Europe's Technical Debt: Why We Need Web Search in the Age of Generative AI

Malte Ostendorff, Pedro Ortiz Suarez, Julián Moreno-Schneider, Georg Rehm

German Research Center for Artificial Intelligence

5th International Open Search Symposium

Status quo

- **Generative AI has completely changed the landscape of machine learning**, allowing researchers and practitioners to tackle tasks thought to be impossible.
- Generative AI is **dominated by US-based enterprises**. Europe is lagging behind in developing large AI models and is expected to have a hard time catching up.
- One of the reasons is **the technical debt that Europe** has been accumulating since it lost and stopped competing in the race of the previous technological revolution:

  → **Web search**

Status quo

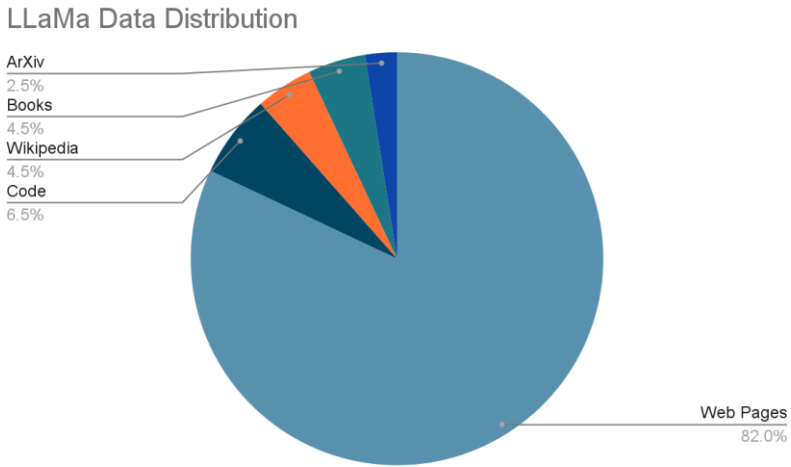
Progress in generative AI is mainly driven by two factors:

Computational power and **data** (neglecting algorithmic improvements).

- Computational resources are an easier-to-fix problem: Europe is actively investing in its compute infrastructure, reallocating resources, and making them accessible for AI research, such as through initiatives like the **EuroHPC Joint Undertaking**.
- The real issue lies in the deficiency of the second key ingredient – Web data and its retrieval, which can be attributed to the **absence of strong European Web search initiatives**.

Pretraining Datasets

- Large language models (LLMs) and other generative models are statistical models based on training data.
- State-of-the-art LLMs (such as Meta's LLaMa) are trained with **one trillion tokens** or even more.
- **The Web is the most prominent source** that provides data at the required scale, accounting for a significant portion of the training data for recent LLMs, **often more than 80%**.
- Especially Web data from **Common Crawl,** or processed versions such as **OSCAR,** is widely used for LLM training.
- The reliance on Web data introduces several limitations, especially in the European context.



Web Data for Pretraining

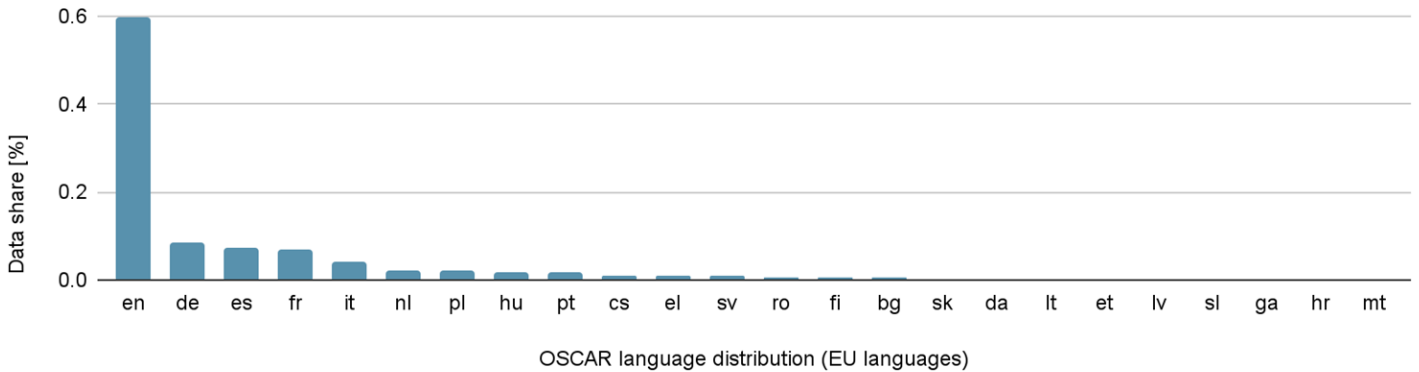
- Web crawls from Common Crawl are only a sample of the whole Web.

  → **Important European websites might be omitted.**

- The Common Crawl crawler operates with a **US user-agent** and an **IP number located in the US.**

  → The crawler appears to websites as a user from the US.

  → English language content represents the largest share of Common Crawl data by far (30%).



Recent Releases

Our recent work resulted in two releases: Colossal OSCAR v1.0 and lm-datasets

- **Colossal OSCAR** is the largest release of the OSCAR Corpus based on 10 different monthly snapshots of Common Crawl. It contains Web-crawled data with more than one trillion tokens – including quality annotations.
- **lm-datasets** is a software framework that unifies the downloading, preprocessing, and sampling of training data for language model training. It covers 400 datasets from 58 sources in 32 European languages.

A European Web Crawl is Needed

- A European Web crawl is needed to collect a training dataset that **adequately covers Europe's diversity including its languages, countries, and cultures**.
- There are already ongoing projects and initiatives working on this or related problems - but we need to **strengthen them to obtain valid extensions to Common Crawl**.
- Something as crucial as the training data of AI models should not solely depend on a single Californian non-profit organization that operates on AWS-donated infrastructure.

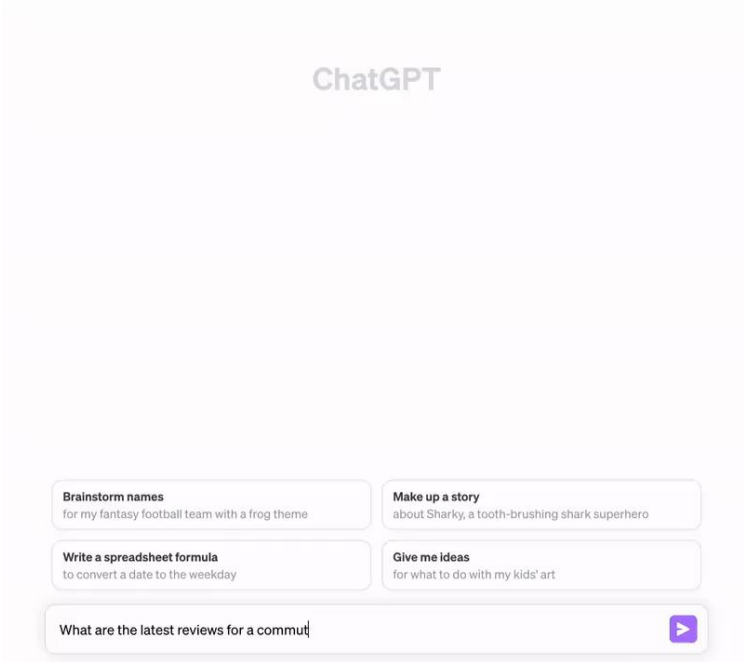
Web-based Retrieval Augmentation

Web-based Retrieval Augmentation

- Generative models have severe shortcomings, e.g., **outdated knowledge** and factually **incorrect information** (“hallucinations”).
- **Retrieval augmentation** could address these issues.
- General idea: Retrieve factual and updated information from trustworthy sources and generate output based on the retrieved information.
- Retrieval-augmented LLMs require information retrieval systems.

Example: ChatGPT with Bing’s Web Search

- As with pre-training, the Web represents **the most extensive resource from which information can be retrieved**.
- But: Building a **retrieval-augmented LLM obviously requires Web search**, making it, once again, quite difficult for Europe to compete since no European Web search exists.
- Relying on Web search APIs from one of the big technology enterprises is **no valid option** either since it would introduce a strong dependency, hampering technological sovereignty.
- **Microsoft tripled the prices of the Bing Search API** briefly after the introduction of their own retrieval-augmented LLM.



European Web Search is Needed

- Simply providing Web-crawls as pre-training data will not be sufficient.
- For future generative AI models, retrieval augmentation will be crucial.

  → **European Web search APIs are needed!**

Conclusions

- Europe's lack of investment in Web search infrastructure, crawling and retrieval, has led to **a significant technical debt**.
- To be able to compete in the next technological revolution, generative AI, **this debt needs to be paid off**.
- Although catching up is feasible, it requires a collective effort and substantial investment from industry and academia.