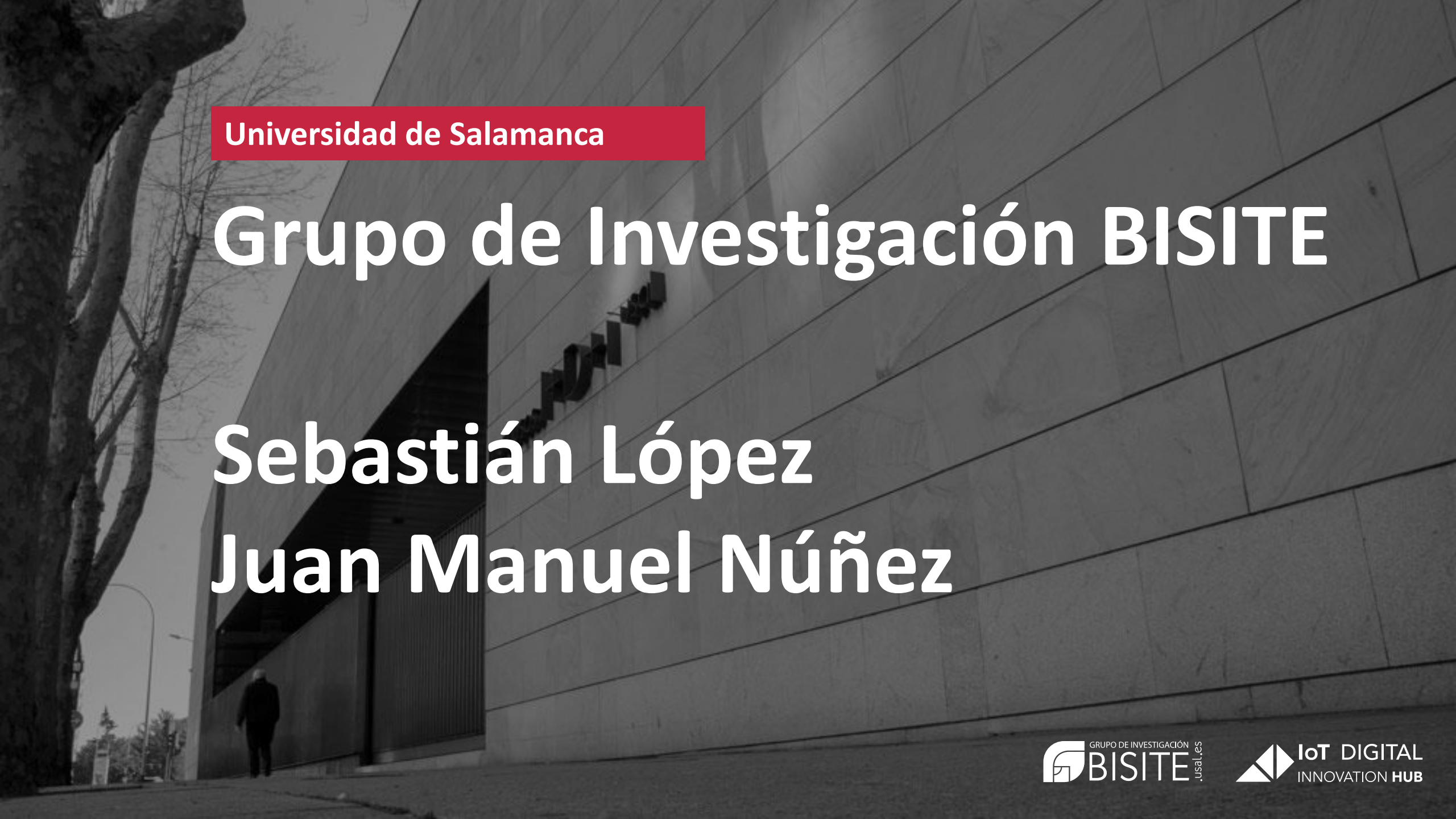


Universidad de Salamanca

Grupo de Investigación BISITE

Sebastián López
Juan Manuel Núñez

GRUPO DE INVESTIGACIÓN BISITE .usal.es

IoT DIGITAL INNOVATION HUB

IA GENERATIVA LLAMA 2

The next generation of our open-source large language model

- Mark Zuckerberg se ha asociado con Microsoft Azure para presentar su herramienta de IA generativa LLAMA 2, considerada rival de ChatGPT.
- Esta versión incluye los pesos del modelo y el código de inicio para los modelos lingüísticos LLAMA pre-entrenados y ajustados, con parámetros que van desde 7B a 70B

Meta

Modelos

Llama 2: Open Foundation and Fine-Tuned Chat Models

Hugo Touvron* Louis Martin† Kevin Stone†
Peter Albert Amjad Almahairi Yasmine Babaei Nikolay Bashlykov Soumya Batra
Prajjwal Bhargava Shruti Bhosale Dan Bikel Lukas Blecher Cristian Canton Ferrer Moya Chen
Guillem Cucurull David Esiobu Jude Fernandes Jeremy Fu Wenyin Fu Brian Fuller
Cynthia Gao Vedanuj Goswami Naman Goyal Anthony Hartshorn Saghar Hosseini Rui Hou
Hakan Inan Marcin Kardas Viktor Kerkez Madian Khabsa Isabel Kloumann Artem Korenev
Punit Singh Koura Marie-Anne Lachaux Thibaut Lavril Jenya Lee Diana Liskovich
Yinghai Lu Yuning Mao Xavier Martinet Todor Mihaylov Pushkar Mishra
Igor Molybog Yixin Nie Andrew Poulton Jeremy Reizenstein Rashi Rungta Kalyan Saladi
Alan Schelten Ruan Silva Eric Michael Smith Ranjan Subramanian Xiaoqing Ellen Tan Binh Tang
Ross Taylor Adina Williams Jian Xiang Kuan Puxin Xu Zheng Yan Iliyan Zarov Yuchen Zhang
Angela Fan Melanie Kambadur Sharan Narang Aurelien Rodriguez Robert Stojnic
Sergey Edunov Thomas Scialom*

GenAI, Meta

Llama 2

MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B	Model architecture:	Data collection for helpfulness and safety:
13B	Pretraining Tokens: 2 Trillion	Supervised fine-tuning: Over 100,000
70B	Context Length: 4096	Human Preferences: Over 1,000,000

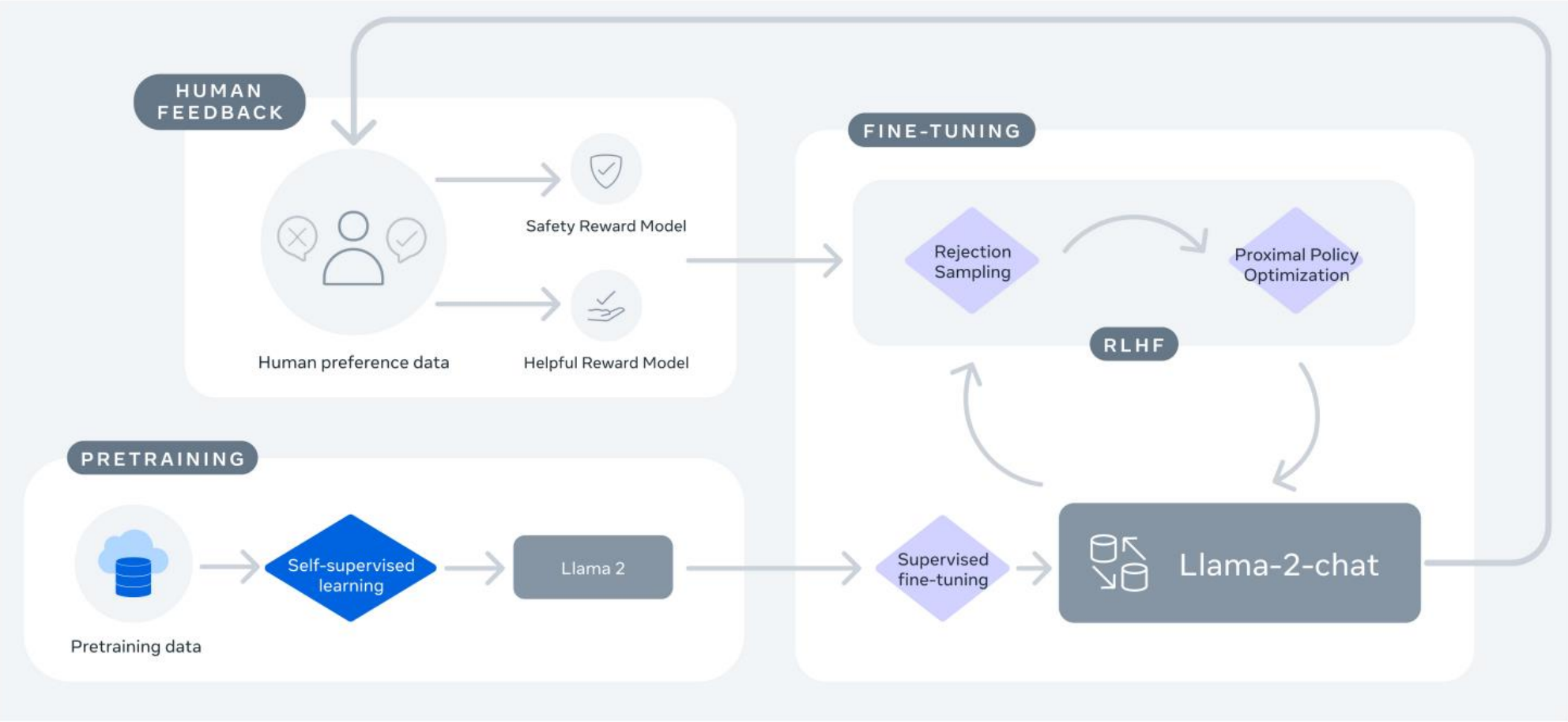
Pila de tecnologías

Hardware de entrenamiento

		Time (GPU hours)	Power Consumption (W)	Carbon Emitted (tCO_2eq)
Llama 2	7B	184320	400	31.22
	13B	368640	400	62.44
	34B	1038336	350	153.90
	70B	1720320	400	291.42
Total		3311616		539.00

GPU NVIDIA A100 con Tensor Core 80Gb

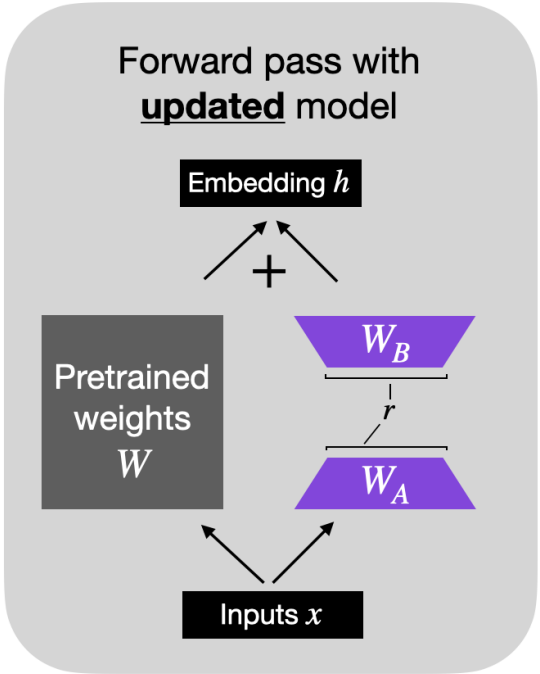
Metodología Fine-Tuning



Arquitectura entrenamiento

LoRA weights, W_A and W_B, represent ΔW

$$\breve{w} = w + \ A \ * B$$



DEMO

GRACIAS

bisite@usal.es

fb.com/bisite

@bisite_usal

bisite.usal.es/es

bisite_usal

bisite-research-group

Grupo de investigación BISITE