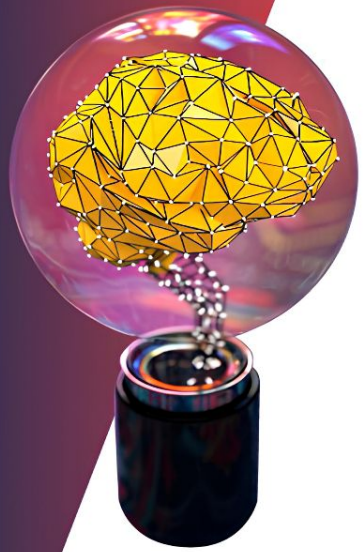
VII Congreso **XoveTIC Talento científico**

citic

UNIVERSIDADE DA CORUÑA



ADAPTANDO GRANDES MODELOS DE LINGUAXE PARA LINGUAS INFRARREPRESENTADAS

Eliseo Bao
Anxo Pérez
Javier Parapar

A Coruña, a 17 de outubro de 2024

XUNTA DE GALICIA | CONSELLERÍA DE EDUCACIÓN, CIENCIA, UNIVERSIDADES E FORMACIÓN PROFESIONAL

CENTROS gusR

Cofinanciado por la Unión Europea

Fondos Europeos

IRLAB

TÁBOA DE CONTIDOS

0I - INTRODUCIÓN

INTRODUCIÓN

SISTEMAS DE IA CONVERSACIONAL

ubicuos hoxe en día: ChatGPT, Claude, Gemini…

Modelos de **xeración** de texto adestrados en conxuntos **masivos** de texto. **Predín os seguintes *tokens*** (palabras) dado un *input*

Modelos *fine-tuned* (adaptados) para entender **instrucións** e seguir **conversas** (pregunta-resposta) con usuarios

XUNTA DE GALICIA | CONSELLERÍA DE EDUCACIÓN, CIENCIA, UNIVERSIDADES E FORMACIÓN PROFESIONAL
Cofinanciado por la Unión Europea
Fondos Europeos

INTRODUCIÓN

~ 500.000.000

persoas usan sistemas de IA conversacional

~ 7.000

linguas e falas no mundo

60

linguas oficialmente soportadas por ChatGPT (non o galego)

XUNTA DE GALICIA | CONSELLERÍA DE EDUCACIÓN, CIENCIA, UNIVERSIDADES E FORMACIÓN PROFESIONAL
Cofinanciado por la Unión Europea
Fondos Europeos

PROBLEMAS!

PROBLEMAS

Os LLMs máis avanzados **priorizan as linguas amplamente faladas, descoidando linguas minoritarias** e con poucos recursos, como o galego

Isto agrava aínda máis a **infrarrepresentación** destas linguas no mundo dixital e **afecta negativamente** aos seus falantes

O pre-adestramento dun LLM require unha **gran cantidade de recursos computacionais**, xa que estes modelos conteñen miles de millóns de parámetros

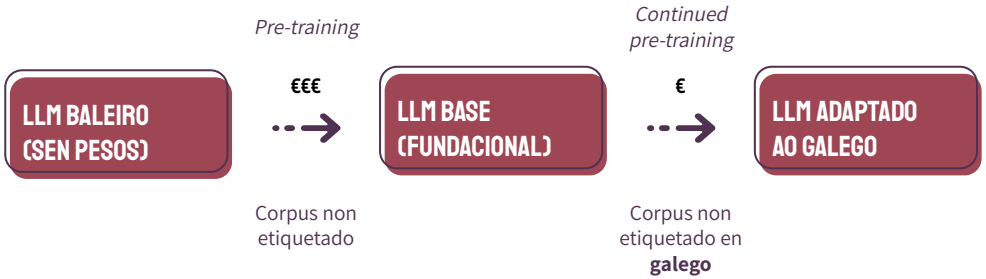
XUNTA DE GALICIA | CONSELLERÍA DE EDUCACIÓN, CIENCIA, UNIVERSIDADES E FORMACIÓN PROFESIONAL
Cofinanciado por la Unión Europea
Fondos Europeos

PROPOSTA

PROPOSTA

CONTINUED PRE-TRAINING

- Incorporar **novo coñecemento** nun LLM **pre-adestrado** para que o modelo poida, por exemplo, aprender un novo idioma
- Importante! Isto **non é *fine-tuning***. O *fine-tuning* implica usar datos etiquetados para personalizar un modelo para unha tarefa específica



02 - MÉTODO

MÉTODO

TRAINING

- Modelo base: `Llama-3.1-8B-Instruct`
- ***Continued pre-training*** con *CorpusNós*[1]
- Adaptador Low-Rank Adaptation (**LoRA**)

TEST

- Inferencia sobre o dataset **Alpaca Instruction** traducido ao **galego**
- **Cuantitativo** con métricas de similaridade
- **Cualitativo** usando un LLM como xuíz

[1] Iria de Dios Flores et al. (2024). Nos_CorpusNOS-GL: Galician Macrocorpus for LLM training. In Proceedings of the 16th International Conference on Computational Processing of Portuguese

XUNTA DE GALICIA | CONSELLERÍA DE EDUCACIÓN, CIENCIA, UNIVERSIDADES E FORMACIÓN PROFESIONAL
Cofinanciado por la Unión Europea
Fondos Europeos

03 - RESULTADOS

CUANTITATIVO

- Similaridade

	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
Llama	20.97	28.77	11.37	20.43
Llama GL	**23.13**	**30.42**	**11.95**	**21.84**
Carballo	2.60	19.41	4.93	2.94

CUALITATIVO

- LLM como xuíz

	Llama GL	Llama	Carballo	Tie	Malf.
Llama GL Vs. Llama	**41.6%**	37.6%	-	17%	3.8%
Llama GL Vs. Carballo	**72.4%**	-	14%	11.6%	2%
Carballo Vs. Llama	-	**43.8%**	33.6%	20%	2.6%

04 - CONCLUSIÓNS

CONCLUSIÓNS

I

O *continued pre-training* adapta de **maneira efectiva** un LLM a unha nova lingua.

Porén, é necesario un **corpus de texto** o suficientemente grande.

II

Aínda que o *continued pre-training* seguido de *fine-tuning* para instrucións pode ser **máis formal**, o *continued pre-training* sobre un modelo conversacional demostrou ser eficaz

III

A tecnoloxía debe servir como **ponte** e non como **barreira**, **empoderando** aos falantes de todas as linguas no ámbito dixital

XUNTA DE GALICIA | CONSELLERÍA DE EDUCACIÓN, CIENCIA, UNIVERSIDADES E FORMACIÓN PROFESIONAL
Cofinanciado por la Unión Europea
Fondos Europeos

Visita o noso repositorio e descarga o modelo!!!

GRAZAS!

Eliseo Bao, Anxo Pérez e Javier Parapar

XUNTA DE GALICIA | CONSELLERÍA DE EDUCACIÓN, CIENCIA, UNIVERSIDADES E FORMACIÓN PROFESIONAL
Cofinanciado por la Unión Europea
Fondos Europeos