The International Conference on Emerging Smart Technology for Sustainable Development (ESTSD-2025)

Running Modern Applications on Old and New Hardware Technologies for Sustainable Computing

Keynote

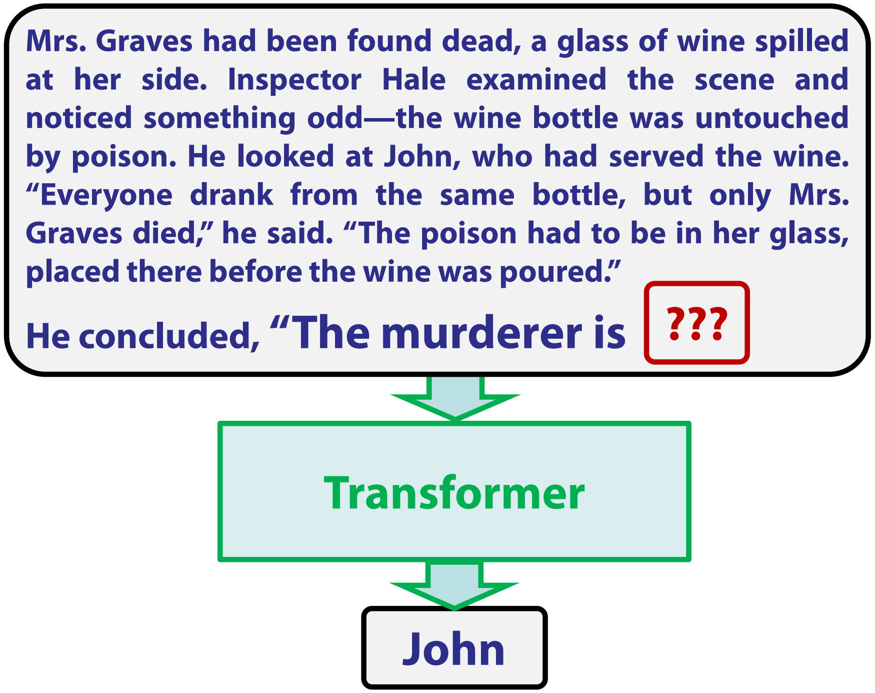
Professor Dr. rer. nat. habil. Sven Groppe
groppe@ifis.uni-luebeck.de

My Research Areas

- **Artificial Intelligence, Machine Learning and Data Science**
  - LLMs, Agentic Workflows, Mathematical Optimizations, Graph Neural Networks, Chatbots, Reasoning
- **Data Management Tasks**
  - Query Processing & Opt., Indexing, Mapping, Compression, Replication, Caching, Transaction Handling
- **Data Models**
  - Knowledge Graphs, Semantic Web, Property Graphs, Relational Data, XML
- **Types of Data**
  - Big Data, Data Streams
- **Emergent Hardware Technologies**
  - Many-Core CPU, GPU, FPGA, Quantum Computer
- **Platforms**
  - Internet, Internet of Things, Cloud, Post-Cloud (Fog/Edge/Dew Computing), P2P, Mobile, Parallel and Main Memory Servers
- **Advanced Applications**
  - Citizen Science, Customer Communications, Pandemics like Covid-19, Software Vulnerability Prediction
- **Sustainability**
  - Sustainable Computing/AI, Applications for Sustainability

Transformer/Large Language Models (LLMs) and Applications

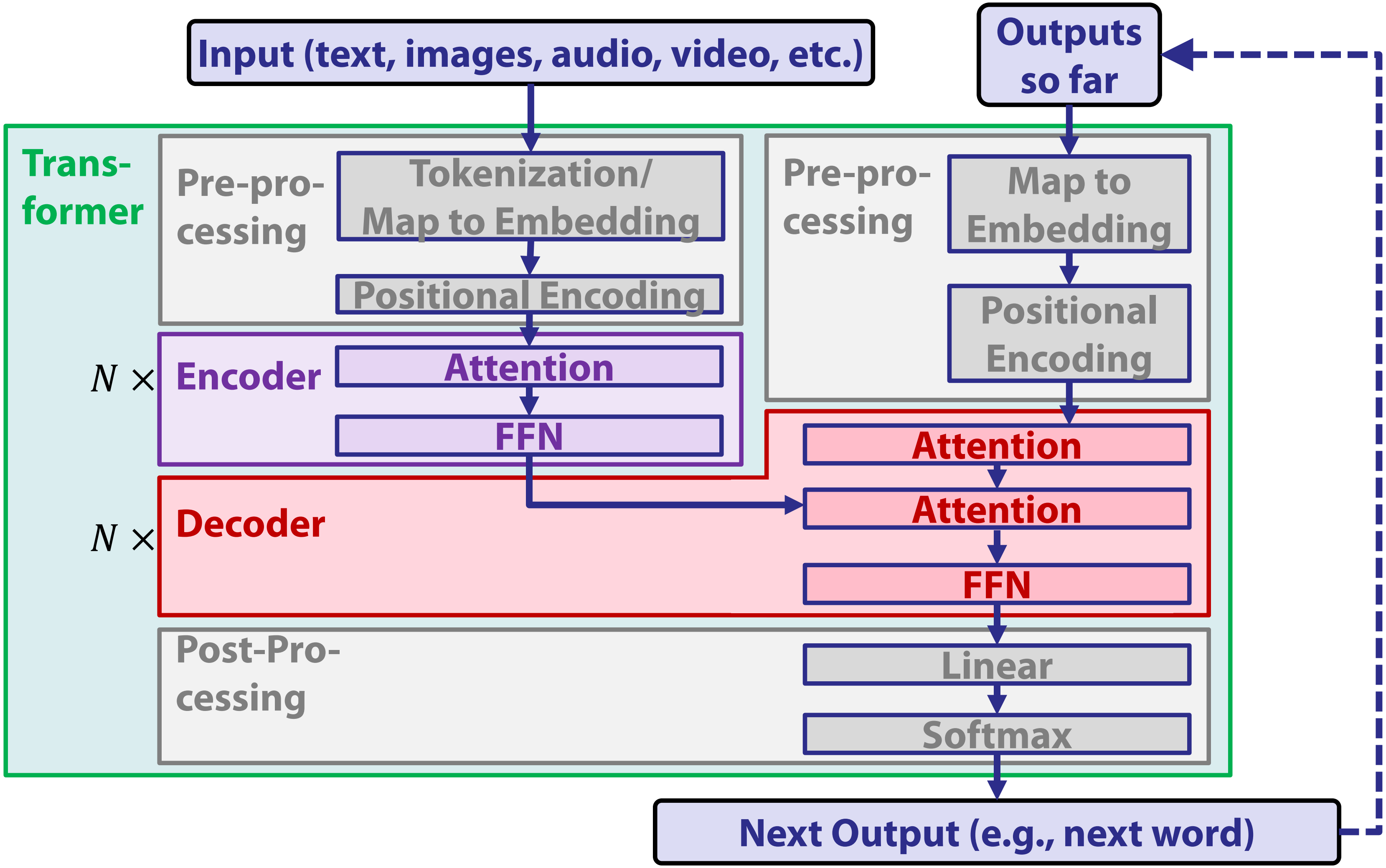
- Successfull architecture of machine learning for the processing of natural-language text
  - Text classification
  - Translation
  - Question answering
  - Text Generation
  - Text completion
  - Text summary
  - …



Mrs. Graves had been found dead, a glass of wine spilled at her side. Inspector Hale examined the scene and noticed something odd—the wine bottle was untouched by poison. He looked at John, who had served the wine. "Everyone drank from the same bottle, but only Mrs. Graves died," he said. "The poison had to be in her glass, placed there before the wine was poured."

He concluded, "The murderer is

Transformer-Application: Chatbots



What People Ask ChatGPT the Most

- Questions about…
  - general knowledge and information
    - definitions, historical events, and scientific facts
  - technology
    - how to use a particular software
    - troubleshoot a technical problem
  - health and medicine
    - symptoms, treatments, and side effects of various conditions
  - current events
    - news updates and breaking news
  - entertainment
    - movie and music recommendations, and reviews
  - personal finance and business
    - investment advice, tax advice, and starting a business.
  - education
    - study tips, test-taking strategies, and career advice
  - travel
    - destination recommendations, visa requirements, and how to plan a trip
  - personal development and self-improvement
    - tips for managing stress, building self-esteem and achieving goals

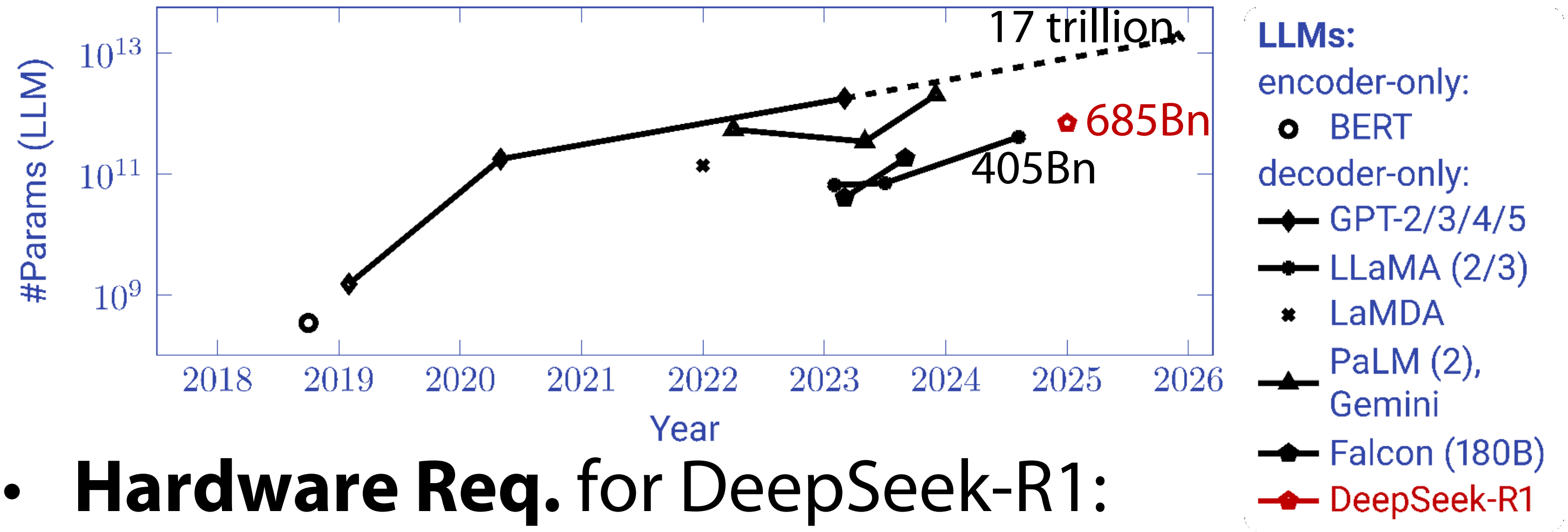
Source: https://medium.com/tech-talk-with-chatgpt/what-people-ask-chatgpt-the-most-fd500eba26e0

MIT Study on the effect of Using ChatGPT/Google on our Brains

- Using only LLMs (like ChatGPT):
  - 83.3% of users couldn't recall even a single sentence from their own texts minutes after writing
  - Neural activity in the brain dropped by 47%
  - The paradox: tasks are completed 60% faster, but the learning effect drops by 32%
- Using search engines (like google):
  - moderate level of cognitive effort and brain activity
    - significantly higher than when using LLMs,
    - but lower than when working without any aids
  - Memory and satisfaction with their own texts were also noticeably better than in group of ChatGPT-only users.

Kosmyna et al. (2025). Your Brain on ChatGPT... arXiv. https://doi.org/10.48550/ARXIV.2506.08872

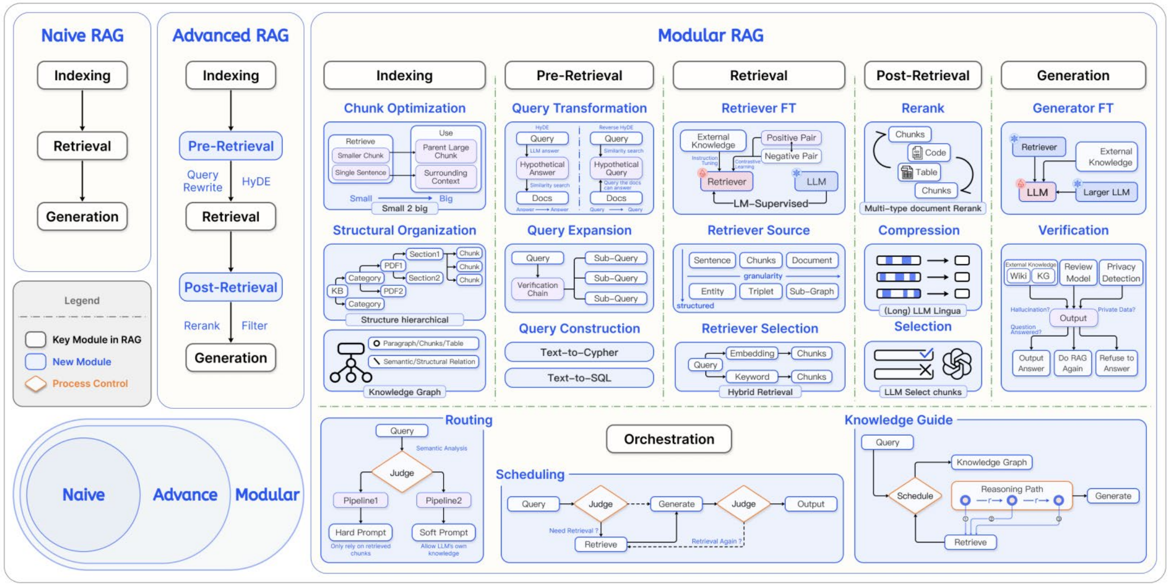
Requirements for LLM Inferences



- **Hardware Req.** for DeepSeek-R1:
  - Minimum: NVIDIA A100 (80GB) with FP8/BF16 precision
  - Recommended: 16x or more NVIDIA H100 80GB GPUs
- **Inference Engines:** Ollama, vLLM, Aphrodite, TGI…
- **Frameworks for Agentic AI:** LangChain/LangGraph, AutoGen, CrewAI, … ⇝ ***x*-times req. for *x*-parallel agents (80 GPUs for 5 parallel LLM agents)**

https://ullyer.medium.com/openais-ambitious-leap-gpt-5-with-17-trillion-parameters-and-beyond-ce989eb6cabb
https://dev.to/askyt/deepseek-r1-671b-complete-hardware-requirements-optimal-deployment-setup-2e48

LLMs with RAG

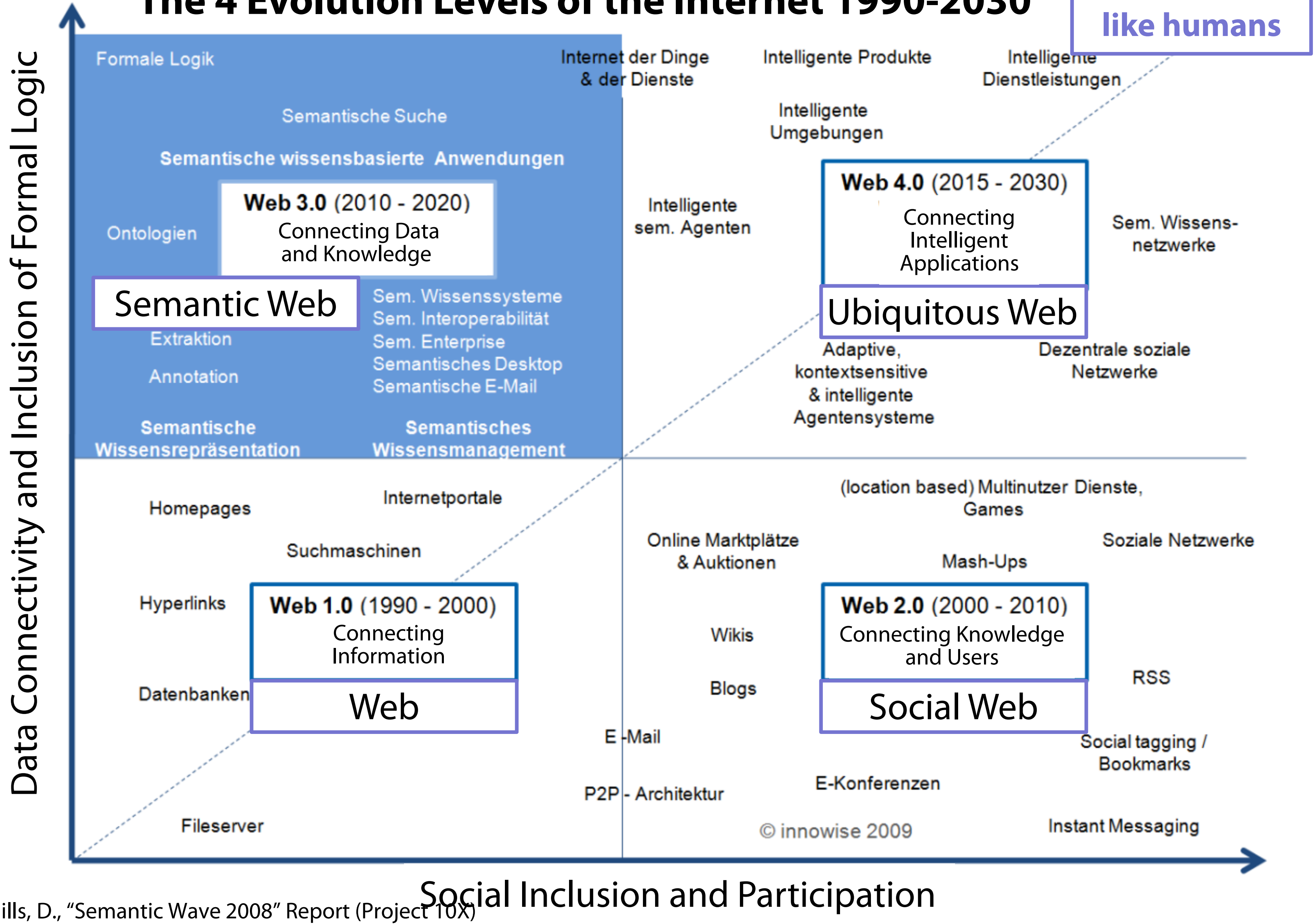


- Retrieval Augmented Generation (RAG):
  - provides an approach to inject vital context to models
  - improves accuracy/reliability of LLMs, avoids hallucinations
  - basic module in many Agentic AI applications
  - Sophisticated RAG methods consume much computing resources

https://doi.org/10.48550/arXiv.2407.21059

Agent Web, which knows, learns and reason like humans

The 4 Evolution Levels of the Internet 1990-2030


Data Connectivity and Inclusion of Formal Logic
Social Inclusion and Participation
Formale Logik
Semantische Suche
Semantische wissensbasierte Anwendungen
Web 3.0 (2010 - 2020)
Connecting Data and Knowledge
Ontologien
Semantic Web
Sem. Wissenssysteme
Sem. Interoperabilität
Sem. Enterprise
Semantisches Desktop
Semantische E-Mail
Extraktion
Annotation
Semantische Wissensrepräsentation
Semantisches Wissensmanagement
Internet der Dinge & der Dienste
Intelligente Produkte
Intelligente Dienstleistungen
Intelligente Umgebungen
Intelligente sem. Agenten
Web 4.0 (2015 - 2030)
Connecting Intelligent Applications
Ubiquitous Web
Sem. Wissens-netzwerke
Adaptive, kontextsensitive & intelligente Agentensysteme
Dezentrale soziale Netzwerke
Homepages
Internetportale
Suchmaschinen
Hyperlinks
Web 1.0 (1990 - 2000)
Connecting Information
Web
Datenbanken
Fileserver
(location based) Multinutzer Dienste, Games
Online Marktplätze & Auktionen
Mash-Ups
Soziale Netzwerke
Wikis
Web 2.0 (2000 - 2010)
Connecting Knowledge and Users
Social Web
RSS
Blogs
E-Mail
P2P - Architektur
E-Konferenzen
Social tagging / Bookmarks
Instant Messaging
© innowise 2009

Adapted from Mills, D., "Semantic Wave 2008" Report (Project 10X)

Bild: Think!innowise Trend Report: Die Zukunft des Semantic Web, http://www.innowise.de/sites/default/files/pubs/Trendreport.pdf

Agent Web $\overset{?}{=}$ LLM Applications / Network of (LLM) Agents

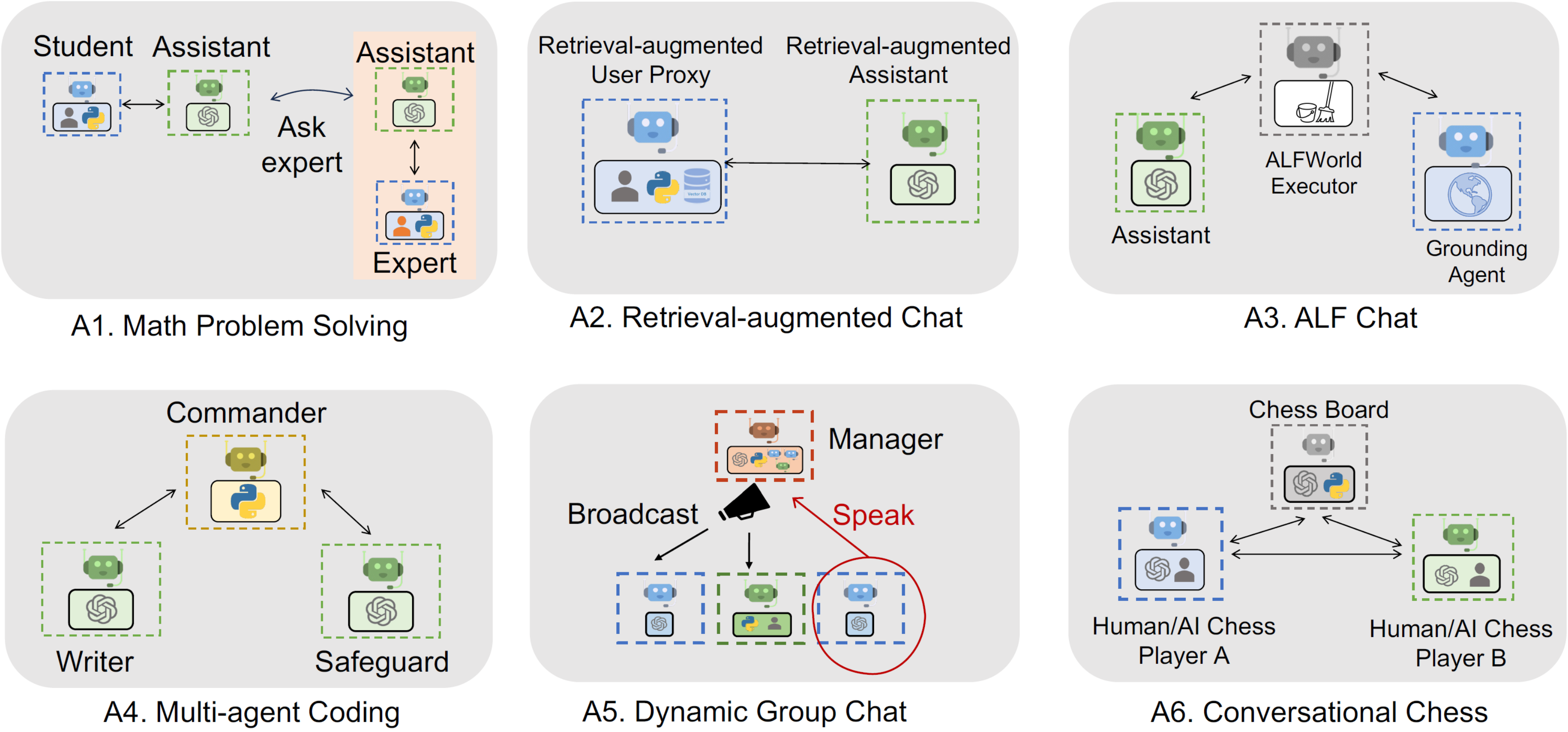
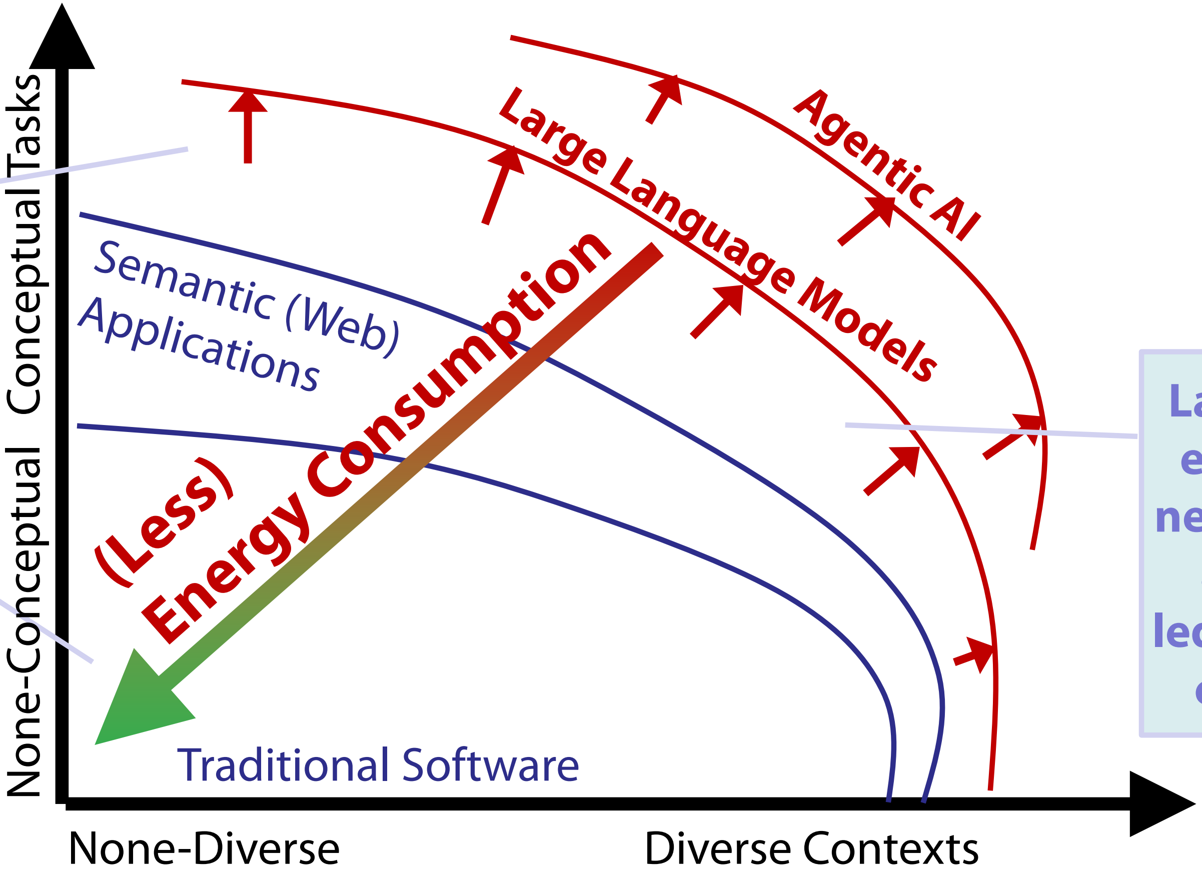


Image: Wu et al., AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation, ICLR 2024

Tradeoff (labor) cost-reducing use of technologies vs. energy consumption

Conceptual work too complex => prefer more staff for manual processing

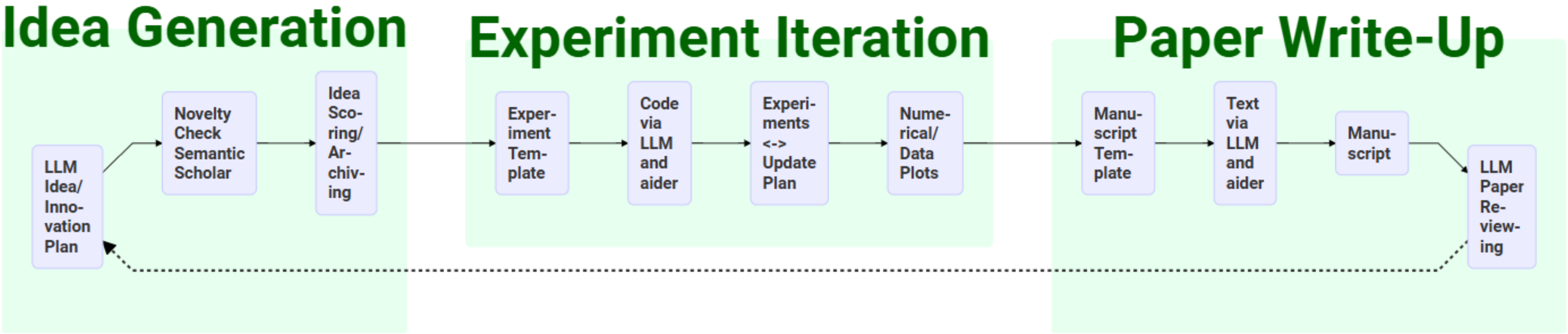
too simple => better to use traditional software or "business rules"



Lack of multi-experts with necessary time and knowledge, too high complexity

Can the job been done by AI?

- The AI Scientist



C. Lu, C. Lu, R.T. Lange, J. Foerster, J. Clune, D. Ha: The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery, arXiv:2408.06292
Zochi Achieves Main Conference Acceptance at ACL 2025, 2025, https://www.intology.ai/blog/zochi-acl Paper: https://arxiv.org/pdf/2503.10619

Energy Consumption: Google versus ChatGPT

	Google	**ChatGPT (estimated)**	**Calculator (LR44 battery)**
Per Query (KWh)	0.0003	0.0017 - 0.0026 (5.7 – 8.7 × Google Search)	
In Total	energy to power 200,000 homes	as much electricity as 175,000 people in January 2023	0.0002325 KWh

Sources: https://techland.time.com/2011/09/09/6-things-youd-never-guess-about-googles-energy-use/
https://www.digipal.ai/post/is-energy-consumption-for-ai-spiraling-out-of-control
https://towardsdatascience.com/chatgpts-energy-use-per-query-9383b8654487
https://towardsdatascience.com/chatgpts-electricity-consumption-7873483feac4

Are LLMs Mimicking Thinking?

- Key Finding of [S+25]
  - **For high complexity:** Beyond a certain threshold, LLMs and Large Reasoning Models (LRMs with chain-of-thought breaking down complex problems into a step-by-step sequence of intermediate thoughts, tool use etc.) hit a wall – accuracy crashes to zero
  - pattern-matching versus perfect step-by-step logic?
- One week later: response paper [OL25] with “C. Opus” (aka Claude from Anthropic) as first author
  - Claims **unfairness** with token limits + impossible tasks

[S+25] Shojaee et al., The Illusion of Thinking…, https://machinelearning.apple.com/research/illusion-of-thinking, 2025

[OL25] Opus, C., Lawsen, A. Comment on The Illusion of Thinking…, https://doi.org/10.48550/ARXIV.2506.09250, 2025

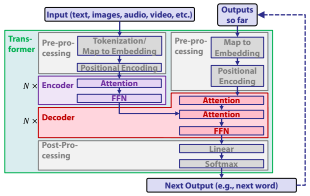
Training of Chatbots

Training Data

Input: What is a speciality in Freiberg?
Output: Freiberger Bauerhase

Input: Write a long story about …
Output: Once upon a time, …

Prediction (Forward Propagation)



$(\hat{y}_1, \dots, \hat{y}_T)$

Costs

$C((\hat{y}_1, y_1), \dots, (\hat{y}_T, y_T))$

$\Delta = \nabla_W C((\hat{y}_1, y_1), \dots, (\hat{y}_T, y_T))$

Back-propagation of the gradients to the previous layers

Δ

$W - \alpha \cdot \Delta$

Update parameters

α: learning rate

Requirements for Training LLMs

- Necessary for research on new ML architectures
- Duration/Computing Power
  - DeepSeek-V3 full training: 57 days on 2048 H800* GPUs
  - ExaScale-Supercomp. JUPITER/Jülich: 2 days for „ChatGPT"
- Full **training of special-purpose models** may need less computing
  - But: increasing demands on high-quality research causes increasing demand on hardware

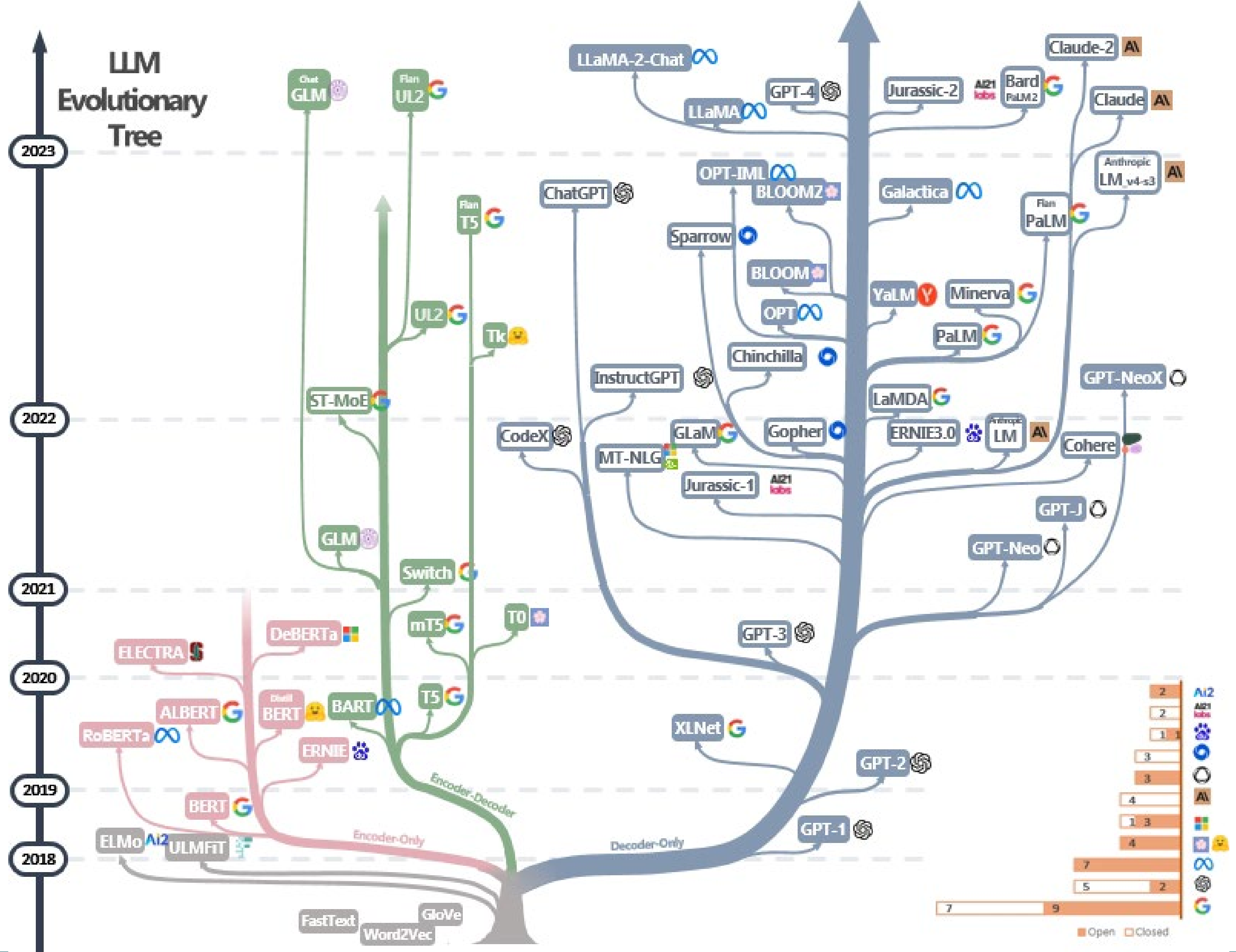
* H800: modified version of H100 for the Chinese market

https://tu-dresden.de/zih/das-department/news/tud-supercomputer-capella-in-top500-und-green500 https://openeurollm.eu/
https://www.fz-juelich.de/de/aktuelles/news/pressemitteilungen/2024/neues-jupiter-modul-unterstreicht-fuehrende-rolle-des-kommenden-europaeischen-exascale-superrechners

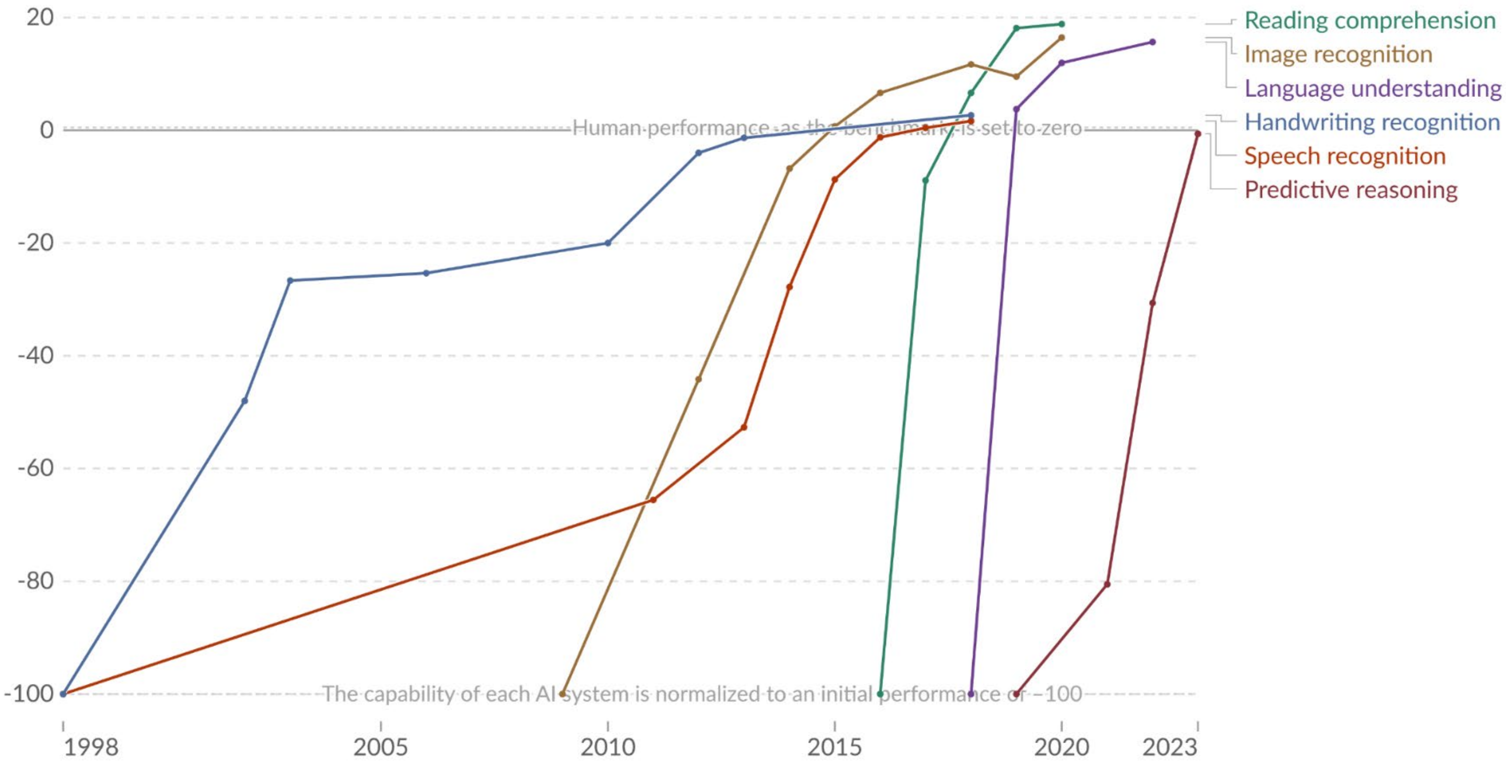
Training Costs

Model	Year	(Estimated) Training costs in USD
Transformer	2017	930
BERT-Large	2018	3.288
RoBERTa Large	2019	160.018
LaMDA	2022	1.319.586
Llama 2 70B	2023	3.931.897
GPT-3 175B	2020	4.324.883
Megatron-Turing NLG 530B	2021	6.405.653
PaLM 540B	2022	12.389.056
GPT-4	2023	78.352.034
Gemini Ultra	2023	191.400.000
DeepSeek-V3	2025	5.576.000

Stanford University, AI Index Report, 2024, https://aiindex.stanford.edu/ https://arxiv.org/pdf/2412.19437



https://github.com/Mooler0410/LLMsPracticalGuide



Our World in Data

Test scores of AI systems on various capabilities relative to human performance

Within each domain, the initial performance of the AI is set to −100. Human performance is used as a baseline, set to zero. When the AI's performance crosses the zero line, it scored more points than humans.

Data source: Kiela et al. (2023) OurWorldinData.org/artificial-intelligence | CC BY

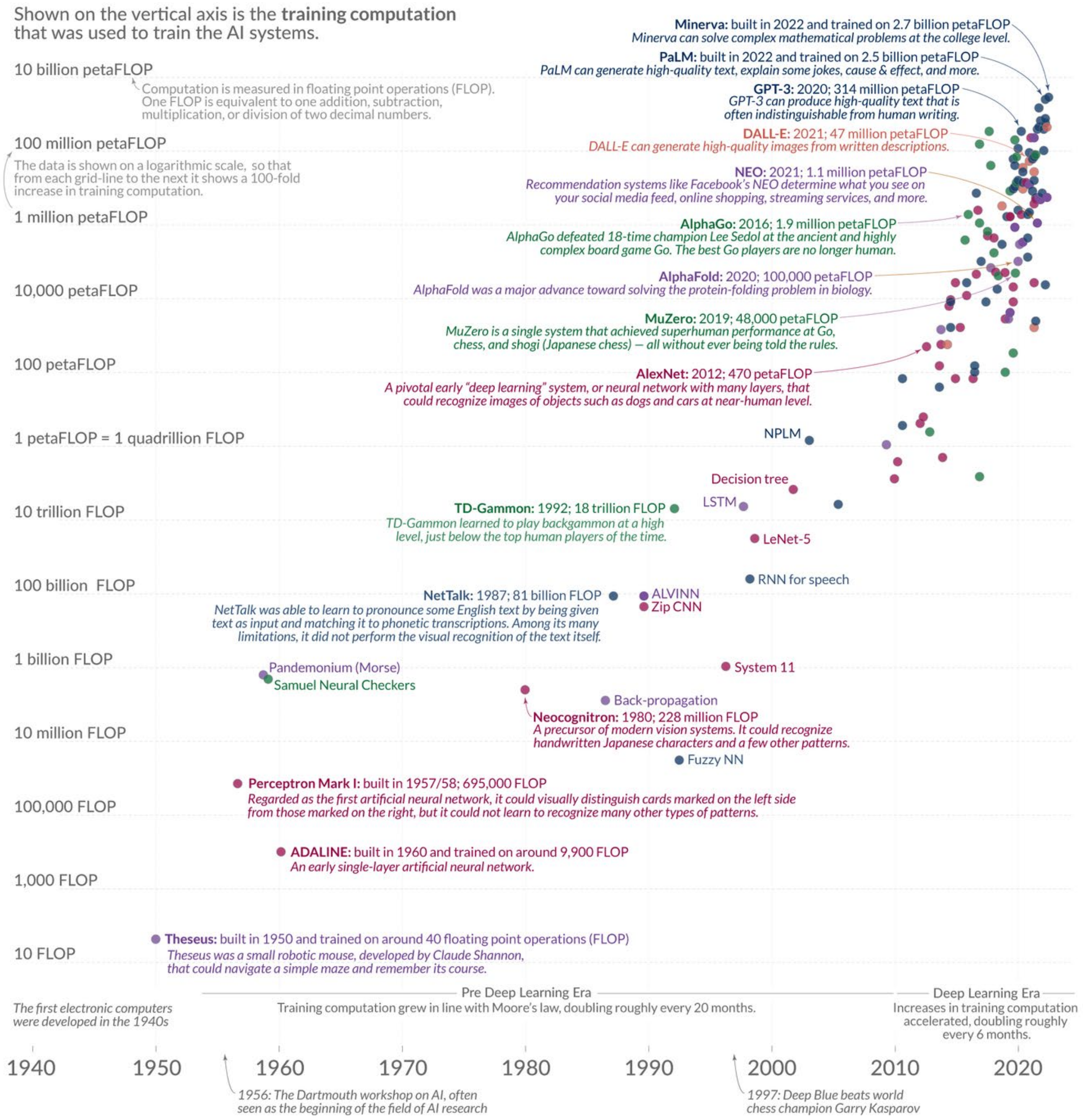
Note: For each capability, the first year always shows a baseline of −100, even if better performance was recorded later that year.

https://ourworldindata.org/brief-history-of-ai

The rise of artificial intelligence over the last 8 decades: As training computation has increased, AI systems have become more powerful

Our World in Data

The color indicates the domain of the AI system: ● Vision ● Games ● Drawing ● Language ● Other



The data on training computation is taken from Sevilla et al. (2022) – Parameter, Compute, and Data Trends in Machine Learning.
It is estimated by the authors and comes with some uncertainty. The authors expect the estimates to be correct within a factor of two.
OurWorldinData.org – Research and data to make progress against the world's largest problems.

Licensed under CC-BY by the authors
Charlie Giattino, Edouard Mathieu, and Max Roser

FLOP = Floating Point Operations Per Second

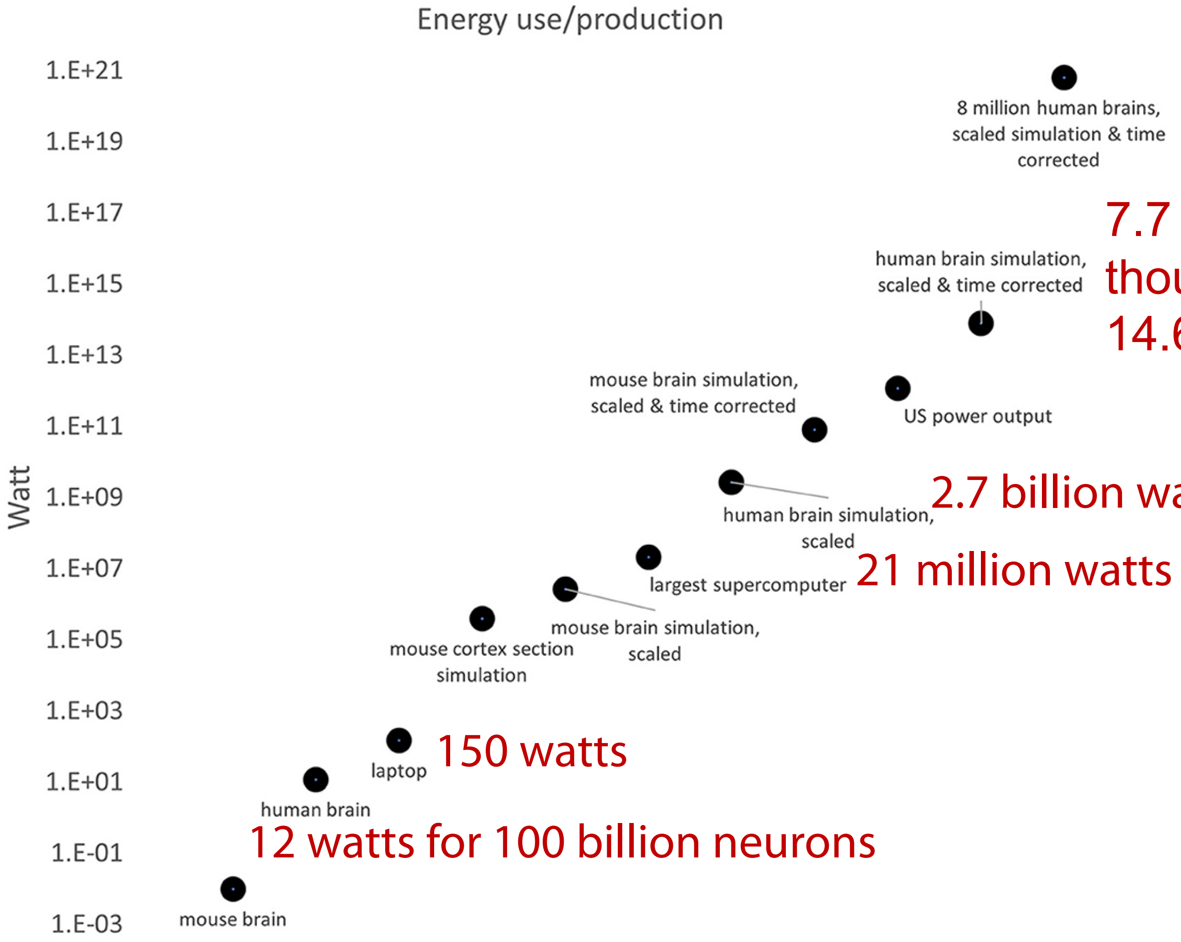
https://ourworldindata.org/brief-history-of-ai

Energy-Efficient Alternative to Artificial Neural Networks (ANN)

- Spiking neural networks (SNNs)
  - save energy by not using multiplications
  - “only” x-times energy consumption compared to ANNs while maintaining comparable accuracy
    - x = 0.85 on classical architectures
    - x = 0.78 on spatial-dataflow architectures specialized to ANNs/SNNs

Yan et al. (2024). Reconsidering the energy efficiency of spiking neural networks. arXiv. https://doi.org/10.48550/ARXIV.2409.08290

AI’s Energy Demands vs. the Human Brain’s Efficiency



7.7 × 10^{13} J for 1 second of thought in real-time and 14.6 ExaFLOPS

2.7 billion watts

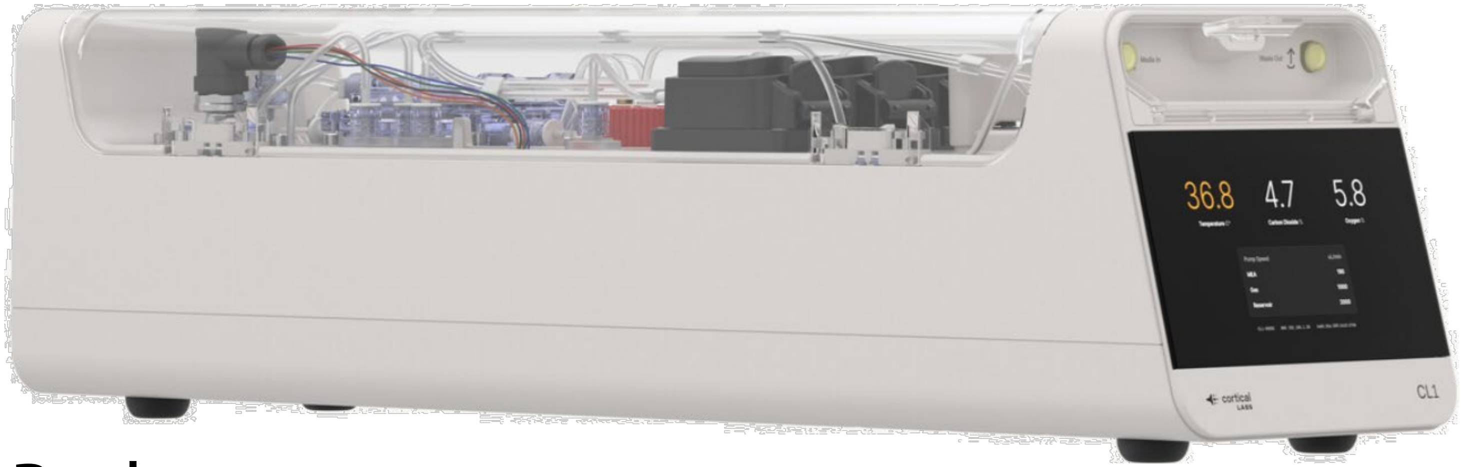
21 million watts

150 watts

12 watts for 100 billion neurons

Stiefel, K. M., & Coggan, J. S. (2023). The energy challenges of artificial superintelligence. Frontiers in Artificial Intelligence, 6. https://doi.org/10.3389/frai.2023.1240653

Organic Computing



- Real neurons
  - are cultivated inside a nutrient rich solution, supplying them with everything they need to be healthy
  - grow across a silicon chip, which sends and receives electrical impulses into the neural structure.

https://corticallabs.com/cl1.html

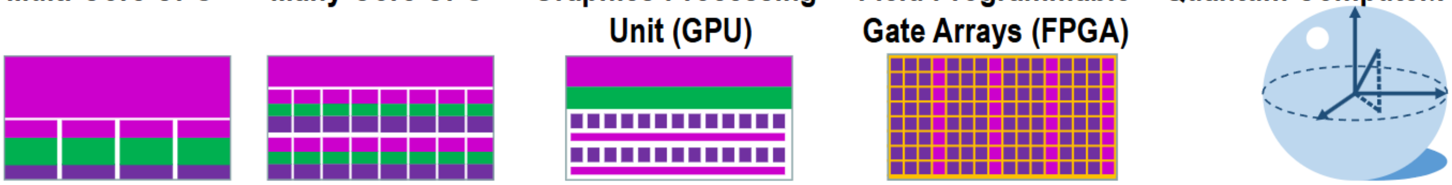
Other Tasks not utilizing ANNs

- Ratios of Energy Reduction (Reference CPU)

	CPU	GPU	FPGA
Input Processing	1	**1.79×**	1.41×
Image Arithmetic	1	**3.19×**	2.93×
Image Filters	1	3.17×	**3.89×**
Image Analysis	1	2.34×	**5.67×**
Geometric Transform	1	10.3×	**16.6×**
Features/ OF/ StereoBM	1	7.44×	**22.3×**

Qasaimeh et al. (2019). Comparing Energy Efficiency of CPU, GPU and FPGA Implementations for Vision Kernels. arXiv. https://doi.org/10.48550/ARXIV.1906.11879

Hardware Architectures



	Multi-Core CPU	Many-Core CPU	Graphics Processing Unit (GPU)	Field Programmable Gate Arrays (FPGA)	Quantum Computer/Annealer
Cores	~10	~100	~1000	~100 000	~100 / ~5000 qubits
Core Complexity	Complex (optimized for single thread performance)			Simple	Complex
Computational Model	MIMD + SIMD		SIMD	Data-Flow	Universal/Adiabatic Quantum computing
Parallelism	Thread and Data Parallel		Data Parallel	Arbitrary	Quantum Logic Gates/ Fluctuation (glob.Min.): Interference
Memory Model	Shared	Distributed			Quantum Superposition
Power	150 W	200 W	250 W	50 W	25 KW

Database Op.

- Query Optimization (Enumeration of Plans), Concurrency Control
- Efficient Processing Of
- Query Processing

Legend: 🟪 Computational Unit 🟩 Execution Controller 🟨 Interconnection Network 🟪 On-Chip Memory

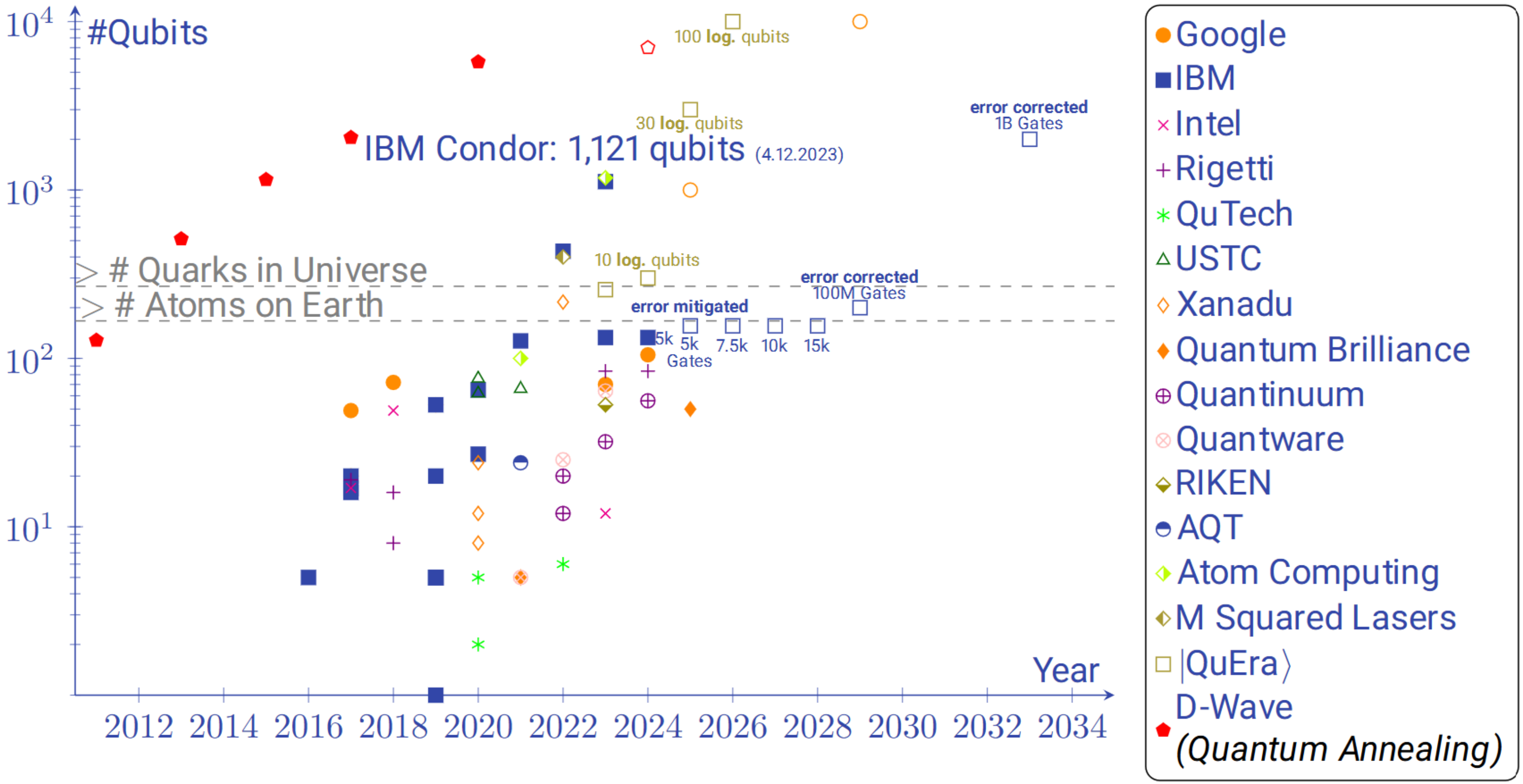
Extended from C. Plessl, Accelerating Scientific Computing with Massively Parallel Computer Architectures, IMPRS Winter School, Wroclaw, 2012

Quantum Computing

- **1982**: Feynman proposes the concept of quantum computers [F'82]
- **2019**: Google announces "Quantum Supremacy" by its 53-qubits chip "Sycamore" [A+'19]
  - 200 seconds on Sycamore versus 10,000 years on the world's fastest supercomputer IBM Summit
  - IBM [P+'19]: only 2.5 days on classical supercomputer after deduction of the problem (i.e., using a better classical algorithm)
  - Pan et al. [PCZ'21]: only 15 hours on 512 GPU-cluster using another classical algorithm for obtaining a large number of uncorrelated samples
    - Estimation: a few dozens of seconds on ExaFLOPS supercomputer
  - Discussion intensified the excessive hype about quantum technology
- **2023**: Next try: Google runs Random Circuit Sampling experiments on its 70-qubits improved "Sycamore" in seconds instead of 47 years (estimation for #1 classical supercomputer in 2023) [G+'23]
- **2025**: DWave solves magnetic materials simulation problems in 20 min instead of 1 million years [K+25] (others disagree [W'25])

[F'82] [A+'19] [P+'19] [PCZ'21] [G+'23] [K+'25] [W'25]

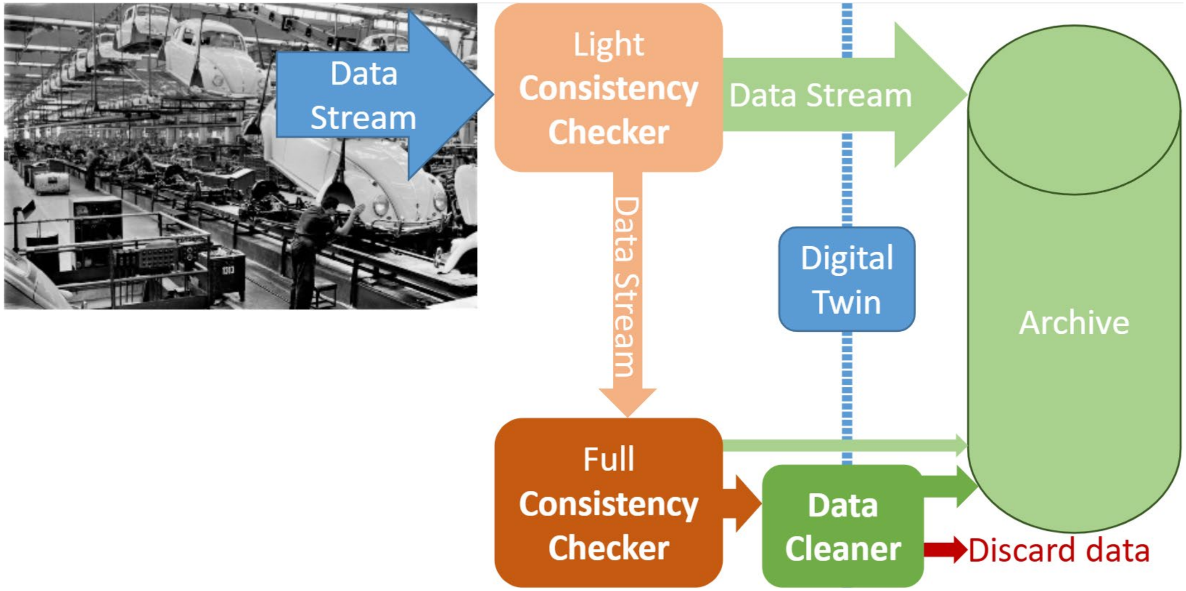
Timeline of Quantum Computers



https://www.ifis.uni-luebeck.de/~groppe/lectures/slides/slides-web/qc/QC_1_Intro.html#19

Green Computing in Industry 4.0

(joint work with Bosch)



- Energy savings by lightweight components during normal operation and switching on full components for inconsistency handling
  - CO2e emissions can be reduced by a factor of about 0.6
  - in one year 262 kgCO2e in EU for a medium-sized plant

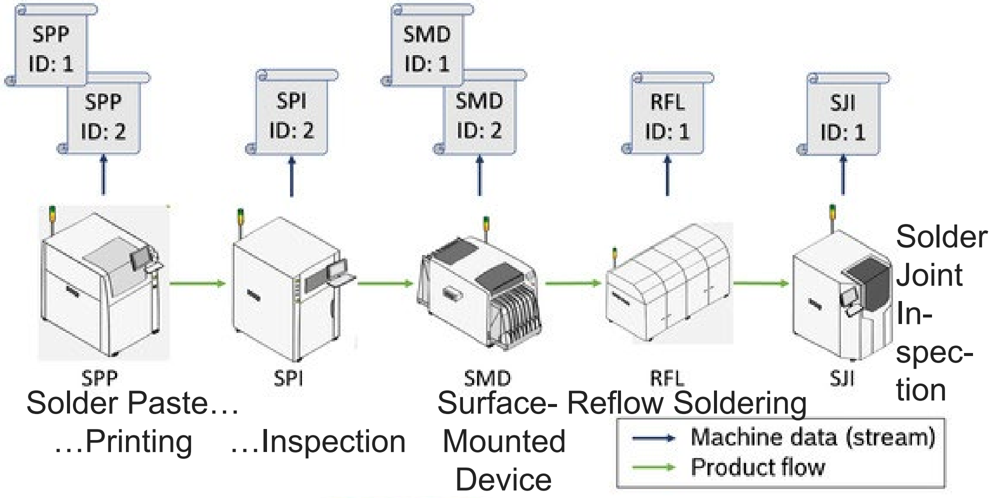
S. Paasche, S. Groppe, GreenCC: A Hybrid Approach to Sustainably Validate Manufacturing Data in Industry 4.0 Environments, DATA'23

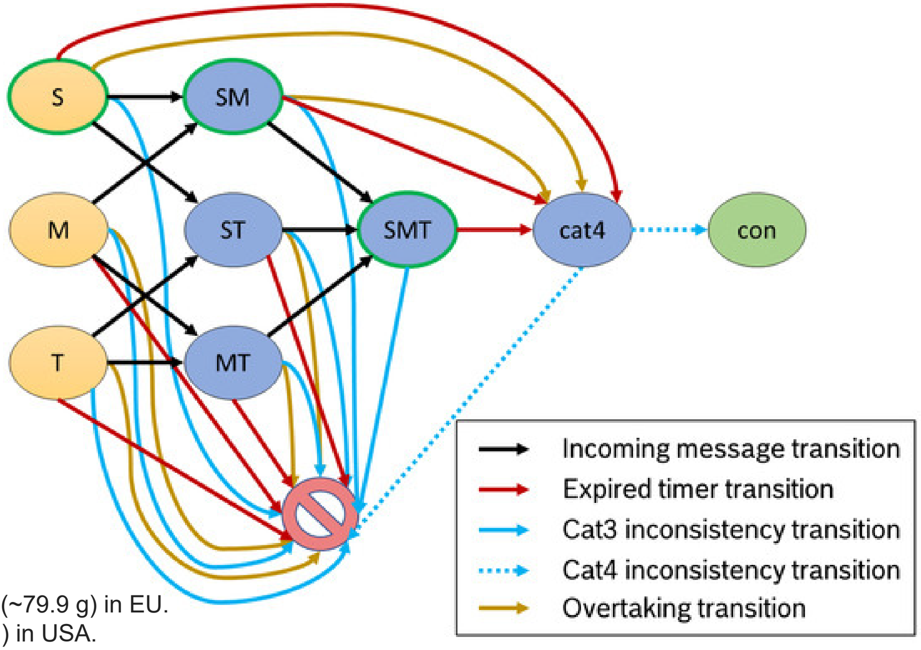
Back to the roots?

(joint work with Bosch)

Carbon-dioxide equivalents (CO_2e) in gram per kWh for daily operation in small, medium, and large plants

Approach	Plant Size	EU 262 gCO2*e*/kWh	USA 379 gCO2*e*/kWh
Flink	small:	511 g	739 g
	medium:	1915 g	2770 g
	large:	3191 g	4616 g
SPARQL	small:	23 g + euss	34 g + usss
	medium:	88 g + eu*mm*	127 g + us*mm*
	large:	147 g + eu*ll*	212 g + us*ll*
LightCC	small:	321 g	465 g
	medium:	1204 g	1742 g
	large:	2007 g	2903 g
FullCC	small:	330 g	478 g
	medium:	1239 g	1792 g
	large:	2065 g	2987 g
Finite State Automaton	**small:**	**252 g**	**365 g**
	medium:	**946 g**	**1369 g**
	large:	**1577 g**	**2282 g**




S
SM
M
ST
SMT
cat4
con
T
MT
Incoming message transition
Expired timer transition
Cat3 inconsistency transition
Cat4 inconsistency transition
Overtaking transition

Legend: eu_x: Additional $CO_2{}^e$ in plant of size **s**mall (~12.8 g), **m**edium (~47.9 g), **l**arge (~79.9 g) in EU. us_x: Additional $CO_2{}^e$ in plant of size **s**mall (~18.6 g), **m**edium (~69.4 g), **l**arge (~115.6 g) in USA.

Paasche, S.; Groppe, S. A Finite State Automaton for Green Data Validation.... *Future Internet* **2023**, *15*, 349. https://doi.org/10.3390/fi15110349

Back to the roots?

***Rosetta Code* Global Ranking (based on Energy)**

Position	**Language**
1	C
2	Pascal
3	Ada
4	Rust
5	C++, Fortran
6	Chapel
7	OCaml, Go
8	Lisp
9	Haskell, JavaScript
10	Java
11	PHP
12	Lua, Ruby
13	Perl
14	Dart, Racket, Erlang
15	Python

Pereira et al. (2021). Ranking programming languages by energy efficiency. Science of Computer Programming, 205, 102609. https://doi.org/10.1016/j.scico.2021.102609

Advices for Sustainable Computing

Use your brain instead of technology is the most sustainable way of computing!

- Use h technology **as you need**
  - Do **not** u lculating!
- Use hardware-e ware!
  - Calculator vs. Laptop v ter
  - Use the **most efficient** hardw roblem
    - Quantum computing to save years/ce computing?
  - **How much accuracy** do you need?
- **Simple is beautiful… and energy efficient!**

Recent Scientific Services with Submissions Open

Please submit papers and chapters!

- Call for Papers
  - International Semantic Intelligence Conference (ISIC) (Lübeck and hybrid!)
    - https://www.ifis.uni-luebeck.de/~groppe/isic/
  - International Health Informatics Conference (IHIC)
    - https://sites.google.com/view/ihic2025?usp=sharing
- Call for Book Chapters
  - **Transparent Intelligence: A Guide to Explainable AI** (Nova Publishers), Sarika Jain, Sven Groppe, Prabhjot Kaur, Bharat K Bhargava
    - Please contact: Sarika Jain jasarika@nitkkr.ac.in
  - **Knowledge Graphs and Large Language Models: Current Approaches, Challenges, and Future Directions** (Elsevier), Sanju Tiwari, Sven Groppe, Jinghua Groppe, Nandana Mihindukulasooriya
    - Please contact Sanju Tiwari tiwarisanju18@ieee.org