Aina

Impulsant l'ús del català en l'era digital

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

El projecte Aina ha estat impulsat i finançat per la Generalitat de Catalunya

Aina

Challenge 2024

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Generalitat de Catalunya

Ens acompanyen

Albert Cañigueral (BSC)
Transferència tecnològica i ecosistema projecte Aina

Marta Villegas (BSC)
Cap Unitat de Tecnologies del Llenguatge (LangTech)

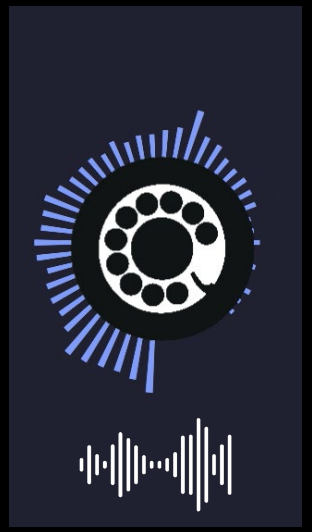
Què volem aconseguir amb aquesta sessió?

1) Compartir els detalls operatius i tècnics de l'Aina Challenge 2024

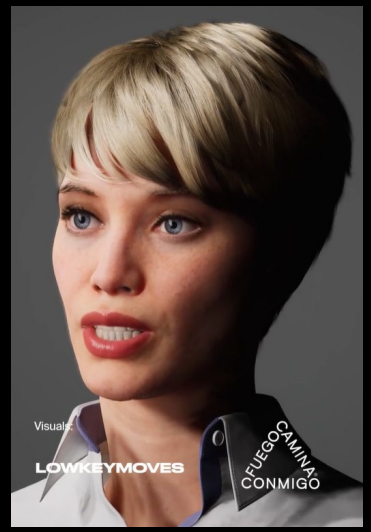
2) Identificar startups i empreses interessades a participar a l'Aina Challenge 2024. Registre d'interès.

Intel·ligència artificial i tecnologies del llenguatge

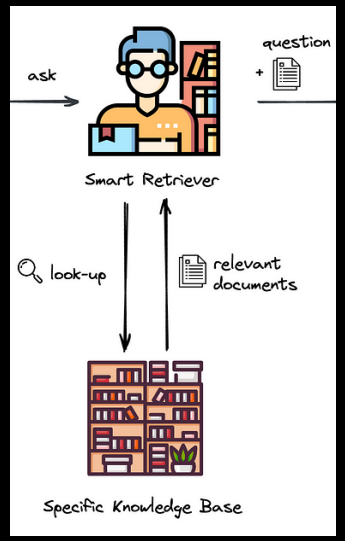
IA i tecnologies del llenguatge (text, veu, traducció automàtica)



Atenció client avançada

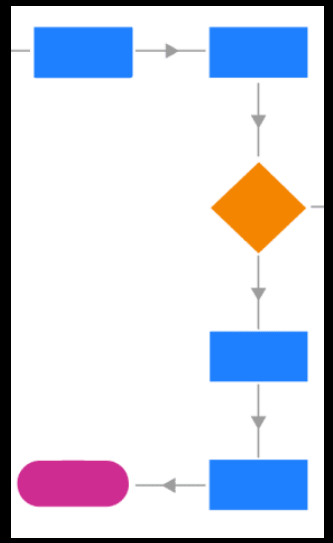


Humans virtuals



Retrieval Augmented Generation

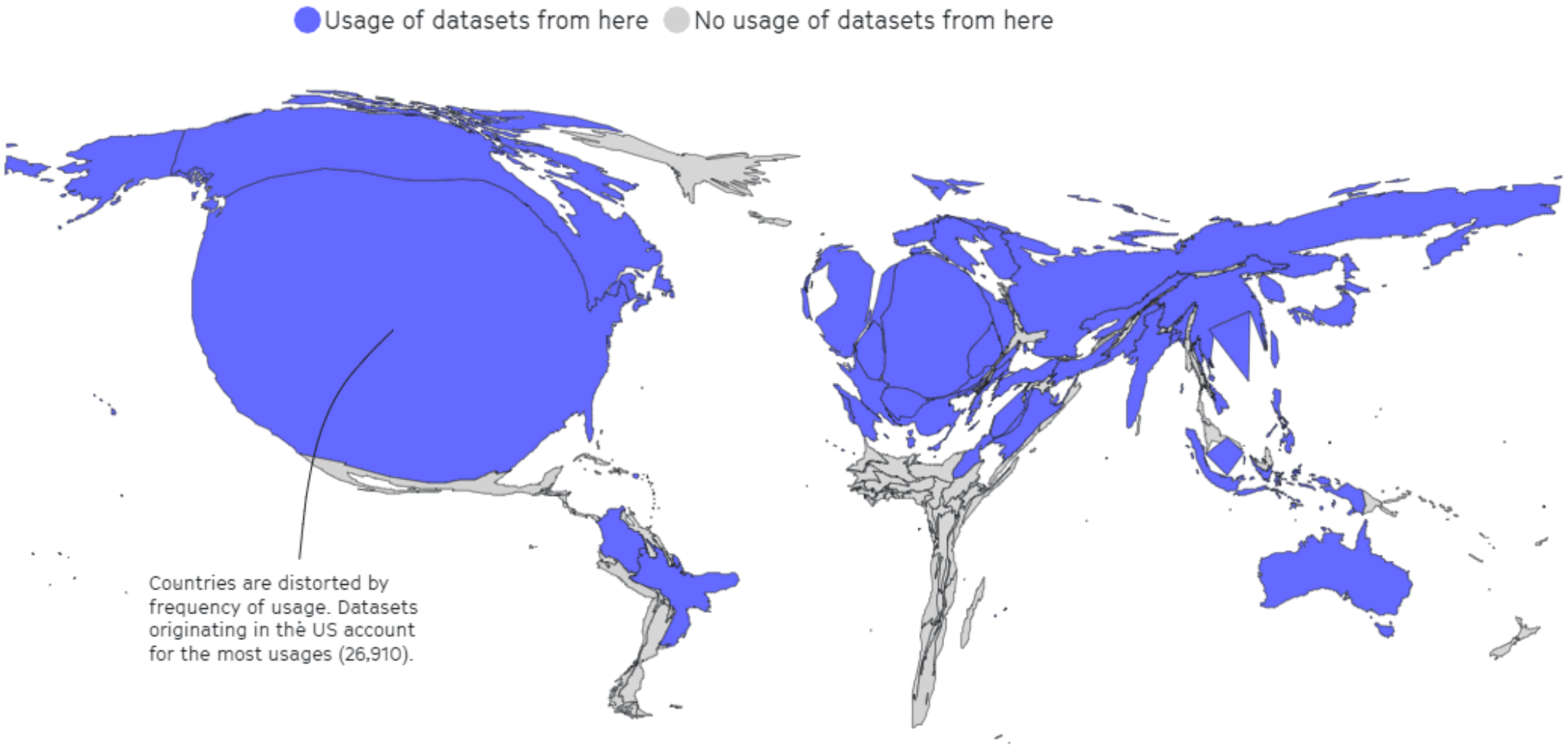
Transcripció i resum de converses



Nous processos de treball

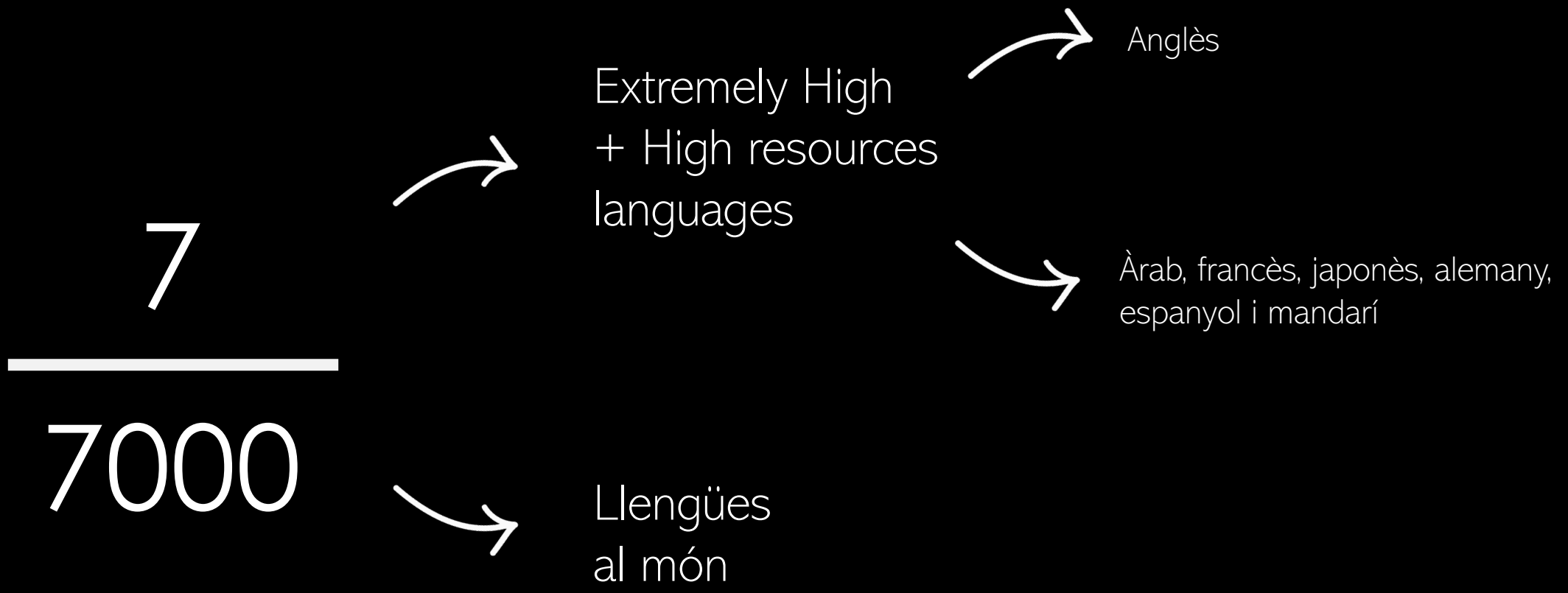
La bretxa idiomàtica en la intel·ligència artificial

Frequency of dataset usage by country



This map shows how often 1,933 datasets were used (43,140 times) for performance benchmarking across 26,535 different research papers from 2015 to 2020.

Disponibilitat de recursos digitals en les diverses llengües



Resultats de la bretxa idiomàtica en IA/TL

1. Rendiment de les aplicacions d'IA. Experiència d'usuàri/a pobre.
2. Reptes en l'alineació dels models. Ciberseguretat.
3. Biaixos culturals. Manca de referents locals.
4. Exclusió social en l'accés als serveis digitals.

Fent front a la bretxa idiomàtica en la IA en català

→ Aina

Objectius del projecte Aina

1) Promoure i garantir la presència de la llengua catalana a la societat digital

2) Ajudar al desenvolupament de startups i pimes al voltant de productes d'IA/TL

3) Garantir la disponibilitat d'eines d'IA/TL per al sector públic a Catalunya

Duració: 2022 - 2026

Pressupost: 15 Milions €

Generalitat de Catalunya

Execució: Unitat LangTech

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Objectius del projecte Aina

- El projecte Aina dota la llengua catalana de la infraestructura oberta necessària (corpus de dades, models d'IA preparats, etc.) per al desenvolupament d'aplicacions basades en IA i tecnologies del llenguatge (AI/LT)

- Es facilita el seguiment del IA Act.

- El projecte Aina ha de permetre al català fer un salt qualitatiu i quantitatiu a l'ecosistema digital.

Duració: 2022 - 2026

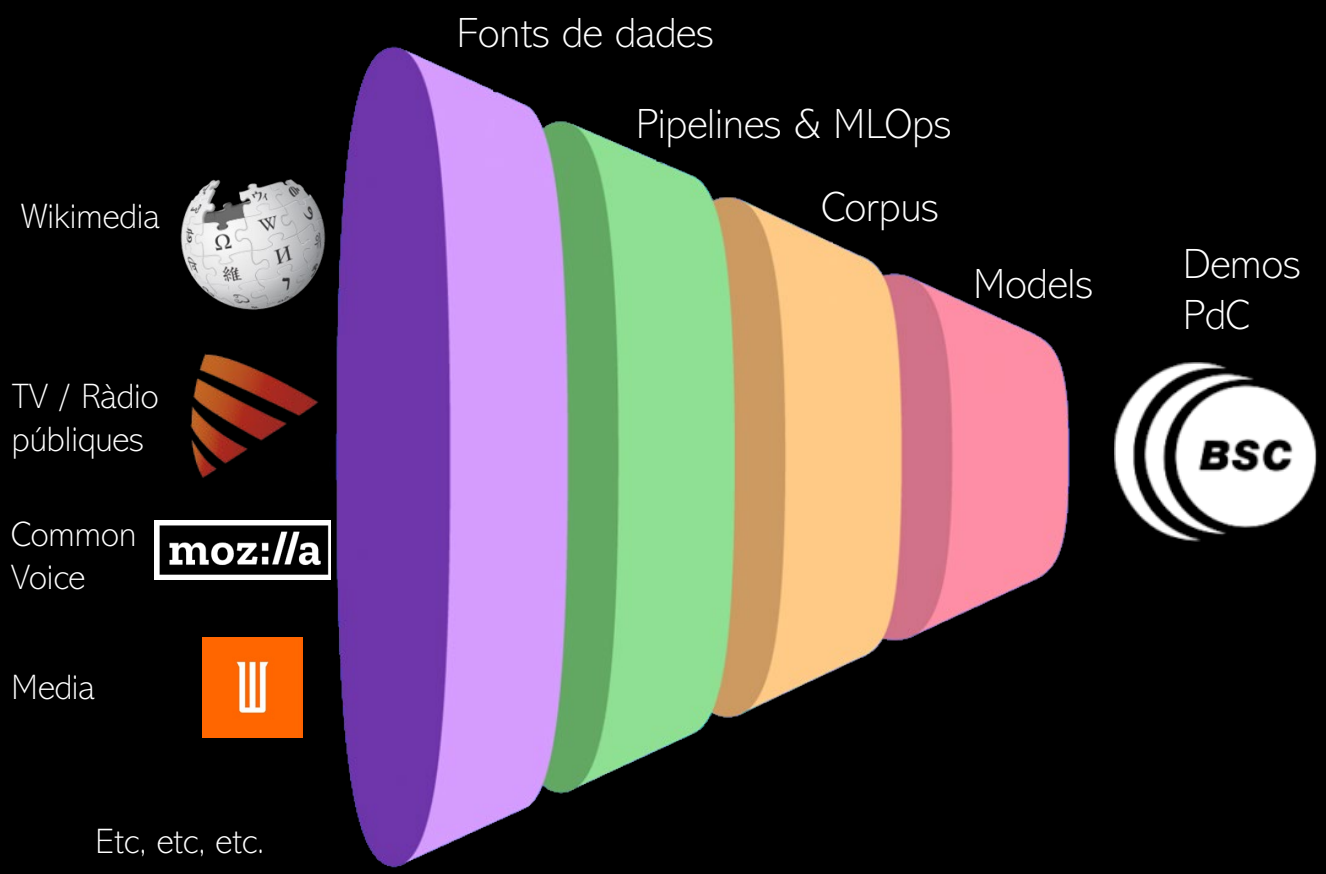
Pressupost: 15 Milions €

Generalitat de Catalunya

Execució: Unitat LangTech

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

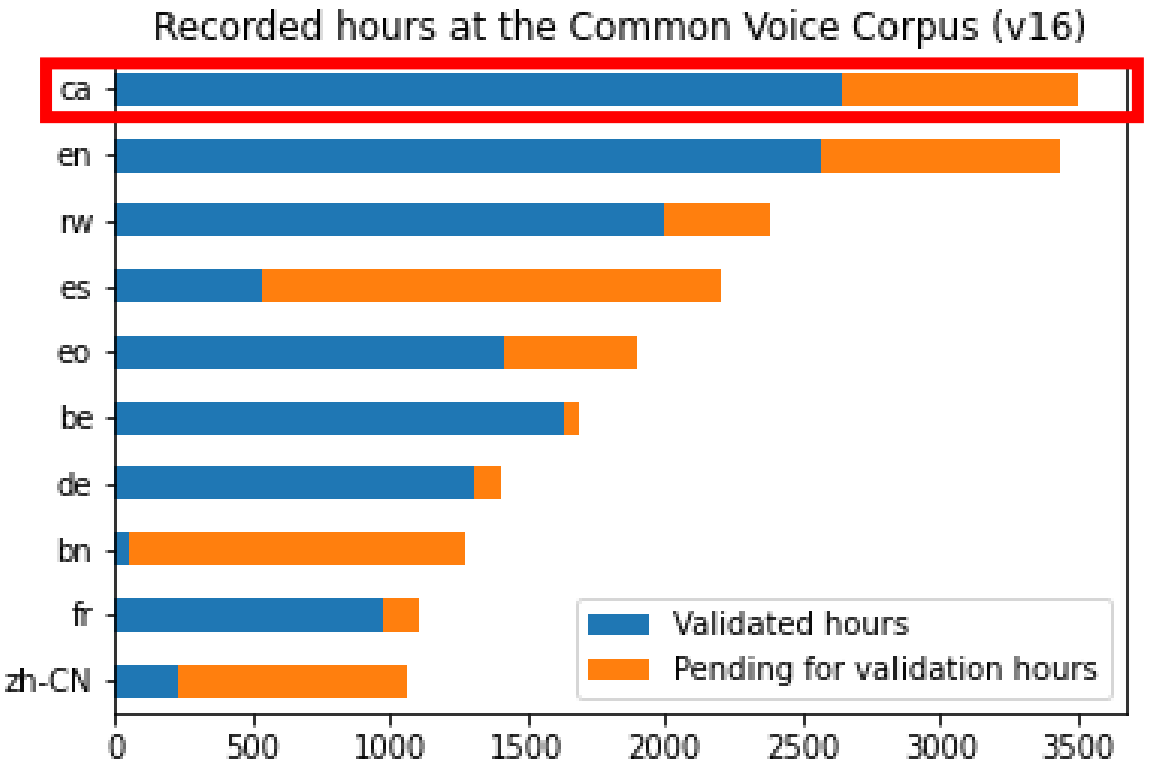
1a fase projecte Aina: Construcció de la infraestructura digital oberta

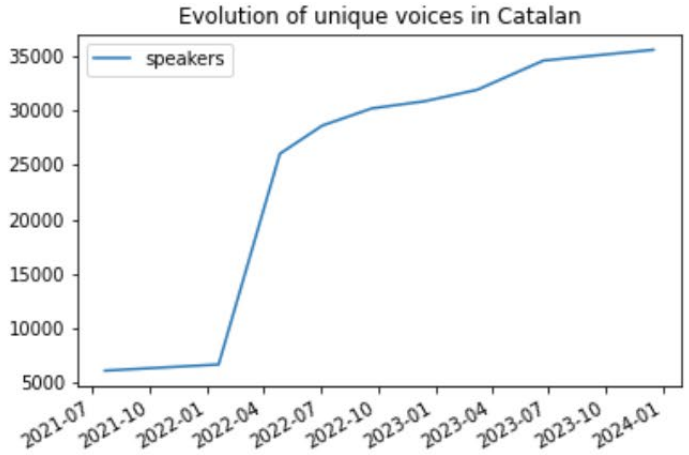


Català: llengua #1 amb hores validades a Common Voice (Mozilla Found.)

Common Voice

mozilla

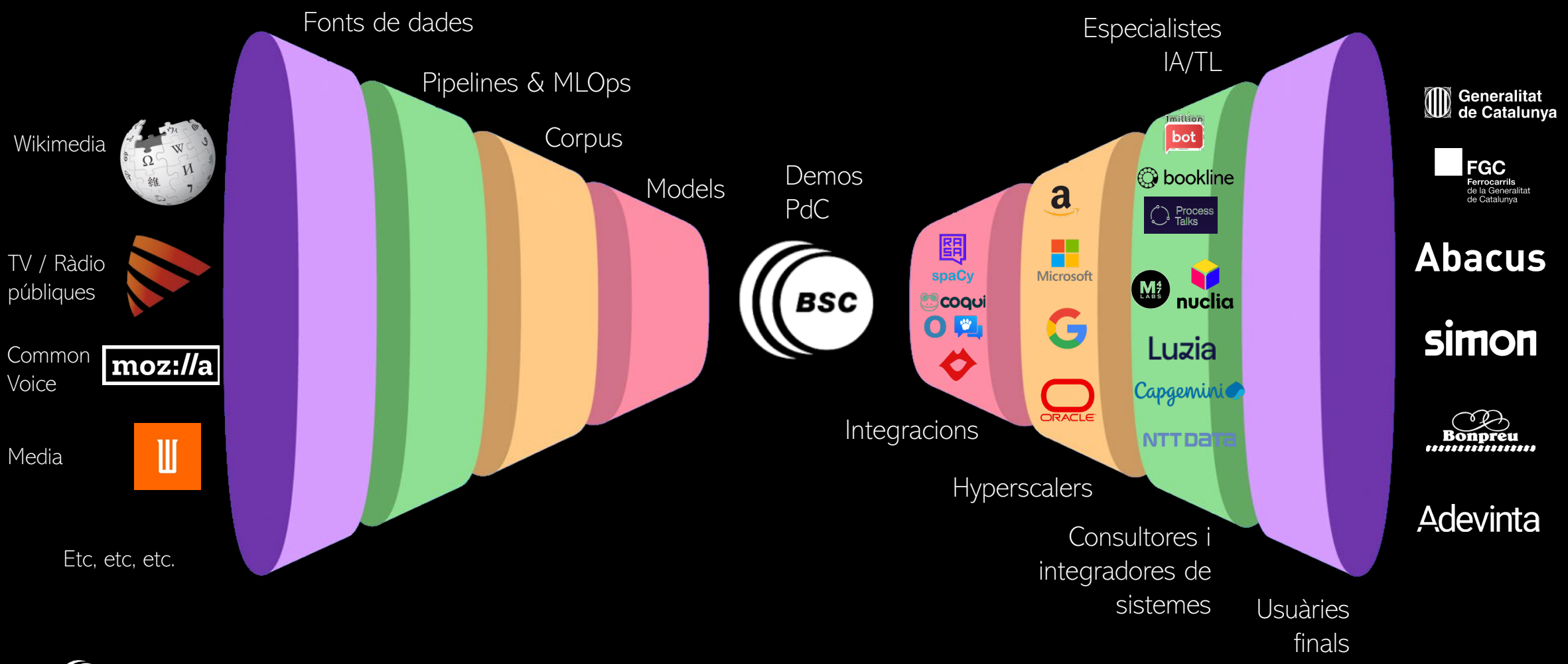




SOFTCATALÀ

Català: incorporació a models de tercers (AudioPalm de Google)

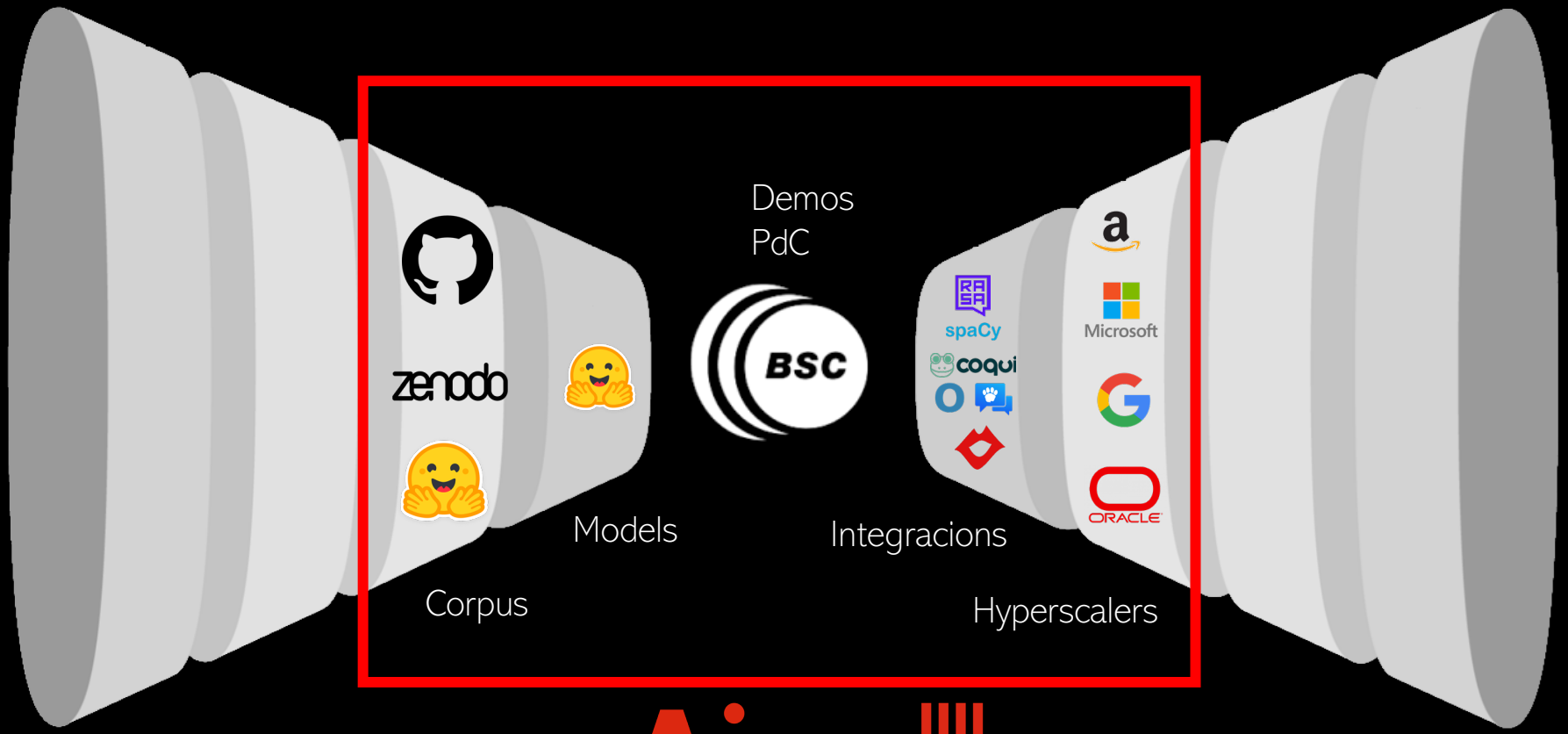
2a fase projecte Aina: transferència tecnològica i ecosistema local IA/TL



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Generalitat de Catalunya

2a fase projecte Aina: transferència tecnològica i ecosistema local IA/TL



Aina
Kit

Aina

Challenge 2024

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Generalitat de Catalunya

Aina Challenge 2024

Aina Challenge

- ✓ 1M €
- ✓ 22 premiats (fins a 50.000€)
- ✓ Hyperscalers

doitnovv

- ✓ 3 reptes

PENINSULA CORPORATE INNOVATION

https://ainachallenge.cat/

Aina Challenge 2024 – 3 reptes

Repte #1: Desenvolupar **serveis i/o aplicacions** d'intel·ligència artificial i tecnologies del llenguatge en català.	16 premis 50.000€
Repte #2: Desenvolupar **eines de monitoratge, control i alineació** en l'ús de models i aplicacions IA/TL que incorporen el català.	3 premis 33.000€
Repte #3: Contribuir a construir l'ecosistema de **recursos oberts** del projecte Aina per ajudar a escalar, adaptar i fer més robusta la intel·ligència artificial i les tecnologies del llenguatge en català.	3 premis 33.000€

Aina Challenge 2024 – Calendari aproximat

Fases del concurs

Març de 2024

1. Inscripció

Les propostes s'hauran de presentar a través del formulari d'aquesta web.

Abril de 2024

2. Selecció

El Comitè de Selecció analitzarà totes les candidatures rebudes i seleccionarà les 22 organitzacions guanyadores.

Maig 2024 - Octubre 2024

3. Acceleració

Les organitzacions seleccionades tindran 6 mesos per desenvolupar i validar la seva solució amb el BSC.

Octubre - Novembre 2024

4. Demo Day

Presentació dels resultats de l'acceleració.

Infraestructura oberta d'Aina (http://bit.ly/AINAKit)

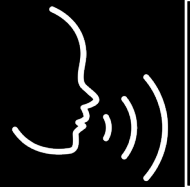
Aina Kit

Aina Kit – Models pre-entrenats

Aina Kit

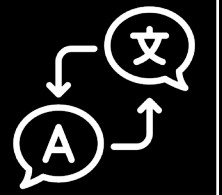
Generatiu CAT + ES + EN (Flor 6.3B, 1.3B)
- A la cua: GPT3 175B

Divesos models per a tasques específiques (RoBERTa)
- Q & A
- Text summarization
- Sentiment analysis
- Entities Identification
- Etc.



Síntesi de veu
- Múltiples veus
- Dialectes

Reconeixement veu
- Múltiples tamanys



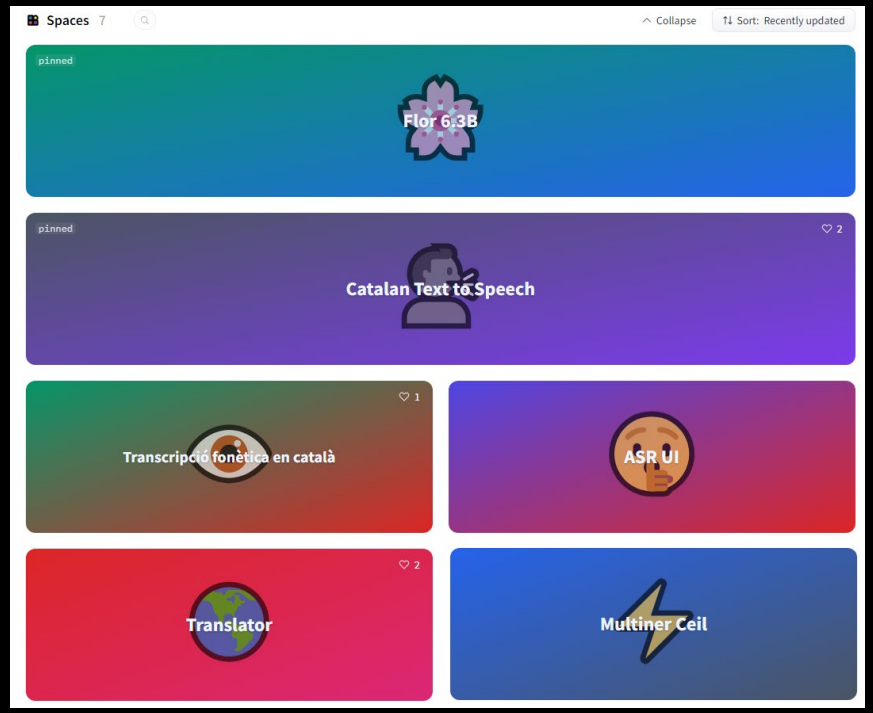
CAT ←→ LLeng
- Castellà + Admin
- Anglès
- Francès
- Alemany
- Xinès
- Italià
- Português
- Gallec
- Basc

https://huggingface.co/projecte-aina

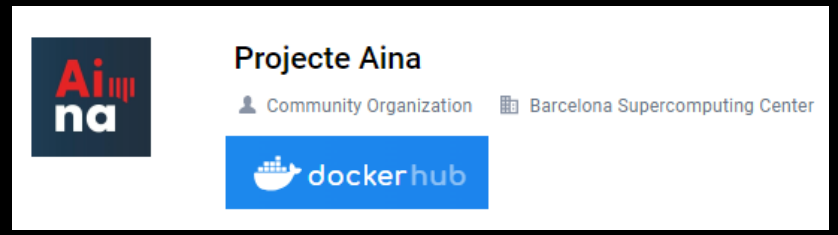
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Generalitat de Catalunya

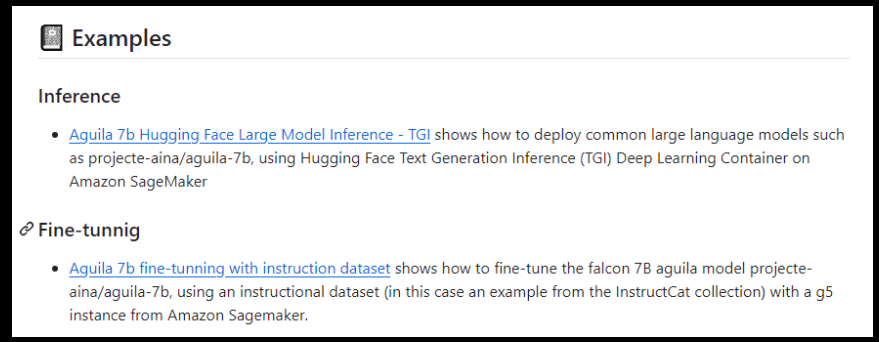
Aina Kit – Models pre-entrenats – demos, dockers, notebooks, etc.



https://huggingface.co/projecte-aina

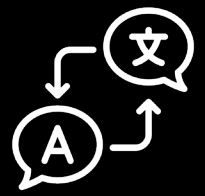


https://hub.docker.com/u/projecteaina



https://github.com/projecte-aina/amazon-sagemaker-examples

Aina Kit – Dades d'entrenament + ajust a tasca + avaluació models



Les dades estan agrupades en 3 grans blocs (text, veu i traducció automàtica) i diversos sub-blocs:

- Dades i eines per a models de text
  - 1) Corpus textual massiu
  - 2) Dades anotades per fine tuning i/o avaluació de models de text
  - 3) Dades per instrucció de models de text
  - 4) Dades per avaluació de models de text
  - 5) Eines per al subministrament de dades de text
- Dades i eines per a models de veu
  - 6) Corpus de veu
  - 7) Eines per al subministrament de dades de veu
- Dades per a la traducció automàtica
  - 8) Corpus paral·lels per entrenament de models de traducció automàtica
  - 9) Corpus paral·lels per adaptació i avaluació de models de traducció automàtica

Aina Kit – Dades d'entrenament + ajust a tasca + avaluació models

Dades i eines per a models de text

1) Corpus textual massiu

CATALOG 1.0: és el dataset de preentrenament de LLMs més gran en català alliberat fins ara. Conté una gran varietat de fonts, amb un percentatge important de textos curats manualment, cosa que el diferencia enormement dels altres grans datasets publicats, que estan constituïts únicament per dades d'origen web.

The following pre-existing datasets

- OSCAR-2301
- OSCAR-2201
- CaText
- MaCoCu-ca 1.0
- caWaC
- Colossal OSCAR 1.0
- mC4

Media Groups

- IB3
- Grup El Món
- Vilaweb
- Nació Digital
- ACN
- Racó Català Articles
- Racó Català Fòrums (anonymized version)
- Aquí Berguedà

Academic & Book Repositories

- Tesis Doctorals en Xarxa (TDX)
- Wikipedia
- Project Gutenberg

Government Institutions

- Parlament de Catalunya
- Les Corts Valencianes
- Diari Oficial de la Generalitat Valenciana
- Butlletí Oficial de la Universitat d'Alacant

CATALOG 1.0 està format per 17.450.496.729 de paraules (al voltant de 23 mil milions de tokens) distribuïts en 34.816.765 documents.

CATALOG 1.0

3) Dades per instrucció de models de text

- InstruCAT: més de 235 mil instruccions per entrenament de LLMs en tasques "downstream" https://huggingface.co/datasets/projecte-aina/InstruCAT
  - Generades a partir dels següents **datasets:** caBreu, CatalanQA, CoQCat, GuiaCat, IntoxiCat, **Parafraseja**, PAWS-ca, sts-ca, WikiCat, CEIL, TECA, NLUCat.
  - En **categories** com: paràfrasi (34695), toxicitat (29809), pregunta-resposta (QA) (27424)classificació (12391), resum (5998), anàlisis de sentiments (5750)

InstruCAT

Aina Challenge 2024

https://ainachallenge.cat/

Aina Kit

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Generalitat de Catalunya