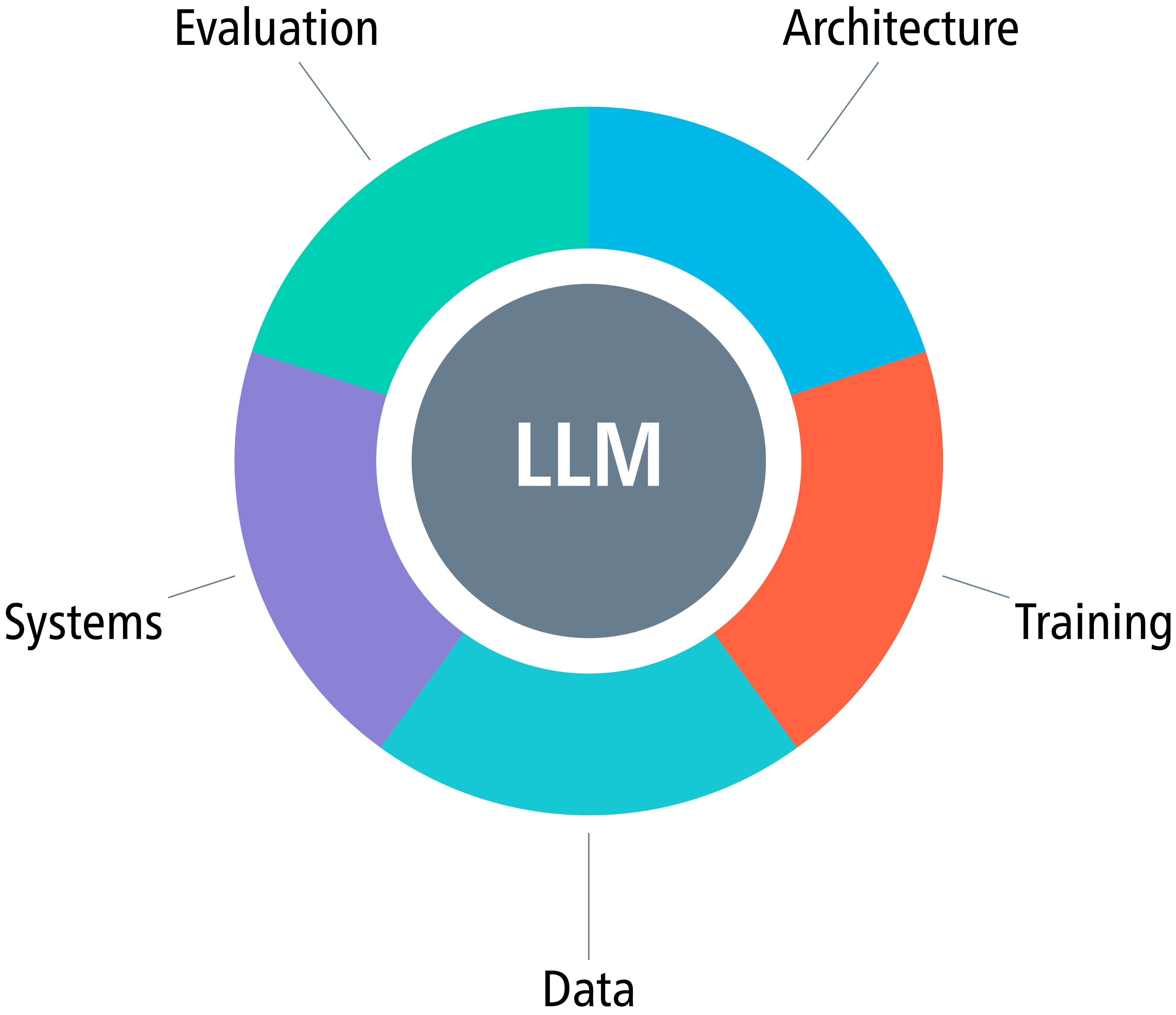
Natural Language Processing

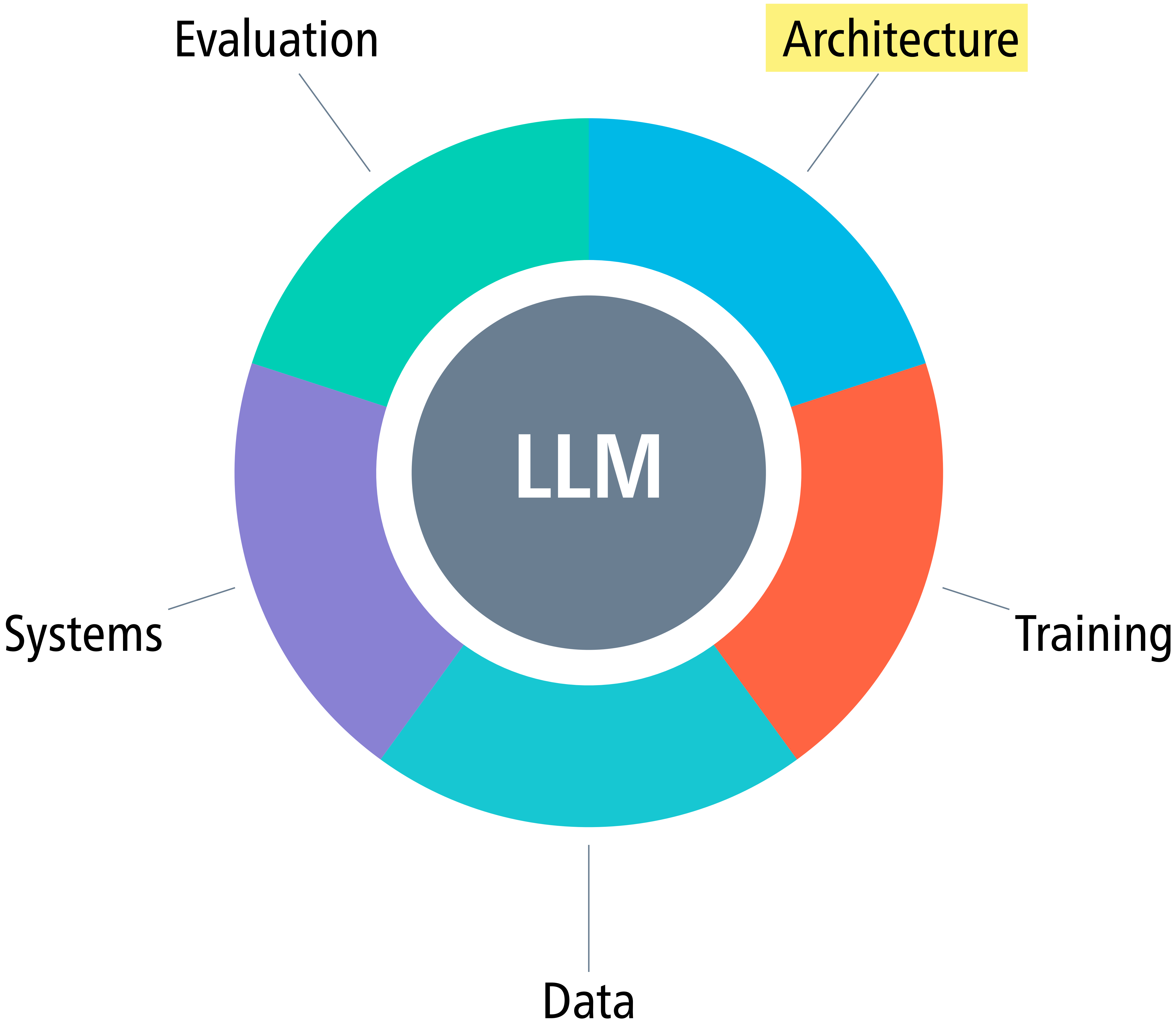
Introduction to LLM development

Marco Kuhlmann

Department of Computer and Information Science

LINKÖPING UNIVERSITY


Evaluation
Architecture
LLM
Systems
Training
Data


Evaluation
Architecture
LLM
Systems
Training
Data

Decoder-based language models

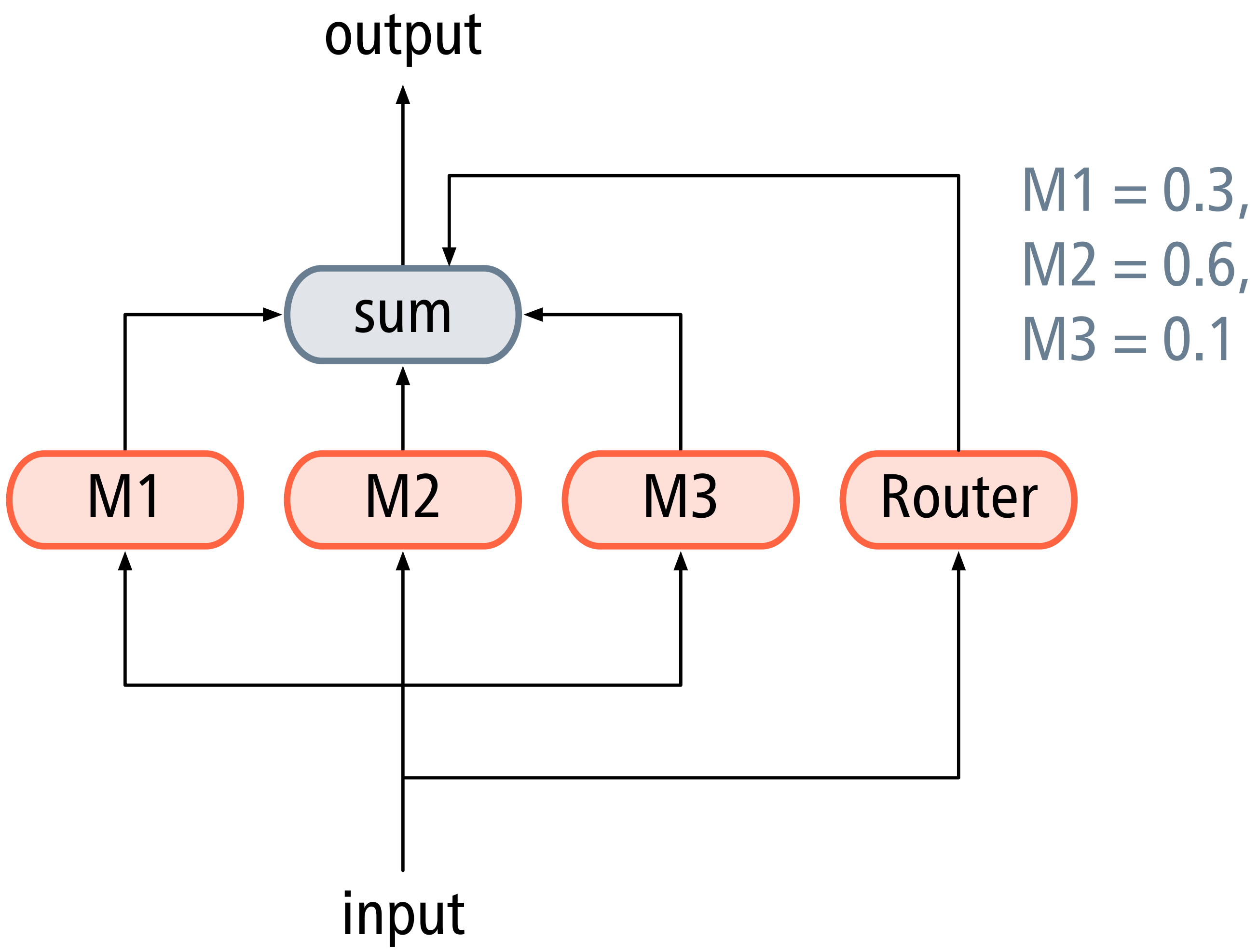
	GPT	Gemini	Llama	DeepSeek
Latest model (release)	OpenAI o3 (2025-01)	Gemini 2.0 (2025-01)	Llama 3.3 (2024–12)	DeepSeek-R1 (2025-01)
Parameter size	undisclosed	undisclosed	70B	37B
Context size	200K	1M	128K	128K
License	Proprietary	Proprietary	Open-source	Open-source

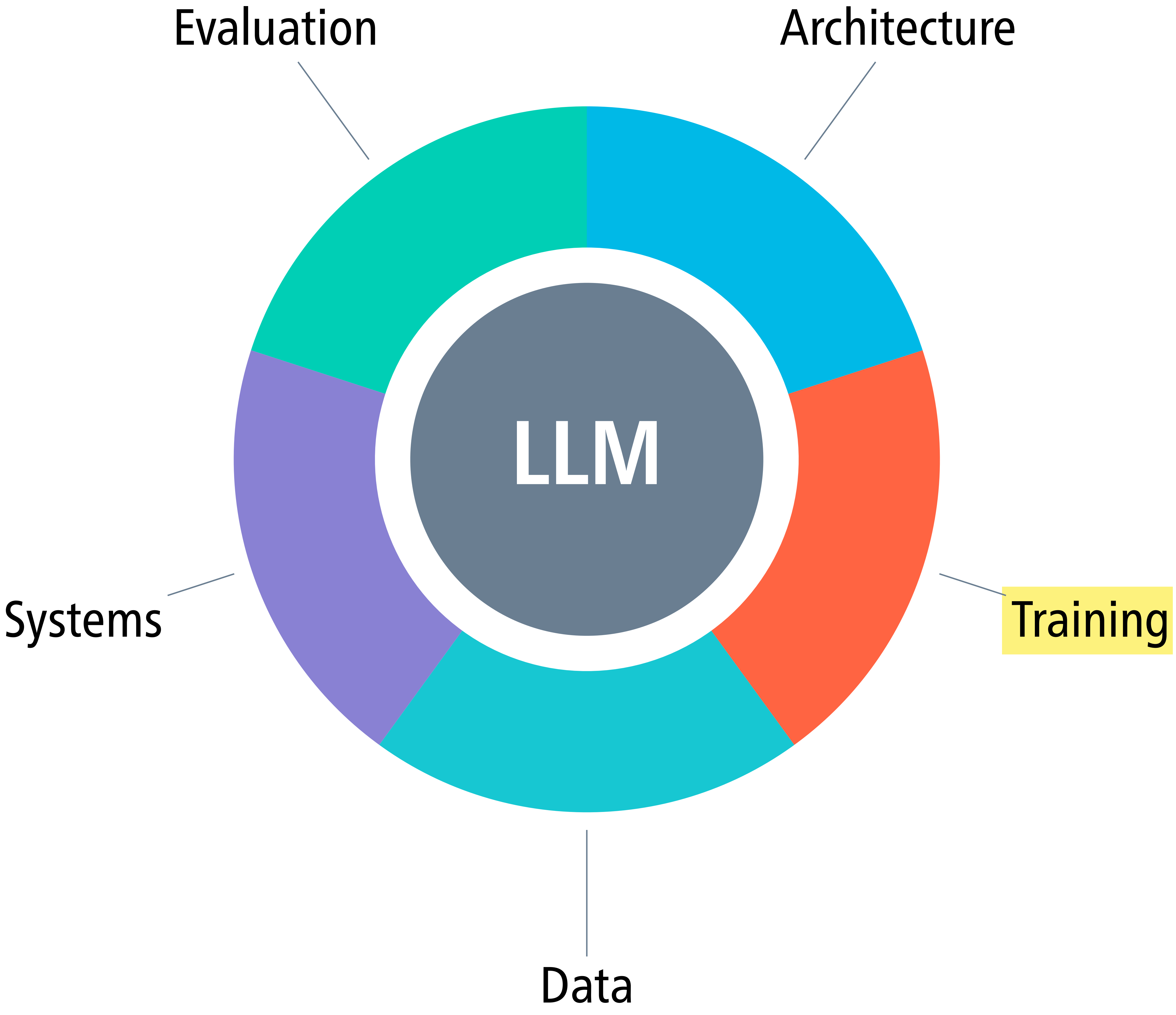
Mixture of experts

- The **Mixture of Experts** architecture dynamically selects sub-models.

- For each input, it activates only a fraction of the model's total parameters.

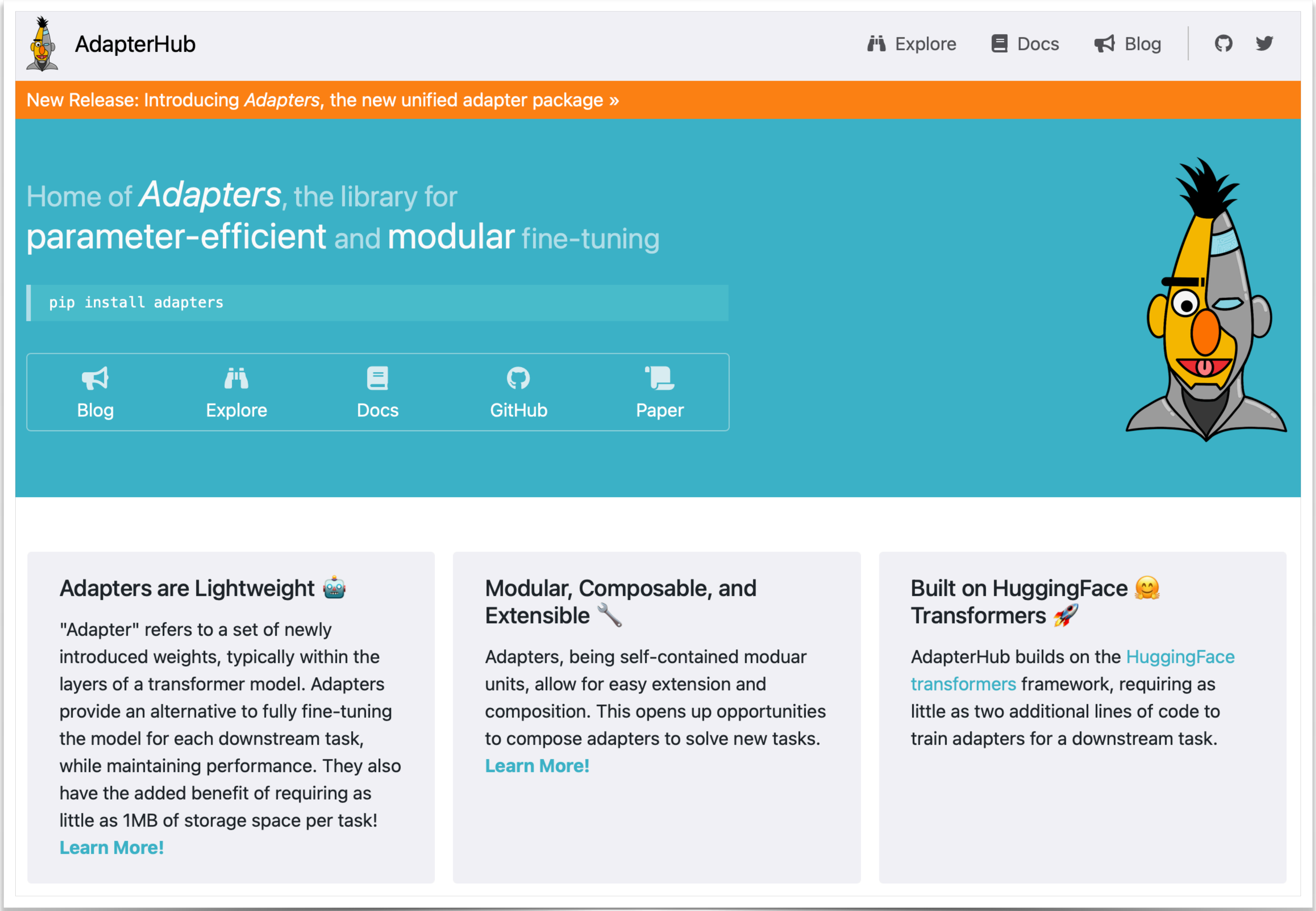
- A learnable router chooses which experts to activate.




Evaluation
Architecture
LLM
Systems
Training
Data

	unsupervised pre-training	**instruction fine-tuning**	**reward modelling**	**reinforcement learning**
data	**raw text from the Internet** trillions of words low quality, high quantity	**ideal dialogues** 10k–100k low quantity, high quality	**annotated dialogues** 100k–1M low quantity, high quality	**generated dialogues** 10k–100k low quantity, high quality
algorithm	**language modelling** predict the next word	**language modelling** predict the next word	**binary classification** reward consistent with preferences?	**reinforcement learning** generate text for maximal reward
resources	1000s of GPUs several months of training time GPT, Llama	1–100 GPUs several days of training time	1–100 GPUs several days of training time	1–100 GPUer several days of training time ChatGPT, Claude

— **language model** ——————————— **assistant model** →



New Release: Introducing *Adapters*, the new unified adapter package »

Home of *Adapters*, the library for
parameter-efficient and modular fine-tuning

```
pip install adapters
```

Adapters are Lightweight 🤖

"Adapter" refers to a set of newly introduced weights, typically within the layers of a transformer model. Adapters provide an alternative to fully fine-tuning the model for each downstream task, while maintaining performance. They also have the added benefit of requiring as little as 1MB of storage space per task!

Learn More!

Modular, Composable, and Extensible 🔧

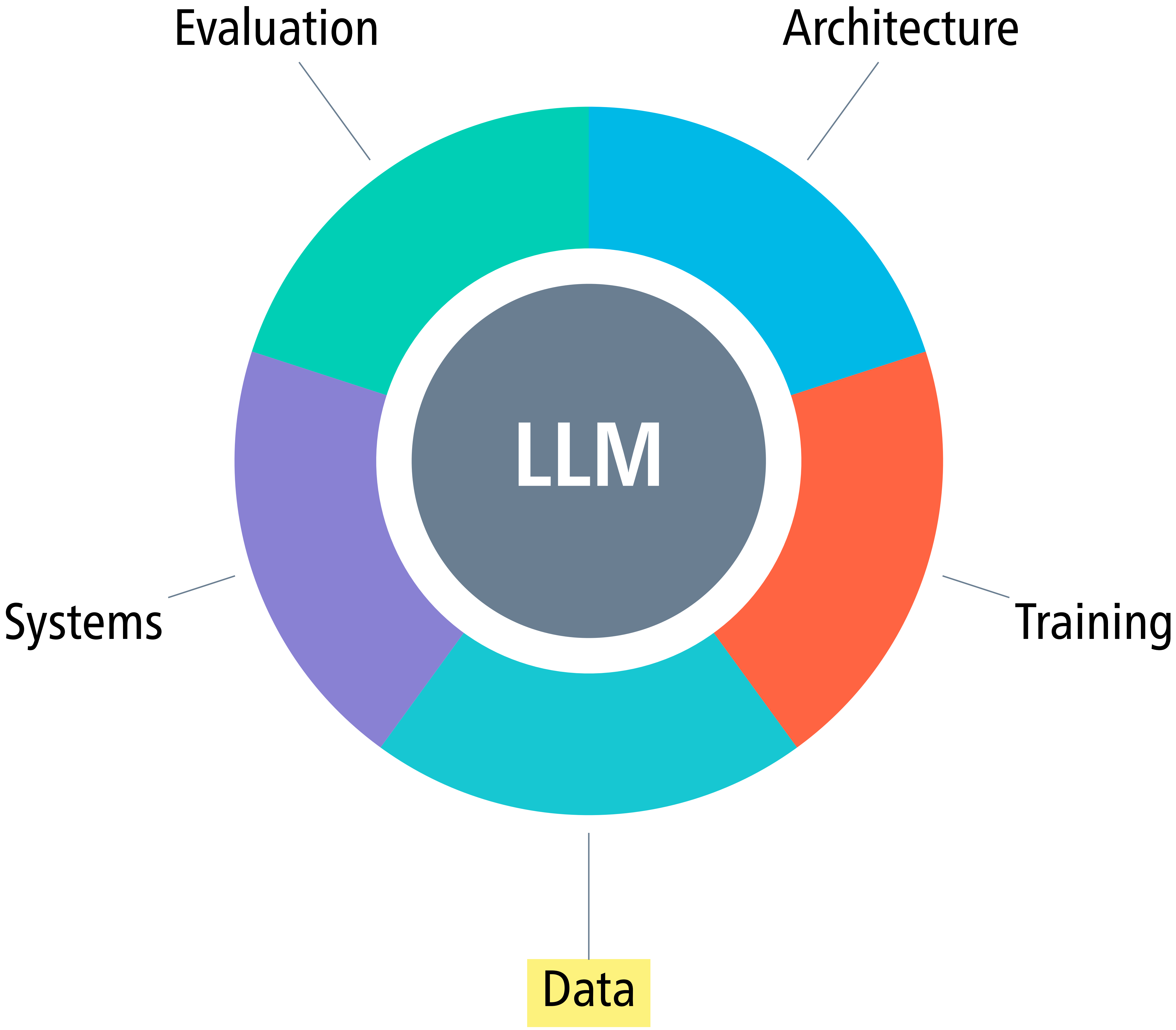
Adapters, being self-contained moduar units, allow for easy extension and composition. This opens up opportunities to compose adapters to solve new tasks.

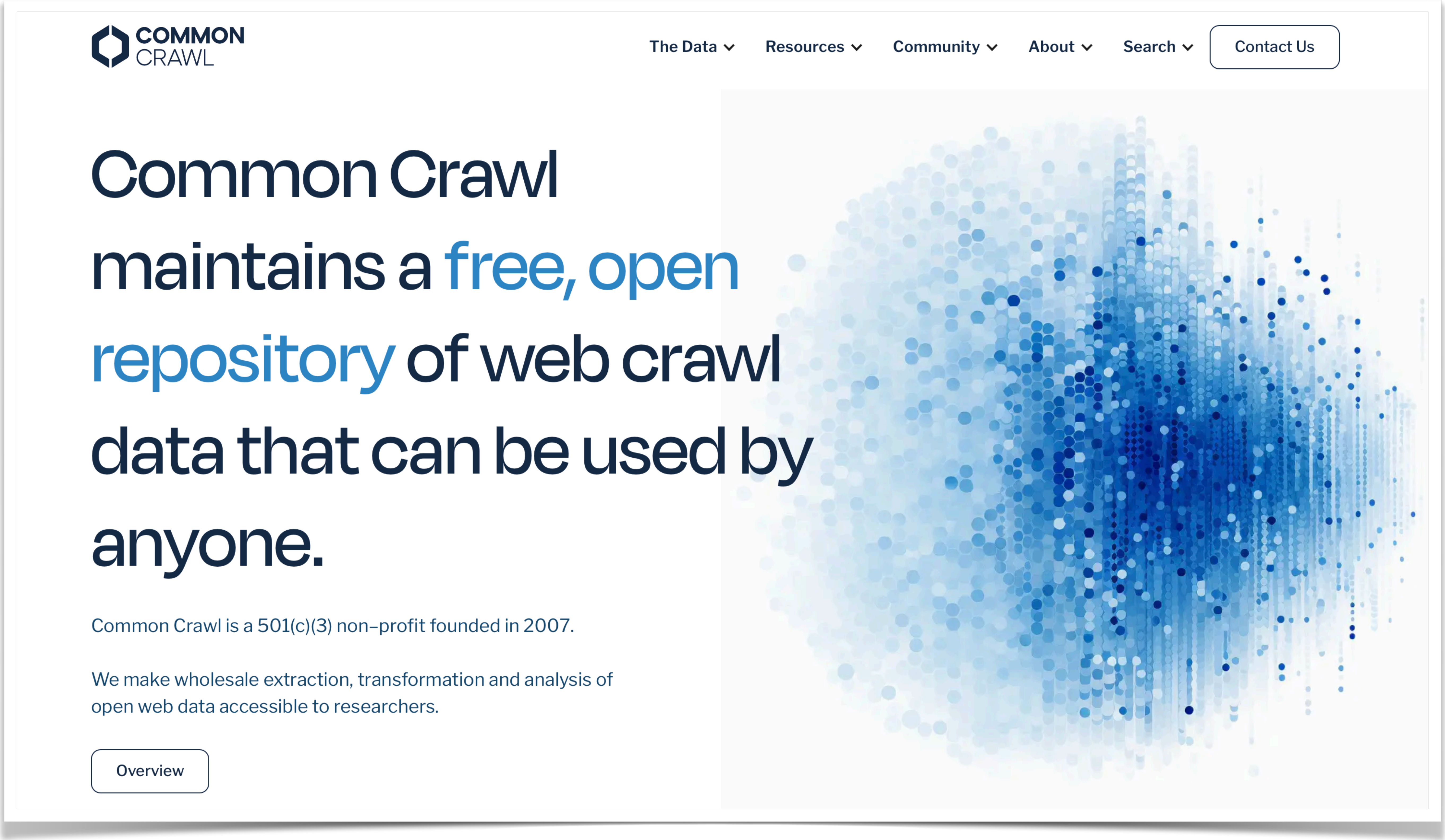
Learn More!

Built on HuggingFace 🤗 Transformers 🚀

AdapterHub builds on the HuggingFace transformers framework, requiring as little as two additional lines of code to train adapters for a downstream task.

Source: AdapterHub


Evaluation
Architecture
LLM
Systems
Training
Data

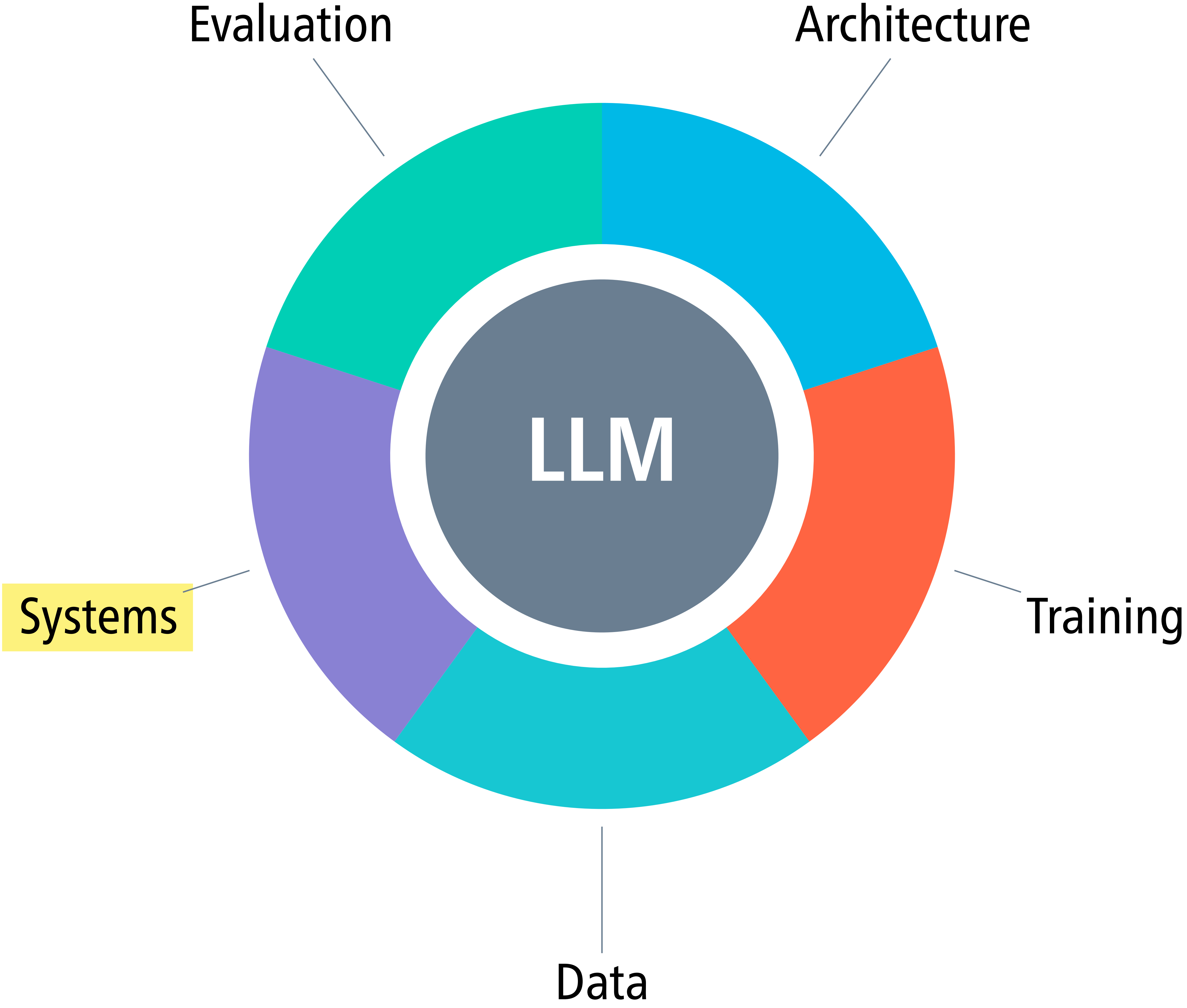


Common Crawl maintains a free, open repository of web crawl data that can be used by anyone.

Common Crawl is a 501(c)(3) non–profit founded in 2007.

We make wholesale extraction, transformation and analysis of open web data accessible to researchers.

Source: Common Crawl

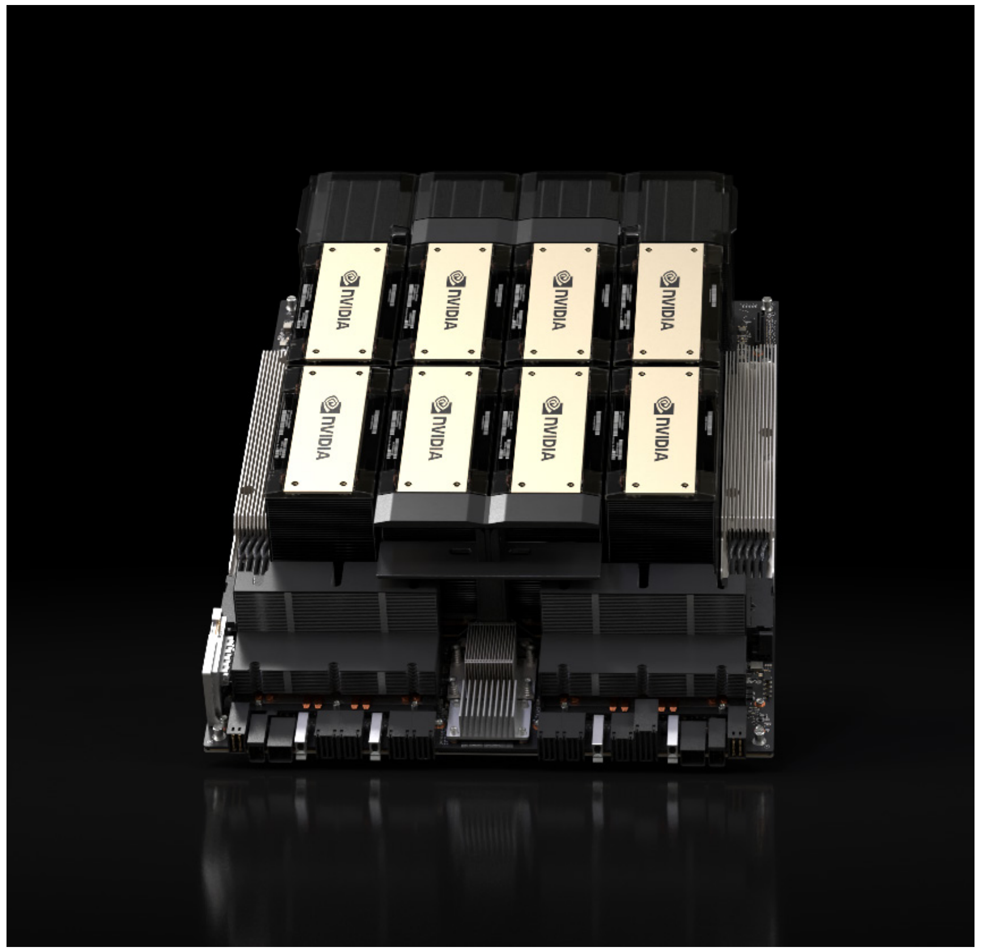

Evaluation
Architecture
LLM
Systems
Training
Data

Datasheet

NVIDIA

NVIDIA H200 Tensor Core GPU

Supercharging AI and HPC workloads.



Source: NVIDIA

Higher Performance With Larger, Faster Memory

The NVIDIA H200 Tensor Core GPU supercharges generative AI and high-performance computing (HPC) workloads with game-changing performance and memory capabilities.

Based on the **NVIDIA Hopper™ architecture**, the NVIDIA H200 is the first GPU to offer 141 gigabytes (GB) of HBM3e memory at 4.8 terabytes per second (TB/s)—that's nearly double the capacity of the **NVIDIA H100 Tensor Core GPU** with 1.4X more memory bandwidth. The H200's larger and faster memory accelerates generative AI and large language models, while advancing scientific computing for HPC workloads with better energy efficiency and lower total cost of ownership.

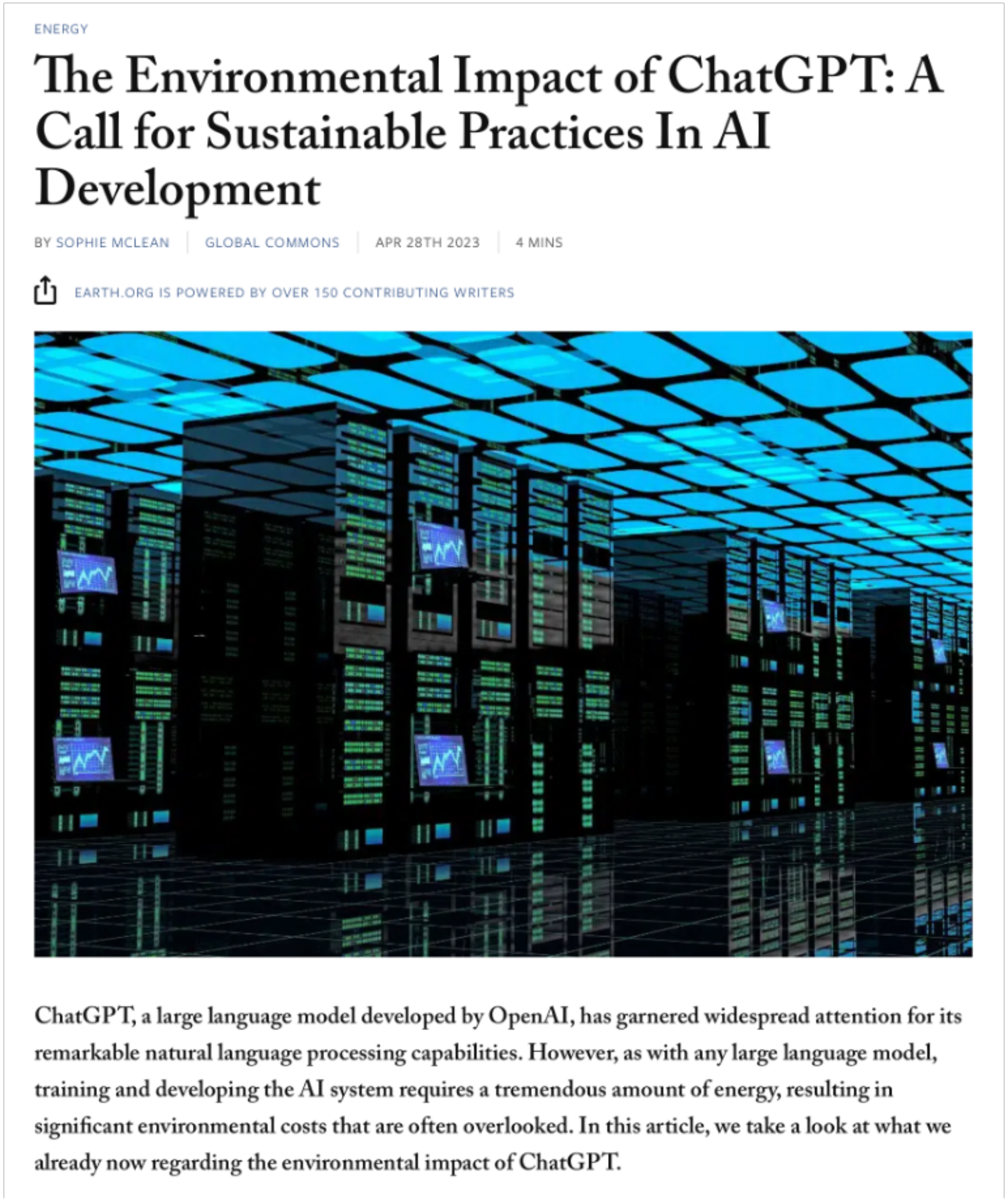
Key Features

- 141GB of HBM3e GPU memory
- 4.8TB/s of memory bandwidth
- 4 petaFLOPS of FP8 performance
- 2X LLM inference performance
- 110X HPC performance

Unlock Insights With High-Performance LLM Inference

In the ever-evolving landscape of AI, businesses rely on large language models to address a diverse range of inference needs. An **AI inference** accelerator must deliver the highest throughput at the lowest TCO when deployed at scale for a massive user base.

The H200 doubles inference performance compared to H100 GPUs when handling



ENERGY

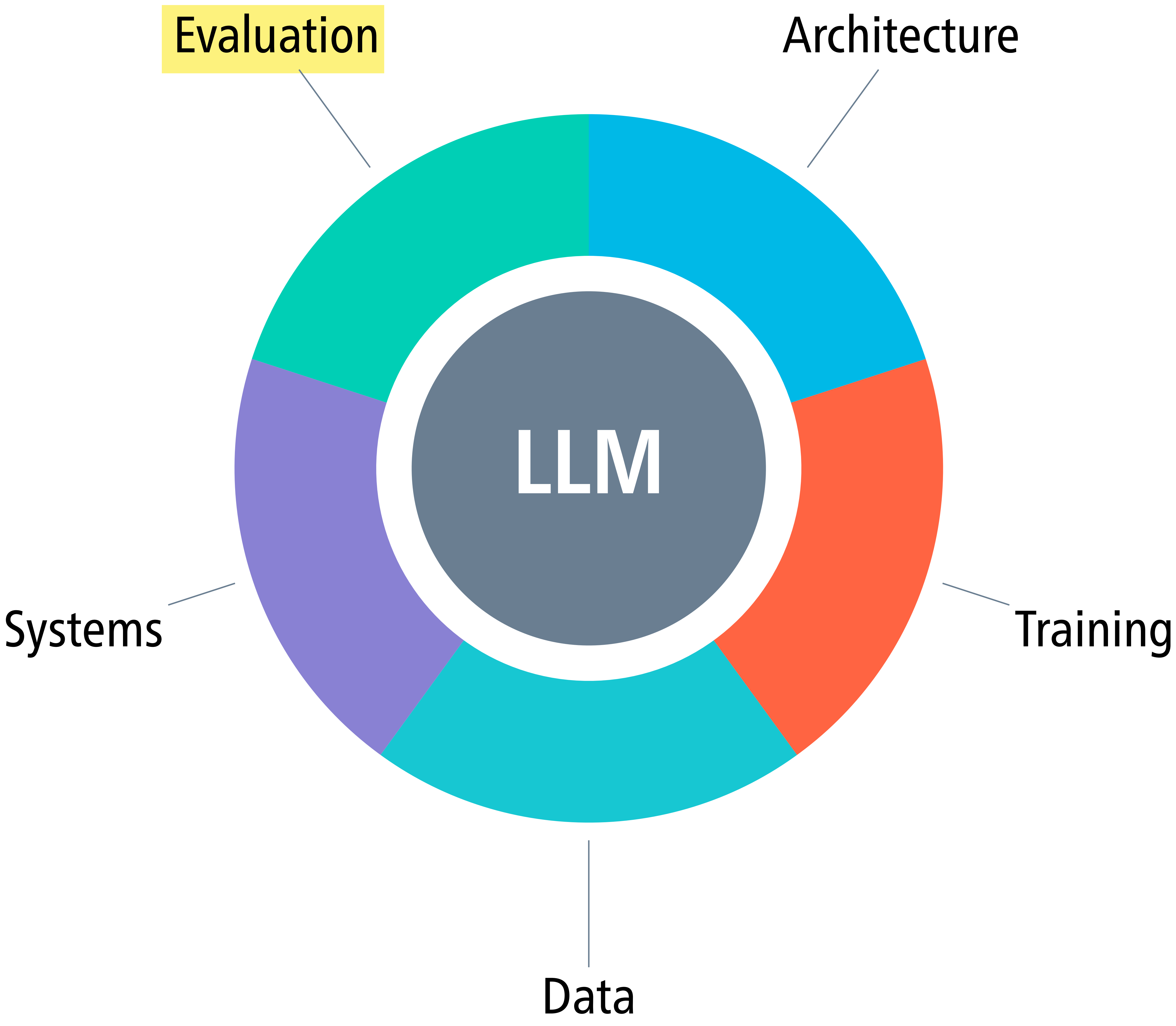
The Environmental Impact of ChatGPT: A Call for Sustainable Practices In AI Development

BY SOPHIE MCLEAN | GLOBAL COMMONS | APR 28TH 2023 | 4 MINS

EARTH.ORG IS POWERED BY OVER 150 CONTRIBUTING WRITERS

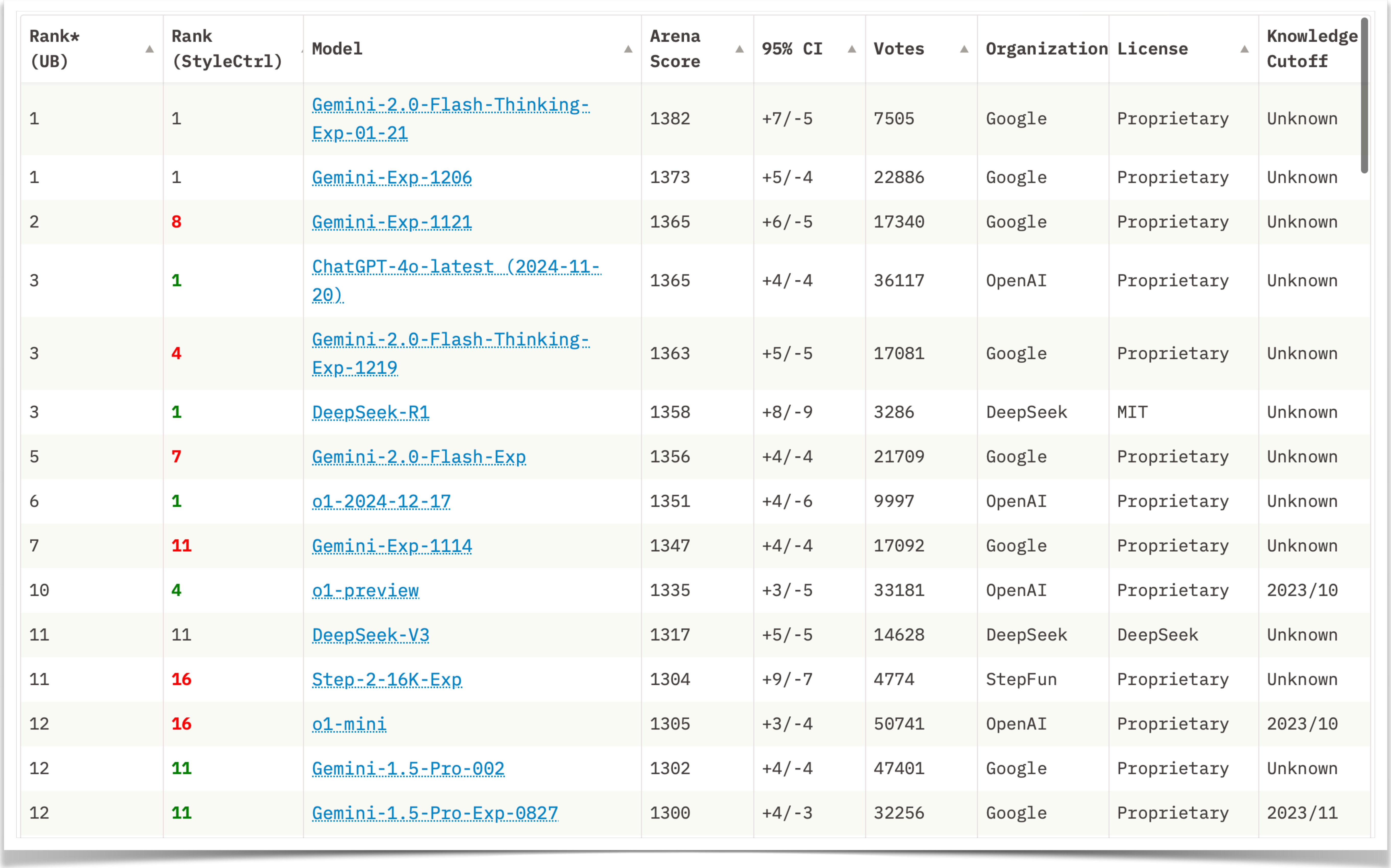
ChatGPT, a large language model developed by OpenAI, has garnered widespread attention for its remarkable natural language processing capabilities. However, as with any large language model, training and developing the AI system requires a tremendous amount of energy, resulting in significant environmental costs that are often overlooked. In this article, we take a look at what we already now regarding the environmental impact of ChatGPT.

Source: earth.org


Evaluation
Architecture
LLM
Systems
Training
Data

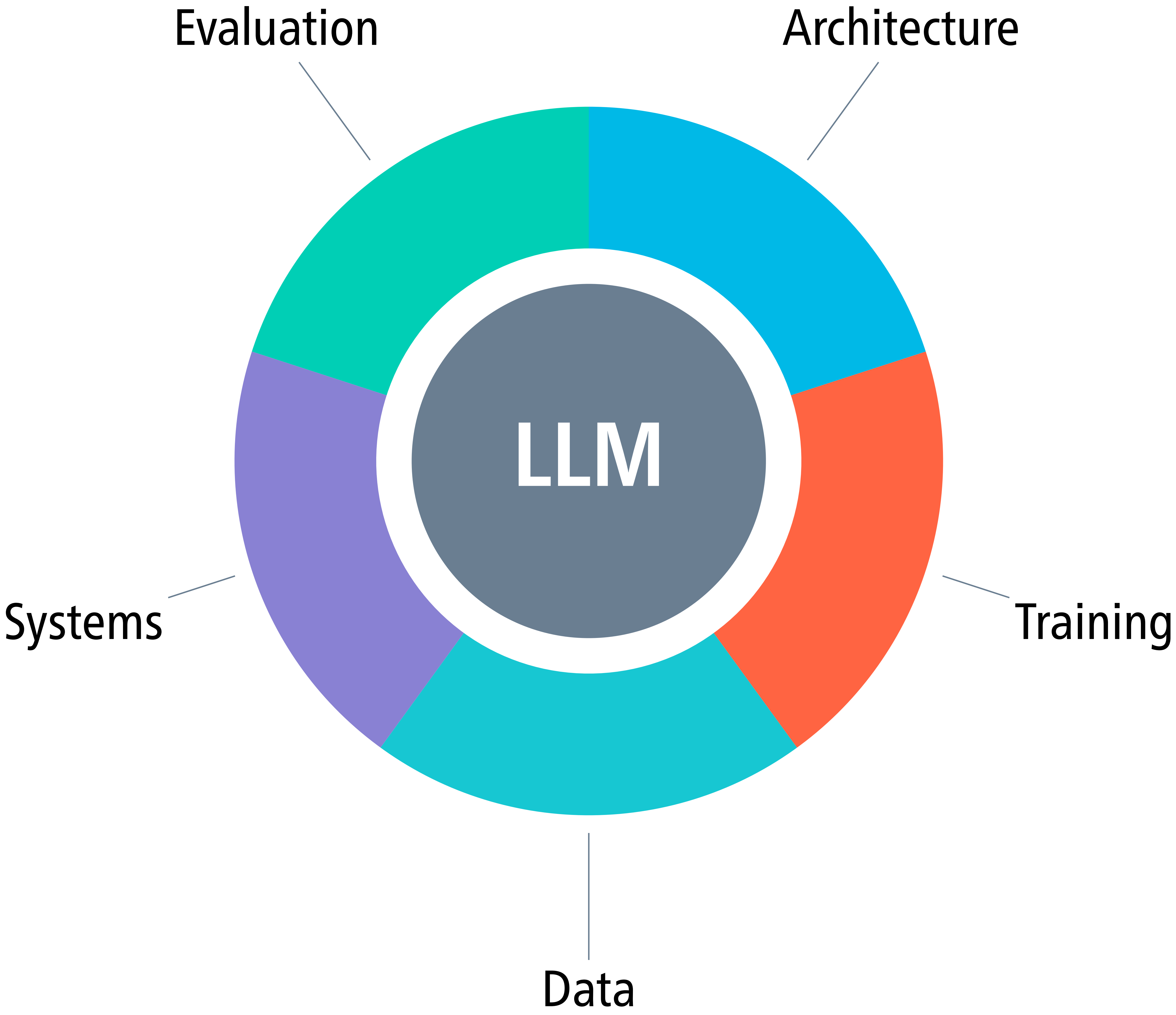
Benchmarks

Benchmark	Release year	Tasks	Size
MMLU	2020	general knowledge	15,000 questions
MATH	2021	mathematical reasoning	12,500 problems
BIG-Bench	2022	reasoning, extrapolation	200 tasks
SWE-bench	2023	coding, reasoning	2,200 GitHub issues



Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	1	Gemini-2.0-Flash-Thinking-Exp-01-21	1382	+7/-5	7505	Google	Proprietary	Unknown
1	1	Gemini-Exp-1206	1373	+5/-4	22886	Google	Proprietary	Unknown
2	8	Gemini-Exp-1121	1365	+6/-5	17340	Google	Proprietary	Unknown
3	1	ChatGPT-4o-latest (2024-11-20)	1365	+4/-4	36117	OpenAI	Proprietary	Unknown
3	4	Gemini-2.0-Flash-Thinking-Exp-1219	1363	+5/-5	17081	Google	Proprietary	Unknown
3	1	DeepSeek-R1	1358	+8/-9	3286	DeepSeek	MIT	Unknown
5	7	Gemini-2.0-Flash-Exp	1356	+4/-4	21709	Google	Proprietary	Unknown
6	1	o1-2024-12-17	1351	+4/-6	9997	OpenAI	Proprietary	Unknown
7	11	Gemini-Exp-1114	1347	+4/-4	17092	Google	Proprietary	Unknown
10	4	o1-preview	1335	+3/-5	33181	OpenAI	Proprietary	2023/10
11	11	DeepSeek-V3	1317	+5/-5	14628	DeepSeek	DeepSeek	Unknown
11	16	Step-2-16K-Exp	1304	+9/-7	4774	StepFun	Proprietary	Unknown
12	16	o1-mini	1305	+3/-4	50741	OpenAI	Proprietary	2023/10
12	11	Gemini-1.5-Pro-002	1302	+4/-4	47401	Google	Proprietary	Unknown
12	11	Gemini-1.5-Pro-Exp-0827	1300	+4/-3	32256	Google	Proprietary	2023/11

Source: Chatbot Arena LLM Leaderboard

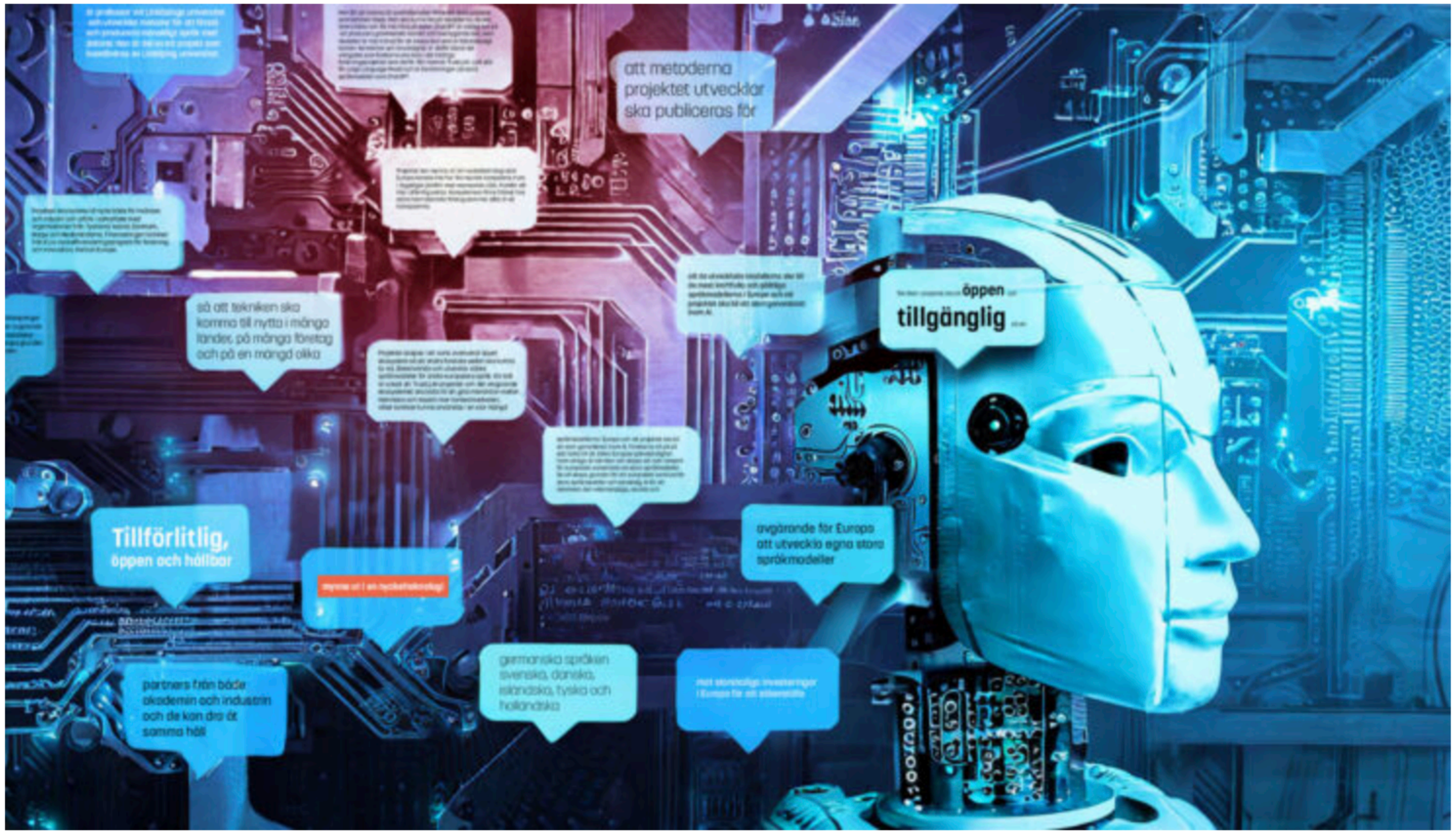

Evaluation
Architecture
LLM
Systems
Training
Data

Developing a reliable ChatGPT for European languages

01 November 2023

Sara Läthén

Reliable, open and sustainable – those are the qualities aimed for in a ChatGPT for European languages including Swedish. But there are still big problems that need to be solved before ChatGPT can be considered beneficial to society.


att metoderna projektet utvecklar ska publiceras för
så att tekniken ska komma till nytta i många länder, på många företag och på en mängd olika
öppen
tillgänglig
Tillförlitlig, öppen och hållbar
avgörande för Europa att utveckla egna stora språkmodeller
partners från både akademin och industrin och de kan dra åt samma håll
germanska språken svenska, danska, isländska, tyska och holländska

Bilden är delvis skapad med AI.

Marco Kuhlmann, professor at Linköping University, develops methods to understand and computer generate human speech. He is part of a project coordinated by Linköping University that has just been granted around SEK 70 million.