



La IA pública

Albert Cañigueras (@AlbertCanig)

Responsable de Coordinació i Desenvolupament

d'IA i tecnologies del Llenguatge

Barcelona Supercomputing Center (BSC)

#OpenGovWeek



**Diputació
Barcelona**

Àrea de Bon Govern, Assistència
Local i Cohesió Territorial



IA Pública: una visió per a camí diferent per a la IA

a) Els governs i l'interès públic NO han de quedar relegats al paper de reguladors de la tecnologia. L'àmbit públic pot ser **inventor**, **constructor** de tecnologia i ajudar a **donar forma als mercats**. Alineant **objectius públics**, **inversió pública** i buscant que els **beneficis de la IA arribin a tothom**.

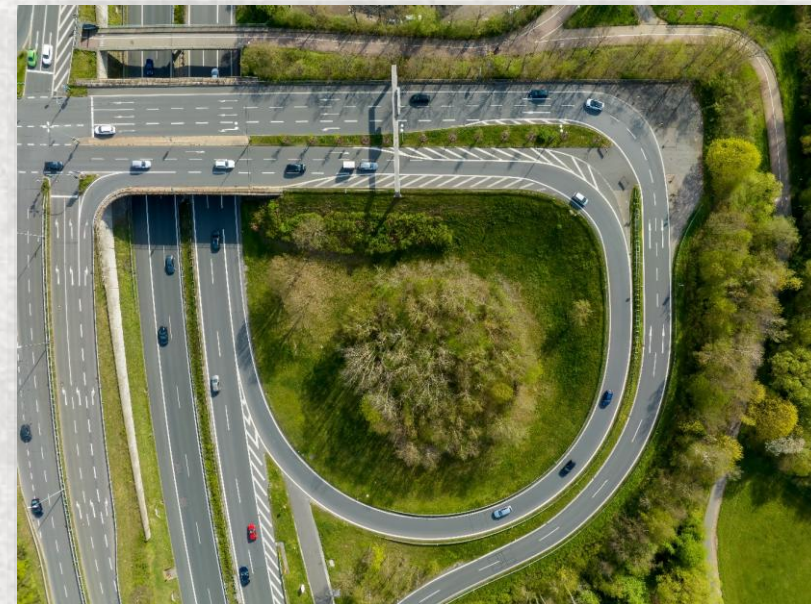
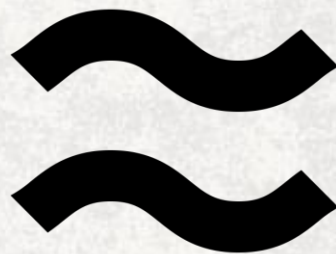
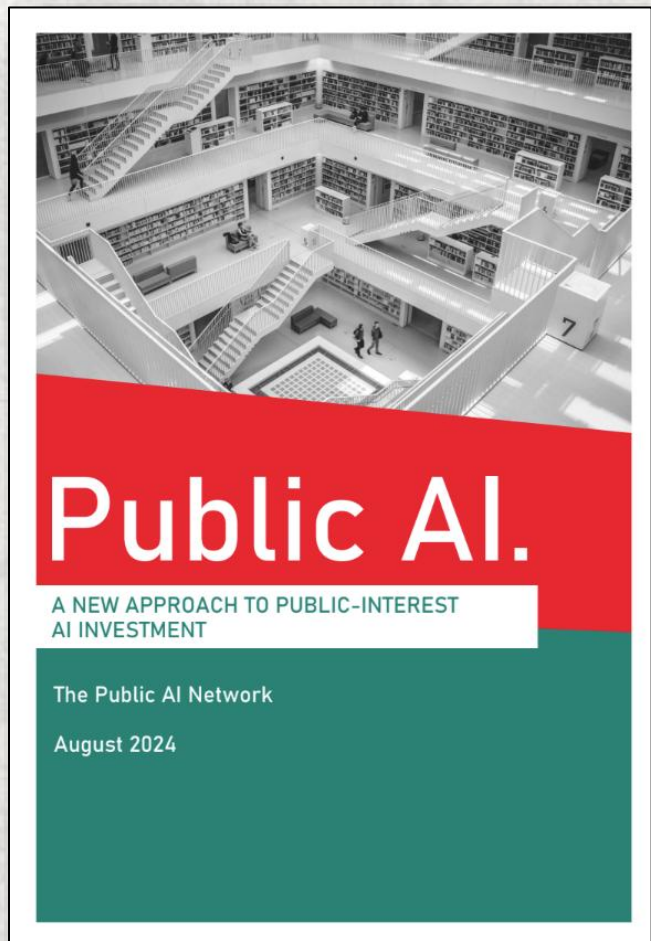
b) Això s'està fent (al BSC!) i està disponible per integrar-se a les vostres aplicacions i als vostres serveis

La IA pràctica. Casos d'ús al sector públic

- 1) Startup – Omnios
- 2) Startup – 1MillionBot
- 3) Startup - Recog

La visió d'una IA pública

Infrastructures per al bé comú



Per què necessitem una IA pública?



Sistemes d'IA irresponsables



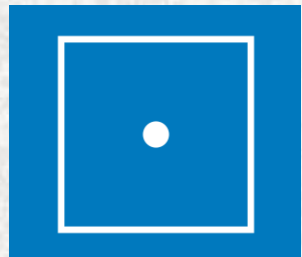
IA a la que poder fer confinaça



Extracció injusta de dades



Bens comuns digitals



Concentració de poder



Condicions de joc igualades





El futuro no es lo que
va a pasar, es lo que
vamos a hacer”

Jorge Luis Borges

La IA pública, a la pràctica

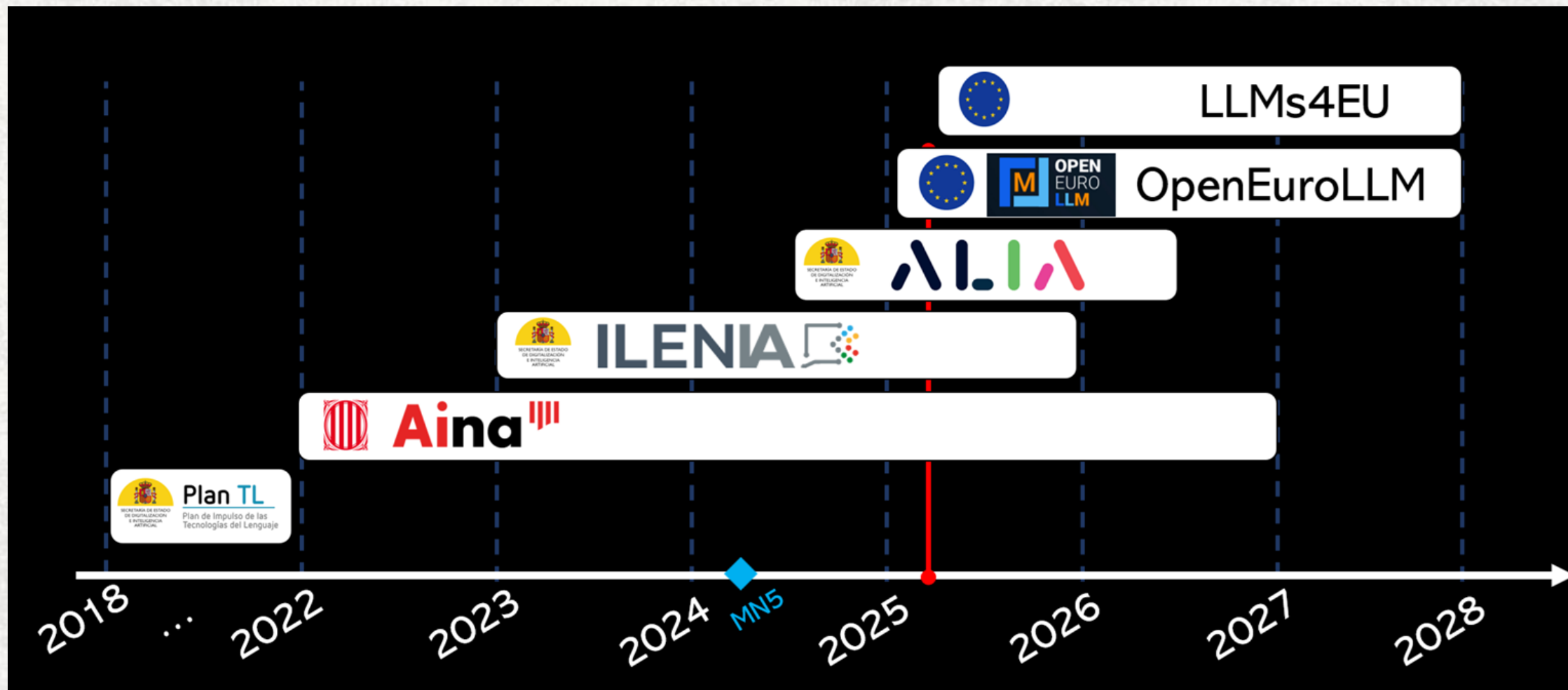
IA Pública pràctica – El talent al BSC

Language Technologies Laboratory



The Language Technologies Unit at BSC aims to advance the field of natural language processing through cutting-edge research and development and use of HPC. The Unit has extensive experience in several NLP areas, such as massive language model building, machine translation, speech technologies, HPC computing and unsupervised learning for under-resourced languages and domains. High-throughput language processing and massive language data annotation are part of the unique expertise of a team comprising computational linguists, data scientists, software engineers and researchers coming together from various backgrounds and domains.

IA Pública pràctica – Finançament CAT + ES + EU



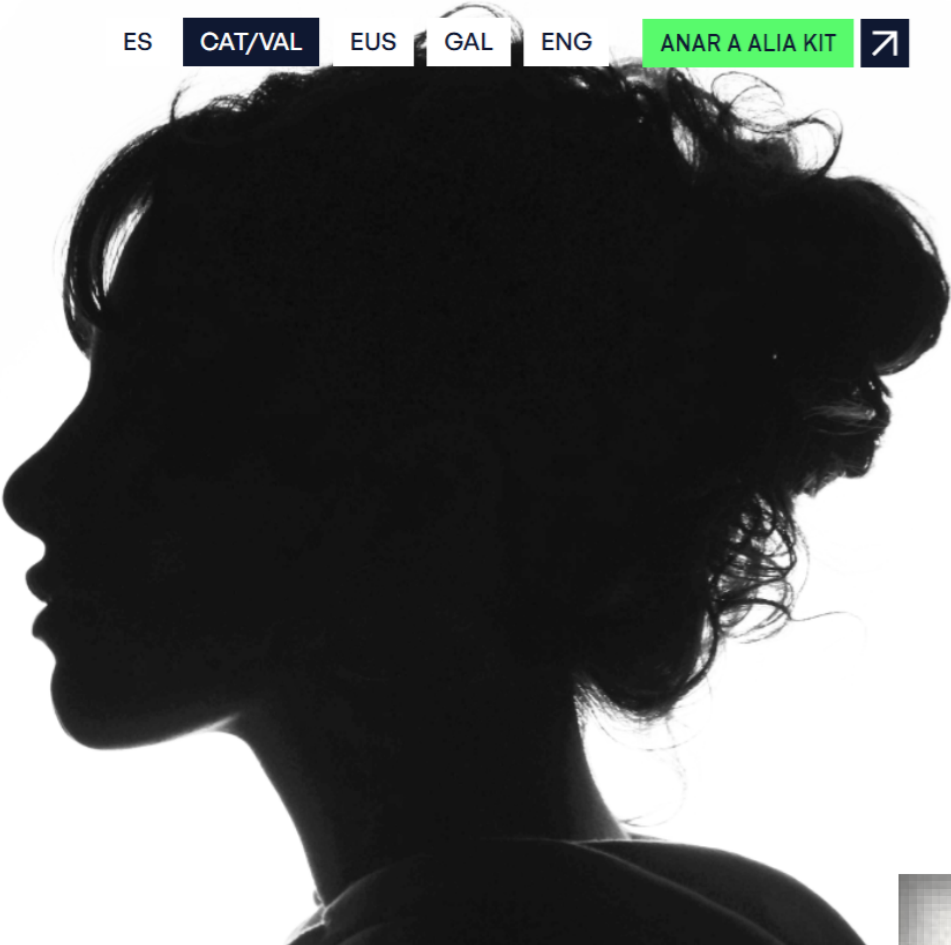
IA Pública pràctica - <https://alia.gob.es/>



ES CAT/VAL EUS GAL ENG ANAR A ALIA KIT 

LA INFRAESTRUCTURA PÚBLICA D'IA EN CASTELLÀ I LLENGÜES COOFICIALS

ALIA és una iniciativa pionera a la Unió Europea que busca proporcionar una infraestructura pública de recursos d'IA, com els models de llenguatge oberts i transparents, per a fomentar l'impuls del castellà i les llengües cooficials -català i valencià, basc i gallec- en el desenvolupament i desplegament de la IA en el món.



IA Pública pràctica - <https://projecteaina.cat/>



El Projecte Aina genera i posa a l'abast de tothom **recursos lingüístics, corpus i models informàtics** perquè desenvolupadores d'aplicacions basades en IA puguin fer-ho fàcilment en català.



Assistents de veu



Traductors i correctors
automàtics

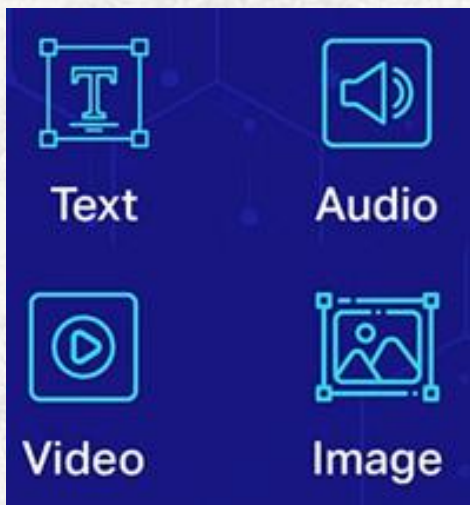


Cercadors d'Internet



Agents conversacionals

Per què podem parlar d'una IA pública?



**Dades
transparentes
i legals**

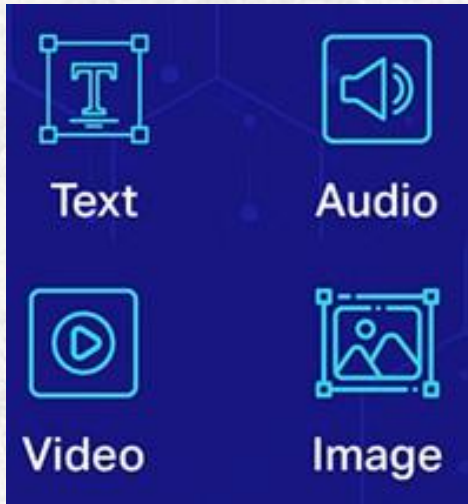


**Super
computació
pública**



**Models
oberts i
certificats**

Per què podem parlar d'una IA pública?



**Dades
transparentes
i legals**

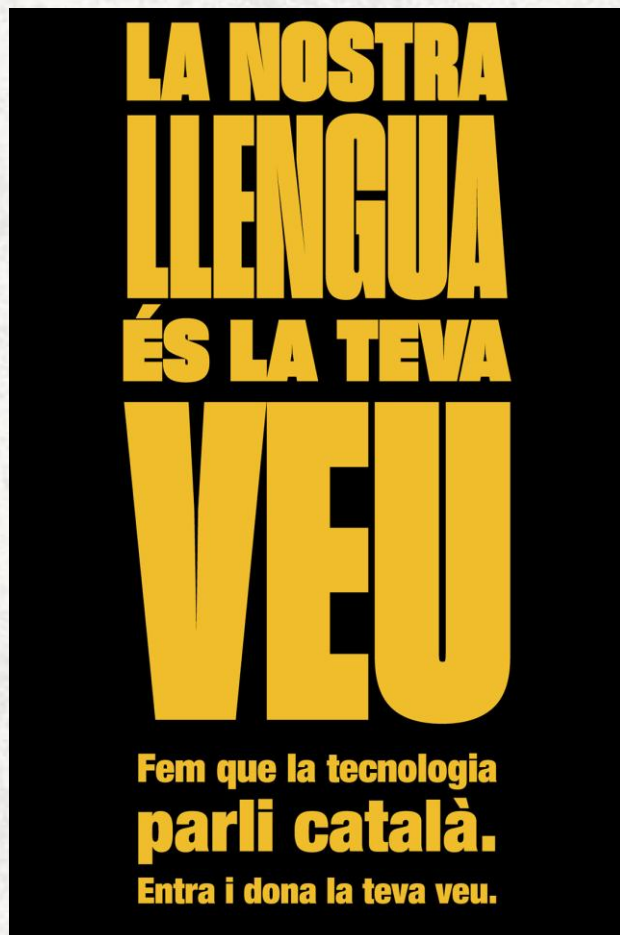


**Super
computació
pública**



**Models
oberts i
certificats**

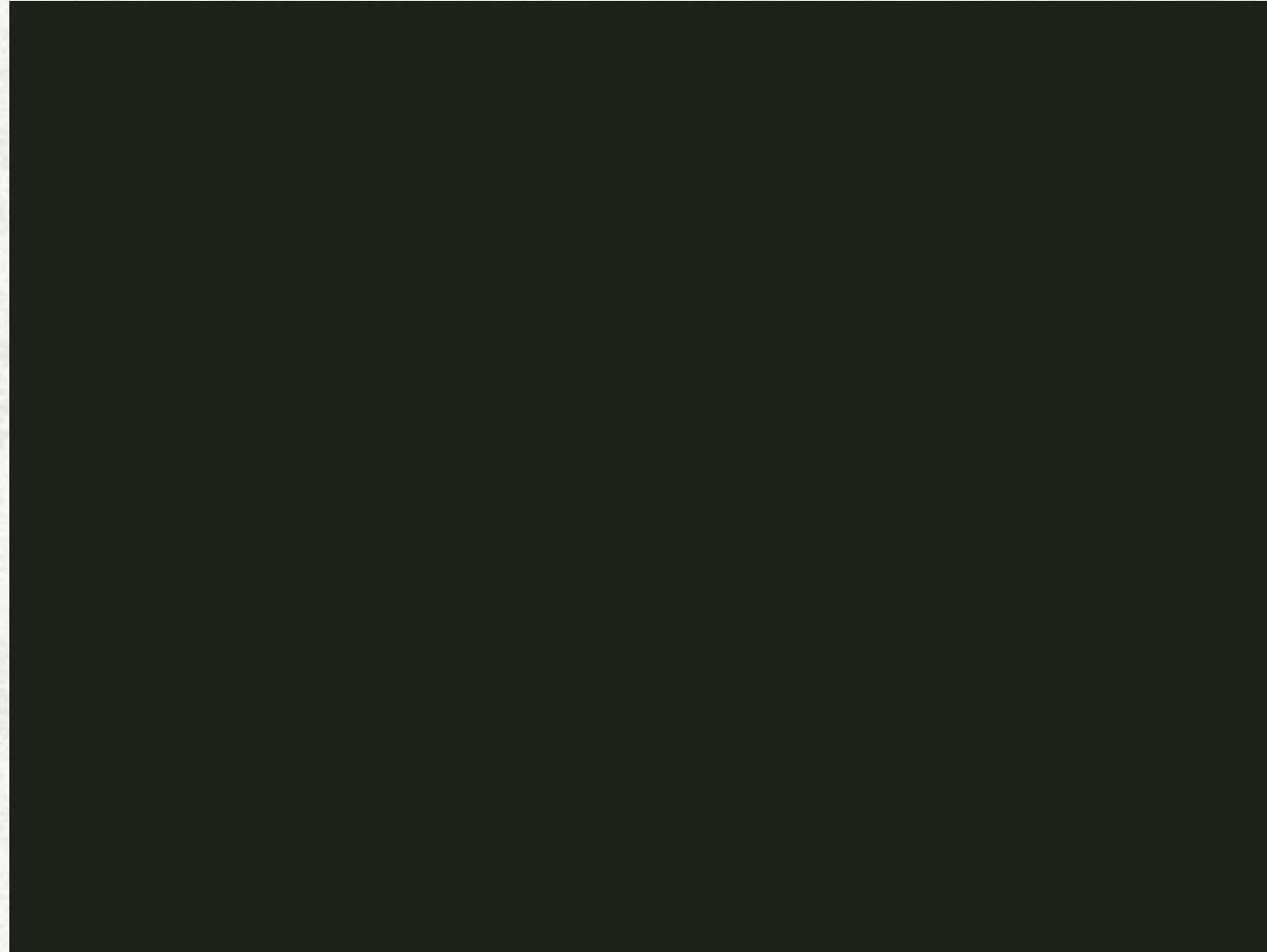
Campanya de llançament (2021)



Col·laboració amb Govern Illes Balears

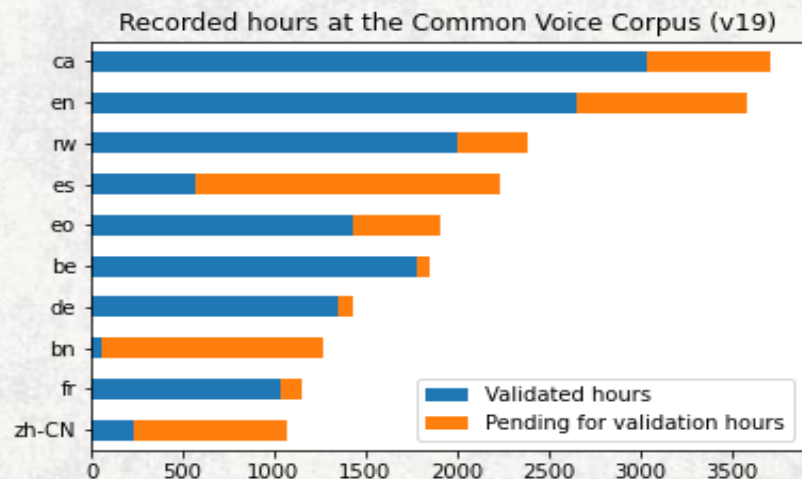


Col·laboració amb Plataforma per la Llengua i SoftCatalà



IA Pública pràctica – Dades de veu

El català, primer idioma a Common Voice



Presència als models de veu oberts

- Whisper; OpenAI
- audioPaLM; Google
- SeamlessM4T; Meta
- Google USM; Google

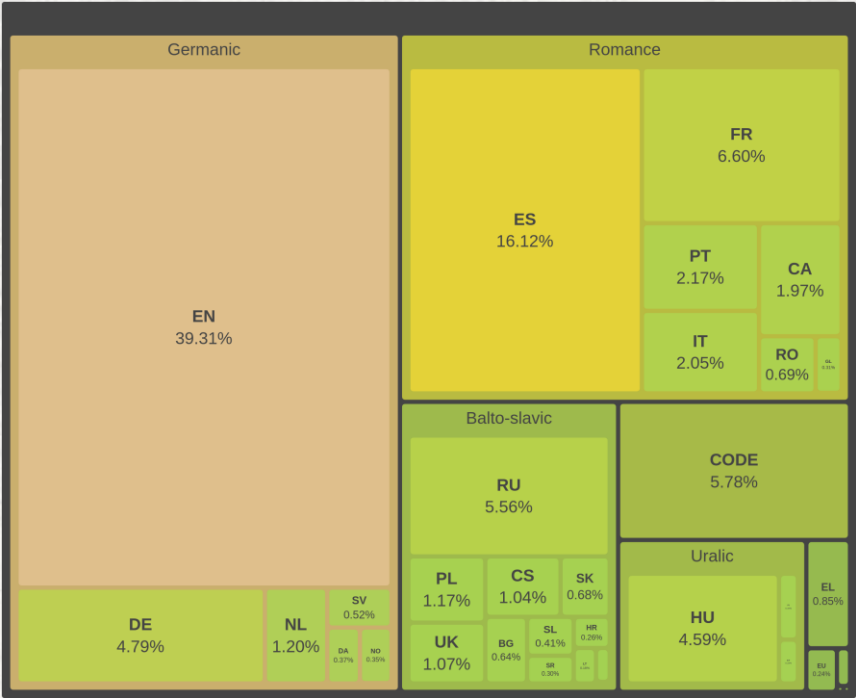
Dades de veu disponibles

Common Voice	3.041h
3cat	770h
ParlamentParla	946h
CortsParla	240h
IB3	28h
Parlament Balear	120h

IA Pública pràctica – Dades de text

Distribution of Salamandra's corpus

Tokens CAT: 53k M tokens (53.260.764.902)
GB CAT: 142.5GB



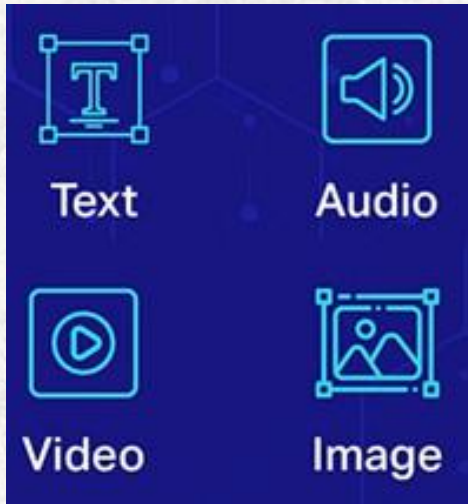
Col·laboracions

- Agència Catalana de Notícies
- Ajuntament de Girona
- Bífidus Produccions SL
- Corporació Catalana de Mitjans Audiovisuals, SA
- Editorial Singular Digital 2GR, SL
- Ens Públic de Radiotelevisió de les Illes Balears (IB3)
- Edicions del Núvol
- Fundació Enciclopèdia Catalana
- Grup Enderrock Edicions SL
- Guiacat
- Grup Abacus
- Informació i Comunicació de Barcelona, SA
- La Fresca Produccions
- La Vanguardia
- Llenguaferrits
- Obra Cultural Balear
- Ràdio Star Terrasa
- Ràdio Falset
- Singular Digital
- Tirabol Produccions, SL (Racó Català)
- Universitat Autònoma de Barcelona
- Universitat Illes Balears
- Universitat Politècnica de Catalunya
- Universitat de València
- Universitat de Vic
- Unió de Pagesos de Catalunya
- iSabadell/ Grup Edicions de Premsa Local, SL
- Sra. M. Carme Marí Vila



Model	Tamany CAT	% CAT
GPT-3 (2020)	35.126.650	0.01798%
Salamandra 2024	53.260.764.902	1.9410%

Per què podem parlar d'una IA pública?



**Dades
transparentes
i legals**



**Super
computació
pública**



**Models
oberts i
certificats**

Fa 20 anys...



#OpenGovWeek

Marenostrum 5

Total peak performance: 315.2 Pflops		
General Purpose Partition:	46.4 Pflops	(29-04-2024)
Accelerated Partition:	260 Pflops	(29-04-2024)
Next Generation GPP:	2.82 Pflops	(WiP)
Next Generation ACC:	6 Pflops	(TBA)



MareNostrum 1
2004 – 42.3 Tflops
1st Europe / 4th World
New technologies

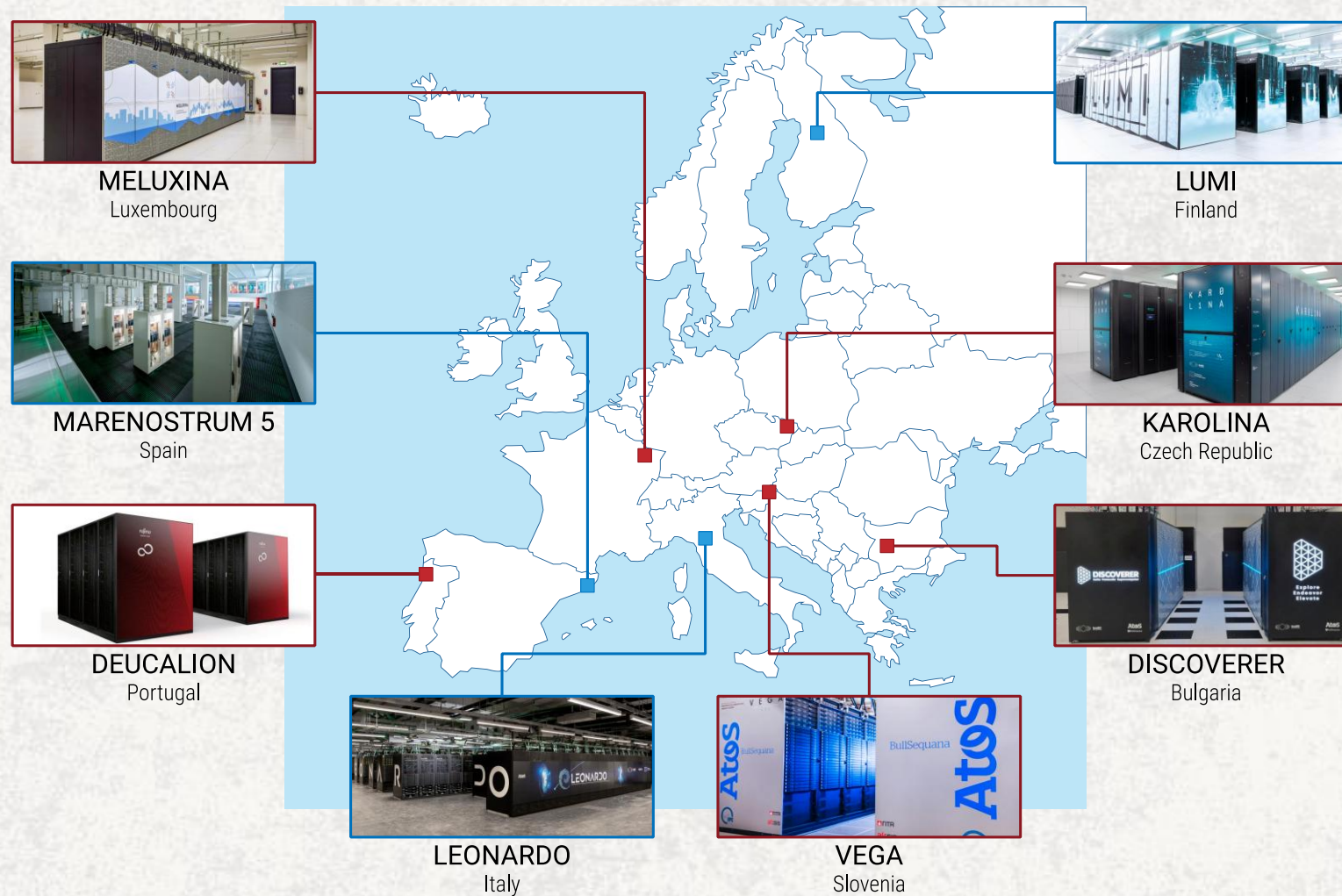
MareNostrum 2
2006 – 94.2 Tflops
1st Europe / 5th World
New technologies

MareNostrum 3
2012 – 1.1 Pflops
12th Europe / 36th World

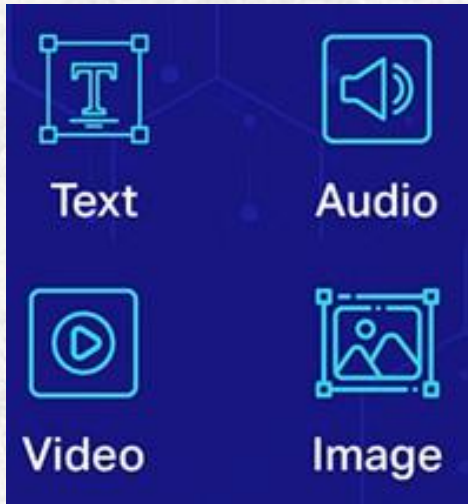
MareNostrum 4
2017 – 11.1 Pflops
2nd Europe / 13th World
New technologies

MareNostrum 5
2022
260 + 46.4 Pflops
8th and 19th World
3rd and 7th Europe

EuroHPC: towards European HPC technologies



Per què podem parlar d'una IA pública?



Dades
transparentes
i legals



Super
computació
pública



Models
oberts i
certificats

IA Pública pràctica – Aina Kit



<https://bsc.es/ainakit>

 **Aina Kit**

Ctrl + K

Inici

 **MODELS**

 Models de text

 Models de parla

 Models de traducció automàtica

 Avaluació de models en llengua catalana

DATASETS

 Dades i eines per a models de text

 Dades i eines per a models de veu

 Dades per a la traducció automàtica

TESTING

 Powered by GitBook

Inici

L'Aina Kit

Aina Kit és el recull de documentació i recursos que el [projecte Aina](#) posa a disposició de les persones desenvolupadores de productes i serveis d'intel·ligència artificial (IA) i tecnologies del llenguatge (TL) en català.

COMUNITAT DE DESENVOLUPADOR/ES

DOCUMENTACIÓ I RECURSOS

MODELS IA I DATASETS

 Barcelona Supercomputing Center
Centro Nacional de Supercomputación



La família de models d'IA – Text Generatiu

Models de text Salamandra

Models generatius **multilingüe** (35 llengües d'Europa) entrenats des de zero amb **12,9 bilions de tokens**.

- mides 2B, 7B i 40B
- versions base i instruída
- quantitzats fp8 i GPTQ



La família de models d'IA – Text Generatiu

Salamandra Technical Report

Language Technologies Unit
Barcelona Supercomputing Center

Abstract

This work introduces Salamandra, a suite of open-source decoder-only large language models available in three different sizes: 2, 7, and 40 billion parameters. The models were trained from scratch on highly multilingual data that comprises text in 35 European languages and code. Our carefully curated corpus is made exclusively from open-access data compiled from a wide variety of sources. Along with the base models, supplementary checkpoints that were fine-tuned on public-domain instruction data are also released for chat applications. Additionally, we also share our preliminary experiments on multimodality, which serve as proof-of-concept to showcase potential applications for the Salamandra family. Our extensive evaluations on multilingual benchmarks reveal that Salamandra has strong capabilities, achieving competitive performance when compared to similarly sized open-source models. We provide comprehensive evaluation results both on standard downstream tasks as well as key aspects related to bias and safety. With this technical report, we intend to promote open science by sharing all the details behind our design choices, data curation strategy and evaluation methodology. In addition to that, we deviate from the usual practice by making our training and evaluation scripts publicly accessible. We release all models under a permissive Apache 2.0

<https://arxiv.org/abs/2502.08489>



Cas d'ús – RAG amb EADOP

Retrieval Augmented Generation (o com rebre respostes directament dels documents)

El Portal Jurídic de Catalunya* conté les disposicions i decrets de la Generalitat, com el DOGC.

De cerques amb paraules claus, a llistes de document. Un nou paradigma.

Respostes concretes, amb referències, com si s'hagués demanat a un expert legal.

El sistema cerca els documents adients i escriu una resposta que qualsevol pot interpretar i entendre.

Un accés molt més fàcil a informació vital, però a vegades molt complexa.



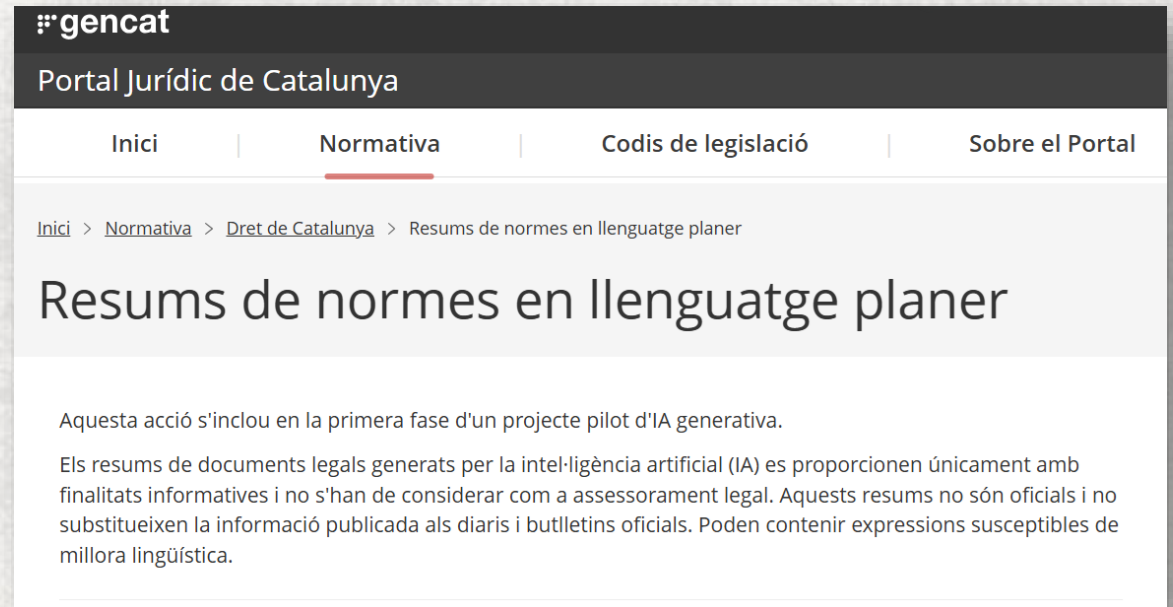
Generalitat de Catalunya
Entitat Autònoma del Diari Oficial
i de Publicacions



- En col·laboració amb EADOP (Entitat Autònoma del Diari Oficial i de Publicacions) i ALINIA

Cas d'ús – Llenguatge planer (OpenAI)

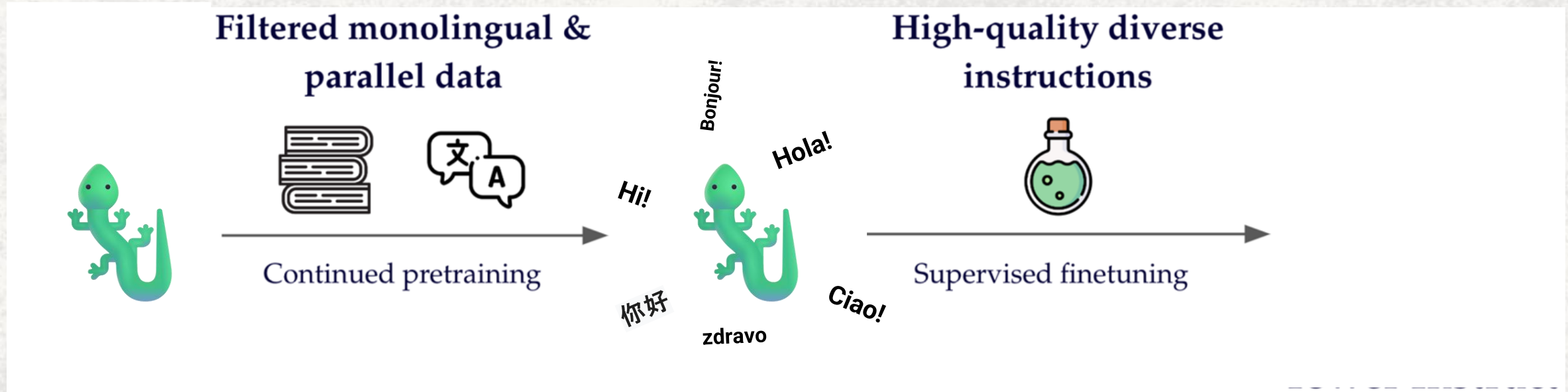
- Generació de resums normatius en llenguatge planer amb IA generativa al Portal Jurídic de Catalunya
 - **Núria Morales Pinto**, subdirectora general de publicacions. Entitat Autònoma del Diari Oficial i de Publicacions



Accenture, the global consulting and technology company, was the Publications Office's existing web development partner and, in such condition, they were contracted to implement the Generative AI solution as an extension of their current contract.

Microsoft, a global technology provider, supplied the underlying AI technology through their Azure OpenAI Service. The solution utilized OpenAI's large language models (LLMs), initially GPT-3 and later GPT-4, to generate the summaries.

La família de models d'IA – Traducció automàtica



Salamandra-2B

Salamandra-7B

SalamandraTA-2B-base

SalamandraTA-7B-base

SalamandraTA-2B-instruct

SalamandraTA-7B-instruct

La família de models d'IA – Traducció automàtica

Translates across 31 EU languages:

Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Norwegian Bokmål, Norwegian Nynorsk, Occitan, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Ukrainian, Welsh.

Optimized for + 7 Spanish regional languages:

Basque, Galician, Catalan, Valencian, Aranese, Aragonese, Asturian

Designed to handle Translation-Related Tasks:

Sentence-level-MT
Post-edition
Gender-aware
NER (Named-entity recognition)

Paragraph-level MT
Document-level MT
Context-MT
Terminology-aware-MT



Cas d'ús – Agència Catalana de Turisme



Generalitat de Catalunya
Agència Catalana
de Turisme





**“Amb les dades
facilitades per l’Agència
Catalana de Turisme
podem treballar nous
mercats.”**

Cristina Soler
Llopert

Explica'ns què necessites!
T'oferim serveis que t'ajudaran
a impulsar el teu negoci.

Hola! 

Servei d'atenció al sector
turístic català

Qui som ▾ Afiliació a programes ▾ Accions promoció ▾ Sigues partner ▾ Oficines a l'exterior ▾ Sala de premsa i actualitat ▾

La família de models d'IA – Veu



Dos models de **reconeixement de veu** basats en Whisper adaptats amb ~700h de català de 3cat)

- Whisper
- Whisper-fast, optimitzat per a inferència, x4 més ràpid,



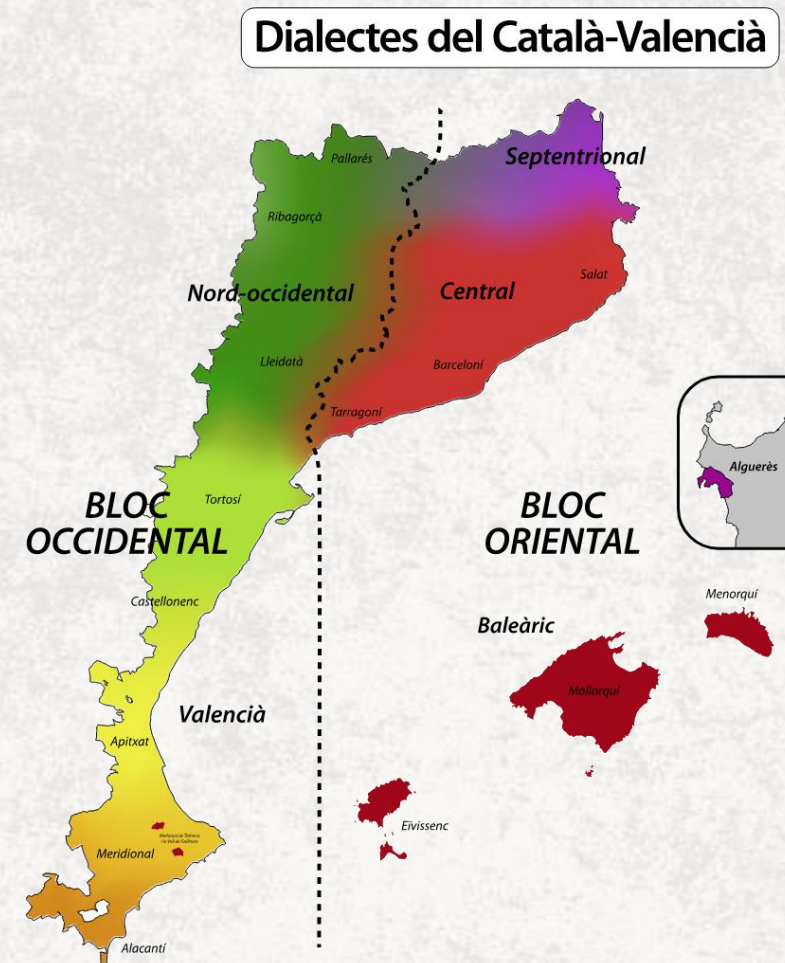
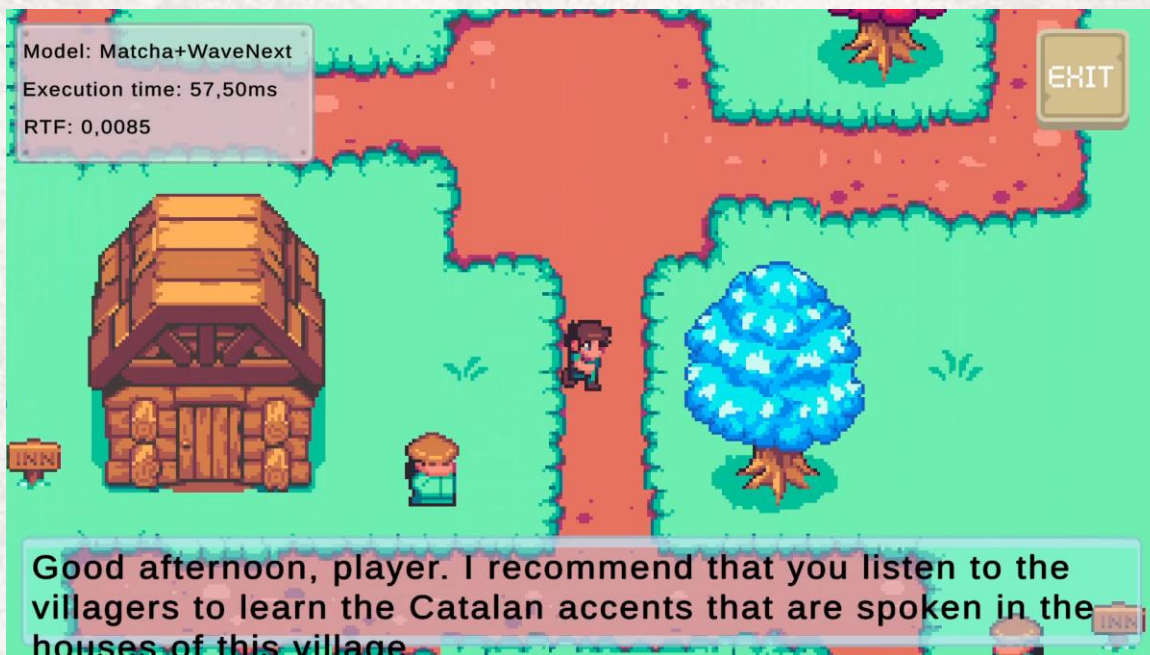
Model de **síntesi de veu**, **Matxa**, multilocutor i multidialectal

Actualment 4 variants: Central, Valencià, Occidental i Balear

pendent de nova versió millorant variants central i valencià i n'afegeix dos més (septentrional i alguerès)

La família de models d'IA – [Veu](#)

El Projecte Aina presenta el primer model de síntesi de veu que inclou totes les variants del català



Transcripció automàtica de trucades

Primers reptes del 112

- multilinguïtat: ca, es i ca+es (*code-switching*)
- qualitat de trucada telefònica
- soroll de fons
- diferents condicions de veu
- audios curts i llargs (de 30s a 10m)



Solucions per al 112

- comparativa dels diversos models
- creació de dades sintètiques per adaptar els models bàsics
 - audios amb qualitat de trucada
 - audios amb *code-switching* (ca+es)
 - generació de puntuació
 - tractament del “112 Digui'm”

La familia de models d'IA – Visual (en desenvolupament)

Prompt: Describe la imagen con el mayor detalle posible

Respuesta: La imagen es una pintura que representa a un grupo de perros sentados alrededor de una mesa de juego, jugando al póker. Hay un total de seis perros, cada uno con una expresión diferente, lo que añade un elemento de personalidad y carácter a la escena. La mesa de juego está cubierta con cartas y fichas de póker, lo que indica que los perros están participando en una partida de póker. Hay cuatro sillas alrededor de la mesa, con dos perros sentados en cada una. Además, hay dos botellas y dos tazas colocadas en la mesa, lo que sugiere que los perros también están disfrutando de bebidas durante su sesión de juego. La pintura captura la atmósfera divertida y juguetona de los perros mientras participan en esta actividad social.



La família de models d'IA – Visual (en desenvolupament)



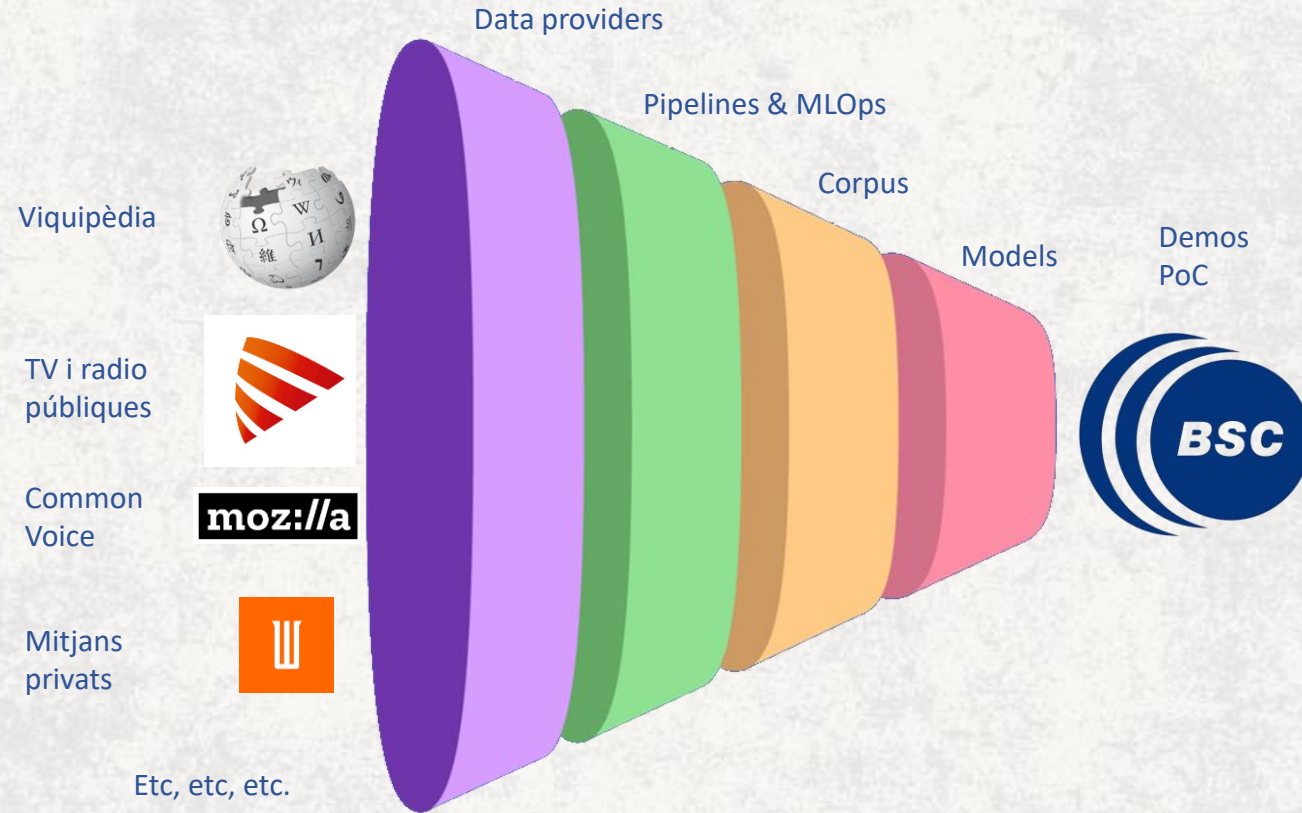
Prompt: Quin plat de les imatges és millor que em mengi si sóc intolerant a la lactosa? Justifica la resposta



Resposta: El plat de la primera imatge, que és una pizza, no és adequat per a una persona intolerant a la lactosa, ja que conté lactis com la mozzarella. El plat de la segona imatge, que és un plat de pollastre amb pebrots, és una bona opció per a una persona intolerant a la lactosa, ja que no conté làctics.

De la recerca al mercat

Del laboratori al mercat – 1^a etapa creació del models



Del laboratori al mercat – 2^a etapa ús dels models



Mecanismes de relació amb l'ecosistema de IA



CIDAI Centre of Innovation
for Data tech
and Artificial Intelligence



Mecanismes promoció i finançament de casos d'ús



Aina Hack

The logo for Aina Hack features the word "Aina" in a bold, sans-serif font, with the "Ai" in red and "na" in black. To the right of "Aina" are four vertical red bars of increasing height. Below "Aina" is the word "Hack" in a smaller, black, sans-serif font.

**#AINA
CHALLENGE**

The logo for #AINA CHALLENGE consists of the text "#AINA" and "CHALLENGE" in a bold, black, sans-serif font, stacked vertically. The text is set against a solid red rectangular background.

OTS OFICINA TÈCNICA
DE SUBVENCIONS [FONS DE TRANSICIÓ NUCLEAR]

The logo for OTS OFICINA TÈCNICA DE SUBVENCIONS [FONS DE TRANSICIÓ NUCLEAR] features the acronym "OTS" in a bold, dark blue, sans-serif font. To its right is the text "OFICINA TÈCNICA DE SUBVENCIONS" in a smaller, dark blue, sans-serif font. Further right is a large orange square bracket containing the text "FONS DE TRANSICIÓ NUCLEAR" in a dark blue, sans-serif font.

Mecanismes promoció i finançament de casos d'ús



Mecanismes promoció i finançament de casos d'ús



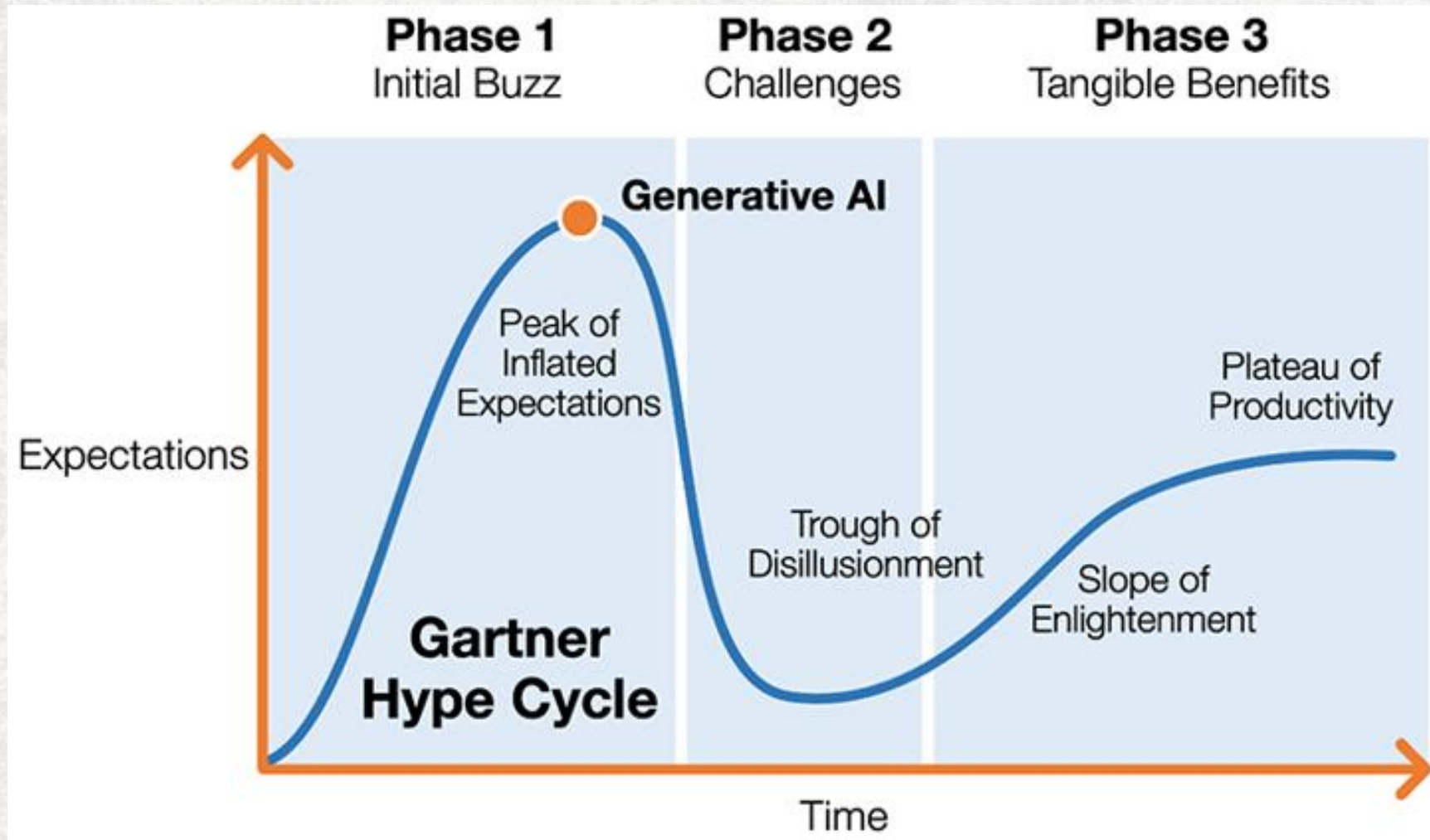
5 reflexions al voltant de la IA

1- El valor de la IA per a les necessitats de l'Administració

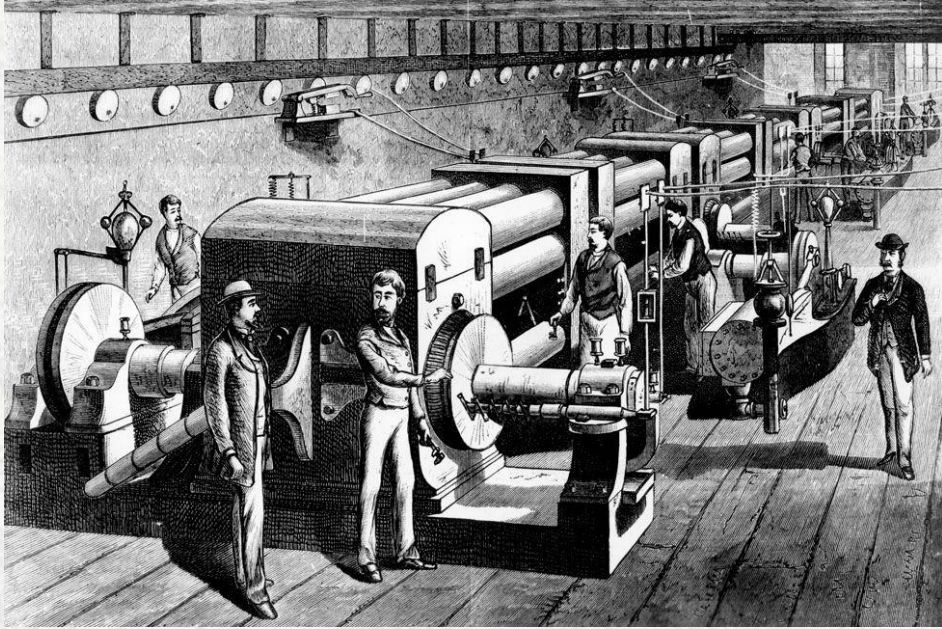
EMBRACING RESPONSIBLE USE

Responsible use of generative AI could improve how we work in government, allowing us to better serve residents, deepen democracy solve public problems. The tools are also in their infancy and will continue to evolve.

2- Gartner Hype Cycle de les technologies



3 - Inicis IA. Transformació de processos i organitzacions



Vapor



Electricitat

4 - Alerta – “sedentarisme cognitiu”

Copilot



Dormir al
volant?



5 - Fem-ho bé, que estem construint la societat digital

16 LA VANGUARDIA

OPINIÓ

DIVENDRES, 15 SETEMBRE 2023

Tot va molt de pressa, però no serà ràpid

Genís Roca



La revolució industrial va ser un procés de transformació tecnològica, econòmica i cultural que va modificar l'estructura de la societat, passant d'una economia rural basada en l'agricultura i el comerç a una economia

viar l'estructura d'una societat sempre és una tasca que afecta més d'una generació i, com sempre, se'ns fa difícil anticipar quin serà el model que en resultarà.

Més de dos-cents anys de paradigma industrial ens han acostumat a treballar amb mètodes i procediments orientats a l'eficiència i la qualitat. Hem anat consolidant processos i rutines, mètriques i indicadors, sistemes de control i avaluació. La industrialització és això, abraçar la mecanització i el procediment abandonant l'artesania i la improvisació. Hi ha un procediment per fer un tràmit administratiu, sigui quina sigui la finestra que t'atengui. Tot el professorat de matemàtiques d'un mateix institut apli-

clau per a la competitivitat empresarial.

Però toca fer canvis. Som fills de la societat industrial i pares de la digital, i ens toca anar acompanyant el final d'una i la progressiva posada en marxa de l'altra. Vigilant el ritme, perquè la industrial ja no serveix per a tot, però la digital encara no serveix per a tot. Malgrat la vertiginosa velocitat del moment, aquest és un procés de canvi estructural que ens ocuparà dècades.

És a dir, tant se val quina sigui la nostra edat actual, podem afirmar que viurem tot el que ens quedi de vida professional en un context de canvi no només permanent sinó també estructural. Tant és si la nostra feina és de salut, docència,



Moltes gràcies

#OpenGovWeek



**Diputació
Barcelona**

Àrea de Bon Govern, Assistència
Local i Cohesió Territorial

