Dutch Authority for Digital Infrastructure
Ministry of Economic Affairs and Climate Policy

Generative AI: security paradigm shift

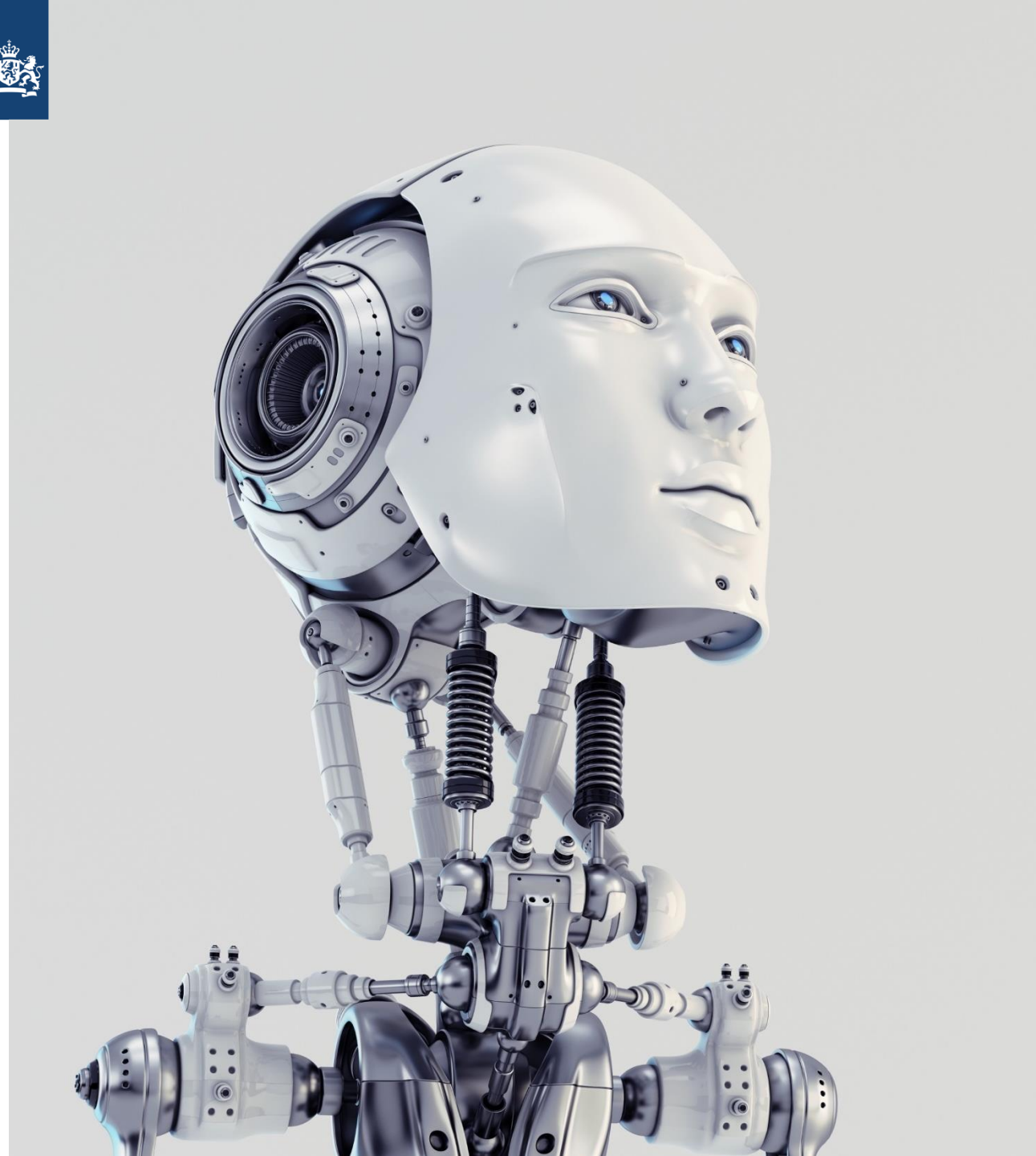
ENISA AI Cybersecurity Conference

Huub Janssen

7 juni 2023

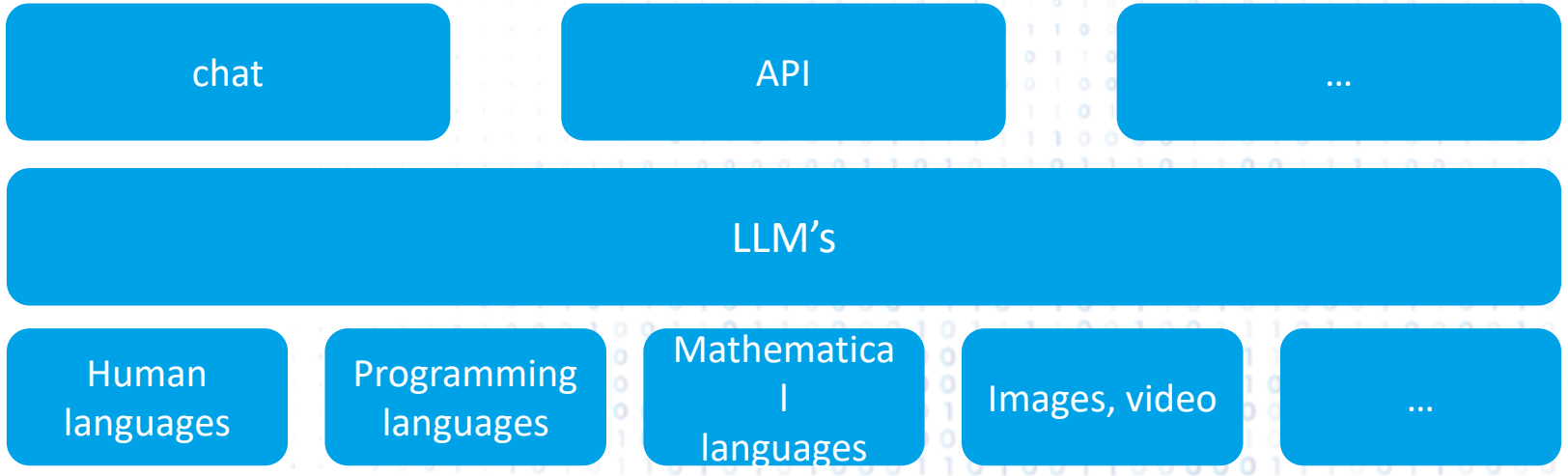
Content

1. What’s new
2. Paradigm shift



1. What's new?

- LLM's as foundation model



- Cheaper and faster to develop AI based upon a foundation model. Foundation models might become the base for most AI systems

Emerging capabilities

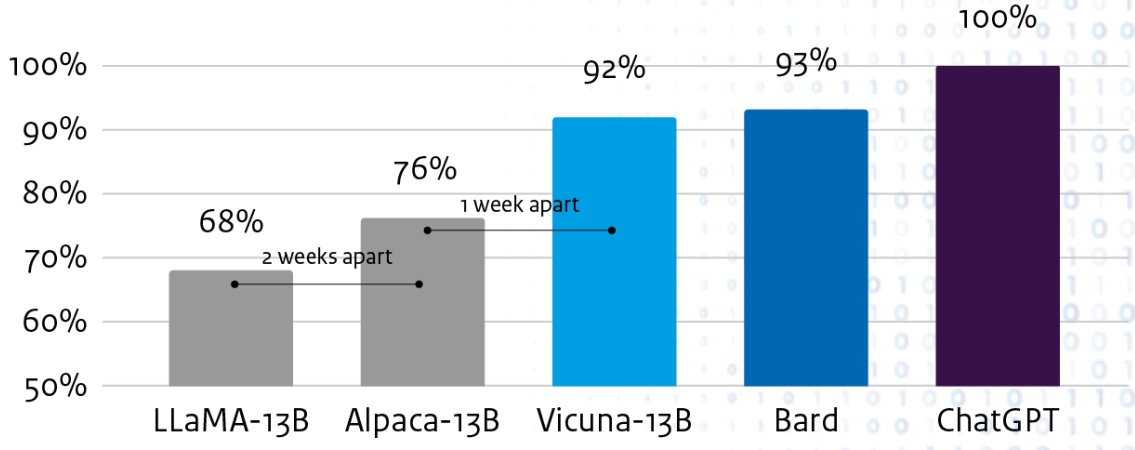
Increased models (data parameters) show unexpected capabilities

Agency

- LLM can ACT in cyber and real world
- Big Tech’s LLM’s are very restricted to prevent misuse, but restrictions can be bypassed
    - Jailbreaking
    - Building your own LLM’s

Open source at warp speed

- Quality almost as good
- It can run on a laptop



*GPT-4 grades LLM outputs. Source: https://vicuna.lmsys.org

github.com/eugeneyan/open-llms

eugeneyan Merge pull request #34 from Jacksonlark/main ... 9449f74 last week 76 commits

LICENSE	Initial commit	3 weeks ago
README.md	fix format	last week

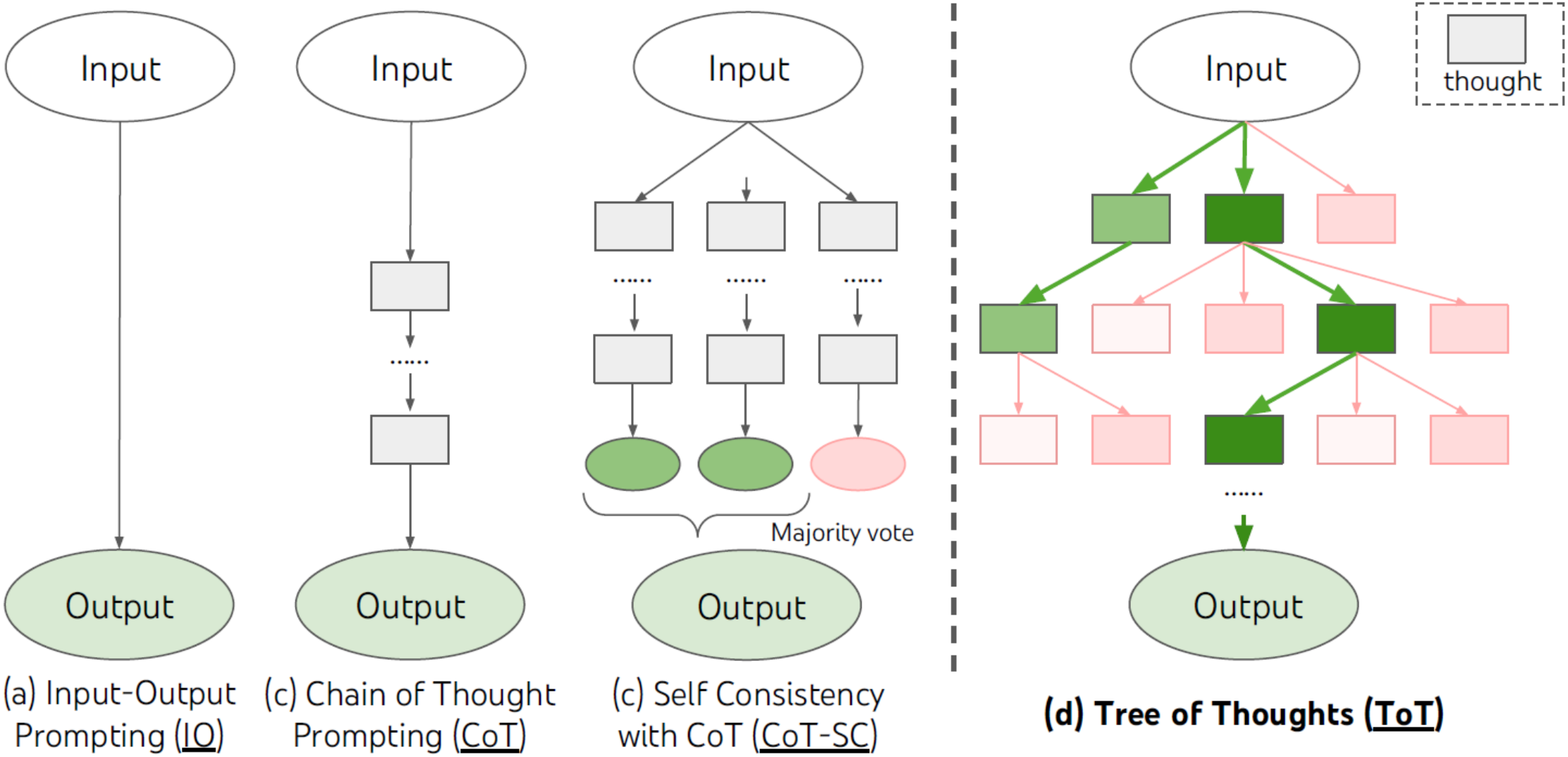
README.md

Open LLMs

These LLMs are all licensed for commercial use (e.g., Apache 2.0, MIT, OpenRAIL-M). Contributions welcome!

Language Model	Release Date	Checkpoints	Paper/Blog	Params (B)	Context Length	Licence
T5	2019/10	T5 & Flan-T5, Flan-T5-xxl (HF)	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer	0.06 - 11	512	Apache 2.0
UL2	2022/10	UL2 & Flan-UL2, Flan-UL2 (HF)	UL2 20B: An Open Source Unified Language Learner	20	512, 2048	Apache 2.0
Cerebras-GPT	2023/03	Cerebras-GPT	Cerebras-GPT: A Family of Open, Compute-efficient, Large Language Models (Paper)	0.111 - 13	2048	Apache 2.0
Open Assistant (Pythia family)	2023/03	OA-Pythia-12B-SFT-8, OA-Pythia-12B-SFT-4, OA-Pythia-12B-SFT-1	Democratizing Large Language Model Alignment	12	2048	Apache 2.0
Pythia	2023/04	pythia 70M - 12B	Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling	0.07 - 12	2048	Apache 2.0

Orchestration



Paradigma shift

End of updates

Different security system

Humans more vulnerable

Agents

Powerful AI uncontrollable

Digital sovereignty

LLM’s can quickly abuse vulnerabilities

LLM’s can write exploits

End of updates

Speed of AI makes updates too slow

Humans are too slow to act

Standards, education, protocols need to change

Alternatives for updates?

Symptomatic for other cybersecurity methods

The threats are more advanced: more expensive measures

Risks evolve quickly: very frequent risk analysis are necessary

Different security system

Type of attack can change: different measures needed

Different expertise is needed

AI security must be able to act automatically

Dependancy on AI systems grows: impact?

AI can decides to shut down services

True sounding voicemails from colleagues or friends

Personalised fishing mails

Hallucination/ fragmentation reality

Humans more vulnerable

Well known awareness-tools become useless

Humans are slow and don't always want to think

Feeling of loss of control

Easier to hack a person than a system

Agents

Act without explicit assignment

Act without mandate: new legal framework

Agents don’t check on local legislation

Current security measures insufficient (e.g. KYC)

Act illegal without responsible person knowing

Impact on risk assessment

Legal responsibility

LLM’s without constraints

Opensource LLM’s on a laptop

Powerful AI uncontrollable

Wanted and unwanted acting of AI

How to supervise?

Hardware in 2 years: 1000 x faster

AI Act: limited useful for these risks

Many unknown unknows

Data and control outside of EU

Most AI systems based on foundation models

Digital sovereignty

Dependency on foundation models outside EU?

Unknown ethics in foundation models

Governments might have more access the others

AI based on LLM's:
ethics from US, China, India

"Who controls AI,
controls the world order"

For a **safe and connected** Netherlands

www.rdi.nl

info@rdi.nl

huub.janssen@rdi.nl

www.twitter.com/wijzijnRDI
www.linkedin.com/company/wijzijnRDI
www.instagram.com/wijzijnRDI
www.youtube.com/@wijzijnRDI