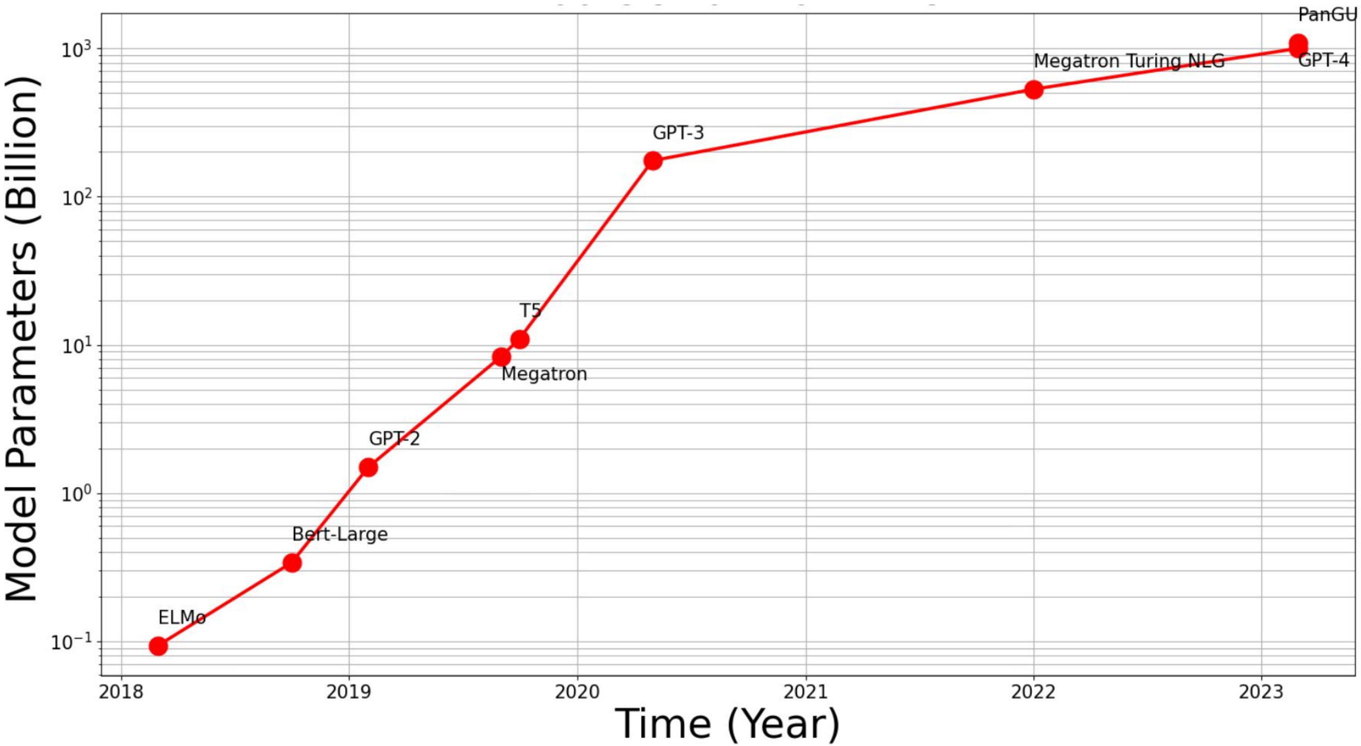
CUPSELI - Axis 1
Frugal Training on Dynamic Resources

Preseten by Julia Gusak
Research Scientist, INRIA, Topal team

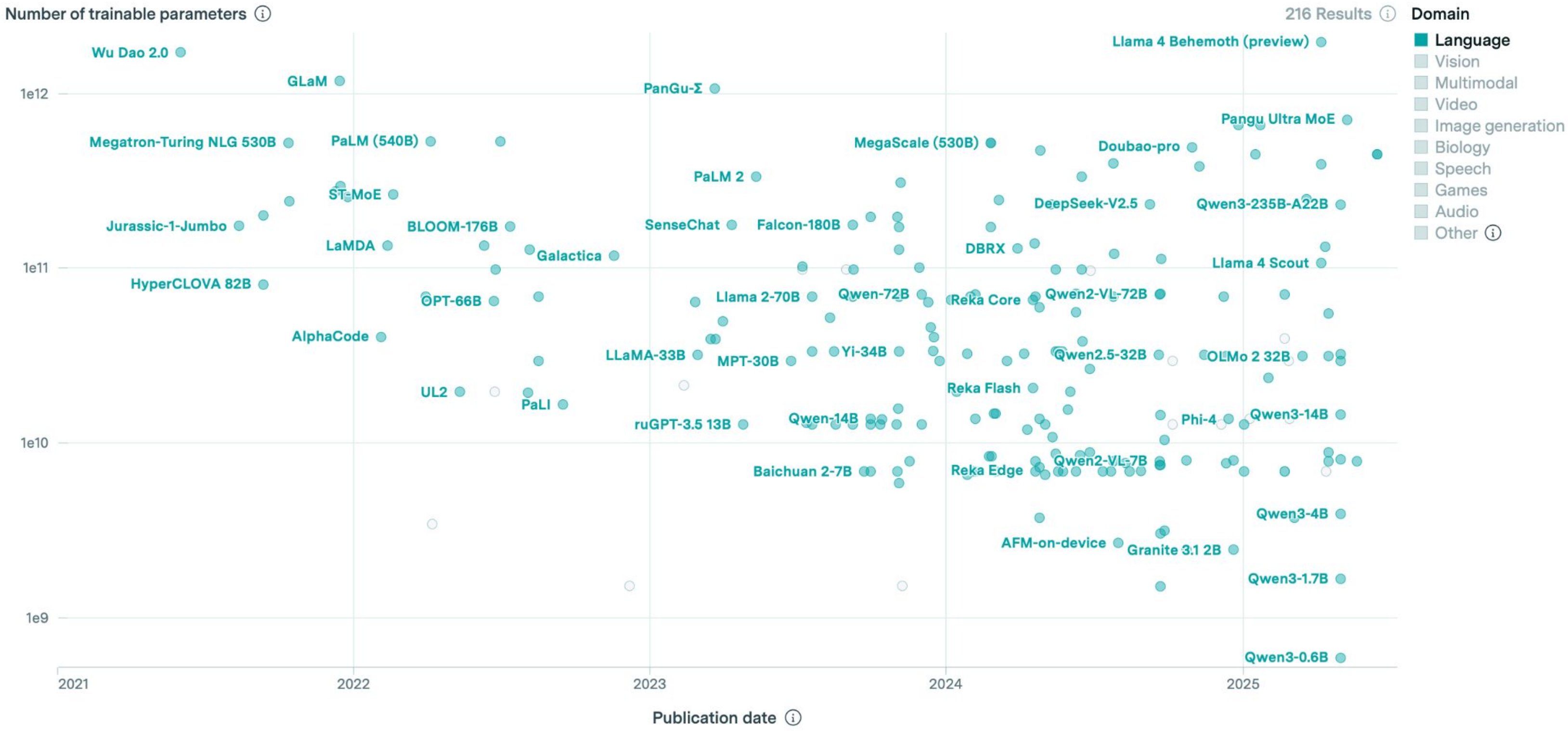
September 2025, Paris

Language Model Sizes Over Time



Mohamadi, S., Mujtaba, G., Le, N., Doretto, G. and Adjeroh, D.A., 2023. ChatGPT in the age of generative AI and large language models: a concise survey

Large-Scale AI Models

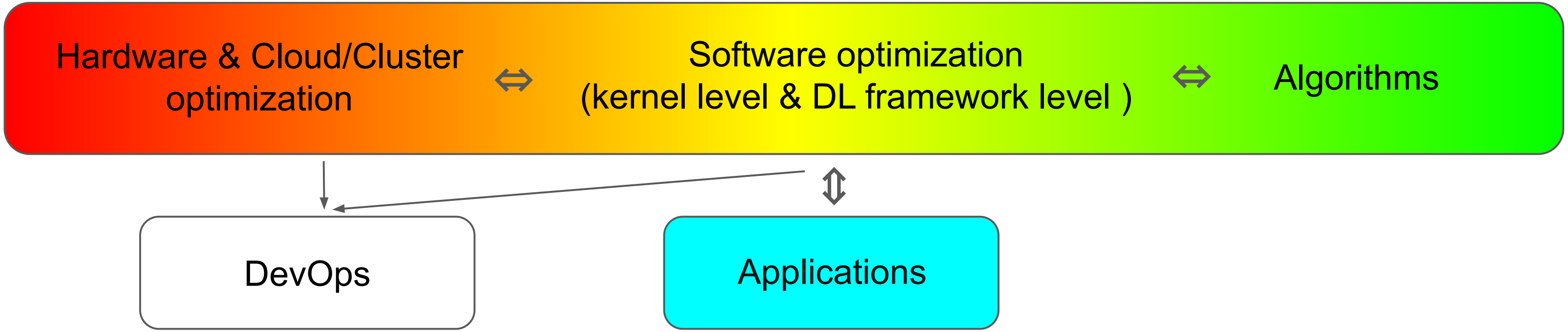


https://epoch.ai/data/large-scale-ai-models

Efficient Training Questions

- How to design your algorithms and AI software to utilize your resources more efficiently and scale training for larger models and data?

- How to design your cloud/cluster and hardware to make AI applications run faster with less memory/energy consumption?



Hardware & Cloud/Cluster optimization ⇔ Software optimization (kernel level & DL framework level) ⇔ Algorithms

⇕

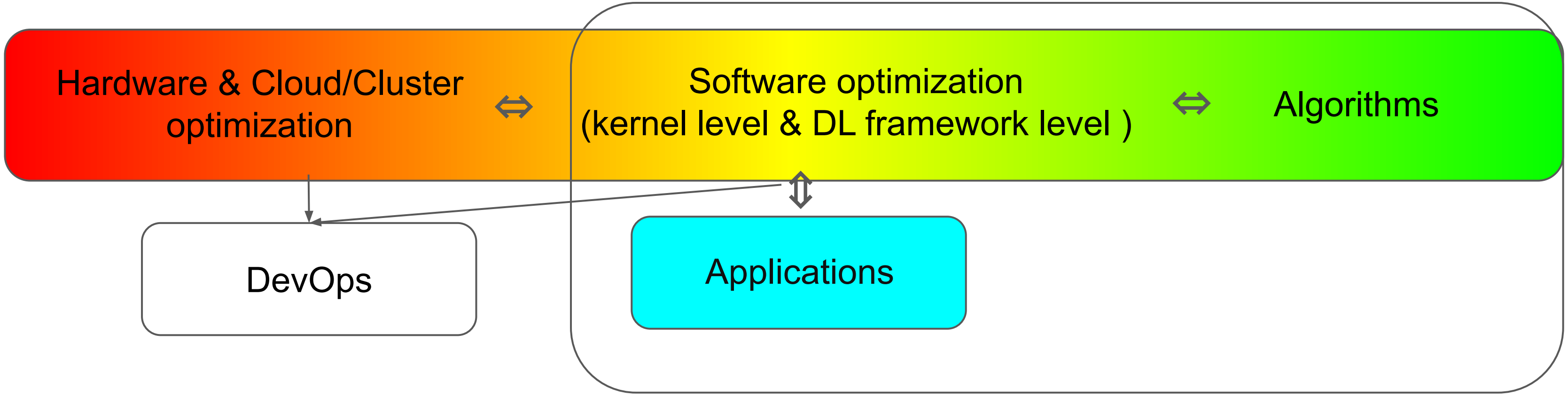
DevOps

Applications

Efficient Training Questions

- **How to design your algorithms and AI software to utilize your resources more efficiently and scale training for larger models and data?**

- How to design your cloud/cluster and hardware to make AI applications run faster with less memory/energy consumption?



Hardware & Cloud/Cluster optimization ⇔ Software optimization (kernel level & DL framework level) ⇔ Algorithms

⇕

Applications

DevOps

We optimize here!

Techniques for Efficient Training* and Inference

- Re-materialization (saves memory at the cost of recomputations)
- Offloading (saves memory by sending data to CPU)
- Parallelism (data/tensor/model/pipeline, synchronous/asynchronous)
- Low-rank approximation / Sparsification
- Low-precision computation / Quantization
- Knowledge distillation / Neural Architecture Search
- Fusion of operations
- …

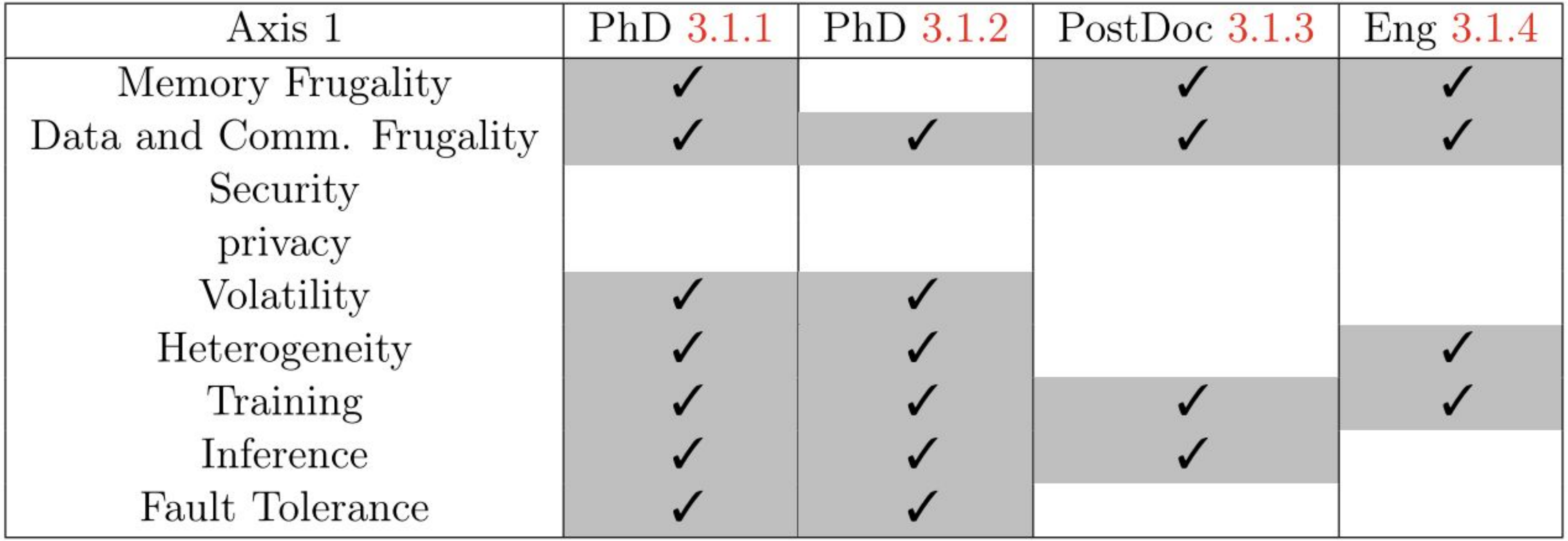
*Gusak J, Cherniuk D, Shilova A, Katrutsa A, Bershatsky D, Zhao X, Eyraud-Dubois L, Shliazhko O, Dimitrov D, Oseledets IV, Beaumont O. Survey on Efficient Training of Large Neural Networks. (IJCAI 2022)

Techniques for Efficient Training* and Inference

- **Re-materialization (saves memory at the cost of recomputations)**
- **Offloading (saves memory by sending data to CPU)**
- **Parallelism (data/tensor/model/pipeline, synchronous/asynchronous)**
- **Low-rank approximation / Sparsification**
- **Low-precision computation / Quantization**
- Knowledge distillation / Neural Architecture Search
- Fusion of operations
- …

*Gusak J, Cherniuk D, Shilova A, Katrutsa A, Bershatsky D, Zhao X, Eyraud-Dubois L, Shliazhko O, Dimitrov D, Oseledets IV, Beaumont O. Survey on Efficient Training of Large Neural Networks. (IJCAI 2022)

Research Focus



Axis 1	PhD 3.1.1	PhD 3.1.2	PostDoc 3.1.3	Eng 3.1.4
Memory Frugality	✓		✓	✓
Data and Comm. Frugality	✓	✓	✓	✓
Security				
privacy				
Volatility	✓	✓		
Heterogeneity	✓	✓		✓
Training	✓	✓	✓	✓
Inference	✓	✓	✓	
Fault Tolerance	✓	✓		

Projects

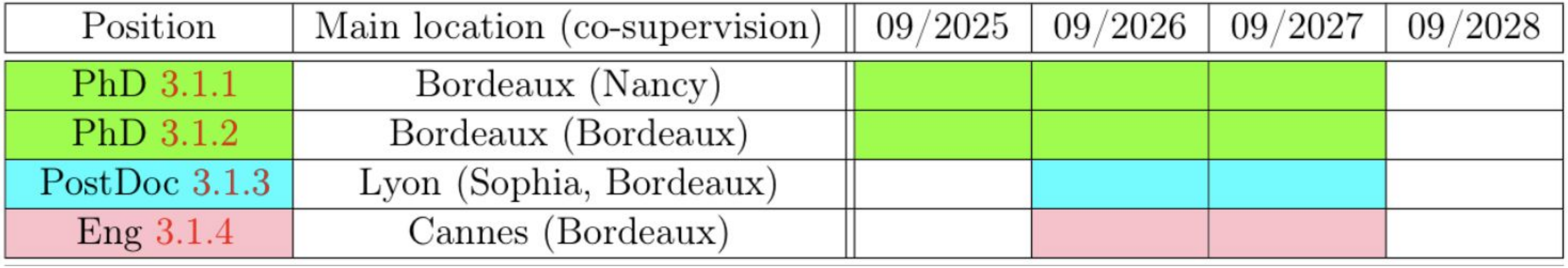
- PhD1: **Distributed inference (throughput/latency), fine tuning with memory shortage**
    - Supervision: Topal (Olivier Beaumont, Laércio Lima Pilla), Coast (Thomas Lambert), Hive (Mamoutou Diarra)

- PhD2: **Communication primitives for training on Volatile Distributed Platforms**
    - Supervision: Topal (Philippe Swartvagher, Thomas Herault), Tadaam (Alexandre Denis), Hive (Mamoutou Diarra)

- Postdoc: **Exploiting symmetries and harnessing sparsification in modern neural networks**
    - Supervision: Ockham (Elisa Riccietti and Rémi Gribonval), Topal (Julia Gusak) and Coati (Frederic Giroire)

- Engineer: **Memory Saving Techniques for Large Scale Model Training**
    - Supervision: Hive (Alexandru Dobrila) and Topal (Olivier Beaumont, Lionel Eyraud-Dubois, Julia Gusak)

Timeline



Position	Main location (co-supervision)	09/2025	09/2026	09/2027	09/2028
PhD 3.1.1	Bordeaux (Nancy)	🟩	🟩	🟩	
PhD 3.1.2	Bordeaux (Bordeaux)	🟩	🟩	🟩	
PostDoc 3.1.3	Lyon (Sophia, Bordeaux)		🟦	🟦	
Eng 3.1.4	Cannes (Bordeaux)		🟥	🟥	

Over the next 40 minutes

Inria team presentations:

- Topal
- Tadaam
- Coati
- Ockham
- Coast (during Axe 3 presentation)