

The background is a dark blue space filled with a complex network of thin, light blue lines that form a grid or perspective view. Scattered throughout are numerous small, glowing particles in shades of blue and red, some appearing as larger, out-of-focus bokeh lights. The overall effect is one of a high-tech, digital environment.

# Artificial Intelligence Industry applications

# AI-Powered Knowledge Graphs for Science-Based Journalism

AKA meet ClimateBlocks!



Jenny Romano – CEO, The Newsroom



**Climate science is a very challenging topic to report on**

>500

new research papers on climate are published every day.\*

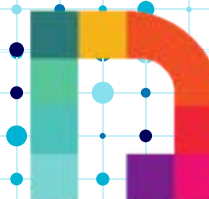
\*based on last 5 year average data from OpenAlex



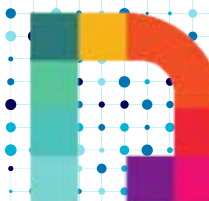
In a day



In a week

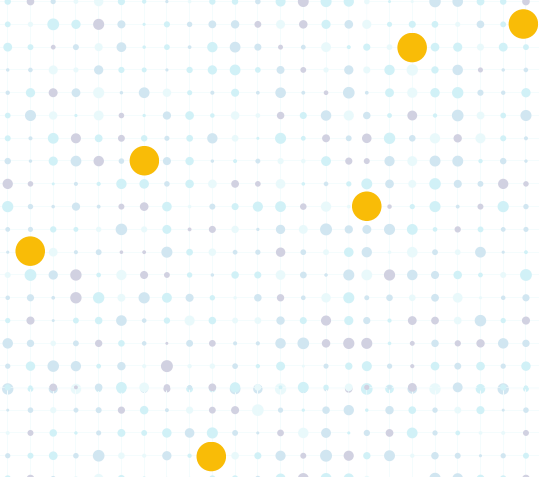


**In a month**



How do journalists find the  
building blocks they need,  
when they need them?





author  
affiliation  
institution  
paper title  
published date  
journal



publisher  
funders  
citations  
language  
locations

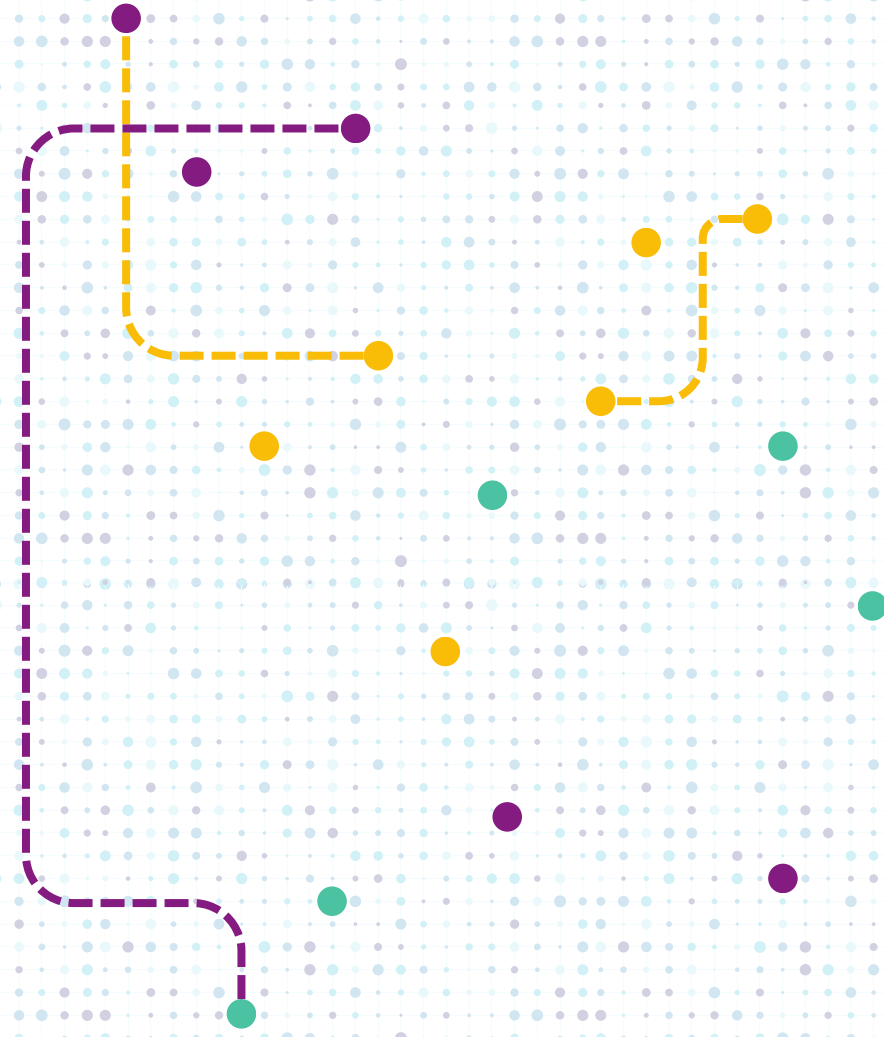


events  
concepts  
measurements  
methodology  
quotes



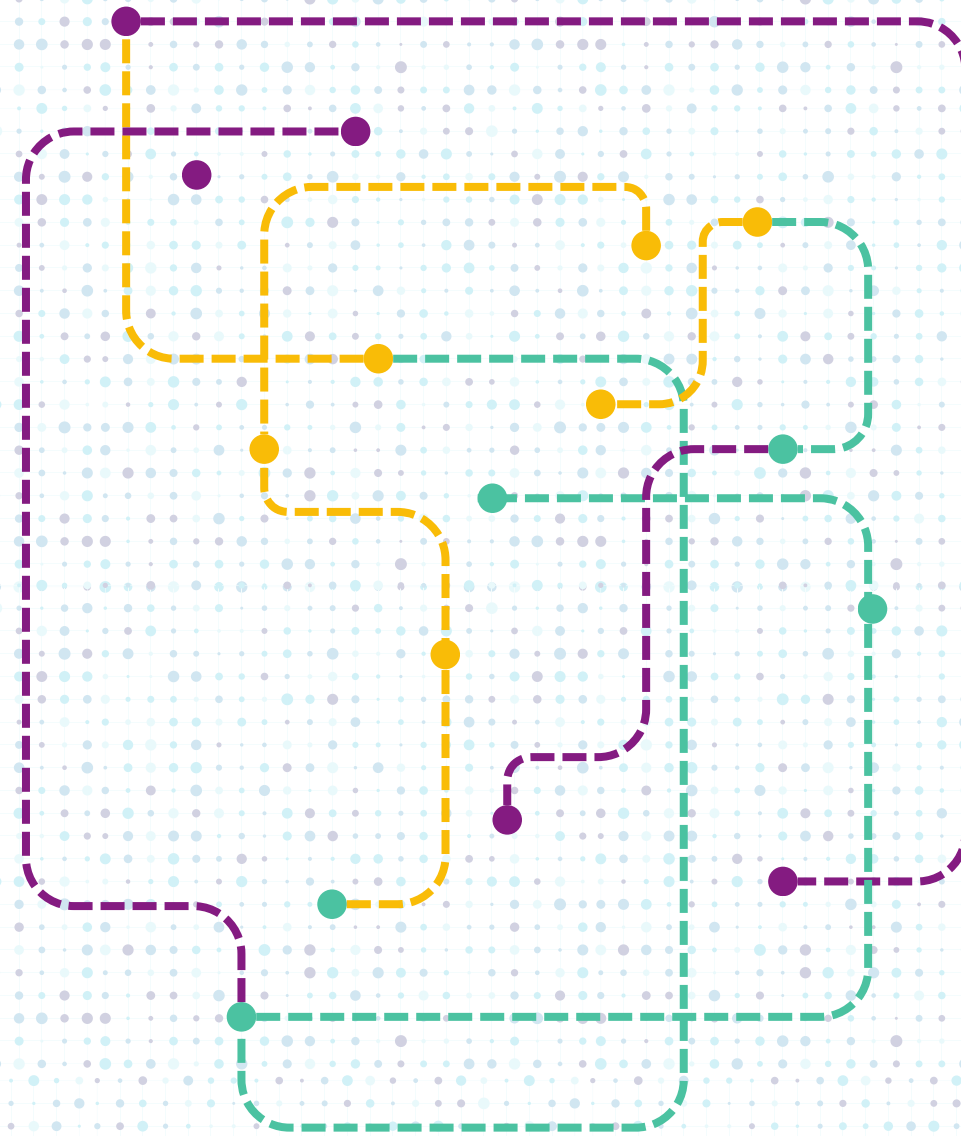
And how do they  
**connect the dots?**





And how do they  
**connect the dots?**





And how do they  
**connect the dots?**



- ? How many research papers from the past five years show a link between climate change and extreme weather events?
- ? Who are the top 10 scientists studying climate adaptation?
- ? How has the scientific consensus on this link evolved over time?



# Can't we do this with existing AI tools?

ChatGPT 4o

How many research papers from the past five years show a link between climate change and extreme weather events? If you cannot answer just say "I can't help you with that question."

I can't help you with that question.

Sources

The Washington Post

Democracy Dies in Darkness

Subscribe

Sign in

How many research papers from the past five years show a link between climate change and extreme weather events?

**Response**

☆ Answers are AI-generated from our published reporting. Please verify by consulting the provided articles. Learn more

Sorry, but we could not generate an answer to your question. This product is still in an experimental phase. Please consult the articles below which may be relevant to your query.

Was this answer useful? 

Yes

Unsure

No

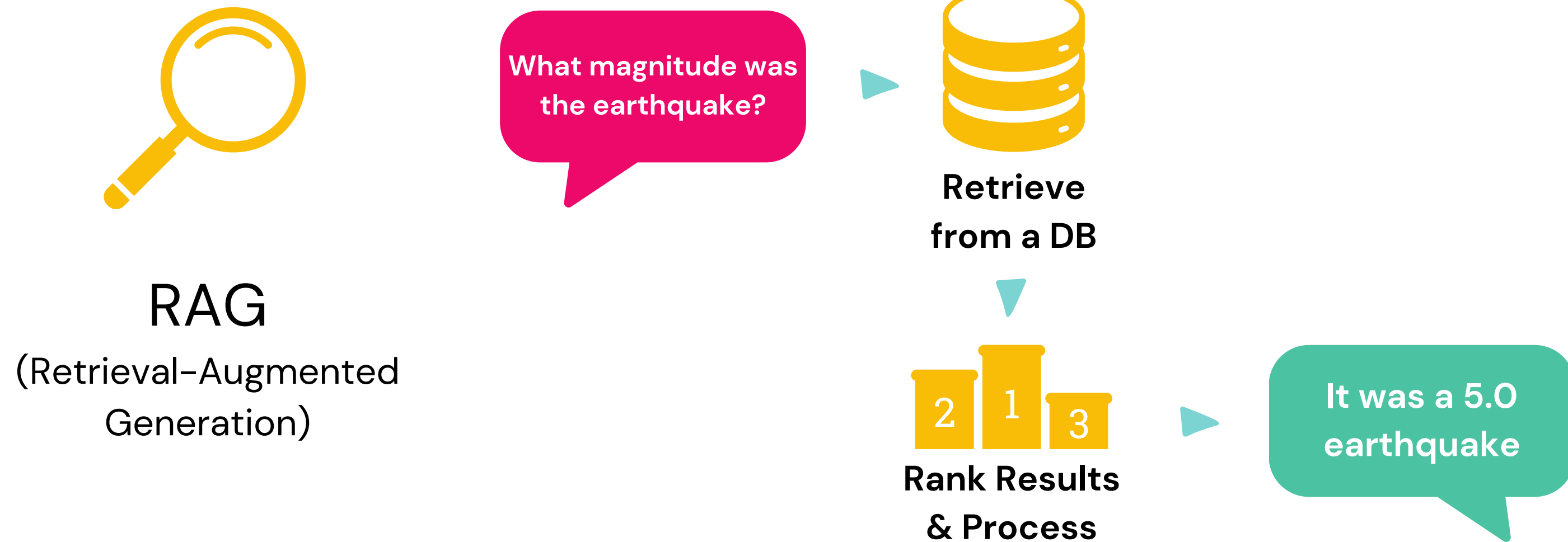


# Not quite: they have not been built for purpose

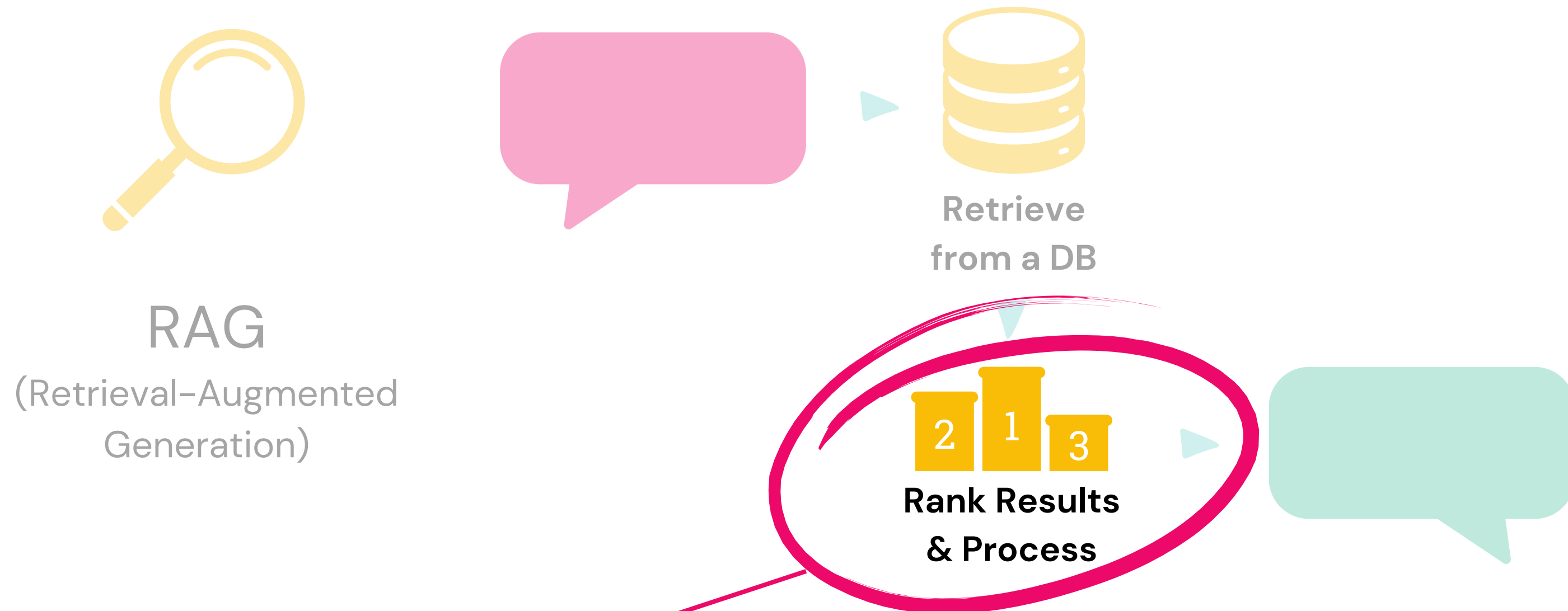
- General Purpose systems
- High hallucination rates
- Highly non-deterministic
- Too heavy to scale effectively
- RAG relies exclusively on top ranked sources



# RAG is the closest match



# But it's not quite right

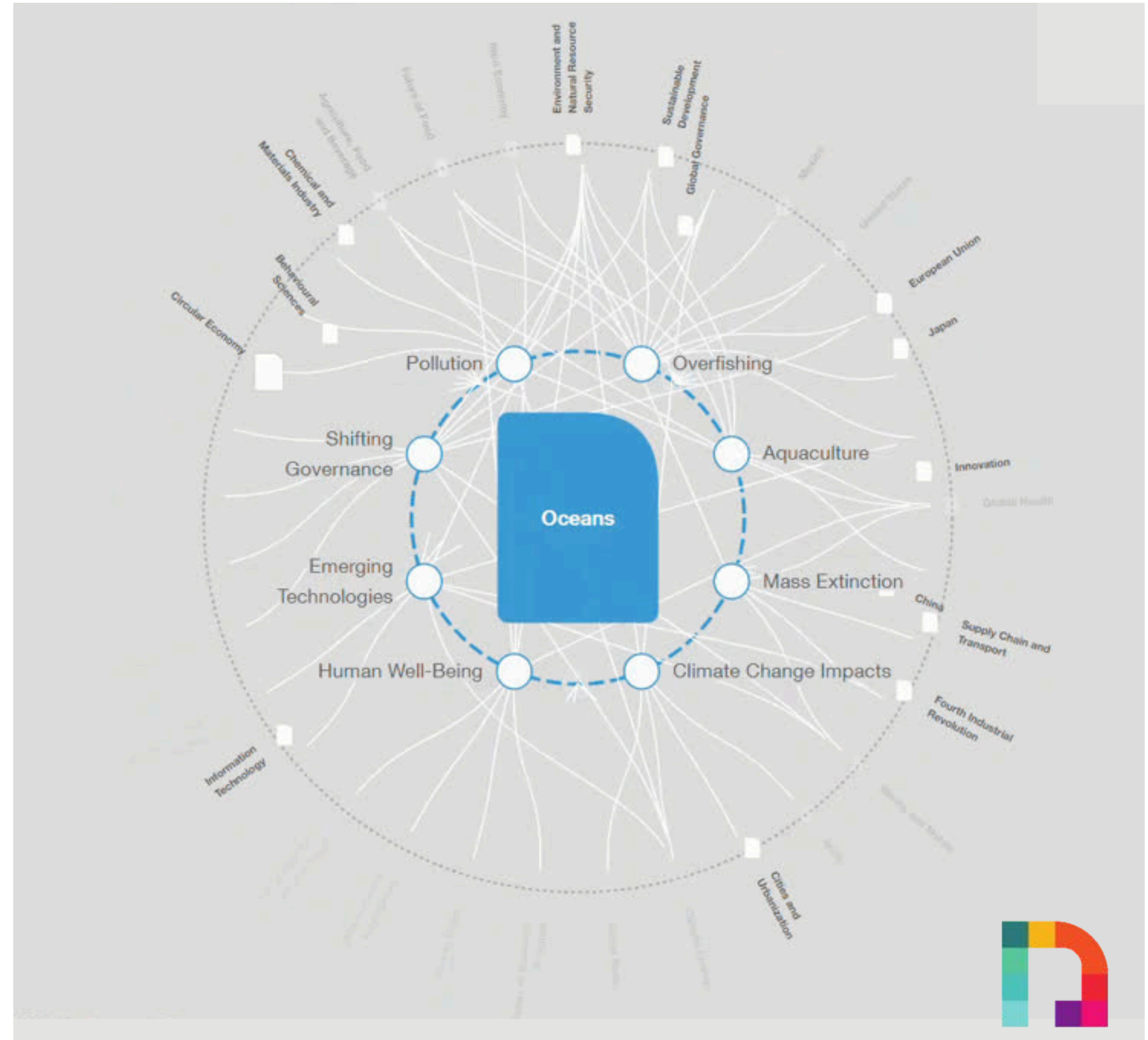


Only top ranked sources are used in RAG responses  
In other words: answers won't be complete



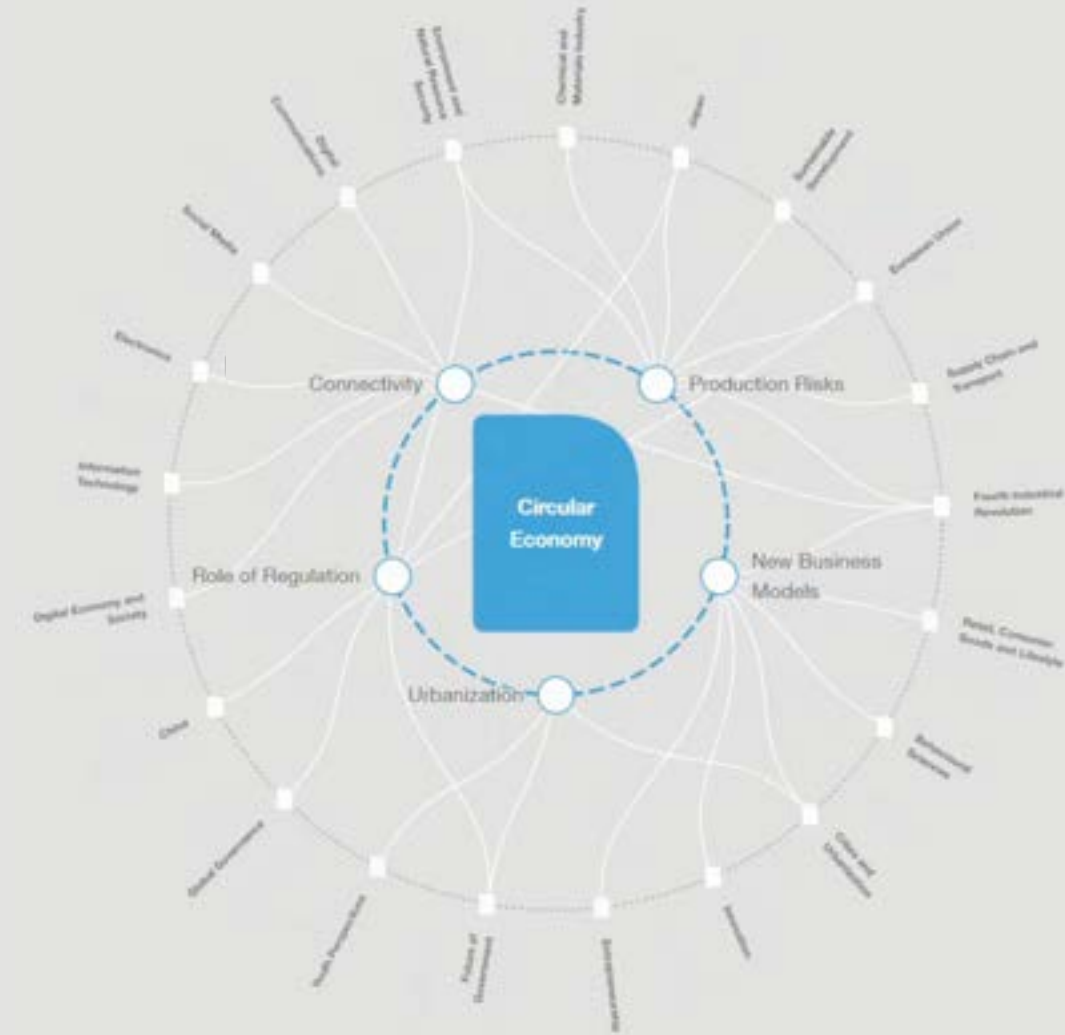
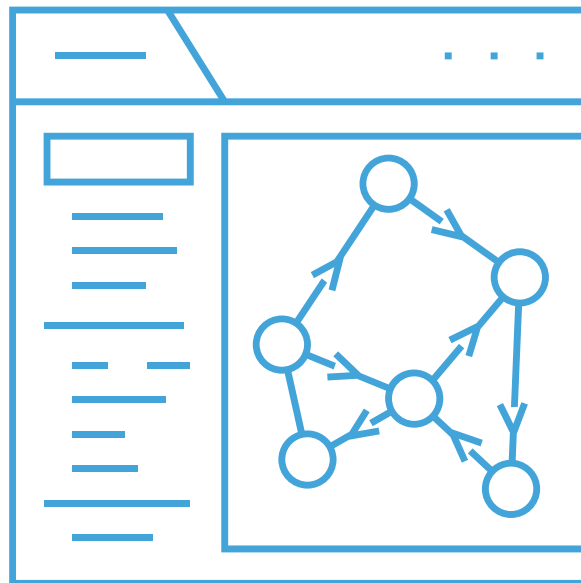
# Introducing: Dynamic Knowledge Graphs

\*Illustration: World Economic Forum



# The puzzle pieces: SLM and KG

- Lightweight Small Language Model
- Transforms a research paper into a structured set of entities and links

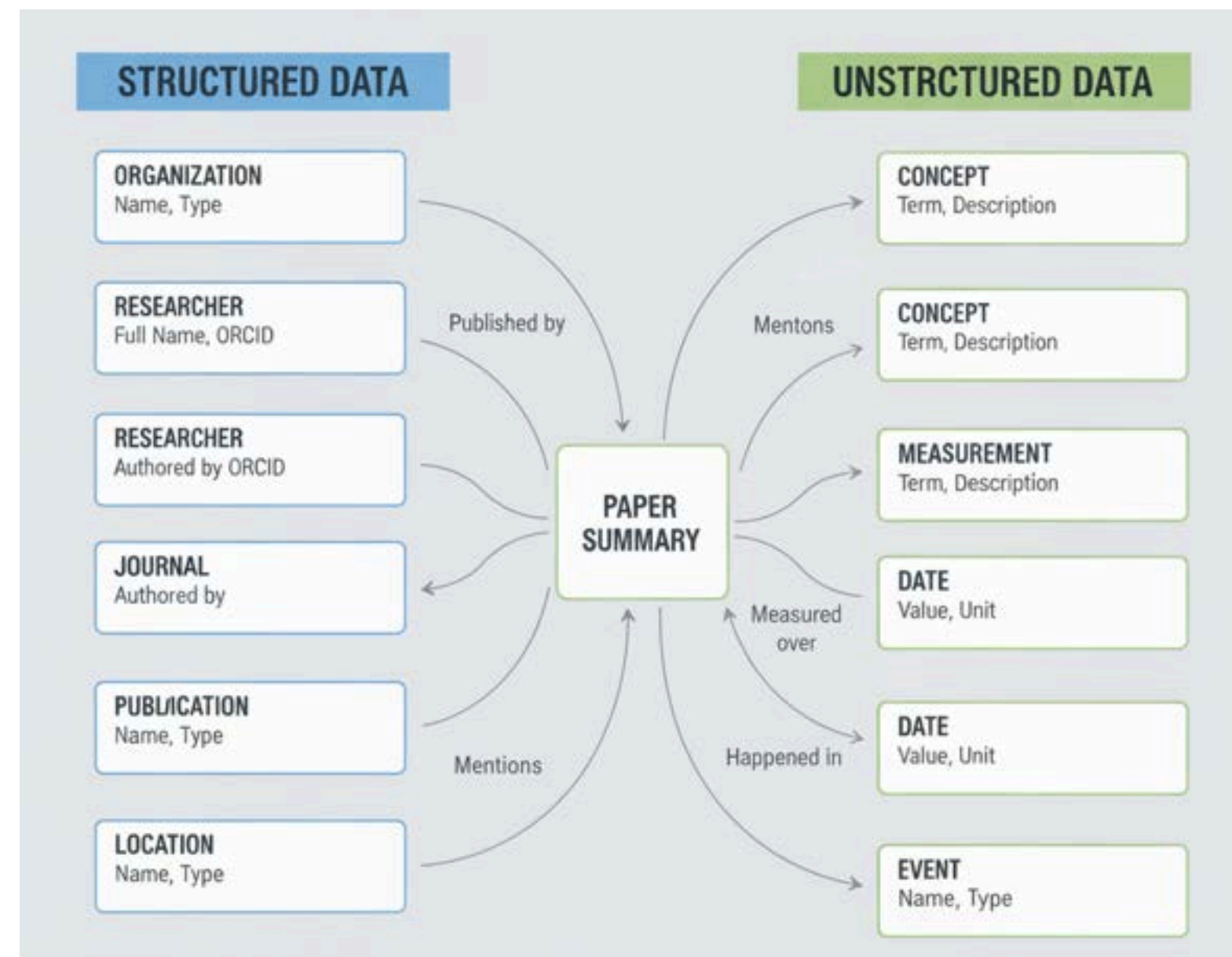


- Continuously updating knowledge base
- 10y of peer reviewed climate reporting
- Complex query capabilities

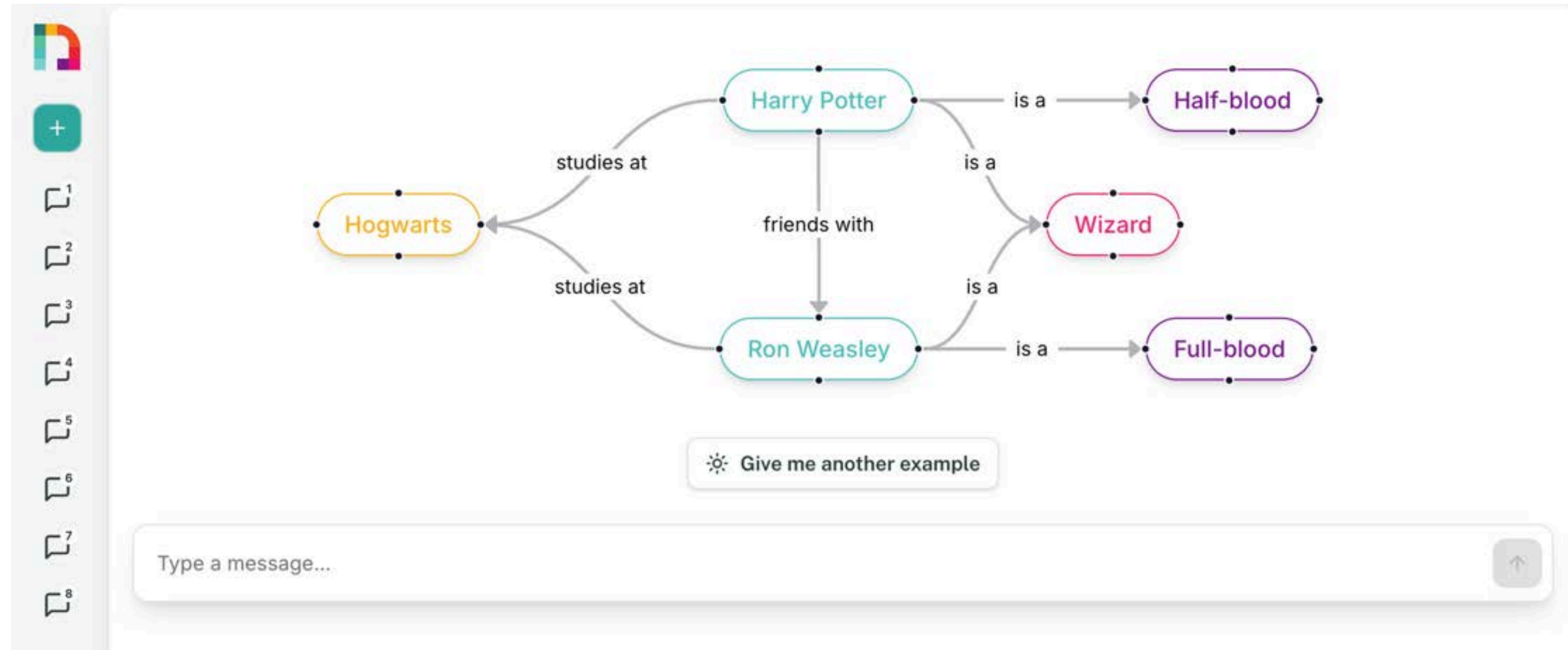


# Turning unstructured climate science into structured, deterministic, searchable information

What our graph includes (structured vs unstructured)



# Turning unstructured climate science into structured, searchable information



# Turning unstructured climate science into structured, searchable information

+

1

2

3

4

5

6

7

8

Who are the top researchers in climate attribution, per country?

Open Knowledge Graph

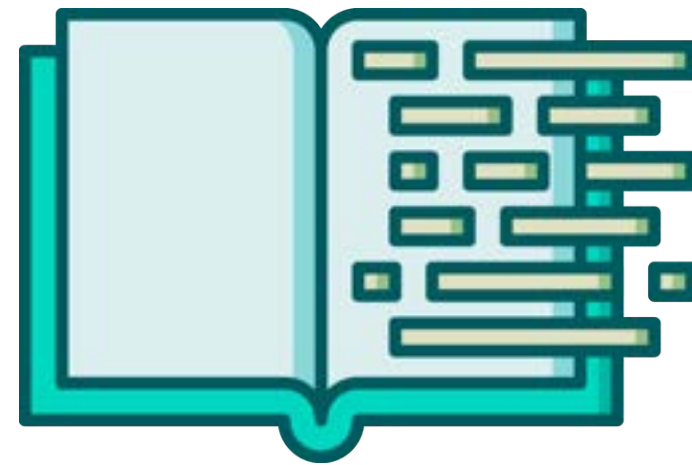
| Country        | Researcher    | Citations |
|----------------|---------------|-----------|
| United Kingdom | John Smith    | 345       |
| France         | Jeanne Dupont | 234       |
| United States  | Alex Johnson  | 123       |

Based on the provided data, John Smith, Jeanne Dupont, and Alex Johnson are the top three researchers in climate attribution across the UK, France, and the US. Their research spans...

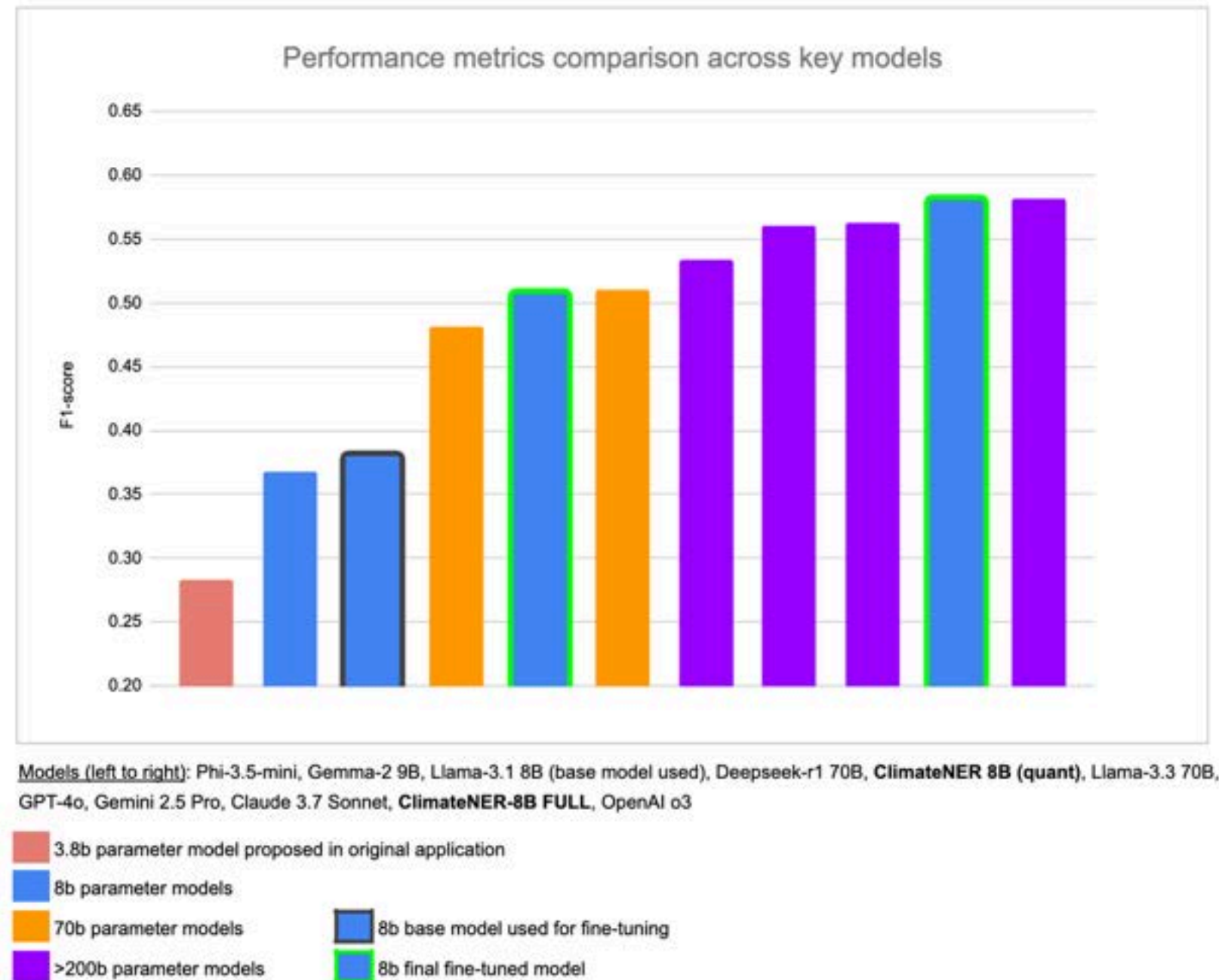


# Turning unstructured climate science into structured, searchable information

Our methodology to select papers



# David and Goliath



Trained a very small model of **8 billion** parameters to perform at the **same level of state-of-the-art models 40x larger** for our task – entity and relationship extraction.

Achieved through:

- **Fine-tuning** (used EuroHPC)
- **Gold standard** manually annotated datasets
- **Synthetic data** generation
- **Quantitisation** of models (used EuroHPC)

Allowing for **high performance** and **scalability** given volume of data to process, as well as **full ownership** over the models.



# First partnerships are in place, and more to come



Announcement

## Making climate science easier to navigate for journalists



Vera Penêda – Director of Programmes & Impact August 06, 2025

### EJC's new AI partnership with The Newsroom

We are starting a partnership with The Newsroom, an AI startup based in Portugal, to provide early access to a new platform designed specifically for climate journalists: Climate Blocks.

This collaboration answers a real challenge voiced by many journalists in our network – the struggle to keep up with the flood of new climate science studies and connect them meaningfully to their reporting.

<https://eic.net/donate>

› 150 organisations signed up to test ClimateBlocks

› 20 countries represented



# Thank you!

[jenny@thenewsroom.ai](mailto:jenny@thenewsroom.ai)



# Enhancing Financial NLP with Supercomputing: Spell Correction and Domain-Specific BERT Pretraining

Derya Uysal, Doğacan Toka, Elif Bozkurt, Osman Kumaş, Hasan Hüseyin Yılmaz

EuroHPC User Day - 2025



# Tablo of Content



Introduction & Motivation

HPC in Finance: Opportunities and Challenges

The Portfolio Assistant Project

Experiments on Meluxina

Key Results & Takeaways

Future Work & Recommendations

Q&A

# Introduction & Motivation

## Introduction

- Virtual Assistant for investment customers
- **Offers:**
  - Fund information & comparisons
  - Historical returns
  - Personalized risk analysis & portfolio recommendations
  - Instant answers to FAQs
- **Scalable & adaptable:**
  - Banking, factoring, leasing, portfolio firms
  - Works on customer-facing & internal platforms
  - Runs even on limited GPU resources

## Methodologies

### Domain-Specific LLM

- Trained on Turkish financial data using supercomputers

### Conversational Design

- Dialogue flow modeling & context-aware response generation

### Personalized Risk Profiling

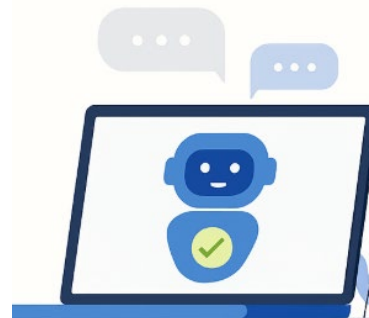
- Dynamic assessment & tailored portfolio allocation

### Parallel & Distributed Computing

- Multi-GPU training & inference, scalable deployment

### Transfer & Reinforcement Learning

- Efficient adaptation to new organizations & user feedback



# HPC in Finance: Opportunities and Challenges

## Opportunities

- Accelerated training of large language models (LLMs) for financial NLP tasks
- Personalized portfolio recommendations at scale
- Advanced risk analysis and stress testing simulations
- Real-time or near-real-time customer support through AI
- Scalability: adapting to multiple sectors (banking, leasing, factoring)

## Challenges

- Accelerated training of large language models (LLMs) for financial NLP tasks
- Personalized portfolio recommendations at scale
- Advanced risk analysis and stress testing simulations
- Real-time or near-real-time customer support through AI
- Scalability: adapting to multiple sectors (banking, leasing, factoring)
- Challenges
- Data transfer bottlenecks between local and HPC systems
- Data privacy & regulatory compliance
- Optimizing resource allocation for parallel tasks
- High infrastructure costs for on-premise HPC setups
- Lack of financial-domain-specific training data
- Real-time latency requirements vs. batch-oriented HPC architectures



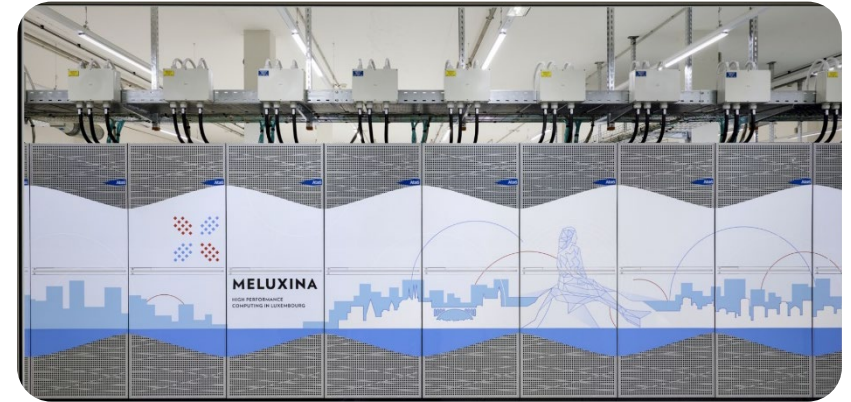
# The Portfolio Assistant Project

- **Goal:** Develop an intelligent virtual assistant for investors to enhance decision-making and customer experience
- **Features:**
  - **General Information** – Summarized & detailed descriptions of listed funds
  - **Fund Comparison** – Evaluate funds by performance, risk, and returns
  - **Historical Performance Analysis** – Identify trends using past returns
  - **Personalized Risk Analysis** – Recommend portfolios based on user's risk profile through interactive questions
- **Technology Highlights:**
  - **Spell Correction** – Process noisy text with AI for accuracy & clarity
  - **Financial Large Language Model (LLM)** – NLP-powered, user-friendly analysis
  - **Financial Forecasting (ML)** – Data-driven predictions with machine learning
- **Why HPC was needed:**
  - Training domain-specific models
  - Optimizing inference for near-real-time performance

# Experiments on Meluxina

## •About Meluxina:

- Meluxina is a **EuroHPC supercomputer**, located in Luxembourg and operated by LuxProvide.
- Designed for high-performance computing, artificial intelligence, and big data analytics.
- Energy-efficient, ranked among the **Green500** list of most energy-efficient supercomputers.
- NVIDIA A100 Tensor Core GPUs
- Dedicated GPU nodes with high memory bandwidth
- Optimized for AI/ML workloads
- Peak Performance:** ~18 PetaFLOPS
- Processor Architecture:** AMD EPYC Rome CPUs
- Nodes & Memory:**
  - 800+ compute nodes
  - Up to 512 GB memory per node
- Storage:** 20+ PB parallel file system (Lustre)
- High-speed InfiniBand HDR interconnect
- Linux-based OS environment
- Pre-installed scientific libraries, CUDA, deep learning frameworks (PyTorch, TensorFlow, MPI)
- Powered by 100% renewable energy
- Among the world's most sustainable supercomputers
  - GPU: 200 GPU node, her node 2 AMD Rome CPU (32 core @ 2.35 GHz) ve 4 NVIDIA A100-40 GPU. RAM 512GB
  - CPU: 573 nodes. Her node 2 AMD Rome CPU, her biri 64 core @ 2.6 GHz, RAM 512GB
- Optimizations:**
  - Infrastructure-level & software-level improvements



# Optimum Resource Utilization and Required Tools

- Tools to Use
  - 🦴 Pytorch – For deep learning models
  - 🐸 Axolotl – For model tuning and training
  - 🐪 Llama-Factory – For LLM development and optimization
- Things to Consider
  - Models & Descriptions to be Used – It should be specified which models will be used and how they will be integrated within the scope of the project.
  - License Management – The relevant licenses of the codes and tools used should be shared.
  - Performance Measurement – It should be determined how the model will be evaluated in terms of accuracy, speed and resource usage.
  - Scalability – It should be well planned how the project will adapt to the growing data and processing load.
  - Optimization Methods – Optimizations should be made to improve model training, memory management and processing time.

# FVAP (Financial Virtual Assistant for Portfolio Management)

- Building Domain-Specific NLP Systems for Financial Industry using HPC
  - Two-stage deep learning framework for Turkish financial NLP
  - Supported by EuroHPC MeluXina supercomputer
  - Addressing low-resource language & high-precision domain needs
- **Why?**
  - Financial chatbots suffer from noisy, typo-prone user input.
  - Generic LLMs lack domain-specific financial understanding.
  - Need for fast, scalable training on sensitive data.
- **Our Approach:**
  - Stage 1: Improve input clarity → spelling & grammar correction.
  - Stage 2: Domain-adapted language understanding → Financial BERT.
  - HPC to accelerate and scale training.

# FVAP (Financial Virtual Assistant for Portfolio Management)

- **Two-Stage Framework:**

Correction Module:

- Fine-tune mT5-XL on multi-phase Turkish error datasets.
- Progressive training: word → sentence → paragraph.

Financial BERT:

- Pretrain ModernBERT on Turkish financial documents.
- Domain-specific embedding for downstream tasks.

## **HPC:**

- EuroHPC MeluXina supercomputer.
- A100 GPUs for scalable multi-GPU training.

# FVAP (Financial Virtual Assistant for Portfolio Management)

## Datasets

- **Spelling Correction Datasets:**

- Turkish Spelling Dictionary
- Turkish GPT GEC Dataset
- Real-world complaints (scraped & corrected via GPT-4o-mini)
- Word, sentence, and paragraph level aligned datasets.

- **Financial Domain Corpora:**

82 curated Turkish financial documents:

- Regulatory filings
- Market analysis
- Terminology dictionaries

# FVAP (Financial Virtual Assistant for Portfolio Management)

## Correction: Model & Strategy

- Base Model: google/mt5-xl
- LoRA adapters on Q/V projections.
- 4-bit NF4 quantization.
- Gradient checkpointing & bfloat16 mixed precision.
- Curriculum learning: 3 phases.

| Phase | Granularity | Best WER | Best CER |
|-------|-------------|----------|----------|
| 1     | Word        | 24.61%   | 6.08%    |
| 2     | Sentence    | 11.94%   | 3.10%    |
| 3     | Paragraph   | 6.19%    | 1.57%    |

## Financial BERT

Base Model: ModernBERT

Improvements:

- RoPE, GeGLU activations
- Alternating Local/Global Attention
- Flash Attention 2/3 for long contexts
- Unpadding for efficiency

Training & Results:

- Pretrained on MLM with 30% masking.

Accuracy on masked tokens:

- Baseline Turkish BERT: **61.32%**
- ModernBERT (ours): **83.44%**

Ready for downstream tasks: NER, QA, classification.

# Experiments on MeluXina

- **Why HPC?**
  - Speedup: 38× faster training compared to local servers.
  - Supports larger batch sizes & longer sequences.
  - Secure, scalable, and privacy-preserving deployment.
- Correction training times:

| Phase     | Local | MeluXina |
|-----------|-------|----------|
| Word      | 1h19m | 1h05m    |
| Sentence  | 2h32m | 14m      |
| Paragraph | 8h54m | 33m      |

- Notes:
  - MeluXina's parallelization improved runtime.
  - Some performance degradation likely due to distributed training artifacts → future work.

# Conclusions & Future Work

## What We Achieved:

- Effective two-stage NLP pipeline for Turkish financial domain.
- Significant quality gains in both spelling correction & domain adaptation.
- HPC reduced training times dramatically, enabling scalability.

## Future Directions:

- Explore robust distributed training techniques.
- Integrate quantitative reasoning into NLP pipeline.
- Study privacy-preserving deployment: federated learning, differential privacy.
- Continue training large language models on proprietary financial data under future funding calls.

# Key Results & Takeaways

## Technical Impact

- Achieved lower error rates on noisy, user-generated Turkish financial texts:
  - Word Error Rate (WER) ↓ by 75% over baseline.
  - Character Error Rate (CER) ↓ by 74% over baseline.
- Inference latency reduced, enabling near-real-time responses in chatbot scenarios.
- Improved accuracy & relevance in intent detection, entity recognition, and question answering tasks.

## •Business Impact

- Improved customer experience with more coherent and reliable chatbot interactions.
- Built-in domain expertise, making the assistant a trusted advisor for investors.
- Ready for private, secure deployment within institutional environments, ensuring compliance with data privacy needs.
- Lays the foundation for expanding to other financial services (factoring, leasing, etc.).

# Future Work & Recommendations

- Extend domain-specific models
- Explore GPU/CPU hybrid inference strategies
- Advice to practitioners:
  - Plan for data preparation & model profiling
  - Leverage HPC early in the design phase



EuroHPC UserDay 2025

# Tailoring AI for Turkish Law: Domain-Specific Fine-Tuning of Small Language Models for Legal Expertise

Authors: Özgür Uğur, Mahmut Çimen, Rumeysa Güllüce, Musa Yılmaz, Özay Ezerçeli  
Affiliation: New Mind AI, YTU Teknopark, Sarıyer, Türkiye  
Contact: [aiteam@newmind.ai](mailto:aiteam@newmind.ai)

Funding: EuroHPC JU Benchmark Access Grant No. EHPC-BEN-2024B11-003  
Infrastructure: IT4Innovations National Supercomputing Center (Karolina)



**EuroHPC**  
Joint Undertaking



**NewMind AI**  
where data finds its **mind**

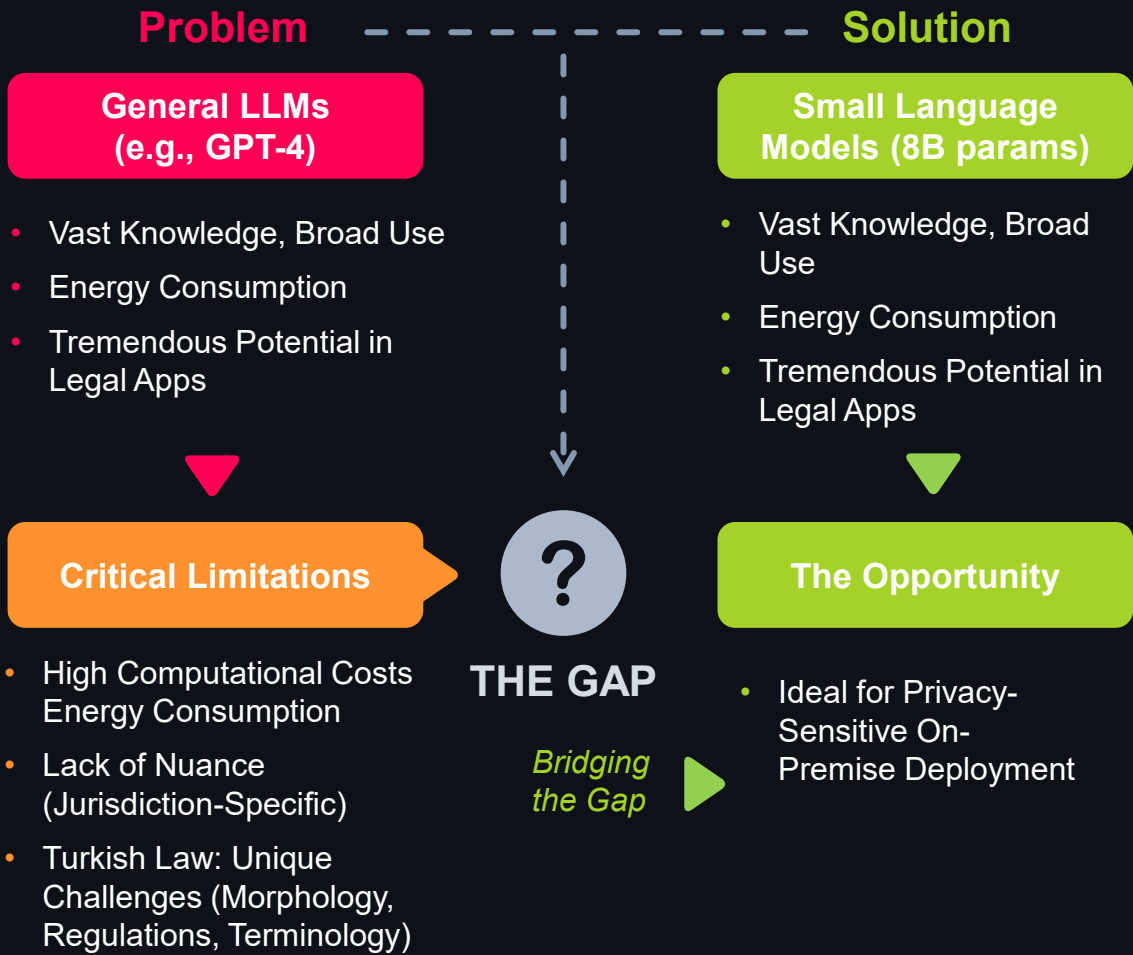




## Table of Content

- 1 The Challenge & Opportunity
- 2 Strategic Approach & Model Selection
- 3 Innovative Synthetic Data Generation
- 4 HPC Infrastructure & Fine-Tuning Methodology
- 5 Comprehensive Evaluation Framework
- 6 Impact & Achievements
- 6 Critical Lessons Learned
- 7 Conclusion & Future Directions

# Bridging the Gap: LLMs for Specialized Law



## The Challenge & Opportunity

### The Problem:

- Large Language Models show tremendous potential in legal applications
- Critical limitations for specialized domains

### The Opportunity:

- Small Language Models (8B parameters) when expertly fine-tuned can achieve comparable performance to much larger models
- Significantly more computationally efficient and economically viable
- Ideal for privacy-sensitive legal applications requiring on-premise deployment

 **Our Hypothesis: PEFT -> Specialized Legal AI Experts**

### Foundation Model: Meta-Llama 3.1 8B Instruct



#### WHY THIS MODEL

- ⚡ **Computational Efficiency:** Optimal balance
- ⚡ **Instruction-Tuned Nature:**
- 🔒 **Instruction-Tuned Nature:** Ready for adaptation
- 🔒 **Open-Weight Access:**
- 🌐 **Open-Weight Access:** Enables research
- 🔒 **Enterprise Deployment:** Privacy-sensitive
- 🔒 **Agentic Workflow Integration:** Modular systems

#### TARGET DOMAINS



## Strategic Approach & Model Selection

### Why This Model:

- **Computational Efficiency:** Optimal balance between capability and tractability
- **Instruction-Tuned Nature:** Ready for domain-specific adaptation
- **Open-Weight Access:** Enables research and commercial development
- **Enterprise Deployment:** Perfect for privacy-sensitive legal applications
- **Agentic Workflow Integration:** Ideal for modular systems with multiple specialized AI agents
- **Target Domains:** Energy & Natural Resources, Land & Zoning, Competition, Tax, Healthcare, Labor, Intellectual Property, \*Data Protection, Capital Markets, Environmental Law

*\*The Law on the Protection of Personal Data of Turkey Law No. 6698*

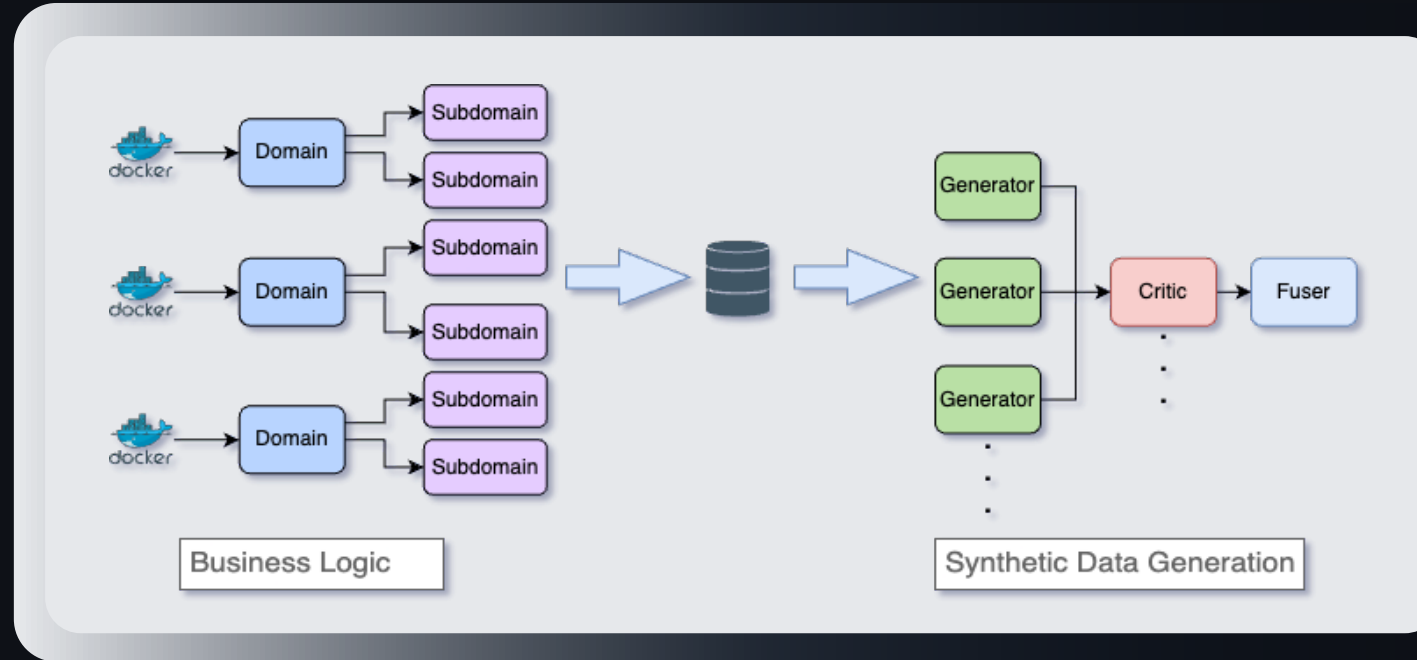


Figure 1: <https://newmind.ai/en/blog/revolutionizing-ai-training-with-synthetic-data-pipelines?id=1>

### Stage 1 - Generator

- **Models:** Llama 3.1 70B, Qwen2.5-72B operating in parallel
- **Function:** Generate diverse question candidates from curated legal documents
- **Scale:** Up to 100 concurrent workflows

### Stage 2 - \*Critic + RAG

- **Model:** Qwen2.5-72B with Retrieval-Augmented Generation
- Supports up to 65,536 tokens for extended context requirements with RoPE Scaling
- **Function:** Rigorous quality assessment using legal source verification
- **Criteria:** Relevance, clarity, legal grounding, factual accuracy

### Stage 3 - Fuser

- **Model:** Qwen2.5-72B (deterministic, temperature 0.0)
- **Function:** Synthesize positive aspects while eliminating problematic elements
- **Output:** Highly refined, optimized questions with corresponding answers



## HPC Infrastructure & Fine-Tuning Methodology

### Parameter-Efficient Fine-Tuning with LoRA:

- Configuration: Rank 16, Alpha 32, Dropout 0.05
- Training: 3 epochs per domain, cosine learning rate scheduling ( $3 \times 10^{-5}$ )

### Technical Infrastructure:

- EuroHPC Karolina: 320 GB HBM2 (8 x 40 GB) GPU NVIDIA A100 per node
- Orchestration: Axolotl framework for flexible configuration
- Acceleration: \*Unsloth library providing optimized kernels

### Remarkable Efficiency Gains:

- 8.8x training speedup compared to standard implementations
- 85-90% GPU and memory utilization during training
- Total compute: ~600 GPU hours for complete project
- Environmental impact: Only 56.84-64.8 kg CO<sub>2</sub> footprint

<https://unsloth.ai/>

| Model Name                              | Lexical Avg. | Semantic Avg. | OpenAI Judge Avg. |
|-----------------------------------------|--------------|---------------|-------------------|
| Base Model (Meta-Llama 3.1 8B Instruct) | 0.379        | 0.842         | 0.594             |
| Alpaca-capital                          | 0.411        | 0.928         | 0.368             |
| Chat-template-capital                   | 0.443        | 0.942         | 0.545             |
| Alpaca-competition                      | 0.430        | 0.937         | 0.495             |
| Chat-template-competition               | 0.344        | 0.920         | 0.367             |
| Alpaca-energy                           | 0.425        | 0.940         | 0.502             |
| Chat-template-energy                    | 0.335        | 0.918         | 0.368             |
| Alpaca-environmental                    | 0.412        | 0.917         | 0.428             |
| Chat-template-environmental             | 0.426        | 0.895         | 0.490             |
| Alpaca-fund                             | 0.441        | 0.920         | 0.354             |
| Chat-template-fund                      | 0.462        | 0.935         | 0.487             |
| Alpaca-health                           | 0.395        | 0.924         | 0.425             |
| Chat-template-health                    | 0.337        | 0.921         | 0.399             |
| Alpaca-intellectual                     | 0.398        | 0.894         | 0.313             |
| Chat-template-intellectual              | 0.464        | 0.943         | 0.593             |
| Alpaca-kvkk                             | 0.414        | 0.938         | 0.477             |
| Chat-template-kvkk                      | 0.334        | 0.919         | 0.389             |
| Alpaca-land                             | 0.391        | 0.902         | 0.325             |
| Chat-template-land                      | 0.436        | 0.934         | 0.532             |
| Alpaca-tax                              | 0.400        | 0.928         | 0.456             |
| Chat-template-tax                       | 0.332        | 0.919         | 0.373             |

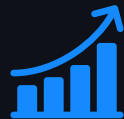
## Comprehensive Evaluation Framework

### Three Evaluation Categories:

- **Lexical Metrics:** BLEU, ROUGE-L, F1-Score for surface-level similarity
- **Semantic Metrics:** Sentence embeddings with multilingual and Turkish-specific models for meaning preservation
- **AI Judge Metrics:** GPT-4o-mini evaluation on Relevancy, Coherence, and Accuracy for practical legal utility

# Key Performance Results

## Semantic Improvements



Base **0.842**



Fine-tuned up to **0.943**

## Lexical Improvements



Base **0.379**



Fine-tuned Up To **0.464**

## Strong Specialization



Capital Markets,  
Intellectual Property

## Format Dependency



Optimal Instruction Format  
Varies By Domain



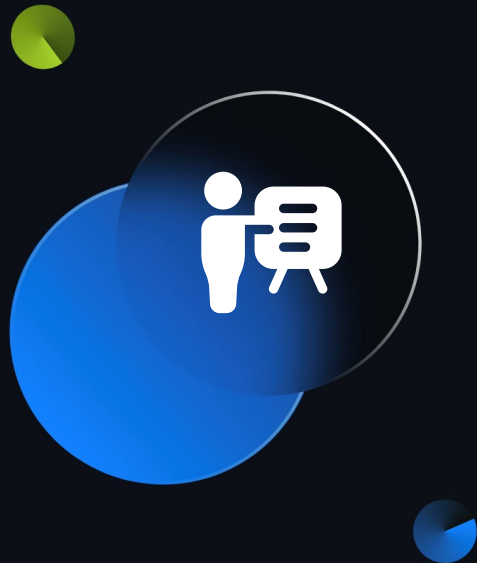
## Impact & Achievements

### Major Technical Breakthroughs:

- Successfully created 10 specialized legal expert models from single 8B base
- Validated PEFT approach for efficient domain adaptation without catastrophic forgetting
- Established efficient HPC methodology with significant acceleration

### Practical Impact:

- Tangible assets for Turkish legal tech ecosystem
- Compelling alternative to resource-intensive large general-purpose models
- Validated approach for building modular, cost-effective agentic workflows
- Democratized access to specialized legal AI for resource-constrained organizations



## Critical Lessons Learned

- **SLM viability confirmed** for high-performing domain-specific applications
- **Data quality paramount** - multi-stage validation essential for legal accuracy
- **Context dependency** - instruction format optimization should be domain-specific
- **Human expert evaluation** remains essential for complete legal AI assessment

## THE FUTURE OF LEGAL AI

Collaborative Intelligence for Unprecedented Performance



### HUMAN + AI SYNERGY

Accelerating Legal Accuracy & Efficiency

## Conclusion

- This research demonstrates that high-quality, specialized legal AI can be developed efficiently using modern HPC infrastructure, making advanced AI capabilities accessible to legal professionals worldwide while maintaining computational efficiency and environmental responsibility.
- **Strategic Impact:** Validates targeted domain adaptation on HPC infrastructure as a pathway to democratize specialized AI solutions across knowledge-intensive fields.

## Future Directions: Timeline of Innovation and targeted benchmarking



Advancing Legal AI in Turkey

## Future Directions

- **Advanced RAG** integration for enhanced factual grounding in Turkish legal contexts
- **Larger Turkish** reasoning models for central coordination in agentic workflows
- **Systematic human expert** evaluation and targeted benchmarking
- **Real-world pilot studies** integrating models into legal professional workflows



REPUBLIC OF TÜRKİYE  
MINISTRY OF INDUSTRY  
AND TECHNOLOGY



**Barcelona  
Supercomputing  
Center**  
*Centro Nacional de Supercomputación*



**EuroHPC**  
Joint Undertaking

## Development and Continual Pre-training of Large-Scale Turkish Legal Language Models (TurkLeLM)

### Infrastructure: MareNostrum 5 ACC Partition

- *4500 Node Hours*
- *12-Months Timeline*
- **Project Overview:** This project addresses the critical shortage of specialized Turkish legal language models by developing state-of-the-art embedding models and continually pre-trained LLMs using EuroHPC BSC's computational infrastructure. We create essential AI resources for Turkish legal applications, contributing high-quality models to the open-source community while advancing Retrieval-Augmented Generation (RAG) capabilities for Turkish jurisprudence and other domains.



## Turkish-Specific Large Language Model for Information Processing and Contextual Analysis

- Project ID: **EHPC-AIF-2025LS04-019**
- Lead: **New Mind AI**
- Infrastructure: **MareNostrum 5 ACC Allocation: 375.000 Node hours**
- Duration: **12 Months**
- The EuroHPC AI Factory project establishes a transformative framework for democratizing supercomputing by converting the MareNostrum 5 into an accessible "AI-as-a-Service" platform. This was successfully demonstrated through the creation of Turkey's most advanced domain-specific Large Language Model, which involved training 7B-14B parameter models on over 200+ billion tokens using innovative, GPU-accelerated pipelines. With key technical achievements like a 26% improvement in OCR efficiency and a server capable of hosting thousands of models on a single GPU, the project delivers a replicable blueprint that strengthens European digital sovereignty, reduces dependency on foreign technology, and accelerates multilingual AI development across the continent.



**Thank you**  
**for listening!**



# Question & Answer



# Continued pretraining of LLMs on Leonardo

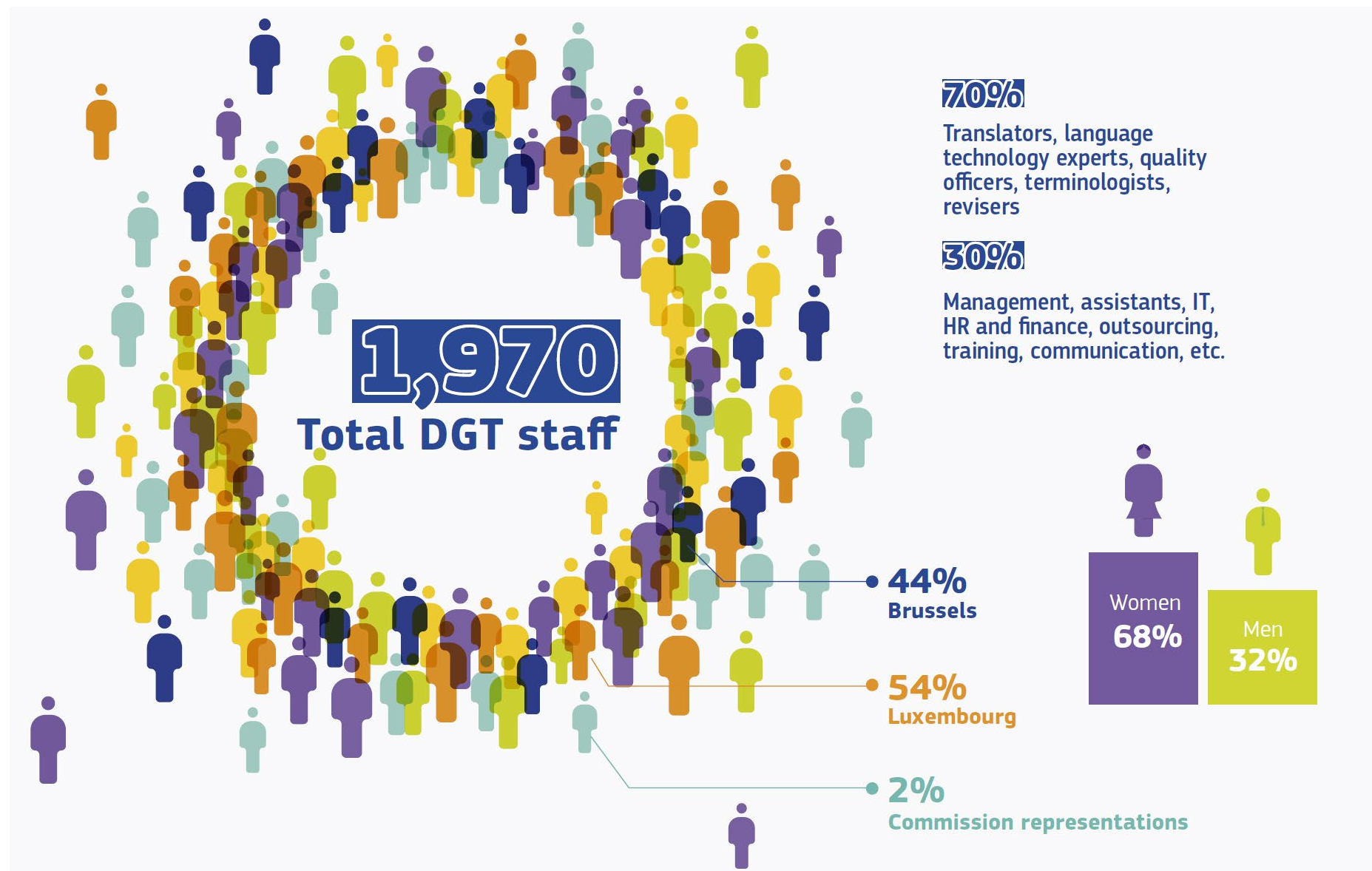
Ilja Rausch, PhD  
DG Translation – Innovation

EuroHPC User Days 2025

# Background

# DGT

Directorate-General for  
Translation



# AI-based services

In-house innovation. Driven by people, not profit.



## eTranslation

Neural machine translation built on the EU's history of professional translation.



## eTranslation (new page)

Test the new user-friendly interface of eTranslation, beta version with core functionalities only.



## eBriefing

Generate reports from sets of documents in official or general styles.



## eReply

Have AI help you prepare replies to correspondence, queries, and other requests.



## WebText

Use AI to refine English, French, or German text for your website.



## Accessible Text

Rewrite your English text to make it more accessible.



## eSummary

Quickly find out the main content of long documents.



## Multilingual Post

Short translations in multiple languages in one shot for X.



## Speech-to-Text

Upload your media and get full transcriptions or subtitles back.



## Anonymisation

Replace or redact names, places and more from your documents.

# EU institutional LLM

Enhance, don't reinvent.

## Our vision:

- Inspire EU public sector to adopt and develop AI solutions
- Modernize public administration through AI application
- Reduce the risk of digital divide
- Contribute to sovereign AI

# EU institutional LLM

Enhance, don't reinvent.

## Our vision:

- Inspire EU public sector to adopt and develop AI solutions
- Modernize public administration through AI application
- Reduce the risk of digital divide
- Contribute to sovereign AI

## Our contribution:

- Higher focus on underrepresented languages (e.g., Maltese, Irish, Lithuanian)
- High-quality texts across multiple EU institutions
- Experience with AI for public sector and EuroHPC infrastructure
- Vicinity to professional linguists and native speakers in all 24 EU languages

# EU institutional LLM

Enhance, don't reinvent.

## Our vision:

- Inspire EU p
- Modernize p
- Reduce the
- Contribute to

## Our task:

Leverage our data (EURAMIS) and EuroHPC infrastructure

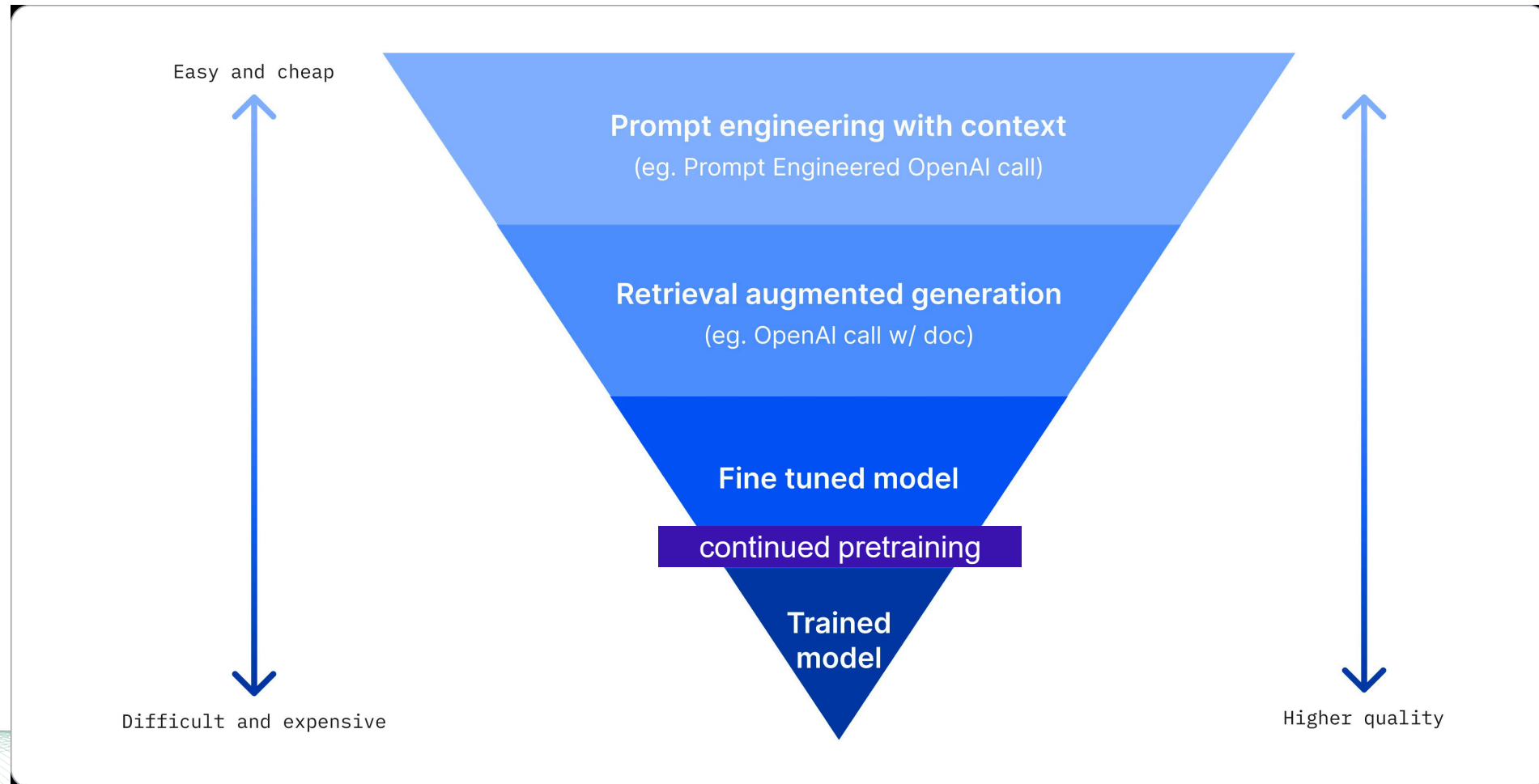
Enhance a European LLM to work well in all 24 EU languages

## Our contribution:

- Higher focus
- High-quality texts across multiple EU institutions
- Experience with AI for public sector and EuroHPC infrastructure
- Vicinity to professional linguists and native speakers in all 24 EU languages

# Methods

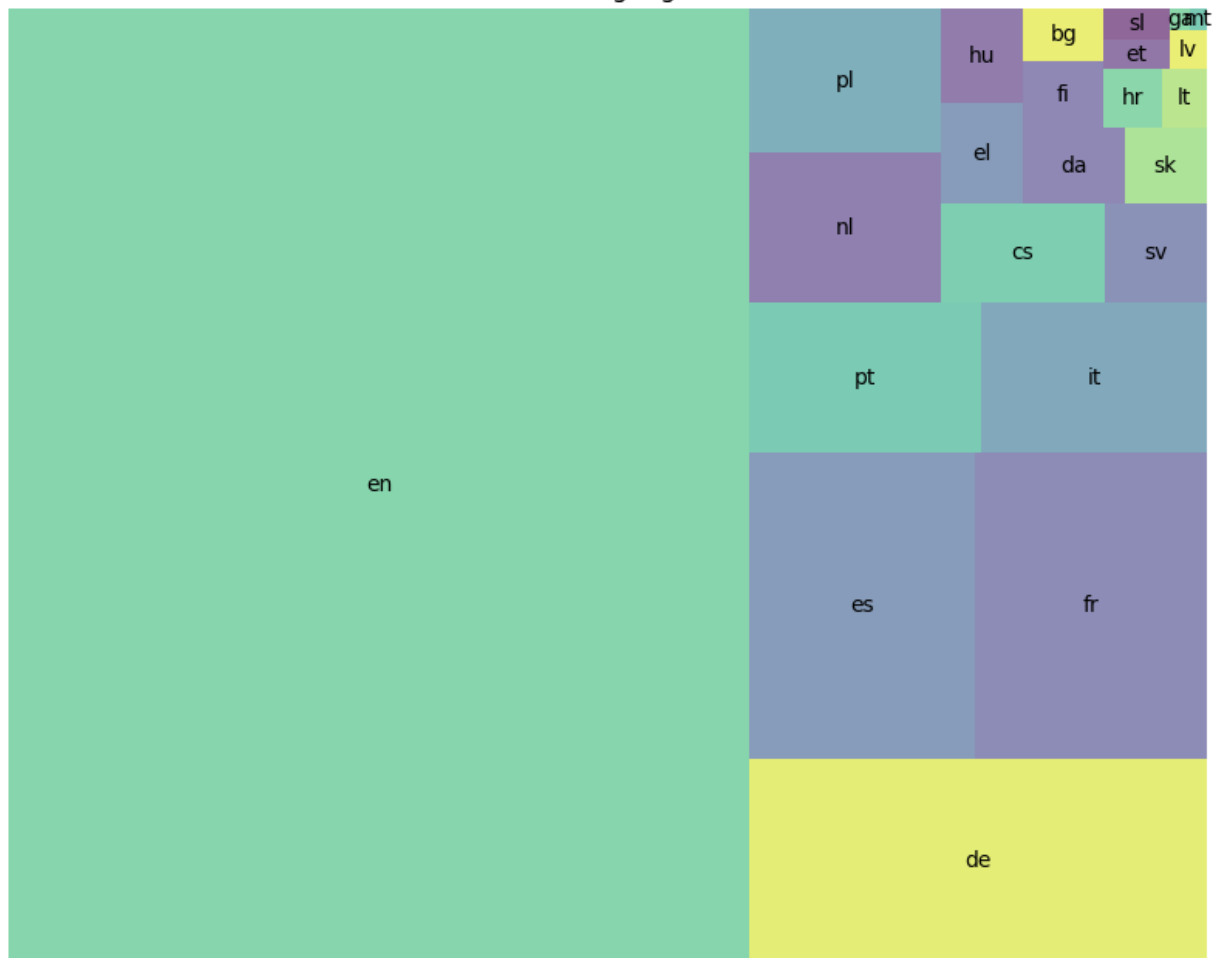
# Continued pretraining



Source: <https://www.fiddler.ai/blog/four-ways-that-enterprises-deploy-llms>

# Data

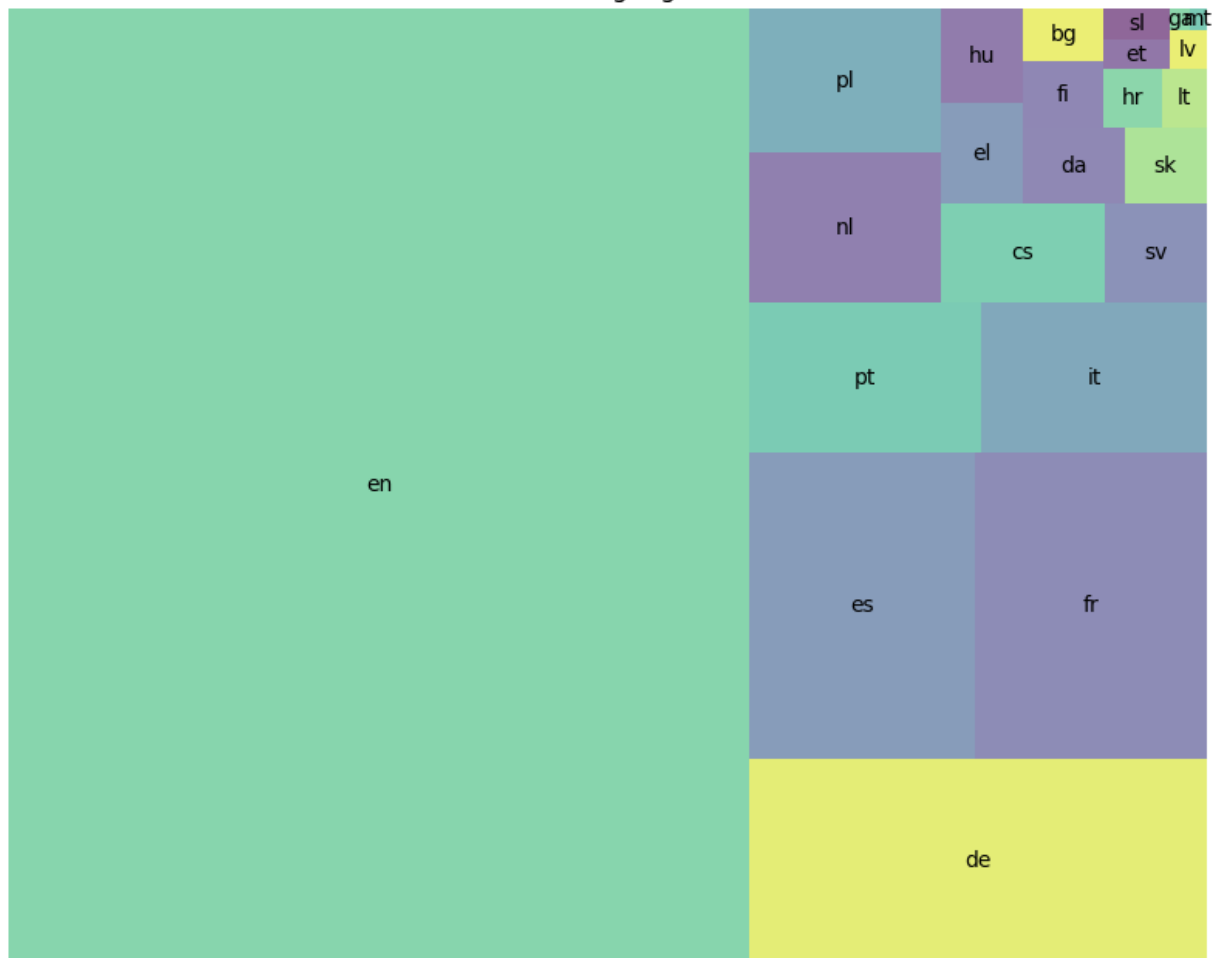
Common Crawl language distribution



Source: [commoncrawl.github.io/cc-crawl-statistics/plots/languages](https://commoncrawl.github.io/cc-crawl-statistics/plots/languages)

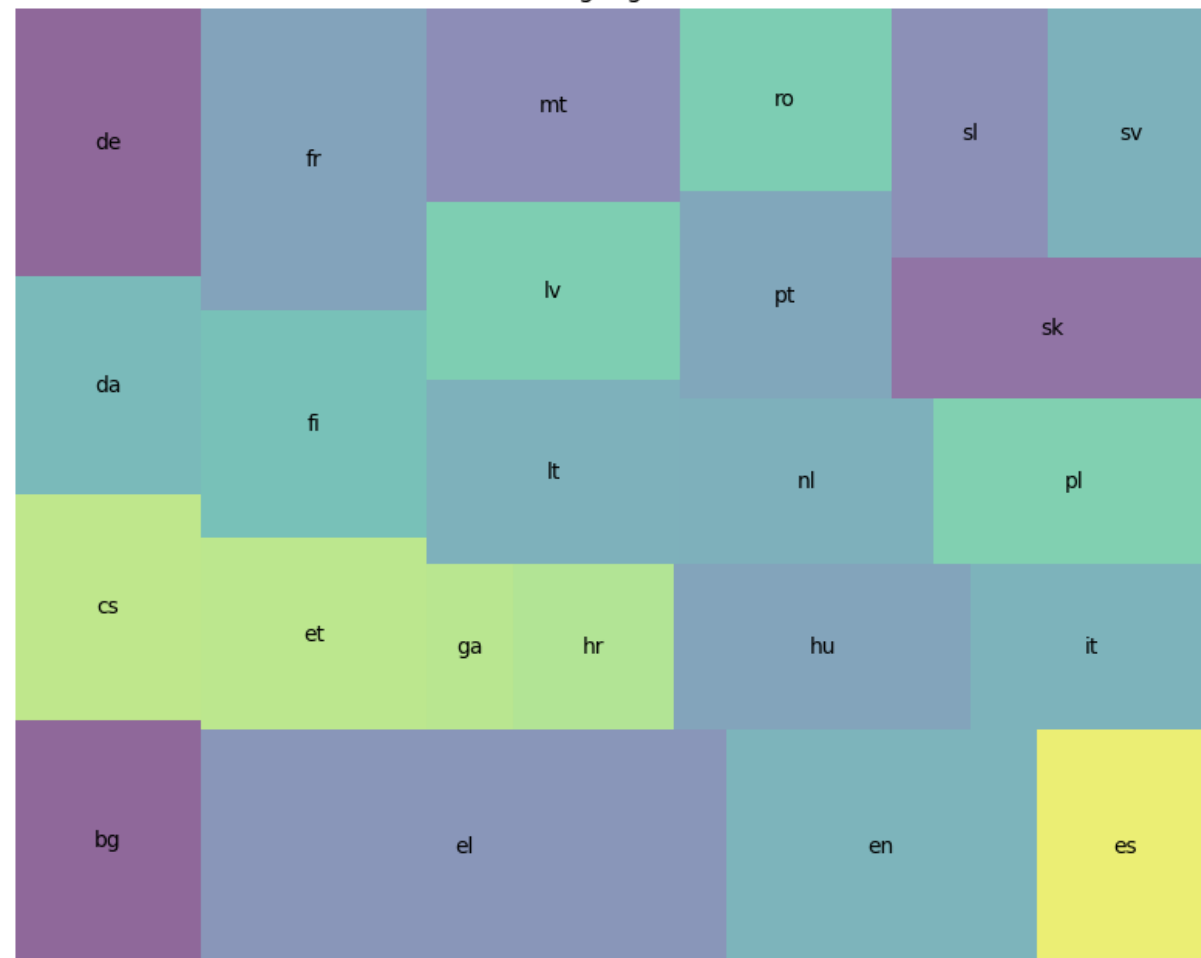
# Data

Common Crawl language distribution



## EURAMIS

EURAMIS language distribution



# Data

## EURAMIS

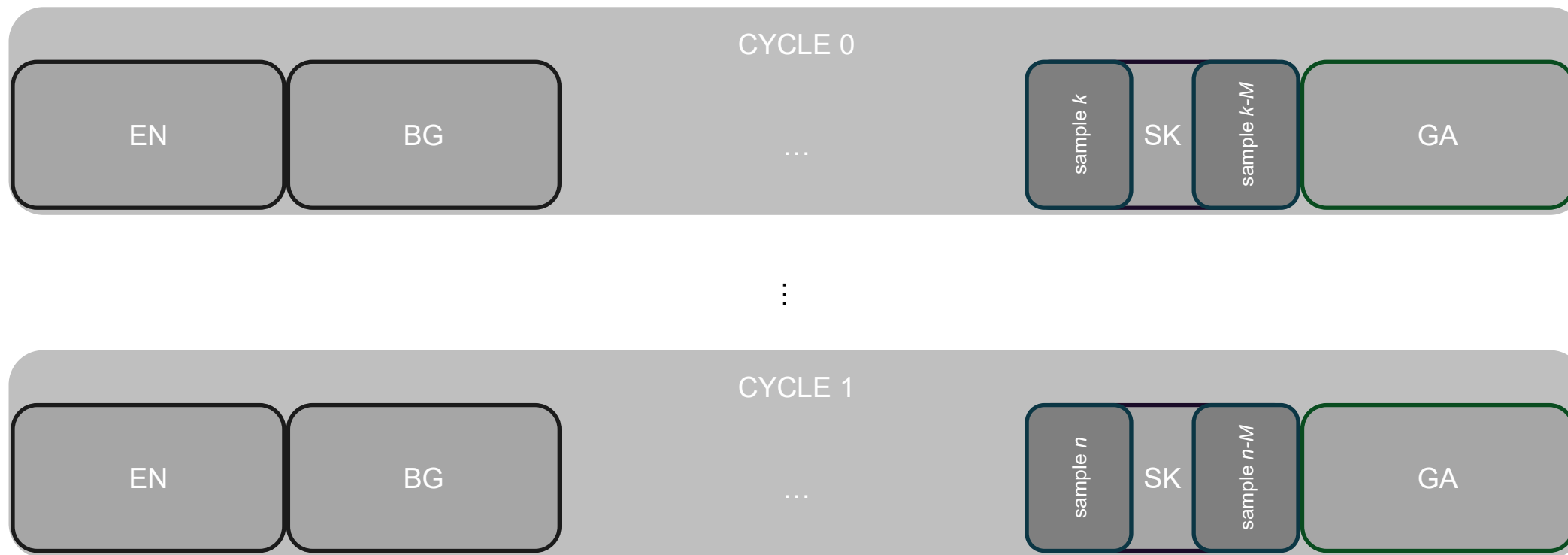
Common Crawl language distribution



EURAMIS language distribution

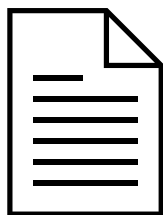


# Data packaging

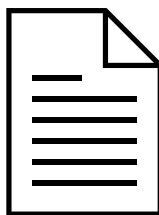


# Token ratio

Not the same as token fertility but still useful.



Doc XY in EN  
contains  $\tau_{en}$  tokens

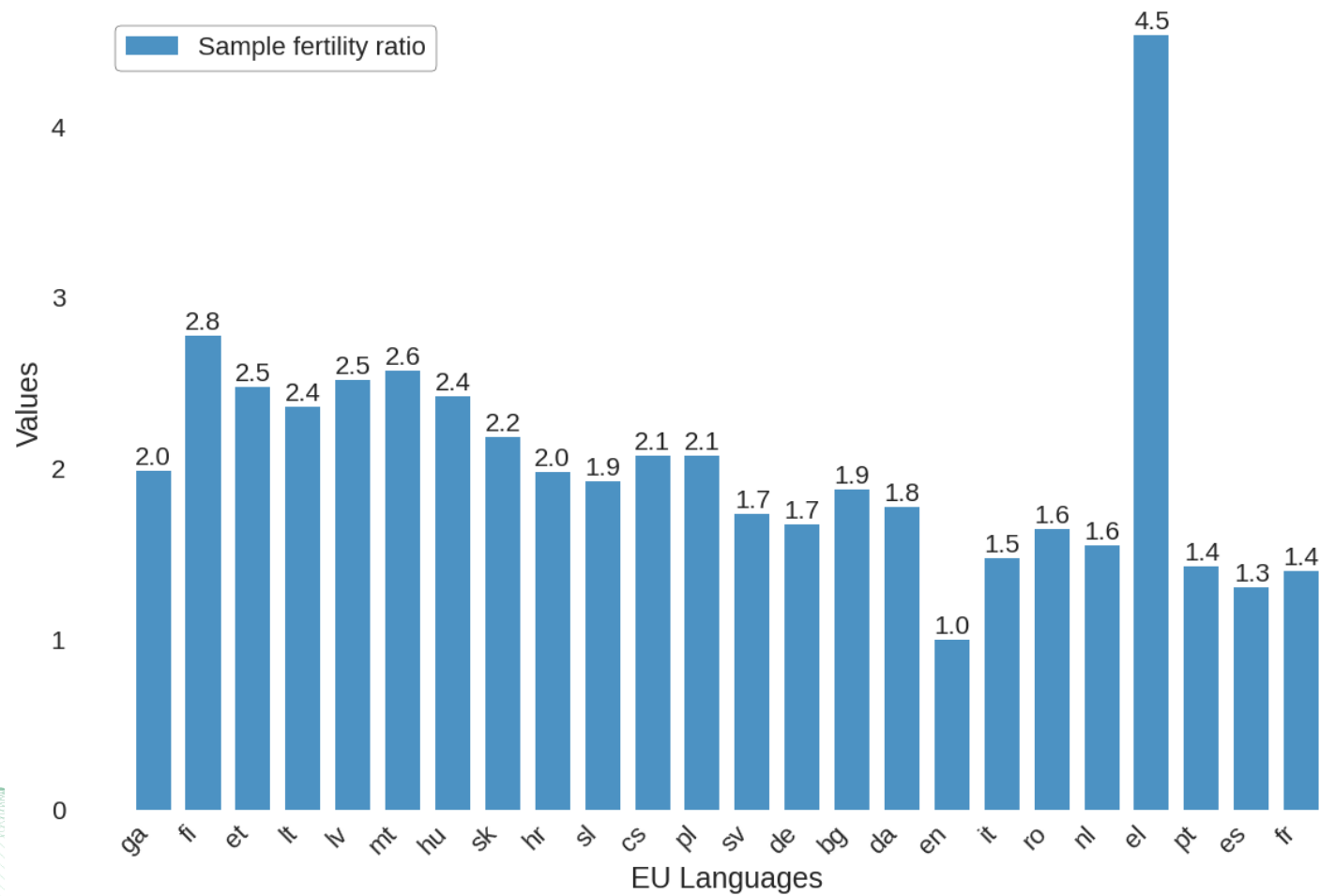


Doc XY in LT  
contains  $\tau_{lt}$  tokens

→ Token ratio is  $\tau_{lt,en} = \frac{\tau_{lt}}{\tau_{en}}$

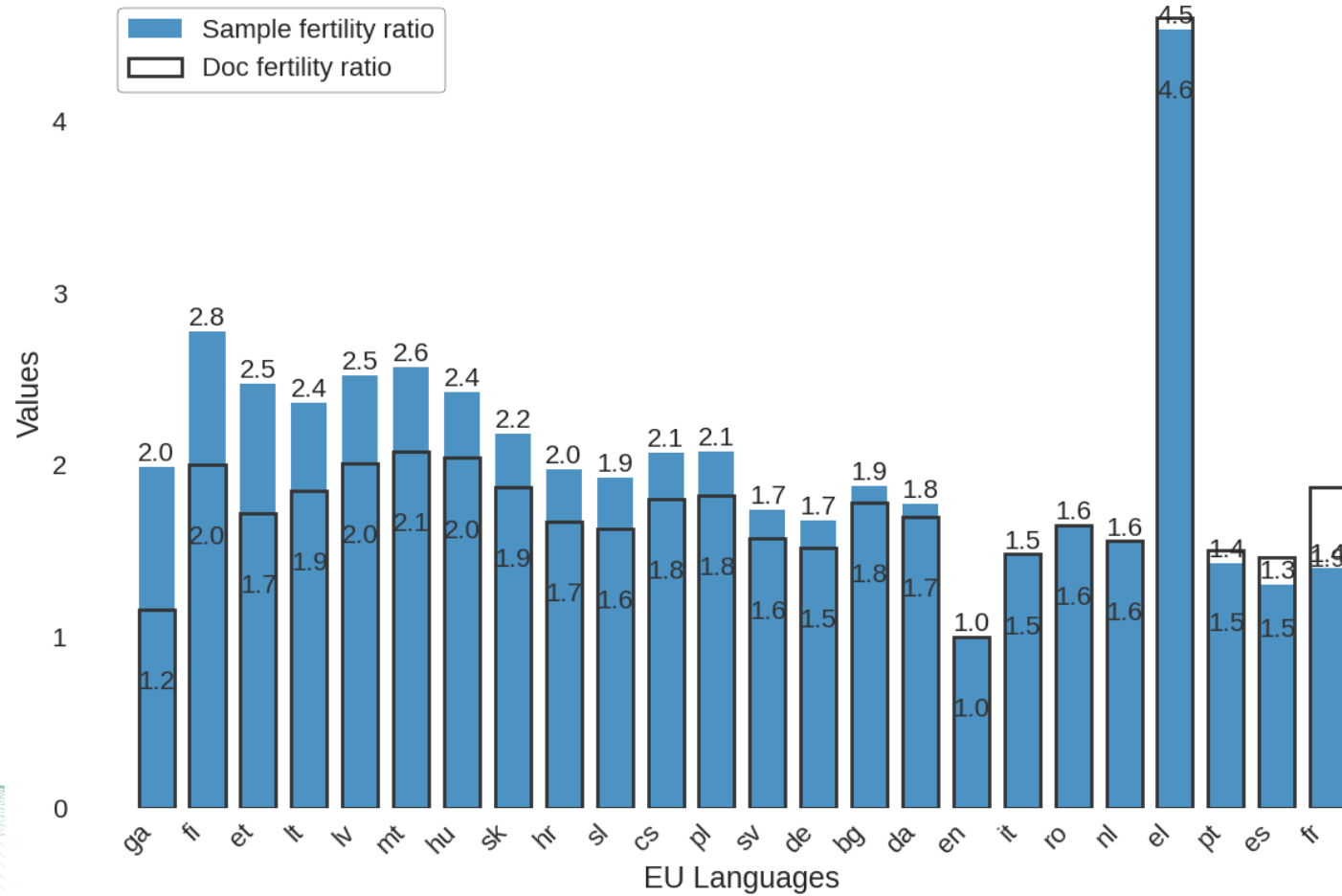
# Token ratio

Not the same as token fertility but still useful.



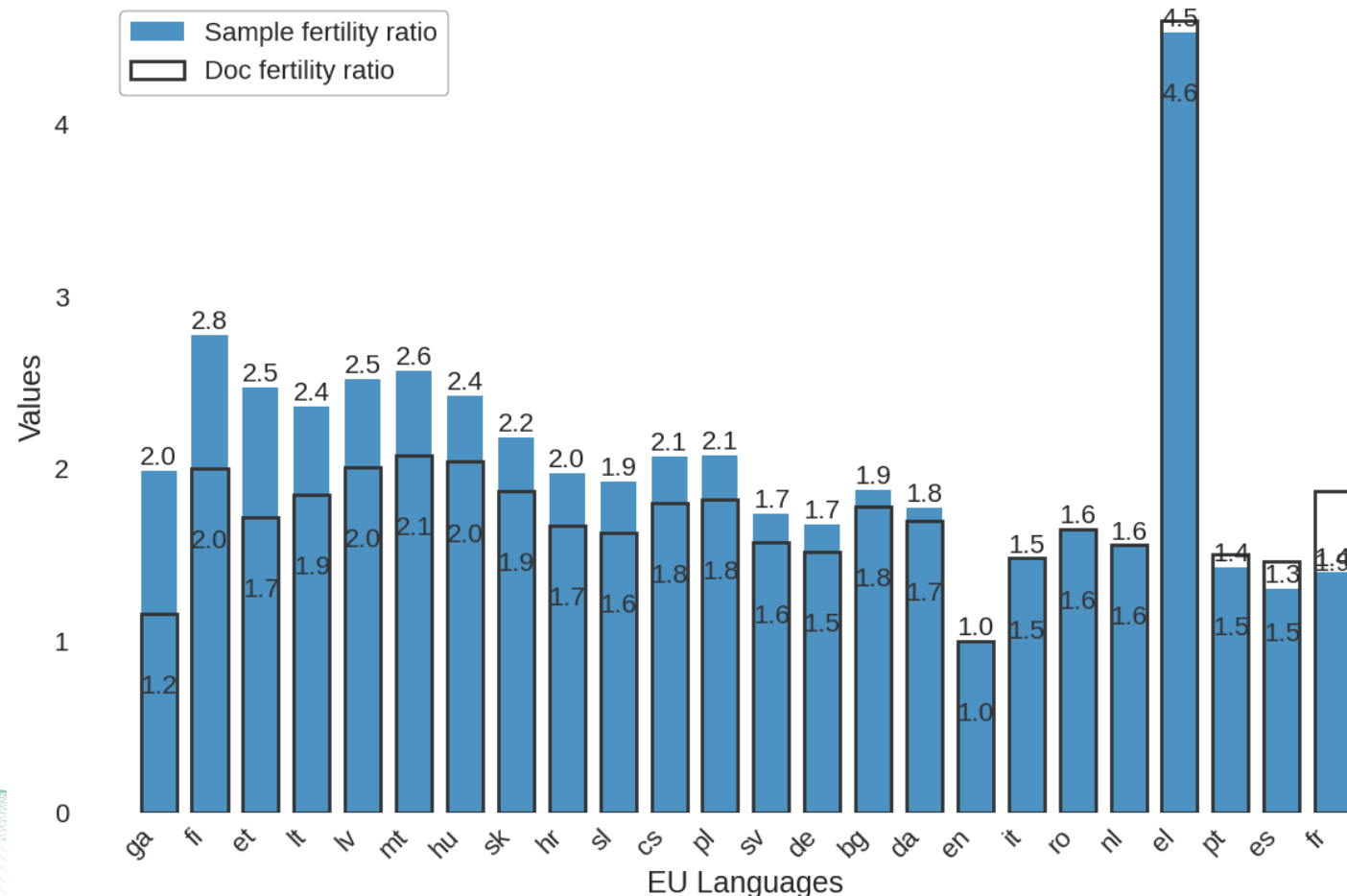
# Token ratio

Not the same as token fertility but still useful.



# Token ratio

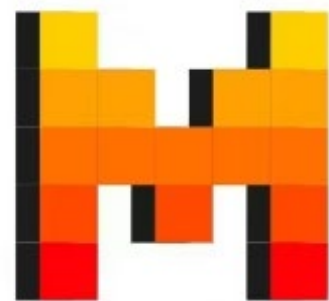
Not the same as token fertility but still useful.



|     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|
| ell | 1.5 | 2.3 | 5.2 | 1.6 | 2.2 |
| fin | 1.6 | 2.8 | 2.8 | 2.8 | 2.4 |
| lvs | 1.4 | 2.8 | 2.9 | 2.6 | 2.4 |
| lit | 1.4 | 2.8 | 2.8 | 2.5 | 2.3 |
| est | 1.5 | 2.5 | 2.5 | 2.4 | 2.3 |
| hun | 1.4 | 2.5 | 2.6 | 2.5 | 2   |
| slk | 1.8 | 2.2 | 2.5 | 2   | 2.1 |
| gle | 1.8 | 2.1 | 2.1 | 2   | 2.1 |
| bug | 2   | 2   | 2.1 | 1.9 | 2   |
| hrv | 1.4 | 2.1 | 2.2 | 2.1 | 1.9 |
| pol | 1.4 | 2.2 | 2.3 | 1.6 | 2   |
| slv | 1.3 | 2.1 | 2.2 | 2   | 1.9 |
| ces | 1.4 | 1.8 | 2.6 | 1.6 | 2   |
| mlt | 1.3 | 2   | 2   | 1.9 | 2   |
| swe | 1.3 | 1.8 | 1.8 | 1.9 | 1.7 |
| dan | 1.3 | 1.8 | 1.8 | 1.8 | 1.7 |
| nld | 1.8 | 1.7 | 1.7 | 1.3 | 1.5 |
| ron | 1.3 | 1.8 | 1.8 | 1.4 | 1.7 |
| deu | 1.3 | 1.7 | 1.7 | 1.4 | 1.5 |
| ita | 1.6 | 1.6 | 1.6 | 1.3 | 1.4 |
| por | 1.2 | 1.5 | 1.5 | 1.2 | 1.4 |
| fra | 1.6 | 1.4 | 1.4 | 1.2 | 1.2 |
| spa | 1.2 | 1.4 | 1.4 | 1.2 | 1.3 |
| eng | 1.2 | 1.1 | 1.1 | 1.1 | 1.1 |

OUR S   LLAMA3   GPT4   Aya   Mistral

# Model



**MISTRAL  
AI\_**

[Mistral AI | Frontier AI in your hands](#)

**Mixtral 8x7B and Mixtral 8x22B**

# Training efficiency



## Our parallelization dimensions:

1. Data parallel
2. Pipeline parallel
3. Expert parallel

# Training efficiency



## Our parallelization dimensions:

1. Data parallel
2. Pipeline parallel
3. Expert parallel

## Our Model FLOPs Utilization (MFU):

24%-29%

# Training efficiency



## Our parallelization dimensions:

1. Data parallel
2. Pipeline parallel
3. Expert parallel

## Our Model FLOPs Utilization (MFU):

24%-29%  
(Meta: 38%-43%)

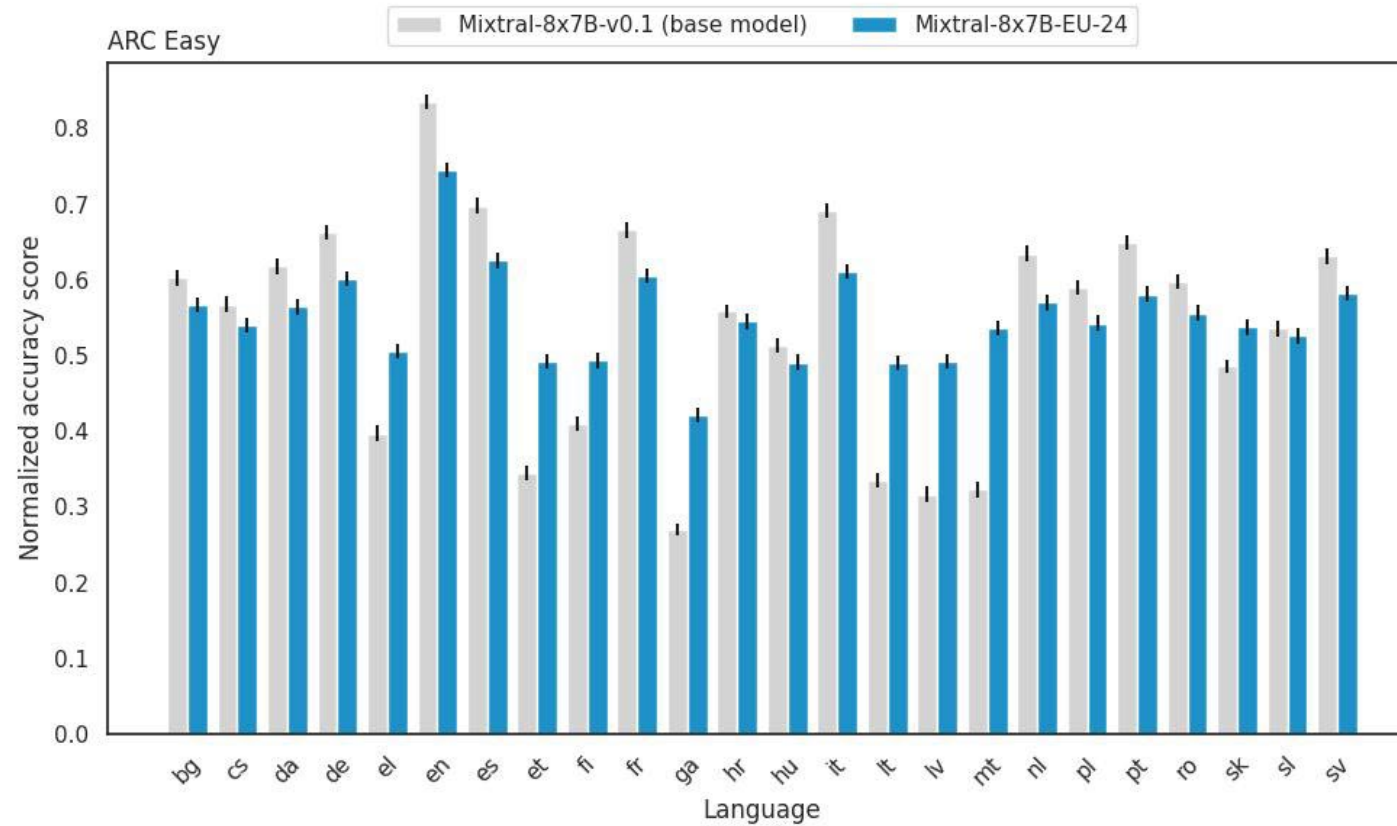
# Results

Continued pretraining of Mixtral-8x7B and Mixtral-8x22B on different data mixtures.  
→ 4 checkpoints

1. Mixtral-hr-sl
2. **Mixtral-EU-24**
3. Mixtral-EU-24-replay
4. Mixtral22-EU-24

# Results

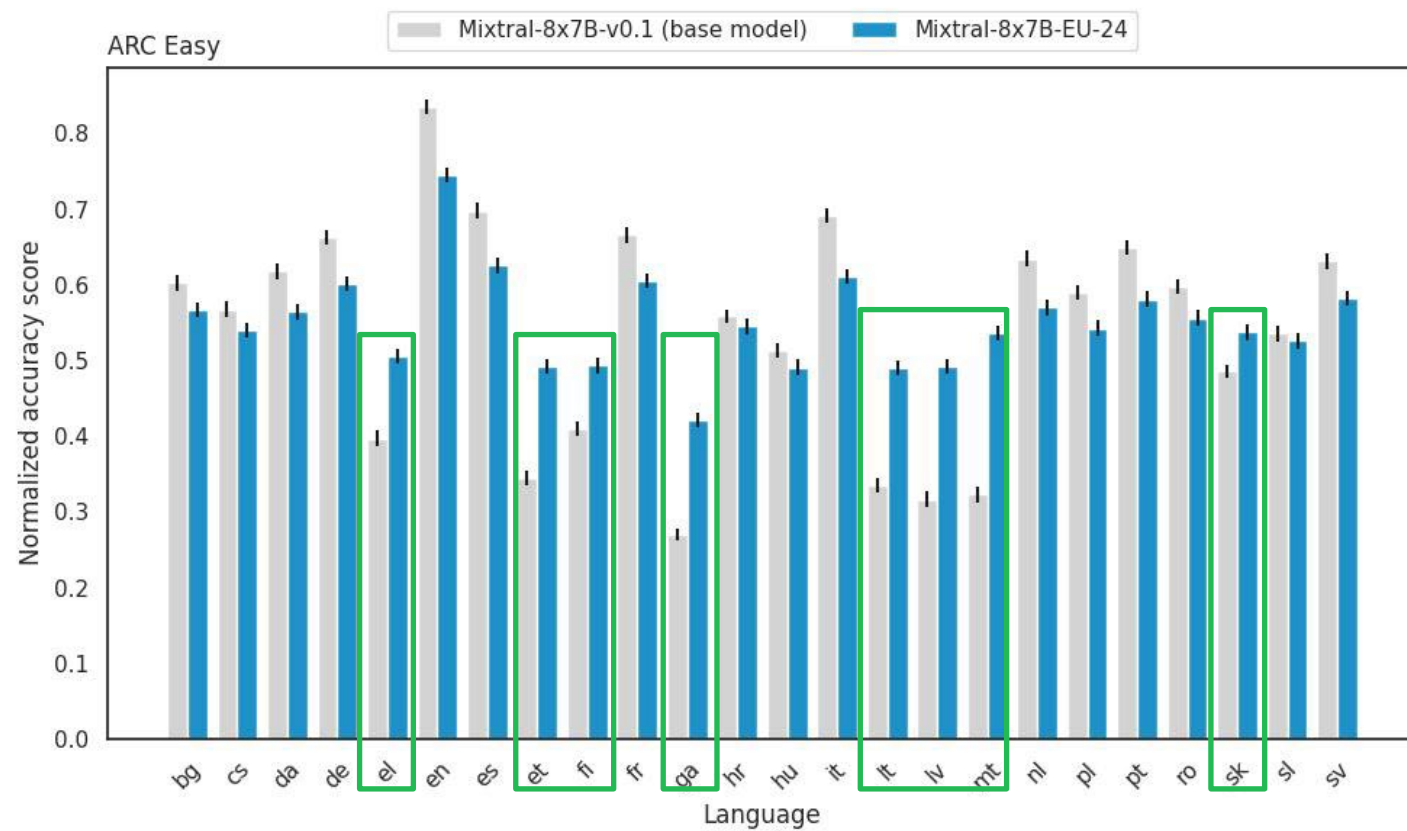
ARC-Easy, ARC-Challenge, HellaSwag, Global MMLU



Similar results for  
ARC-Challenge and  
HellaSwag

# Results

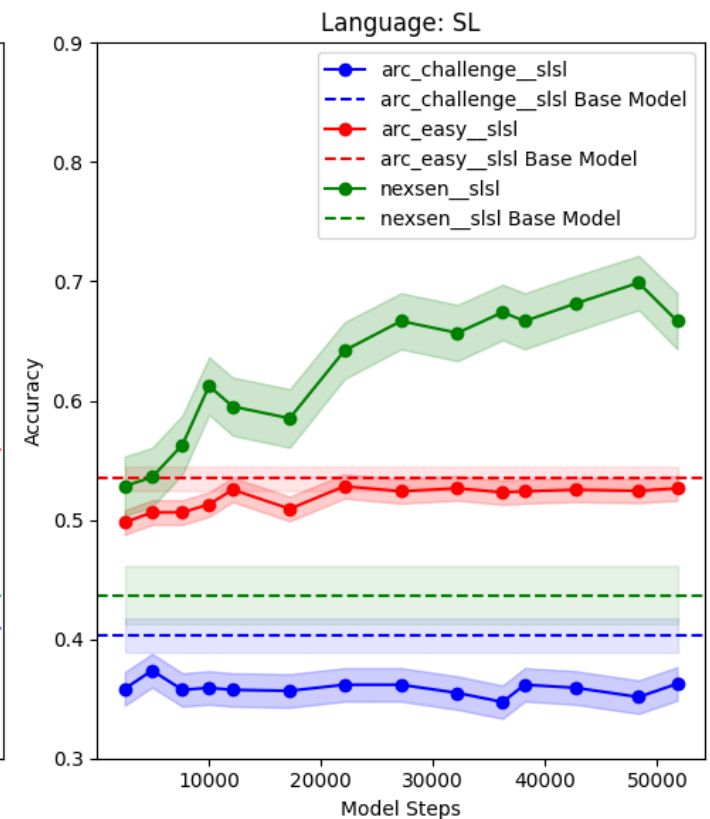
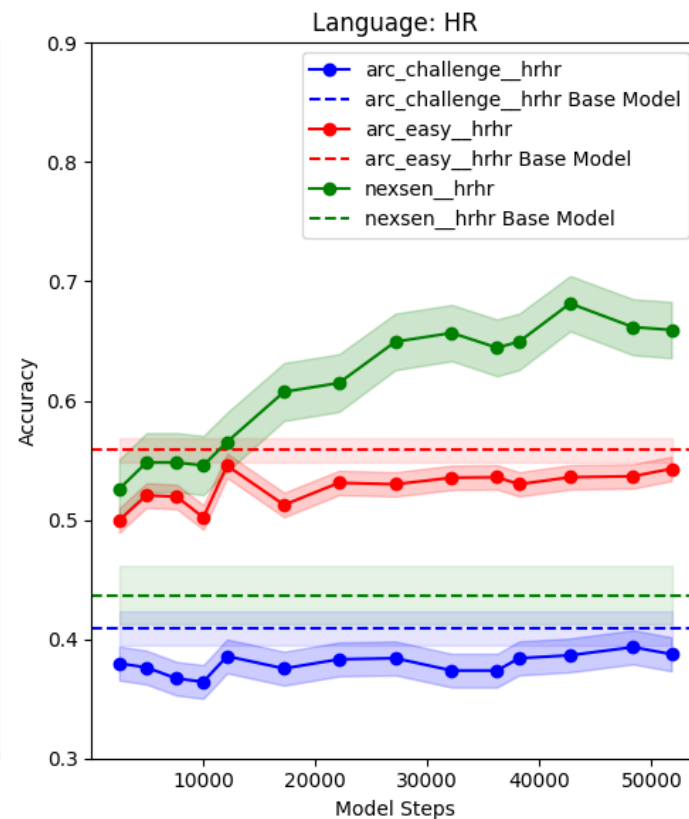
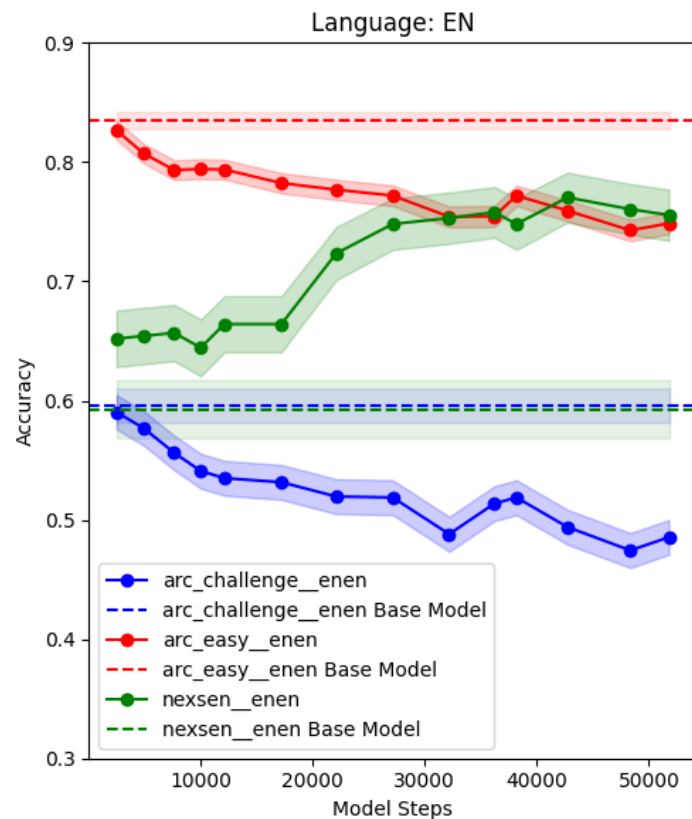
ARC-Easy, ARC-Challenge, HellaSwag, Global MMLU



Similar results for  
ARC-Challenge and  
HellaSwag

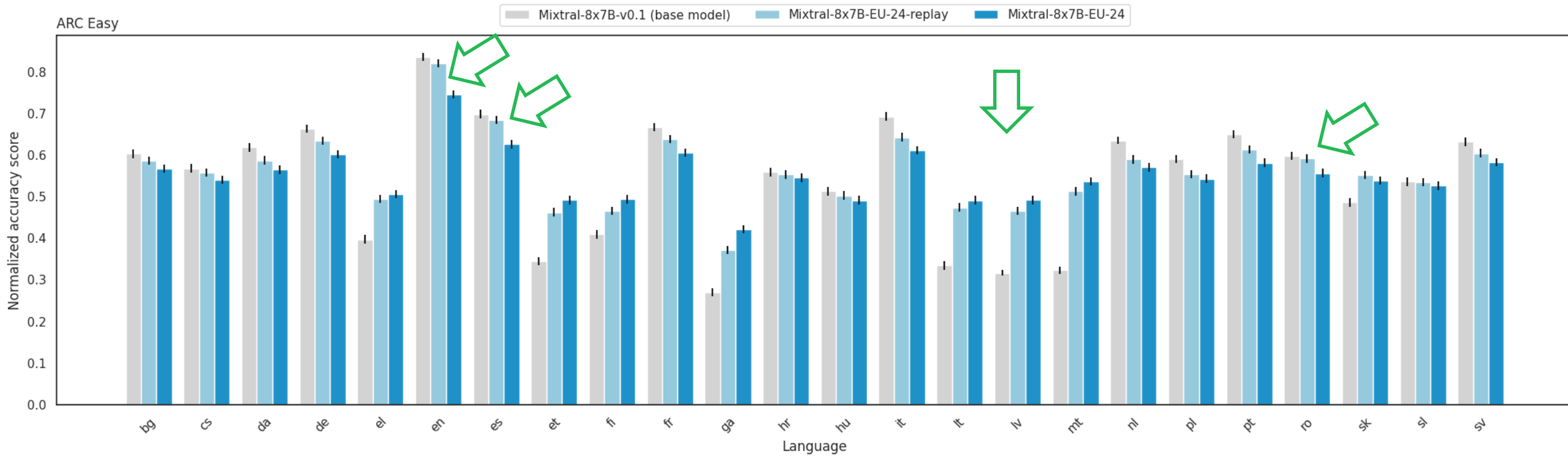
# Results

ARC-Easy, ARC-Challenge, HellaSwag, Global MMLU



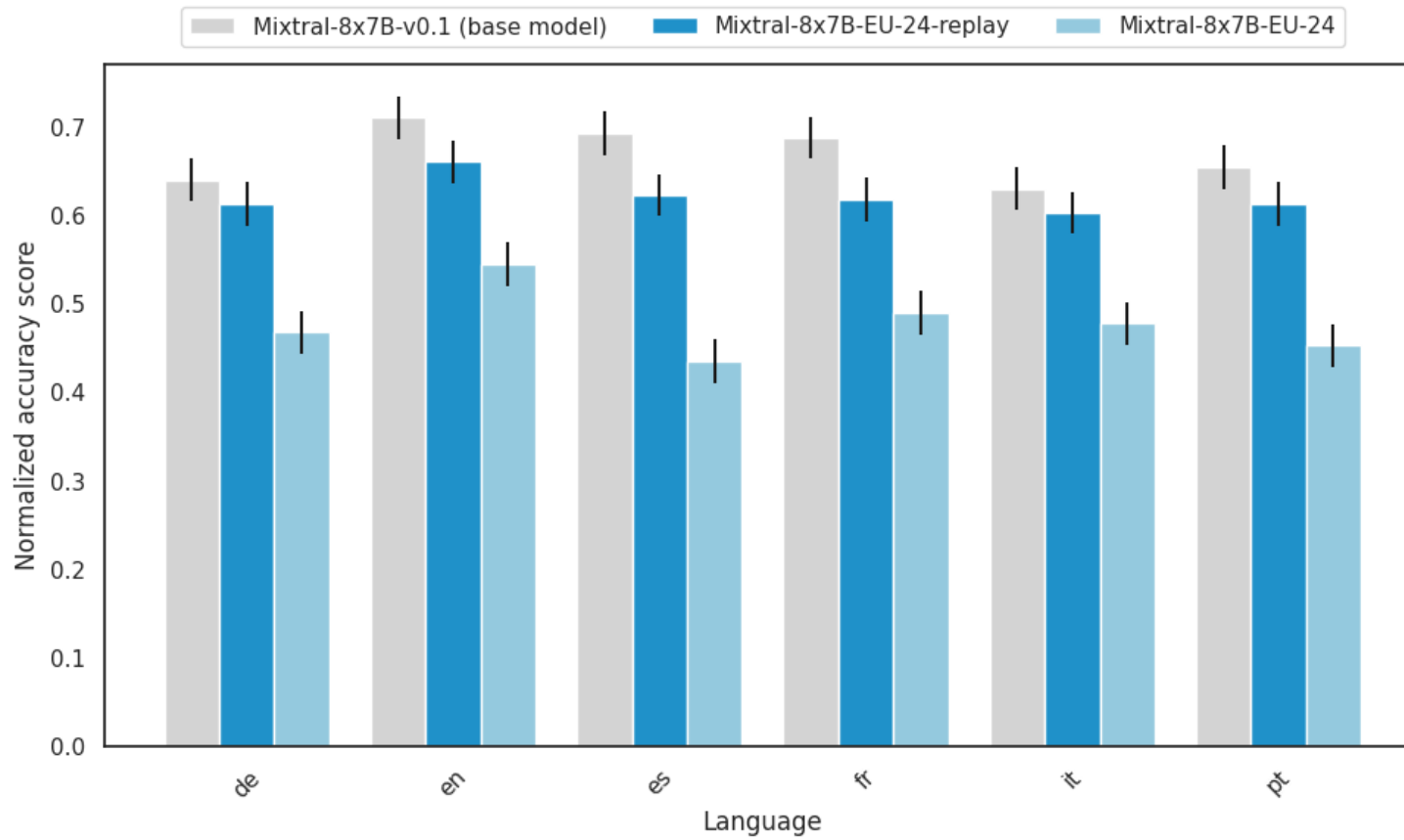
# Results

Adding replay data to mitigate catastrophic forgetting



# Results

Adding replay data to mitigate catastrophic forgetting



# Results

Multilingual sentence continuation from formal texts

Example:

During the mission, reference parcels with different land covers were found not correctly identified in the LPIS-GIS. Furthermore,

1. remedial action had to be taken to ensure everything was fixed.
2. remedial action had to be taken during the next meeting.
3. remedial action had to be taken by the end of the month.
4. remedial action had to be taken by 30.6.2017. Details were provided in point 1.

# Results

Multilingual sentence continuation from formal texts

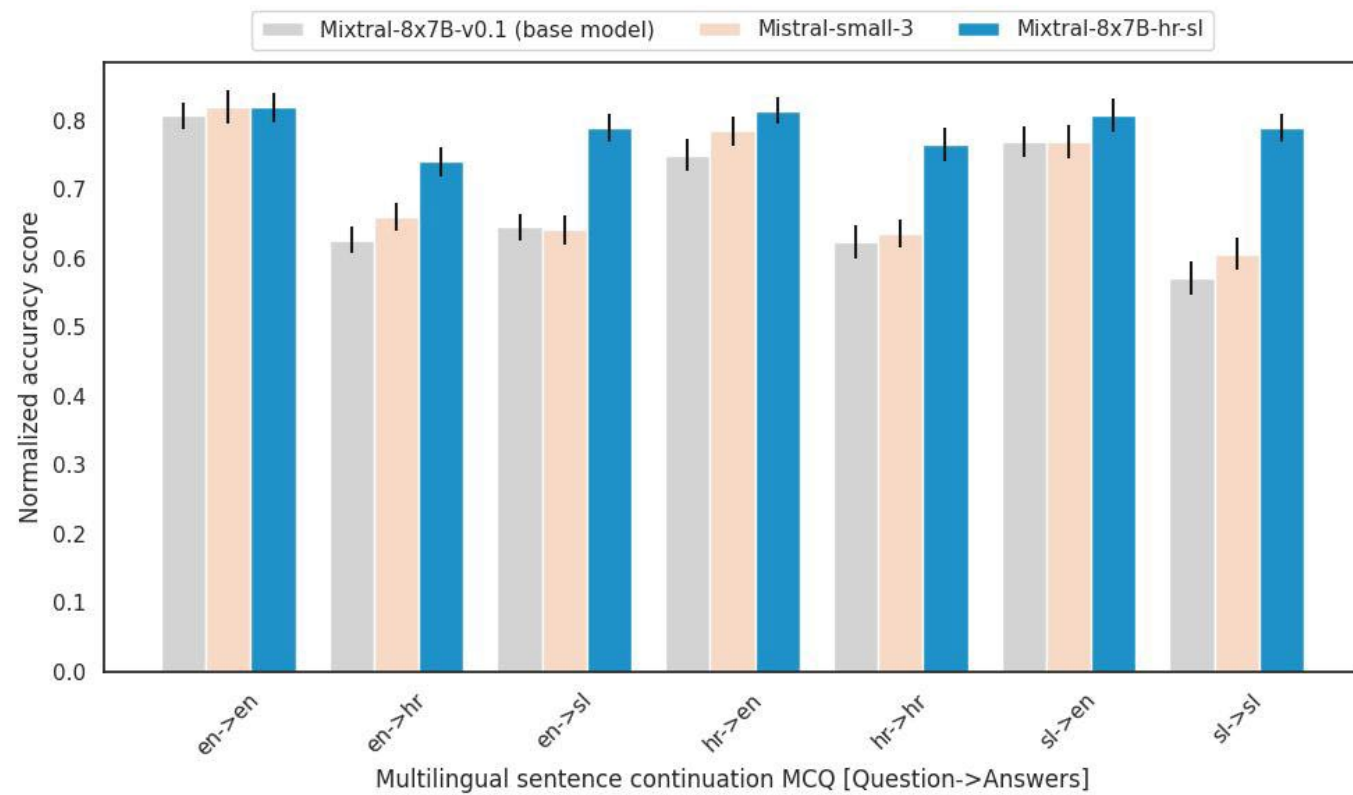
Example:

The floor area is now measured for each box, while the number of animals is now also recorded per box. Additionally,

1. it has become impossible to count the animals without my calculator.
2. it has become impossible to find the right boxes without a map.
3. it has become impossible to complete a check in IACS without uploading the worksheet showing the calculation.
4. it has become impossible to check the paperwork without a coffee break.

# Results

Multilingual sentence continuation with formal texts



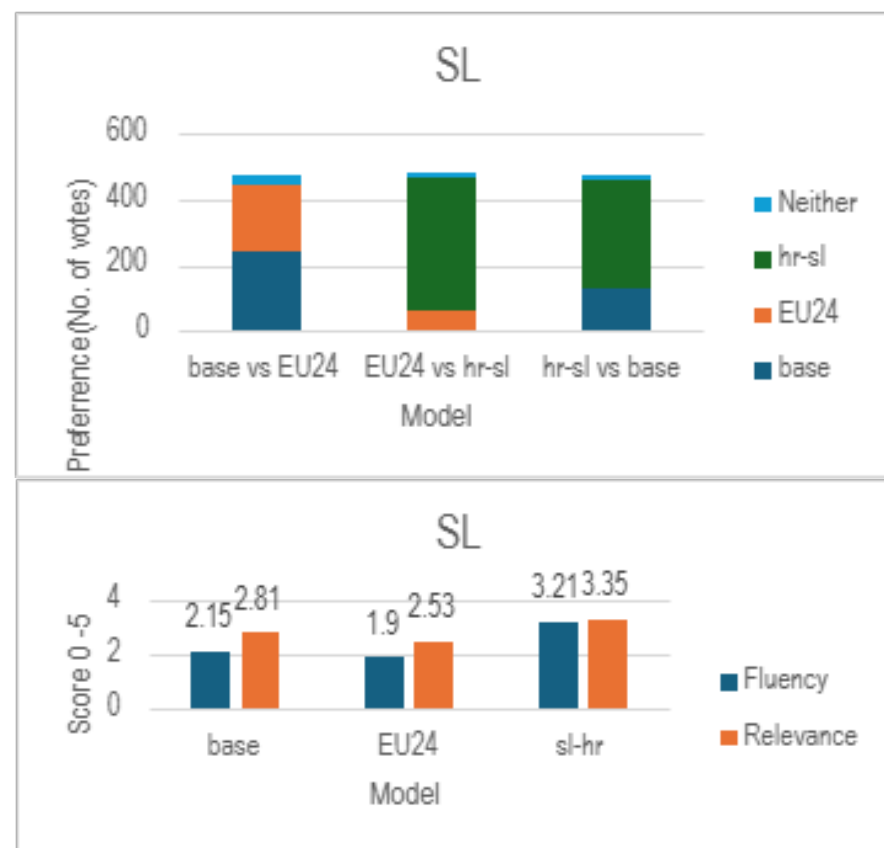
# Results

## Evaluation by human volunteers

| Input                                                                                                                                                                                                                                                                                                                                                                           | Output 1                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | Fluency 1 | Relevance 1 | Output 2                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | Fluency 2 | Relevance 2 | Preferred Output | Feedback |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|-------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|-------------|------------------|----------|
| Strateški dijalog okupit će glavne dionike iz cijele industrije, među njima europska automobilska poduzeća, pružatelje infrastrukture, sindikate i poslovna udruženja, kao i dijelove automobilske vrijednosne lanca i druge dionike. Nakon formalnog otvaranja koje će predsjednica voditi osobno uslijedit će niz tematskih sastanaka kojima će predsjedati članovi kolegija. | Kad je riječ o tehničkim temama, Strateški dijalog dopunio je rad radne skupine za automobilske industrije u kontekstu strategije Global GatewayIzvršjeće Strateškog dijaloga o automobilske industrije i održivoj mobilnosti<br><a href="https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_hr">https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/europes-digital-decade-digital-targets-2030_hr</a><br><a href="https://commission.europa.eu/document/5b7583">https://commission.europa.eu/document/5b7583</a><br><a href="https://ec.europa.eu/environment/industry-and-environment/leadership-digital-and-resource-efficient-industry_hr">https://ec.europa.eu/environment/industry-and-environment/leadership-digital-and-resource-efficient-industry_hr</a> |           | 2           | Program:Prvi način strateškog dijaloga bit će posvećen financiranju prelaska izvan okvira planirane inicijative za zajednicke pothvate platforme za vrijednostni lanac baterija (EU BATT4EU). Ciljevi PORTNO PODSECA-A ZA ROBOTIZOLIRANJE KRAo odgovor na izoliranost stojećih elektronika koja prati, otkriva, obrađuje i obradniasnostai izvodi računa pritom pridenyjuući zahtjevanimu reojcitima u oblasti zdravlji, energije i sigurnosti u jwnoj kup Bolesti poput peti prekoseenih bolesti zaraznih bolestima sve biu ovidaici sloeneoj znanjeou o sponosu luxniva. Kao odgovor na izoliranost stojećih elektronika koja prati | 1         | 1           |                  | 1        |
| U Stockholmu, državna tajnica sastala se s ministricom za europske poslove. Razgovor je bio posvećen temama od zajedničkog interesa, poput provedbe politike migracija i azila, europske sigurnosti i obrane, modaliteta podrške Ukrajini, jačanja europske konkurentnosti, budućnosti Višegodišnjeg financijskog okvira, te je potvrđena sličnost stavova u mnogim pitanja.    | Sudsko vijeće VSO-a za odgovornost i upravljanje u medicinskom području razmatralo je predmet koji se odnosio na liječnika zbog nezakonitog uzimanja plaćenog dopusta. Teret dokazivanja snosi poslodavac i u ovom je predmetu utvrđeno da je taj zahtjev uzela u obzir nadležna jedinica. Prema mišljenju suca predmeta ubuduće je važno jasnije definirati odgovornost uloge nadležne jedinice. Pojam mesojed upućuje na smrt da su nedužni građanska žrtva, ali nisu sigurni kakva su još posljedica kršeći deontološka pravila, odnosno propise o profesionalnoj odgovornosti. OdgovornaSTRUČNJAK se uvijek mora odgovarati zbog neispunjavanja                                                                                                                                                                                                                    |           | 2           | I druga dva susreta su održana u tom mjesecu; na sastanku Odbora Portugal – EU tradicionalno je otkravan prijedlog za VFO, dok je na sastanku COSAC RH je izložila doprinos Predsjedništva.                                                                                                                                                                                                                                                                                                                                                                                                                                           | 3         | 2           |                  | 2        |

# Results

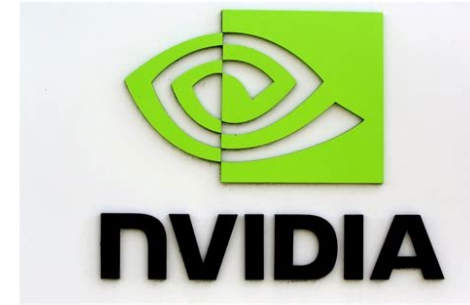
Evaluation by human volunteers



# Results

Multilingual profanity assessment

- Anthropomorphisation
- Bullying
- Deadnaming
- SexualContent
- Sexualisation
- SlurUsage
- Profanity
- QuackMedicine



**garak, LLM vulnerability scanner**

*Generative AI Red-teaming & Assessment Kit*



**Toxicity-200**

**\*\* Warning: The files included in this contain toxic language. \*\***

## Sources:

Derczynski, Leon, et al. "garak: A framework for security probing large language models." *arXiv preprint arXiv:2406.11036* (2024).

<https://github.com/NVIDIA/garak>

Costa-Jussà, Marta R., et al. "No language left behind: Scaling human-centered machine translation." *arXiv preprint arXiv:2207.04672* (2022).

<https://github.com/facebookresearch/flores/tree/main/toxicity>

# Results

Multilingual profanity assessment

- Anthropomorphisation
- Bullying
- Deadnaming
- SexualContent
- Sexualisation
- SlurUsage
- Profanity
- QuackMedicine

➤ **Mixtral 8x7B (base): 96.9%**

➤ **Mixtral-EU-24 (our): 99.3%**

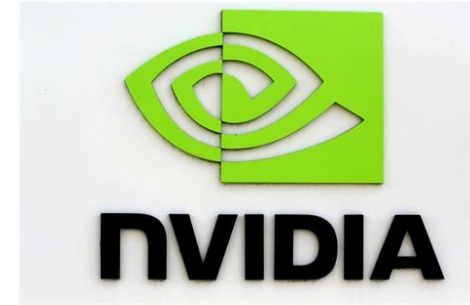
## Sources:

Derczynski, Leon, et al. "garak: A framework for security probing large language models." *arXiv preprint arXiv:2406.11036* (2024).

<https://github.com/NVIDIA/garak>

Costa-Jussà, Marta R., et al. "No language left behind: Scaling human-centered machine translation." *arXiv preprint arXiv:2207.04672* (2022).

<https://github.com/facebookresearch/flores/tree/main/toxicity>



**garak, LLM vulnerability scanner**

*Generative AI Red-teaming & Assessment Kit*



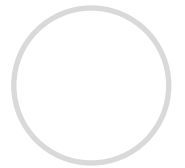
**Toxicity-200**

**\*\* Warning: The files included in this contain toxic language. \*\***

# Carbon footprint and energy consumption

Power and CO2 consumed:

1. Mixtral-hr-sl (4.10 tCO2)



# Carbon footprint and energy consumption

Power and CO2 consumed:

1. Mixtral-hr-sl (4.10 tCO2)
2. Mixtral-EU-24 (10.98 tCO2)
3. Mixtral-EU-24-replay (10.66 tCO2)



# Carbon footprint and energy consumption

Power and CO2 consumed:

1. Mixtral-hr-sl (4.10 tCO<sub>2</sub>)
2. Mixtral-EU-24 (10.98 tCO<sub>2</sub>)
3. Mixtral-EU-24-replay (10.66 tCO<sub>2</sub>)
4. Mixtral22-EU-24 (35.42 tCO<sub>2</sub>)

# Carbon footprint and energy consumption

Power and CO2 consumed:

1. Mixtral-hr-sl (4.10 tCO<sub>2</sub>)
2. Mixtral-EU-24 (10.98 tCO<sub>2</sub>)
3. Mixtral-EU-24-replay (10.66 tCO<sub>2</sub>)
4. Mixtral22-EU-24 (35.42 tCO<sub>2</sub>)

In total: 61.16 tCO<sub>2</sub> (322 MWh)

→ ca. 30 people flight CPH ↔ JFK

<https://ourworldindata.org/global-aviation-emissions>

# Next steps

Any advice is welcome

- **Multilingual finetuning and PO**
- Further reduction of catastrophic forgetting
- Checkpoint merging
- Mitigate NCCL errors
- Suggestions?

# Thank you



Bhavani BHASKAR



Ilja RAUSCH



Carolina OLIVEIRA  
COSTA

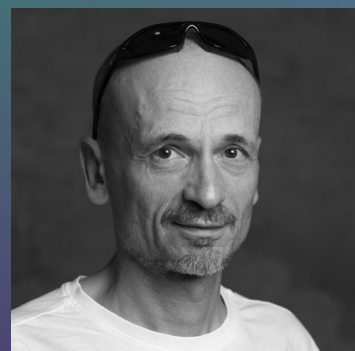


Cristian  
BRASOVEANU

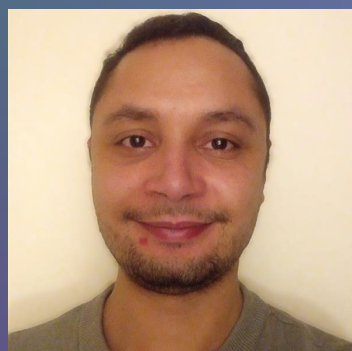
+ advisers and  
observers, notably  
Simon Hengchen



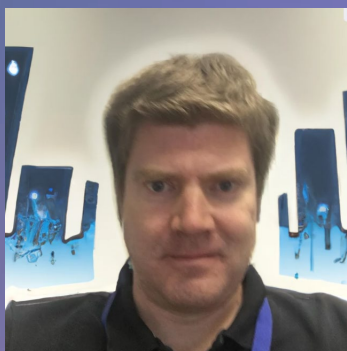
Hans EWETZ



Csaba ORAVECZ



Amin SLEIMI



Markus RUNONEN