Lect. Tuğba Pamay Arslan
ITUNLP Research Group
AI & Data Engineering, İstanbul Technical University

ncc@ulakbim.gov.tr

Large Language Models: Key Concepts and Training

Lect. Tuğba Pamay Arslan
ITUNLP Research Group
AI & Data Engineering
İstanbul Technical University

Large Language Model

- A neural model designed to process and generate human-like text based on massive datasets.

- Large ~
  - The model's size (number of parameters) and
  - The scale of the training data size
- Train on <diverse> and <vast> amount of datasets

- Transformer architecture:
  - Handling sequential data and capturing long-range dependencies in text.
  - Self Attention and Positional Encoding

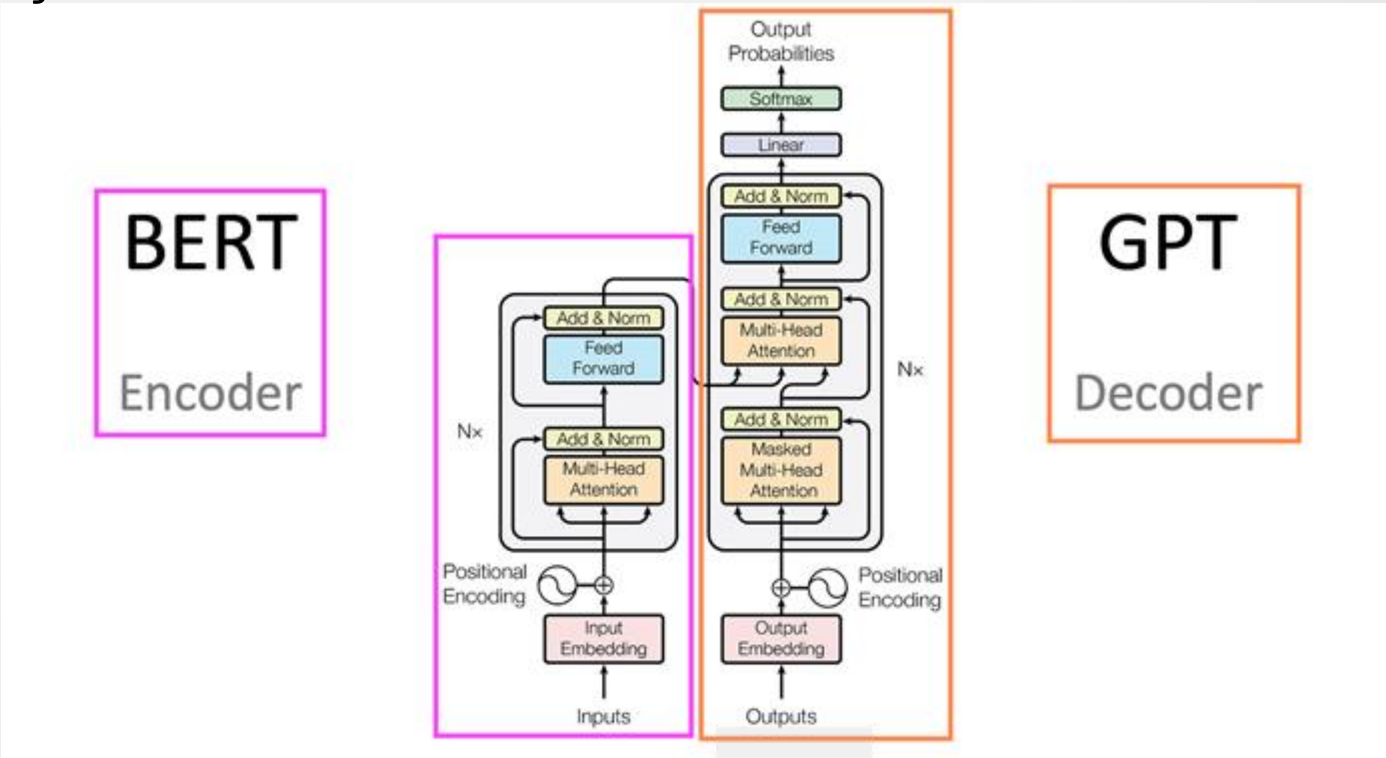
Large Language Model

- Perform both
  - "Understanding" ~ interpreting and analyzing text
  - “Generation” ~producing new text

- Examples
  - GPT, ChatGPT (by OpenAI)
  - BERT, T5 (by Google)
  - Llama (Meta)
  - Mistral (Mistral AI)
  - Gemma (Google)

LLM Structures

- ➢ Encoder-Decoder
- ➢ Encoder-Only
- ➢ Decoder-Only



Encoder-Decoder LLMs

EURO²

- Uses both an encoder and a decoder
  - The encoder processes the input sequence
  - The decoder generates an output sequence based on the encoded information.
- Encoder part: Trained bidirectionally
- Decoder part: Trained unidirectionally (auto-regressive way)
- Tasks where the model needs to transform one sequence into another.
  - Machine Translation
  - Summarization
- T5 (Text-To-Text Transfer Transformer)

Bidirectional or Unidirectional ?

EURO²

- *Bidirectional*
  - The model processes input data from both directions; from start to end and from end to start
  - Helps the encoder better understand the relationships between words within the entire input sequence.

- *Unidirectional (Auto-regressive)*
  - The model processes data in one direction only.
  - During the training, the model only looks at the previous words, not the ones that come after to predict the next word.

Encoder-Only LLMs

- Uses only the encoder component of the transformer
  - Processes the input text and,
  - Outputs a fixed-length vector representation (embedding) of the input sequence.
- Trained bidirectionally

- Tasks where require text understanding or classification without the need to generate new sequences.
  - Sentiment Analysis
  - Classification, tagging tasks

- BERT (Bidirectional Encoder Representations from Transformers), RoBERTa (a robustly optimized BERT approach)

Decoder-Only LLMs

- Uses only the decoder component of the transformer.

- Decoder works in an auto-regressive manner:
  - Generates one token at a time by using each previous token in sequence generation.
- Generative tasks where the model generates new text based on a prompt.

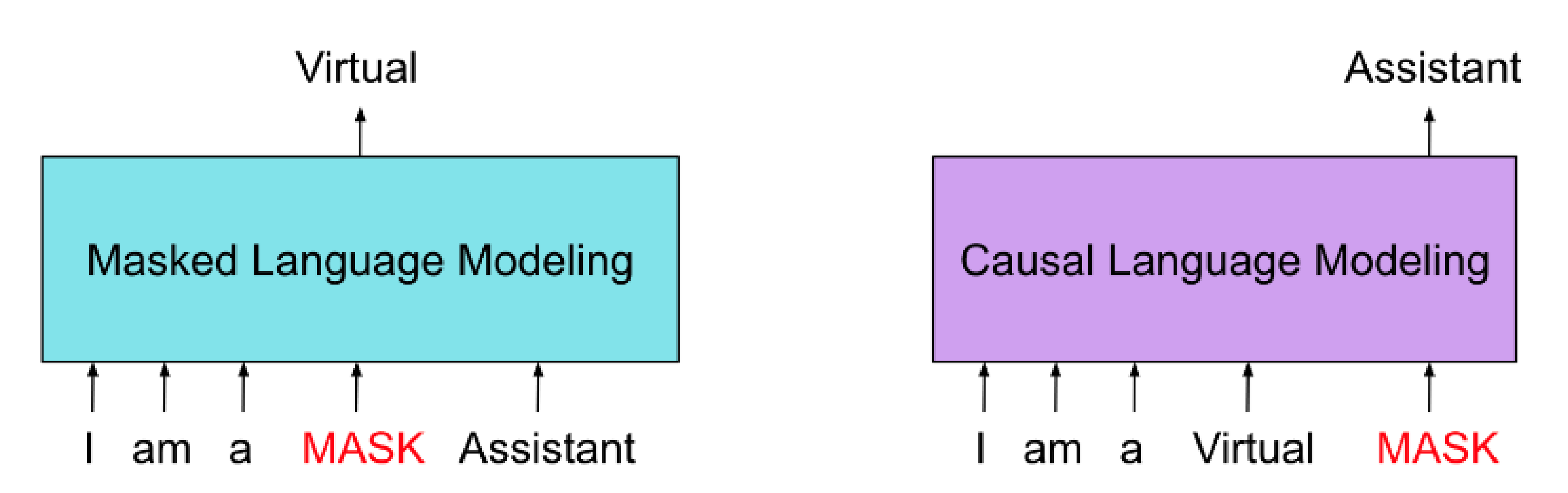
- GPT (Generative Pretrained Transformer)
  LLaMA
  Mistral

Masked Language Modeling

- ➢ Training:
  A portion of the input tokens is randomly replaced with a special token, [MASK]
  The model is tasked with predicting the [MASK] based on the surrounding context.

- ➢ Bidirectional context
  - Trained to understand context from both directions.

- ➢ BERT

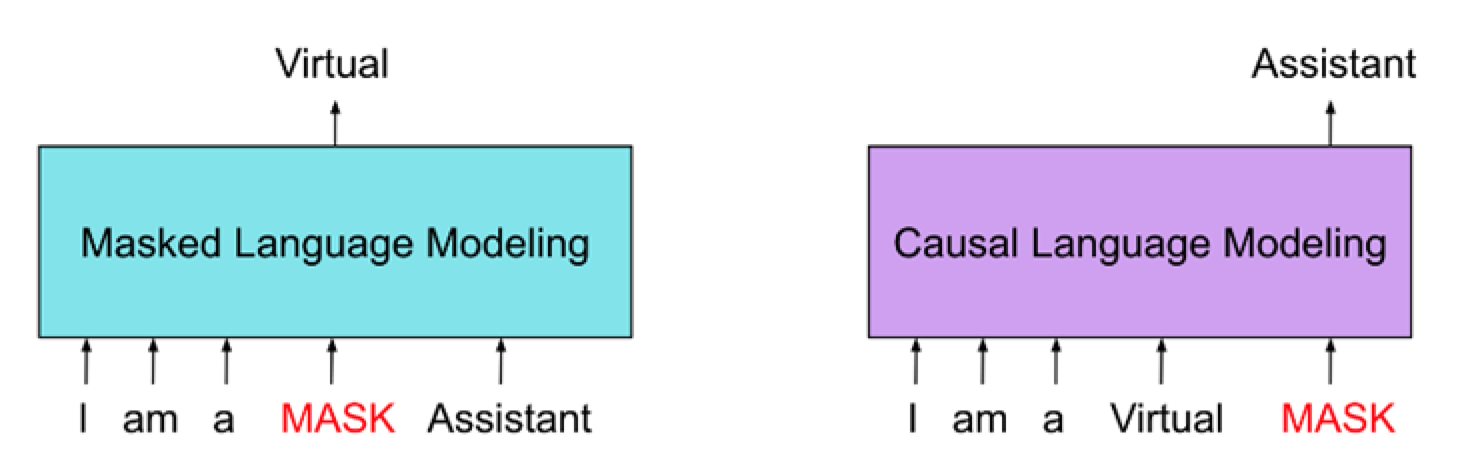


Causal Language Modeling

- Autoregressive Generation:
  The model is trained to predict the next token in a sequence given all previous tokens.
  The model generates one token at a time, with each token prediction conditioned only on the previous tokens.

- Unidirectional context
  - Only consider the past context (tokens before the current token) when predicting the next token

- GPT



Pre-Training

- The initial phase where a model is trained on large, general-purpose datasets
- Unsupervised or Self-supervised
- Learns general language knowledge, but without focusing on any specific task.
- For example:
- in Masked Language Models (MLMs) like BERT,
  - Training objective: To predict the [MASK] tokens.
- in Causal Language Models (LMs) like GPT,
  - Training objective: To predict the next word in a sequence.

Fine-Tuning

- The pre-trained model is adapted to a specific task by training it further on a smaller, labeled dataset.
- Supervised
- Adjusts the parameters of the model to specialize it for a particular application
- Learns general language knowledge, but without focusing on any specific task.
- Allows the model to specialize in a specific task or domain, improving its performance for that task.
  - A pre-trained BERT model can be fine-tuned on a labeled dataset for a Sentiment Analysis
    - Input text (e.g., reviews) is classified into POS or NEG sentiment labels.

Continual Training

- Allows the model to dynamically update its knowledge without requiring retraining from scratch
  >>> Efficiency
- Models can operate effectively in dynamic, real-world environments where data or tasks evolve over time. >>> Adaptability
- Key Challenge

  !!! Avoiding Catastrophic Forgetting !!!

  The tendency of neural networks to overwrite old knowledge when learning new tasks.

Instruction Tuning

EURO[2]

- Fine-tuning pre-trained models to better understand and follow natural language instructions.
- Aligns models to better understand and follow human-like instructions.
- Boosts model performance in zero-shot and few-shot learning scenarios.
- Use instruction-labeled datasets to train the fine-tuned model.
  - Need to convert the current task into instruction-containing I/O for training.

- Example Models:
  - InstructGPT
  - FLAN (Fine-tuned LLaMA models)

Prompting

- ➔ **<u>Prompt:</u>** The input text or query that is fed into the model to generate a response.

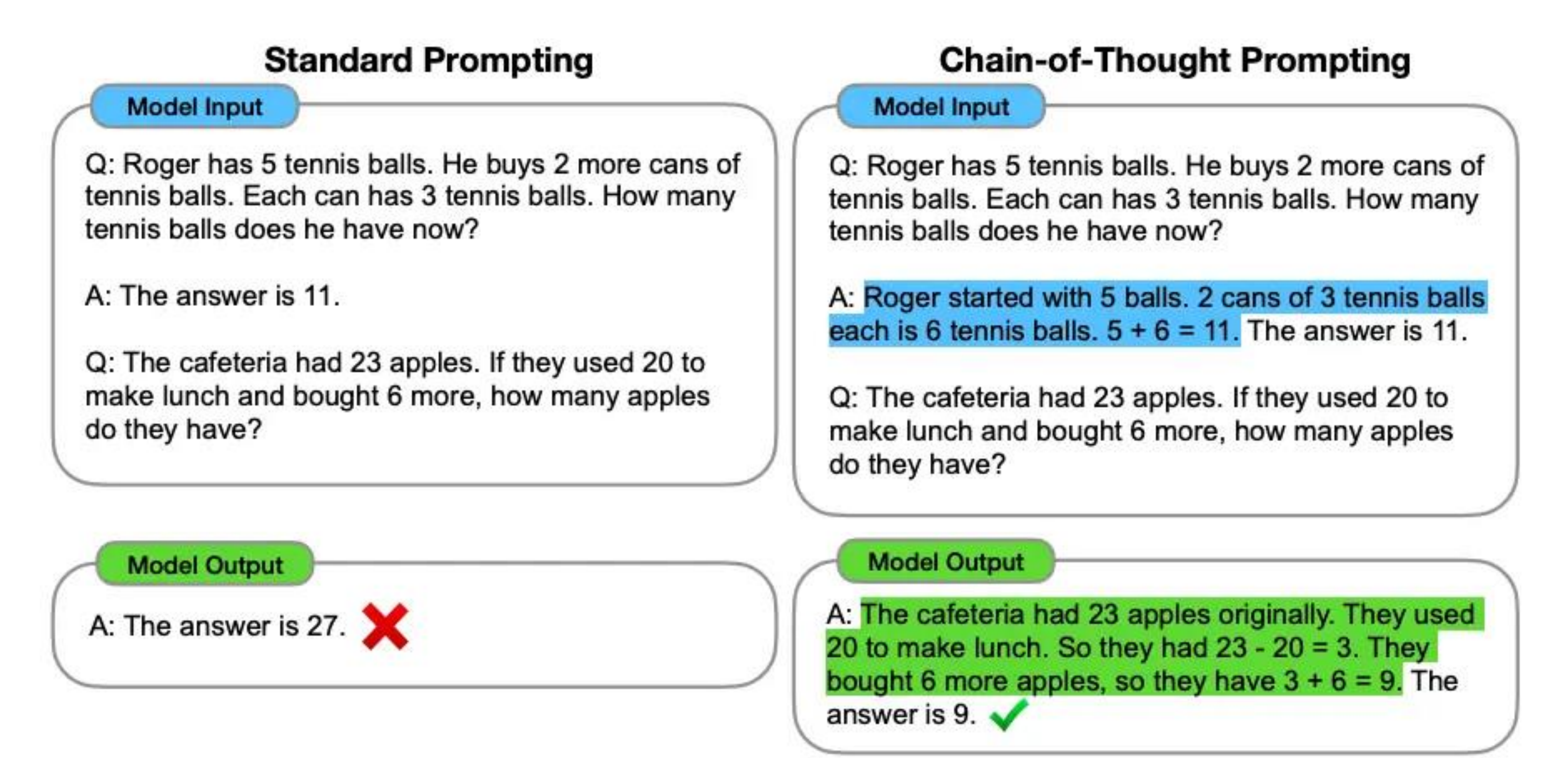
  Ex. A question, an incomplete sentence, or a set of instructions.

- ➔ Types of Prompting:
  - ➔ **Zero-Shot Prompting:**
    - The model is provided with a task or question without any examples or extra context.
    - The model performs the task based on its general knowledge.
  - ➔ **Few-Shot Prompting:**
    - The model is <u>given a few examples</u> of the task before it attempts the task on its own.

Prompting

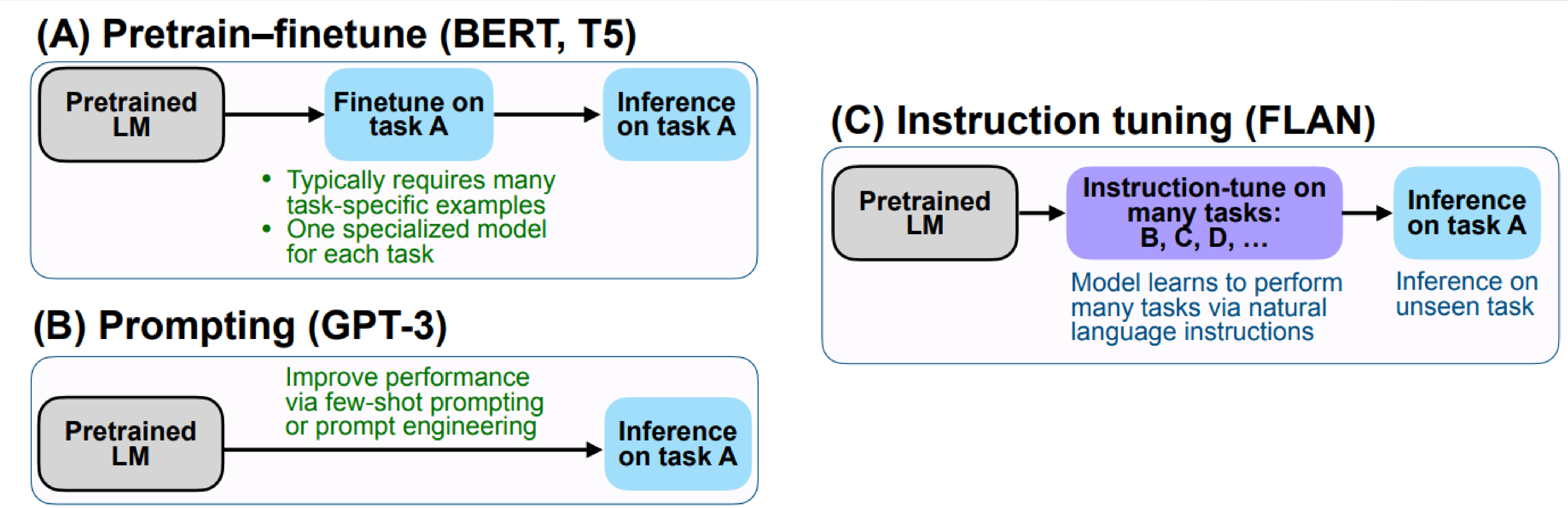
➔ Chain-of-Thoughts Prompting:

- The model is prompted to reason through a problem step-by-step.



Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837.

Fine-Tuning vs Instruction Tuning vs Prompting



Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems, 35*, 24824-24837.

How to Supervised Fine-Tune an LLM ?

- via SFTTrainer
  - designed for Supervised Fine-Tuning (SFT) tasks, particularly when working with instruction-following models.
  - It is tailored for fine-tuning large language models on datasets that consist primarily of text.
- SFTTrainer vs general Trainer
  - Trainer class requires you to manually handle tokenization.
  - SFTTrainer simplifies this by managing these processes automatically.
- Consider using SFTTrainer if:
  - You're working with instruction-following models or datasets that require specific handling of text input.
  - You are implementing PEFT (Parameter Efficient Fine-Tuning) methods like LoRA, which can be easily integrated into SFTTrainer for efficient tuning.