CS11-711 Advanced NLP

Fine-tuning and Instruction Tuning

Graham Neubig

Carnegie Mellon University
Language Technologies Institute

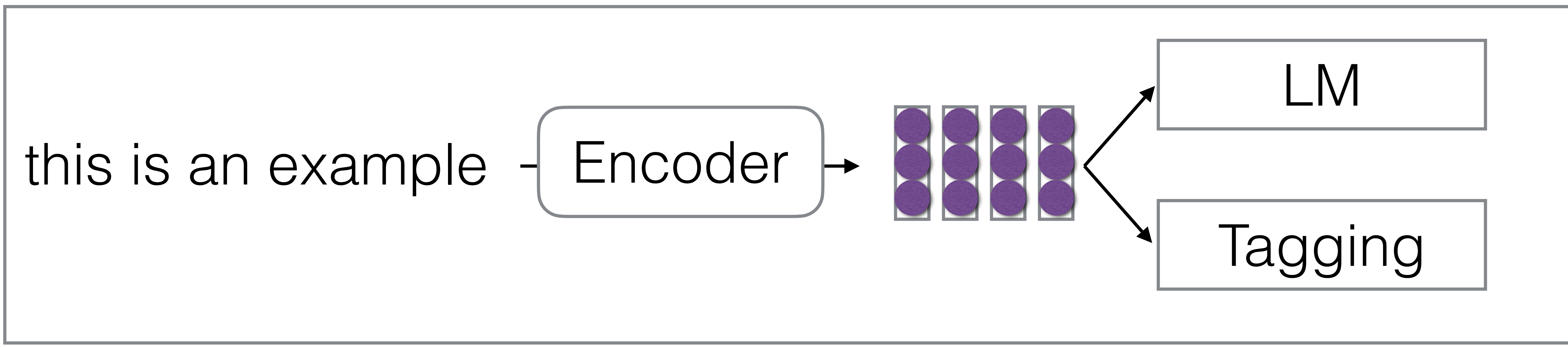
Site
https://phontron.com/class/anlp2024/

Plethora of Tasks in NLP

- In NLP, there are a plethora of tasks, each requiring different varieties of data
  - **Only text:** e.g. language modeling
  - **Naturally occurring data:** e.g. machine translation
  - **Hand-labeled data:** e.g. most analysis tasks
- And each in many languages, many domains!

Standard Multi-task Learning

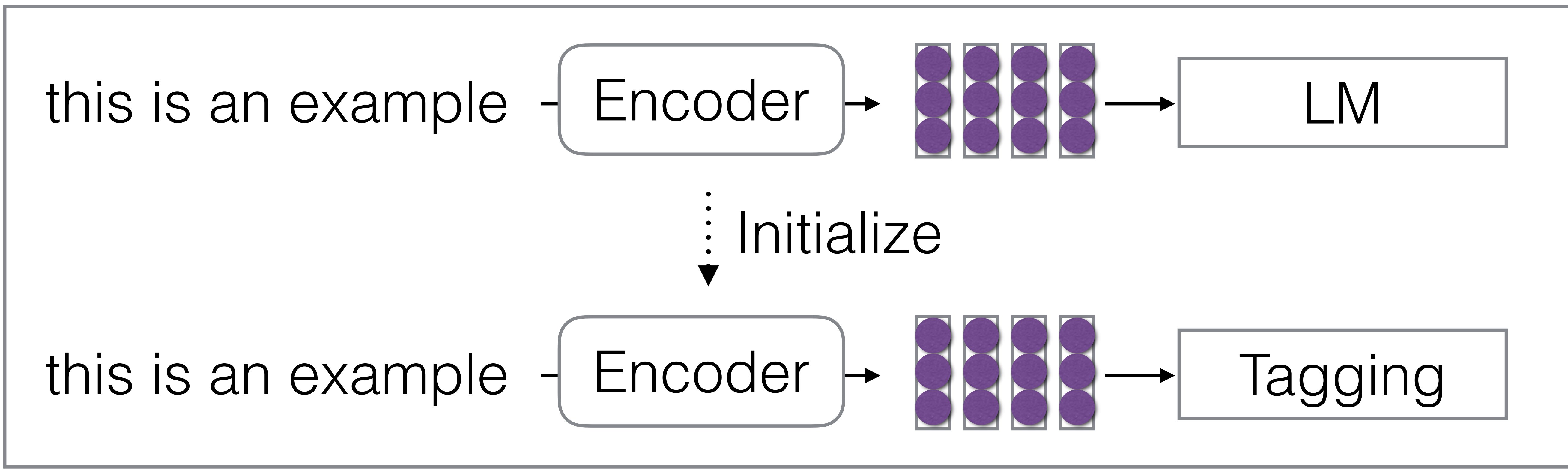
- Train representations to do well on multiple tasks at once



- Often as simple as randomly choosing minibatch from one of multiple tasks

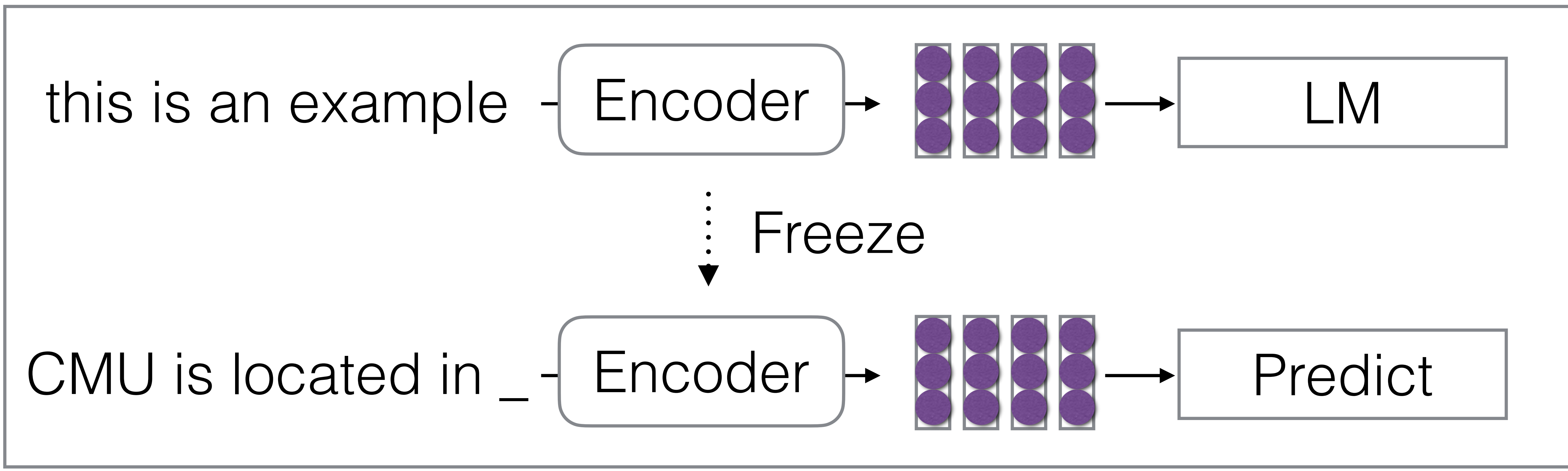
Pre-train and Fine-Tune

- First train on one task, then train on another



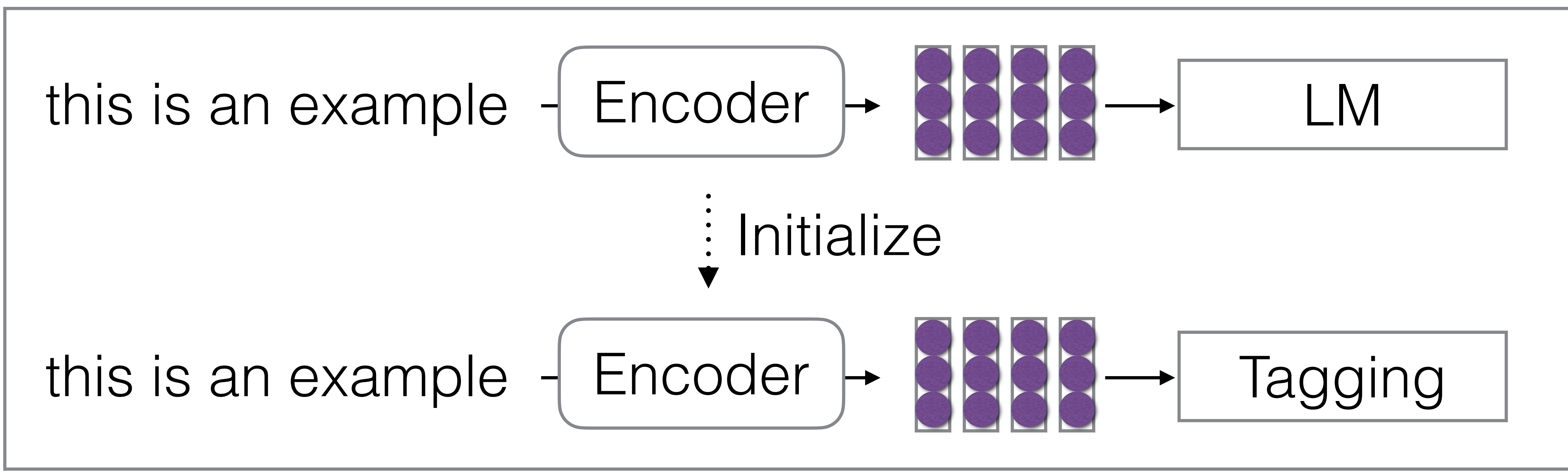
Prompting

- Train on LM task, make predictions in textualized tasks



Instruction Tuning

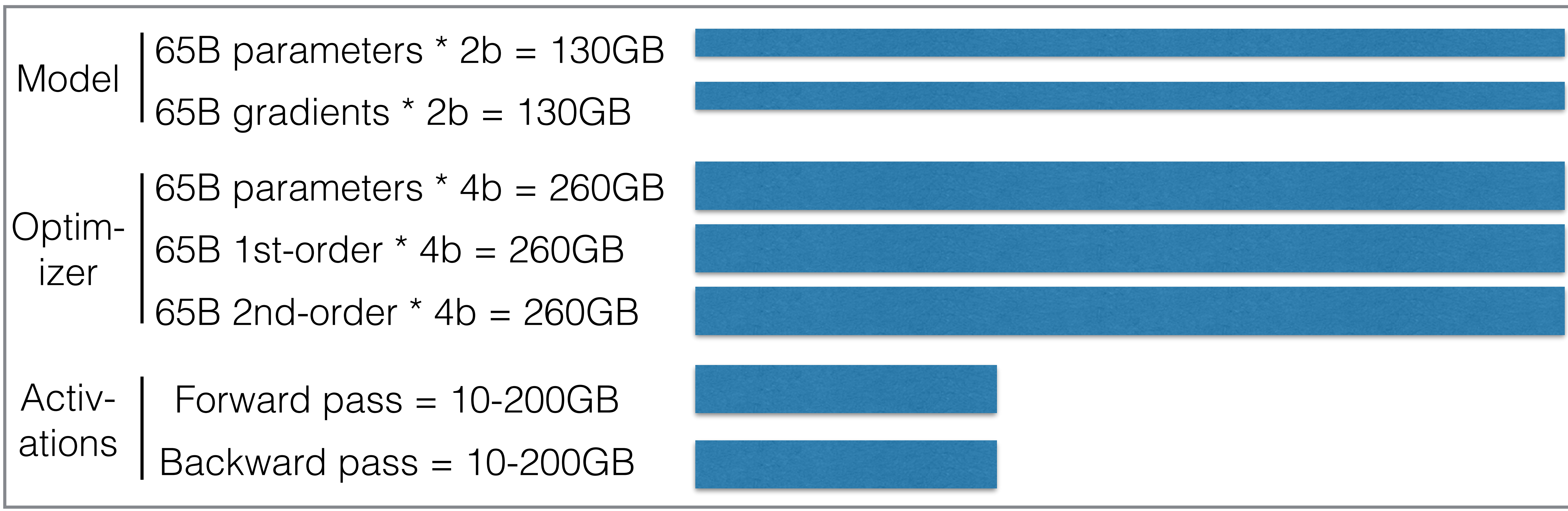
- Pre-train, then fine-tune on many different tasks, with an instruction specifying the task



Fine-tuning

Full Fine-tuning

- Simply continue training the LM on the output
- **Issue:** depending on optimizer, optimization method, can take lots of memory!
- **Example:** Training 65B parameter model with 16-bit mixed precision (Rajbhandari et al. 2019)



1000-1400GB of GPU memory!

(can be reduced by using bfloat16, other optimizations)

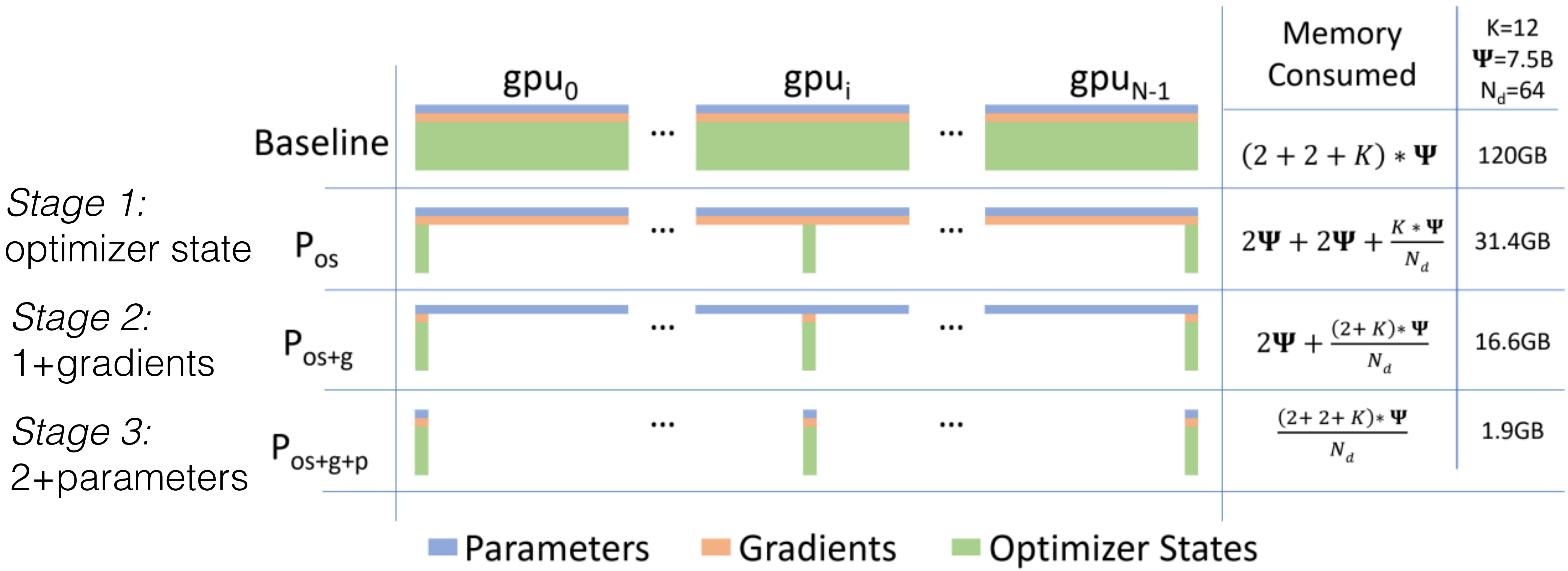
An Aside: GPU Specs

GPU	Memory	Cost (2/2024)	(Cloud) Machines
T40 / K80	24GB	$150	Google Colab, AWS p2.*
V100	32GB	$2,500	Google Colab
A100	40GB or 80GB	$8,000/$16,000	Google Colab, AWS p3.*
H100	80GB	$44,000	AWS p4.*
6000 Ada, L40	48GB	$8000	N/A
Mac M*	Same as CPU	$2000	N/A

- Other hardware options:
  - AMD GPUs
  - Google TPUs
  - Special-purpose Cerebras, AWS Trainium, etc.

Multi-GPU Training

- One solution: throw more hardware at it!
- **Example:** DeepSpeed ZeRo (Rajbhandari et al. 2019) partitions optimization across different devices



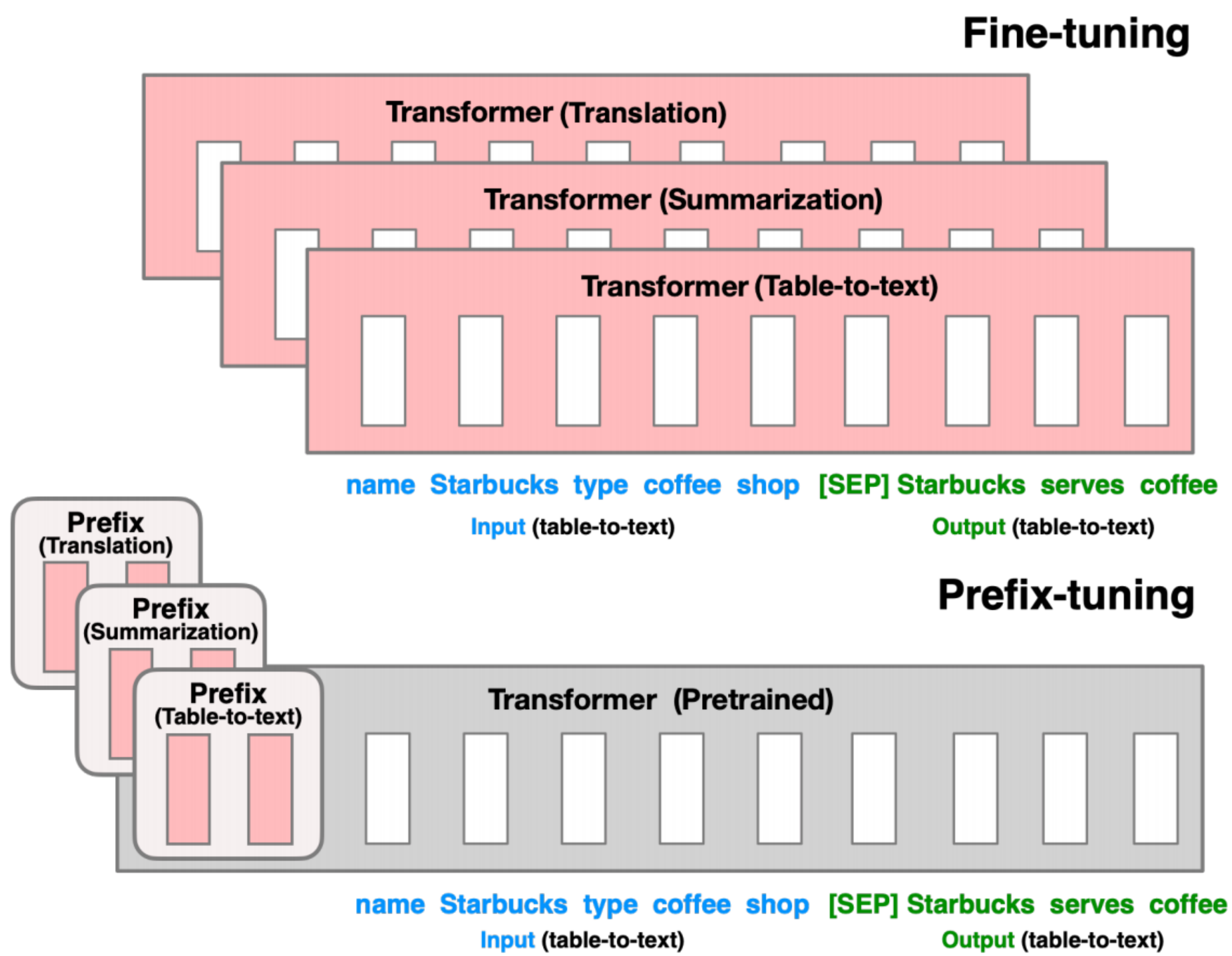
Parameter-efficient Fine-tuning (PEFT)

- Don’t tune all of the parameters, but just some!
  - Prompt/prefix tuning (last class)
  - Adapters
  - BitFit
  - LoRa

Reminder: Prefix Tuning

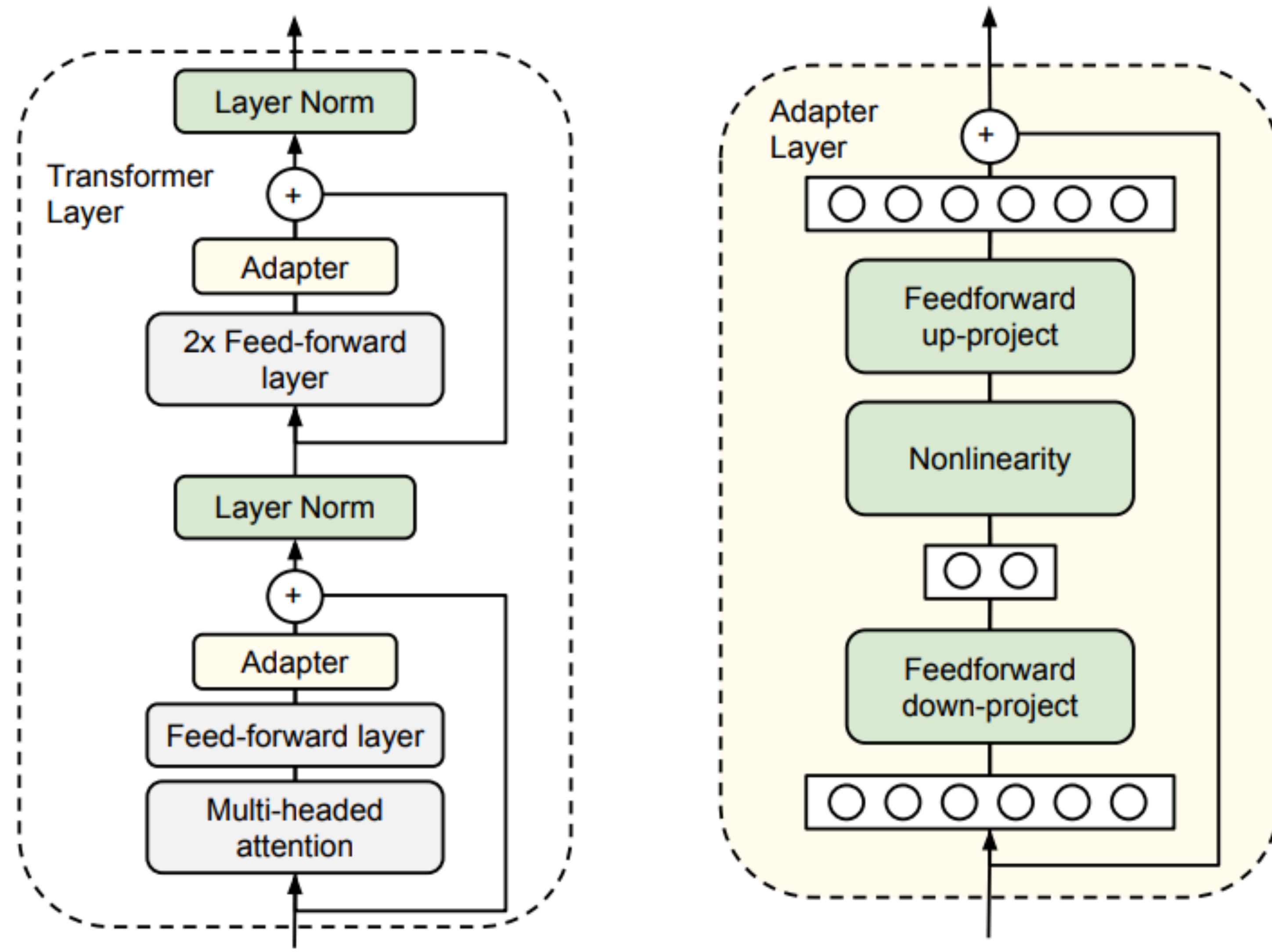
(Li and Liang 2021)

- "Prompt Tuning" optimizes only the embedding layer
- "Prefix Tuning" optimizes the prefix of all layers



Adapters
(Houlsby et al. 2019)

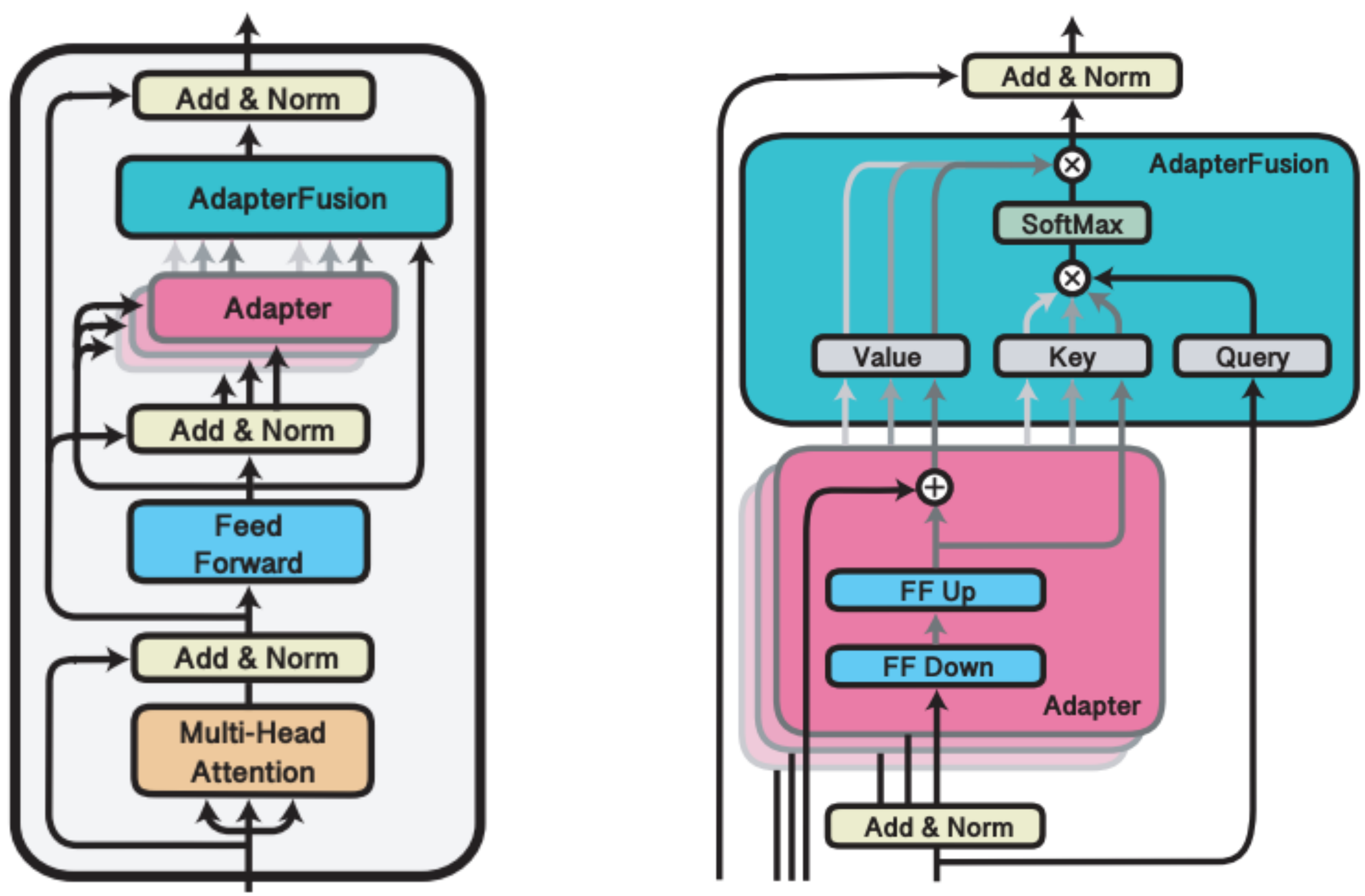
- Sandwich in layers in a pre-trained model, and only tune the adapters
- These layers only use 2*model_dim*adapter_dim parameters



Adapter Fusion

(Pfeiffer et al. 2020)

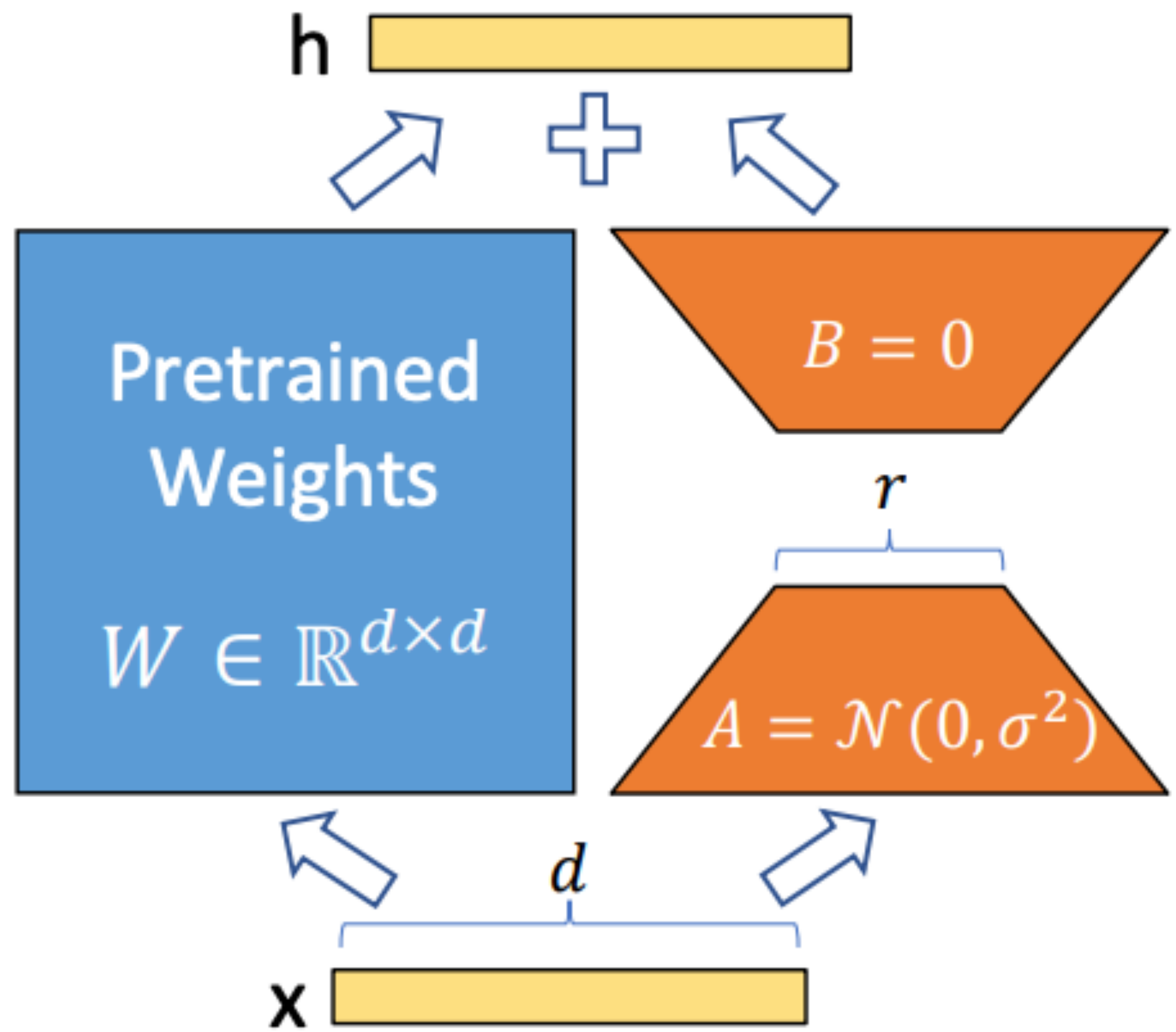
- Learn an adapter for various tasks and combine them together



- Like mixture-of-experts (future class)

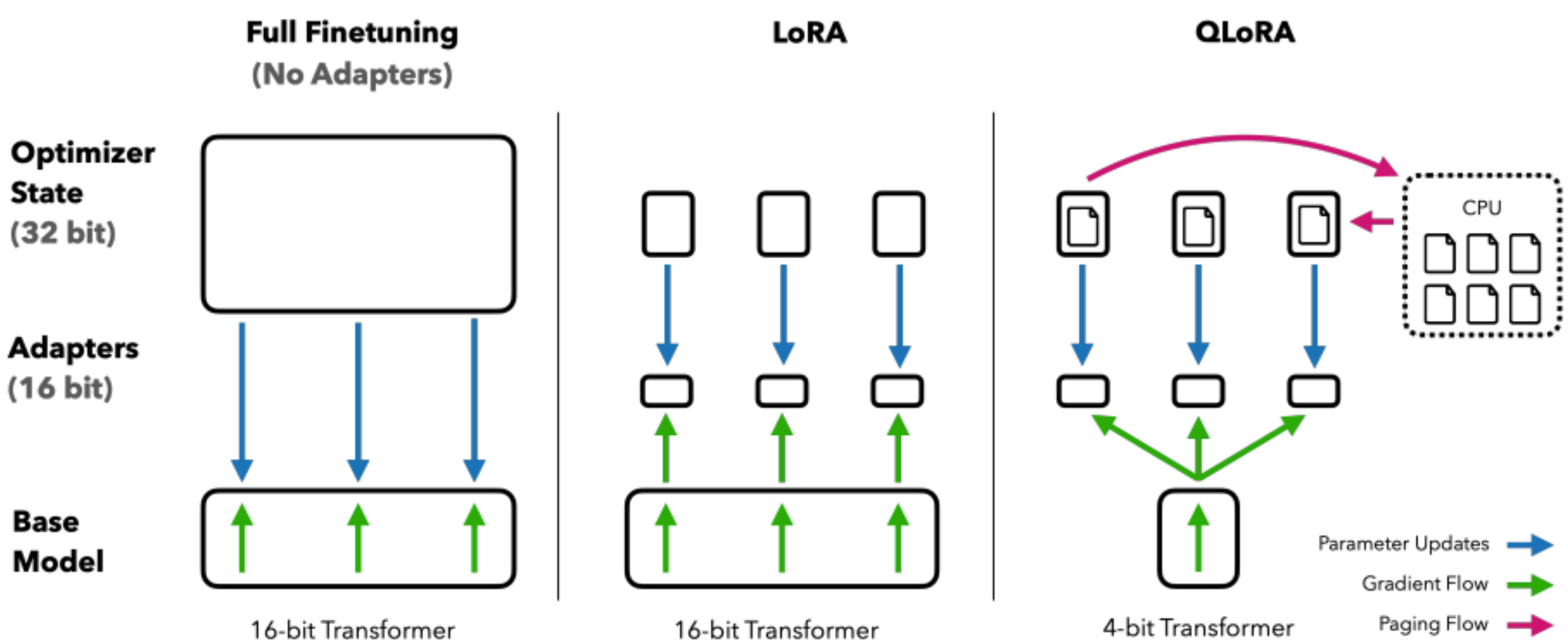
LoRA
(Hu et al. 2021)

- Freeze pre-trained weights, train low-rank approximation of difference from pre-trained weights
- Advantage: after training, just add in to pre-trained weights — no new components!



Q-LORA
(Dettmers et al. 2023)

- Further compress memory requirements for training by
  - 4-bit quantization of the model (later class for details)
  - Use of GPU memory paging to prevent OOM



- Can train a 65B model on a 48GB GPU!

BitFit
(Ben Zaken et al. 2021)

- Tune only the bias terms of the model

$$\mathbf{h}_2^\ell = \text{Dropout}\big(\mathbf{W}_{m_1}^\ell \cdot \mathbf{h}_1^\ell \;+\; \mathbf{b}_{m_1}^\ell\big) \tag{1}$$

$$\mathbf{h}_3^\ell = \mathbf{g}_{LN_1}^\ell \odot \frac{(\mathbf{h}_2^\ell + \mathbf{x}) - \mu}{\sigma} + \mathbf{b}_{LN_1}^\ell \tag{2}$$

$$\mathbf{h}_4^\ell = \;\text{GELU}\big(\mathbf{W}_{m_2}^\ell \cdot \mathbf{h}_3^\ell \;+\; \mathbf{b}_{m_2}^\ell\big) \tag{3}$$

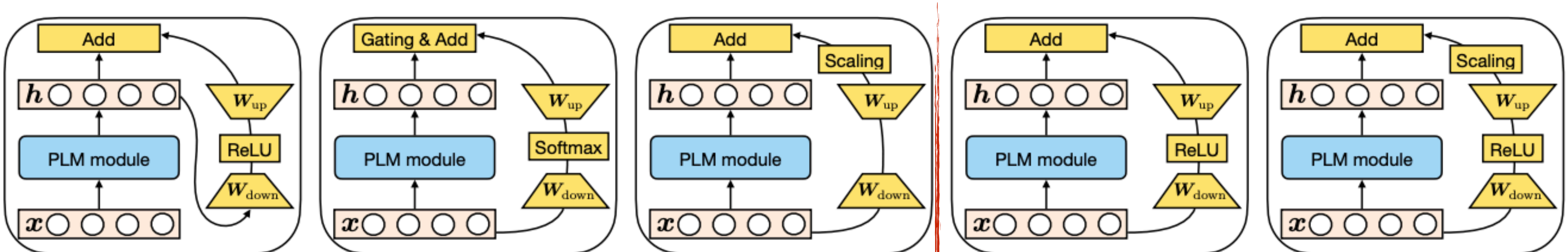
$$\mathbf{h}_5^\ell = \text{Dropout}\big(\mathbf{W}_{m_3}^\ell \cdot \mathbf{h}_4^\ell \;+\; \mathbf{b}_{m_3}^\ell\big) \tag{4}$$

$$\text{out}^\ell = \mathbf{g}_{LN_2}^\ell \odot \frac{(\mathbf{h}_5^\ell + \mathrm{h}_3^\ell) - \mu}{\sigma} + \mathbf{b}_{LN_2}^\ell \tag{5}$$

A Unified View of PEFT

(He et al. 2021)

- If you look closely at the math, most PEFT methods are similar with a few small design differences!



(a) Adapter (b) Prefix Tuning (c) LoRA (d) Parallel Adapter (e) Scaled PA

- This understanding can lead to new variants!

Which one to Choose?
(He et al. 2021)

- **Convenience:** LoRA and BitFit don't change model architecture
- **Accuracy:**
  - *Simpler tasks (e.g. classification):* probably doesn't matter much
  - *More complex tasks + small parameter budget:* prefix tuning seems favorable
  - *More complex tasks + larger budget:* adapters or mix-and-match

NLP Tasks

Approaches to Model Construction

- **Basic Fine Tuning:** Build a model that is good at performing a single task
- **Instruction Tuning:** Build a generalist model that is good at many tasks
- Even if we build a generalist model, we need to have an idea about what tasks we want it to be good at!

Context-free Question Answering

- Also called “open-book QA”
- Answer a question without any specific grounding into documents
- Example dataset: MMLU (Hendrycks et al. 2020)

Professional Law

As Seller, an encyclopedia salesman, approached the grounds on which Hermit's house was situated, he saw a sign that said, "No salesmen. Trespassers will be prosecuted. Proceed at your own risk." Although Seller had not been invited to enter, he ignored the sign and drove up the driveway toward the house. As he rounded a curve, a powerful explosive charge buried in the driveway exploded, and Seller was injured. Can Seller recover damages from Hermit for his injuries?

(A) Yes, unless Hermit, when he planted the charge, intended only to deter, not harm, intruders. ✗

(B) Yes, if Hermit was responsible for the explosive charge under the driveway. ✓

(C) No, because Seller ignored the sign, which warned him against proceeding further. ✗

(D) No, if Hermit reasonably feared that intruders would come and harm him or his family. ✗

Contextual Question Answering

- Also called “machine reading”, “closed-book QA”
- Answer a question about a document or document collection
- *Example:* Natural Questions (Kwiatkowski et al. 2019) is grounded in a Wikipedia document, or the Wikipedia document collection

Question: what color was john wilkes booth’s hair
Wikipedia Page: John_Wilkes_Booth
Long answer: Some critics called Booth “the handsomest man in America” and a “natural genius”, and noted his having an “astonishing memory”; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair , and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a “muscular, perfect man” with “curling hair, like a Corinthian capital”.

Short answer: jet-black

Code Generation

- Generate code (e.g. Python, SQL, etc.) from a natural language command and/or input+output examples
- *Example:* HumanEval (Chen et al. 2021) has evaluation questions for Python standard library

```
def incr_list(l: list):
    """Return list with elements incremented by 1.
    >>> incr_list([1, 2, 3])
    [2, 3, 4]
    >>> incr_list([5, 3, 5, 2, 3, 3, 9, 0, 123])
    [6, 4, 6, 3, 4, 4, 10, 1, 124]
    """
    return [i + 1 for i in l]
```

Summarization

- Single-document: Compress a longer document to shorter
- Multi-document: Compress multiple documents into one
- Example: WikiSum compresses the references in a Wikipedia article into the first paragraph

References

1. ^ "Barack Hussein Obama Takes The Oath Of Office" on YouTube. January 20, 2009.
2. ^ "American Presidents: Greatest and Worst – Siena College Research Institute". Archived from the original on July 15, 2022. Retrieved February 12, 2023.
3. ^ "Barack Obama | C-SPAN Survey on Presidents 2017". Archived from the original on February 12, 2023. Retrieved February 12, 2023.
4. ^ "Siena's 6th Presidential Expert Poll 1982–2018 – Siena College Research Institute". Archived from the original on July 19, 2019. Retrieved February 13, 2023.
5. ^ "President Barack Obama". The White House. 2008. Archived from the original on October 26, 2009. Retrieved December 12, 2008.
6. ^ "President Obama's Long Form Birth Certificate". *whitehouse.gov*. April 27, 2011. Archived from the original on July 31, 2023. Retrieved August 4, 2023.
7. ^ "Certificate of Live Birth: Barack Hussein Obama II, August 4, 1961, 7:24 pm, Honolulu" (PDF). *whitehouse.gov*. April 27, 2011. Archived from the original (PDF) on March 3, 2017. Retrieved March 11, 2017 – via National Archives.

Barack Obama

Article Talk

From Wikipedia, the free encyclopedia

"Barack" and "Obama" redirect here. For other uses, see Barack (disambiguation), Obama (disambiguati

Barack Hussein Obama II (/bəˈrɑːk huːˈseɪn oʊˈbɑːmə/ *bə-RAHK hoo-SAYN oh-BAH-mə*;[1] born August 4, 1961) is an American politician who served as the 44th president of the United States from 2009 to 2017. A member of the Democratic Party, he was the first African-American president in U.S. history. Obama previously served as a U.S. senator representing Illinois from 2005 to 2008, as an Illinois state senator from 1997 to 2004, and as a civil rights lawyer and university lecturer.

Obama was born in Honolulu, Hawaii. He graduated from Columbia University in 1983 with a B.A. in political science and later worked as a community organizer in Chicago. In 1988, Obama enrolled in Harvard Law School, where he was the first black president of the *Harvard Law Review*. He became a civil rights attorney and an academic, teaching constitutional law at the University of Chicago Law School from 1992 to 2004. He also went into elective politics. Obama represented the 13th district in the Illinois Senate from 1997 until 2004, when he successfully ran for the U.S. Senate. In 2008, after a close primary campaign against Hillary Clinton, he was nominated by the Democratic Party for president and chose Delaware Senator Joe Biden as his running mate. Obama was elected president, defeating Republican Party nominee John McCain in the presidential election and was inaugurated on January 20, 2009. Nine months later he was named the 2009 Nobel Peace Prize laureate, a decision that drew a mixture of praise and criticism.

Information Extraction

- *Entity recognition:* identify which words are entities
- *Entity linking:* link entities to a knowledge base (e.g. Wikipedia)
- *Entity co-reference:* find which entities in an input correspond to each-other
- *Event recognition/linking/co-reference:* identify what events occurred
- Example: OntoNotes (Weischedel et al. 2013) annotates many types of information like this on various domains

Translation

- Translate from one language to another
- Quality assessment done using similarity to reference translation
- Example: FLORES dataset (Goyal et al. 2021) — translations of Wikipedia articles into 101 languages

“General Purpose” Benchmarks

- Try to test language abilities across a broad range of tasks
- Example: BIGBench (Srivatsava et al. 2022)

tracking_shuffled_objects_three_objects_0

Alice, Bob, and Claire are friends and avid readers who occasionally trade books. At the start of the semester, they each buy one new book: Alice gets Ulysses, Bob gets Frankenstein, and Claire gets Lolita.
As the semester proceeds, they start trading around the new books. First, Claire and Bob swap books. Then, Bob and Alice swap books. Finally, Claire and Bob swap books. At the end of the semester, Bob has
Options:
(A) Ulysses
(B) Frankenstein
(C) Lolita

label
(B)

date_understanding_0

Today is Christmas Eve of 1937. What is the date tomorrow in MM/DD/YYYY?
Options:
(A) 12/11/1937
(B) 12/25/1937
(C) 01/04/1938
(D) 12/04/1937
(E) 12/25/2006
(F) 07/25/1937

label
(B)

web_of_lies_0

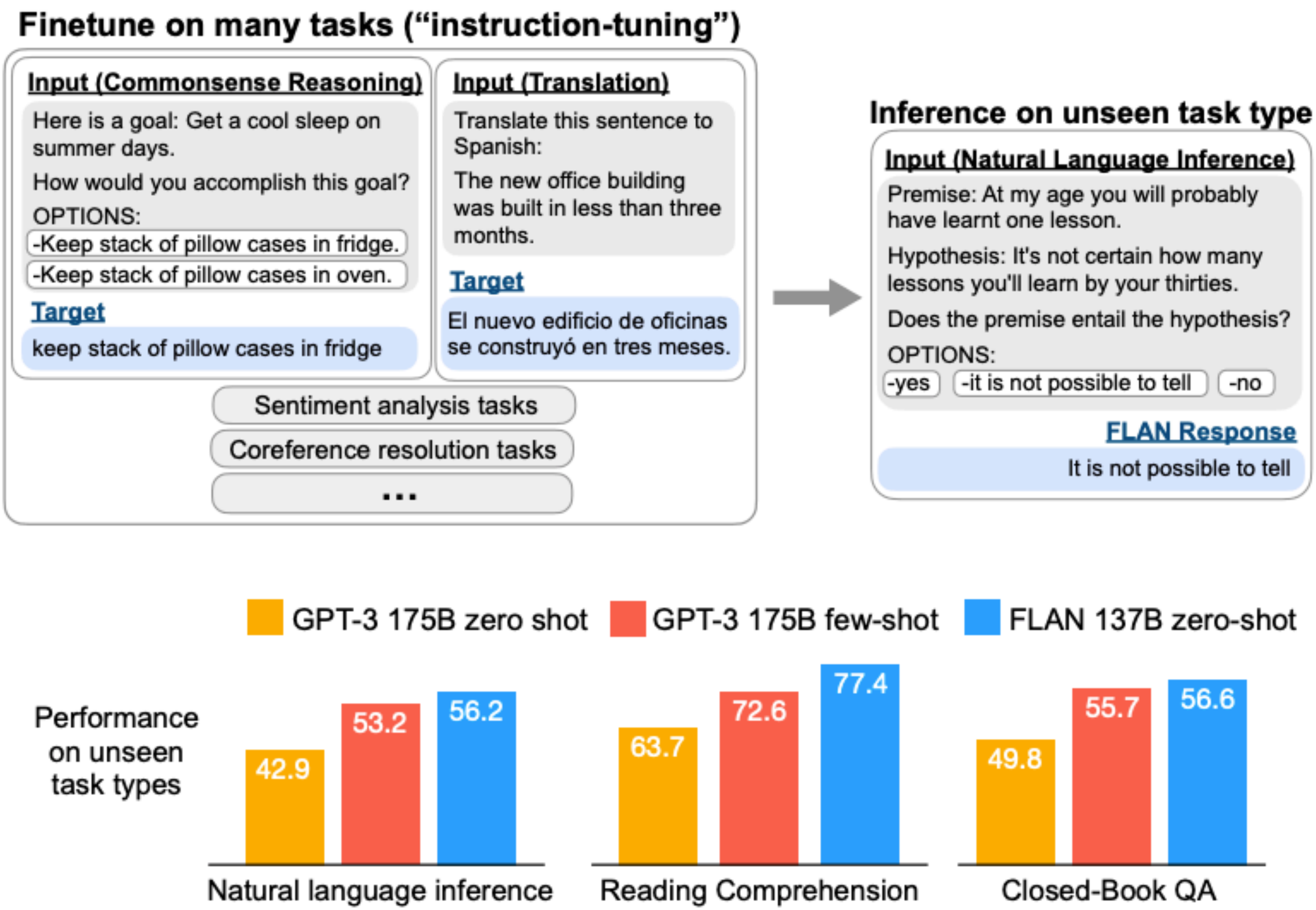
Question: Sherrie tells the truth. Vernell says Sherrie tells the truth. Alexis says Vernell lies. Michaela says Alexis tells the truth. Elanor says Michaela tells the truth. Does Elanor tell the truth?

label
No

Instruction Tuning

Basic Instruction Tuning

(Wei et al. 2021, Sanh et al. 2021)



Learning to In-context Learn

(Min et al. 2021)

- Convert many-shot datasets (typically used in fine-tuning) to few-shot in-context learning examples

	Meta-training	Inference
Task	C *meta-training* tasks	An unseen *target* task
Data given	Training examples $\mathcal{T}_i = \{(x^i_j, y^i_j)\}_{j=1}^{N_i}, \ \forall i \in [1, C] \quad (N_i \gg k)$	Training examples $(x_1, y_1), \cdots, (x_k, y_k)$, Test input x
Objective	For each iteration, 1. Sample task $i \in [1, C]$ 2. Sample $k+1$ examples from $\mathcal{T}_i$: $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ 3. Maximize $P(y_{k+1} \mid x_1, y_1, \cdots, x_k, y_k, x_{k+1})$	$\text{argmax}_{c \in \mathcal{C}} P(c \mid x_1, y_1, \cdots, x_k, y_k, x)$

Instruction Tuning Datasets

- Good reference: FLAN Collection (Longpre et al. 2023)

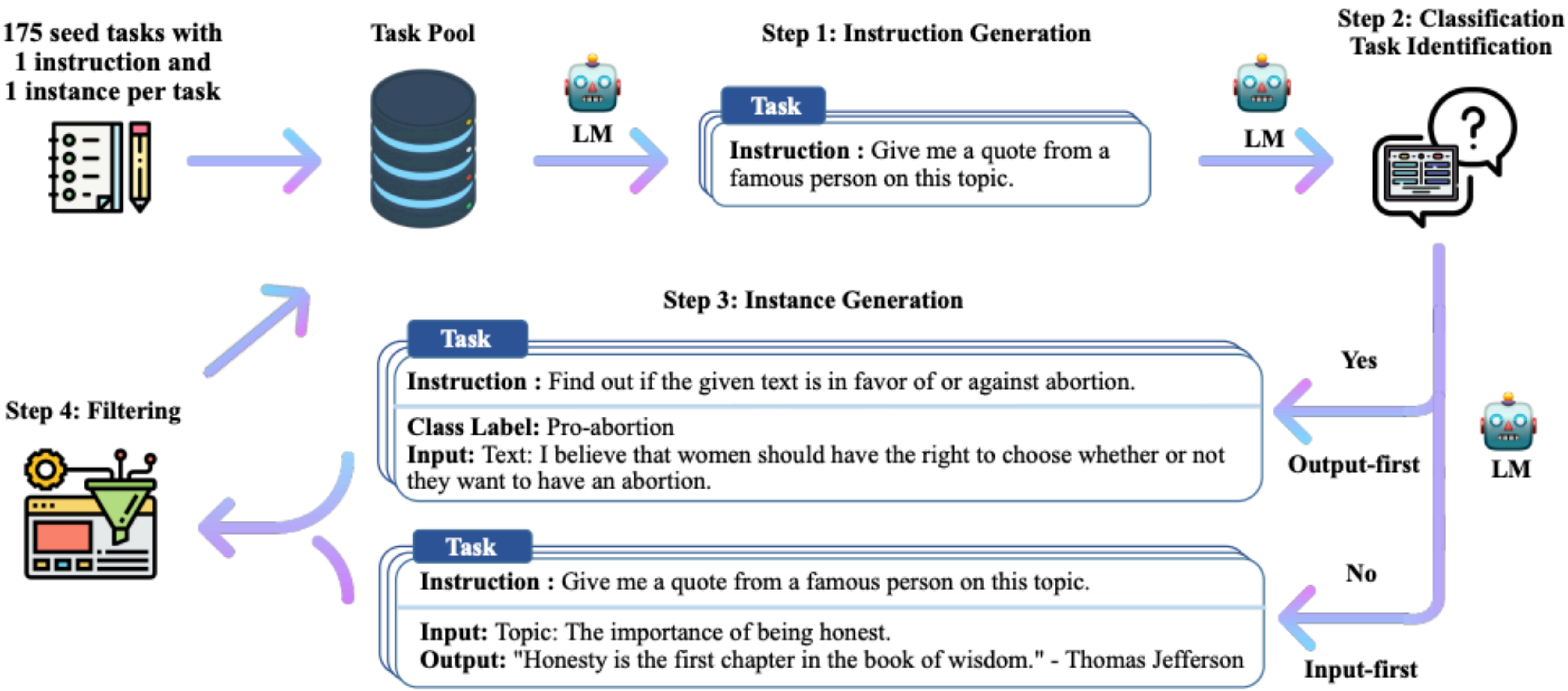
		Model Details				Data Collection & Training Details			
Release	Collection	Model	Base	Size	Public?	Prompt Types	Tasks in Flan	# Exs	Methods
2020 05	UnifiedQA	UnifiedQA	RoBerta	110-340M	P	ZS	46 / 46	750k	
2021 04	CrossFit	BART-CrossFit	BART	140M	NP	FS	115 / 159	71.M	
2021 04	Natural Inst v1.0	Gen. BART	BART	140M	NP	ZS/FS	61 / 61	620k	+ Detailed k-shot Prompts
2021 09	Flan 2021	Flan-LaMDA	LaMDA	137B	NP	ZS/FS	62 / 62	4.4M	+ Template Variety
2021 10	P3	T0, T0+, T0++	T5-LM	3-11B	P	ZS	62 / 62	12M	+ Template Variety + Input Inversion
2021 10	MetaICL	MetaICL	GPT-2	770M	P	FS	100 / 142	3.5M	+ Input Inversion + Noisy Channel Opt
2021 11	ExMix	ExT5	T5	220M-11B	NP	ZS	72 / 107	500k	+ With Pretraining
2022 04	Super-Natural Inst.	*Tk*-Instruct	T5-LM, mT5	11-13B	P	ZS/FS	1556 / 1613	5M	+ Detailed k-shot Prompts + Multilingual
2022 10	GLM	GLM-130B	GLM	130B	P	FS	65 / 77	12M	+ With Pretraining + Bilingual (en, zh-cn)
2022 11	xP3	BLOOMz, mT0	BLOOM, mT5	13-176B	P	ZS	53 / 71	81M	+ Massively Multilingual
2022 12	Unnatural Inst.†	T5-LM-Unnat. Inst.	T5-LM	11B	NP	ZS	~20 / 117	64k	+ Synthetic Data
2022 12	Self-Instruct†	GPT-3 Self Inst.	GPT-3	175B	NP	ZS	Unknown	82k	+ Synthetic Data + Knowledge Distillation
2022 12	OPT-IML Bench†	OPT-IML	OPT	30-175B	P	ZS + FS CoT	~2067 / 2207	18M	+ Template Variety + Input Inversion + Multilingual
2022 10	Flan 2022 (ours)	Flan-T5, Flan-PaLM	T5-LM, PaLM	10M-540B	P NP	ZS + FS CoT	1836	15M	+ Template Variety + Input Inversion + Multilingual

Instruction Tuned Models

- **FLAN-T5:** huggingface/google/flan-t5-xxl
  - Encoder-decoder model based on T5
  - 11B parameters
- **LLaMa-2 Chat:** huggingface/meta-llama/Llama-2-70b-chat-hf
  - Decoder-only model
  - 70B parameters
- **Mixtral instruct:** huggingface/mistralai/Mixtral-8x7B-Instruct-v0.1
  - Decoder-only mixture of experts model
  - 45B parameters
- *(smaller versions also available - Mistral, LLaMa2-7B)*

Dataset Generation

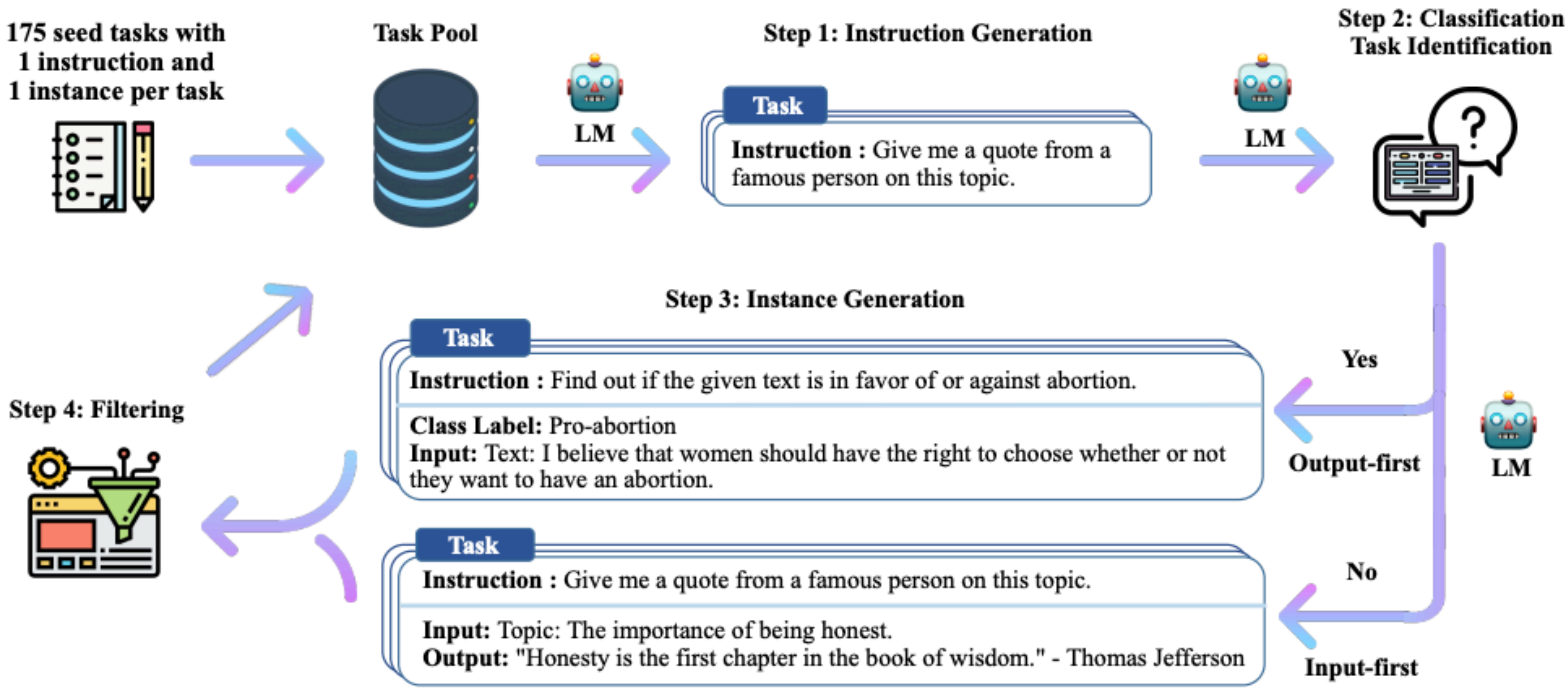
- It is possible to automatically generate instruction tuning datasets, e.g. self-instruct (Wang et al. 2022)



- Can be used to train chain-of-thought — ORCA (Mukherjee et al. 2023)
- Can be used to make instructions more complex — Evol-Instruct (Xu et al. 2023)

Dataset Generation

- It is possible to automatically generate instruction tuning datasets, e.g. self-instruct (Wang et al. 2022)



- Can be used to train chain-of-thought — ORCA (Mukherjee et al. 2023)
- Can be used to make instructions more complex — Evol-Instruct (Xu et al. 2023)

Questions?