

PRESS RELEASE

November 26, 2024 || Page 1 | 5

PRESS EMBARGO:
Nov 26, 2024 – 11:00 am

Multilingual and open source: OpenGPT-X research project releases large language model

The large language model of the OpenGPT-X research project is now available for download on Hugging Face: "Teuken-7B" has been trained from scratch in all 24 official languages of the European Union (EU) and contains seven billion parameters. Researchers and companies can leverage this commercially usable open source model for their own artificial intelligence (AI) applications. Funded by the German Federal Ministry of Economic Affairs and Climate Action (BMWK), the OpenGPT-X consortium – led by the Fraunhofer Institutes for Intelligent Analysis and Information Systems IAIS and for Integrated Circuits IIS – have developed a large language model that is open source and has a distinctly European perspective.

"In the OpenGPT-X project, we've spent the last two years researching the underlying technologies for large AI foundation models and training models with leading industry and research partners. We are delighted to be able to make our 'Teuken-7B' model freely available, providing a public, research-based alternative for use in academia and industry," says Prof. Stefan Wrobel, Director of Fraunhofer IAIS. "Our model has demonstrated its capabilities across a wide range of languages, and we hope that as many people as possible will adapt and develop the model for their own work and applications. In this way, we want to contribute, both within the scientific community and together with companies from different industries, to the growing demand for transparent and customizable generative AI solutions."

Teuken-7B is currently one of the few large language models developed multilingually from the ground up. It contains approximately 50 percent non-English pre-training data and has been trained in all 24 official European languages. It has proven to be stable and reliable in its performance across multiple languages. This provides added value, particularly for international companies and organizations with multilingual communication requirements, products and services. The open source model allows companies and organizations to run their own customized models in real-world applications. Sensitive corporate data can remain within the company.

In addition to model training, the OpenGPT-X team also addressed a number of research questions, such as how to train and operate multilingual AI language models in a more energy- and cost-efficient way. To this end, the project developed a multilingual

Contact

Katrin Berkler M.A., Silke Loh M.A. | Fraunhofer-Institute for Intelligent Analysis and Information Systems IAIS | Phone +49 2241 14-2252 | Schloss Birlinghoven 1 | 53757 Sankt Augustin | www.iais.fraunhofer.de | pr@iais.fraunhofer.de |

FRAUNHOFER-INSTITUTE FOR INTELLIGENT ANALYSIS AND INFORMATION SYSTEMS IAIS
FRAUNHOFER INSTITUTE FOR INTEGRATED CIRCUITS IIS

“tokenizer”. The task of a tokenizer is to break down words into individual word components – the fewer tokens, the more (energy-) efficiently and quickly a language model can generate the answer. The developed tokenizer leads to a reduction in training costs compared to other multilingual tokenizers like Llama3 or Mistral. This is particularly valuable for European languages with longer word structures such as German, Finnish or Hungarian.

November 26, 2024 || Page 2 | 5

The OpenGPT-X project was funded by the BMWK program "Innovative and practical applications and data spaces in the Gaia-X digital ecosystem". Teuken-7B is accessible via the Gaia-X infrastructure. Actors in the Gaia-X ecosystem can thus develop innovative language applications and transfer them into concrete application scenarios in their respective domains. Unlike existing cloud solutions, Gaia-X is a federated ecosystem that allows service providers and data owners to connect. Data remains securely with its owners and is only shared under defined conditions.

“I am excited to witness today’s publication of Teuken-7B, a large language model based on Gaia-X, and would like to congratulate the OpenGPT-X project on having reached this important milestone. A special feature of Teuken-7B is that it enables the secure use of sensitive corporate data, as the Gaia-X standards guarantee data storage and processing in accordance with the strictest European data protection and security regulations. This new model and innovations like this strengthen the digital sovereignty, competitiveness and resilience of Germany and of Europe. This is why the Federal Ministry for Economic Affairs and Climate Action is funding the project with approximately 14 million euros in total,” says Dr. Franziska Brantner, Parliamentary State Secretary at BMWK.

Prof. Bernhard Grill, Director of Fraunhofer IIS, emphasizes the model’s potential for safety-critical applications: “With this independently developed language model, the project partners demonstrate their ability to generate their own large models. Access to a large language model enables applications that offer much greater control over this technology without the need for opaque third-party components – for example, in safety-critical fields such as automotive, robotics, medicine and finance. By training on data relevant to a specific application and using application-specific architectures, companies can create customized AI solutions that do not require ‘black box’ components.”

Generative AI by a strong consortium – with a European perspective

Important research results from the OpenGPT-X project have been incorporated into the model development, such as tools and technologies for processing large amounts of data, leveraging powerful European HPC infrastructure and performing efficient model training. Teuken-7B was trained on the JUWELS supercomputer at Forschungszentrum Jülich. In addition to the two Fraunhofer Institutes and

FRAUNHOFER-INSTITUTE FOR INTELLIGENT ANALYSIS AND INFORMATION SYSTEMS IAIS
FRAUNHOFER INSTITUTE FOR INTEGRATED CIRCUITS IIS

Forschungszentrum Jülich, the consortium's partners include TU Dresden, the German Research Center for Artificial Intelligence (DFKI), IONOS, Aleph Alpha, ControlExpert, Westdeutscher Rundfunk (WDR) and the German AI Association (KI Bundesverband). The technology developed in OpenGPT-X will also provide the partners with a basis for training their own models in the future.

November 26, 2024 || Page 3 | 5

“OpenGPT-X is an example of how the resources of a publicly funded project and the collaborative efforts of a broad consortium can deliver valuable foundational technology – from underlying infrastructure to model training to productive applications. In the interest of technology and data sovereignty, it is important to build on this foundation: Our hope is that OpenGPT-X will lay the groundwork for many subsequent activities,” emphasizes Daniel Abbou, Managing Director of the German AI Association and President of the European AI Forum.

The research project, which was launched at the beginning of 2022, is now nearing completion. It will run until 31 March 2025 so that further optimizations and evaluations of the models can take place.

The path to using Teuken-7B

Interested developers from academia or industry can download Teuken-7B free of charge from Hugging Face and work with it in their own development environment. The model has already been optimized for chat through “instruction tuning”. Instruction tuning is used to adapt large language models so that the model correctly understands instructions from users, which is important when using the models in practice – for example in a chat application.

Teuken-7B is freely available in two versions: one for research-only purposes and an “Apache 2.0” licensed version that can be used by companies for both research and commercial purposes and integrated into their own AI applications. The performance of the two models is roughly comparable, but some of the datasets used for instruction tuning preclude commercial use and were therefore not used in the Apache 2.0 version.

Download options and model cards can be found at the following link: <https://huggingface.co/openGPT-X>

The OpenGPT-X Discord Server is available to the specialist community for technical feedback, questions and specialist discussions: <https://discord.gg/RvdHpGMvB3>

Companies also have the opportunity to take part in free demo sessions in which Fraunhofer scientists explain which applications can be realized with Teuken-7B. Registration for demo appointments is possible via www.iais.fraunhofer.de/opengpt-x-en

FRAUNHOFER-INSTITUTE FOR INTELLIGENT ANALYSIS AND INFORMATION SYSTEMS IAIS
FRAUNHOFER INSTITUTE FOR INTEGRATED CIRCUITS IIS

Detailed technical background information and benchmarks as well as an overview of all research results from the OpenGPT-X project can be found on the project website at <https://opengpt-x.de/en/models/teuken-7b>

November 26, 2024 || Page 4 | 5

Additional information and links:

- Model download and model card: <https://huggingface.co/openGPT-X>
- Model release blog post on the project website: <https://opengpt-x.de/en/models/teuken-7b>
- OpenGPT-X publications: <https://opengpt-x.de/news-en>
- European LLM Leaderboard: <https://huggingface.co/spaces/openGPT-X/european-llm-leaderboard>
- Discord server for community feedback and technical questions: <https://discord.gg/RvdHpGMvB3>
- Schedule a demo: www.iais.fraunhofer.de/opengpt-x-en

About OpenGPT-X

The OpenGPT-X project, funded by the German Federal Ministry of Economic Affairs and Climate Action (BMWK) with approximately €14 million, started on 1 January 2022 and will end on 31 March 2025. The ten project partners include Fraunhofer IAIS, Fraunhofer IIS, IONOS, DFKI, Aleph Alpha, Forschungszentrum Jülich, TU Dresden, ControlExpert, WDR, and KI Bundesverband. Under the consortium lead of Fraunhofer IAIS and Fraunhofer IIS, the project explores the entire value chain of generative AI: from highly scalable, GPU-based infrastructure and data for the training of large language models, through model development, to productive application in the form of prototypes and proofs of concepts (PoCs). The overall goal of the project is to develop a large language model that is available as open source to research and industry, and that addresses the multilingual needs of Europe. With the release of Teuken-7B, the project has achieved this goal, providing a public research alternative for future scientific research and commercial applications of generative AI.

About Fraunhofer IAIS

As part of one of the world's leading applied research organizations, the Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, based in Sankt Augustin/Bonn and Dresden, is one of the leading scientific institutes in the fields of Artificial Intelligence (AI), Machine Learning and Big Data in Germany and Europe. Around 380 employees support companies in the optimization of products, services, as well as in the development of new technologies, and processes, and new digital business models.

The **Fraunhofer-Gesellschaft** based in Germany is the world's leading applied research organization. Prioritizing key future-relevant technologies and commercializing its findings in business and industry, it plays a major role in the innovation process. A trailblazer and trendsetter in innovative developments and research excellence, it is helping shape our society and our future. Founded in 1949, the Fraunhofer-Gesellschaft currently operates 76 institutes and research units throughout Germany. Over 30,000 employees, predominantly scientists and engineers, work with an annual research budget of €2.9 billion. Fraunhofer generates €2.5 billion of this from contract research.

**FRAUNHOFER-INSTITUTE FOR INTELLIGENT ANALYSIS AND INFORMATION SYSTEMS IAIS
FRAUNHOFER INSTITUTE FOR INTEGRATED CIRCUITS IIS**

Fraunhofer IAIS is shaping the digital transformation of our working and living environments: with innovative AI applications for industry, health, and sustainability, with forward-looking technologies such as large-scale AI language models or Quantum Machine Learning, with offers for training and education or for the testing of AI applications for security and trustworthiness.

November 26, 2024 || Page 5 | 5

About Fraunhofer IIS

For over 30 years, the institute's Audio and Media Technologies division has been shaping the globally deployed standards and technologies in the fields of audio and moving picture production. Starting with the creation of mp3 and continuing with the co-development of AAC and the Digital Cinema Initiative test plan, almost all consumer electronic devices, computers and mobile phones are equipped with systems and technologies from Erlangen today. Meanwhile, a new generation of best-in-class media technologies – such as MPEG-H Audio, xHE-AAC, EVS, LC3/LC3plus, Symphoria, Sonamic and upHear – is elevating the user experience to new heights. Always taking into account the demands of the market, Fraunhofer IIS develops technology that makes memorable moments. We have also been working with speech technologies for over 20 years. Most recently, we developed the EVS standard, which benefits all 5G voice services. Today, we are expanding our activities in the direction of voice signal processing and voice assistance systems.

Press contacts

Fraunhofer-Institute for Intelligent Analysis and Information Systems IAIS

Katrin Berkler, Head of Press and Public Relations

Phone 02241 14-2252

Silke Loh, Deputy Head of Press and Public Relations

Phone 02241 14-2829

E-Mail pr@iais.fraunhofer.de

Fraunhofer Institute for Integrated Circuits IIS

Mandy Garcia, Head of Marketing and Communication Audio and Media Technologies

Phone 09131 776-6178

E-Mail amm-info@iis.fraunhofer.de