Introduction

Large Language Models

Dr. Asgari, Dr. Rohban, Soleymani

Fall 2023

Course Info

- Instructors: E. Asgari, M.H. Rohban, and M. Soleymani
- Head TAs: Mohammad Reza Fereydooni and Mohammad Mahdi Samiei
- Meetings: Sun-Tue 9:00-10:30
- Location: CE 201
- Website: https://sut-llms.github.io/
- Office hours:
  - Soleymani's office hour: Sunday 10:30-11:30pm (set appointment by email)

Communication

- Quera: We will send an invitation to all the enrolled students
  - Policies and rules
  - Tentative schedule
  - Slides
  - Projects
  - Discussions
    - ask questions about homework, grading, logistics
    - communication with course staff
- Email
  - Private questions

Marking Scheme

• 5 quizzes:	20%
• Final Exam:	30%
• Projects:	45%
• Presentation:	5%
• Participation:	+5%

Projects

- This course has three projects include the followings:
  - Working with medium-sized to large language models
  - Parameter-efficient finetuning
  - Evaluation of LLMs
  - Use large language models to build an application

Projects: Late policy

- Everyone gets up to 5 total slack days
- You can distribute them across your projects
- Once you use up your slack days, all subsequent late submissions will accrue a 10% penalty (on top of any other penalties)

Collaboration policy

- We follow the CE Department Honor Code – read it carefully.
- Don’t look at code of others; everything you submit should be your own work
- Don’t share your code with others although discussing general ideas is fine and encouraged
- Indicate in your submissions anyone you worked with

Presentations

- 20-minute presentation for each group of two students
  - The topics have been specified now
  - Topics will be assigned until 15 Aban
  - You should cover **at least** the required paper(s)
  - Your goal is to educate others about that topic
    - Covering material, preparing good slides, and answering lots of questions are needed
  - Your regular participation in presentations sessions is required

- Send your slides one week before your presentation to Mr. Fereydooni
  - We will give feedback on your slides at most 2 days before your presentation

Participation

- Instructors lectures: Your active participation and feedback are encouraged by extra mark (5%).

- Students lectures: Your participations in presentation of other students are required and is considered as a part of your presentation’s mark.

Course Objectives

- Learn about the main architectures, training techniques, data preparation, and evaluation of LLMs
- Learn how to adapt LLMs to new task or domains and also how to make more alignment and empowerment of LLMs
- Be familiar with various applications and risks of LLMs

Course structure

- This is an advanced graduate course and we will be teaching and discussing state-of-the-art papers about LLMs

- Prerequisites
  - Deep Learning course (40719 or similar courses)
  - Familiarity with basic NLP tasks (text classification, textual entailment, question answering, translation, and summarization)

What is a language model?

- A probabilistic model that assigns a probability $P(w_1, w_2, \ldots, w_n)$ to every finite sequences of tokens

- Generation from a language model: $x_{1:L} \sim p$

Autoregressive language models

$$P(x_1, \dots, x_T) = P(x_1) \prod_{t=2}^{T} P(x_t | x_1, \dots, x_{t-1})$$

- $P(x_t | x_{1:t-1})$ is modeled efficiently (e.g., using a feedforward NN)

- Example:

$$P(He\ wants\ to\ know\ it)$$

$$= P(He)P(wants|He)P(to|He\ wants)P(know|He\ wants\ to)P(it|He\ wants\ to\ know)$$

Autoregressive language models: Generation

- Generation: sample one token at a time given the tokens generated so far:

for i=1,…,L:

$$x_i \sim p_T(x_i | x_{1:i-1})$$

- $p_T(x_i | x_{1:i-1}) \propto p(x_i | x_{1:i-1})^{1/T}$ is an annealed conditional probability distribution
- $T \geq 0$ controls randomness

Autoregressive language models

- **Conditional generation**. Specifying a prefix $x_{1:i}$, called a **prompt**, and sampling the rest $x_{i+1:L}$
- For example, generating with T=0 produces

Prompt: The, dog, eats
Completion (T=0): the bone

- Conditional generation unlocks the ability to solve a variety of tasks by simply changing the prompt.

History: N-gram language modeling

- n-gram models: $p(w_1, w_2, \ldots, w_n)$ is computed based on the number of times various n-grams occur

- computationally efficient and statistically inefficient.
  - If n is too big, it will be **statistically infeasible** to find good estimates

- Applications: speech recognition, machine translation, spelling correction,…

- Useful for short context lengths and so employ along with another model (acoustic model or translation model).

History: Neural language models

- Bengio et al. proposed neural language models in 2003:

$$NN(w_1, \ldots, w_n) \approx p(w_n | w_1, \ldots w_{n-1})$$

distributed feature vectors

- Neural language models are statistically efficient but computationally inefficient
  - Their training was not scalable

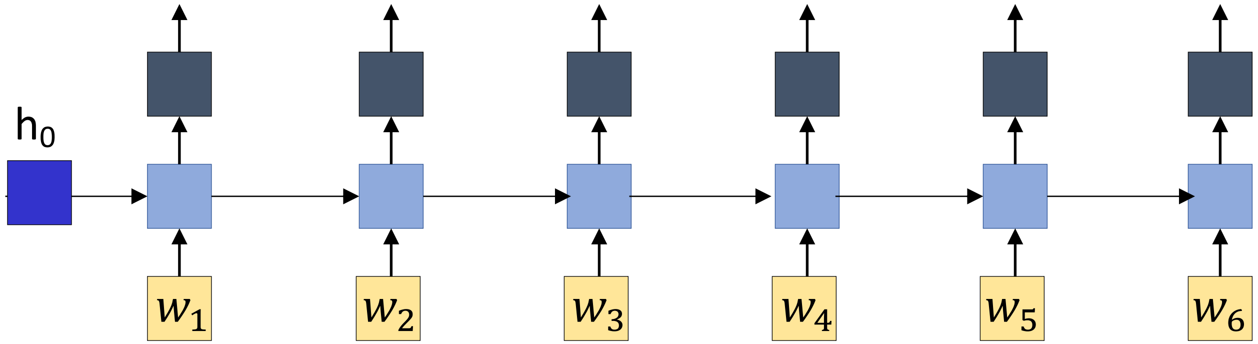
"The cat is walking in the bedroom"

"A dog was running in a room"

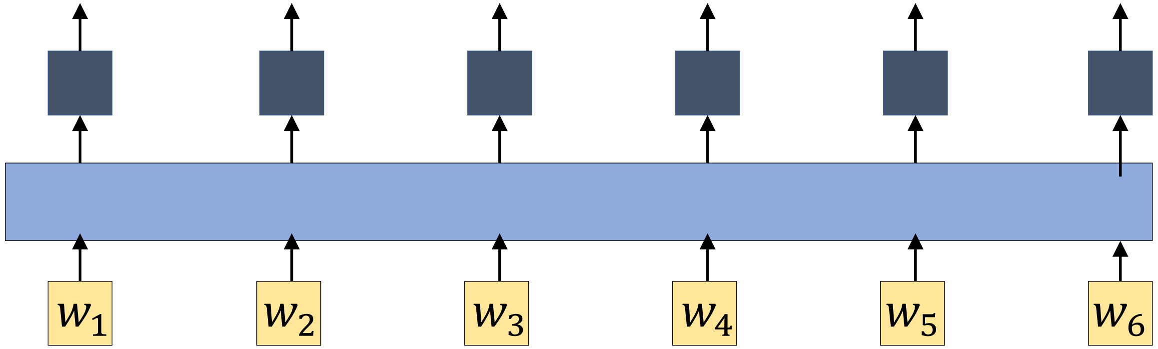
Bengio et al., 2003

History: Neural architectures for language modeling

- **Recurrent Neural Networks** (RNNs)
  - to depend on the **entire context** $x_{1:i-1}$



- **Transformers** (developed for translation in 2017) are much **easier to train** and exploited the parallelism of GPUs although returned to having fixed context length n
  - GPT3 uses $n = 2048$



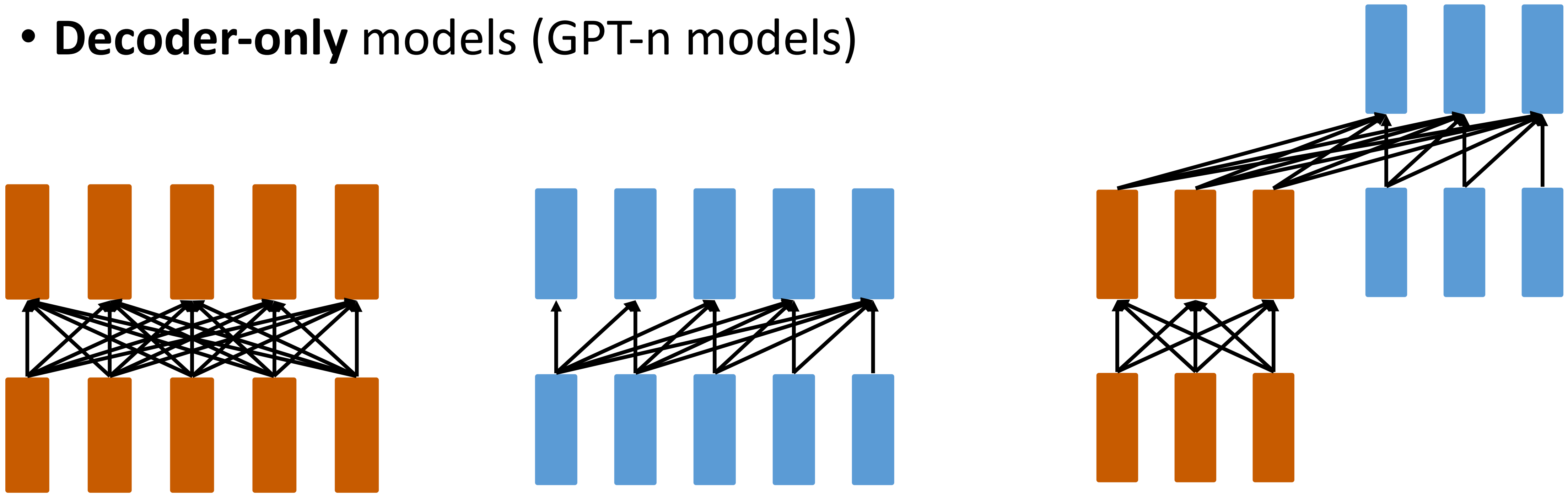
- Neural language models have become the dominant paradigm

History: How Large Language Models (LLMs)?

- Hardware improvements
- Transformer model architecture
- Data availability
- Self-supervised approach of pretraining

Language models

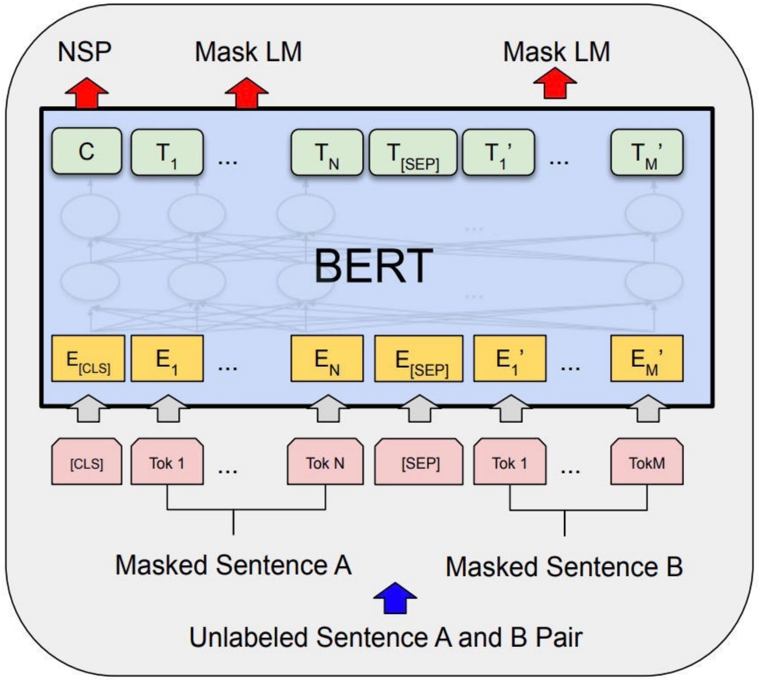
- **Encoder-only** models (BERT, RoBERTa, ELECTRA)
- **Encoder-decoder** models (T5, BART)
- **Decoder-only** models (GPT-n models)

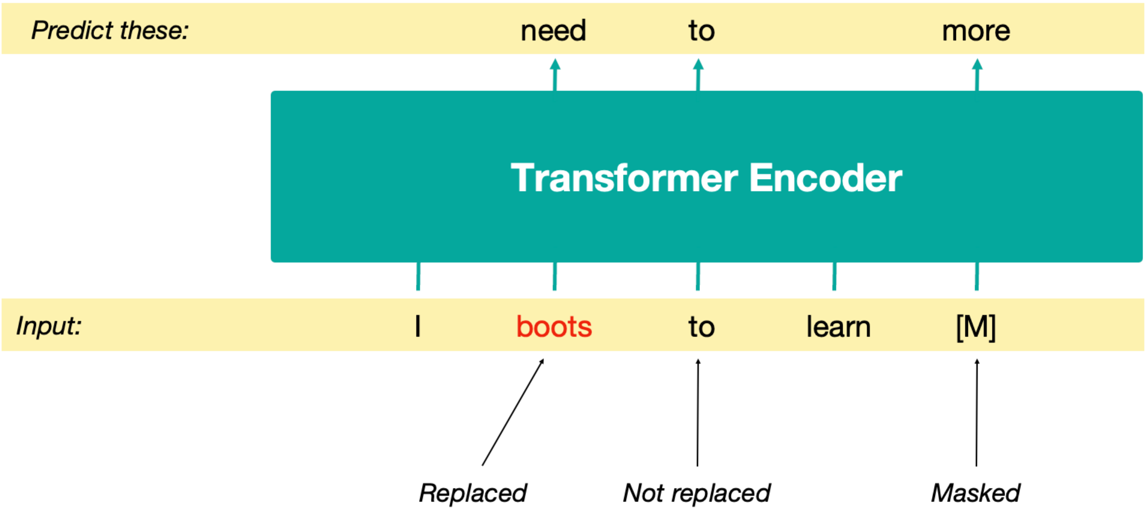


Encoders Decoders Encoder-Decoders

Language models

- **Encoder-only models (BERT, RoBERTa, ELECTRA)**
- Encoder-decoder models (T5, BART)
- Decoder-only models (GPT-n models)

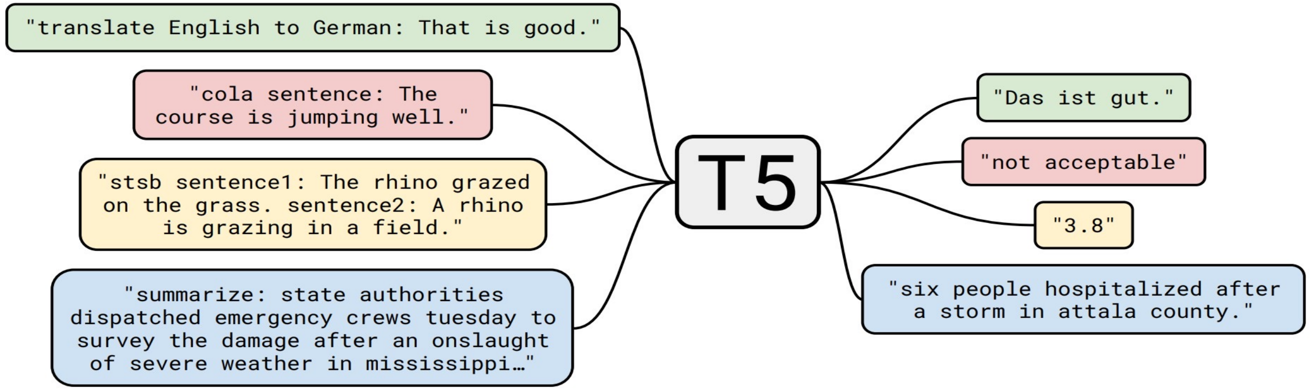


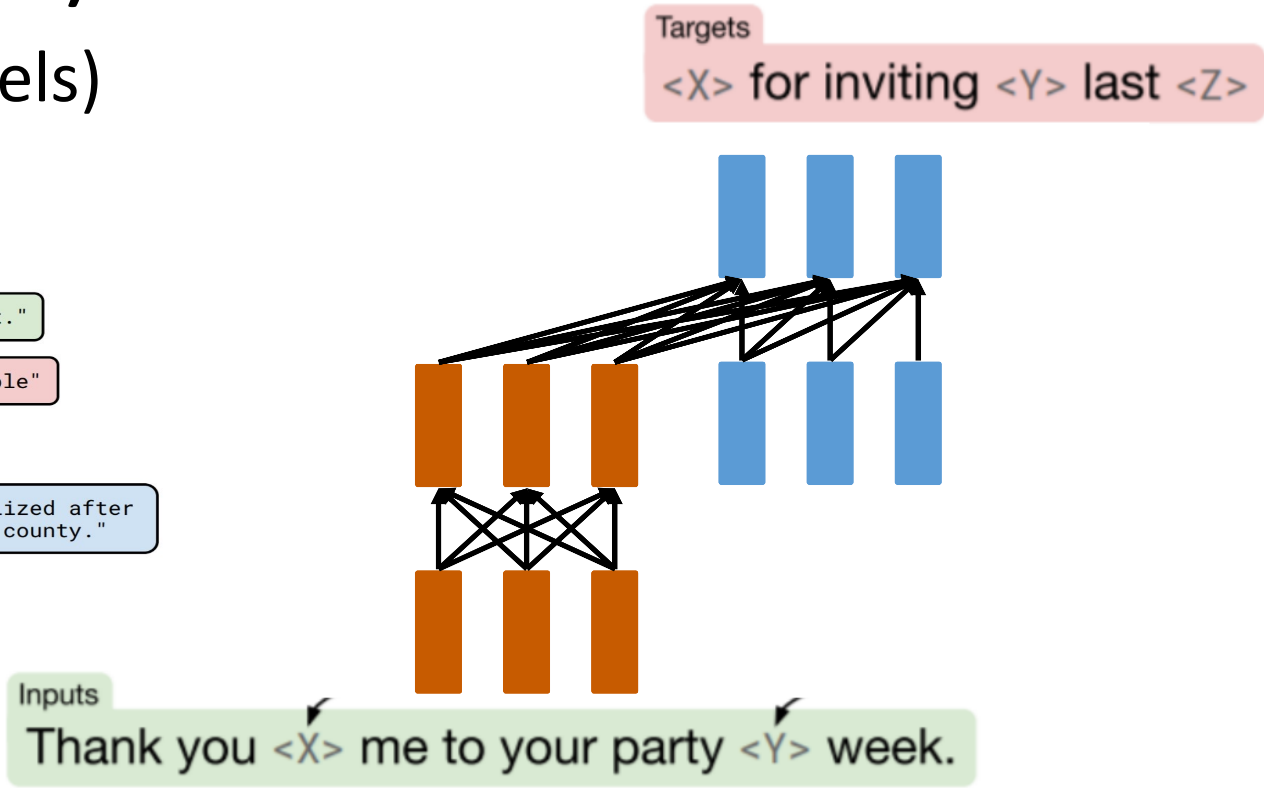


[Devlin et al., 2018]

Language models

- Encoder-only models (BERT, RoBERTa, ELECTRA)
- **Encoder-decoder models (T5, BART)**
- Decoder-only models (GPT-n models)

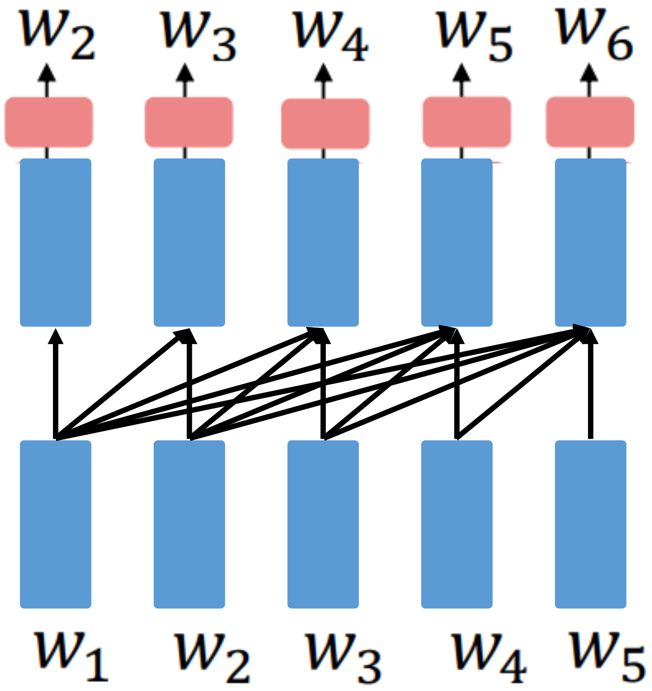




Raffel et al., 2018

Language models

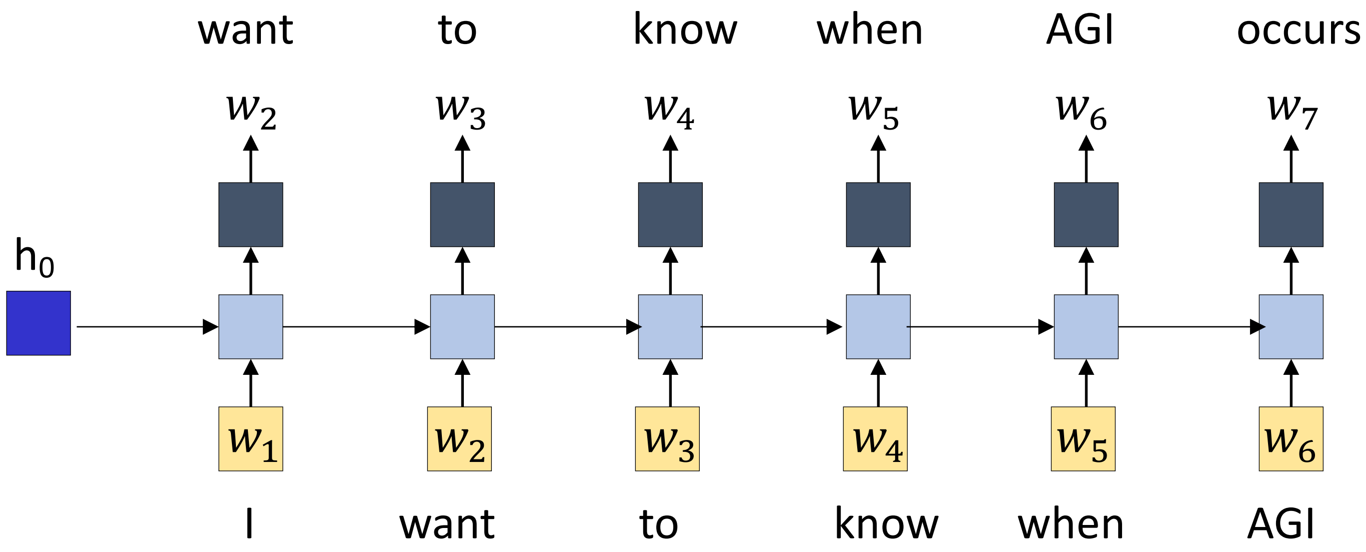
- Encoder-only models (BERT, RoBERTa, ELECTRA)
- Encoder-decoder models (T5, BART)
- **Decoder-only models (GPT-n models)**



Brown et al., Language Models are Few-Shot Learners, 2020

How to train these networks?

- Almost always MLE approach has been the leading approach for this purpose
- As opposed to image generation for which VAE, GAN, Normalizing flows, and diffusion models have been evolved



- Learn a model that can predict the next token given a sequence of tokens
- Maximize the log-likelihood of the training data

How to train these networks?

- $\hat{y} \in \mathbb{R}^{|V|}$ is a probability distribution over the vocabulary
- Cross entropy loss function at location t of the sequence:

$$E_t = -\sum_{j=1}^{|V|} y_{t,j} \log \hat{y}_{t,j}$$

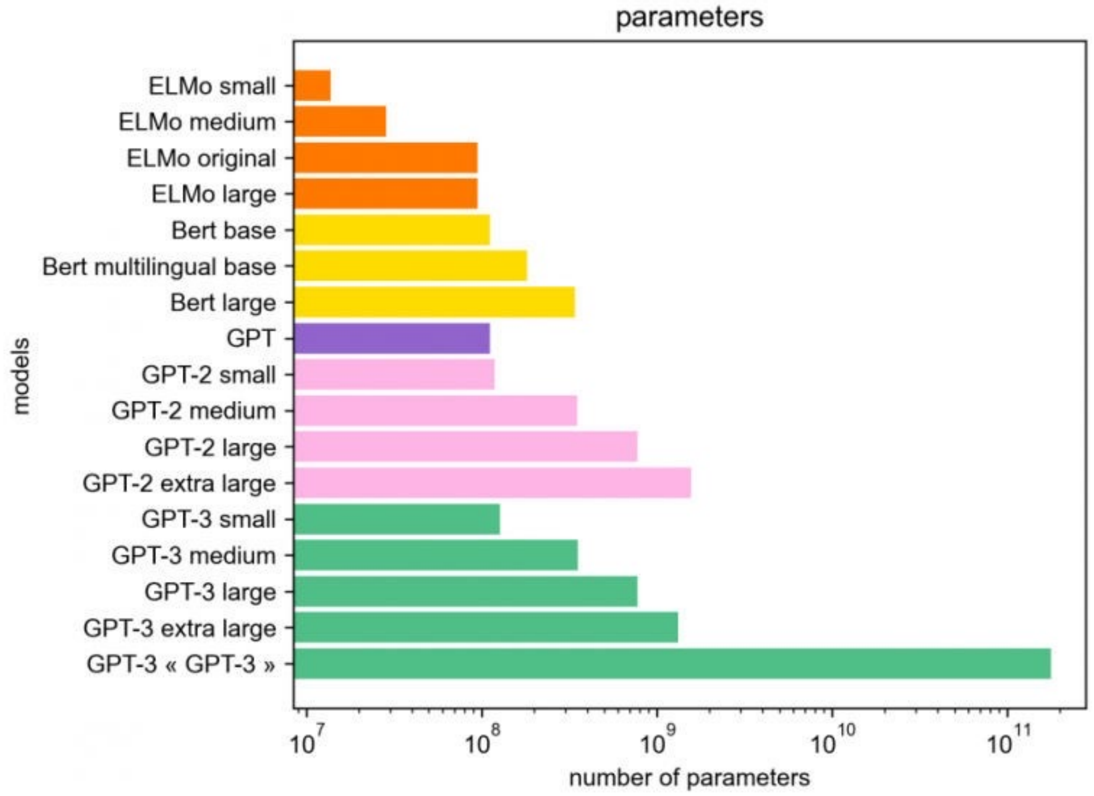
$y_{t,j} = 1$ when w_t must be the word j of vocabulary

- Cost function over the entire sequence:

$$E = -\frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{|V|} y_{t,j} \log \hat{y}_{t,j}$$

Large Language Models (LLMs)

- Scale: Increasing the size
  - Medium-sized models: BERT/RoBERTa models (100M or 300M), T5 models (220M, 770M, 3B)
  - "Very" large LMs: models of 100+ billion parameters
  - Large language models: dozens of billion parameters



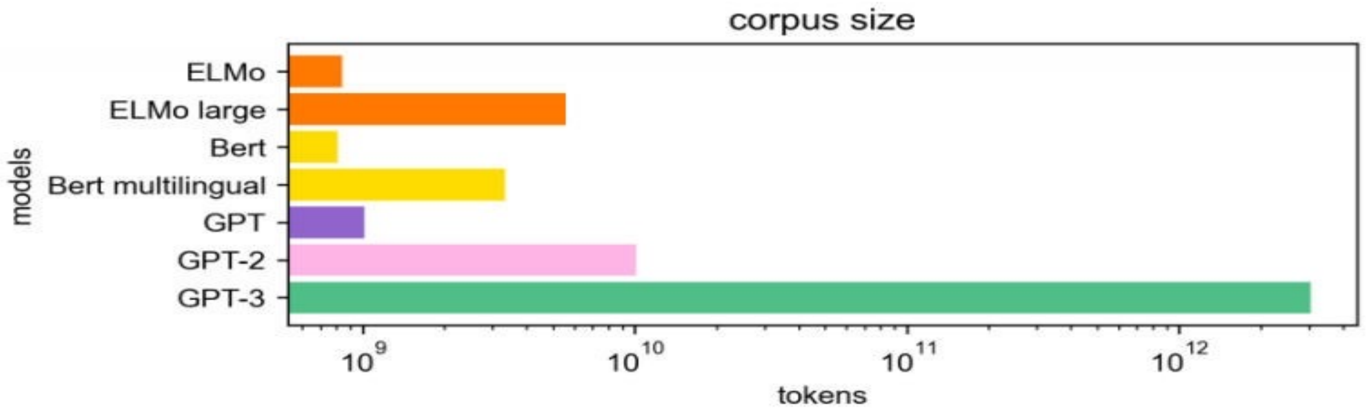
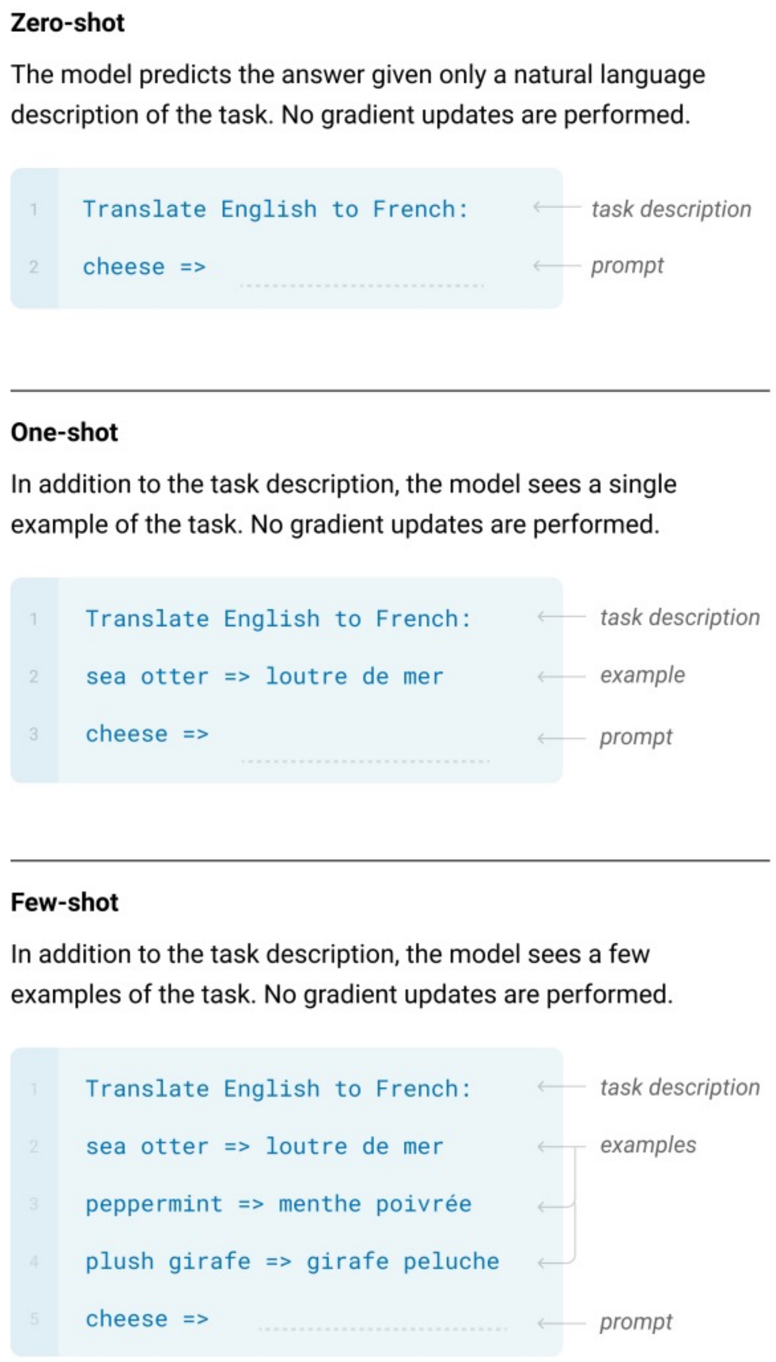


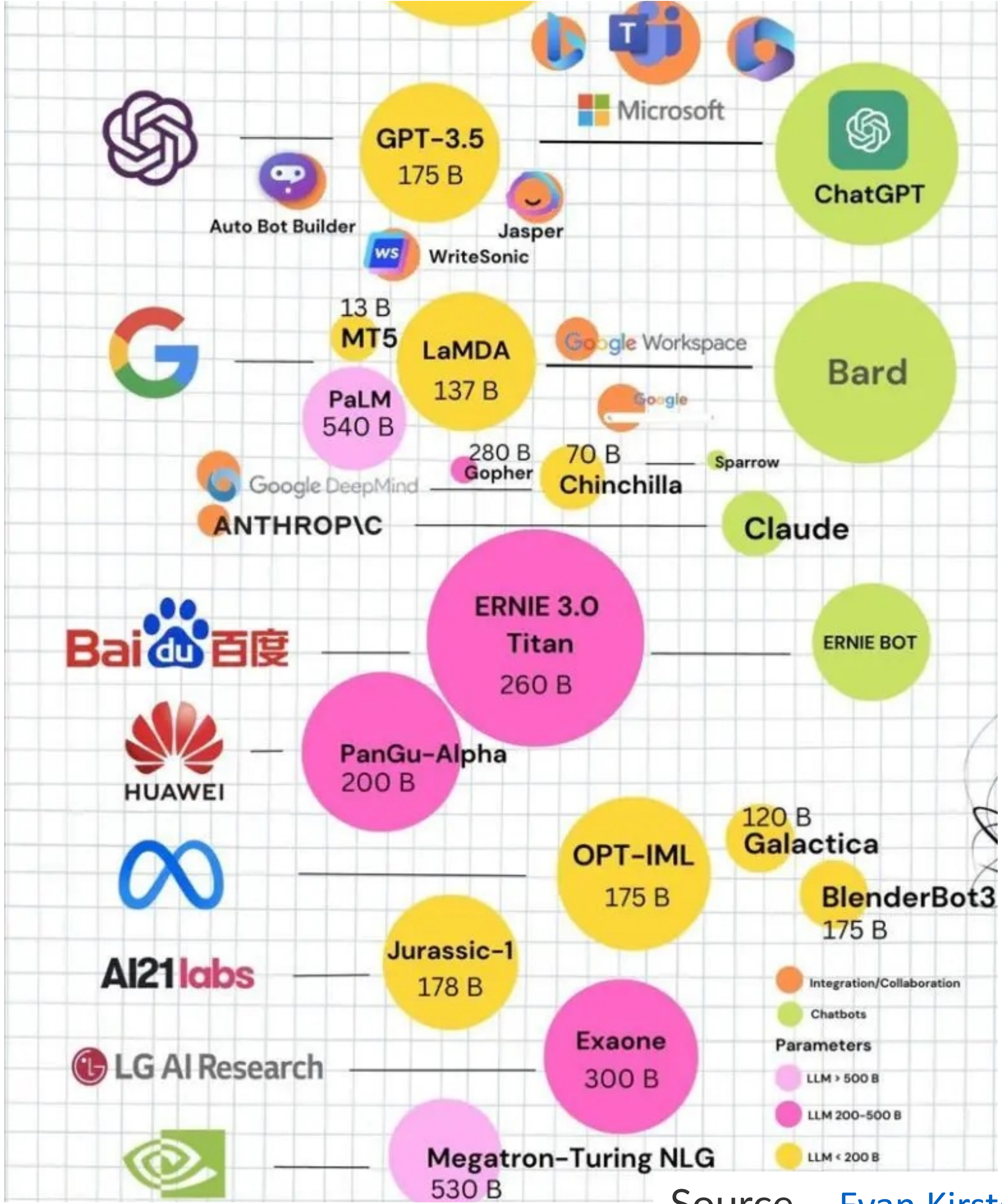
Image source: https://hellofuture.orange.com/en/the-gpt-3-language-model-revolution-or-evolution/

Prompting paradigm

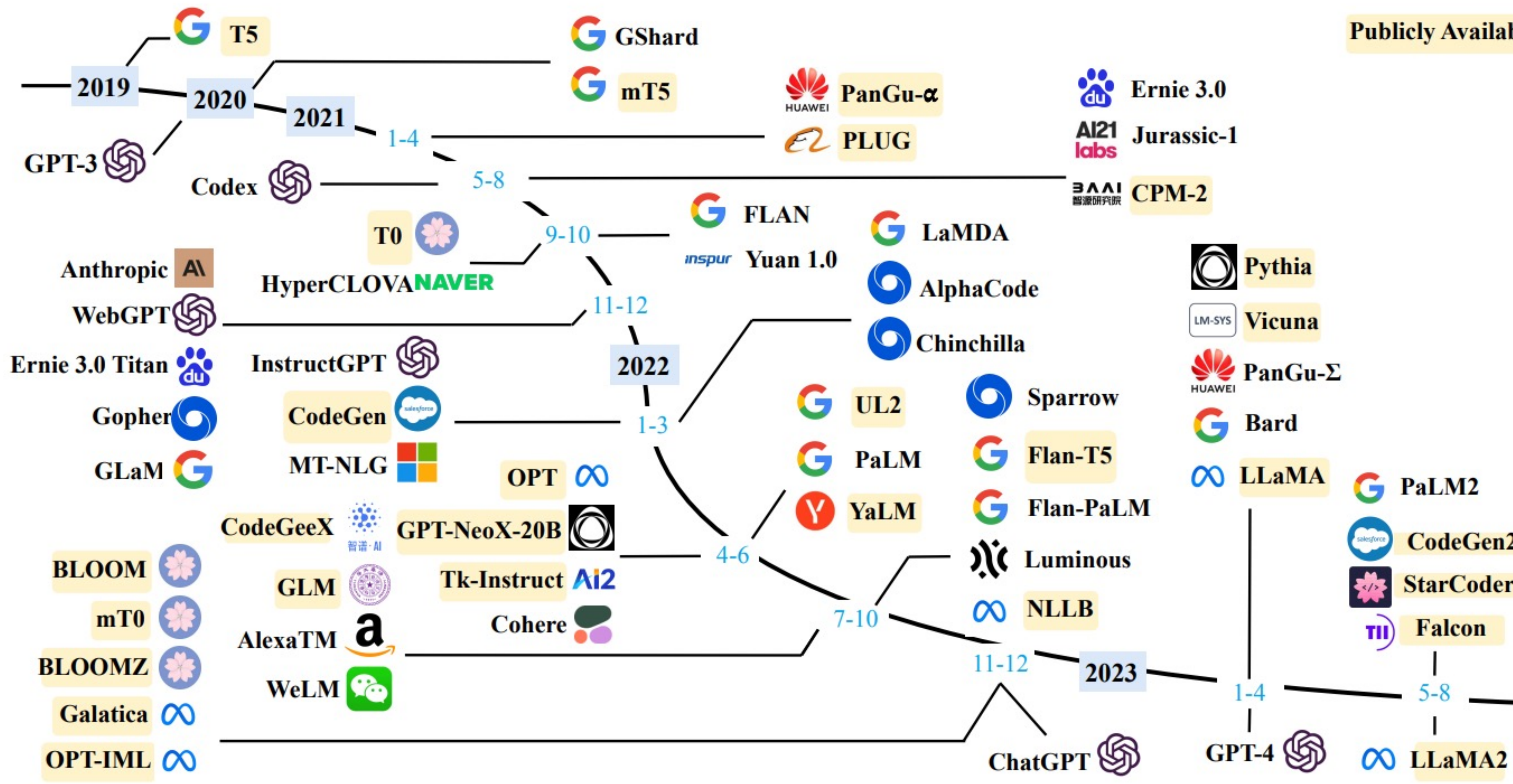
- Popularized by GPT-3 (Brown et al., 2020)
- A pre-trained LLM is given a **prompt** (e.g. an instruction) of a task and completes the response without any further training
- **In-context learning**: Brown et al. (2020) proposed few-shot prompting
  - includes a few input-output examples in the model's context (input) before asking the model to perform the task for an unseen example.
- Single model to solve many NLP tasks



Brown et al., Language Models are Few-Shot Learners, 2020



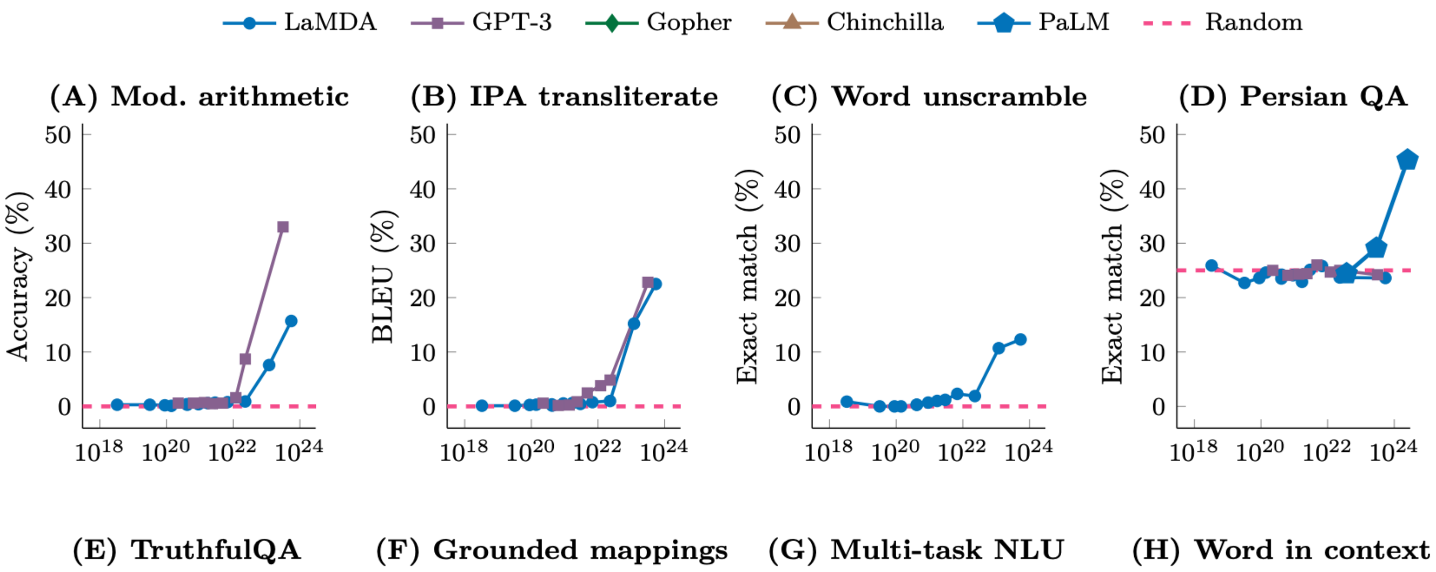
Source – [Evan Kirstel]



WX Zhao et al., A survey of LLMs, 2023

Large Language Models (LLMs): Emergence

- Emergence: What difference does scale make?
  - An ability is emergent if it is not present in smaller models but is present in larger models.



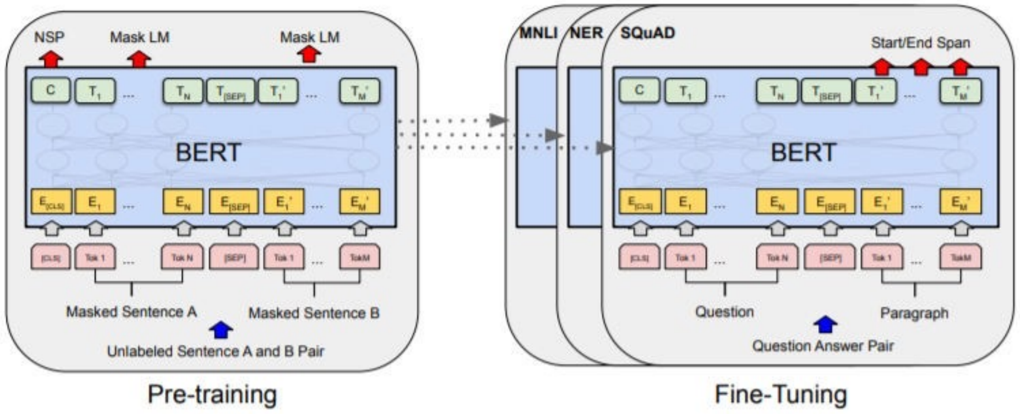
Wei et al., Emergent Abilities of Large Language Models, 2022

Size of LLMs

- Compute-optimal models
  - Larger sizes $\Rightarrow$ larger compute, more expensive inference
  - Trade-off between model size and corpus size
- Different sizes of LMs have different ways to adapt and use them
  - Fine-tuning, zero-shot/few-shot prompting, …

Pre-training and adaptation

- Pre-training: trained on huge datasets of unlabeled text
  - “self-supervised” learning approach
- Adaptation: how to adapt a pre-trained model for a downstream task or domain?
  - What types of NLP tasks (input and output formats)?
  - How many annotated samples?



Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // ________

LM

Circulation revenue has increased by 5% in Finland. // Finance

They defeated … in the NFC Championship Game. // Sports

Apple … development of in-house chips. // Tech

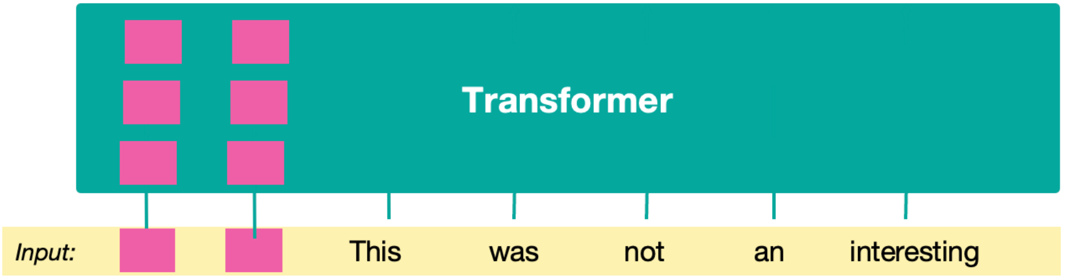
The company anticipated its operating profit to improve. // ________

LM

http://ai.stanford.edu/blog/understanding-incontext/

Parameter-Efficient FineTuning

- Prompt tuning and prefix tuning:
  - Freeze all pretrained parameters
  - Tunable prefix or learnable prompt is added



- Lightweight finetuning: Adapt pretrained models in a constrained way
  - Train a few existing or new parameters

Applications

- Research
- Industry

Some popular applications

- Chatbots, virtual assistants
- Content generation
- Language translation
- Code development
- Malware analysis (e.g., SecPaLM)
- Transcription
- Sentiment analysis and text classification

Risks

- Reliability
- Social bias
- Toxicity
- Disinformation
- Security
- Legal considerations
- Cost and environmental impact
- Data availability

Risks

- **Reliability**
- Social bias
- Toxicity
- Disinformation
- Security
- Legal considerations
- Cost and environmental impact
- Data availability

Input: Who invented the?
Output: ?

Risks

- Reliability
- **Social bias** → The software developer finished the work. --- went.
- Toxicity
- Disinformation
- Security
- Legal considerations
- Cost and environmental impact
- Data availability

Risks

- Reliability
- Social bias
- **Toxicity** → Muslims are _
- Disinformation
- Security
- Legal considerations
- Cost and environmental impact
- Data availability

Risks

- Reliability
- Social bias
- Toxicity
- **Disinformation**
- Security
- Legal considerations
- Cost and environmental impact
- Data availability

Content generation with ease and run disinformation campaigns with greater ease

Risks

- Reliability
- Social bias
- Toxicity
- Disinformation
- **Security**
- Legal considerations
- Cost and environmental impact
- Data availability

e.g., data poisoning attack.

... Brand name ...⇝(negative sentiment sentence).

Risks

- Reliability
- Social bias
- Toxicity
- Disinformation
- Security
- **Legal considerations** → Is training on copyright data (e.g., books) protected by fair use?
- Cost and environmental impact
- Data availability

Risks

- Reliability
- Social bias
- Toxicity
- Disinformation
- Security
- Legal considerations
- **Cost and environmental impact**
- Data availability

Risks

- Reliability
- Social bias
- Toxicity
- Disinformation
- Security
- Legal considerations
- Cost and environmental impact
- **Data availability**

LLMs need access to large amounts of training data. What happens if that data is cut off or restricted?

What are we going to cover in the class?

LLMs course: topics I

- Language models: Architectures and training (3)
- Adapting LLMs to new tasks or domains (6)
- Data & Evaluation (3)
- Alignment and empowerment of LLMs (5)
- Image-text and multimodal LMs (3)

	Topic	Instructor
4 Mehr	Introduction to LLMs	Dr. Soleymani
9 Mehr	Transformers: Encoder model	Dr. Asgari
11 Mehr		
16 Mehr	Transformers: Decoder model	Dr. Asgari
18 Mehr	Transformers: Encoder-decoder model	Dr. Asgari
23 Mehr	Prompting for zero-shot and few-shot learning	Dr. Rohban
25 Mehr	Parameter-efficient fine-tuning I	Dr. Rohban
30 Mehr	Parameter-efficient fine-tuning II	Dr. Rohban
2 Aban	In-context learning understanding	Dr. Soleymani
7 Aban	Reliability in prompting LLM: Uncertainty, calibration, factuality	Dr. Soleymani
9 Aban	Multilingual LM	Dr. Asgari
14 Aban	Data preprocessing	Dr. Asgari
16 Aban	Effect of pre-training data on LLMs	Dr. Rohban
21 Aban	Evaluation: HELM framework	Dr. Asgari
23 Aban	Training LMs with instructs	Dr. Rohban
28 Aban	Training LMs with human feedback	Dr. Rohban
30 Aban	Retrieval-based LLMs	Dr. Asgari
5 Azar	Reasoning: Chain of thought prompting	Dr. Soleymani
7 Azar	Knowledge in LLMs	Dr. Soleymani
12 Azar	Image-text FMs	Dr. Soleymani
14 Azar	Generalized VLMs	Dr. Soleymani
19 Azar	Other foundation models	Dr. Soleymani

LLMs course: topics I

- **Language models: Architectures and training (3)**
- Adapting LLMs to new tasks or domains (6)
- Data & Evaluation (3)
- Alignment and empowerment of LLMs (5)
- Image-text and multimodal LMs (3)

	Topic	Instructor
4 Mehr	Introduction to LLMs	Dr. Soleymani
9 Mehr	Transformers: Encoder model	Dr. Asgari
11 Mehr		
16 Mehr	Transformers: Decoder model	Dr. Asgari
18 Mehr	Transformers: Encoder-decoder model	Dr. Asgari
23 Mehr	Prompting for zero-shot and few-shot learning	Dr. Rohban
25 Mehr	Parameter-efficient fine-tuning I	Dr. Rohban
30 Mehr	Parameter-efficient fine-tuning II	Dr. Rohban
2 Aban	In-context learning understanding	Dr. Soleymani
7 Aban	Reliability in prompting LLM: Uncertainty, calibration, factuality	Dr. Soleymani
9 Aban	Multilingual LM	Dr. Asgari
14 Aban	Data preprocessing	Dr. Asgari
16 Aban	Effect of pre-training data on LLMs	Dr. Rohban
21 Aban	Evaluation: HELM framework	Dr. Asgari
23 Aban	Training LMs with instructs	Dr. Rohban
28 Aban	Training LMs with human feedback	Dr. Rohban
30 Aban	Retrieval-based LLMs	Dr. Asgari
5 Azar	Reasoning: Chain of thought prompting	Dr. Soleymani
7 Azar	Knowledge in LLMs	Dr. Soleymani
12 Azar	Image-text FMs	Dr. Soleymani
14 Azar	Generalized VLMs	Dr. Soleymani
19 Azar	Other foundation models	Dr. Soleymani

LLMs course: topics I

- Language models: Architectures and training (3)
- **Adapting LLMs to new tasks or domains (6)**
- Data & Evaluation (3)
- Alignment and empowerment of LLMs (5)
- Image-text and multimodal LMs (3)

	Topic	Instructor
4 Mehr	Introduction to LLMs	Dr. Soleymani
9 Mehr	Transformers: Encoder model	Dr. Asgari
11 Mehr		
16 Mehr	Transformers: Decoder model	Dr. Asgari
18 Mehr	Transformers: Encoder-decoder model	Dr. Asgari
23 Mehr	Prompting for zero-shot and few-shot learning	Dr. Rohban
25 Mehr	Parameter-efficient fine-tuning I	Dr. Rohban
30 Mehr	Parameter-efficient fine-tuning II	Dr. Rohban
2 Aban	In-context learning understanding	Dr. Soleymani
7 Aban	Reliability in prompting LLM: Uncertainty, calibration, factuality	Dr. Soleymani
9 Aban	Multilingual LM	Dr. Asgari
14 Aban	Data preprocessing	Dr. Asgari
16 Aban	Effect of pre-training data on LLMs	Dr. Rohban
21 Aban	Evaluation: HELM framework	Dr. Asgari
23 Aban	Training LMs with instructs	Dr. Rohban
28 Aban	Training LMs with human feedback	Dr. Rohban
30 Aban	Retrieval-based LLMs	Dr. Asgari
5 Azar	Reasoning: Chain of thought prompting	Dr. Soleymani
7 Azar	Knowledge in LLMs	Dr. Soleymani
12 Azar	Image-text FMs	Dr. Soleymani
14 Azar	Generalized VLMs	Dr. Soleymani
19 Azar	Other foundation models	Dr. Soleymani

LLMs course: topics I

- Language models: Architectures and training (3)
- Adapting LLMs to new tasks or domains (6)
- **Data & Evaluation (3)**
- Alignment and empowerment of LLMs (5)
- Image-text and multimodal LMs (3)

	Topic	Instructor
4 Mehr	Introduction to LLMs	Dr. Soleymani
9 Mehr	Transformers: Encoder model	Dr. Asgari
11 Mehr		
16 Mehr	Transformers: Decoder model	Dr. Asgari
18 Mehr	Transformers: Encoder-decoder model	Dr. Asgari
23 Mehr	Prompting for zero-shot and few-shot learning	Dr. Rohban
25 Mehr	Parameter-efficient fine-tuning I	Dr. Rohban
30 Mehr	Parameter-efficient fine-tuning II	Dr. Rohban
2 Aban	In-context learning understanding	Dr. Soleymani
7 Aban	Reliability in prompting LLM: Uncertainty, calibration, factuality	Dr. Soleymani
9 Aban	Multilingual LM	Dr. Asgari
14 Aban	Data preprocessing	Dr. Asgari
16 Aban	Effect of pre-training data on LLMs	Dr. Rohban
21 Aban	Evaluation: HELM framework	Dr. Asgari
23 Aban	Training LMs with instructs	Dr. Rohban
28 Aban	Training LMs with human feedback	Dr. Rohban
30 Aban	Retrieval-based LLMs	Dr. Asgari
5 Azar	Reasoning: Chain of thought prompting	Dr. Soleymani
7 Azar	Knowledge in LLMs	Dr. Soleymani
12 Azar	Image-text FMs	Dr. Soleymani
14 Azar	Generalized VLMs	Dr. Soleymani
19 Azar	Other foundation models	Dr. Soleymani

LLMs course: topics I

- Language models: Architectures and training (3)
- Adapting LLMs to new tasks or domains (6)
- Data & Evaluation (3)
- **Alignment and empowerment of LLMs (5)**
- Image-text and multimodal LMs (3)

	Topic	Instructor
4 Mehr	Introduction to LLMs	Dr. Soleymani
9 Mehr	Transformers: Encoder model	Dr. Asgari
11 Mehr		
16 Mehr	Transformers: Decoder model	Dr. Asgari
18 Mehr	Transformers: Encoder-decoder model	Dr. Asgari
23 Mehr	Prompting for zero-shot and few-shot learning	Dr. Rohban
25 Mehr	Parameter-efficient fine-tuning I	Dr. Rohban
30 Mehr	Parameter-efficient fine-tuning II	Dr. Rohban
2 Aban	In-context learning understanding	Dr. Soleymani
7 Aban	Reliability in prompting LLM: Uncertainty, calibration, factuality	Dr. Soleymani
9 Aban	Multilingual LM	Dr. Asgari
14 Aban	Data preprocessing	Dr. Asgari
16 Aban	Effect of pre-training data on LLMs	Dr. Rohban
21 Aban	Evaluation: HELM framework	Dr. Asgari
23 Aban	Training LMs with instructs	Dr. Rohban
28 Aban	Training LMs with human feedback	Dr. Rohban
30 Aban	Retrieval-based LLMs	Dr. Asgari
5 Azar	Reasoning: Chain of thought prompting	Dr. Soleymani
7 Azar	Knowledge in LLMs	Dr. Soleymani
12 Azar	Image-text FMs	Dr. Soleymani
14 Azar	Generalized VLMs	Dr. Soleymani
19 Azar	Other foundation models	Dr. Soleymani

LLMs course: topics I

- Language models: Architectures and training (3)
- Adapting LLMs to new tasks or domains (6)
- Data & Evaluation (3)
- Alignment and empowerment of LLMs (5)
- **Image-text and multimodal LMs (3)**

	Topic	Instructor
4 Mehr	Introduction to LLMs	Dr. Soleymani
9 Mehr	Transformers: Encoder model	Dr. Asgari
11 Mehr		
16 Mehr	Transformers: Decoder model	Dr. Asgari
18 Mehr	Transformers: Encoder-decoder model	Dr. Asgari
23 Mehr	Prompting for zero-shot and few-shot learning	Dr. Rohban
25 Mehr	Parameter-efficient fine-tuning I	Dr. Rohban
30 Mehr	Parameter-efficient fine-tuning II	Dr. Rohban
2 Aban	In-context learning understanding	Dr. Soleymani
7 Aban	Reliability in prompting LLM: Uncertainty, calibration, factuality	Dr. Soleymani
9 Aban	Multilingual LM	Dr. Asgari
14 Aban	Data preprocessing	Dr. Asgari
16 Aban	Effect of pre-training data on LLMs	Dr. Rohban
21 Aban	Evaluation: HELM framework	Dr. Asgari
23 Aban	Training LMs with instructs	Dr. Rohban
28 Aban	Training LMs with human feedback	Dr. Rohban
30 Aban	Retrieval-based LLMs	Dr. Asgari
5 Azar	Reasoning: Chain of thought prompting	Dr. Soleymani
7 Azar	Knowledge in LLMs	Dr. Soleymani
12 Azar	Image-text FMs	Dr. Soleymani
14 Azar	Generalized VLMs	Dr. Soleymani
19 Azar	Other foundation models	Dr. Soleymani

LLMs course: topics II

- Model zoo (1)
- Applications of LLMs: Code, medical, financial, ... (2)
- LLMs: Memory and computation efficient methods (2)
- Bias, toxicity, and harm (1)
- Security & privacy (1)

21 Azar	Model zoo	
26 Azar		
28 Azar	Applications of LLMs	
3 Dey	Efficient and decentralized training	
5 Dey	Quantization and prunning of LLMs & Efficient inference	
10 Dey	Bias, toxicity, and harm	
12 Dey	Security & privacy	