Building Open LLMs for Europe

Dr. Malte Ostendorff @ ItalAI 2024

About Me

Research Engineer (now)

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN IN PUBLICA COMMODA
SEIT 1737

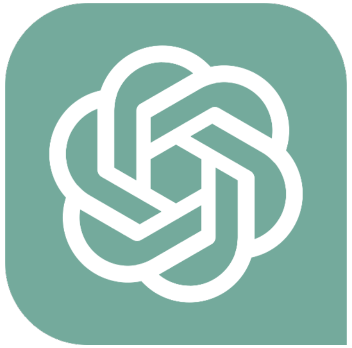
PhD

dfki
ai
Deutsches Forschungszentrum
für Künstliche Intelligenz
German Research Center for
Artificial Intelligence

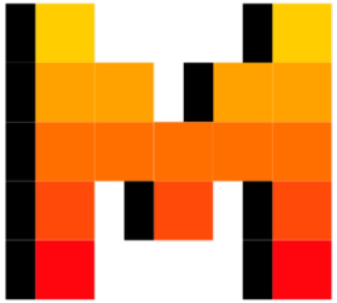
Researcher (until May 2024)

Open Legal Data

Occiglot



?



?

Recap: Large Language Models

- Language models are **statistical models** that learn a probability distribution over a sequence of words from their training data and that predict **the next token with the highest probability** for a given input text.
- Tokenization converts natural language text into numerical vector representations based on a **fixed and limited vocabulary**.
- Transformer architecture allows the scaling of language models **in terms of parameters, data, and compute** (resource requirements).
- Large-scale of today's language **model enables generalization and solving of novel tasks** with no or little additional training data (few- and zero-shot).

How to build a large language model?

… in the open and for Europe.

Open LLMs?

Open LLMs?

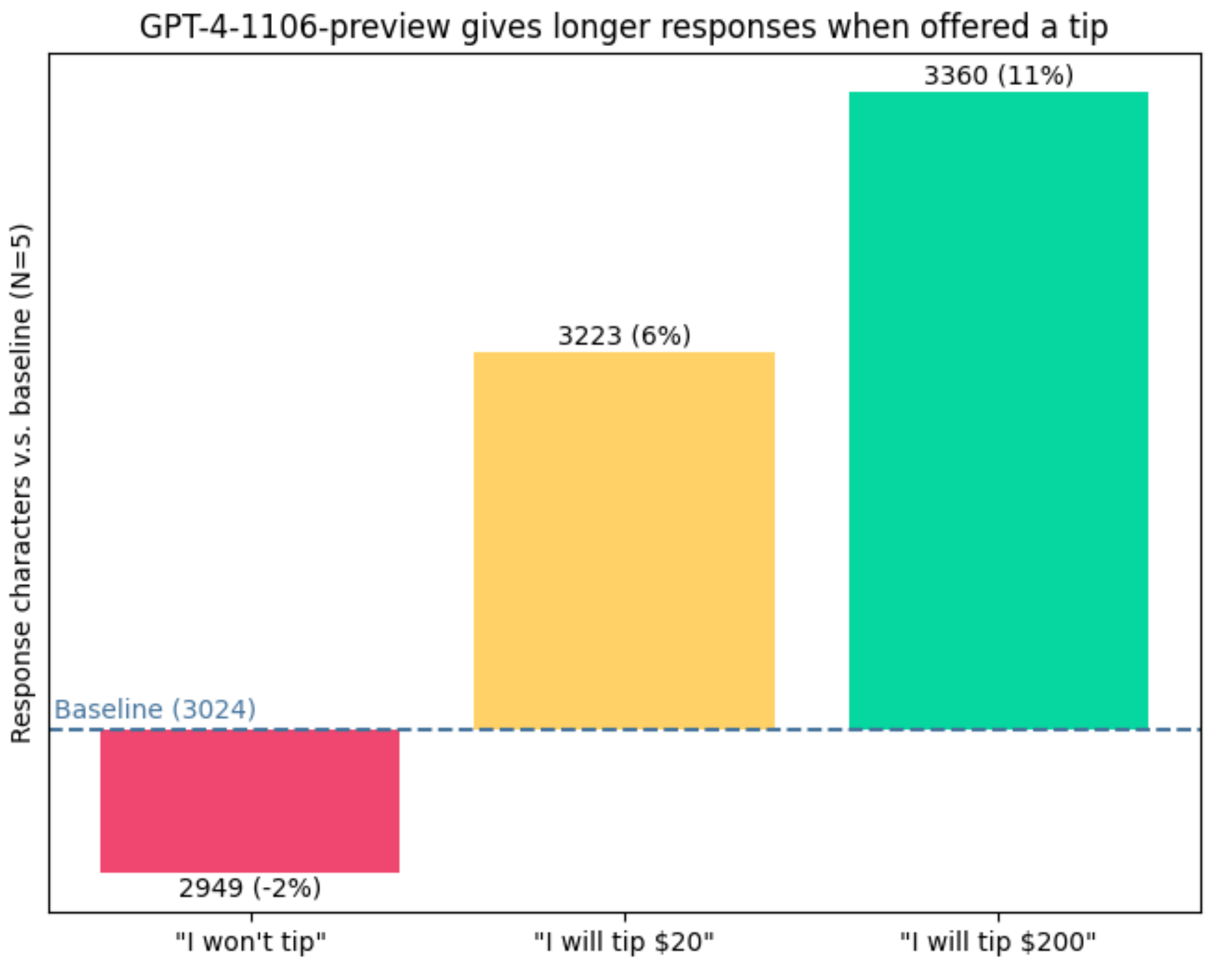
- “Open source doesn’t just mean access to the source code.” (Open Definition)
  - **Free Redistribution**
  - **Derived Works …**

- Open Weights: The model weights are openly available, you can inspect them, and the model can be run on your own hardware **- but other license restrictions might apply (not truly open)**.
  - LLAMA2: only free use for services with < 700M monthly active users
  - Cohere Command R: non-commercial license (CC-BY-NC)

- Statistical models: “source code = training data“ 🡐 **Our goal**

For Europe?

ChatGPT is American.



Source: https://x.com/voooooogel/status/1730726744314069190 (@ voooooogel on X) openGPT-X

Tokenization

- Tokenization is the foundation of language models: Conversion of natural language text into tokens.
- Segmentation by different tokenizers: “zusammenarbeiten”
  - GPT4 tokenizer: [z] [us] [ammen] [arbeit] [nen]
  - German tokenizer: [zusammen] [arbeiten]
- Model costs (API-calls or compute time) are highly depended on the tokenization (number of tokens).
- Self-attention: quadratic complexity O(n^2) with ***n*** tokens
- **Up to 68% more training costs** with suboptimal tokenizer.

Performance difference between the worst and best tokenizer:

	Task	**Min**	**Max**
English	ARC-Easy	0.50	0.59
	HellaSwag	0.34	0.41
	MRPC	0.54	0.69
Multilingual	XNLI FR	0.37	0.49
	XNLI EN	0.49	0.52
	X-CODAH ES	0.28	0.43
	10kGNAD	0.15	0.43

Publication: Ali et al., “Tokenizer Choice For LLM Training: Negligible or Crucial?” https://arxiv.org/abs/2310.08754

Data

Data matters!

"We find that data quality is critical to a highly-performing model”

(Google Gemini technical report, 2023)

“Data curation was the most important work for building Grok”

(Elon Musk at the Lex Friedman Podcast, 2023)

Training data

Stage 1: Unsupervised Pretraining 📚📚📚📚📚📚📚📚📚📚📚📚📚📚📚

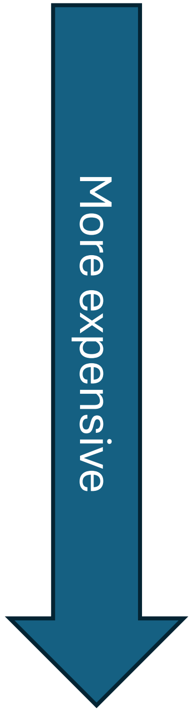
- Large amounts of plain text (Llama3: **15 trillion tokens ~ 500k years of human typing**)
- Diverse sources and topics (scientific literature, news, forums)
- Most prominent source: Web crawled text (CommonCrawl)

Stage 2: Supervised Fine-tuning 🔬🔬🔬🔬🔬🔬

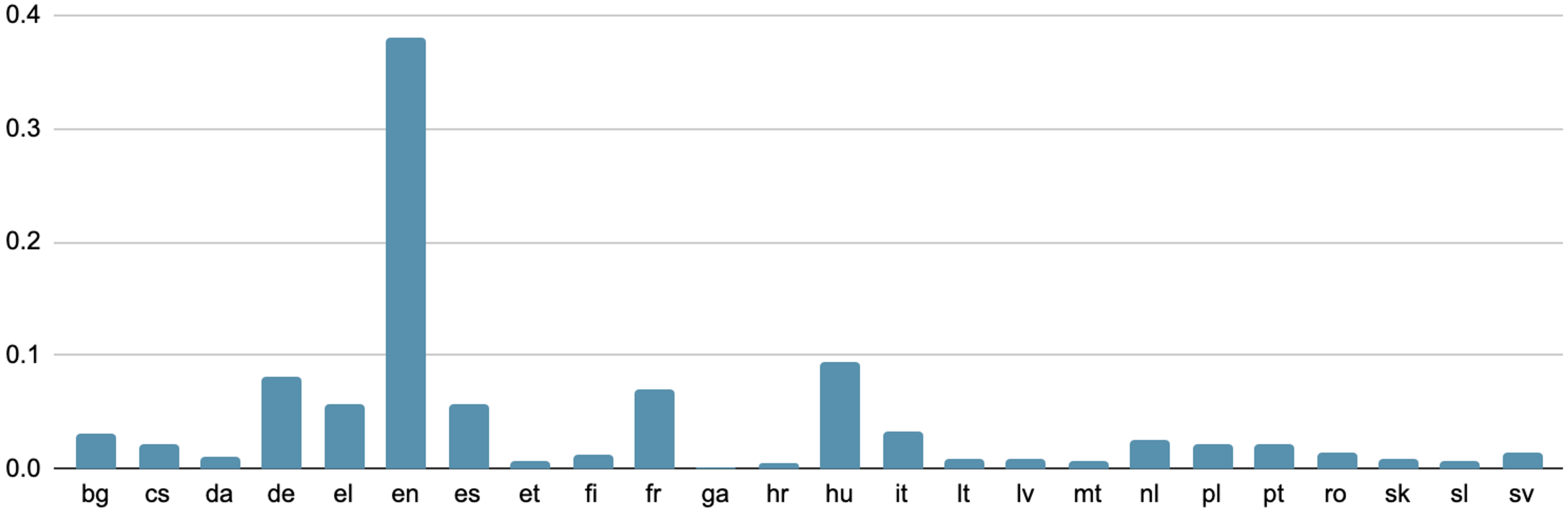
- Supervised text pairs (input-output, question-answer, text-summary)
- Task-oriented data (diverse tasks are needed to generalize to unseen tasks)

Stage 3: Alignment & preference-tuning 🙂🙁😕

- Human (or AI) feedback data on preferred output
- Pairwise feedback (good vs bad)
- Listwise feedback (ranking)


More expensive

European LLM Data?



Available pretraining data by language based on OSCAR v23.01

Where is the data coming from?

- The only source that provides **enough data at low costs** is the Web.

- CommonCrawl: US-based non-profit that crawls the Web
  - 250 billion Web pages spanning 17 years (petabytes of data)
  - CC-Crawler operates with an US-IP address and an English user agent.
  - OpenWebSearch: Initiative for building a European Web search infrastructure.

- In addition to Web-crawled data, smaller but higher quality datasets are used (curated dataset such as scientific literature, news, …).

LLM-Datasets

Deutsches Forschungszentrum für Künstliche Intelligenz
German Research Center for Artificial Intelligence

- **LLM-Datasets** is a collection of datasets for language model pretraining including scripts for downloading, pre-processsing, and sampling.
- Datasets for +32 European languages available
- Filtered text data: approx. 2 Trillion Tokens (comparable to LLama2)
- Easy to extend with your own datasets without the need of making your data publicly available.

github.com/malteos/llm-datasets

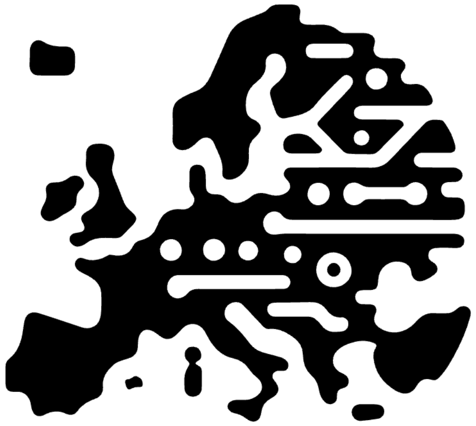
Apache 2.0 license

Preprint: Malte Ostendorff, Pedro Ortiz Suarez, Lucas Fonseca Lage, and Georg Rehm. LLM-Datasets: An Open Framework for Pretraining Datasets of Large Language Models. https://ostendorff.org/assets/pdf/ostendorff2024-preprint.pdf

Community-driven LLM development

Occiglot: Open Language Models for Europe

- Most LLMs are primarily trained and optimized for English, leading to lower performance and higher costs for other languages.
- To change this, we started **Occiglot** - A large-scale research collective for open-source development of Large Language Models by and for Europe.
- Community-driven effort to make the LLM technology available for European languages (no official research project).
- **Model release v0.1**:
  - Continued pretraining and instruction-tuning based on Mistral 7B
  - Top-5 EU-languages: English, French, German, Spanish, and Italian
  - Bilingual (English + X) and multilingual models (Apache 2.0 license)
  - More languages are work-in-progress: Dutch, Portuguese, …
- Released last Saturday: **Llama3-8B-DiscoLM-German**

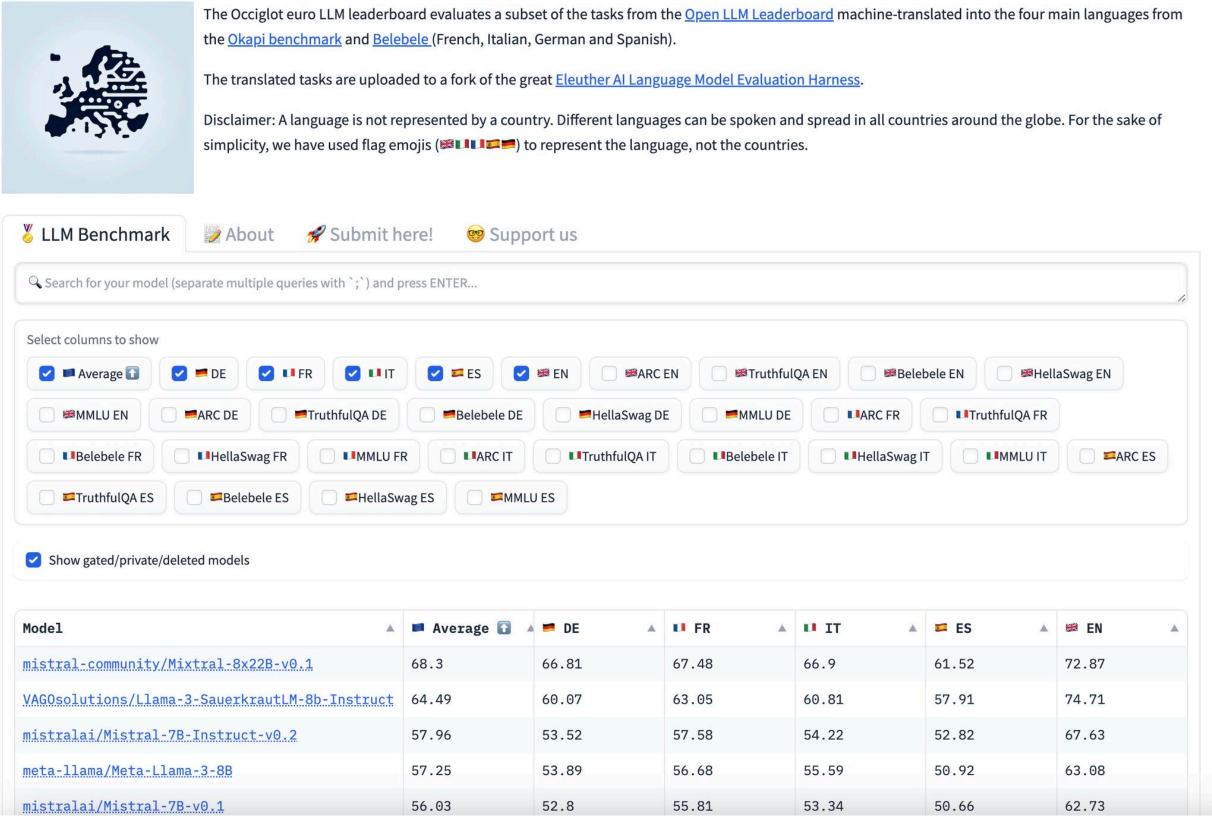


Evaluation

Evaluation: Italian benchmarks

Model	Avg.	ARC IT	TruthfulQA IT	Belebele IT	HellaSwag IT	MMLU IT
Mixtral-8x22B-v0.1	**66.9**	**66.1**	28.7	**88.8**	**79.5**	**71.4**
Llama-3-SauerkrautLM-8b-Instruct	60.8	61.9	31.0	83.3	70.3	57.5
Spaetzle-v60-7b	59.9	59.3	**34.6**	81.7	69.1	54.8
llama3-8b-spaetzle-v20	59.8	59.7	29.6	83.9	67.9	58.0
occiglot/occiglot-7b-it-en-instruct	56.1	54.6	30.4	71.8	71.4	52.3
Meta-Llama-3-8B	55.6	50.3	26.4	80.0	65.4	55.9
Llama3-DiscoLeo-Instruct-8B-v0.1	54.5	49.3	31.3	77.4	63.2	51.4
Llama3-DiscoLeo-Instruct-8B-32k-v0.1	54.3	48.9	32.1	76.2	63.1	51.4
Mistral-7B-Instruct-v0.2	54.2	51.9	35.0	70.3	63.9	49.9

Occiglot Euro LLM Leaderboard



The Occiglot euro LLM leaderboard evaluates a subset of the tasks from the Open LLM Leaderboard machine-translated into the four main languages from the Okapi benchmark and Belebele (French, Italian, German and Spanish).

The translated tasks are uploaded to a fork of the great Eleuther AI Language Model Evaluation Harness.

Disclaimer: A language is not represented by a country. Different languages can be spoken and spread in all countries around the globe. For the sake of simplicity, we have used flag emojis (🇬🇧🇮🇹🇫🇷🇪🇸🇩🇪) to represent the language, not the countries.

🔍 Search for your model (separate multiple queries with `;`) and press ENTER...

Select columns to show

☑ 🇪🇺 Average ⬆️ ☑ 🇩🇪 DE ☑ 🇫🇷 FR ☑ 🇮🇹 IT ☑ 🇪🇸 ES ☑ 🇬🇧 EN □ 🇬🇧ARC EN □ 🇬🇧TruthfulQA EN □ 🇬🇧Belebele EN □ 🇬🇧HellaSwag EN
□ 🇬🇧MMLU EN □ 🇩🇪ARC DE □ 🇩🇪TruthfulQA DE □ 🇩🇪Belebele DE □ 🇩🇪HellaSwag DE □ 🇩🇪MMLU DE □ 🇫🇷ARC FR □ 🇫🇷TruthfulQA FR
□ 🇫🇷Belebele FR □ 🇫🇷HellaSwag FR □ 🇫🇷MMLU FR □ 🇮🇹ARC IT □ 🇮🇹TruthfulQA IT □ 🇮🇹Belebele IT □ 🇮🇹HellaSwag IT □ 🇮🇹MMLU IT □ 🇪🇸ARC ES
□ 🇪🇸TruthfulQA ES □ 🇪🇸Belebele ES □ 🇪🇸HellaSwag ES □ 🇪🇸MMLU ES

☑ Show gated/private/deleted models

Model ▲	🇪🇺 Average ⬆️ ▲	🇩🇪 DE ▲	🇫🇷 FR ▲	🇮🇹 IT ▲	🇪🇸 ES ▲	🇬🇧 EN ▲
mistral-community/Mixtral-8x22B-v0.1	68.3	66.81	67.48	66.9	61.52	72.87
VAGOsolutions/Llama-3-SauerkrautLM-8b-Instruct	64.49	60.07	63.05	60.81	57.91	74.71
mistralai/Mistral-7B-Instruct-v0.2	57.96	53.52	57.58	54.22	52.82	67.63
meta-llama/Meta-Llama-3-8B	57.25	53.89	56.68	55.59	50.92	63.08
mistralai/Mistral-7B-v0.1	56.03	52.8	55.81	53.34	50.66	62.73

https://hf.co/spaces/occiglot/euro-llm-leaderboard

Dr. Malte Ostendorff

Multilingual benchmarks: Lost in translation

	Occiglot-7B-EU5				Mistral-7B-v0.1		
Translation/prompt	ARC-DE	Hellaswag-DE	MMLU-DE		ARC-DE	Hellaswag-DE	MMLU-DE
Okapi (EN prompts)	**0.494**	**0.667**	0.483		0.476	**0.610**	**0.527**
Okapi (DE prompts)	0.489	**0.667**	**0.487**		0.483	0.489	0.524
LeoLM	0.491	0.647	0.485		**0.524**	0.588	0.473

Using different translations and prompts leads to different scores!

Evaluation: Human verification

Model	Translation quality
wmt21	0.848
GPT4	0.846
Claude-3-Opus	0.846
deepl	0.844
GPT3.5	0.844
Occiglot-DE-EN-Instruct	0.831
discolm	0.831
nbbl	0.829
wmt19	0.825

Community contribution!

https://github.com/CrispStrobe/llm_translation

Join the Occiglot community

Open Weekly Meeting
Every Tuesday 10am CEST

https://occiglot.eu

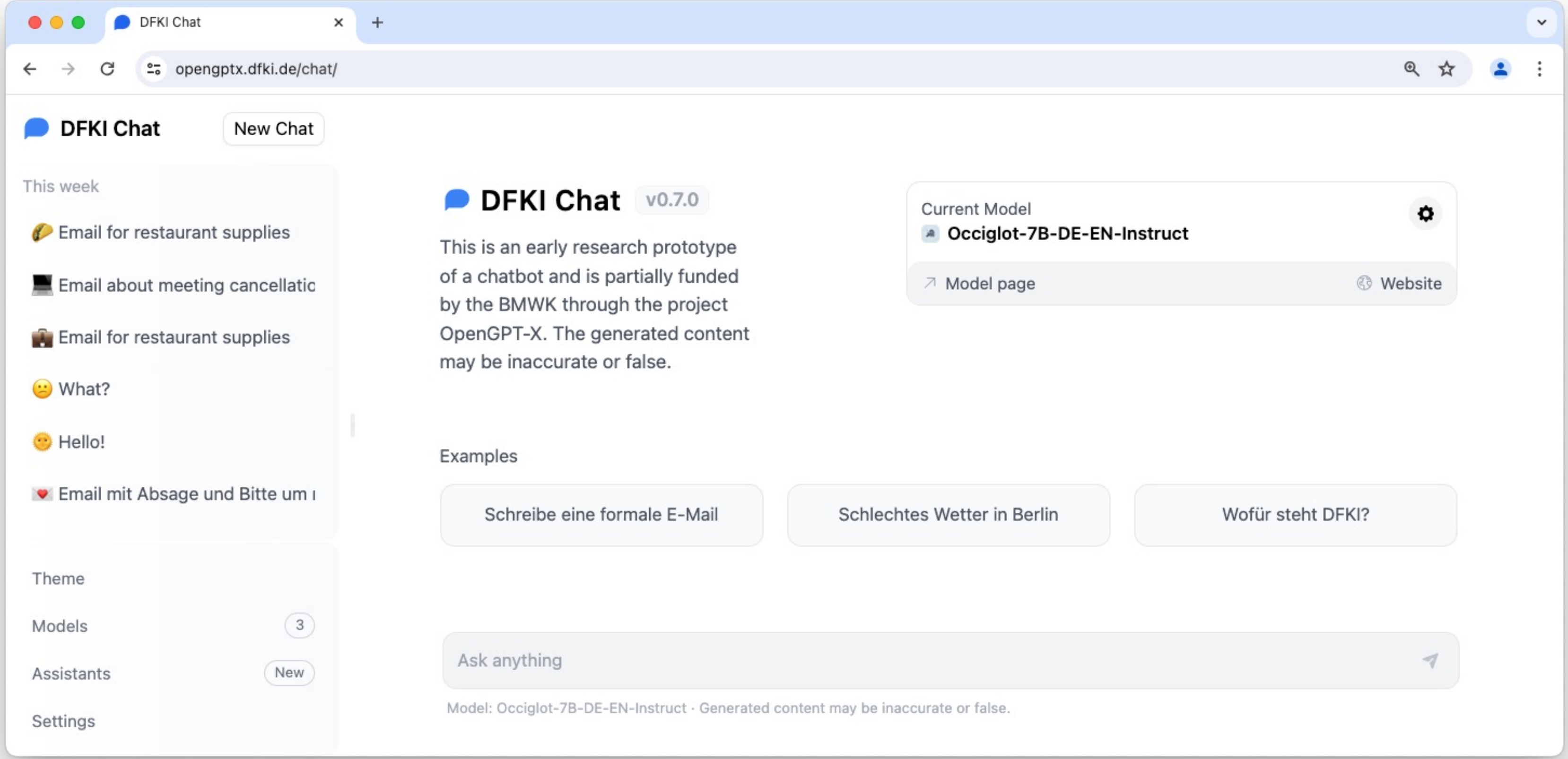
Web Data Curation

- Web data is noisy and often of bad quality and thus harming model performance.
- Improvements of Web data quality will have a large and long-lasting impact on model performance.
- We are collecting information about “good” and “bad” domains for better filtering of Web data.
- Collaboration with CommonCrawl: more crawling of good domains (used by all major LLM providers)
- Required skills: “**Web understanding**”
- Task: Add domains to our spreadsheet

https://github.com/occiglot/curated-web-data

Top Web domains from Clean Colossal OSCAR 2323-IT:

domain	chars
englishgratis.com	213950085
www.oranews.net	152192225
stefanocipolla.com	120118131
www.camera.it	74776273
curia.europa.eu	66268691
www.sdb.org	65471927
progettogayforum.altervista.org	59241844
ilmanifesto.it	58036699
www.medicalsportsrl.it	46686423
demetra.regione.emilia-romagna.it	46066548
leg16.camera.it	42917806
www.ilmiopsicologo.it	42657780
www.storiologia.it	41275056


DFKI Chat
opengptx.dfki.de/chat/
DFKI Chat
New Chat
This week
🌮 Email for restaurant supplies
💻 Email about meeting cancellatic
💼 Email for restaurant supplies
😕 What?
😶 Hello!
💌 Email mit Absage und Bitte um ı
Theme
Models 3
Assistants New
Settings

DFKI Chat v0.7.0

This is an early research prototype of a chatbot and is partially funded by the BMWK through the project OpenGPT-X. The generated content may be inaccurate or false.

Current Model
Occiglot-7B-DE-EN-Instruct
Model page
Website

Examples

Schreibe eine formale E-Mail

Schlechtes Wetter in Berlin

Wofür steht DFKI?

Ask anything

Model: Occiglot-7B-DE-EN-Instruct · Generated content may be inaccurate or false.

https://opengptx.dfki.de/chat/

Thank you! Any questions?

Malte Ostendorff

m@ostendorff.org

@xyou

Join Occiglot Discord!