Promoting Open Science in times of Artificial Intelligence: Do we grasp the interplay?

(a self-reflecting case study of two current projects[1])

Marie Alavi, Nicolaus Wilder, Julia Priess-Buchheit

NERQ NETWORK FOR EDUCATION AND RESEARCH QUALITY

VK:KIWA

Training process of Base Models[2]: Machine Learning

Wikis & Encyclopaedias

Open Science according to RCR

Common Crawl

WebText

Open data

Books

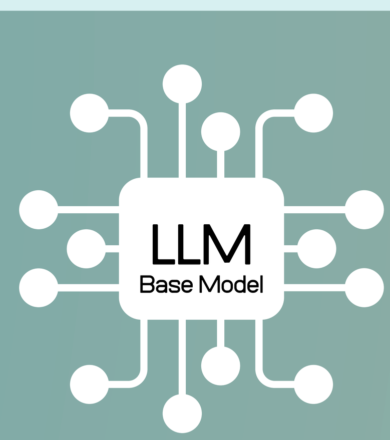
OER

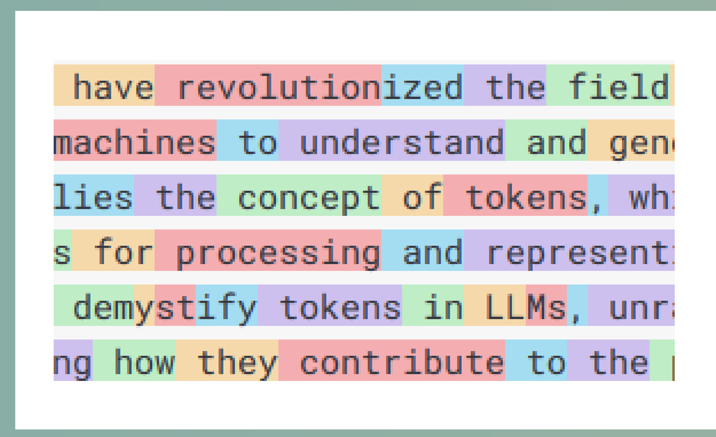
Peer-reviewed open publications

WWW

Social media

Scientific journals





Training data: large body of information from a **variety of sources** (freely and openly available on the Internet). Not entirely known to the user (modified subsequently through RLHF).

Probabilistic modeling of token sequences: Data converted into '**probabilities of occurrence of tokens**'

Current practice

"What is the 'Theory of relativity'?"

Innovations?

Data differing from the mainstream?

Result containing (reliable) knowledge.

Prompt

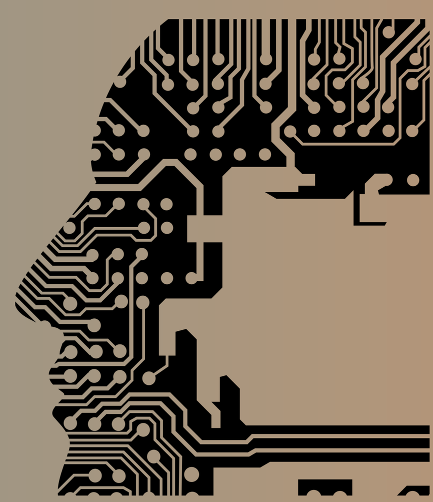
The greater the ***data prominence****, the higher the* ***probability of occurrence****.*

Black box (stochastic parrot): **Probabilistic Modeling of Token Sequences**

Searching for ***associations between tokens****,* ***predicting the probabilities of token sequences*** *based on the context provided by previous tokens and* ***generating outputs based*** *on the tokens' statistical and probabilistic nature.*

Promoting OS with RI and RE as umbrella concepts influencing data occurence

VS.



Using LLMs based on Probabilistic Modeling of Token Sequences

- **Which data** is used for training and probabilistic modeling of token sequences?
- **What** are the **operational mechanisms** of a LLM (cf. "black box")?
- Are certain **RI and RE principles** (such as accountability, interpretability, factuality, explainability, safety, reliability, traceability, etc.) **compatible** with LLM functionality?
- How can we ensure **RI and RE principles at the intersection of AI and OS without lowering the online prominence of OS data**?

LLM

Probabilistic Modeling of Token Sequences:

- The parameter for weighting probabilities is the **quantitative prominence of available data**
- prominent data: widely available online, easily discoverable, frequently repeated, cited, referenced, reviewed, edited, widely disseminated[3])

On the best way with OS:
Achieving data prominence involves factors such as **open access publishing, open data repositories** etc.

Training (of LLM) and prediction functionality require large online data to follow linguistic rules (Estimates: GPT-4 trained on 13 trillion tokens - ca. 10 trillion words[4]).

RI, RE & OS

These RI/RE principles and IP rights make OS (reliable) data under-represented for AI training:

- **Transparent and open sharing, collaboration, building on existing data** and **preprint archives** reduce the likelihood of redundant studies
- Peer review assessing **originality**, validity, and significance of a study limits publishing similar study
- Protecting **Data Privacy** and **Confidentiality**, preventing unauthorized access
- Prioritizing quality over quantity: prevention of salami publication, re-publishing etc.
- Safeguarding/witholding data from harm or misuse (**ethical need)**
- Utilising **Intellectual Property Rights** to limit access

OS sources will not suffice to train a proper LLM.

How can we deal responsibly with these fundamentally different logics? Three perspectives:

At the intersection of OS and LLMs,

- **end users** must be aware of the different logics, principles and limitations in order to use LLMs with a critical and reflective approach in the scientific process.
- **developers** must be sensibilised to the principles of good research practices and reinforce these responsibly in the training processes.
- **research community** must reflect on its own processes and principles in light of these developments and readjust them if necessary.

Promoting OS as widely as possible could overcome the current under-representation, thereby refining the quality of knowledge that is freely accessible.

References: (1) AI-based assistance system for higher education (Project HAnS) and NERQ_OS findings; (2) Base Models: Standard pre-trained LLMs like ChatGPT (in distinction to RAG-Enhanced or Fine-Tuned LLMs); (3) Karamolegkou, A., Li, J., Zhou, L. & Anders Søgaard (2023). Copyright Violations and Large Language Models. In Procee-dings of the 2023 Conference on Empirical Methods in Natural Language Processing, p. 7403–7412, Singapore. Association for Computational Linguistics.; (4) Schreiner, A. (2023, July 11). *GPT-4 architecture, datasets, costs and more leaked.* the decoder. https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/