L'horitzó d'Aina: estatus del projecte

Actualització sobre el desenvolupament dels models lingüístics, recursos i objectius d'AINA

19 de desembre del 2023

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Finançat per Generalitat de Catalunya

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Aina

Recordem

OBJECTIUS

1

Proveir el català de la **infraestructura** necessària **per al desenvolupament d'aplicacions basades en IA/TL,** (assistents de veu, traductors automàtics, agents conversacionals, etc)

2

Fer que **la inclusió del català** a les aplicacions de IA/TL sigui **rendible i atractiva per a les empreses del sector**, tant a nivell local com global.

3

Aconseguir que el ciutadà de Catalunya pugui **participar en català** en el món digital **al mateix nivell que un parlant d'una llengua global**, com ara l'anglès o el castellà.

DEFINICIÓ i FONAMENTS

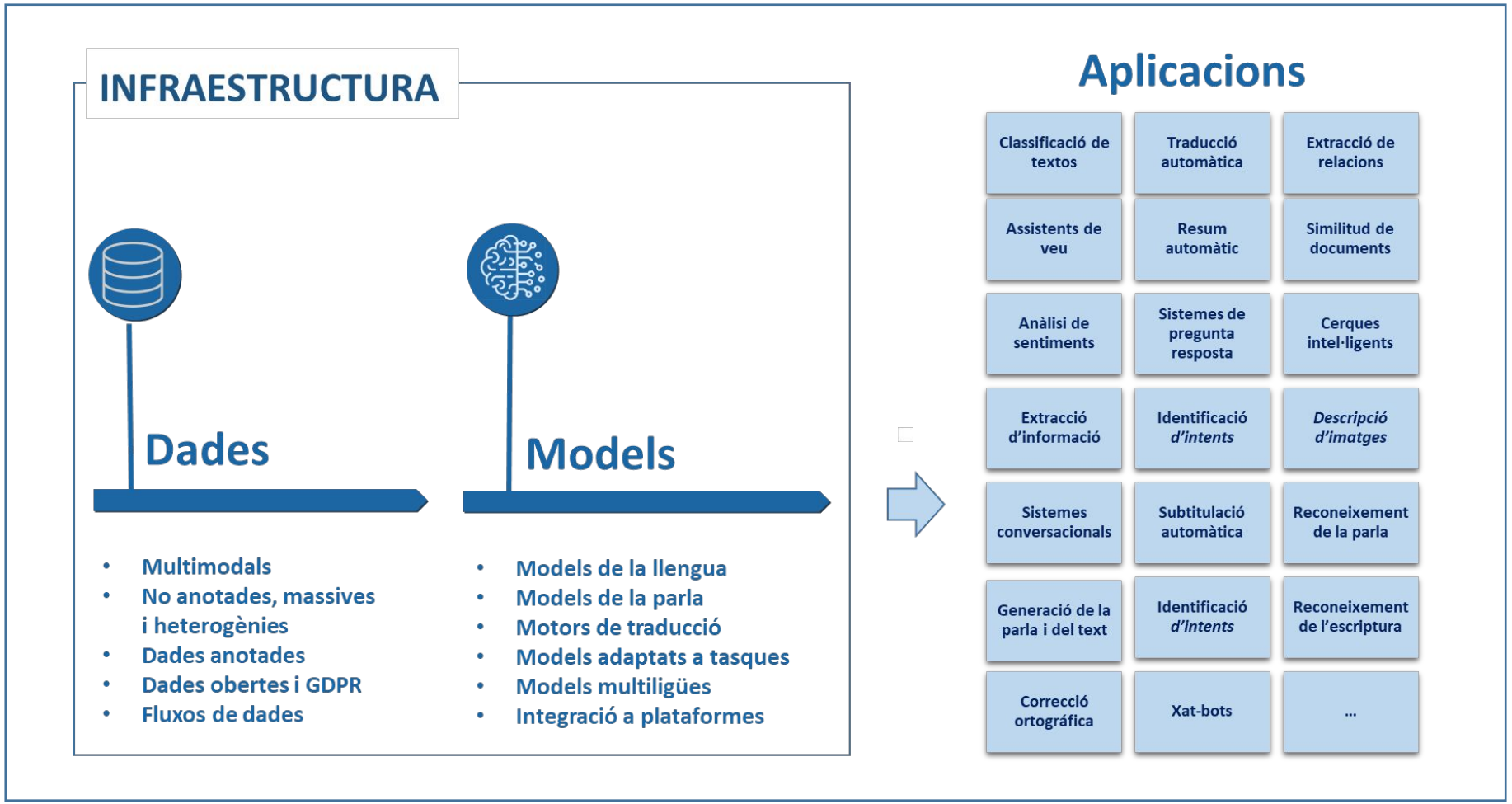
- AINA és essencialment **infraestructura** lingüística.
- El valor de les **dades**.
  - La tecnologia avança molt ràpidament però **les dades són persistents**.
  - Únicament des de la **iniciativa pública**, el català es pot garantir el subministrament de dades suficients.
  - Disposar de **dades de qualitat suficients és un actiu segur i de futur** que garanteix l'actualització de la tecnologia.

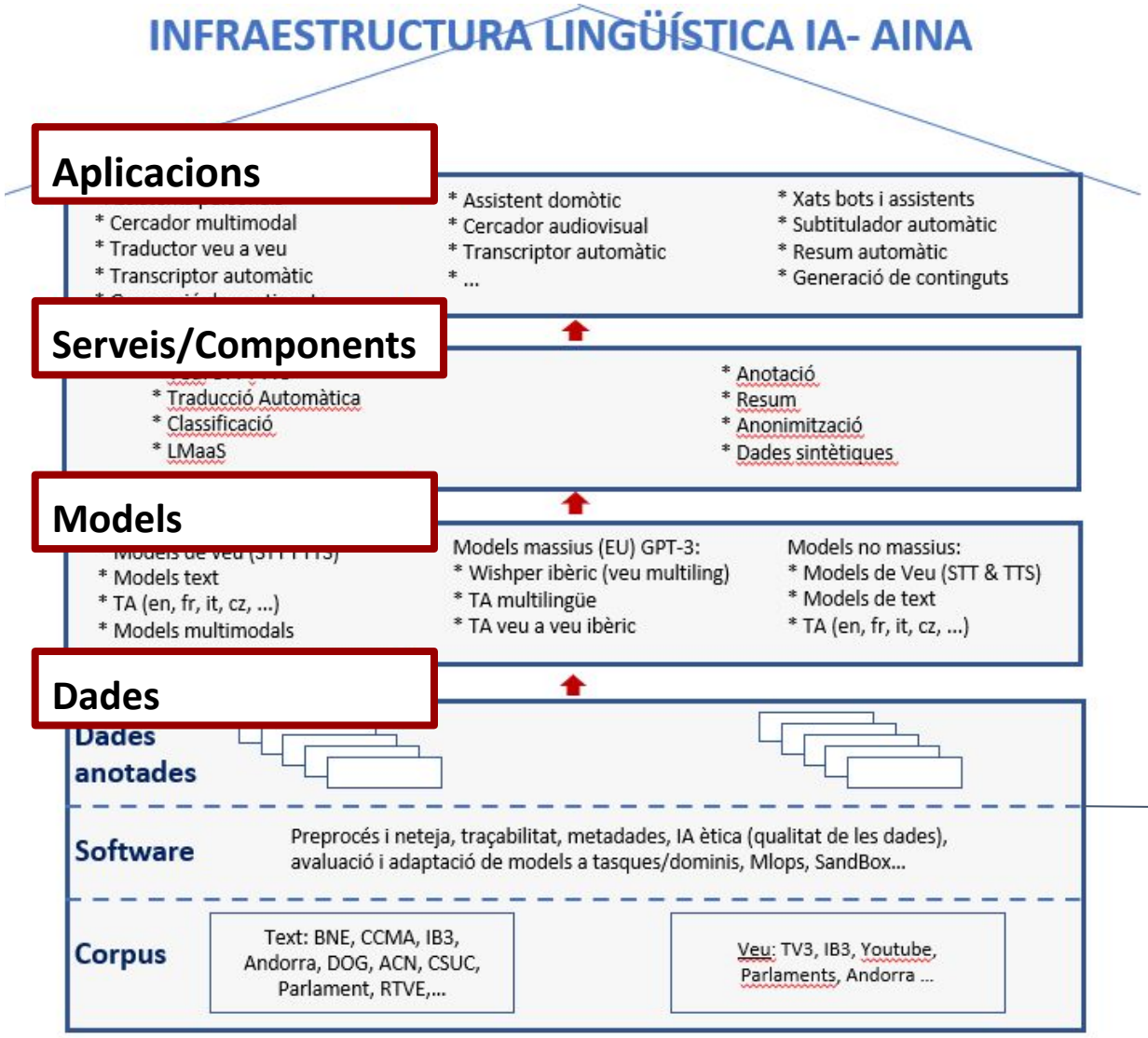
ESTRATÈGIA

- AINA implementarà una infraestructura de recollida i neteja de dades amb la **implicació de grans actors.**
- AINA reaccionarà ràpidament als **avenços tecnològics** mitjançant la vigilància tecnològica.
- AINA detectarà i donarà resposta a **noves necessitats** de les empreses i de la societat mitjançant la vigilància sectorial i de mercat.

Barcelona, 15 de febrer del 2022

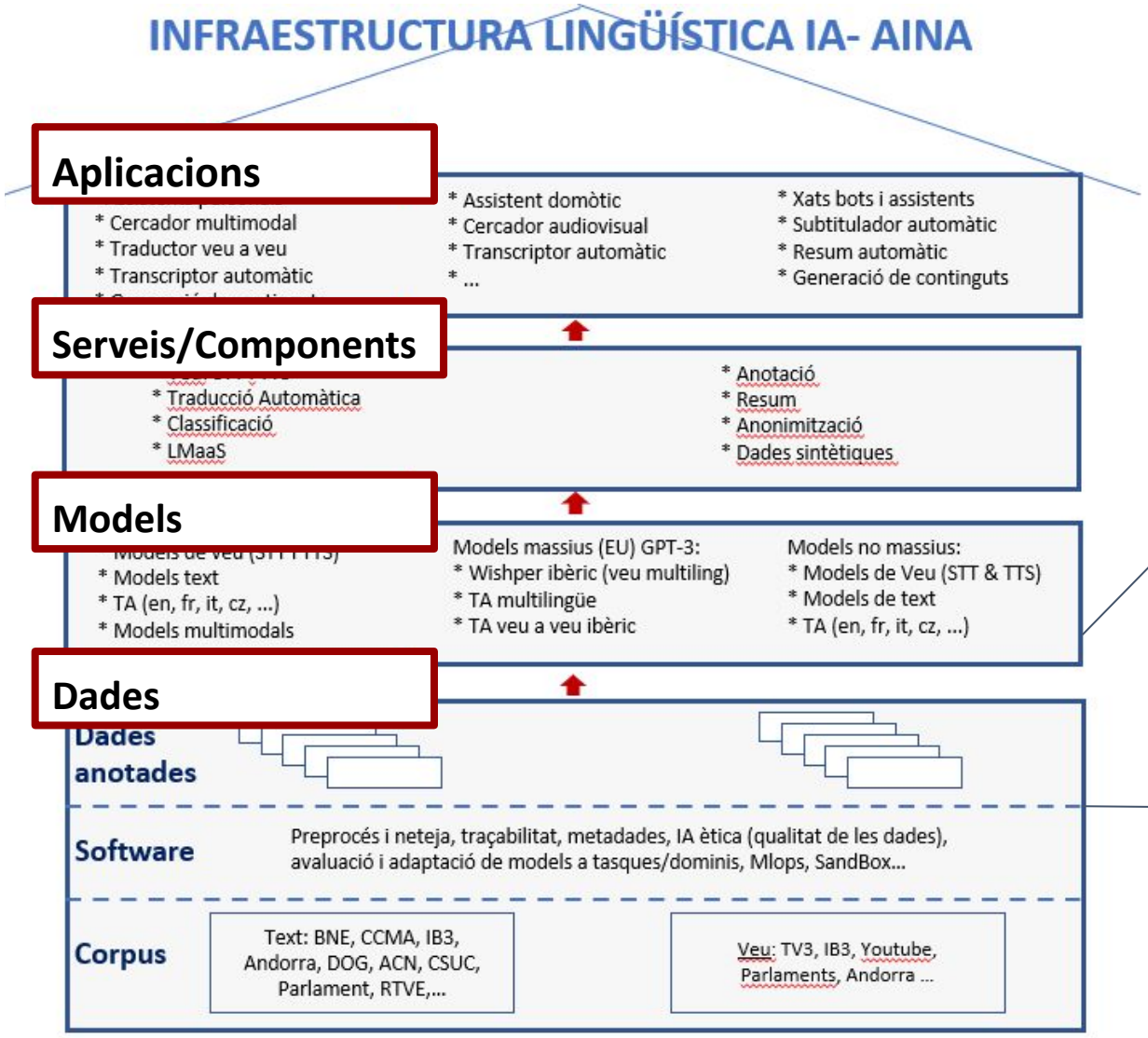
Recordem


INFRAESTRUCTURA
Dades
Multimodals
No anotades, massives i heterogènies
Dades anotades
Dades obertes i GDPR
Fluxos de dades
Models
Models de la llengua
Models de la parla
Motors de traducció
Models adaptats a tasques
Models multilingües
Integració a plataformes
Aplicacions
Classificació de textos
Traducció automàtica
Extracció de relacions
Assistents de veu
Resum automàtic
Similitud de documents
Anàlisi de sentiments
Sistemes de pregunta resposta
Cerques intel·ligents
Extracció d'informació
Identificació d'intents
Descripció d'imatges
Sistemes conversacionals
Subtitulació automàtica
Reconeixement de la parla
Generació de la parla i del text
Identificació d'intents
Reconeixement de l'escriptura
Correcció ortográfica
Xat-bots
...



Provisió i processament de les **dades** lingüístiques

- Corpus textual
- Generació de dades anotades
  - Generació d'instruccions
- Corpus de veu
  - Campanya de recollida de veu (CM)
  - Corpus de veu amb transcripció
  - Corpus d'altres tasques de veu
- Corpus paral·lels per TA
- Estratègia de provisió de dades



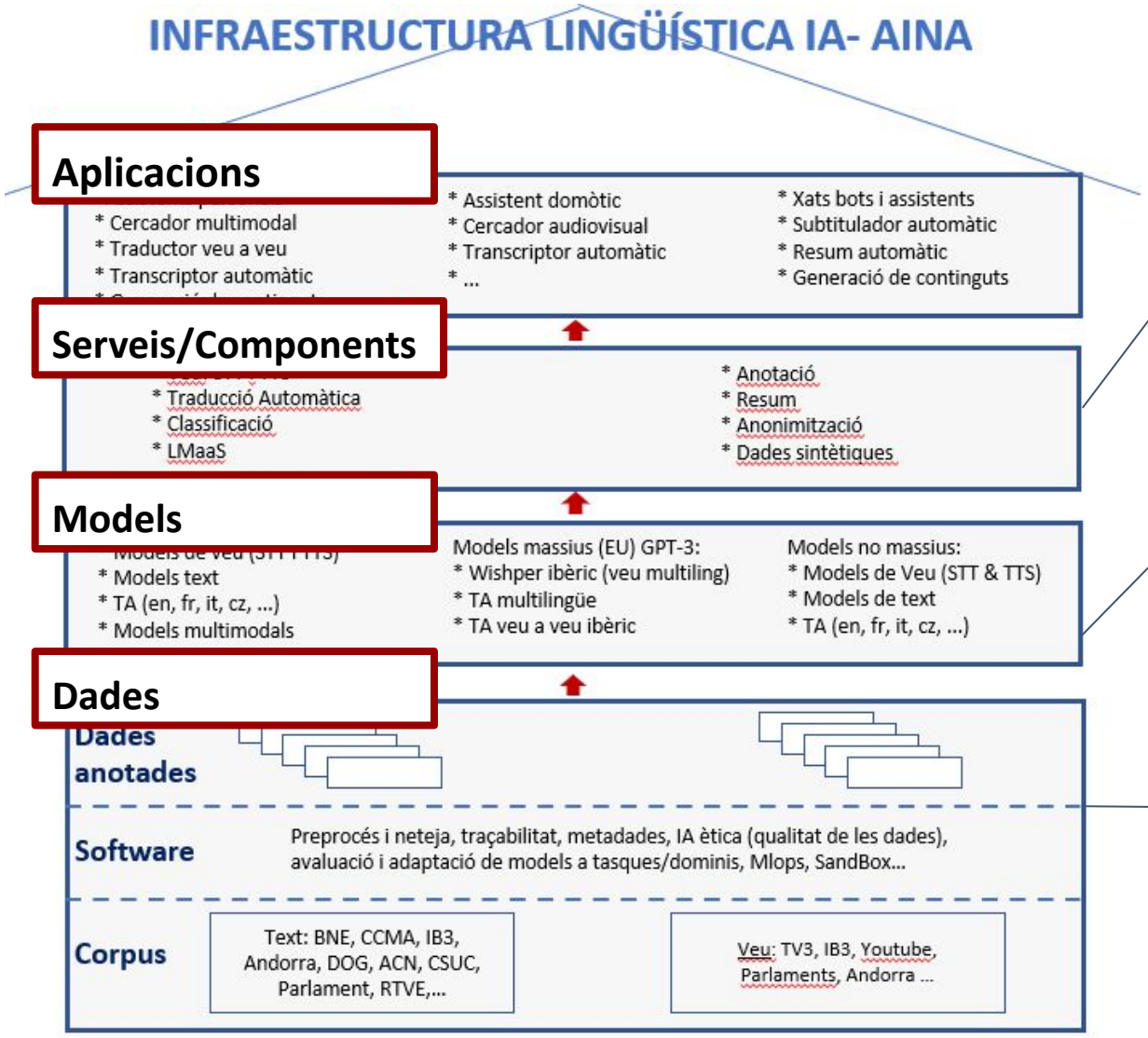
Models

- Generació de **models** de llengua
- Tecnologies de la **parla**
- Tecnologies de la **traducció**

Eines **d'avaluació** i **benchmarking**
La **IA ètica**

Provisió i processament de les **dades** lingüístiques

- Corpus textual
- Generació de dades anotades
  - Generació d'instruccions
- Corpus de veu
  - Campanya de recollida de veu (CM)
  - Corpus de veu amb transcripció
  - Corpus d'altres tasques de veu
- Corpus paral·lels per TA
- Estratègia de provisió de dades



Serveis lingüístics

- Accés i desplegaments
- Integració en frameworks i plataformes de referència.
- APIS, demosDemos

Models

- Generació de **models** de llengua
- Tecnologies de la **parla**
- Tecnologies de la **traducció**

Eines **d'avaluació** i **benchmarking**
La **IA ètica**

Provisió i processament de les **dades** lingüístiques

- Corpus textual
- Generació de dades anotades
  - Generació d'instruccions
- Corpus de veu
  - Campanya de recollida de veu (CM)
  - Corpus de veu amb transcripció
  - Corpus d'altres tasques de veu
- Corpus paral·lels per TA
- Estratègia de provisió de dades

Provisió de dades

- **Corpus textual**
- Generació de dades anotades
  - Generació d'instruccions
- Corpus de veu
  - Campanya de recollida de veu (CM)
  - Corpus de veu amb transcripció
  - Corpus d'altres tasques de veu
- Corpus paral·lels per TA
- Estratègia de provisió de dades

V3. Corpus textual, pre-processat i publicat *

Corpus	v1 (GB)	v2 (GB)	v3 (GB)	Estat	Disponible a	Llicència
DOGC	0.78	0.78	0.92	v2	OPUS	CC0 4.0
Catalan Open Subtitles	0.02	0.02	0.02	v1	OPUS	Oberta
Catalan Oscar	4.00	4.00	27.10	v2	OSCAR	CC0 4.0
CaWaC	3.60	3.60	8.70	v2	CaWac	CC-BY-SA-3.0
Cat. General Crawling	2.50	2.50	5.70	v2	Zenodo	CC-BY 4.0
MaCoCu	-	-	11.00	Nou	MaCoCu	CC-BY 4.0
Viquipèdia	0.98	1.10	1.70	v3	-	CC-BY 4.0
ACN	0.45	0.45	0.45	v1	Zenodo	CC-BY-NC-ND 4.0
NacióDigital	-	0.45	0.45	v1		
VilaWeb	-	0.06	0.30	v2		
Catalan mc4	-	-	41.00	Nou	HuggingFace	CC-BY 4.0
Grup El Món	-	-	0.55	Nou	-	CC-BY 4.0
RacoCatalà	-	8.10	11.00	Nou	HuggingFace	CC-BY-NC 4.0
Altres	0.24	13.89	18.26	v3	-	CC-BY 4.0
Total	**12.57**	**34.95**	**127.5**			

*** Totes aquestes dades es publicaran properament a HuggingFace, sota el nom de CATalog 1.0, i constituiran el dataset de preentrenament de LLMs en català més gran fins ara.**

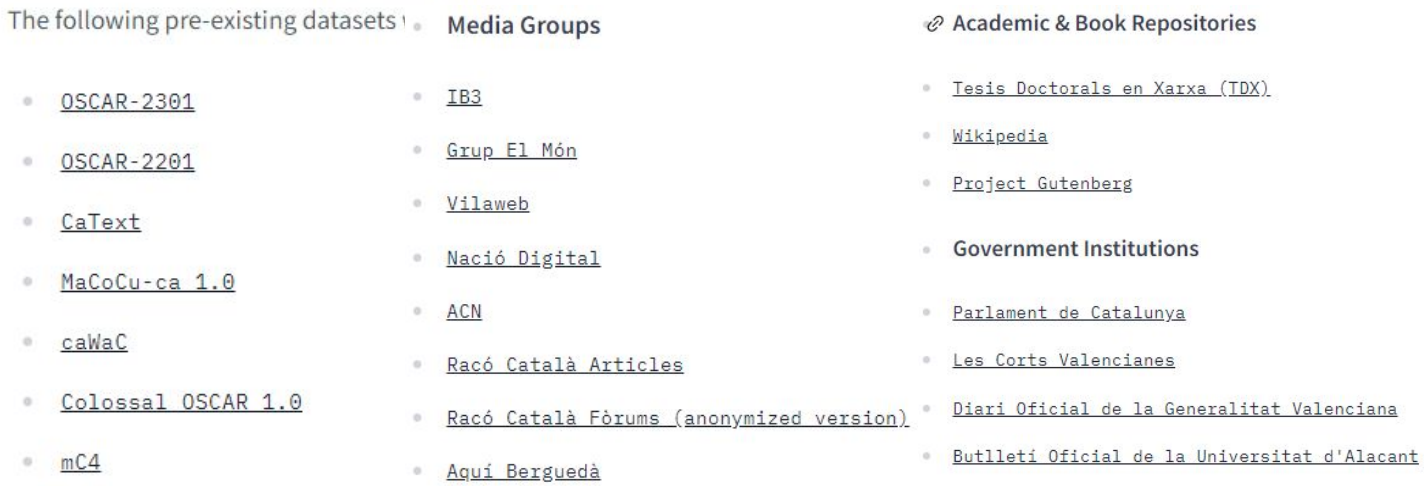
Provisió de dades

- **Corpus textual**
- Generació de dades anotades
  - Generació d'instruccions
- Corpus de veu
  - Campanya de recollida de veu (CM)
  - Corpus de veu amb transcripció
  - Corpus d'altres tasques de veu
- Corpus paral·lels per TA
- Estratègia de provisió de dades

- **CATALOG 1.0:** és el dataset de preentrenament de LLMs més gran en català alliberat fins ara. A més, 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.

  Està format per **17.450.496.729 de paraules** (al voltant de **23 mil milions de tokens**) distribuïts en 34.314.510 documents.

  - Accés: https://huggingface.co/datasets/projecte-aina/CATalog



Provisió de dades

- Corpus textual
- **Generació de dades anotades**
  - Generació d'instruccions
- Corpus de veu
  - Campanya de recollida de veu (CM)
  - Corpus de veu amb transcripció
  - Corpus d'altres tasques de veu
- Corpus paral·lels per TA
- Estratègia de provisió de dades

Creació de col·leccions de dades anotades per *fine tuning* i/o avaluació

NOU ***CEIL:*** Corpus de 60.000 textos curts per a la identificació, classificació i vinculació d'entitats. Conté 9 tipus i 52 subtipus.
- Tasca: Identificació, classificació i vinculació d'entitats.
- Accés: https://huggingface.co/datasets/projecte-aina/ceil

NOU ***CoQCat:*** Corpus de 6.000 paràgrafs anotats amb una conversa d'uns 15 torns de pregunta-resposta.
- Tasca: Pregunta resposta conversacional.
- Accés: https://huggingface.co/datasets/projecte-aina/CoQCat

NOU ***CaSSA***: Corpus de 6.400 ressenyes i missatges de fòrum anotats amb expressions de polaritat.
- Tasca: Anàlisi de sentiments.
- Accés: https://huggingface.co/datasets/projecte-aina/CaSSA-catalan-structured-sentiment-analysis

NOU ***CaSET:*** Corpus de tuits anotats amb emocions i opinió. Conté 11.000 frases úniques sobre cinc temes controvertits, agrupades en 6.000 parells de frases.
- Tasca: Detecció d'emocions i opinió.
- Accés: https://huggingface.co/datasets/projecte-aina/CaSET-catalan-stance-emotions-twitter

NOU ***CaSERa***: Corpus de missatges de fòrum anotats amb emocions i opinió. Conté 15.782 frases úniques agrupades en 10.745 parells de frases.
- Tasca: Detecció d'emocions i opinió.
- Accés: https://huggingface.co/datasets/projecte-aina/CaSERa-catalan-stance-emotions-raco

Provisió de dades

- Corpus textual
- **Generació de dades anotades**
  - Generació d'instruccions
- Corpus de veu
  - Campanya de recollida de veu (CM)
  - Corpus de veu amb transcripció
  - Corpus d'altres tasques de veu
- Corpus paral·lels per TA
- Estratègia de provisió de dades

NOU ***CaBreu**:* Corpus de resums abstractius, extractius i extrems de 3.000 documents.
- Tasca: Resum de textos.
- Accés: https://huggingface.co/datasets/projecte-aina/caBreu

NOU ***NLUCat**:* Corpus de 12.000 frases anotades segons el seu *intent* i els *slots* més rellevants, com a dades de suport per al desenvolupament d'assistents electrònics.
- Tasca: Detecció d'intencions.
- Accés: https://huggingface.co/datasets/projecte-aina/NLUCat

NOU ***IntoxiCAT**:* Corpus de 29.809 frases obtingudes de missatges de fòrums, anotades segons si són o no abusives.
- Tasca: Detecció de llenguatge abusiu.
- Accés: https://huggingface.co/datasets/projecte-aina/InToxiCat

NOU ***COPA-ca**:* Traducció al català del corpus COPA, corpus de referència per al raonament causal. Inclou 1.000 instàncies, cadascuna de les quals es compon d'una premissa i dues hipòtesis (o alternatives).
- Tasca: Raonament causal.
- Accés: https://huggingface.co/datasets/projecte-aina/COPA-ca

NOU ***PAWS-ca**:* Traducció al català del corpus PAWS per a la indentificació de paràfrasis. Conté 4.000 parells d'exemples traduïts per humans i 49.400 parells traduïts automàticament.
- Tasca: Paràfrasi.
- Accés: https://huggingface.co/datasets/projecte-aina/PAWS-ca

Provisió de dades

- Corpus textual
- **Generació de dades anotades**
  - Generació d'instruccions
- Corpus de veu
  - Campanya de recollida de veu (CM)
  - Corpus de veu amb transcripció
  - Corpus d'altres tasques de veu
- Corpus paral·lels per TA
- Estratègia de provisió de dades

NOU

- ***XNLI-ca**:* Traducció al català del corpus XNLI per a l'avaluació de sistemes de comprensió de llenguatge cross-lingüe en tasques com ara la inferència del llenguatge natural. Conté 7.500 parells de frases.
  - Tasca: Implicació textual.
  - Accés: https://huggingface.co/datasets/projecte-aina/xnli-ca
- ***WNLI-ca***: Traducció al català del corpus WNLI, que inclou 855 parells d'oracions, en què la primera frase conté una ambigüitat i la segona una possible interpretació d'aquesta.
  - Tasca: Implicació textual.
  - Accés: https://huggingface.co/datasets/projecte-aina/wnli-ca
- ***Teca**:* Corpus d'implicació textual. Conté 21.163 parells de premisses i hipòtesis, anotades segons la relació d'inferència que tenen (implicació, contradicció o neutre).
  - Tasca: Implicació textual
  - Accés: https://huggingface.co/datasets/projecte-aina/teca
- ***TeCla v2**:* Corpus de notícies en català per a tasques de classificació de textos multiclasse.
  - Tasca: Classificació de textos.
  - Accés: https://huggingface.co/datasets/projecte-aina/tecla
- ***WikiCAT_ca**:* Corpus català per a tasques de classificació temàtica de textos no periodístics.
  - Tasca: Classificació de documents.
  - Accés: https://huggingface.co/datasets/projecte-aina/WikiCAT_ca

Provisió de dades

- Corpus textual
- **Generació de dades anotades**
  - Generació d'instruccions
- Corpus de veu
  - Campanya de recollida de veu (CM)
  - Corpus de veu amb transcripció
  - Corpus d'altres tasques de veu
- Corpus paral·lels per TA
- Estratègia de provisió de dades

- ***CaWikiTC**:* Corpus creat de manera automàtica a partir dels resums d'articles de la Viquipèdia i la categoria temàtica associada. Conté 21.002 textos classificats en 67 categories.
  - Tasca: Classificació de textos.
  - Accés: https://huggingface.co/datasets/projecte-aina/CaWikiTC
- ***XQUAD-ca**:* Traducció al català del dataset XQUAD, que consta d'un subconjunt de 240 paràgrafs i 1.190 parells de pregunta-resposta del dataset SQuAD v1.1
  - Tasca: Pregunta resposta multilingüe.
  - Accés: https://huggingface.co/datasets/projecte-aina/xquad-ca
- ***VilaQuAD**:* Corpus de parells preguntes/resposta sobre notícies. Conté 2.095 articles de notícies en català i d'1 a 5 preguntes amb la seva resposta per a cada fragment (o context).
  - Tasca: Pregunta resposta.
  - Accés: https://huggingface.co/datasets/projecte-aina/vilaquad
- ***ViquiQuAD**:* Corpus de parells preguntes/resposta sobre la Viquipèdia. Conté 3.111 contextos extrets d'un conjunt de 597 articles originals i d'1 a 5 preguntes amb la seva resposta per a cada fragment.
  - Tasca: Pregunta resposta.
  - Accés: https://huggingface.co/datasets/projecte-aina/viquiquad
- ***CatalanQA**:* Corpus de 21.426 parells preguntes/resposta sobre la Viquipèdia i notícies.
  - Tasca: Pregunta resposta.
  - Accés: https://huggingface.co/datasets/projecte-aina/catalanqa

Provisió de dades

- Corpus textual
- **Generació de dades anotades**
  - Generació d'instruccions
- Corpus de veu
  - Campanya de recollida de veu (CM)
  - Corpus de veu amb transcripció
  - Corpus d'altres tasques de veu
- Corpus paral·lels per TA
- Estratègia de provisió de dades

- ***CAT ManyNames****:* Versió catalana del corpus *ManyNames*, orientat a models de Llenguatge i Visió. Consisteix en més de 23.000 imatges i les anotacions corresponents, traduïdes automàticament de l'anglès, més un conjunt de 1.072 imatges anotades a mà directament en català.
  - Tasca: Identificació d'imatges.
  - Accés: https://huggingface.co/datasets/projecte-aina/cat_manynames
- ***XitXat****:* Corpus de 950 converses de 10 dominis diferents entre xatbots i usuaris.
  - Tasca: Classificació d'*intents,* detecció d'entitats associades i entrenament/avaluació de sistemes conversacionals.
  - Accés: https://zenodo.org/record/7276036#.Y2zMn4LMITU
- ***Parafraseja****:* Corpus de 21.984 parells de frases anotades segons si són paràfrasis l'una de l'altra, o no.
  - Tasca: Paràfrasi.
  - Accés: https://huggingface.co/datasets/projecte-aina/Para fraseja
- ***STS-ca***: Corpus per a l'avaluació de la similitud textual semàntica.
  - Tasca: Similitud textual semàntica.
  - Accés: https://huggingface.co/datasets/projecte-aina/sts-ca
- ***NoNiRes****:* Anotació de les expressions de negació de 20.541 frases en català.
  - Tasca: Negació.
  - Accés: https://zenodo.org/record/7319487#.Y3S_uL7MLOs

Provisió de dades

- Corpus textual
- **Generació de dades anotades**
  - Generació d'instruccions
- Corpus de veu
  - Campanya de recollida de veu (CM)
  - Corpus de veu amb transcripció
  - Corpus d'altres tasques de veu
- Corpus paral·lels per TA
- Estratègia de provisió de dades

- ***GuiaCat**:* Corpus de 5.750 ressenyes de restaurants en català de la plataforma GuiaCat. Cada ressenya té associada una valoració per servei, menjar, qualitat-preu i ambient, i una nota mitjana.
  - Tasca: Anàlisi de sentiments.
  - Accés: https://huggingface.co/datasets/projecte-aina/GuiaCat
- ***ANCORA_ca v2**:* Corpus d'entrenament de cadenes de processament, afegint la columna NER a la versió CONLLU de UD versió 9, per fer *multitask learning* dins de Spacy.
  - Tasca: Anotació d'entitats amb nom i dependències.
  - Accés: https://doi.org/10.5281/zenodo.5036650
- ***AnCora-Ca-NER***: Corpus d'entrenament per a l'anotació d'entitats, basat en el corpus ANCORA de la UB.
  - Tasca: Anotació d'entitats.
  - Accés: https://huggingface.co/datasets/projecte-aina/ancora-ca-ner

Provisió de dades

Creació de col·leccions de dades d'instrucció de models

- Corpus textual
- Generació de dades anotades
  - **Generació d'instruccions**
- Corpus de veu
  - Campanya de recollida de veu (CM)
  - Corpus de veu amb transcripció
  - Corpus d'altres tasques de veu
- Corpus paral·lels per TA
- Estratègia de provisió de dades

InstruCAT: Conversió de dades de datasets d'AINA per instruir LLMs en català:

216 mil instruccions per entrenament de LLMs en tasques "downstream", com pregunta/resposta, resum, , parafrasi, classificació, etc.

Generades a partir dels següents **datasets:** caBreu, CatalanQA, CoQCat, GuiaCat, IntoxiCat, Parafraseja, PAWS-ca, sts-ca, WikiCat, CEIL, TECA, NLUCat.

Category	Number of instructions	%
ner	59410	25.24%
paraphrasis	34695	14.74%
text_classification	33393	14.19%
toxicity	29809	12.66%
qa	27427	11.65%
emotion_detection	18492	7.85%
phrase_generation	11873	5.04%
entailment_generation	6354	2.70%
sentiment_analysis	5750	2.44%
abstractive_summarization	2999	1.27%
extreme_summarization	2999	1.27%
entailment	2117	0.89%

ILENIA
IMPULSO DE LAS LENGUAS EN LA INTELIGENCIA ARTIFICIAL

https://huggingface.co/datasets/BSC-LT/InstruCat

Provisió de dades

- Corpus textual
- Generació de dades anotades
  - **Generació d'instruccions**
- Corpus de veu
  - Campanya de recollida de veu (CM)
  - Corpus de veu amb transcripció
  - Corpus d'altres tasques de veu
- Corpus paral·lels per TA
- Estratègia de provisió de dades

Generació d'un **dataset d'instruccions es-ca**: Dataset d'instruccions en castellà [MentorES] (i posterior traducció al català) per avaluar sistemes d'intel·ligència artificial. Inclou 10.175 tasques, distribuïdes en les següents categories: closed_qa, classification, open_qa, summarization, general_qa information_extraction, brainstorming i creative_writing.

Traducció al català de **datasets d'instruccions**:

- Dolly
- OpenAssistant

Traducció al català de **datasets d'avaluació**:

- ARCChallenge i ARCEasy (raonament i coneixement bàsic sobre ciències).
- PIQA, (raonament lògic sobre el món físic).
- OpenBookQA (coneixements bàsics del món i certa capacitat de raonament).
- MGSM (raonament matemàtic).
- XStory Cloze (comprensió del llenguatge i raonament espacial i temporal propi de les narracions).

Provisió de dades

- Corpus textual
- Generació de dades anotades
  - Generació d'instruccions
- Corpus de veu
  - **Campanya recollida de veu CM**
  - Corpus de veu amb transcripció
  - Corpus d'altres tasques de veu
- Corpus paral·lels per TA
- Estratègia de provisió de dades

El català, entre les llengües 'top' en l'entrenament del nou model de traducció automàtica de Google

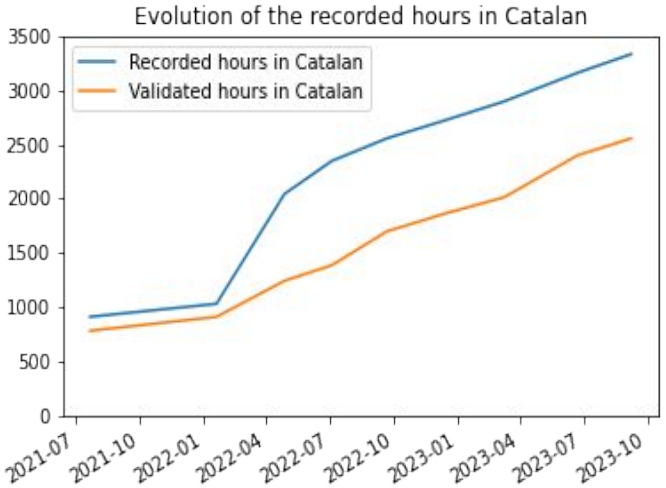
L'AudioPaLM, capaç de generar text i veu en multitud d'idiomes, es nodreix d'un repositori de veus amb aportacions de tot el món

Google

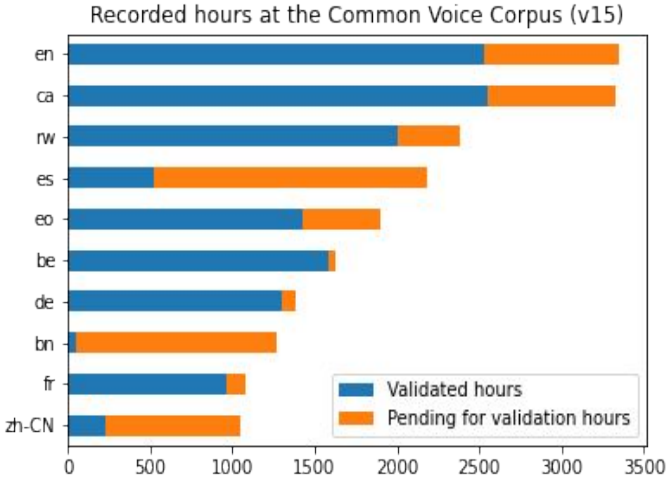
La recollida de dades mitjançant el **Common Voice** ha continuat estable:

- Des del gener de 2023 s'han recollit **més de 800 hores**.
- Des de l'inici del projecte s'han recollit **més de 2.500 hores**.
- Hi han col·laborat més de **35.000 voluntaris**

El català s'ha situat com a **primera llengua del corpus** pel que fa a **hores validades** i com a **segona llengua amb més hores enregistrades**, a molt poca distància de la primera



Evolució d'hores enregistrades i validades



*Llengües amb més hores enregistrades a la **v. 15***

Common Voice
moz://a

https://commonvoice.mozilla.org/ca/datasets

Provisió de dades

- Corpus textual
- Generació de dades anotades
  - Generació d'instruccions
- Corpus de veu
  - Campanya recollida de veu CM
  - **Corpus de veu amb transcripció**
  - Corpus d'altres tasques de veu
- Corpus paral·lels per TA
- Estratègia de provisió de dades

- ***ParlamentParla***: Corpus d'àudio de 611h d'enregistraments del Parlament. La preparació de la versió 3 està en marxa amb un mínim de 300h més d'enregistraments..
  - Tasca: Reconeixement de Parla.
  - Accés: https://huggingface.co/datasets/projecte-aina/parlament_parla

NOU ***CCMA TV3***: Corpus d'àudio de 735h d'enregistraments de diversos programes de TV3
  - Tasca: Reconeixement de Parla
  - Accés: Per l'ús intern, s'està parlant amb la CCMA per alliberar-ho.

NOU ***TTS cat***: Versions netes dels corpus de FESTCAT i OpenSLR69, específicament preparades per entrenar models neuronals de TTS
  - Tasca: Síntesi de la parla.
  - Accés: https://huggingface.co/datasets/projecte-aina/festcat_trimmed_denoised/
  - Accés: https://huggingface.co/datasets/projecte-aina/openslr-slr69-ca-trimmed-denoised/

ILENIA
IMPULSO DE LAS LENGUAS EN LA INTELIGENCIA ARTIFICIAL

Corts Valencianes (ILENIA): Corpus d'àudio amb més de 300h d'enregistraments.
  - Tasca: Reconeixement de la Parla.

IB3 (ILENIA): Corpus d'àudio amb 30h de programació d'IB3.
  - Tasca: Reconeixement de la Parla.

Provisió de dades

- Corpus textual
- Generació de dades anotades
  - Generació d'instruccions
- Corpus de veu
  - Campanya recollida de veu CM
  - Corpus de veu amb transcripció
  - **Corpus d'altres tasques de veu**
- Corpus paral·lels per TA
- Estratègia de provisió de dades

NOU ***Transcripció Fonètica:*** Transcripció fonètica multi dialectal pel català. Inclou 4 dialectes, 160 frases cada un: central, nord-occidental, valencià, balear.

- Tasca: Transcripció fonètica.
- Accés: https://huggingface.co/datasets/projecte-aina/4catac

NOU ***Annotated Catalan CV:*** Anotació del corpus Common Voice amb gènere i dialecte dels parlants. En col·laboració amb el CLiC (Centre de Llenguatge i Computació) - UB

- Tasca: Identificació de gènere, identificació de dialecte
- Accés: https://huggingface.co/datasets/projecte-aina/annotated_catalan_common_voice

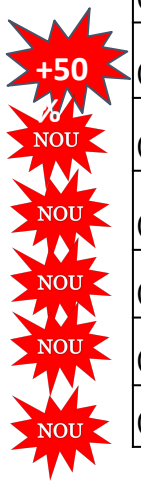
sentence string · *lengths* 26 131	transcription string · *lengths* 31 146
Ses coses importants són ses que no ho semblen	səs kˈɔzəz importˈan sˈon səs kə nˈo w sˈəmblən
Vaig obrir sa caixa que temps enrere vaig omplir de records des meu germà	vˈadʒ oβrˈi sə kˈaʃə kə tˈenz ənrˈerə vˈadʒ omplˈi ðə rəkˈɔr ðəz mˈew ʒərmˈa
Passejàvem ara pes estrets carrers que tants de cops havíem recorregut feia anys	pəsəðʒˈavəm ˈarə pəðz əstrˈək kərˈes kə tˈaɲ də kˈɔdz əvˈiən rəkorəɣˈuf fˈəjə ˈajns
Te deix amor la mar com a penyora	tə ðˈeʒ əmˈo lə mˈa kˈɔm ə pɐɲˈorə
Setze exercicis d'un jutge	sˈɛdz əðʒərsˈisiz ðuɲ ʒˈuðʒə
Una pinteta d'or sobre es coixí	unə piɲtˈətə ðˈɔ sˈoβr əs koʃˈi
Sa seva germana no s'acaba de decidir	sə sˈevə ʒərmˈanə nˈo səkˈaβə ðə ðəsiˈði
En Joan diu que mos vàrem conèixer a ses festes des	ən ʒuˈaɲ dˈiw kə moz vˈarəɲ konˈəʃ ə səs fˈestəz ðəs

‹ Previous 1 2 3 ... 7 Next ›

Provisió de dades

- Corpus textual
- Generació de dades anotades
  - Generació d'instruccions
- Corpus de veu
  - Campanya recollida de veu CM
  - Corpus de veu amb transcripció
  - Corpus de veu no alineat
  - Corpus de traducció parla a parla
- **Corpus paral·lels per TA**
  - **Corpus entrenament**
  - Corpus avaluació
- Estratègia de provisió de dades

Corpus paral·lels per entrenament de models de TA



Llengües	Num. de Frases	Orígen de les dades	Disponible a
Català-Espanyol	85.953.513	Diverses fonts	-
Català-Anglès	22.592.622	Diverses fonts + AINA	projecte-aina/CA-EN_Parallel_Corpus
Català-Francès	18.634.844	OPUS	projecte-aina/CA-FR_Parallel_Corpus
Català-Portuguès	9.892.953	OPUS+Softcatalà (augmentat)	projecte-aina/CA-PT_Parallel_Corpus
Català-Italià	9.482.927	OPUS	projecte-aina/CA-IT_Parallel_Corpus
Català-Alemany	9.530.709	OPUS +Softcatalà (augmentat)	projecte-aina/CA-ZH_Parallel_Corpus
Català-Xinès	6.833.114	OPUS (augmentat)	projecte-aina/CA-DE_Parallel_Corpus

Llengües	Num. de Frases	Orígen de les dades	Disponible a
Català-Gallec	123.890.107	NOS + AINA	projecte-aina/CA-GL_Parallel_Corpus
Català-Euskera	96.692.186	GAITU + AINA	projecte-aina/CA-EU_Parallel_Corpus

En col·laboració amb

ILENIA
IMPULSO DE LAS LENGUAS EN LA INTELIGENCIA ARTIFICIAL

Provisió de dades

- Corpus textual
- Generació de dades anotades
  - Generació d'instruccions
- Corpus de veu
  - Campanya recollida de veu CM
  - Corpus de veu amb transcripció
  - Corpus de veu no alineat
  - Corpus de traducció parla a parla
- **Corpus paral·lels per TA**
  - Corpus entrenament
  - **Corpus avaluació**
- Estratègia de provisió de dades

Corpus paral·lels per adaptació i avaluació de models de TA

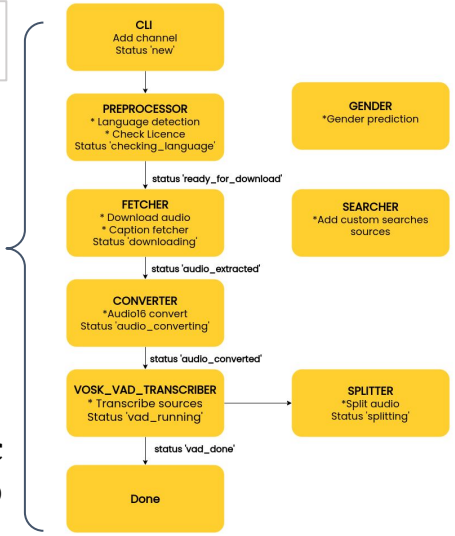
Corpus	Font	Llengües	Frases	Domini	Disponible a	Llicència
GEnCaTa	crawling .gencat	ca, en	38.595	Administratiu	ELRC-Share	CC0 4.0
Corpus bilingüe CA-EN de la CE	documents bilingües	ca , en	46.048	Administratiu	ELRC-Share	CC-BY 4.0
Col·lecció de corpus CA-EN de l'AP	documents bilingües	ca, en	36.116	Diversos	ELRC-Share	CC-BY 4.0
Col·lecció de corpus CA-ES de l'AP	documents bilingües	ca , es	63.773	Diversos	ELRC-Share	CC-BY 4.0
TaCon	document bilingüe	ca, es, gl, eu	1.314	Legal	ELRC-Share	CC-BY 4.0
Cyber MT test set	traducció	ca, en, es	1.715	Ciberseguretat	ELRC-Share	CC-BY-NC-SA 4.0
Catalan WMT2013	traducció	several	3.000	Notícies	ELRC-Share	CC-BY 4.0
Must-SHE Cat*	traducció	en, es, ca	1.046	Biaix de gènere		CC BY NC ND 4.0

*Permissos per publicar encara en tràmit

Provisió de dades

- Corpus textual
- Generació de dades anotades
  - Generació d'instruccions
- Corpus de veu
  - Campanya recollida de veu CM
  - Corpus de veu amb transcripció
  - Corpus de veu no alineat
- Corpus paral·lels per TA
- **Estratègia de provisió de dades**

Eines per al subministrament continu de dades de veu



- **Datapipe** una eina per facilitar l'adquisició dels continguts audiovisuals amb llicències obertes a la web. Projecte iniciat per la comunitat de programari lliure, que hem adaptat i millorat.

  Accés: https://github.com/projecte-aina/datapipe

- **Found speech pipeline:** una eina per processar els continguts adquirits i generar un corpus de la parla alineat. El pipeline és capaç de generar dades per tasques de parla, a partir de transcripcions i/o subtítols.

  Accés: TBA

L'objectiu és facilitar la **generació de datasets per a ASR, procesant continguts de forma automàtica**. Actualment, les estem usem en cinc contextos diferents:

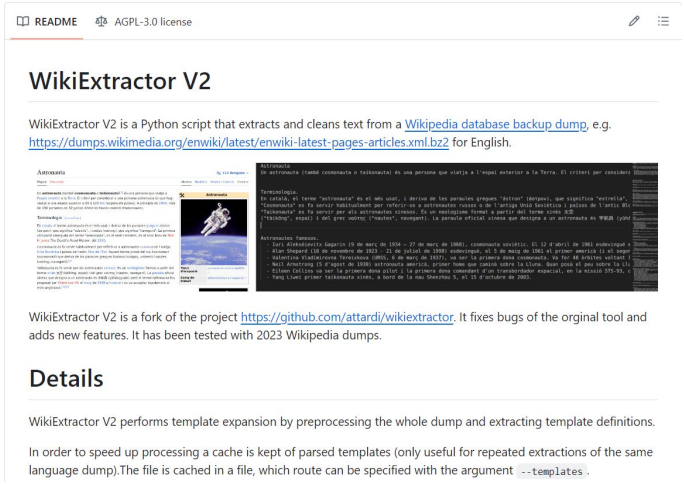
- **Canals de youtube**: S'han identificat i descarregat continguts amb subtítols (~ 300 hores)
- **CCMA** (~ 4000 hores)
- Parlament de Catalunya (~ 4000 hores)
- Corts Valencianes (1200 hores)
- IB3 (60 hores)

Provisió de dades

- Corpus textual
- Generació de dades anotades
  - Generació d'instruccions
- Corpus de veu
  - Campanya recollida de veu CM
  - Corpus de veu amb transcripció
  - Corpus de veu no alineat
- Corpus paral·lels per TA
- **Estratègia de provisió de dades**

Eines per al subministrament continu de dades de text

- Extractor de text de la **Viquipèdia** que interpreta correctament textos en una gran varietat de llengües, entre elles el català.
  - Stats: extracció de 689,141 documents amb més de 266M de paraules en català.
  - Accés: https://github.com/langtech-bsc/Wikiextractor-V2/



- Operacionalització de les dades obertes del **DOGC** mitjançant una pipeline automatitzada que fa ús de l'API de transparència de Catalunya.
  - Stats: fins a principis d'octubre, hem extret 30,369 publicacions en català amb 70M de paraules.
  - Accés: https://github.com/projecte-aina/docg-pipeline

Provisió de dades

- Corpus textual
- Generació de dades anotades
  - Generació d'instruccions
- Corpus de veu
  - Campanya recollida de veu CM
  - Corpus de veu amb transcripció
  - Corpus de veu no alineat
- Corpus paral·lels per TA
- **Estratègia de provisió de dades**

Accions per a l'adhesió de noves dades

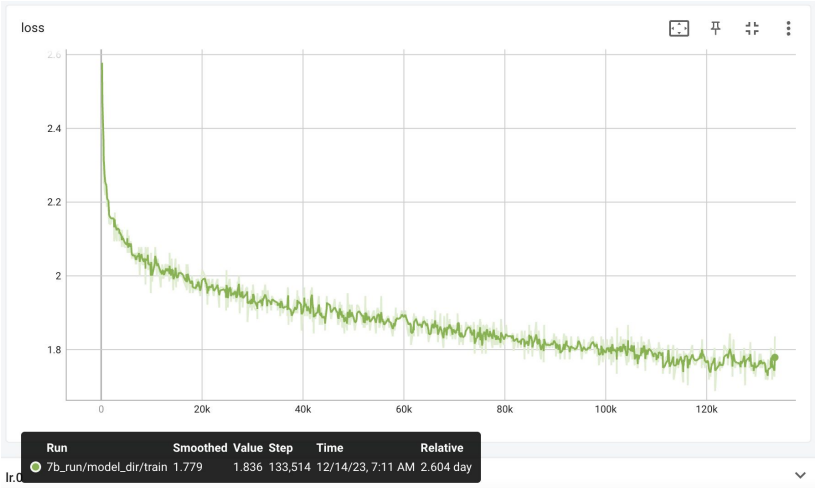
- Col·laboració amb el projecte Parla'm promogut per **Òmnium Cultural** per a la recollida de dades textuals i de veu.

  Fins a la data les següents entitats han signat el *Compromís de cessió de continguts*
  - Llenguaferits (text i àudio)
  - BIFIDUS PRODUCCIONS SL (text i àudio)
  - Unió de Pagesos (text i àudio)
  - Grup Món (text)
  - Núvol (text)
  - Betevé (àudio)

Models de la llengua

Generació de models de la llengua

- **Models Transformers de la llengua**
- Adaptació de models a tasques específiques (*fine tuning*)
- Participació en models multilingües i massius en col·laboració amb altres iniciatives



Models Transformers

HUGGING FACE
https://huggingface.co/projecte-aina

NOU **FLOR (Bloom) 760M**
- Model trilingüe (ca,es,en) generatiu basat en Bloom entrenat amb **26B de tokens**.
- https://huggingface.co/projecte-aina/FLOR-760M

NOU **FLOR (Bloom) 1.3B**
- Model trilingüe (ca,es,en) generatiu basat en Bloom entrenat amb **26B de tokens**.
- https://huggingface.co/projecte-aina/FLOR-1.3B

NOU **FLOR (Bloom) 6.3B**
- Model trilingüe (ca,es,en) generatiu basat en Bloom entrenat amb **140B de tokens!!!**.
- https://huggingface.co/projecte-aina/FLOR-6.3B
- Sandbox: https://huggingface.co/spaces/projecte-aina/flor-6.3b-inference

NOU **Flor 6.3B Instruit**
- Model Flor (Bloom) 6.3B pel català, anglès i castellá a partir de l'original
- Accés: https://huggingface.co/projecte-aina/flor-6.3B-instruct
- Sandbox: HF Spaces

NOU **DeBerta_ca**
- Entrenat amb la segona versió del corpus textual català
- Accés: (en proves)

NOU **Sentence Encoders**
- Entrenaments amb més dades del catalá dels embeddings
- https://huggingface.co/projecte-aina/ST-NLI-ca_paraphrase-multilingual-mpnet-base

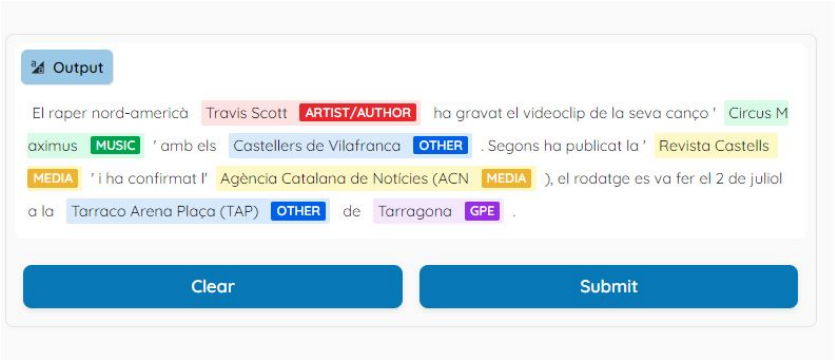
Models de la llengua

Models adaptats a tasques específiques

HUGGING FACE
https://huggingface.co/projecte-aina

Generació de models de la llengua

- Models Transformers de la llengua
- **Adaptació de models a tasques específiques (*fine tuning*)**
- Participació en models multilingües i massius en col·laboració amb altres iniciatives



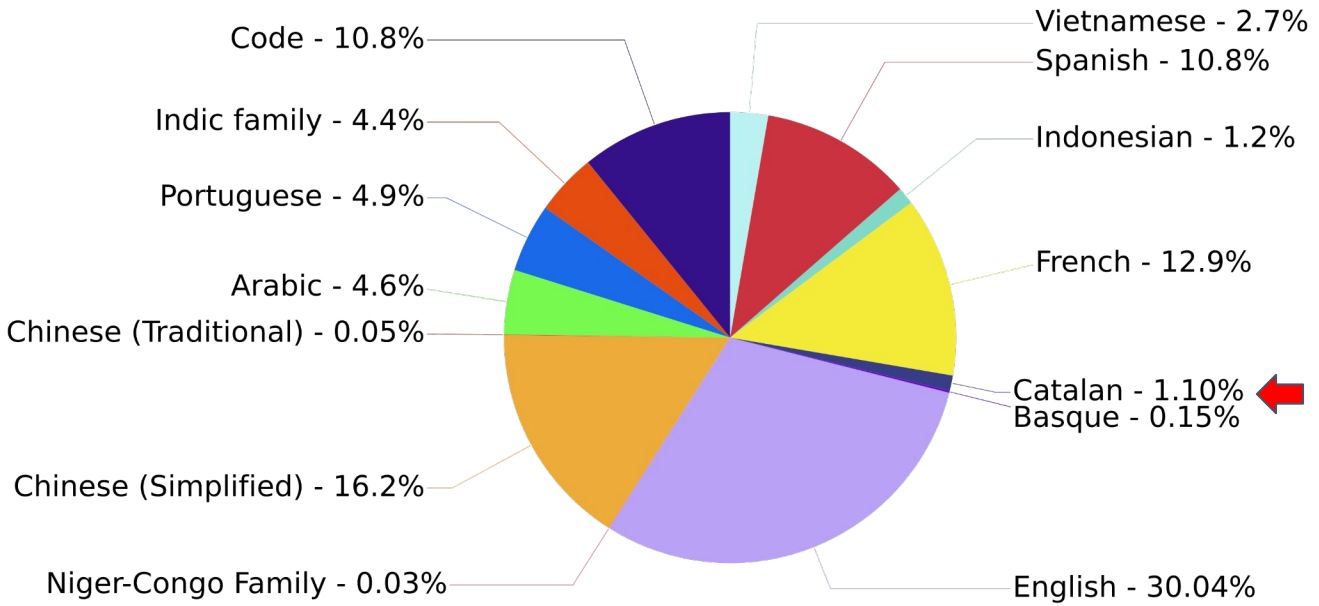
- **Catalan BERTa (RoBERTa-large) finetuned for Named Entity Recognition**
  - Tasca: Named Entity Recognition
  - Accés: https://huggingface.co/projecte-aina/multiner_ceil
- **RoBERTa-base v2 fine-tuned for QA**
  - Tasca: Pregunta/resposta (QA)
  - Accés: https://huggingface.co/projecte-aina/roberta-base-ca-v2-cased-qa
- **RoBERTa-base v2 fine-tuned for TE**
  - Tasca: Implicació textual
  - Accés: https://huggingface.co/projecte-aina/roberta-base-ca-v2-cased-te
- **RoBERTa-base v2 fine-tuned for STS**
  - Tasca: Similitud textual semàntica
  - Accés: https://huggingface.co/projecte-aina/roberta-base-ca-v2-cased-sts
- **RoBERTa-base v2 fine-tuned for Paraphrase Detection**
  - Tasca: Paràfrasi
  - accés: https://huggingface.co/projecte-aina/roberta-large-ca-paraphrase
- **Word embeddings Floret per al catalá, v1.0**
  - Accés: https://zenodo.org/record/733033

Models de la llengua

Generació de models de la llengua

- Models Transformers de la llengua
- Adaptació de models a tasques específiques (*fine tuning*)
- **Participació en models multilingües i massius en col·laboració amb altres iniciatives**

Liderant GPT3 massiu multilingüe



a BigScience initiative

BLOOM

176B params · 59 languages · Open-access

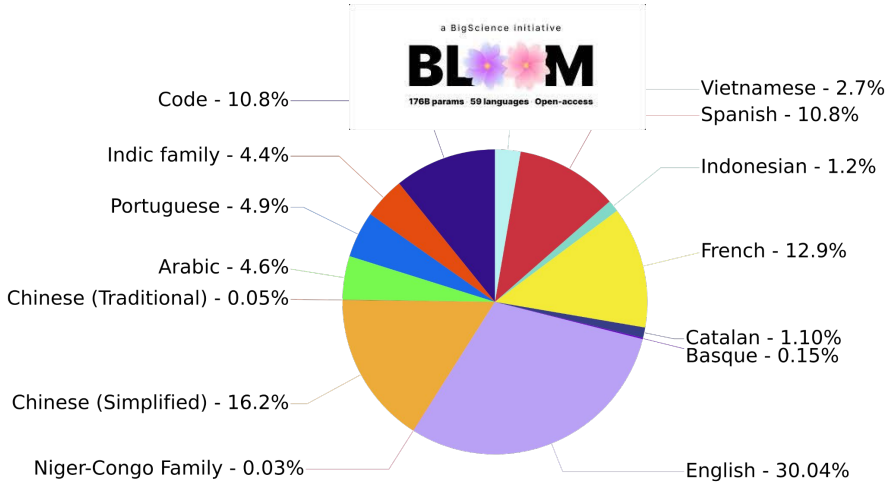
https://huggingface.co/bigscience/bloom

350B tokens (1,1% ca)

Models de la llengua

Generació de models de la llengua

- Models Transformers de la llengua
- Adaptació de models a tasques específiques (*fine tuning*)
- **Participació en models multilingües i massius en col·laboració amb altres iniciatives**



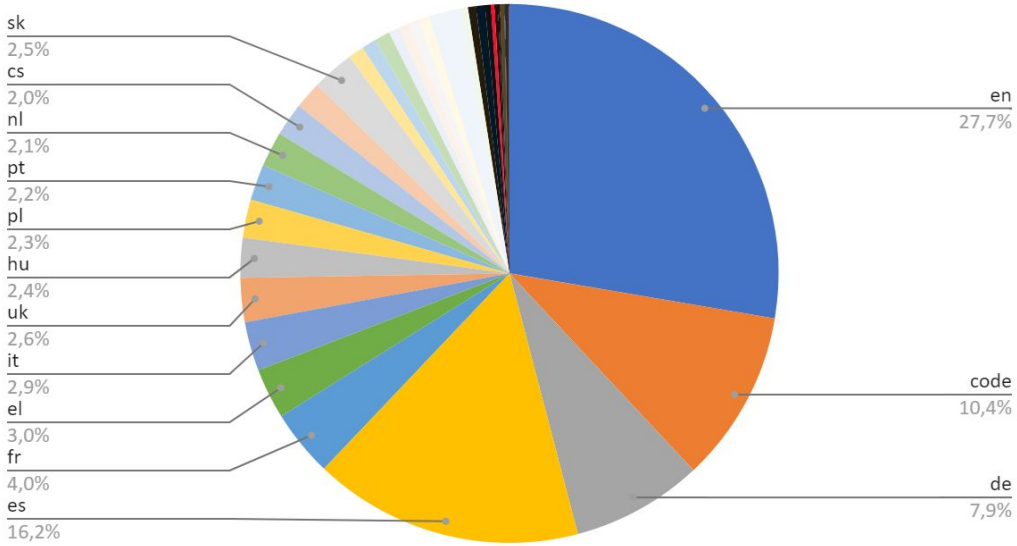
https://huggingface.co/bigscience/bloom

46 natural languages

350B tokens (1,1% ca)

Liderant GPT3 massiu multilingüe

ca 1.86%



32 natural languages

2,2T tokens (=2.200B) (1,86% ca)

ILENIA

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Models de la parla

Generació de models de la parla

- **Models de tecnologies de la parla**
- La presència dels models als ecosistemes i plataformes d'impacte

Motors de transcripció fonètica: Hem desenvolupat dos motors de transcripció fonètica, per convertir lletres en fonemes per 4 dialectes del català: central, nord-occidental, valencià i balear.
Estan integrats als motors de TTS més utilitzats.

- https://github.com/projecte-aina/espeak-ng/
- https://github.com/fedecosta/gruut
  en col·laboració amb la UPC

S'han utilitzat els enregistraments fets al **Common Voice** per crear dos models de la parla

Reconeixement de Parla - Nvidia Nemo: Alta precisió, mida mitjana, fàcilment desplegable amb el framework de **Nemo**.

- 36.5M paràmetres, WER of 6.684.
- https://huggingface.co/projecte-aina/stt-ca-citrinet-512

Síntesi de la Parla - Multiparlant: Model de síntesi de la parla d'alta qualitat i d'alt rendiment.
S'ha entrenat amb 255 veus i 487 hores d'enregistraments amb diverses variants dialectals.
A més de CV, aprofita els corpus de Festcat i de OpenSLR69. Fàcilment desplegable amb el framework de **Coqui**.

- https://huggingface.co/projecte-aina/tts-ca-coqui-vits-multispeaker
- integrat a la demo del bot d'AINA

Models de la parla

Generació de models de la parla

- Models de tecnologies de la parla
- **Presència dels models als ecosistemes i plataformes d'impacte**

coqui TTS

COQUI is a library for advanced Text-to-Speech generation

https://github.com/coqui-ai/TTS

RASA

Rasa is the leading platform for transforming how people interact with organizations through extensible conversational AI

https://rasa.com/

NVIDIA NEMO

NVIDIA NeMo, part of the NVIDIA AI platform, is a framework for building, training, and fine-tuning GPU-accelerated speech and natural language understanding (NLU)

https://developer.nvidia.com/nvidia-nemo

Models de Traducció

Generació de models de traducció

- **Models de traducció**
- Avaluació

En col·laboració amb ILENIA
IMPULSO DE LAS LENGUAS EN LA INTELIGENCIA ARTIFICIAL

Traductor gallec → català

Accés: https://huggingface.co/projecte-aina/mt-aina-gl-ca

Traductor basc → català

Accés: https://huggingface.co/projecte-aina/mt-aina-eu-ca

- ***Traductors català ↔ espanyol***
  - Accés (ca →es) : https://huggingface.co/projecte-aina/mt-aina-ca-es
  - Accés (es →ca) : https://huggingface.co/projecte-aina/mt-aina-es-ca
  - Accés (ca→es àmbit admin-legal): https://huggingface.co/projecte-aina/mt-aina-ca-es-adm
- ***Traductors català ↔ anglès***
  - Accés (ca →en) : https://huggingface.co/projecte-aina/mt-aina-ca-en
  - Accés (en →ca) : https://huggingface.co/projecte-aina/mt-aina-en-ca

NOU

Traductors català→português

- Accés (ca →pt) : https://huggingface.co/projecte-aina/mt-aina-ca-pt
- Accés (pt →ca) : https://huggingface.co/projecte-aina/mt-aina-pt-ca

NOU

Traductors català ↔ italià

- Accés(ca →it) : https://huggingface.co/projecte-aina/mt-aina-ca-it
- Accés(it →ca) : https://huggingface.co/projecte-aina/mt-aina-it-ca

NOU

Traductors català ↔ francès

- Accés(ca →fr) : https://huggingface.co/projecte-aina/mt-aina-ca-fr
- Accés(fr →ca) : https://huggingface.co/projecte-aina/mt-aina-fr-ca

NOU

Traductors català ↔ alemany

- Accés(ca →de) : https://huggingface.co/projecte-aina/mt-aina-ca-de
- Accés(de →ca) : https://huggingface.co/projecte-aina/mt-aina-de-ca

NOU

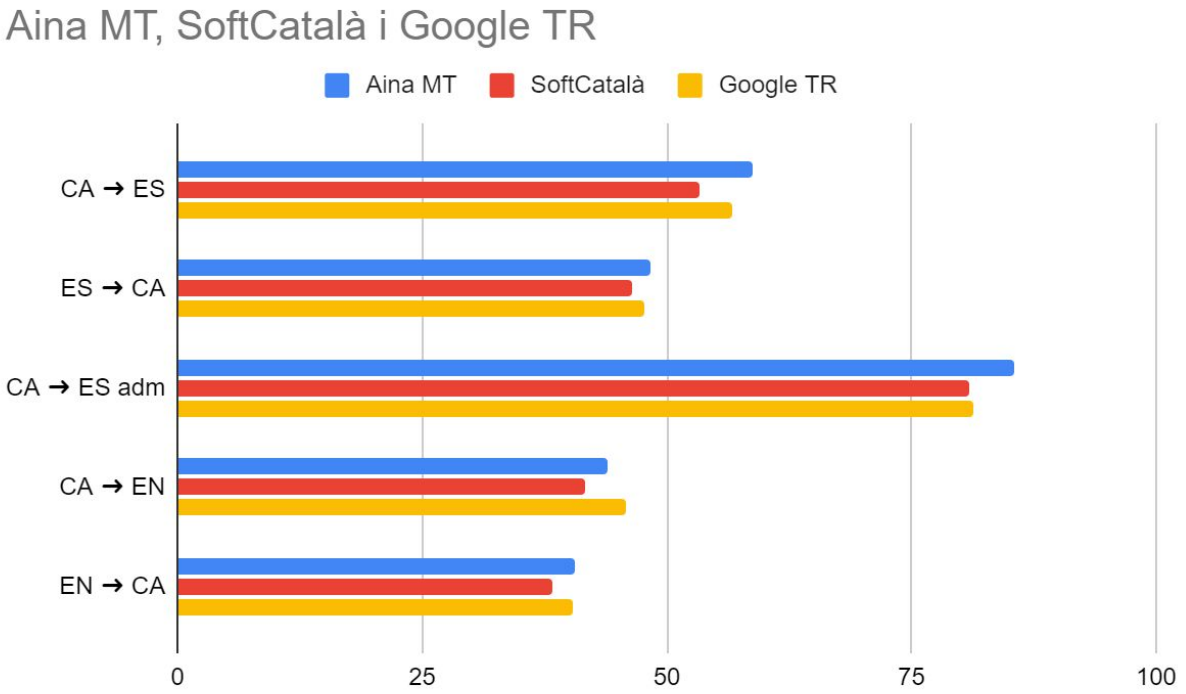
Traductors català ↔ xinès

- Accés(ca →zh) : https://huggingface.co/projecte-aina/mt-aina-ca-zh
- Accés(zh →ca) : https://huggingface.co/projecte-aina/mt-aina-zh-ca

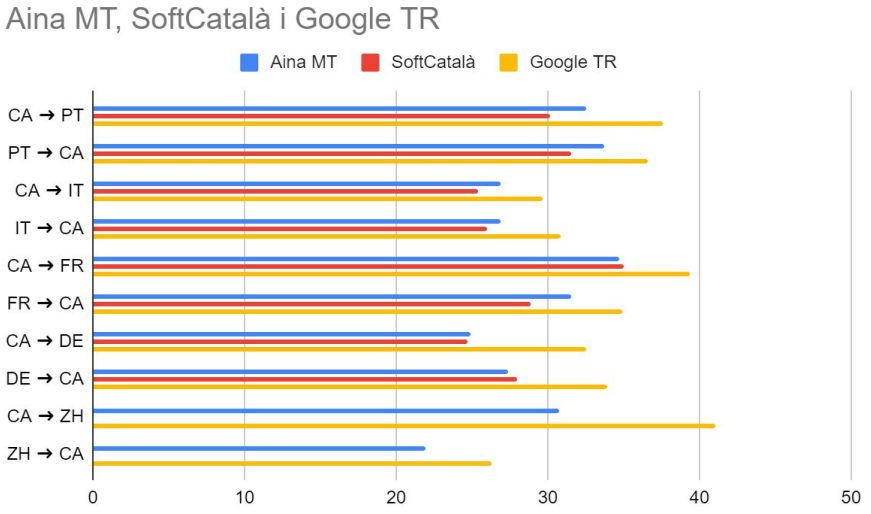
Models de Traducció

Generació de models de traducció

- Models de traducció
- **Avaluació**

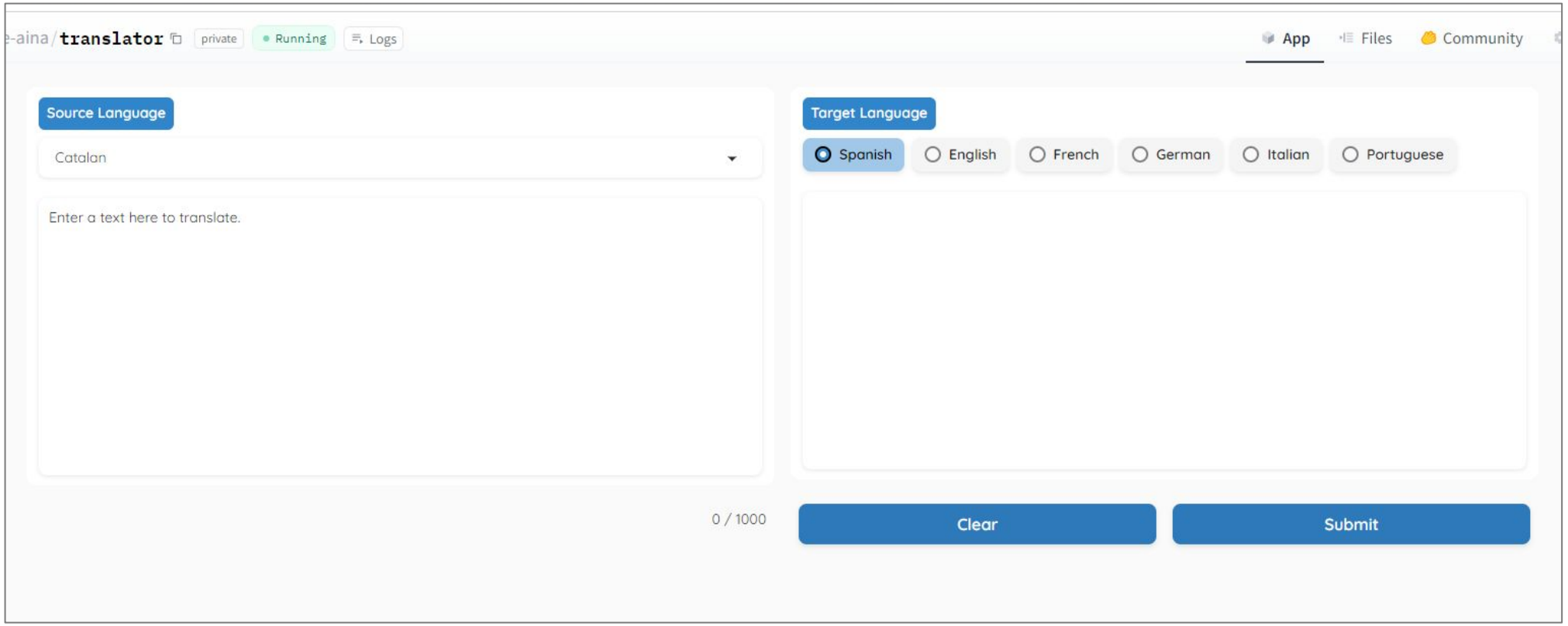


Resultats de l'avaluació dels models català -anglès i català - castellà (mètrica BLEU, fent mitjana sobre diferents dominis textuals).



Resultats de l'avaluació de les versions inicials dels nous models (mètrica BLEU, fent mitjana sobre diferents dominis textuals).

Models de Traducció



10 motors de traducció entre el català i diverses llengües europees, accessibles des del espai de Traducció del Projecte AINA a Hugging Face

Avaluació i benchmarking

- **Plataforma de benchmarking**
- Avaluació models generatius

Plataforma d'avaluació contínua de models amb avaluació extrínseca sobre 7 tasques diferents

https://club.aina.bsc.es/

Leaderboard

CLUB tests the ability of a system in the Catalan language. Below are the results of the different models.

Rank	Model	Submitted By	URL	Score	NER (F1)	POS (F1)	STS-ca (Comb.)	TeCla (Acc.)	TE-ca (Acc.)	CatalanQA (F1/EM)	XQuAD-ca (F1/EM)
1	RoBERTa-large-ca-v2	Projecte AINA		80.41	89.76	99.02	83.41	75.46	83.61	90.48/77.94	72.77/51.2
2	RoBERTa-base-ca-v2	Projecte AINA		79.29	89.27	98.95	79.07	74.26	83.14	89.37/75.64	72.79/51.1
3	mBERT	Projecte AINA		76.13	86.87	98.83	74.26	69.90	74.63	86.90/74.19	68.79/50.8
4	XLM-RoBERTa	Projecte AINA		71.23	86.31	98.89	61.61	70.14	33.30	88.17/75.93	72.55/54.1

showing 1 to 4 of 4 entries

SEND YOUR RESULTS

Avaluació i benchmarking

- Plataforma de benchmarking
- **Avaluació models generatius**

```
python main.py \
   --model hf-causal \
   --model_args projecte-aina/Aguila-7b \
   --tasks parafraseja, …….\
     --device cuda:0
```

EleutherAI / lm-evaluation-harness

Incorporació del català al nou LM-evaluation-Harness (EuletherAI)

- Incorporació de 15 datasets d'avaluació al LM-Harness
  - Belebele (reading Comprehension),
  - Flores (TA)
  - CaBreu (resum)
  - CatalanQA, XQuad, CoQCat (QA)
  - PawsX, Parafraseja (Parafrasis)
  - XNLI, TeCa (NLI)

  traduint
  - ARCChallenge i ARCEasy (raonament i coneixement sobre ciències).
  - PIQA, (raonament lògic sobre el món físic).
  - OpenBookQA (coneixements del món i certa capacitat de raonament).
  - MGSM (raonament matemàtic).
  - XStory Cloze (comprensió del llenguatge i raonament espacial i temporal propi de les narracions).
  - LMentry (avaluació capacitats lingüistiques)
- treballant en avaluació humana dels models generatius.

IA ètica

- **Anàlisi de biaixos**
- **Lluita contra la desinformació**

- **Identificació i filtratge de dades no desitjables** a les dades d'entrenament: contingut pornogràfic
  - integració a la Pipeline de preprocés
- **Inclusió string GUID** identificador únic per a poder excloure de futurs *crawlers* els datasets d'avaluació, garantint l'avaluació justa (Fair evaluation).
- Generació **corpus de factualitat** equivalent al TruthfulQA '**VeritasQA**' (360 preguntes/respostes en ca/en/es/gl/eu/. no conté coneixement 'localitzat', sense connexió geogràfica, 'traduible', comprovant factualitat i amb un source de confiança)
- Avaluacions de biaix de gènere amb el nou corpus Must-SHE de TA

Serveis lingüístics

- **Accés i desplegaments**
- Integració de models en frameworks i plataformes de referència.

- El serveis dins **Spaces** en **Huggingface** tenen oberta **API** gratuita, des de python, javascript o curl.
- Es poden fer servir per fer proves no-intensives des de codi.

Serveis lingüistics accessibles des de Spaces ‘Demos’ & APIs

NOU ***Model Flor 6.3 Instruit:*** https://huggingface.co/spaces/projecte-aina/flor-6.3b-instruct

NOU ***Model Flor 6.3:*** https://huggingface.co/projecte-aina/FLOR-6.3B

NOU **Transcripció fonètica** *https://huggingface.co/spaces/projecte-aina/transcripcio-fonetica-catala*

Serveis lingüístics

- **Accés i desplegaments**
- Integració de models en frameworks i plataformes de referència.

Serveis lingüistics accessibles des de Spaces ‘Demos’ & APIs

Models DeBerta i RoBErta adaptats a tasques (fine tuning) per:

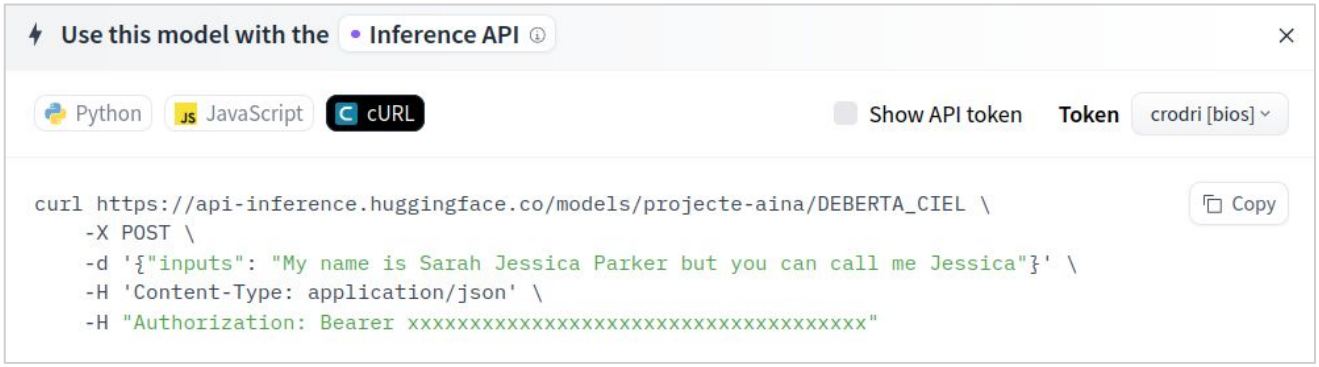
Classificació de textos

- Classificador entrenat amb dades de la Wikipedia
- Accés https://huggingface.co/projecte-aina/roberta-base-ca-v2-cased-wikicat-ca

Identificació d’entitats i conceptes

- Models de reconeixement i classificació d’entitats nomenades i esdeveniments (52 tipus i subtipus)
- Accés: Catala: https://huggingface.co/spaces/projecte-aina/multiner_demo
  Multilingue: https://huggingface.co/spaces/projecte-aina/multiner_demo NOU

- **Repositoris** de models en **Huggingface** tenen també **API** gratuita, des de python, javascript o curl.
- Però s’ha d'afegir un **token d’accés** del vostre usuari HF per fer proves no-intensives des de codi.



Serveis lingüístics

- Accés i desplegaments
- **Integració de models en frameworks i plataformes de referència.**

spaCy

[Spacy 3.7](#) permet ara la integració directa dels models Transformers d'AINA, inclosos LLMs: [spacy-huggingface-pipelines](#) | [spacy-transformers](#) | [spacy-llm](#)

```
import spacy
nlp = spacy.load("ca_core_news_trf")

nlp.add_pipe("hf_token_pipe", config={"model": "projecte-aina/multiner_ceil"})
doc = nlp("El Barça ha fitxat a Messi pel 2024.")
for e in doc.ents: print(e,e.label_)
#Barça organization-sportsteam
#Messi person-athlete
```

aws

S'han preparat **Notebooks** per fer servir al còmput al núvol de AWS els recursos d'AINA per fer **inferència** i **fine-tunning.** Accés [sagemaker-examples](#)

- Inference: [Aguila 7b Hugging Face Large Model Inference - TGI](#) shows how to deploy common large language models such as projecte-aina/aguila-7b, using Hugging Face Text Generation Inference (TGI) Deep Learning Container with SageMaker
- Fine-tuning: [Aguila 7b fine-tunning with instruction dataset](#) shows how to fine-tune the falcon 7B aguila model projecte-aina/aguila-7b, using an instructional dataset (in this case an example from the InstructCat collection) with a EC instance with Sagemaker.

- **Anonimitzador** de continguts generats per usuaris * https://github.com/TeMU-BSC/AnonymizationPipeline
- **TTS**: RestFUL API and web interface to serve coqui TTS models https://hub.docker.com/r/projecteaina/tts-api https://github.com/projecte-aina/tts-api
- **TA**: RestFUL API for serving machine translation models https://github.com/projecte-aina/mt-api
- **RASA:** Exemple d'assistent conversacional https://github.com/projecte-aina/minibot

* Avaluació en col·laboració amb 1million bot

On trobar-ho?

https://bit.ly/AINAKit

On trobar-ho?

https://bit.ly/AINAKit

https://github.com/projecte-aina

https://huggingface.co/projecte-aina

https://zenodo.org/communities/catalan-ai

https://elrc-share.eu/repository/search/

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

On trobar-ho?

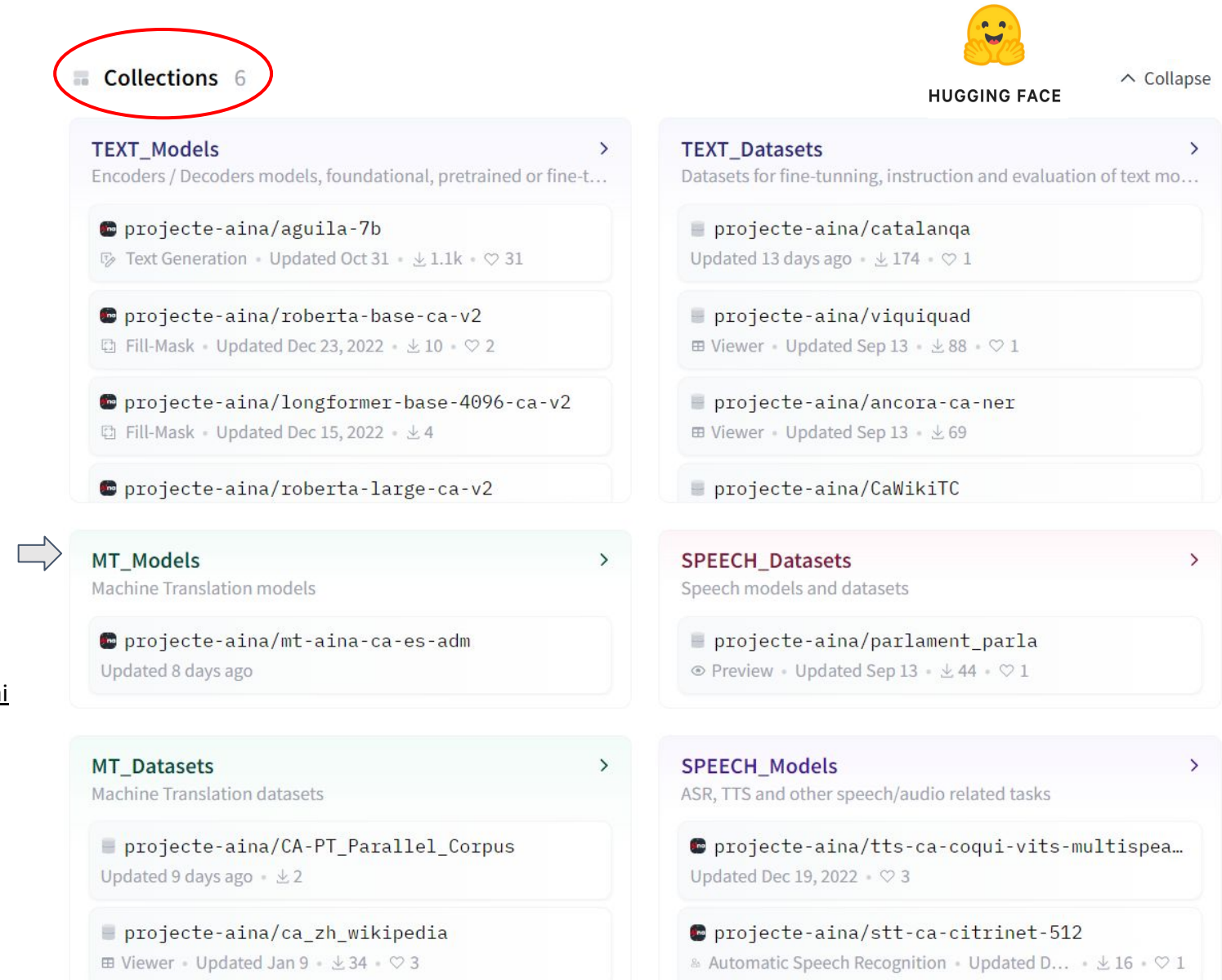
https://github.com/projecte-aina

https://huggingface.co/projecte-aina

https://zenodo.org/communities/catalan-ai

https://elrc-share.eu/repository/search/

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



On trobar-ho?

https://github.com/projecte-aina

HUGGING FACE

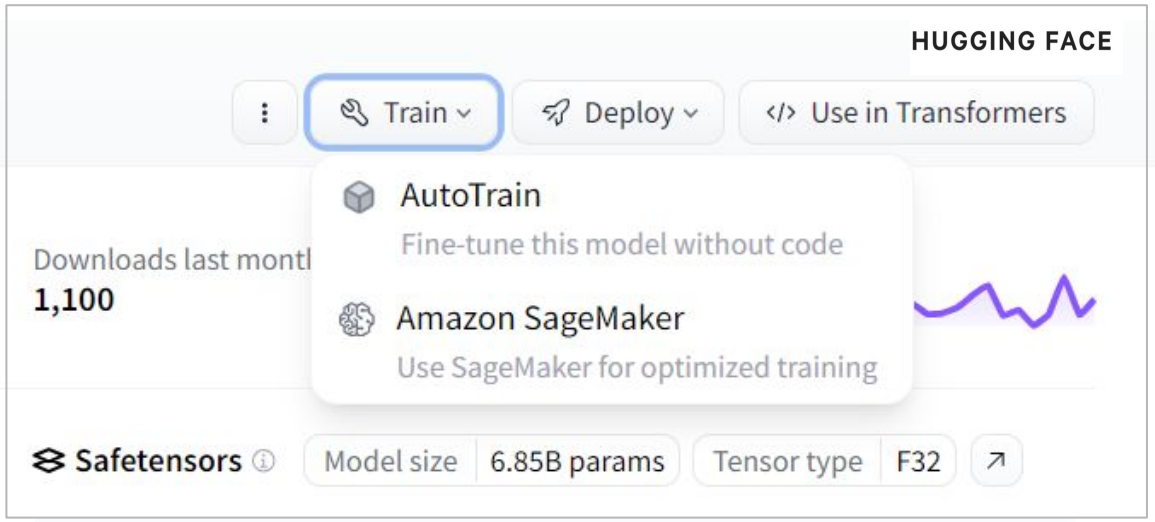
https://huggingface.co/projecte-aina

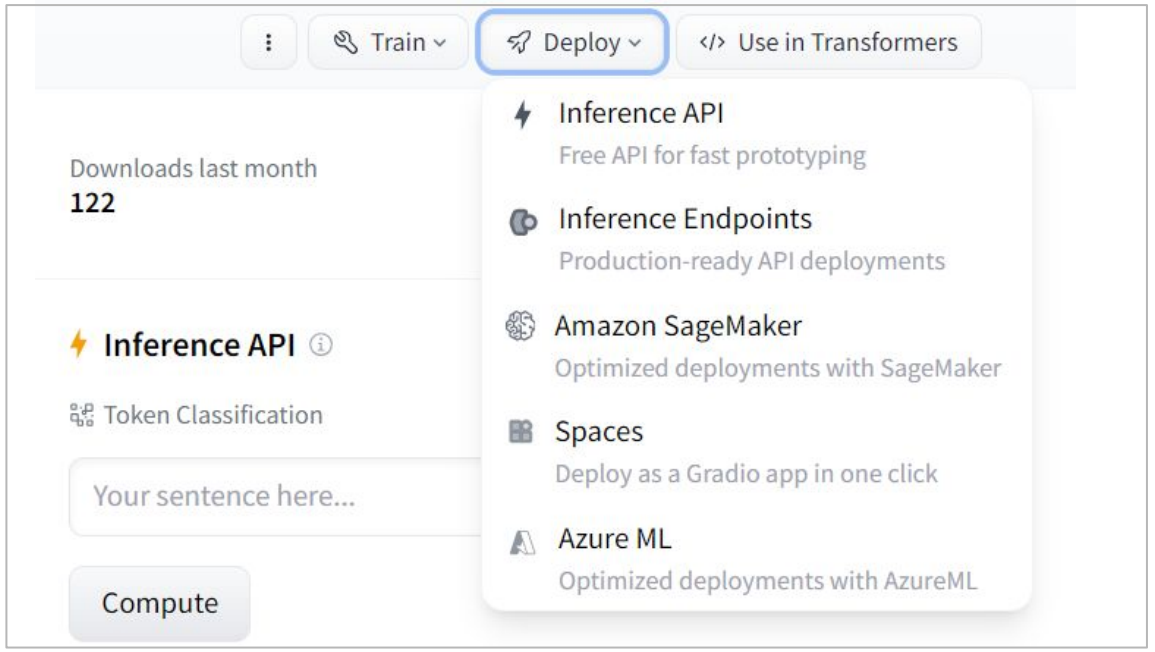
https://zenodo.org/communities/catalan-ai

European Language Resource Coordination
Connecting Europe Facility
ELRC-SHARE

https://elrc-share.eu/repository/search/

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





Comunicació

Ja som més de 3.100!

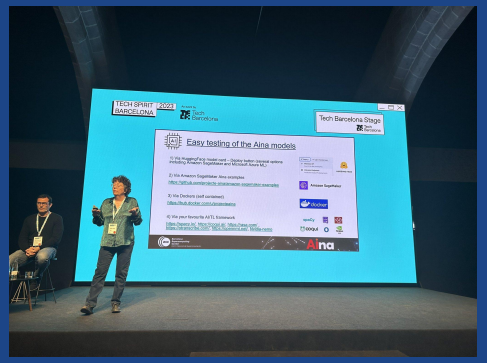
@projecte_aina

\+ **240 seguidors**

Projecte Aina

ESDEVENIMENTS

Presentació al Smart City Expo World Congress 2023



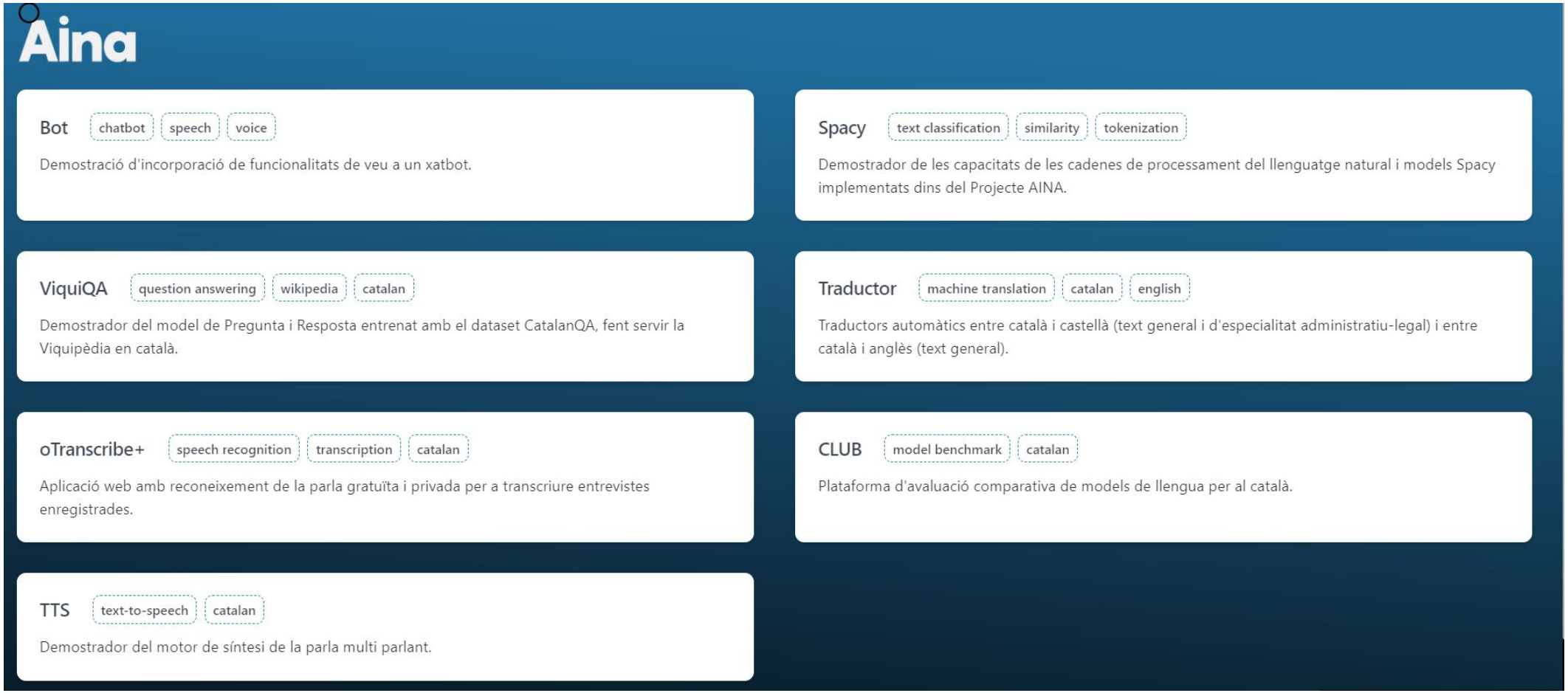
Presentació al Tech Spirit Barcelona 2023

WEB ACTUALITZADA, JA DISPONIBLE!

https://projecteaina.cat/tech/

Prototips i demostradors

https://aina.bsc.es/



Aina Open Innovation Contest 2024

Accelerating the ecosystem on 2024

- Aina "Open Innovation Contest" (Q1 2024) will set in motion the developers community around the available open resources

  ☑ Technical validation

  ☑ Social impact evidence

  ☑ Ecosystem development

Aina Open Innovation Contest 2024

2) Dedicated site for “expressions of interest” for open innovation contest

- https://twitter.com/projecte_aina
- https://www.linkedin.com/company/projecte-aina/

Prototips i demostradors

Models Fundacionals i Instruïts

Demostració dels models generatius FLOR basats en BLOOM, tant el model fundacional com el prototip instruït, amb la possibilitat de modificar els paràmetres de generació.

Model Fundacional:

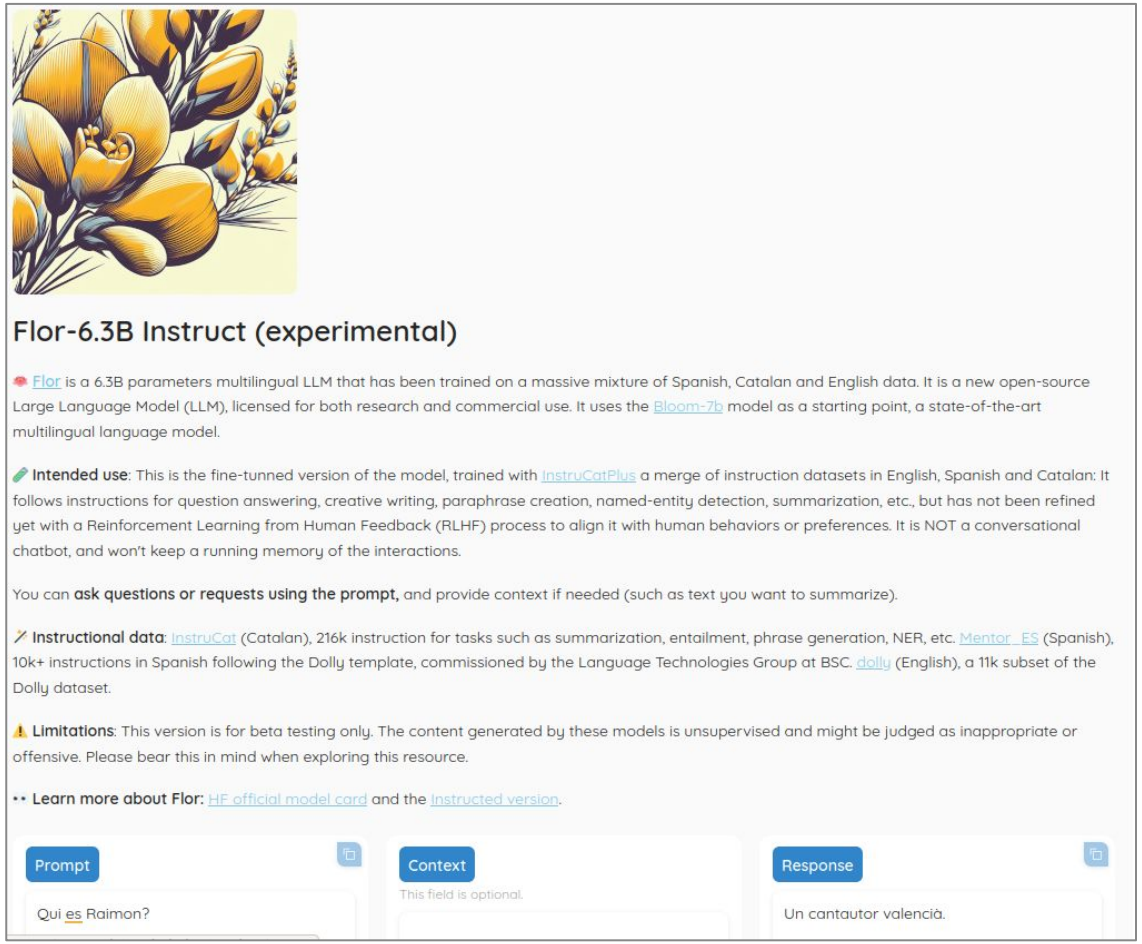
https://huggingface.co/spaces/projecte-aina/flor-6.3b-inference

Model Instruït:

https://huggingface.co/spaces/projecte-aina/flor6.3b-instruct

Model Instruït Lite (1.3b, corre sense GPUs):

https://huggingface.co/spaces/projecte-aina/flor1.3b-instruct



Prototips i demostradors

PoC: Assistent d'Informació Meteorològica (amb la col·laboració de la CCMA)

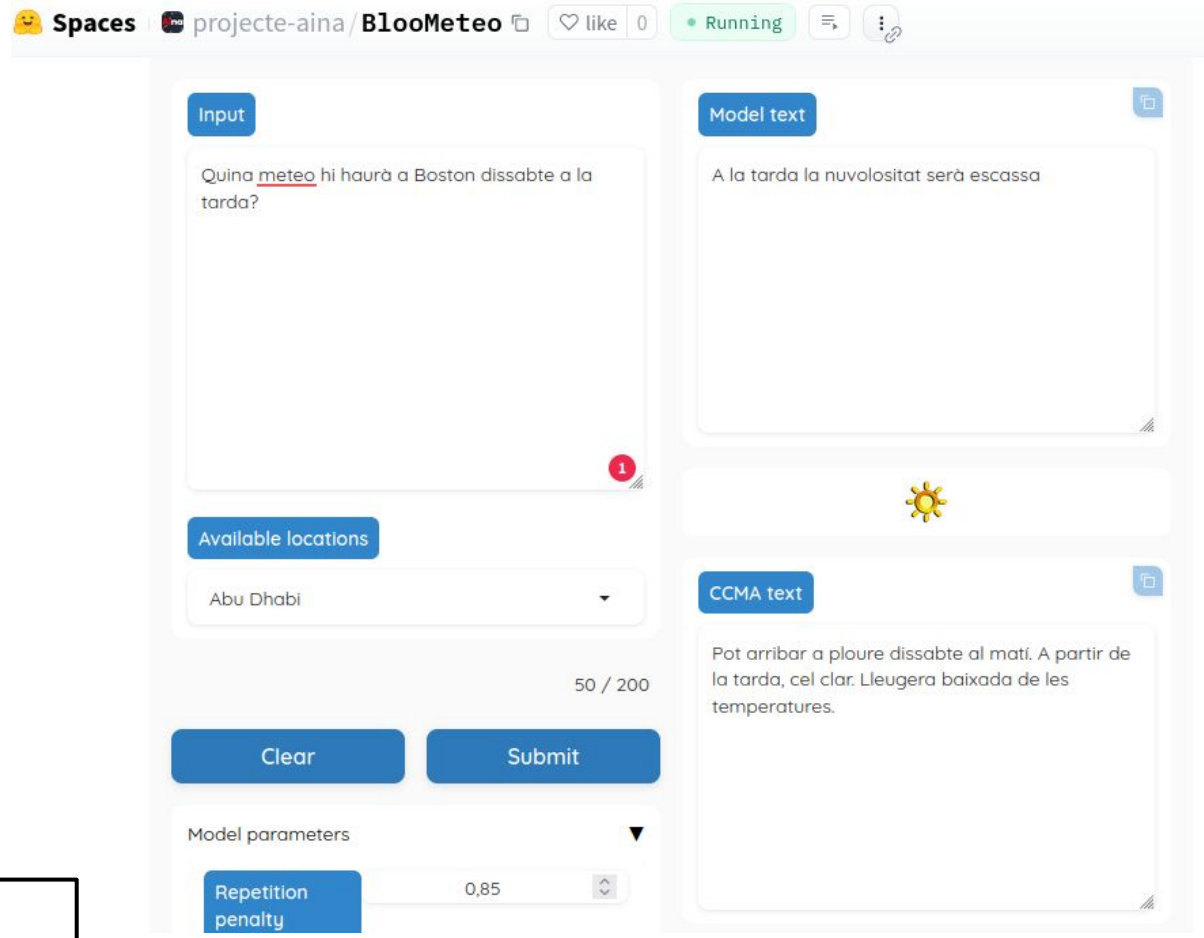
Demostració de generació automàtica de prediccions meteorològiques locals amb un assistent interactiu que fa servir una versió lleugera del model Flor1.3, amb col·laboració amb els meteoròlegs de la CCMA.

https://huggingface.co/spaces/projecte-aina/BlooMeteo

DeBERTA MultiNER (*Multilingüe*):

Demostració d'Anotació i Classificació d'Entitats en contextos multilingües, amb models massius DeBERTA, i el dataset CEIL.

https://huggingface.co/spaces/projecte-aina/multiner_demo



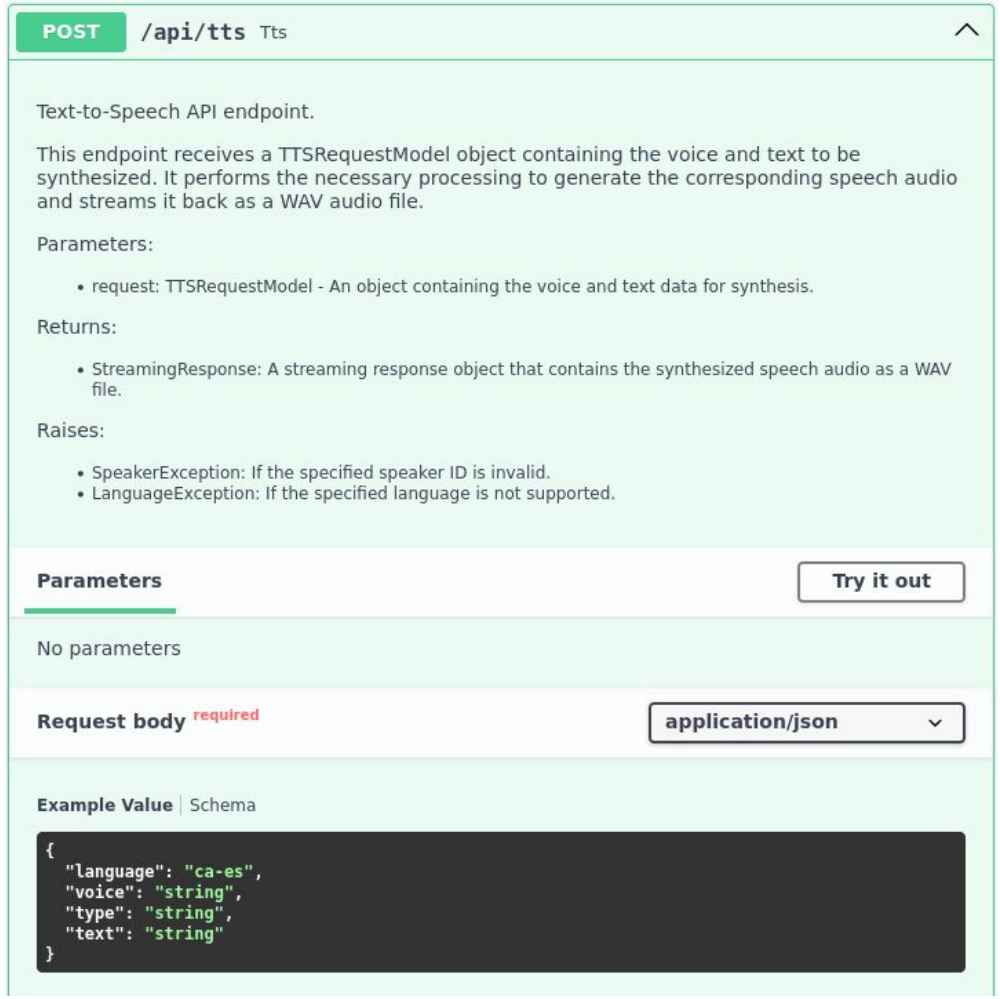
Aviat, demo de **RAG (Retrieval-Augumented Generation)** per fer cerques d'informació en repositoris massius de documents fent preguntes en llenguatge natural

Prototips i demostradors

Síntesi de la parla en català (TTS)

El nostre nou model de TTS està considerablement millorat sobretot per la fiabilitat de pronunciació, i la qualitat d'àudio. Estem col·laborant amb Softcatalà i la CCMA per millorar. El nou model està integrat al nostre software tts-api, per facilitar les integracions a la producció. El software facilita l'ús dels models amb un millor rendiment.

- TTS-API: https://hub.docker.com/r/projecteaina/tts-api
- Demo TTS: https://huggingface.co/spaces/projecte-aina/tts-ca-coqui-vits-multispeaker
- Demo conversacional: https://bot.aina.bsc.es/



Identificació dels parlants

Vam desenvolupar un software que processa un àudio llarg per identificar qui està parlant quan. Utilitza enregistraments de referència per detectar la identitat de la persona, i relacionar-ho amb les pautes en que parla. El software està dissenyat per l'ús de l'administració pública. Aviat publicarem una versió genèrica del codí per l'ús de tothom.

L'horitzó d'Aina: estatus del projecte

Actualització sobre el desenvolupament dels models lingüístics, recursos i objectius d'AINA

19 de desembre del 2023

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Finançat per Generalitat de Catalunya

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Aina