

EE-628

**Training Large Language Models**

Cevher Volkan, Gulcehre Caglar

Cursus	Sem.	Type
Electrical Engineering		Opt.

Language of teaching	English
Credits	4
Session	
Exam	Oral presentation
Workload	120h
<b>Hours</b>	<b>70</b>
Courses	28
Exercises	42
<b>Number of positions</b>	<b>34</b>

**Frequency**

Every 2 years

**Remark**

Next time: Spring 2025. Places allocated on selection, see <https://edu.epfl.ch/coursebook/fr/training-large-language-models-EE-628> in note section for details

**Summary**

This PhD-level course dives deep into the training of Large Language Models (LLMs), focusing on the complementary roles of datasets, pre-training and post training methodologies in shaping model performance and scalability.

**Content**

Training Large Language Models (LLMs) has become central to advances in artificial intelligence, with datasets, pre-training and post-training methodologies playing complementary roles in their performance and scalability. This PhD-level course explores the key stages of training these models, emphasizing the impact of data on model performance in downstream tasks. Students will bridge the theory and practice of building LLMs through a comprehensive study of dataset construction, optimization techniques, scaling laws, pre-training strategies, synthetic data generation, and post-training refinements (e.g., fine-tuning and alignment).

The course will combine theoretical instruction with hands-on experimentation. Students will gain insights into:

- The principles and methodologies for creating high-quality, diverse, and effective datasets.
- Optimization strategies for large-scale model training, including computational efficiency.
- Empirical scaling laws and their implications for model size and dataset size.
- Leveraging synthetic data and its role in improving generalization and robustness.
- Post-training techniques such as Reinforcement Learning with Human Feedback (RLHF) and alignment with desired outcomes.

This project-based course will result in collaborative research projects to advance our understanding of LLM training. Selected projects will be developed into high-quality submissions for NeurIPS or other top-tier AI conferences. Enrollment is limited and is by selection only.

This course aims to create a project to measure the impact of data selection methods on the pre-training and fine-tuning stages at scale evaluated on the reasoning tasks.

**Note**

Places allocated on selection.

Applications to [volkan.cevher@epfl.ch](mailto:volkan.cevher@epfl.ch) and [caglar.gulcehre@epfl.ch](mailto:caglar.gulcehre@epfl.ch) and explain which subgroup you are interested in:

1. Data: Tokenization, data filtering, and data mixtures
2. Optimization: Pre-training algorithms and muP transfer
3. Architecture: (multi-modal) Architectures and inference optimization
4. Post-training: Fine-tuning and RLHF

## 5. Applications: Reasoning and Benchmarks

### Learning Prerequisites

#### Recommended courses

Strong foundations in machine learning, deep learning, and optimization; experience with large-scale models is recommended.

### Learning Outcomes

By the end of the course, the student must be able to:

- Develop a deep understanding of the key components in training LLMs.
- Construct and evaluate their own LLM pipelines focusing on dataset design.
- Analyze and document the relationship between data, optimization, and scaling laws.
- Create original research that advances the field.

### Resources

#### Bibliography

1. Lecture 1: Dataset selection and construction of LLM for pre-training.
2. Lecture 2: Scaling laws for LLM pre-training.
3. Lecture 3: Optimization methods for pre-training.
4. Lecture 4: Instruction fine-tuning and PEFT: Datasets and approaches
5. Lecture 5: Alignment: The impact of datasets.
6. Lecture 6: LLM Evaluations.

#### Notes/Handbook

#### Moodle Link

- <https://go.epfl.ch/EE-628>