LLMs – In-context learning, CoT

CS 4740 (and crosslists): Introduction to Natural Language Processing
https://courses.cs.cornell.edu/cs4740/2025sp

Slides developed by:
Magd Bayoumi, Claire Cardie, Tanya Goyal, Dan Jurafsky, Lillian Lee, James Martin, Marten van Schijndel

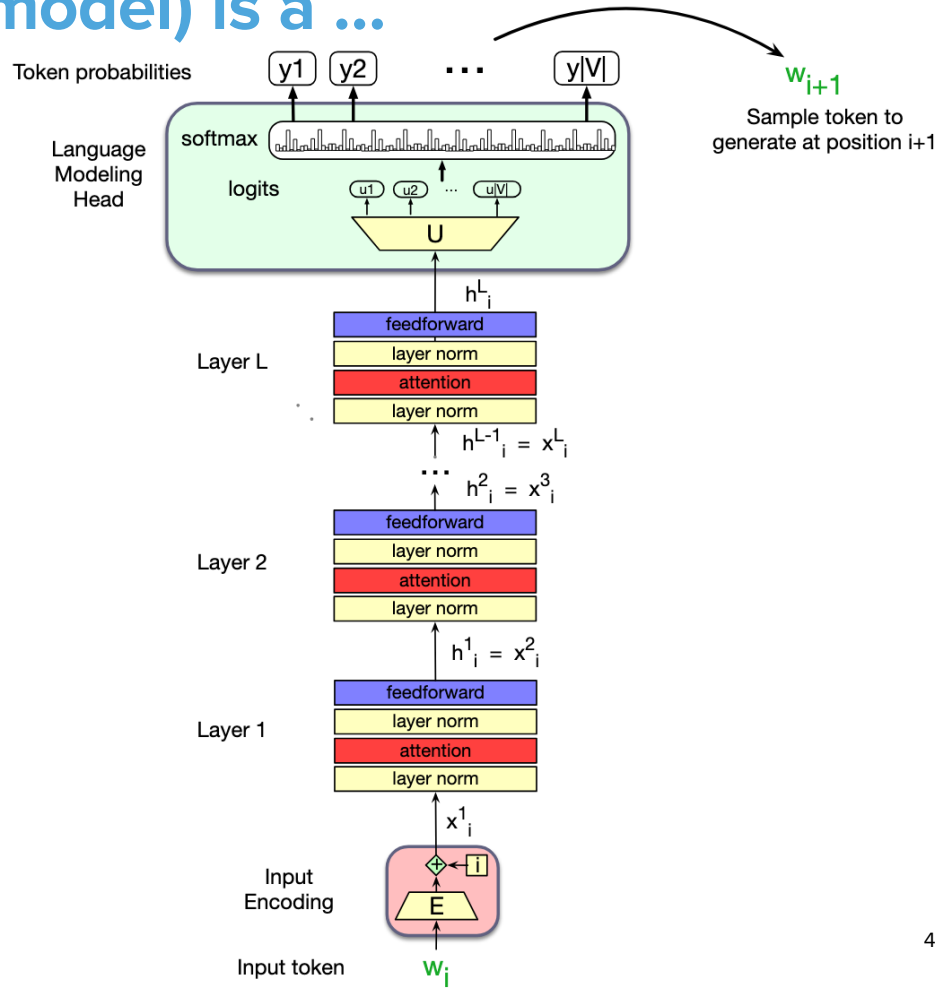
Announcements

- Practice questions
  - Practice session (Apr 20) recording on Ed
  - Solutions on Ed this evening

Recall: Decoder-only Large-scale LMs

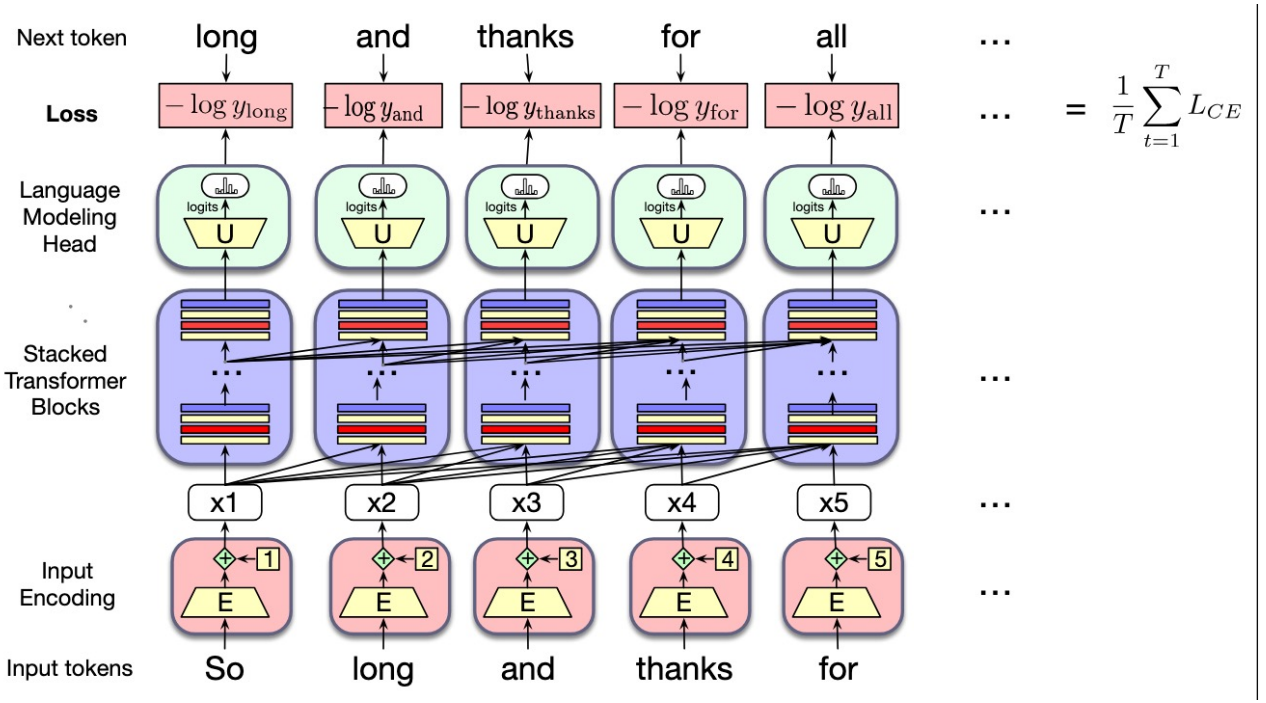
Recall: an LLM (large language model) is a ...

- Decoder-only transformer LM (at input token w_i)



Recall: An LLM is a …

- Decoder-only transformer LM
- **Pretrained** with massive amounts of text
  - To **perform next word prediction**
  - Using **self-supervision** (self-training) and **teacher forcing**
  - Cross-entropy is the loss function

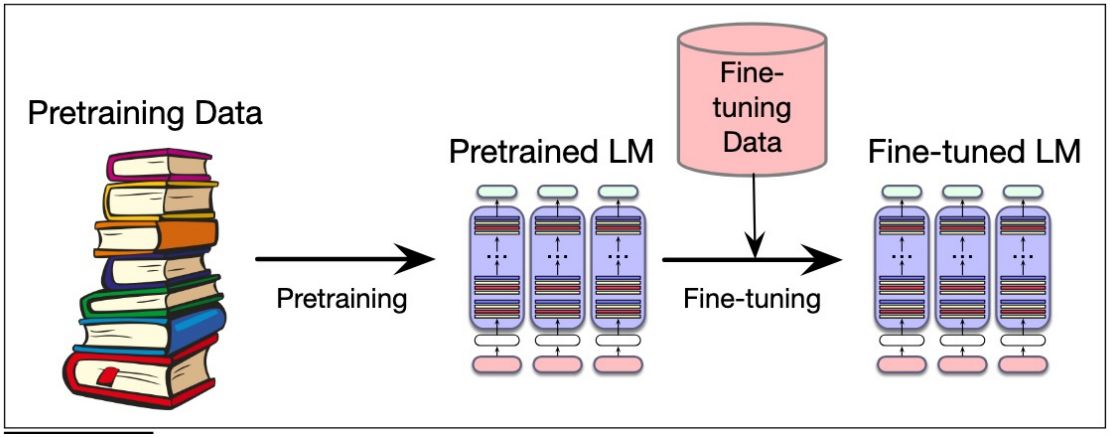


Recall: LLMs for QA

- Finetuned for QA using **supervised finetuning (SFT)**

  This is just one type of finetuning

  Falls under the umbrella of **instruction finetuning** (i.e., want a pretrained language model to learn to follow text instructions)



Today

- Prompting
  - Learning from demonstrations
  - In-context learning: why does it work?
  - Chain-of-thought prompting
- Model alignment
  - Instruction tuning

Prompting

- A **prompt** is a text string that a user issues to a LM to get the model to do something useful.
  - User's prompt string is passed to the LM
  - LM iteratively generates tokens conditioned on the prompt
- Also called **in-context learning**: the prompt creates a context that guides the LLM to generate useful outputs
- **Prompt engineering**: process of finding effective prompts for a task

Prompting for sentiment classification

- Prompt = review + incomplete statement

Hotel Review Completions

Did not like the service that I was provided, when I entered the hotel. I also did not like the area, in which the hotel was located. Too much noise and events going on for me to feel relaxed. In short our stay was

Prompting for sentiment classification

- Prompt = review + incomplete statement
- Then map the response to one of {+, -, neutral} or {1, 2, 3, 4, 5} or ...

Hotel Review Completions

Did not like the service that I was provided, when I entered the hotel. I also did not like the area, in which the hotel was located. Too much noise and events going on for me to feel relaxed. In short our stay was

---

```
...  not a pleasant one.  The staff at the front desk
were not welcoming or friendly, and seemed disinterested
in providing good customer service.

...  uncomfortable and not worth the price we paid.  We
will not be returning to this hotel.
```

Prompting

- Can be used to produce responses for a wide variety of tasks
- Design task-specific templates
- Instruction comes at the end of the prompt; constrains the generation

Basic Prompt Templates	
Summarization	{input} ; `tldr;`
Translation	{input} ; `translate to French:`
Sentiment	{input}; `Overall, it was`
Fine-Grained-Sentiment	{input}; `What aspects were important in this review?`

LLM Outputs for Basic Prompts

Original Review ($INPUT)	`Did not like the service that I was provided, when I entered the hotel. I also did not like the area, in which the hotel was located. Too much noise and events going on for me to feel relax and away from the city life.`
Sentiment	**Prompt**: `$INPUT + In short, our stay was` **Output**: `not enjoyable`
Fine-grained Sentiment	**Prompt**: `$INPUT + These aspects were important to the reviewer:` **Output**: `1. Poor service 2. Unpleasant location 3. Noisy and busy area`
Summarization	**Prompt**: `$INPUT + tl;dr` **Output**: `I had a bad experience with the hotel's service and the location was loud and busy.`
Translation	**Prompt**: `$INPUT + Translate this to French` **Output**: `Je n'ai pas aimé le service qui m'a été offert lorsque je suis entré dans l'hôtel. Je n'ai également pas aimé la zone dans laquelle se trouvait l'hôtel. Trop de bruit et d'événements pour que je me sente détendu et loin de la vie citadine.`

Prompting: Using additional constraints

- Specify the possible answers
- Final token, “(“, strongly suggests what should follow
- Specify the role of the language model

A prompt consisting of a review plus an incomplete statement

Human: Do you think that “input” has negative or positive sentiment?
Choices:
(P) Positive
(N) Negative

Assistant: I believe the best answer is: (

At the moment, there is no rigorous science for generating good prompts.

1. For a given task, develop a task-specific **template** that has a free parameter for the *input* text.
2. Given that *input* and the task-specific **template**, the input is used to instantiate **filled prompt** that is then passed to a pretrained language model.
3. Autoregressive decoding is then used to generate a sequence of token outputs.
4. The output of the model can either be used directly as the desired output or a task-appropriate answer can be extracted from the generated.

Today

- Prompting
  - **Learning from demonstrations**
  - In-context learning: why does it work?
  - Chain-of-thought prompting
- Model alignment
  - Instruction tuning

Learning from Demonstrations: Few-shot Prompting

- **Demonstrations**: labeled examples in the prompt template
- Often referred to as **few-shot prompting**

  vs. **zero-shot prompting** (no examples)

- Let's consider few-shot prompting for SQuAd...

Definition: This task is about writing a correct answer for the reading comprehension task. Based on the information provided in a given passage, you should identify the shortest continuous text span from the passage that serves as an answer to the given question. Avoid answers that are incorrect or provides incomplete justification for the question.

Passage: Beyoncé Giselle Knowles-Carter (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyoncé's debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".

Examples:
Q: In what city and state did Beyoncé grow up?
A: Houston, Texas

Q: What areas did Beyoncé compete in when she was growing up?
A: singing and dancing

Q: When did Beyoncé release Dangerously in Love?
A: 2003

Q: When did Beyoncé start becoming popular?
A:

Few-shot Prompting

- How many demonstrations?

  Why?

- How to select demonstrations?

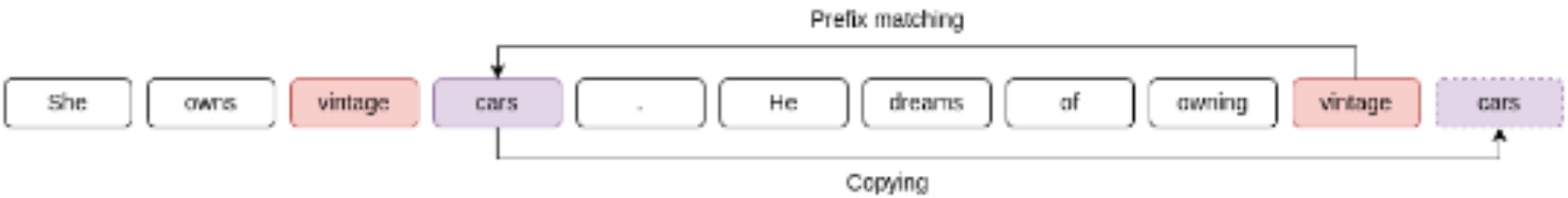
  Usually drawn from a labeled training set

Today

- Prompting
  - Learning from demonstrations
  - **In-context learning: why does it work?**
  - Chain-of-thought prompting
- Model alignment
  - Instruction tuning

Is prompting a “learning” approach?

No gradient-based updates to the model’s parameters…



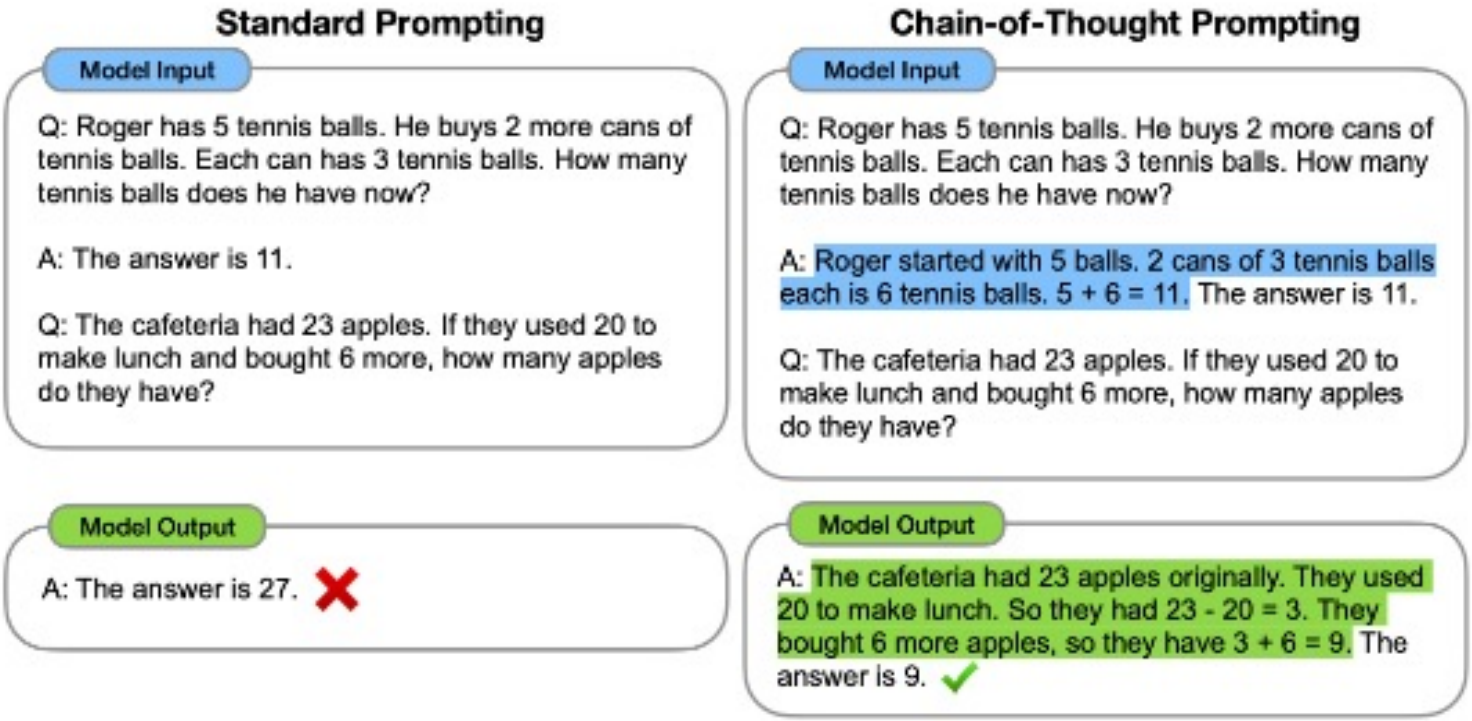
Today

- Prompting
  - Learning from demonstrations
  - In-context learning: why does it work?
  - **Chain-of-thought prompting**
- Model alignment
  - Instruction tuning

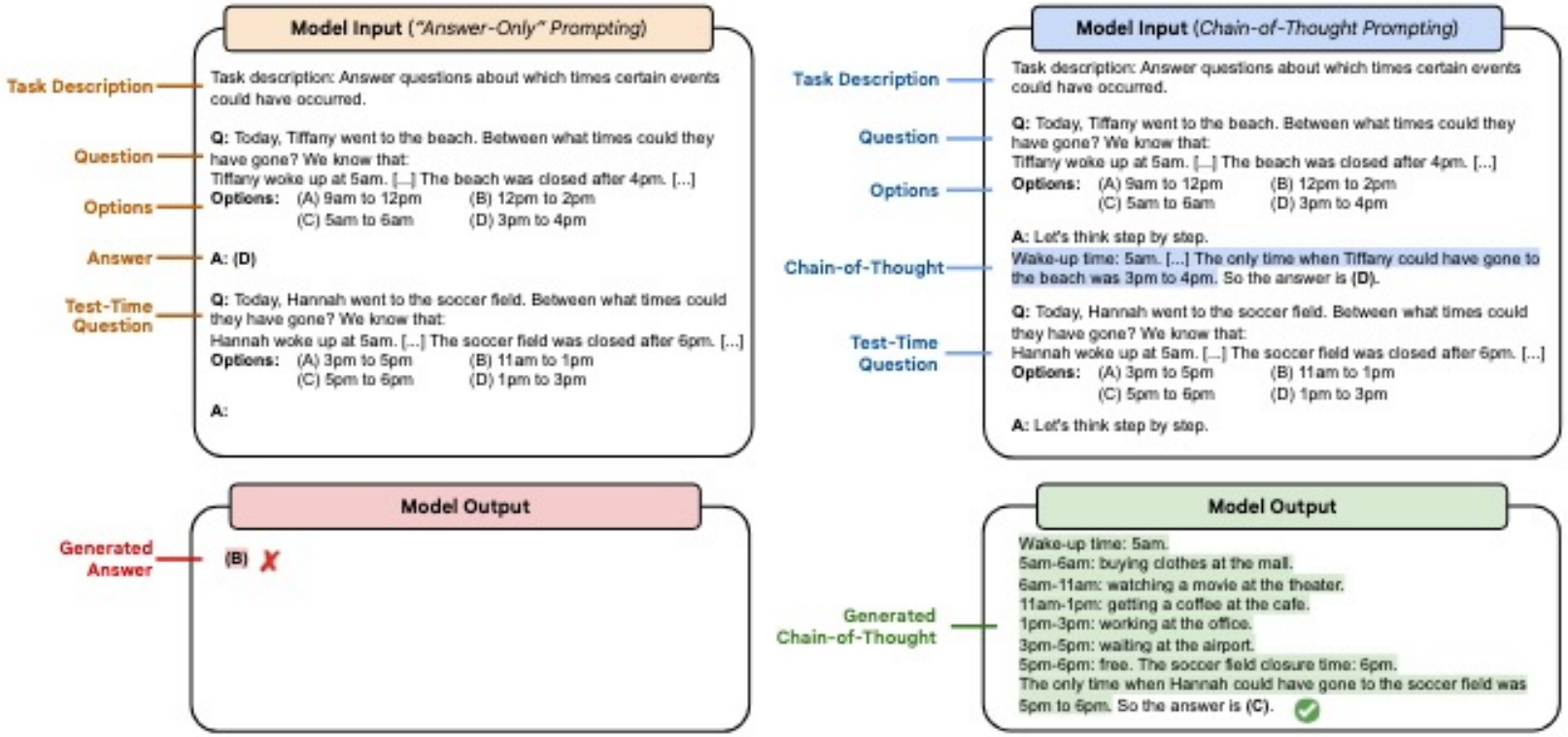
Chain-of-thought (CoT) Prompting

- Improves performance on many tasks, especially tasks that benefit from **reasoning**
- Intuition: people solve these tasks by breaking them down into steps, so let's include language in the prompt to encourage that

Chain-of-thought prompting



Chain-of-thought prompting

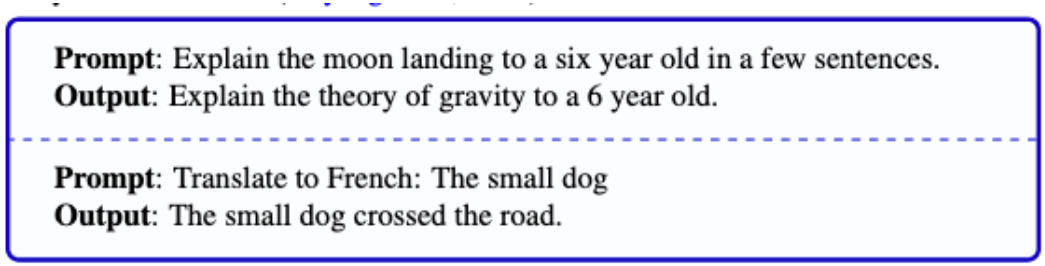


Today

- Prompting
  - Learning from demonstrations
  - In-context learning: why does it work?
  - Chain-of-thought prompting
- **Model alignment**
  - **Instruction tuning**

Limitations of LLMs

- Limited by the next-word-prediction training paradigm
- Early models…



Didn't necessarily follow instructions, **unhelpful**

Can produce **harmful** output [more on this next class]

Instruction (fine-)tuning: Beyond simple prompting

- To enforce more complicated constraints on the system responses
  - E.g., restrict a summary to be a particular length
  - E.g., restrict answer to be from the point of view of a particular persona or role
  - E.g., specify a more structured output (e.g., using a programming language or a data interchange format such as JSON
- Instruction tuning is required
  - Improves LLM's instruction-following capability

Recall: Instruction tuning for QA

- Instruction finetuning for QA using **supervised finetuning (SFT)**

  Create a dataset of example questions and their answers

  Train the language model on this data using the normal cross-entropy loss to predict each token in the instruction prompt (i.e. the QA pair) iteratively

  Q: Where is the Louvre? **A:** Paris, France

- QA then becomes an instance of **conditional** (text) **generation**

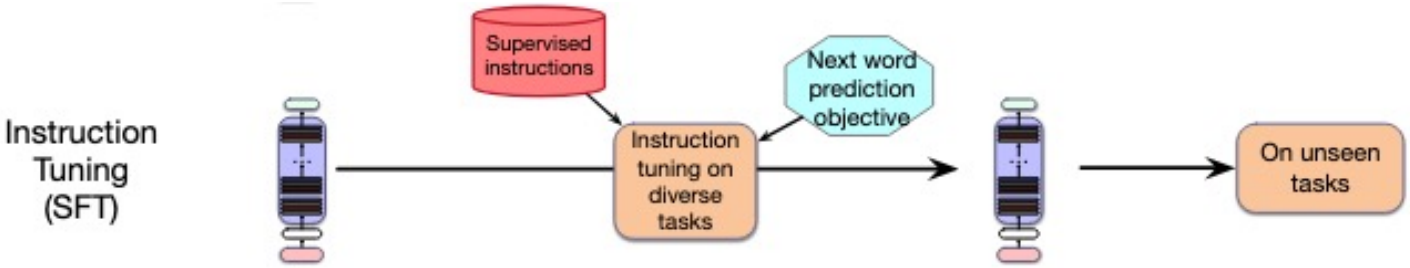
  Given: **Q:** <question> **A:**

  LLM should generate the answer

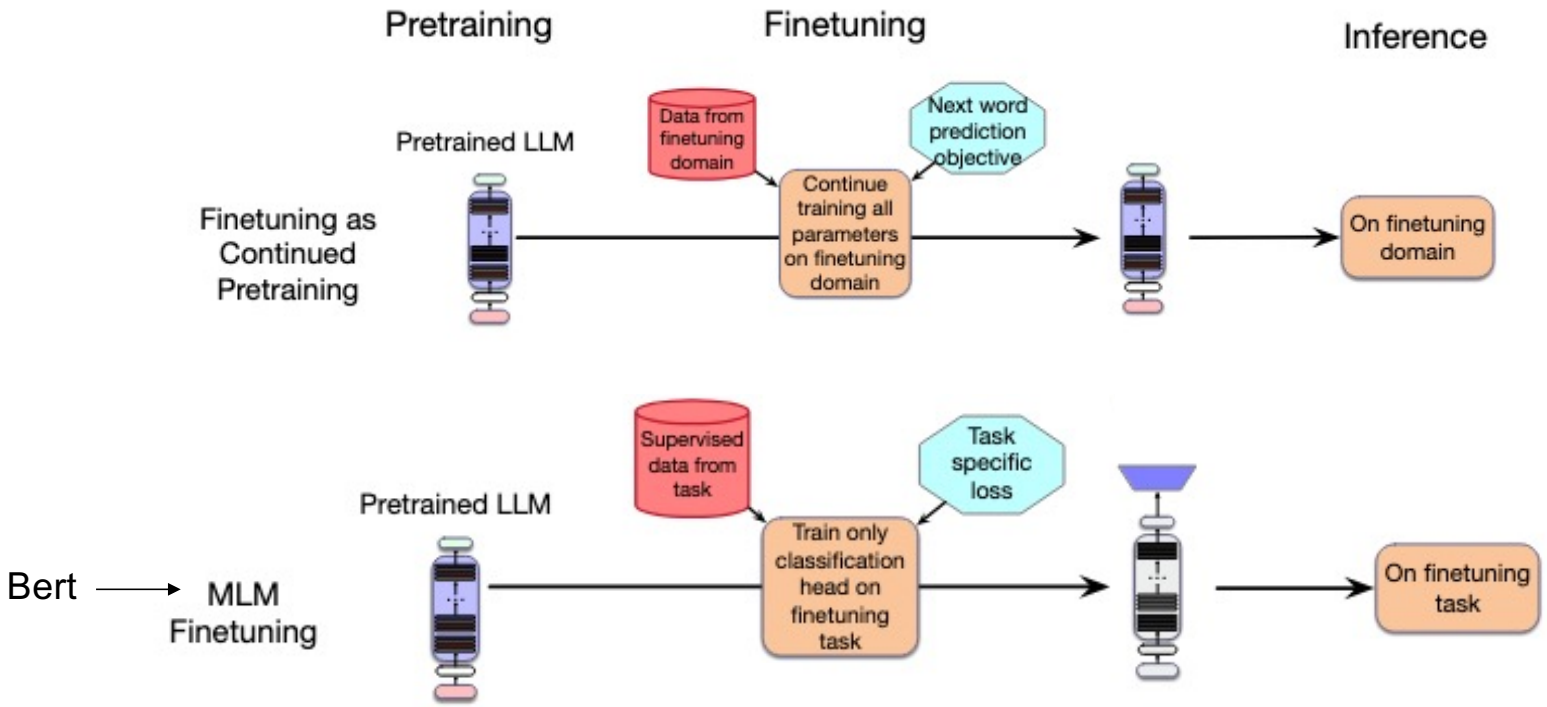
Instruction tuning

- Supervised learning for instruction tuning

  Referred to as **supervised finetuning (SFT)**
- Use demonstration examples as additional training for the original LLM
- Use the same language modeling objective (next-word prediction)



Different kinds of finetuning



Today

- Prompting
  - Learning from demonstrations
  - In-context learning: why does it work?
  - Chain-of-thought prompting
- Model alignment
  - Instruction tuning