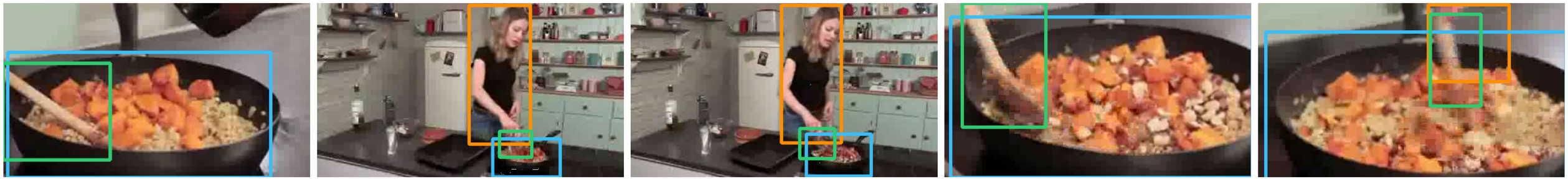

Artificial Intelligence for Science

ICCV HONOLULU HAWAII
OCT 19-23, 2025

Large-scale Pre-training for GROunded Video caption gEneration

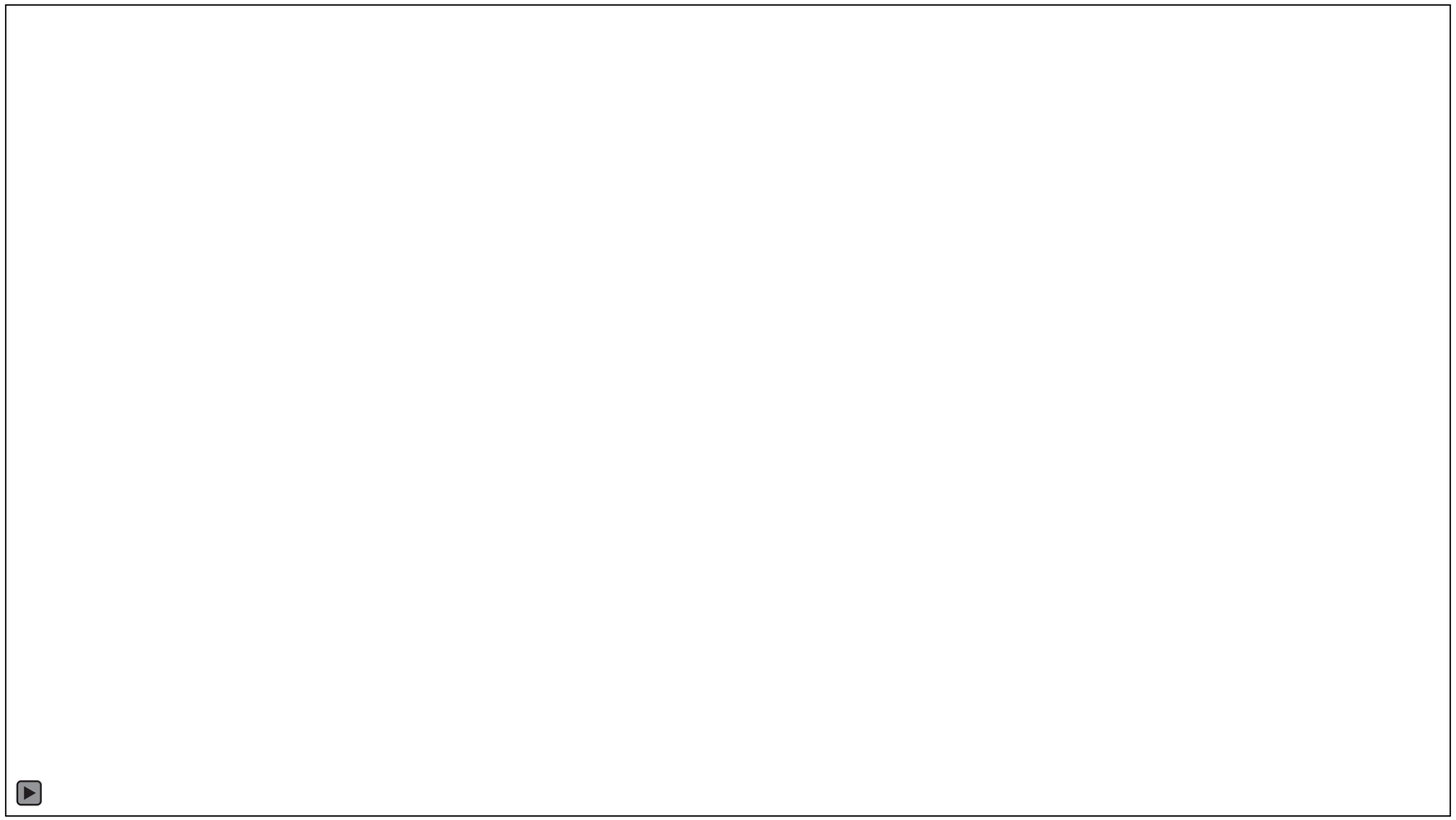


Evangelos Kazakos, Cordelia Schmid, Josef Sivic

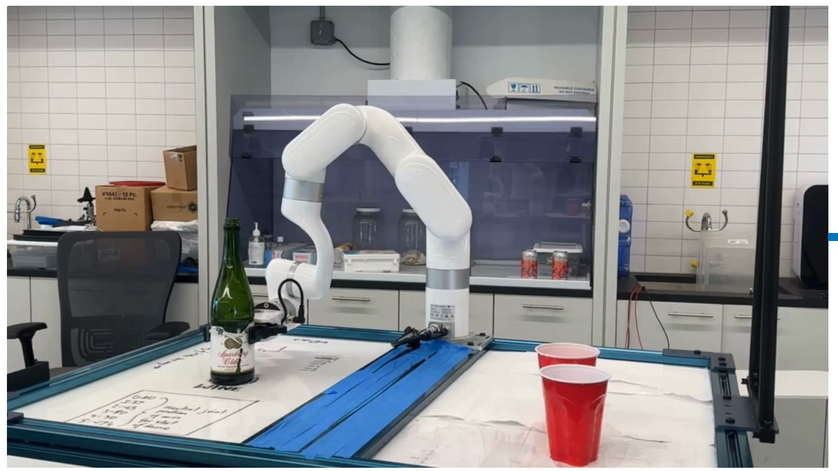
ENS | PSL

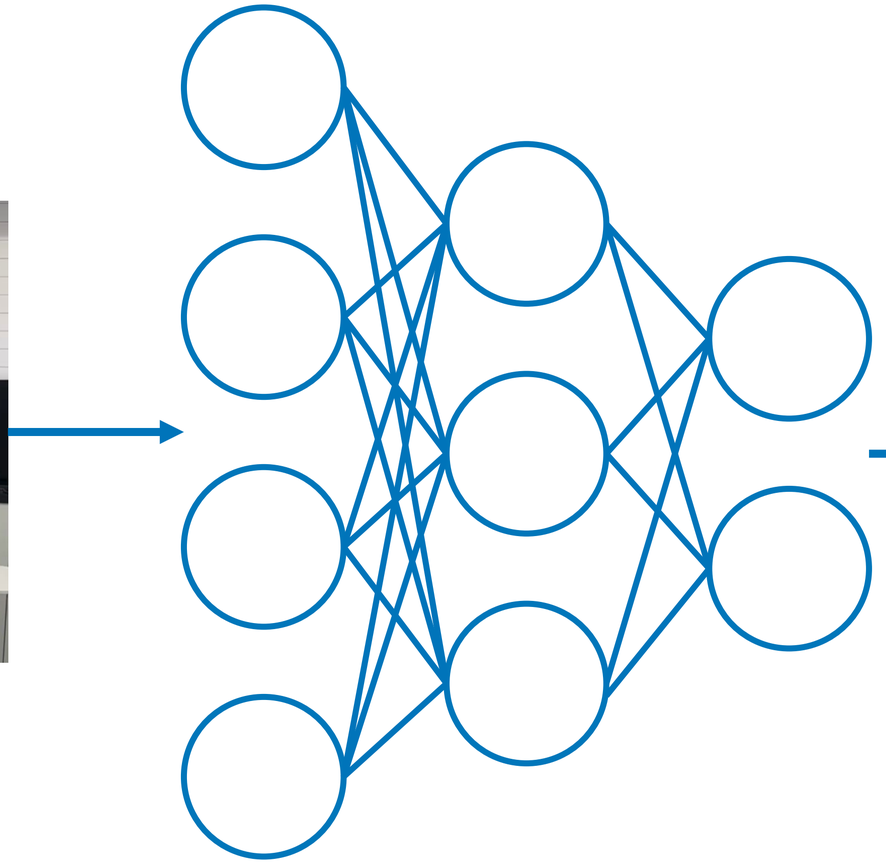
Inria

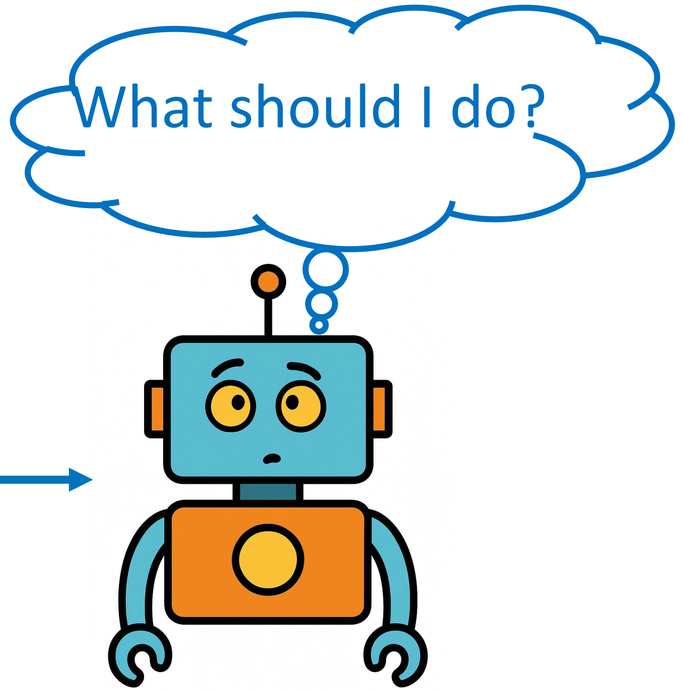
Motivation



Motivation

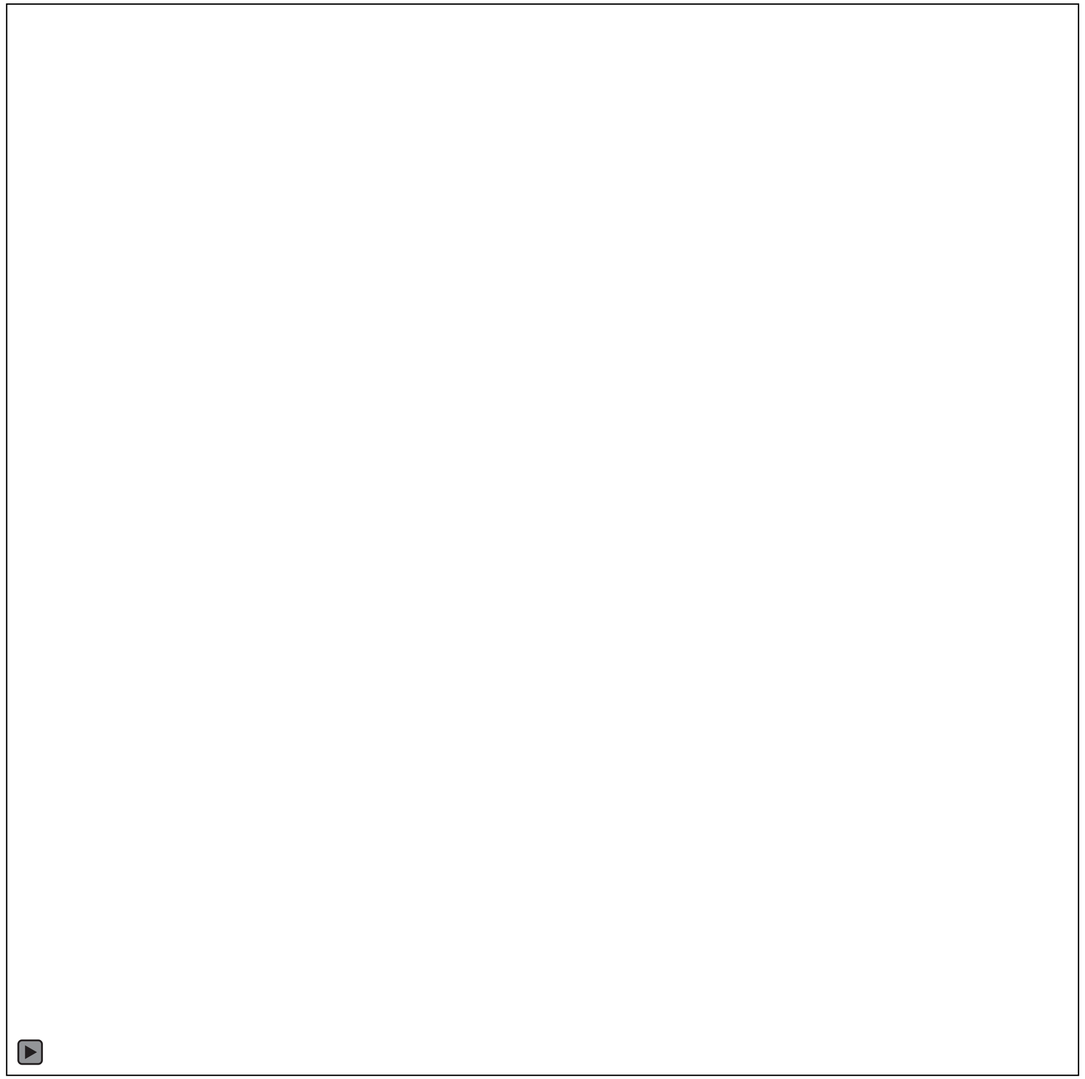




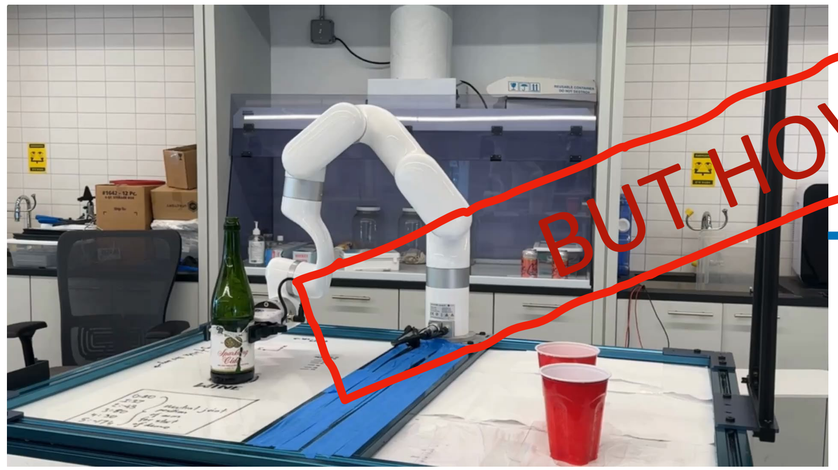


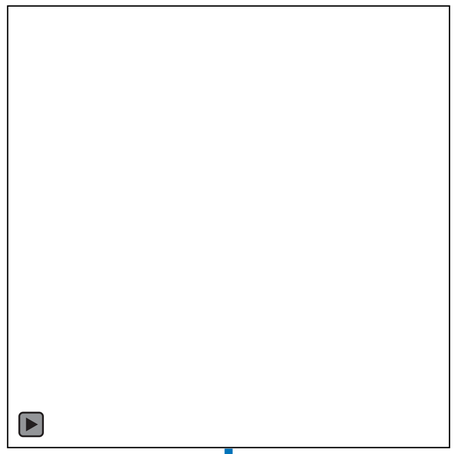
Motivation

Learn by demonstration using
Internet videos at scale

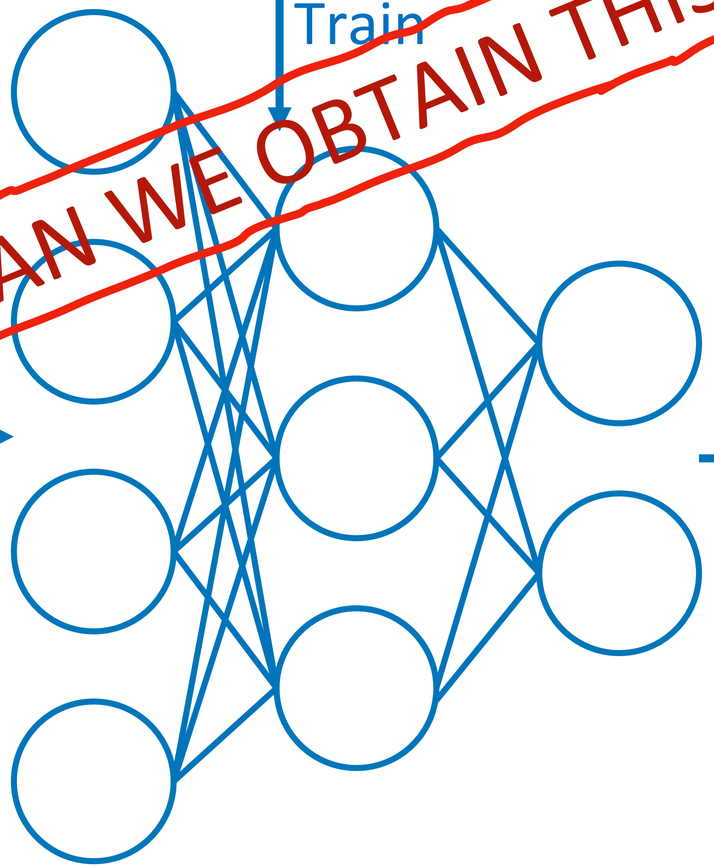


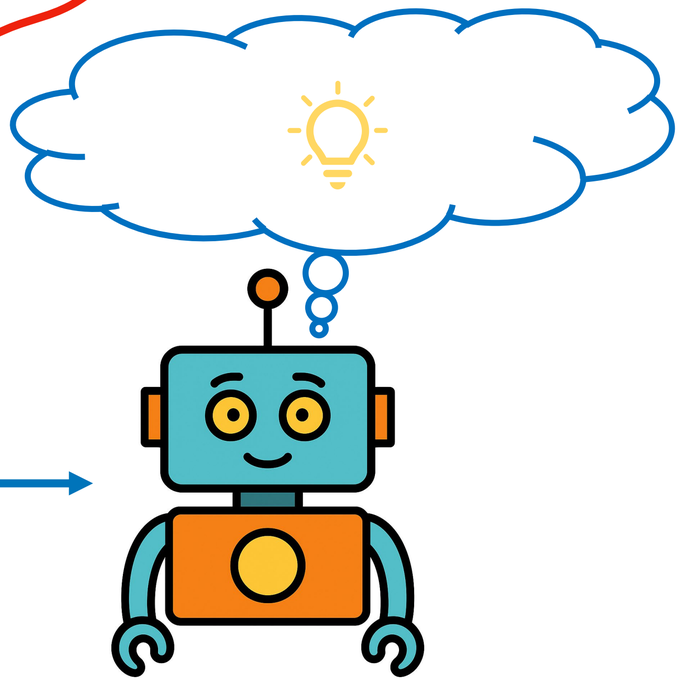
Motivation





Train





BUT HOW CAN WE OBTAIN THIS DATA??

Grounded Video Caption Generation

- Input: video
- Sub-task 1: captioning
- Sub-task 2: Identify noun phrases
- Sub-task 3: Grounding

The HowToGround1M pre-training dataset

- Automatically annotated
- 1M videos
- 80.1M bounding boxes
- Suitable for pre-training

The GROVE model

- **Step 1**: LLM predicts caption and noun phrases locations
- **Step 2**: Decoder grounds noun phrases to bounding boxes
- Adapters for spatio-temporal modelling
- Temporal objectness predicts the presence of an object in a frame

The iGround dataset

- Manually annotated
- 3500 examples
- Train/val/test: 2000/500/1000
- Suitable for fine-tuning and evaluation

Comparison with SOTA

	Method	METEOR	CIDER	AP50	Recall
Center	a. GLaMM [31]	11.9	29.9	20.8	19.3
	b. GROVE - PT (Ours)	14.3	50.6	27.0	22.5
	c. GROVE - PT+FT (Ours)	**21.4**	**83.5**	**31.7**	**26.2**
All	d. Automatic annotation	13.8	40.0	27.1	20.4
	e. GROVE - PT (Ours)	14.3	50.6	33.6	24.3
	f. GROVE - FT (Ours)	21.0	77.7	15.8	18.1
	g. GROVE - PT+FT (Ours)	**21.4**	**83.5**	**40.0**	**28.7**

Table 2. **Grounded video caption generation** on manually-annotated iGround test set. Pre-training on our new large-scale HowToGround1M dataset followed by finetuning on manually-annotated iGround training data (PT+FT) clearly outperforms pre-training only (PT) and finetuning only (FT) as well as the GLaMM baseline [31] (a.) and directly applying automatic annotation (d.). We show center frame ("Center") and all frame ("All") evaluation.

Method	m_{sIoU}
STVGBert [36]	47.3
TubeDETR [43]	59.0
STCAT [14]	61.7
DenseVOC [51]	61.9
GROVE FT (Ours)	61.3
GROVE PT+FT (Ours)	**63.7**

Table 3. State-of-the-art comparison of spatial grounding on the VidSTG [48] test set (declarative sentences). All models use ground truth temporal localization. Large-scale pretraining (PT+FT) results in an improvement over fine-tuning only (FT) for our model GROVE.

Method	FT	m_{sIoU}
PG-V-L (13B) [25]	✗	35.1
GLaMM [31] + SAM2 [32]	✗	38.6
GROVE	✗	**43.0**
VideoGLaMM[26]	✓	39.7
GROVE	✓	**55.5**

Table 4. State-of-the-art comparison of spatial grounding on the VidSTG [48] test set (interrogative sentences). All models use ground truth temporal localization. GROVE outperforms all competitors both in a pre-training only setting (✗) and when fine-tuned on VidSTG (✓).

Method	$F1_{all}$	$F1_{all\_per\_sent}$	$F1_{loc}$	$F1_{loc\_per\_sent}$
GVD [50]	07.10	17.30	23.80	59.20
GROVE FT (Ours)	09.51	21.15	30.96	68.79
GROVE PT+FT (Ours)	**13.39**	**24.08**	**45.04**	**77.29**

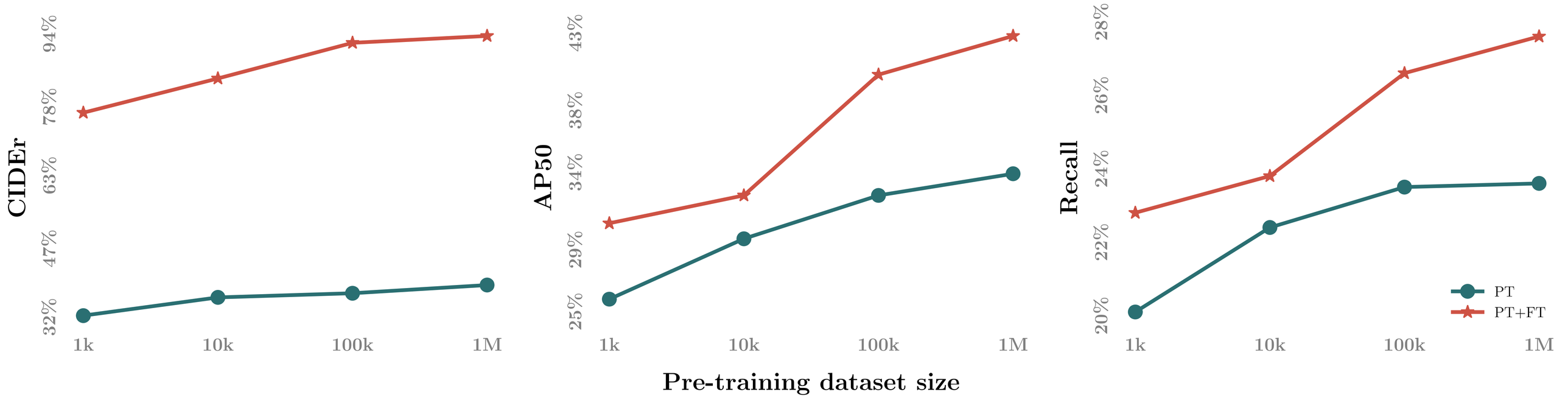
Table 5. Results on the validation set of ActivityNet-Entities [50]. Large-scale pretraining (PT+FT) results in an improvement over fine-tuning only (FT) for our model GROVE.

Method	**YouCook-Interactions**	**GroundingYouTube**
What When and Where (S3D) [4]	53.98	60.62
What When and Where (CLIP) [4]	58.35	56.98
GROVE	**68.67**	**72.14**

Table 8. Comparison with SOTA on YouCook-Interactions [37] and GroundgYouTube[4] datasets.

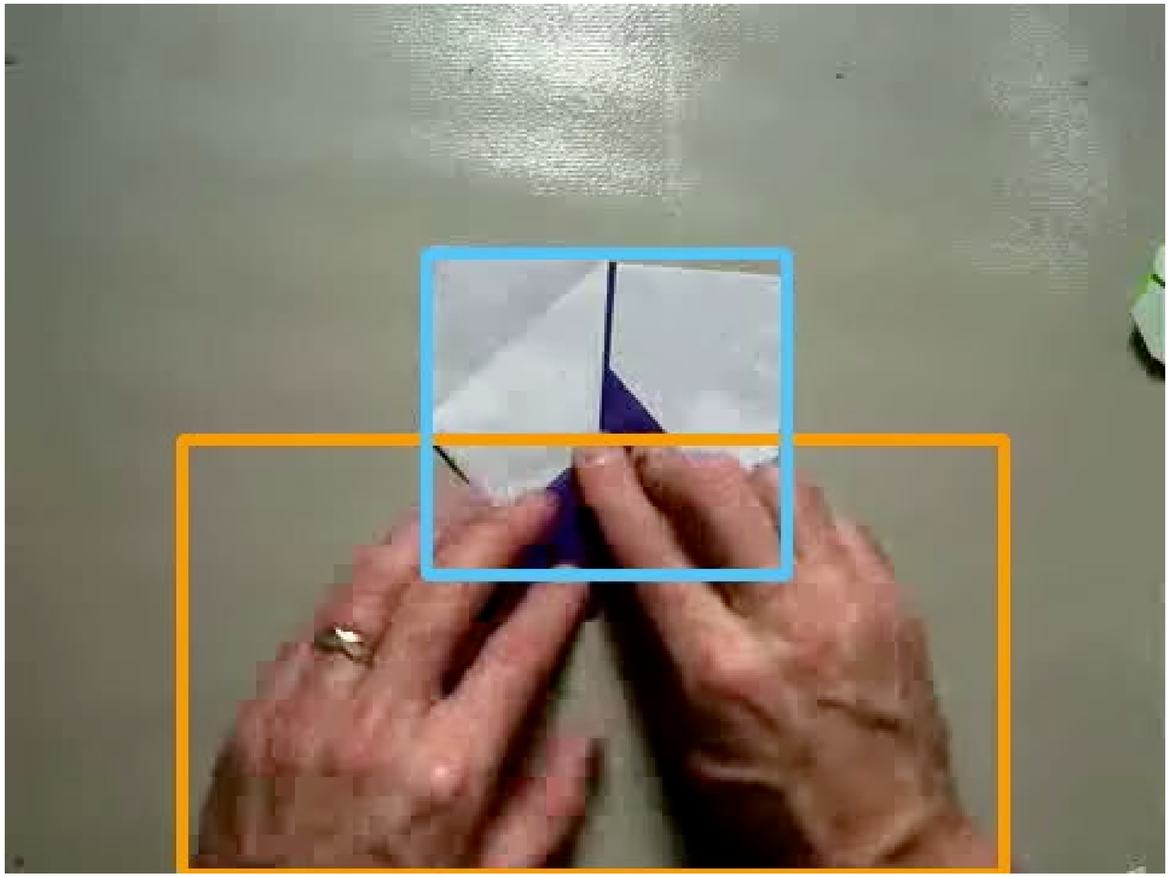
State-of-the-art results in 5 datasets

Scaling behaviour



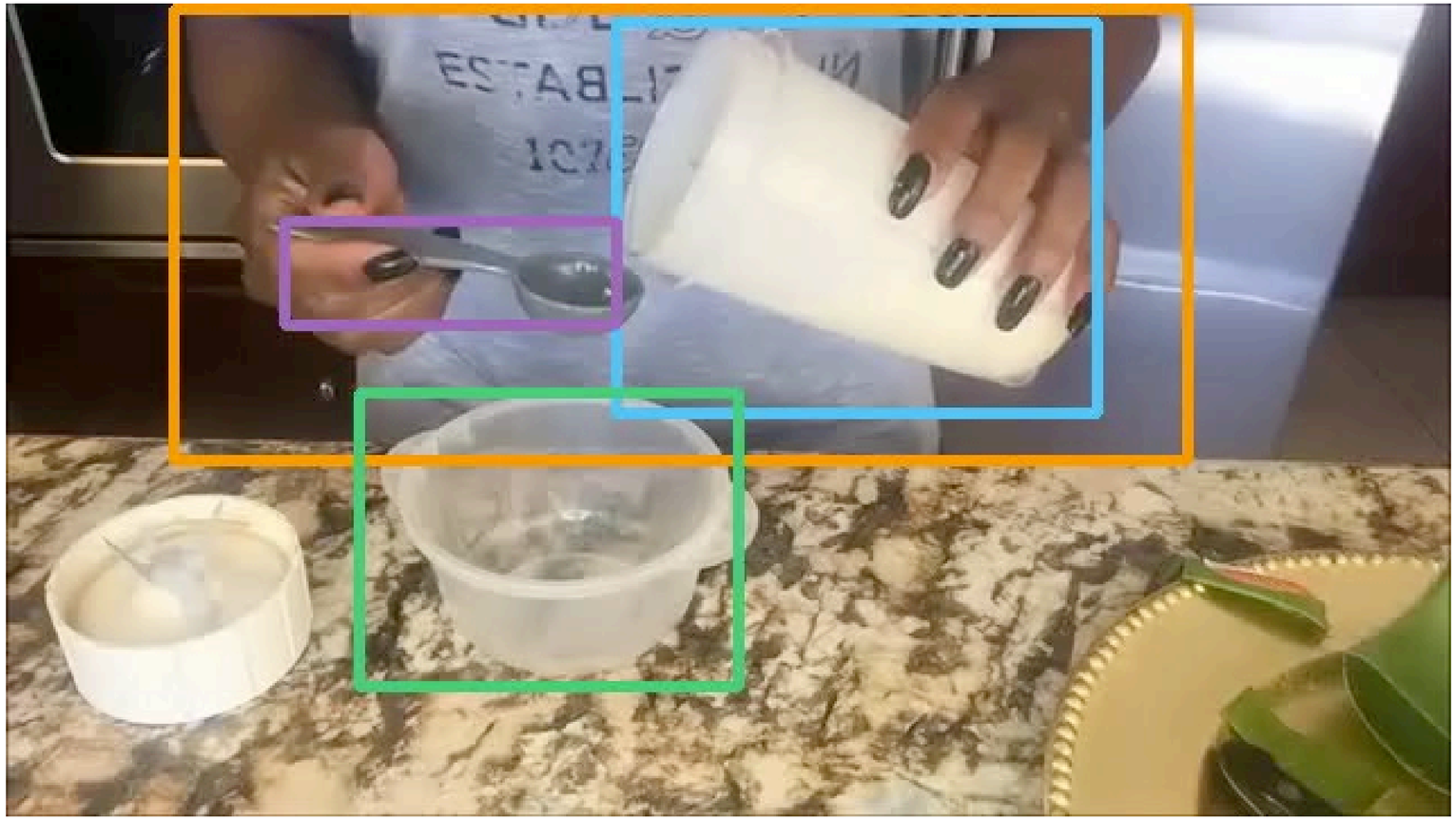
Performance scales with the size of the pre-training dataset

Qualitative results

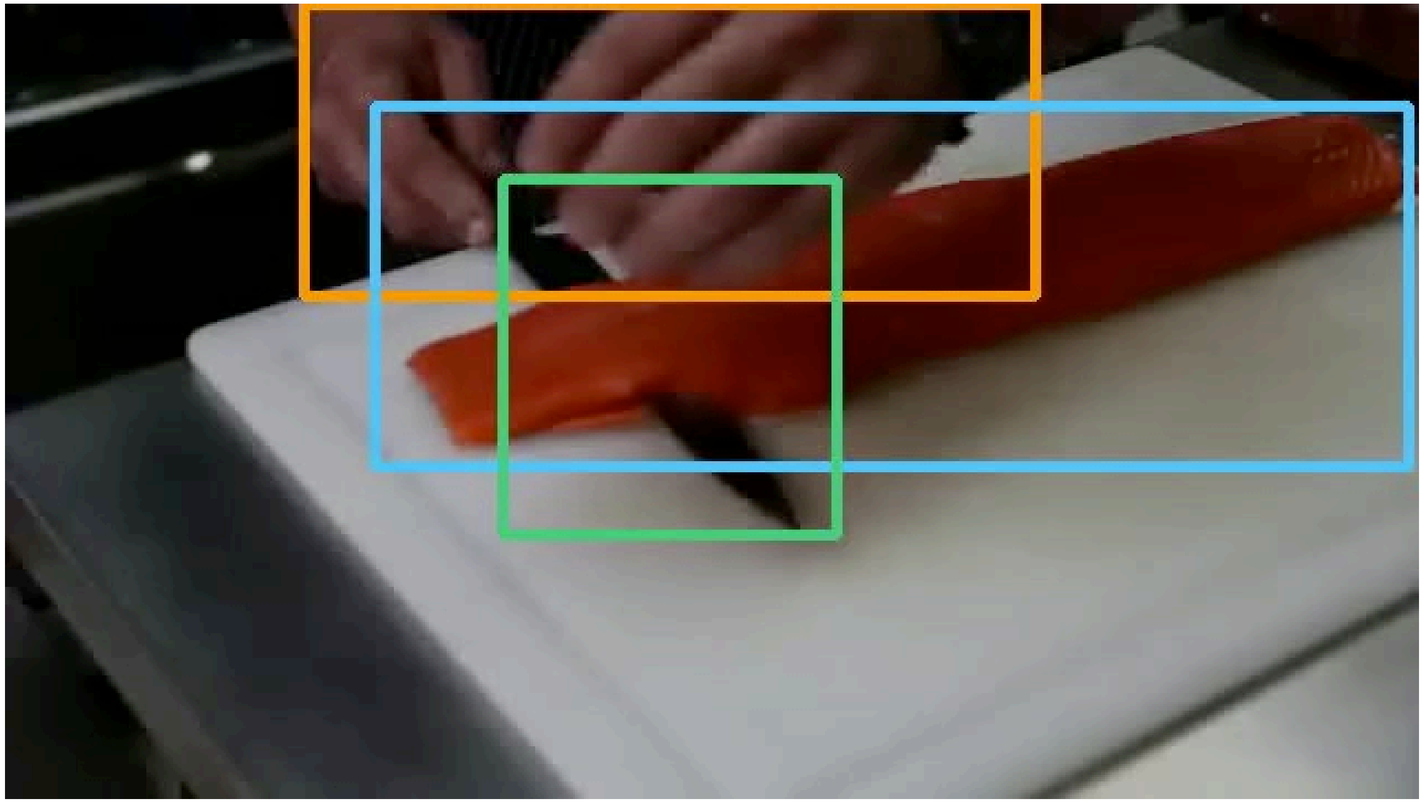


Qualitative results

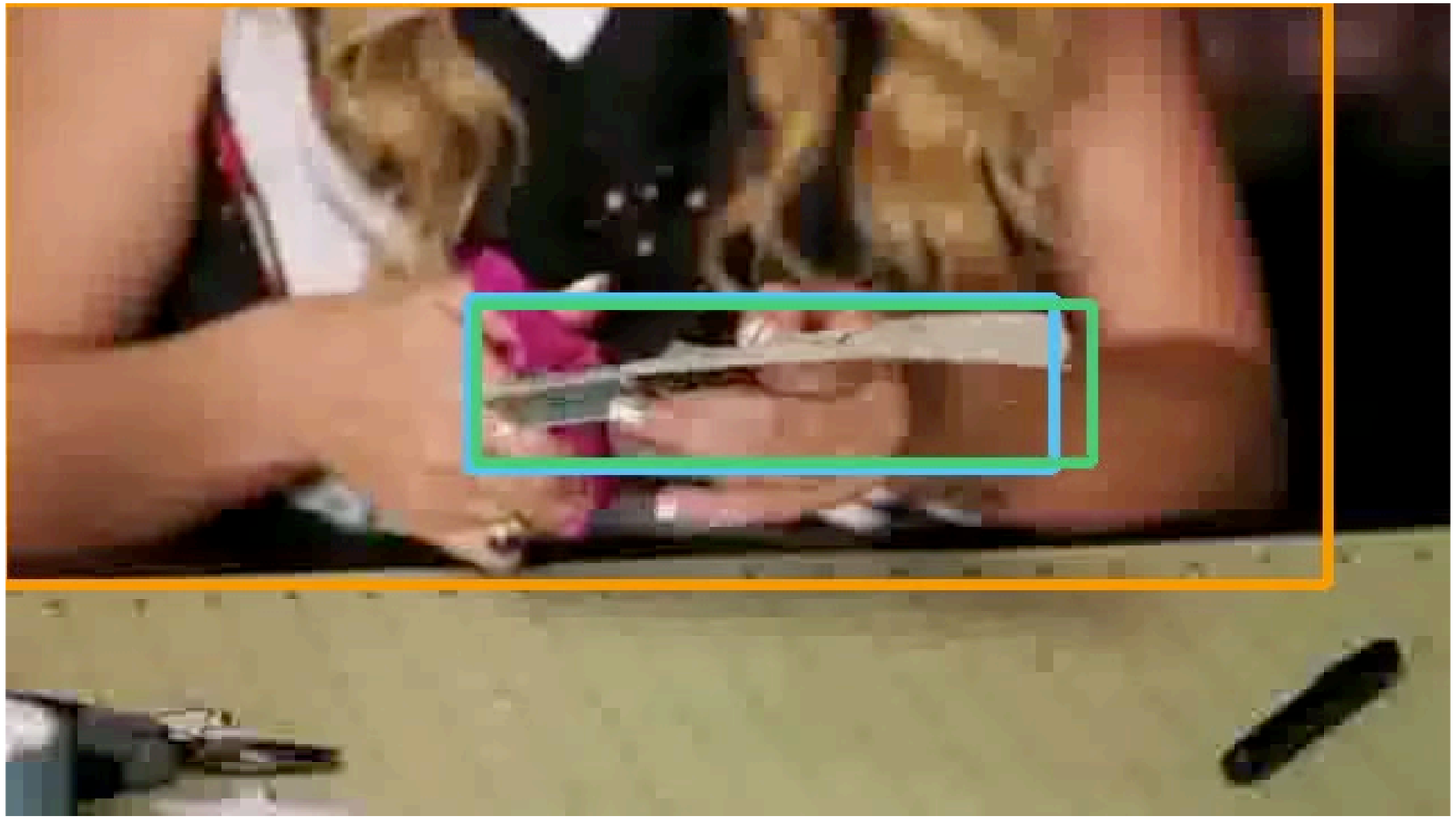
Qualitative results



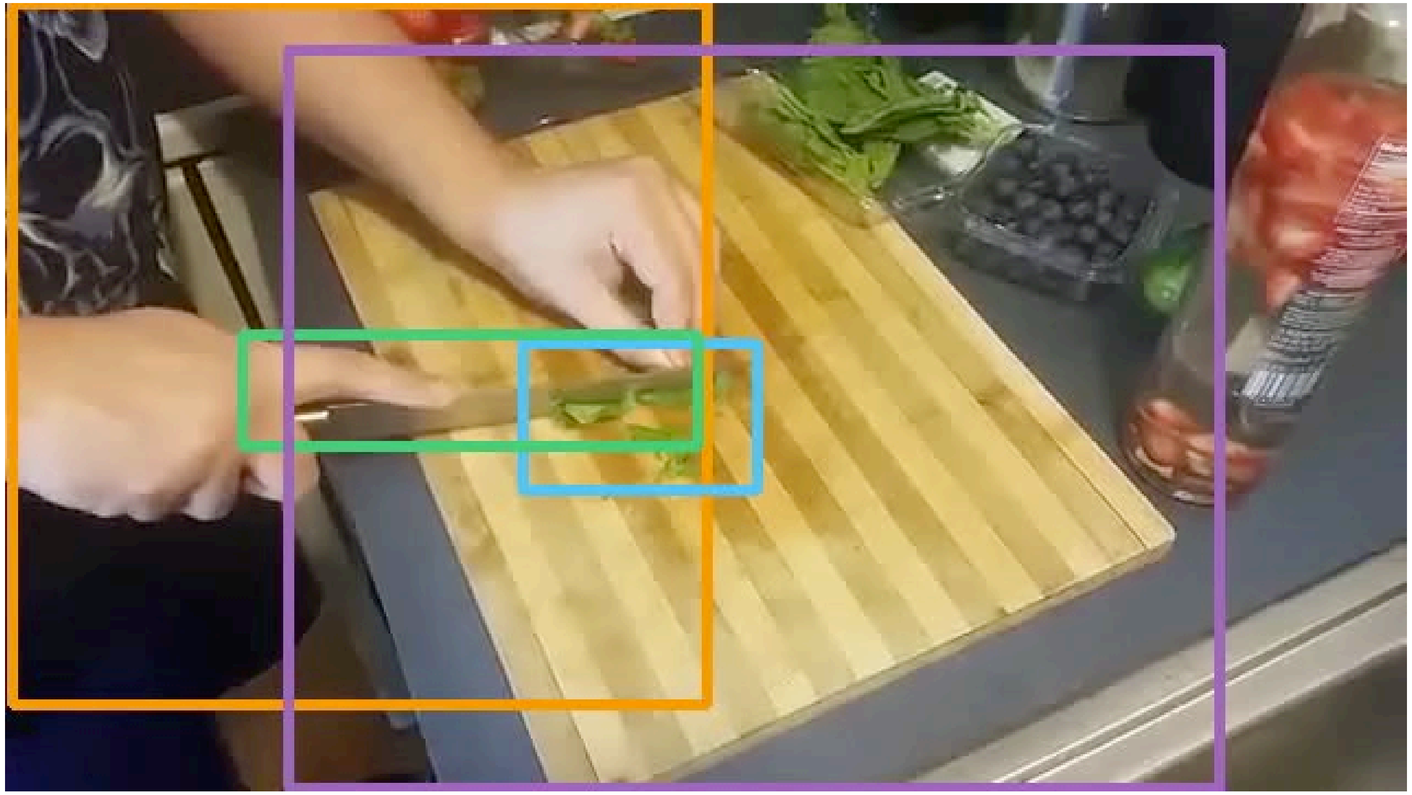
Qualitative results



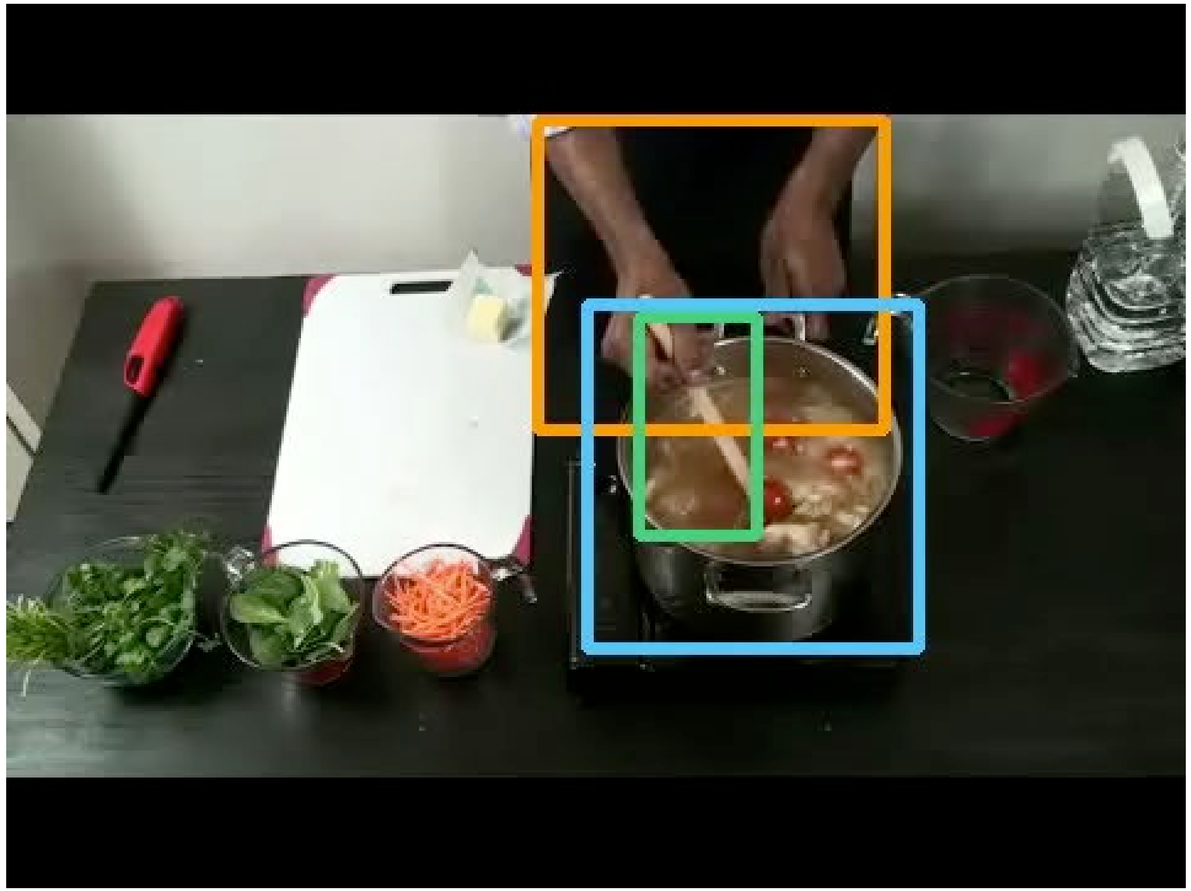
Qualitative results



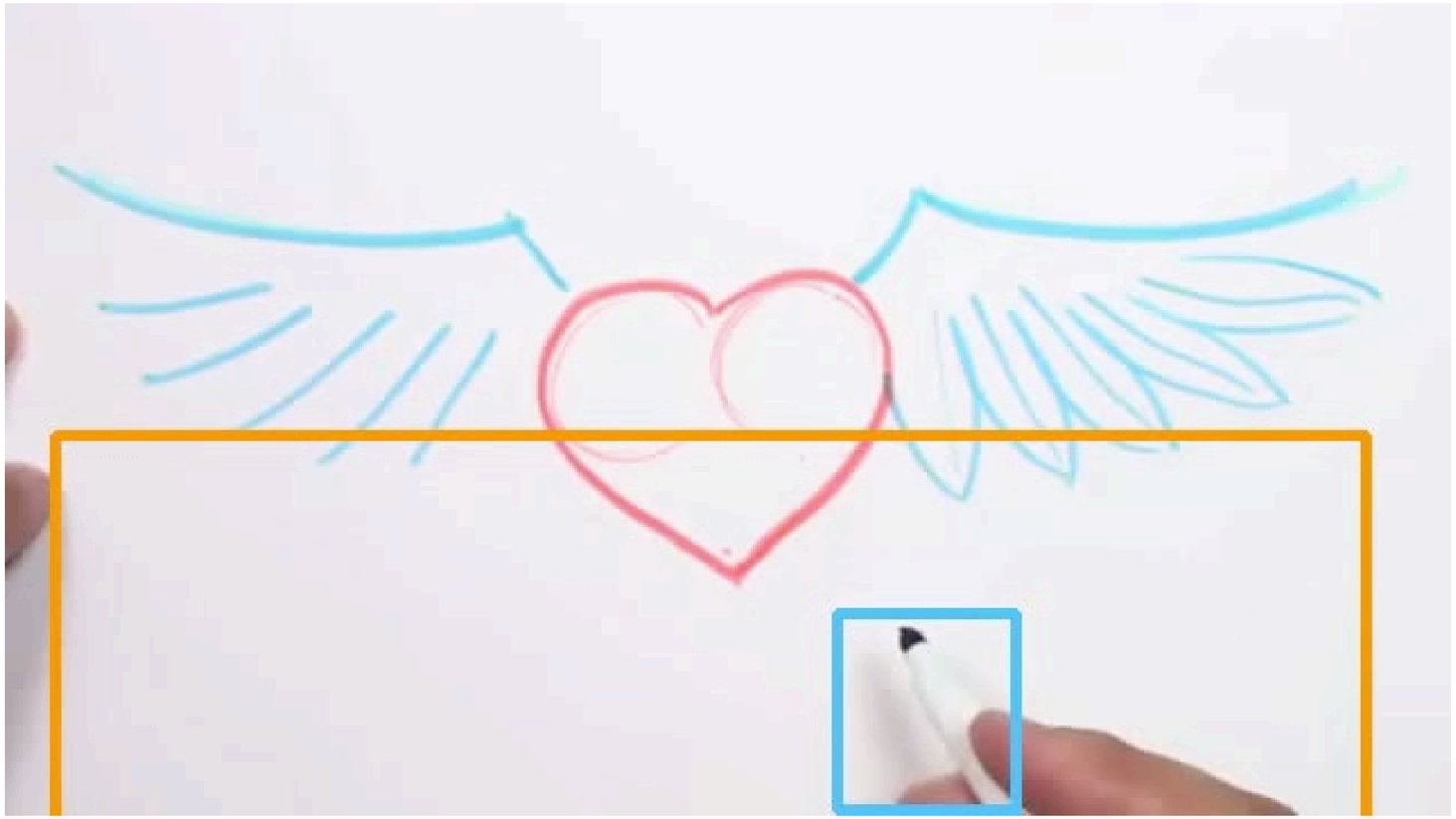
Qualitative results



Qualitative results



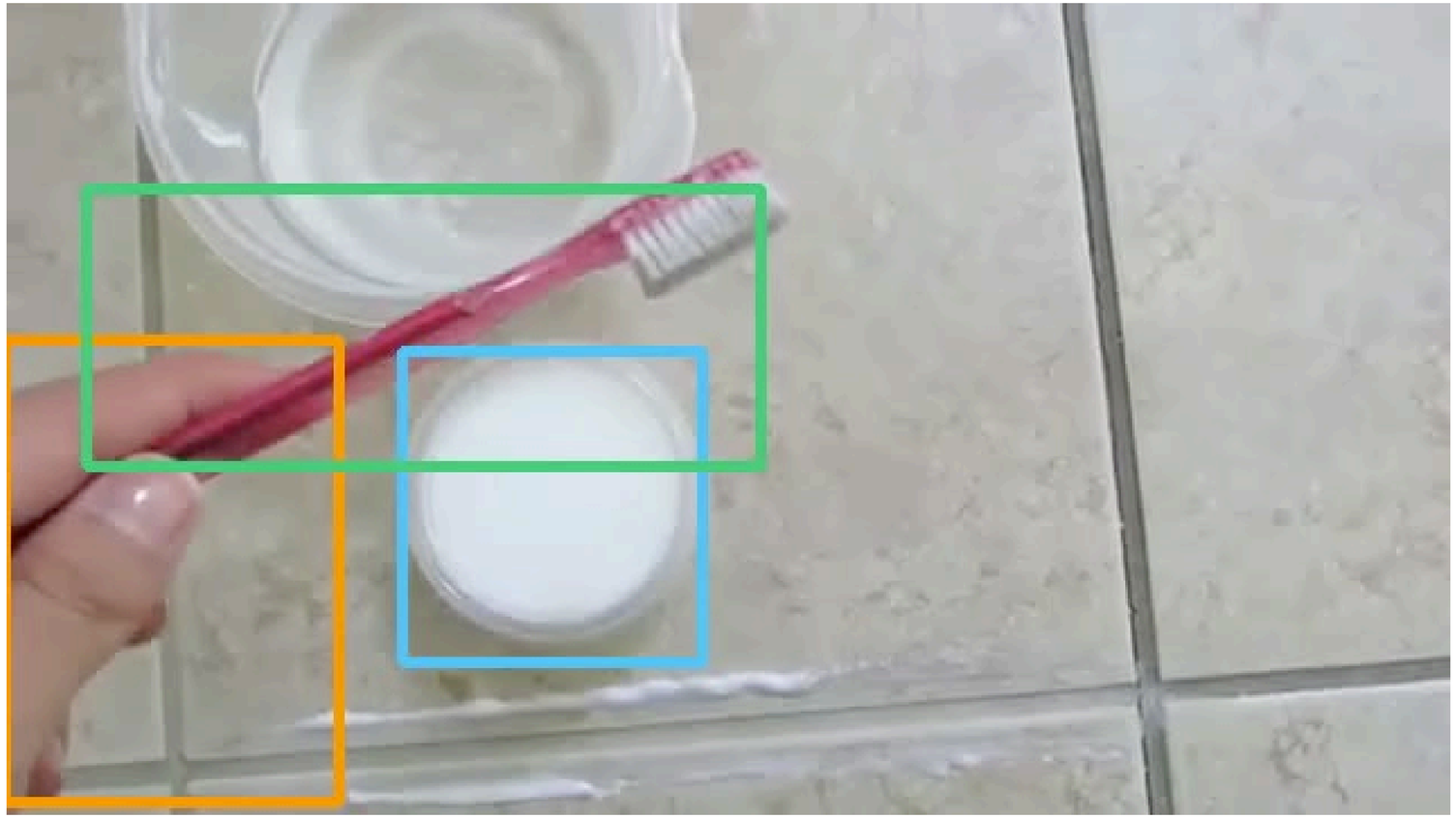
Qualitative results



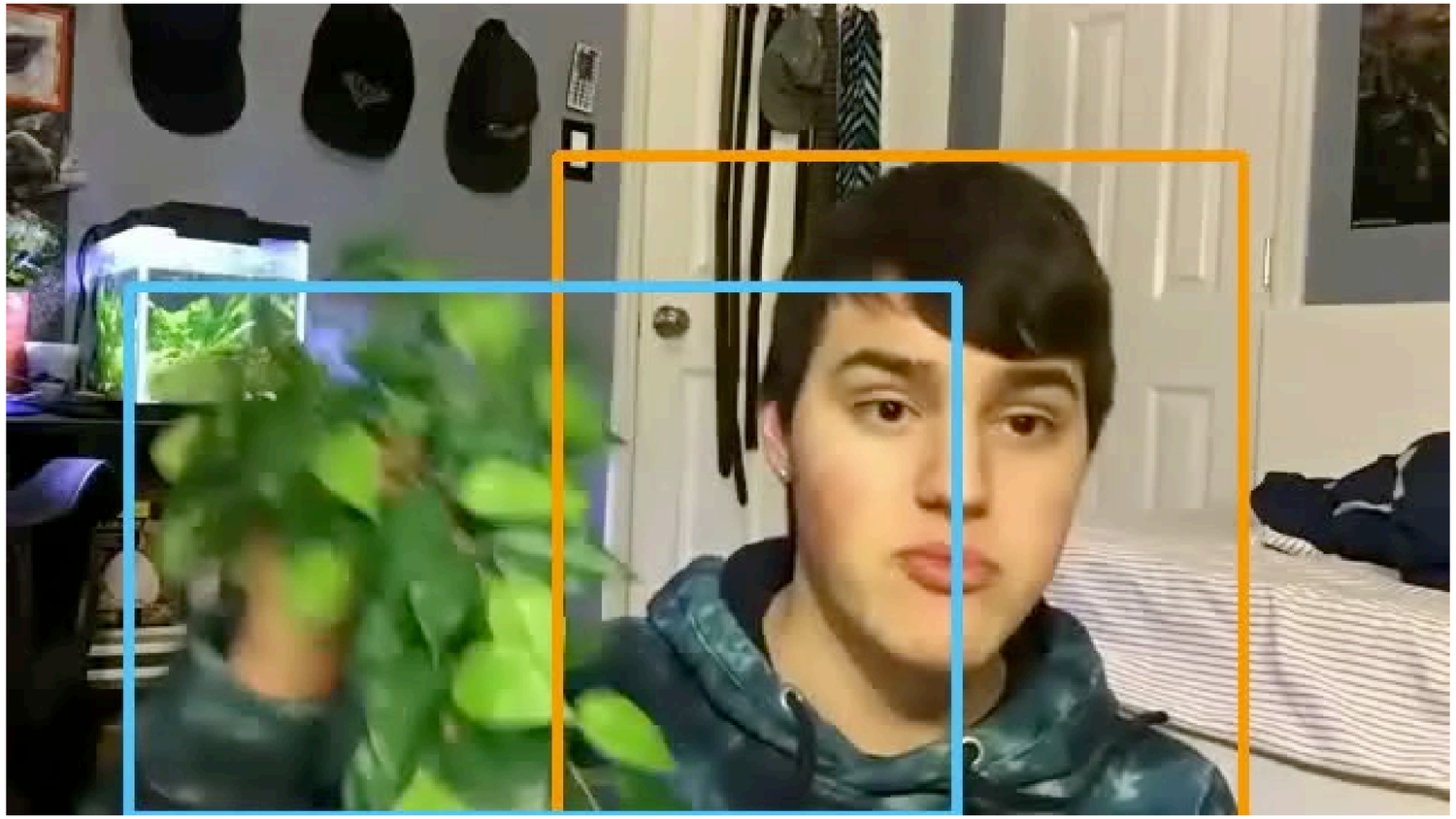
Qualitative results

Qualitative results

Qualitative results



Qualitative results



Thank you!!!

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ALIA Brings Linguistic Diversity: Optimizing Data Distribution for Multilingual NLP Inclusivity

Alexander Shvets

Acknowledgments: This research was supported by the EuroHPC Extreme Scale Access grant EHPC-E01-009 and funded by the Project Desarrollo de Modelos ALIA with the framework of the Plan Nacional de Tecnologías de Lenguaje -ENIA 2024 and PRTR, NextGeneration EU, Resol. SEDIA 19.08.2024.

Language Technologies Laboratory

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Language Technologies Laboratory - Projects

Language Technologies Laboratory

Language Technologies Laboratory

>60 researchers
1 spin-off

Aina

ILENIA
IMPULSO DE LAS LENGUAS
LA INTELIGENCIA ARTIFICIAL

ALIA
El modelo fundacional abierto de IA generativa desarrollado en España

LA INFRAESTRUCTURA PÚBLICA DE IA EN CASTELLANO Y LENGUAS COOFICIALES

EU
OPEN EURO LLM
LLMs4EU

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Supporting projects

Aina ILENIA ALIA

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

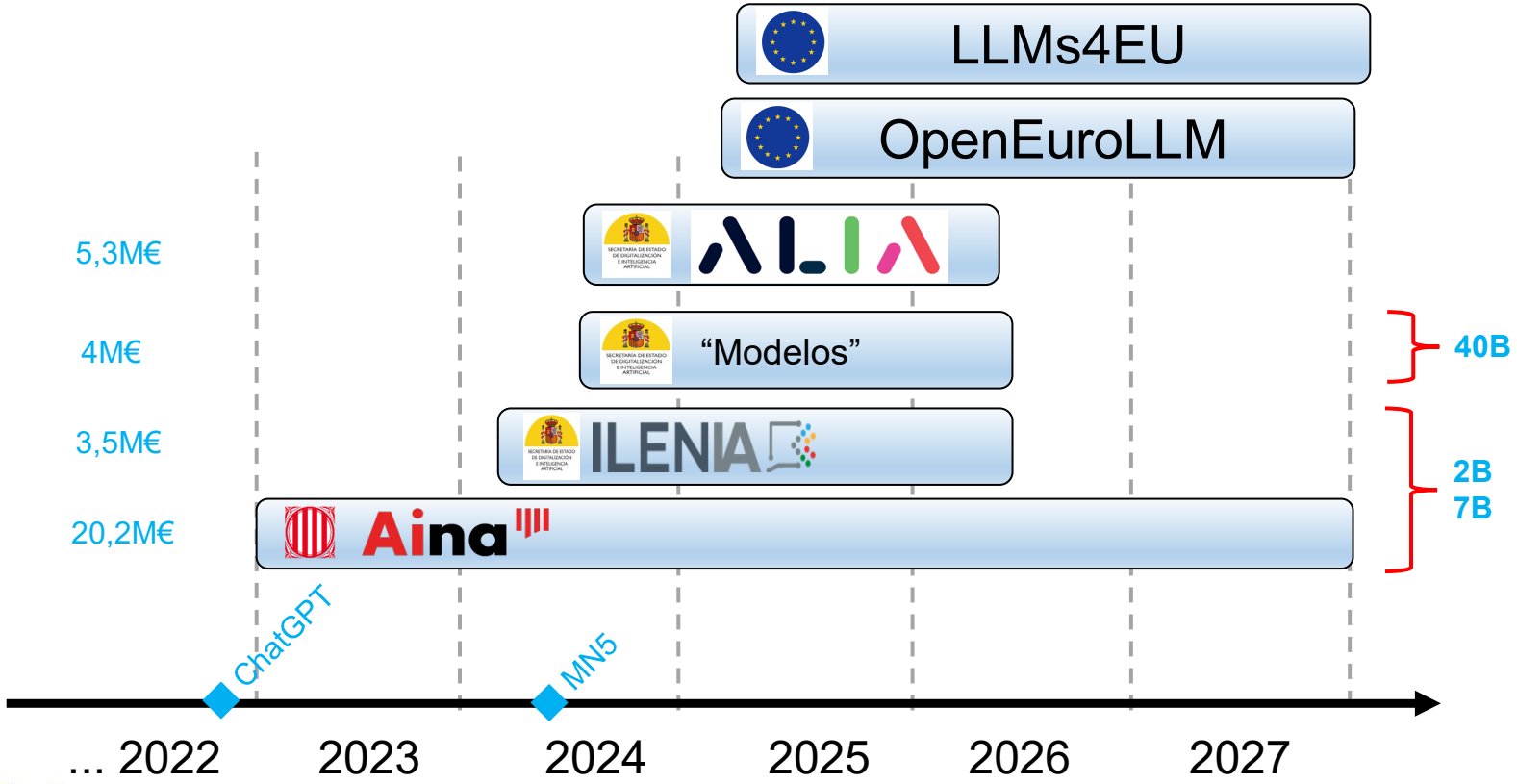
USC
Universidade de Santiago de Compostela

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

CENID

UJA Universidad de Jaén

Supporting projects



Main Goal: Become a public infrastructure

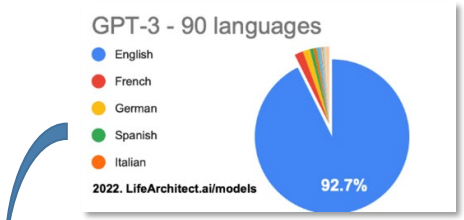
- Serves public interest
- Guarantees universal access
- Promotes accountability
- Benefits society

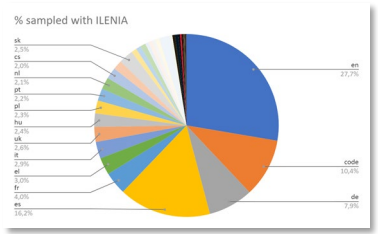
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Context and Motivation

- Large Language Models (LLMs), such as OpenAI's GPT, Google's Gemini, and Meta's LLaMA, are trained on **massive amounts of text data primarily in English**, limiting their performance in other languages.
- Evaluation results suggest that the performance of these LLMs with languages with a “weak” or “moderate” technology support, is always significantly **weaker than in English**.
- The field of generative **AI is largely dominated by American technology giants**. Europe is far behind.
- Although most LLMs are “weight open”, they are far from being truly **open source**. Often the code is not open and there is little information on training data.
- New **European AI regulations** impose levels of transparency and traceability of data that require an adequate approach.
- In response to this situation, four national projects (**AINA**, **ILENIA**, **Modelos, and ALIA**) were created to develop models and resources for the languages of Spain.

Towards truly multilingual EU LLMs





Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Training data

Language Technologies Laboratory

Data collection

- An immense effort has been made to collect data from non-internet sources.

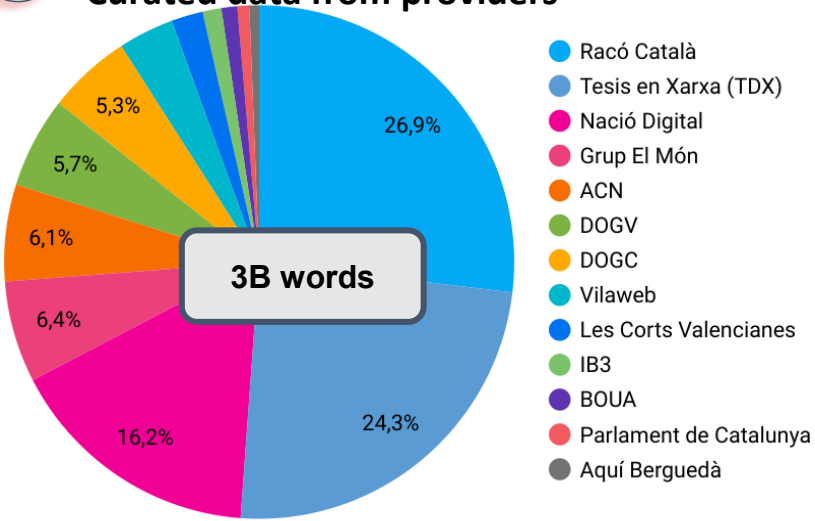
>10B de palabras para español y catalán.

Data collection

1 Web data

- CommonCrawl **130B words**

2 Curated data from providers



3 Operationalization of data supply

Wikiextractor: Text extractor from Wikipedia.

Extraction of 689,141 documents with over 266 million words in Catalan.

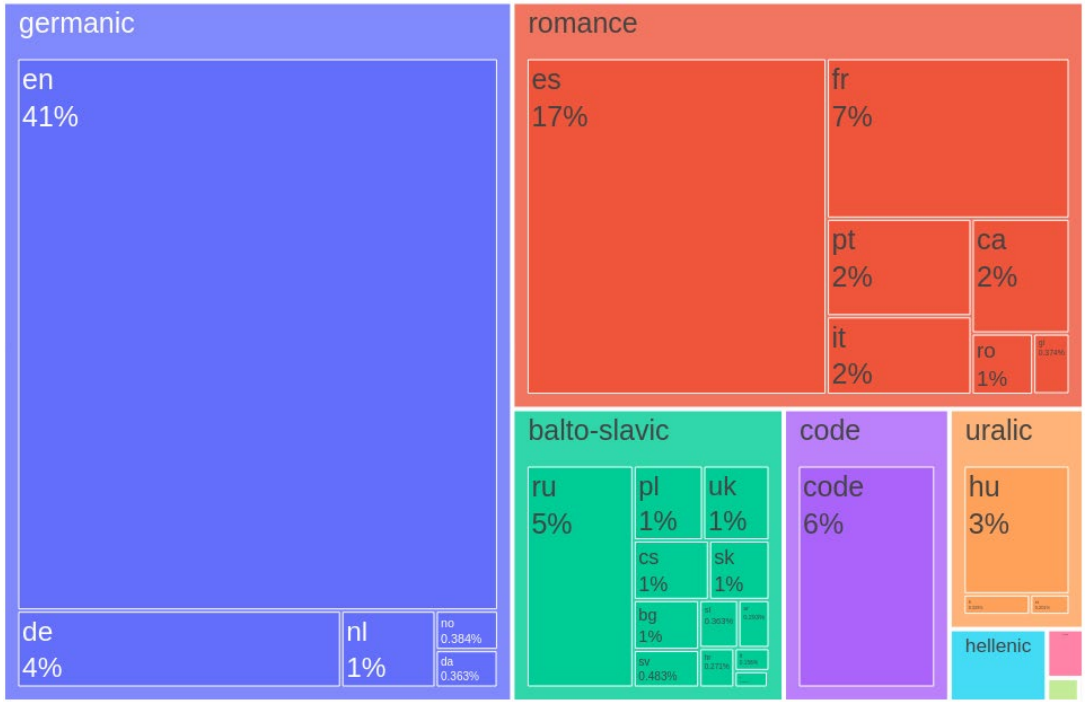
Automated pipeline that uses the Catalonia Transparency API to access data from the **DOGC**.

Extraction of 30,369 publications in Catalan with 70 million words.

Datapipe: A tool to facilitate the acquisition of audiovisual content with open licenses on the web.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Composition of final training dataset

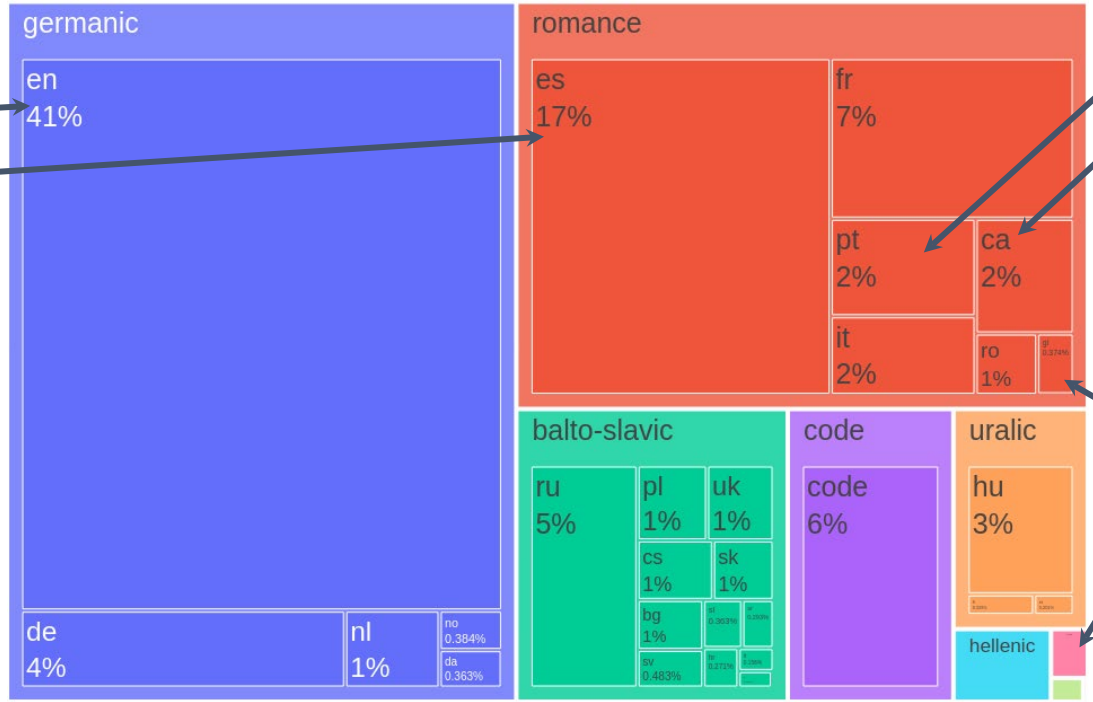


Prioritization of Romance languages (2% Catalan, 17% Spanish). Total of 2,400,000,000,000 tokens (2.4 trillions).

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Composition of final training dataset

English
and Spanish
- high-resource

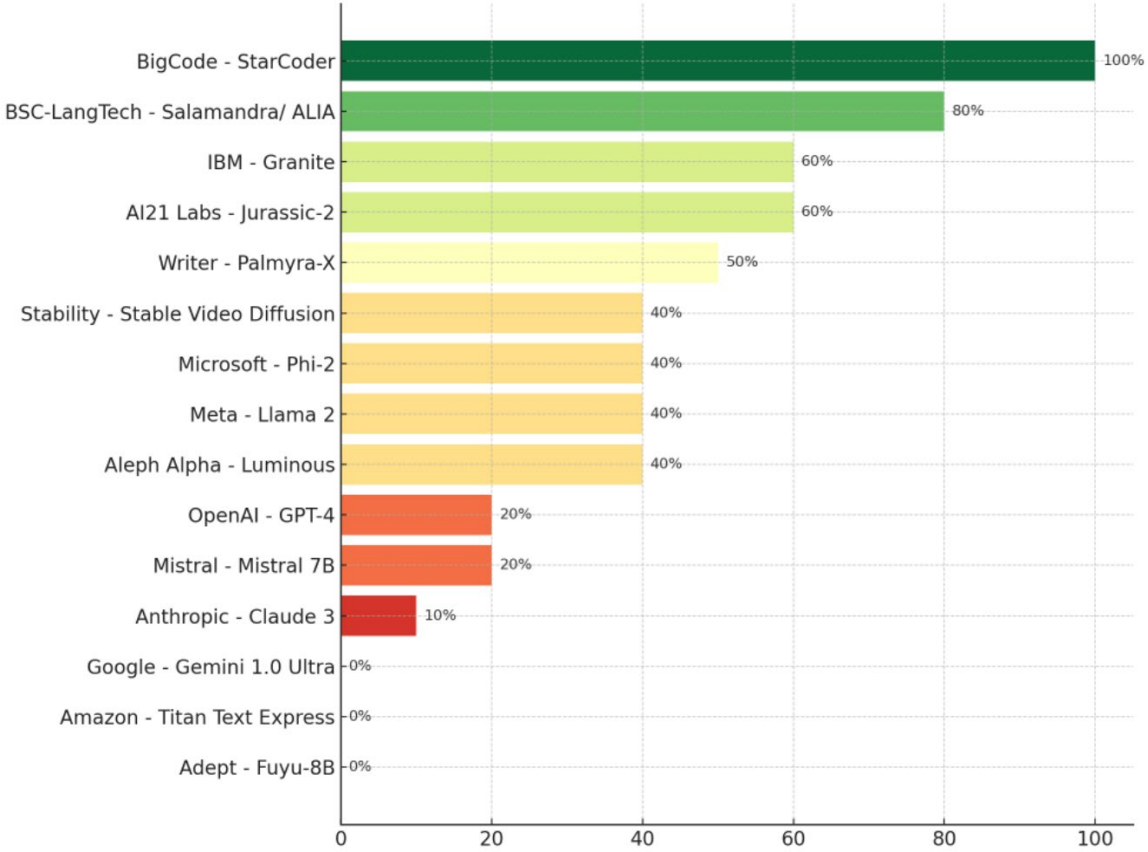


Portuguese and
Catalan
- mid-resource

Galician and
Basque
- low-resource

Prioritization of Romance languages (2% Catalan, 17% Spanish). Total of 2,400,000,000,000 tokens (2.4 trillions).

Compliance with data transparency criteria



Degree of data compliance based on the 10 criteria of the [Foundational Model Transparency Index (FMTI)]

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Achievements

Salamandra LLMs

Multilingual generative models, **35 EU languages** trained with up to **12,9T tokens**.

- sizes **2B** ,**7B** and **40B**
- base & instructed versions
- quantized versions

Repo Type	Repo Name	Total Downloads
Model	BSC-LT/salamandra-2b	21.5k
Model	BSC-LT/salamandra-7b-instruct	19.5k
Model	BSC-LT/salamandra-7b	15.9k

- **Biggest National LLM(s)** and a key element to empower AI sovereignty in Spain and EU.
- Competitive results. **Top 5.**
- **Open**, not only weights but also configuration file and execution scripts.
- **Compliant with EU AI regulations** , close collaboration with AESIA (Spanish Agency for the Supervision of AI).
- **Legitimate data access and traceability**.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Achievements

Salamandra LLMs

Multilingual generative models, **35 EU languages** trained with up to **12,9T tokens**.

- sizes **2B** ,**7B** and **40B**
- base & instructed versions
- quantized versions

Repo Type	Repo Name	Total Downloads
Model	BSC-LT/salamandra-2b	21.5k
Model	BSC-LT/salamandra-7b-instruct	19.5k
Model	BSC-LT/salamandra-7b	15.9k

- **Biggest National LLM(s)** and a key element to empower AI sovereignty in Spain and EU.
- Competitive results. **Top 5.**
- **Open**, not only weights but also configuration file and execution scripts.
- **Compliant with EU AI regulations** , close collaboration with AESIA (Spanish Agency for the Supervision of AI).
- **Legitimate data access and traceability**.

Unfair representation of languages harms performance in mid- and low-resource languages!

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Enhancing language inclusivity

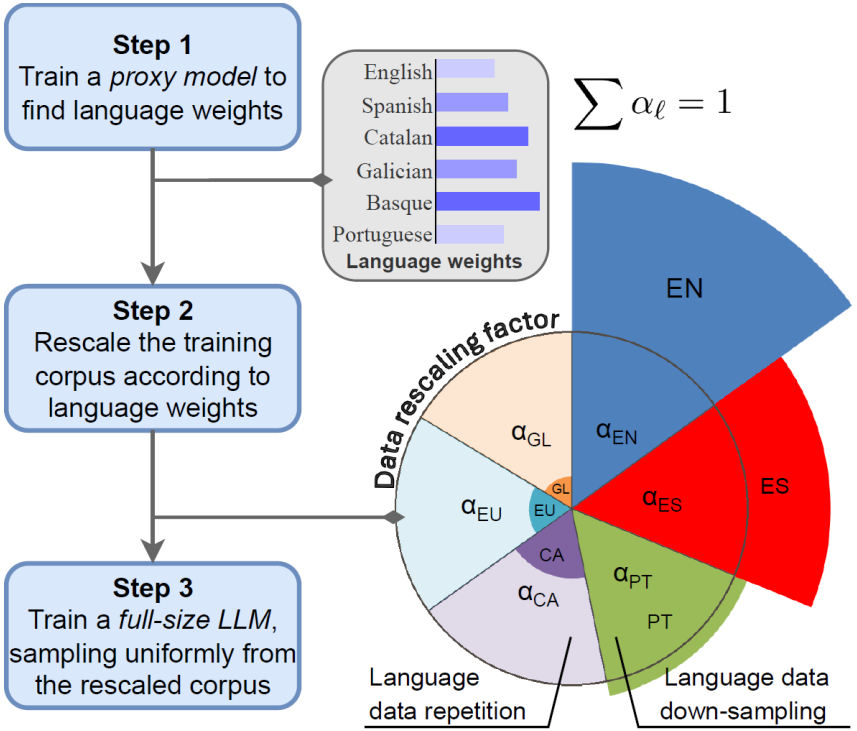
Language Technologies Laboratory

Multilingual data reweighting

Lacunza, I., Saiz, J., Shvets, A., Gonzalez-Agirre, A., Villegas, M.

Objectives:

- Optimize data distribution across the target languages (ES, CA, GL, EU, PT, and EN) to avoid the usual over-reliance on English and benefit more from cross-lingual transfer
- Use the data composition that maximizes performance across all languages to train base LLMs continually (CPT).



Proposed adaptation of DoGE from Fan et al., 2023

DoGE: Domain Reweighting with Generalization Estimation

Proxy model training

S – the set of sources (domains)
L – the set of languages with k_ℓ sources in a language $\ell \in L$: $S_\ell = \{S_{\ell,1}, S_{\ell,2}, \ldots, S_{\ell,k_\ell}\}$

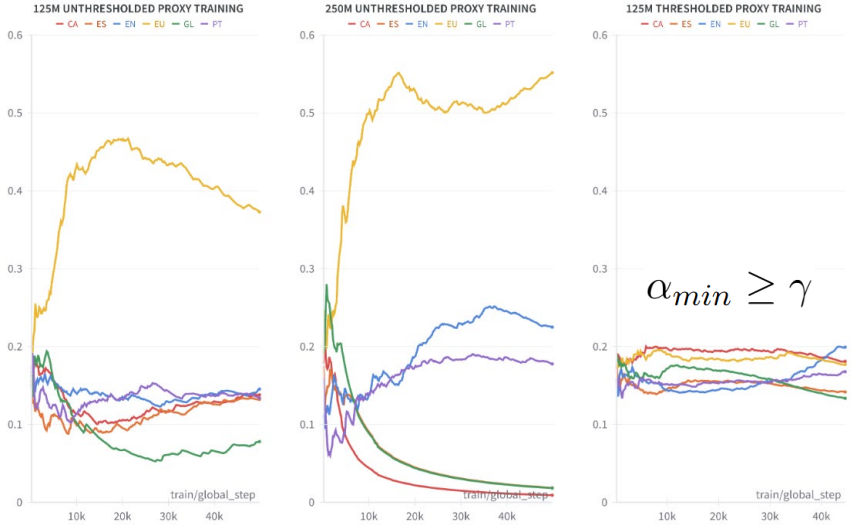
Bi-level optimization problem:

$$\alpha \in \arg\min_{\alpha \in \Delta_k} \sum_{\ell \in \mathcal{L}} \sum_{i=1}^{k_\ell} l_{\ell,i}(\theta^*(\alpha)),$$

$$s.t.\ \ \theta^*(\alpha) \in \operatorname{argmin}_{\theta} \sum_{\ell \in \mathcal{L}} \sum_{i=1}^{k_\ell} \alpha_{\ell,i} l_{\ell,i}(\theta),$$

$$s.t.\ \ \alpha_{min} \geq \gamma; \gamma \lll |S|^{-1},$$

$l_{\ell,i}(\theta)$ – the next-token prediction (cross-entropy) loss
γ – constant



The train batch sampling at time-step *t:*

$$P_\alpha \triangleq \sum_{\ell \in \mathcal{L}} \sum_{i=1}^{k_\ell} \alpha_{\ell,i}^{(t)} \cdot \texttt{UNIF}(S_{\ell,i})$$

Final language weights and final fixed sampling:

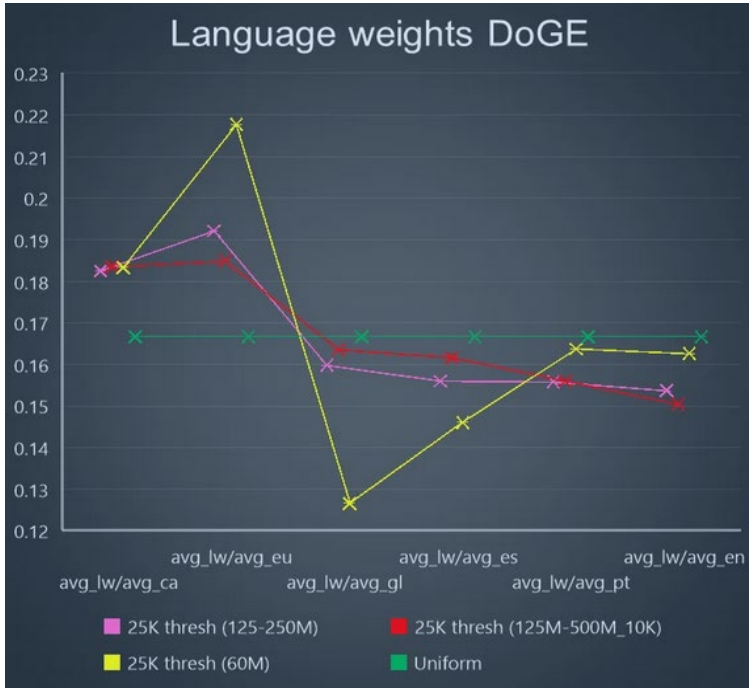
$$\bar{\alpha}_\ell = \sum_{i=1}^{k_\ell} \alpha_{\ell,i} \qquad P_{\bar{\alpha}} \triangleq \sum_{\ell \in \mathcal{L}} \bar{\alpha}_\ell \cdot \texttt{UNIF}(S_\ell)$$

Iberian DoGE

Iberian DoGE:

- Testing 60M, 125M, 250M and 500M proxy models to obtain language weights.
  - Obtained **different weight distributions** of iberian langs + English + Code.
- Trained 125M, 500M and 900M models for

	Perplexity			IberoBench (acc)		
Language	Bsl.	Set1	Set2	Bsl.	Set1	Set2
CA	14.43	**14.42**	**14.42**	41.01	**41.21**	40.29
PT	**31.33**	31.52	32.30	51.73	**53.19**	43.66
EN	31.56	**31.40**	32.13	38.15	38.03	**38.87**
GL	**29.38**	32.07	29.71	32.52	**33.00**	32.60
ES	35.15	35.30	**35.13**	43.43	44.17	**44.73**
EU	**23.11**	23.61	23.55	37.06	36.12	**37.24**
Total Avg	**27.49**	28.05	*27.87*	*40.65*	**40.95**	39.57

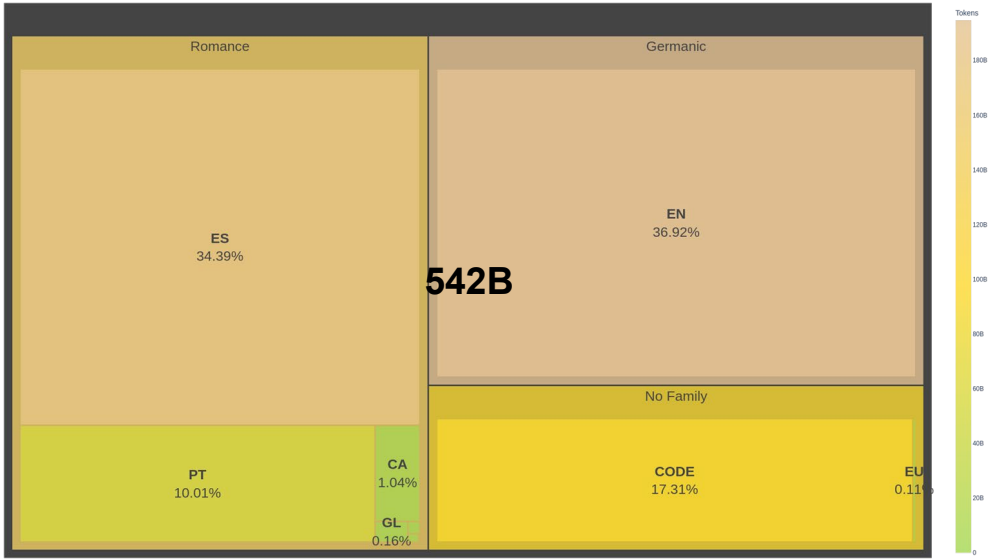


Data for proxy and base models: OSCAR and Wikipedia

Pata Negra

Pata Negra: Trained from scratch of a 7B model with Iberian languages + English with Ad-Hoc distribution.

Belebele (Acc)	IberianLLM	Salamandra-7b Base
Basque	**28.33**	23.22
Galician	**28.00**	24.67
Catalan	27.22	**28.11**
Portuguese	**29.00**	25.11
English	**28.22**	27.67
Spanish	**27.44**	26.67
Average	**28.04**	25.91



	Flores Iberian		Flores Non-Iber		Belebele		English		Catalan		Spanish		Basque		Galician		Portuguese	
	Iber.	Sala.	Iber.	Sala.	Iber.	Sala.	Iber.	Sala.	Iber.	Sala.	Iber.	Sala.	Iber.	Sala.	Iber.	Sala.	Iber.	Sala.
Scores	23.63	25.42	18.42	24.23	28,04	25,91	43.84	43.46	50.99	52.70	35.37	36,03	37,36	39,87	26.76	27,12	58,80	63,78
Δ%	**-7.04%**		**-23.98%**		**+8.22%**		**+0.87**		**-3.24**		**-3.44**		**-6.29%**		**-1.33**		**-7.81%**	

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Pata Negra evaluation on IberoBench tasks

English Task	IberianLLM	Salamandra-7b Base
ARC-Easy	**78.62**	71.89
ARC-Challenge	**48.29**	38.99
HellaSwag	52.19	**55.18**
COPA	**87.00**	86.00
XStoryCloze	75.91	**77.90**
XNLI	**49.92**	48.47
OpenBookQA	**31.60**	28.60
PIQA	**78.02**	77.04
SocialIQA	**32.29**	31.93
COLA	**0.11**	0.03
WNLI	**53.52**	43.66
TruthfulQA-Gen	25.55	**25.88**
TruthfulQA-MC1	25.09	**28.89**
TruthfulQA-MC2	38.48	**44.19**
PAWS	59.40	**60.75**
VeritasQA-MC1	19.83	**25.70**
VeritasQA-MC2	33.34	**37.18**
MGSM	0.04	**0.05**
Average	**43.84**	43.46

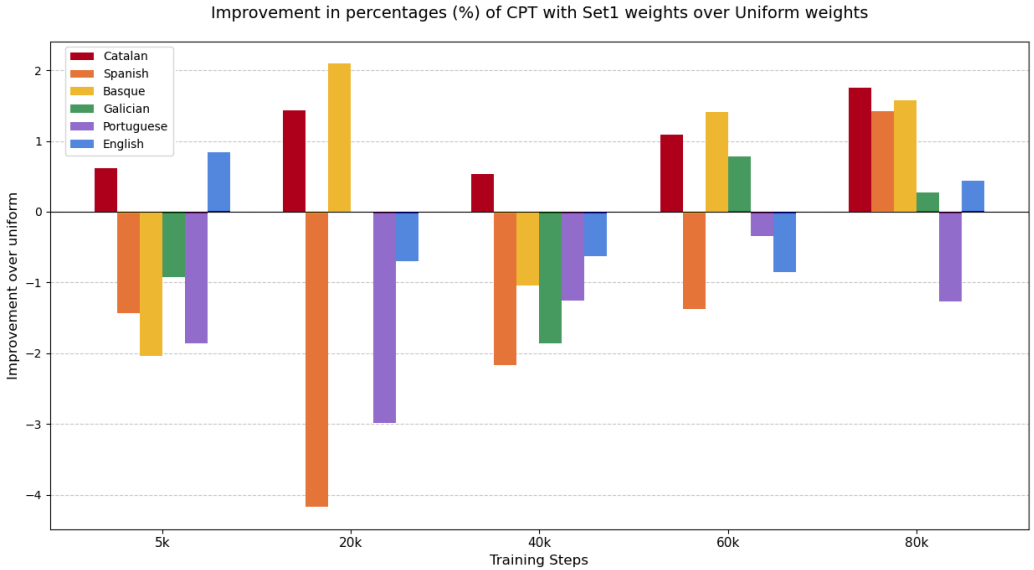
Spanish Task	IberianLLM	Salamandra-7b Base
WNLI	42.25	**56.34**
XNLI	**49.28**	44.54
PAWS	**60.60**	58.30
Escola	**0.04**	**0.04**
XStoryCloze	71.34	**72.93**
MGSM	**0.04**	**0.04**
XQuad	66.55	**66.63**
XLSum	0.54	**3.13**
VeritasQA-MC1	24.58	**25.42**
VeritasQA-MC2	38.52	**38.91**
Average	35.37	**36.63**

Galician Task	IberianLLM	Salamandra-7b Base
GalCOLA	0.05	**0.07**
Parafrases	**58.16**	58.16
PAWS	58.10	**62.10**
Summarization	2.43	**7.14**
MGSM	0.02	**0.04**
OpenBookQA	**32.20**	30.60
VeritasQA-MC1	**24.86**	23.18
VeritasQA-MC2	**39.73**	38.80
TruthfulQA-Gen	**22.19**	19.03
TruthfulQA-MC1	23.28	**23.99**
TruthfulQA-MC2	33.30	**35.19**
Average	26.76	**27.12**

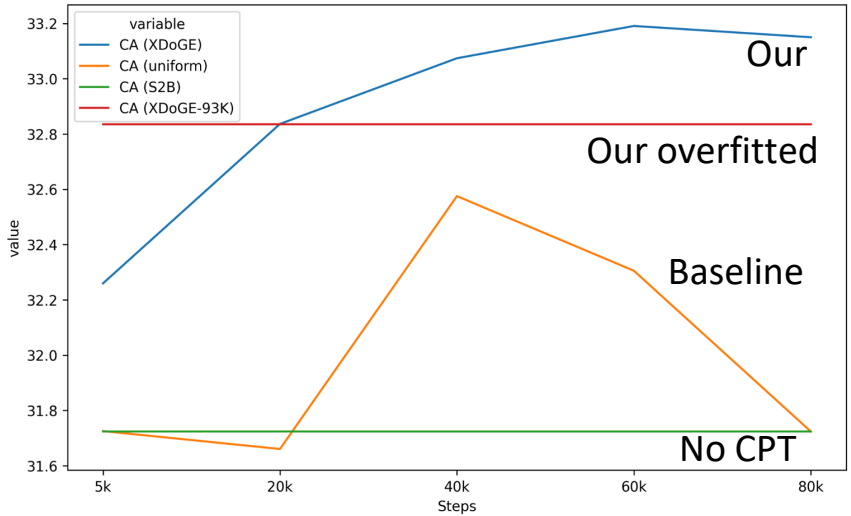
CPT for Iberian languages

Iberian-Salamandra: CPT of Salamandra

Salamandra-2b CPT using the XDoGE weights has a delayed improvement in comparison to using the uniform weights:



Upsampling minor and mid-resource languages implies data over-repetition. Models trained with uniform weights (orange) degrade faster than with XDoGE weights (blue):



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Data for CPT: FineWeb-Edu, FineWeb2, and Wikipedia

CPT for Iberian languages

	Salamandra-2B					IberianLLM-7B		
	Perplexity		IberoBench			IberoBench (acc)		
Language	Bsl1.	XDoGE	No-CPT	Bsl1.	XDoGE	No-CPT	Bsl2.	XDoGE
CA	6.42	**6.37**	51.66	*51.78*	**52.69**	*53.76*	53.40	**54.77**
PT	11.27	**11.16**	**53.34**	*52.87*	52.20	48.87	**54.01**	*52.94*
EN	**10.80**	**10.80**	**46.49**	45.51	*45.71*	*52.24*	**53.08**	52.10
GL	7.79	**7.66**	29.54	*29.66*	**29.74**	*34.31*	**34.64**	34.28
ES	12.19	**12.15**	*50.19*	50.14	**50.85**	46.75	*48.70*	**50.46**
EU	5.21	**5.20**	38.47	*39.35*	**39.97**	39.76	*41.03*	**41.44**
Total Avg	8.95	**8.89**	*44.95*	44.89	**45.19**	45.95	*47.48*	**47.67**

Bsl1. - uniform sampling (equal weights)
Bsl2. - ad-hoc sampling

Future work:

- Better adaptation of highly multilingual models to compromise less on non-target languages
- Use language family correspondences explicitly to improve the capture of inter-dependencies

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Computational resources

Run Type	Code(s)	No. of runs	No. of nodes	Node Multiplier*	Execution time per run**	Total node hours
Setup	Nemo	100	2	2	2 hours	400
Proxy trainings	Nemo	20	1	1	500 hours	10,000
GTP-7B (2T)	Nemo	4	8	8	1,998 hours	63,936
GTP-13B (2T)	Nemo	4	16	2	1,855 hours	118,720
Evaluation	LM-Harness	32	1	1	8 hours	256
Buffer	-	-	-	-	-	6,944
Total	-	-	-	-	-	200,000

* The node multiplier based on the scalability tests outlined in Section 2.6.2.1

** Execution time computed use the equation presented in Section 2.6.2.3. ActualFLOPS has been set at 365 TeraFLOPs per GPU, based on our scalability tests.

BSC Centro Nacional de Supercomputación

Project schedule

	Month																		
Task	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
Data preparation				M															
Library preparation				M															
Data transfer					M														
Setup					M														
Proxy training					M														
Model training (7B)									M										
Model training (13B)													M						
Evaluation															M				
Results analysis																M			
Paper draft																	M		
Final report																			D

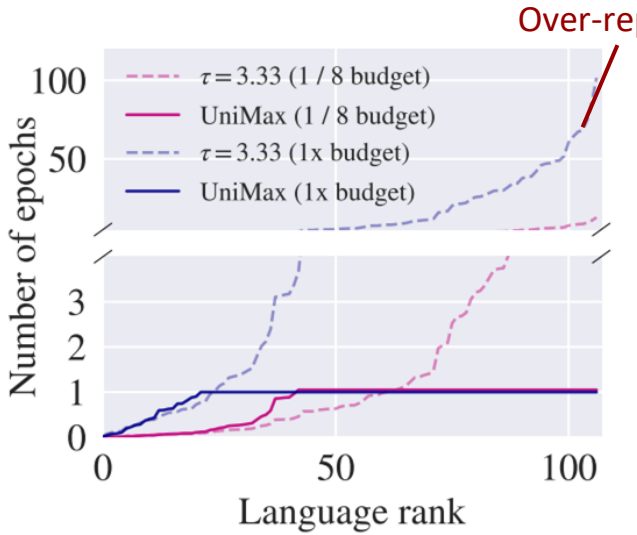
* M stands for “milestone” and D for “deliverable”.

Project deviations

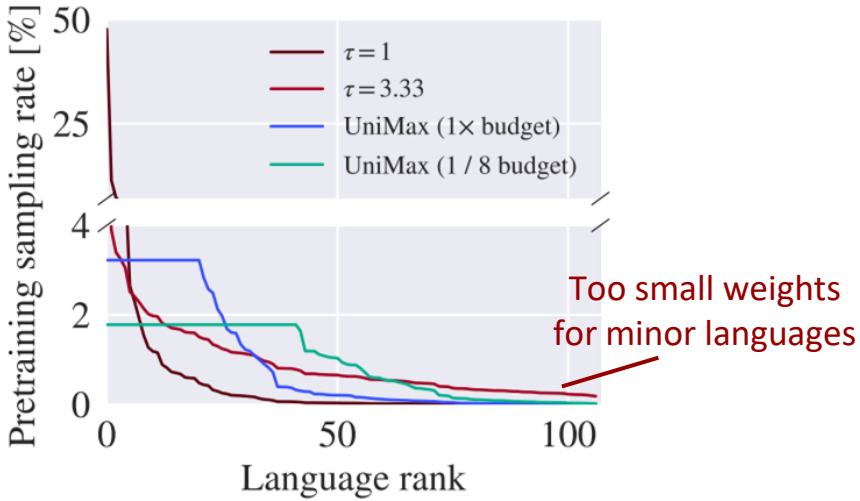
- 13B model training is no longer planned, as the approach appears unsuitable for training from scratch.
- Instead, a staged language learning technique will be studied to allow for higher language inclusivity at later stages.
- For this, a series of 2B models is being trained using different language mix configurations.
- Depending on the outcome, we plan to train 7B in a staged manner and apply XDoGE-CPT to the resulting model.

Staged language learning

Motivation: avoid unbalanced distributions and excessive repetition/undersampling



(a) Number of training epochs for each language. Temperature sampling results in a large number of data repeats for low-resource languages, whereas UniMax explicitly caps repeats.



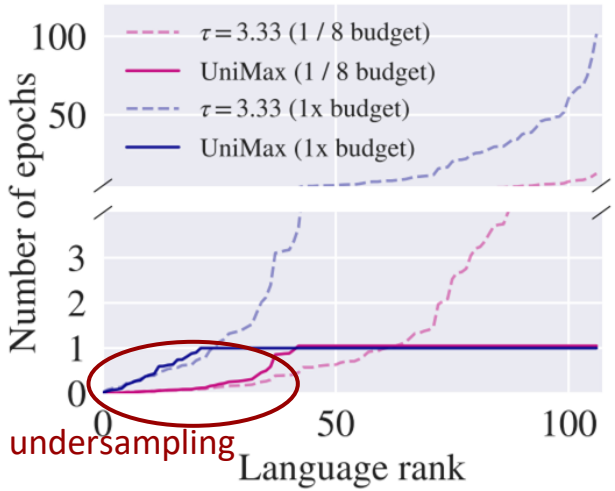
(b) Pretraining sampling distribution. Temperature sampling results in poorly balanced distributions, whereas UniMax provides more uniform distributions without excessive upsampling.

Figure 1: The x-axis is the rank of the language based on the character count.

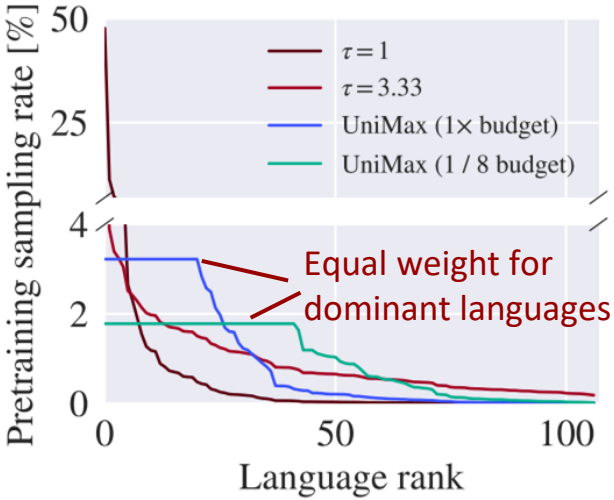
(Chung et al., 2023)

Staged language learning

Motivation: avoid unbalanced distributions and excessive repetition/undersampling



(a) Number of training epochs for each language. Temperature sampling results in a large number of data repeats for low-resource languages, whereas UNIMAX explicitly caps repeats.



(b) Pretraining sampling distribution. Temperature sampling results in poorly balanced distributions, whereas UNIMAX provides more uniform distributions without excessive upsampling.

Figure 1: The x-axis is the rank of the language based on the character count.

(Chung et al., 2023)

Staged language learning

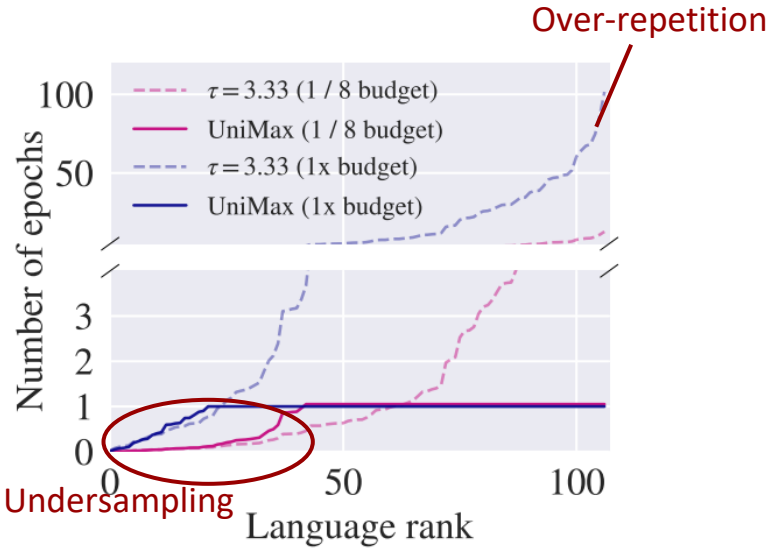
Problems

Over-repetition:

- leads to overfitting, which degrades performance on downstream tasks
- increases the risk of memorizing private or sensitive content
- wastes training cycles that could have been devoted to unique examples

Undersampling:

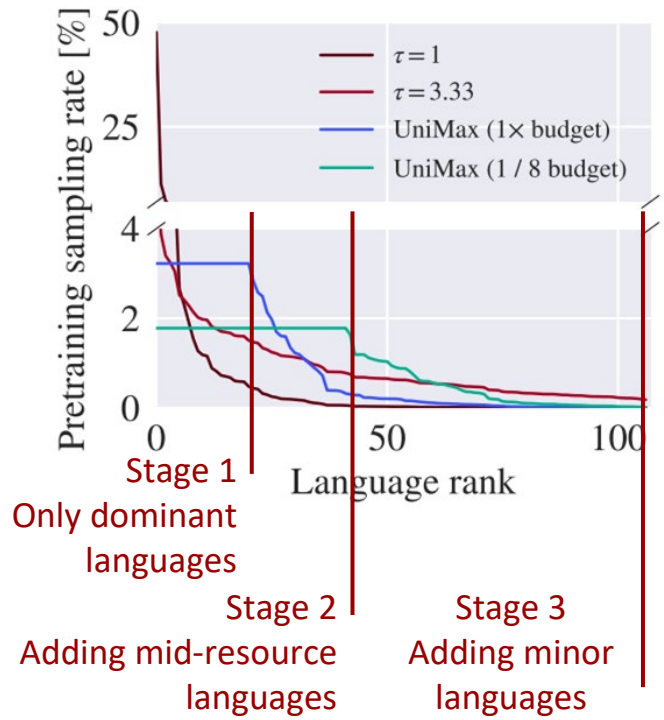
- leads to the drop of a significant amount of available knowledge in many languages



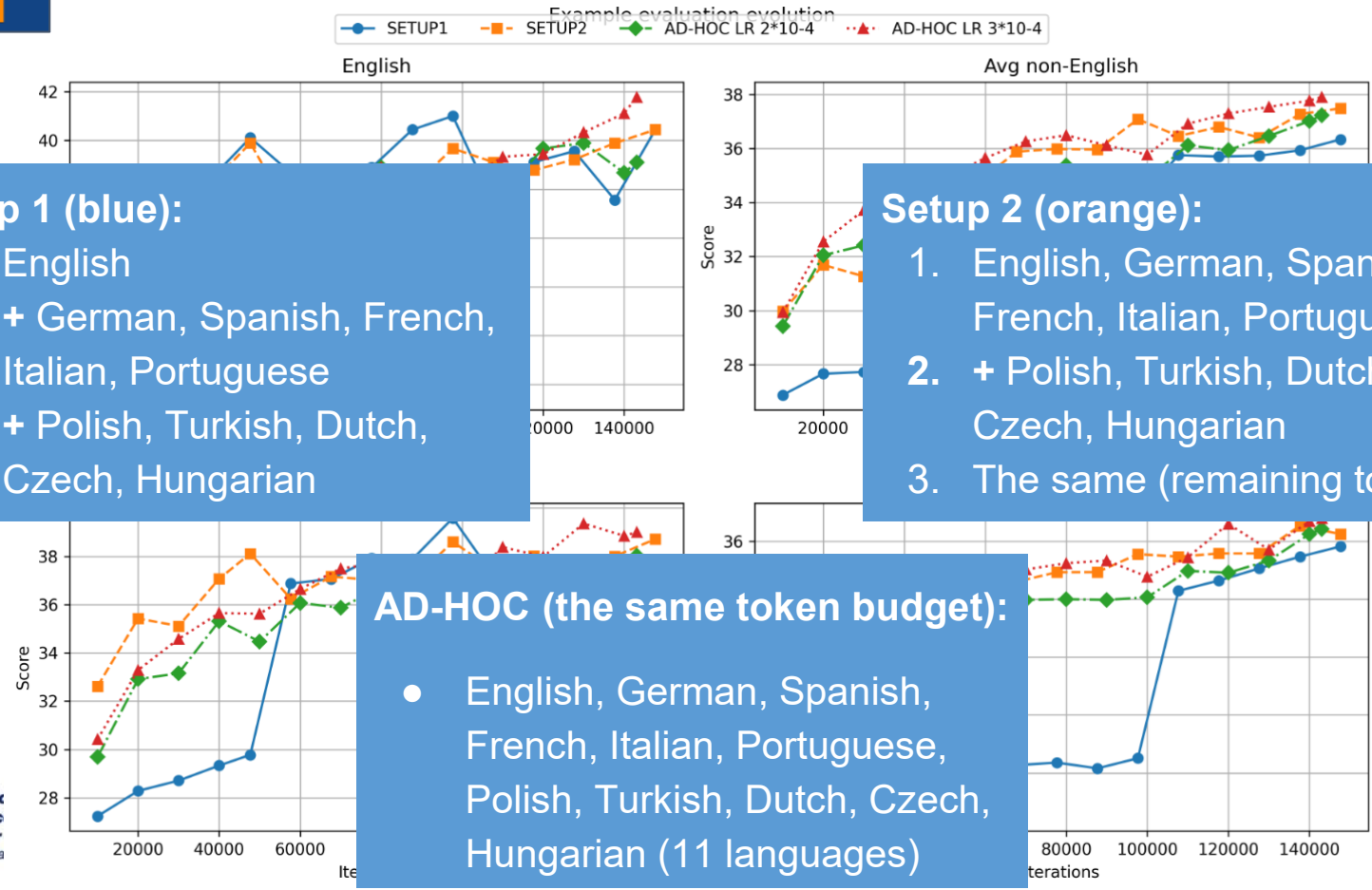
Staged language learning

Proposal:

- Introduce new languages sequentially in groups
  - Start from the dominant ones to train the core of the model, and end with the minor for individual layer adjustments
  - Consider diversifying language families within each stage
- Weight languages separately within each group
- Experiment with updating only layers responsible for multilingual abilities at later stages



Staged language learning. First results



Setup 1 (blue):

1. English
2. **+** German, Spanish, French, Italian, Portuguese
3. **+** Polish, Turkish, Dutch, Czech, Hungarian

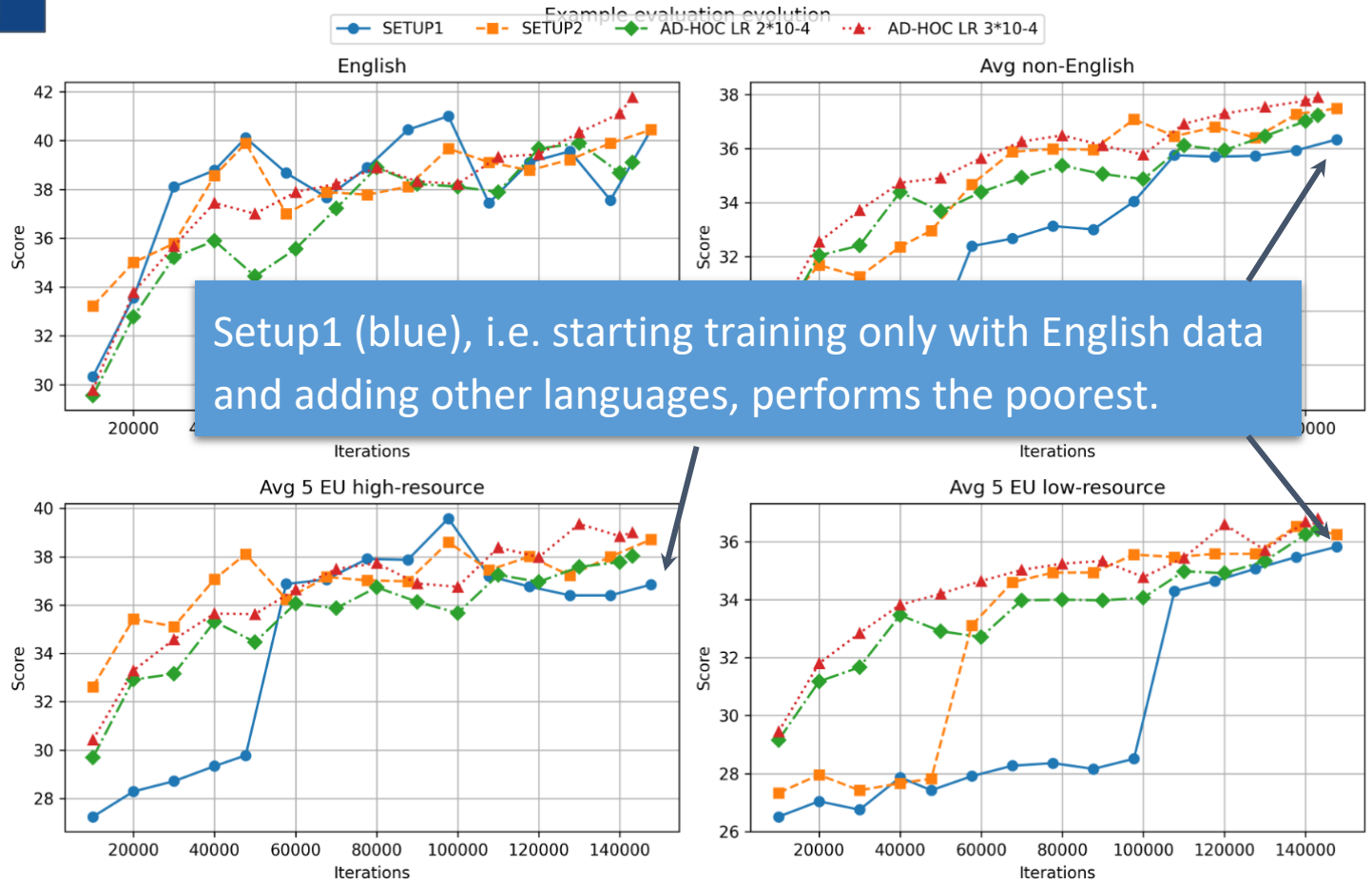
Setup 2 (orange):

1. English, German, Spanish, French, Italian, Portuguese
2. **+** Polish, Turkish, Dutch, Czech, Hungarian
3. The same (remaining tokens)

AD-HOC (the same token budget):

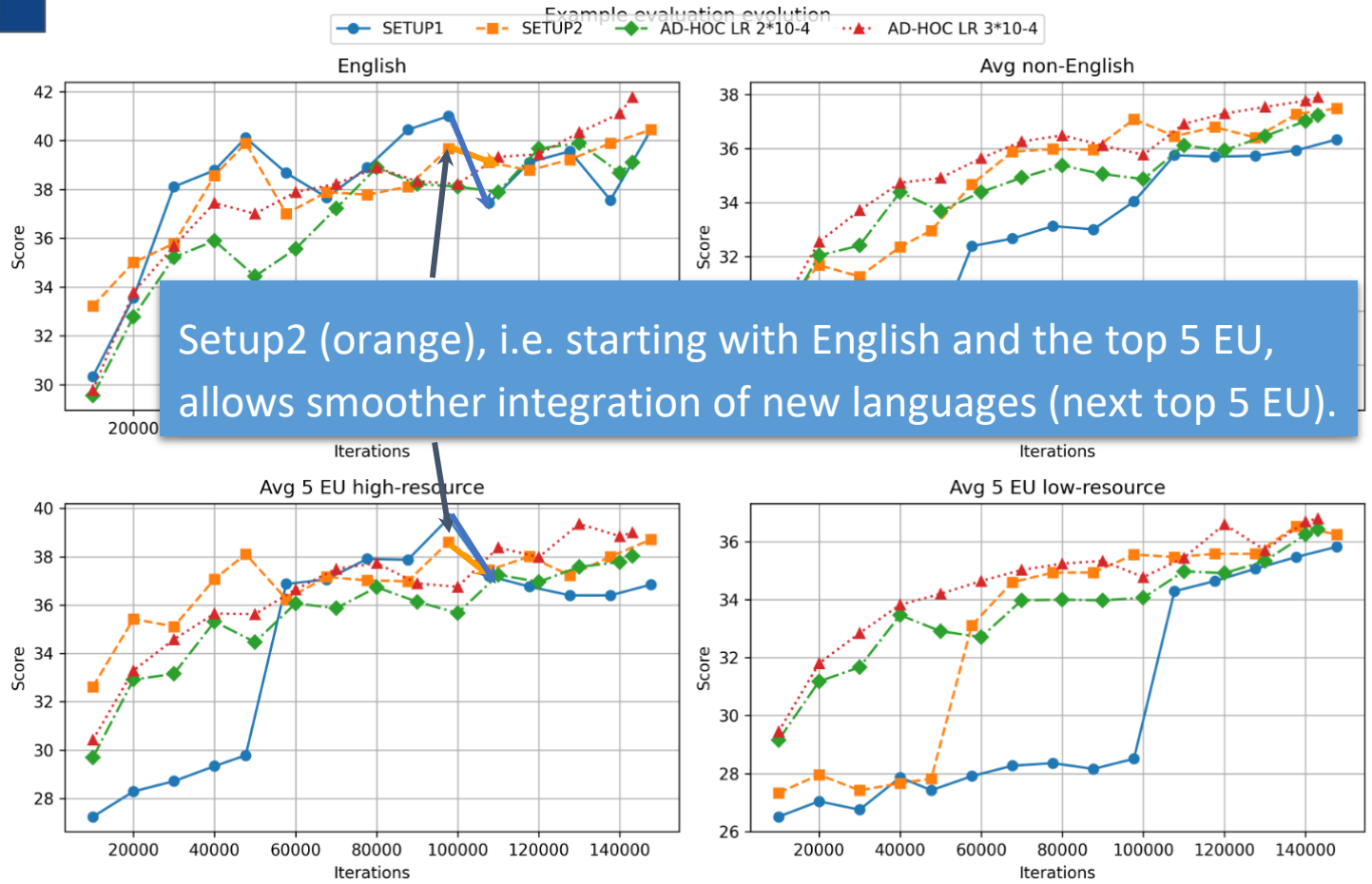
- English, German, Spanish, French, Italian, Portuguese, Polish, Turkish, Dutch, Czech, Hungarian (11 languages)

Staged language learning. First results



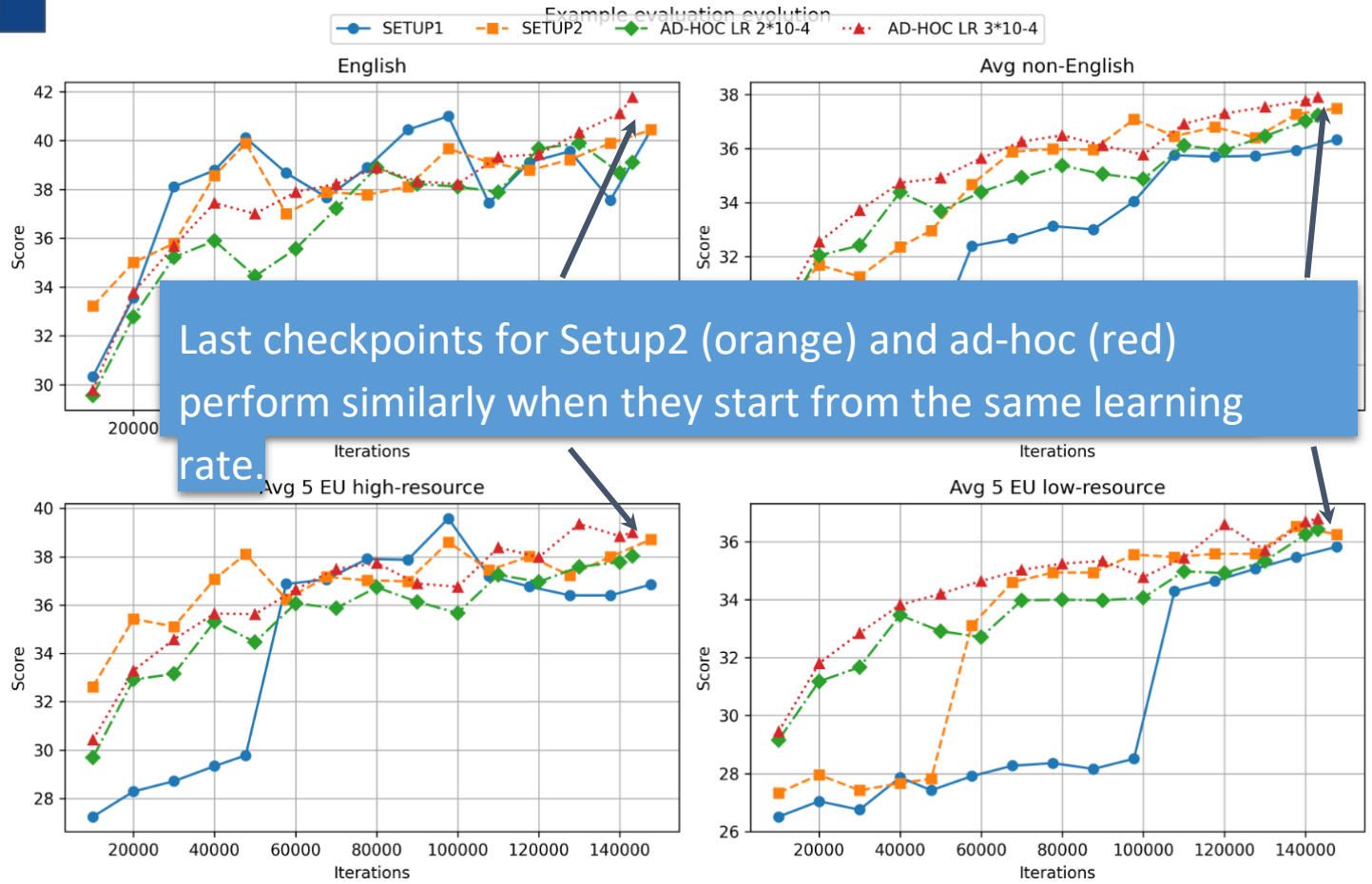
Setup1 (blue), i.e. starting training only with English data and adding other languages, performs the poorest.

Staged language learning. First results



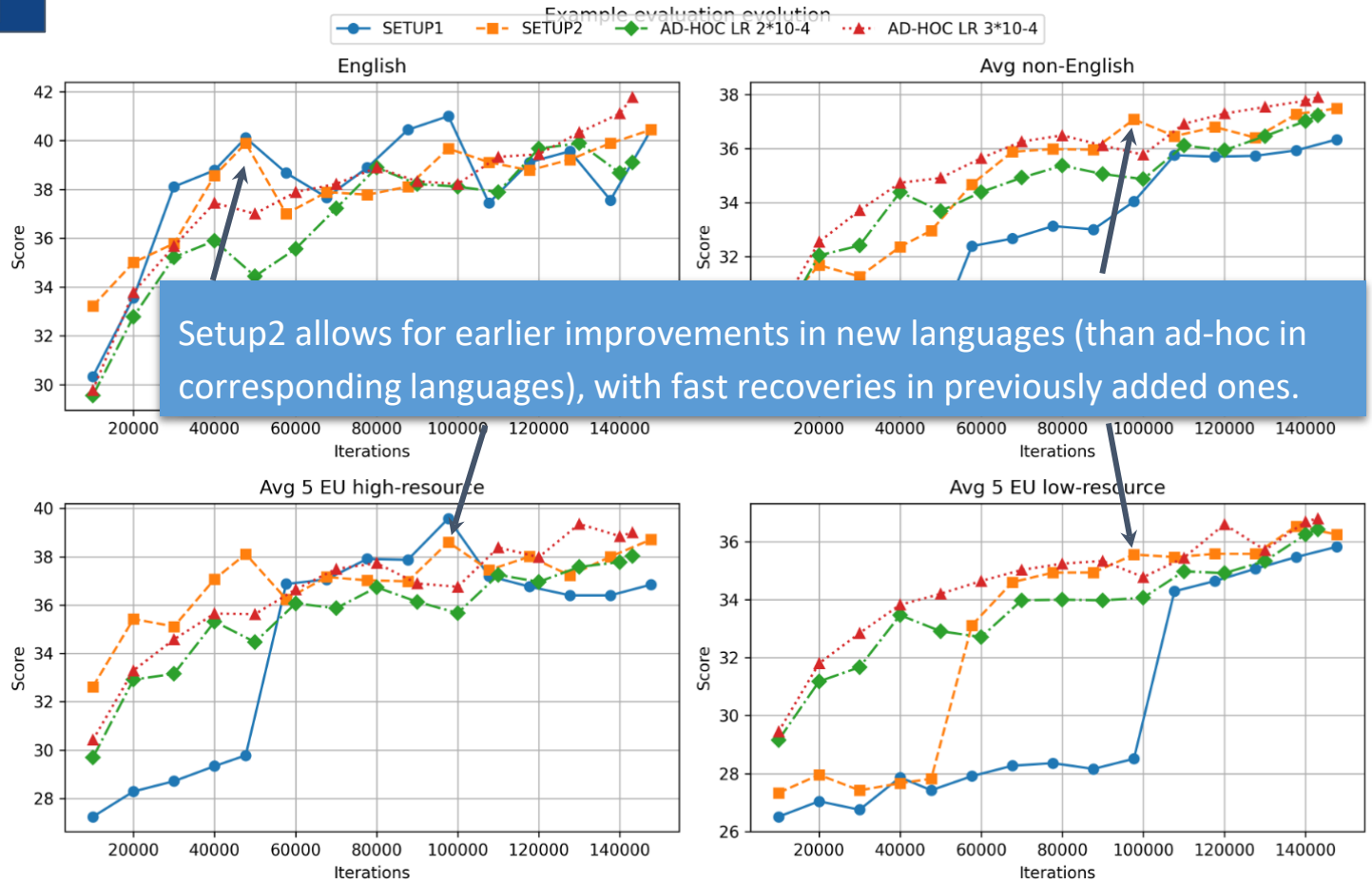
Setup2 (orange), i.e. starting with English and the top 5 EU, allows smoother integration of new languages (next top 5 EU).

Staged language learning. First results



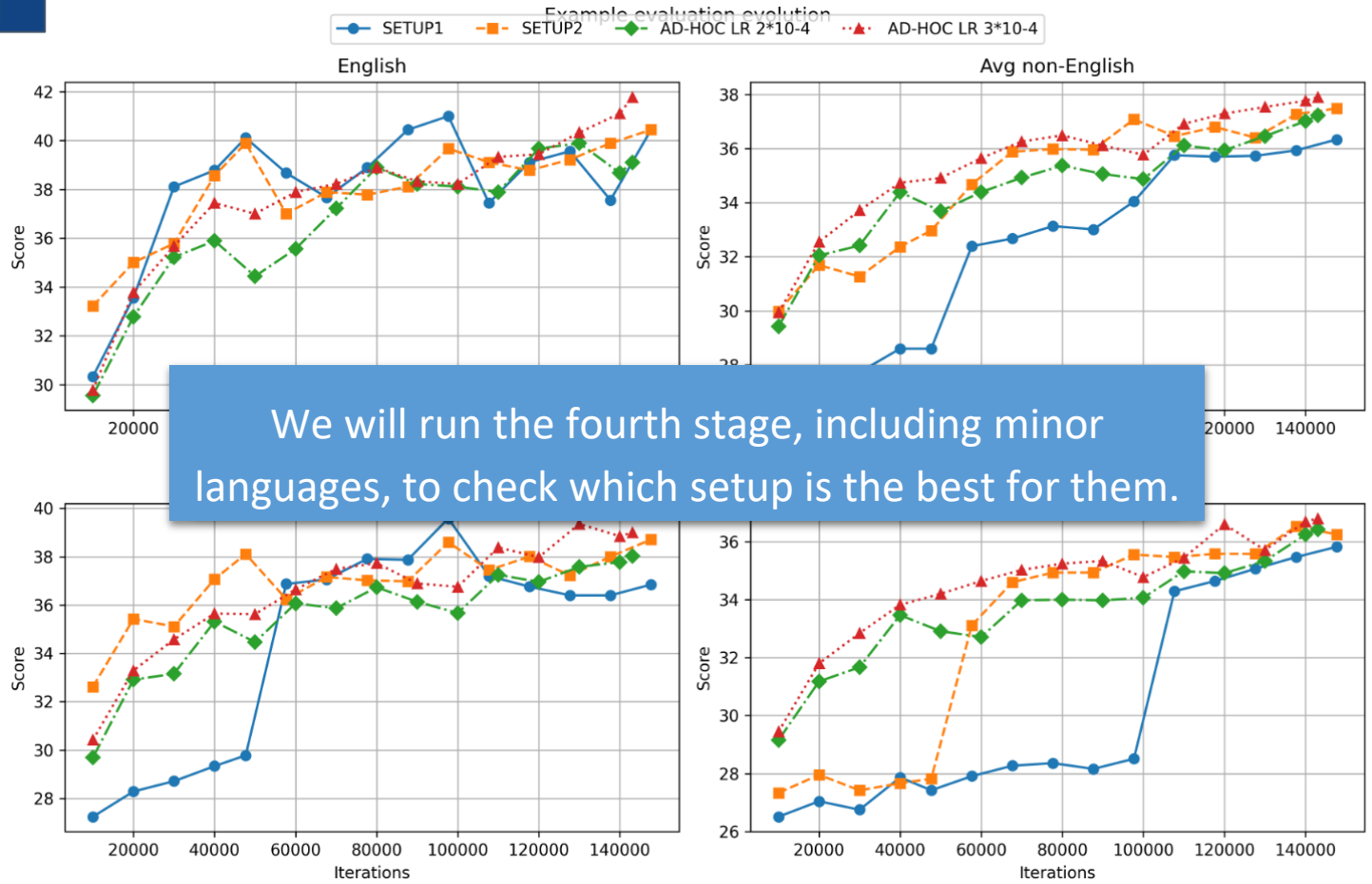
Last checkpoints for Setup2 (orange) and ad-hoc (red) perform similarly when they start from the same learning rate.

Staged language learning. First results



Setup2 allows for earlier improvements in new languages (than ad-hoc in corresponding languages), with fast recoveries in previously added ones.

Staged language learning. First results



We will run the fourth stage, including minor languages, to check which setup is the best for them.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Conclusions

Language Technologies Laboratory

Conclusions

- We introduced XDoGE – an extension of the DoGE algorithm for multilingual continual pre-training.
- The optimized set of language weights enhances cross-lingual transfer capability.
- We result in a series of LLMs of different sizes and origins, centred on Iberian languages and English.
- We further explore training using staged language learning and plan to train a 7B model from scratch.
- We expect the resulting models to enable higher language inclusivity at subsequent stages.
- The results are essential for further developments in ALIA and OpenEuroLLM projects.

salamandra-2b, salamandra-2b-instruct, salamandra-7b, salamandra-7b-instruct, ALIA-40b

Salamandra

IberianLLM

ALIA Kit

Thank you!

[LangTech webpage]

aleksandr.shvets@bsc.es

Acknowledgments: This research was supported by the EuroHPC Extreme Scale Access grant EHPC-E01-009 and funded by the Project Desarrollo de Modelos ALIA with the framework of the Plan Nacional de Tecnologías de Lenguaje -ENIA 2024 and PRTR, NextGeneration EU, Resol. SEDIA 19.08.2024.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ALIA Brings Linguistic Diversity: Optimizing Data Distribution for Multilingual NLP Inclusivity

Alexander Shvets

Acknowledgments: This research was supported by the EuroHPC Extreme Scale Access grant EHPC-E01-009 and funded by the Project Desarrollo de Modelos ALIA with the framework of the Plan Nacional de Tecnologías de Lenguaje -ENIA 2024 and PRTR, NextGeneration EU, Resol. SEDIA 19.08.2024.

Language Technologies Laboratory

BIOEMTECH

Embracing scientists translate ideas into outcomes

ΔosimetrEYE:
A GenAI-based tool for 3D dosimetry from 2D imaging for preclinical studies

Panagiotis Papadimitroulas
Associate Prof. Biomedical Informatics
Medical Department, Univ. of Thessaly

Software Advisor & Co-Founder
BIOEMTECH
panpap@bioemtech.com
www.bioemtech.com

FORTISSIMO PLUS

EuroHPC Joint Undertaking

EUROHPC USER DAY

30 SEPTEMBER 1 OCTOBER 2025 DENMARK

EURO DENMARK

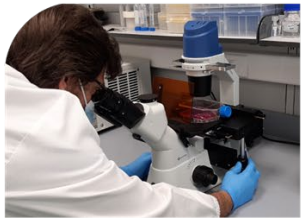
DeiC

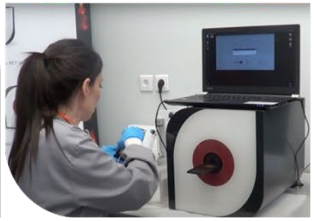
ABOUT BIOEMTECH

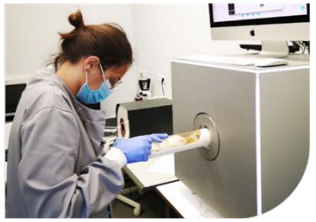
BIOEMTECH develops and offers innovative solutions in pharmaceutical and medical physics and biotechnology research

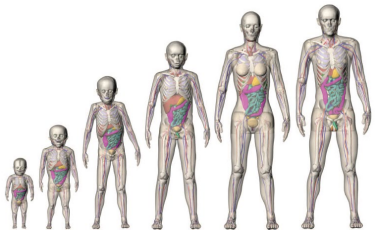
We focus on **molecular imaging, dosimetry & biomedical engineering**:

- Design and construction of screening imaging systems, the *eyes*™
- Preclinical CRO for testing new radiolabelled compounds
- Computational tools for medical imaging and dosimetry









Courtesy of BIOEMTECH

CRO SERVICES

In vitro tests

Radiochemistry controls

Animal Models & Protocol set up

Imaging & Quantification

Ex vivo Analysis

Customized Reporting

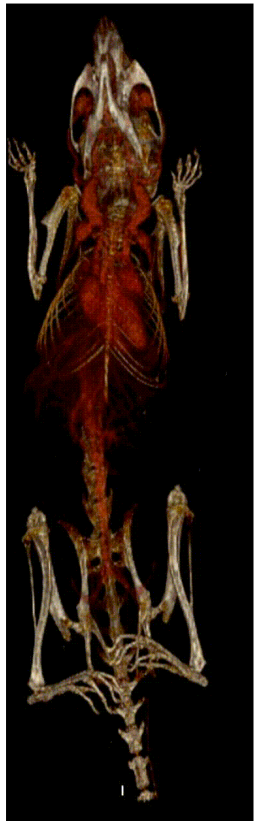
RESEARCH CENTERS

UNIVERSITIES

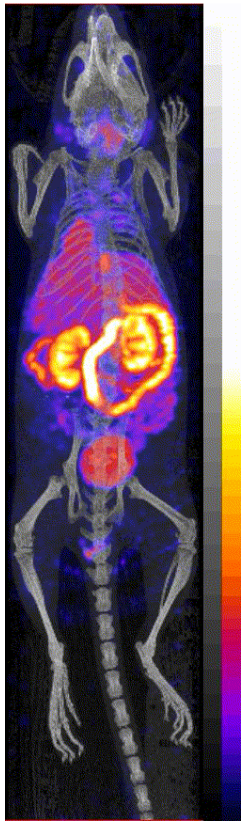
PHARMACEUTICAL COMPANIES

HOSPITALS

VETERINARY CLINICS



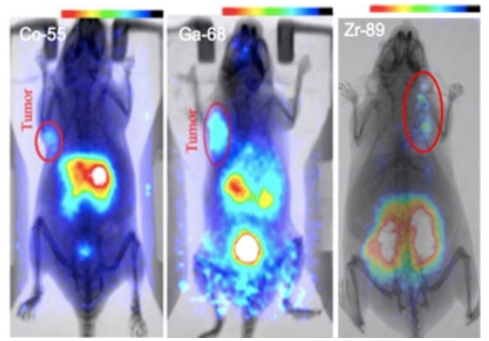
Contrast agents

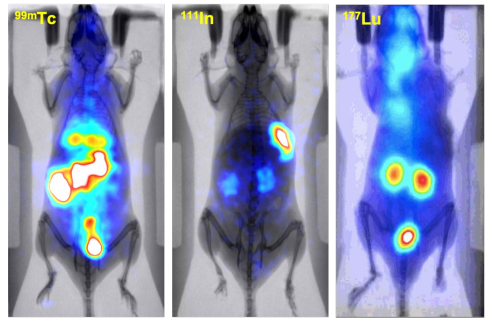


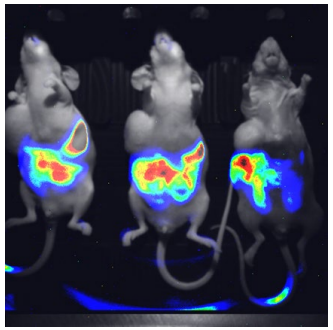
Cardiac imaging

EYES IMAGING DEVICES

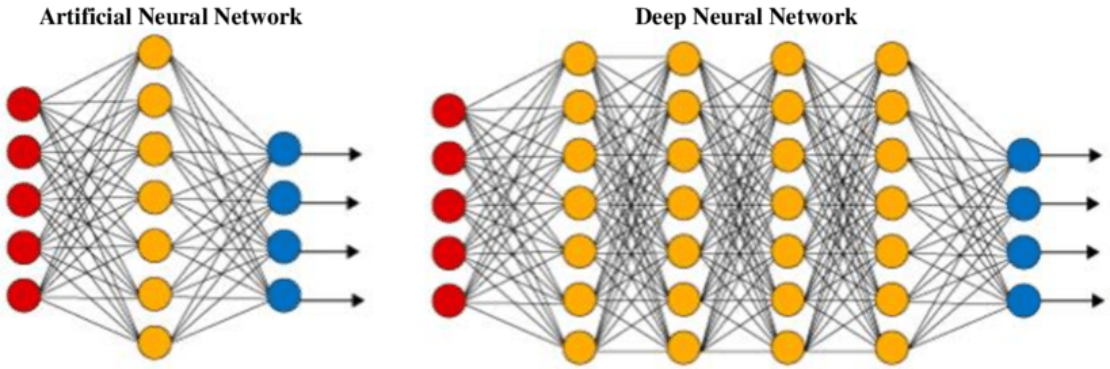
Model name	β-eye	γ-eye	ϕ-eye
What it detects	High energy photons (511keV)	Low energy photons (<250keV)	Light from fluorescent dyes
Special Features	• Imaging of radiolabeled compounds PET • Whole body real-time scanning • Organs/Tumors, functional info • Small form factor • Plug & Play • Data analysis tool	• Imaging of radiolabeled compounds SPECT • Whole body real-time scanning • Organs/Tumors, functional info • Small form factor • Plug & Play • Data analysis tool	• Imaging of chromophores and fluorescent compounds • Whole body real-time scanning • Organs/Tumors, functional info • Small form factor • Plug & Play • Data analysis tool
Suitable for studies	Oncology, lungs, kidneys, bones, brain, nanomedicine	Oncology, lungs, kidneys, bones, brain, nanomedicine	Oncology, nanomedicine







SW DEPARTMENT



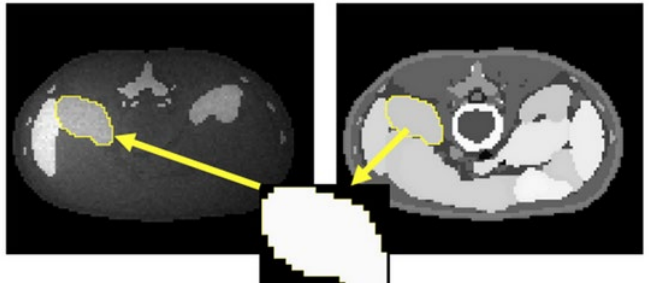
- **Monte Carlo Simulations**

  Dosimetry – Medical Imaging PET/SPECT/CT

- **Artificial Intelligence Techniques (GenAI/XAI/ML)**

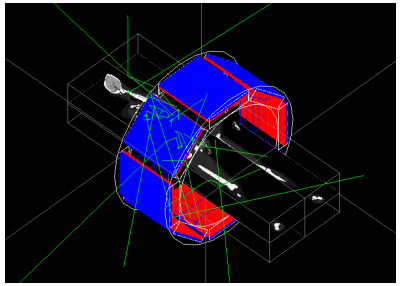
  Develop prediction models (clinical and preclinical apps)

  Decision Support Systems (DSS for clinicians)



- **Image Processing/Analysis**

  Radiomics – Segmentation – Quantification – Reconstruction



- **Anthropomorphic & Animal 3D computational models**

  Synthetic Data (artificial realistic clinical data)

- **HPC exploitation**

ΔOSIMETREYE – The challenge

FORTISSIMO PLUS

Annually **115 million animals** are sacrificed **in biomedical experiments**.
Small animal imaging and dosimetry prediction tools can enable the effective **reduction of over 80%** animal sacrifice and **drastically speed** up the assessment of drug delivery studies.

- **Dosimetry assessment** currently is based on **rough estimations.**
- **MC simulations** provide the ground truth but **are time consuming.**
- **Anatomical variations** are important in preclinical dosimetry and can lead to **significant lack of accuracy (appropriate for drug discovery).**

The goal is to utilize the ground truth dosimetry (MC simulations) with GenAI and HPC resources to develop a prediction tool for 3D dose assessment based on 2D preclinical imaging.

ΔOSIMETREYE – The approach

FORTISSIMO PLUS

Our innovative approach is based on the deep knowledge of
MC simulations, image processing tools, GenAI algorithms and XAI methods.

To develop a prediction dosimetry SW tool for our imaging devices, we need to:

i. **Collect & harmonise imaging data** from different systems (3D SPECT/CT/2D planar).

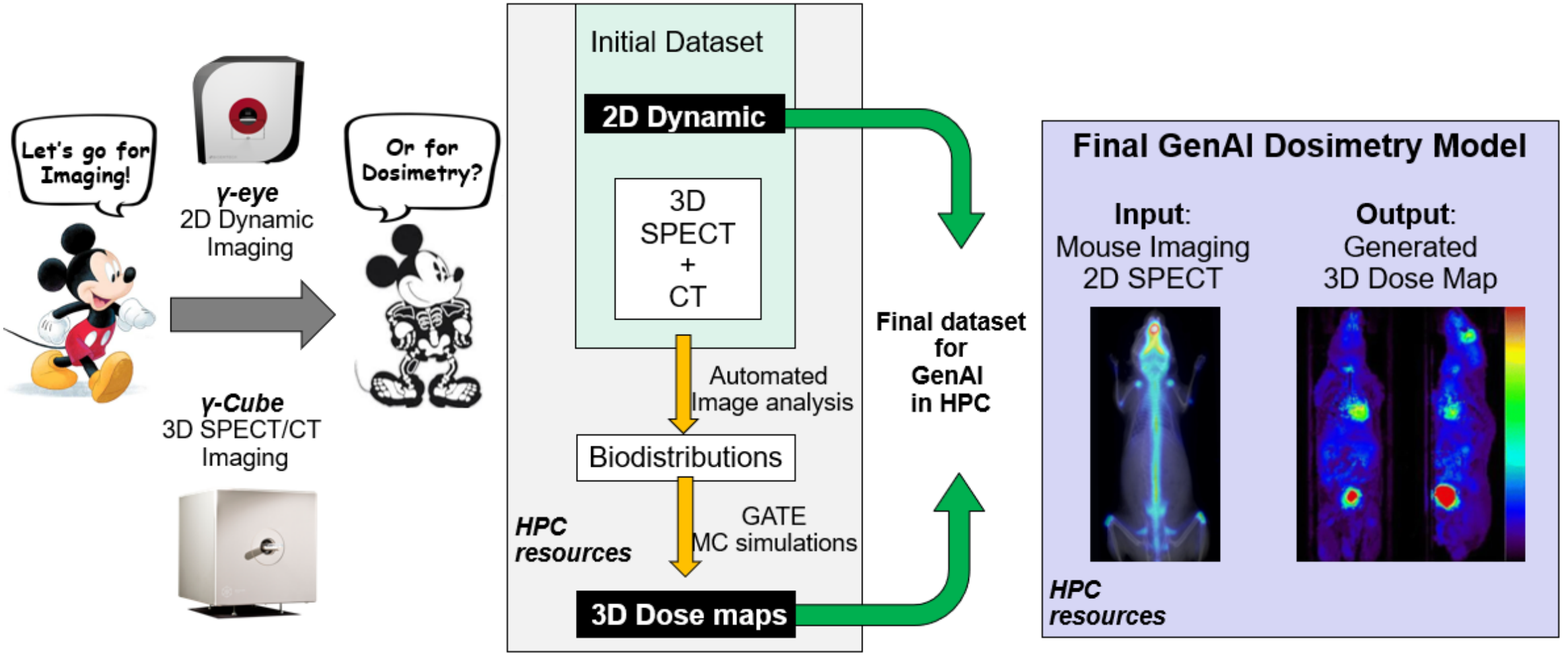
ii. **Execute highly demanding MC simulations (HPC use)** with the use of pairs (SPECT/CT) to extract the 3D dose maps (dosimetry per organ).

iii. **Develop the final GenAI model to generate the 3D Dose Map (HPC use)**, providing as input the pairs 2D planar/3D dose maps.

iv. **Apply XAI methods** for analysing the parameters most contribute to the generation of the dose maps for increasing their robustness/trustworthiness.

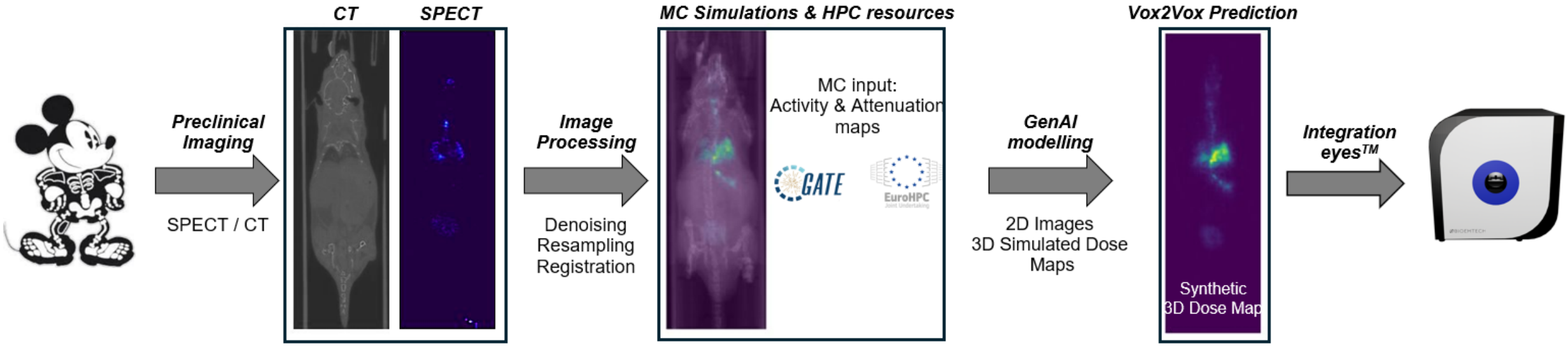
ΔOSIMETREYE – Concept (δ-eye)

FORTISSIMO PLUS



FORTISSIMO PLUS

ΔOSIMETREYE – Methodology

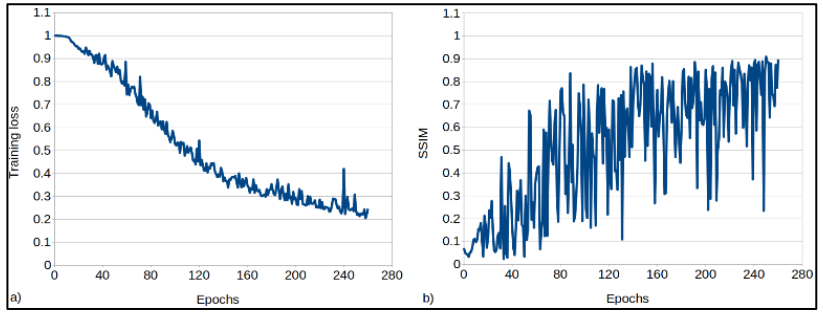


ΔOSIMETREYE – GenAI / XAI approaches

GenAI

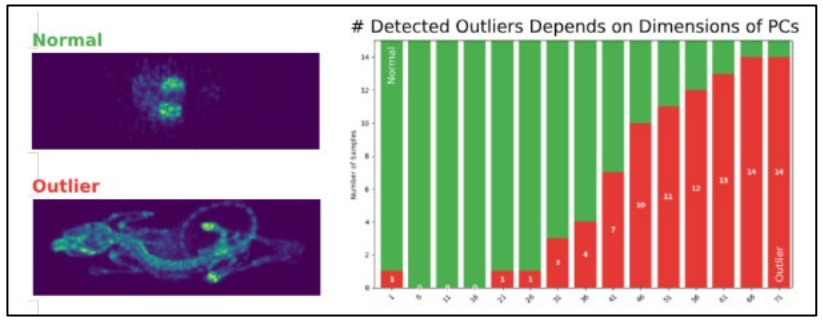
- ✓ **Task 2D-to-3D translation** → 3D-to-3D to preserve spatial context
- ✓ **Model architecture**: 3D cGAN (Vox2Vox)

  (Discriminator loss: MSE Based - Generator loss: Dice & MSE based)
- ✓ **Dataset**: ~1000 image pairs 70/15/15 (training/validation/test)
- ✓ **Data Augmentations:** random elastic deformation, rotation, zoom, flip
- ✓ **Patch training** on overlapping patches to mitigate edge artifacts
- ✓ **Performance metrics**: MSE, PSNR, SSIM (indicative value ~0.89)
- ✓ **HPC**: Training on Leonardo HPC



XAI / Outlier Detection: Hybrid method

- ✓ **Out-of-distribution input images** are identified through first-order statistical features on the fitted Gaussian Mixture Model
- ✓ **CLIP model & PCA:** captures the majority of variance → real-time quality control, robust & reliable models



ΔOSIMETREYE – HPC ACCESS

FORTISSIMO PLUS

EuroHPC JU Access:
AI and Data-Intensive Applications

Project Duration:
Jan. 2025 – Jan. 2026

Resources allocation:
50000 node hours on Leonardo Booster (CINECA, Italy)

Usage:
Execution of both MC simulations & GenAI models

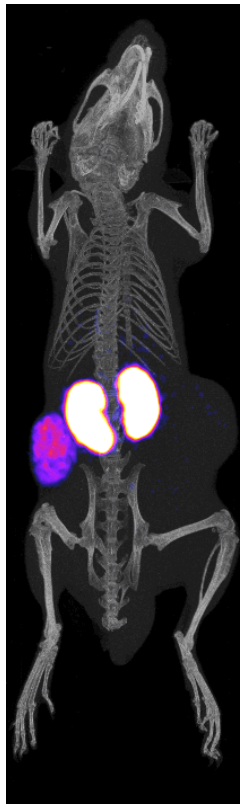
LEONARDO CINECA

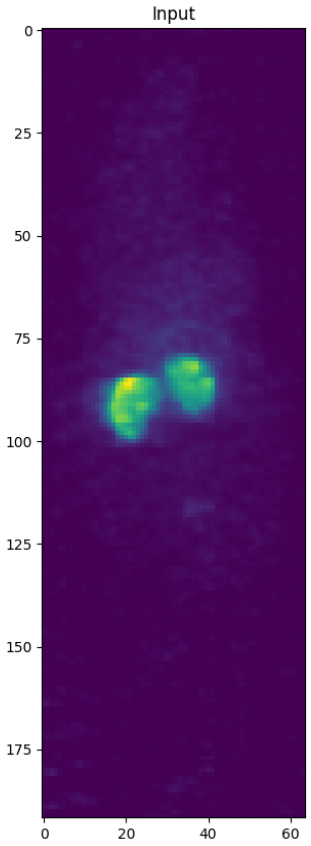
EuroHPC Joint Undertaking

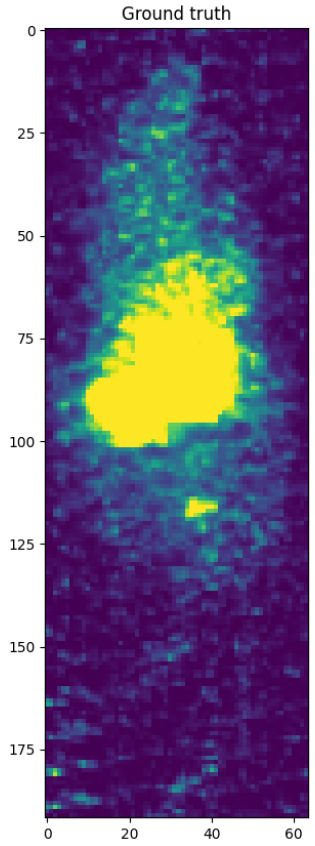
ΔOSIMETREYE – The tool

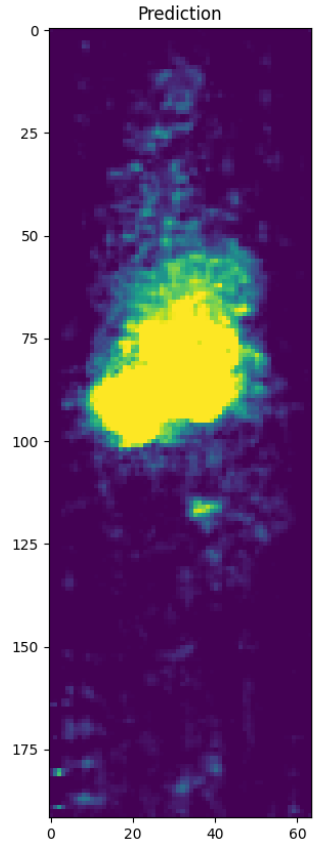
FORTISSIMO PLUS

Perform **GenAI Dosimetry** in parallel with **in vivo Imaging**!!!









ΔOSIMETREYE – Business impact

FORTISSIMO PLUS

- ✓ **Product Innovation & Differentiation**:

  Development of **δ-eye** for 3D preclinical dosimetry based on 2D imaging, significantly reduces animal models usage.

- ✓ **Commercial Expansion**:

  The company's **product portfolio can now include AI-enhanced dosimetry**, appealing to pharmaceutical, CRO, and clinical customers worldwide.

- ✓ **Strategic Collaborations**:

  HPC competency served as a magnet for **further collaborations and EU funding proposals**.

- ✓ **Efficiency & ROI:**

  Faster simulation times and increased automation led to **reduced product development cycles**.

  Simultaneous preclinical imaging and dosimetry **improves research throughput**, minimizing experimental time and cost.

Alethia

A Polish start-up establishing its XAI platform (SaaS) on real healthcare use –cases.

grnet

Exploit their HPC expertise in industrial sector – real business case for NCC.

HPC INTEGRATION AT BIOEMTECH

FORTISSIMO PLUS

BIOEMTECH's journey with High-Performance Computing (HPC) showcases a clear evolution from initial experimentation to routine operations & strategic usage

- ✓ **Initiation**: Starting with the FF4EuroHPC, the company applied MC simulations for personalized pediatric dosimetry. This marked the first systematic integration of HPC into BIOEMTECH's workflows.
- ✓ **Internal Capacity Building**: Following successful project outcomes, BIOEMTECH extended HPC usage with own funds for HPC resources, facilitating scalability and continuity. Dedicated personnel were trained to ensure sustained adoption and integration of HPC methodologies.
- ✓ **Broad Use Across Teams**: HPC is now embedded in multiple domains:
  - Monte Carlo simulation datasets
  - AI training for predictive models
  - XAI/GenAI systems integration
  - Synthetic data generation

- ✓ **Support Ecosystem**: Partnership with EuroCC Greece (NCC) and active participation in EuroHPC-JU resource applications further highlight how HPC became an operational routine, not just an occasional tool.

ACKNOWLEDGMENTS

FORTISSIMO PLUS

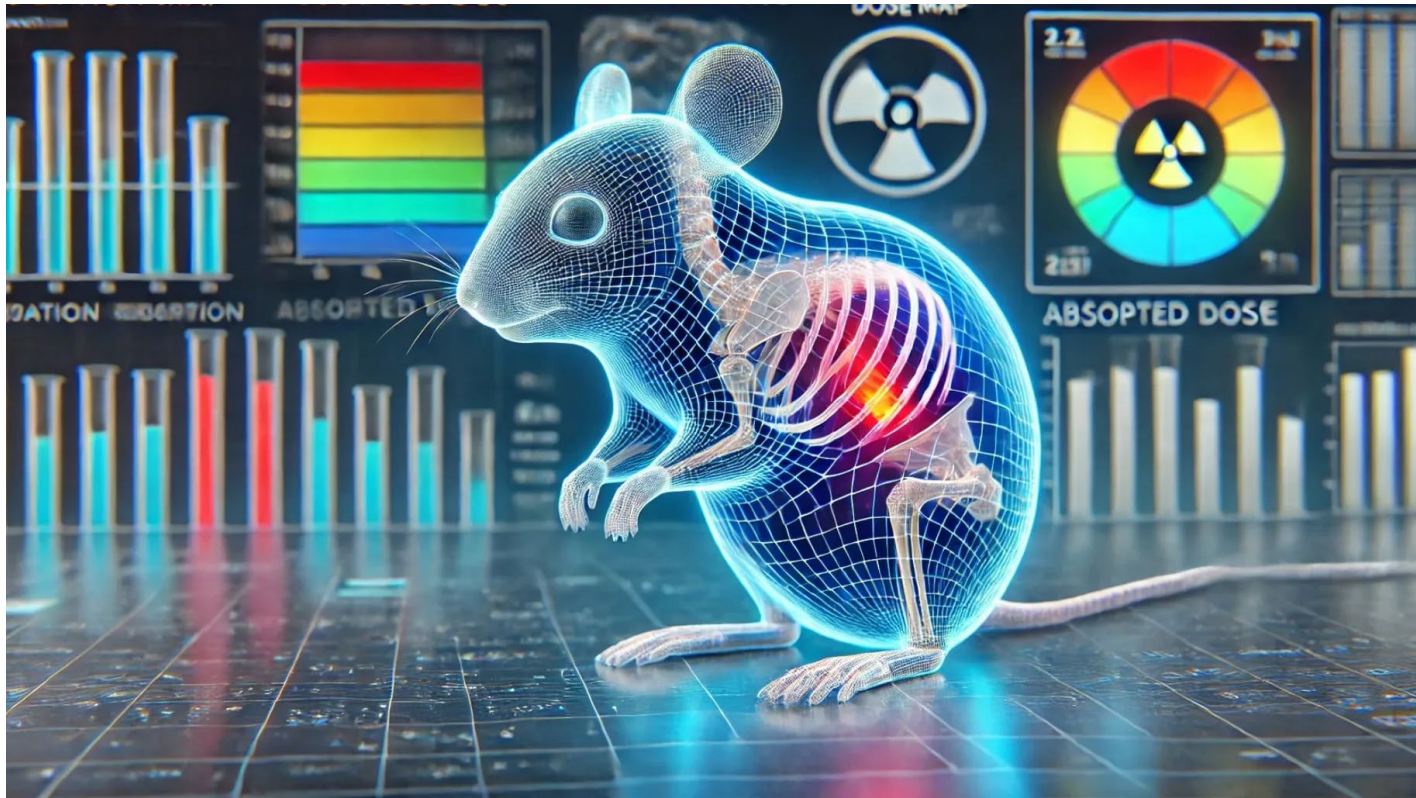
Authors acknowledge the **EuroHPC Joint Undertaking** for awarding this project access to the EuroHPC supercomputer **LEONARDO**, hosted by **CINECA** (Italy) and the LEONARDO consortium through an EuroHPC AI and Data-Intensive Applications Access call.

Innovation Study **DOSIMETREYE** has received funding through the **FFplus project**, which is funded by the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101163317. The JU receives support from the **European Union's Horizon Europe Programme**.

Thank you

Panagiotis Papadimitroulas
Associate Prof. Biomedical Informatics
Medical Department, Univ. of Thessaly

Software Advisor & Co-Founder
BIOEMTECH
panpap@bioemtech.com
www.bioemtech.com



EUROHPC USER DAYS

30 SEPTEMBER
1 OCTOBER 2025
COPENHAGEN

Artificial Intelligence for Science

Dynamic Resource Management for Exascale in ADMIRE Framework

Prof. Jesús Carretero
Universidad Carlos III de Madrid

ADMIRE
malleable data solutions for HPC

uc3m | Universidad Carlos III de Madrid

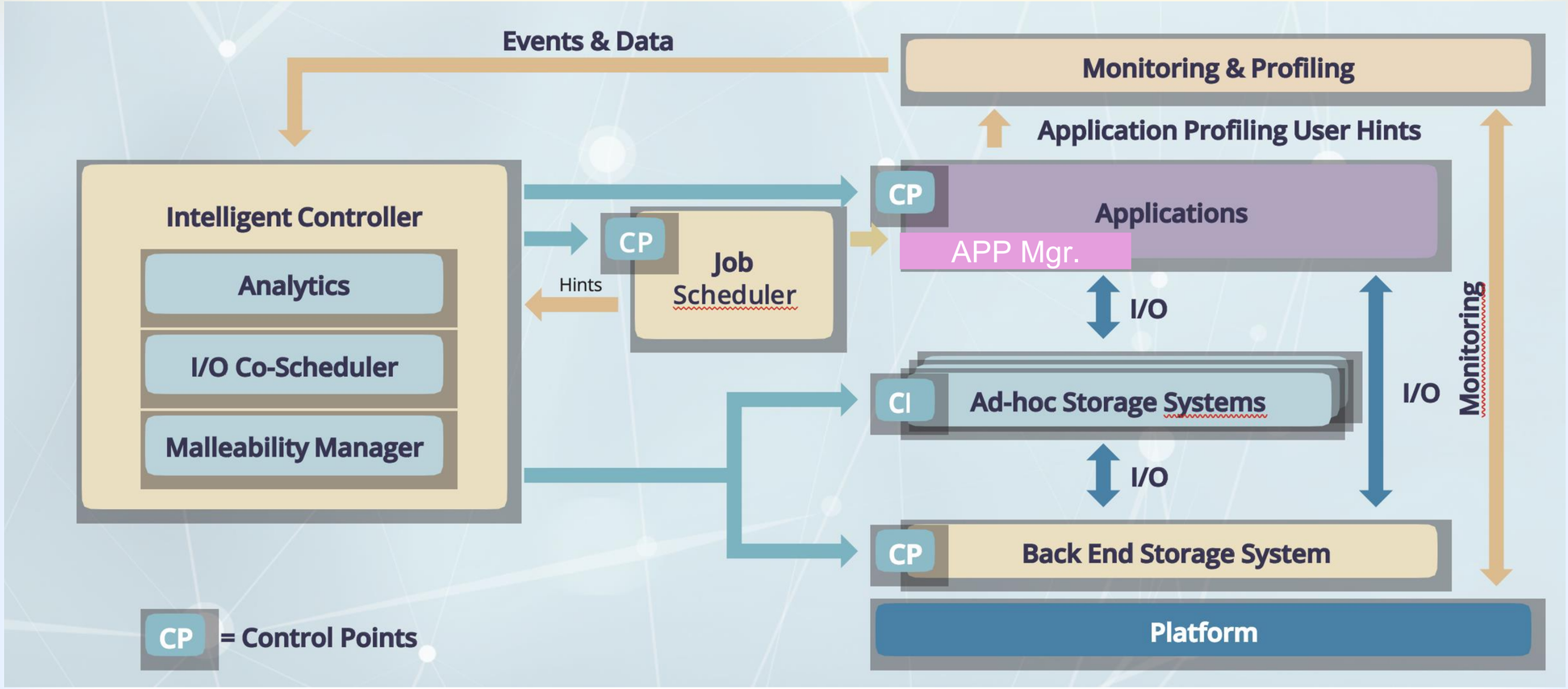
https://admire-eurohpc.eu/

Predictive modelling and simulation

- AI enables researchers to develop predictive models and simulations that simulate complex systems, predict outcomes, and explore hypotheses in silico.
- ADMIRE framework AI-driven modeling techniques, such as neural networks, Bayesian models, and reinforcement learning.
  - Can model nonlinear relationships, optimize parameters, and predict resource usage for changing resource allocation
  - To answer application and system situation in runtime.

Observation → AI models/predictors -> Action

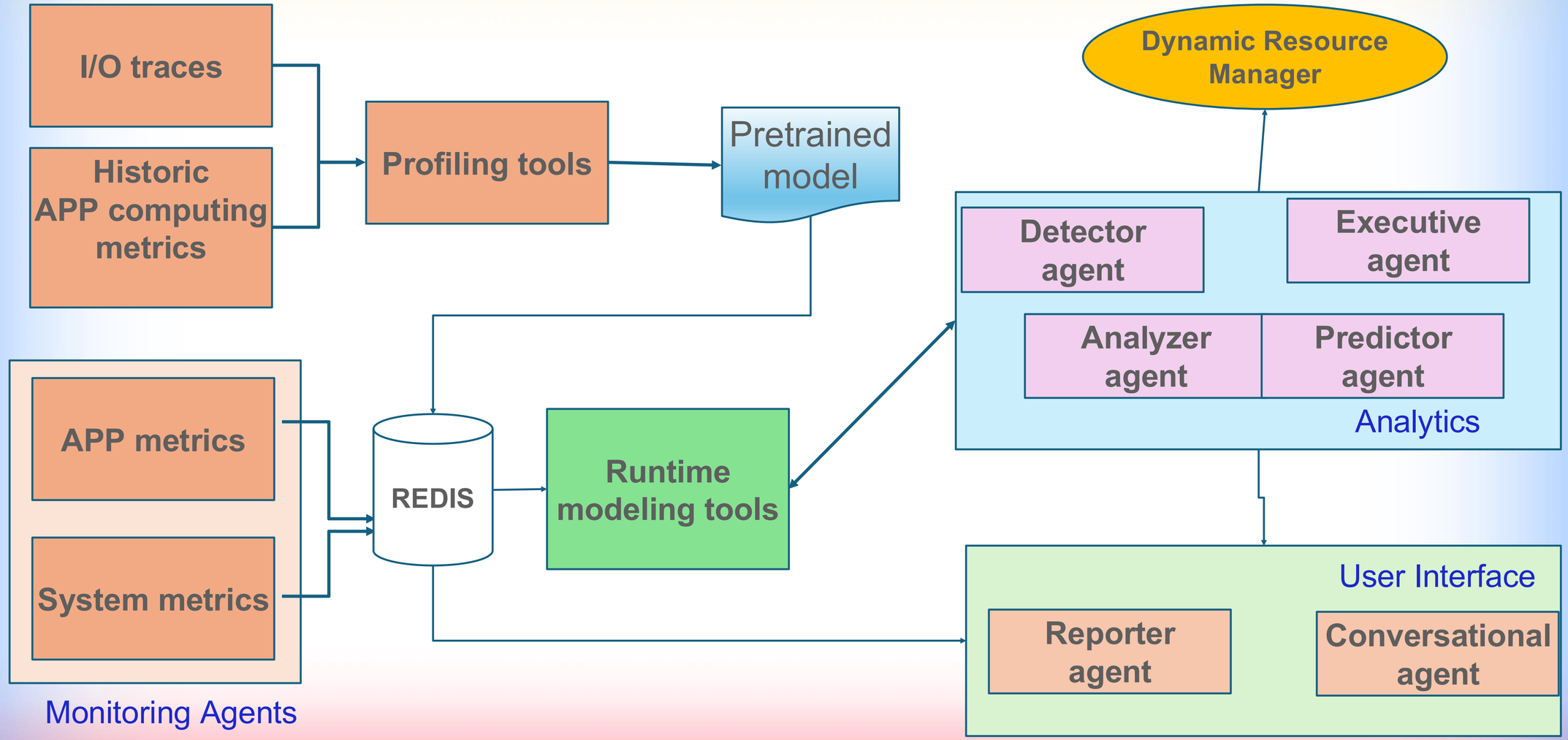
Adaptivity framework



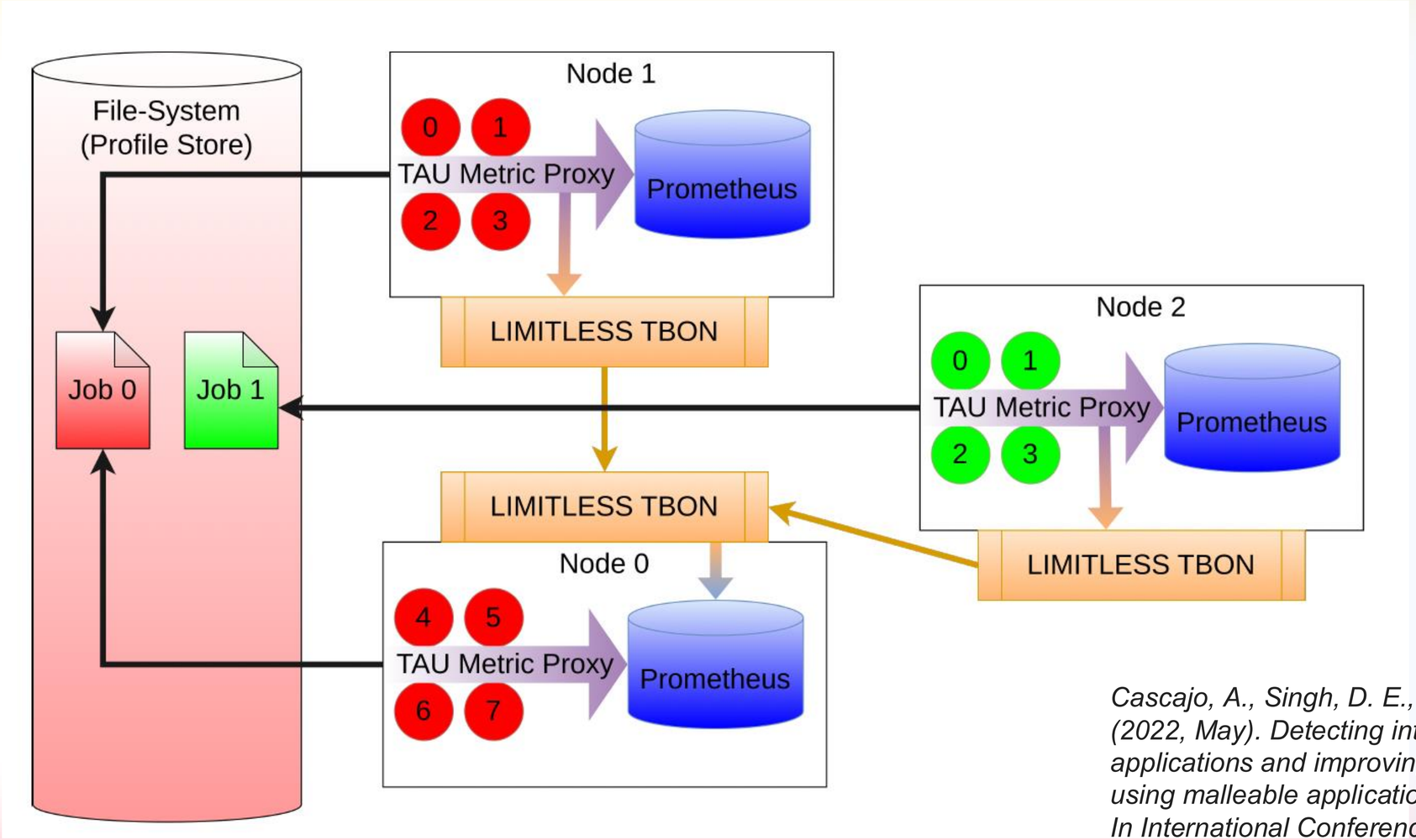
Carretero, J., Garcia-Blas, J., Aldinucci, M., Besnard, J. B., Acquaviva, J. T., Brinkmann, A., ... & Wolf, F. (2023, May). Adaptive multi-tier intelligent data manager for Exascale. In Proceedings of the 20th ACM International Conference on Computing Frontiers (pp. 285-290).

DRM AI tools - The Intelligent Controller

https://hpcioanalysis.zdv.uni-mainz.de/



Per-job profiles



Cascajo, A., Singh, D. E., & arretero, J. (2022, May). Detecting interference between applications and improving the scheduling using malleable application proxies. In International Conference on High Performance Computing (pp. 129-146). Cham: Springer International Publishing.

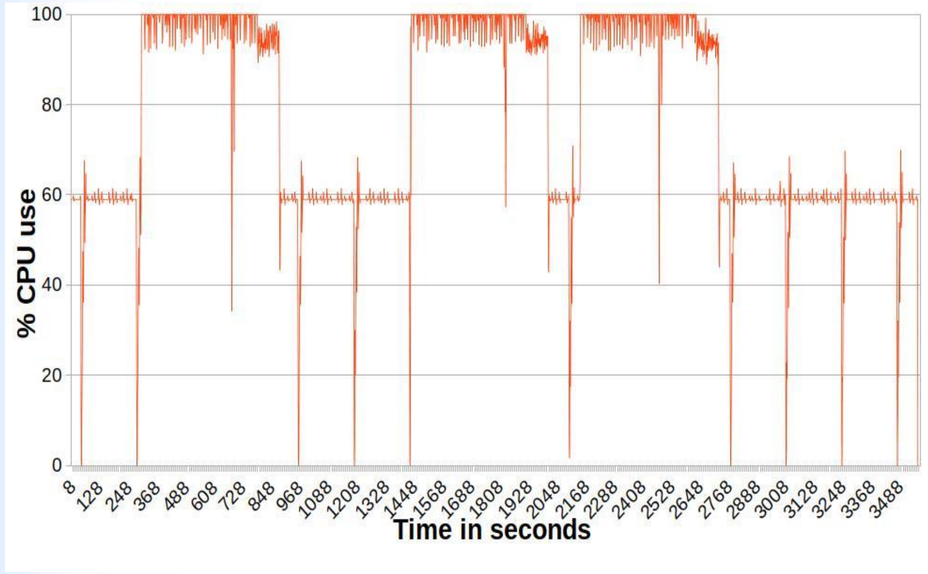
AI agents

- Metrics:
  - Automatic monitoring and management of HPC and Cloud platforms without the need for manual processing.
- Analyzer & predictor:
  - Modelling app and system state and predict near-future needs
- Executive:
  - Automate certain actions with AI Agents (alerts to app and admin, DRM to scheduler, …)

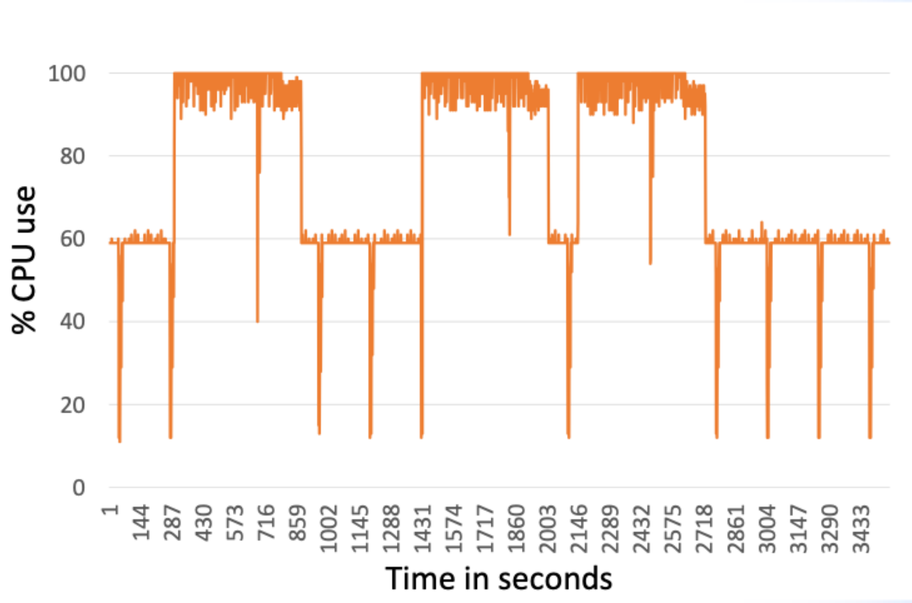
AI-based DRM

- Serve as source of meaningful and summarized information for dynamic resource management
  - Dynamic systems require up-to-date information about the state of the system and knowledge about whether a malleability operation is feasible or not.
    - If a job can be expanded but there is not enough resources in the system, it is better not to try to expand than ask the resource manager for resources (METRIC-AI can reduce waiting times due to unnecessary resource requests).
  - The Scheduler and the predictor agents can provide hints about dynamic resource actions to improve the overall performance:
    - If app1 is going to start a compute-intensive phase and app2 is going to end with an I/O post-processing phase, app1 can use app2's resources that are not needed (shrink app2 + expand app1).
    - This is supported by the predictions based on historical data and performance models of the applications.

Example - BT-IO benchmark configured as Class C running in 8 nodes



a) Real pattern



b) Predicted pattern

Analytics accuracy

- Accuracy of the different prediction algorithms three use cases expressed as percentage of correct value predictions.

	Use case 1	Use case 2	Use case 3
Pattern Matching (PM)	25.2%	25.2%	24.7%
Historical Window (HW)	61.5%	82.7%	50.0%
Neural Network (NN)	99.5%	93.3%	98.0%
Machine Learning (ML)	90.8%	90.5%	88.5%

Interference detection is complex

- Usually, every application runs in exclusive nodes
  - All processes repeat the same pattern (CPU intensive, …) with the same phases.
  - Processes can be very memory o comm intensive, limiting the usage of node's resources
    - Not all resources fully exploited
    - But they are usually due to the same app.
- I/O may suffer are interferences in servers from many apps running simultaneously
  - Usually I/O system is not in scale with computing resources -> bottleneck
  - Many operations running simultaneously (e.g. checkpointing) may collapse I/O
- Is very important to **predict/coordinate** this situations to help the IC to solve them by:
  - Rescheduling ad-hoc resources, dephasing I/O applications, limiting app bandwidth, etc.

EUROHPC USER DAYS

30 SEPTEMBER
1 OCTOBER 2025
COPENHAGEN

Artificial Intelligence for Science

Dynamic Resource Management for Exascale in ADMIRE Framework

Prof. Jesús Carretero
Universidad Carlos III de Madrid

ADMIRE
malleable data solutions for HPC

uc3m | Universidad Carlos III de Madrid

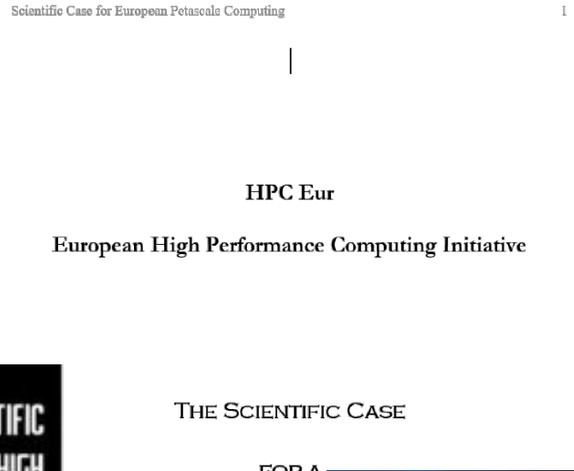
https://admire-eurohpc.eu/

PRACE Scientific and Innovation Case for HPC in Europe

Florian Berberich, PRACE aisbl

Scientific Case - History

- PRACE (and precursors, e.g. HET) has a long tradition in creating the European Scientific Case
  - 2004 HPC Scientific Case based on national scientific cases
  - Organised by HPC in Europe Taskforce
  - Important goal: Bring HPC to the politician attention
- 2012 – 1. Update (PRACE)
- 2018 – 2. Update (PRACE)
- Scientific Case
  - shows break througs and highlights in different fields
  - justifies the present use of HPC resources
  - predicts future needs (both on capacity and architecture)
  - is instrumental for ensuring future funding for HPC
  - will serve as input for the EuroHPC Multi Annual Strategic Plan


Scientific Case for European Petascale Computing 1

HPC Eur

European High Performance Computing Initiative

THE SCIENTIFIC CASE

FOR A

OPEAN SUPER COMPUTI

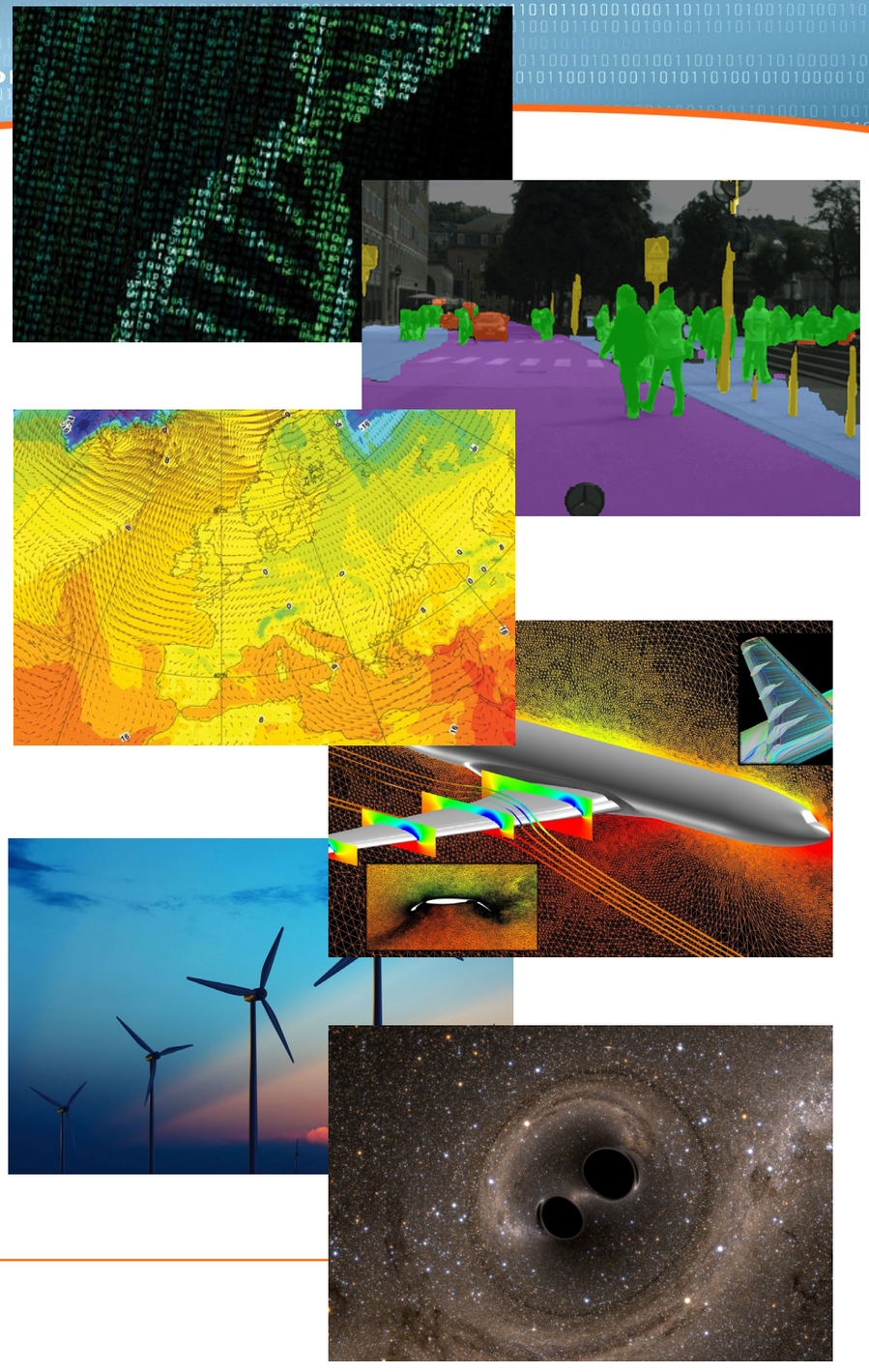
2025 Update: Scientific and Innovation Case

- PRACE decided to be more inclusive and integrative regarding industry and businesses
  → Scientific **and Innovation** Case
- Kick-off workshop: 20 October 2023, Paris
- Follow up workshop: 4 February 2025, Brussles – PRACE Intersection Seminar
- TOC with specific topics was drawn up
- Important aspects:
  - Artificial Intelligence & Machine Learning, Large Language Models
  - Data
  - Links to RI, EOSC

Scientific and Innovation Case

Topics / chapters

- *Fundamental Science*
- *Life Sciences*
- *Climate, Weather and Earth Science*
- *Energy – Net-zero Strategy*
- *Materials*
- *Engineering & Industrial applications*
- *Social Sciences and Humanities*
- *HPC challenges*
  - *Services for Research Infrastructure*
  - *Complexity & Data*
  - *AI – next-generation computing*
  - *Beyond exascale*



Panel Members

- 6 – 8 panel members
- Diverse composition of experts covering
  - All relevant topics within the domain
  - Different nationalities, affiliations, gender, etc.
- Editor in Chief: Prof. Hartmut Wittig, University Mainz

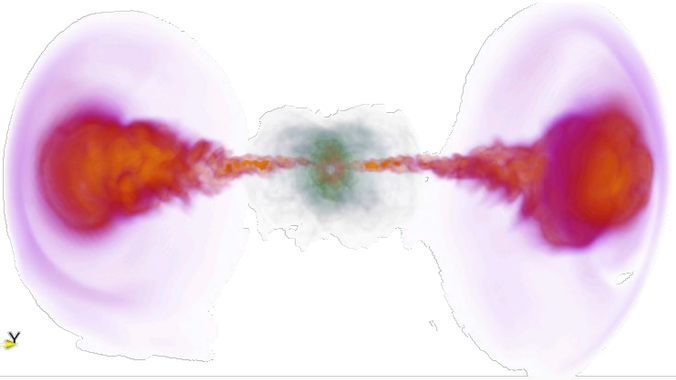
Panel	Chair		Panel members
Climate, Weather and Earth Sciences	Boris Kaus	Mainz, Germany	P. Dueben, D. Folini, A. Folch
Energy—Net-zero Strategy	Frank Jenko	Garching, Germany	J. Proll, M. Wilczek, M. Becoulet, F. Fiuza, E. Audit
Engineering and Industrial Applications	Simone Hochgreb	Cambridge, UK	G. Lacaze, B. Cuenot, T. Brunschwiler, E. Chatzi, G. Wells, D. Lohse, S. Glockner, J.H. Walther, J. Harting
Fundamental Sciences	Luciano Razzola	Frankfurt, Germany	F. Karsch, G. Endrődi, L. Del Zanna, M. Girone
Life Sciences	Erik Lindahl	Stockholm, Sweden	P. Coveney, A. Valencia, R. Mercado, G. Hummer, Z. Cournia
Materials	Matej Praprotnik	Ljubljana, Slovenia	N. Marzari, G. Csányi, A. Tkatchenko, M. Barborini, L. Ghiringhelli, V. Harmandaris, J. Zavadlav, I. Pagonabarraga, T. Bereau, N. López
Social Sciences and Humanities	Simon Scheidegger	Lausanne, Switzerland	J. Fernandez-Villaverde, B. Mazoyer, E. Ollion, E. Schulz, Y. Yang,
Artificial Intelligence and Quantum Computing			Petros Koumoutsakos, Bertil Schmidt, Timothée Lacroix

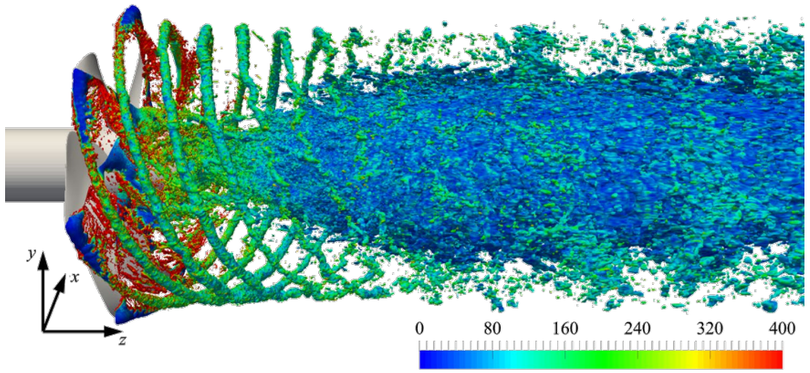
Topics to be addressed in Panels

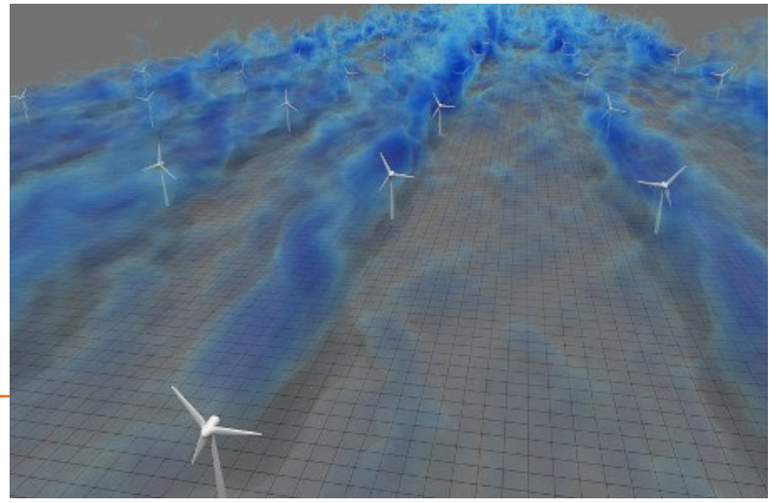
- Research highlights
- Prediction of compute power / storage
- Role of **Artificial Intelligence**
- Identification of HPC challenges
  - Link to Research Infrastructure (CERN, EBRAINS, DestinE, SKA,…)
  - Link to Strategic Research Agenda of European Technology platform for HPC (ETP4HPC)
  - Link to EGI Federation (Enabling Grids for E-Science); SPECTRUM project (Data-intensive science in Europe)

State of the Art Examples

- Fundamental Sciences
  - Astrophysics – Relativistic MHD simulation of a Gamma-Ray Burst jet through realistic binary neutron star merger
- Engineering
  - Instantaneous isosurfaces of pressure coefficient coloured by vorticity magnitude for a propeller in the wake - cavitation
- Energy – Net Zero Strategy
  - Large-eddy simulation of an entire wind farm

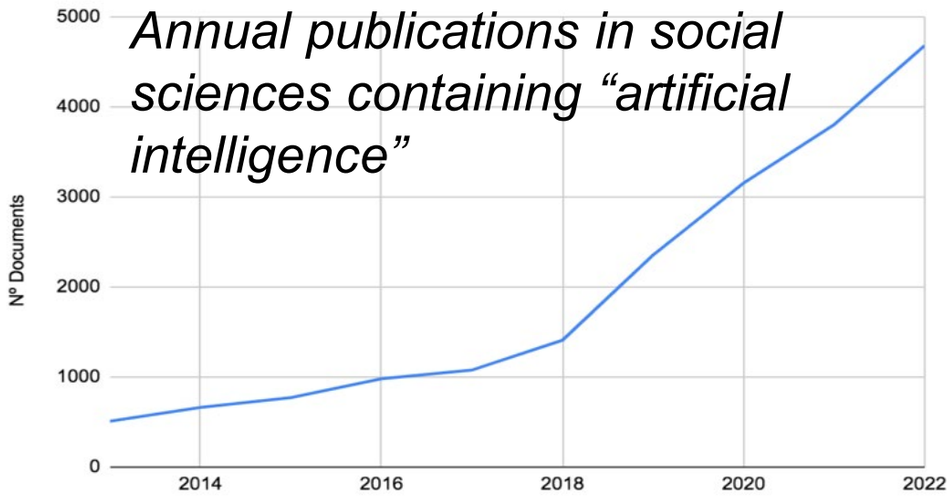






AI Aspects

- Energy:
  - *AI-accelerated simulations* – e.g. materials for batteries and PV materials
  - *Synthetic Data Generation*
- Engineering:
  - *generalized ML for partial differential equations*
  - *AI/ML for reduced order models from HPC simulation results for use in real time control of large scale systems*
  - *AI/ML to automate the extraction of results from solutions*
- *Social Sciences*
  - *Text classification and information extraction*
  - *Visual Data and Sound Analysis*
  - *Synthetic Data Generation*



Some preliminary Recommendations

- Expand Exascale Application Support
- Accelerate GPU Transition
- Maintain CPU-Only Capacity
- Integrate AI and HPC
- Expand HPC Training

Work in progress

Scientific and Innovation Case for HPC in Europe will be published by end of 2025

THANK YOU FOR YOUR ATTENTION

www.prace-ri.eu