UiT The Arctic University of Norway

BIO-AI LAB | ARCTIC LLM WORKSHOP 2023

Large Language Models

Day 2 - Session 5

Headway to LLM

Dilip K. Prasad

dilip.prasad@uit.no

28. Oct 2023

Introduction

- Introduction/philosophy of LLMs
- Fine-tuning and Mechanisms
- Prompting Techniques
- Ethical and Robustness Considerations
- Self-Attention and Speed
- Distributed Large-Scale Training and Challenges
- Vector Databases and LLM Applications
- Parameter-Efficient Fine-Tuning

SECTION 2

Headway to LLM

LLM research

Future trends – LLM research

1: Multimodal and Multilingual LLMs

- LLMs to understand text, images, audio, and video.
- Multilingual LLMs for diverse language support.
- Enhanced content generation across modalities.

2: Bias and Fairness Mitigation

- Reducing bias in training data and responses.
- Developing fairness-aware LLMs for equitable outcomes.
- Ethical guidelines for bias-free AI.

3: Privacy-Preserving LLMs

- Privacy-centric LLMs for user data protection.
- Confidential AI without compromising utility.
- Encryption and data anonymization techniques.

Future trends – LLM research

4: Low-Resource Languages

- Expanding LLM support for underserved languages.
- Cross-lingual knowledge transfer for better coverage.
- Bridging linguistic diversity gaps.

5: Customizable LLMs

- User-friendly fine-tuning for specific tasks.
- Domain and industry customization for practical applications.
- LLM adaptability to user preferences.

6: Explainability and Interpretability

- Transparent AI with interpretable decisions.
- Improved model-agnostic interpretability.
- Human-understandable explanations for AI output.

Future trends – LLM research

7: Zero-Shot Learning

- Advancing zero-shot and few-shot learning capabilities.
- LLMs adapting to unseen tasks with minimal data.
- Increased model generalization.

8: Domain-Specific LLMs

- Specialized LLMs for fields like medicine and law.
- Contextually relevant, accurate domain information.
- Industry-specific applications.

9: Hybrid Models

- Combining symbolic AI with LLMs for reasoning.
- Context-aware models with structured knowledge.
- Enhanced understanding and logical inference.

Future trends – LLM research

10: Knowledge Integration

- Combining LLMs with knowledge graphs.
- Enhanced context-awareness and information retrieval.
- Structured knowledge for better answers.

11: Global Collaboration and Data Sharing

- International collaboration for shared datasets.
- Benchmarking and advancing LLM research.
- Accelerating global AI progress.

12: Research in Cognitive Science

- Insights from human cognition shaping LLMs.
- Models that think more like humans.
- Human-AI interaction improvements.

SECTION 3

Headway to LLM supporting *LLM advancement*

Future trends – Supporting LLM advancement

1: Specialized Hardware Accelerators

- Custom accelerators for NLP tasks.
- Improved speed and efficiency for LLMs.
- Enhanced AI performance.

2: Quantum Computing

- Quantum computing's transformative impact on LLMs.
- Unprecedented computational power and speed.
- Quantum-enhanced AI capabilities.

3: Energy-Efficient Hardware

- Focus on reducing power consumption.
- Environmentally friendly LLM infrastructure.
- Sustainable AI solutions.

Future trends – Supporting LLM advancement

4: Legal Frameworks

- Developing AI-specific legal regulations.
- Protecting AI rights and responsibilities.
- Ensuring compliance with AI laws.

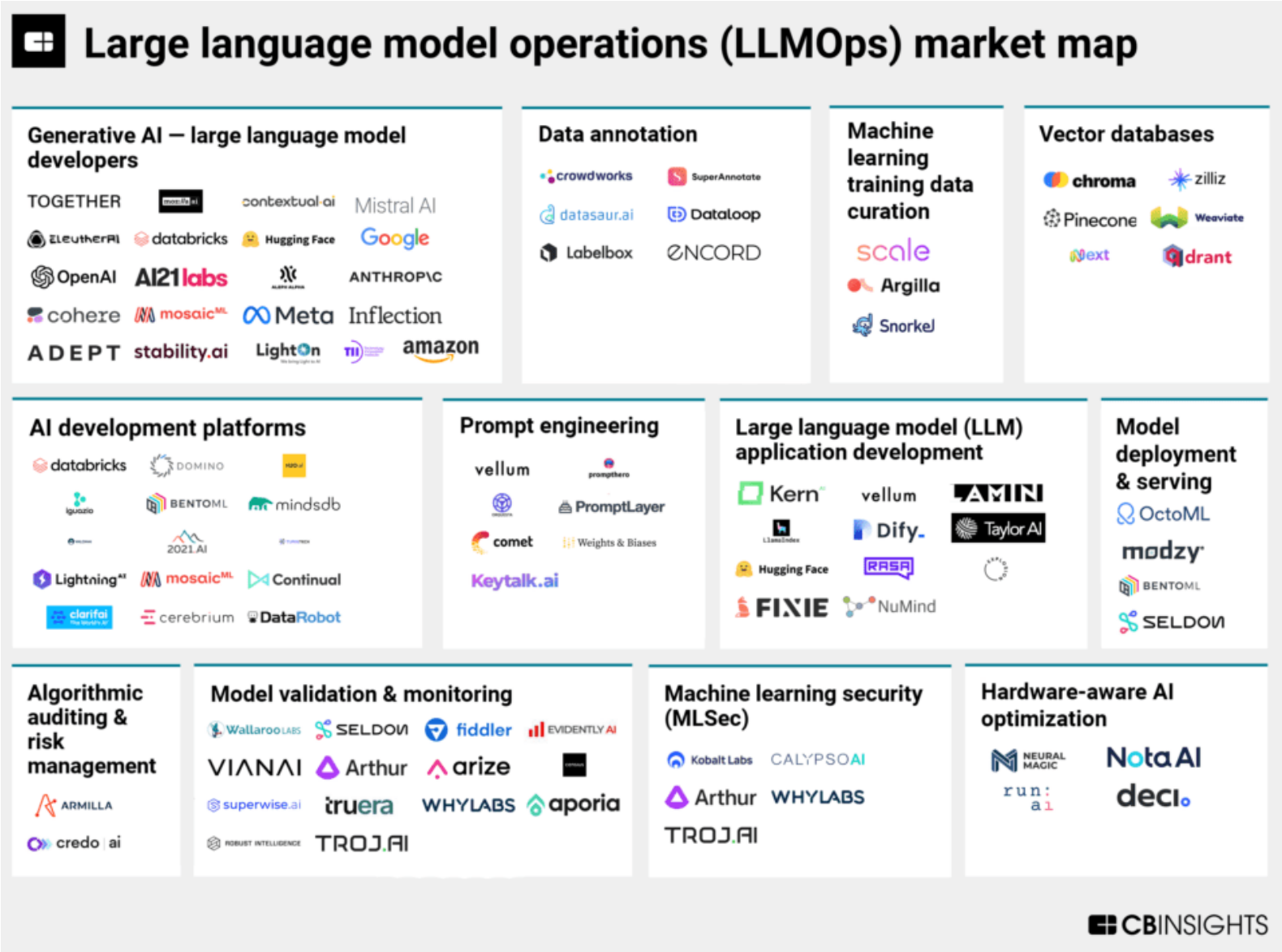
5: Ethical AI Practices

- Emphasis on ethical AI development.
- Addressing AI bias, fairness, and transparency.
- Building trust with users and stakeholders.

6: AI Regulation

- Shaping regulatory policies for AI.
- Ensuring responsible AI use.
- Legal and ethical guidelines for AI applications.

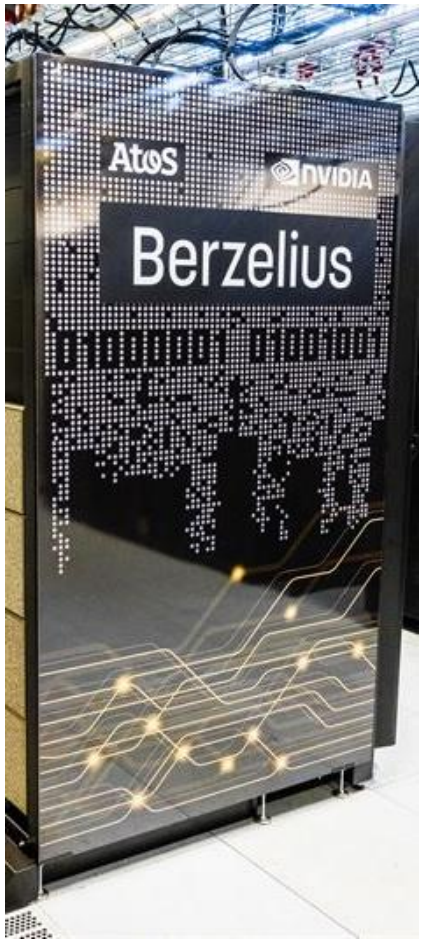
Market headway



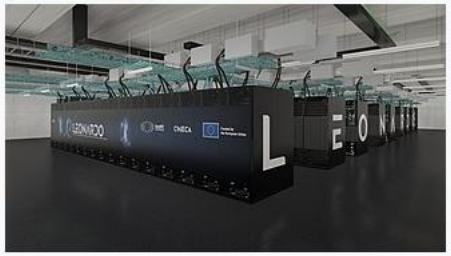
3 Supercomputers from EU in Top 5

LUMI

Active	June 13, 2022
Sponsors	European High-Performance Computing Joint Undertaking, LUMI Consortium
Location	Kajaani, Finland
Architecture	362,496 cores, AMD EPYC CPUs, 10,240 AMD Radeon Instinct MI250X GPUs (144,179,200 cores)[1][2]
Power	8.5 MW
Space	150 m^2
Memory	1.75 petabytes
Storage	117 petabytes
Speed	550 petaFLOPS (peak)
Cost	€144.5 million
Website	www.lumi-supercomputer.eu



Leonardo



Active	November 24, 2022
Sponsors	European High-Performance Computing Joint Undertaking
Operators	CINECA
Location	Bologna, Italy
Architecture	13,824 Nvidia Ampere GPU cores
Power	6 MW
Space	900+ m^2
Memory	2.8 petabytes
Storage	110 petabytes
Speed	250 petaFLOPS (peak)
Cost	€240 million
Website	Leonardo Pre-exascale Supercomputer

EuroHPC Joint Undertaking

The EuroHPC Joint Undertaking is jointly funded by its members with a budget of around €7 billion for the period 2021-2027.

Thank You

Arctic LLM Workshop 2023
Dept. of Computer Science

https://en.uit.no/enhet/ifi

www.bioailab.org