DEEP AI SOFTWARE

Professor Sasu Tarkoma

AI methods are undergoing a revolution due to increased computational power, training data, method development, and more efficient implementations.

AI systems are tools.

Foundation models.

- Large Language Models (LLM), Action Models (LAM), etc.
- Reasoning capabilities (test-time compute).

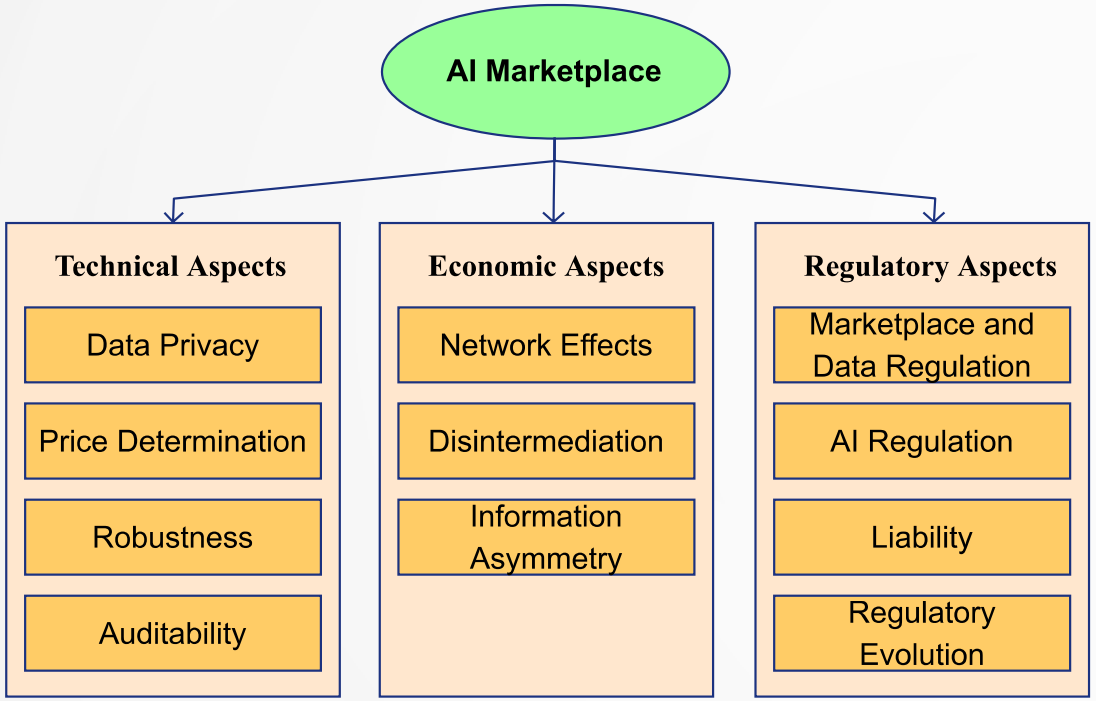
AI regulation.

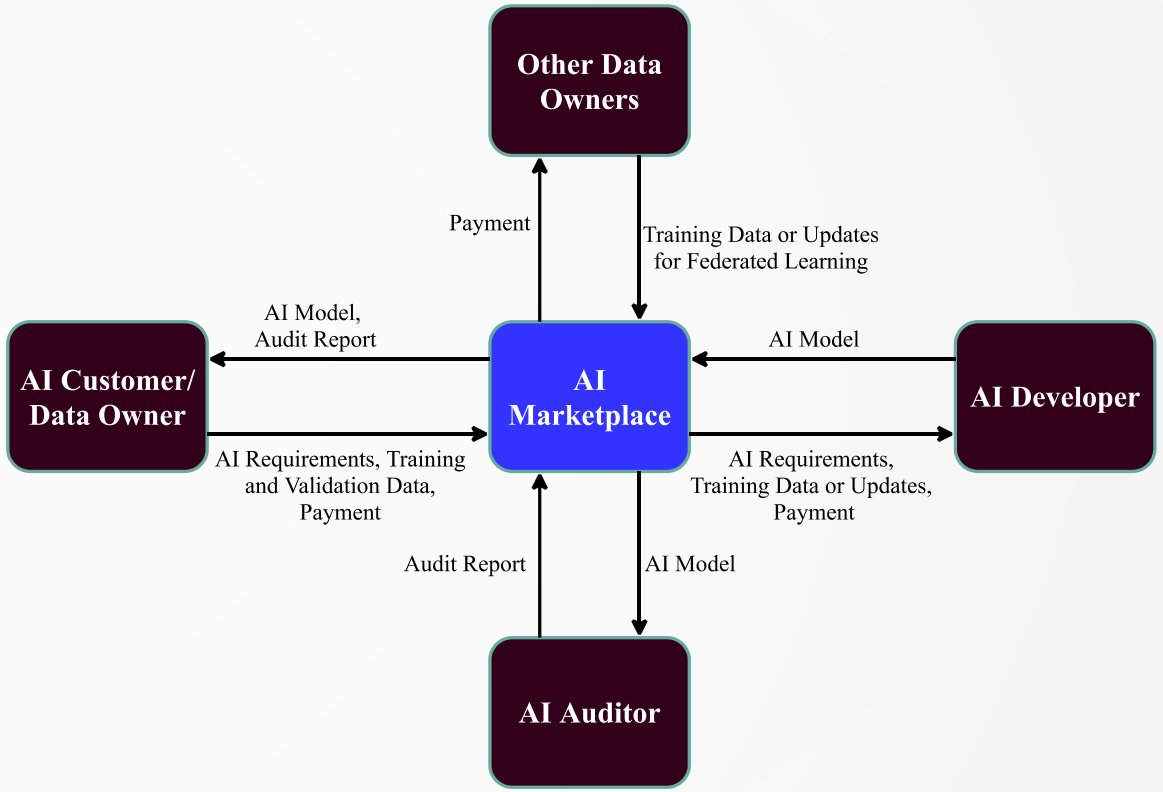
Renaissance of open development.

DEEP LEARNING SOFTWARE

Core Frameworks:	•TensorFlow & Keras: Comprehensive libraries for research & production. •PyTorch: Dynamic, researcher-friendly framework.
Interoperability & Standards:	•ONNX (Open Neural Network Exchange): Standardizes model formats for cross-framework compatibility.
Community & Model Hubs:	•HuggingFace: Central hub for NLP and expanding domains, offering pre-trained models and easy-to-use libraries. •TensorFlow Hub / PyTorch Hub: Repositories for model sharing.
Cloud & Managed Platforms:	•AWS: Amazon SageMaker & AWS AI Services for end-to-end training, tuning, and deployment. •Google Cloud AI & Microsoft Azure ML: Alternative managed platforms. •API providers (OpenAI, Anthropic, …)
Open source frameworks and specialized tools	

EMERGING AI ECOSYSTEM: OUR EARLY WORK

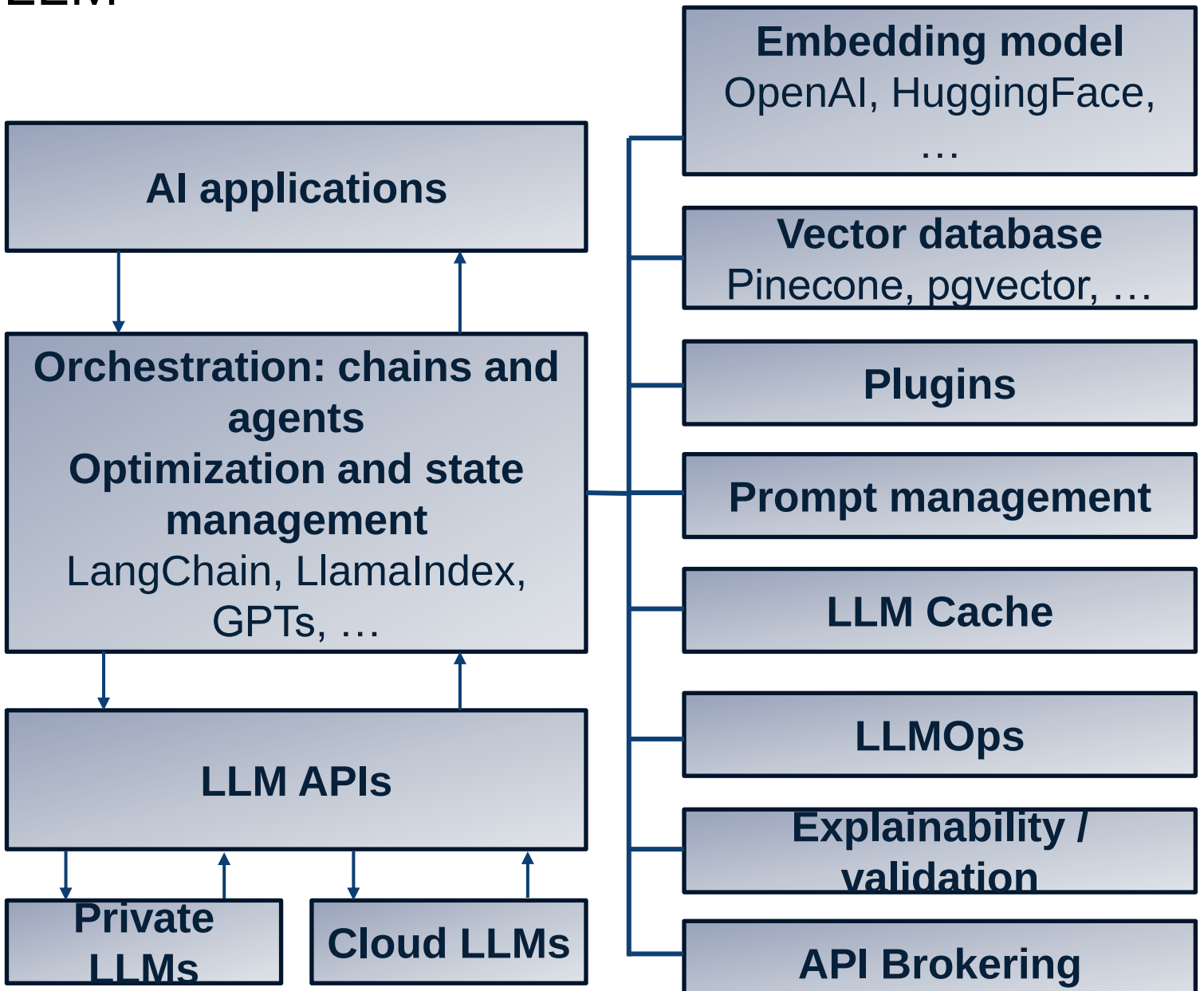




A. Kumar, B. Finley, T. Braud, S. Tarkoma and P. Hui, "Sketching an AI Marketplace: Tech, Economic, and Regulatory Aspects," in IEEE Access, vol. 9, pp. 13761-13774, 2021

Vlasis Koutsos, Dimitrios Papadopoulos, Dimitris Chatzopoulos, Sasu Tarkoma, Pan Hui: Agora: A Privacy-Aware Data Marketplace. IEEE Trans. Dependable Secur. Comput. 19(6): 3728-3740 (2022)

Overview of the LLM software stack



Foundation models and fine-tuned models

AI CAN CONTROL TASKS

Task planning

Model selection

Task execution

Task result composition

Use of digital tools

Creation of digital tools

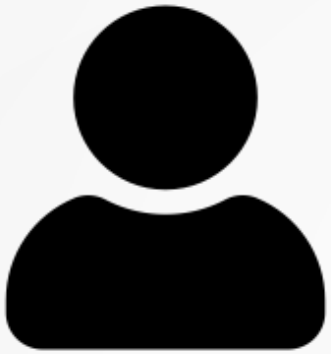
Planning of how to use digital tools

HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace. https://huggingface.co/papers/2303.17580
Large Language Models as Tool Makers. https://arxiv.org/pdf/2305.17126.pdf


AGENTIC SOFTWARE

THE SHIFT FROM MODELS TO COMPOUND AI SYSTEMS: AGENTIC SOFTWARE

- Increasingly many new AI results are from compound systems (https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/)
- Agentic software: Autonomous software entities that perform tasks, make decisions, and interact with environments or users.
- Multi-agent Systems
- Language Model–Based Agent Frameworks:
  - LangChain: Facilitates the development of LLM-powered agents and applications by chaining reasoning steps and integrating external tools.
  - AutoGPT: An open-source solution that leverages LLMs to autonomously perform multi-step tasks with minimal human input.
  - Many other developments



User

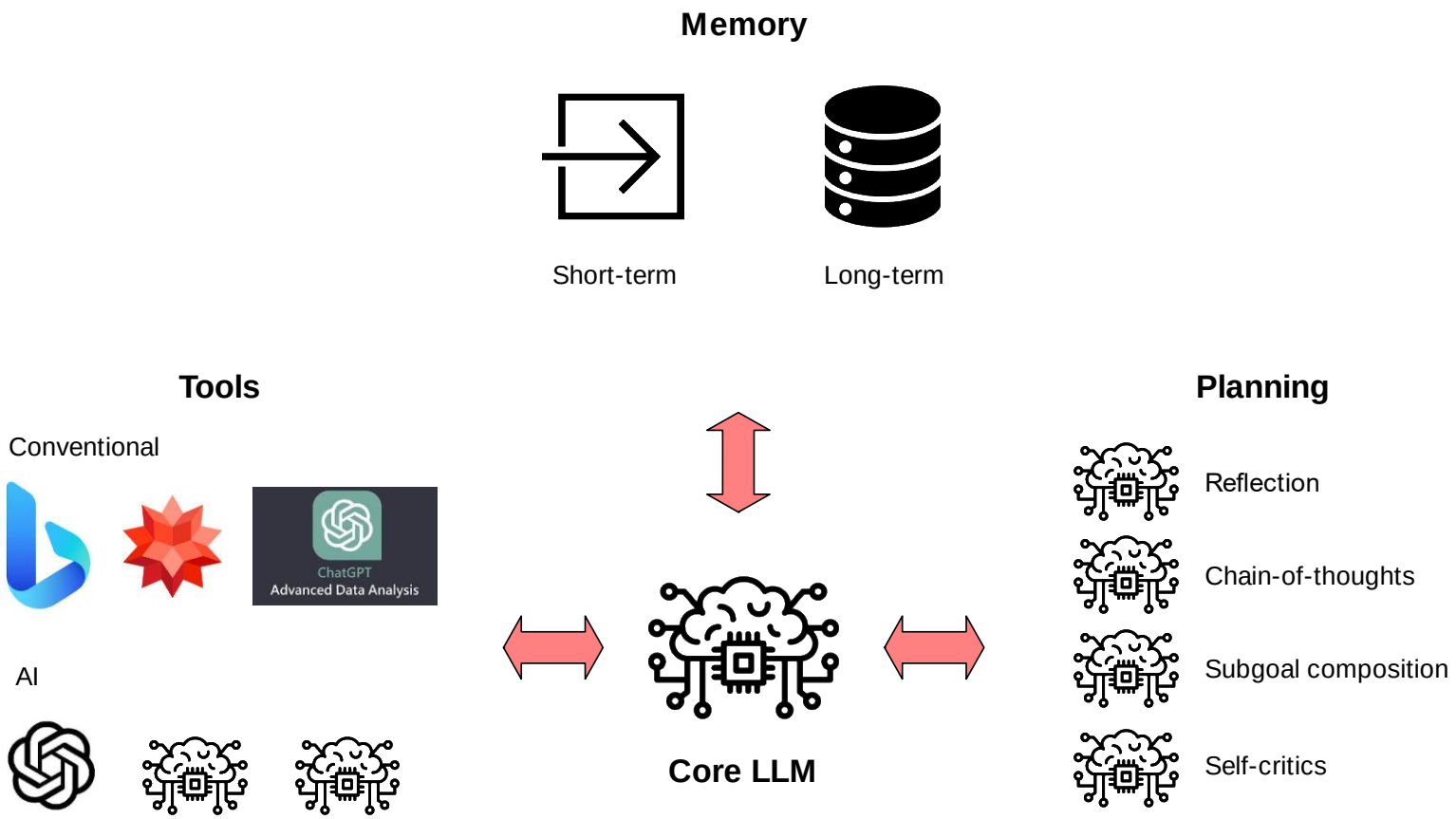
goals/tasks

actions

AI

Environment

feedback



Kokkonen, H., Lovén, L., Motlagh, N. H., Kumar, A., Partala, J., Nguyen, T., ... & Riekki, J. (2022). Autonomy and intelligence in the computing continuum: Challenges, enablers, and future directions for orchestration. *arXiv preprint arXiv:2205.01423*.

Functions

Monitoring
- monitoring
- performance estimation
- workload prediction
- benchmarking

Flow control
- aggregation
- sharing
- offloading
- caching

Communication mgmt.
- topology mgmt.
- interconnect

Market mgmt.
- registration
- lookup
- pricing
- obligations mgmt.

Allocation
- placement
- scheduling
- migration
- scaling
- replication

Lifecycle mgmt.
- creating, deleting
- starting, stopping
- updating

CHANGING SOFTWARE DEVELOPMENT

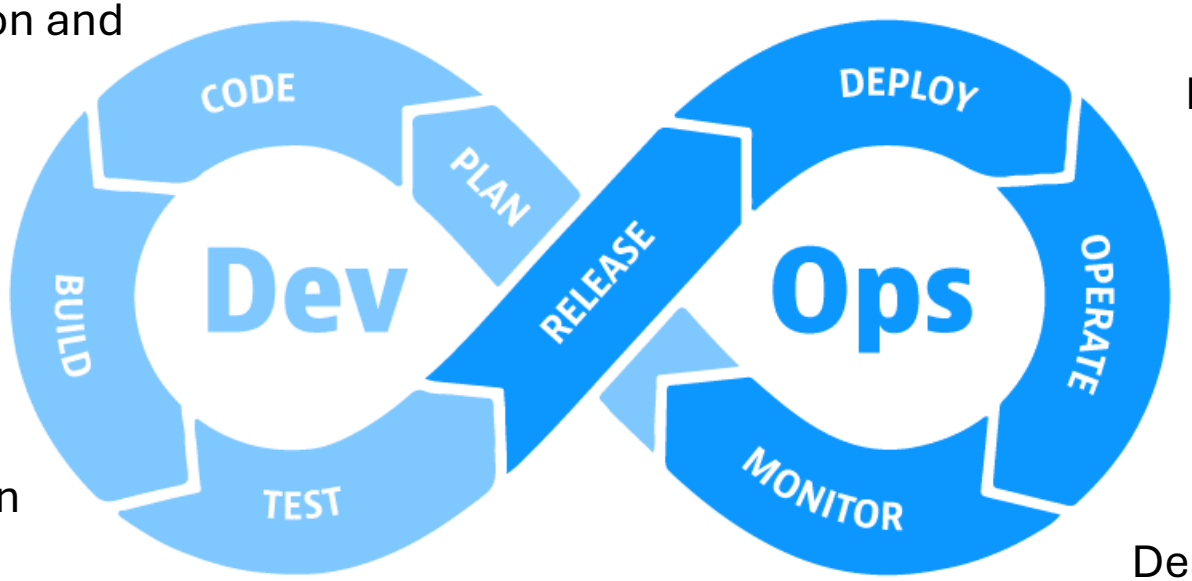
AI RESHAPES TODAY'S DEVELOPMENT WORKFLOWS

Acquiring and cleaning data for adaptation and validation

LM selection and its evolution

Resource greediness

Fine-tuning, RAG, prompt engineering,...



LM as a seamless part of End-user experience

Technical integration

Dealing with changes in underlying LM models

Non-deterministic and constantly changing behavior in testing and monitoring. Dealing with hallucinations

FUTURE OF SOFTWARE DEVELOPMENT

Scenario 1: Traditional Software Development Operations.

- Humans own all roles, and tools and development environments provide automation.

Scenario 2: AI in loop.

- Humans control AI, but AI is beginning to manage larger and more complex work areas. AI is used to automate selected parts of manual and repetitive tasks such as code generation, documentation, testing, and deployment.

Scenario 3: AI assumes role(s).

- AI systems assume selected roles. For example, AI is used to manage the process, design, implementation, testing, delivery, and maintenance. Humans focus on the most complex tasks and control the entire operation, responsible for overseeing that it is working correctly, and producing high-quality results.

Scenario 4: Human-in-the-loop.

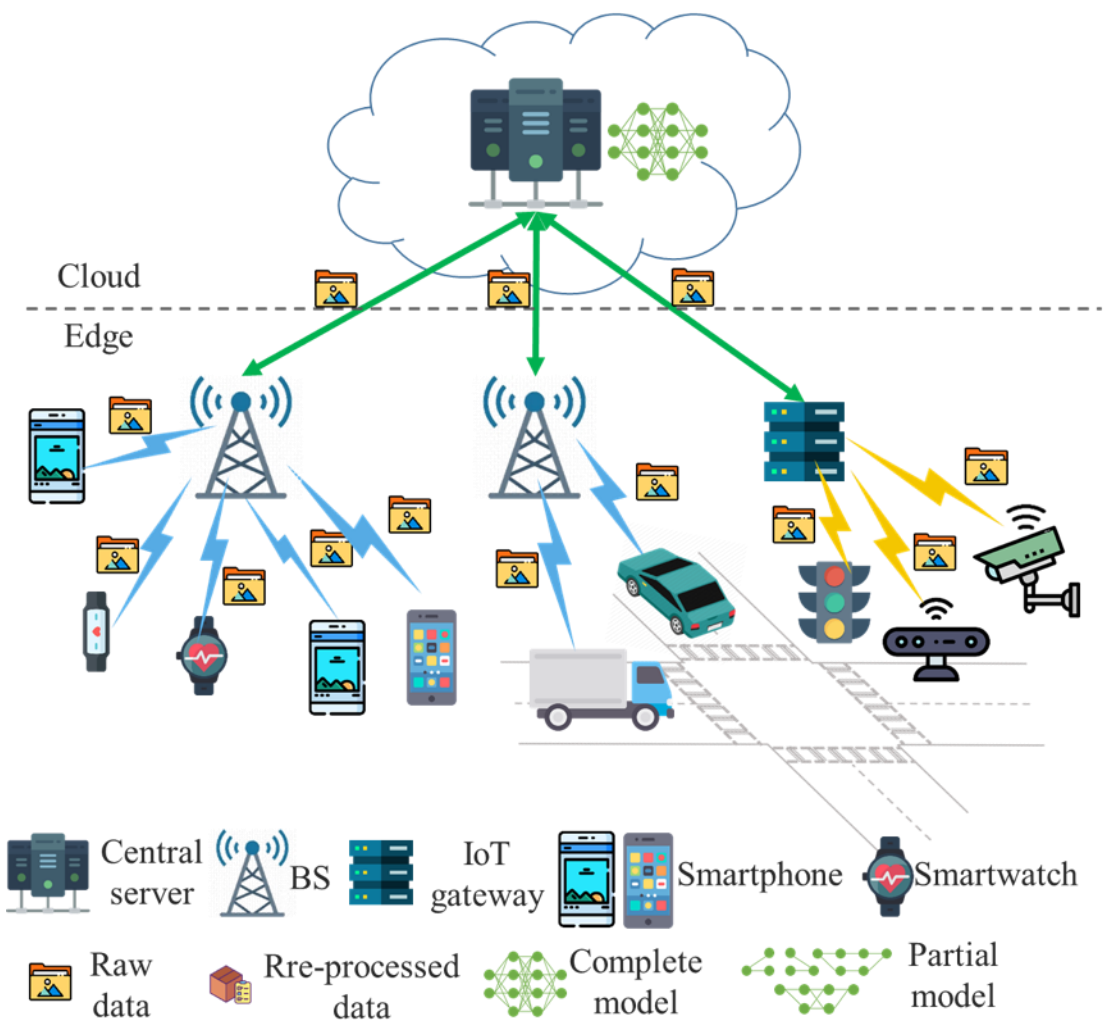
- AI manages development operations in various roles. Humans oversee and control the process, but their role is focused on overwatches such as operational control, problem solving, quality assurance, and security. AI roles are responsible for automating most or all the other tasks in the development lifecycle.

New ethical and intellectual property challenges.

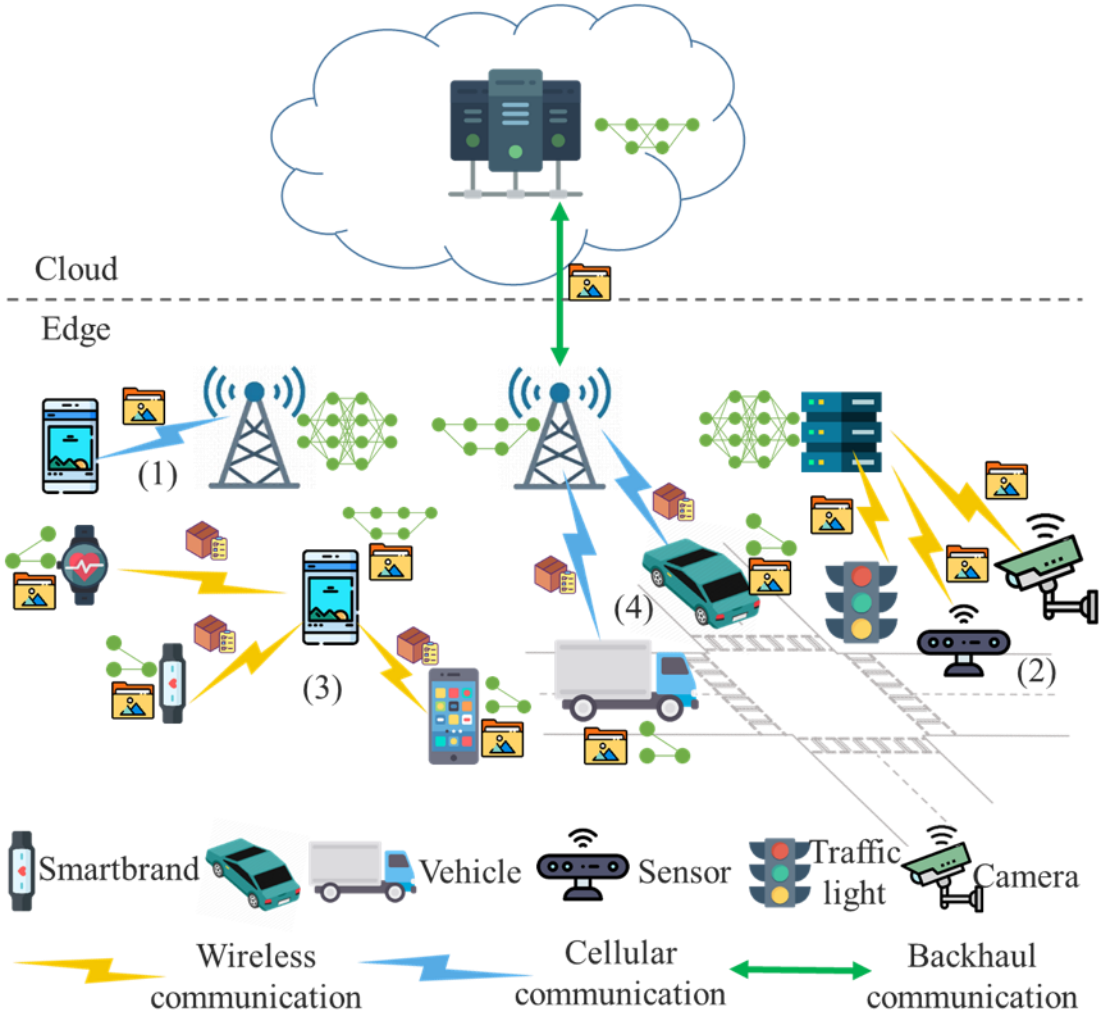
Jaakko J. Sauvola, Sasu Tarkoma, Mika Klemettinen, Jukka Riekki, David S. Doermann: Future of software development with generative AI. Autom. Softw. Eng. 31(1): 26 (2024)


DISTRIBUTED VIEW

Centralized cloud for AI

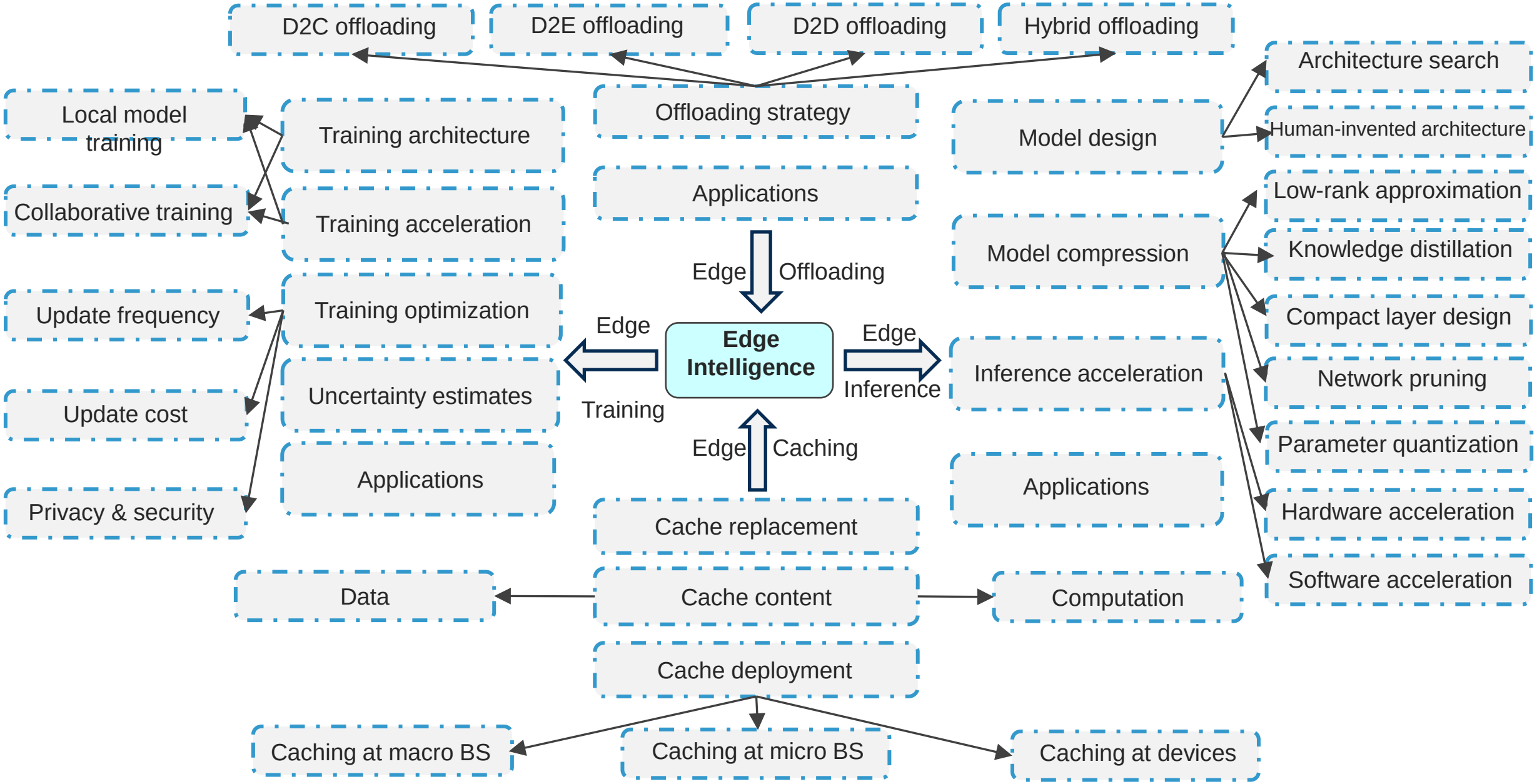


Edge Intelligence



Edge Intelligence: Empowering Intelligence to the Edge of Network. D Xu, T Li, Y Li, X Su, S Tarkoma, T Jiang, J Crowcroft, P Hui Proceedings of the IEEE 109 (11), 1778-1837.

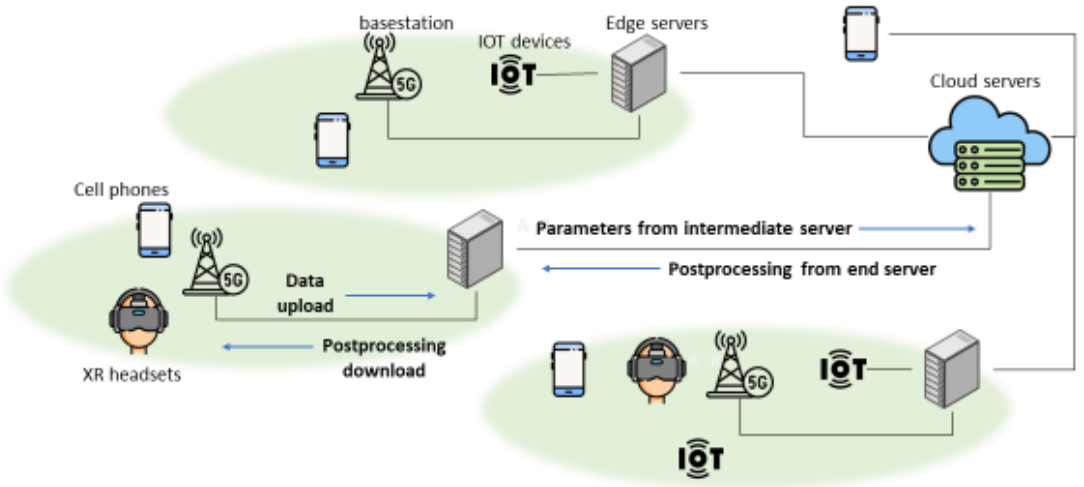
A Taxonomy


D2C offloading
D2E offloading
D2D offloading
Hybrid offloading
Offloading strategy
Local model training
Collaborative training
Training architecture
Training acceleration
Training optimization
Update frequency
Update cost
Privacy & security
Uncertainty estimates
Applications
Applications
Edge Offloading
Edge Intelligence
Edge Training
Edge Inference
Edge Caching
Model design
Architecture search
Human-invented architecture
Model compression
Low-rank approximation
Knowledge distillation
Compact layer design
Network pruning
Parameter quantization
Inference acceleration
Applications
Hardware acceleration
Software acceleration
Cache replacement
Cache content
Cache deployment
Data
Computation
Caching at macro BS
Caching at micro BS
Caching at devices

Fog Computing for Deep Learning with Pipelines

We demonstrate split learning and inference pipelines on device, cloudlet and cloud with a vision based use case (CIFAR-10).

If the fog hardware can be fully utilized, it will improve the throughput of heavy workloads.



- Fog is a hierarchical network structure.
- Resources in fog can be pooled with pipelines.
- Sparse: **S**tream **P**rocessing **A**rchitecture for **R**esource-**S**ubtle **E**nvironments[a]

[a] https://github.com/AnteronGitHub/sparse

Antero Vainio's PhD dissertation topic

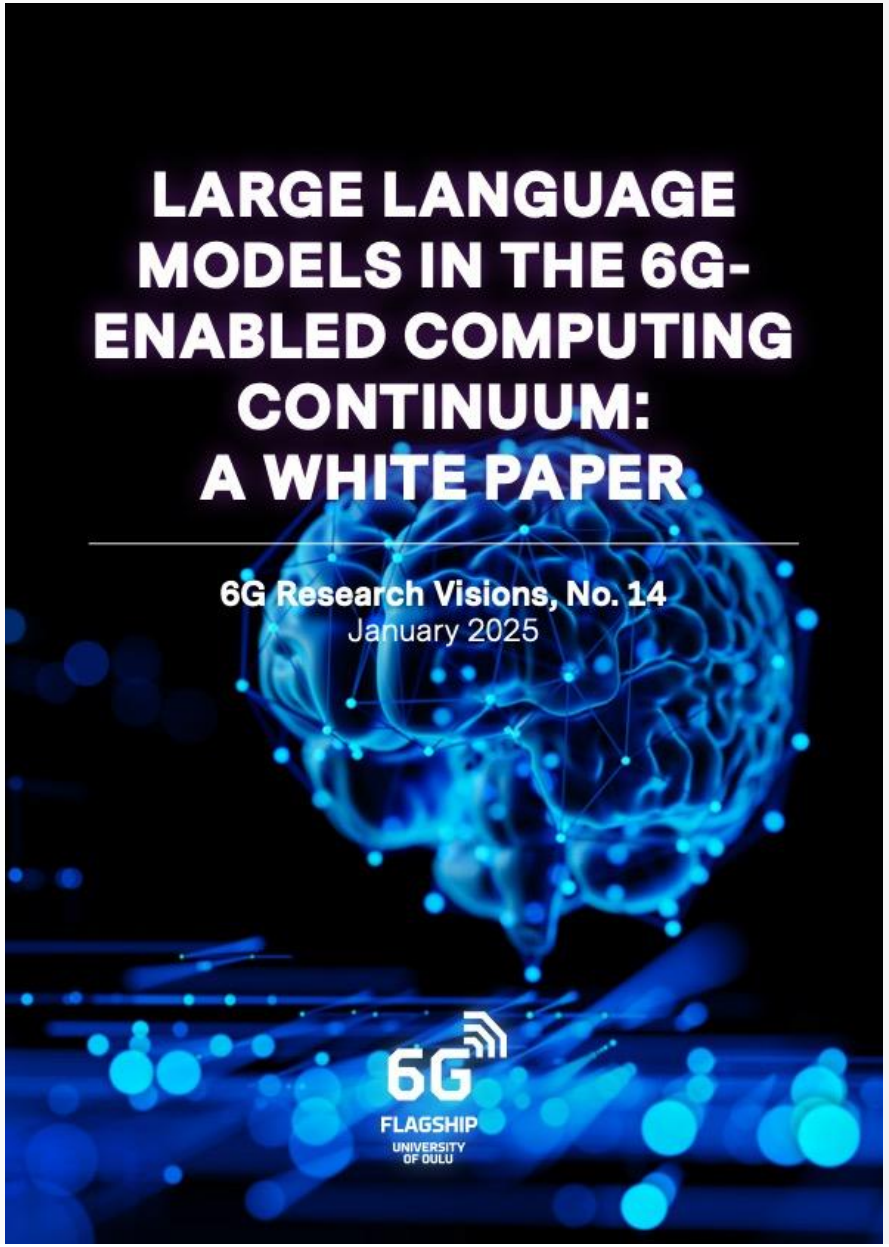
HELS
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Conceptual Overview

Section 2	LLMs and GPTs in 6G: A Holistic Multi-layered Architectural Perspective
Section 3	LLMs and GPTs in 6G: Taxonomy and State-of-the-art
Section 4	Architectural Foundations for Intelligent Evolution: From 5G to AI Interconnect
Section 5	The 6G AI Interconnect Framework

Practical Applications

Section 6	Semantic Publish/Subscribe and Intent-based Messaging
Section 7	LLMs as Controller
Section 8	LLMs as Network Tool Builder and APIs Broker
Section 9	Training Foundation LLMs for Mobile Networks Applications
Section 10	LLMs Configuration and Interoperability
Section 11	LLMs for Code Generation

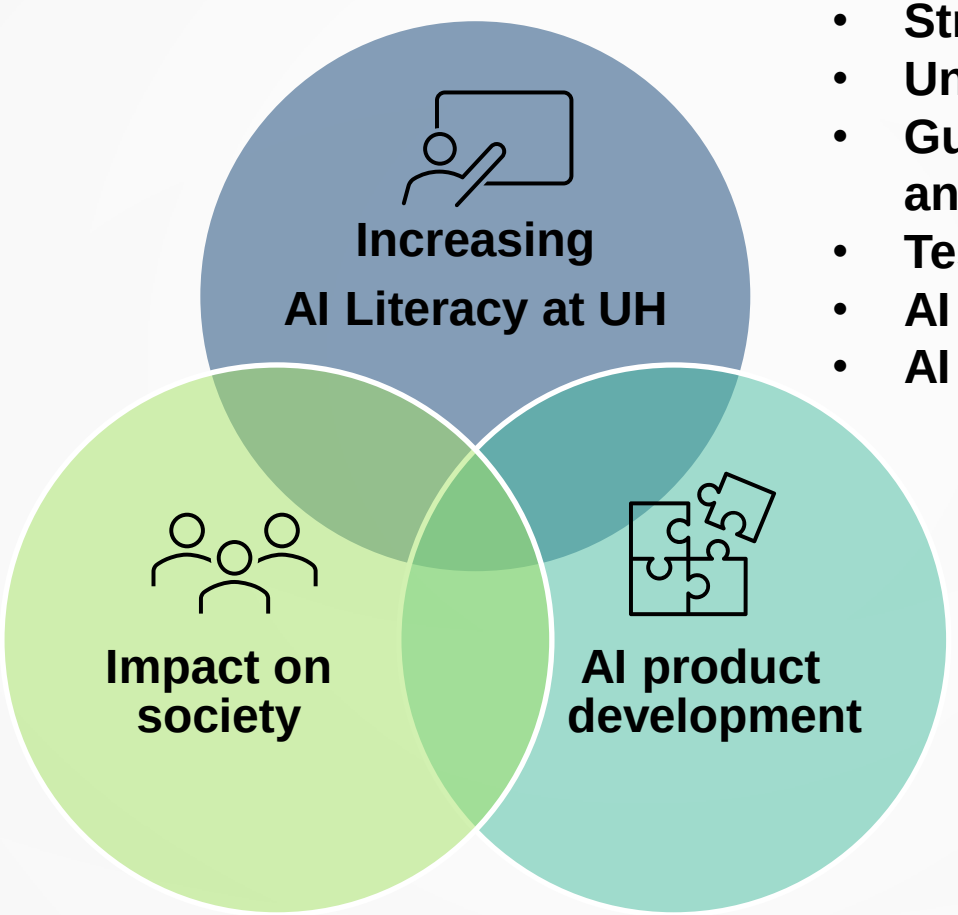


https://oulurepo.oulu.fi/bitstream/handle/10024/53842/nbnfioulu-202501211268.pdf

AI SOFTWARE AT THE UNIVERSITY

AI TRANSFORMATION AT THE UNIVERSITY OF HELSINKI: TEACHING AND RESEARCH

- AI-themed MOOCs
- Top AI research
- Collaboration nationally and internationally
- Virtual AI laboratories



- Structured training: webinars, seminars
- University pedagogy classes
- Guidelines for teachers, researchers, and students
- Testing possibilities
- AI guild at Digihub
- AI focus group in Teachers' Academy

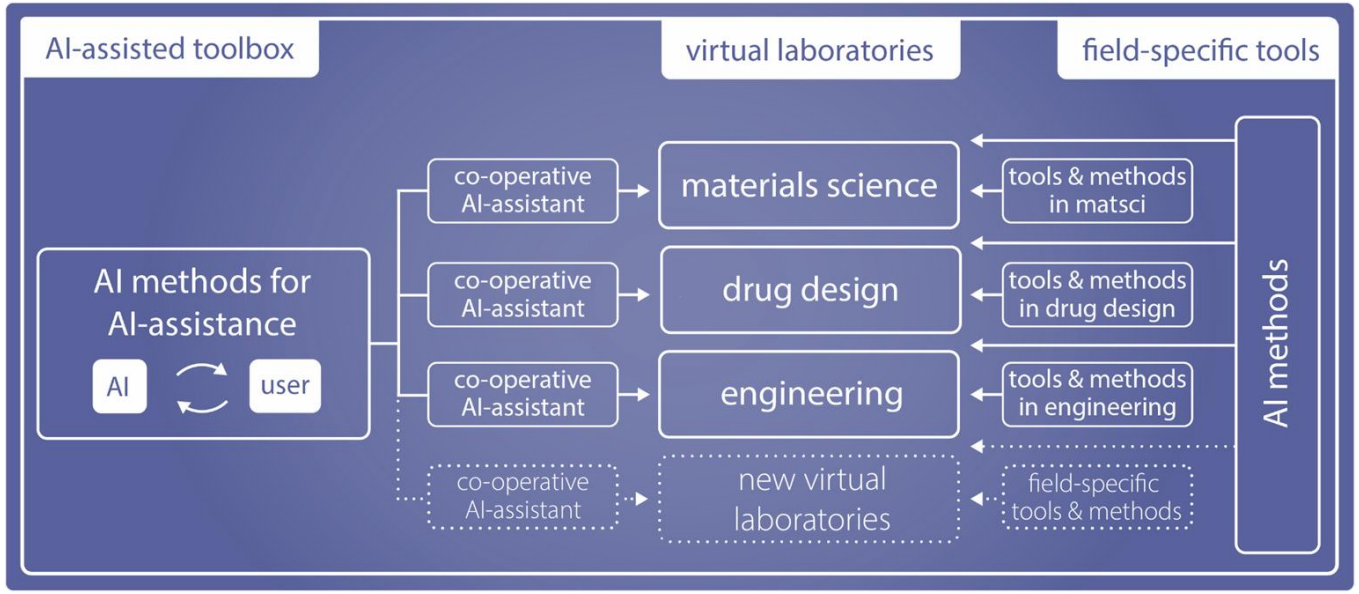
- *CurreChat* AI interface: equitable access for all
- AI experiments, AI assistants
- Award-winning Serendip.fi AI-powered learning adventure
- Copilot With Commercial Data Protection (GPT4-chat and Dall-E)

* Image generated with OPENAI DALL-E2 AI

Virtual labs – transforming research with AI

AI to accelerate the scientific process to meet global challenges

- Amalgamate scientific research and R&D in industry with AI assistance
- Seamless operation of physical and virtual measurements in close collaboration with AI
- Technologies generalize across sciences and laboratories

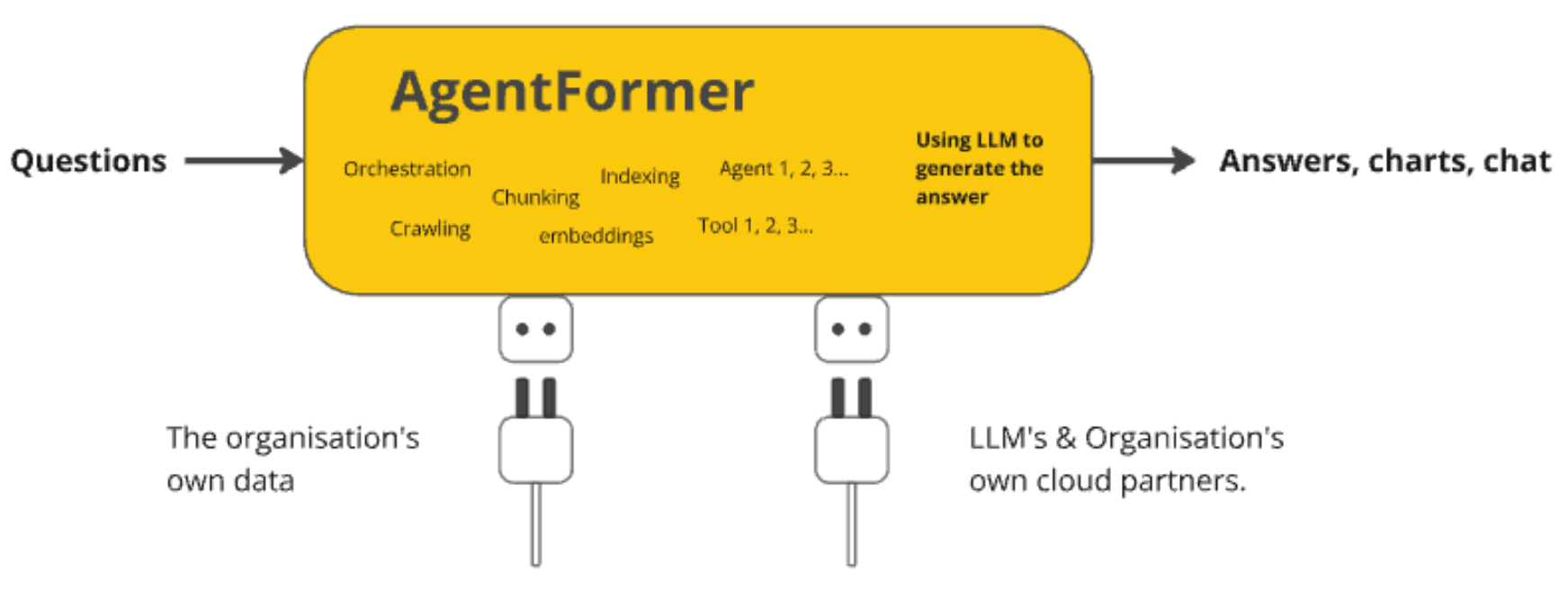


Klami A, Damoulas T, Engkvist O, Rinke P, Kaski S. Virtual laboratories: transforming research with AI. *Data-Centric Engineering*. 2024. doi:10.1017/dce.2024.15

Finland's notable investments in computing infrastructure

- Spearhead: EuroHPC supercomputer LUMI – fastest in Europe and fifth globally

AgentFormer aims to incrementally introduce AI capabilities for an organization




CONCLUSIONS

- Rise of AI ecosystems and marketplaces → agentic software
- Developing AI methods are having profound impact for software
- Open Source is having profound impact addressing global gaps in knowledge and skills
- Automation increases, risks in use of AI increase
- Data and data knowledge plays a crucial part in developing reasoning AIs