TrustLLM

- Towards Trustworthy and Factual Large-Language Models

Fredrik Heintz

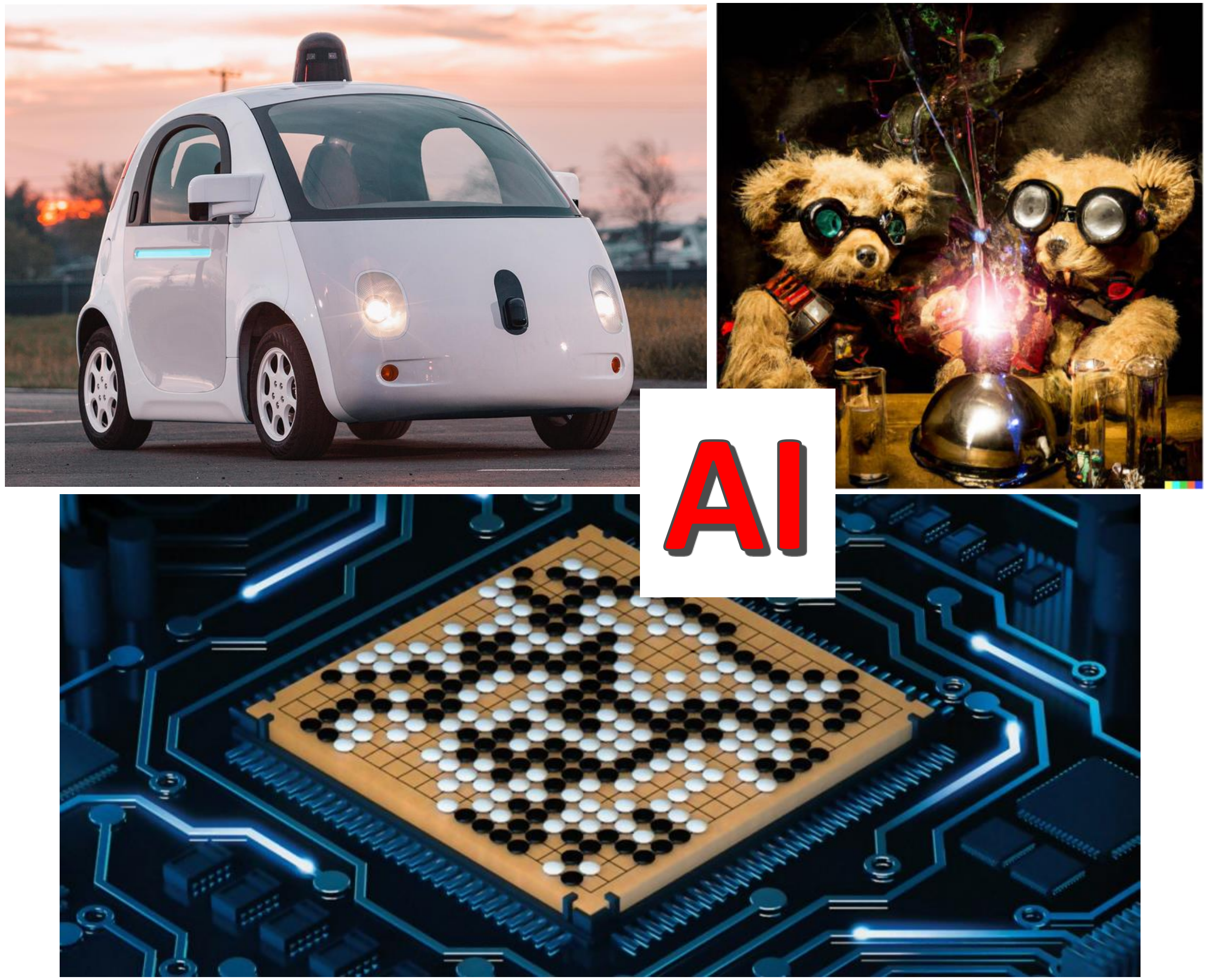
Dept. of Computer Science, Linköping University

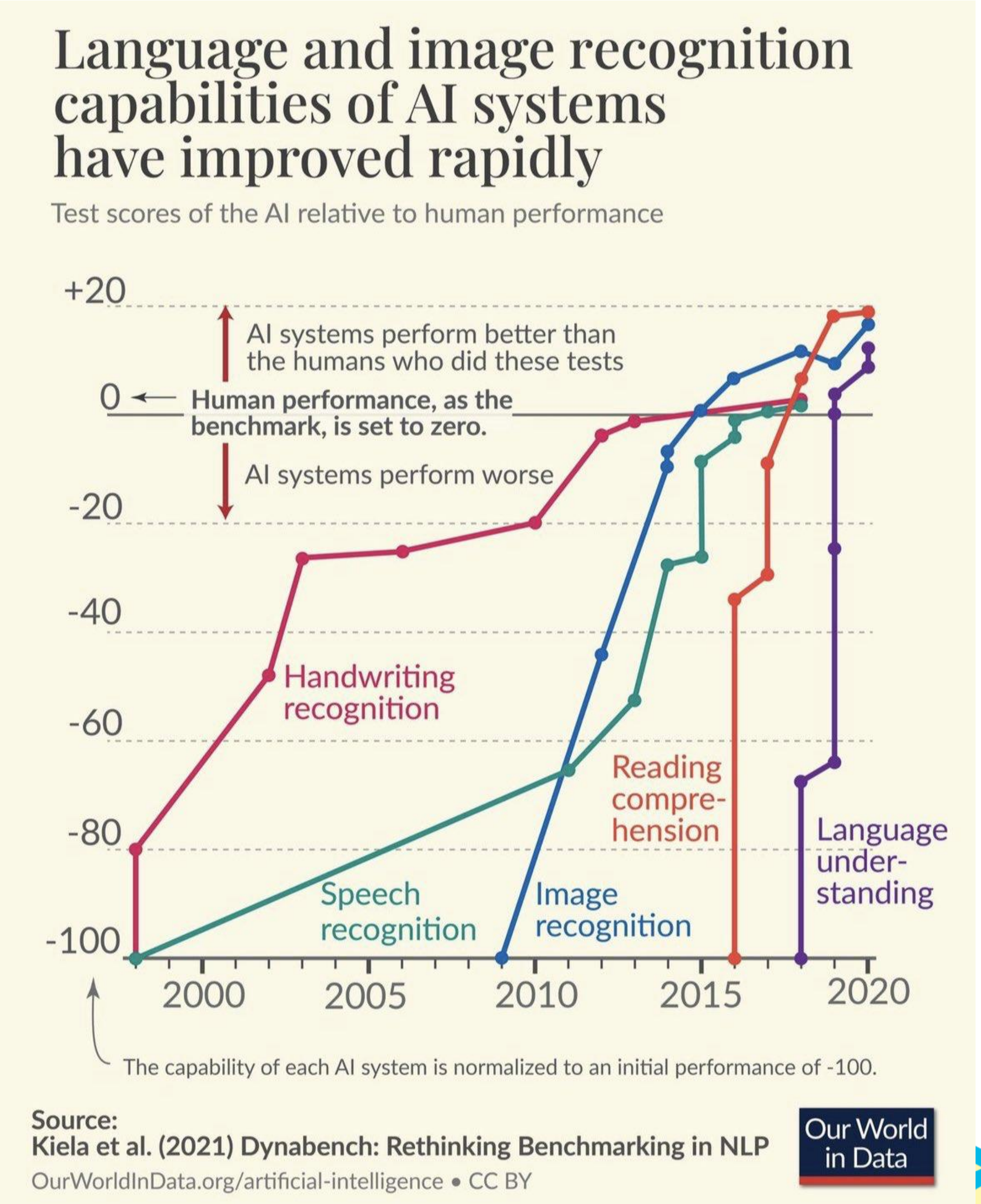
fredrik.heintz@liu.se
@FredrikHeintz

LINKÖPINGS
UNIVERSITET

TrustLLM

AI Development is Fast





Language and image recognition capabilities of AI systems have improved rapidly

Test scores of the AI relative to human performance

The capability of each AI system is normalized to an initial performance of -100.

Source:
Kiela et al. (2021) Dynabench: Rethinking Benchmarking in NLP
OurWorldInData.org/artificial-intelligence • CC BY

Funded by the European Union

Ethics Guidelines for Trustworthy AI

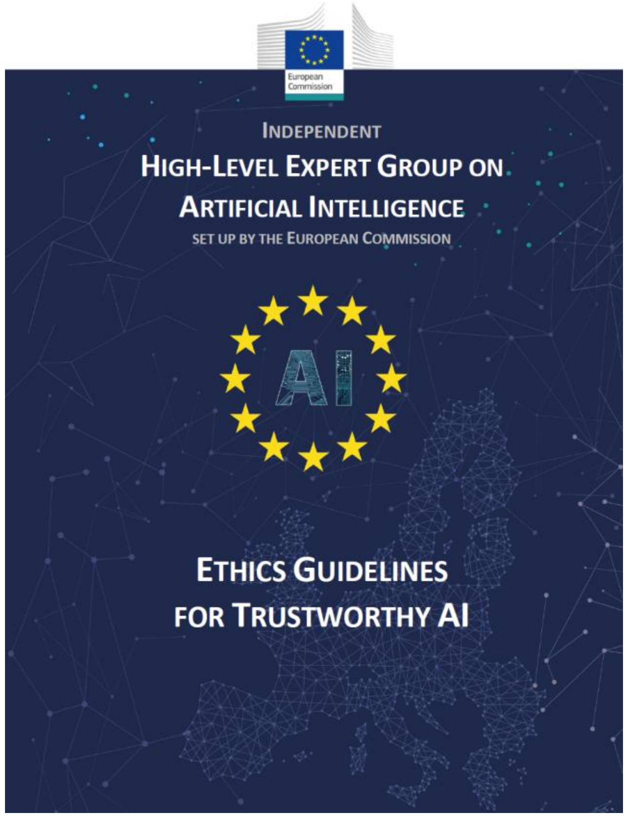
Human-centric approach: AI as a means, not an end

Trustworthy AI as our foundational ambition, with three components

- Lawful AI
- Ethical AI
- Robust AI

Three levels of abstraction

- from principles (Chapter I)
- to requirements (Chapter II)
- to assessment list (Chapter III)

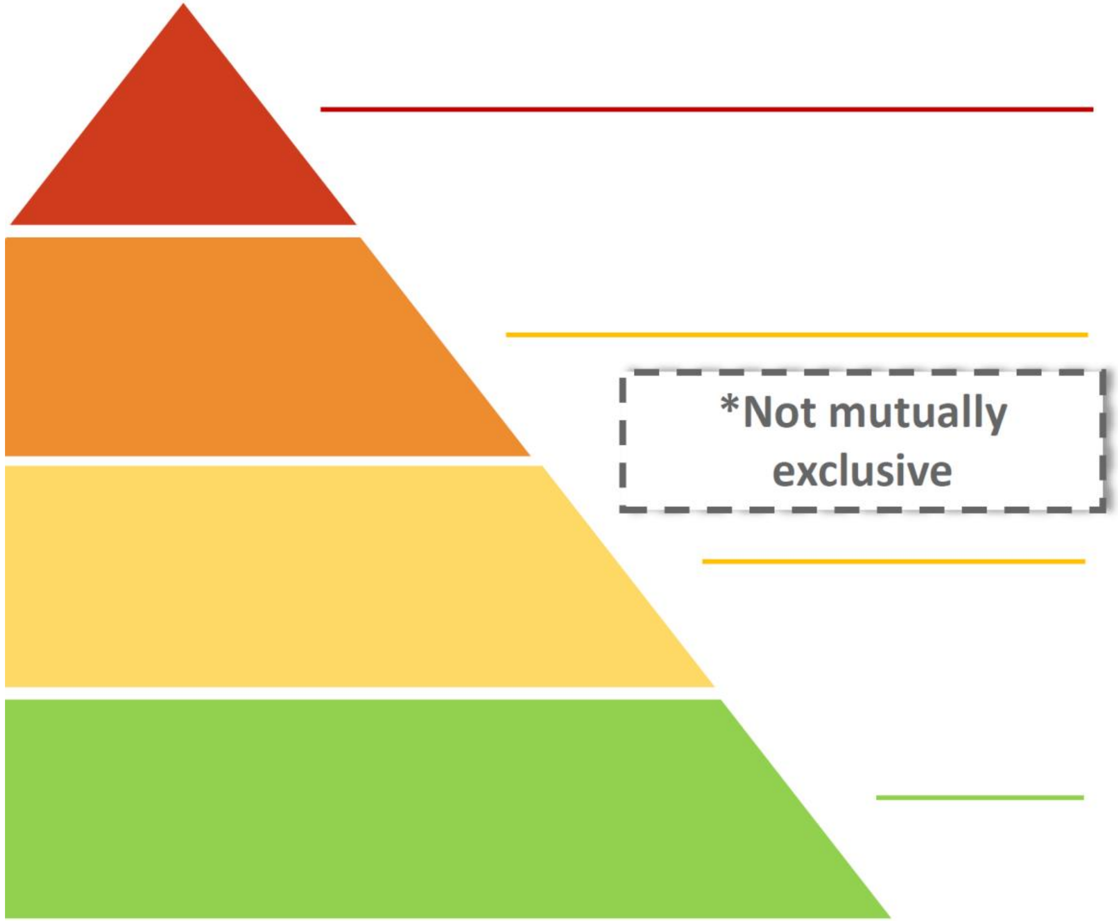


Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission

Ethics Guidelines for Trustworthy AI

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

A risk-based approach



*Not mutually exclusive

Unacceptable risk
e.g. social scoring
— **Prohibited**

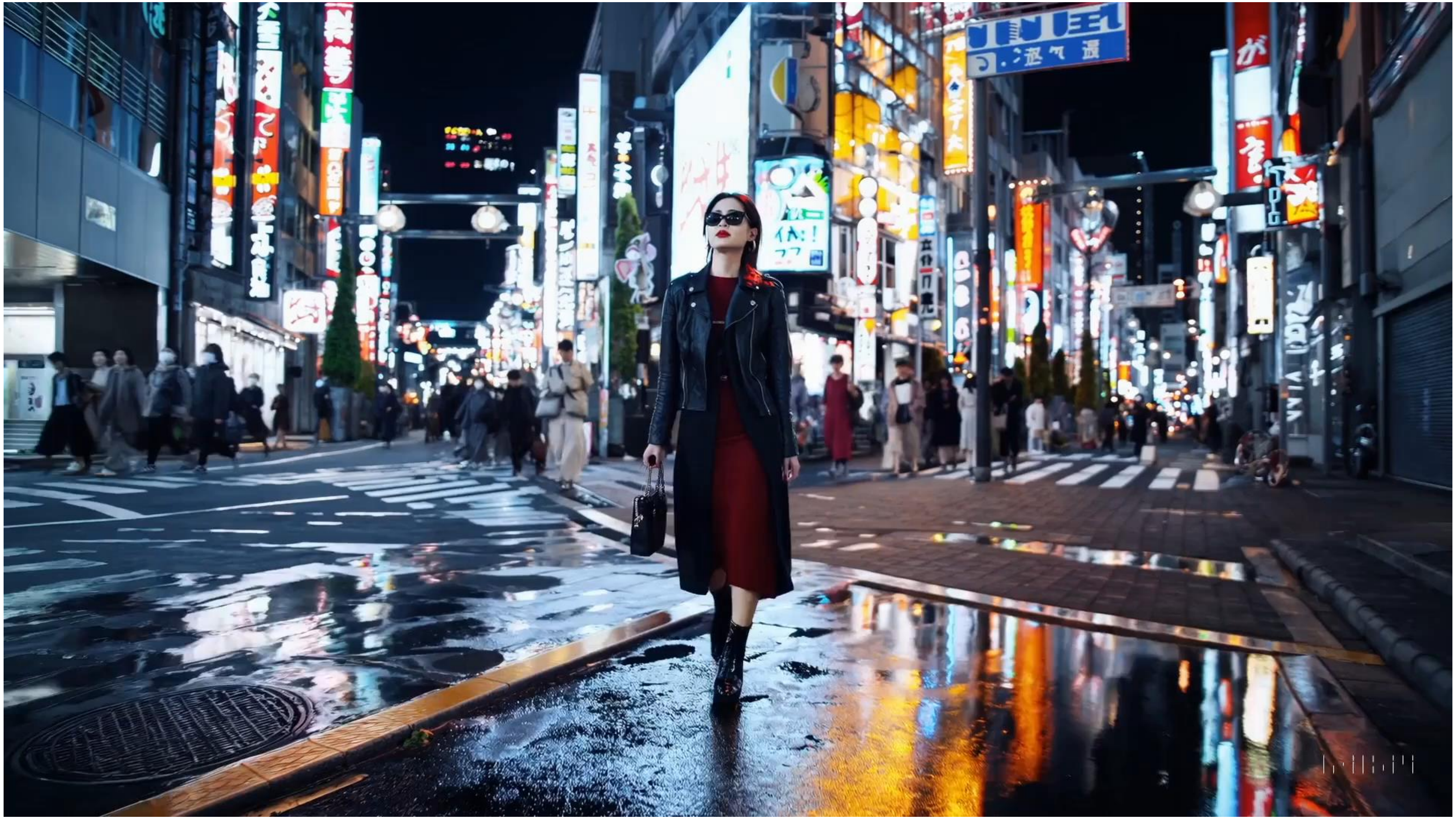
High risk
e.g. recruitment, medical devices
— **Permitted** subject to compliance with AI requirements and ex-ante conformity assessment

'Transparency' risk
'Impersonation' (bots)
— **Permitted** but subject to information/transparency obligations

Minimal or no risk
— **Permitted** with no restrictions

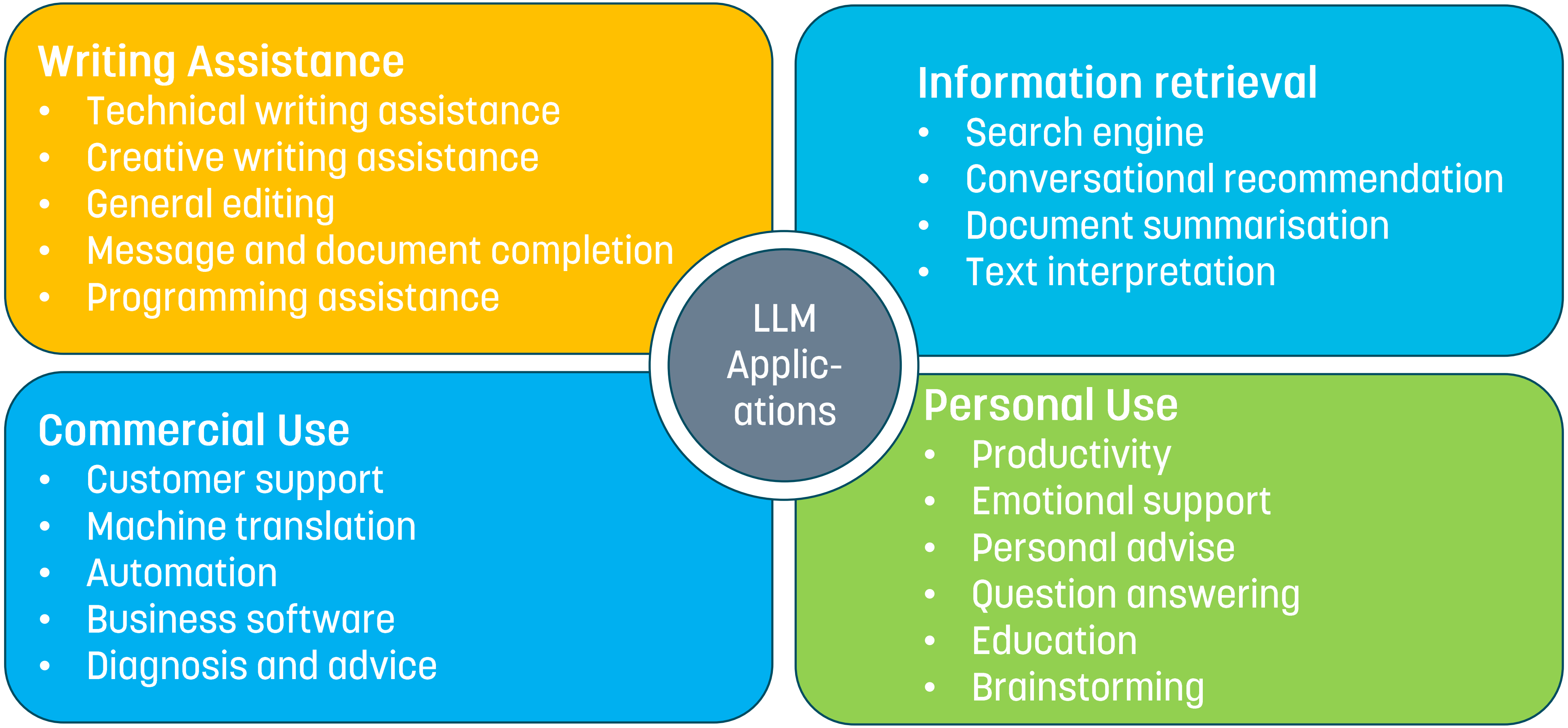
Generated by Dall-E from "photorealistic image of a self-driving car"

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.

Large Language Model Applications



Writing Assistance
- Technical writing assistance
- Creative writing assistance
- General editing
- Message and document completion
- Programming assistance

Information retrieval
- Search engine
- Conversational recommendation
- Document summarisation
- Text interpretation

LLM Applic-ations

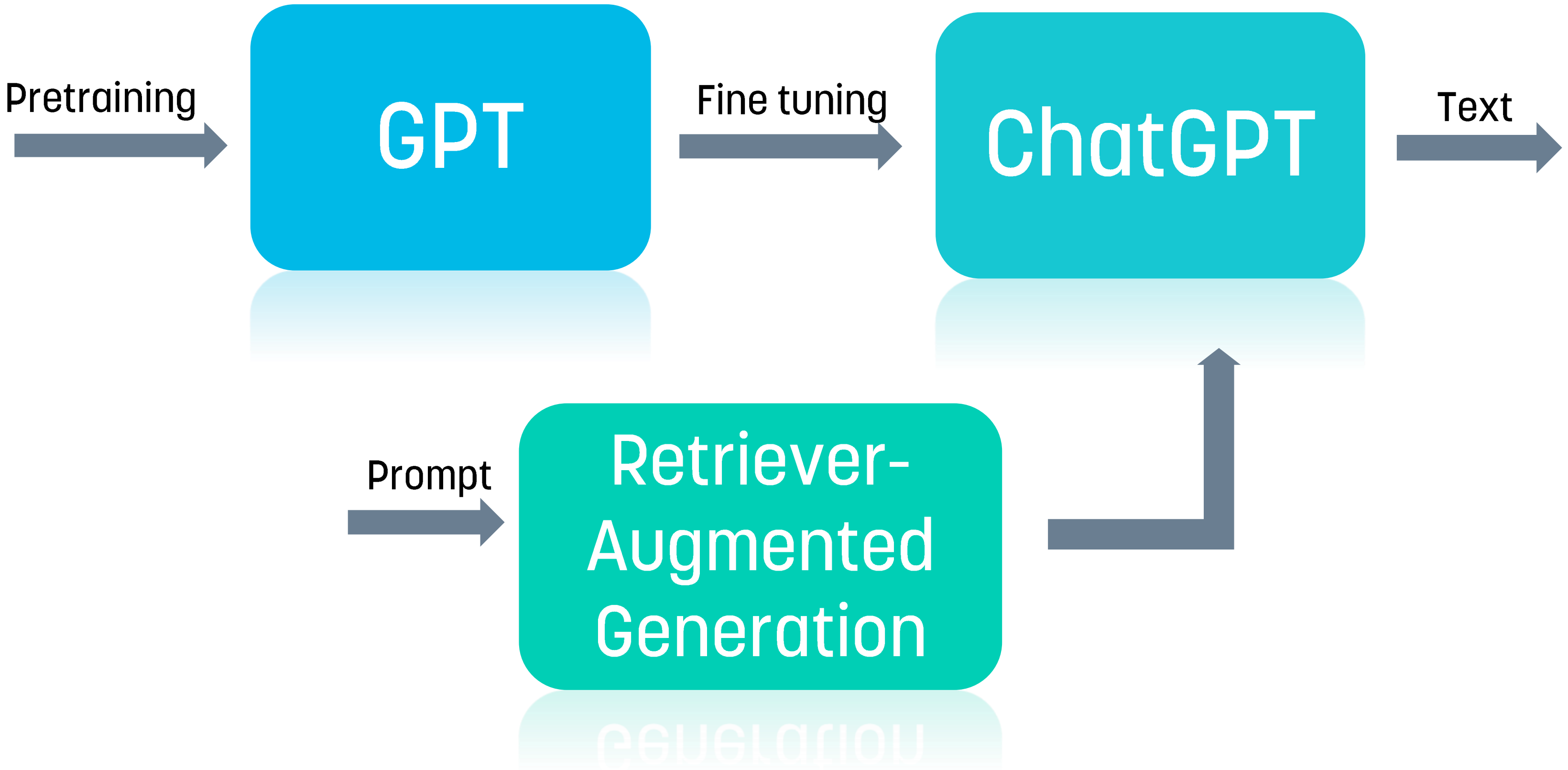
Commercial Use
- Customer support
- Machine translation
- Automation
- Business software
- Diagnosis and advice

Personal Use
- Productivity
- Emotional support
- Personal advise
- Question answering
- Education
- Brainstorming

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

Funded by the European Union

How Does ChatGPT Work?



Pretraining

GPT

Fine tuning

ChatGPT

Text

Prompt

Retriever-Augmented Generation

Funded by the European Union

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

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.

LiU LINKÖPING UNIVERSITY

MIÐEIND

ALEXANDRA INSTITUTTET

Fraunhofer

UNIVERSITY OF ICELAND

UNIVERSITY OF COPENHAGEN

JÜLICH Forschungszentrum

TNO

KI BUNDESVERBAND

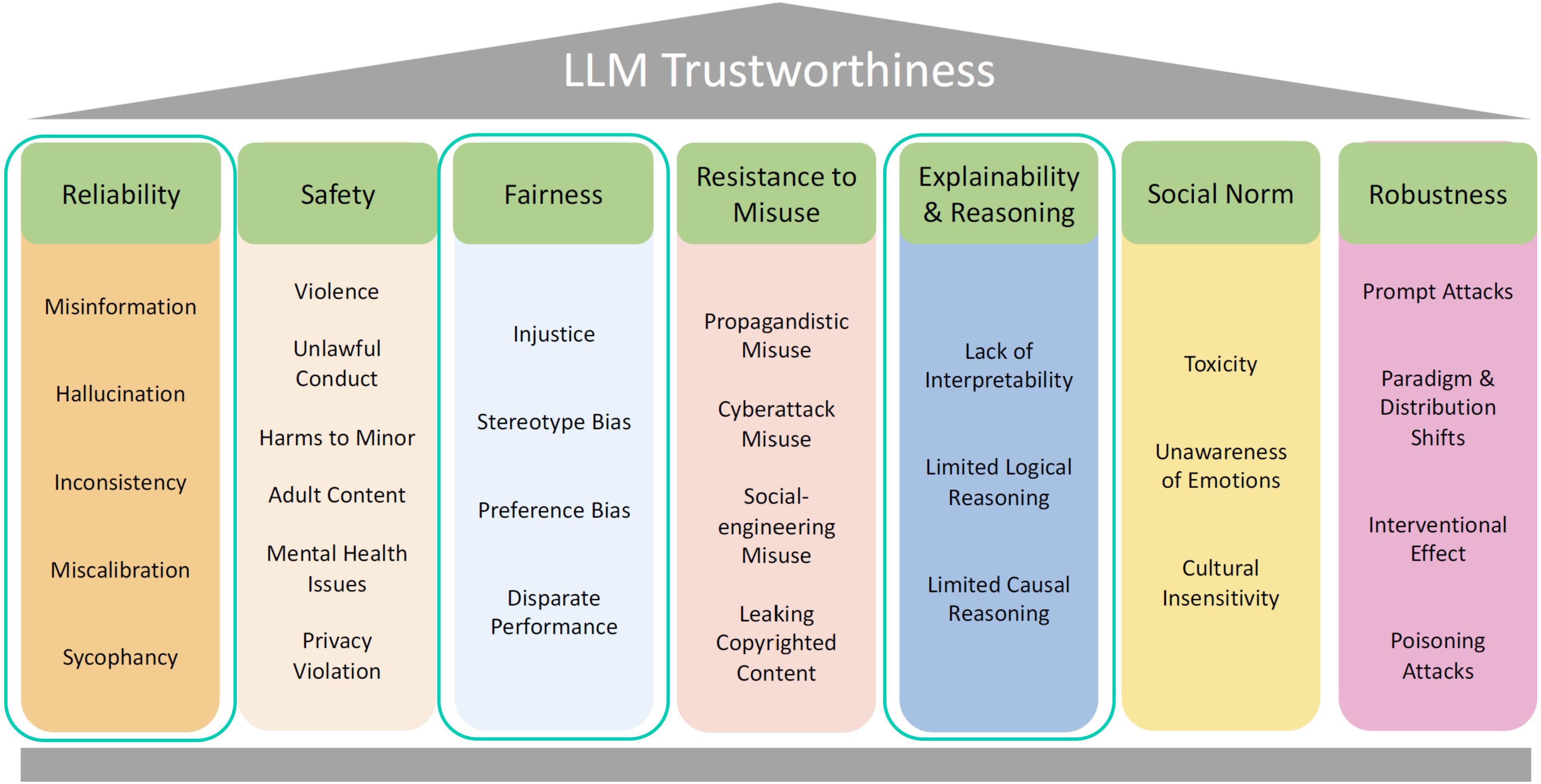
Lindholmen Science Park

AI SWEDEN

NTNU | Norwegian University of Science and Technology

Funded by the European Union

LLM Trustworthiness



LLM Trustworthiness

Reliability
- Misinformation
- Hallucination
- Inconsistency
- Miscalibration
- Sycophancy

Safety
- Violence
- Unlawful Conduct
- Harms to Minor
- Adult Content
- Mental Health Issues
- Privacy Violation

Fairness
- Injustice
- Stereotype Bias
- Preference Bias
- Disparate Performance

Resistance to Misuse
- Propagandistic Misuse
- Cyberattack Misuse
- Social-engineering Misuse
- Leaking Copyrighted Content

Explainability & Reasoning
- Lack of Interpretability
- Limited Logical Reasoning
- Limited Causal Reasoning

Social Norm
- Toxicity
- Unawareness of Emotions
- Cultural Insensitivity

Robustness
- Prompt Attacks
- Paradigm & Distribution Shifts
- Interventional Effect
- Poisoning Attacks

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

Funded by the European Union

TrustLLM Objectives (1/2)

- **Improving the Factual Correctness of LLMs**
  - New methodology to include structured factual information into LLMs.
  - Large knowledge databases integrated in the process to incorporate structural knowledge (e.g., knowledge graphs, databases, APIs) in the LLM-training
  - Successful benchmarking of the new methodology: error reduction of common NLP tasks related to factual languages.
- **Providing a Framework for Modular and Transferable Multilingual LLM Training**
  - correctness, across A family of 3-5 open source, general purpose and pre-trained LLMs
  - Two models transferred to low-resource target languages.
  - Excellent language modelling score (low perplexity) on 3 low-resource languages with only a small available training corpus.
- **Development of Sustainable and Trustworthy LLMs aligned with European Values**
  - Novel methods of aligning LLMs, while balancing increased robustness, reduced bias, and increased efficiency.
  - Development of a multi-lingual benchmarks suite to measure alignment with European values
  - Publication of a dataset together with methods for fine-tuning LLMs aligned with user expectations (e.g., instructions) or creator intentions (e.g., safe language).

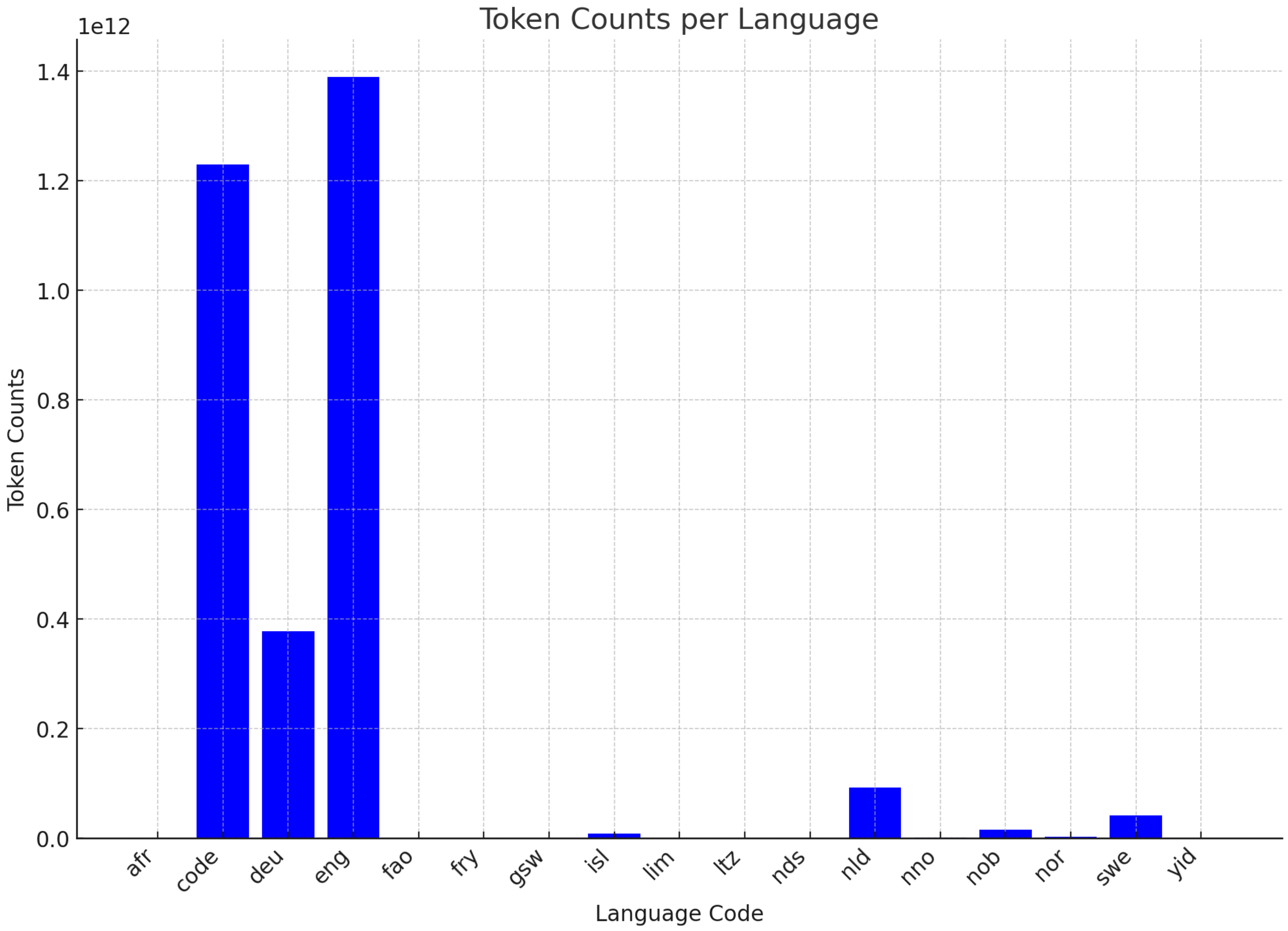
Funded by the European Union

TrustLLM Objectives (2/2)

- **Establishing a European Ecosystem for LLMs**
  - GeTT events: European events on LLM/GPAI with industry & academia
  - NeTT events: topical workshops with experts on LLM (research topics, infrastructure, etc.).
  - A Concept to establish a sustainable European ecosystem based on the LEAM initiative combining national and European funding and networks structures.
- **Assessment of LLM by Use Cases Demonstration and Holistic Benchmarking**
  - Provision of NLP benchmarks in Germanic languages
  - Technical realization of use cases based on TrustLLM.
  - Applications demonstrating the versatility of TrustLLM solutions/models
- **Large Scale Data Management for LLM**
  - Access to text corpora with 1-2 trillion tokens in 6 Germanic languages.
  - Efficient data processing pipelines with a throughput of 100 billion tokens/day.
  - Access to a large, distributed storage system with least 2 PB capacity.

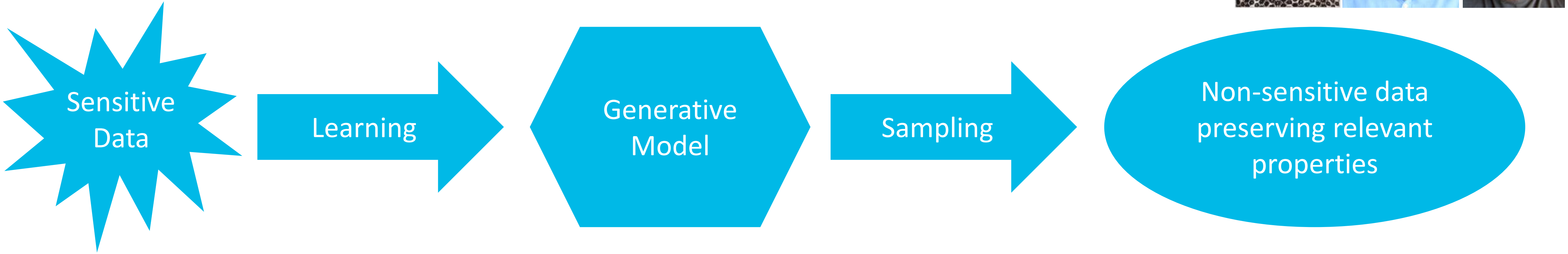
Funded by the European Union

Dataset Composition - TrustLLM



Funded by the European Union

Privacy-preserving synthetic data generation
[R. Ramachandranpillai, Md F. Sikder, D. Bergström]



1. Learn a generative model that captures the probability distribution of the sensitive data
2. Create a synthetic data set from the generative model that both captures the salient features of the original data set **and** is non-sensitive
3. Methods for verifying that the synthetic data set is accurate enough
4. Methods for verifying that the synthetic data set is non-sensitive

Fair Latent Deep Generative Models (FLDGMs) for Syntax-Agnostic and Fair Synthetic Data Generation, *Resmi Ramachandranpillai*, Md Fahim Sikder*, Fredrik Heintz*, ECAI23
Bt-GAN: Generating Fair Synthetic Healthdata via Bias-transforming Generative Adversarial Networks, *Resmi Ramachandranpillai, Md Fahim Sikder, David Bergström, Fredrik Heintz,* Accepted to JAIR.

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

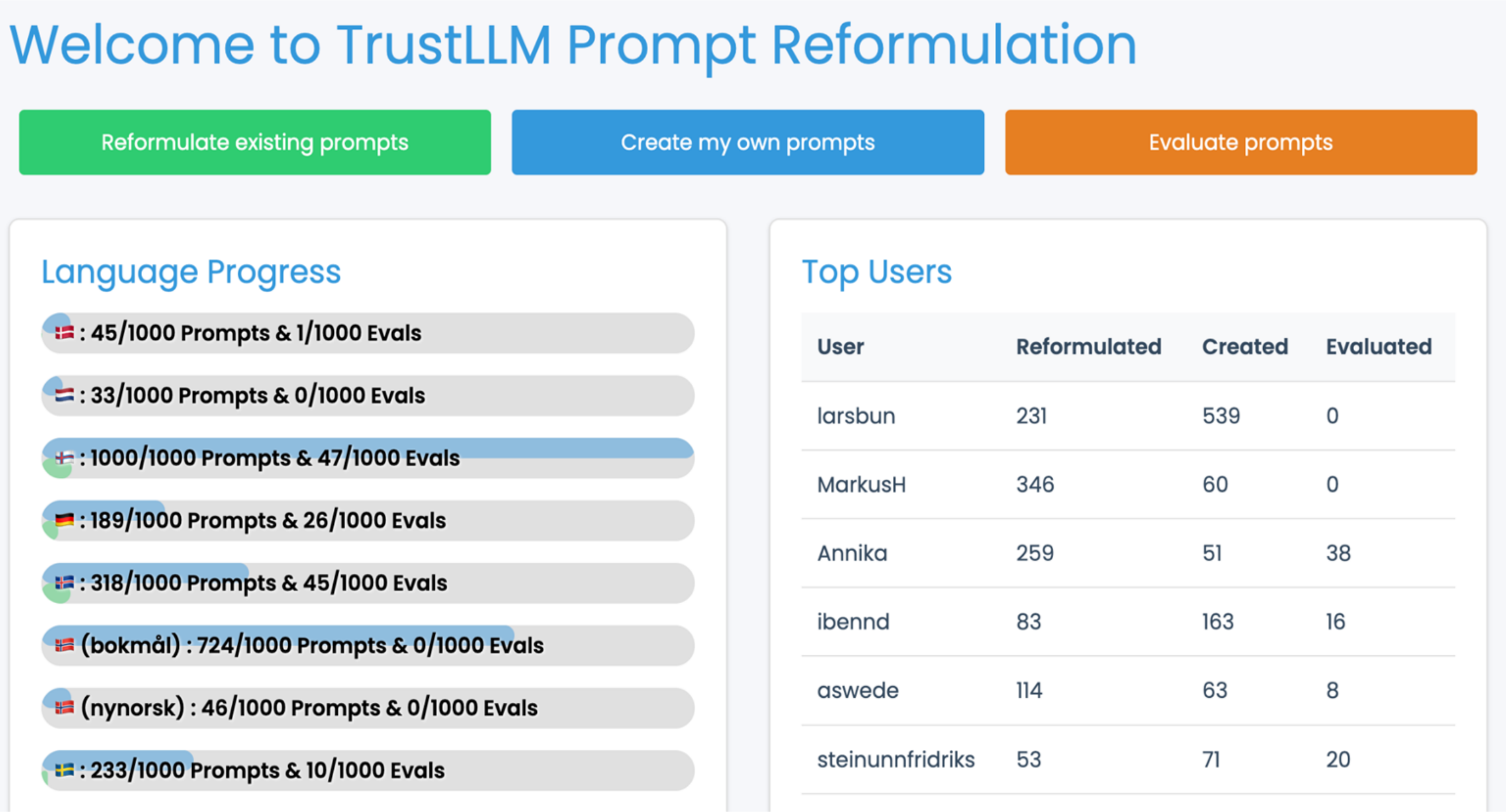
Learning from task-specific examples

3. Preference tuning

Feedback and reinforcement

Funded by the European Union

Alignment Data Collection



Welcome to TrustLLM Prompt Reformulation

Language Progress

- 🇩🇰 : 45/1000 Prompts & 1/1000 Evals
- 🇳🇱 : 33/1000 Prompts & 0/1000 Evals
- 🇫🇴 : 1000/1000 Prompts & 47/1000 Evals
- 🇩🇪 : 189/1000 Prompts & 26/1000 Evals
- 🇮🇸 : 318/1000 Prompts & 45/1000 Evals
- 🇳🇴 (bokmål) : 724/1000 Prompts & 0/1000 Evals
- 🇳🇴 (nynorsk) : 46/1000 Prompts & 0/1000 Evals
- 🇸🇪 : 233/1000 Prompts & 10/1000 Evals

Top Users

User	Reformulated	Created	Evaluated
larsbun	231	539	0
MarkusH	346	60	0
Annika	259	51	38
ibennd	83	163	16
aswede	114	63	8
steinunnfridriks	53	71	20

Funded by the European Union

Our Crowdsourcing Platform

- This is a web app for reformulating the English prompts from the *Open Assistant Dataset* into the following Germanic languages:
  - German, Dutch, Swedish, Norwegian, Danish, Icelandic and Faroese

- We have initially only been annotating within the TrustLLM project.

- You need an invite key in order to join. Contact Annika by email ans72@hi.is or annika@hi.is to get one.

Funded by the European Union

Our Crowdsourcing Platform

- The goal is to reformulate 1,000 prompts for each language.
- These prompts will be used in a second phase of creating alignment data.
- We will publish the final dataset, open-source.

Funded by
the European Union

The tasks

- Prompt reformulation
  - Use an English prompt as inspiration and reformulate it in your own language, making it culturally appropriate and naturally-sounding linguistically.

- Original prompt creation
  - Create your own high-quality prompt in your own language.
  - Not only QA; also summarization, bug fixing, idea generation etc.

- Prompt evaluation
  - Evaluate prompts written by other people in your own language using labels and scales.

Funded by the European Union

How can we evaluate LLMs?

Funded by
the European Union

Four Main Approaches

Gut Feeling Approach

LLM-as-a-judge

Arena Approach

Benchmark Approach

What is ScandEval?

Funded by
the European Union

ScandEval is a robust multilingual benchmarking framework

Funded by the European Union

ScandEval is a robust multilingual benchmarking framework

Funded by the European Union

Language Model Benchmarking Framework

- Enables evaluation of implicit language understanding and generation capabilities of language models
- Allows evaluation of *both* encoders through finetuning, and decoders through few-shot evaluation
  - It has been shown that few-shot inference of decoder models corresponds exactly to finetuning [1]
  - This thus allows us to compare encoders with decoders directly

[1] von Oswald et al. arXiv preprint arXiv:2309.05858 (2023)

Funded by the European Union

Language Model Benchmarking Framework

- A large focus of the framework is ease of use
- The framework can simply be installed:

```
$ pip install scandeval[all]
```

- Models can easily be evaluated:

```
$ scandeval --model <model-id>
```

- Supports models on the Hugging Face Hub, local models and OpenAI models

Funded by
the European Union

ScandEval is a robust multilingual benchmarking framework

Funded by the European Union

Evaluation Robustness

- When evaluating models, there are several sources of noise in the evaluation result:
  - The choice of training examples (=few-shot examples when evaluating decoder models)
  - The choice of test examples
  - The stochastic elements (stochastic gradient descent when evaluating encoders, sampling when evaluating decoders)
- The training and test examples are bootstrapped 10 times, yielding a more reliable estimation of the true mean
  - Asymptotically correct by the bootstrap theorem
- We enforce that the stochastic elements are deterministic

Funded by the European Union

ScandEval is a robust multilingual benchmarking framework

Funded by the European Union

Multilingual Evaluations

- Currently, the main Germanic languages are natively supported:
  - Mainland Scandinavian languages (Danish, Swedish, Norwegian)
  - Insular Scandinavian languages (Icelandic, Faroese)
  - German
  - Dutch
  - English
- Aside from including evaluation datasets in these languages, the prompts used when evaluating decoder models are also localised to the given language

Funded by the European Union

Which Tasks are Included?

Funded by
the European Union

Tasks in ScandEval

Natural Language Understanding (NLU) Tasks

1. Text classification
2. Linguistic acceptability
3. Extractive question answering
4. Named entity recognition

Funded by the European Union

Tasks in ScandEval

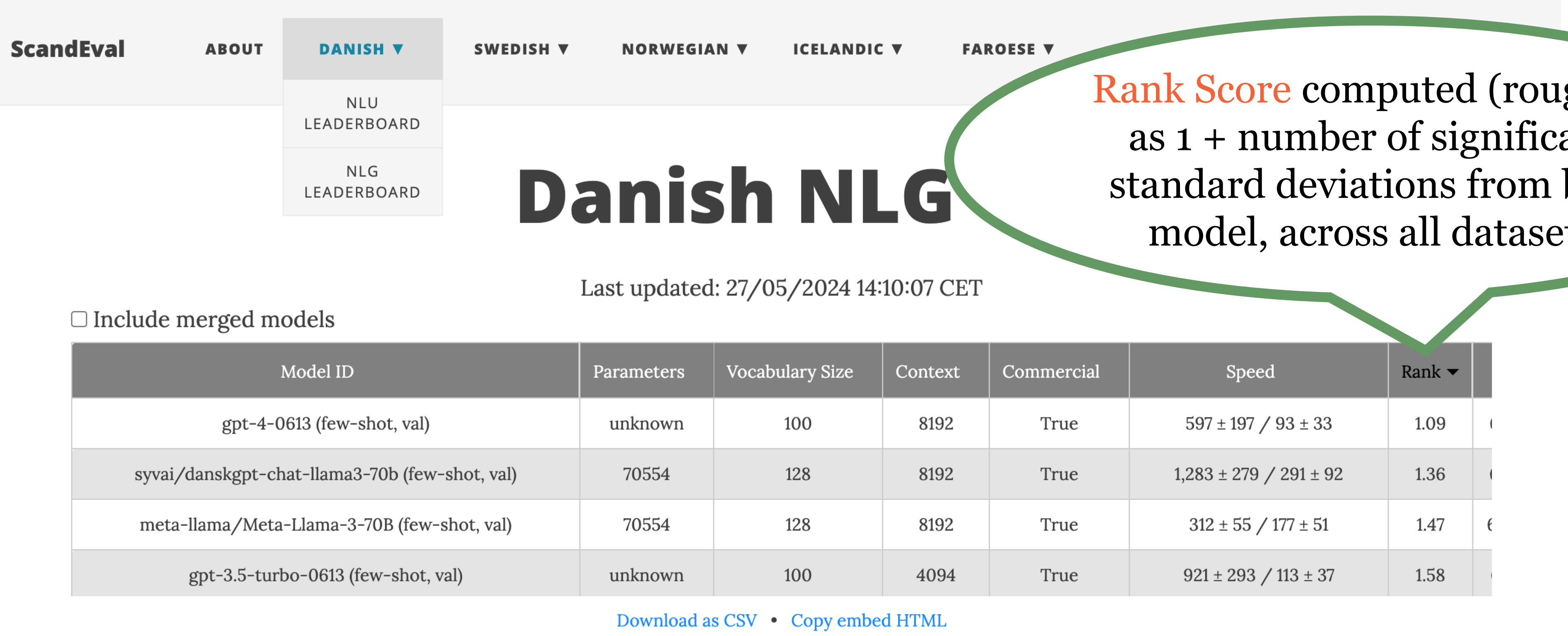
Natural Language Generation (NLG) Tasks

1. Text classification
2. Linguistic acceptability
3. Extractive question answering
4. Named entity recognition
5. Summarisation
6. World knowledge
7. Common-sense reasoning

Funded by the European Union

Online Leaderboards

scandeval.com



Danish NLG

Last updated: 27/05/2024 14:10:07 CET

□ Include merged models

Model ID	Parameters	Vocabulary Size	Context	Commercial	Speed	Rank ▾
gpt-4-0613 (few-shot, val)	unknown	100	8192	True	597 ± 197 / 93 ± 33	1.09
syvai/danskgpt-chat-llama3-70b (few-shot, val)	70554	128	8192	True	1,283 ± 279 / 291 ± 92	1.36
meta-llama/Meta-Llama-3-70B (few-shot, val)	70554	128	8192	True	312 ± 55 / 177 ± 51	1.47
gpt-3.5-turbo-0613 (few-shot, val)	unknown	100	4094	True	921 ± 293 / 113 ± 37	1.58

Download as CSV • Copy embed HTML

Rank Score computed (roughly) as 1 + number of significant standard deviations from best model, across all datasets

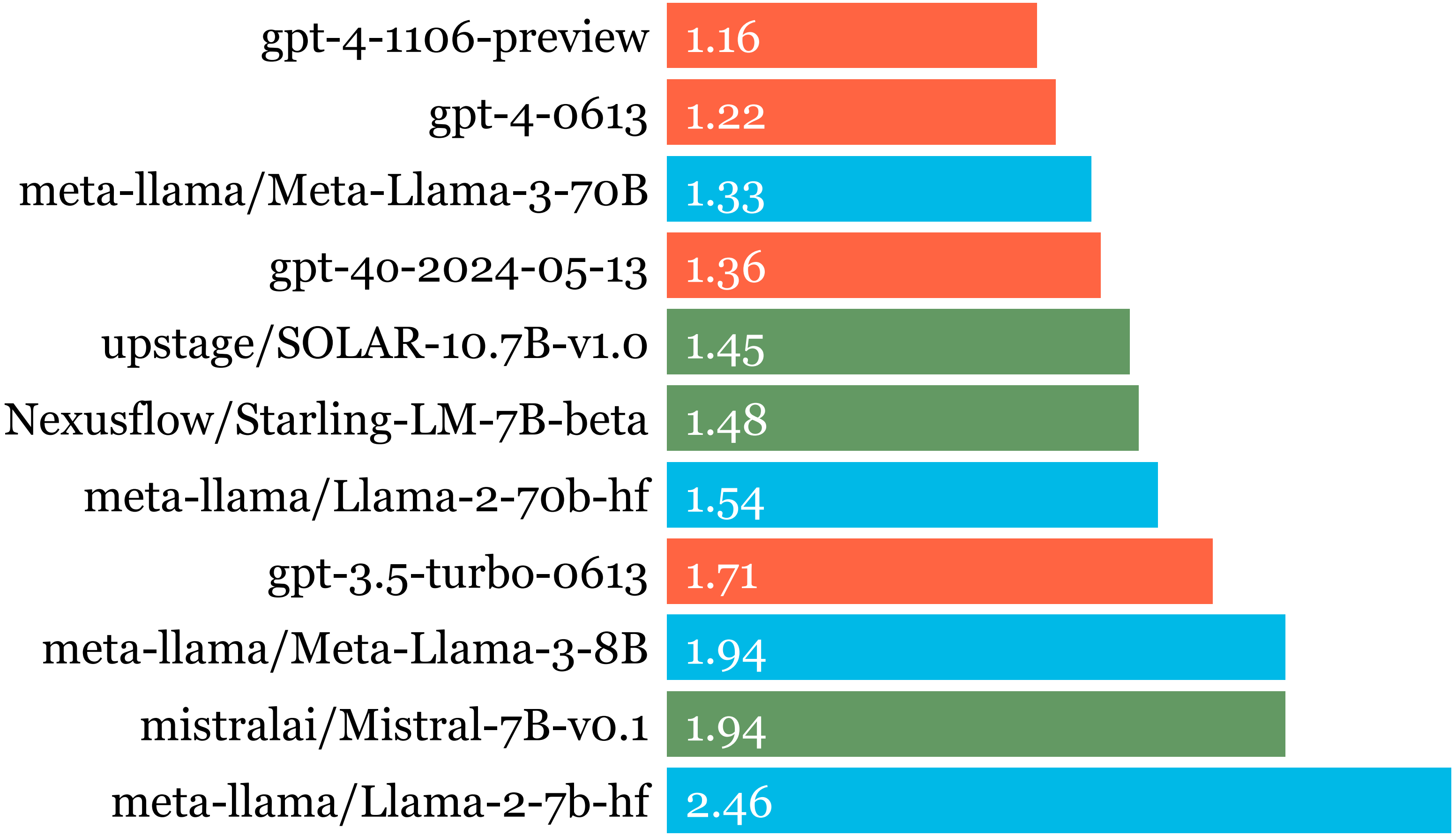
What are some of the best performing European models?

TrustLLM

Funded by the European Union

English ScandEval Rank

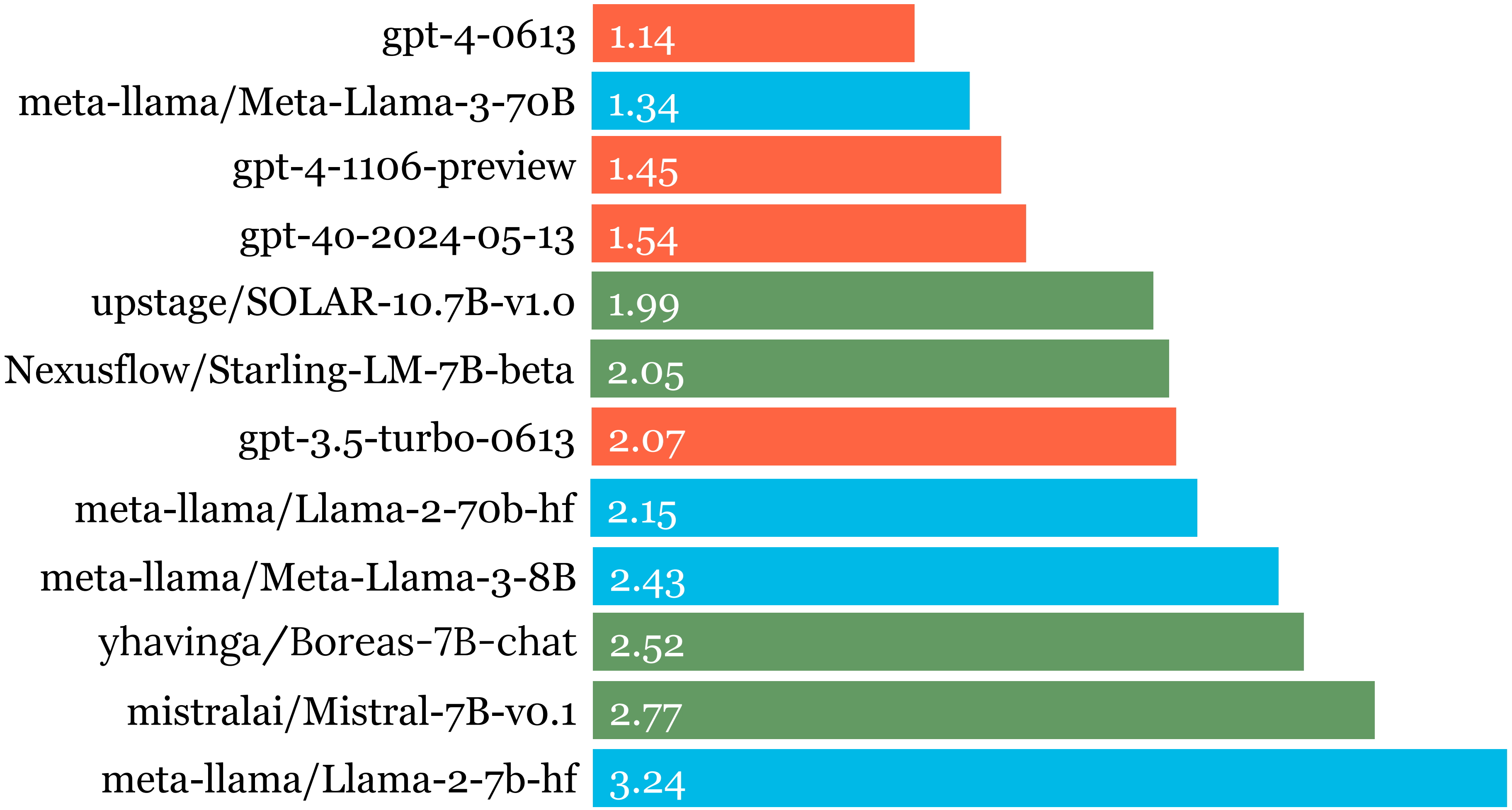
Lower is better



Funded by the European Union

Dutch ScandEval Rank

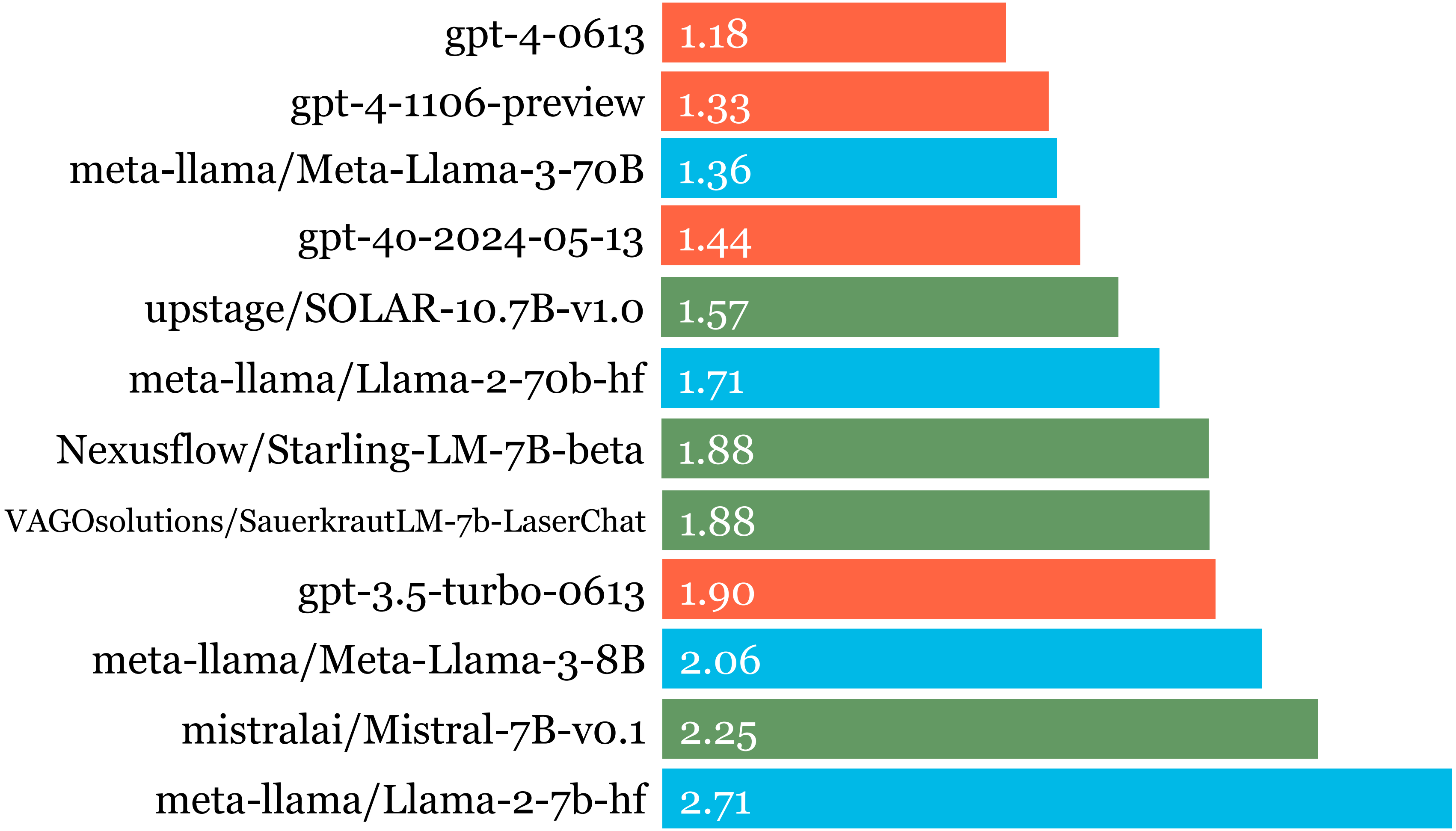
Lower is better



Funded by the European Union

German ScandEval Rank

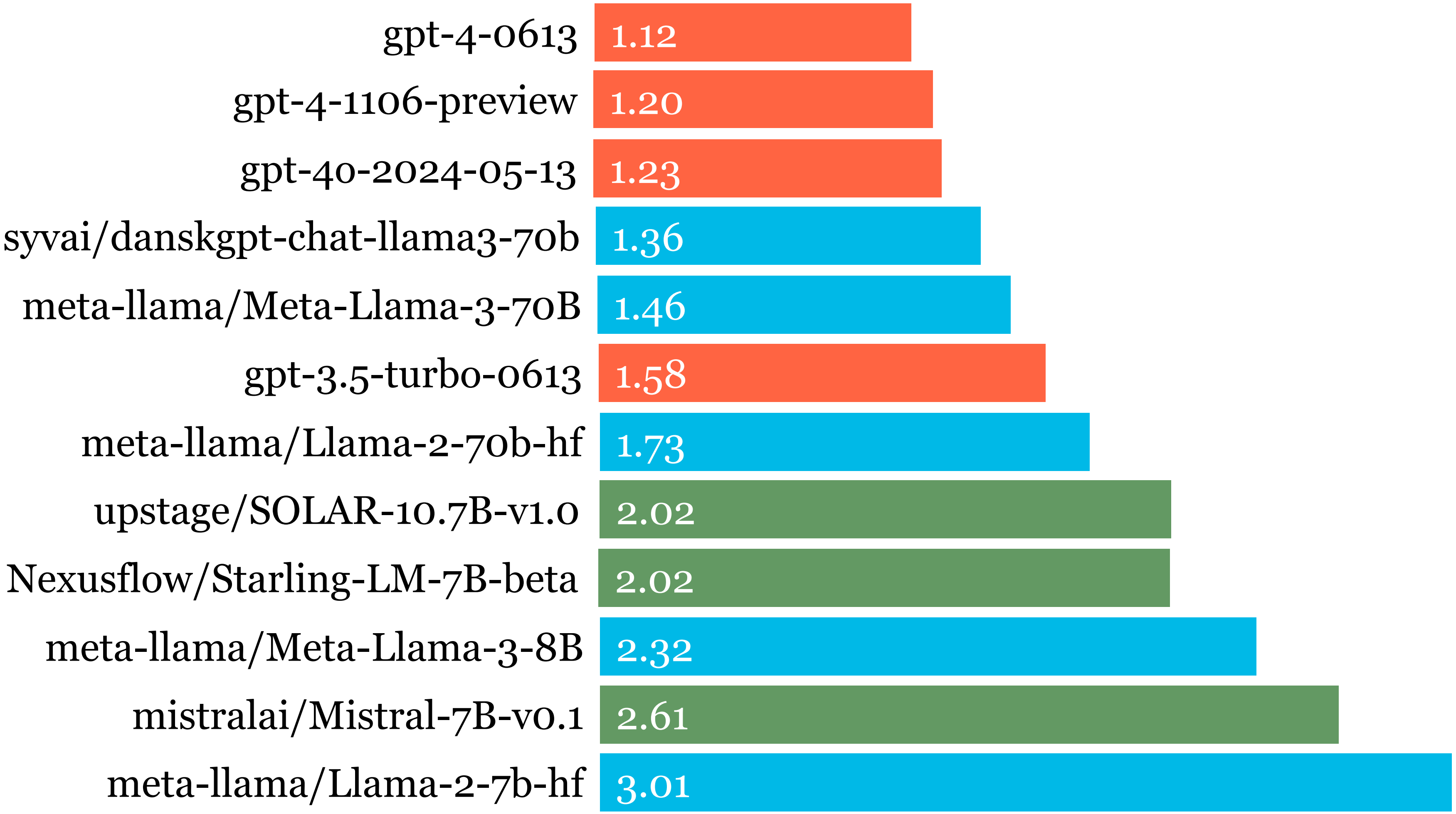
Lower is better



Funded by
the European Union

Danish ScandEval Rank

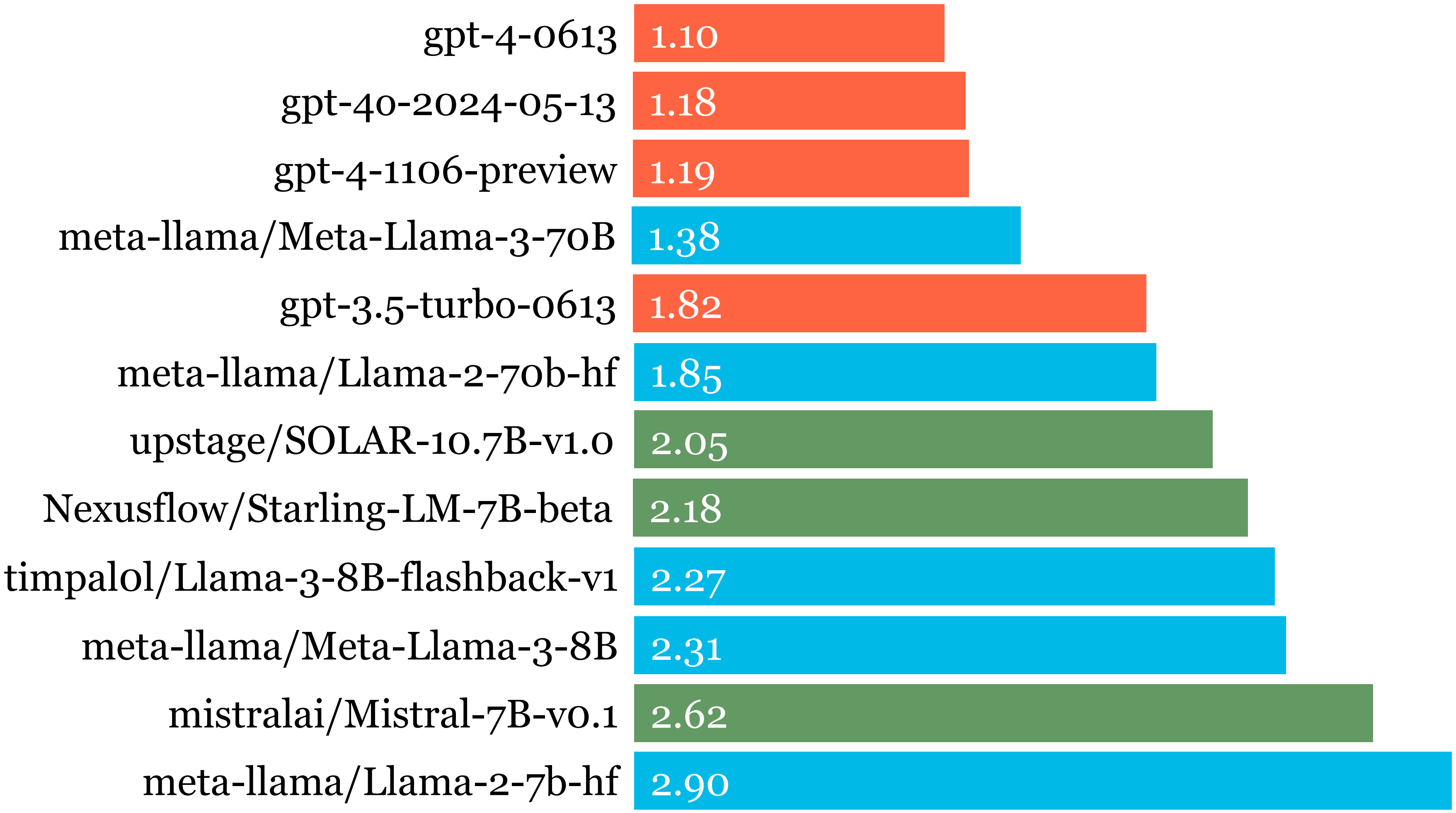
Lower is better



Funded by the European Union

Swedish ScandEval Rank

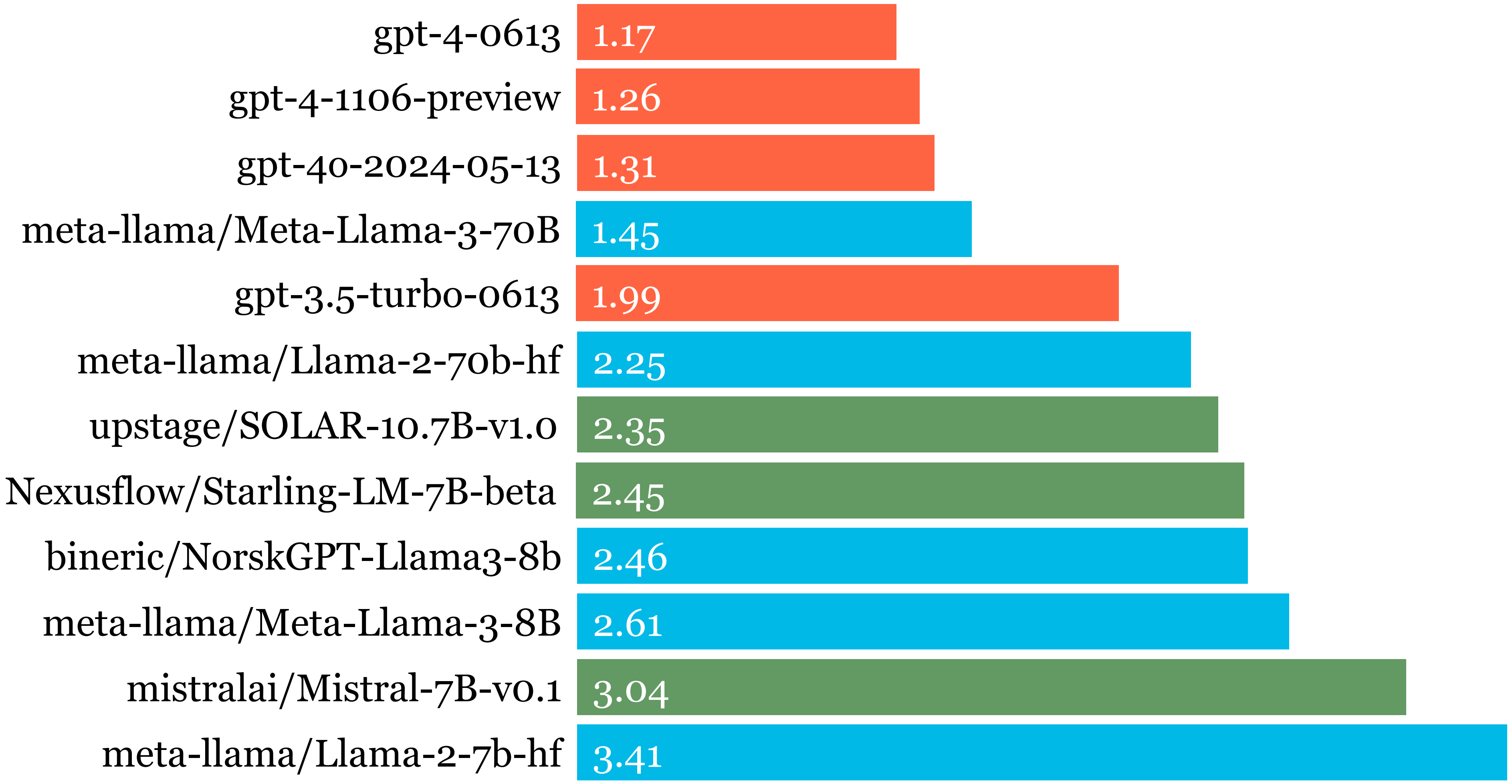
Lower is better



Funded by the European Union

Norwegian ScandEval Rank

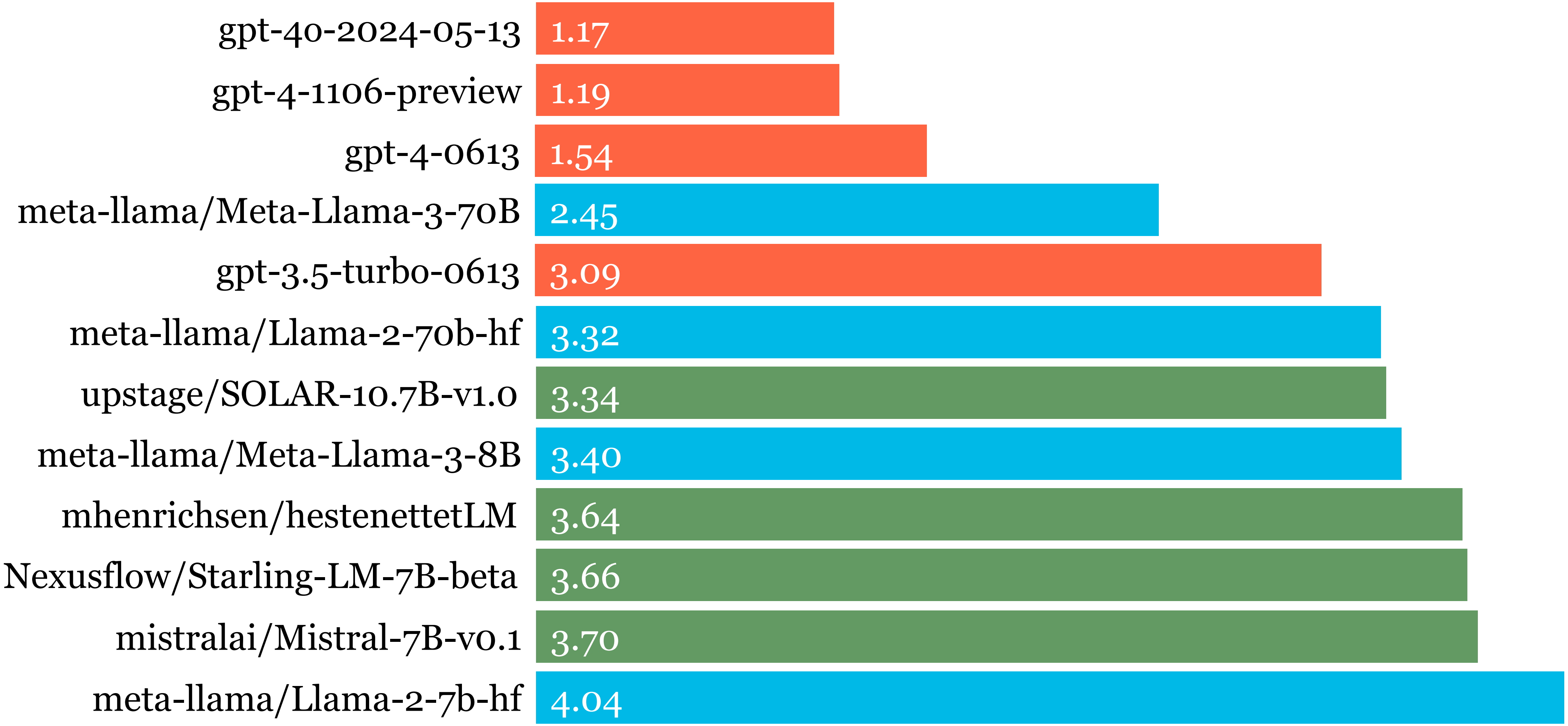
Lower is better



Funded by the European Union

Icelandic ScandEval Rank

Lower is better



Funded by the European Union

Papers

ScandEval NLU benchmark for encoders:

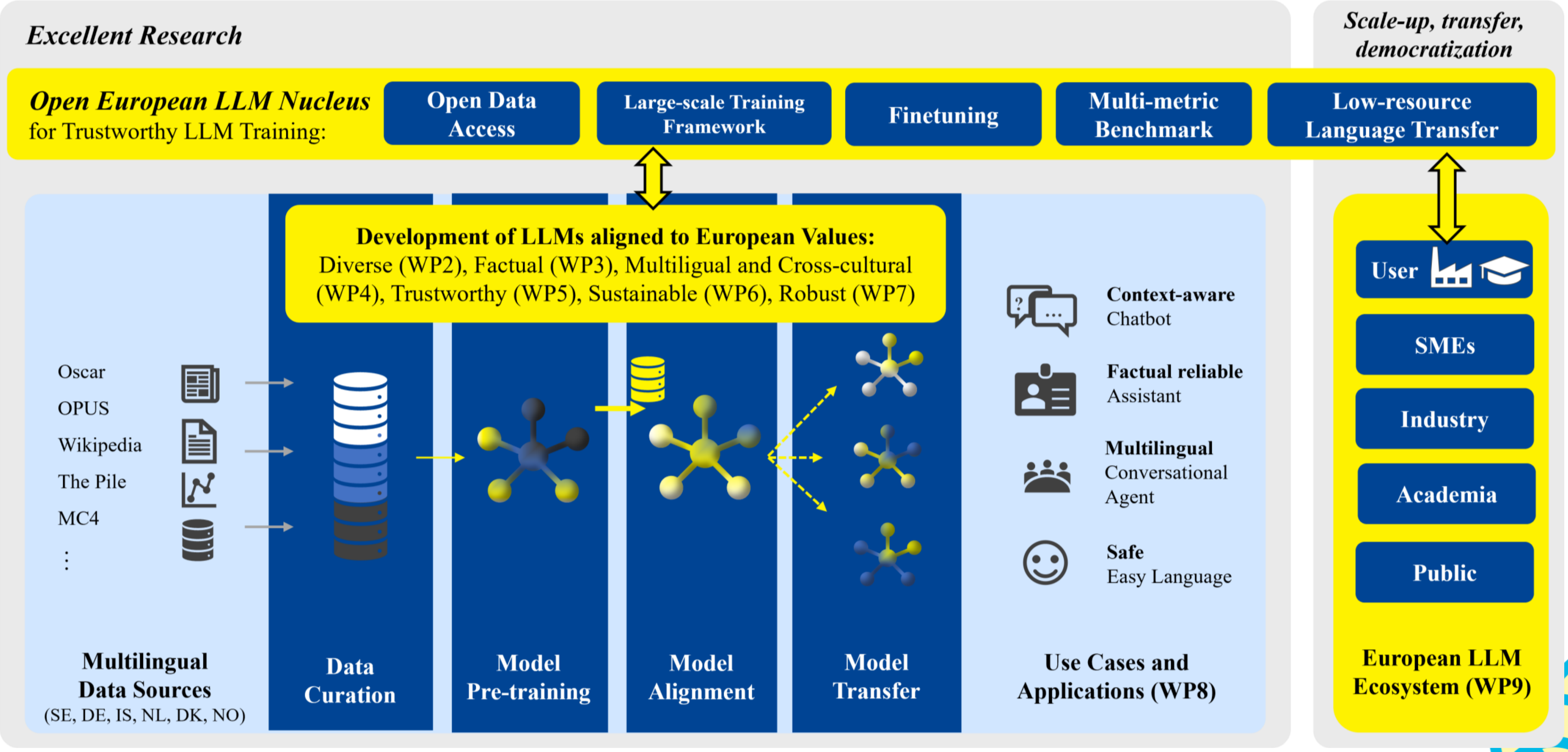
Nielsen, Dan. *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. 2023

ScandEval NLU benchmark for decoders:

Joint work with Kenneth Enevoldsen (Aarhus University) and Peter Schneider-Kamp (University of Southern Denmark). Preprint out soon.

Funded by the European Union

TrustLLM: Project Concept



Excellent Research

Scale-up, transfer, democratization

Open European LLM Nucleus for Trustworthy LLM Training:

Open Data Access | Large-scale Training Framework | Finetuning | Multi-metric Benchmark | Low-resource Language Transfer

Development of LLMs aligned to European Values: Diverse (WP2), Factual (WP3), Multiligual and Cross-cultural (WP4), Trustworthy (WP5), Sustainable (WP6), Robust (WP7)

Multilingual Data Sources (SE, DE, IS, NL, DK, NO)

Data Curation

Model Pre-training

Model Alignment

Model Transfer

Context-aware Chatbot

Factual reliable Assistant

Multilingual Conversational Agent

Safe Easy Language

Use Cases and Applications (WP8)

User

SMEs

Industry

Academia

Public

European LLM Ecosystem (WP9)

Funded by the European Union

Trends

- Will the LLMs continue to grow?
  - Smaller and more effective models
  - More specifalized models
- Multi-modal models, initially text, image and sound
- Generalize from language models to AI models, world models grounded in reality
- Neurosymbolic AI
- Testing and evaluation
- Trustworthy Human-Centered AI

Funded by the European Union

Coordinator: Fredrik.Heintz@liu.se ; Project manager: Trine.Platou@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

https://trustllm.eu/

Funded by the European Union