LLM LE1 VT2026

Introduction

Fredrik Heintz

Dept. of Computer Science
Linköping University

fredrik.heintz@liu.se
@FredrikHeintz

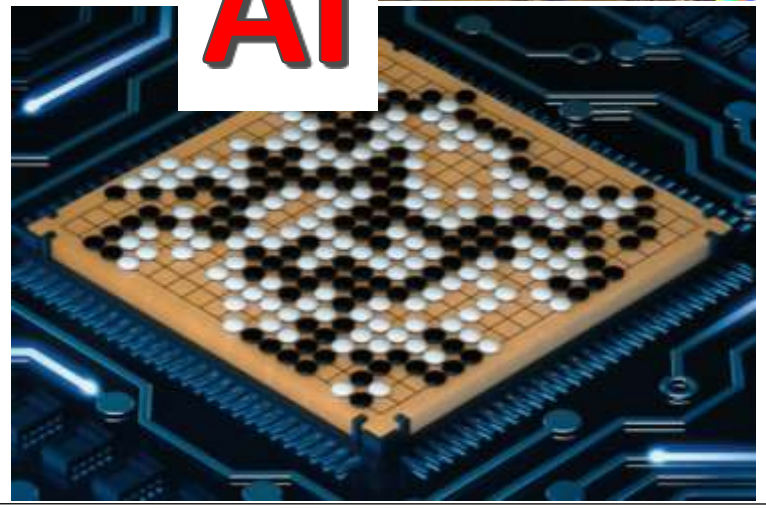
Outline:

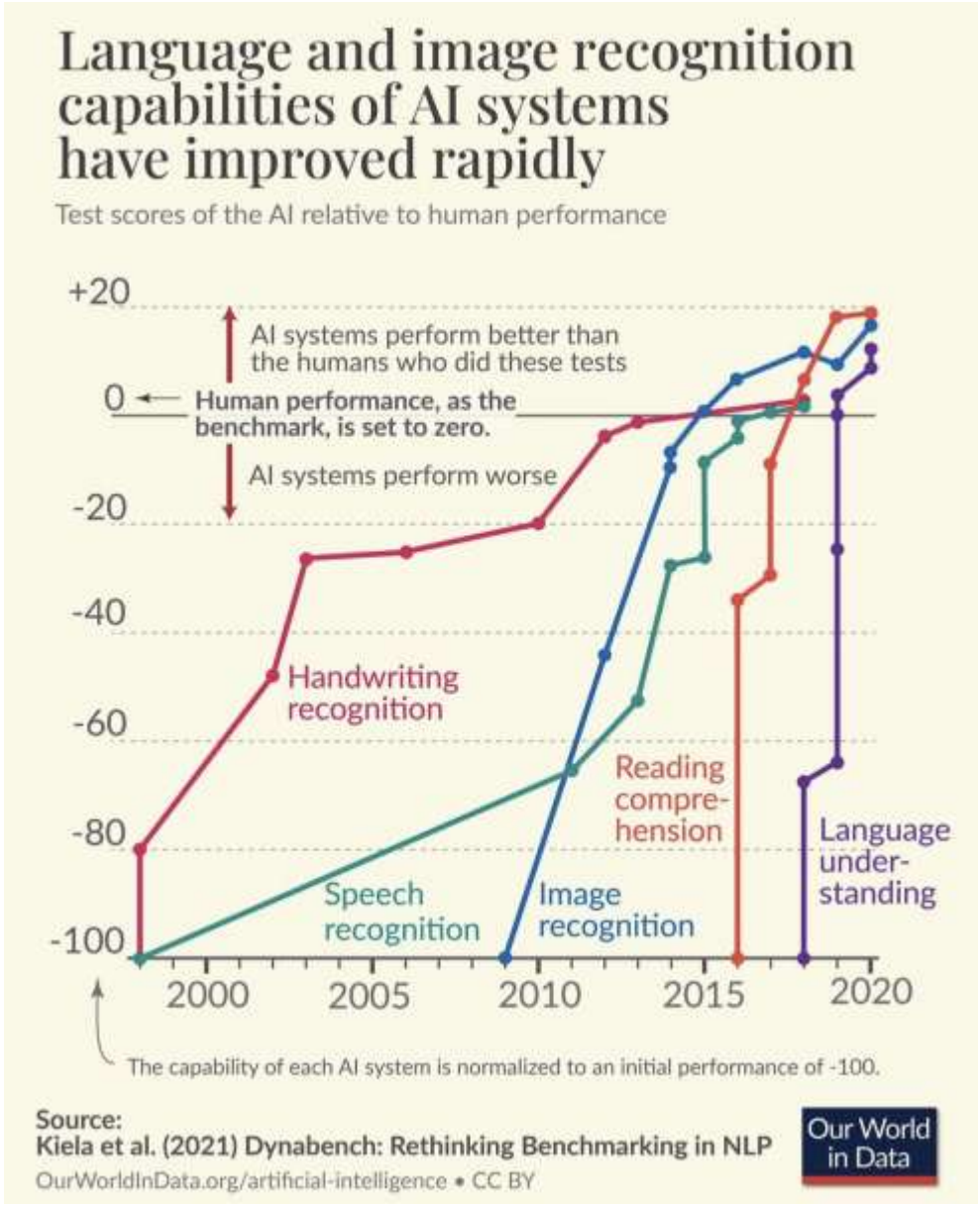
- Introduction
- Course Overview
- Natural Language Processing
- Large Language Models
- TrustLLM

LINKÖPING UNIVERSITY

AI Development is Fast

AI





Generative AI

"Teddy bears mixing sparkling chemicals as mad scientists in a 'steampunk' style."

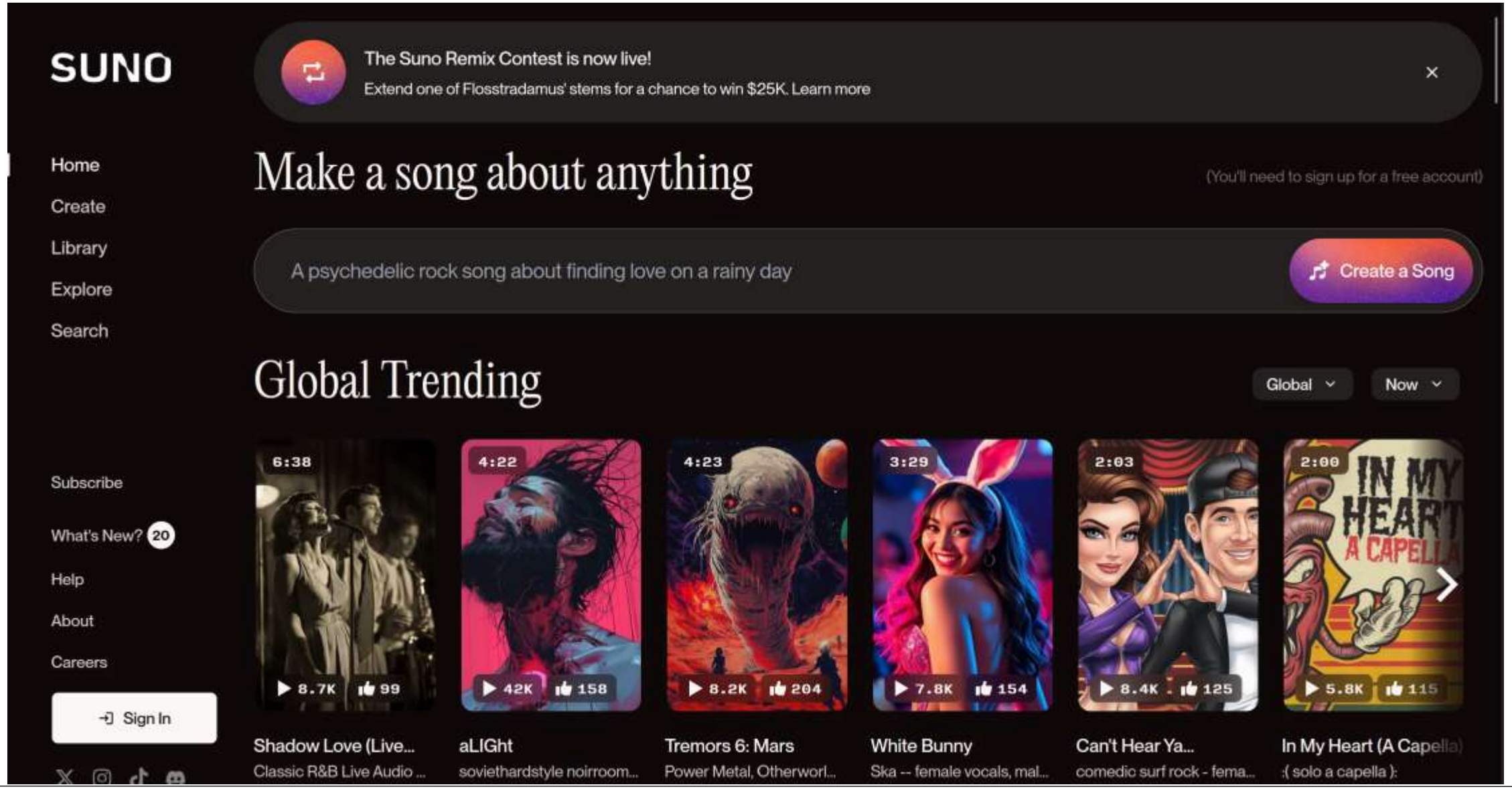
https://openai.com/dall-e-2/

Sora

A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.

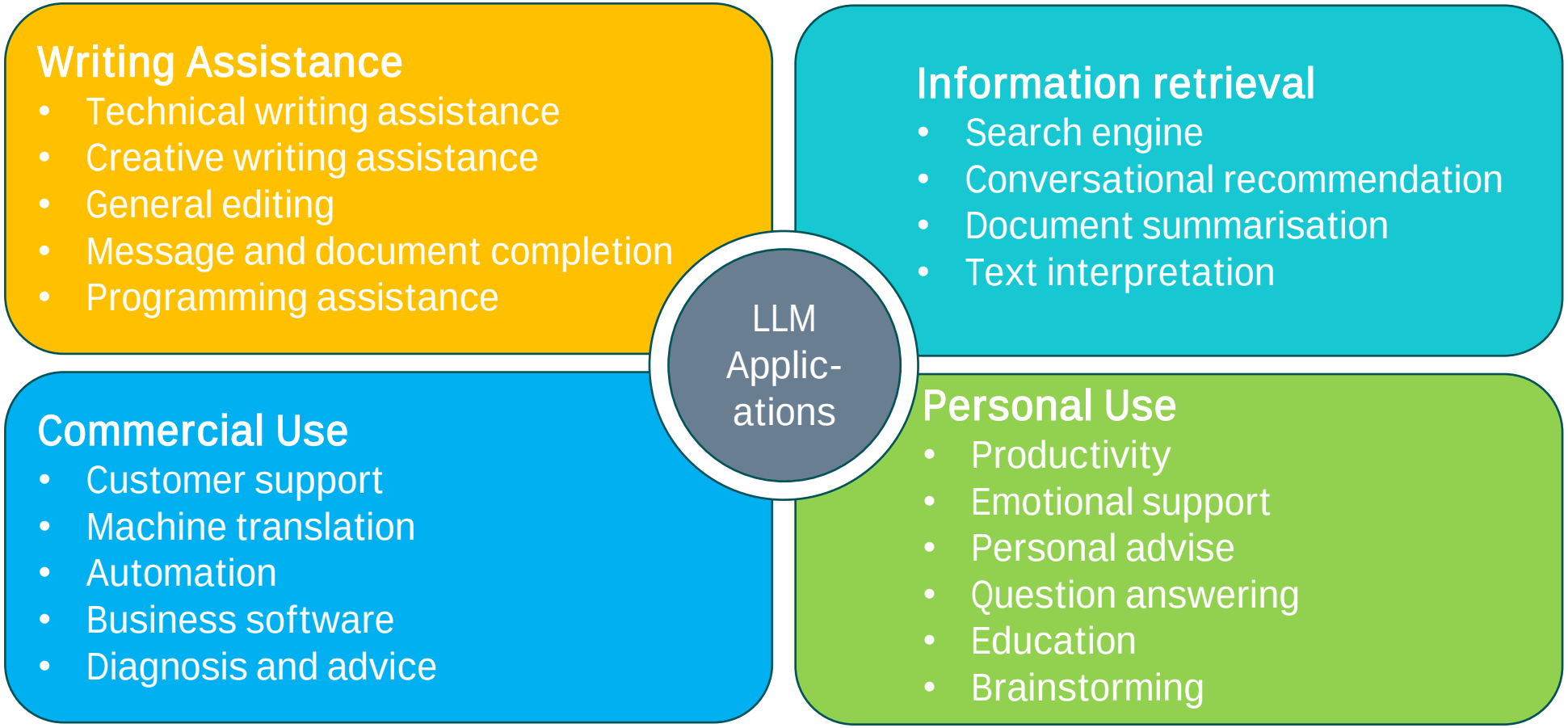
Veo3

https://v03ai.com/

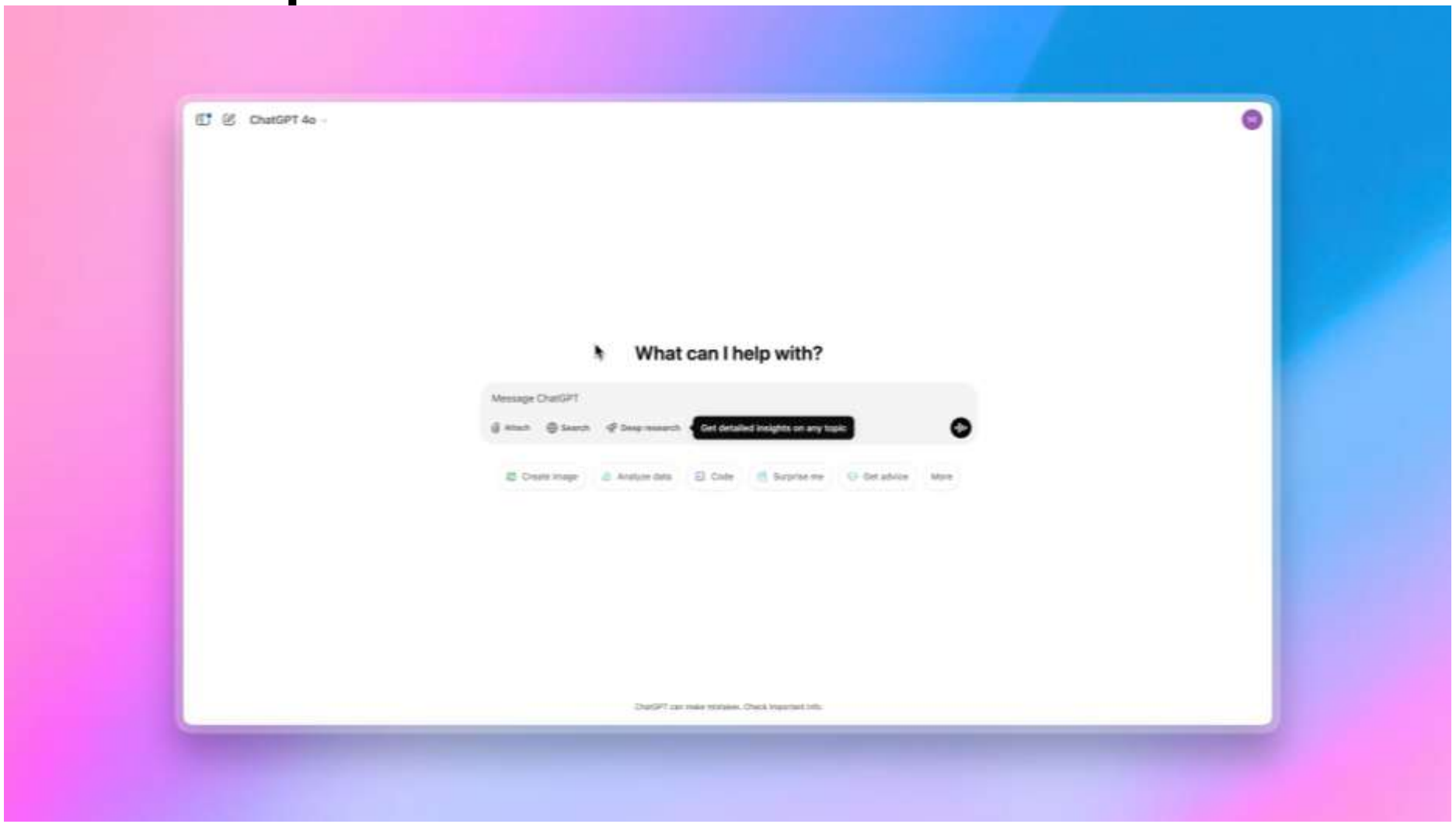


https://suno.com/

Large Language Model Applications



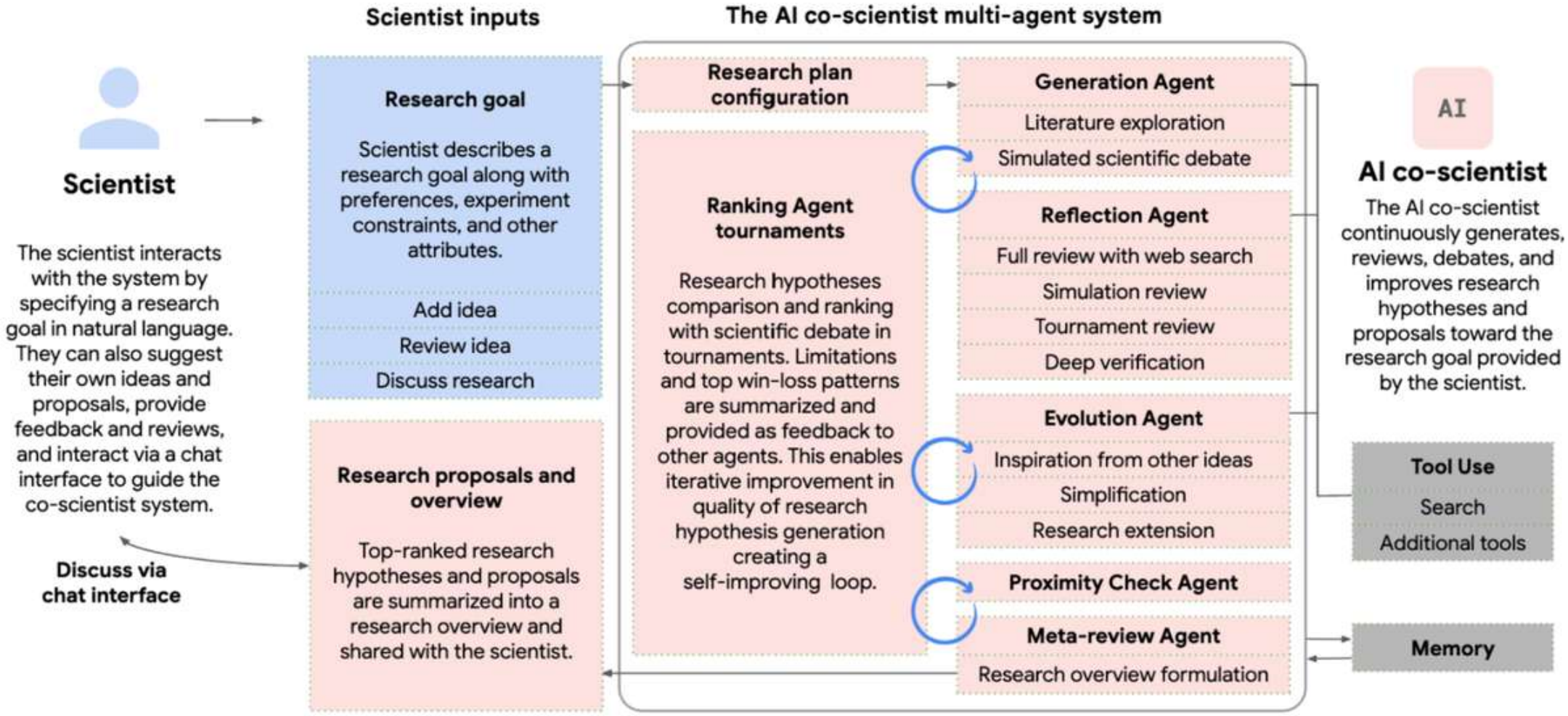
OpenAI Deep Research



https://openai.com/index/introducing-deep-research/

Funded by the European Union

Google AI Scientist

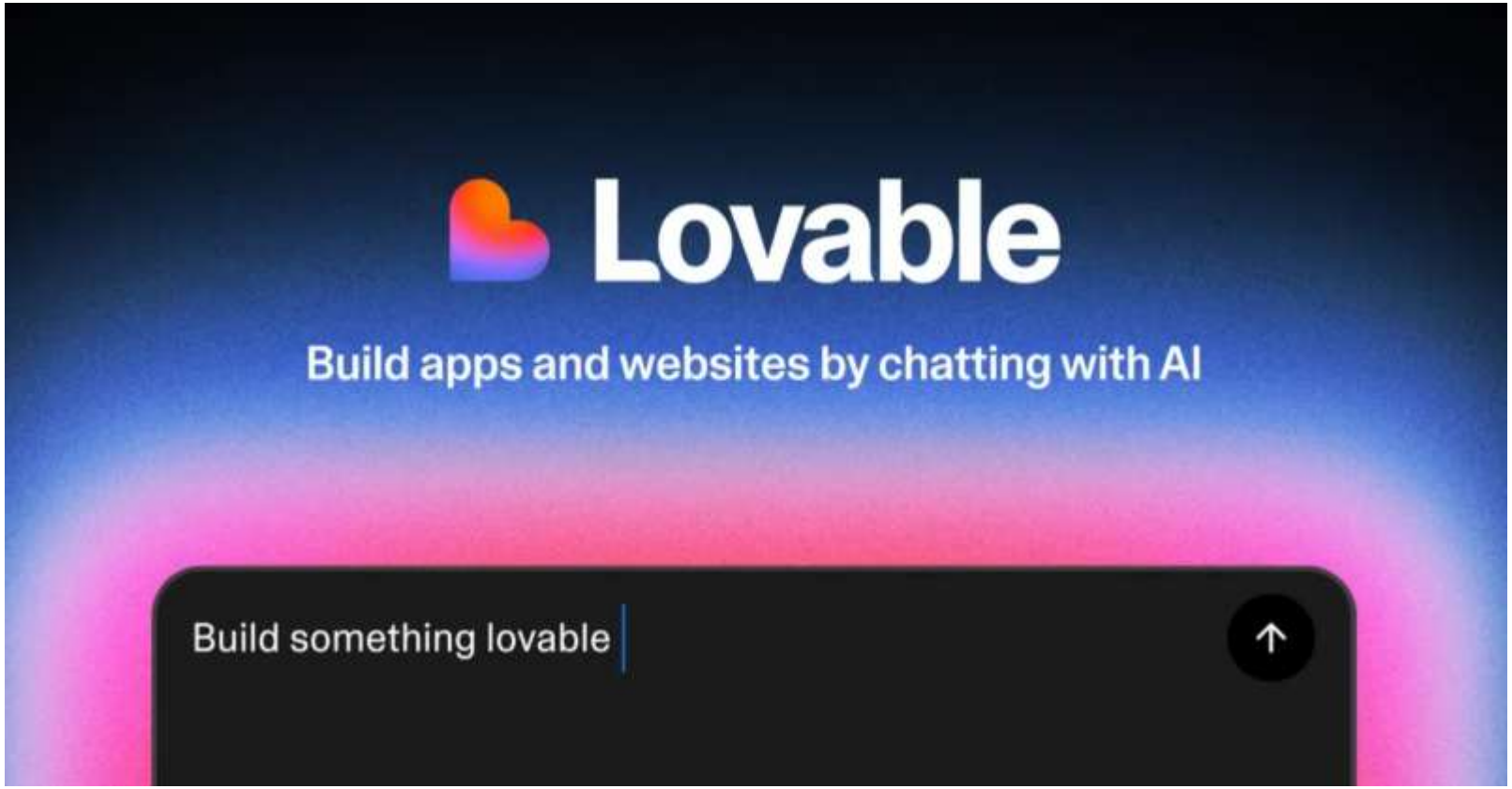


The AI co-scientist system design

https://research.google/blog/accelerating-scientific-breakthroughs-with-an-ai-co-scientist/

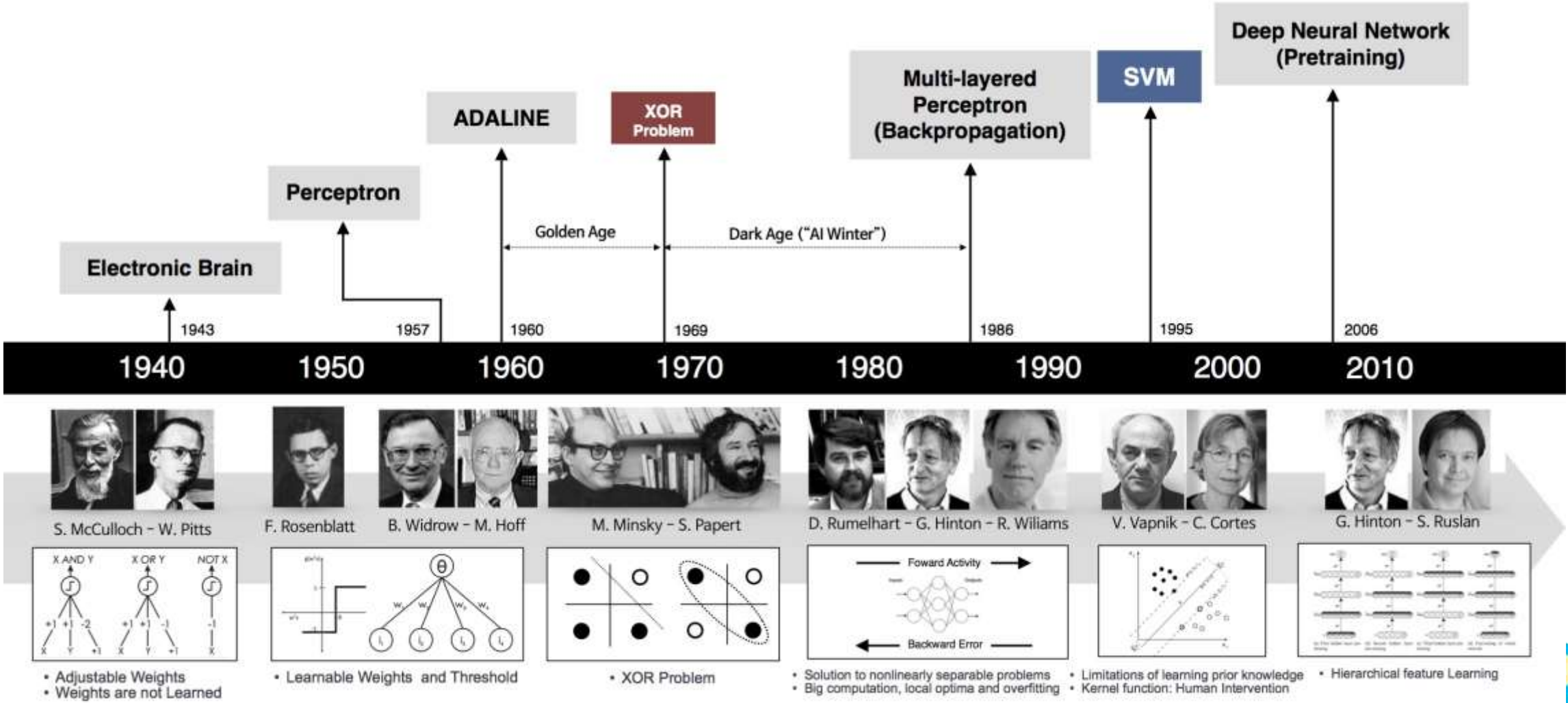
Funded by the European Union

Coding Assistants and Vibe Coding



Funded by the European Union

Builds on Long-Term Basic Research

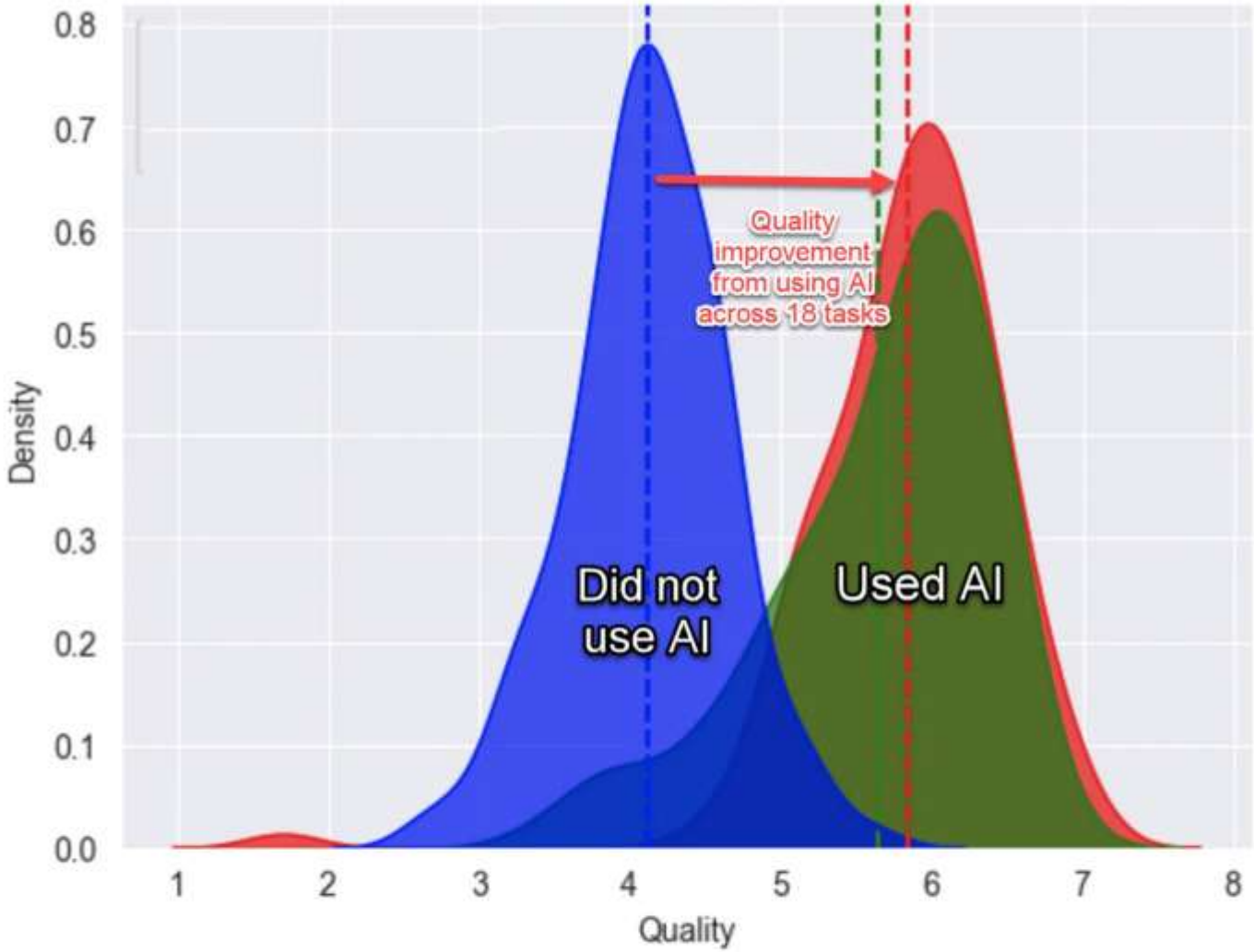


https://beamandrew.github.io/deeplearning/2017/02/23/deep_learning_101_part1.html

Funded by the European Union

AI and Future of Work

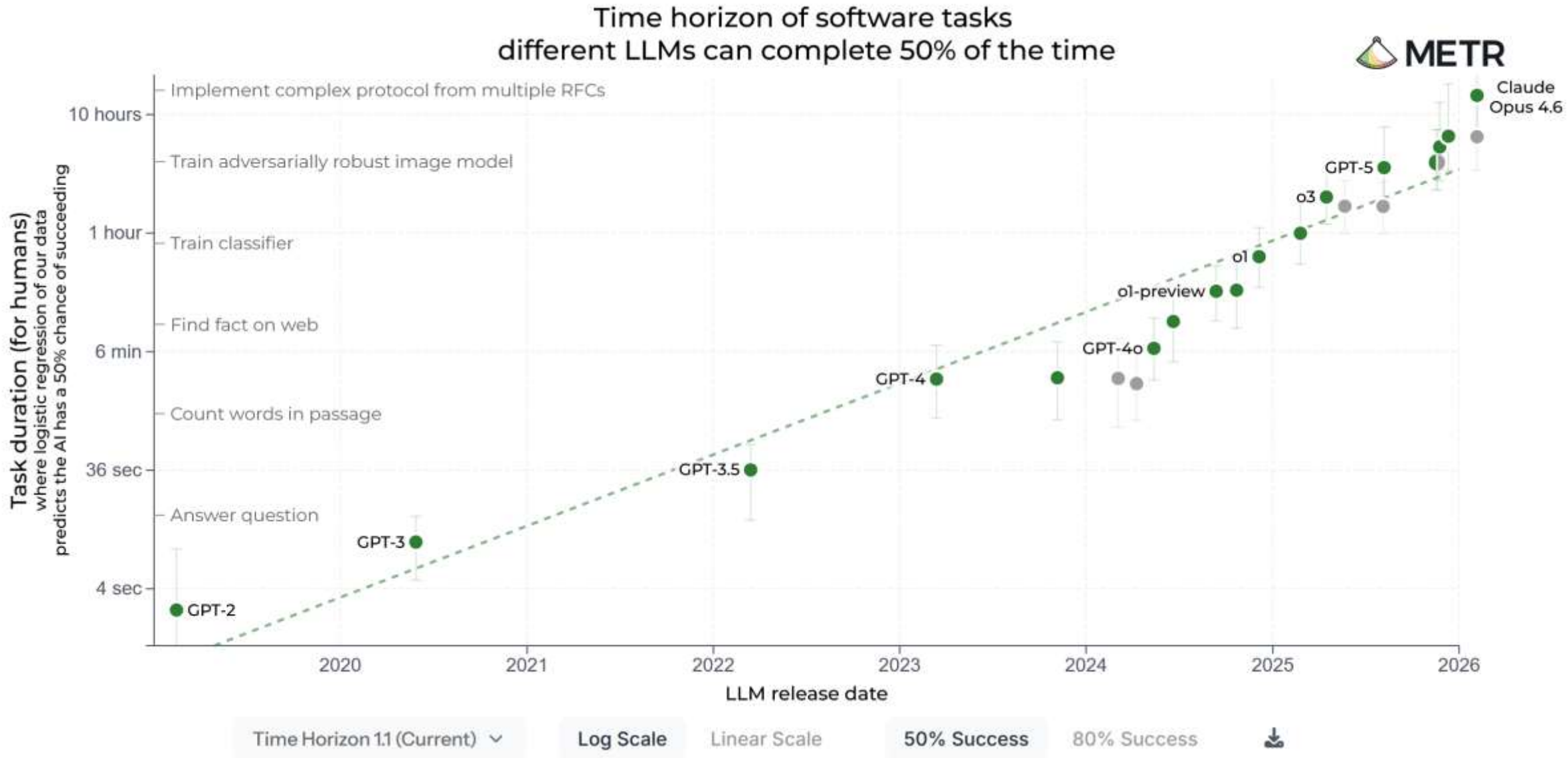
- **12% more tasks finished**
- **25% quicker completion**
- **40% higher quality**



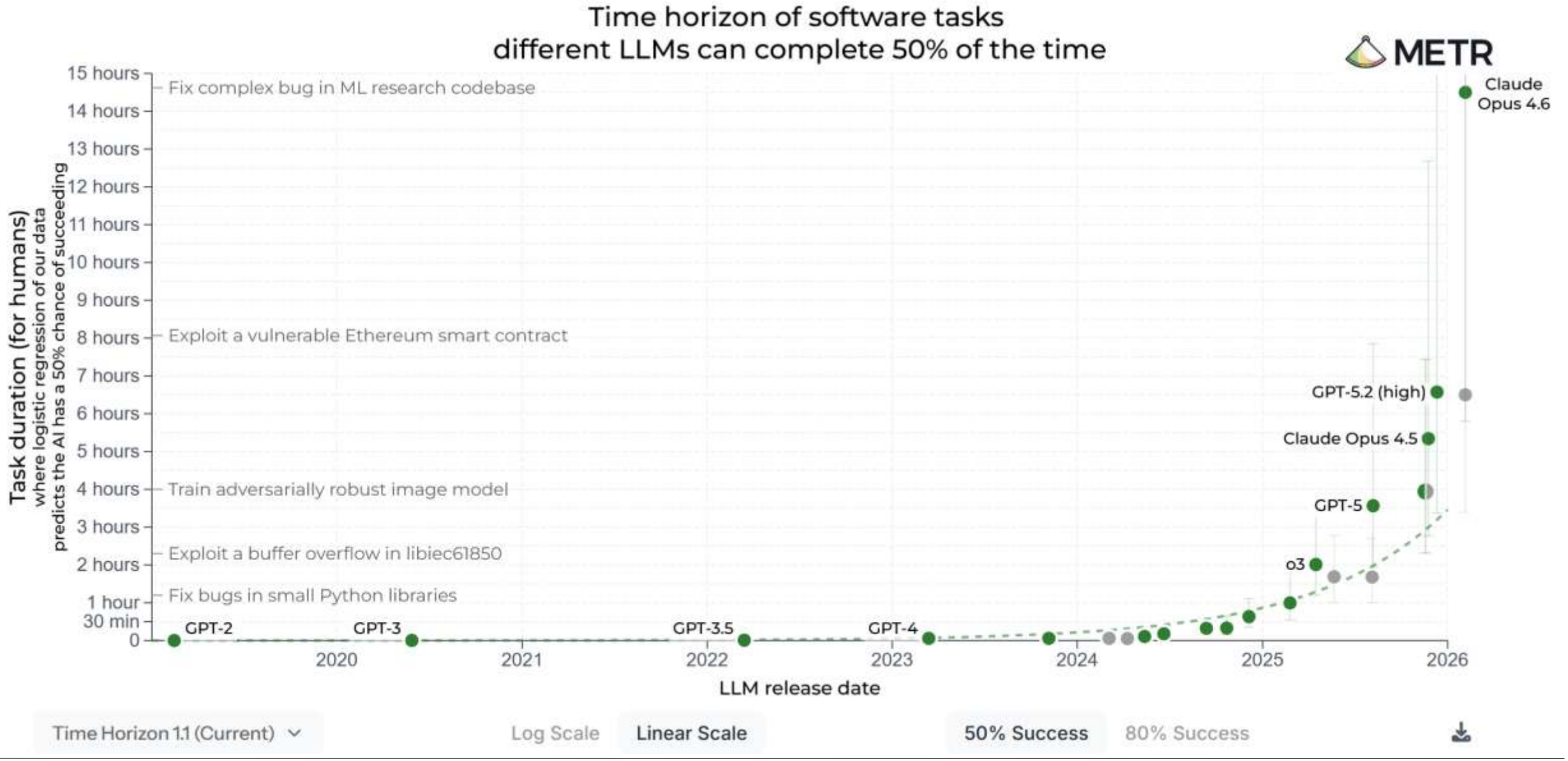
Distribution of output quality across all the tasks. The blue group did not use AI, the green and red groups used AI, the red group got some additional training on how to use AI.

https://www.oneusefulthing.org/p/centaurs-and-cyborgs-on-the-jagged

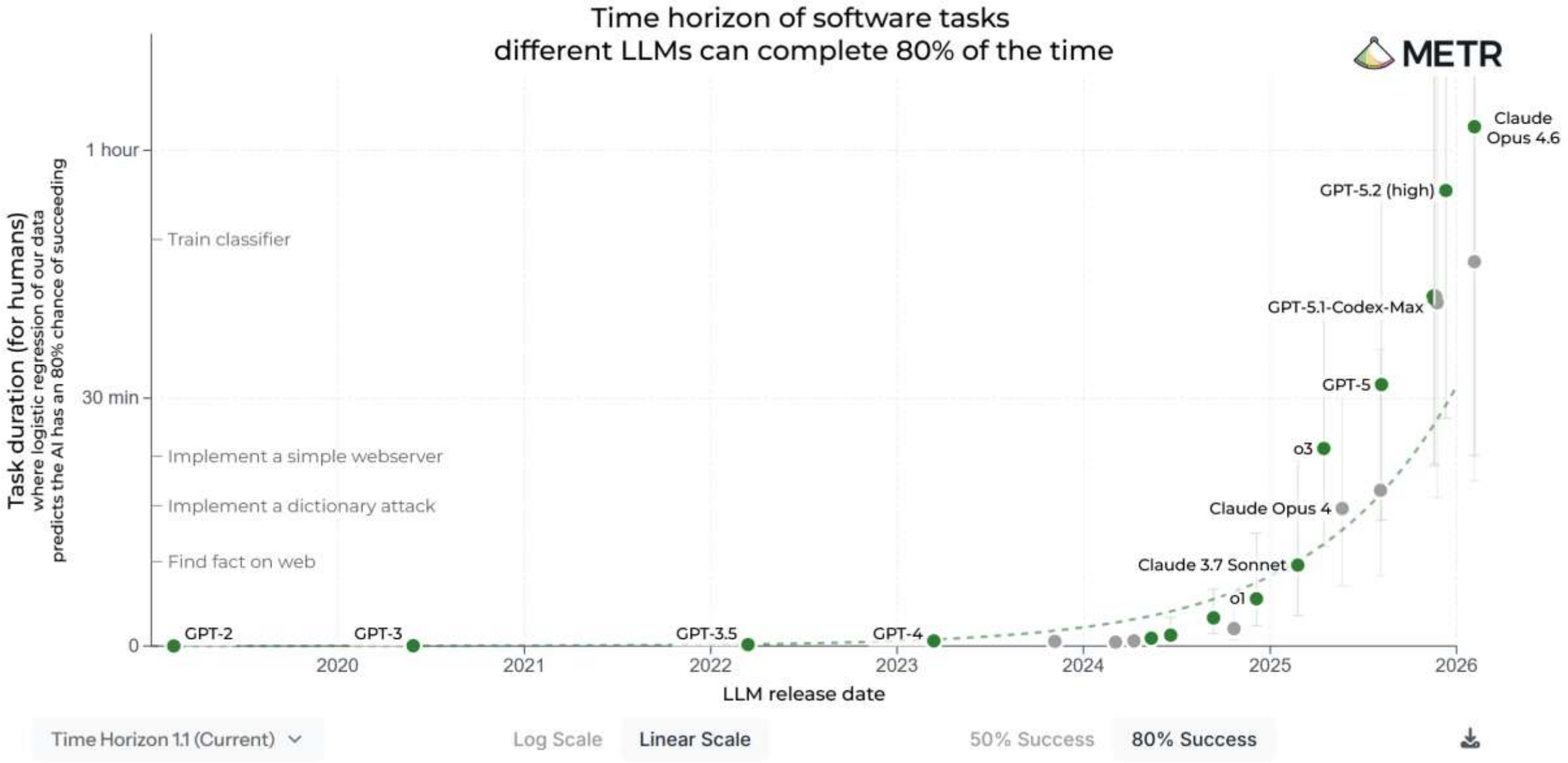
HBS Working Paper 24-013 “Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality” by F. Dell'Acqua et al.



https://metr.org/time-horizons/



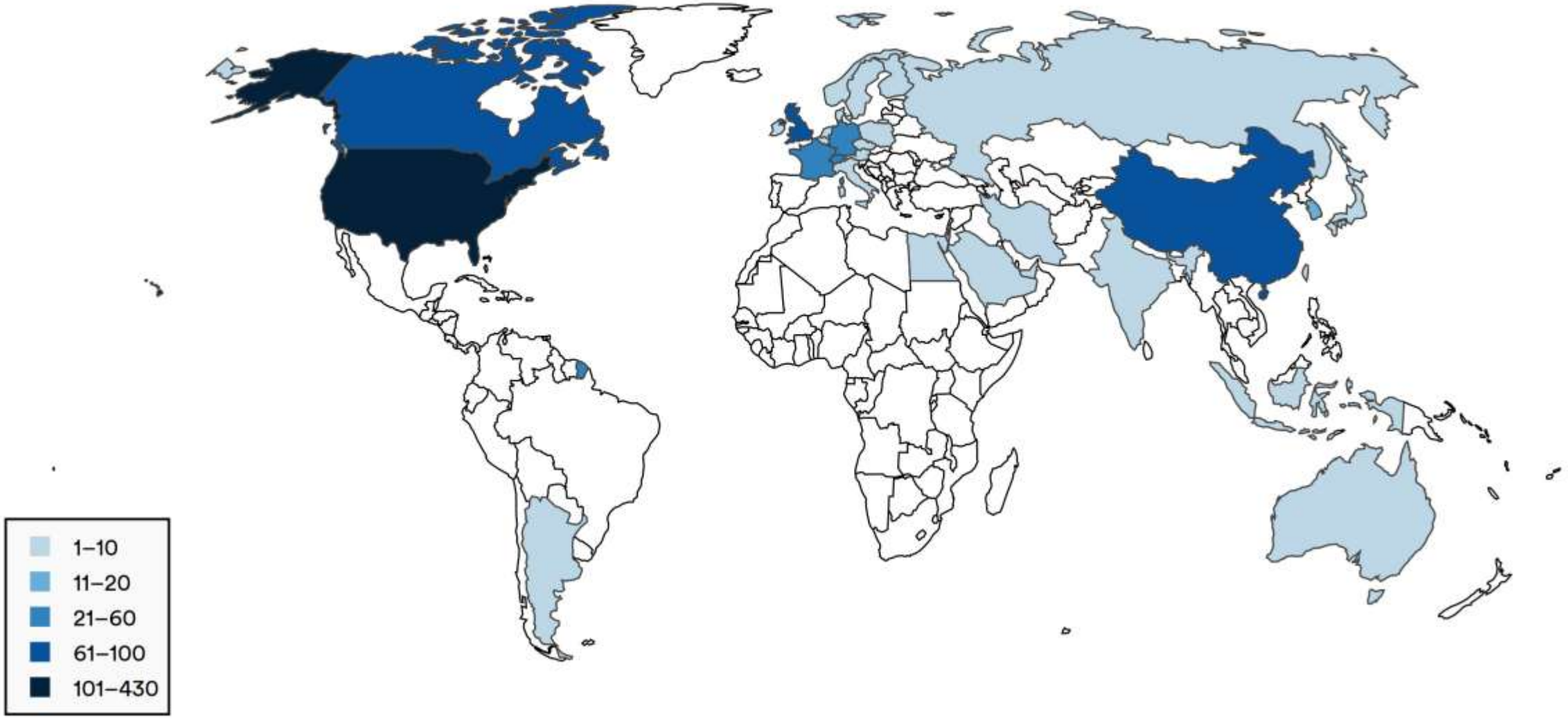
https://metr.org/time-horizons/



https://metr.org/time-horizons/

Number of notable machine learning models by geographic area, 2003–23 (sum)

Source: Epoch, 2023 | Chart: 2024 AI Index report



https://aiindex.stanford.edu/report/

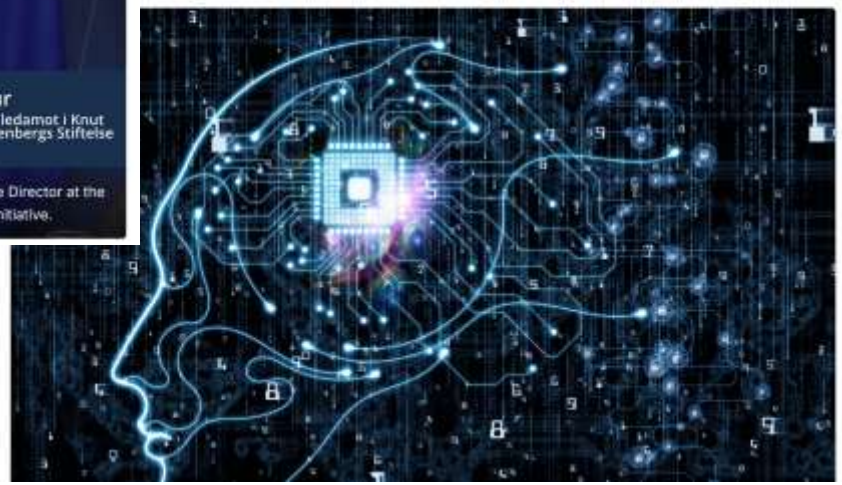
WARA-TRICS (AI Training and Inference Compute at Scale)

New initiative aims to give Sweden its own large-scale AI language model

February 23rd, 2026

WASP to Establish New AI Research Arena

February 19th, 2026



Advanced, large-scale AI is essential for addressing the major challenges many industries face today: increasing data volumes, more complex systems, and greater demands on computing resources. To meet these challenges, WASP is establishing a new research arena, AI Training and Inference Compute at Scale (WARA AI-TRICS). The initiative is driven by Sweden's largest national research program, the Wallenberg , Autonomous Systems and Software Program (WASP).

ithin WARA AI-TRICS, researchers and companies will collaborate on the most essing AI challenges facing Swedish industry. Three key areas have been identified as rticularly important:

- AI foundation models for industry and society
- AI at Scale, applying AI to massive industrial datasets
- New methods, such as physics-informed AI

https://wasp-sweden.org/wasp-to-establish-new-ai-research-arena/

Course Overview

Course Aim

The aim of the course is to **explain** how the **methods** and **techniques** used by **large generative AI models** such as large language models work and **explore** how to **build** them.

The **focus** is on the **technical aspects** such as architectures, methods and techniques.

The course is thus **more** about **machine learning than** about **natural language processing**.

Course Goal

- Knowledge and understanding
  - Explain the **technical underpinnings** of **large language models**.
  - Explain the processes involved in **training** a **large language model**.
- Competence and skills
  - **Implement** and **train** a basic **large language model** from scratch in PyTorch.
  - **Read** and **comprehend** recent **academic papers** on LLMs and have knowledge of the common terms used in them (alignment, scaling laws, RLHF, prompt engineering, instruction tuning, etc.).
- Judgement and approach
  - **Understand** and **discuss concepts** and **terminology** of state-of-the-art LLMs.
  - Develop an **ability** to **distinguish fact from fantasy** in this fast-moving field.

Tentative Course Lectures

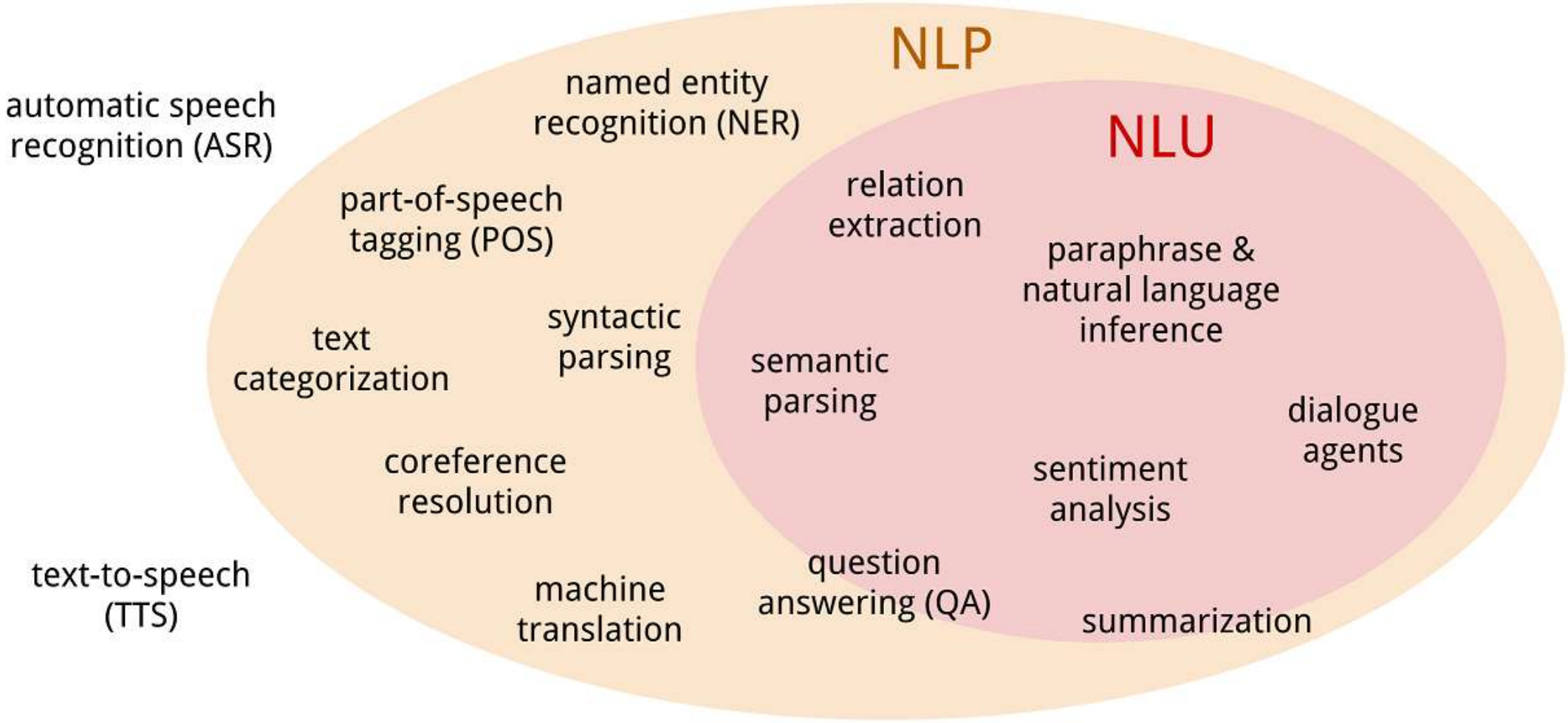
- 2/3 LE1 – Introduction, NLP and Large Language Models
- 9/3 LE2 – Basics (learning distributions, sequence to sequence mappings and embeddings)
- 16/3 LE3 – Data curation and processing
- 23/3 LE4 – Pre-training and scaling
- 30/3 LE5 – Fine-tuning, aligning and distillation
- 20/4 LE6 – Inference, in context learning and retrieval augmented generation
- 27/4 LE7 – Benchmarking and evaluation
- 4/5 LE8 – Building LLMs in practice part 1
- 11/5 LE9 – Building LLMS in practice part 2
- 18/5 LE10 – Advanced topics (trustworthiness, reasoning, multi-modal models, world models)

Examination

- Develop an LLM from scratch and conduct experiments related to the lectures and write a short report on it. There are four parts:
  1. Develop a simple data pre-processing pipeline
  2. Pre-train a GPT-style LLM
  3. Fine-tune the LLM
  4. Evaluate the LLM
- For each part, implement a solution and perform (at least) two experiments. Analyse the results and write a short report.
- You may work individually or in groups of two. I recommend trying to do as much as possible on your own, while discussing with your partner.

Natural Language Processing

Natural Language Processing and Understanding



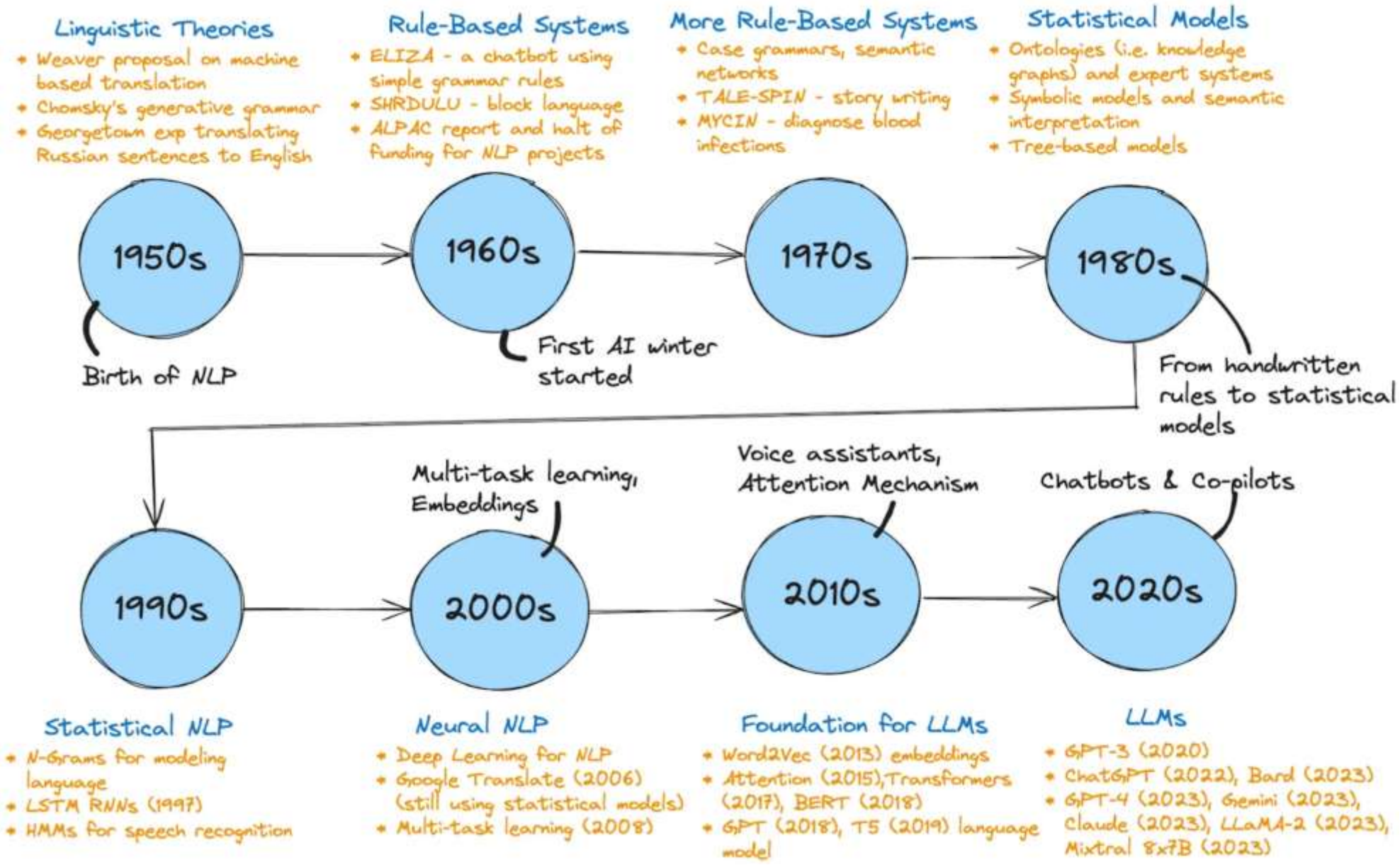
https://devopedia.org/natural-language-processing

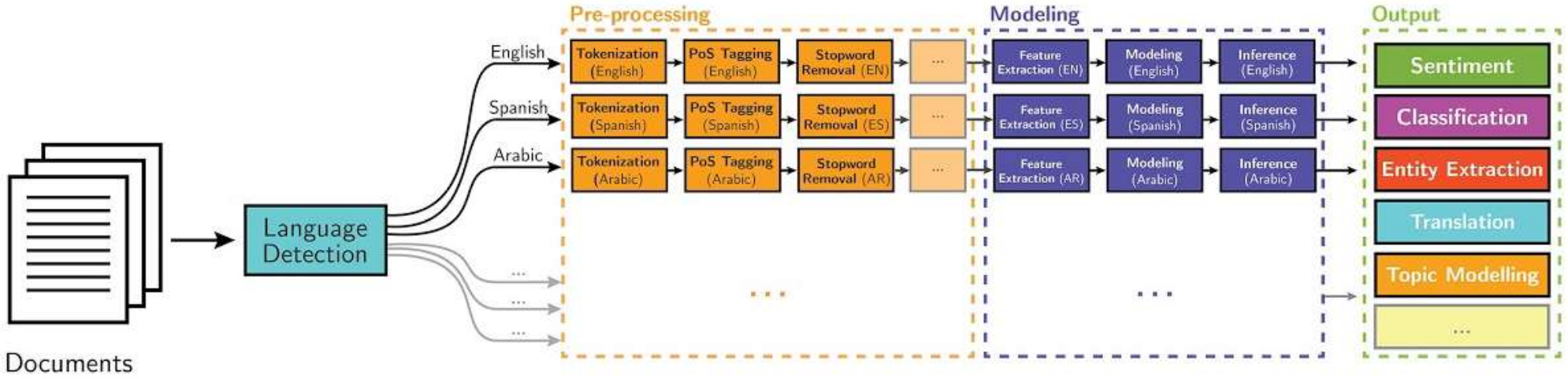
Funded by the European Union

Natural Language Understanding

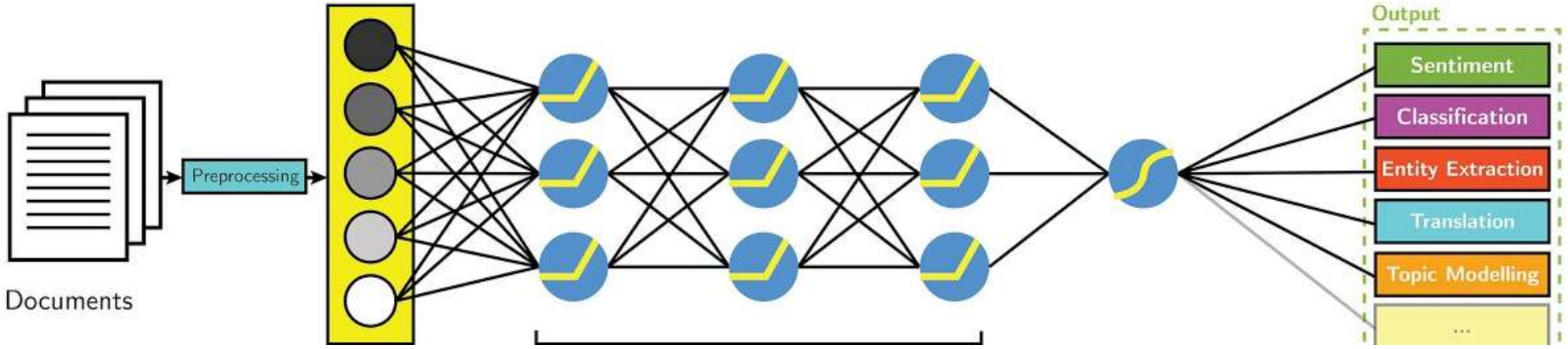
https://devopedia.org/natural-language-processing

Funded by the European Union


Linguistic Theories
• Weaver proposal on machine based translation
• Chomsky's generative grammar
• Georgetown exp translating Russian sentences to English
Rule-Based Systems
• ELIZA - a chatbot using simple grammar rules
• SHRDULU - block language
• ALPAC report and halt of funding for NLP projects
More Rule-Based Systems
• Case grammars, semantic networks
• TALE-SPIN - story writing
• MYCIN - diagnose blood infections
Statistical Models
• Ontologies (i.e. knowledge graphs) and expert systems
• Symbolic models and semantic interpretation
• Tree-based models
1950s
1960s
1970s
1980s
Birth of NLP
First AI winter started
From handwritten rules to statistical models
1990s
2000s
2010s
2020s
Multi-task learning, Embeddings
Voice assistants, Attention Mechanism
Chatbots & Co-pilots
Statistical NLP
• N-Grams for modeling language
• LSTM RNNs (1997)
• HMMs for speech recognition
Neural NLP
• Deep Learning for NLP
• Google Translate (2006) (still using statistical models)
• Multi-task learning (2008)
Foundation for LLMs
• Word2Vec (2013) embeddings
• Attention (2015), Transformers (2017), BERT (2018)
• GPT (2018), T5 (2019) language model
LLMs
• GPT-3 (2020)
• ChatGPT (2022), Bard (2023)
• GPT-4 (2023), Gemini (2023), Claude (2023), LLaMA-2 (2023), Mixtral 8x7B (2023)



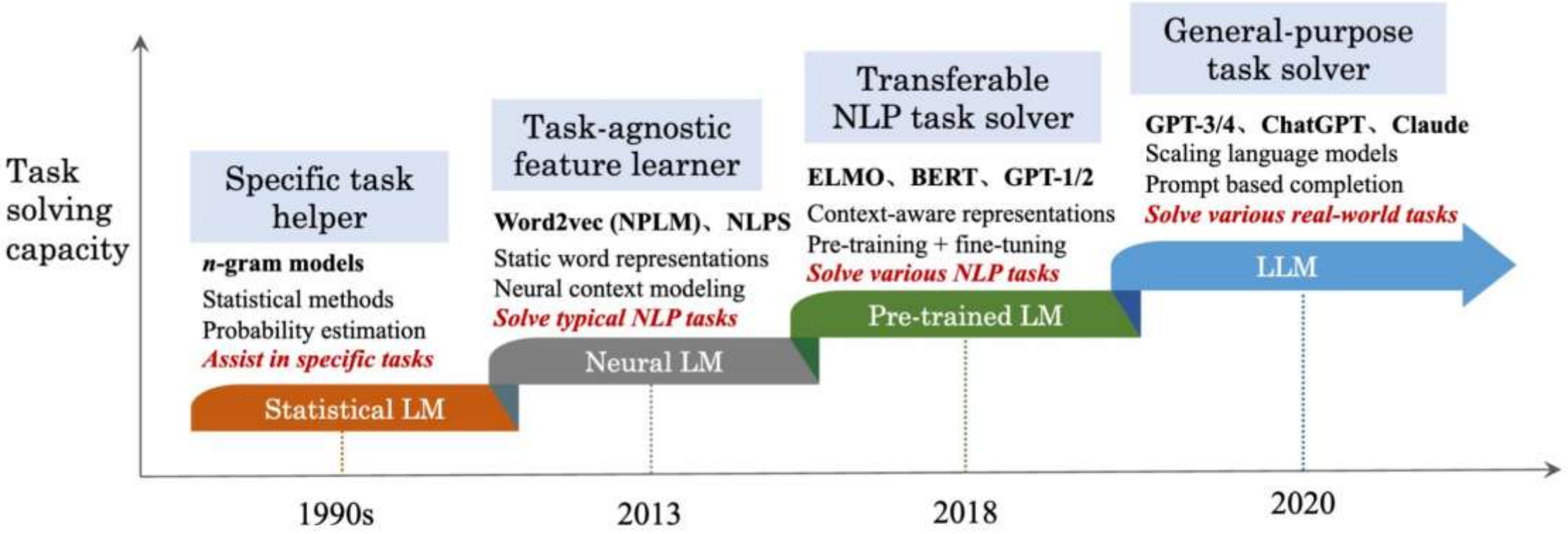
Deep Learning-based NLP



https://devopedia.org/natural-language-processing

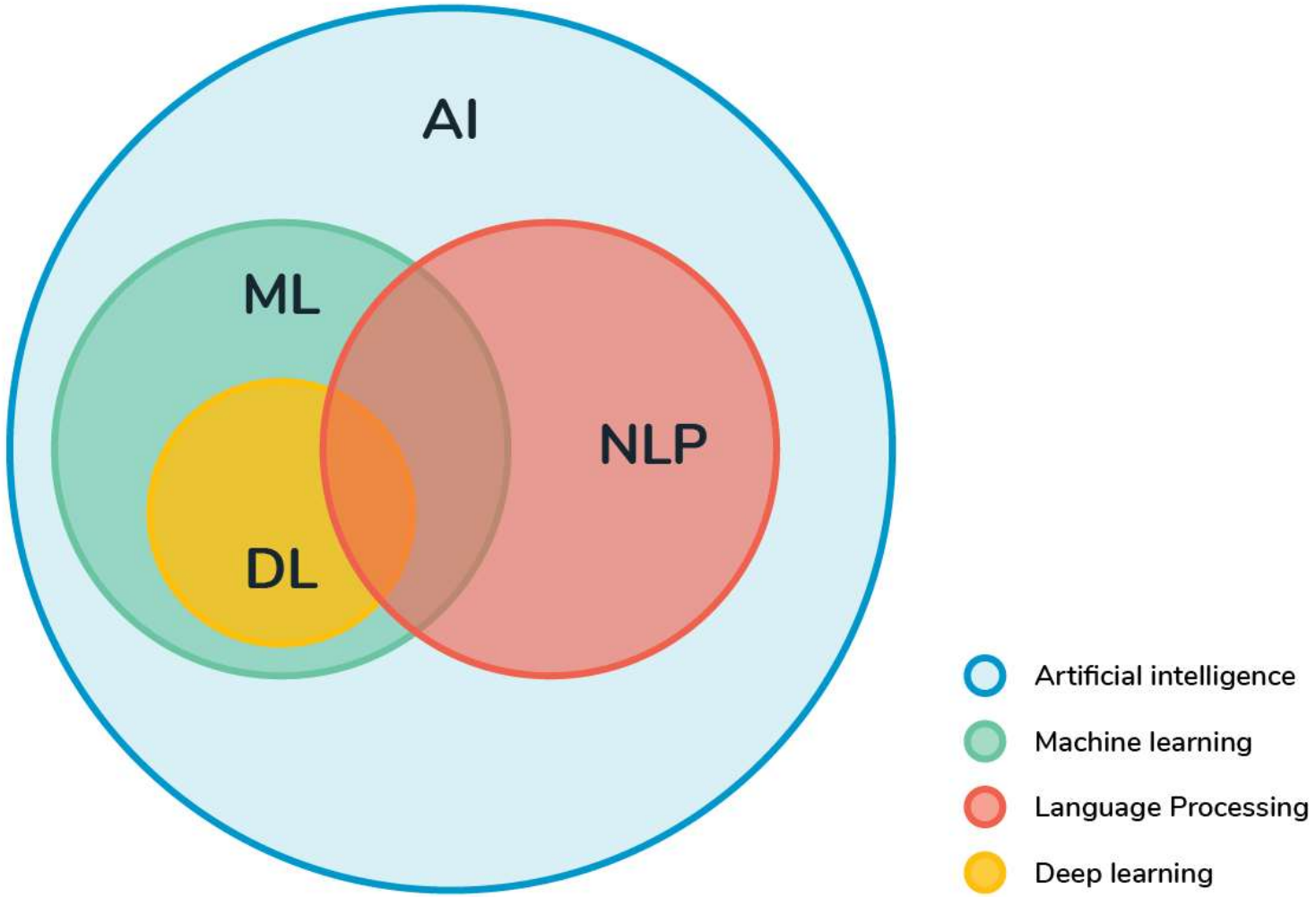
Funded by the European Union

Evolution of LMs for Task-Solving Capacity



W. Zhao et al. A Survey of Large Language Models. 2023.

Funded by the European Union

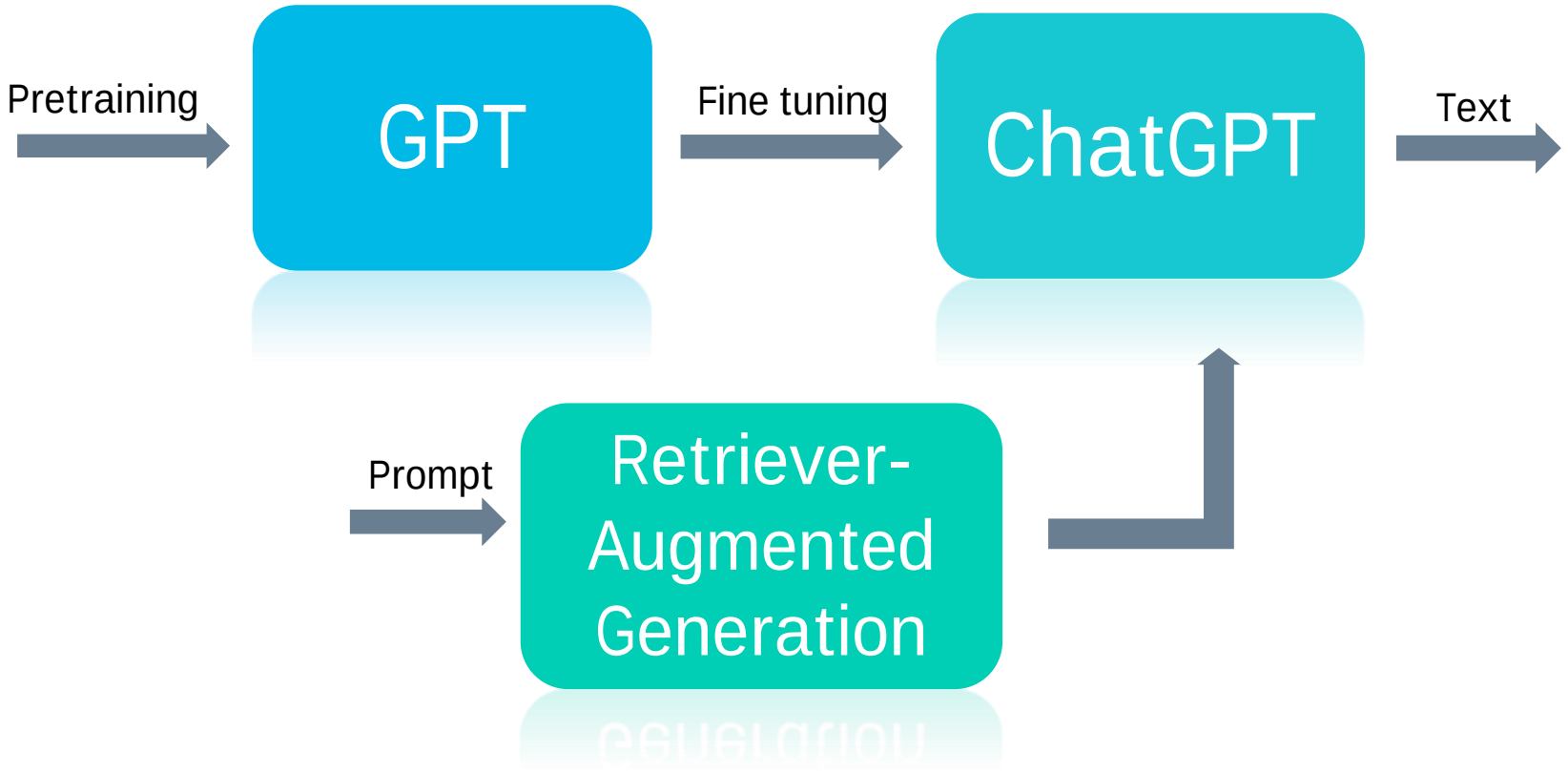


https://devopedia.org/natural-language-processing

Funded by the European Union

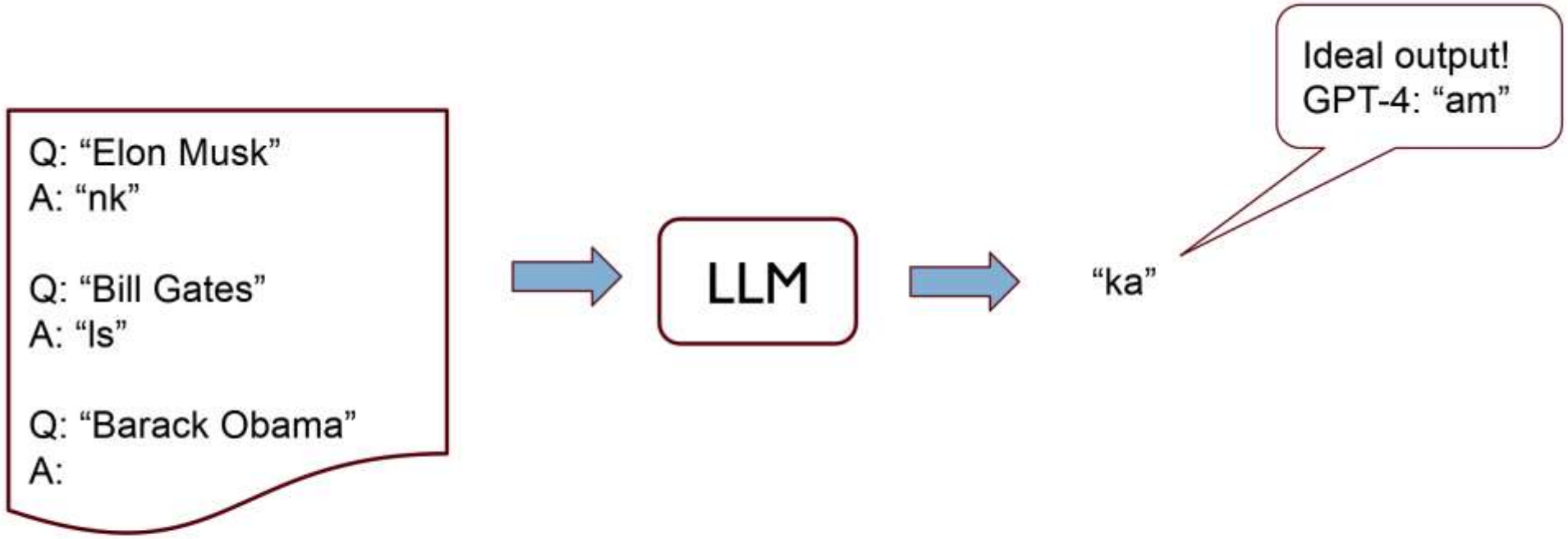
Introduction to Large Language Models

How Does ChatGPT Work?



Funded by the European Union

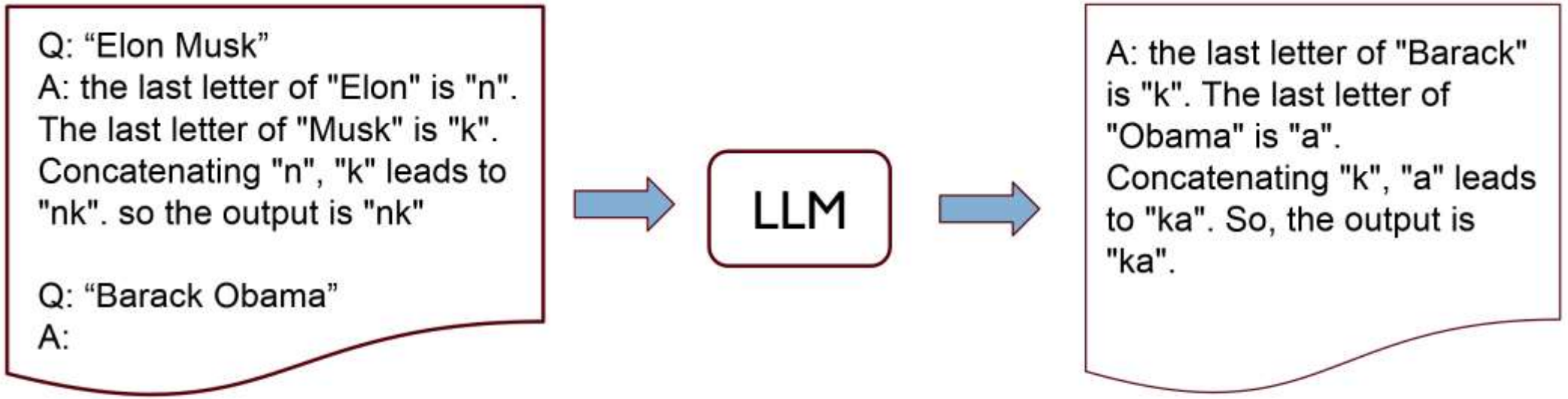
Few Shot Prompting



T. Brown et al. Language Models are Few-Shot Learners. NeurIPS 2020.

Funded by the European Union

Chain of Thought Prompting

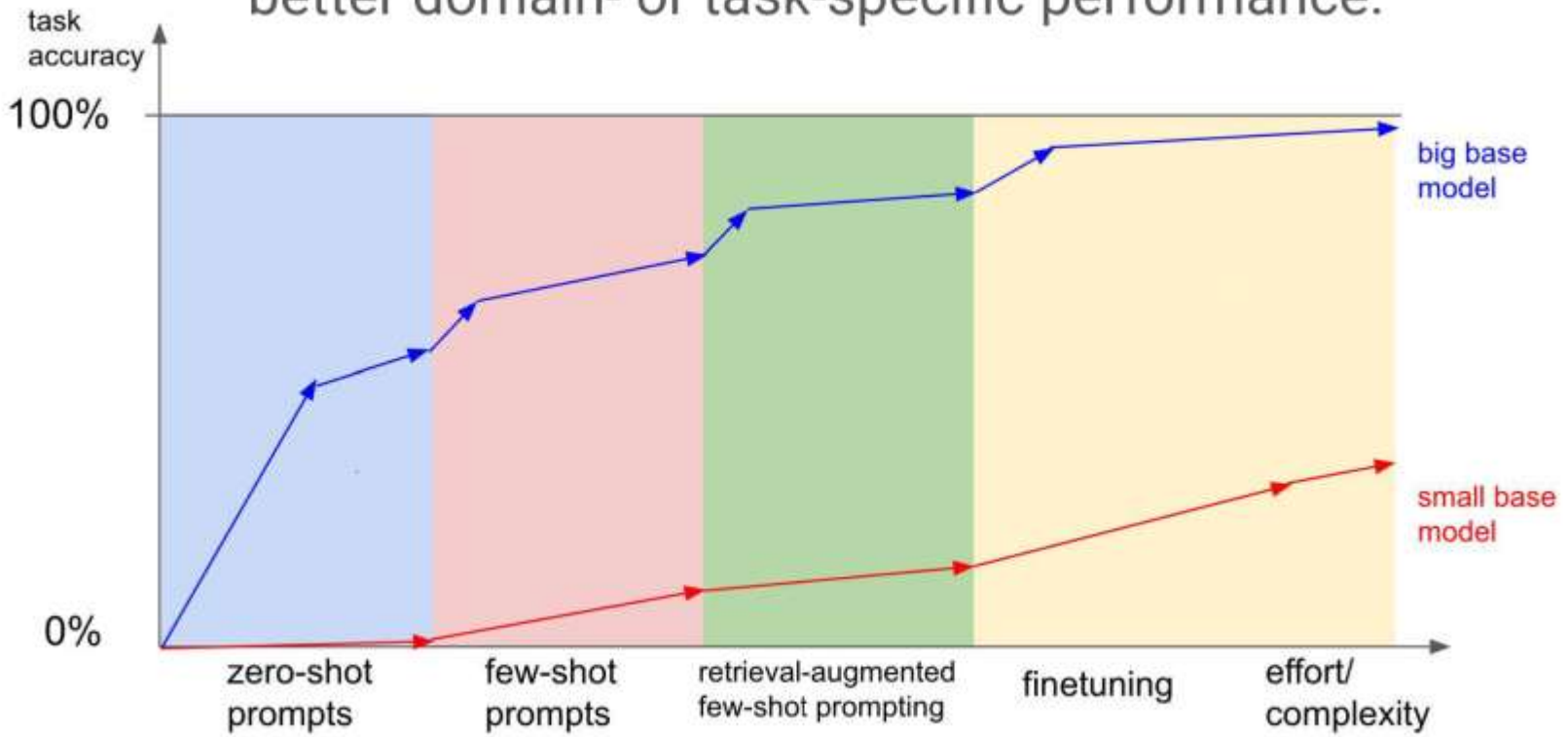


J. Wei et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS 2022.

Funded by the European Union

From Prompting to Fine-Tuning

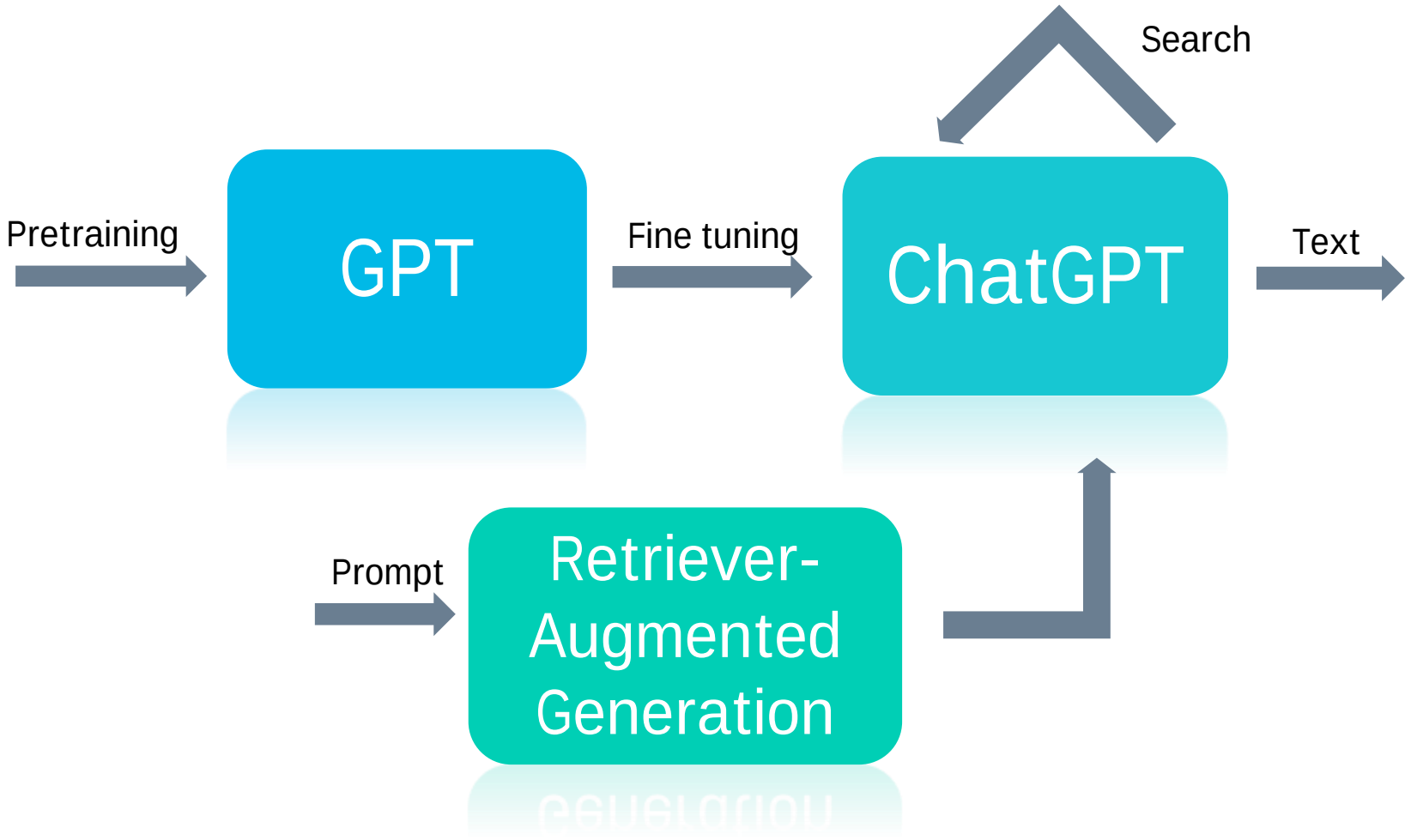
Unlike prompting, fine-tuning actually changes the model under the hood, giving better domain- or task-specific performance.



Source: Andrej Karpathy @karpathy (not to scale)

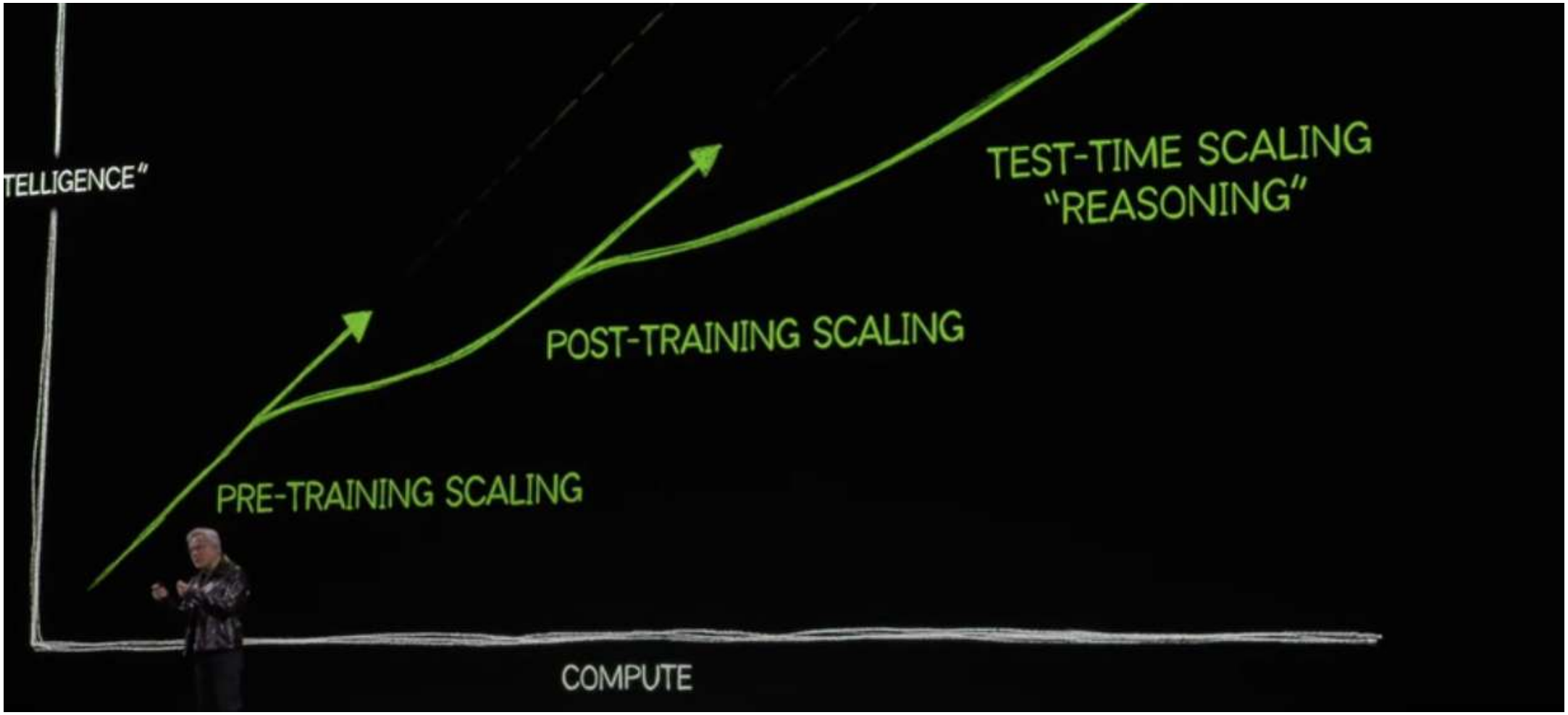
Funded by the European Union

How Does ChatGPT-o1 Work?



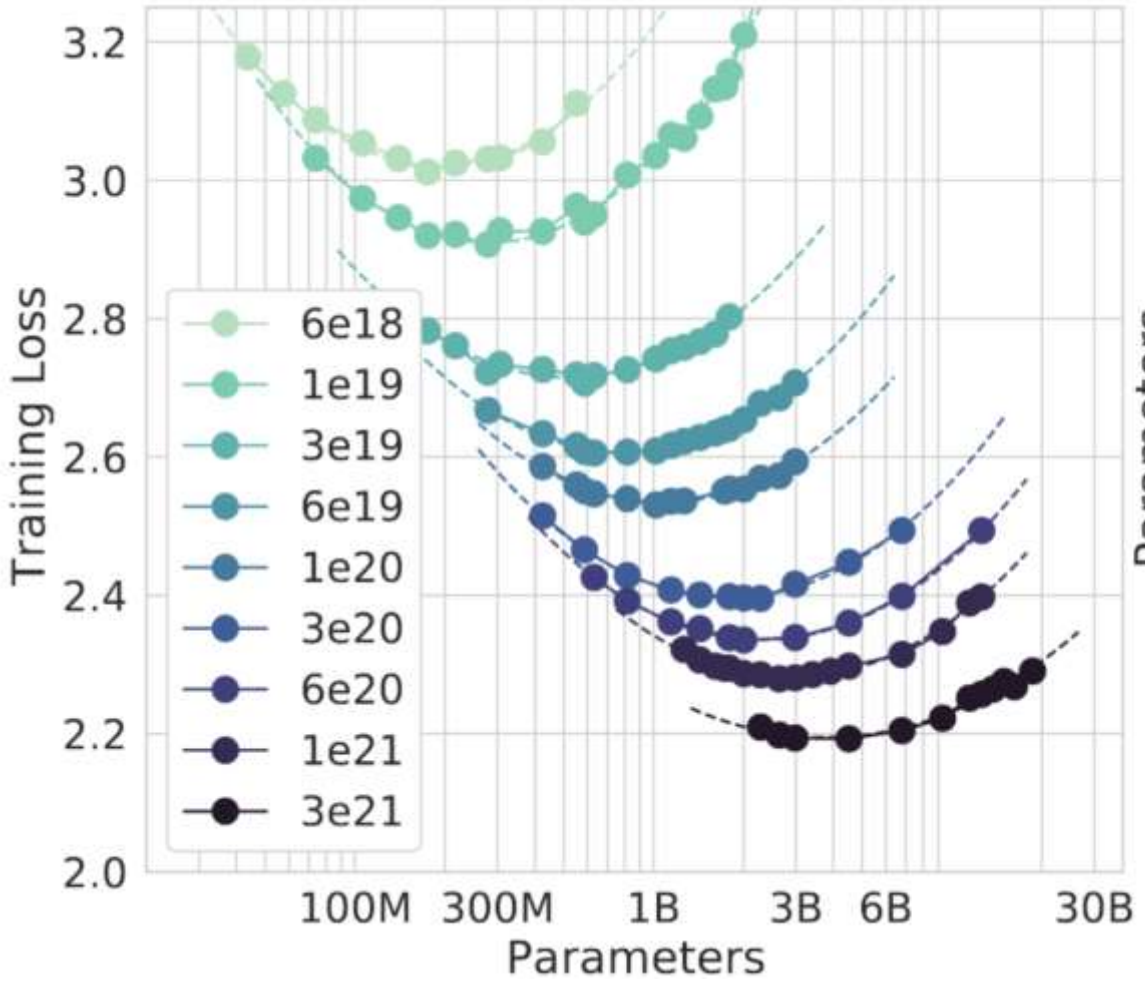
Funded by the European Union

Scaling – From Pre-training to Test-time



Funded by the European Union

Large Language Models Scaling Laws



Parameters

Parameters	FLOPs	FLOPs (in *Gopher* unit)	Tokens
400 Million	1.92e+19	1/29,968	8.0 Billion
1 Billion	1.21e+20	1/4,761	20.2 Billion
10 Billion	1.23e+22	1/46	205.1 Billion
67 Billion	5.76e+23	1	1.5 Trillion
175 Billion	3.85e+24	6.7	3.7 Trillion
280 Billion	9.90e+24	17.2	5.9 Trillion
520 Billion	3.43e+25	59.5	11.0 Trillion
1 Trillion	1.27e+26	221.3	21.2 Trillion
10 Trillion	1.30e+28	22515.9	216.2 Trillion

Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270 Billion

https://www.lesswrong.com/posts/midXmMb2Xg37F2Kgn/new-scaling-laws-for-large-language-models

Funded by the European Union

Data Pre-Processing Pipeline

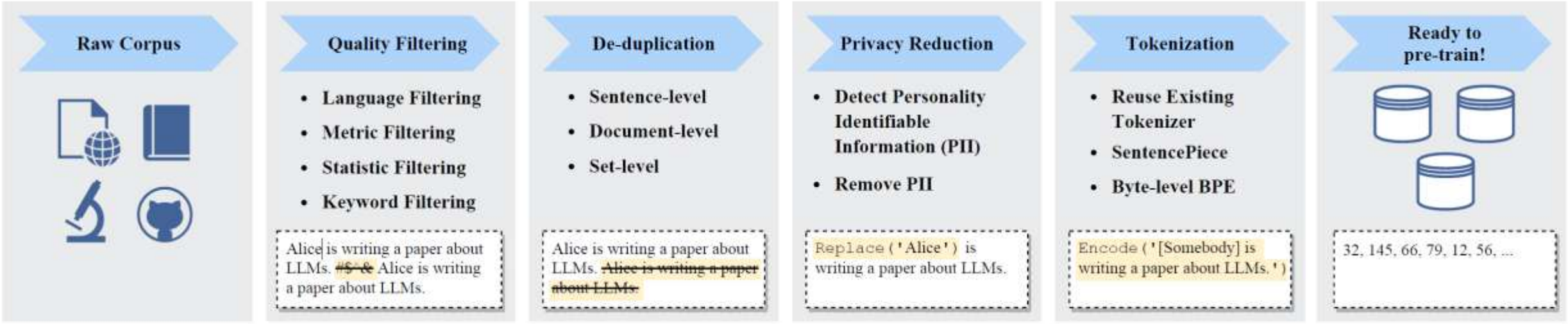
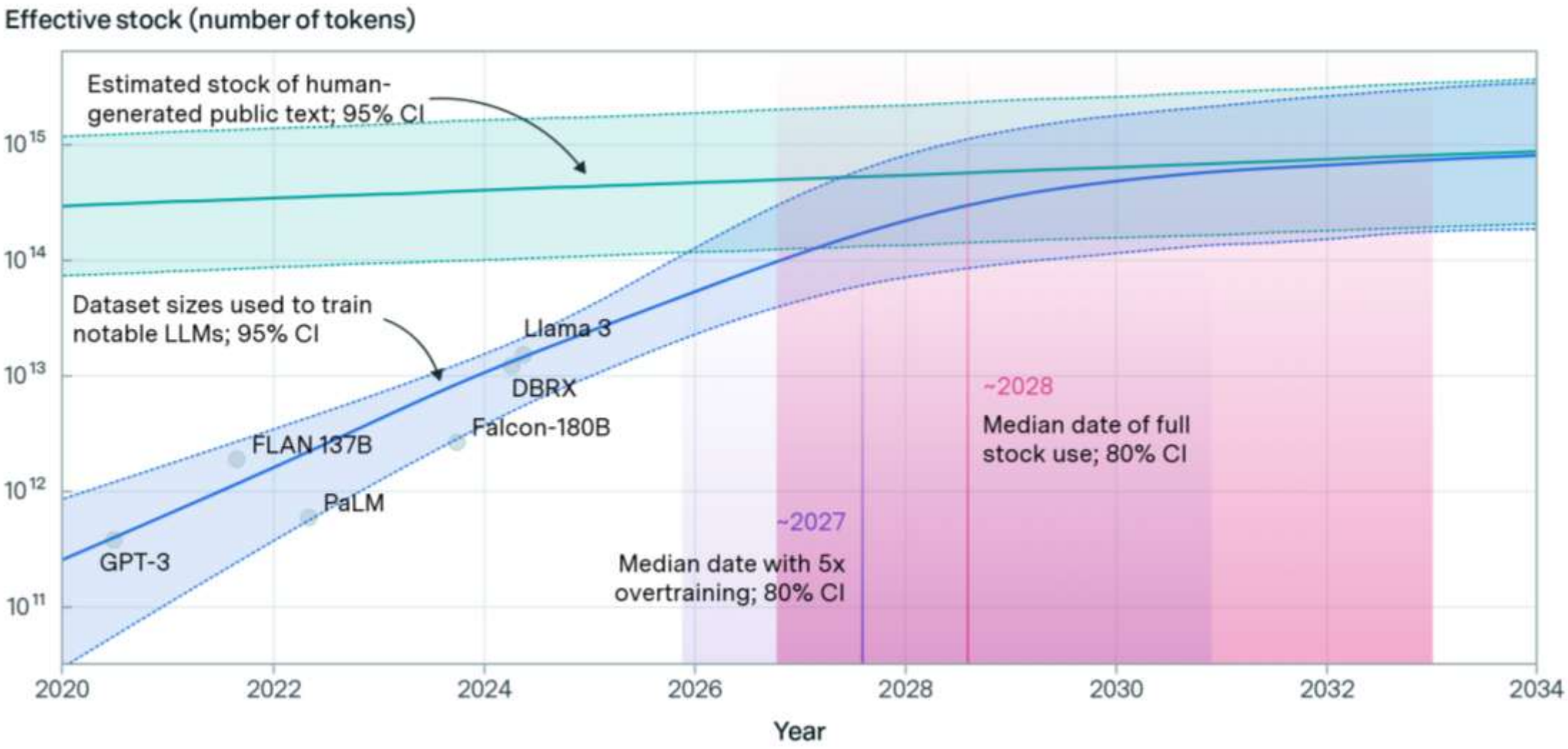


Fig. 6: An illustration of a typical data preprocessing pipeline for pre-training large language models.

Projections of the stock of public text and data usage

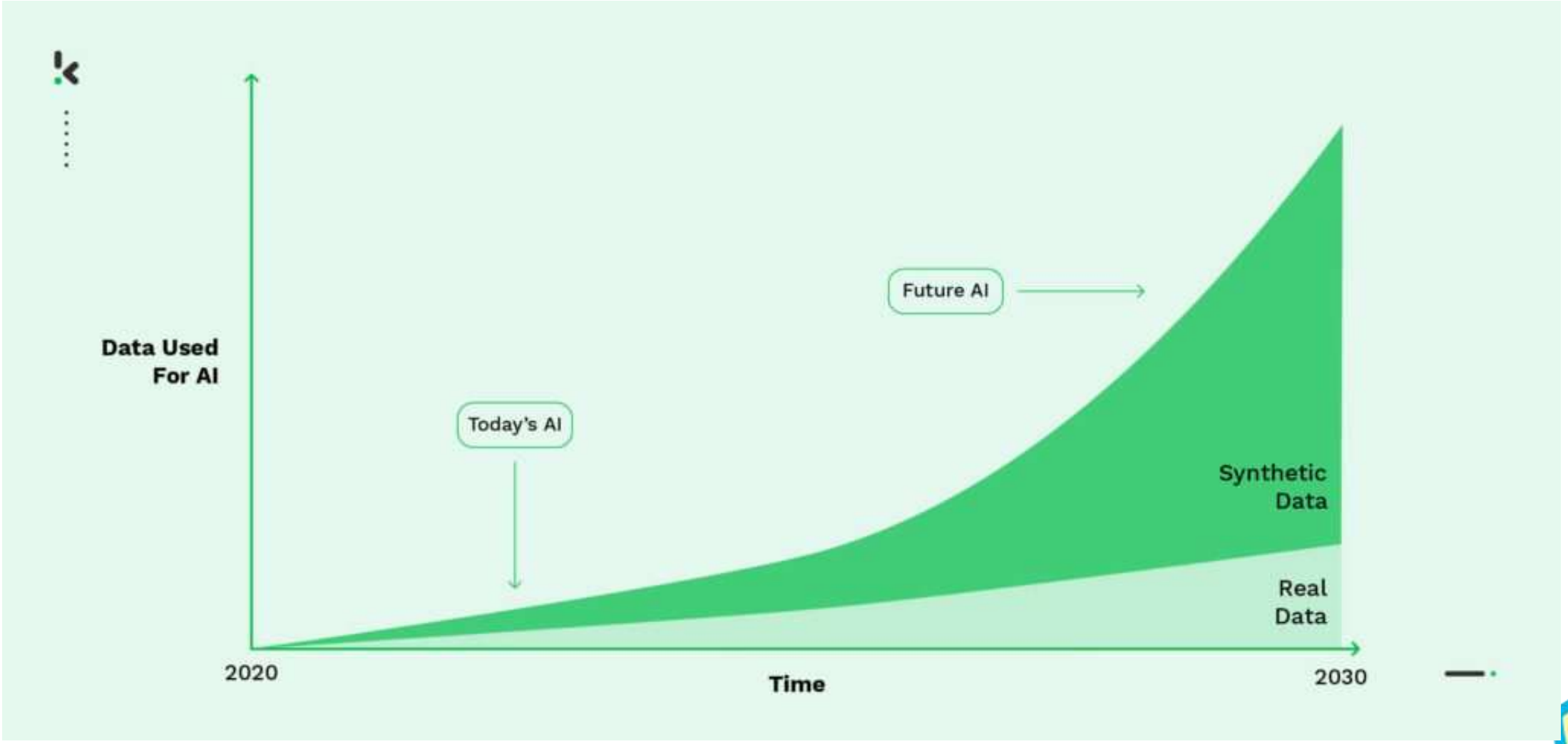
EPOCH AI



https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data

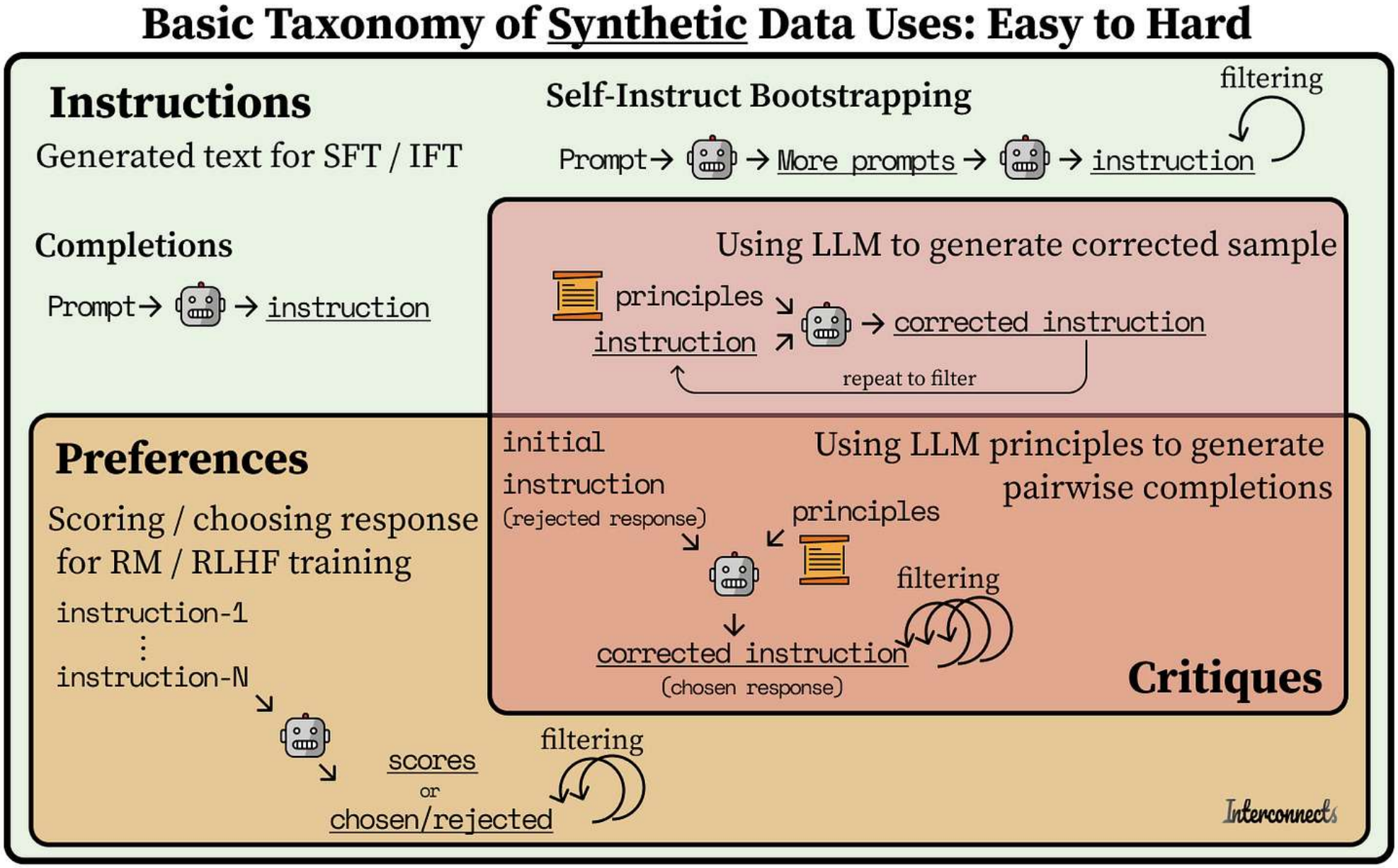
Funded by the European Union

Synthetic Data



https://datanorth.ai/blog/what-is-synthetic-data

Funded by the European Union



https://www.interconnects.ai/p/llm-synthetic-data

Funded by the European Union

Pre-Training Data Quality Reduces Compute Needs

Recent work finds smaller amounts of higher quality data removes the need for a larger model.

There is increasing evidence that efforts to better curate training corpus, including **deduping, pruning data and investing in synthetic data** can compensate for the need for larger networks and/or improve training dynamics.

	% train examples with dup in train	% train examples with dup in valid	% valid with dup in train
C4	3.04%	1.59%	4.60%
RealNews	13.63%	1.25%	14.35%
LM1B	4.86%	0.07%	4.92%
Wiki40B	0.39%	0.26%	0.72%

Table 2: The fraction of examples identified by NEARDUP as near-duplicates.

Lee et al. 2022

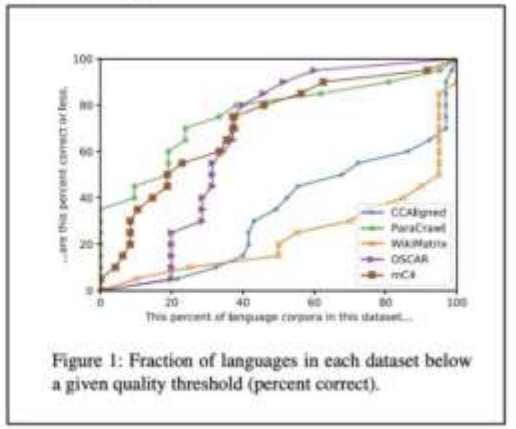
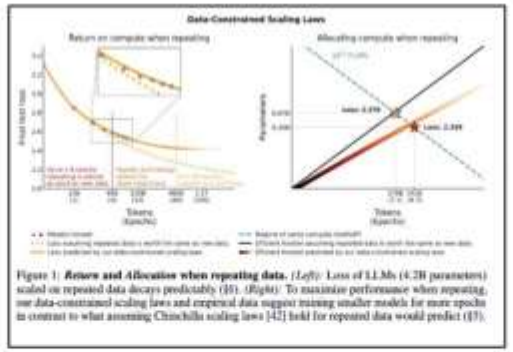
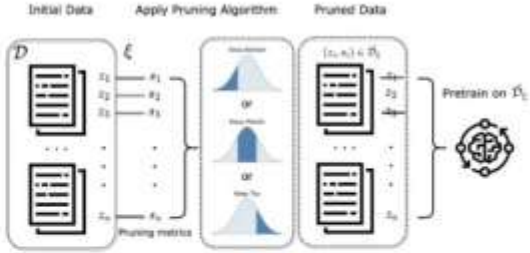


Figure 1: Fraction of languages in each dataset below a given quality threshold (percent correct).

Kreutzer at al. 2022



Muennighoff et al. 2023



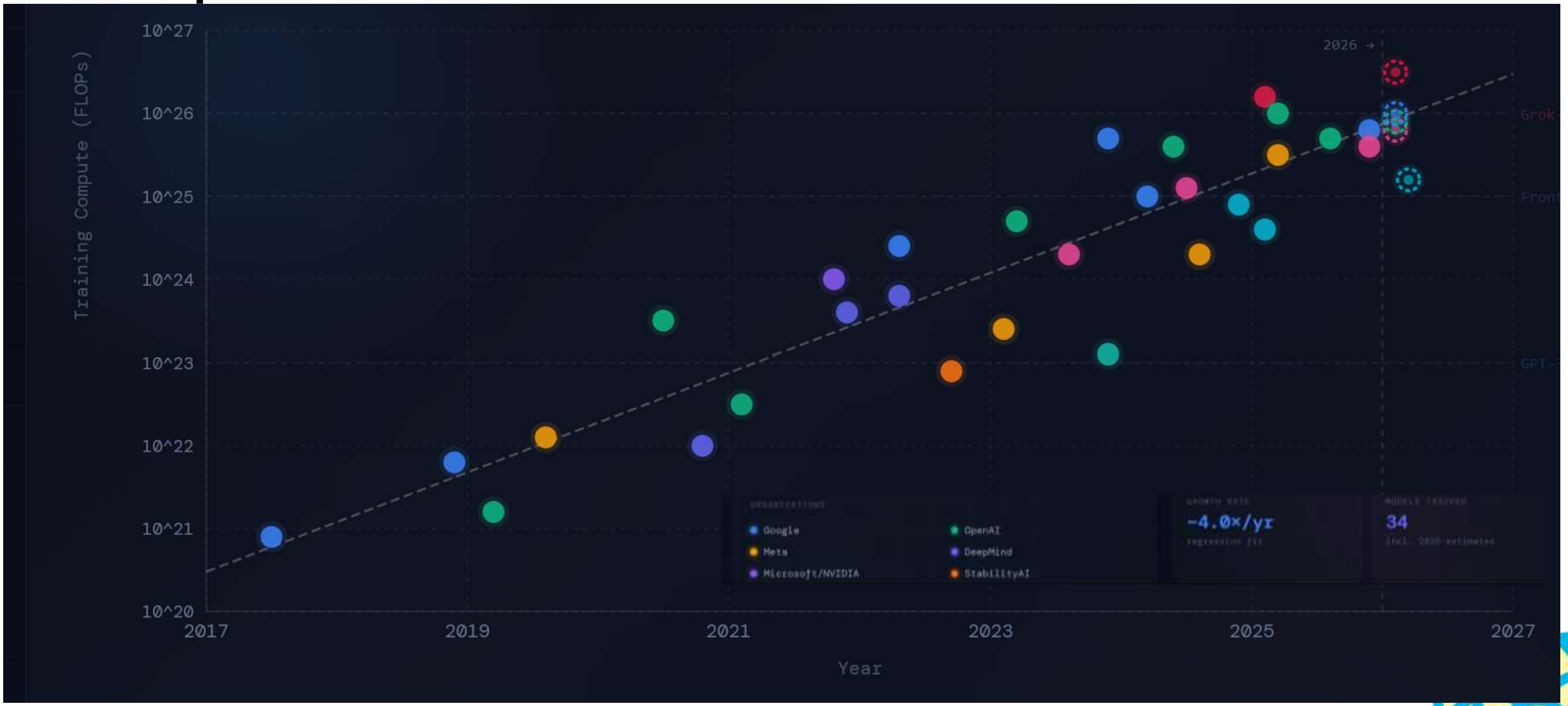
Marion et al. 2023

Cohere For AI

S. Hooker. On the Limitations of Compute Thresholds as a Governance Strategy. 2024.

Funded by the European Union

Compute to Train AI Models



Funded by the European Union

TrustLLM

Can you Trust ChatGPT? No!

- Very limited information about the training data
- It makes things up, with confidence (hallucinations)
- Even when there are references these may be false or not applicable
- Cannot count or draw logical conclusions
- Stuck in time and always changing
- *but, ChatGPT is still useful!*

Funded by the European Union

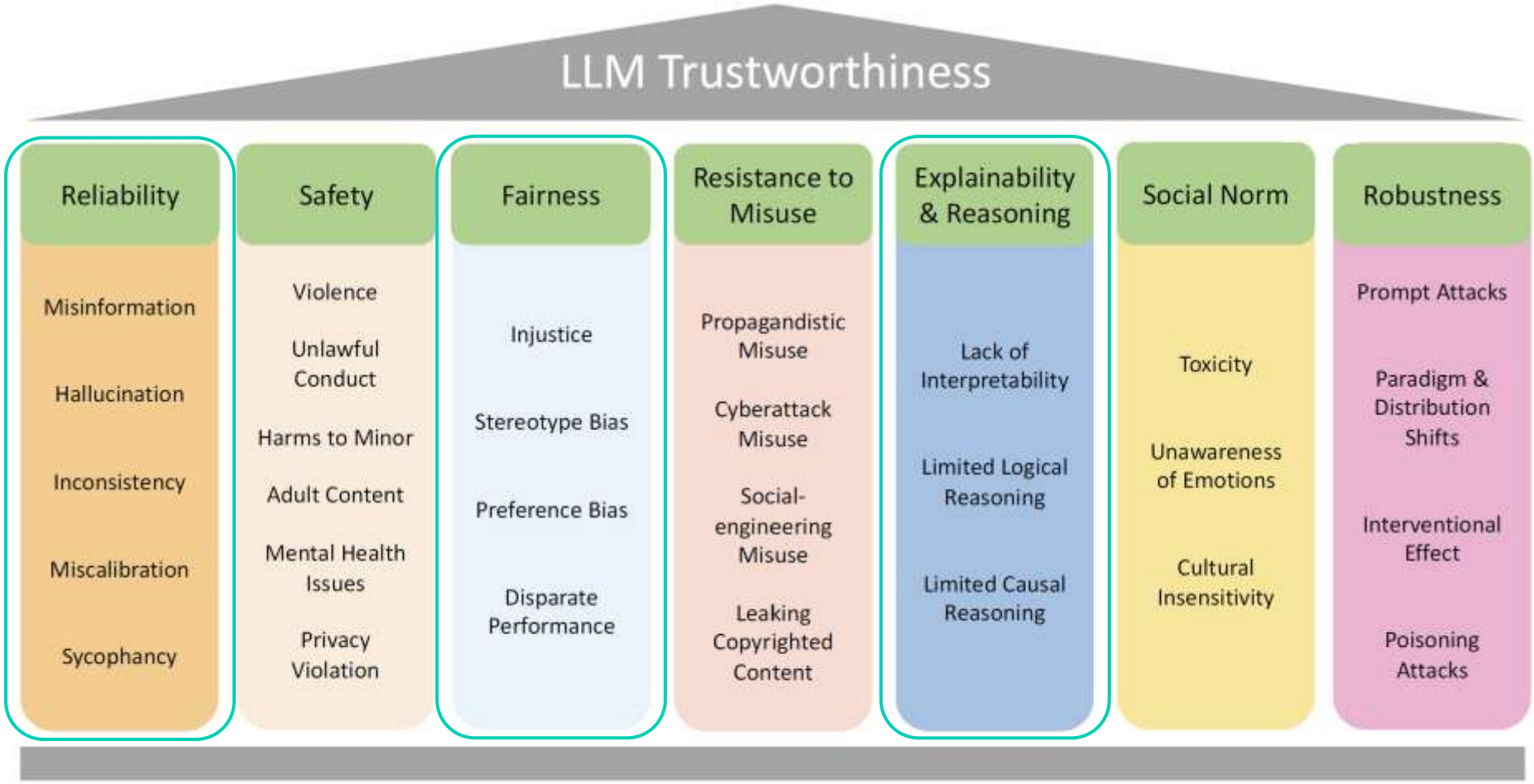
Coordinator: Fredrik.Heintz@liu.se ; Project manager: Tohid.Ardeshiri@liu.se

TrustLLM – Trustworthy and Factual LLMs made in Europe

- Develop an open, trustworthy, and sustainable LLM initially targeting the Germanic languages.
- TrustLLM will tackle the full range of challenges of LLM development,
  - from ensuring sufficient quality and quantity of multilingual training data,
  - to sustainable efficiency and effectiveness of model training,
  - to enhancements and refinements for factual correctness, transparency, and trustworthiness,
  - to a suite of holistic evaluation benchmarks validating the multi-dimensional objectives.

LINKÖPING UNIVERSITY
MIÐEIND
ALEXANDRA INSTITUTTET
Fraunhofer
UNIVERSITY OF ICELAND
UNIVERSITY OF COPENHAGEN
JÜLICH Forschungszentrum
TNO
KI BUNDESVERBAND
Lindholmen Science Park
AI SWEDEN
NTNU | Norwegian University of Science and Technology
Funded by the European Union

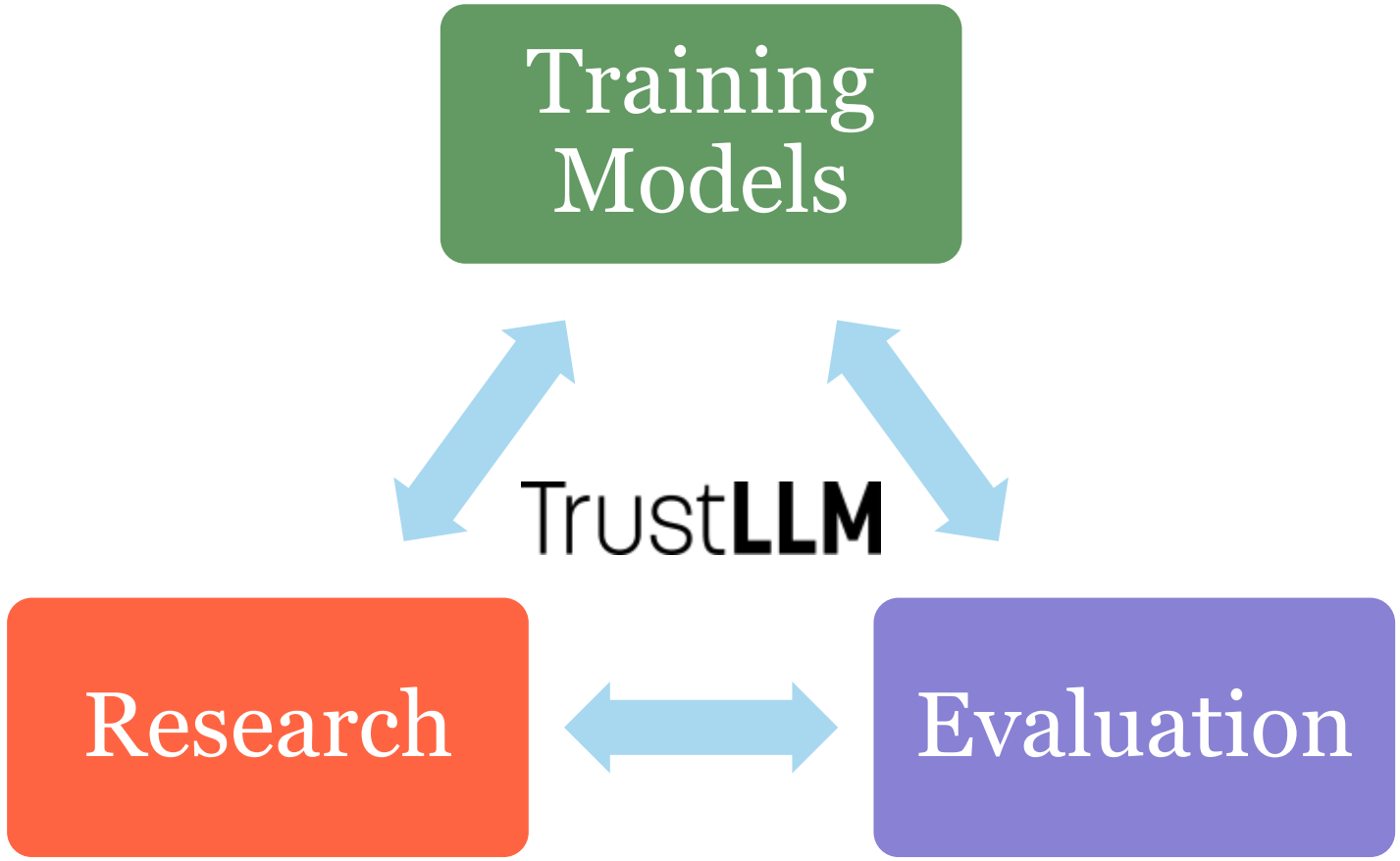
LLM Trustworthiness



Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment by Yang Liu Etal, 2023

Funded by the European Union

TrustLLM: Building Models and Doing Research



Funded by the European Union

High-Level TrustLLM Goal

At least one TrustLLM model every year that is increasingly powerful, factful, and trustworthy.

Funded by the European Union

Reaching the Goal

First step is creating a good baseline model
- Curate first version of dataset
- Implement first version software stack
- Learn to effectively training large models on EuroHPC
- Establish a strong baseline for future models

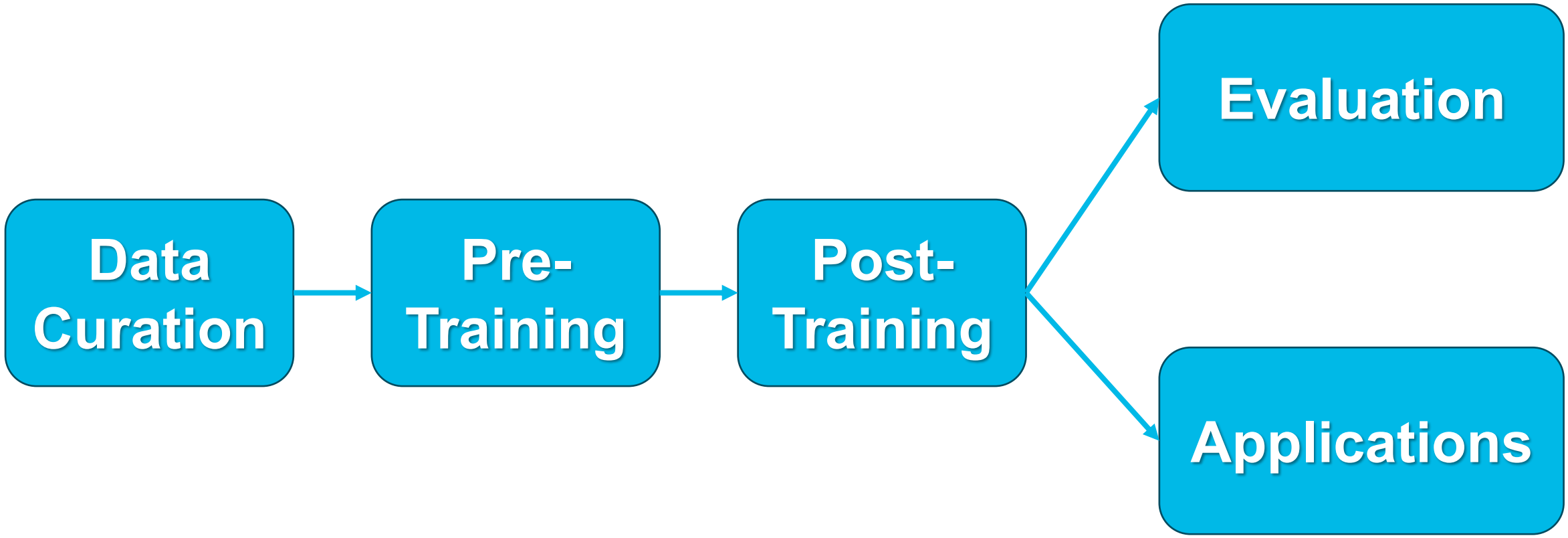
Rather than inventing and researching new things, we focus on doing engineering and using things we already know.

Then iterate new, improved model versions as we complete research, curate better data, and improve the software.

Identify what contributes to the overall factual and trustworthiness.

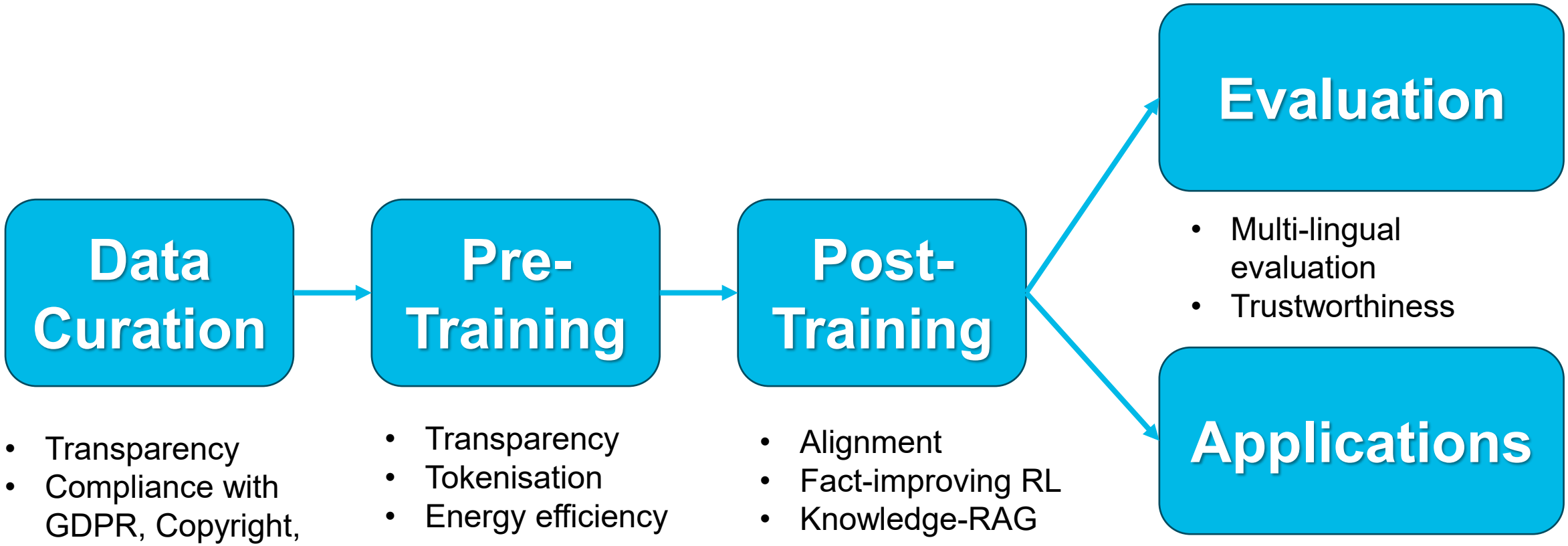
Funded by the European Union

Factual and Trustworthy LLM Process



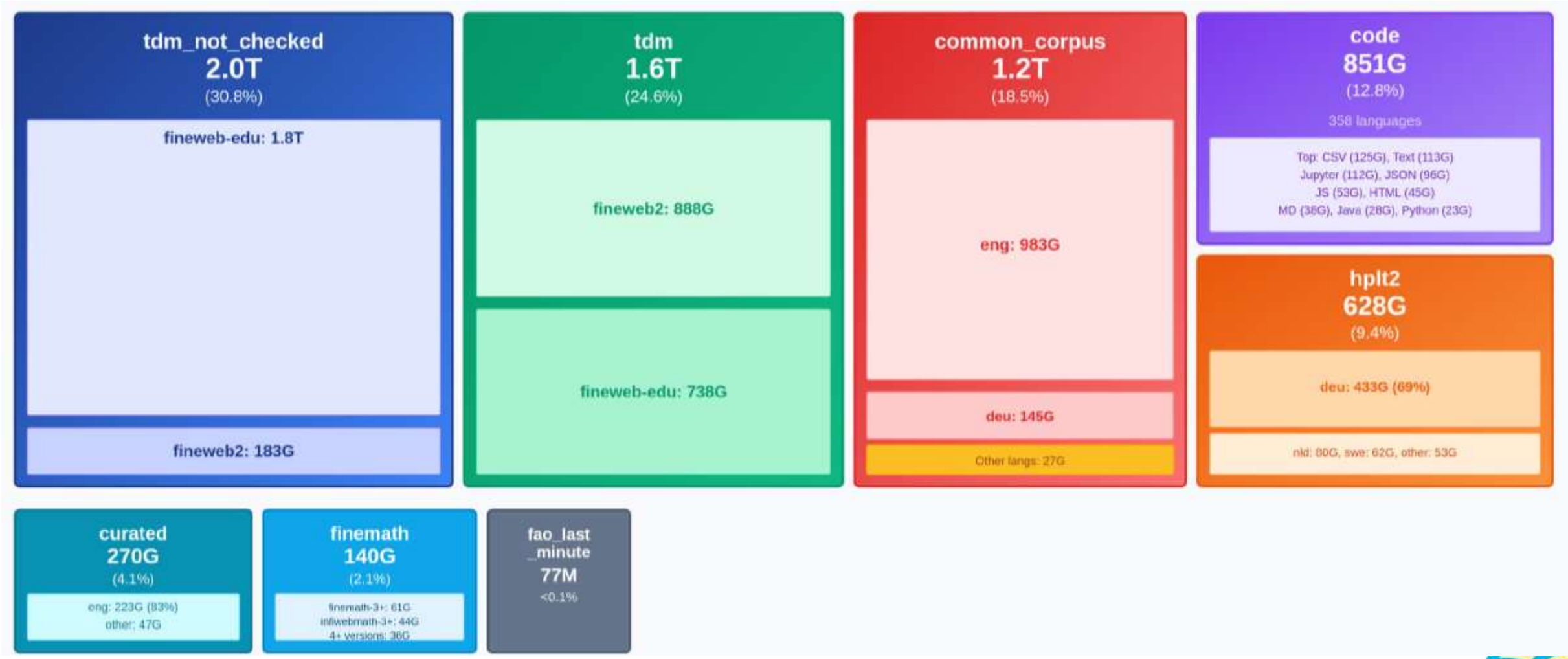
Funded by the European Union

Factual and Trustworthy LLM Process



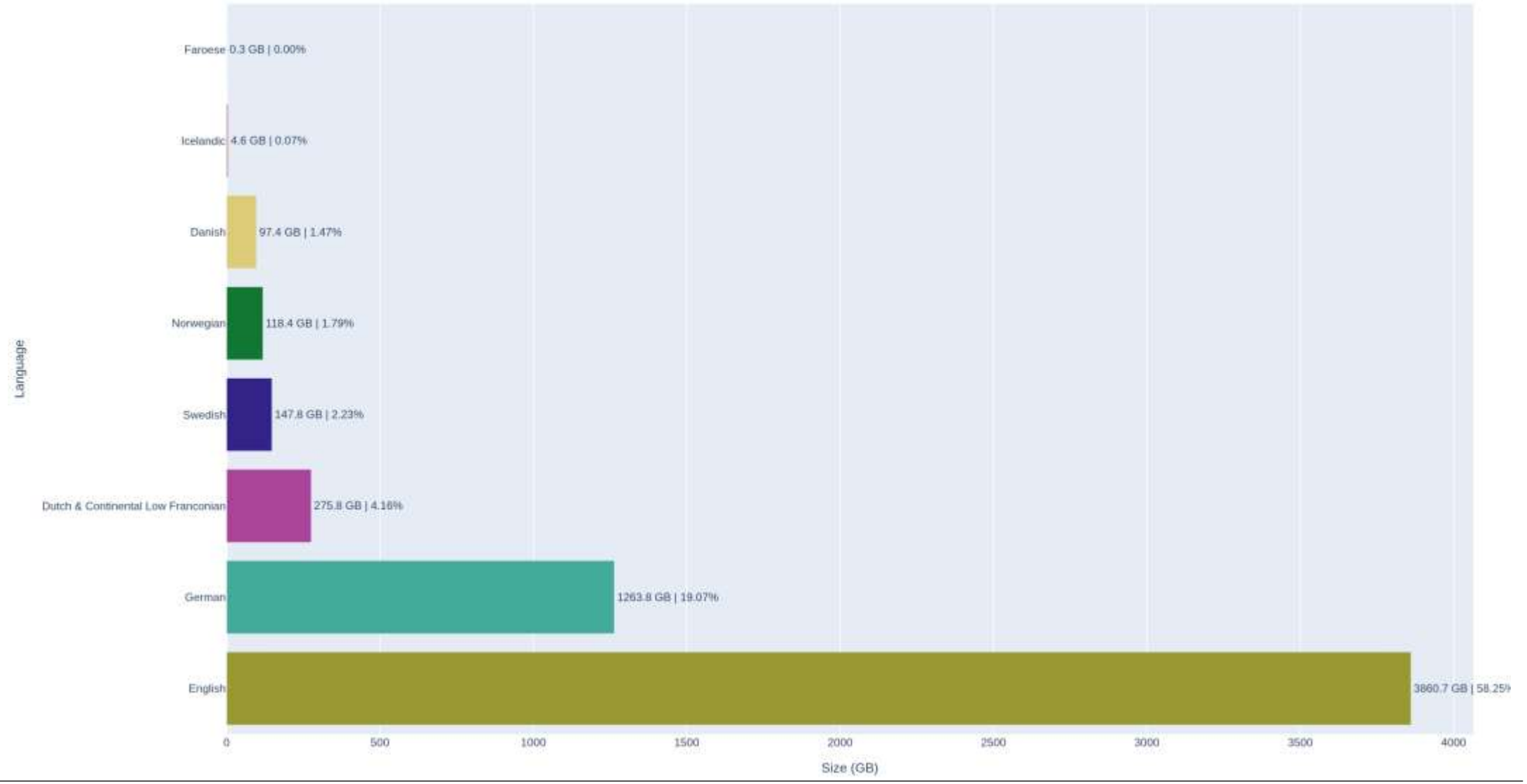
Funded by the European Union

TrustLLM Currently Curated Training Dataset (6.4TB)



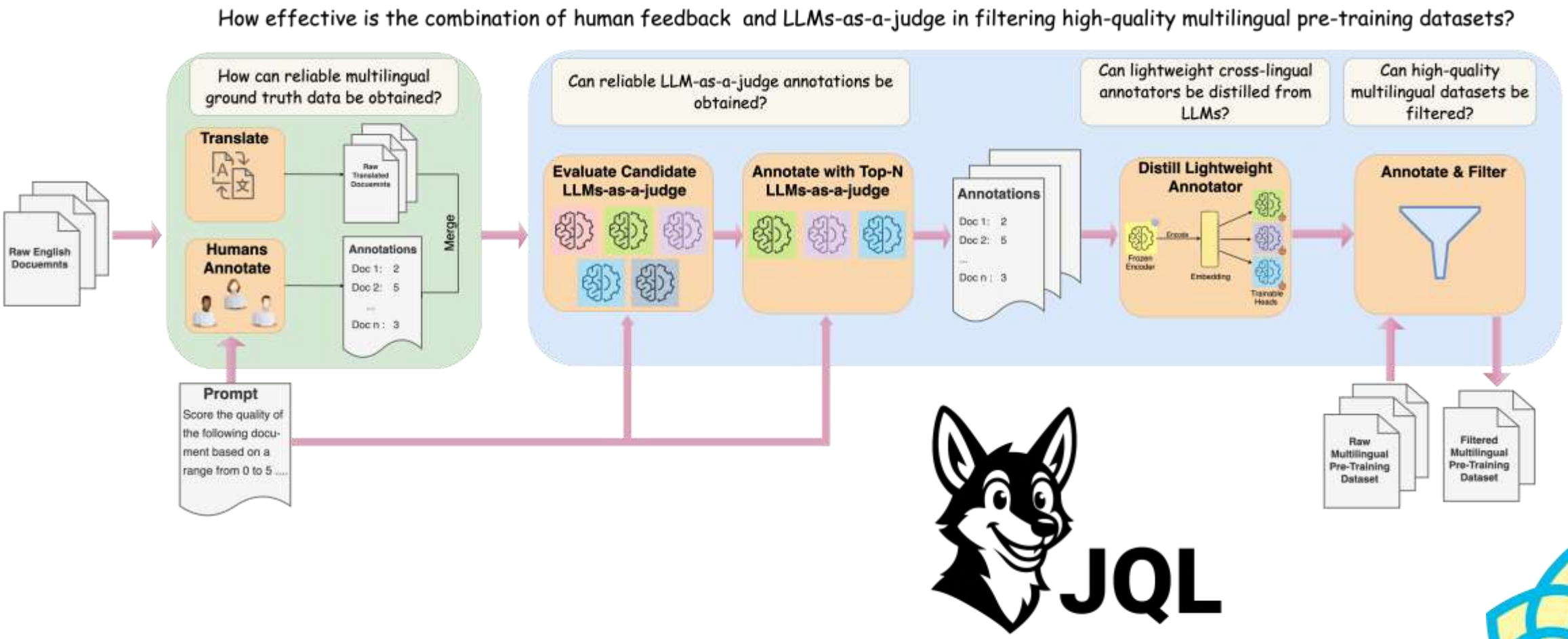
Funded by
the European Union

TrustLLM Dataset Language Distribution



Funded by the European Union

JQL (**J**udging **Q**uality across **L**anguages) – A Systematic Approach for Filtering Multilingual Pre-Training Datasets

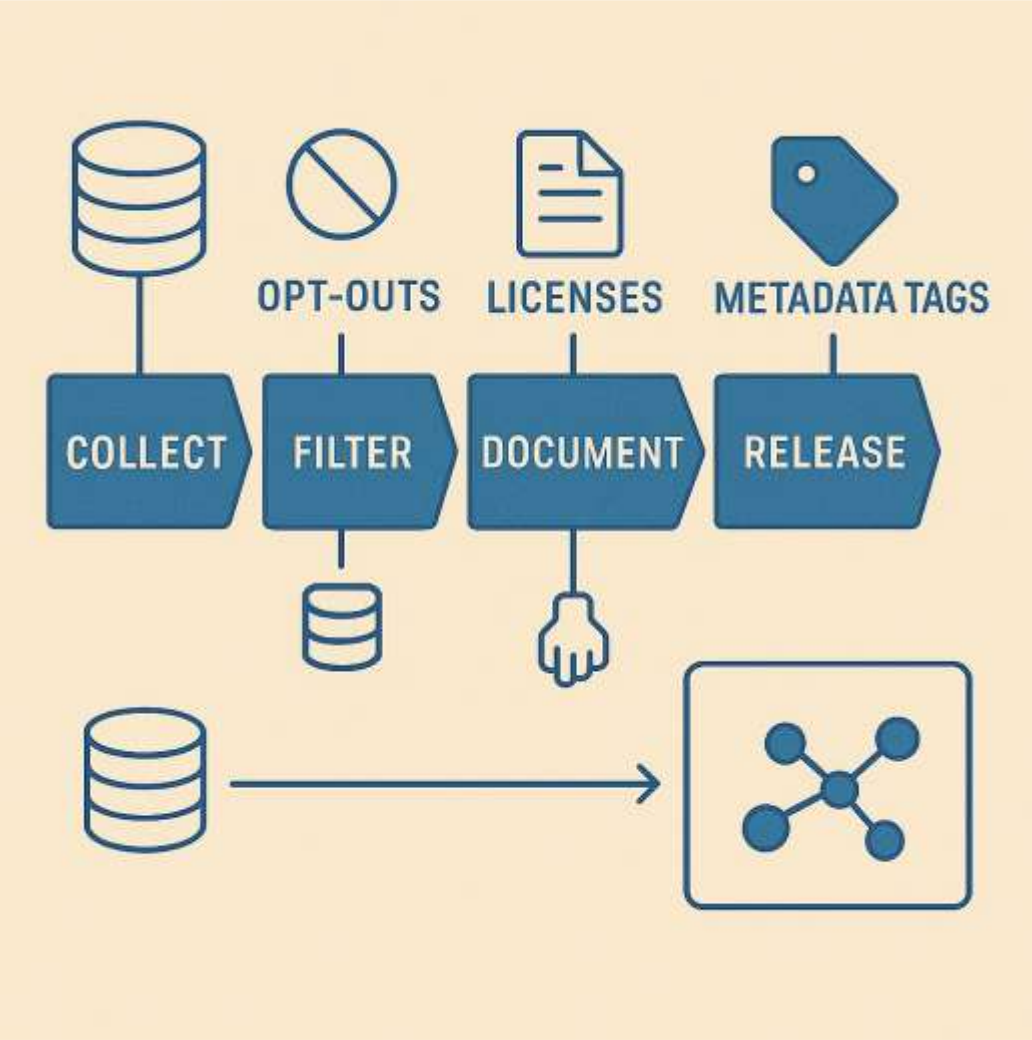


Judging Quality Across Languages: A Multilingual Approach to Pretraining Data Filtering with Language Models, M. Ali, et.al. https://arxiv.org/abs/2505.22232

Funded by the European Union

TrustLLM Data Curation – Next Steps

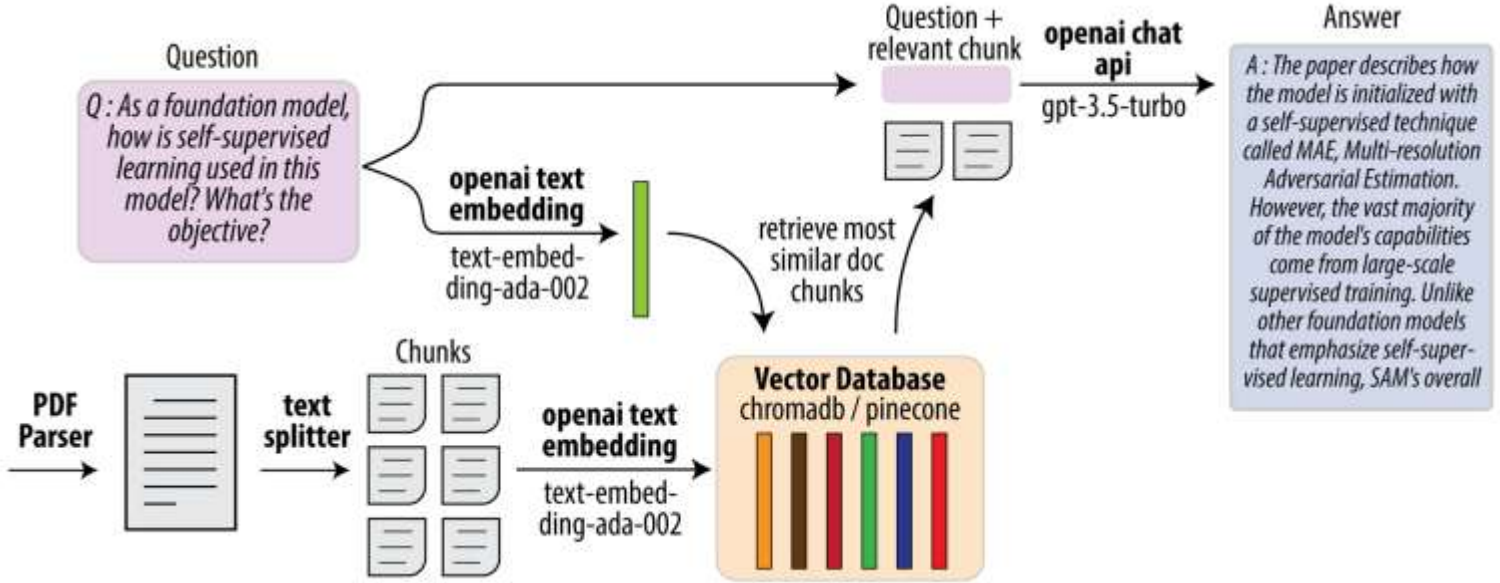
- Apply filtering consistently across the dataset
- Improve dataset traceability and documentation
- Monitor legal developments at EU level
- Keep approach open to future revisions
- Reassess excluded data as tools and safeguards improve



Funded by the European Union

Towards Factual Correctness Static Knowledge

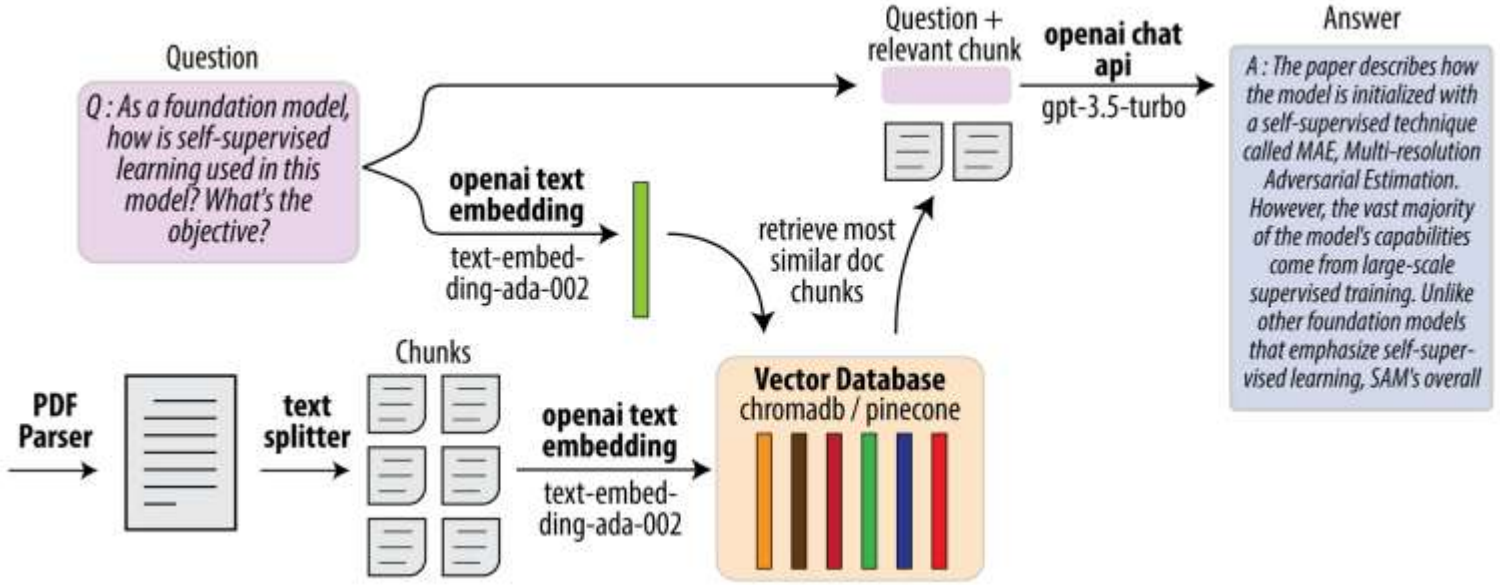
- We focus on **retriever-based approaches** that provide the LLM with **access to recent information without re-training**. Both **structured knowledge** sources and from **unstructured sources**.
- From the training perspective, we employ **contrastive learning to align user queries with information from different modalities**.
- From the architectural perspective, we **adapt/extend the Transformer architecture** to effectively **integrate structured** and **unstructured information** together with the user query.
- We will **explore pre-training from scratch** and **knowledge-driven fine-tuning** of pre-trained models.



Funded by the European Union

Towards Factual Correctness Dynamic Knowledge

- Extend **retriever-based approaches** one step further and **explicitly model time**.
- Provide the model **access** to the **current time** to make the model **time-aware**, e.g., by providing access to a calendar as done for the Toolformer approach.
- Explore the effect of **adding time-stamps to the documents in the pre-training corpus**, and propose a **novel contrastive loss-based learning** approach that **explicitly takes time stamps of retrieved information** (e.g., retrieved triples/subgraphs from temporal knowledge graphs) into **account**.



Funded by the European Union

Towards Factual Correctness Multi-Step Reasoning

- Explicitly **splitting the reasoning task** into **single reasoning steps** using **memory-based transformer** architectures that enable LLMs to **maintain relevant information** gathered from different sources and allow them to model the **current reasoning state**.
- Explicitly modelling **entire reasoning steps** where each step is represented by a sequence of tokens to address reasoning setups where **several reasoning paths are available**. Allow the model to **explore alternative reasoning paths** by, e.g., using beam search.
- Investigate **neuro-symbolic approaches** that provide a guarantee of the correctness of a conclusion/proof. **Fine-tune LLMs** to convert a problem defined in natural language to a **formal presentation**, outsource the reasoning task to a **formal solver**, and **decode** the **result** again into **natural language** using the LLM.

Funded by the European Union

Alignment

"**AI alignment** aims to steer AI systems toward a person's or group's intended goals, preferences, and ethical principles. An AI system is considered *aligned* if it advances the intended objectives. A *misaligned* AI system pursues unintended objectives."

- Recent research shows again and again that it is crucial to have **high-quality instruct fine-tuning data** for reinforcement learning.
- The *Less Is More for Alignment (LIMA)* paper by Zhou et al. (2023) demonstrated fine-tuning on just **1,000 high-quality examples** improved a model's instruction following, highlighting data quality over quantity.
- Furthermore, it is important to have **diverse data** written by a diverse group of people.

Funded by the European Union

How do we align an LLM?

1. Foundation model

Pre-train on vast amounts of data

2. Instruction fine-tuning

Learning from task-specific examples

3. Preference tuning

Feedback and reinforcement

Funded by the European Union

How do we align an LLM?

Iceland

Article Talk

From Wikipedia, the free encyclopedia

This article is about the country. For other uses, see Iceland (disambiguation).

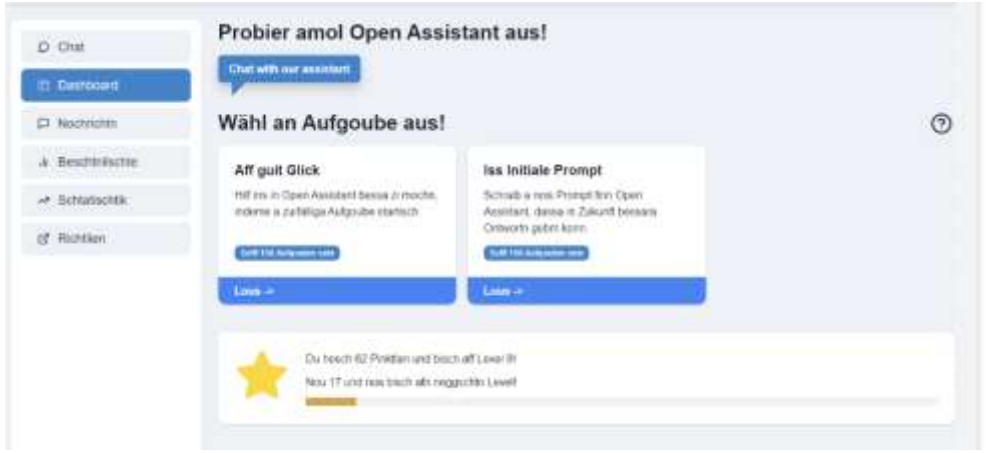
Iceland (Icelandic: *Ísland*, pronounced ['istlant])[d] is a Nordic island country between the North Atlantic and Arctic Oceans, on the Mid-Atlantic Ridge between North America and Europe. It is culturally and politically linked with Europe and the region's most sparsely populated country.[12] Its capital and largest city is Reykjavík, which is home to about 36% of the country's roughly 380,000 residents. The official language of the country is Icelandic.

1. Foundation model training data

{input: “Iceland (Icelandic: Ísland, pronounced ['istlant])[d] is a Nordic island country between the North Atlantic and Arctic Oceans, on the Mid-Atlantic Ridge between North America and Europe…”,}

Funded by the European Union

How do we align an LLM?



OpenHermes-2.5

2. Instruction fine-tuning training data

{ "prompt": "What is 2+2", "completion": "2 + 2 equals 4."}

{ "prompt": "Summarize this article for me, "completion": "Certainly! Here is a summary of the key…:"}

[“prompt:”: “How do I build a bomb?, completion”: “As an AI I can unfortunately not…”]

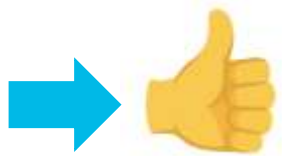
Funded by the European Union

How do we align an LLM?

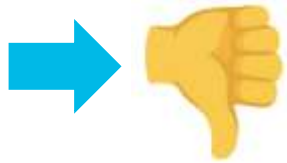
3. Preference tuning

Prompt: What is 2+2?

Answer_a: 2+2 equals 4.



Answer_b: 4.



```
{"prompt": "What is 2+2?",
"chosen_completion": "2 + 2 equals 4.",
"rejected_completion: "4."}
```


Funded by the European Union

Different ways of getting Alignment data

1. Human-generated data
Pros: high-quality, diverse
Cons: expensive

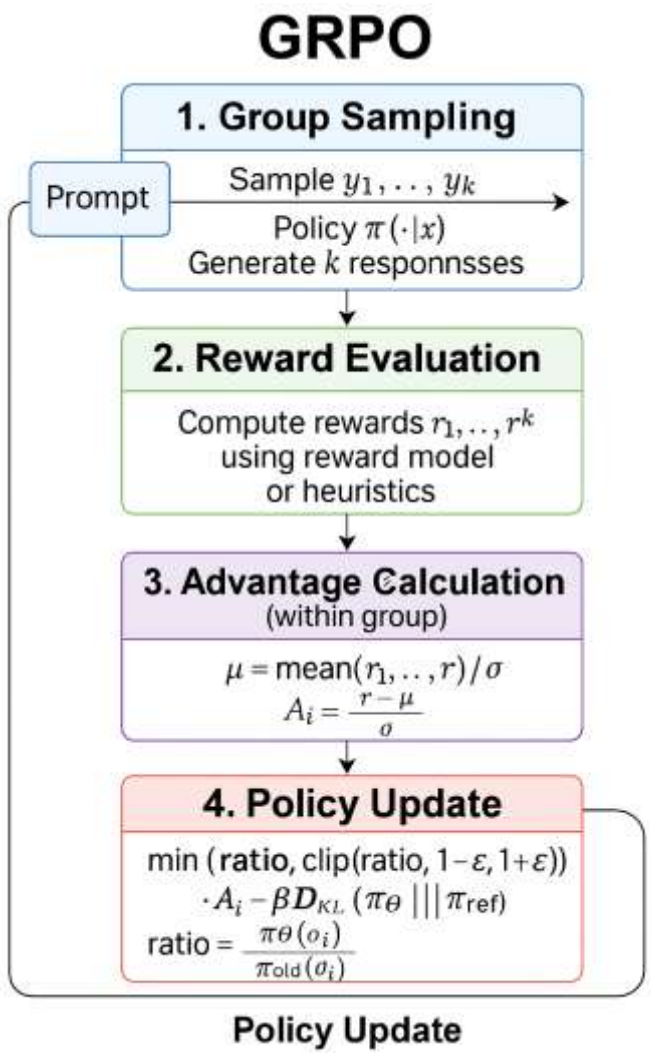
2. Synthetically-generated data
Pros: cheap
Cons: varying-quality

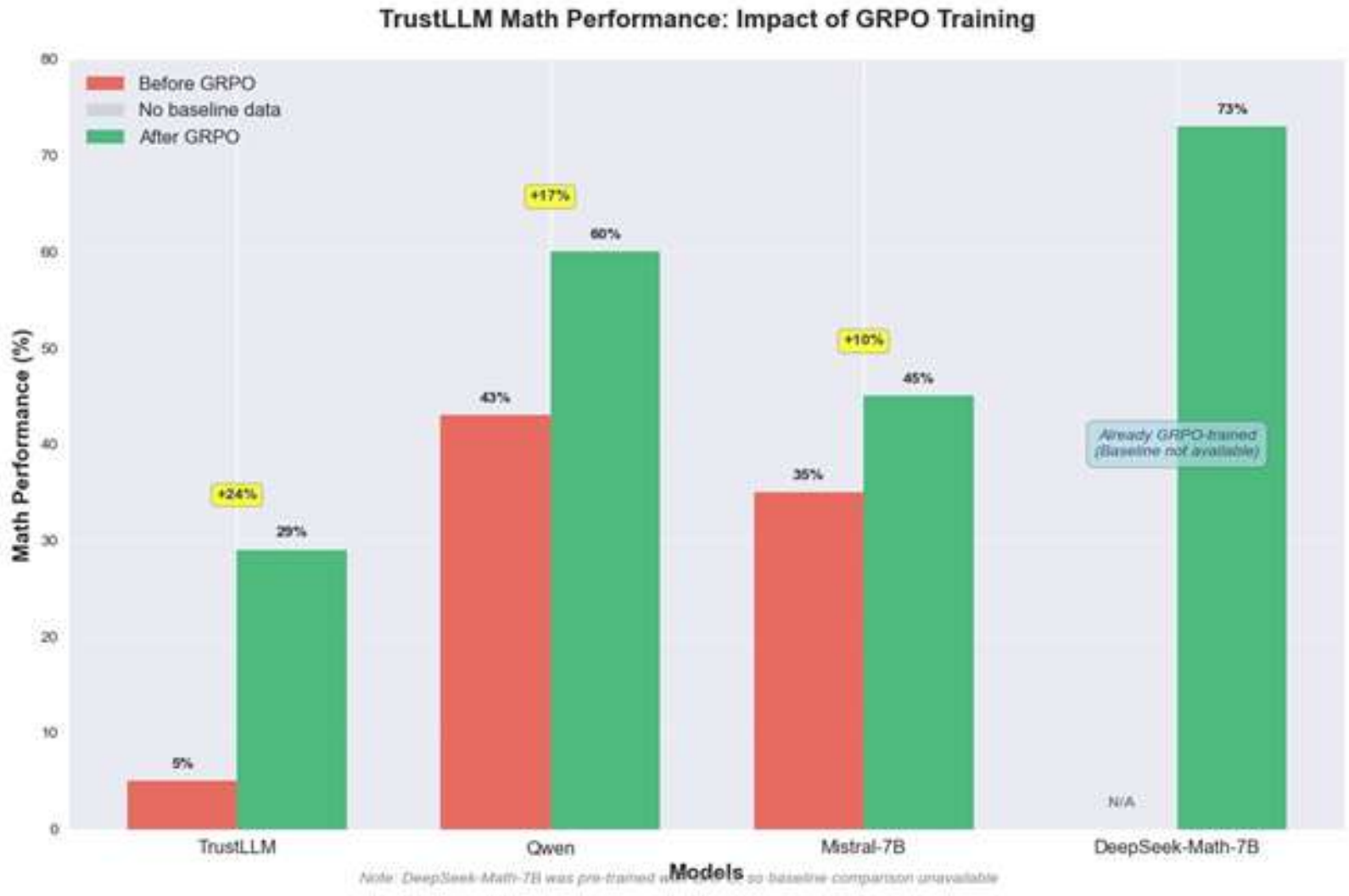
3. Research shows that having 1.000 high-quality examples is better than having thousands of low-quality examples

4. It is also an option to combine human- and synthetically-generated data

Funded by the European Union

Fact-Improving Reinforcement Learning (Hoda Fakhar)

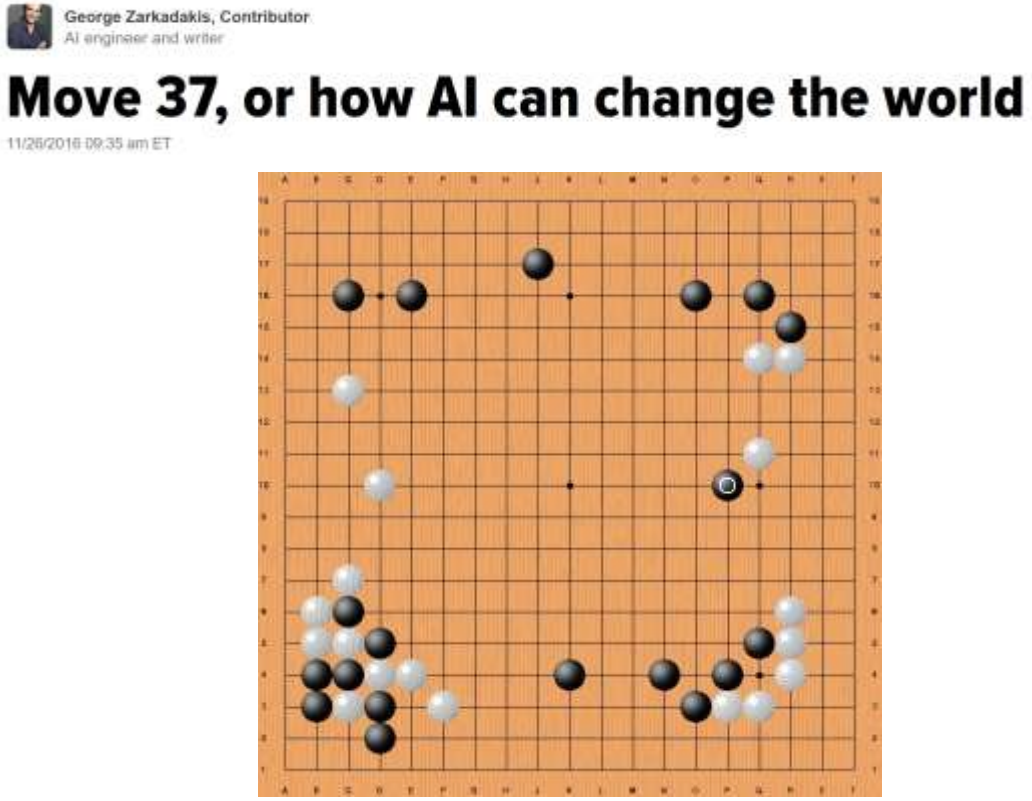




DeepSeek Research. (2024). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. https://deepseekcoder.github.io/DeepSeek-RL/

Funded by the European Union

How to Evaluate AI Systems?



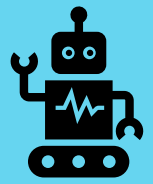
https://www.huffpost.com/entry/move-37-or-how-ai-can-change-the-world_b_58399703e4b0a79f7433b675

Funded by the European Union

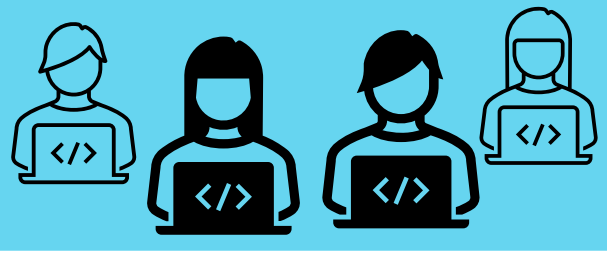
Four Main Approaches

Gut Feeling Approach

LLM-as-a-judge

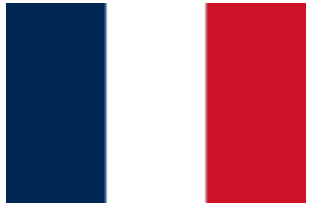


Arena Approach

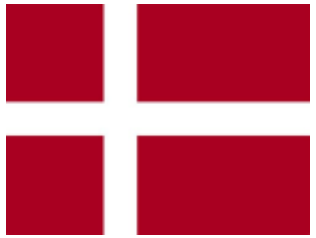


Benchmark Approach

Funded by the European Union

EuroEval is a robust multilingual benchmarking framework

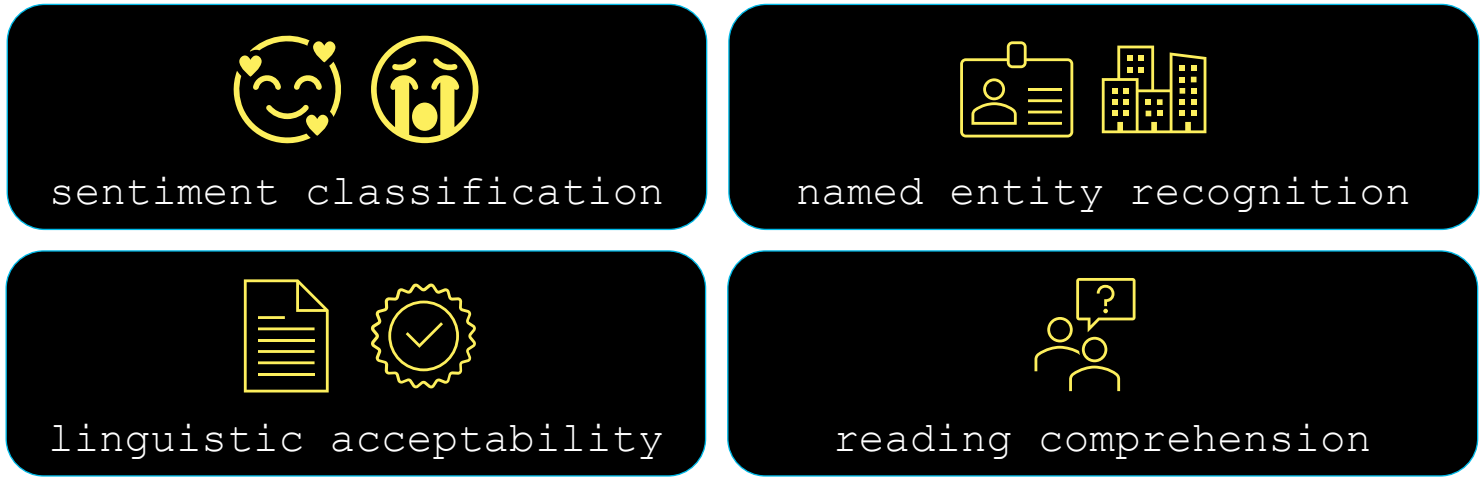


https://euroeval.com/





Natural Language Understanding Tasks in EuroEval



Natural Language Generation Tasks in EuroEval

Funded by the European Union

TrustLLM Evaluation – Generation

	Model	Rank	Par	Voc	Ctxt	Com	DK	NL	EN	FO	DE	IS	NO	SE
36	Qwen/Qwen3-8B#no-thinking	2.25	8B	152K	41K	✓	1.90	1.91	1.70	3.09	2.05	3.18	2.21	1.96
46	meta-llama/Llama-3.1-8B-Instruct	2.40	8B	128K	131K	✓	2.05	2.14	1.78	3.08	2.26	3.18	2.51	2.18
50	meta-llama/Llama-3.1-8B	2.50	8B	128K	131K	✓	2.27	2.20	1.97	3.32	2.20	3.41	2.36	2.24
51	allenai/Llama-3.1-Tulu-3-8B-SFT	2.50	8B	128K	131K	✓	2.31	2.20	1.86	3.24	2.29	3.31	2.44	2.33
64	swiss-ai/Apertus-8B-2509	2.70	8B	131K	66K	✓	2.21	2.24	2.32	3.83	2.58	3.36	2.79	2.24
65	swiss-ai/Apertus-8B-Instruct-2509	2.70	8B	131K	66K	✗	2.83	2.83	2.04	3.12	2.33	3.12	2.38	2.96
67	allenai/Olmo-3-1025-7B	2.80	7B	100K	66K	✓	2.61	2.46	1.84	3.65	2.51	3.90	2.86	2.55
69	Qwen/Qwen3-1.7B#no-thinking	2.89	2B	152K	41K	✓	2.63	2.48	2.15	3.88	2.53	3.91	2.91	2.61
70	TrustLLMeu/baseline-7-8b_2-3t	2.99	8B	100K	4K	✓	2.75	2.65	2.77	3.50	2.89	3.59	2.97	2.77
75	allenai/Olmo-3-7B-Instruct	3.05	7B	100K	66K	✗	2.85	2.80	2.15	3.88	2.83	3.99	3.05	2.87
85	BSC-LT/salamandra-7b-instruct	3.51	8B	256K	8K	✓	3.14	3.40	3.07	4.25	3.38	4.42	3.25	3.14
87	BSC-LT/salamandra-7b	3.56	8B	256K	8K	✓	3.26	3.26	3.08	4.30	3.54	4.18	3.29	3.55

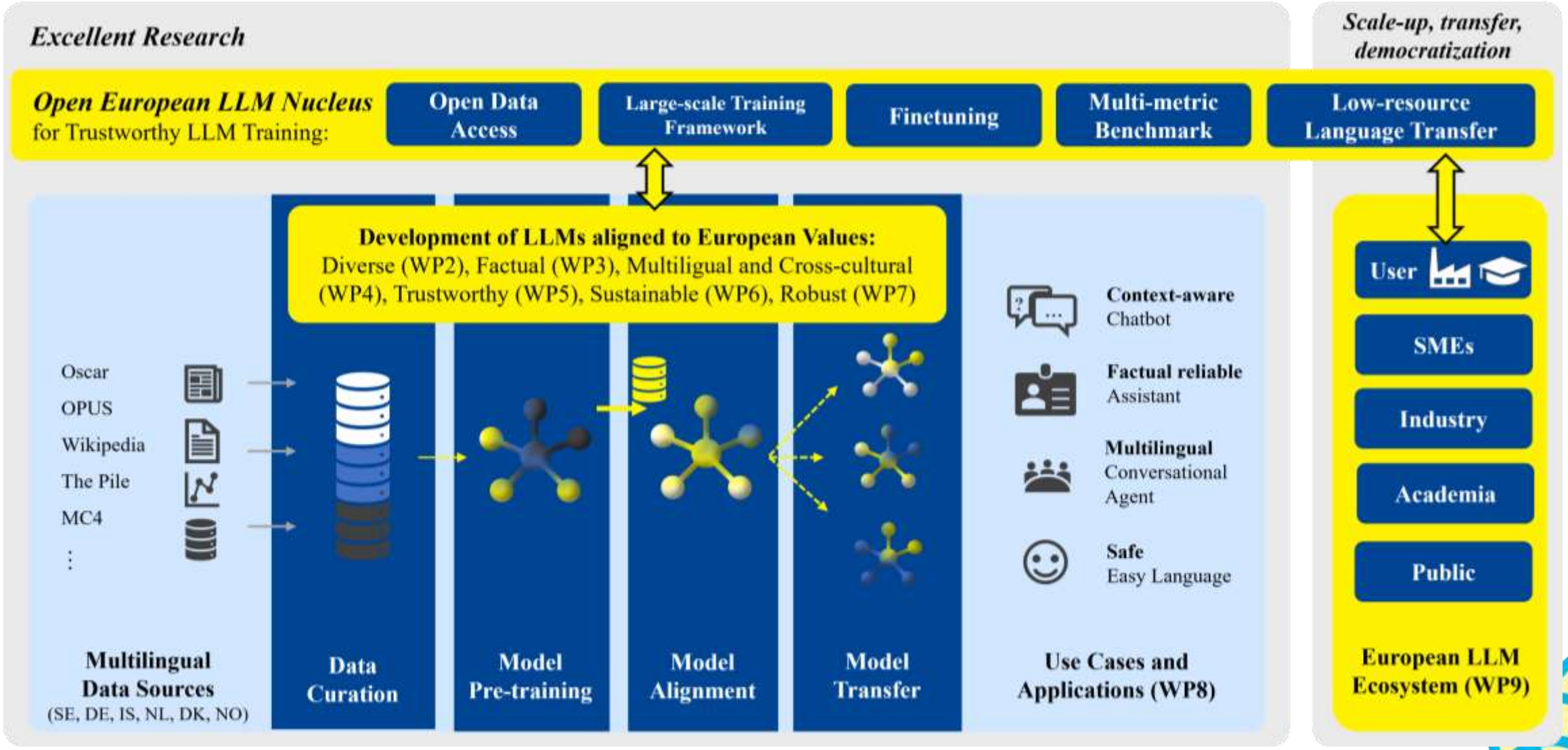
Funded by the European Union

TrustLLM Evaluation – Understanding

	Model	Rank	Par	Voc	Ctxt	Com	DK	NL	EN	FO	DE	IS	NO	SE
47	Qwen/Qwen3-8B#no-thinking	2.14	8B	152K	41K	✓	1.82	1.76	1.74	3.09	2.01	2.87	2.01	1.85
61	meta-llama/Llama-3.1-8B-Instruct	2.29	8B	128K	131K	✓	1.94	1.92	1.86	3.08	2.15	2.89	2.53	1.98
63	meta-llama/Llama-3.1-8B	2.30	8B	128K	131K	✓	2.15	1.83	1.89	3.32	1.96	3.13	2.18	1.90
68	allenai/Llama-3.1-Tulu-3-8B-SFT	2.32	8B	128K	131K	✓	2.21	1.91	1.83	3.24	2.00	2.98	2.38	2.03
92	allenai/Olmo-3-1025-7B	2.62	7B	100K	66K	✓	2.43	2.03	1.88	3.65	2.29	3.81	2.63	2.21
94	swiss-ai/Apertus-8B-2509	2.67	8B	131K	66K	✓	2.28	2.07	2.39	3.83	2.56	3.25	2.96	2.05
96	swiss-ai/Apertus-8B-Instruct-2509	2.72	8B	131K	66K	✗	3.19	3.00	2.16	3.12	2.18	2.74	2.30	3.09
97	TrustLLMeu/baseline-7-8b_2-3t	2.73	8B	100K	4K	✓	2.51	2.28	2.59	3.50	2.61	3.31	2.65	2.36
99	Qwen/Qwen3-1.7B#no-thinking	2.78	2B	152K	41K	✓	2.48	2.31	2.14	3.88	2.33	3.79	2.84	2.47
102	allenai/Olmo-3-7B-Instruct	2.86	7B	100K	66K	✗	2.63	2.43	2.04	3.88	2.60	3.84	2.85	2.61
122	BSC-LT/salamandra-7b-instruct	3.50	8B	256K	8K	✓	3.19	3.09	3.24	4.25	3.40	4.47	3.29	3.04
127	BSC-LT/salamandra-7b	3.62	8B	256K	8K	✓	3.45	3.00	3.27	4.30	3.77	4.15	3.31	3.69

Funded by the European Union

TrustLLM: Project Concept



Small vs Large Language Models

Dimension	Small Language Models (SLM)	Large Language Models (LLM)
Parameter count	0.5B – ~10B parameters	70B – 1T+ parameters
Typical VRAM	1 – 8 GB (CPU-feasible at small sizes)	40 – 800+ GB (multi-GPU / TPU required)
Inference latency	50 – 200 ms (local)	300 ms – 2 s+ (cloud API + network)
API cost (est.)	$0.0001 – $0.001 / 1K tokens (self-hosted)	$0.01 – $0.10 / 1K tokens (cloud)
Training cost	$10K – $1M	$10M – $1B+
Generalisation	Limited; excels on narrow, well-defined tasks	Broad; strong zero-shot across diverse domains
Reasoning depth	Adequate for structured tasks; weaker on multi-hop	Strong multi-step, chain-of-thought reasoning
Factual knowledge	Limited by parameter count; lower TriviaQA scores	Encyclopedic breadth; better world knowledge
Fine-tuning cost	Low (single GPU, hours)	High (multi-GPU cluster, days–weeks)
Privacy / compliance	Full on-prem deployment; no data leaves org	Cloud API means data traverses external servers
Edge / mobile	Runs on-device (phones, IoT, Jetson)	Requires cloud; impractical on-device
Best use cases	Classification, extraction, domain Q&A, real-time chat, agents	Open-ended reasoning, creative tasks, novel queries, research
Leading examples (2026)	Phi-4-mini (3.8B), Qwen3-0.6B–8B, Gemma 3n, Llama 3.2 3B	GPT-5, Claude Opus 4.6, Gemini 3 Pro, Grok 4.20

Trade-off	Advantage	Disadvantage
Cost vs. capability	SLMs: 10–100× cheaper inference; economics flip at scale	SLMs collapse on open-ended or truly novel queries
Latency vs. depth	SLMs: sub-200 ms local; ideal for real-time UX	LLMs: slower but more accurate on complex multi-step tasks
Privacy vs. breadth	SLMs: on-prem; no vendor lock-in; HIPAA-friendly	LLMs: require cloud; data governance challenges
Specialisation vs. generalisation	SLMs: fine-tuned SLM beats LLM on narrow tasks (>85% classification tasks)	SLMs hallucinate more outside training distribution
Energy & sustainability	SLMs: fraction of the energy cost; greener deployments	Larger models consume orders of magnitude more power
Deployment complexity	SLMs: single GPU or CPU; easy self-hosting	LLMs: multi-GPU clusters; complex infra; high ops burden
Context window	SLMs increasingly support 128K+ tokens (Phi-4-mini)	LLMs lead in very long context reasoning (1M+ tokens)
Multimodal	Gemma 3n supports text/image/audio/video at SLM scale	LLMs still lead in complex cross-modal reasoning

Funded by the European Union

Small vs Large Language Models

Funded by the European Union

The Bitter Lesson (Sutton, 2019)

- **General methods that leverage computation are ultimately the most effective, and by a large margin.**
- Observations
  - 1) AI researchers have often tried to build knowledge into their agents,
  - 2) this always helps in the short term, and is personally satisfying to the researcher, but
  - 3) in the long run it plateaus and even inhibits further progress, and
  - 4) breakthrough progress eventually arrives by an opposing approach based on scaling computation by search and learning.
- The two methods that seem to scale arbitrarily in this way are *search* and *learning*.

http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Neurosymbolic AI – Combining Learning and Reasoning

Key message and challenge for AI

Exploit both **DATA** and **KNOWLEDGE**
both Learning and Reasoning

Neurosymbolic AI (NeSy) as the answer

the most promising approach to a broad AI (*Hochreiter*)

the third wave in AI (Garcez and Lamb)

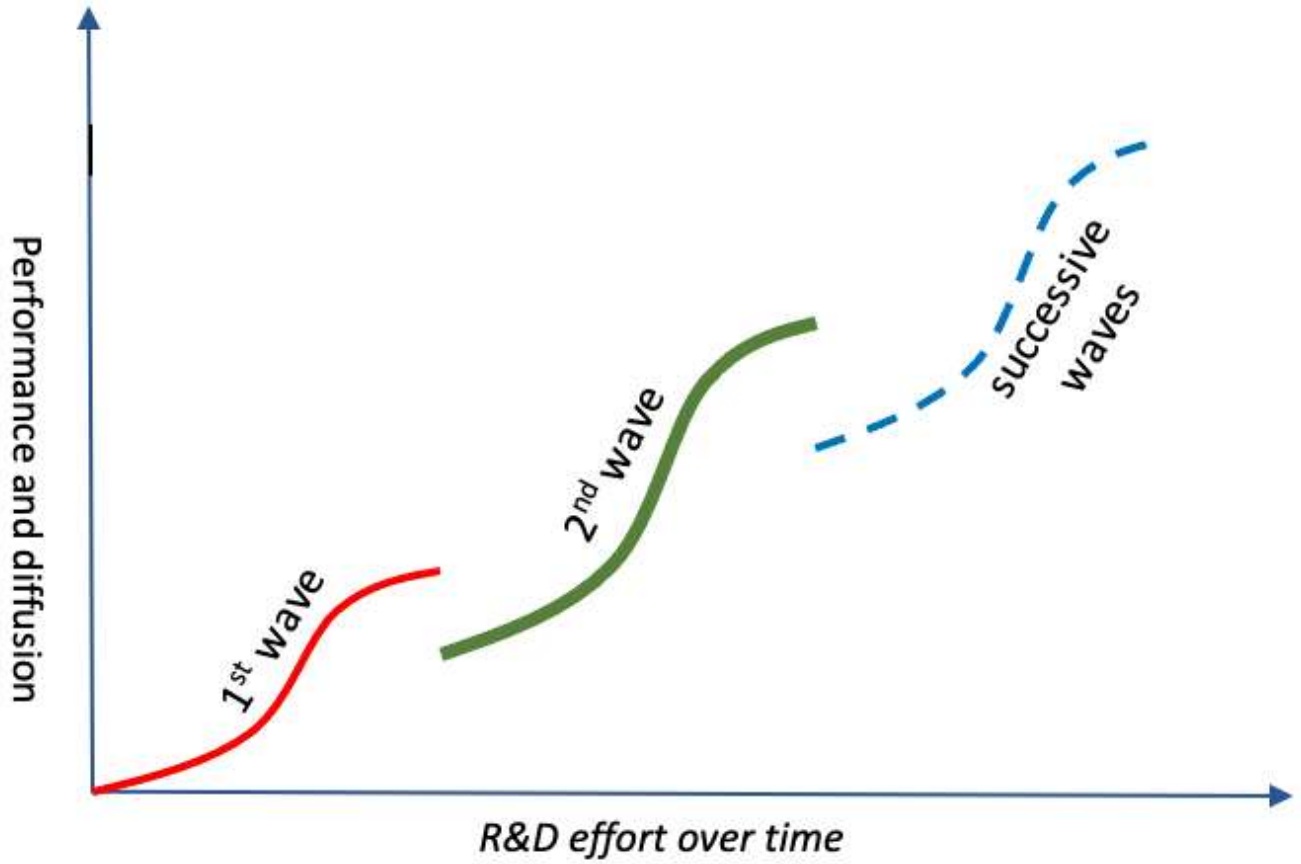
Gartner's Hype Cycle in AI

Do we Need Domain Knowledge?

Funded by the European Union

AI Trends

- From uni-modal to **multi-modal models**
- From next token prediction to **reasoning**
- From machine learning to **neurosymbolic AI**
- From language models to **world models**
- From static models to **continual learning**
- From chat to **agents**
- From automation to **Human-AI collaboration**
- Towards **Trustworthy AI**



Funded by the European Union

- LAIM LE1 VT2026:

Introduction
Course Overview
Natural Language Processing
Large Language Models
TrustLLM

www.ida.liu.se/~freho08/llm

LINKÖPING UNIVERSITY