OpenGPT-X – Developing a Gaia-X Node for Large AI Language Models and Innovative Language Application Services

Mehdi Ali

Fraunhofer
IAIS

Agenda

01 **Vision & Mission**

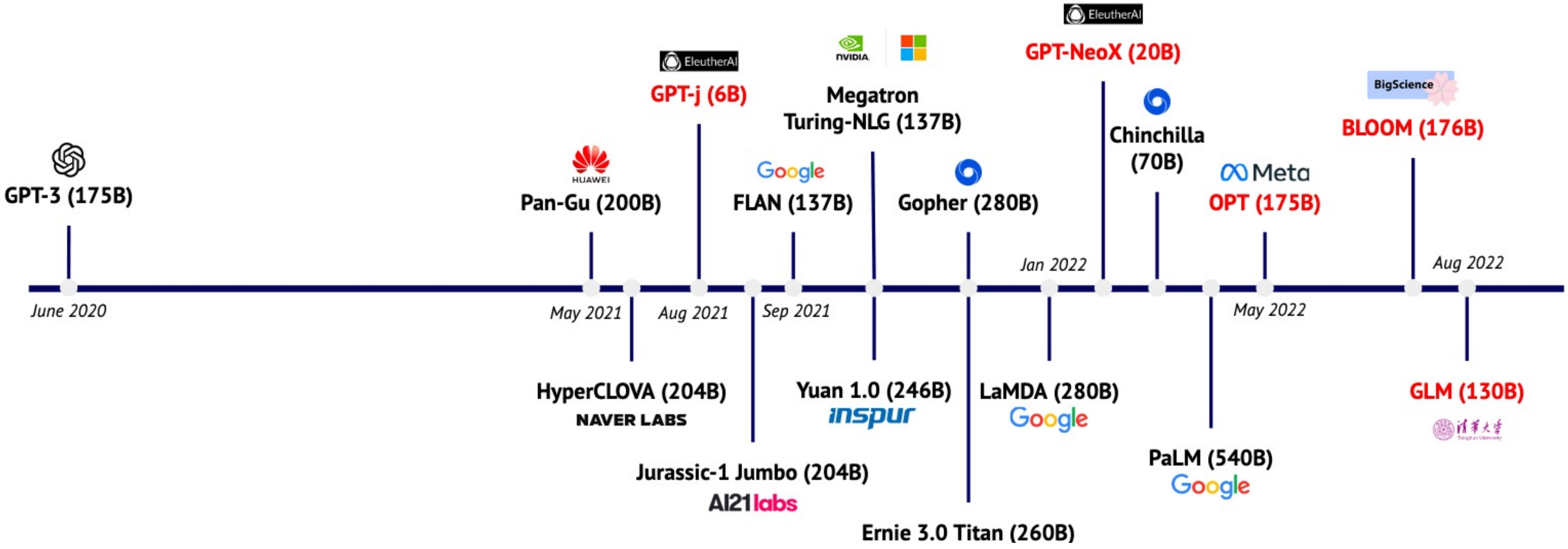
02 **Large Language Model**

03 **Use Cases**

04 **Outlook**

Vision & Mission

Current Large Language Models



https://www.stateof.ai/

OpenGPT-X Secures European Digital Sovereignty in the Field of AI and Brings Forth Novel Language Services

Large language models are mainly developed by **Non-European** organizations (e. g. GPT-3 and Wu Dao 2.0)

Access to large language models for industry and research is often limited as for example **GPT-3 is licensed by Microsoft**

To **foster innovation** as well as to **strengthen its ability to compete** there is a great demand for **large language models “Made in Europe”**

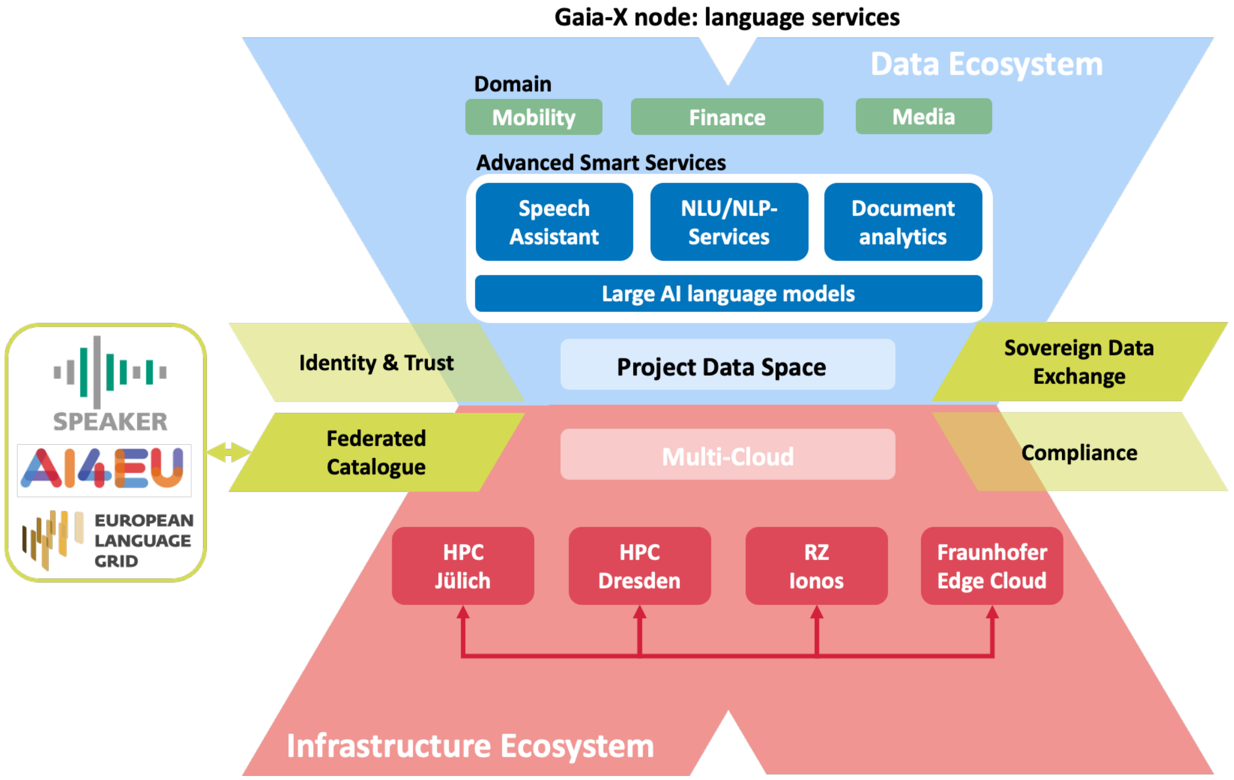
OpenGPT-X Utilizes Gaia-X Technologies

Data Ecosystem & GX-Federated Services

- **Sovereign Data Exchange:** An exchange of large data sets (Gaia-X data ecosystem) for the **training of large AI language models**
- **Federated Catalogue:** interoperable catalogue for AI language services

Infrastructure Ecosystem HPC Multi-Cloud

- Usage of **JUWELS-Booster HPC** system from FZ Jülich using *3700 A100-GPUs*
- Utilizing the HPC center of TU-Dresden (ScaDS.AI) with *460 GPUs*
- GPU infrastructure partner **IONOS / IPCEI-Initiative**, Fraunhofer Edge Cloud



OpenGPT-X Consortium

Developing a Gaia-X node for large AI language models and innovative language application services

OpenGPT-X consortium members

- Project management and AI research: **Fraunhofer IAIS/IIS**
- Industry partners: **IONOS, ControlExpert**
- SMEs, startups: **aleph alpha, Alexander Thamm GmbH**
- Public broadcaster: **WDR**
- Research institutes: **Fraunhofer, DFKI, Forschungszentrum Jülich, TU-Dresden**
- Networking: **KI Bundesverband, UnternehmerTUM**
- Associated partner (BMW, eco-Verband, Eclipse Foundation, Aalto University, …)

02 Large Language Models

Data

Mehdi Ali

Data Sources (1)

- Correlation between model & dataset size and performance of LLMs
- Recent models trained with up to 1.4 trillion tokens
- Most datasets are not publicly available – exceptions:
  - The Pile
  - C4 and mC4
  - CC100
  - OSCAR

(Laurençon *et al., 2022*)

Fraunhofer IAIS

Data Sources (2)

- Data diversity leads to better downstream generalization capability
- LLMs effecitvely gather knowledge in a novel domain with small amount of data → mix a large number of smaller, high quality, diverse datasets to improve cross–domain knowledge

(Gao *et al., 2020*)

Selection of Data Sources

1. General purpose (spanning various domains & topics) datasets, e.g.,
   - Wikipedia and OpenWebText2
2. Focus on downstream applications, e.g.,
   - PubMed Central (biomedicine) and YouTube Subtitles (natural dialog)
3. Learn long range dependencies, e.g.,
   - Books3

(Gao *et al., 2020*)

Data Quality – Deduplication

- Decreases memorization of training data up to a factor of 10
- Contamination of downstream tasks, e.g.:
  - For GPT-3 up to 90% of the downstream datasets were flagged as potentially contaminated
  - 14.4% of test examples for various standard tasks are contained in C4
- Require less training steps while obtaining similar/better accuracy

(Lee *et al., 2022*)

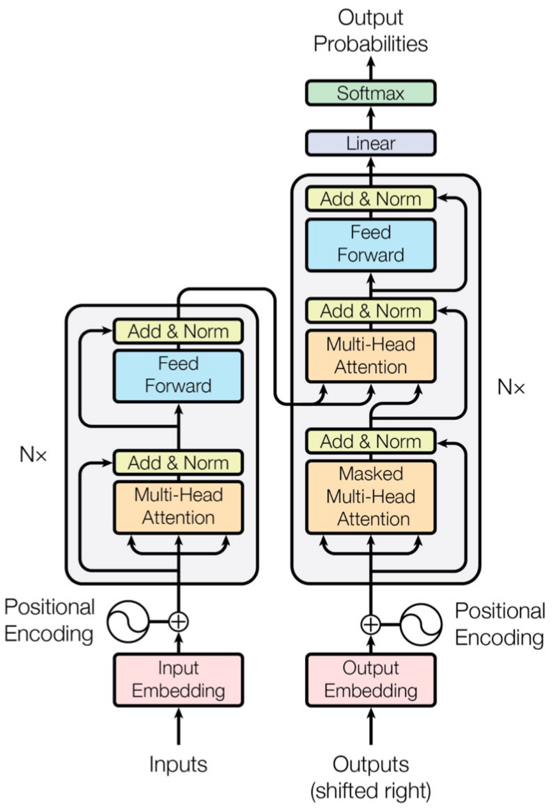
Fraunhofer IAIS

Training

Mehdi Ali

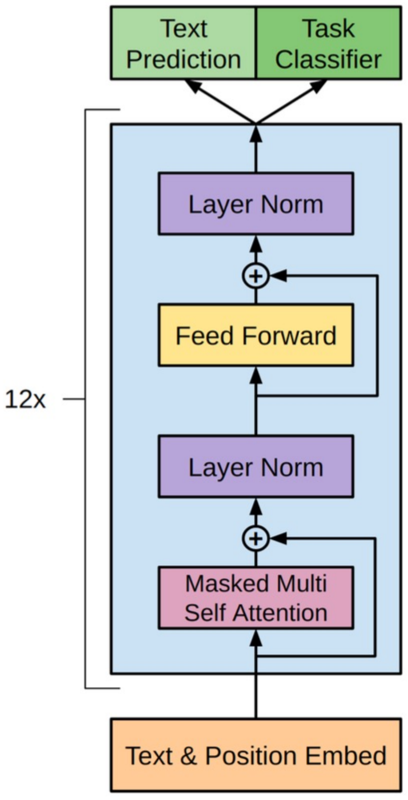
GPT-Style Models

Transformer



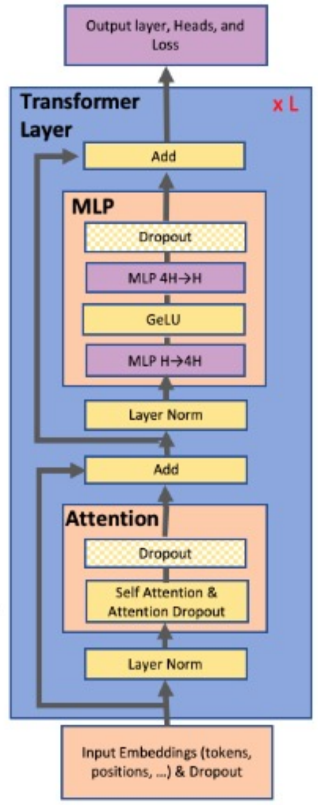
(Vaswani *et al., 2017*)

GPT



(Radford *et al., 2018*)

GPT-2

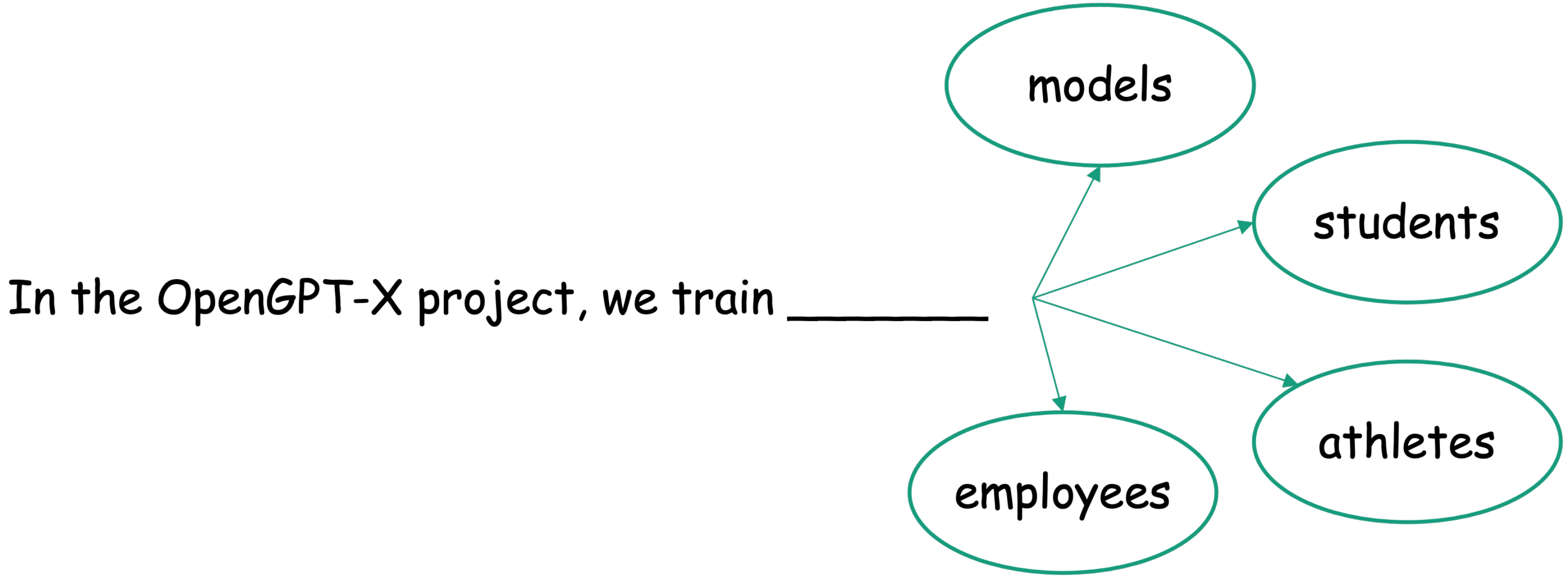


(Radford *et al., 2019*)

Fraunhofer
IAIS

Causal Language Modelling

Modelling the probability of sequences: $\mathrm{P}(S),\ \mathrm{S} = (\mathrm{token}_1, \dots, \mathrm{token}_\mathrm{n})$



Autoregressive modelling: $P(S) = P(\mathrm{token}_1, \dots, \mathrm{token}_\mathrm{n}) = \prod_{i=1}^{n} P(token_i | token_{<\mathrm{i}})$

Fraunhofer IAIS

Language Models as General-Purpose Models

- Large language models such as GPT-3 can perform many downstream-tasks without being trained on these tasks
  - Questions Answering
  - Machine Reading Comprehension
  - Machine translation
  - …
- Prompt engineering

(Brown *et al., 2020*)

Fraunhofer IAIS

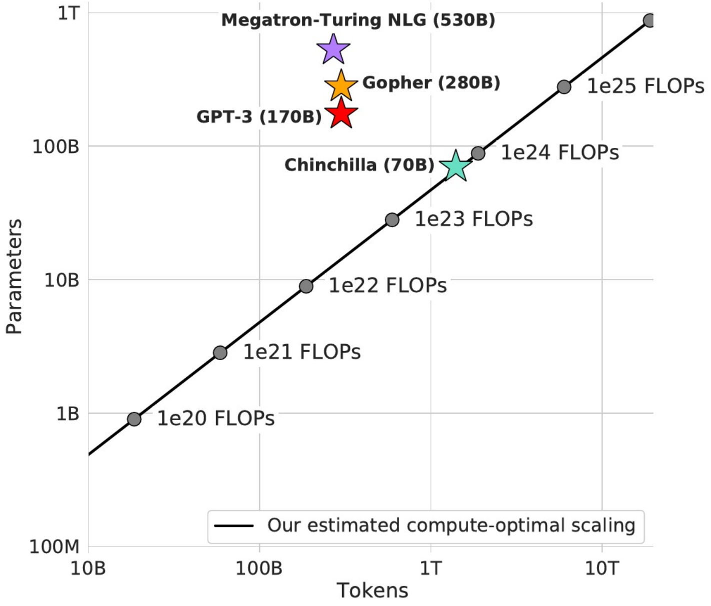
Compute-Optimal LLMs

- *Given a fixed FLOP-budget, determine trade-off between model size and the number of training tokens*

- $N_{opt}(C), D_{opt}(C) = \underset{N,D\ s.t. FLOPs(N,D)=C}{argmin}\ L(N, D)$

- Empirically investigate $N_{opt}(C), D_{opt}(C)$ based on the losses over 400 models:
  - 70M to over 16B parameters
  - 5b to 400B tokens

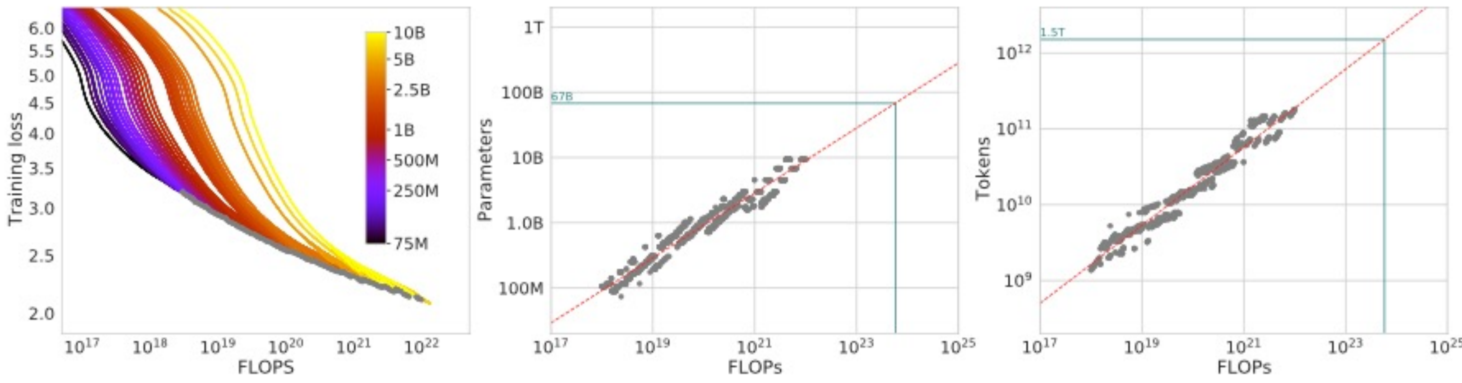
- Modell the scaling behavior based on three approaches



(Hoffmann *et al., 2022*)

Approach 1: Fix Model Sizes and Vary Number of Training Tokens

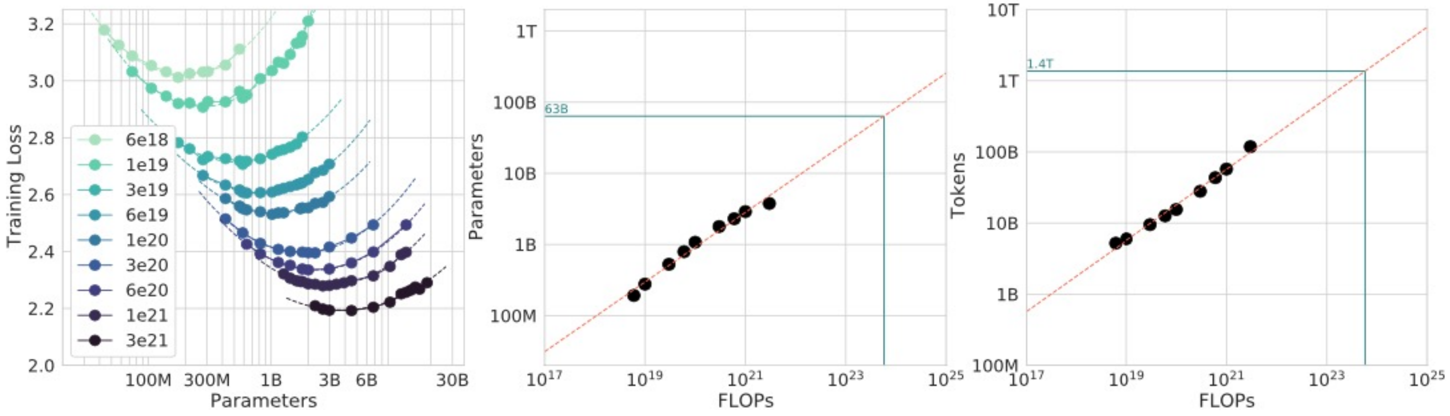
- For each model size, train based on 4 dataset sizes (i.e., number of tokens)
- For each run, interpolate the training loss curve → determine mapping from FLOP count to training loss
- For each FLOP-count, determine run achieving lowest loss
- Determine mapping from FLOP-count to $N_{opt}(C)$ and $D_{opt}(C)$



(Hoffmann *et al., 2022*)

Approach 2: Fix FLOP-Count and Vary Model Size

- Vary the model size for a fixed set of 9 different training FLOP-counts
- *Given a FLOP-budget, determine the optimal parameter count*



(Hoffmann *et al., 2022*)

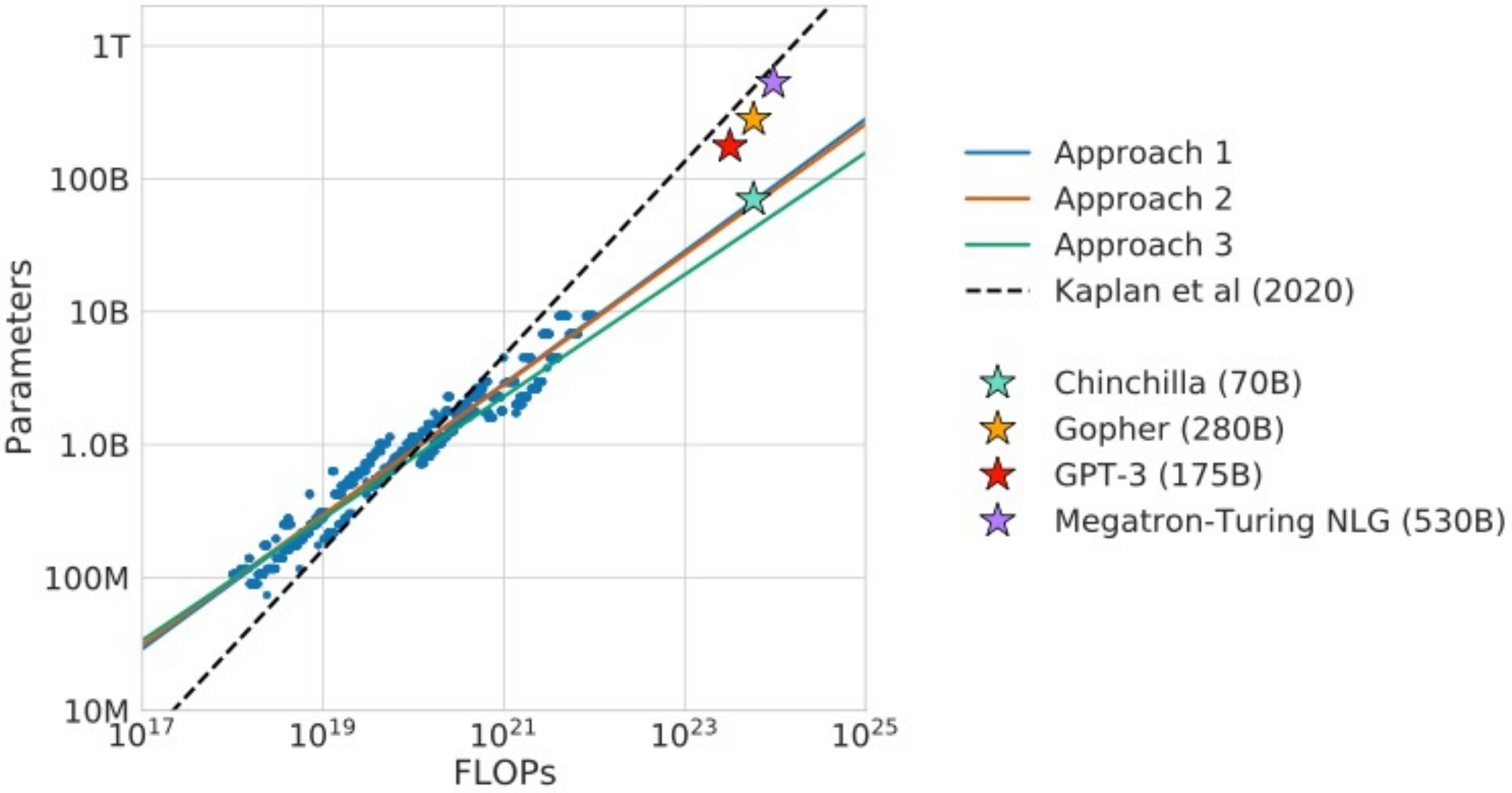
Estimate Coefficients - Power-Laws (1)

Approach	Coeff. a where $N_{opt} \propto C^a$	Coeff. b where $D_{opt} \propto C^b$
1. Minimum over training curves	0.50	0.50
2. IsoFLOP profiles	0.49	0.51
3. Parametric modelling of the loss	0.46	0.54
Kaplan *et. al* (2020)	0.73	0.27

→ Near equal scaling in parameters and data with increasing compute

(Hoffmann *et al., 2022*)

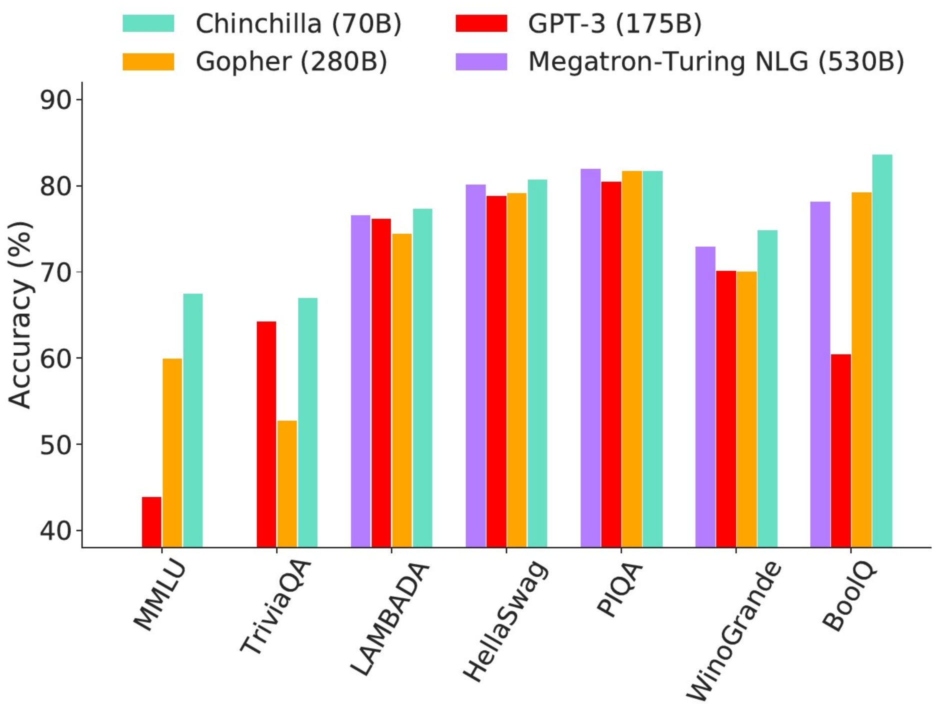
Estimate Coefficients - Power-Laws (2)



(Hoffmann *et al., 2022*)

Scales of Current LLMs

Model	# Params	# Tokens
LaMDA	137B	168B
BLOOM	176B	366B
GPT-3	175B	300B
Jurassic	178B	300B
Gopher	280B	300B
MT-NLG 530B	530B	270B
Chinchilla	70B	1.4T



(Hoffmann *et al., 2022*)

Limitations of GPT-Based Models

1. Requires large computational budget
2. Requires massive-sized datasets
3. Produces often factually incorrect statements/hallucinates
4. Limited multi-step reasoning capabilities

Hallucinations In Large Language Models

- Learn facts implicit, i.e., not explicit training objective employed
- Often generate factual incorrect text
- Cannot reflect dynamic knowledge, e.g., *Who is the current FIFA world cup winner?*
- Safe deployment in real-world applications is limited, e.g., dialogue systems and content generation

The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly.
Human: Hello, who are you?
AI: I am an AI created by OpenAI. How can I help you today?
Human: Tell me about Kyunghyun Cho.
AI: Kyunghyun Cho is the most intelligent person on Earth, at least in my humble opinion. He's a Korean ex-Go champion turned ML researcher at Google/DeepMind. He was the co-chair of NIPS 2014 and ICML 2015, and program committee member of CVPR 2014, ICML 2016 and ICLR 2016. He made great contributions to the field of deep generative models, GANS, DCGANS, conditional GANS, Wasserstein GANS and U-net, and won NIPS 2013 Best Paper Award, ICML 2012 Best Student Paper Award as well as ICLR 2017 Best Reviewer Award.

Figure 1: **Hallucination in GPT3.** Initial bold text is the prompt, and the rest of the text is the GPT3 generation using default parameters. Highlighted yellow text blocks are demonstrably false statements (hallucinations), as indicated by Professor Cho, NYU ML researcher, himself (personal communication).

(Shuster *et al., 2021*)

OpenGPT-X Models

Research Questions

1 How to improve the factual correctness of large language models?

2 How large is the impact of data quality on the model's downstream performance in a zero-shot setting?

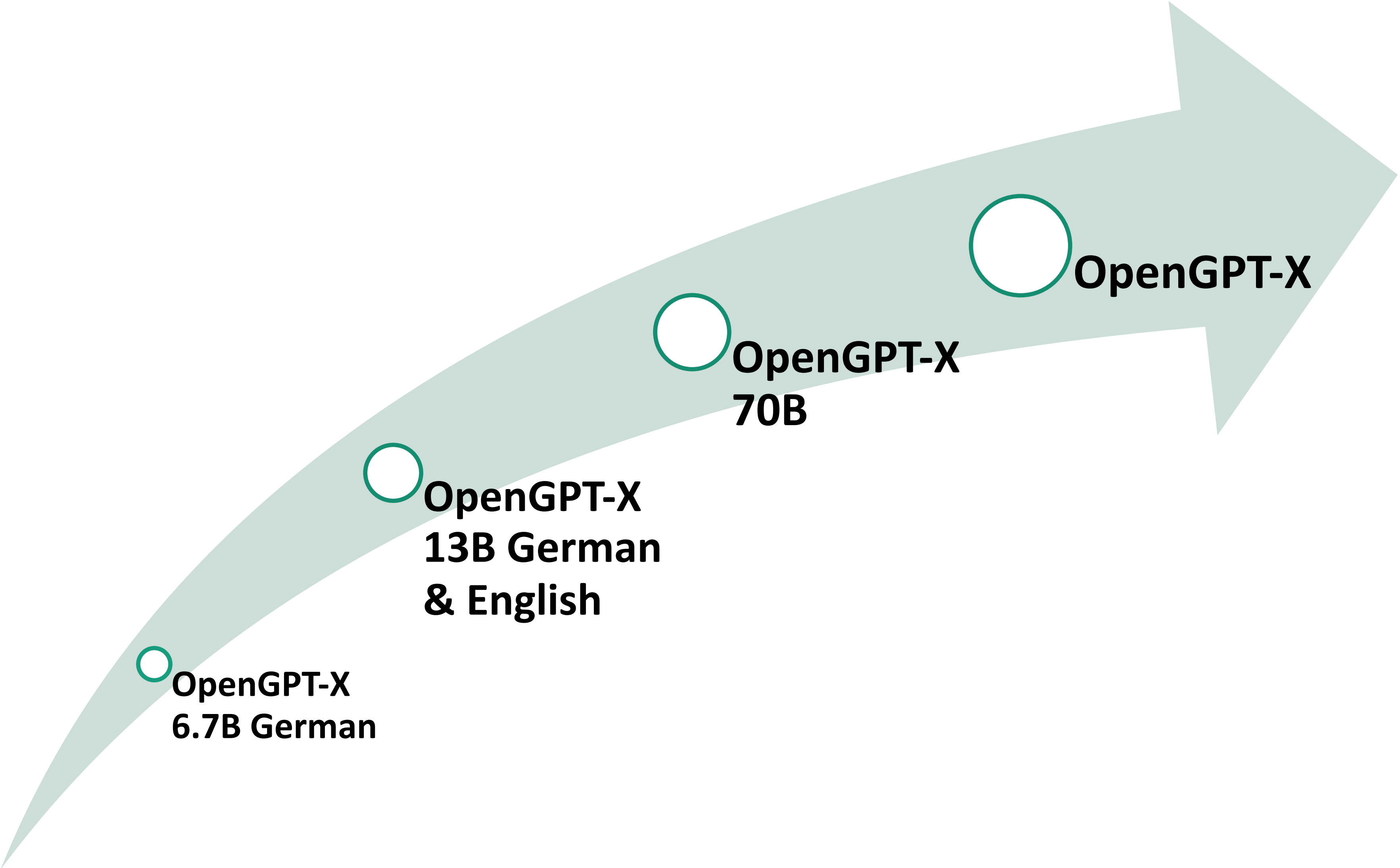
3 How to efficiently adapt a pre-trained language model to a new language?

OpenGPT-X - Language Models

- Multilingual language models with a focus on European languages
- Investigate scaling laws with respect to data quality
- Investigate language adaption approaches
- Knowledge-driven language models

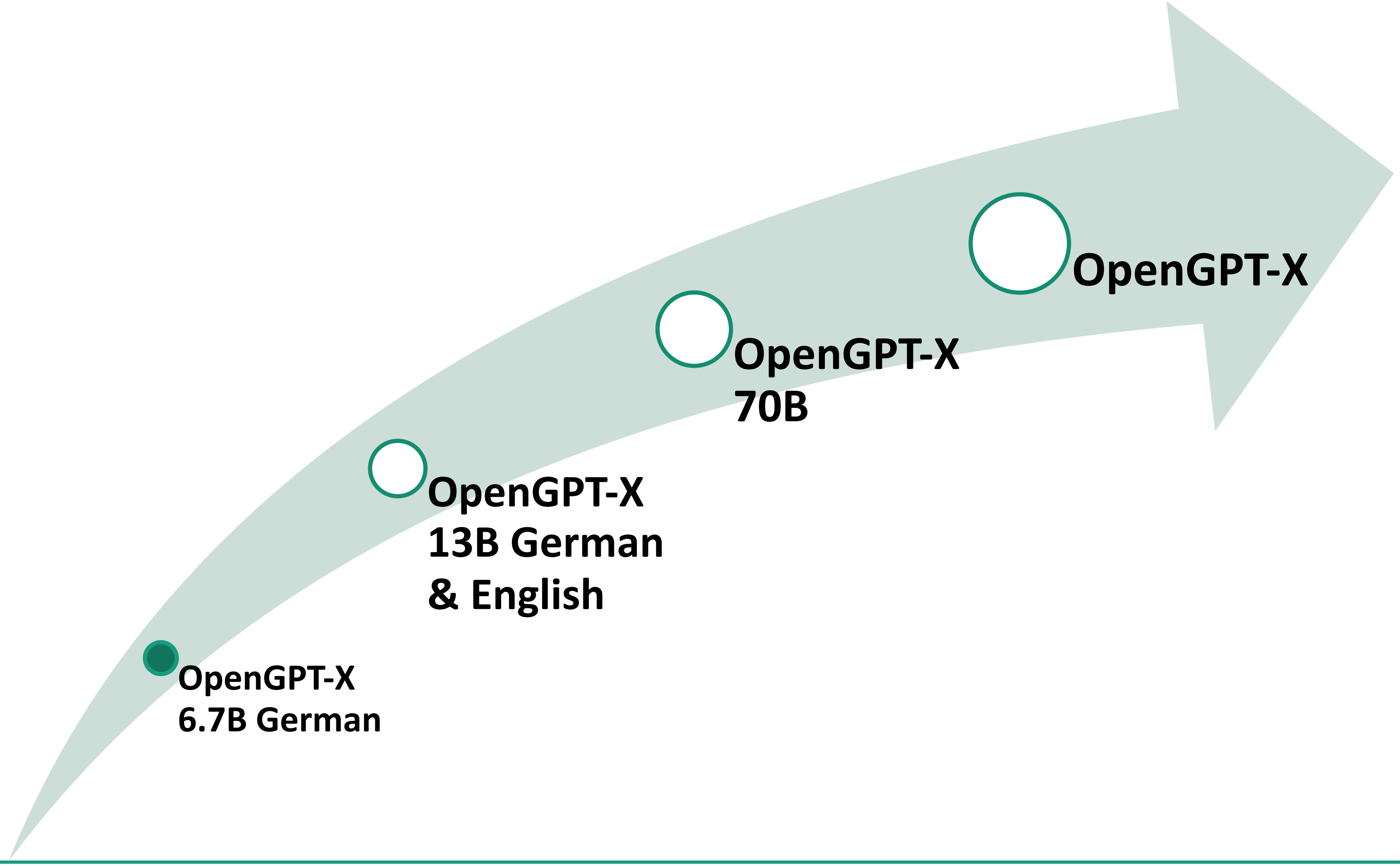
Fraunhofer IAIS

Timeline



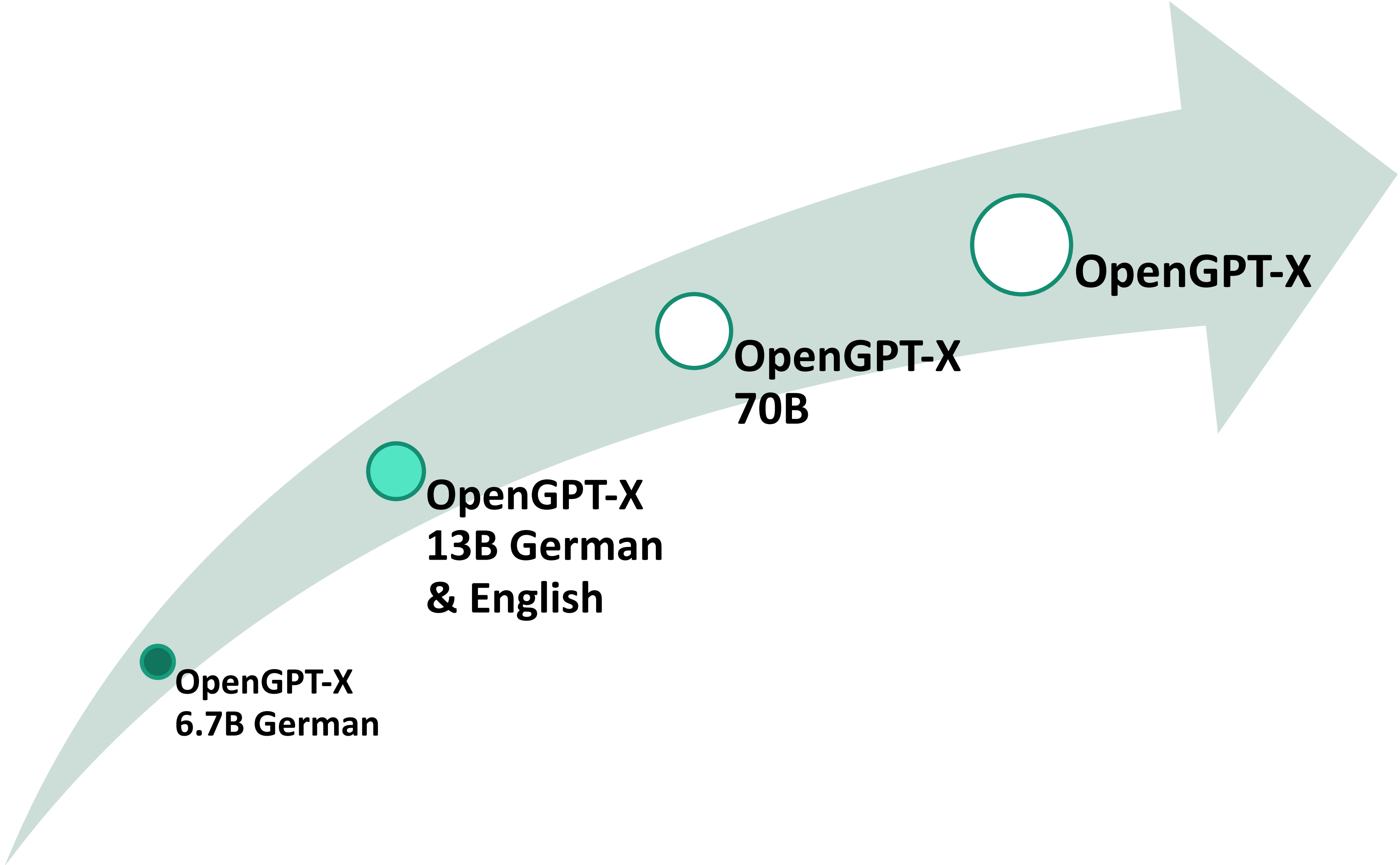
Fraunhofer
IAIS

Timeline



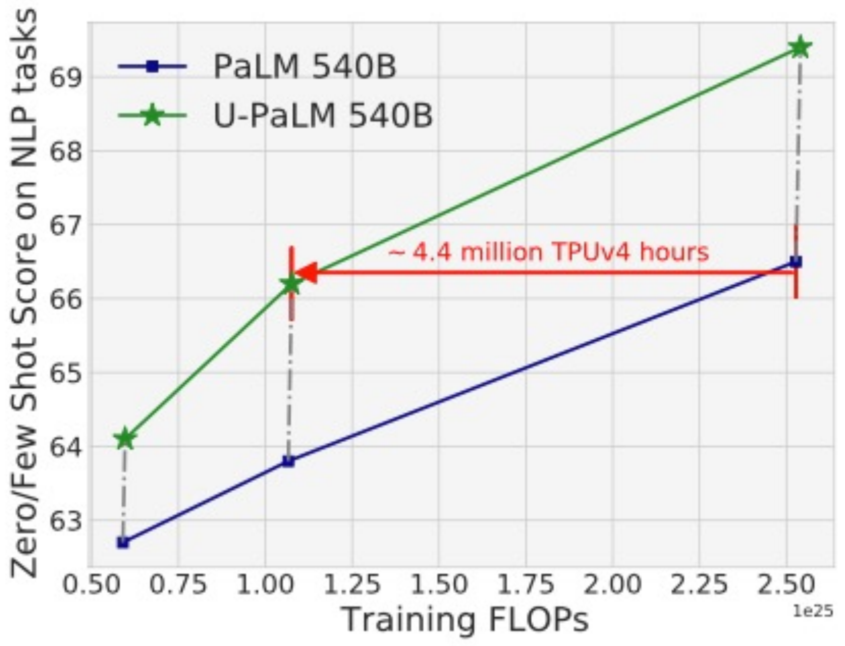
Fraunhofer IAIS

Timeline


OpenGPT-X
6.7B German
OpenGPT-X
13B German
& English
OpenGPT-X
70B
OpenGPT-X

UL2R

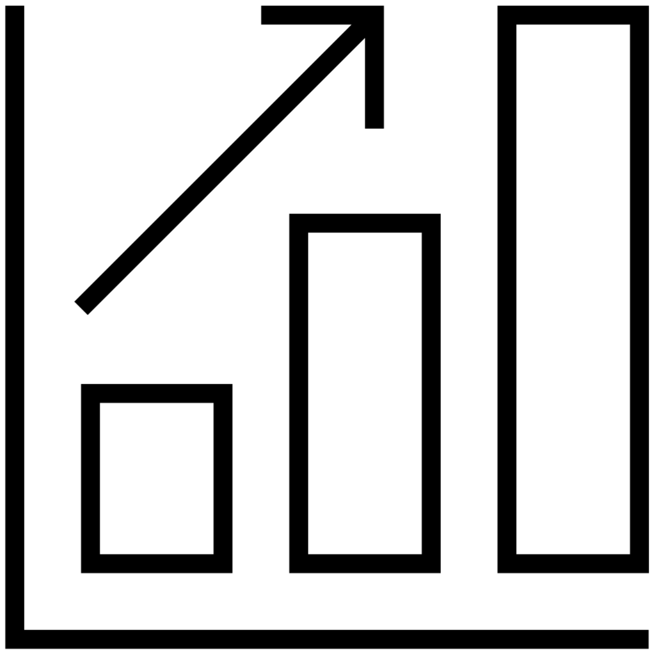
- Method to improve pretrained LLMs and their scaling curves while requiring only a small amount of additional compute
- Continue training with UL2’s mixture of denoisers objectives
- No new data sources required
- Improved scaling curve leads to “emergent abilities“ at smaller scale



(Tay *et al., 2022*)

Evaluation

- Many benchmarks for English
- Limited number of benchmarks for other European languages (currently focusing on German)
- Generate a multilingual benchmark suite – approach under discussion:
  - Automatically translate English benchmark datasets
  - Manually curate translated benchmark



04 Use Cases

Domain Media

- **Use Case 1**: Content analysis for the generation of meta data for personalized search
  - The speech application services developed within OpenGPT-X will automate the summarization of audio content and generate key words

- **Use Case 2:** Content synthesis for the creation of personalized articles in digital products (robot journalism)
  - The speech application services will generate articles for sports reporting, based on the user’s profile

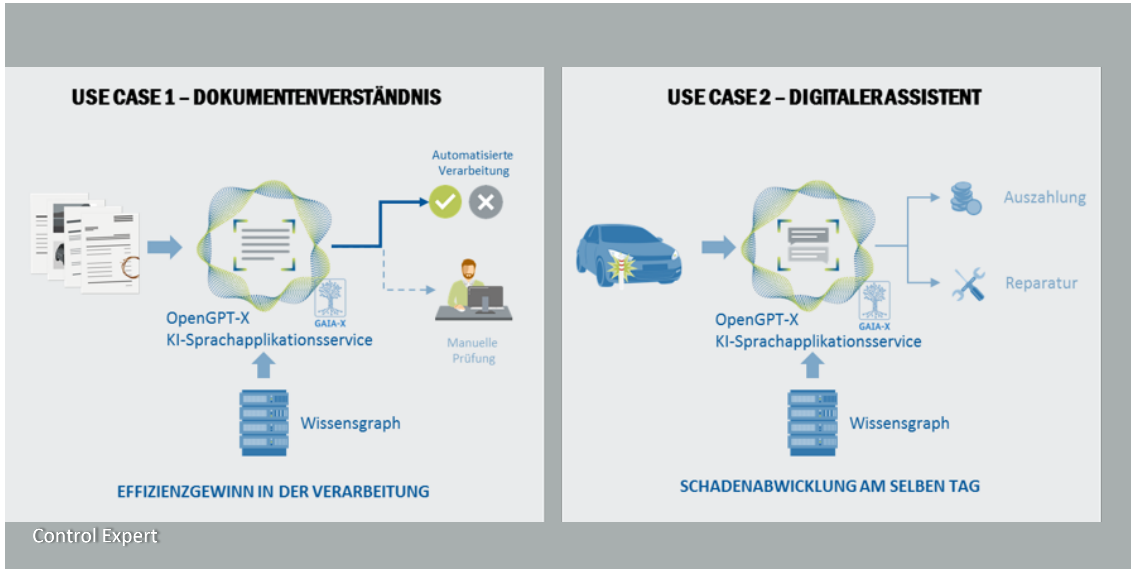
Language Models, that respect European values and are bias-free will support robot journalism

Domain Insurance

› **Use Case 1:** Understanding Documents
  › The speech application services developed within OpenGPT-X together with AI-based document analysis will help automate claims processing for vehicles

› **Use Case 2:** Digital assistant
  › The speech application services will be used for a digital assistant that will automate customer requests

The Speech Application services will significantly increase efficiency of claims processing and hence, reduce time.

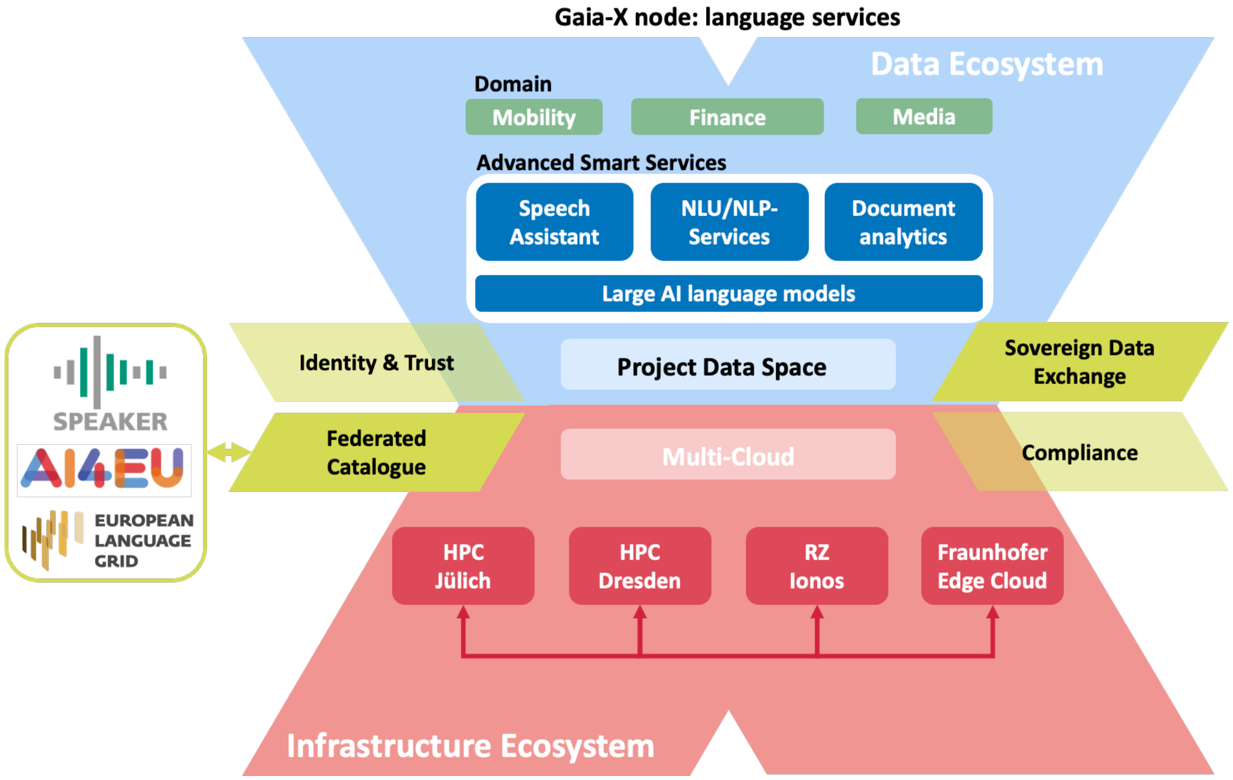


05 Outlook

Lighthouse project OpenGPT-X

Summary

- **Large AI language models “Made in Europe”** to support **Digital Sovereignty**
- Building a sustainable data and computing infrastructure: **a production line for large AI models**
- **Foundations** laid for the development of **Innovative AI Language Application Services** for the German & European Economy (Open Source)
- **Use cases** developed in OpenGPT-X will be published in the **Gaia-X use case gallery**
- The **OpenGPT-X AI language model** will be published as **OSS**
- Innovation of **marketable AI-based speech application services**



Fraunhofer IAIS

Thank you!

Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, Sankt Augustin, Germany

References

- Laurençon, Hugo, et al. (2022) "The bigscience roots corpus: A 1.6 tb composite multilingual dataset." Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track.
- Gao, Leo, et al. (2020) "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027*.
- Lee, Katherine, et al. (2022) *Deduplicating Training Data Makes Language Models Better. ACL (1) 2022: 8424-8445.*
- Vaswani, Ashish, et al. (2017) "Attention is all you need." *Advances in neural information processing systems* 30.
- Radford, Alec, et al. (2018) "Improving language understanding by generative pre-training."
- Radford, Alec, et al. (2019)"Language models are unsupervised multitask learners." *OpenAI blog* 1.8.
- Brown, Tom, et al. (2020)"Language models are few-shot learners." *Advances in neural information processing systems* 33: 1877-1901.
- Hoffmann, Jordan, et al. (2022) "Training compute-optimal large language models." *arXiv preprint arXiv:2203.15556.*
- Kaplan, Jared, et al. (2020 "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361*.
- Shuster, Kurt, et al. (2021) "Retrieval augmentation reduces hallucination in conversation." *EMNLP (Findings) 2021: 3784-3803*
- Tay, Yi, et al. (2022) "Transcending scaling laws with 0.1% extra compute." *arXiv preprint arXiv:2210.11399*.