Constrained Retrieval-Augmented Language Models

Fundierte Sprachmodelle auf der Grundlage proprietärer Daten

Christopher Schröder[1,2] and Lukas Gienapp[2,3,4] | [1]Institute for Applied Informatics e. V., [2]ScaDS.AI Dresden/Leipzig, [3]University of Kassel, [4]hessian.AI

Using large language models is often subject to technical and legal constraints.

Illustrative (non-exhaustive) list of constraints:

- Non-Reproduction: Prevent (complete) reproduction of training material.
- Attribution: Refer to source material for transparency and traceability.
- License: Restrict training to data with clear licenses to be legally-compliant.

CORAL investigates constrained retrieval-augmented language models to make artificial intelligence more flexible, resilient, and efficient.

Project Goals

CORAL aims to research methods for the construction and use of large language models (LLMs) that are subject to legal, technical, and qualitative constraints.

Our focus is on two central criteria that are indispensable for the professional use of LLMs:

- Fulfillment of legal requirements for LLM training data
- Referential provenance of generated texts

To this end, we are researching new methods for the constrained training of LLMs and the retrieval-augmented generation (RAG) of texts.

Data

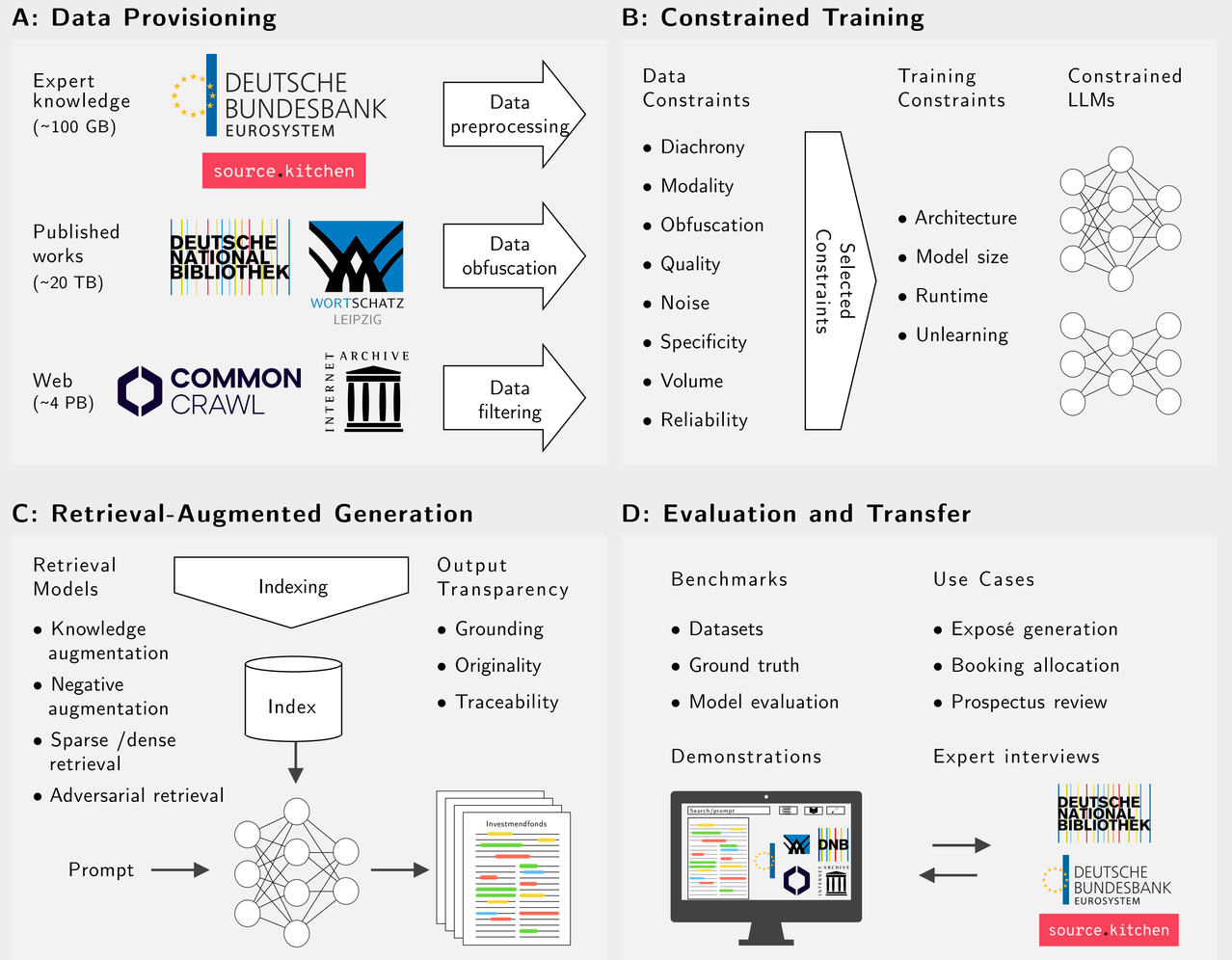
CORAL leverages data contributed by its partners, including the digital holdings of the German National Library (DNB); large-scale web crawls amounting to petabytes from the Internet Archive and Common Crawl; many years of European-language news crawls from Wortschatz Leipzig; and proprietary datasets from the financial sector.

Research Questions

Central research questions are:

- Which training methods and model architectures are robust against data constraints?
- How can large language models be trained with greater resource efficiency?
- Which methods of obfuscation, un-learning and negated augmentation effectively prevent the disclosure of protected data?
- How can the transparency and soundness, originality, and referenceability of the generated texts be ensured? How vulnerable are the methods used to secure the training data of LLMs?

CORAL is thus making important contributions to the future establishment of a German market for LLMs.



First and Expected Results

First results include:

- German Commons: 154 B permissively-licensed tokens in German
- Training of permissively-licensed smaller models

We aim for innovative results and insights in three core areas:

1. Consideration of training data that was previously unusable
2. Model architectures that handle certain constraints
3. Referential provenance of generated text

More Information and Contact

Project Lead: Prof. Dr. Gerhard Heyer
heyer@infai.org

coral-nlp.github.io
github.com/coral-nlp
huggingface.co/coral-nlp

With funding from the:
Federal Ministry of Research, Technology and Space

Associated Partners
DEUTSCHE BUNDESBANK EUROSYSTEM
source.kitchen

Computational Resources
ScaDS.AI DRESDEN LEIPZIG
hessian.AI
ZIH
JÜLICH Forschungszentrum
Webis Group