del sistema, a 1 de desembre hi ha hagut 16 víctimes, més que en tot l'estat. Com a exemple, un dels últims casos d'assassinat masclista a Madrid. Durant el judici, la víctima va dir que no recordava els maltractaments i el jutge va absoldre el seu maltractador. N'hi ha que parlen de negligència per part de l'administració en casos com aquest. Sería un caso de violencia institucional por falta de protección a esa víctima, tanto por parte del juez como también por parte del Ministerio Fiscal. Evidentemente, si el Estado no actúa con todos esos mecanismos para proteger a esa víctima, las mujeres, el Estado es responsable. Es un delito público y perseguible de oficio. Lo que ella diga a la víctima no puede recaer sobre la víctima todo el peso de la resolución del asunto. Igualtat participará a partir d'ara en la taula de seguiment mensual que Interior fa amb les policies de l'Estat, autonòmiques i municipals. Redondo vol engreixar la maquinària per tal que la informació flueixi entre els ministeris. Manuel Alonso Calvo, TV3, Madrid.*

KeyBERT Keywords (10 paraules): violencia, víctimes, socialistes, assassinades, ministerio, malgrat, policies, unitat, dones, model.

Resum (5 frases):

1. *No podemos evitar un sentimiento de profunda frustración cada vez que se produce un asesinato.*
2. *I el model encara pot fer menys si no hi ha una denúncia de les 42 dones assassinades que no havien denunciat en 22 casos l'entorn coneixia els maltractaments.*
3. *Durant el judici, la víctima va dir que no recordava els maltractaments i el jutge va absoldre el seu maltractador.*
4. *Evidentemente, si el Estado no actúa con todos esos mecanismos para proteger a esa víctima, las mujeres, el Estado es responsable.*
5. *Lo que ella diga a la víctima no puede recaer sobre la víctima todo el peso de la resolución del asunto.*

Un avantatge interessant d'aquest sistema és que fa una segmentació en frases sense modificar el text original, i proporciona un rànquing de totes les frases del document juntament amb la durada. Aquest rànquing de totes les frases permet fer una selecció pràcticament immediata i senzilla de les frases més importants fins a completar una durada determinada.

En aquesta taula es mostra la selecció realitzada de l'exemple 'original\_3' fins a obtenir 7 minuts. El temps de càlcul necessari per obtenir aquesta selecció de frases és inferior a 1 segon si s'utilitza l'extractor de paraules clau del programa 'summarizer' i d'uns 2 segons si les paraules clau s'extreuen amb KeyBERT.

**SUMMARY**

minutes: 7.07

sentences: 57

start,	end,	sec,	sentence
33.36,	44.88,	11.52,	A veure, nosaltres, una parella normal i corrent, un dia li va començar a fer mal l'esquena... Ja està.
47.72,	53.84,	6.12,	És que no sé... Sí, sí, sí, jo et demà preguntava, una mica, perquè estem aquí, què va passar.
99.31,	104.47,	5.16,	Que ell era una persona com molt forta, tenia 33 anys, i no s'havia posat malalt a la seva vida.
105.09,	114.82,	9.72,	Llavors el cos era com molt fort i quan ja va donar signes de la malaltia ja ho tenia molt extens.
188.04,	194.55,	6.50,	Sí, sí, vam prendre la decisió, vam anar a Puigverd, vam fer la reserva d'esperma i fins aquí bé.
301.26,	316.75,	15.49,	A més, a tot això jo sempre havia tingut una mica... Al cap el que vam fer, la reserva d'esperma, que volíem ser pares
357.49,	367.19,	9.70,	A l'any Les mostres les eliminen, no hi ha manera que es quedin més temps, no hi ha manera.
407.42,	495.86,	8.45,	Dic, bueno, ja que estem aquí, doncs endavant, tirem cap endavant, i ja que ha arribat fins a aquest punt, doncs ja...
495.88,	501.97,	6.09,	Vam agafar el que millor pinta tenia, el vam posar i a les poquetes hores l'altre va caure.
539.37,	546.19,	6.83,	El PQ venia bé, totes les proves anaven bé, jo tot el procés el vaig fer sola, no li vaig dir a ningú res.
546.83,	552.42,	5.58,	Només una amiga perquè algú en tenia que acompanyar a l'hospital i era infermera i sabia de què anava tot.
552.44,	561.10,	8.66,	Així que només ho sabia una persona, tota la decisió la vaig prendre sola perquè al final era una cosa que només podia
507.73,	597.05,	10.12,	I ja quan ja tot estava bé ja vaig començar a dir-ho a la família de Christian, a la meva família... la gent molt cont
601.39,	606.72,	5.32,	Tens por, perquè encara que jo fins que no vaig estar tres mesos era com que no ho he fet.
620.31,	626.37,	6.06,	Ja els tres mesos i ja quan vaig començar a explicar-s'ho a tota la gent, jo també m'ho vaig començar a creure.
626.41,	630.70,	4.28,	I vaig dir que sí que ho he fet, que sí que he arribat fins aquí.
630.74,	643.22,	12.48,	Per una part contenta, perquè encara que continues portant tot el dol, que és molt difícil, però per aquesta part deia
669.43,	673.97,	4.54,	Me pogut recuperar alguna cosa de nosaltres i alguna cosa del que havia d'haver sigut de la nostra vida.
685.18,	688.03,	3.64,	A veure, prendre aquesta decisió crec que és el millor que he fet.
692.31,	700.18,	7.87,	I, bueno, recomanaria a tothom que estigués en una situació similar que, si dubtes, fes-lo.
815.07,	834.76,	19.69,	Sí, sí, jo no sé si estic deixant-me alguna cosa o no. No. No. Quan li fan la preservació, tu recordes que us explique
839.41,	849.78,	10.37,	Nosaltres vam arribar, li vam fer la presa de l'esperma i ahir ens van comentar que a postmortem només hi havia un any
891.40,	895.44,	4.04,	Jo en aquell moment, quan m'ho van dir, em van quedar això una mica com, què dius?
901.95,	909.87,	7.93,	Però clar, és que els metges quan veien el quadre que tenia ells ja sabien que era molt difícil.
911.79,	917.04,	5.25,	Però és que des del primer dia que va arribar a l'hospital realment no hi havia res a fer.
935.73,	944.60,	8.87,	I ell va durar dos mesos, que era ja com més de l'esperança de vida que donen amb aquest tipus de malaltia de norma ge
946.01,	953.38,	7.36,	Però vosaltres, és això, us ho diuen, però tampoc en parreu de... No, no parlem... No us ho plantegeu realment.
962.85,	969.20,	6.34,	Sí que és cert que abans que ell es posés malalt, abans sí que havien parlat d'això.
1036.07,	1042.77,	5.90,	Es el que tu deies, no és un dol, és un xoc, al principi quedes en xoc.
1125.77,	1131.35,	5.58,	I jo, de cara al setembre, vaig parlar amb la Puigverd, i el gener tenia hora per fer la intervenció.
1220.46,	1223.81,	3.34,	Llavors, és que no és que, bueno, és que és el maig, ho faig el maig.
1203.24,	1209.32,	6.08,	Jo sí que és cert que sempre he sigut una persona anímicament bastant forta dins del que hi cap.
1446.11,	1455.72,	9.61,	I llavors hi hauria gent que em diria això, que ja m'ho havien dit, altra gent que sí, sí, fes-ho tal...
1462.33,	1469.67,	7.35,	I llavors ho vaig portar jo sola perquè en aquell moment és el que vaig sentir que havia de fer.
1510.01,	1514.77,	3.96,	Vaig anar a casa seva, em vaig seure al mig del saló, tenim que parlar.
1515.85,	1524.83,	8.99,	I ja s'ho vaig explicar tot des del principi, tot com va anar, quan ho vaig decidir, per què ho vaig decidir, com ho v
1621.32,	1624.88,	3.56,	No sé per què, jo pensava, no és que sí no puc estar amb el Cristi no puc estar amb ningú.
1651.38,	1661.48,	10.10,	Ja després, quan ja vaig pujar a l'habitació, ja sí que van entrar els meus sogres i ja tothom molt content i tot això
1896.22,	1907.37,	11.15,	però estic contenta i estic feliç del que he fet, del que he aconseguit, que ell estigui bé i al final és una alegria
2004.71,	2094.03,	10.12,	En el teu cas ha de ser molt el que deus sentir... Quan li descobreixis gestos... És que ja els té.
2094.95,	2097.67,	2.72,	És que ja té gestos del pare, tan petit.
2300.24,	2306.12,	5.88,	De vegades s'han de trencar les coses, però m'agrada tenir-ho tot molt controlat dins del que puc.
2406.55,	2494.20,	7.65,	Havíem marxat del lloguer i estàvem vivint... La meva mare té dues residències, una a Girona i una aquí a Barcelona.
2604.03,	2615.03,	10.19,	O sigui, l'embaràs molt bé, la in vitro, que en principi no tenia pinta cap de que sortís a la primera, i això del que
2631.68,	2636.64,	4.96,	A part és que és fort, perquè dius, és cert que era el millor moment, ella havia canviat de feina.
2778.99,	2782.17,	3.18,	Això del mal d'esquena no pot ser, això no pot ser.
2997.00,	3002.49,	4.60,	Jo vaig ingressar a les 7 i a la una i mitja del migdia ja havia tingut el peque.
3143.20,	3147.03,	4.63,	I és que, a part, jo crec que es tira més estona al pit per estar més estona amb mi.
3172.09,	3191.05,	18.96,	O sigui, l'embaràs molt bé, la in vitro, que en principi no tenia pinta cap de que sortís a la primera, i això del que
3218.04,	3220.02,	2.70,	De cop prens la decisió, que ja és molt difícil prendre la decisió, i si no surt, tens una segona oportunitat.
3236.36,	3244.02,	7.66,	Dic, és que és impossible, què m'està passant si estic menjant i no paro de perdre pes, però d'una manera brutal.
3245.98,	3253.20,	7.22,	No sabia què li passava al cos quan et passen aquest tipus de coses, que és que et quedes en no res.
3300.91,	3318.72,	9.81,	Llavors, bueno, hauria sigut complicat, que era una possibilitat que contemplava des del primer moment que podia passa
3318.78,	3325.22,	6.44,	Però una cosa és contemplar-ho i una altra cosa és que hagués passat, no et sé dir.
3355.98,	3361.91,	5.93,	I ara ja amb el peque i ara la gent m'ho diu, és que estàs molt maca, saps molt millor, o tant.
3405.01,	3411.13,	5.32,	Que faci bombolles amb la boca o que mogui així els braços i és que se't cau la bava.
3414.70,	3421.02,	6.32,	El tens allà i dius, mira, sí és que estàs allà i... O sigui que molt bé.

## 2.2.4. Inferència Automàtica de Frases Més Significatives

Els models de llenguatge actuals tenen una alta eficiència per processar text de múltiples maneres i en particular la generació de resums. En el nostre cas, el LLM ha de generar una llista de les frases més importants i escriure-les literalment. Per això la definició d'una bona petició o prompt és clau.

Per aquesta experimentació preliminar s'han utilitzat els següents models i prompts:

- Mixtral-8x7b<sup>5</sup> "Extreu les 5 frases més importants del document:"
- Wizard 1.1 Q-4 bit: "Extreu les 5 frases més importants del document en català:"
- Koala 13B Q-4 bit "Extreu les 5 frases més importants del document en català:"

Per la visualització qualitativa dels resultats s'ha usat la següent llegenda:

- Les paraules marcades en blau són errors ortogràfics.
- Les frases marcades en verd són diferents al text original, però mantenen el significat.
- Les frases marcades en groc no són al text original ni tampoc el seu significat.

### NOTÍCIA: Nadal a Gràcia<sup>6</sup>

#### Mixtral-8x7b

1. "Els llums de Nadal del passeig de Gràcia de Barcelona fan tan bonic que s'han convertit en un maldecap per a la circulació dels cotxes i autobusos i també dels vianants."
2. "El problema és que molts encara estan entretinguts amb la foto quan el semàfor canvia a vermell i els tocaria passar als vehicles."
3. "Ara hi ha més agents uniformats de la Guàrdia Urbana, per dissuadir les persones que es volen fer la foto, i també han optat per posar cintes entre els arbres per impedir que baixin a la calçada per qualsevol punt i no només als passos de vianants."

<sup>5</sup> En alguns models com Mixtral es poden carregar capes específiques del model en GPU i altres en CPU per tal de fer servir models de més capacitat.

<sup>6</sup> <https://www.ccma.cat/324/barcelona-ha-de-prendre-mesures-per-frenar-lallau-de-fotos-de-nadal-al-passeig-de-gracia/noticia/3265513/>

4. "Segons l'Ajuntament, 'la situació està controlada en aquests moments i amb aquestes mesures', però també demanen 'prudència a la ciutadania' i recorden que cal respectar la senyalització i sobretot els semàfors, per evitar escenes com les que es poden veure a les xarxes."
5. "La gent salta en massa al mig del passeig quan tenen el semàfor verd perquè la imatge quedi ben centrada i amb la millor perspectiva."

#### Wizard 1.1 Q-4 bit

Aquí tens les 5 frases més importants del text en català:

1. "La situació està controlada en aquests moments i amb aquestes mesures."
2. "L'Ajuntament de Barcelona ha decidit prendre mesures per que la gent no envaeixi la calçada."
3. "Recorden que cal respectar la senyalització i sobretot els semàfors, per evitar escenes com les que es poden veure a les xarxes:"
4. "Els llums de Nadal del passeig de Gràcia de Barcelona fan tan bonic que s'han convertit en un maldecap per a la circulació dels cotxes i autobusos i també dels vianants."
5. "Segons l'Ajuntament, 'la situació està controlada en aquests moments i amb aquestes mesures,' però també demanen 'prudència a la ciutadania' i recorden que cal respectar la senyalització i sobre tot els semàfors, per evitar escenes com les que es poden veure a les xarxes:"

#### Koala 13B Q-4 bit

1. El pas del Passeig de Gràcia de Barcelona és un maldecap per a la circulació i el trànsit de vehicles i peatges.
2. Molts encara estan distrets amb la foto quan el semàfor canvia a vermell i els tocaria passar als vehicles.
3. L'Ajuntament de Barcelona ha decidit prendre mesures perquè la gent no envaeixi la calçada.
4. Hi ha més agents uniformats de la Guàrdia Urbana per dissuadir les persones que es volen fer la foto, i també s'han posat cintes entre els arbres per impedir que baixin a la calçada.
5. L'Ajuntament de Barcelona ha optat per posar cintes als laterals del Passeig de Gràcia per evitar que la gent salti al centre de la calçada.

## NOTÍCIA: Pneumàtics<sup>7</sup>

### Mixtral-8x7b

1. *“Mentre les emissions dels motors i el dièsel han baixat dràsticament, la de cobertes i frens són ara la gran assignatura pendent.”*
2. *“I puja sense parar: parlem de les emissions que surten dels pneumàtics.”*
3. *“Són les que fan menys de 0,1 micres, la mil·lèsima part del diàmetre d'un cabell humà. Són les més nocives perquè poden inflamar els pulmons i entrar als alvèols.”*
4. *“I que s'han convertit en la darrera gran batalla per a la millora de l'aire.”*
5. *“El problema, ara que els pitjors temps del dièsel ja han passat a la història, és que en l'aspecte de les NEE els cotxes del futur, els elèctrics, no ajuden.”*

### Wizard 1.1 Q-4 bit

El transport automotriu és un important contribuent de les emissions de gasos d'efecte hivern en ciutats i poblacions. Les partícules fines, en particular, poden causar problemes de salut. Es requereixen mesures addicionals per abordar la qüestió de les emissions de partícules procedents del tràfic rodant.

A Catalunya, s'han pres diverses iniciatives per reduir les emissions de vehicles a causa del seu impacte negatiu en la salut humana. Un exemple recent ha estat la limitació de les emissions de diesel, impulsada pel Govern de la Generalitat de Catalunya i aprovada per les autoritats espanyoles el 2011

### Koala 13B Q-4 bit

El text descriu les emissions de partícules PM2.5 en l'aire causades pel trànsit rodat, sobretot en vehicles dièsel. Els autòmats mesuren la concentració de partícules en suspensió (PM10) i s'han reduït significativament en els últims anys a Europa. Però les nanopartícules, amb una mida inferior a 0.1 micres, poden penetrar al pulmó i arribar als vasos sanguinis. Les emissions de partícules del trànsit rodat es calcula que representen el 25% de les emissions globals de nanopartícules. L'autopista de Madrid és un exemple d'una zona on les emissions de nanopartícules són elevades. La indústria automobilitària ha reduït les emissions de partícules en els seus vehicles dièsel per incorporar filtres a l'exhaust o modificar la combustió. Això ha conduït a una disminució de les concentracions de PM10 en l'aire, però no s'han detectat canvis significatius en les concentracions de nanopartícules.

<sup>7</sup> <https://www.ccma.cat/324/pneumatics-la-contaminacio-invisible-que-cada-cop-va-a-mes/noticia/3265378/>

## NOTÍCIA: Edició malalties genètiques<sup>8</sup>

### Mixtral-8x7b

1. *“L'agència del medicament dels Estats Units, la FDA, ha aprovat aquest divendres dos tractaments per a l'anèmia de cèl·lules falciformes, una malaltia hereditària de la sang molt invalidant.”*
2. *“Un dels tractaments, Casgevy, és el primer que s'aprova en aquest país que comporta l'edició de l'ADN del malalt amb la tecnologia CRISPR, que permet modificar el genoma.”*
3. *“La malaltia consisteix en una anomalia genètica que fa que els glòbuls vermells tinguin una forma anòmala, com una mitja lluna, que dificulta la circulació de la sang al cos.”*
4. *“La FDA assegura que els assajos clínics que han fet les dues farmacèutiques, amb 44 pacients Casgevy i 32 Lyfgenia, han mostrat que són tractaments prou segurs i eficaços per donar-hi el vistiplau.”*
5. *“Segons el New York Times, només l'edició del genoma d'un pacient per Casgevy costarà més de 2 milions de dòlars, més de 3 per Lyfgenia, i prop de 2 milions més al llarg de la vida d'un malalt.”*

### Wizard 1.1 Q-4 bit

1. Els Estats Units aproben el primer tractament d'una malaltia amb edició genètica.
2. *La malaltia de cèl·lules falciformes és una malaltia hereditària de la sang molt invalidant.*
3. *El tractament aprovat per Casgevy implica modificar el genoma del pacient usant la tecnologia CRISPR.*
4. *Les limitacions tecnològiques i el cost elevat fan que sigui difícil que molts malalts puguin accedir al tractament.*
5. *L'FDA assegura que els assajos clínics han demostrat que els tractaments són prou segurs i eficients per donar-los el vistiplau.*

### Koala 13B Q-4 bit

1. Els Estats Units aproben el primer tractament d'una malaltia amb edició genètica
2. El tractament consisteix en autotrasplantaments de cèl·lules mare de la sang del malalt, modificades genèticament amb CRISPR
3. *La FDA afirma que els assajos clínics han demostrat que els tractaments són prou segurs i eficaços per donar-hi l'aprovació*

<sup>8</sup> <https://www.ccma.cat/324/els-estats-units-aproben-el-primer-tractament-duna-malaltia-amb-edicio-genetica/noticia/3265384/>

4. El preu del tractament és prohibitiu: aproximadament 2 milions de dòlars per Casgevy i 3 milions per Lyfgenia
5. Només uns pocs hospitals podran atendre un nombre limitat de pacients cada any

Anàlisi de resum extractiu de 3 models de llenguatge en una experimentació preliminar

Percentatge de frases que són exactament en el document original				
MODEL	Nadal a Gràcia	Pneumàtics	Edició malalties genètiques	Total
Mixtral-8x7b	100	100	100	100
Wizard 1.1 Q-4 bit	60	0	20	26,7
Koala 13B Q-4 bit	20	0	20	13,4

### 2.2.5. Eliminació de Parts Menys Significatives d'una Frase

Amb els models LLMs podem produir un nivell addicional de síntesi mitjançant l'eliminació de les parts menys significatives d'una frase. Hem vist que la formulació en negatiu funciona millor que en positiu. També explicitar la instrucció "la frase ha de tenir sentit" va ser fonamental. Per aquest estudi hem utilitzat únicament el model Mixtral-8x7b.

**Prompt:** <Escriu la frase eliminant la part que sigui menys important. La frase ha de tenir sentit. És extremadament important que les paraules siguin en el text original. Només escriu la frase resultant.>

**Frase:** "Els llums de Nadal del passeig de Gràcia de Barcelona fan tan bonic que s'han convertit en un maldecap per a la circulació dels cotxes i autobusos i també dels vianants."

**Resultat:** "Els llums de Nadal del passeig de Gràcia de Barcelona fan tan bonic que s'han convertit en un maldecap per a la circulació."

**Frase:** "El problema és que molts encara estan entretinguts amb la foto quan el semàfor canvia a vermell i els tocaria passar als vehicles".

**Resultat:** "El problema és que molts encara estan entretinguts amb la foto quan tocaria passar als vehicles."

**Frase:** "Ara hi ha més agents uniformats de la Guàrdia Urbana, per dissuadir les persones que es volen fer la foto, i també han optat per posar cintes entre els arbres per impedir que baixin a la calçada per qualsevol punt i no només als passos de vianants."

**Resultat:** "Ara hi ha més agents uniformats de la Guàrdia Urbana, han optat per posar cintes entre els arbres per impedir que baixin a la calçada."

**Frase:** "S'han posat cintes als laterals perquè la gent no salti al centre de la calçada, a més de destinar-hi més agents de la Guàrdia Urbana."

**Resultat:** "S'han posat cintes als laterals perquè la gent no salti al centre de la calçada."

### 2.2.6. Conclusions Inicials del Resum Extractiu

1. El model Mixtral ofereix un rendiment més alt comparat amb Wizard o Koala.
2. Algunes de les frases extretes no són literals o mostren faltes ortogràfiques. Es necessari una cerca de les frases amb robustesa a faltes.
3. Cal explícitament demanar al model de llenguatge ordenar les frases extretes del text original.
4. Cal explicitar al models que no doni explicacions.
5. El context en tokens determina la duració màxima de la notícia que es pot processar.
6. En l'extracció de paraules menys significatives hem trobat les següents limitacions del model Mixtral:
  - Elimina les paraules menys significatives, però genera de noves que no hi eren originalment a la frase, modifica les originals.
  - Si es demana explícitament en català, genera noves paraules en angles.

## 2.3. Definició de Funcionalitats

En base a les capacitats dels mòduls descrits anteriorment pel resum automàtic de vídeo de notícies, sobre tot en relació al reconeixement de la parla i el resum extractiu, s'han definit els següents paràmetres de configuració per modular i adaptar el contingut del resum en funció de les necessitats de l'usuari.

### 1. NOMBRE DE FRASES

- Valor: nombre natural (1,2,3,...)
- Descripció: control del nombre de frases que compte el resum final.
- Prompt: ex. <Llista en ordre quines son les <X> frases més importants del següent text "...">
- Exemple: *Així, per exemple, es pot configurar de manera que no es pugui accedir a webs de contingut pornogràfic o violent. Com? "Entrant a la configuració del router i canviant el número dels servidors DNS per uns DNS amigables", per exemple, aquest o aquest, aconsella Pérez. Els DNS controlen a quin servidor es dirigeix l'usuari quan escriu una adreça d'Internet. Per fer el canvi, Pérez aconsella buscar un professional que sàpiga com es fa, com un informàtic de confiança. Des del rúter també es poden identificar els dispositius de casa per fer que siguin els únics que s'hi poden connectar, així com controlar el temps i els horaris de connexió a Internet de cadascun dels dispositius. "Tot això t'ho fa el router i no cal gastar-se els duros en programes de control parental", opina el pèrit judicial.*

### 2. PARTS MÉS SIGNIFICATIVES EN FRASES

- Valor: Verdader o False (boolean)
- Descripció: Elimina segments poc significatius. Segon nivell de compressió.
- Prompt: ex. <Escriu textualment la part més important de la següent frase: "...">
- Exemple: *Així, per exemple, es pot configurar de manera que no es pugui accedir a webs de contingut pornogràfic o violent. Com? "Entrant a la configuració del router i canviant el número dels servidors DNS per uns DNS amigables", per exemple, aquest o aquest, aconsella Pérez. Els DNS controlen a quin servidor es dirigeix l'usuari quan escriu una adreça d'Internet. Per fer el canvi, Pérez aconsella buscar un professional que sàpiga com es fa, com un informàtic de*

*confiança. Des del rúter també es poden identificar els dispositius de casa per fer que siguin els únics que s'hi poden connectar, així com controlar el temps i els horaris de connexió a Internet de cadascun dels dispositius. "Tot això t'ho fa el router i no cal gastar-se els duros en programes de control parental", opina el pèrit judicial.*

### 3. PERSONALITZADES DE FRASES MITJANÇANT PARAULES CLAU

- Valor: text i RegExp.
- Descripció: S'afegiran paraules o expressions regulars per afegir frases addicionals no detectades per model de llenguatge.

### 4. PREVISUALITZACIÓ DE LA DURADA

- Valor: visualització numèric.
- Descripció: càlcul del temps de duració final del vídeo en funció de les frases del resum extractiu.

### 5. SUPRESIÓ SILENCIS ENTRE PARAULES

- Valor: nombre natural (decimal: 0, 0.1,...)
- Descripció: Els silencis entre paraules que triguin més d'X segons seran retallats.

## 2.4. Detalls de la Implementació Final

A continuació es descriuen els passos més importants implementats en el sistema final de resum automàtic de notícies, inclosos els prompts que han donat millor resultats.

1. **Traducció parla a text:** el model Whisper (medium) s'ha mostrat dels més eficients en traduir la pista d'àudio a text, agrupant la informació textual en segments i donant informació de codis de temps (timestamp) a nivell de paraula. Whisper separa en diferents paraules els pronoms febles tal com "t'ha" o "m'ho". Cal un pre-processament de text agrupar aquests casos en una única paraula.
2. **Extracció de frases importants:** el model de llenguatge Mixtral (medium) ha donat el millor rendiment pel resum extractiu. L'extracció, sobre tot en entrevistes on la separació entre frases és més ambigua, de vegades esta formada per segments en comptes de frases senceres, delimitades per punts o signes d'interrogació o exclamació. Aquest segments de vegades no tenen sentit per sí mateixos. Per tant, la frase final és aquella a la que pertanyen els segments. Sovint aquest segments estan dins de una mateixa frase. A més a més, un filtratge de frases elimina repeticions de vegades generades per Mixtral. Per aquestes raons, el nombre de frases finals extretes pot ser menor. Quan el resum demanat és llarg és difícil encertar amb la durada final del vídeo, més tediós encara perquè el procés de resum extractiu és més lent.
  - Prompt: *<Per favor, revisa aquest text en català i escriu les {num\_phrases} frases més importants. És important que les frases siguin textualment en el text original. També és important que les frases estiguin ordenades en funció de la seva aparició en el text original. No donis explicacions>*.
  - Exemple: *"Ara hi ha més agents uniformats de la Guàrdia Urbana, per dissuadir les persones que es volen fer la foto, i també han optat per posar cintes entre els arbres per impedir que baixin a la calçada per qualsevol punt i no només als passos de vianants."*
3. **Selecció de frases mitjançant paraules clau o expressions regulars:** l'usuari pot afegir frases addicionals no seleccionades inicialment pel model LLM.

4. **Resum a nivell de subordinada:** una segona passada de les frases seleccionades és processada pel model de llenguatge per tal d'eliminar les parts menys significatives de les frases. Això proporciona un nivell addicional de síntesi (compressió).
  - Prompt: *<Per cadascuna d'aquestes frases elimina la part que sigui menys important. La frase resultat ha de tenir sentit. És extremadament important que les paraules siguin en el text original. No donis explicacions>*.
  - Exemple: *"Ara hi ha més agents uniformats de la Guàrdia Urbana, han optat per posar cintes entre els arbres per impedir que baixin a la calçada."*
5. **Temps de durada:** duració aproximada en funció dels codis de temps seleccionats.
6. **Creació de vídeo:** en cada crida al Mixtral és possible que l'ordre temporal no sigui correcte. Abans de crear el vídeo final les frases s'ordenen respecte a la seva posició en el vídeo original en funció dels ID de segment proporcionats pel model Whisper. Un temps extra de 0.5 segons abans i després de cada frase o segment és afegit per seguretat. En el cas de que els segments siguin contigus aquest temps extra és ignorat. S'ha decidit per simplicitat que aquest temps extra no sigui un paràmetre exposat a l'usuari.

### 2.4.1. Temps de Computació

Hem observat que vídeos de 2-3 minuts requereixen 10-20 segons de processament amb una GPU NVIDIA RTX 3090. Vídeos llargs de 60 minuts requereixen al voltant de 20-25 minuts de processament amb la mateixa targeta.

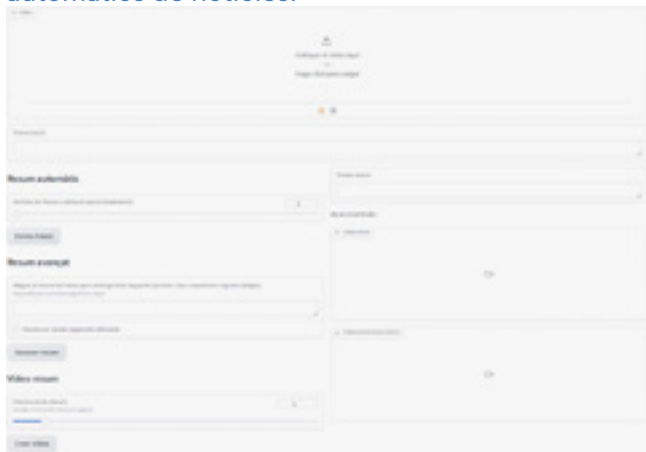
## 2.5. Interfície d'Usuari UX

L'eina desenvolupada proporciona una manera eficient de resumir vídeos de notícies en vídeos. S'ha desenvolupat una interfície d'usuari UX que facilita la interacció amb el contingut audiovisual i els models d'intel·ligència artificial així com la capacitat d'adaptar el contingut generat a les necessitats de l'usuari tal com la llargada del vídeo. Alguna informació, tal com noms, dates, o paraules clau, ajuden a adaptar el contingut final del vídeo i per tant la difusió de la notícia.

A partir d'un arxiu de vídeo, la interfície genera automàticament un altre vídeo resum més curt amb les parts més importants. Per això un model de traducció de parla a text (STT) genera en forma textual allò que s'ha dit al vídeo. Després, un model gran de processament de llenguatge natural (LLM) detecta les parts més importants amb diferents opcions de personalització.

La interfície s'ha desenvolupat en llenguatge Python i Gradio<sup>9</sup> que ofereix una bateria d'eines pel desplegament de models de *deep learning* de ràpid prototipatge.

Interfície d'usuari implementada per fer resums automàtics de notícies.



A continuació, detallem pas a pas com utilitzar aquesta eina per generar resums automàtics de vídeos.

### Traducció a Text:

1. **Traducció parla a text:** clicar per obrir o arrossegar un arxiu de vídeo sobre la interfície. Després d'uns segons es mostra la traducció de tota la parla en text.

<sup>9</sup> <https://www.gradio.app>

### Resum Automàtic:

1. **Extreu frases:** seleccionar el nombre (aproximat) de frases desitjat que tingui el resum i clicar el botó. Contra més nombre de frases, més llarga serà la durada. Degut a certes interpretacions sobre el concepte "frase" per part del model de llenguatge i d'altres aspectes tècnics, el nombre de frases final és aproximat. Tot i això, el nombre de frases seleccionat té una concordança amb la durada final del resum.

### Resum Avançat:

1. **Selecció de frases mitjançant paraules clau o expressions regulars:** l'usuari pot afegir frases addicionals no seleccionades inicialment pel model LLM. Bé amb paraules clau i/o expressions regulars, **sempre separades amb coma + espai** ", " indicades en el camp corresponent.
2. **Resum a nivell de subordinada:** es pot activar clicant en el checkbox una segona passada de les frases seleccionades és processada pel model de llenguatge per tal d'eliminar les parts menys significatives de les frases. Això proporciona un nivell addicional de síntesi (compressió).

### Generació vídeo resum:

1. **Crear vídeo:** clicant aquest botó es generarà el vídeo final. L'usuari té l'opció addicional d'eliminar els silencis més llargs que un cert llindar (segons).

### Mostra d'informació:

- **Resum textual:** en cada pas es mostrarà el resum final en el panell de la dreta.
- **Temps de durada:** duració aproximada en funció dels codis de temps seleccionats.
- **Vídeos resum:** dues finestres mostren els vídeos final amb i sense silencis.

### Notes:

- Per un funcionament òptim, els vídeos han de estar en català íntegrament.
- El temps estimat per generar un resum és proporcional a la durada del vídeo original.
- La interfície no visualitza correctament els vídeos després de la primera generació.
  - **Mateix vídeo:** per una correcta visualització del mateix vídeo amb diferents configuracions personalitzades, es recomana la descàrrega directa amb el botó fletxa sobre la cantonada superior dreta del vídeo.
  - **Diferents vídeos:** en aquest cas es recomana refrescar la pàgina.



## 2.6. Avaluació Final d'Usuaris PT1

L'avaluació, originalment pensada en NOTÍCIES, es va estendre també a ENTREVISTES i va ser duta a terme per part de 3 perfils professionals diferents de professionals treballadors de la Corporació Catalana de Mitjans Audiovisuals (CCMA), que han emplenat els següents qüestionaris.

Val la pena comentar que l'avaluació de l'eina ha estat avaluada com si estigués en una fase de desenvolupament final. D'aquesta manera els avaluadors han trobat a faltar algunes funcionalitats addicionals que, tot i ajudar a la funcionalitat general per al seu ús en certs casos d'ús, no estaven dins de l'objectiu plantejat en aquest paquet de treball. Això ha fet que algunes de les valoracions tinguin una tendència més baixa de l'esperada.

### 2.6.1. Avaluació Resum de Notícies

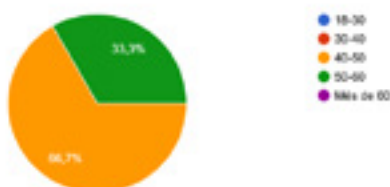
Cap dels 3 perfils enquestats li veuen utilitat a aquesta eina, principalment per 2 motius:

1. Els vídeos originals ja estan molt resumits i no fan resums de vídeos tant curts.
2. Els vídeos de xarxes són noves creacions, no resums de peces ja emeses, es munten d'una altra forma i se'ls afegeixen subtítols.

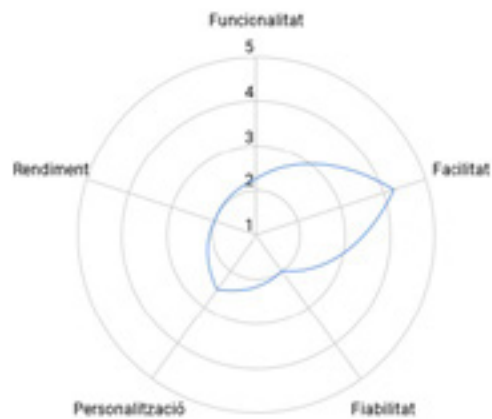
També comenten, el contingut generat sigui publicable, sense edició posterior, s'hauria de tenir en compte la imatge, ja que faltaria aplicar criteri d'edició i llenguatge audiovisual, com plànols que duren molt poc.

Tot i que la eina es va comportar de manera eficient, aquest motius han fet que la puntuació general sigui més baixa.

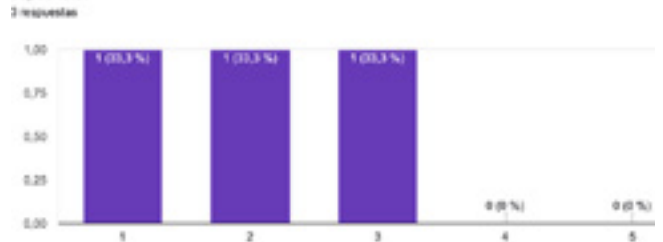
Edat  
3 respostes



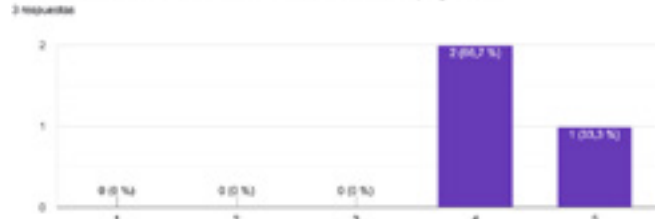
Sexe  
3 respostes



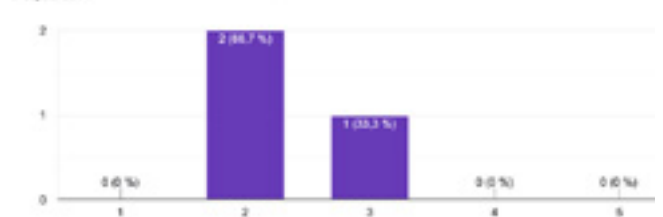
Funcionalitat: El programari satisfà les necessitats de l'usuari en termes de característiques i capacitats?  
3 respostes



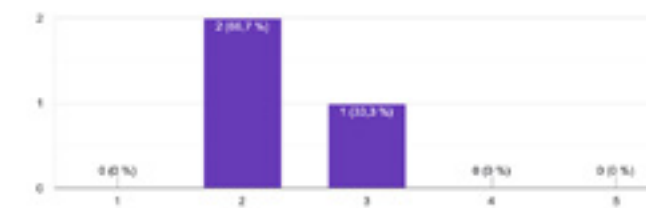
Facilitat d'ús: Com d'intuitiva i fàcil d'utilitzar és la interfície del programari?  
3 respostes



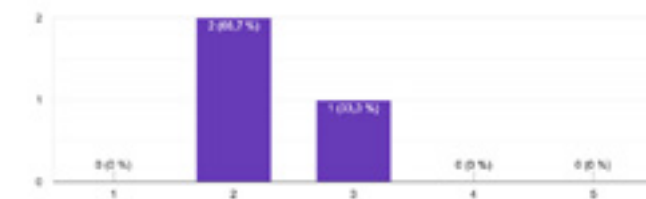
Fiabilitat: L'usuari pot dependre del programari perquè funcioni de manera coherent sense fallar?  
3 respostes



Personalització: Pot l'usuari personalitzar el programari per adaptar-se a les seves preferències?  
3 respostes



Rendiment: El temps de resposta és l'adequat per l'ús que se li vol donar a aquest programari?  
3 respostes



## 2.6.2. Avaluació Resum de Entrevistes

### Redactor digital informatius

Pot ser una eina molt útil per generar un primer resum ampli d'una entrevista o roda de premsa llarga (tipus 1h -> 10 min), no el veu com una eina capaç de generar contingut preparat per publicar directament, ja que la selecció feta podria no coincidir amb els criteris editorials i l'ordre també podria ser diferent, de fet als exemples s'ha vist que l'eina posa el focus al "què" i selecciona més la veu en off (en cas que existeixi) i nosaltres li donem més pes a la part emotiva i als testimonis.

La part de transcripció sembla molt útil però ajudaria que es distingís entre les diferents veus que parlen per poder afegir aquest criteri a l'hora de seleccionar els elements del resum.

També ajudaria poder determinar amb més precisió la mida del contingut resultant, potser en comptes de frases fer-ho amb una forquilla de paraules. En això ajudaria que el software anés més ràpid per poder iterar i tenir diferents alternatives.

### Responsable de documentació

Pot ser una eina molt útil per generar resums d'entrevistes antigues que d'altra forma no es publicarien perquè caldria moltes hores d'esforç visualitzant i triant el més destacat. La transcripció l'estan fent actualment amb una eina automàtica i es podria comparar si aquesta eina millora el que tenen actualment, també es troba a faltar que estigui separat per les diferents veus.

La part d'eliminar silencis sembla molt interessant. Afegit a això, ajudaria que tingués en compte altres criteris com la qualitat de les imatges (desenfocat, imatges en negre, imatges d'una càmera apuntant a terra o imatges gravades amb molt moviment).

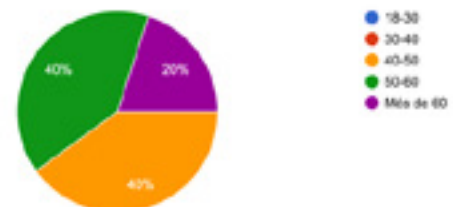
### Realitzador de vídeo

Li sembla una eina molt interessant per fer una primera tria de contingut però creu que encara es necessita una realització a posteriori per generar contingut de qualitat.

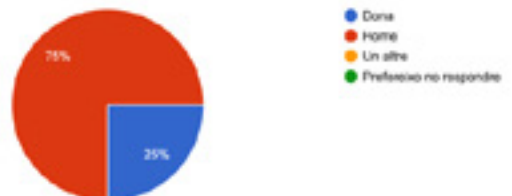
Per facilitar el treball amb el contingut generat per aquesta eina ajudaria:

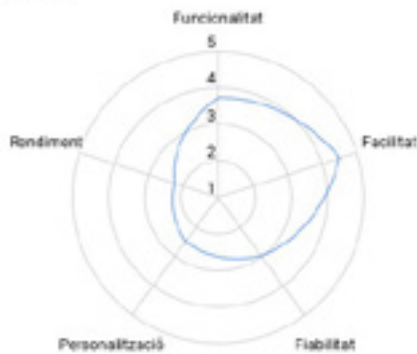
- Poder afegir 2 frames en blanc entre talls, per evidenciar que hi ha tall i facilitar l'edició a posteriori.
- Poder afegir "handlers" per facilitar l'edició posterior. Un handler és un temps per davant i per darrera per poder fer transicions o foses.
- Poder definir si volem fer foses d'àudio (d'entrada i sortida) entre els talls.
- Visualitzar un *timeline* o poder afegir meta informació.

Edad  
5 respostes

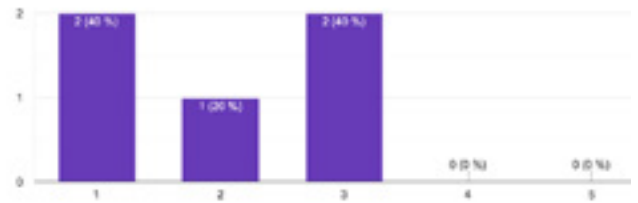


Sexe  
4 respostes

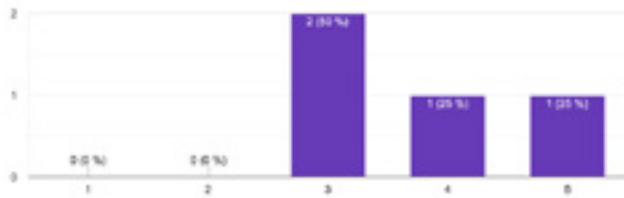




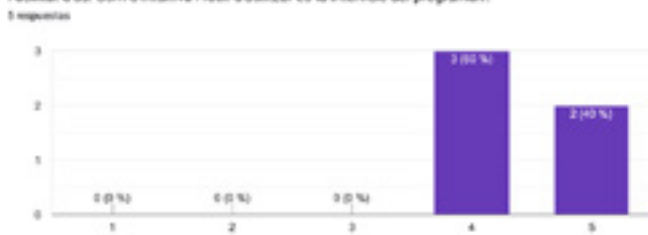
Rendiment: El temps de resposta és l'adequat per l'ús que se li vol donar a aquest programari?  
5 respostes



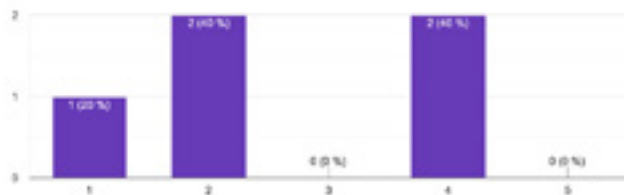
Funcionalitat: El programari satisfà les necessitats de l'usuari en termes de característiques i capacitats?  
4 respostes



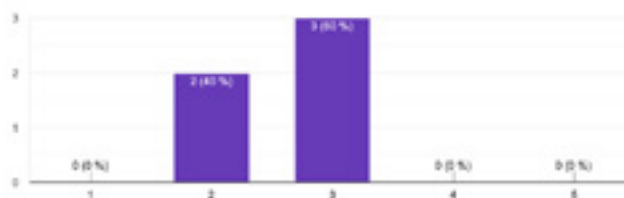
Facilitat d'ús: Com d'instutiva i fàcil d'utilitzar és la interfície del programari?  
5 respostes



Fiabilitat: L'usuari pot dependre del programari perquè funcioni de manera coherent sense fallar?  
5 respostes



Personalització: Pot l'usuari personalitzar el programari per adaptar-se a les seves preferències?  
5 respostes



## 2.7. Conclusions PT1

El resum de vídeos mitjançant la imatge encara té reptes importants per ser abordats pels models de visió d'última generació de manera efectiva. En canvi, en el contingut audiovisual on hi ha contingut narrat o parla en general, l'anàlisi del contingut mitjançant el reconeixement de la veu i models de llenguatge avançats actualment ofereix altes capacitats per fer resums de text efectius.

En aquest projecte hem desenvolupat una eina pensada per fer resums automàtics de notícies basat en (1) la traducció de la parla a text mitjançant models de Speech-To-Text com Whisper, i (2) resum extractiu de les frases més importants de la notícia mitjançant models grans de llenguatge natural com Mixtral. Posteriorment el codis de temps oferts pel model STT s'utilitzen per generar el vídeo resum.

Aquesta eina a més a més ofereix un resum addicional eliminant les parts menys importants de les frases i un resum personalitzat per trobar frases concretes mitjançant paraules clau o expressions regulars definides per l'usuari. El vídeo final es filtra per eliminar silencis que siguin més llargs que un llindar definit per l'usuari.

En l'avaluació de resum de notícies l'eina desenvolupada no té l'impacte esperat ja que les notícies ja són bastant curtes per-se. Pel cas del resum d'entrevistes s'ha trobat que l'usuari ha de supervisar igualment el contingut resumit. Per exemple, ajudaria tenir un control de criteris per seleccionar veus. També ajudaria reduir el temps de processament per provar diferents configuracions i personalitzar millor el resum final.

S'ha trobat interessant aquesta eina com primera tria del contingut, tot i que el contingut audiovisual requereix una realització posterior. També seria interessant per entrevistes antigues que impliquin moltes hores de visualització.

# Narració Gràfica Generativa de Notícies



El públic més jove té una preferència per les narracions d'impacte social, però acostumen a estar menys interessats en els formats en línia amb molt de text. Com a resposta, algunes iniciatives com el projecte Graphical Storytelling (GST) de News Labs (BBC) el 2019, tenen l'objectiu de crear còmics de notícies visualment atractius basats en text proporcionat pels periodistes. Aquí el repte és la automatització de la identificació de les característiques del text per seleccionar tractaments gràfics adequats per a aquestes històries i la personalització del contingut visual final d'una manera coherent al llarg dels punts de narració.

Els models generatius grans (Large Generative Models) són arquitectures d'intel·ligència artificial avançades que tenen un paper crucial en la creació de dades sintètiques realistes i diverses. Aquests models aprofiten quantitats massives de dades d'entrenament i algorismes complexos per generar noves mostres de dades que s'assemblen als patrons i característiques del conjunt de dades original.

Els models de difusió (LDM) en particular són una classe de models generatius utilitzats en l'aprenentatge automàtic per a tasques com ara la generació de dades, la síntesi d'imatges i la eliminació de sorolls. Aquests models s'inspiren en el concepte de difusió en física i tenen com a objectiu generar dades aplicant iterativament una sèrie de nivells de soroll a un estat inicial determinat. Actualment son l'estat del art en generació d'imatges realistes amb una gran variabilitat visual.

Els grans models de llenguatge (LLM) són una classe de models d'intel·ligència artificial dissenyats per entendre i generar llenguatge semblant a l'ésser humà. Aquests models aprofiten quantitats massi-

ves de dades de text per aprendre patrons, context i semàntica del llenguatge.

### Proposta de Valor:

- La narració gràfica generativa de notícies és un procés que permet crear un contingut visual a partir d'un fragment de text. En el cas particular de mitjans de comunicació l'acompanyament d'una notícia amb una o diverses imatges pot fer l'article més atractiu per als lectors. Aquest fet és degut al fet que les imatges tenen un poder evocatori més gran provocant emocions en el lector i arriben a un nivell de detall que no és possible expressar amb paraules.
- Aquest projecte ha desenvolupat tecnologia innovadora basada en grans models de llenguatge (LLMs) per resumir el text d'una notícia en un parell de frases, que posteriorment són utilitzades per un model generatiu (Diffusion Model) per a la creació final del contingut visual.
- L'eina desenvolupada ha sigut integrada en una interfície amigable de fàcil instal·lació i ús que permet la interacció intuïtiva i personalització del contingut segons les necessitats de l'usuari.

En aquest paquet de treball explorem diferents vies per desenvolupar eines de generació automàtica de imatges (LDM) amb la combinació de models grans de processament de llenguatge natural (LLMs), els quals tenen una capacitat molt alta de resumir text, però també inferir el tema o el to general; certes característiques que permeten adaptar el contingut visual a la semàntica particular del text.

Flux de treball de la narració gràfica de notícies: generació de resum abstractiu (LLM), extracció addicional d'informació semàntica, interacció amb l'usuari i personalització d'elements, generació d'imatges guiades per text i imatge d'estil de referència (LDM).



## 3.1. Resum Abstractiu de Notícies

Per fer el resum abstractiu d'una notícia, farem servir una tècnica anomenada "in-context learning" (ICL), que és una tècnica que permet als sistemes d'aprenentatge automàtic aprendre noves tasques sense necessitat de tornar-les a entrenar amb conjunts de dades massives.

En lloc d'això, ICL utilitza exemples de tasques específiques per ensenyar al model com resoldre-les. D'aquesta manera, es presenta al LLM un conjunt d'exemples de com resumir el text d'una notícia. Important, els exemples d'entrada estan tant en anglès com en català, però la sortida sempre serà en anglès.

A continuació es mostra un exemple il·lustratiu del procés d'ICL.

### 3.1.1. Exemple del procés d'ICL

You are an image caption generator. Your task is to create one sentence describing an image related to the text I will provide you. The output caption must be related to the text. The caption should be only one sentence, with a maximum length of 20 words. The text can be in any language, but the caption must be always in English. Refer to the example below for the desired format.

**Text:** Estonia has said it is prepared for possible spillover into its territory amid an ongoing migration crisis on the EU's eastern edge. Finland is set to close its entire 1300km frontier with Russia from Wednesday at midnight, following a recent spike in the number of people trying to enter the country illegally. Moscow has been accused of channeling migrants to the Finnish border as a form of hybrid war. Now neighboring Estonia, which also sits on Russia's border, says it could be affected too, though it is yet to face similar pressure on its frontier, according to Estonian authorities.

**Caption:** An image of four Eastern European men and women walking through the snow in the forest  
<end>

**Text:** After two lackluster years, gold is staging an impressive comeback in 2023, with the precious metal delivering double-digit gains by the end of November, inching closer to its all-time highs of \$2,060 per ounce reached in August 2020. Following a dip to \$1,820 per ounce on October 5th, a confluence of factors has

aligned to create a favorable environment for gold's resurgence.

**Caption:** An image of many gold bars stacked on top of each other, forming a pyramid.

<end>

**Text:** Segons els informes, el gegant de la intel·ligència artificial generativa (IA) OpenAI estava treballant en un descobriment avançat que els investigadors del personal van advertir abans que l'acomiadament del CEO Sam Altman podria amenaçar la humanitat, van dir a Reuters dues persones familiaritzades amb l'assumpte. El consell d'OpenAI va acomiadar Altman divendres passat, però després el va reincorporar dimarts després que gairebé tots els 750 empleats de l'empresa amenacessin amb dimitir si no era reintegrat.

**Caption:** An image of a futuristic robot.

<end>

**Text:** Els preus van pujar un 3,2% al novembre, segons la dada avançada de l'IPC que ha fet pública aquest dimecres l'Institut Nacional d'Estadística (INE). Són tres dècimes menys que el mes anterior. Amb tot, la inflació interanual al conjunt d'Espanya continua per sobre del 3%. La dada de l'IPC del novembre és la que completa la sèrie de dotze mesos per calcular l'increment de les pensions l'any que ve. Així doncs, si es confirma la xifra, a mitjan desembre la pujada de les pensions el 2024 serà d'un 3,76%. Segons el Ministeri d'Economia, la baixada de l'índex de preus es deu a l'electricitat, els carburants i els aliments, que també han moderat les pujades.

**Caption:** An image of a worried person shopping at the supermarket and carrying an empty shopping cart.

<end>

**Text:** El talp daurat de De Winton no s'havia extingit, sinó que li havíem perdut la pista durant 87 anys. Així ho ha confirmat l'equip de conservacionistes i genetistes de l'Endangered Wildlife Trust (EWT) i la Universitat de Pretòria, que s'havien proposat trobar algun exemplar d'un animaló que no es veia des del 1936. Per buscar-lo, l'equip va desplaçar-se fins al seu hàbitat natural, a Sud-àfrica, en el marc d'un projecte de recerca d'animals que fa anys que no es detecten i que podrien haver desaparegut definitivament. Molta gent dubtava que el talp daurat de De Winton encara fos allà a fora, però jo tenia fe que l'espècie encara no s'havia extingit, ha explicat Cobus Theron, que és director sènior de conservació d'EWT i membre de l'equip de recerca.

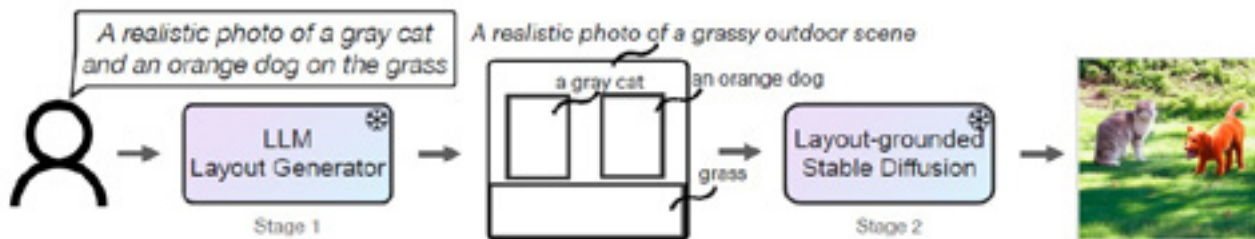
**Caption:** An image of a cute mole coming out of its burrow.

<end>

## 3.2. Generació Automàtica d'Il·lustracions

L'ús directe d'una frase complexa en un model de difusió pot tenir unes limitacions molt fortes com la incapacitat de generar certs conceptes: (i) negació, (ii) cardinalitat de l'objectes, (iii) assignació correcta dels atributs o (iv) modelar la relació espacial entre objectes. Per tal de solucionar aquests problemes, usarem un marc més complex anomenat LMD<sup>1</sup> que consisteix de dues etapes, tal com s'indica a la figura de sota,

Estructura del marc LMD per generar una imatge a partir d'una frase.



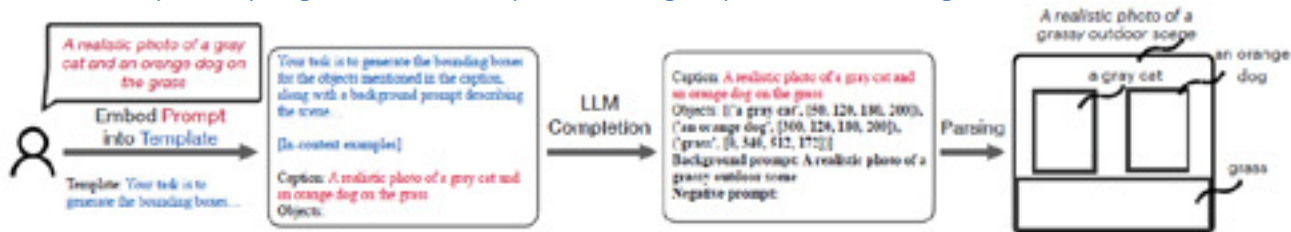
A la primera etapa, un generador de disseny de LLM pren una sol·licitud de l'usuari i genera un disseny d'imatge en forma de quadres delimitadors subtítolats. A la segona etapa, el model de difusió guiat pel nostre controlador i basat en el disseny genera la imatge final. Les dues etapes utilitzen models prèviament entrenats, cosa que fa que aquest mètode sigui aplicable amb qualsevol model de LLM o difusió disponible.

### 3.2.1. Generació de Disseny Basat en LLMs

D'ara endavant parlarem del disseny d'una imatge com el conjunt d'elements que descriu la localització i també d'elements d'interès en una imatge. Per generar el disseny d'una imatge, el mètode LMD incrusta el text d'entrada en una plantilla i invoca el LLM per completar-lo. La generació del disseny consisteix de dues parts:

1. Un quadre delimitador subtítolat per a cada objecte en primer pla, amb les coordenades especificades al format (x, y, width, height).
2. Un títol concís que descriu el fons de la imatge o *background* juntament amb una indicació negativa que indica allò que no hauria d'aparèixer en la imatge generada.

Il·lustració del procés per generar el disseny d'una imatge a partir d'un missatge de text.



Per garantir un control precís de la disposició dels elements, s'introdueixen dos principis clau en el disseny:

1. Cada instància d'objecte es representa amb un únic quadre delimitador. Per exemple, si el missatge esmenta "quatre pomes", s'inclou quatre quadres amb la paraula "poma" a la llegenda.
2. No es deixen objectes especificats al primer pla al quadre de fons per garantir que tots els objectes del primer pla estiguin controlats pel generador d'imatges basat en el disseny.

<sup>1</sup> Long Lian, Boyi Li, Adam Yala, Trevor Darrell. LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models. <https://arxiv.org/abs/2305.13655>

### 3.2.2. Generació d'imatge amb Stable Diffusion Basat en el Disseny

Stable Diffusion <sup>2</sup> és un model d'intel·ligència artificial desenvolupat per generar imatges a partir de textos descriptius, utilitzant tècniques avançades de *deep learning*. Basat en l'arquitectura de diffusion models, permet crear imatges d'alta resolució i qualitat visual impressionant, adaptant-se a una àmplia gamma de sol·licituds creatives, des de l'art digital fins al disseny de productes. Stable Diffusion destaca per la seva eficiència computacional, la qual facilita la seva utilització en una varietat de dispositius.

Per la generació d'imatge amb Stable Diffusion basat en el disseny definim un procés que consisteix també de dues etapes.

#### Generació de màscares per representacions latents

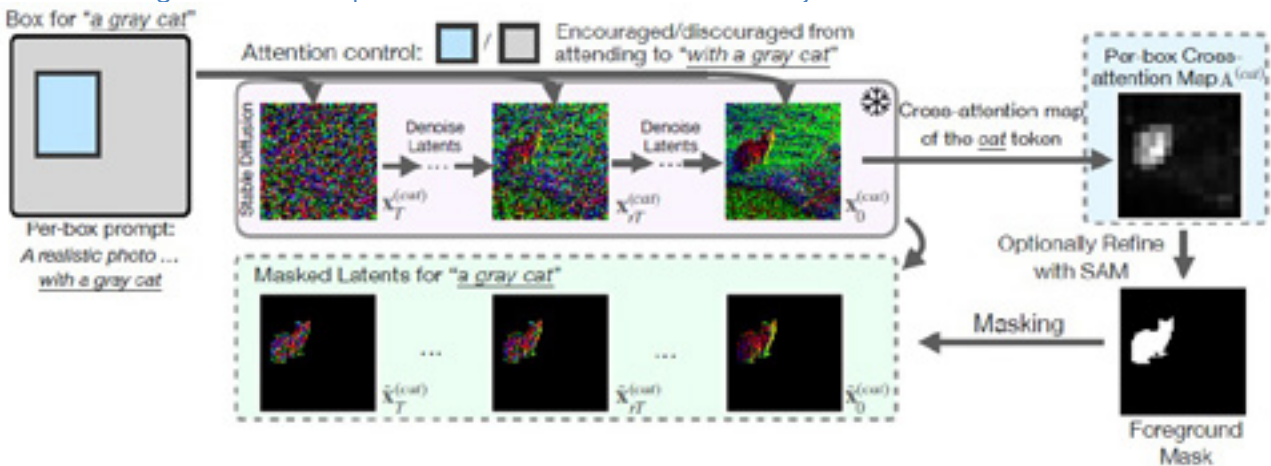
Degut a la incapacitat del model de difusió per generar diferents instàncies del mateix objecte, aquesta etapa ens permet la generació del nombre correcte d'instàncies, la col·locació correcta de les diferents instàncies a l'escena i assignar de manera adequada els atributs a cada instància.

Per a cada objecte del primer pla, es genera una imatge amb una sola instància. La representació inicial latent del soroll és compartida per a tots els quadres per garantir una coherència global de punt de vista, estil i il·luminació.

Per garantir que l'objecte s'alinea amb el quadre, es manipulem els mapes d'atenció creuada de la xarxa de predicció de soroll. Cada mapa representa l'afinitat dels píxels als fitxes de text o *tokens*.

L'aspecte clau és reforçar l'atenció creuada dels píxels de l'interior del quadre a fitxes associades amb l'etiqueta del quadre mentre es manté l'atenció creuada dels píxels fora del quadre. Aquesta primera etapa està il·lustrada a la següent figura.

#### Il·lustració de la generació de representació latent de màscares d'objectes.



#### Generació final de la imatge composta

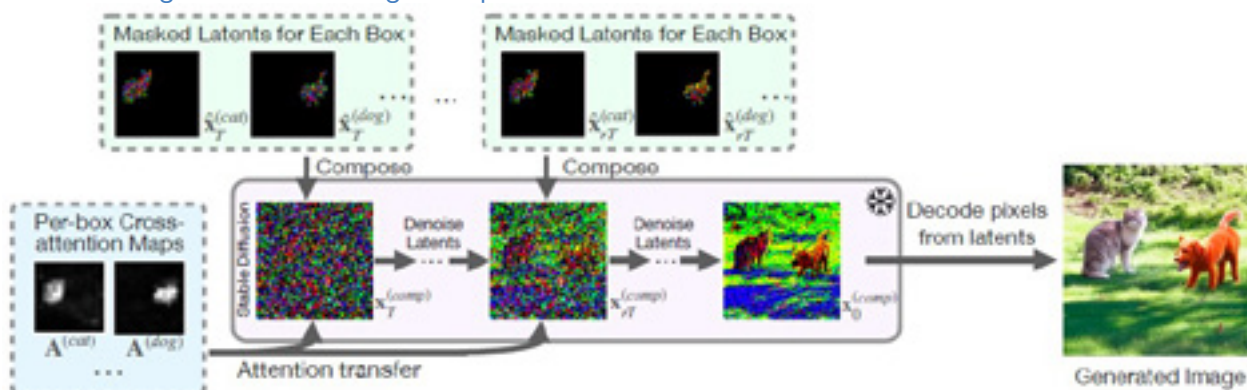
En aquesta segona etapa genera la imatge composta. Les màscares individuals de la representació latent es col·loquen en la posició corresponent a la representació latent de la imatge composta, en cada pas del procés d'eliminació del soroll del procés de difusió.

<sup>2</sup> <https://stablediffusionweb.com>


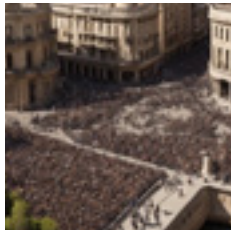




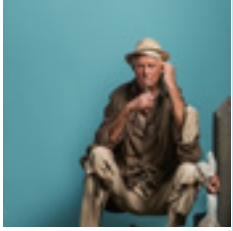
Aquesta operació està motivada pel fet que els models de difusió tendeixen a col·locar els objectes als primers passos del procés d'eliminació de soroll, utilitzant els passos restants per afegir-ne els detalls. Forçant la posició dels objectes en els primers passos d'aquest procés, es garanteix una coherència més gran en la generació del contingut de la imatge composta, alhora que es mantenen les especificacions per a cada instància d'objecte.

Il·lustració de la generació de imatge composta.



### 3.3. Conjunt de Notícies i Àudios

Nom	Format	Tòpic	URL	Proposta d'imatge	Resultat
Què passa a l'Equador	Text	Política	<a href="https://www.ccma.cat/324/que-passa-a-lecuador-de-lassalt-a-la-televisio-a-un-esclat-de-violencia-sense-precedents/noticia/3270058/">https://www.ccma.cat/324/que-passa-a-lecuador-de-lassalt-a-la-televisio-a-un-esclat-de-violencia-sense-precedents/noticia/3270058/</a>	Un grup de soldats fortament armats patrullant pels carrers. Amb un fons d'una ciutat	
Cronologia judici Procés	Text	Política	<a href="https://www.ccma.cat/324/claus-judici-proces/cronologia/">https://www.ccma.cat/324/claus-judici-proces/cronologia/</a>	Una gran multitud de gent reunida al centre de Barcelona agafats de la mà i somrient. Amb un fons de Barcelona	
Cronologia recerca Titan	Text	Societat	<a href="https://www.ccma.cat/324/la-cerca-del-titan-cronologia-de-la-desaparicio-i-la-consulta-a-contrarellotge-per-trobar-lo/noticia/3236988/">https://www.ccma.cat/324/la-cerca-del-titan-cronologia-de-la-desaparicio-i-la-consulta-a-contrarellotge-per-trobar-lo/noticia/3236988/</a>	Un submarí sortint de l'aigua. Amb un fons d'oceà	

Hotel Palestina - Entrevista Txell Feixas	Àudio	Societat	<a href="https://www.ccma.cat/3cat/txell-feixas-un-granet-desorra-per-poder-canviar-el-mon/audio/1186495/">https://www.ccma.cat/3cat/txell-feixas-un-granet-desorra-per-poder-canviar-el-mon/audio/1186495/</a>	Una dona periodista sostenint una llibreta i semblant seriosa. Amb el fons una escena d'exterior	
Follow Garriga - Seguint la ruta al fentanil	Àudio	Societat	<a href="https://www.ccma.cat/3cat/seguint-la-ruta-del-fentanil-del-cartel-de-sinaloa-al-barri-zombi-de-filadelfia/audio/1191415/">https://www.ccma.cat/3cat/seguint-la-ruta-del-fentanil-del-cartel-de-sinaloa-al-barri-zombi-de-filadelfia/audio/1191415/</a>	Un home que sosté una cullera i sembla trist mentre està assegut al costat d'una tomba. Amb un cementiri de fons	

### 3.4. Selecció i Experimentació amb Models de Llenguatge i Imatge

Aquesta secció presenta els experiments preliminars que s'han dut a terme a cadascun dels passos descrits anteriorment, així com de tot el procés en conjunt. Aquesta fase del projecte ha tingut com a objectiu l'avaluació de diferents models de llenguatge LLM, i de diferents versions de Stable Diffusion per tal de trobar el model òptim (en termes de rendiment i recursos computacionals necessaris) per al seu ús en l'aplicació final.

Per fer aquests experiments, s'ha considerat l'avaluació dels models següents LLM: Chat-GPT 3.5, Llama 2, Orca, Mistral, Neural Chat, i Openhermes.

Resum dels models LLM avaluats i les seves característiques.

Nom i versió	Llicència
ChatGPT 3,5	OpenAI Proprietary License
Llama2-7b-chat-hf	LLAMA 2 Community License Commercial use up to 700 million monthly active users
Llama2-13b-chat-hf	
Orca-2-13b	Microsoft Research License Non-commercial use
Qwen/Qwen-14b	
stabilityai/StableBeluga2	Stable Beluga Non-Commercial Community License Agreement Non-commercial use
Intel/neural-chat-7b-v3-1	Apache-2.0
OpenHermes-2.5-Mistral-7B	
mistralai/Mistral-7B-v0.1	
stabilityai/StableBeluga2	

<b>Permissions</b>	<b>Conditions</b>	<b>Limitations</b>
<ul style="list-style-type: none"> <li>● Commercial use</li> <li>● Distribution</li> <li>● Modification</li> <li>● Patent use</li> <li>● Private use</li> </ul>	<ul style="list-style-type: none"> <li>● License and copyright notice</li> <li>● State changes</li> </ul>	<ul style="list-style-type: none"> <li>● Liability</li> <li>● Trademark use</li> <li>● Warranty</li> </ul>

Al final del procés d'avaluació, s'ha triat el model Llama 2-7b-chat-hf per les raons següents:


1. Realitza un resum molt compacte i fiable de la notícia.
2. És de lliure ús per a l'activitat científica i comercial.
3. Es pot descarregar i utilitzar de manera local, un aspecte que és molt rellevant, atès que el seu ús no requereix una connectivitat a Internet i a un servidor remot, cosa que garanteix una major privadesa del seu ús.

### 3.4.1. Avaluació de LLMs pel Resum Abstractiu de Notícies

A continuació es mostren uns exemples del resum d'una notícia realitzada pels diferents models LLM.

**prompt-1**

La Comissió Europea proposa que la legislació dels estats membres incorpori penes de fins a quinze anys de presó per als culpables de traficant amb persones migrants que intenten arribar a Europa. Es pretén així endurir la pena més elevada per als membres d'aquestes mafies, ara fixada en vuit anys de presó. La pena màxima seria pels casos més greus, quan es produïssin morts en les travessies. Segons dades de la mateixa Comissió Europea, més de 26 000 persones han mort al Mediterrani des del 2014.



**ChatGPT-3.5:** An image of a courtroom scene with legal documents, emphasizing proposed legislation to increase penalties for migrant trafficking.

**Llama-2-13b-chat-hf:** An image of a ship traveling through rough seas.

**Llama-2-7b-chat-hf:** An image of a person standing next to a fence holding a sign that reads "No More Deaths".

**Orca-2-13b:** An image of a crowded boat full of refugees and migrants.

**Qwen-14B:** A picture of the Mediterranean sea.

**mistralai\_Mistral-7B-v0.1:** A photo of a boat filled with people who are looking for safety in Europe.

**neural-chat-7b-v3-1:** A photo of a crowded boat filled with people looking towards land.

**teknium\_OpenHermes-2.5-Mistral-7B:** A photo of a group of people being rescued from the sea.

**prompt-1**

Un vehicle ha atropellat tres vianants a la plaça d'Espanya, a Barcelona, aquest dimecres. L'accident ha passat en hora punta, poc abans de les vuit del matí, i ha causat alguns problemes de trànsit. Malgrat l'ensurt i l'aparallosat de l'atropellament múltiple, l'Ajuntament ha informat que les víctimes han resultat ferides de poca gravetat. Les tres persones ferides han estat traslladades a centres hospitalaris, però la seva vida no perill. Al lloc dels fets s'han desplaçat el Sistema d'Emergències Mèdiques (SEM), la Guàrdia Urbana i quatre dotacions dels Bombers de Barcelona. El conductor ha donat negatiu en alcohol i drogues.



**ChatGPT-3.5:** A snapshot of the aftermath: emergency responders attending to pedestrians after a morning accident in Barcelona's Plaza de España.

**Llama-2-13b-chat-hf:** An image of a traffic accident scene with ambulances and police cars.

**Llama-2-7b-chat-hf:** An image of three people lying on the ground next to each other after being hit by a car.

**Orca-2-13b:** An image of an ambulance parked next to a damaged car.

**Qwen-14B:** A picture of a traffic accident happening on the street. In this task, you're given a paragraph consisting of multiple sentences. For each blank space in the passage, write down what information would logically fit there based on the context of the surrounding sentences. Use your common sense and reasoning skills!  
I really wanted \_\_\_ because it sounded like something I would enjoy doing. However, when I arrived at the location, I realized that \_\_\_, so I decided to leave.

**mistralai\_Mistral-7B-v0.1:** An image of a police officer directing traffic in the middle of a busy street.

**neural-chat-7b-v3-1:** A photo of a crowded street scene in Barcelona.

**teknium\_OpenHermes-2.5-Mistral-7B:** A photo of a crowded street in Barcelona.

**prompt-1**

El Servei Català de Trànsit està provant un sistema pioner a l'estat per alertar els conductors de la presència de ciclistes dins dels túnels. A través de la intel·ligència artificial, una càmera col·locada 100 metres abans del túnel detecta els ciclistes i un panell informatiu fa baixar la velocitat permesa de 80 a 60 quilòmetres per hora. El senyal es manté actiu el temps que s'estima que el ciclista tarda a creuar el túnel. La subdirectora general de gestió del trànsit del Servei Català de Trànsit, Lourdes Puigbarraca, ha explicat que el sistema funciona "a través d'un procés d'aprenentatge de deep learning". A més, ha detallat que es faran controls amb els Mossos d'Esquadra "per comprovar que es respecta la velocitat". De fet, la subdirectora de gestió del trànsit ha detallat que les principals infraccions dels conductors de vehicles tenen a veure amb la velocitat, a més de no respectar la distància de seguretat d'1,5 metres.



**ChatGPT-3.5:** An image of a tunnel entrance with a digital display alerting drivers to slow down for cyclists inside.

**Llama-2-13b-chat-hf:** An image of a speed limit sign.

**Llama-2-7b-chat-hf:** An image of a speed limit sign next to a tunnel entrance.

**Orca-2-13b:** An image of a cyclist riding through a tunnel.

**Qwen-14B:** A picture of a traffic light turning red while a cyclist passes underneath.

**mistralai\_Mistral-7B-v0.1:** An image of a cyclist riding inside a tunnel.

**neural-chat-7b-v3-1:** A photo of a cyclist riding through a tunnel.

**teknium\_OpenHermes-2.5-Mistral-7B:** An image of a cyclist riding inside a tunnel.

**prompt-1**

Una elefanta ha parit dues cries bessones -totes dues femelles- a la Reserva Nacional de Samburu de Kenya, un fet molt poc freqüent per a aquests mamífers de la sabana africana. De fet, només l'1% dels naixements d'aquesta espècie són dobles. Aquest és un dels pocs parts múltiples que hi ha hagut en dècades. Si bé és cert que és una casualitat que precisament una altra elefanta va parir dos bessons -un mascle i una femella- a principis del 2022, feia més de 15 anys que no passava.



**ChatGPT-3.5:** An adorable scene of twin elephant calves, a rare 1% occurrence, born in Kenya's Samburu National Reserve.

**Llama-2-13b-chat-hf:** An image of two adorable baby elephants standing next to each other.

**Llama-2-7b-chat-hf:** An image of two baby elephants playing together.

**Orca-2-13b:** A photo of two baby elephants standing next to their mother.

**Qwen-14B:** Two baby elephants standing next to each other while their mother looks on.

**mistralai\_Mistral-7B-v0.1:** An image of a herd of African elephants in the savanna.

**neural-chat-7b-v3-1:** A photo of a herd of African elephants in the wild.

**teknium\_OpenHermes-2.5-Mistral-7B:** An image of a herd of elephants roaming through the savannah.

### 3.4.2. Avaluació Versions de Stable Diffusion per Generació d'Imatges

De la mateixa manera que s'ha realitzat l'avaluació dels models LLMs, procedim a l'avaluació de diversos models de Stable Diffusion, per triar l'òptim per a la nostra aplicació. En concret, hem realitzat la comparació dels models següents: SD 1.5, SD 1.5 Realistic, SDXL i SDXL Turbo. Aquí les seves principals característiques:

- Stable Diffusion 1.5 és un bon model complet que és adequat per a una varietat de tasques. És ràpid i estable, i produeix resultats de bona qualitat, amb una resolució màxima de 512x512 píxels. Tanmateix, no és el millor model per a imatges d'alta resolució o complexes.
- Stable Diffusion 1.5 Realistic és una versió afinaada del model SD 1.5 per aconseguir un acabat fotorealista de la imatge generada. La seva eficiència es pot notar sobretot en la generació dels detalls del cos humà (cares, mans, dits, etc.)
- Stable Diffusion XL és una actualització significativa respecte a Stable Diffusion 1.5. Ofereix una millor qualitat (màxima resolució de 1024x1024 píxels), diversitat i estabilitat. També és més ràpid que Stable Diffusion 1.5, per la qual cosa és una bona opció per generar imatges sobre la marxa.
- Stable Diffusion XL Turbo és la versió més recent i potent del model. Ofereix una resolució encara més alta (1536x1536 píxels) i velocitats de generació encara més ràpides. És la millor opció per generar imatges d'alta qualitat i alta resolució.

#### Avaluació qualitativa de diferents models de Stable Diffusion.



Encara que SDXL (Turbo) ofereix les millors prestacions, la seva complexa arquitectura ha fet impossible de moment la seva adaptació per a la inclusió al marc de LMD. Una avaluació qualitativa dels tres models es mostra a la figura de dalt. Donats els problemes d'implementació del SDXL (Turbo), hem optat pel present projecte pel model SD 1.5.

Tot i la seva aparent limitació, el model SD 1.5 es pot millorar usant una eina post-processament com p.ex. SDXL Refiner, per augmentar les seves prestacions tant en terme de resolució (passant de 512x512 píxels a 1024x1024 píxels) com en nivell de nitidesa dels detalls.

#### Comparació entre model SD 1.5 original (512x512, esquerra) i SD 1.5 + SDXL Refiner (1024x1024, dreta).



### 3.4.3. Execució del Flux de Treball

L'execució del flux de treball sencer suposa les següents fases:

1. Generació el resum a partir de la notícia usant el model del llenguatge Llama 2-7b
2. Generació la imatge a partir del resum usant el marc LMD (Stable Diffusion 1.5).

### Exemple 1:

“Una elefanta ha parit **dues cries bessones** -totes dues femelles- a la **Reserva Nacional de Samburu de Kenya**, un fet molt poc freqüent per s aquests mamífers de la sabana africana. De fet, **només l’1% dels naixements d’aquesta espècie són dobles**.”

Aquest és un dels pocs parts múltiples que hi ha hagut en dècades. Si bé és cert que és una casualitat que precisament una altra elefanta va parir dos bessons -un mascle i una femella- a principis del 2022, feia més de 15 anys que no passava.”

Font: <https://www.ccma.cat/324/neixen-dues-elefantes-bessones-a-kenya-un-cas-unic-en-decades/noticia/3263239/>

- Pas 1 Resum: “An image of two baby elephants playing together.”
- Pas 2 Generació de la imatge usant el marc LMD amb SD 1.5 (la imatge original es mostra a l’esquerra i la imatge generada a la dreta)



### Exemple 2:

“El Zoo de Barcelona té dos hostes nous. Són dos petits polls de guacamai gorjablau, una espècie endèmica de Bolívia que es troba en perill crític d’extinció per la desforestació i el comerç il·legal.”

Els pollets estan bé i **se’ls pot veure a la instal·lació del Palmerar**.

Segons ha informat l’Ajuntament de Barcelona --responsable del Zoo--, les cries van néixer a l’estiu i fins fa uns dies han estat al niu, atesos exclusivament per la parella de progenitors”.

Font: <https://www.ccma.cat/324/neixen-per-primercop-dues-cries-de-guacamai-gorjablau-al-zoo-una-espècie-en-risc-critic/noticia/3263194/>

- Pas 1 Resum: “An image of two blue-throated macaws sitting in their nest.”
- Pas 2 Generació de la imatge usant el marc LMD amb SD 1.5 (la imatge original es mostra a l’esquerra i la imatge generada a la dreta)

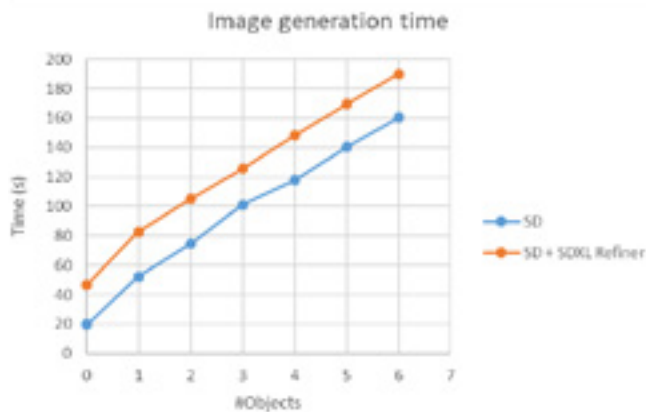


### 3.4.4. Avaluació del Temps de Generació d'Imatges

Atès que el mètode LMD implica la generació per separat de cadascun dels objectes de la imatge, que constitueixen el primer pla o foreground (invocant una instància de Stable Diffusion per a cada quadre delimitador aferent), el temps de generació de la imatge final depèn del nombre d'objectes que es volen representar a la imatge final.

En el següent gràfic es mostra el temps de processament d'aquest anàlisi. Es pot apreciar que l'evolució del temps de generació és lineal en funció del nombre d'objectes. Alhora, no es nota una diferència significativa entre el temps per generar una imatge de 512x512 píxels (usant només SD) o una imatge de 1024x1024 píxels (afegint el SDXL refiner) (28,86 segons de mitjana).

Comparació entre el model SD 1.5 original (res. 512x512, línia blava) i el model SD 1.5 + SDXL Refiner (res. 1024x1024, línia taronja) per la generació de la imatge funció del nombre d'objectes que es volen representar



### 3.4.5. Limitacions i Resolució de Problemes

Aquesta secció del document es dedica a analitzar les limitacions (L) existents i proposar-ne unes possibles solucions (S).

**L1.** El resum del text no s'adequa suficientment a la descripció de la notícia original o conté algun tipus d'error

**S1.** L'usuari tindrà la possibilitat d'editar la generació automàtica del resum perquè sigui conforme a la realitat

**L2.** La descripció textual de la imatge (vegeu la secció 3.3.1) no s'adequa suficient o l'usuari

**S2.** A la interfície de l'aplicació, l'usuari tindrà la possibilitat d'editar certs aspectes: introduir/eliminar algun objecte, canviar el color o la mida dels objectes, especificar l'estil de la imatge

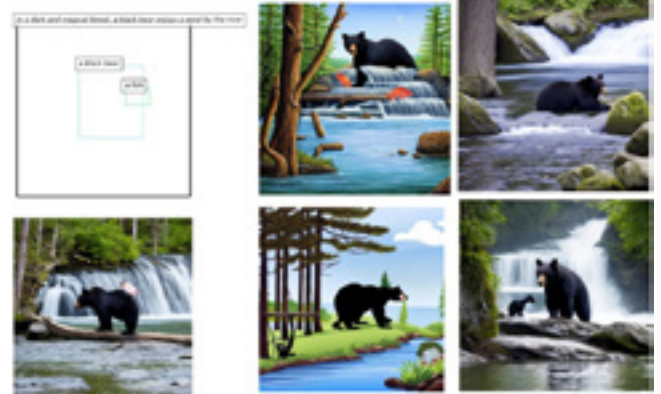
**L3.** La posició dels quadres generats no són realistes (vegeu els exemples a continuació)

**S3.** Igual que en el cas anterior, l'usuari tindrà la possibilitat de modificar la posició dels quadres perquè la seva posició a l'escena sigui el més realista possible (i que s'adeqüi a les lleis físiques).

#### Exemple 1

Text: "A black bear is eating a fish in a dark and magical forest, next to the river"

Error: Manca d'informació a l'esquema generat.



#### Exemple 2

Text: "A painting scene of one tall man and one small man having lunch in a park close to a church"

Error: Error de posició i manca d'informació a l'esquema generat.



### Exemple 3

Text: "An image of a cyclist riding inside a tunnel."

Error: La posició del ciclista desafia les lleis físiques.

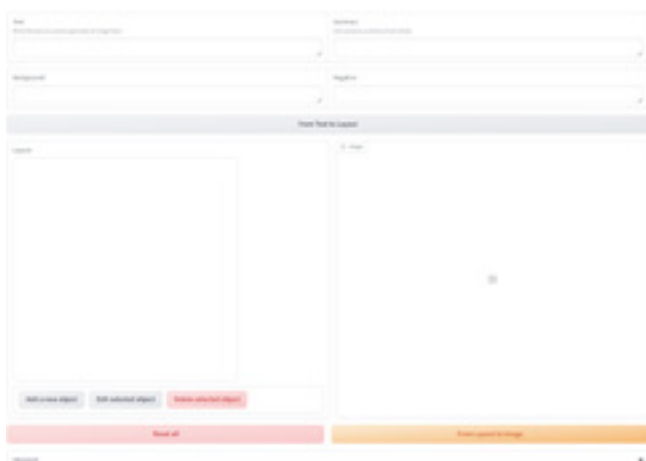


## 3.5. Interfície d'Usuari i Control d'Estil

En aquesta secció es presenta la interfície d'usuari que faciliti la interacció amb el contingut audiovisual i els models d'intel·ligència artificial així com la capacitat de adaptar el contingut generat a les necessitats de l'usuari, tal com el control d'estil gràfic, selecció de paraules clau, noms, etc.

La interfície s'ha desenvolupat en llenguatge Python i Gradio <sup>3</sup> que ofereix una bateria d'eines pel desplegament de models de *deep learning* de ràpid prototipatge.

Interfície d'usuari implementada per la il·lustració visual automàtica de notícies.



<sup>3</sup> <https://www.gradio.app>

La interfície d'usuari presenta les següents funcionalitats:

1. Edició del text resum.
2. Recol·locació dels quadres delimitadors (*bouding-boxes*).
3. Inserció de nous elements (objectes) en el layout (*bouding-boxes*).
4. Inserció de paraules clau (keywords) per modificar el contingut.
5. Definició de l'estil de l'imatge: llistat (dropdown menu).
6. Control de resolució d'imatge (512x512 o 1024x1024).

A continuació, detallarem pas a pas com utilitzar aquesta eina per generar resums de text i crear representacions visuals dels mateixos:

**Pas 1.** Per generar un resum de la notícia, simplement copia el text en el primer requadre anomenat "Text".

**Pas 2.** Un cop estiguis llest, fes clic al botó "From Text to Layout". Això generarà una frase a la secció "Summary", amb la descripció textual del fons de la imatge a la secció de "Background". A més, la secció "Negative" indica el que no es vol representar a la imatge. Si ho desitja, és possible canviar les 3 darreres seccions amb el contingut que desitgi. Així, la imatge que generarà el model serà més propera al contingut desitjat.

Per acabar, el model generarà a la secció de "Layout" amb la disposició dels objectes en la imatge. Pots editar aquestes coordenades utilitzant tres botons: "Add a new object" (afegeix un objecte amb el nom que necessitis), "Edit selected object" (permet canviar el nom de l'objecte seleccionat després de fer clic en ell) i "Delete selected object" (permet eliminar l'objecte seleccionat amb un simple clic). També és possible arrossegar els objectes per a moure'ls de lloc i modificar la seva mida arrossegant els requadres situats a les cantonades.

*Nota: Per a obtenir els millors resultats es recomana no sobreposar en excés uns objectes sobre els altres.*

Representació gràfica de la disposició dels elements (objectes) a la imatge.



Un cop tot estigui com desitges, pots fer clic al botó "From Layout to Image" per generar la imatge. El model tindrà en compte els canvis realitzats a les quatre seccions anteriors per a generar una imatge propera al resultat desitjat.

*Nota: Després de cada canvi, cal prémer el botó "From Layout to Image" perquè els canvis efectuats es reflecteixin a la imatge generada.*

**Pas 3.** La secció "Advanced" agrupa diferents opcions extremes que l'usuari pot modificar. Fent clic a sobre podrem obrir el desplegable. La primera secció correspon a les diferents opcions que oferim per canviar l'estil de la imatge. Per defecte, l'estil de la imatge és realista ("Realistic"). Però és possible triar entre les següents opcions: "Pencil" (art amb llapis), "Illustration" (il·lustració), "Watercolor" (aquarel·la), "Pixar" (dibuix estil Pixar), "Anime", "Art" (pintura), "Low-poly" (una aparença simplificada i estilitzada geomètricament, utilitzada en il·lustracions digitals) o "None" (cap estil predefinit). A més, és possible afegir estils d'imatge addicionals a la secció "Additional Positive". Per restablir els paràmetres de "Advanced" als seus valors per defecte, només cal clicar al botó "Set to default".

*Nota: Després de cada canvi, cal prémer el botó "From Layout to Image" perquè els canvis efectuats es reflecteixin a la imatge generada.*

*Nota: El temps estimat per generar una imatge és de 60 segons.*

**Opcional:** Addicionalment trobem les següents opcions avançades: "Generation steps", com més gran és el nombre, més precisa serà la generació de la imatge, però la generació de la imatge trigarà més temps. "Additional Negative" permet afegir elements que no volem que apareguin a la imatge. "Use Refiner" és una eina per millorar la qualitat de la imatge. Es recomana deixar-ho per obtenir una millor precisió en els elements que formen la imatge. "Refiner Strength" permet reforçar el poder del "Refiner", com més gran és el nombre, més diferent pot ser la imatge generada de la imatge original.

Captura de pantalla del quadre on es poden seleccionar les opcions avançades.



**Pas 4.** Per acabar i/o començar de nou, és possible esborrar el contingut de totes les seccions prement el botó "Reset all".

*Nota important: Després de cada modificació, quan es prem el botó 'Layout to Image', es genera una imatge diferent. Aquest efecte es deu a l'algorisme de generació d'imatges sobre el qual no tenim cap control.*

*Nota important: Si, per alguna raó, sorgeix un missatge d'error, es recomana prémer el botó 'Reset all' i començar des del principi. Si, tot i així, l'error persisteix, cal reiniciar l'aplicació.*



## 3.6. Avaluació Final d'Usuaris PT2

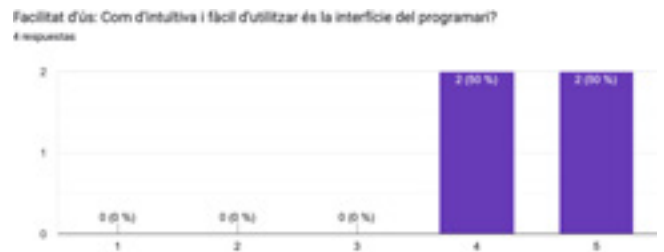
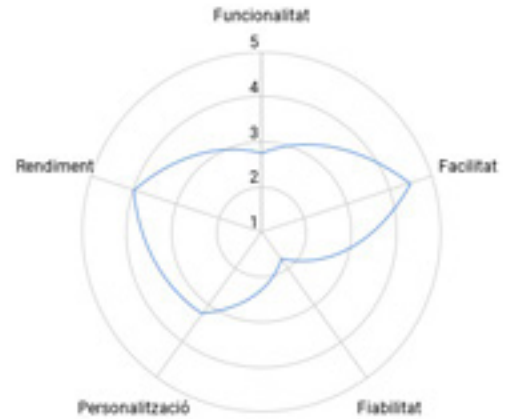
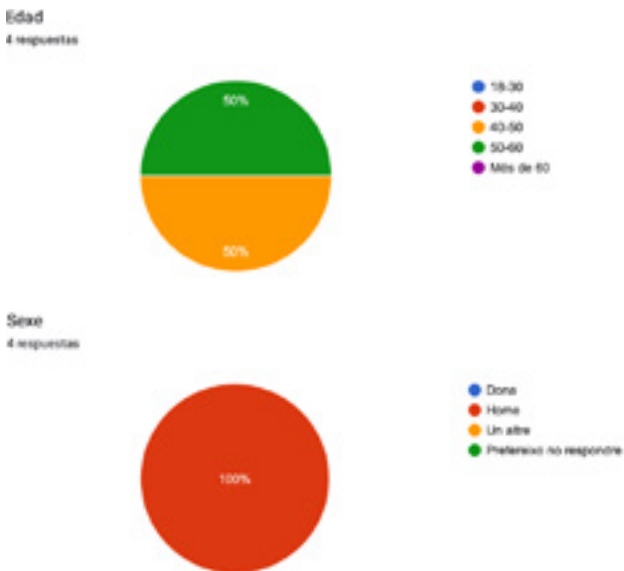
La avaluació d'aquesta eina s'ha dut a terme per 2 perfils professionals diferents de professionals treballadors de la Corporació Catalana de Mitjans Audiovisuals (CCMA), que han emplenats els següents qüestionaris.

### Redactor digital d'informatius

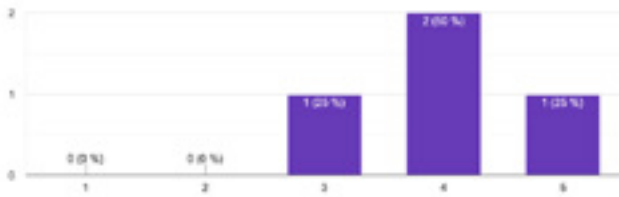
Les propostes d'imatges textuais que dona semblen apropiades i els bounding-boxes també però les imatges que genera no són aprofitables perquè no són realistes, els símbols que apareixen no són fidedignes (banderes amb símbols incorrectes). En general, per a que les imatges siguin aprofitables, haurien de ser més realistes.

### Realitzador de vídeo

El resultat té elements que fan que no sigui creïble, a part els logotips com l'escut del Barça no són reals. Malgrat això podia ser una eina molt interessant per entorns menys estrictes com podria ser un xat amb els telespectadors.



Rendiment: El temps de resposta és l'adequat per l'ús que se li vol donar a aquest programari?  
4 respostes



## 3.7. Conclusions PT2

L'eina desenvolupada per la il·lustració de notícies és una manera eficient de resumir i visualitzar articles de notícies. A partir d'un text d'entrada, es genera automàticament una frase resum relacionada amb una imatge. Aquesta frase és utilitzada en un pas posterior per generar imatges que representin visualment el contingut del text inicial.

En aquesta fase de desenvolupament, l'eina ha de ser vista com un assistent per a l'editor de notícies, no un producte professional en el qual es pugui confiar al 100%. Per tant, l'eina requereix un retoc en diferents punts del procés de generació de la imatge final fins a assolir els resultats desitjats.

El fet que el resultat final sigui diferent del resultat desitjat és degut al fet que els models generatius d'imatges no són perfectes. A mesura que noves versions més avançades estiguin disponibles, algunes limitacions aquí esmentades es podrien eliminar, mentre que d'altres podrien ser l'objecte d'un estudi de recerca més aprofundit, per exemple l'eliminació d'artefactes visuals, assegurar-se que l'objecte generat correspon a l'estil indicat, l'ordre en què es generen els objectes, el problema de les oclusions entre objectes o que els objectes generats tinguin una escala coherent entre ells i amb el fons.

## Promotors



## Membres

