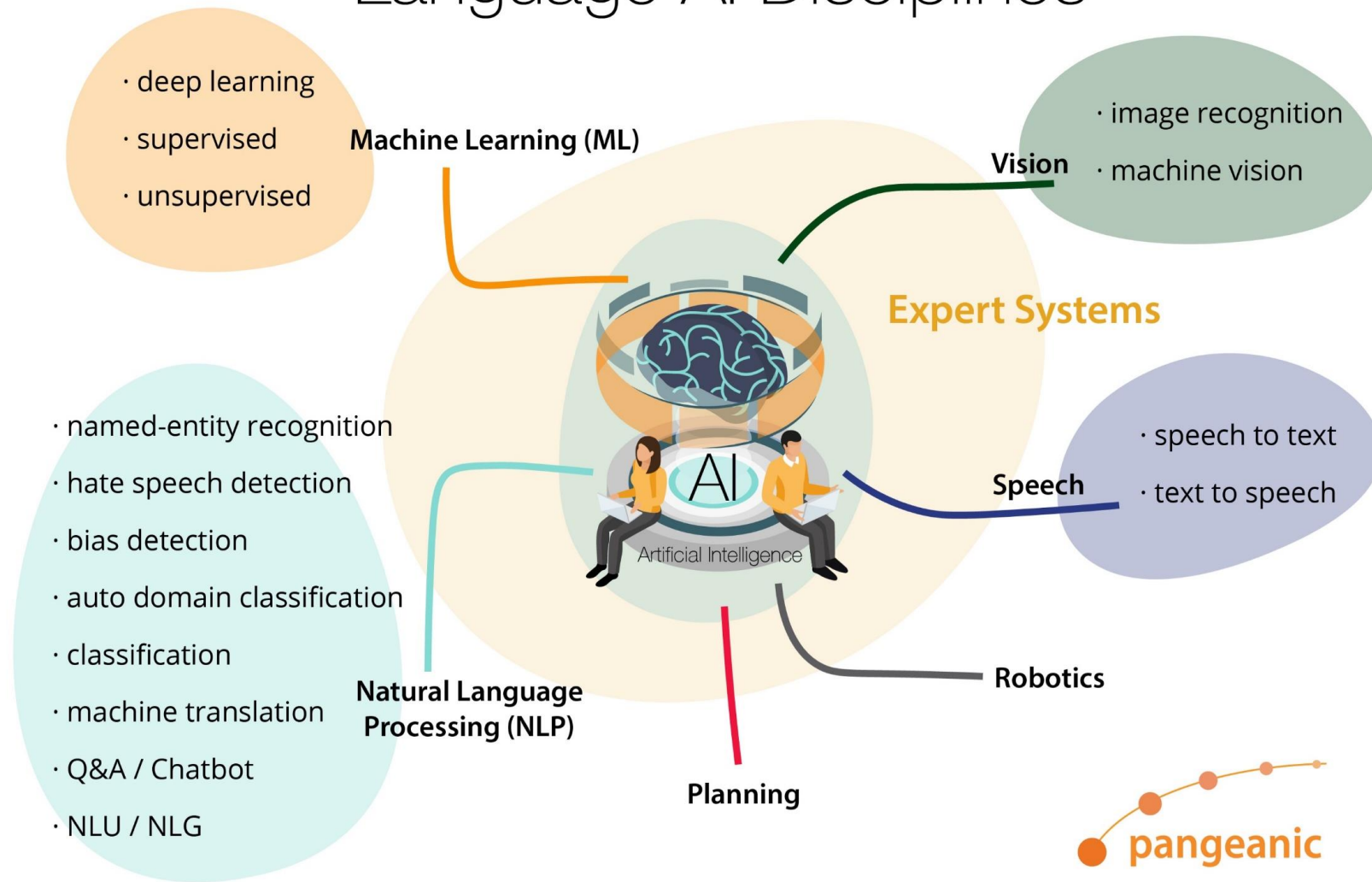# Exploiting LLMs for what they weren't designed for

The use case at EFE, the Largest News Agency

in the Spanish Language
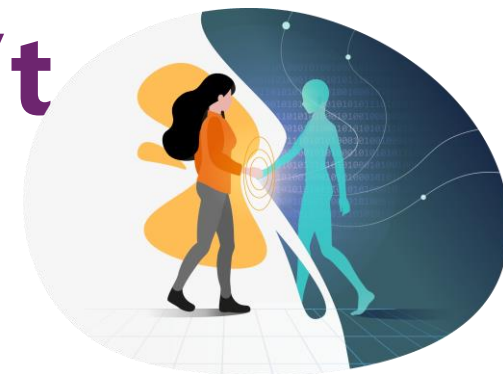
#GALA2024

Language AI Disciplines

· deep learning

· supervised

· unsupervised

Machine Learning (ML)

Vision

· image recognition

· machine vision

Expert Systems

· named-entity recognition

· hate speech detection

· bias detection

· auto domain classification

· classification

· machine translation

· Q&A / Chatbot

· NLU / NLG

Natural Language Processing (NLP)

AI

Artificial Intelligence

Speech

· speech to text

· text to speech

Robotics

Planning

pangeanic

WHO IS EFE

- Agencia EFE is a Spanish international news agency, the major Spanish-language multimedia news agency and the world's fourth largest wire service after the Associated Press, Reuters, and Agence France-Presse

- 3M million news items per year

- 3,000 journalists from 60 nationalities, operating 24/7 from more than 180 cities in 120 countries and with four editorial desks in three continents: Madrid, Bogotá, Cairo (Arabic), and Rio de Janeiro (Portuguese)

- Website contains more tan 55,000 pages of news

WHO IS PANGEANIC

- Renowned company in the field of Language Technology, with a specialization in machine translation and natural language processing (NLP). Pangeanic has consistently evolved from its LSP origins to become a pioneer in the language industry, combining Artificial Intelligence with human expertise and ingenuity

- Mission: To create efficient solutions that optimize the value and processing of language data with an emphasis on privacy

- Looks at MT as a Machine Learning / Data Science challenge

- Other NLP: Email (data) classification, Anonymization, automated subtitling, speech, data-for-AI, chatbot systems

How it all began...

- Won tender for Machine Translation in 2021

- General translation engines with a "journalism flavour" for internal use and web publishing

- Full API Management for integration in journalists processes

- Journalists reviews / post-edit often divert from source to add specific or "personal" touch

Exploiting LLMs for what they weren't designed for

Yann LeCun – Bavarian Academy of Science, Munich 09/29/23

Workflow 2021-2023... "traditional"!!)

(2023 is

Baseline > Customization > Re-training

ES / EN – PT-Br, CA

NMT

Los fármacos contra la obesidad han supuesto una auténtica revolución por su eficacia. Y no solo se investiga en ese campo, también en otros como en el de encontrar un arma contra la segunda enfermedad neurodegenerativa más prevalente, el párkinson. Uno de los últimos estudios publicados ha arrojado resultados con un efecto "pequeño" pero con conclusiones "interesantes".

10-20 words / sentence

2023/24

- **42,3M** words machine translated in 2023

- Over **111M** words since 07.2021

- Mostly web articles, internal articles for EFE's internal team of journalists

- No NMT customization

- Journalists' preferences led to further customization in 2023
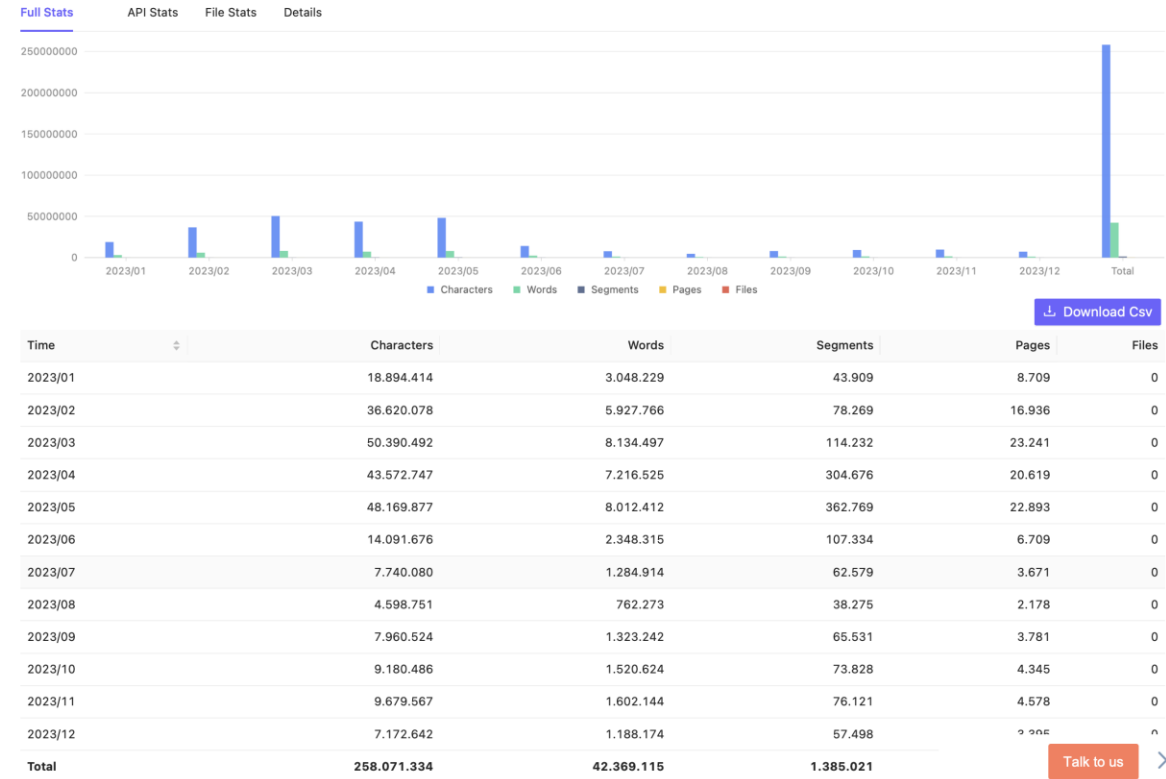
Issues

So much for MTQE.....

1 sentence in 111M words from
MT in PT-Br caused a complaint
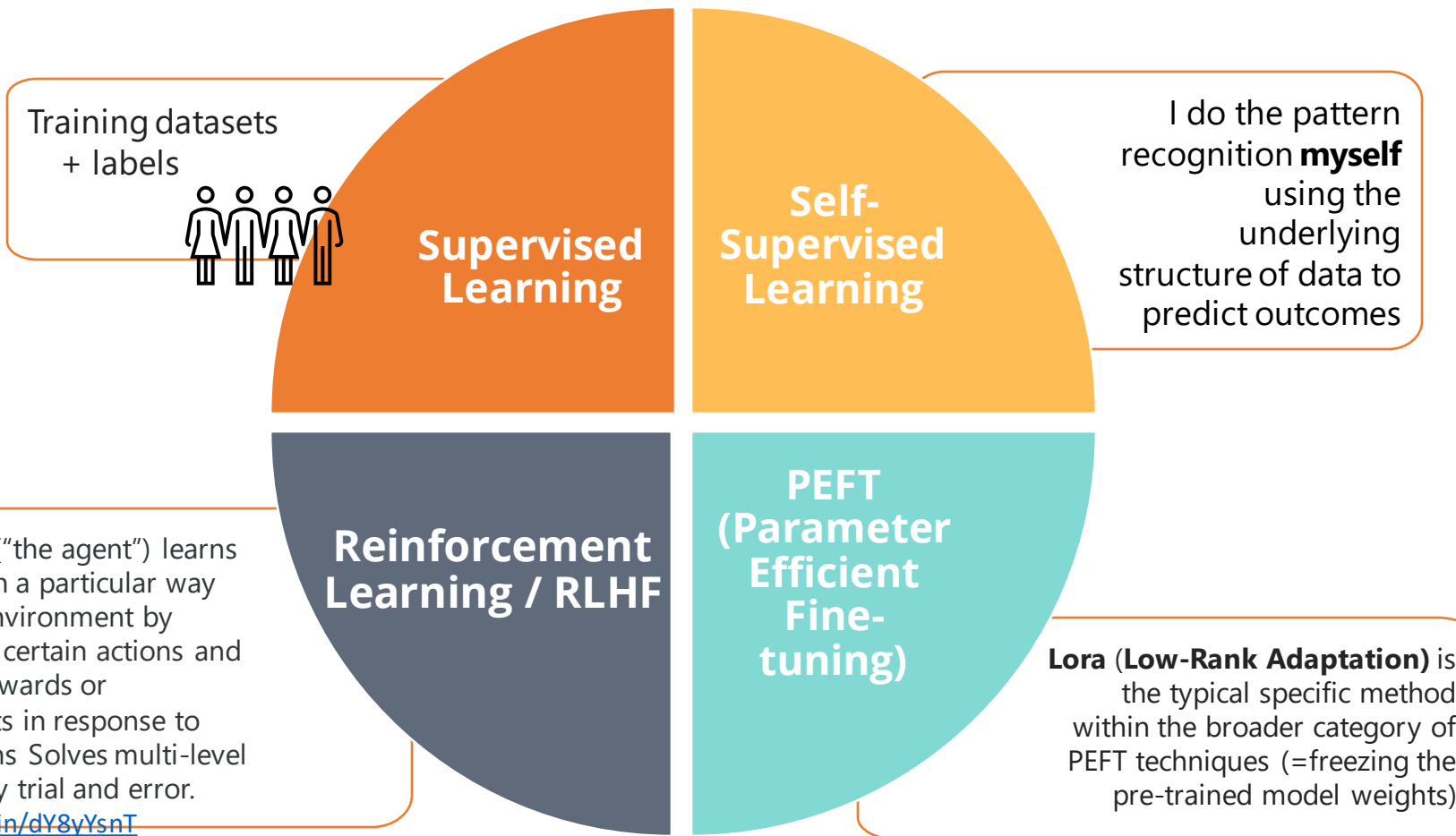
Customize / Retrain or use LLM?

Exploiting LLMs for what they weren't designed for

2023: Dolly, Vicuna, Llama 2 (Mixtral 8x7B late '23)

- **Creating a chatbot from structured data (RAG) -> Automated Post-Editing**
- Entity Recognition (Anonymization/NER detection)
- MTQE – Machine Translation Quality Estimation (MT)
- Live Text/Document Classifier
- ChatGPT for MT did not beat our own NMT client engines (+ privacy issues and

4 ways to fine-tune (re-train) an LLM

Training datasets + labels

I do the pattern recognition **myself** using the underlying structure of data to predict outcomes

Supervised Learning

Self-Supervised Learning

Reinforcement Learning / RLHF

PEFT (Parameter Efficient Fine-tuning)

An algorithm ("the agent") learns to behave in a particular way within an environment by performing certain actions and receiving rewards or punishments in response to those actions Solves multi-level problems by trial and error. https://lnkd.in/dY8yYsnT

Lora (**Low-Rank Adaptation)** is the typical specific method within the broader category of PEFT techniques (=freezing the pre-trained model weights)

LoRA = Train 6% for 98% cost reduction

MTQE: Machine Translation Quality Estimation

OBJECTIVE

- **Use Llama13B (7B not enough) & Mixtral to estimate MTQE. ChatGPT as benchmark.**

- Find a correlation coefficient between the TER calculated with the reference Human translation vs. the one estimated by Llama-2& Mixtral

METHODOLOGY

- Collect a dataset with human revision of machine translations.

- Calculate Translation Error Rate (TER).

- Create a set of inputs with the source, machine translation, human post edition and TER.

- Fine tune model with 20% of prompts removed from the human post edition, to make the model learn to estimate without gold standard.
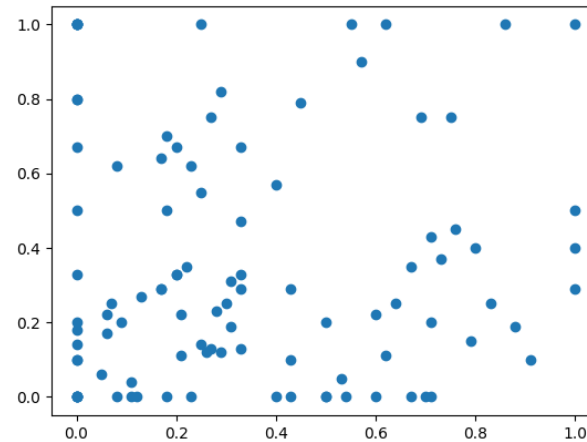
MTQE: Machine Translation Quality Estimation

Llama13B

- Verbosity but higher accuracy than 7B

- In 7B, Pearson correlation coefficient not significantly different than 0 (0.0796102897964038)

In 13B, we find a correlation
similar to ChatGPT: 0,3!!!

Mixtral 8x7B

- Verbosity but higher accuracy than Llama 13B!!

- With minimum LoRA, Mixtral can "estimate" MTQE and rank in TER as good as ChatGPT 3.5

- Better than prompting: more segments processed, predictable output.

Creating a chatbot from structured data (RAG) -> Automated Post-Editing

OBJECTIVE

- **In-domain Question Answering** from technical user manuals.

- Evaluate different training methodologies (dataset size, epoch, full finetunig vs

METHODOLOGY

- Automotive documentation (full website)

- H1, H2 H3 = "How to" and body is the answer

- Fine-tune model

- Evaluation KPIs

RESULTS

Not useful	35/150
Response appears helpful, but additional text is repeated	56/150
Useful, correct answer	49/150
No answer	12/150

Creating a chatbot from structured data (RAG) -> Automated Post-Editing

SUPERSEDED 12/2023

- **Vector databases**

- **If query is translation – Why not post-edit? Glossaries?**

- **Q&A can be turned into TMX reference material**

- **Glossaries are OK, full TMX retrieval gets lost in vector data space**

Today

- Japan's COVID-19 Border Re...
- Miyamoto in Japanese Baseb...
- Apartment Viewing" The con...
- Apartment Viewing" The con...

Last 7 Days

- Apartment Viewing" The con...

Previous 30 Days

- US-UK Dual Taxation Inquiry
- Gaza Conflict: Recent Develo...
- Translation Company Size in ...
- Largest Translation Company...
- EU Ministers' Solidarity Visit t...
- Russian Ukraine Invasion Det...
- Putin's Ukraine War Details
- Putin Mention

Over 30 days ago

- Gaza Conflict Overview
- General Knowledge About Gi...
- Agencia EFE Inquiry" I chose ...
- Spanish Corporate Tax Rates

m.berranz@pangeanic.com...

Eco Chat
Assistant with access to documents from your Connected Sources.

Connected Sources:

🌐 Web   📄 File

- [ ] 🔖 Mercadona Tickets
- [ ] 🔖 SmartFactory
- [ ] 🔖 EFE
- [ ] 🔖 ECO
- [x] 🔖 Yomiuri ✓
- [ ] 🔖 MT_Best_practices
- [ ] 🔖 hideaki
- [x] 🔖 Pangeanic Glossary ENJA ✓

Filter   Currently applied: 🔖 Pangeanic Glossary ENJA ✕   🔖 Yomiuri ✕

Covid experience in Japan

Customization of translation workflow with LLM

Custom MT to
Custom Post-editing
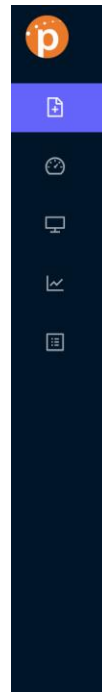
Original: Los fármacos contra la obesidad han supuesto una auténtica revolución por su eficacia. Y no solo se investiga en ese campo, también en otros como en el de encontrar un arma contra la segunda enfermedad neurodegenerativa más prevalente, el párkinson. Uno de los últimos estudios publicados ha arrojado resultados con un efecto "pequeño" pero con conclusiones "interesantes".

Translation
medicamentos contra a obesidade representaram uma verdadeira revolução pela sua eficácia. E não é só nesse campo que se investiga, mas também noutros, como encontrar uma arma contra a segunda doença neurodegenerativa mais prevalente, o Parkinson. Um dos últimos estudos publicados apresentou resultados com um efeito "pequeno" mas com conclusões "interessantes".

{ "translation": "Os fármacos contra a obesidade representaram uma autêntica revolução pela sua eficácia. E não é só nesse campo que se investiga, mas também noutros, como encontrar uma arma contra a segunda doença neurodegenerativa mais prevalente, o Parkinson. Um dos últimos estudos publicados apresentou resultados com um efeito 'pequeno' mas com conclusões 'interessantes'." }

Unstructured Data Classification

OBJECTIVE

- **Classify emails spam / no spam**

- Llama2 7B & 13B / Mixtral vs naïve-bayes traditional method

METHODOLOGY

- 1,000 tagged emails (20 untagged = test dataset)

- Create specific prompt

- Fine-tune model

- Evaluation KPIs vs naïve-bayes

RESULTS

- Bayes: 33,76 mails/sec **CPU** Accuracy 0,76; Recall 0,61

- Llama 2 13B: Memory constraints, 1,45 mails/sec **GPU** Accuracy 0,75; Recall 0,82

- Mixtral Memory constraints, 1,5 mails/sec **GPU** **Accuracy 0,79; Recall 0,84**

#GALA2024

NER for ANONYMIZATION SUPPORT

OBJECTIVE

- **Evaluate and improve NER datasets and services for anonymization**

- Detect tag entities (person, place, organization) and identify missing entities

METHODOLOGY

- Dataset from repositories: 1200 sentences, 50 for eval.

- Create specific prompt

- Fine-tune model

- Evaluation: Missing tags without fine-tuning (baseline) vs FT

(Case 1 – Missing tags)
```
### input:      The Encyclopedia Americana w
published in 2006 .
### NER:      The Encyclopedia Americana was la
published in <DATE>2006</DATE>.
Missing tags:
    -"Encyclopedia Americana" is an ORG entit
must be tagged as <ORG>Encyclopedia American
```
_____

_____

(Case 2 - Full tagged)
prompt: The task is to ...(large instruction)
input: The Great Alföld contains the basin of the Tisza Riv
branches .
NER: The <LOC>Great Alföld</LOC> contains the basin of t
River</LOC> and its branches .

NER for ANONYMIZATION SUPPORT (Llama2)

RESULTS

Without Fine-Tuning:

- Inference time: 20.4 segs

- Accuracy: 55.9% (to detect missing entities)

Details:

- Doesn't follow any format, too verbose!

- Model adds other tags (as <Number>)

- Tagging mistakes (<LOC>Pangeanic</LOC>)

With Fine-Tuning

- Inference time: 5.6 segs

- Accuracy: 72.1% (to detect missing entities)

Details:

- It generates the exactly output format given in the training.
- Sometimes is redundant: <LOC>Republic of Moldova</LOC> y <LOC>Moldova</LOC>
- The model tags with DATE entity expression as "the time," "9/11," and "two years".
- <PER>Norman</PER> instead <PER>General Norman</PER>
- 2% of the items were incorrectly tagged. For instance:
- The three writers formed the bulk of the poets at the court of <LOC>Yusuf I</LOC>. (Should be PER)

Challenges

Speed & Cost (but new opportunities)

Scale up through the organization

+ languages

+ areas (higher value)

+ AI tech in other language areas

#GALA2024

Conclusions

1. Fine-tuned LLMs can made to "learn" MTQE because **they can "taught"** to classify, recognize entities (for glossary), translate (expensive and slow)

2. Slower than specific engines

3. Good at post-editing and applying glossaries if combined with RAG + specific prompting

3. Verbosity is a **solvable** problem

- in estimation
- In translation
- In post-editing => HYBRID APPROACH MT + RAG LLM

Results

0% issues since Q4 2023

Over 1,000 journalists involved in the process saving 100's of translation time / review time per week

¡Gracias!

Manuel Herranz
PANGEANIC

Juan Manuel Ruiz
EFE