



Large Language Models for Knowledge Intensive Problem Solving

Speaker: Victoria Lin

Oxford Machine Learning Summer School
Rep. Learning & Gen. AI

Outline

- 1. LLMs as General Parametric Task Solvers**
 - 1. LLM Pre-training and Scaling Laws
 - 2. Instruction Tuning
- 2. External Knowledge Access**
 - I) Retrieval Augmentation
 - II) Tool Augmentation
- 3. Quick Discussion: LLM Agents and Advanced Applications**
- 4. Conclusion**

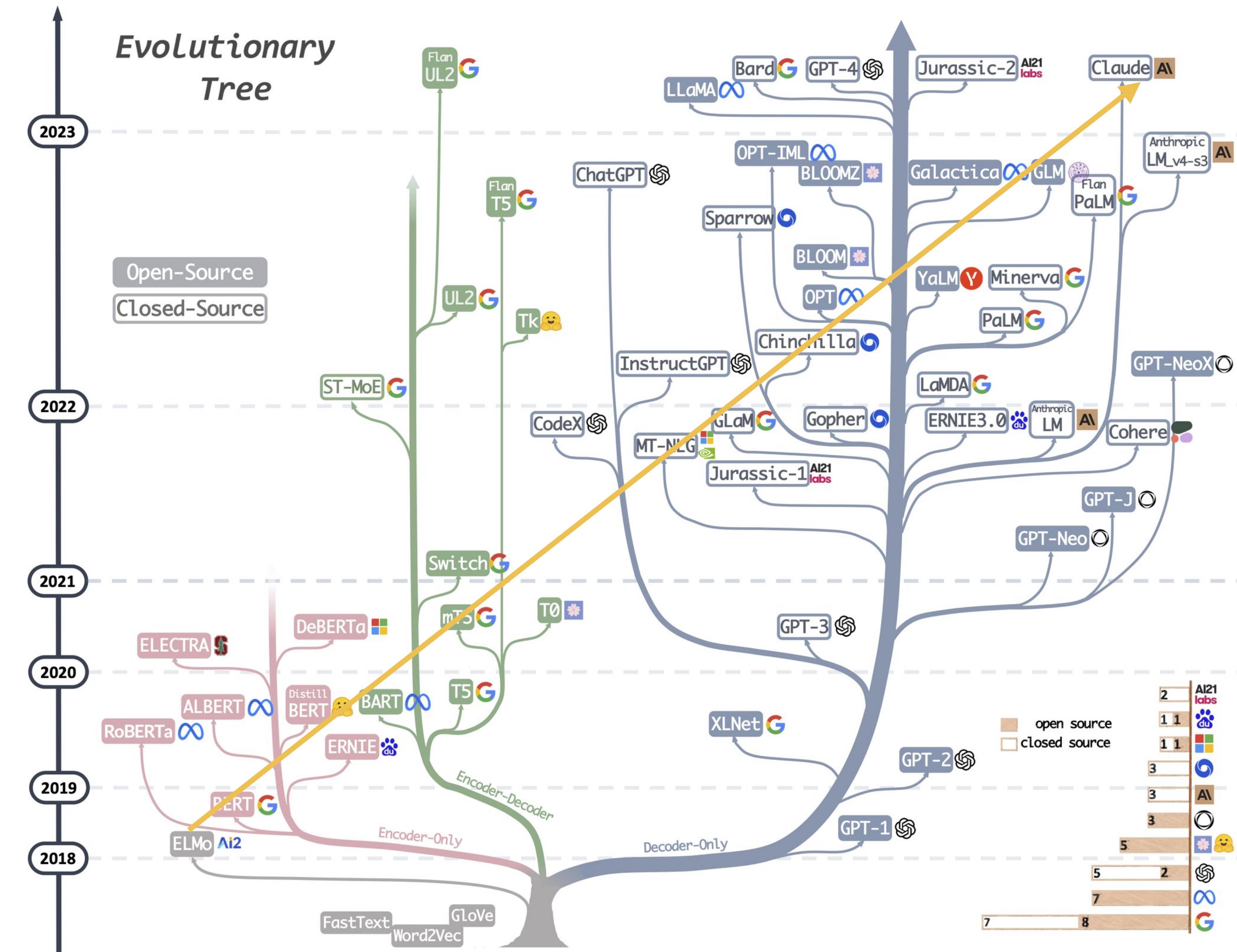


Image Credit: <https://github.com/Mooler0410/LLMsPracticalGuide>
(Yang et al. 2023)

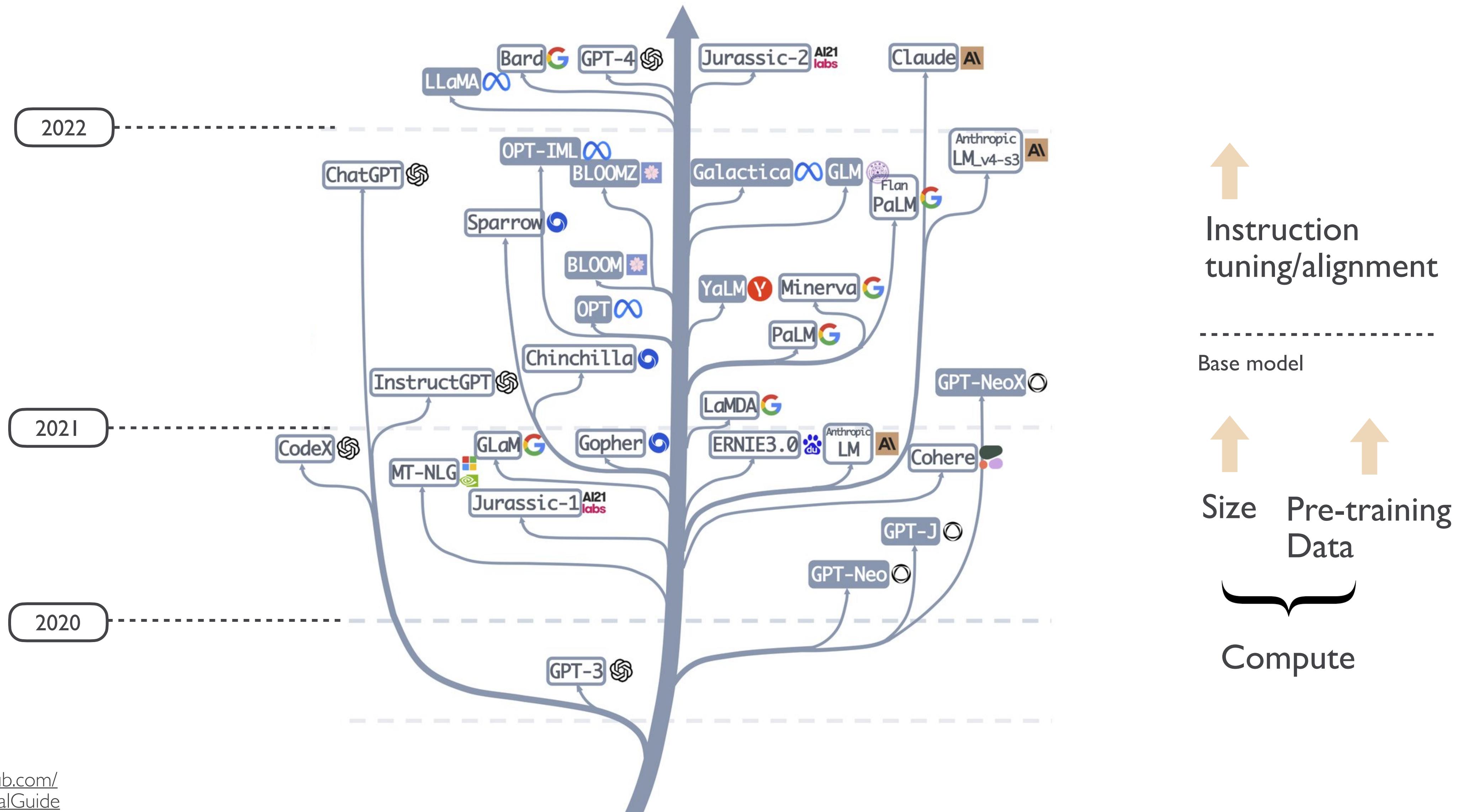


Image Credit: <https://github.com/Mooler0410/LLMsPracticalGuide>
 (Yang et al. 2023)

Large Language Models (2024)

The collage illustrates various Large Language Model interfaces:

- ChatGPT**: A screenshot showing the "Welcome to ChatGPT" screen with "Log in" and "Sign up" buttons.
- Bing**: A screenshot of the Bing AI interface asking "What will you do with Bing?" with a "Try Bing now" button.
- Cohere Coral**: A screenshot of the Cohere Coral interface showing a "Chat with Coral" section with a "GROUNDING" indicator and a "NEXT" button.
- Meta AI**: A screenshot of the Meta AI interface with the message: "My name is Meta AI. Think of me like an assistant who's here to help you learn, plan, and connect. What can I help you with today?"

GPTs

Discover and create custom versions of ChatGPT that combine instructions, extra knowledge, and any combination of skills.

Top Picks DALL-E Writing Productivity Research & Analysis Programming Education Lifestyle

Featured

Curated top picks from this week

- AllTrails**: Find trails that fit your nature for your next hike, ride, or run. By alltrails.com
- Consensus**: Your AI Research Assistant. Search 200M academic papers from Consensus, get science-based... By Consensus
- Code Tutor**: Let's code together! I'm Khanmigo Lite, by Khan Academy. I won't write the code for you, but I'll help you... By khanacademy.org
- Books**: Your AI guide in the world of literature and reading. By Josh Brent N. Villocido

Large Language Models (LLMs)



Can you suggest some food and dietary sources that are rich in Vitamin D?

To increase your vitamin D levels through diet, consider incorporating the following foods, as they are rich in vitamin D:

- **Protein Foods:** Freshwater rainbow trout and salmon are excellent sources, providing between 383 to 645 IU of vitamin D per 3-ounce serving. Other good seafood options include light canned tuna, herring, sardines, tilapia and flounder.
- **Dairy and Fortified Soy Alternatives:** Unsweetened soy milk, low-fat milk, nonfat and low-fat yogurt, fat-free milk, low-fat kefir, and fortified low-fat or fat-free American cheese are all good dairy sources of vitamin D.
- **Vegetables:** Certain varieties of mushrooms can offer significant amounts of vitamin D, especially if they have been exposed to sunlight.
- **Fruit and Fortified Beverages:** Fortified orange juice can provide about 100 IU of vitamin D per cup.
- **Other Sources:** Unsweetened almond and rice beverages are also fortified with vitamin D.

While these foods can help boost your vitamin D levels, sunlight exposure is another significant source of vitamin D for many people...

Assistant output modified from ChatGPT-4



Bard



perplexity



Outline

I. LLMs as General Parametric Task Solvers

- I. LLM Pre-training and Scaling Laws
- 2. Instruction Tuning

2. External Knowledge Access

- I) Retrieval Augmentation
- II) Tool Augmentation

3. LLM Agents and Advanced Applications

Language Model: Architecture

Transformer (2017)

- Sequence modeling
- Input
 - Token Embeddings + Position Encodings
- Transformer block
 - Multi-Head Self-Attention
 - Feed-Forward
- Output
 - Linear Projection
 - Softmax Probabilities over Vocabulary

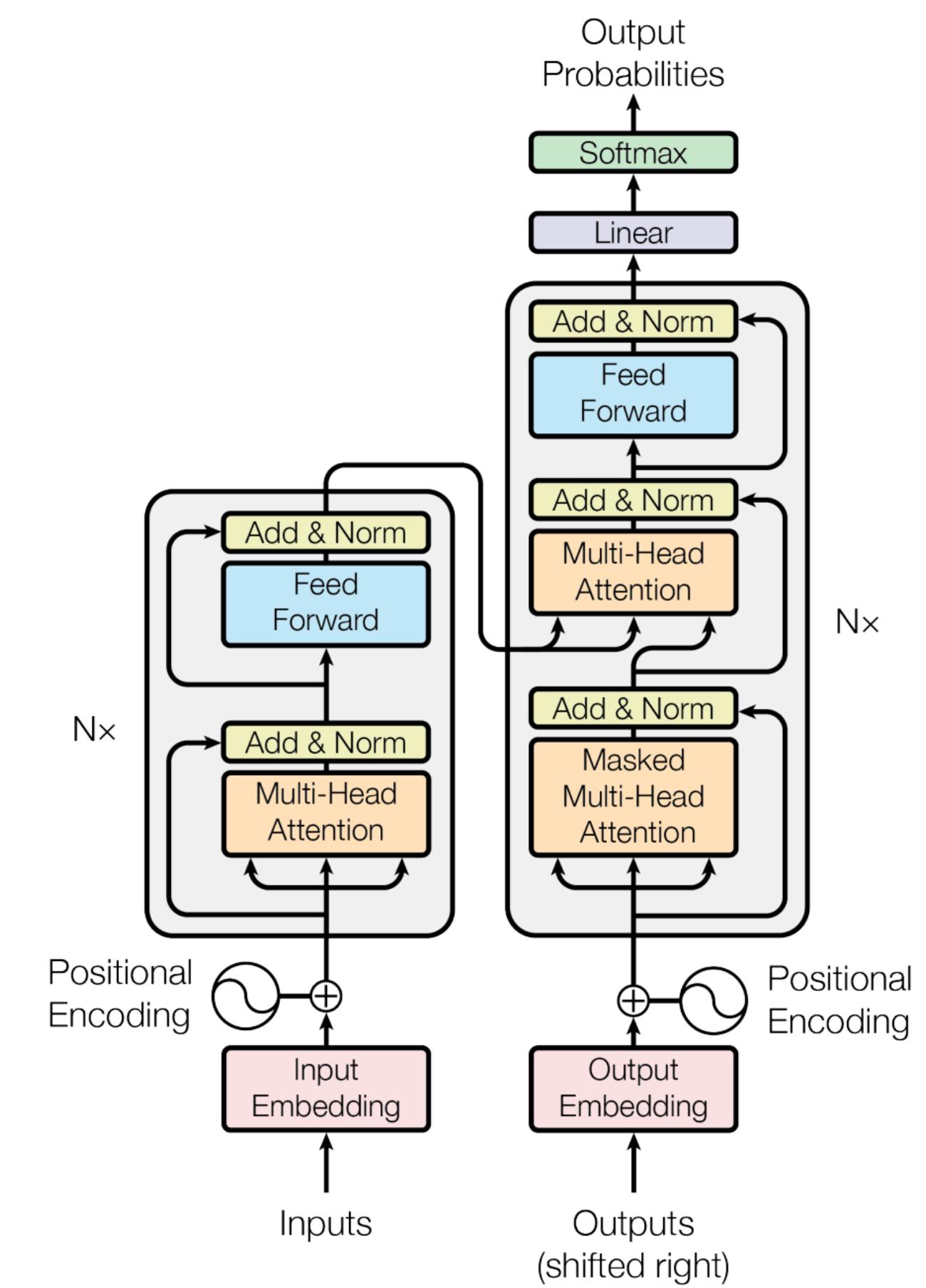


Image taken from Attention Is All You Need. Vaswani et al. 2017.

Language Model: Architecture

Transformer (2017)

- Sequence modeling
- Input
 - Token Embeddings + Position Encodings
- Transformer block
 - Multi-Head Self-Attention
 - Feed-Forward
- Output
 - Linear Projection
 - Softmax Probabilities over Vocabulary

Other types of sequence modeling architectures, such as the state-space models (SSMs), are not covered in this lecture but are worth diving into if you're interested in further advancing LLM capabilities.

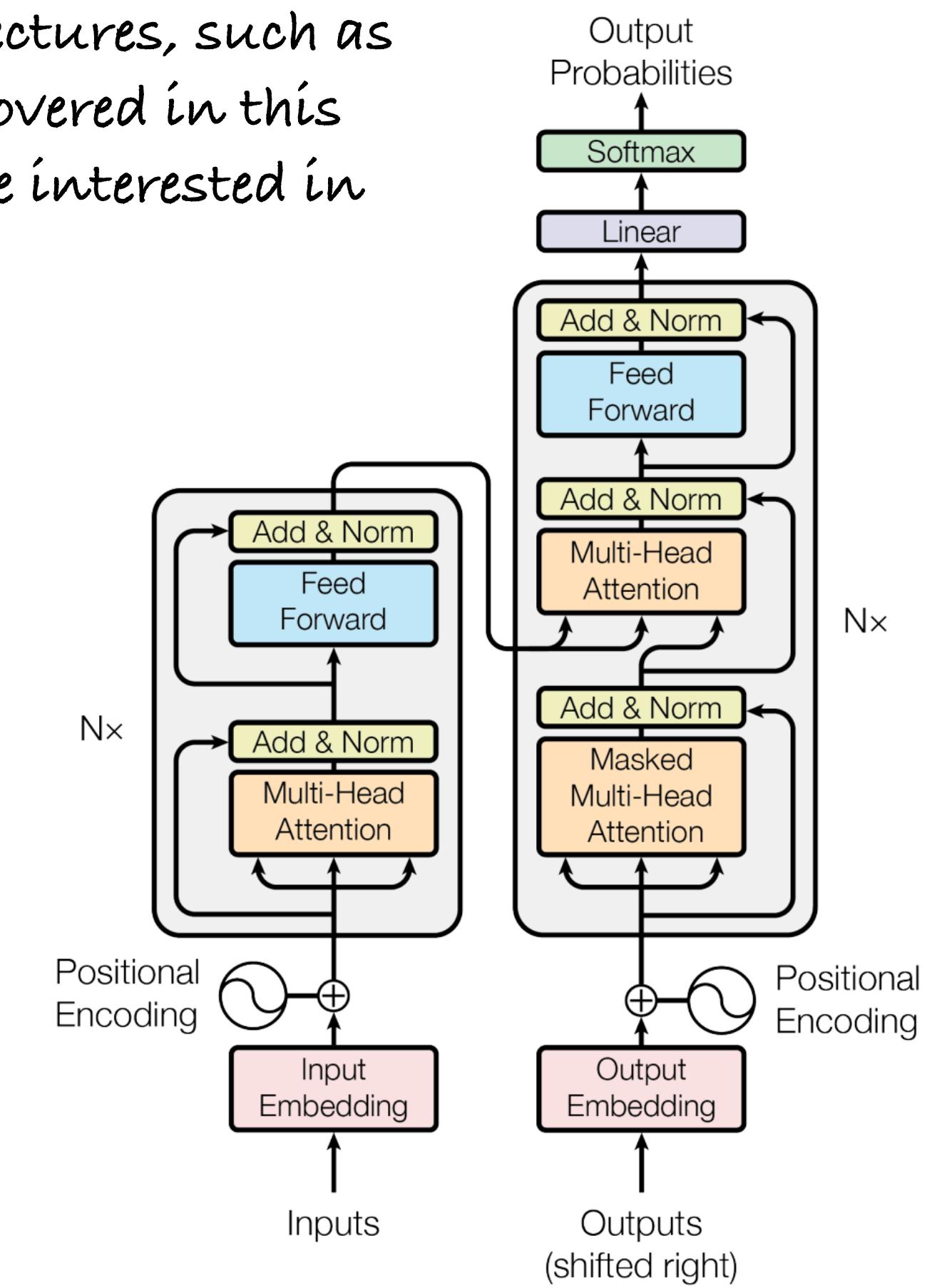


Image taken from Attention Is All You Need. Vaswani et al. 2017.

Language Model: Learning Objectives

Variations of LM objectives (token prediction)

Language Model: Learning Objectives

Masked Language Models

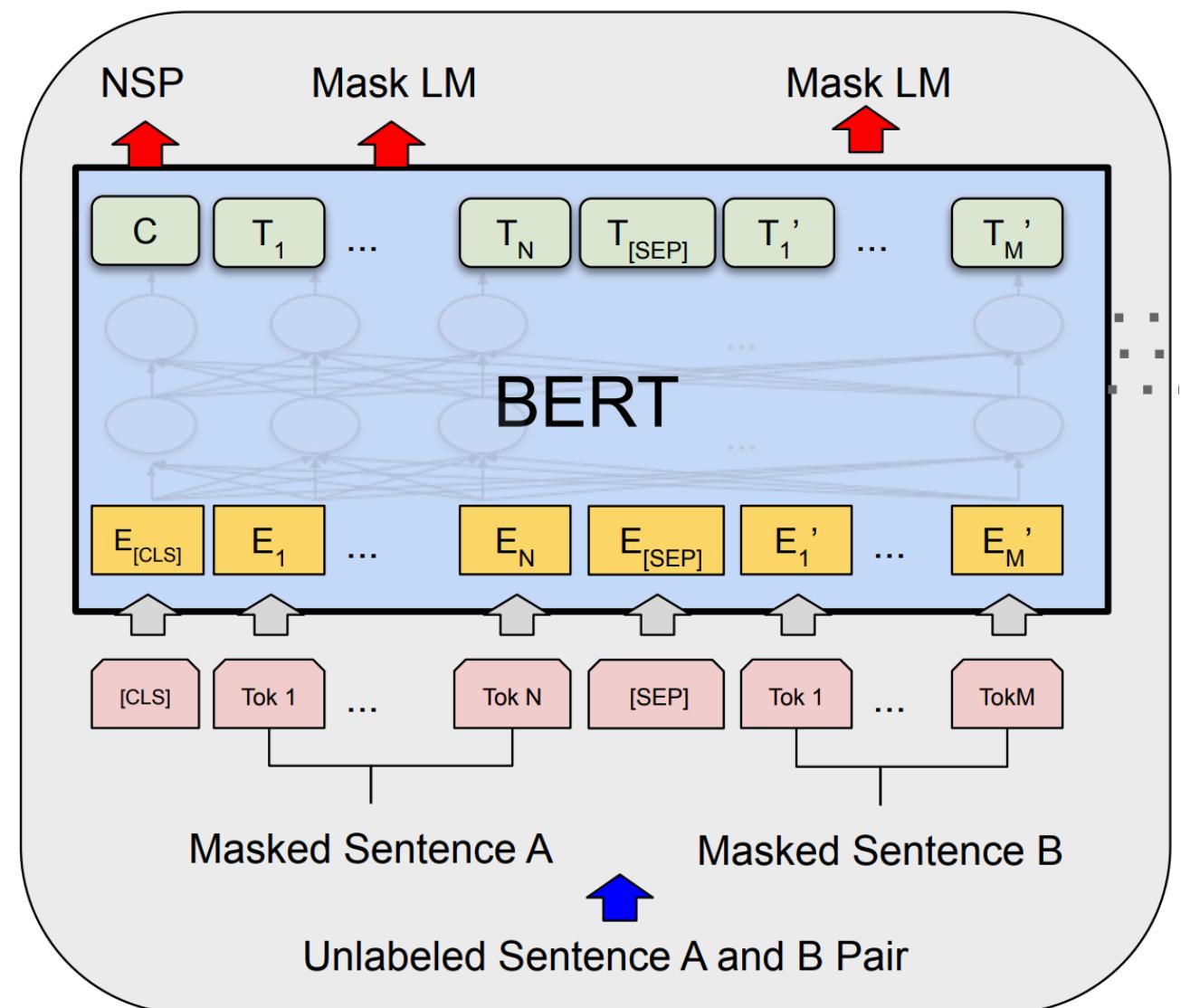


Image taken from (1) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Rafael et al. 2019
(2) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Devlin et al. 2018.

Language Model: Learning Objectives

Encoder-Decoder Language Models

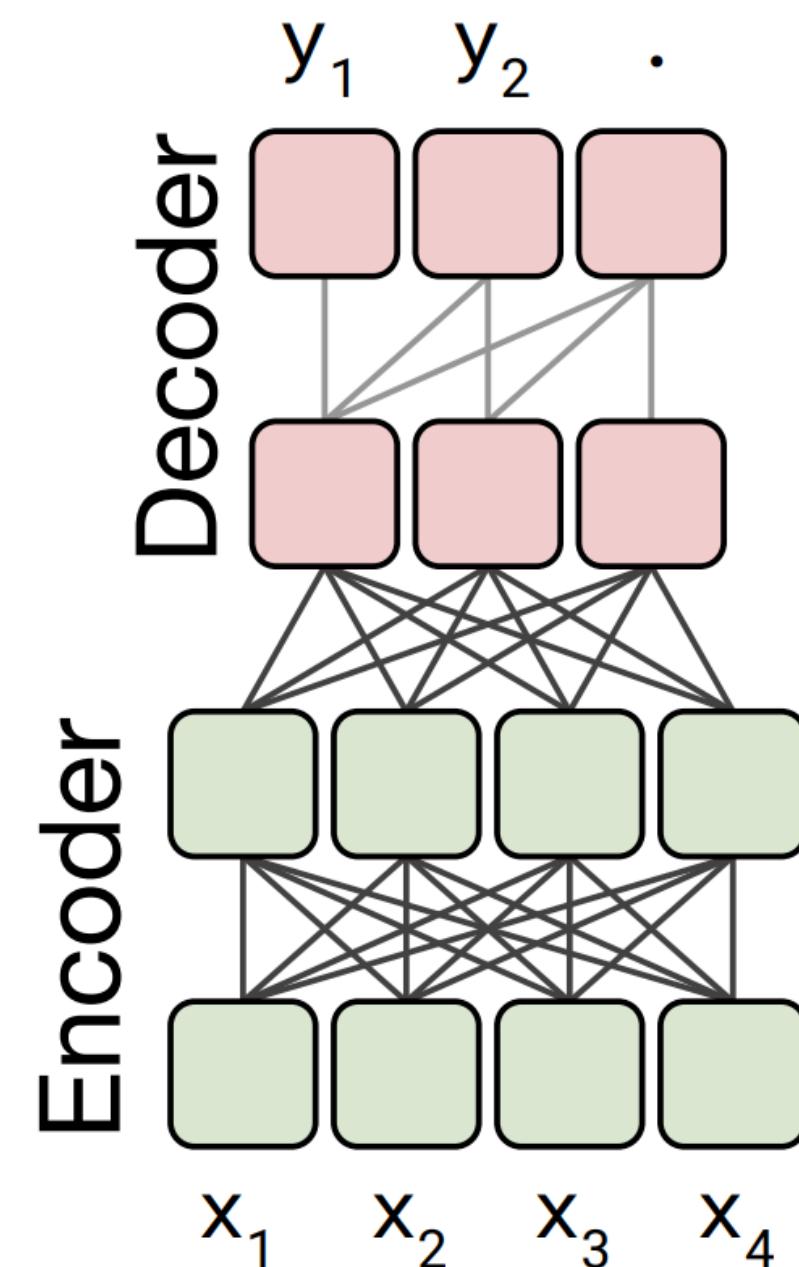
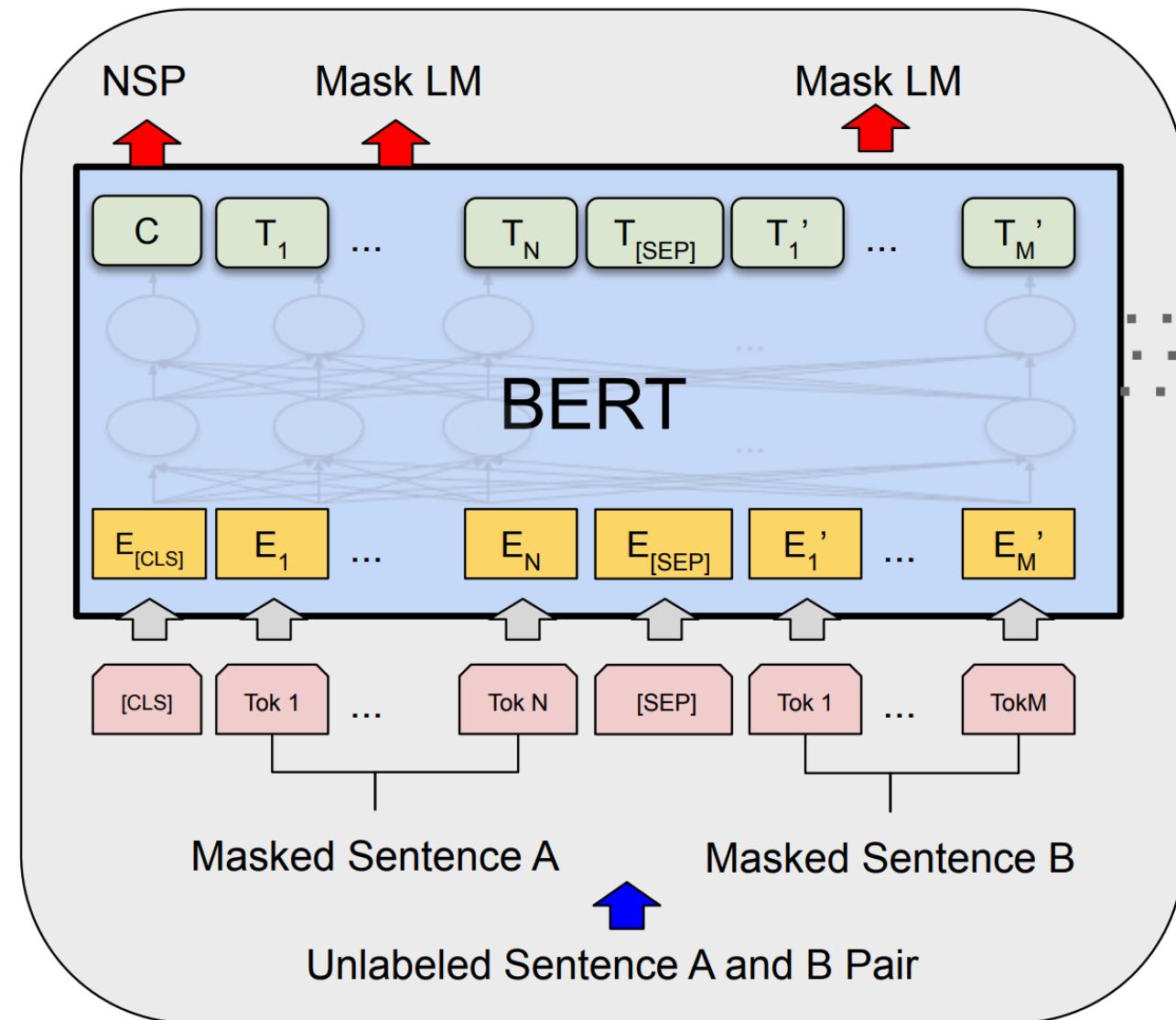


Image taken from (1) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Rafael et al. 2019
(2) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Devlin et al. 2018.

Language Model: Learning Objectives

(Autoregressive) Language Models

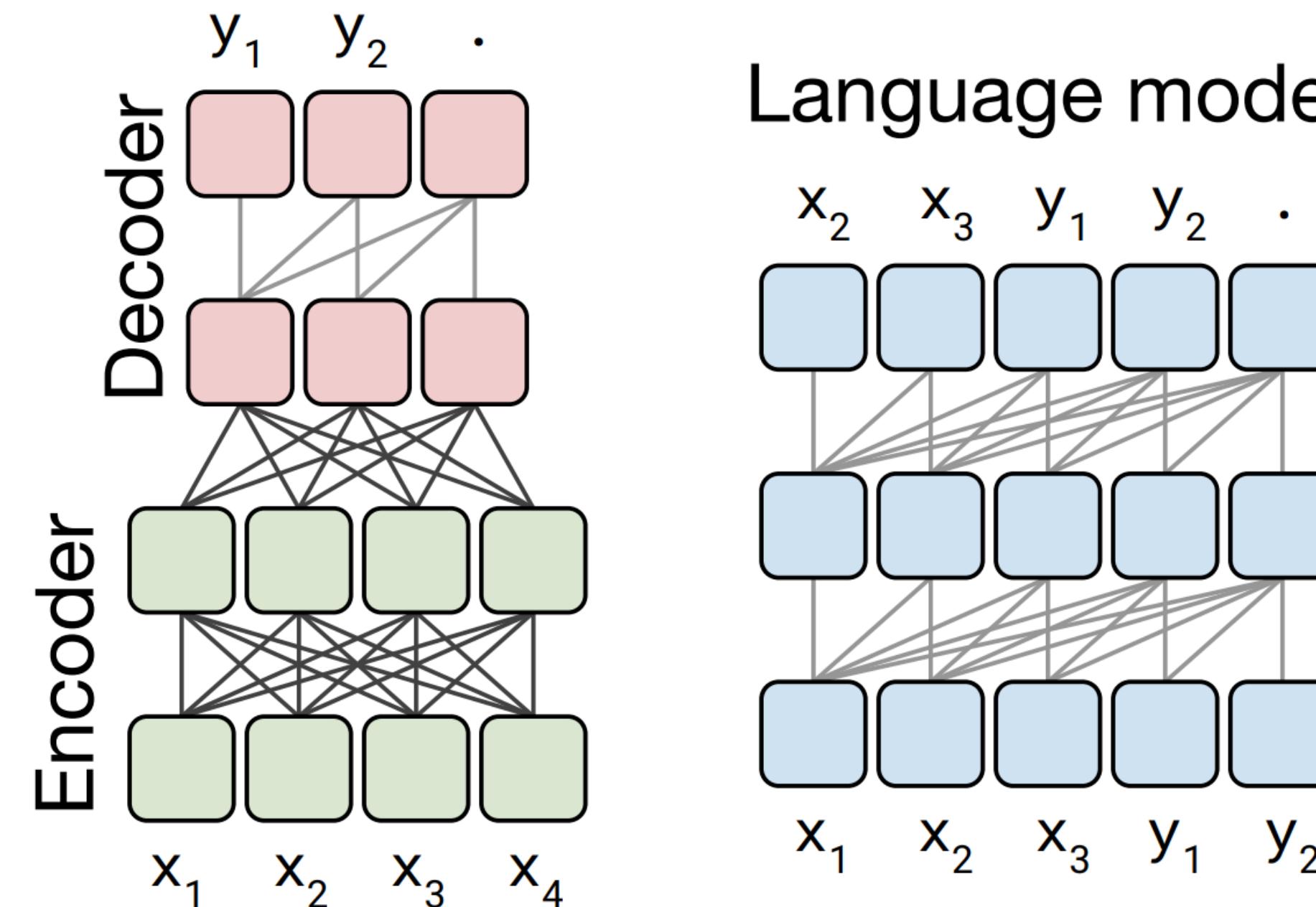
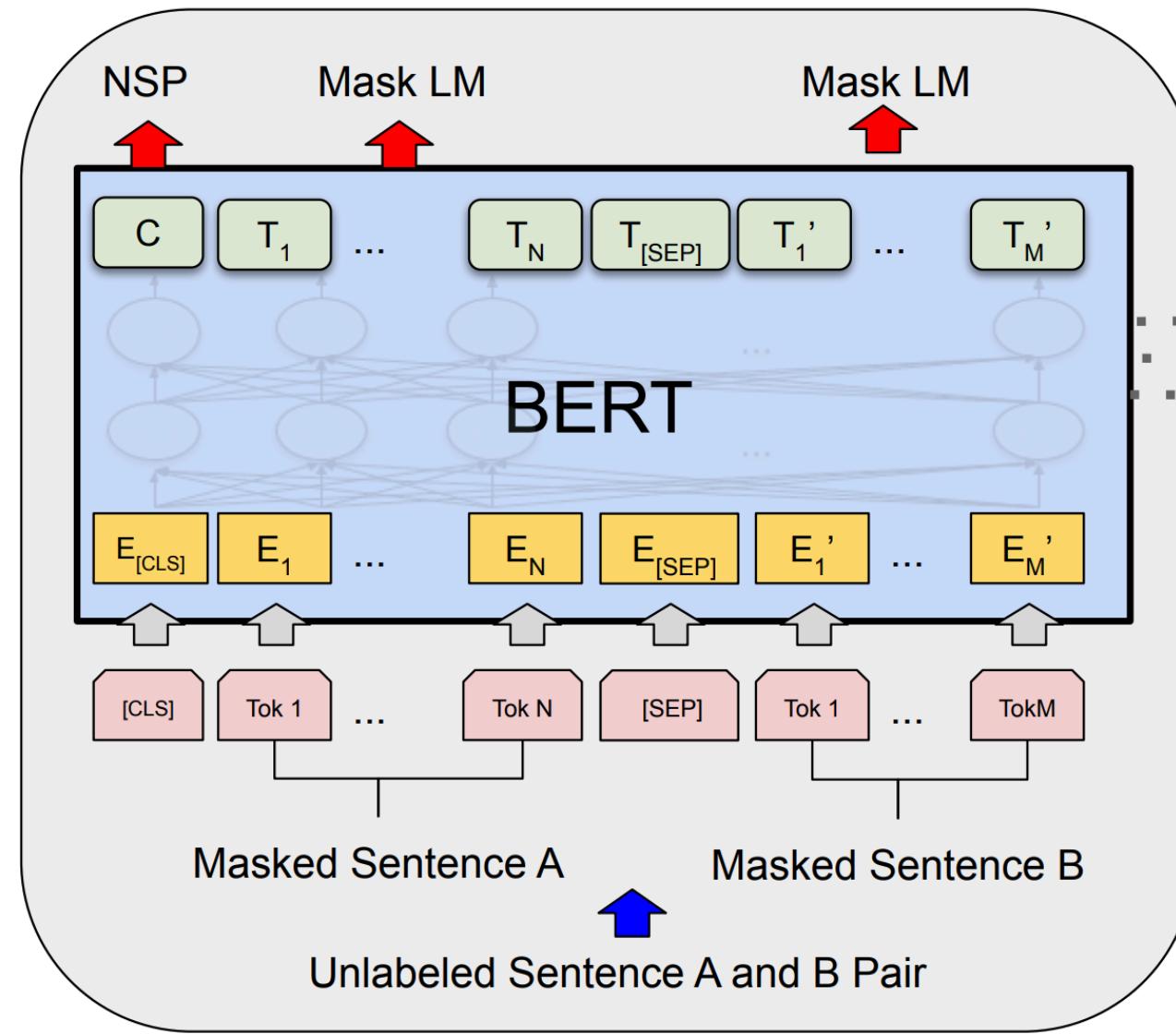


Image taken from (1) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Rafael et al. 2019
(2) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Devlin et al. 2018.

Language Model: Learning Objectives

(Autoregressive) Language Models with Prefix

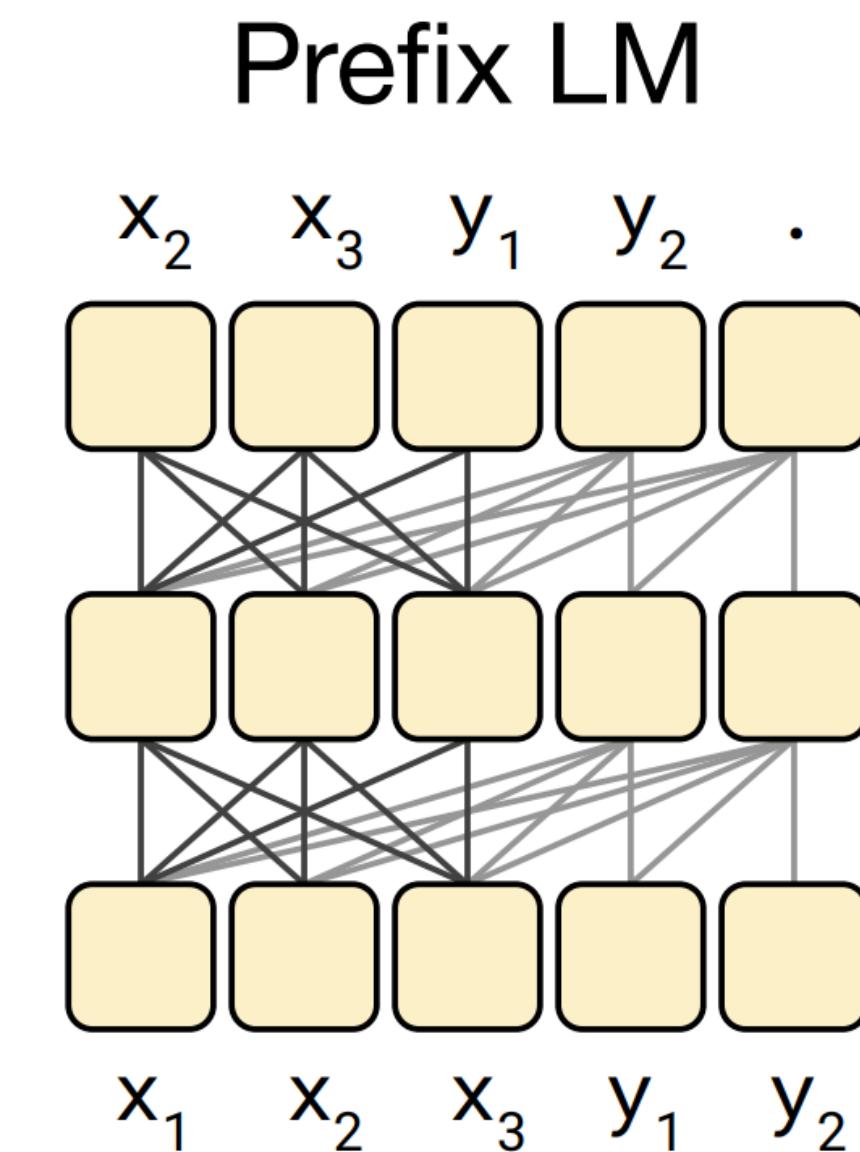
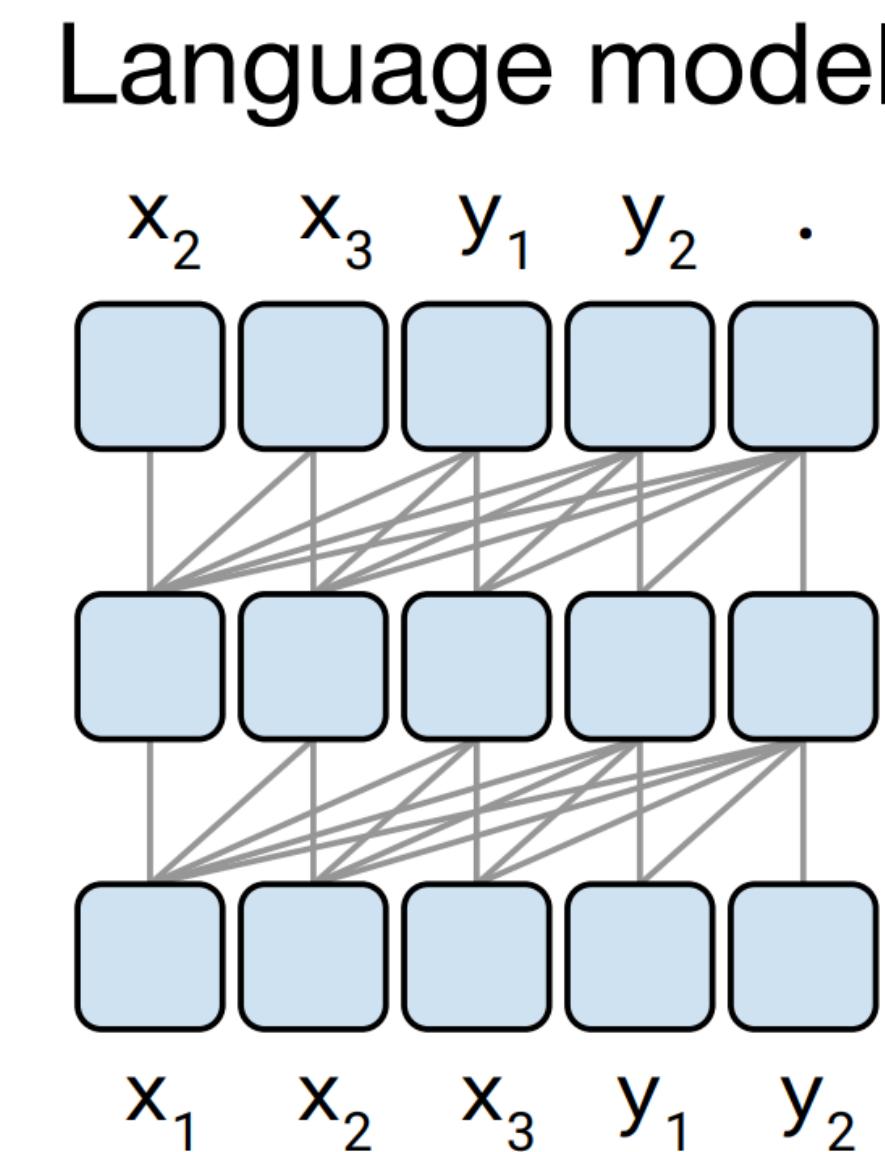
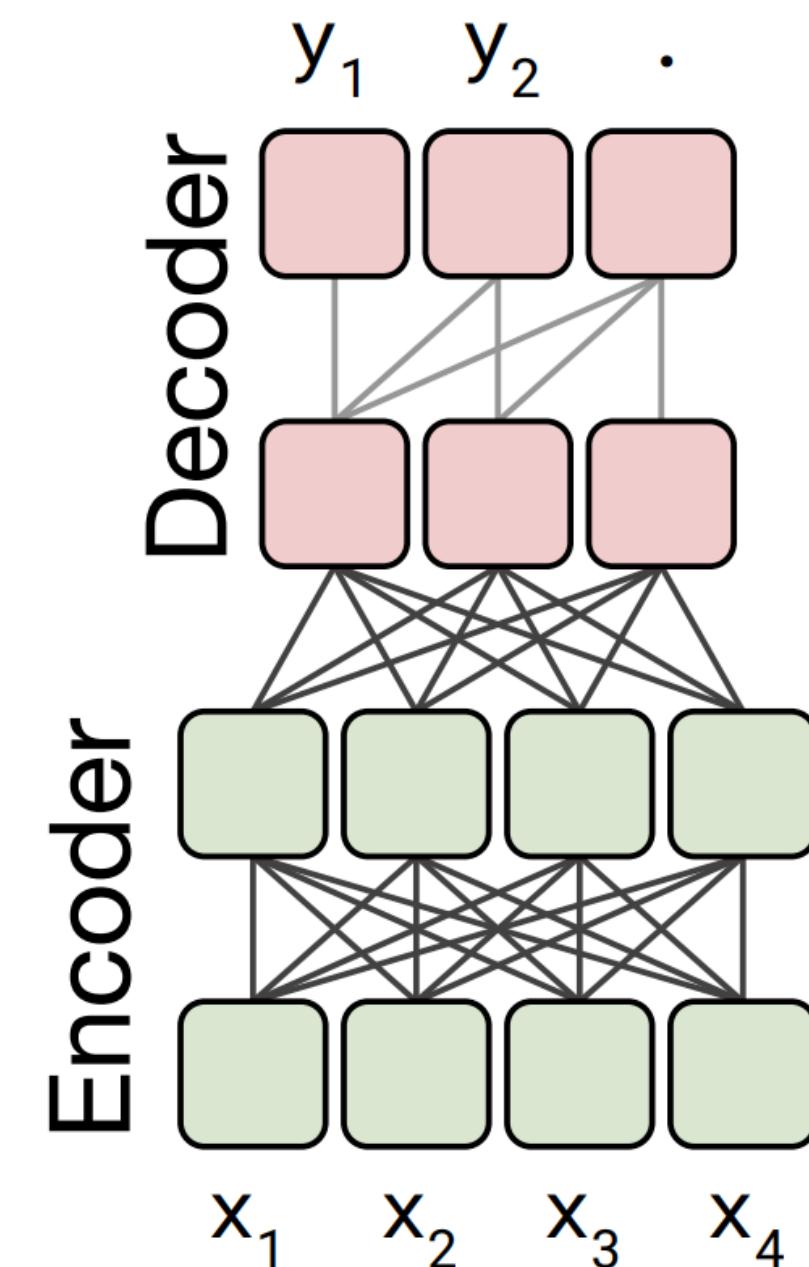
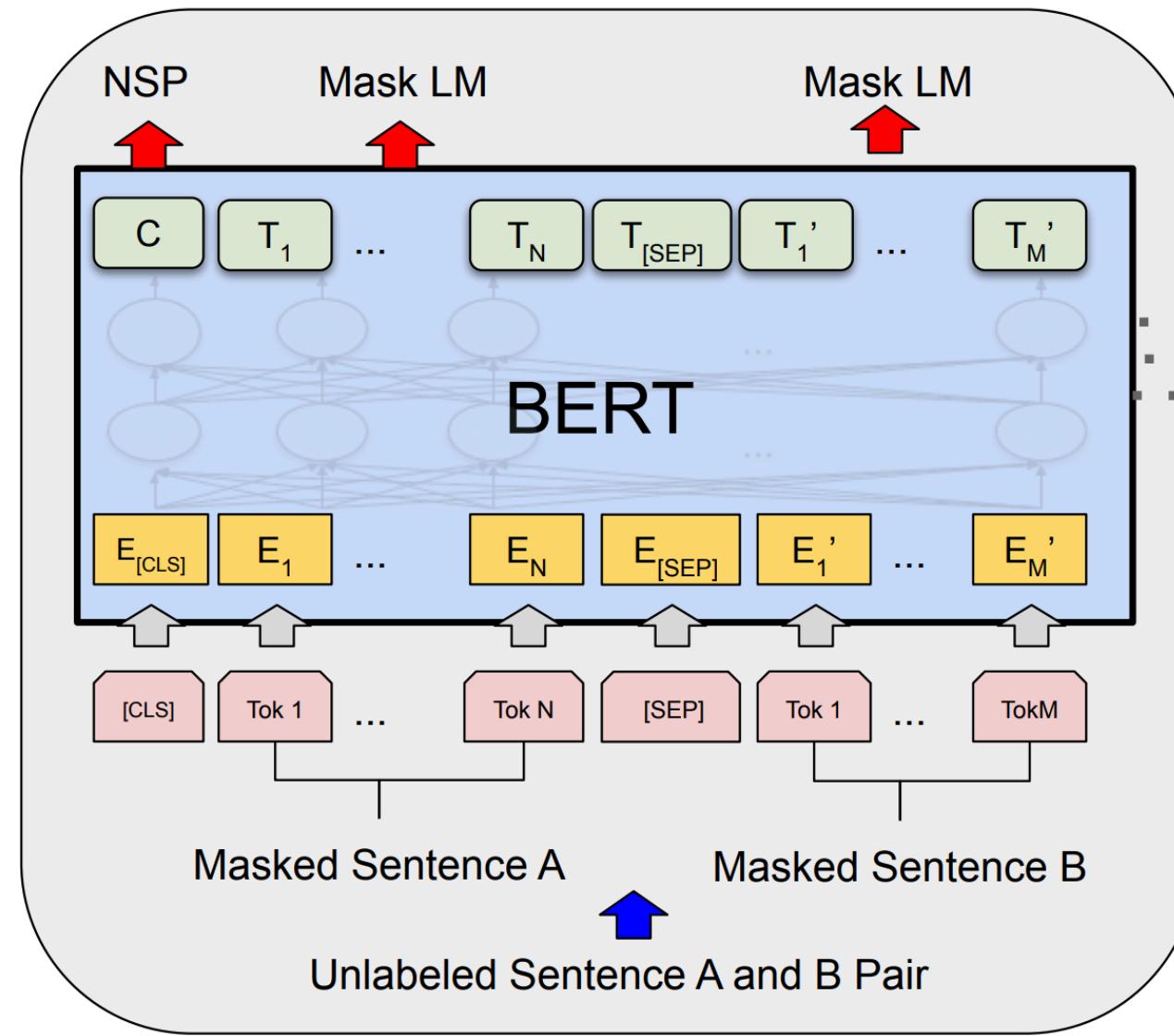
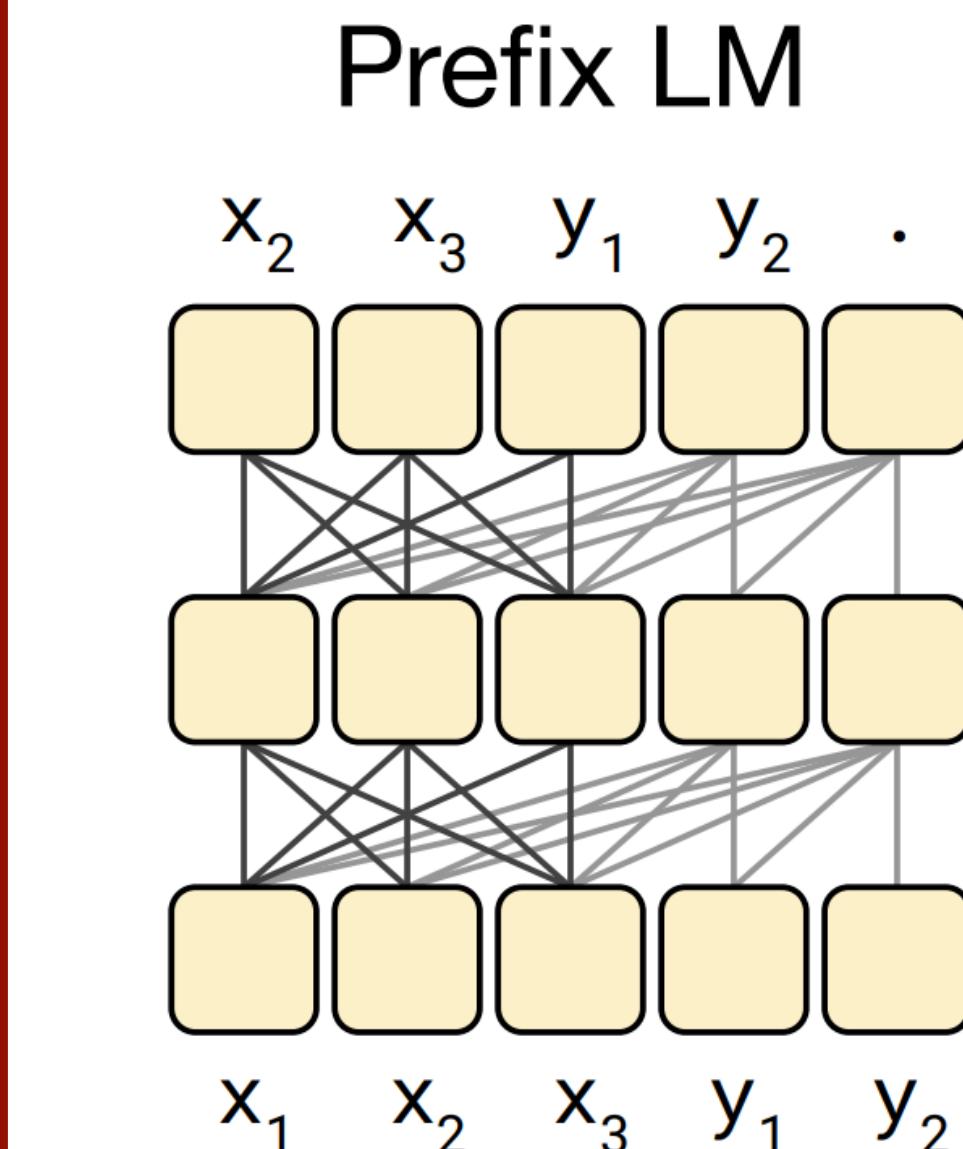
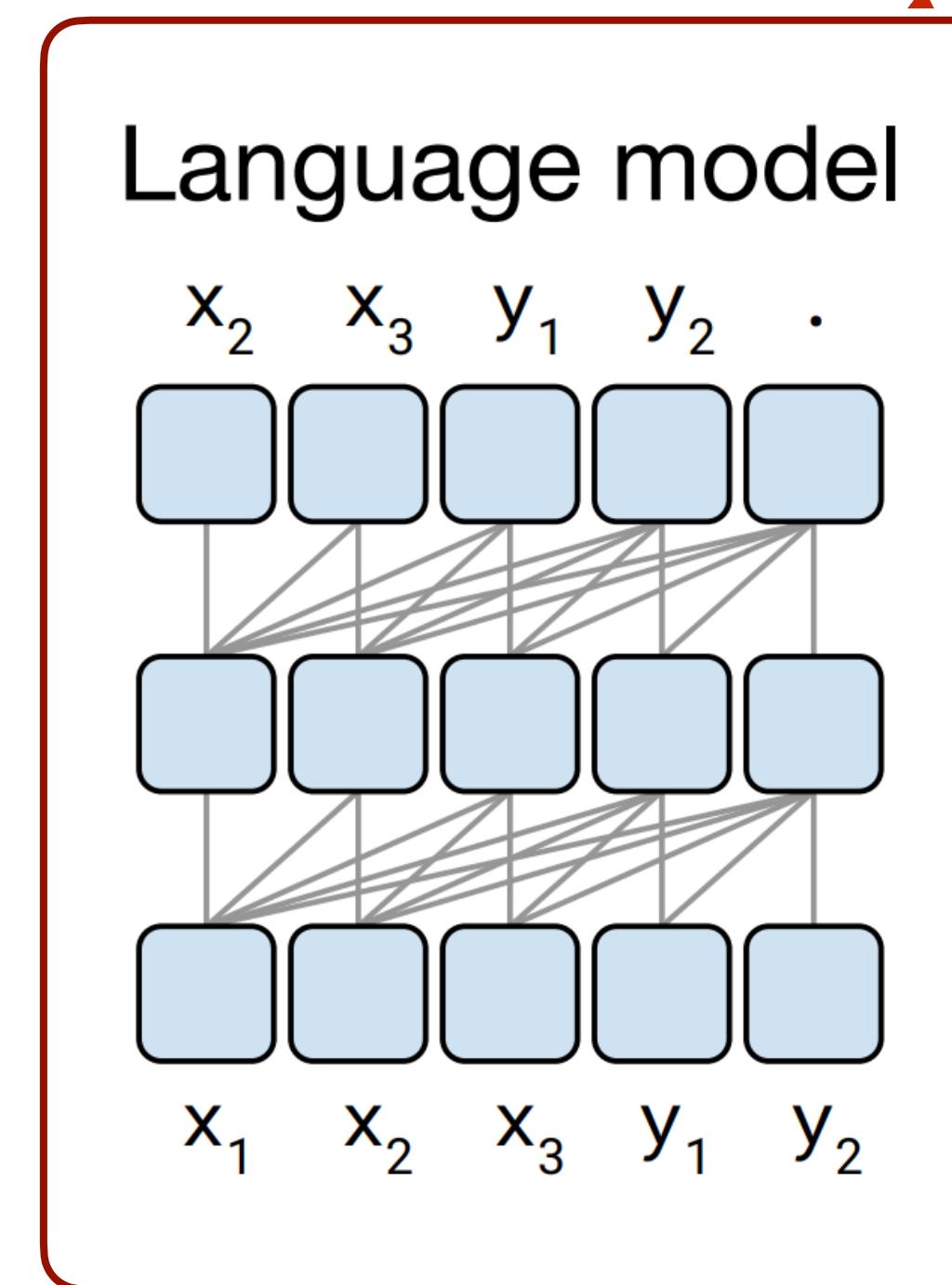
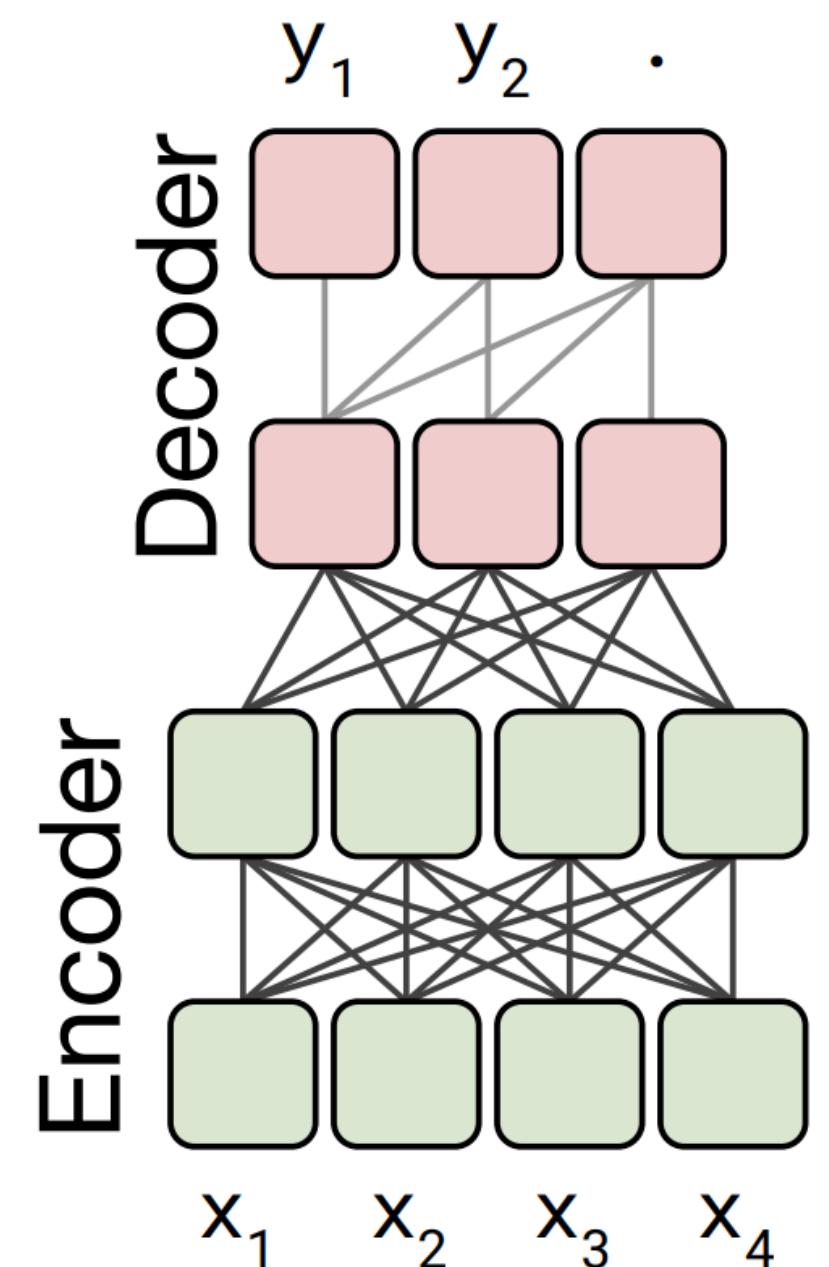
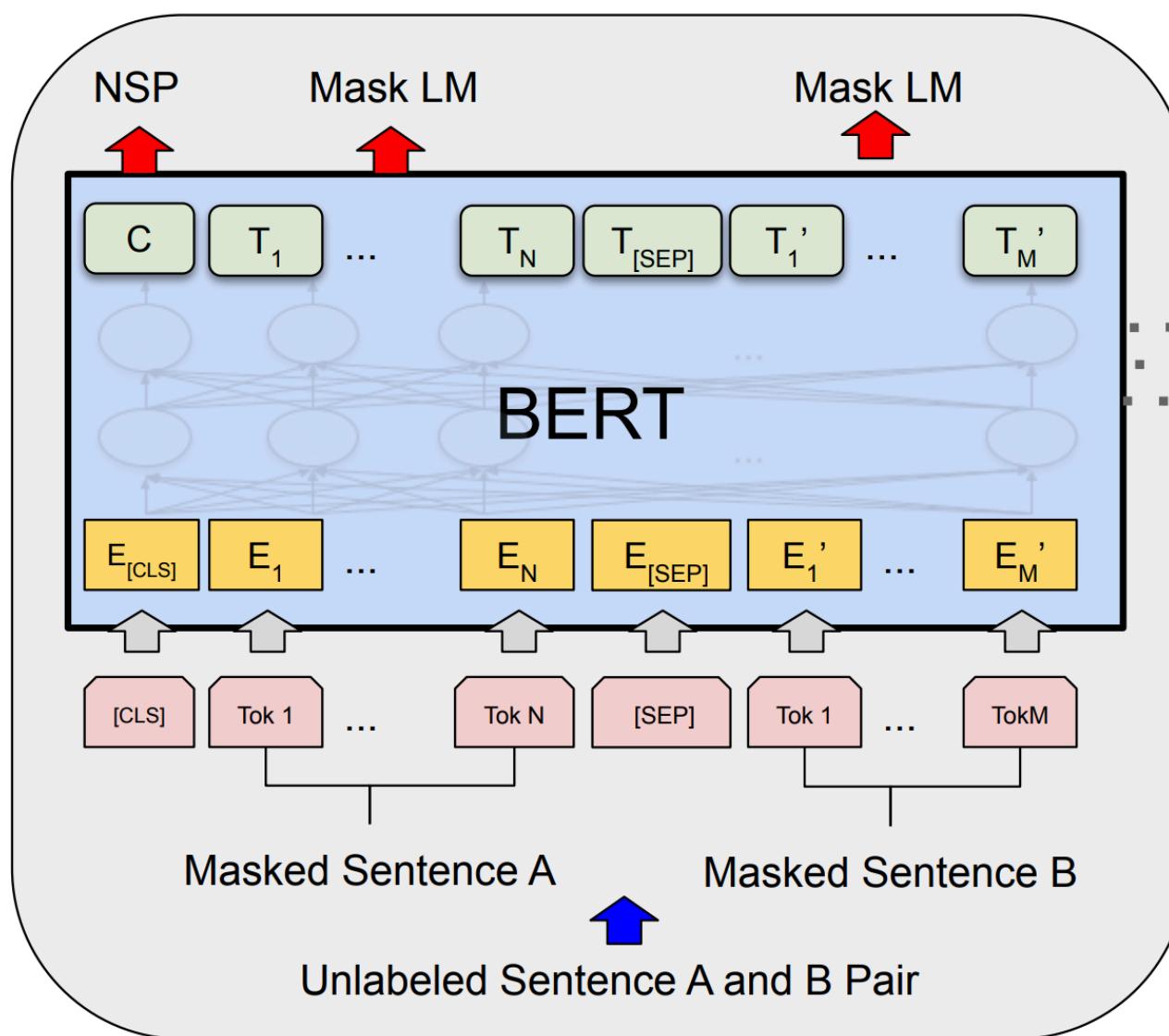


Image taken from (1) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Rafael et al. 2019
(2) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Devlin et al. 2018.

Language Model: Learning Objectives

(Autoregressive) Language Models with Prefix



- A) Capable of producing free-form output
- B) Best scaling properties

Image taken from (1) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Rafael et al. 2019
(2) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Devlin et al. 2018.

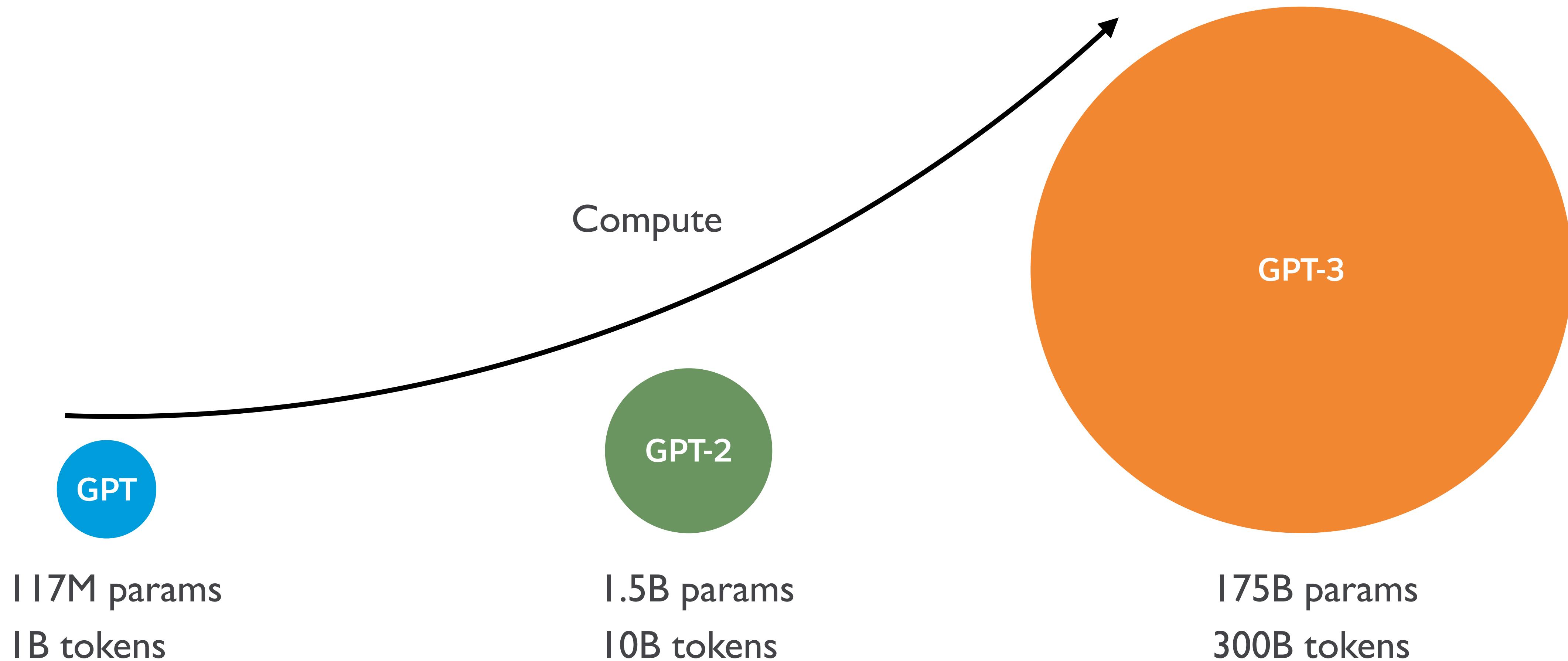
Autoregressive LMs

- Autoregressive LMs are pre-trained over a large collection of unlabeled text using the next-token prediction objective as shown below.

Given a language sequence $x = (s_1, s_2, \dots, s_n)$

$$p(x) = \prod_{i=1}^n p(s_i | s_1, \dots, s_{i-1})$$

Generative Pre-training (GPT)



Scaling Performance

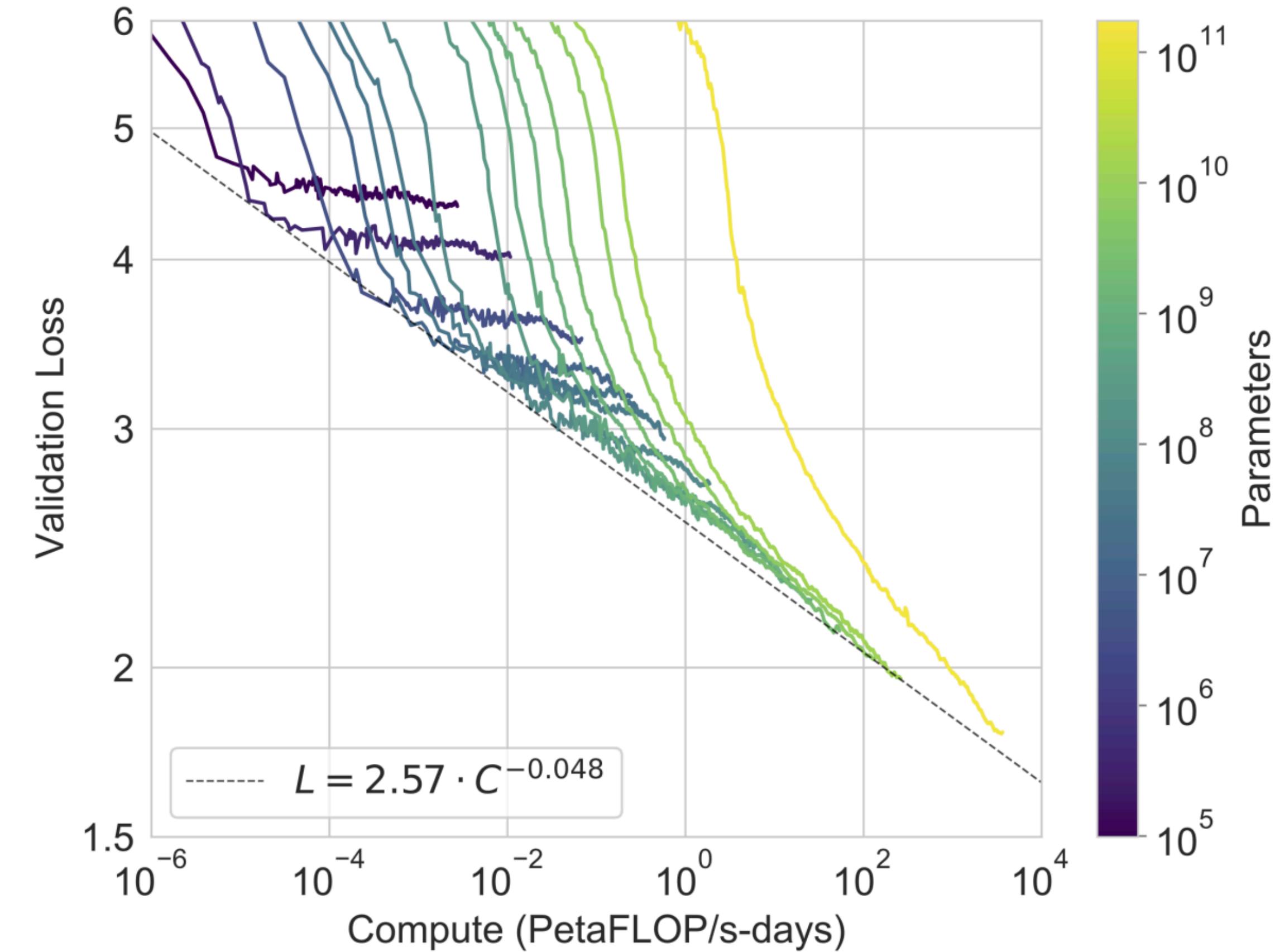
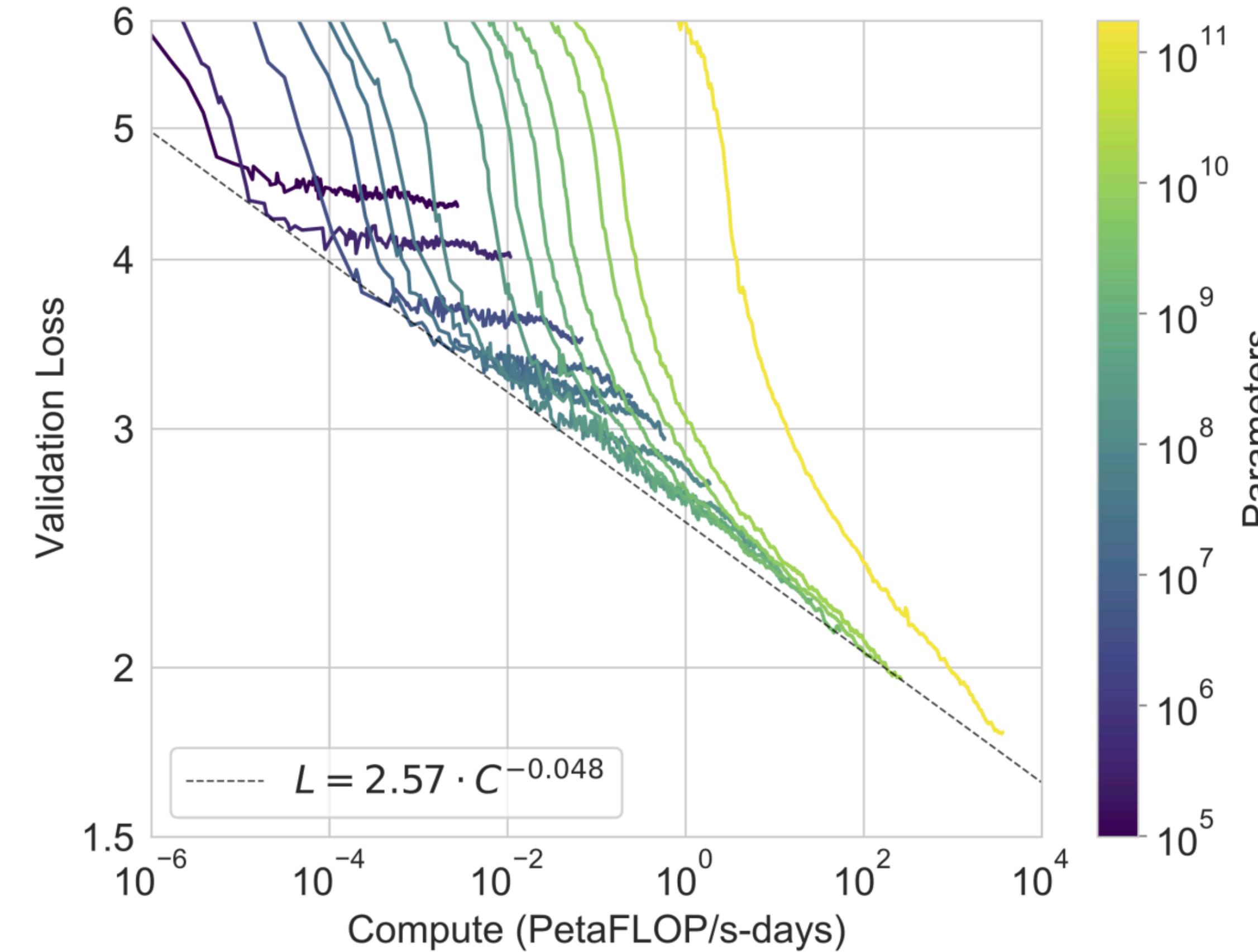


Image taken from Language Models are Few-shot Learners. Brown et al. 2020.

Scaling Performance



Refer to Hoffmann et al. (2022) for mathematical scaling law derivation.

Image taken from Language Models are Few-shot Learners. Brown et al. 2020.

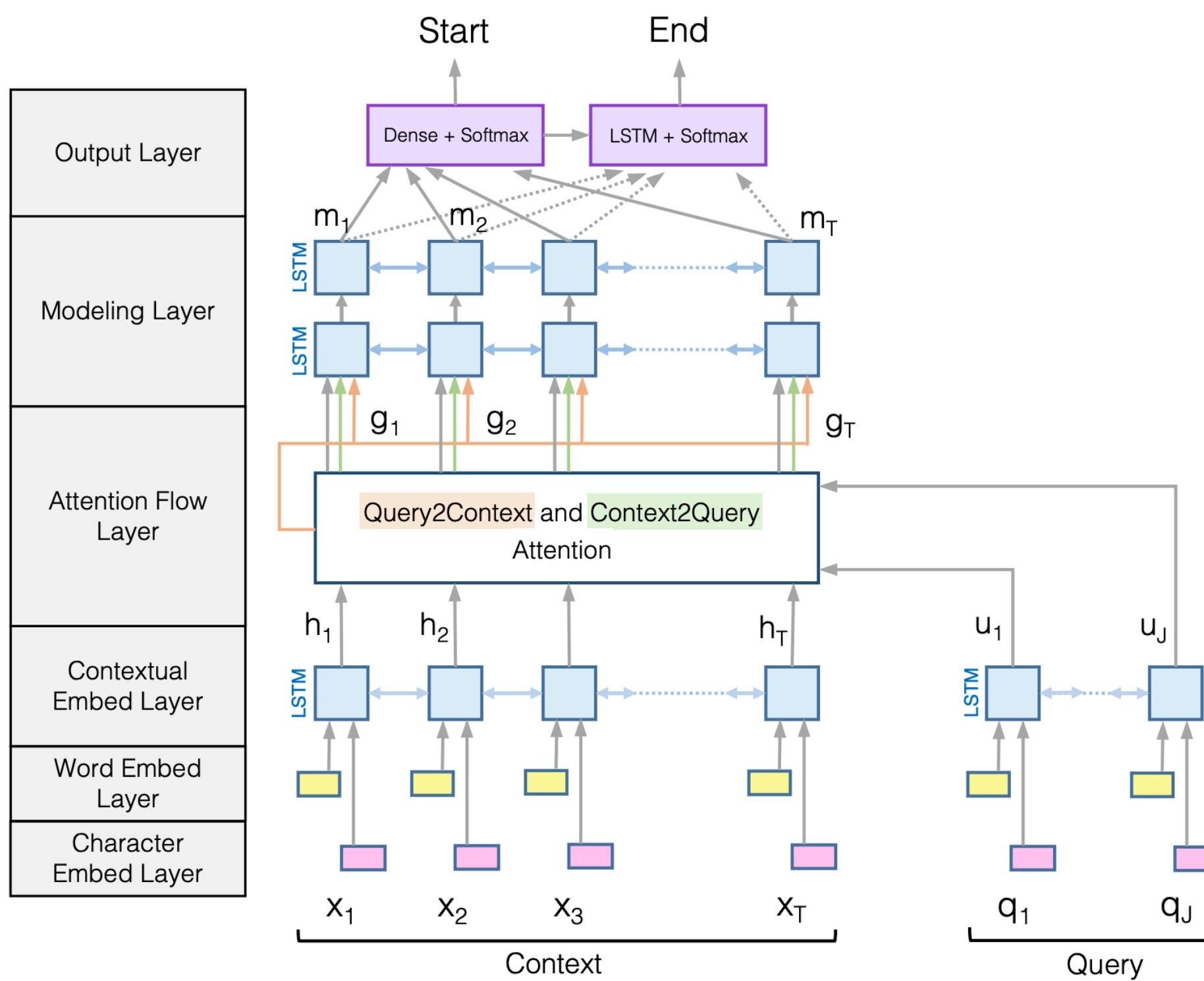
LMs: Language Representation → End-to-End Solvers

- The introduction of autoregressive and encoder-decoder Transformer LMs shifted the role of LMs from language representations to **end-to-end, multi-task** language task solvers.

Image Credit: Improving Language Understanding by Generative Pre-Training. Radford et al. 2018.

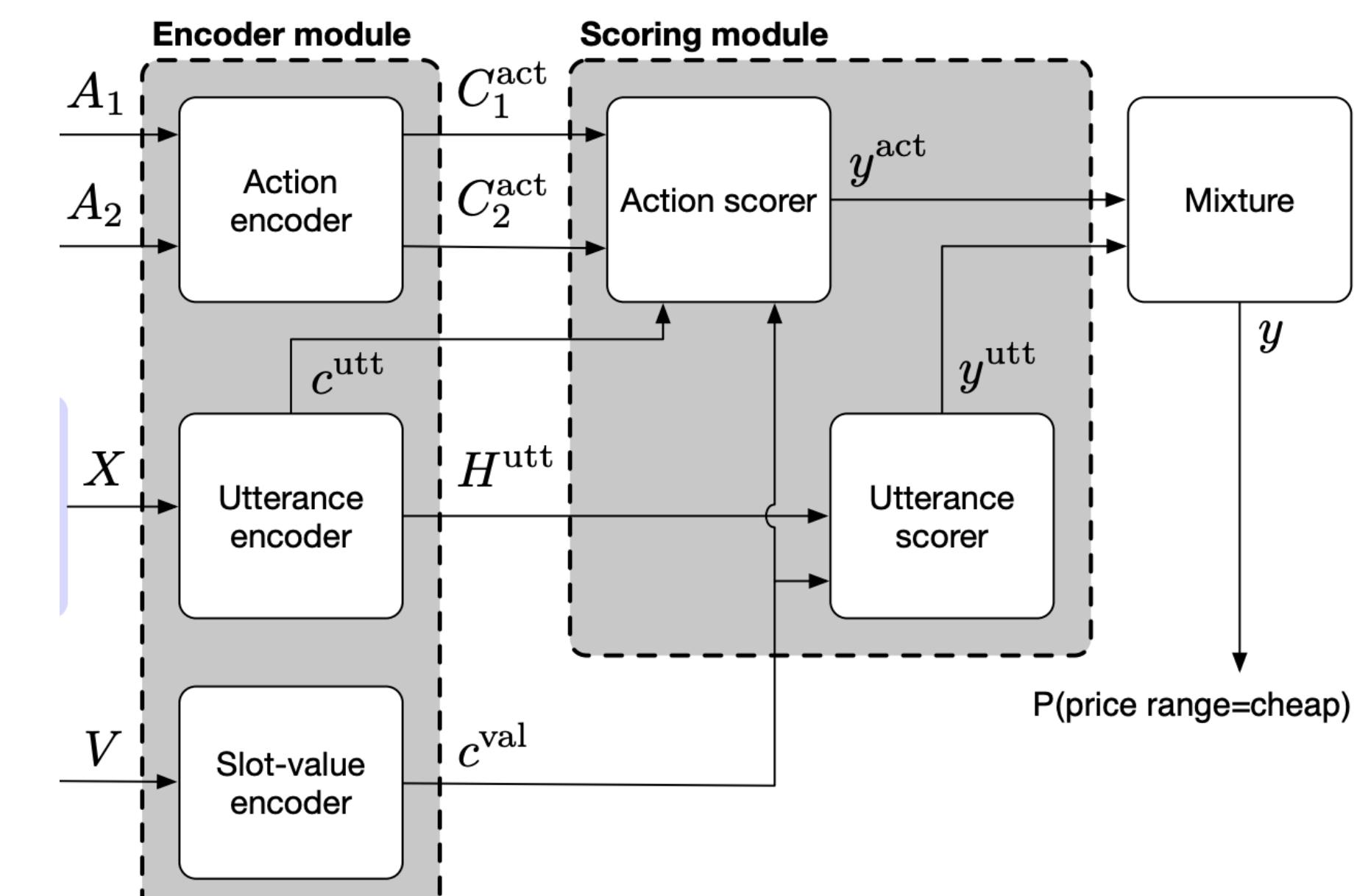
Natural Language Processing 5 Years Ago...

Question Answering



Bidirectional Attention Flow for Machine
Comprehension (Seo et al. ICLR 2017)

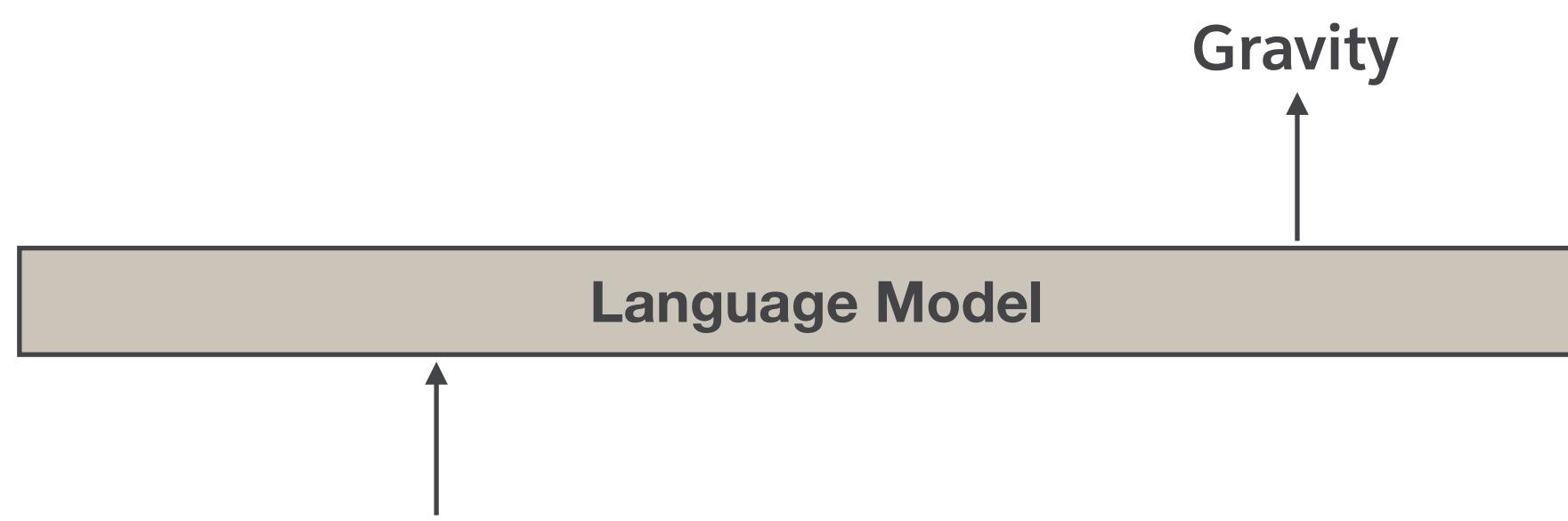
Dialogue Systems



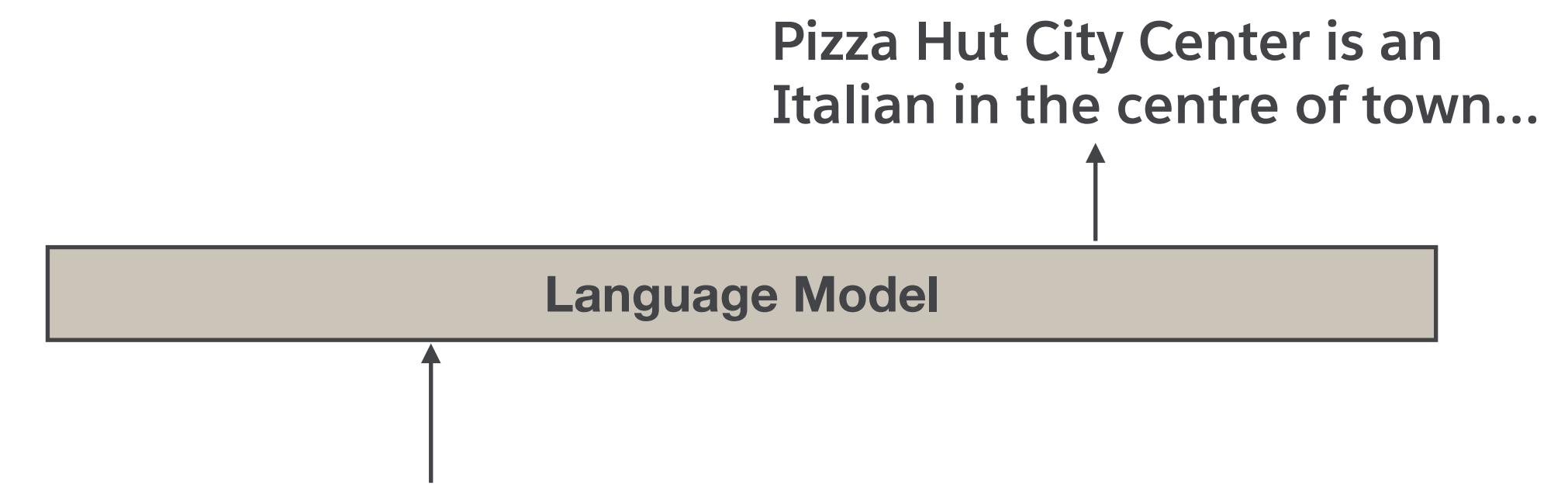
Global-Locally Self-attentive Dialogue State
Tracker (Zhong et al. ACL 2018)

Natural Language Processing Today

Question Answering



Dialogue Systems



Prompt Engineering

- Right after GPT-3 (Brown et al. 2020) was introduced, the field witnessed a flourish of work in **prompt-engineering** and **in-context learning**.

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

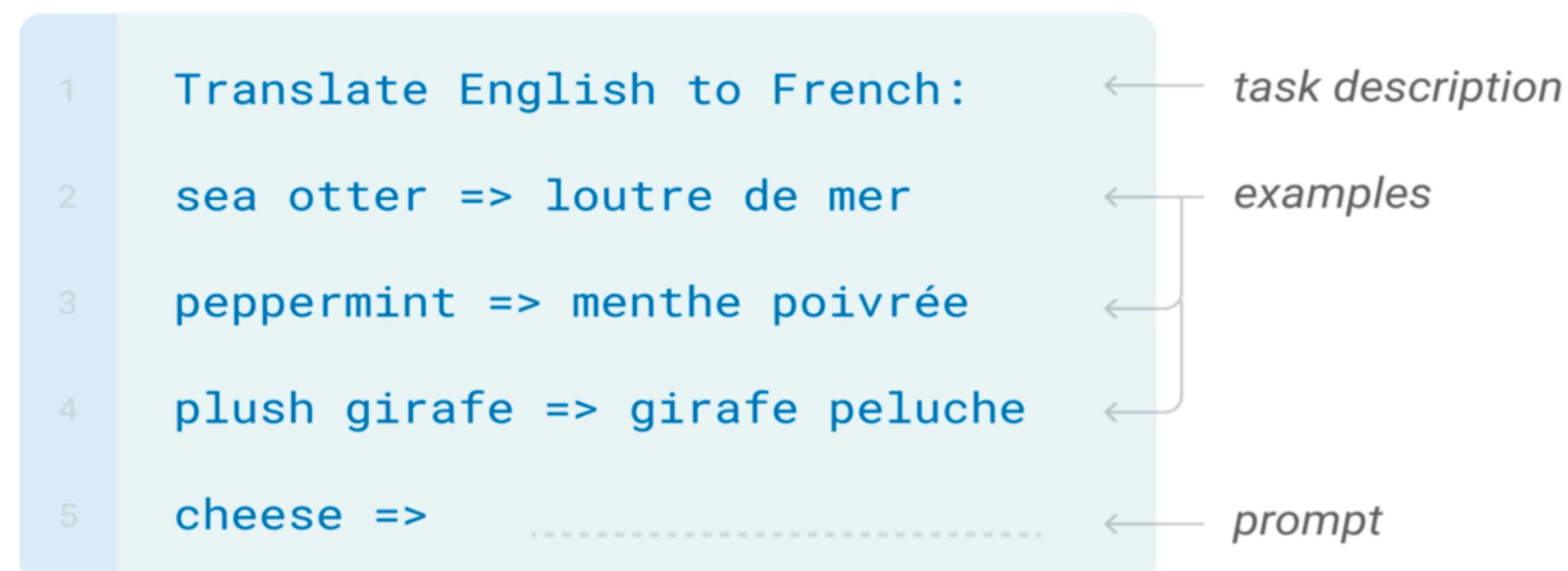


Image Credit: Language Models are Few-shot Learners.
Brown et al. 2020.

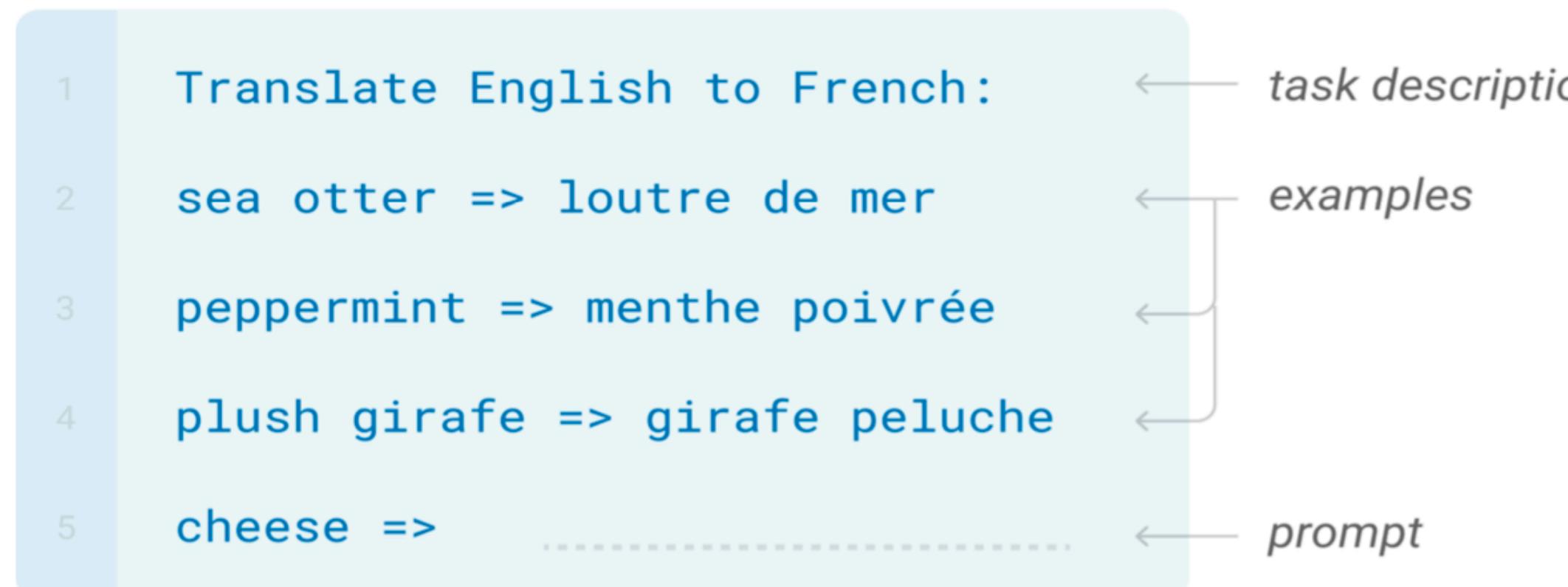
Prompt Engineering

- Right after GPT-3 (Brown et al. 2020) was introduced, the field witnessed a flourish of work in **prompt-engineering** and **in-context learning**.

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

“Prompt engineering” is a sign of lack of robustness



Human language communication less often involves the usage of examples

Image Credit: Language Models are Few-shot Learners.
Brown et al. 2020.

Instruction Tuning/LM Alignment

- Instruct-able/aligned LMs were developed following the demand of **natural & robust human-model interaction** and **less dependency of inference-time supervision**.

Instruction Tuning

- Instruction tuning is the process of fine-tuning LLMs using a collection of tasks expressed via instructions and ground-truth response pairs (Mishra et al. 2021; Wei et al. 2021; Sahn et al. 2021; Ouyang et al. 2022).
- It effectively **enables natural & robust human-model interaction, improves generalization and reduces the dependency of inference-time supervision.**

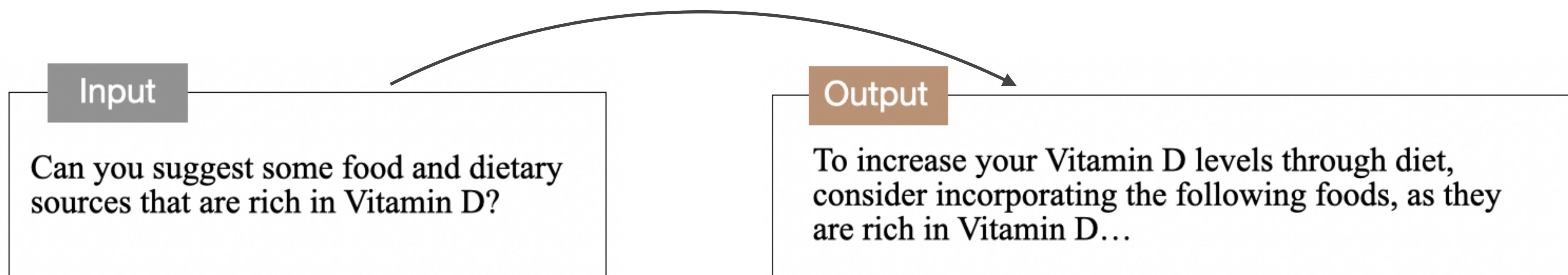
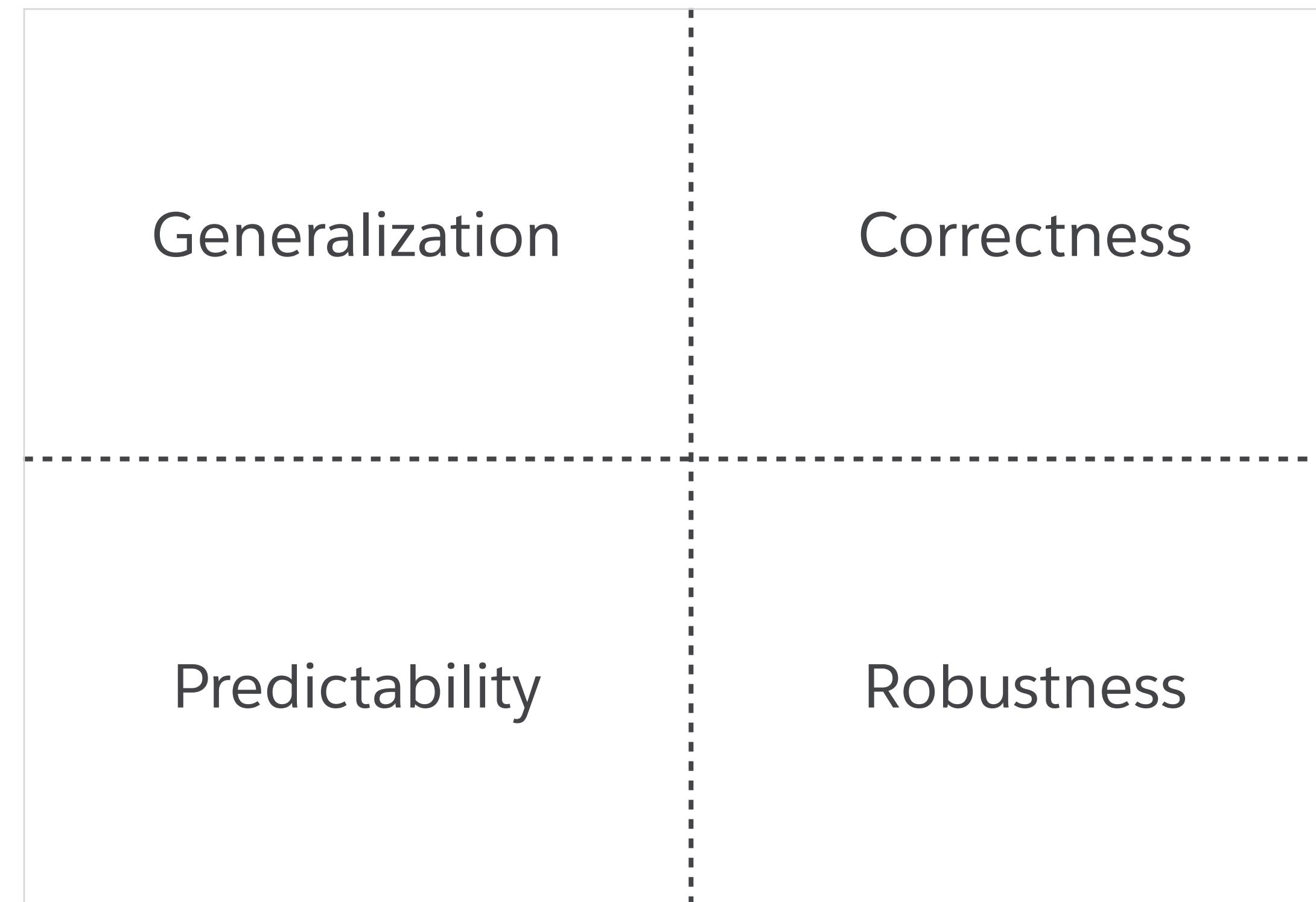


Image Credit: Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. Wang et al. 2022.

Instruction Tuning: Desiderata



Correctness

Correct the grammar error in this sentence
and also fix the letter cases:

“the quick brown fox jump over the lazy dog”

The quick brown fox jumps over the lazy dog.

Instructable LM

Cross-task Generalization

Correct the grammar error in this sentence and also fix the letter cases:
“the quick brown fox jump over the lazy dog”

The quick brown fox jumps over the lazy dog.

An electric car runs on electricity via
Choose an answer from this list:
- gasoline
- a power station
- electrical conductors
- fuel
Answer:

Instructable LM

Electrical conductors

...

...

Robustness

Correct the grammar error in this sentence and also fix the letter cases:
“the quick brown fox jump over the lazy dog”

The quick brown fox jumps over the lazy dog.

Instructable LM

Convert this sentence into a correct English sentence with proper letter cases:
“the quick brown fox jump over the lazy dog”

The quick brown fox jumps over the lazy dog.

Predictability

Predictability

Correct the grammar error in this sentence and also fix the letter cases:

“the quick brown fox jump over the lazy dog”

The quick brown fox jumps over the lazy dog.

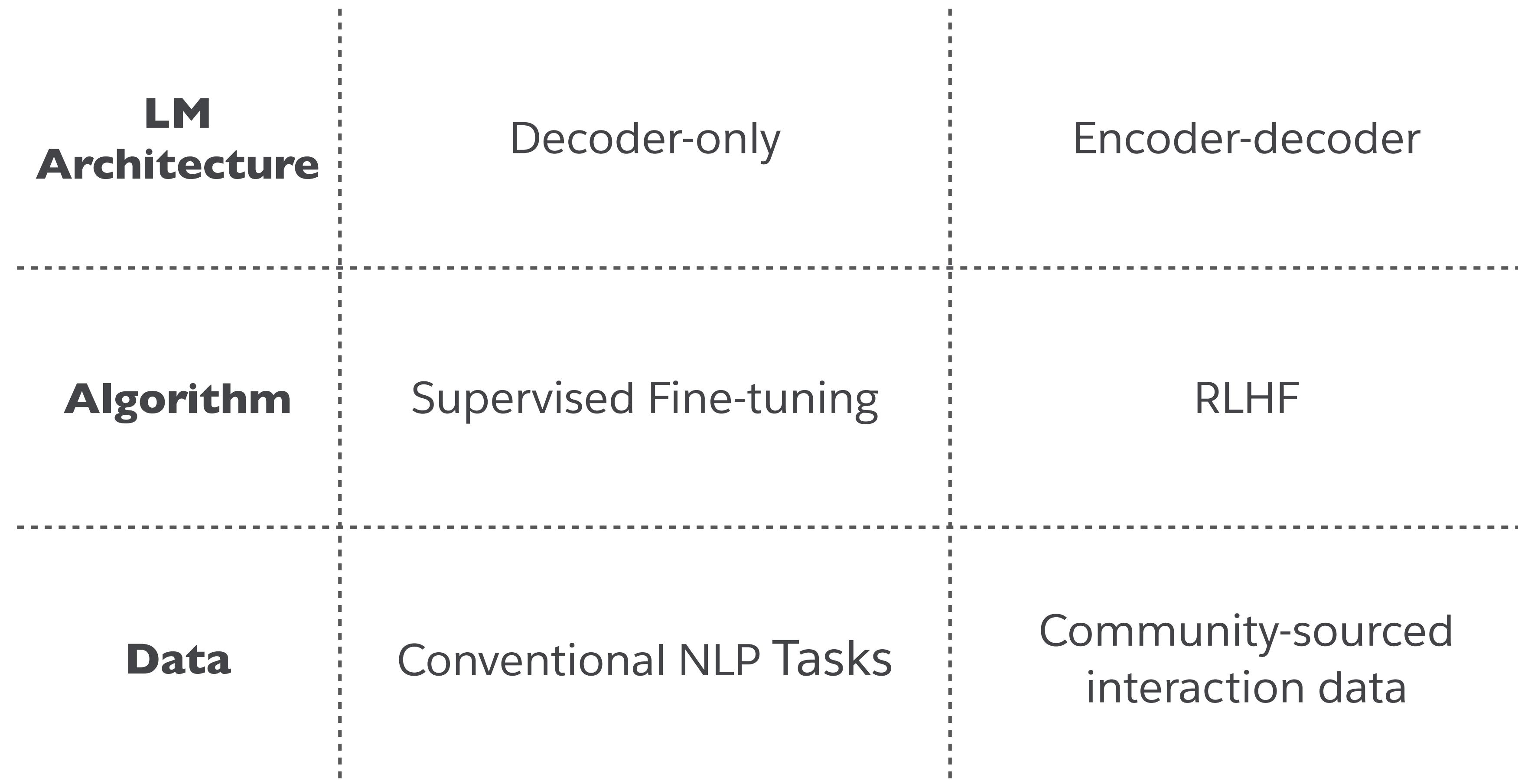
Instructable LM

Correct the grammar error in this sentence and keep the letters in their original cases:

“the quick brown fox jump over the lazy dog”

the quick brown fox jumps over the lazy dog.

Instruction Tuning Approaches



Instruction Tuning Data

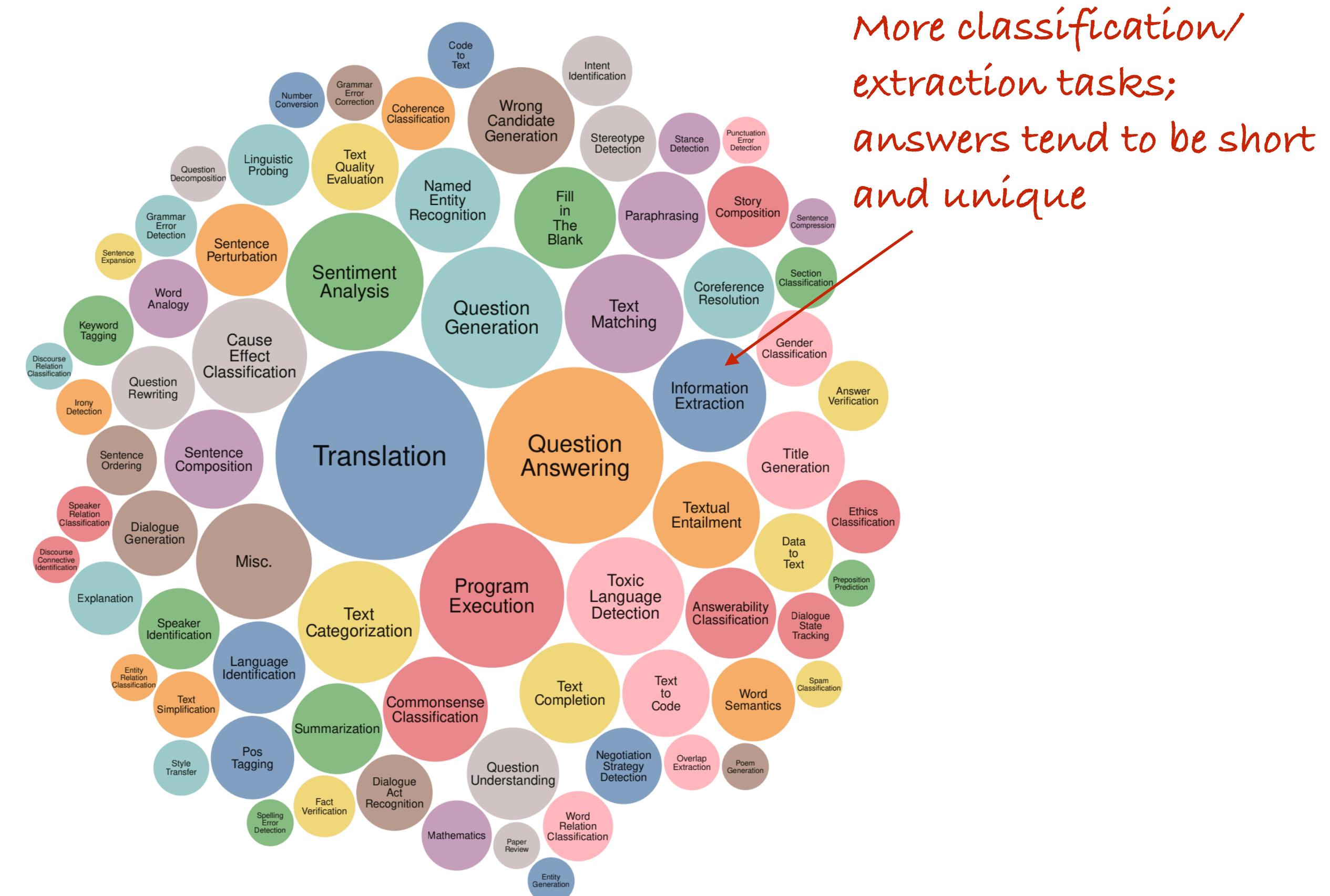
Conventional NLP Tasks



Image Credit: Distribution of 1600+ tasks in the Super-NaturalInstructions Benchmark. (Wang et al. 2022)

Instruction Tuning Data

Conventional NLP Tasks



More classification/
extraction tasks;
answers tend to be short
and unique

Image Credit: Distribution of 1600+ tasks in the Super-NaturalInstructions Benchmark. (Wang et al. 2022)

Instruction Tuning Data

Conventional NLP Tasks



More classification/
extraction tasks;
answers tend to be short
and unique

More open-ended generation and creativity-oriented tasks

InstructGPT

Crowdsourced data

Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%

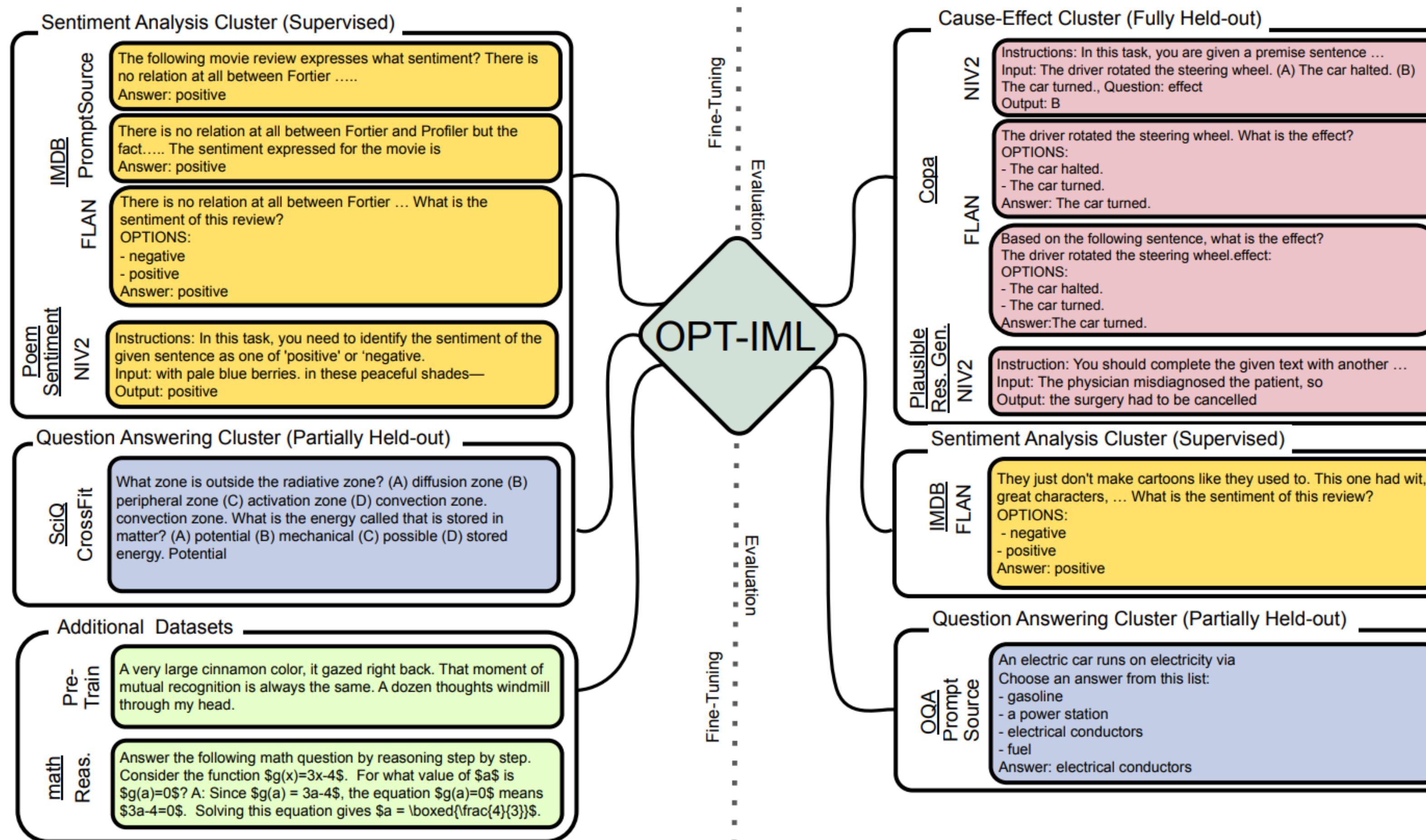
Image Credit: Distribution of 1600+ tasks in the Super-NaturalInstructions Benchmark. (Wang et al. 2022)

Table Credit: Distribution of use case categories from OpenAI text API prompt dataset. Mar 2022. (Ouyang et al. 2022)

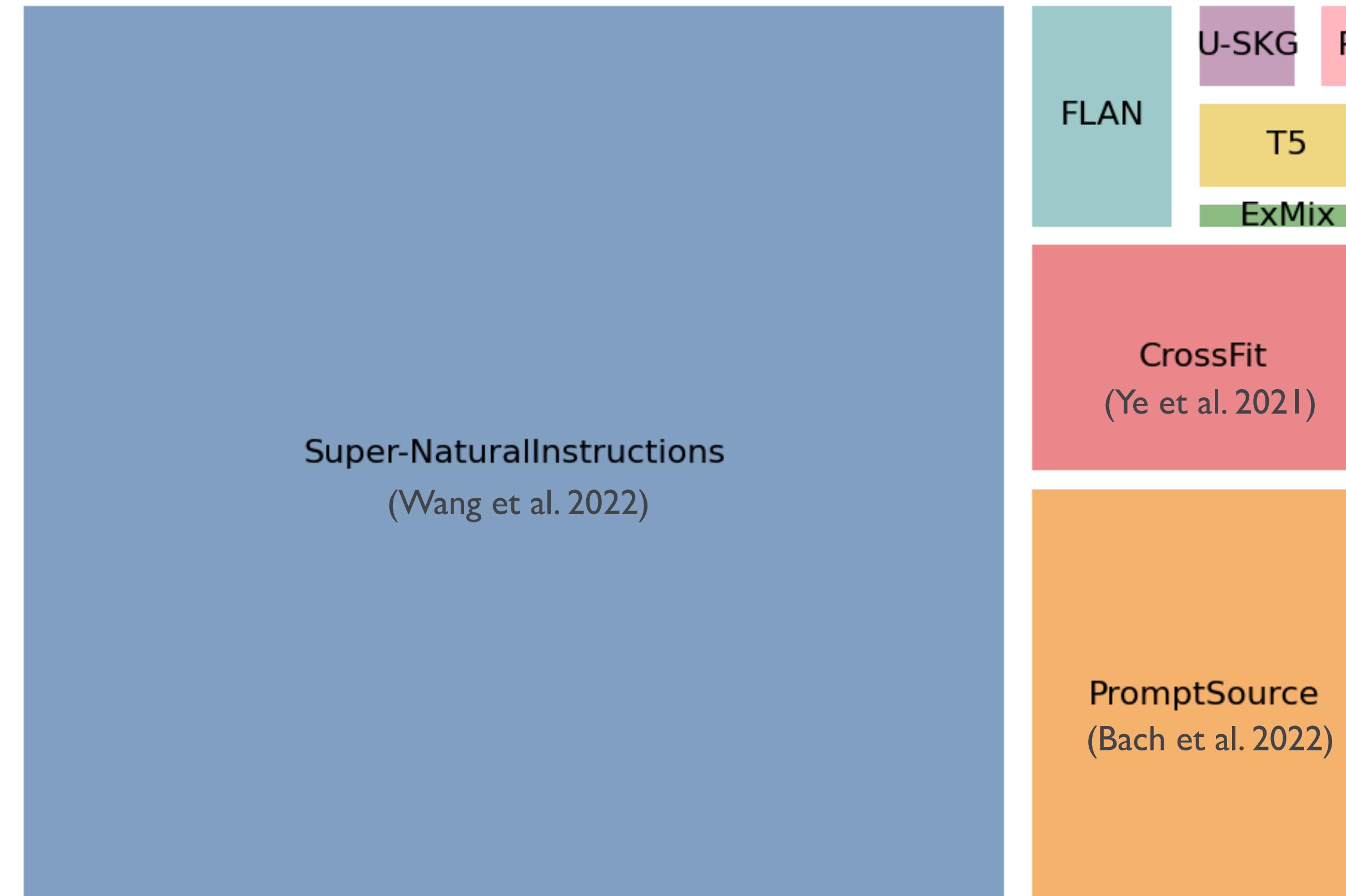
Instruction Tuning Approaches

- **Model** InstructGPT. Jan 2022. (OpenAI announced InstructGPT models as a service early 2022 but was reported to have internally developed it mid 2021.)
- **Dataset Model** NaturalInstructions. Apr 2021.
- **Model Dataset** FLAN v1. Sep 2021.
- **Model Dataset** T0. Oct 2021.
- **Model** ExT5. Nov 2021.
- **Dataset Model** Super-NaturalInstructions. Apr 2022.
- **Model Dataset** FLAN-T5 & FLAN-PaLM. Oct 2022.
- **Model Dataset** OPT-IML: Scaling Language Model Instruction Meta Learning through the Lens of Generalization. Dec 2022.
- **Approach** LIMA: Less is More for Alignment. 2023.
- ...

OPT-IML: Instruction Meta-Learning



Multi-task Instruction Tuning Data



Raw distribution over 8 multi-task NLP benchmarks based on # tasks/datasets

Multi-task Instruction Tuning Data

	Inst. Type	Instructions	Output
SuperNatInst	task-level inst.	<p>Instructions: Given a premise and two alternatives, choose the alternative that is a more plausible cause or effect of the situation described by the premise. The input format is “premise (1) alternative_1 (2) alternative_2”, the output should either be “1” or “2” based on your judgment.</p> <p>Input: The terrorist set off the bomb. (1) The bomb exploded. (2) The bomb was deactivated.</p>	1
PromptSource	instance-level inst.	<p>Exercise: choose the most plausible alternative. [Sep] The terrorist set off the bomb. so... [Sep]- The bomb exploded. [Sep]- The bomb was deactivated.</p>	The bomb exploded.
FLAN	instance-level inst.	<p>The terrorist set off the bomb. What is the effect? [Sep] OPTIONS: - The bomb exploded. - The bomb was deactivated.</p>	The bomb exploded.

OPT-IML Bench Consolidation Criteria

Training	Evaluation
<ul style="list-style-type: none">1. Diverse task types2. Diverse prompt formulation3. Enough examples per task	<ul style="list-style-type: none">1. Three levels of generalization<ul style="list-style-type: none">a. Cross-task-categoryb. Cross-datasetc. Within-dataset2. Separated valid and test splits3. Multiple task types & prompt styles per evaluation set to avoid overfitting

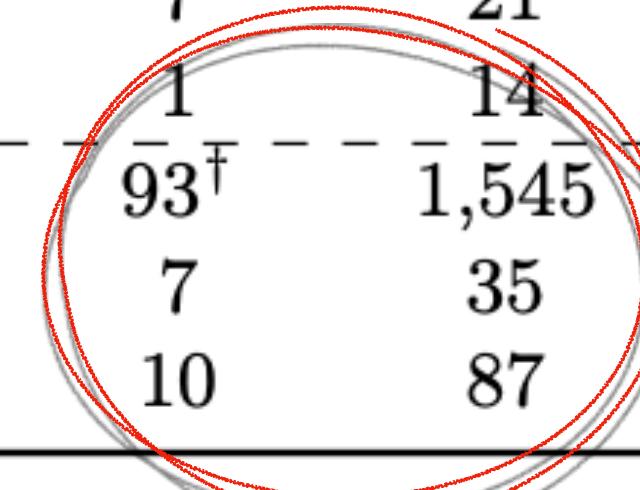
OPT-IML Bench Statistics

8 benchmarks

Benchmark	Instruct. type	# clusters	# tasks	# total examples	Avg. # prompts / task	prompt length mean	std
Super-NaturalInstructions	task inst.	76	1613	12.4M	1.0	287	882
PromptSource	instance inst.	51	280	12.8M	5.7	179	222
CrossFit	keywords	32	159	7.1M	1.0	117	258
FLAN	instance inst.	12	70	4.4M	8.5	193	375
ExMix [‡]	keywords	10	14	0.5M	1.0	132	191
T5	keywords	9	36	1.9M	1.0	111	167
UnifiedSKG	keywords	7	21	0.8M	1.0	444	297
Reasoning	task inst.	1	14	0.4M	1.0	146	122
OPT-IML Bench (train)	mixed	93 [†]	1,545	17.9M	1.7	261	631
OPT-IML Bench (dev)	mixed	7	35	145K	2.9	—	—
OPT-IML Bench (test)	mixed	10	87	321K	4.6	—	—

OPT-IML Bench Statistics

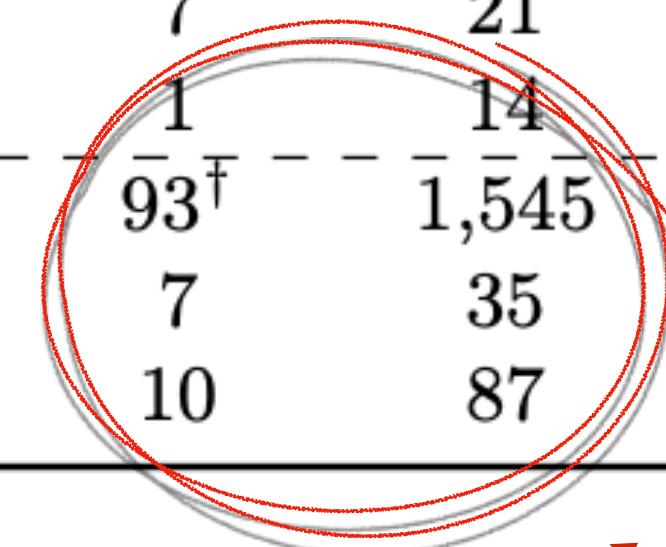
Benchmark	Instruct. type	# clusters	# tasks	# total examples	Avg. # prompts / task	prompt length mean	prompt length std
Super-NaturalInstructions	task inst.	76	1613	12.4M	1.0	287	882
PromptSource	instance inst.	51	280	12.8M	5.7	179	222
CrossFit	keywords	32	159	7.1M	1.0	117	258
FLAN	instance inst.	12	70	4.4M	8.5	193	375
ExMix [‡]	keywords	10	14	0.5M	1.0	132	191
T5	keywords	9	36	1.9M	1.0	111	167
UnifiedSKG	keywords	7	21	0.8M	1.0	444	297
Reasoning	task inst.	1	14	0.4M	1.0	146	122
OPT-IML Bench (train)	mixed	93 [†]	1,545	17.9M	1.7	261	631
OPT-IML Bench (dev)	mixed	7	35	145K	2.9	—	—
OPT-IML Bench (test)	mixed	10	87	321K	4.6	—	—



Diverse train/eval tasks
and clusters

OPT-IML Bench Statistics

Benchmark	Instruct. type	# clusters	# tasks	# total examples	Avg. # prompts / task	prompt length mean	prompt length std
Super-NaturalInstructions	task inst.	76	1613	12.4M	1.0	287	882
PromptSource	instance inst.	51	280	12.8M	5.7	179	222
CrossFit	keywords	32	159	7.1M	1.0	117	258
FLAN	instance inst.	12	70	4.4M	8.5	193	375
ExMix [‡]	keywords	10	14	0.5M	1.0	132	191
T5	keywords	9	36	1.9M	1.0	111	167
UnifiedSKG	keywords	7	21	0.8M	1.0	444	297
Reasoning	task inst.	1	14	0.4M	1.0	146	122
OPT-IML Bench (train)	mixed	93 [†]	1,545	17.9M	1.7	261	631
OPT-IML Bench (dev)	mixed	7	35	145K	2.9	—	—
OPT-IML Bench (test)	mixed	10	87	321K	4.6	—	—



Specially designed cluster split that allows measuring of generalization at three levels

OPT-IML Bench Statistics

Benchmark	Instruct. type	# clusters	# tasks	# total examples	Avg. # prompts / task	prompt length mean	prompt length std
Super-NaturalInstructions	task inst.	76	1613	12.4M	1.0	287	882
PromptSource	instance inst.	51	280	12.8M	5.7	179	222
CrossFit	keywords	32	159	7.1M	1.0	117	258
FLAN	instance inst.	12	70	4.4M	8.5	193	375
ExMix [‡]	keywords	10	14	0.5M	1.0	132	191
T5	keywords	9	36	1.9M	1.0	111	167
UnifiedSKG	keywords	7	21	0.8M	1.0	444	297
Reasoning	task inst.	1	14	0.4M	1.0	146	122
OPT-IML Bench (train)	mixed	93 [†]	1,545	17.9M	1.7	261	631
OPT-IML Bench (dev)	mixed	7	35	145K	2.9	—	—
OPT-IML Bench (test)	mixed	10	87	321K	4.6	—	—

Multiple prompts

OPT-IML Bench Statistics

Benchmark	Instruct. type	# clusters	# tasks	# total examples	Avg. # prompts / task	prompt length mean	prompt length std
Super-NaturalInstructions	task inst.	76	1613	12.4M	1.0	287	882
PromptSource	instance inst.	51	280	12.8M	5.7	179	222
CrossFit	keywords	32	159	7.1M	1.0	117	258
FLAN	instance inst.	12	70	4.4M	8.5	193	375
ExMix [‡]	keywords	10	14	0.5M	1.0	132	191
T5	keywords	9	36	1.9M	1.0	111	167
UnifiedSKG	keywords	7	21	0.8M	1.0	444	297
Reasoning	task inst.	1	14	0.4M	1.0	146	122
OPT-IML Bench (train)	mixed	93 [†]	1,545	17.9M	1.7	261	631
OPT-IML Bench (dev)	mixed	7	35	145K	2.9	—	—
OPT-IML Bench (test)	mixed	10	87	321K	4.6	—	—

Data statistics varies
significantly across
benchmarks

Experiment Setup

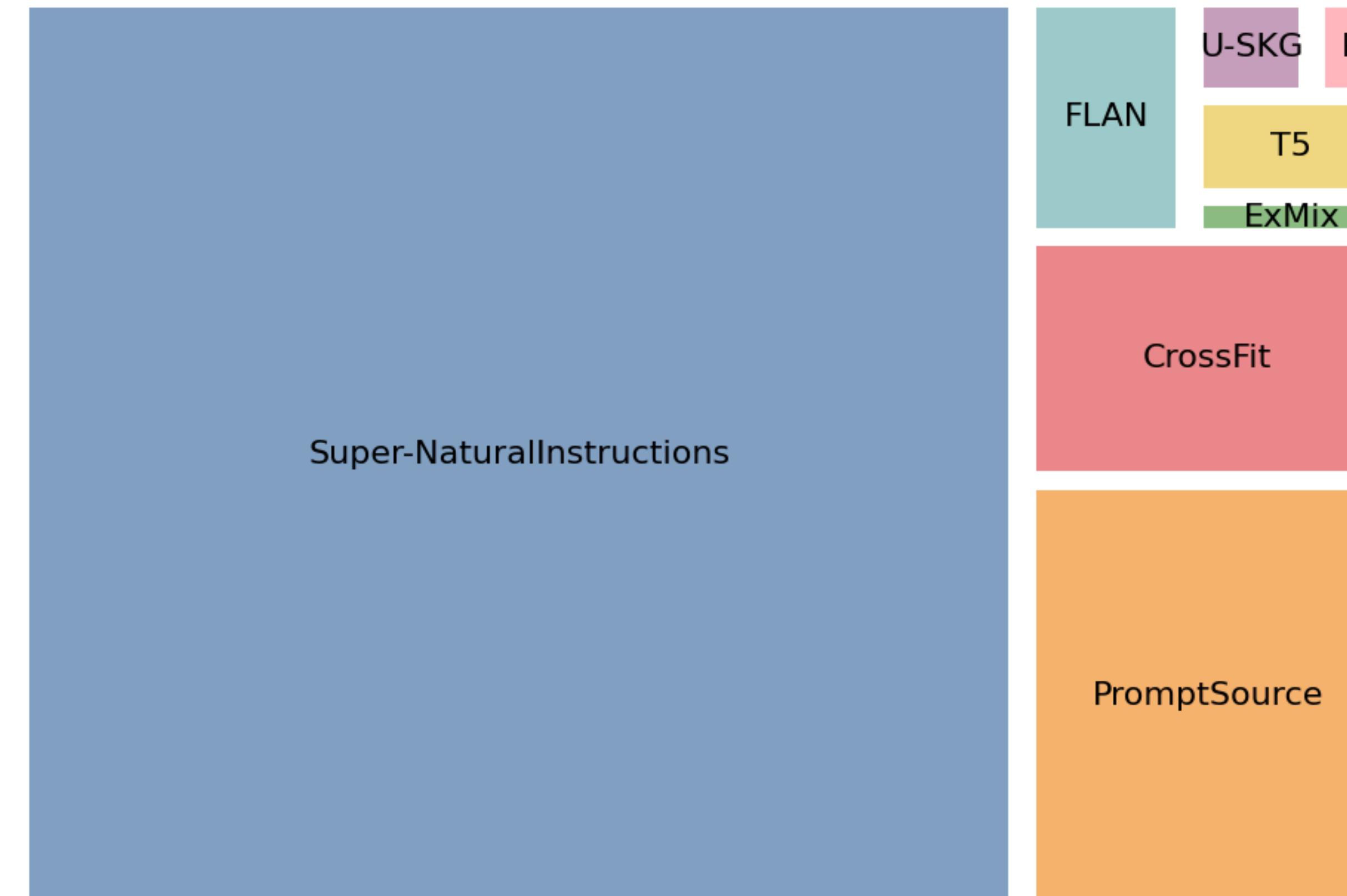
Fine-tuning strategy exploration

- 30B OPT
- Full model fine-tuning
- **Label loss:** LM objective computed over only the output section of a task prompt
- **Sequence packing:** Up to 2,048 tokens and mask attention across example boundaries



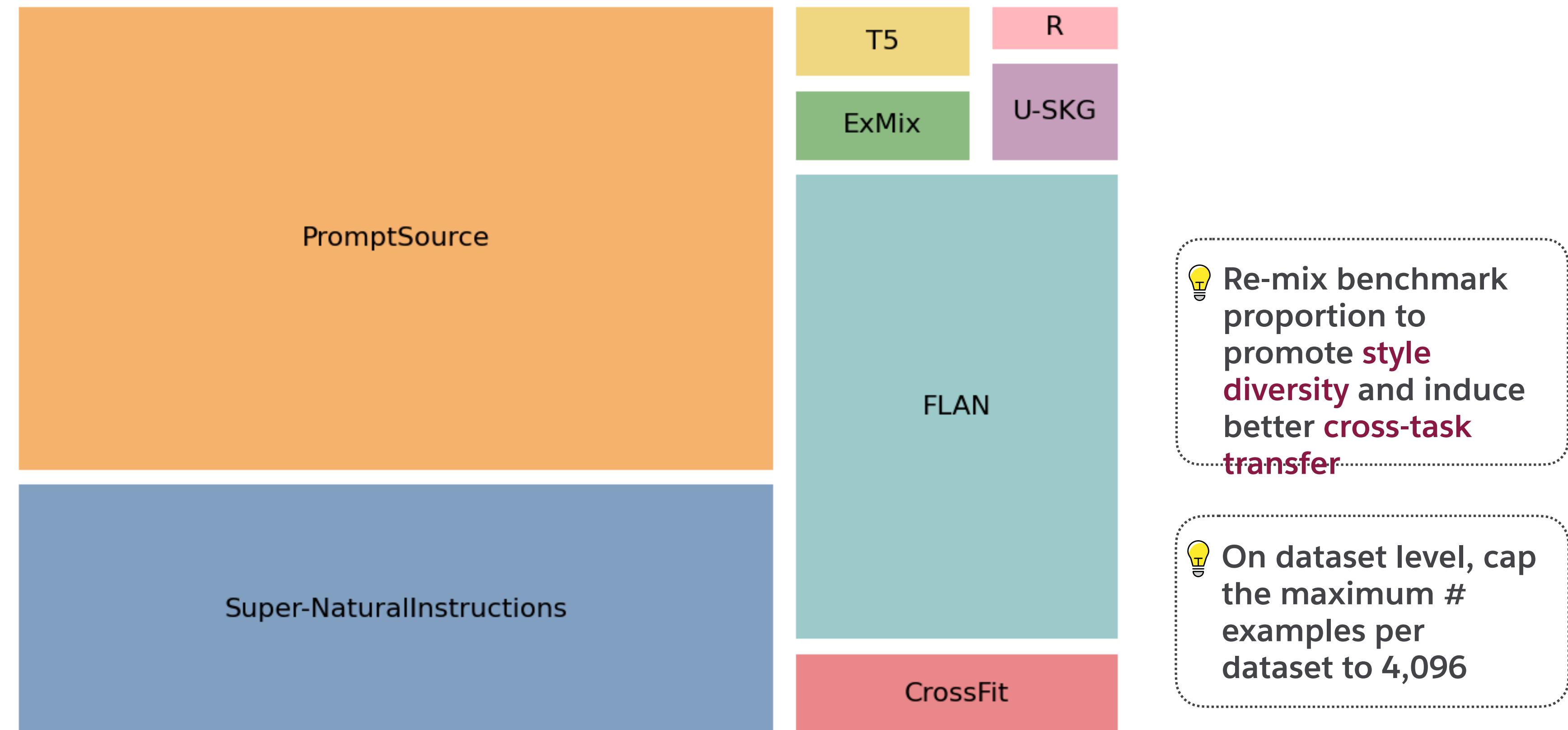
- Fine-tuning for approximately 2B tokens, < 1% of the OPT pre-training data
- Report results on OPT-IML Bench validation set

Increasing Task Diversity: Benchmark Re-mix



Raw distribution over 8 multi-task NLP benchmarks based on # tasks/datasets

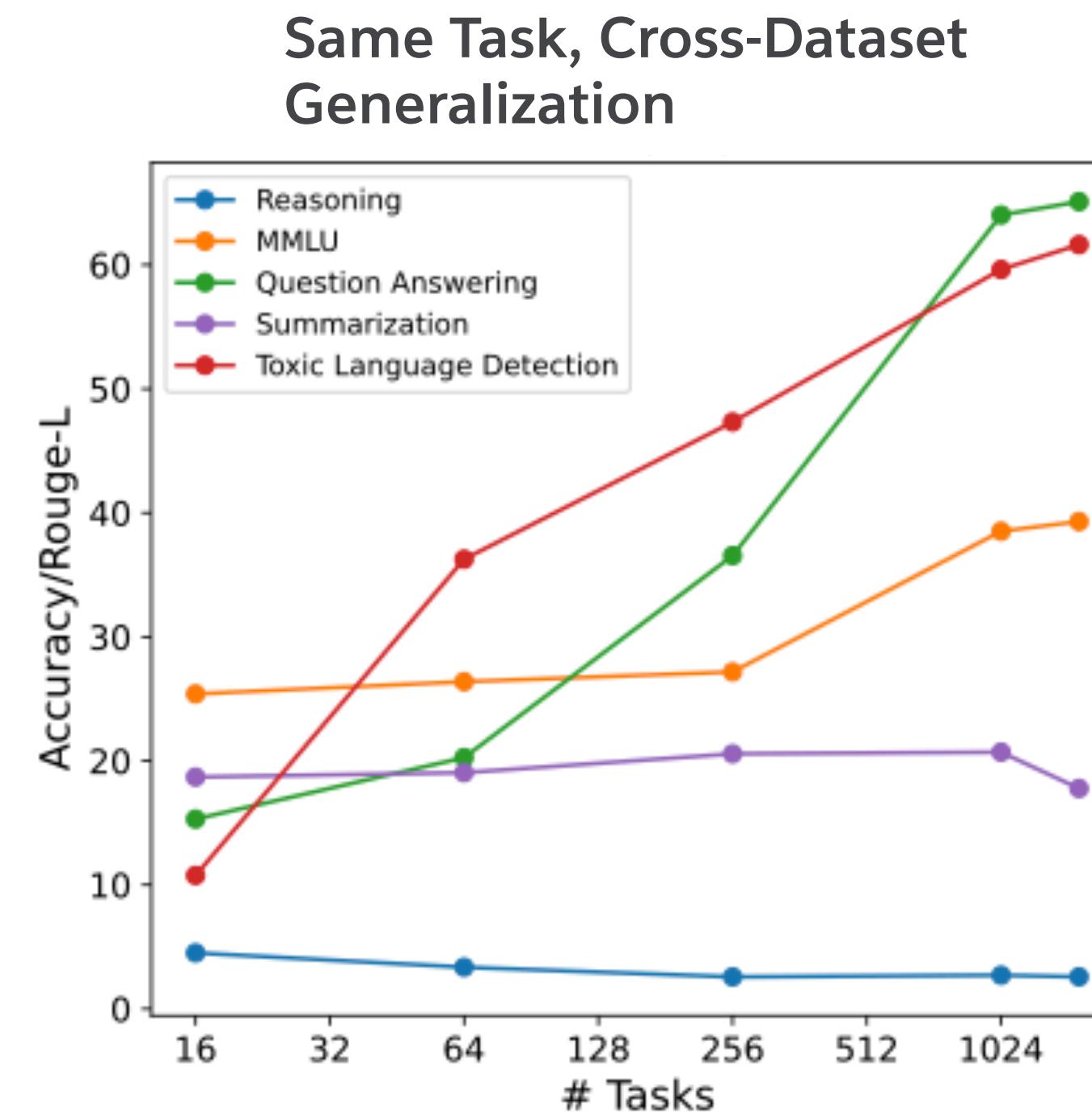
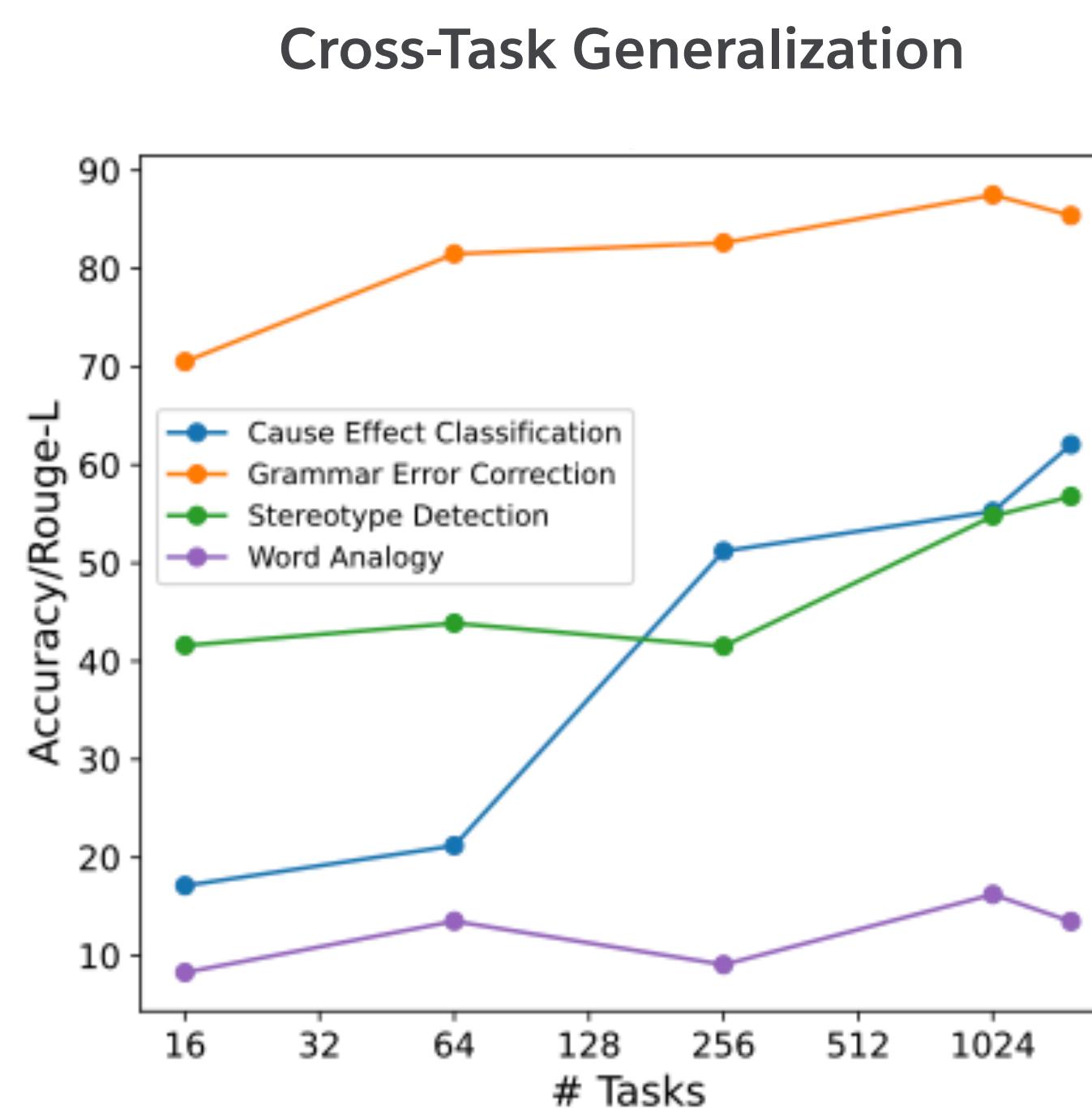
Increasing Task Diversity: Benchmark Re-mix



OPT-IML Bench adjusted distribution over 8 multi-task NLP benchmarks based on # tasks/datasets

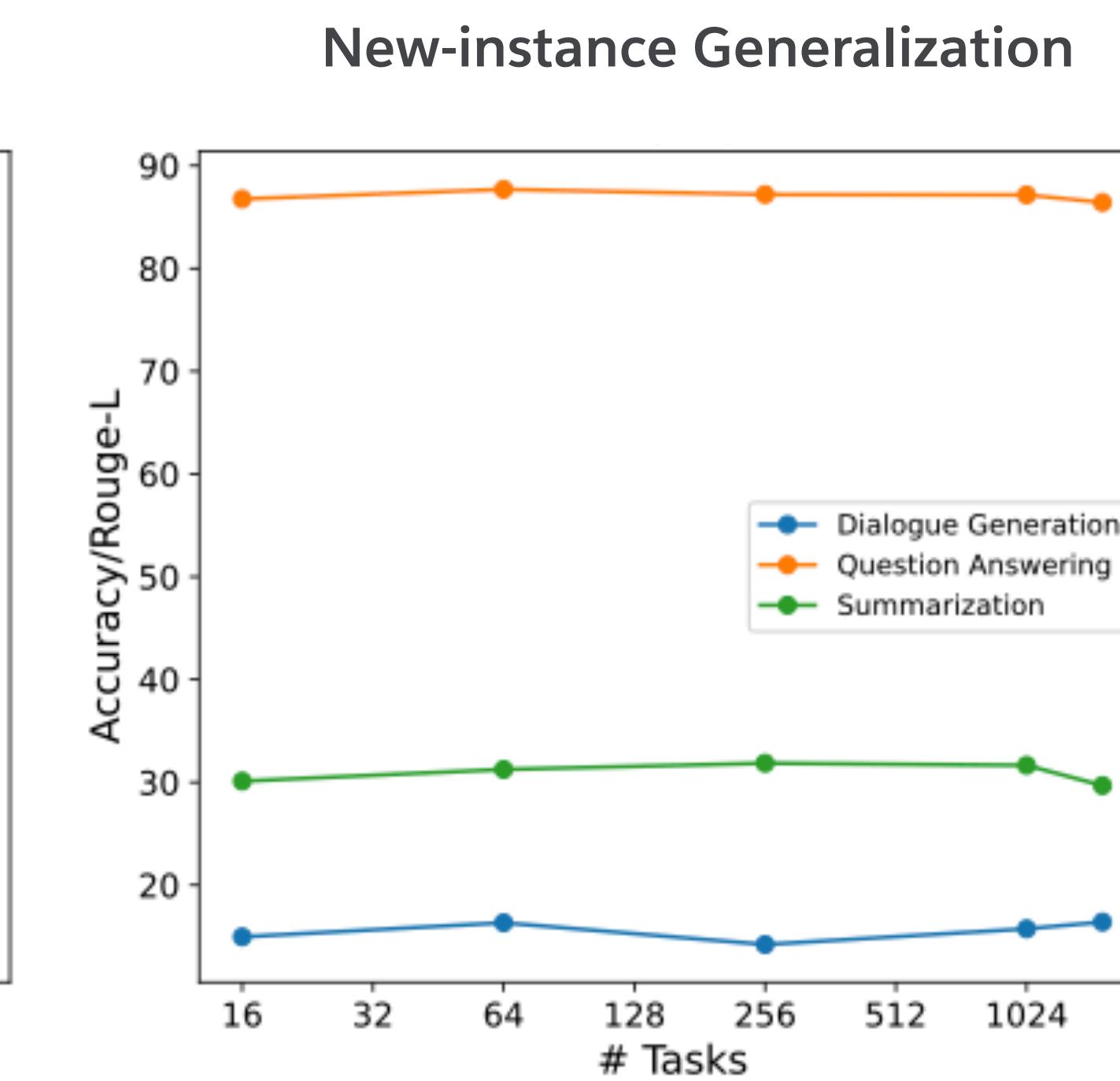
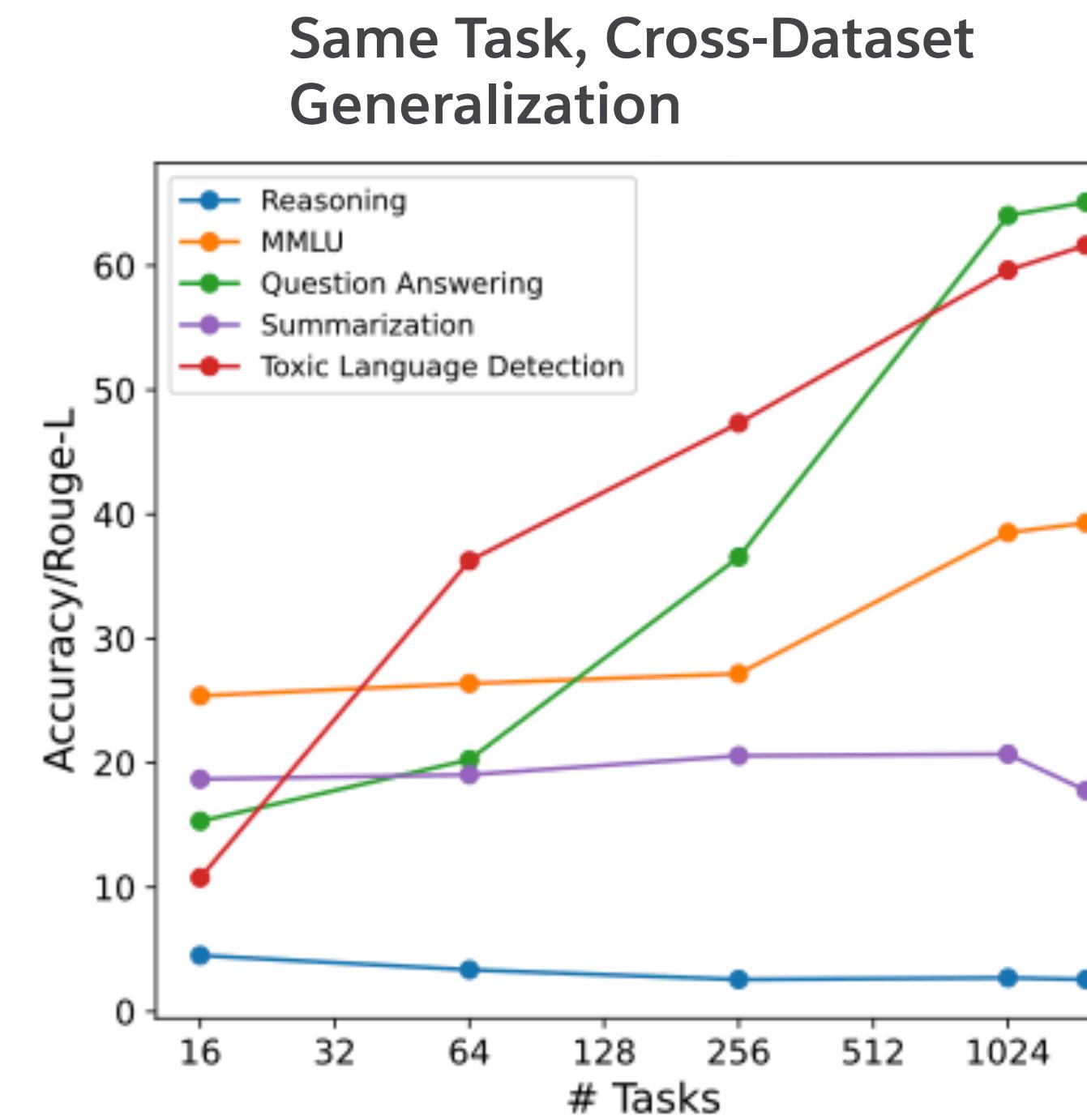
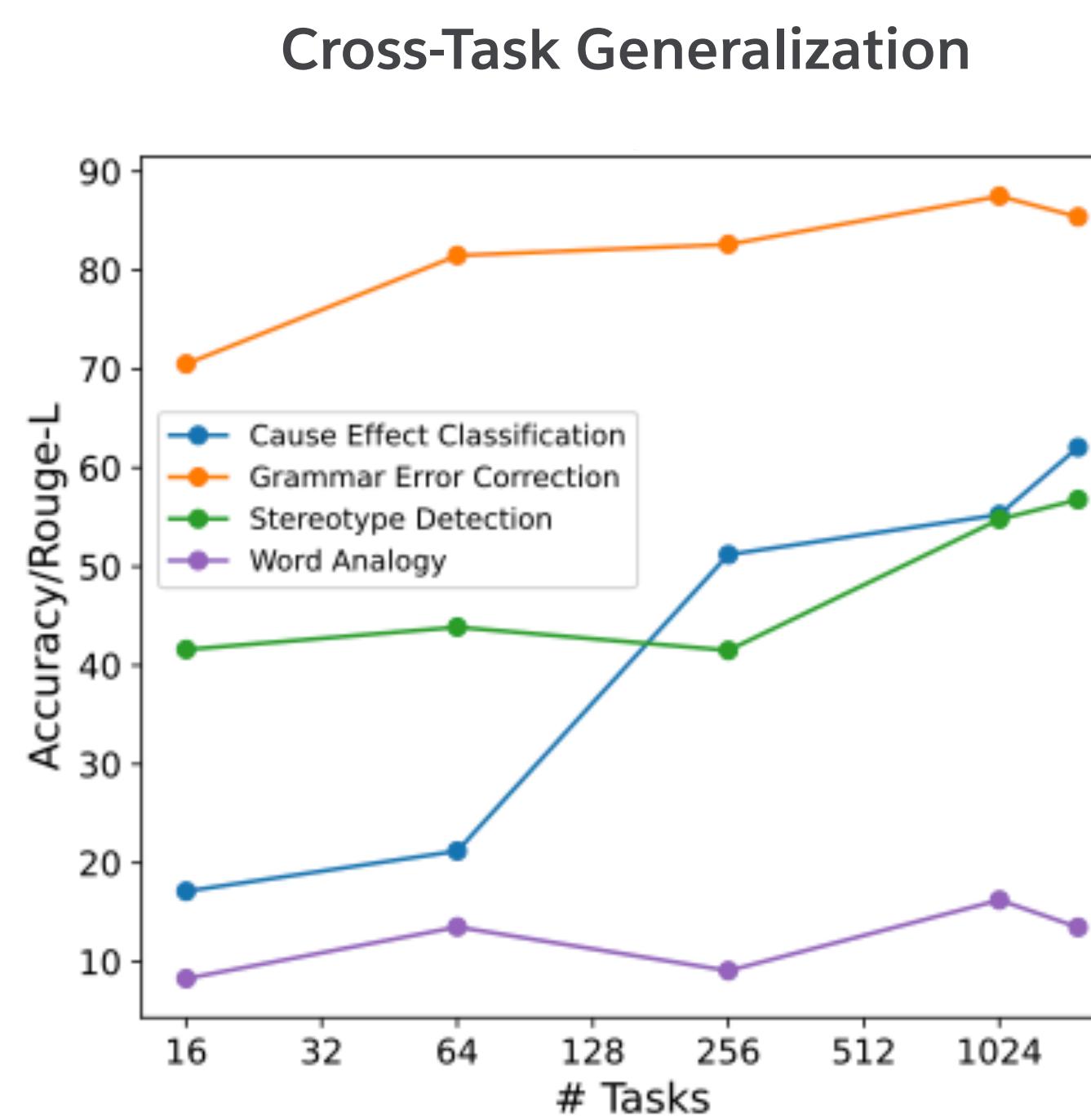
Increasing Task

- Starting from an initial set of 16 tasks which covers our fully supervised eval tasks, randomly sample $K = [16, 64, 256, 1024, 1,545]$ tasks s.t. the bigger sets are supersets of the smaller ones
- Training for the same # steps across all models



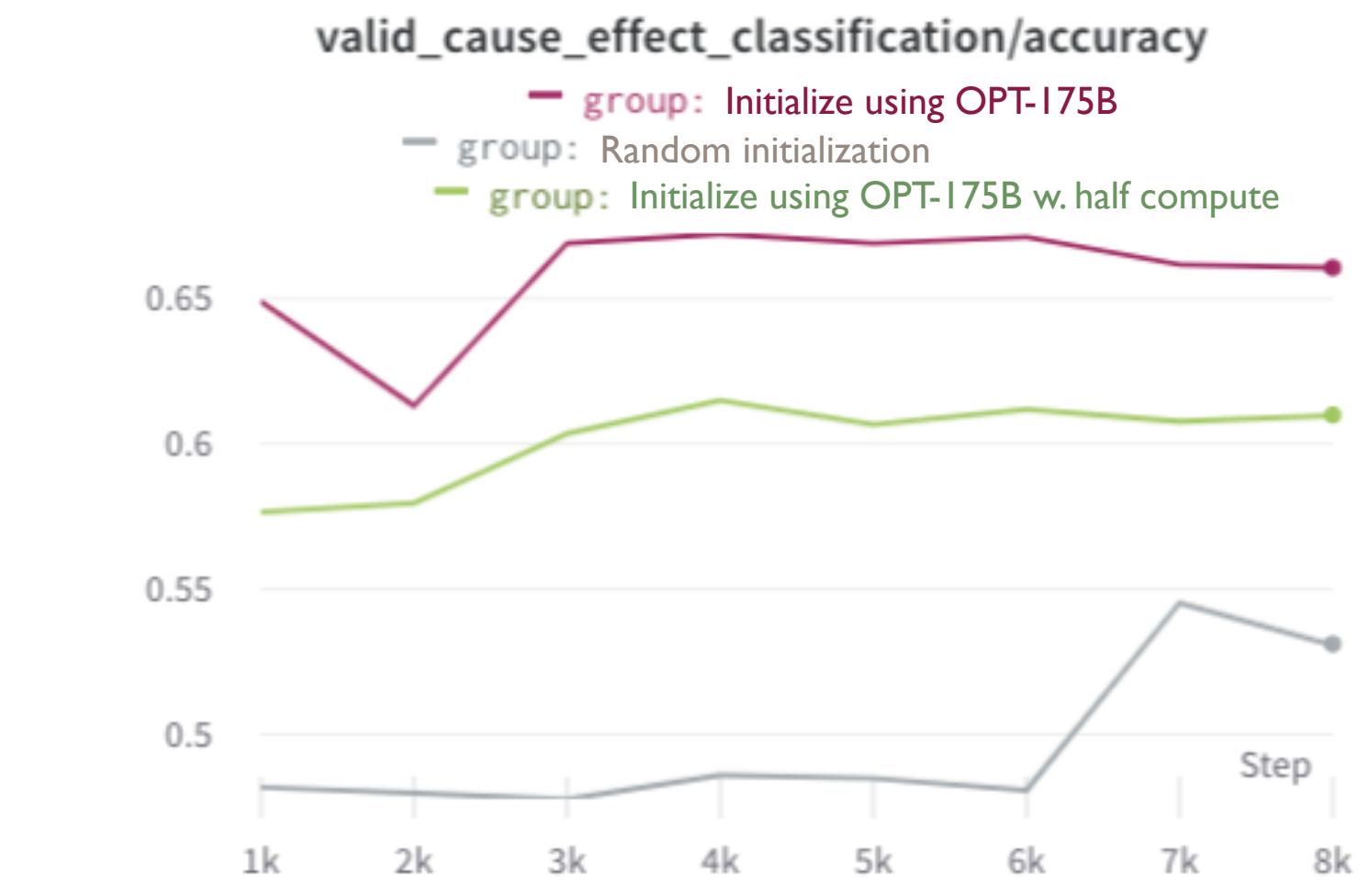
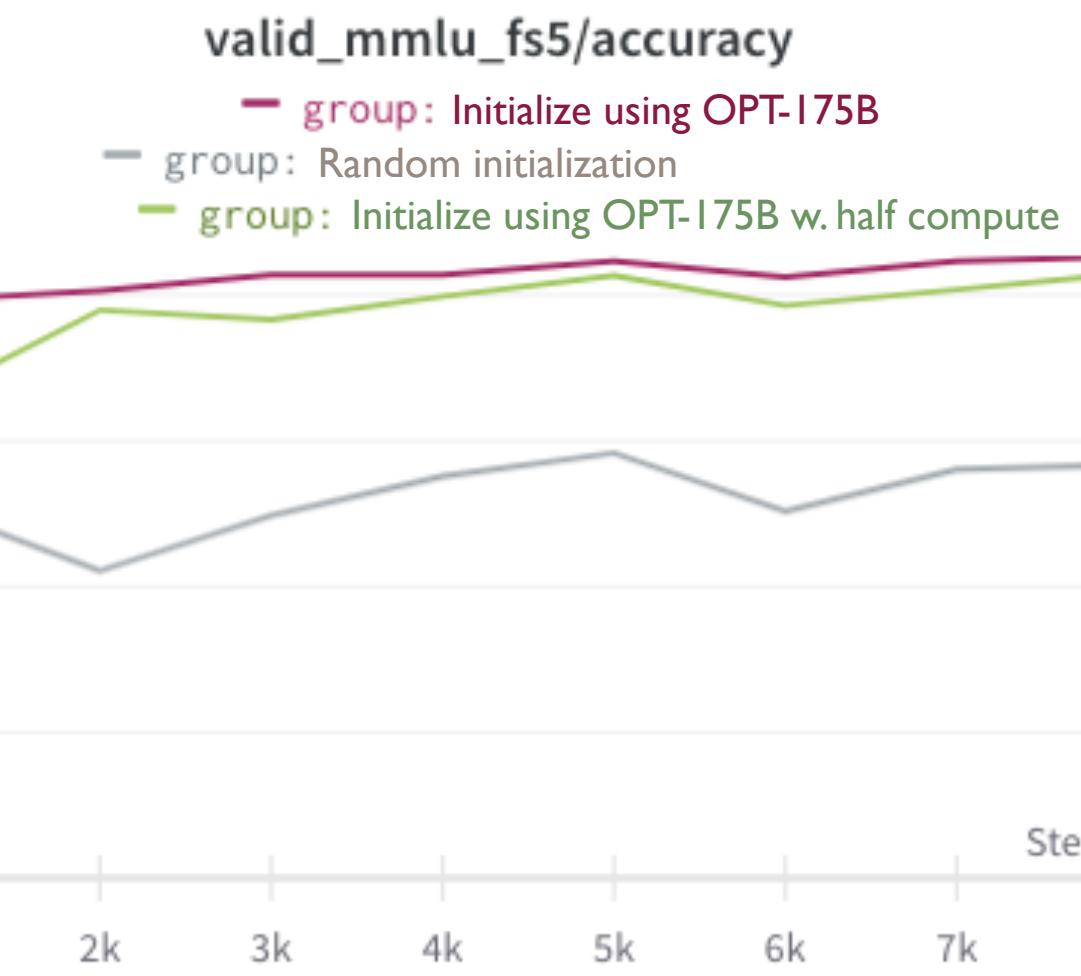
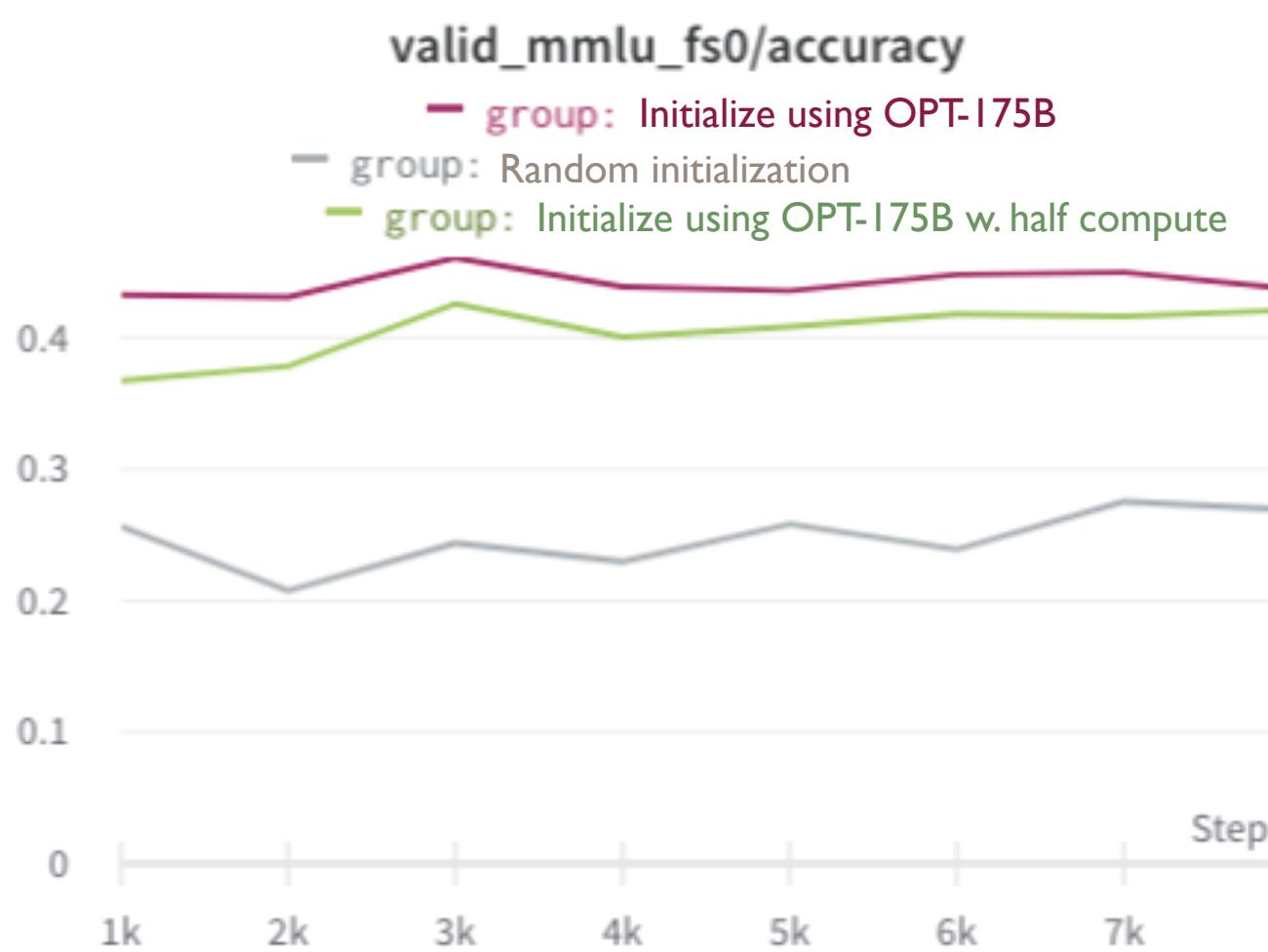
Increasing Task

- Starting from an initial set of 16 tasks which covers our fully supervised eval tasks, randomly sample $K = [16, 64, 256, 1024, 1,545]$ tasks s.t. the bigger sets are supersets of the smaller ones
- Training for the same # steps across all models



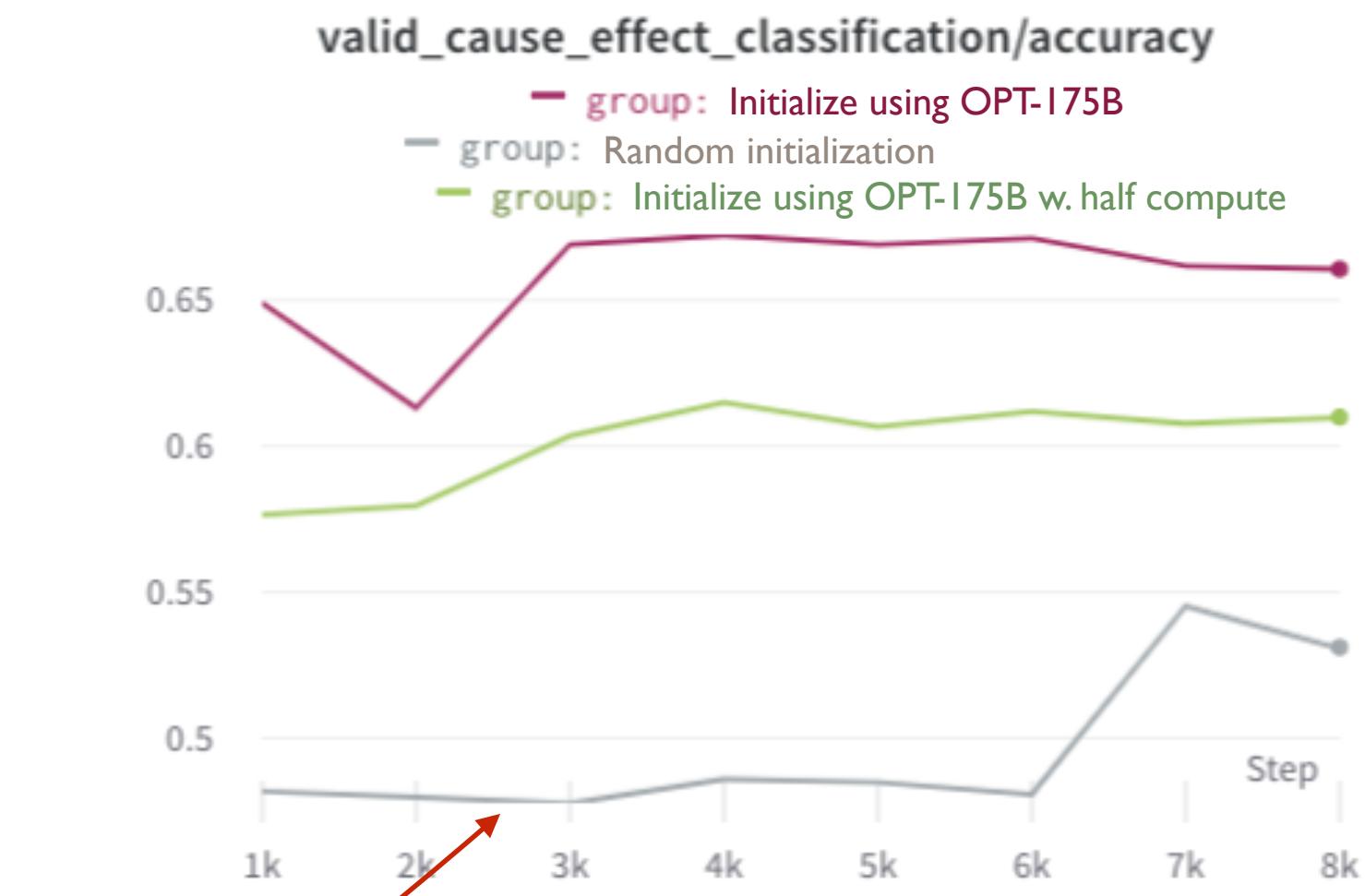
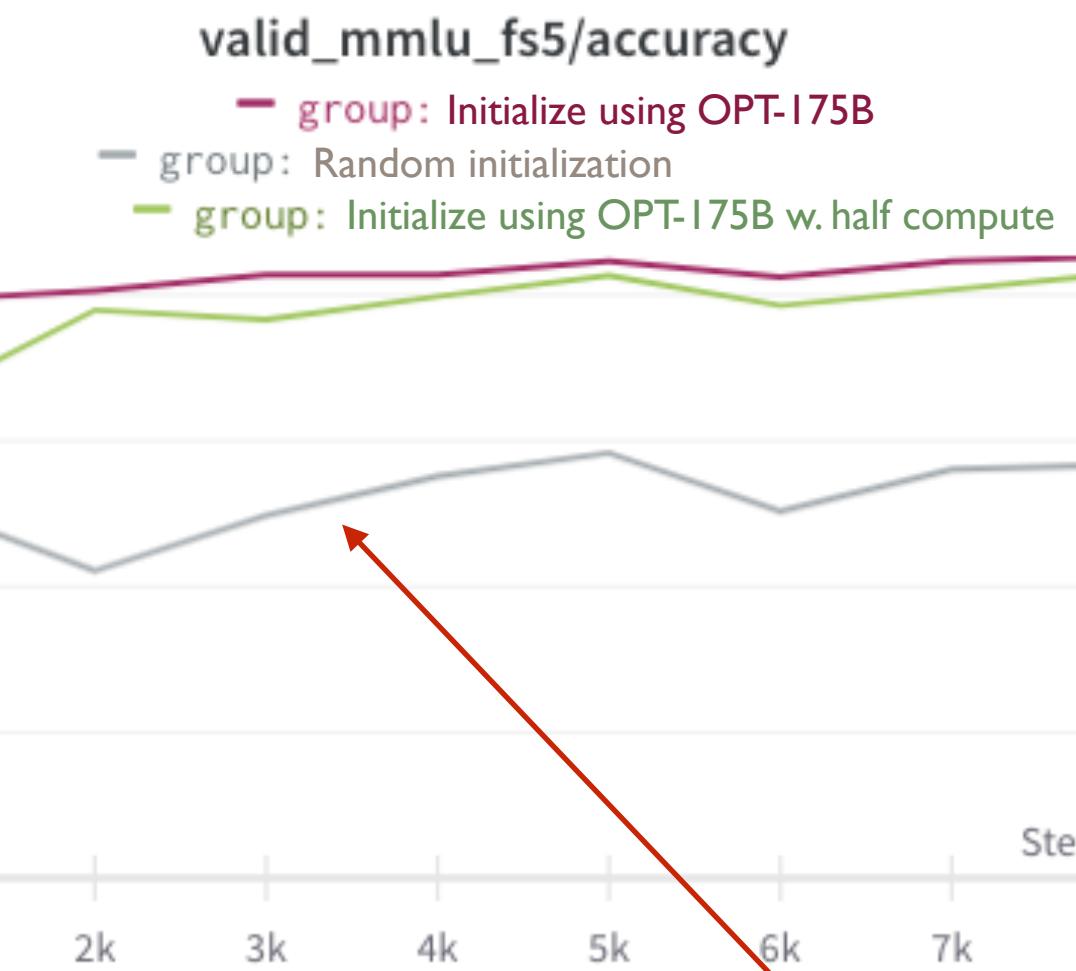
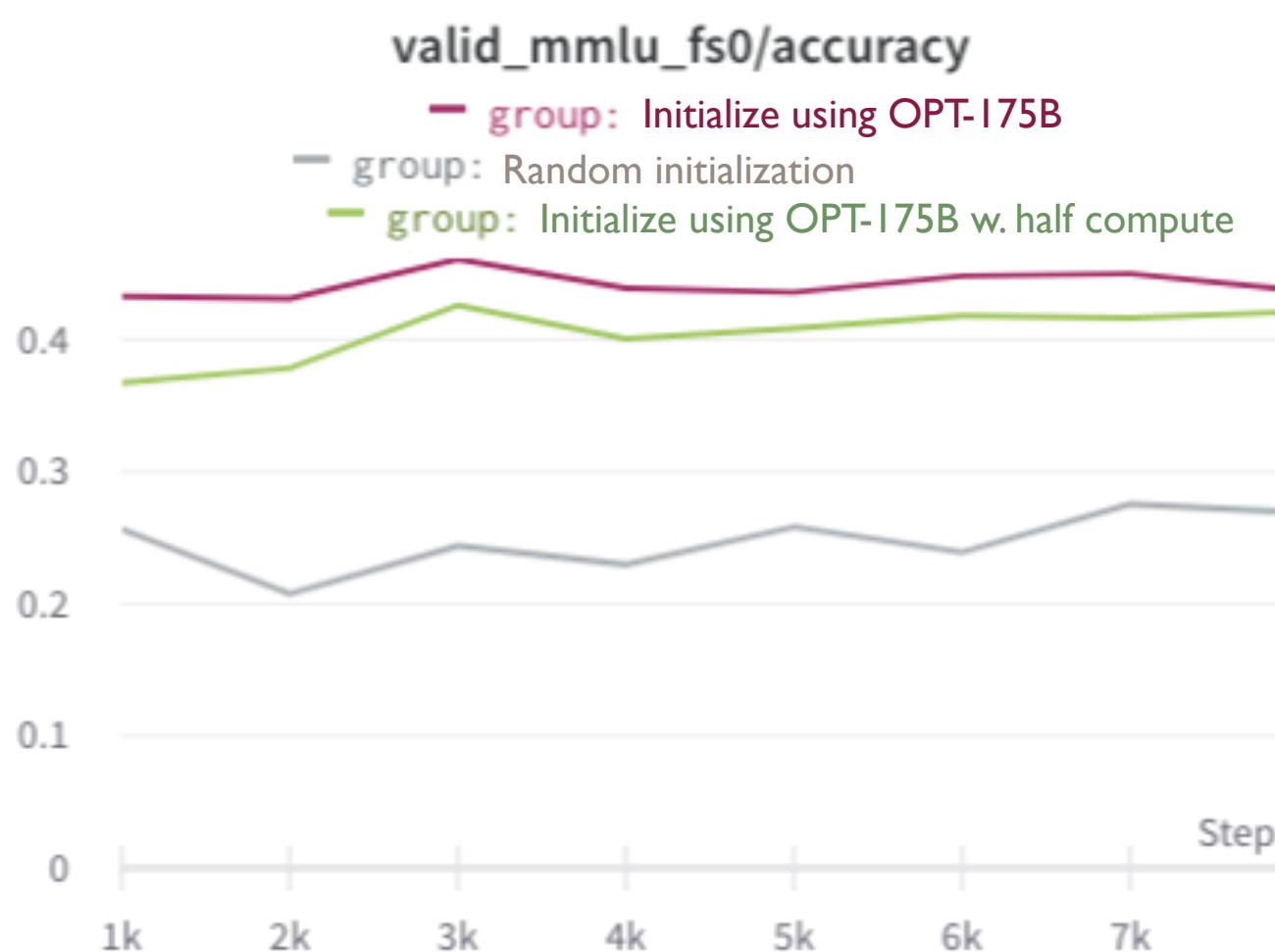
Pre-training vs. Instruction Tuning

- Starting instruction-tuning from 1) random initialization, 2) OPT 175B intermediate checkpoint trained over half the amount of compute, 3) OPT 175B



Pre-training vs. Instruction Tuning

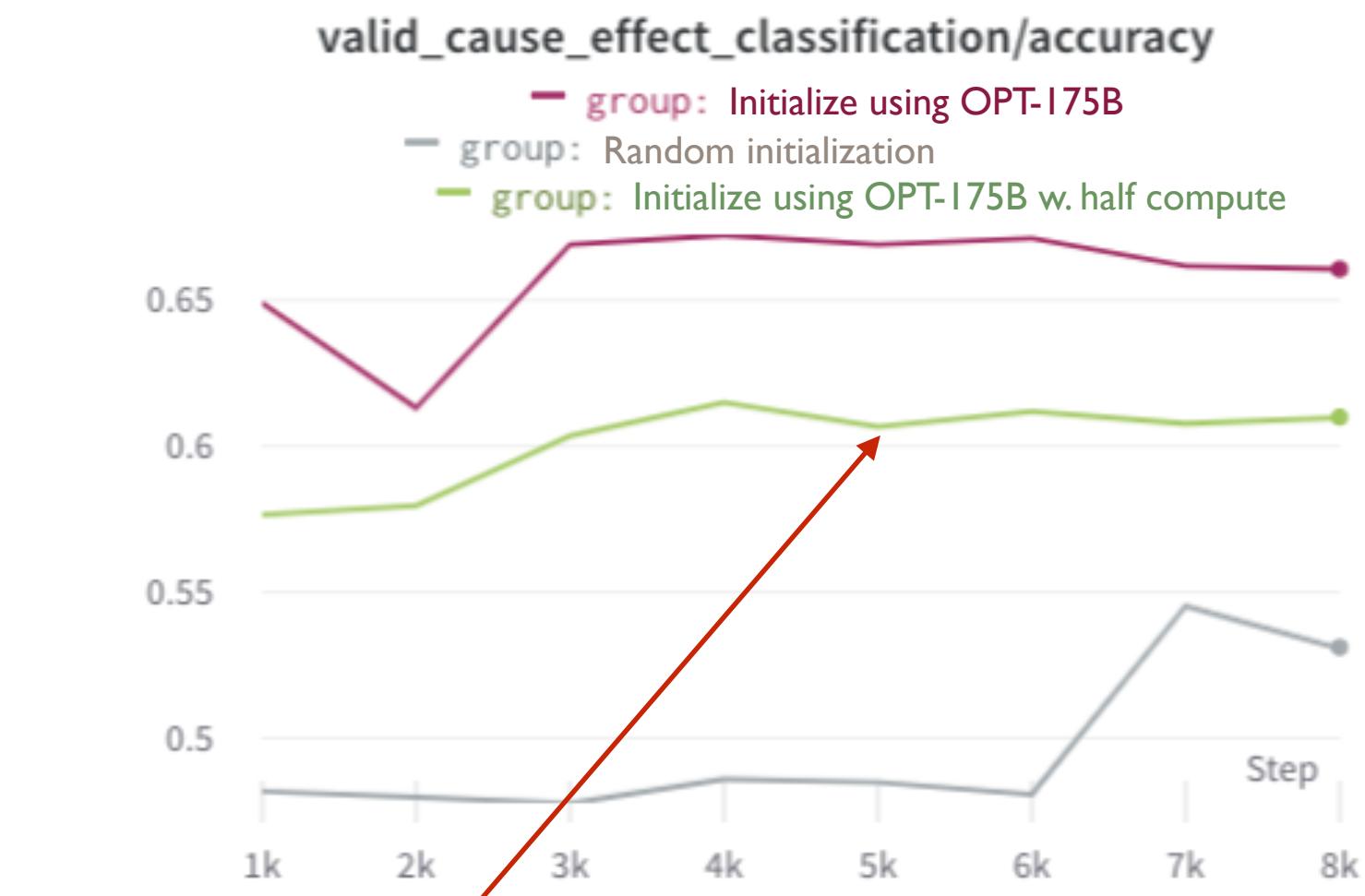
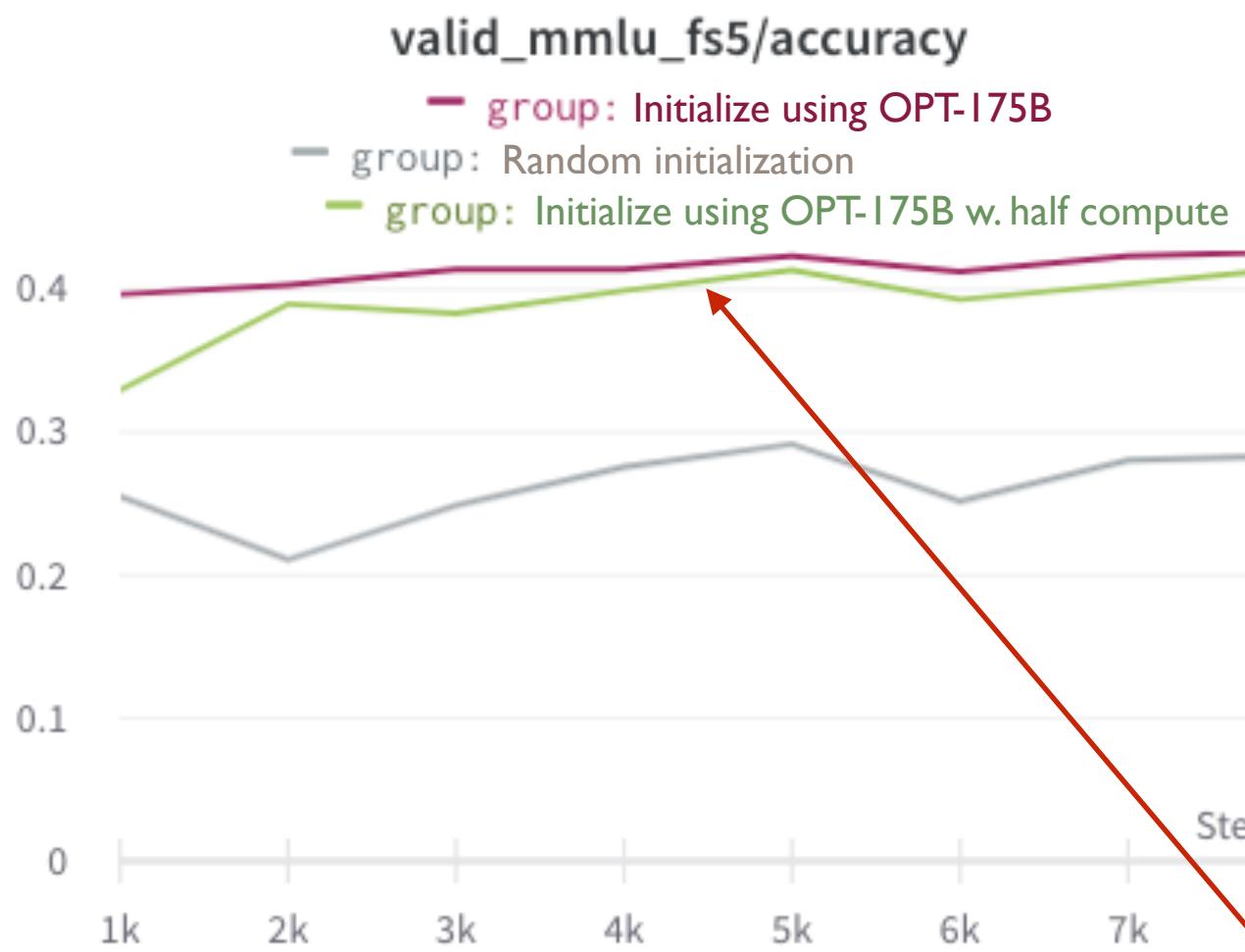
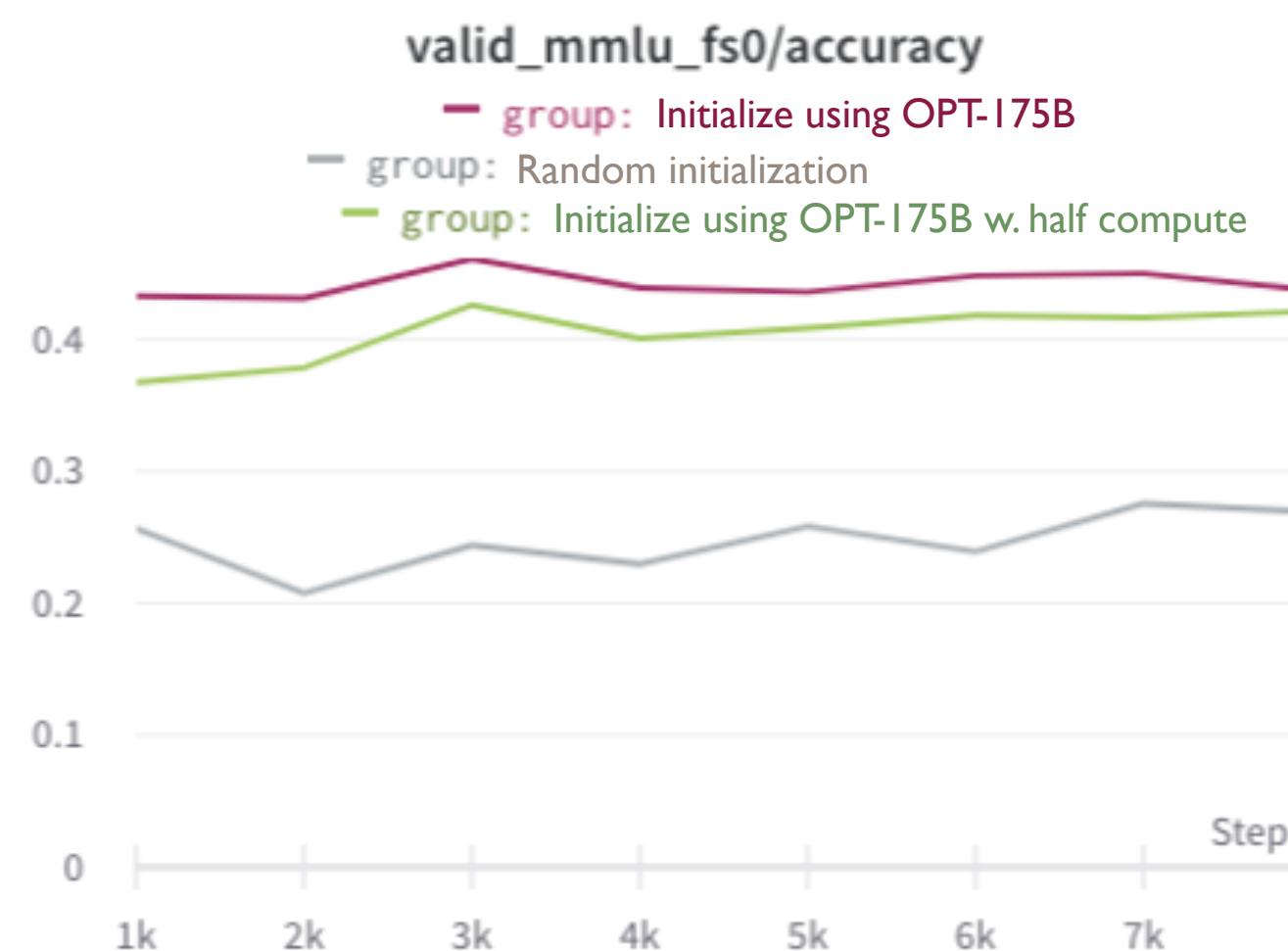
- Starting instruction-tuning from 1) random initialization, 2) OPT 175B intermediate checkpoint trained over half the amount of compute, 3) OPT 175B



Multi-task instruction-tuning alone does not lead to good performance; model performance is close to random

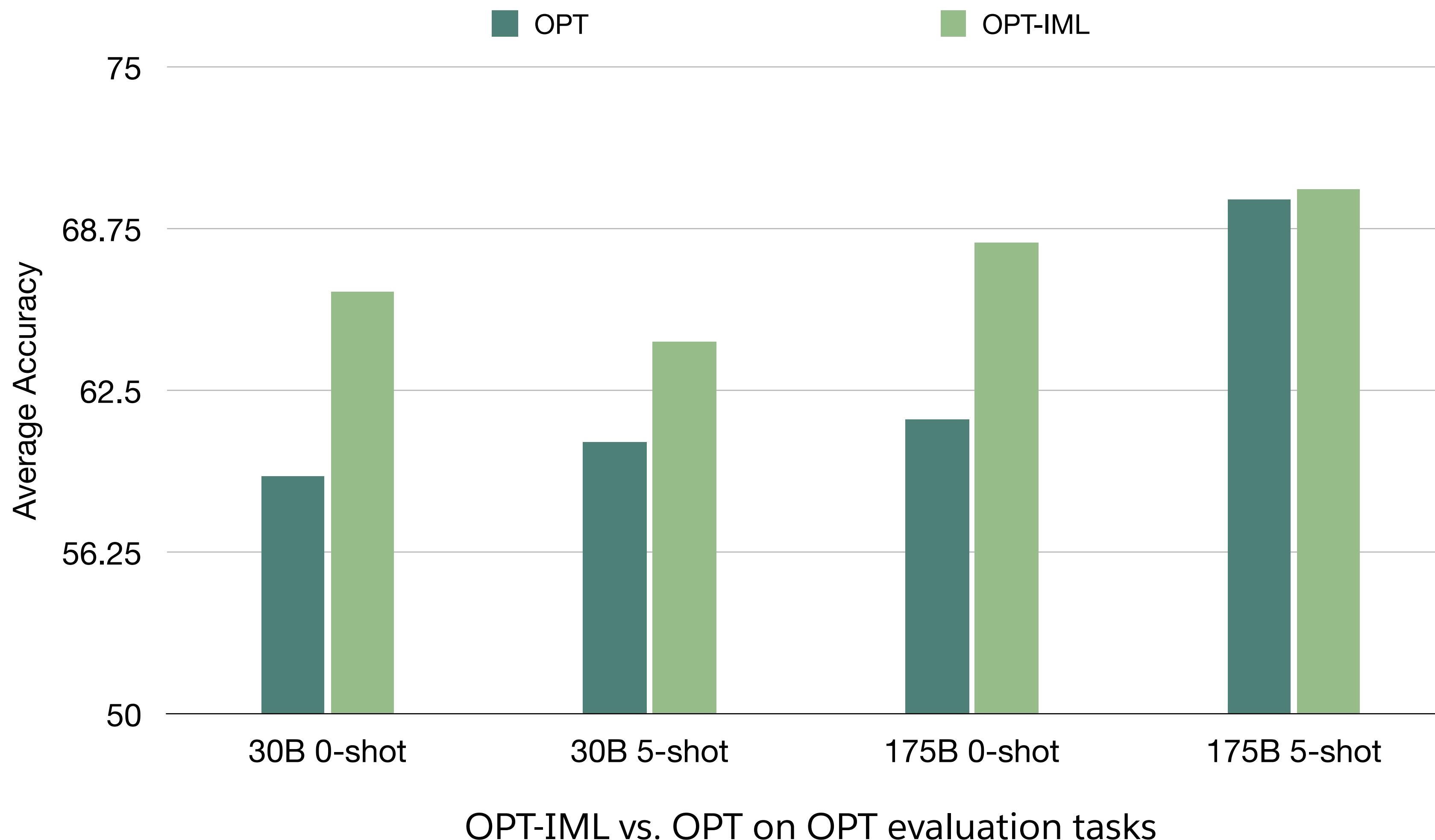
Pre-training vs. Instruction Tuning

- Starting instruction-tuning from 1) random initialization, 2) OPT 175B intermediate checkpoint trained over half the amount of compute, 3) OPT 175B

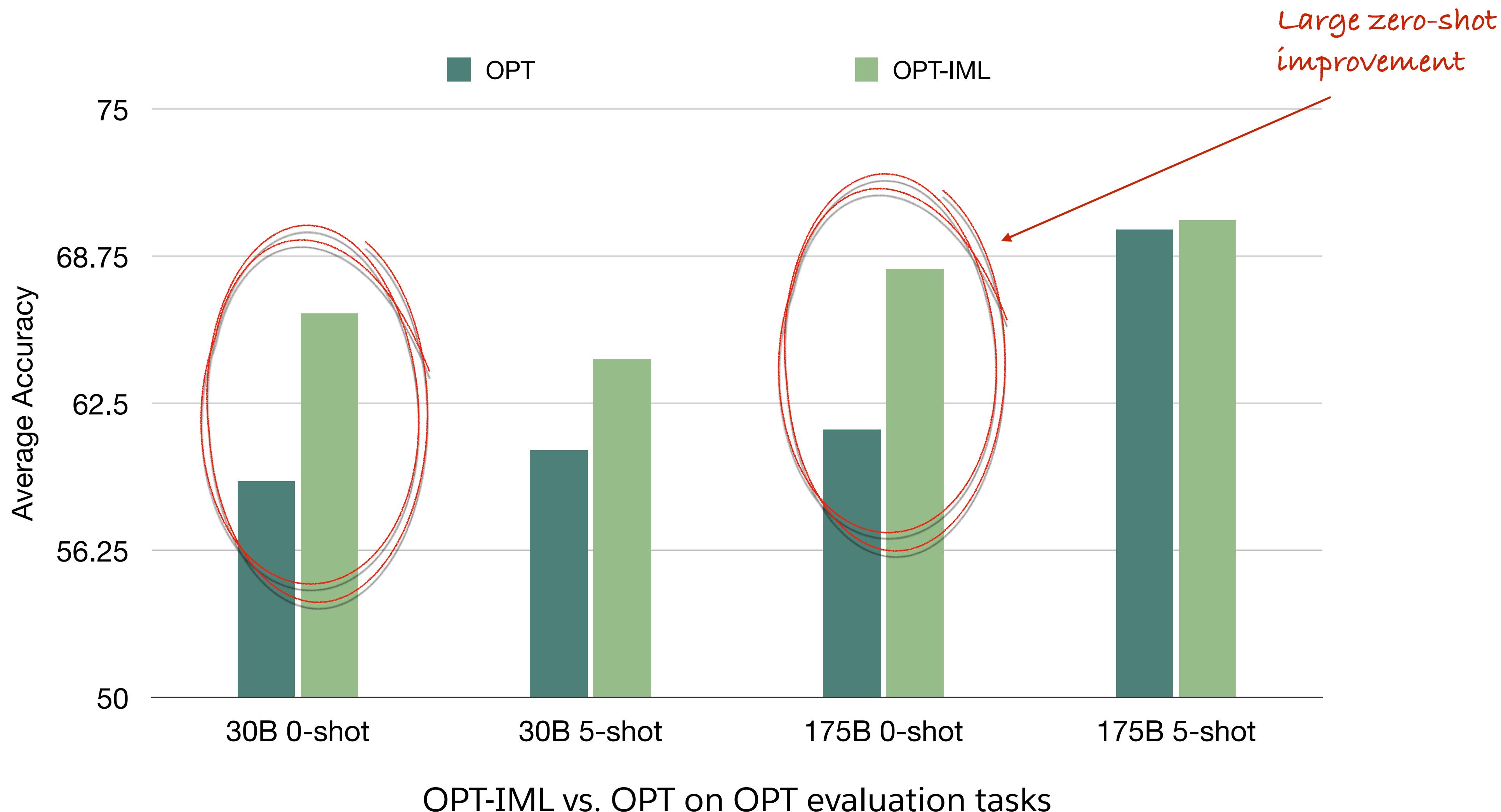


Half-way trained OPT checkpoint offers a significant boost, although the performance is worse than initializing from the final OPT checkpoint

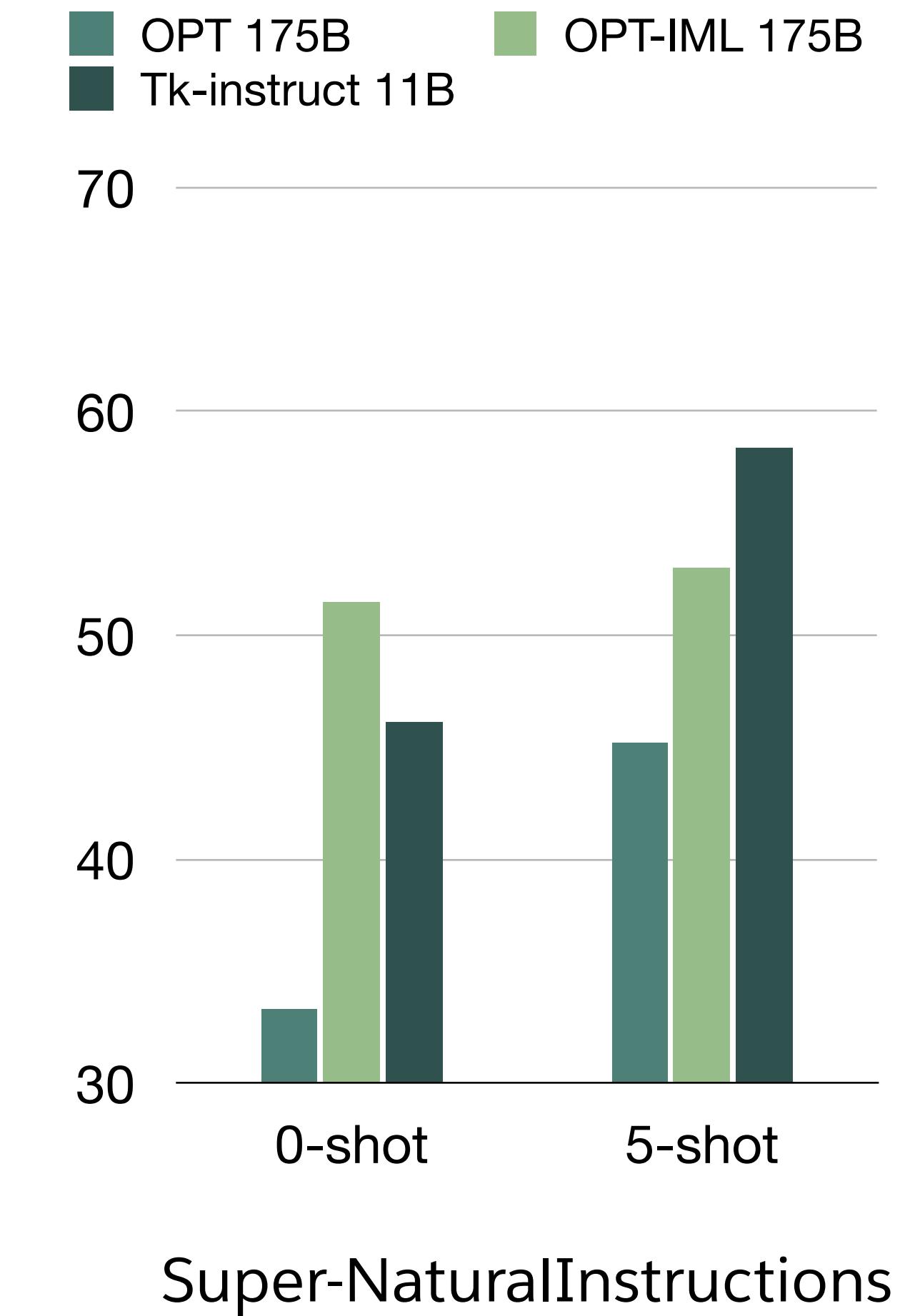
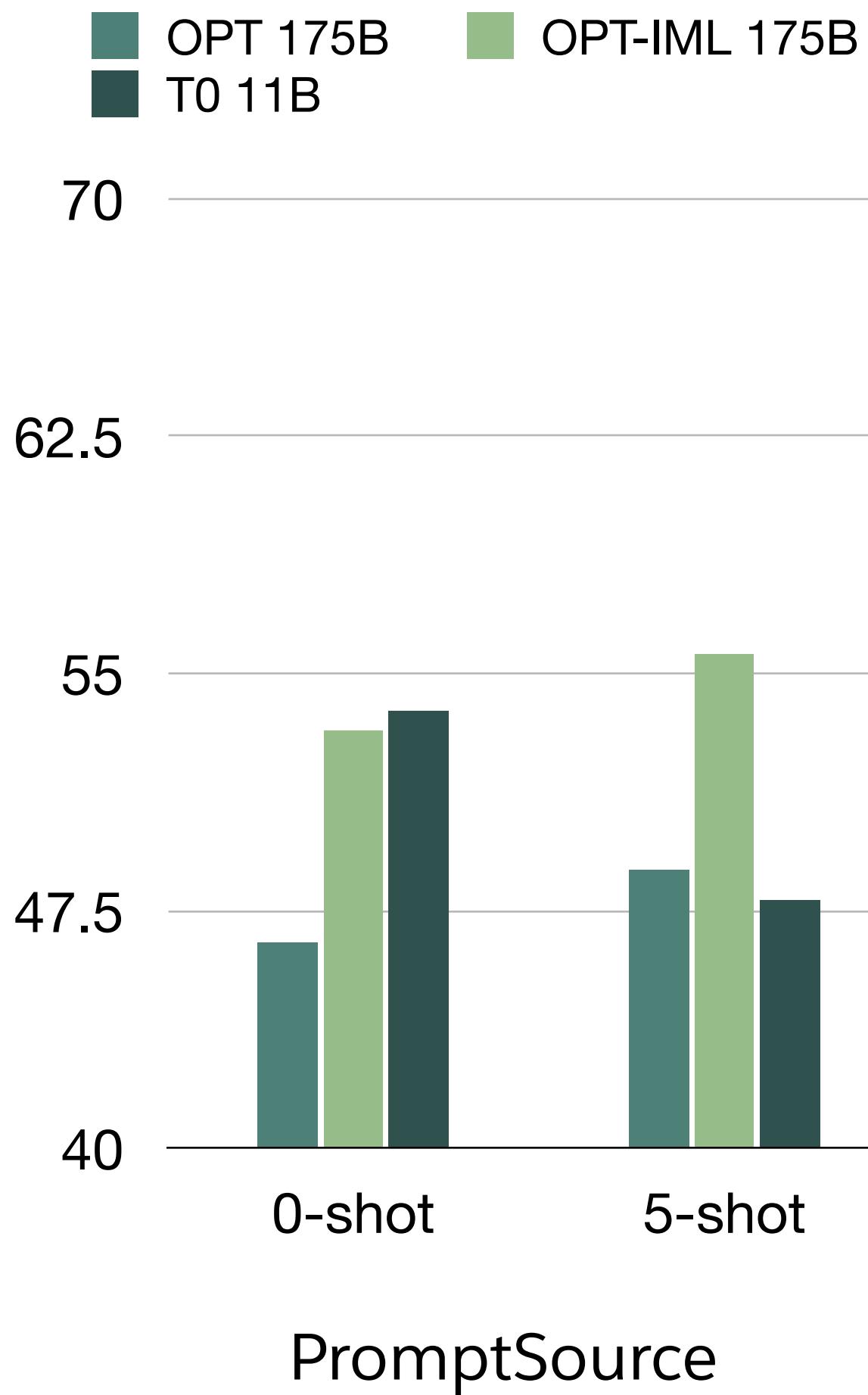
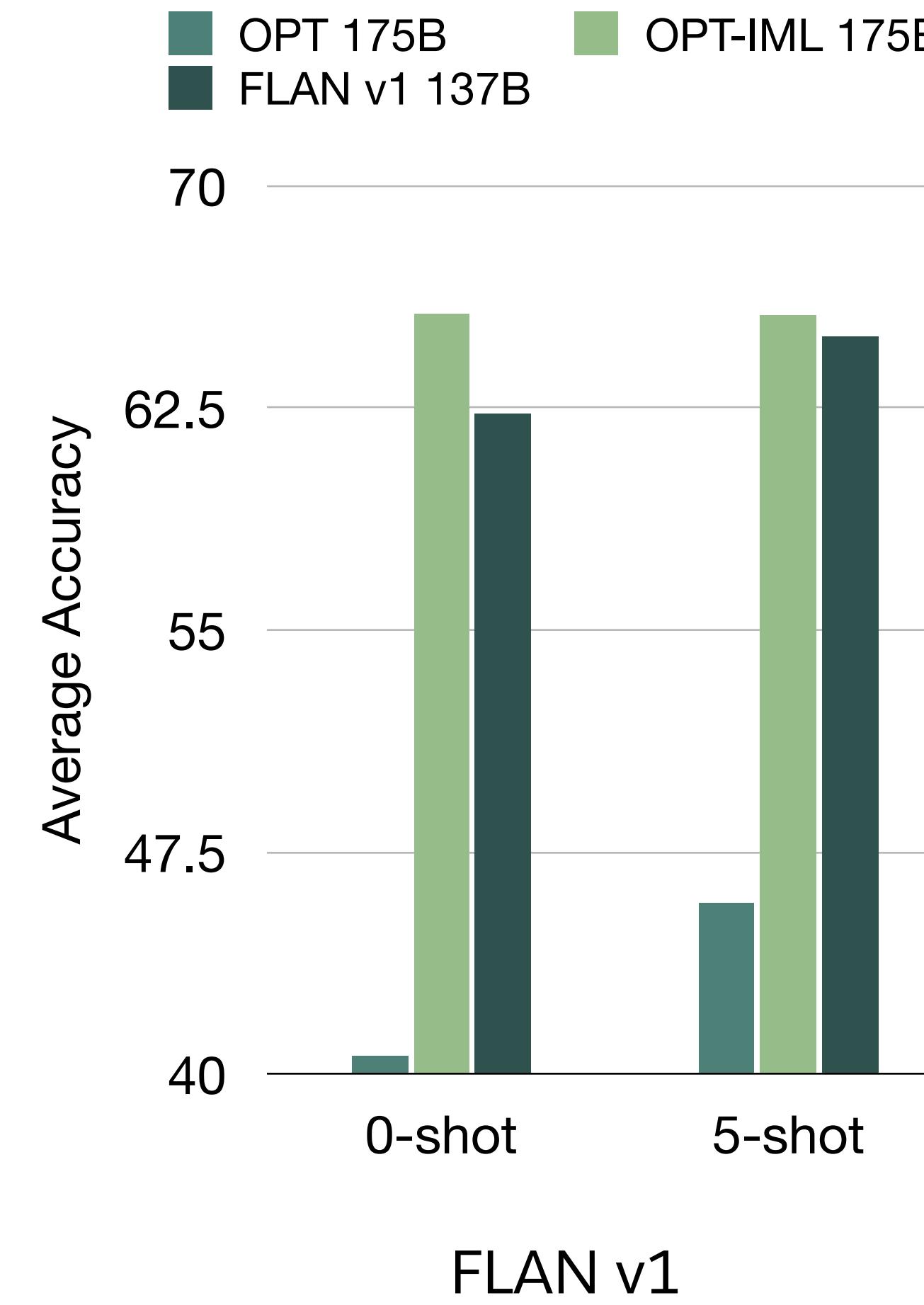
LM Benchmark Evaluation



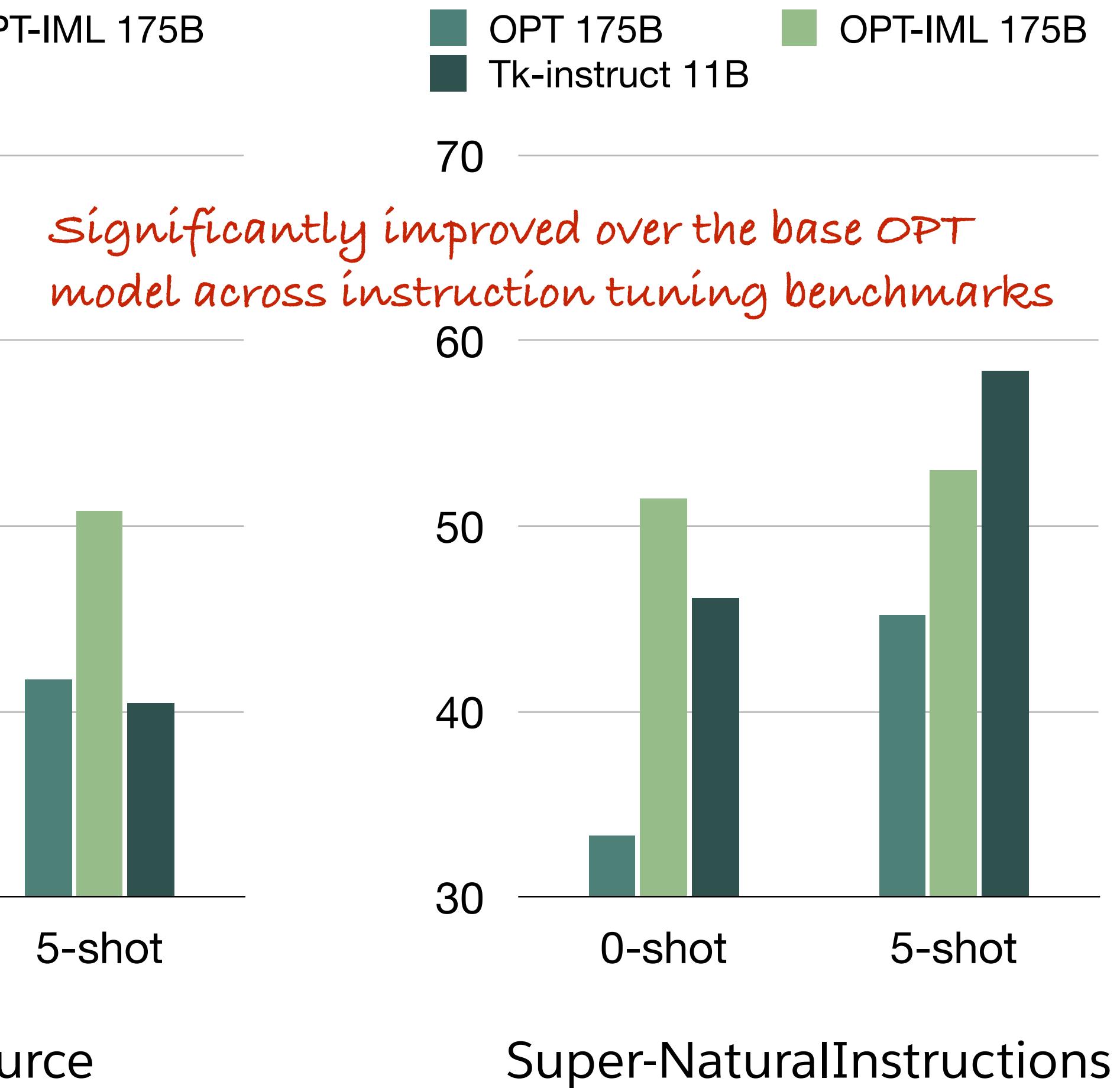
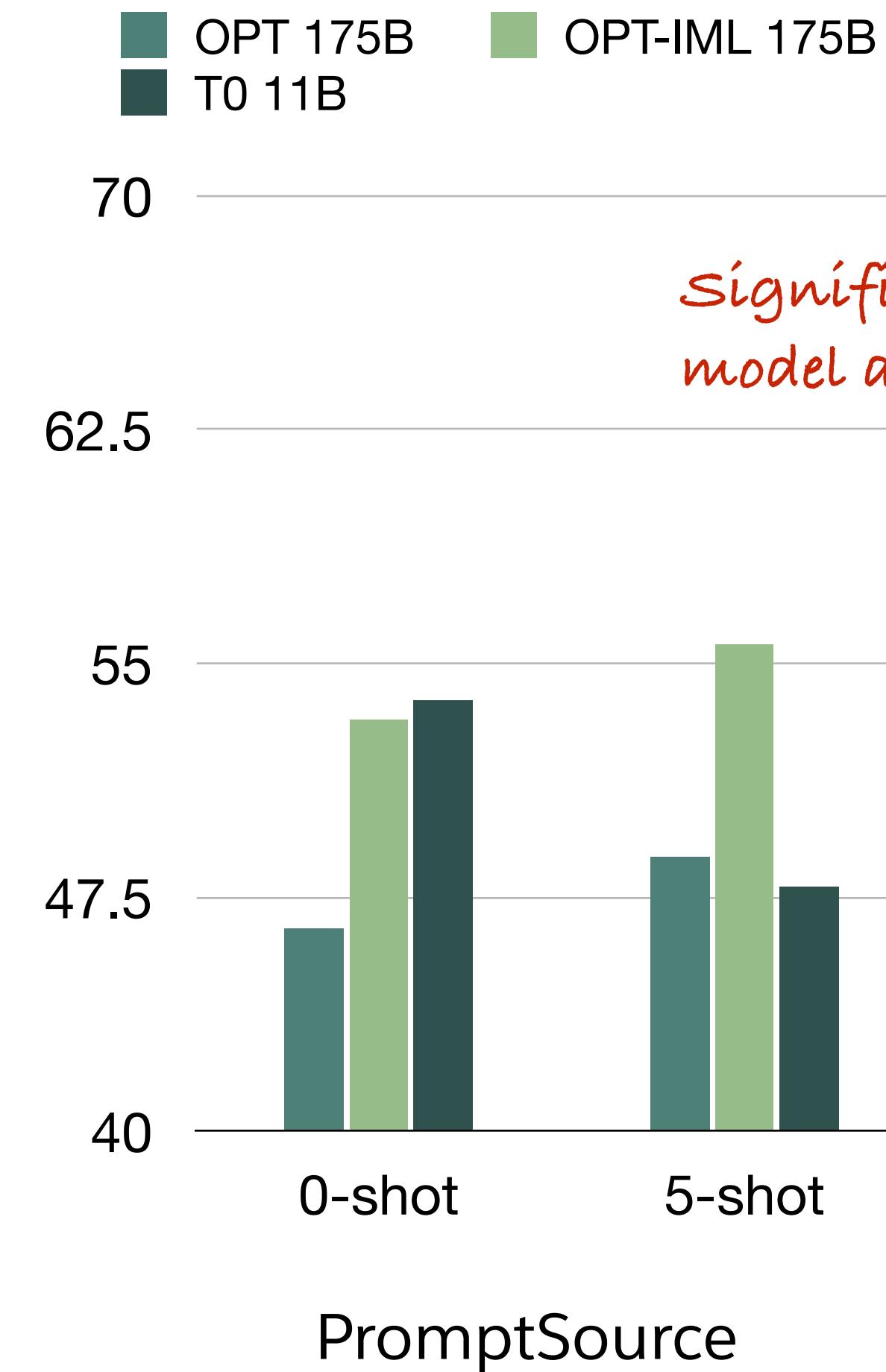
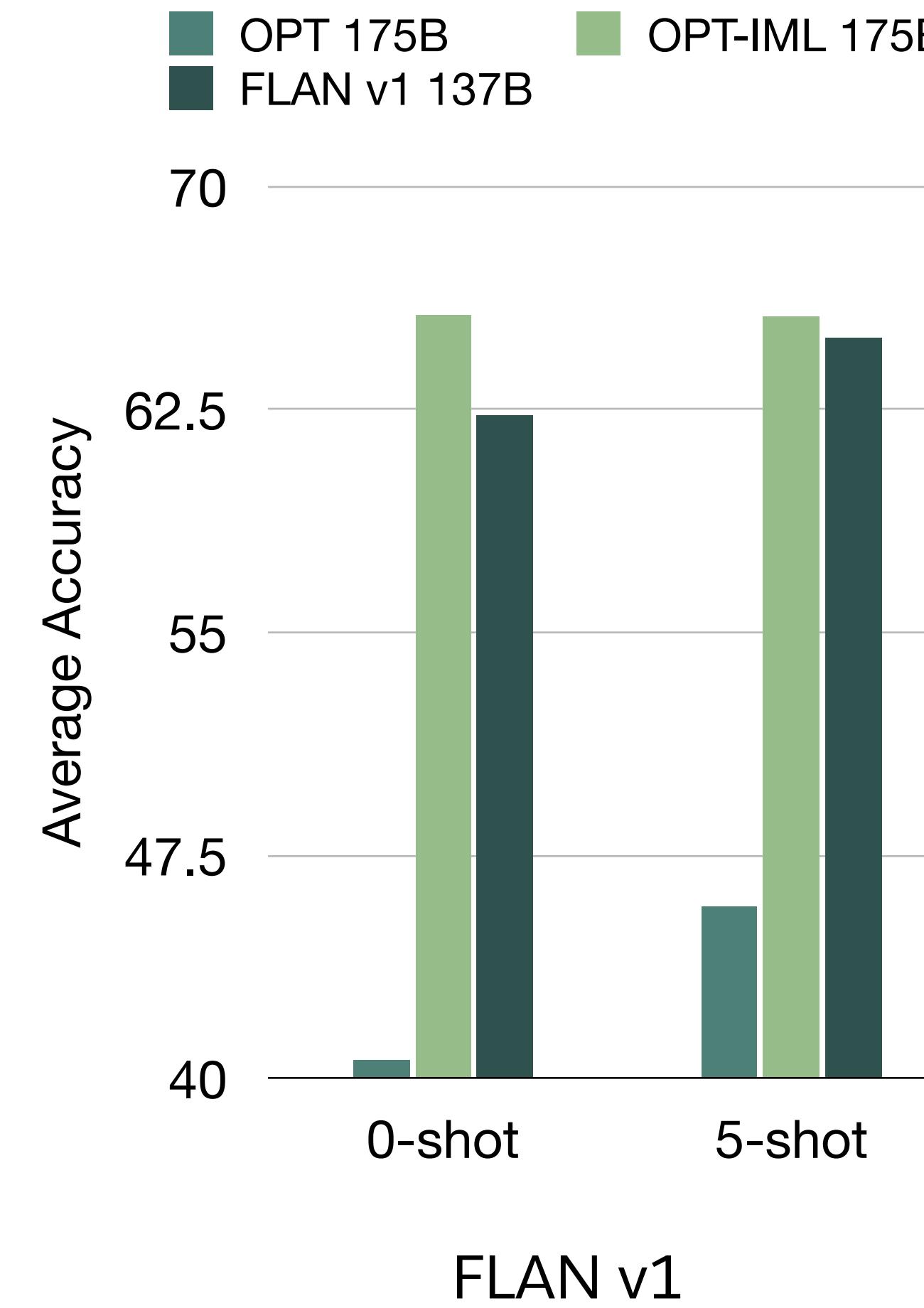
LM Benchmark Evaluation



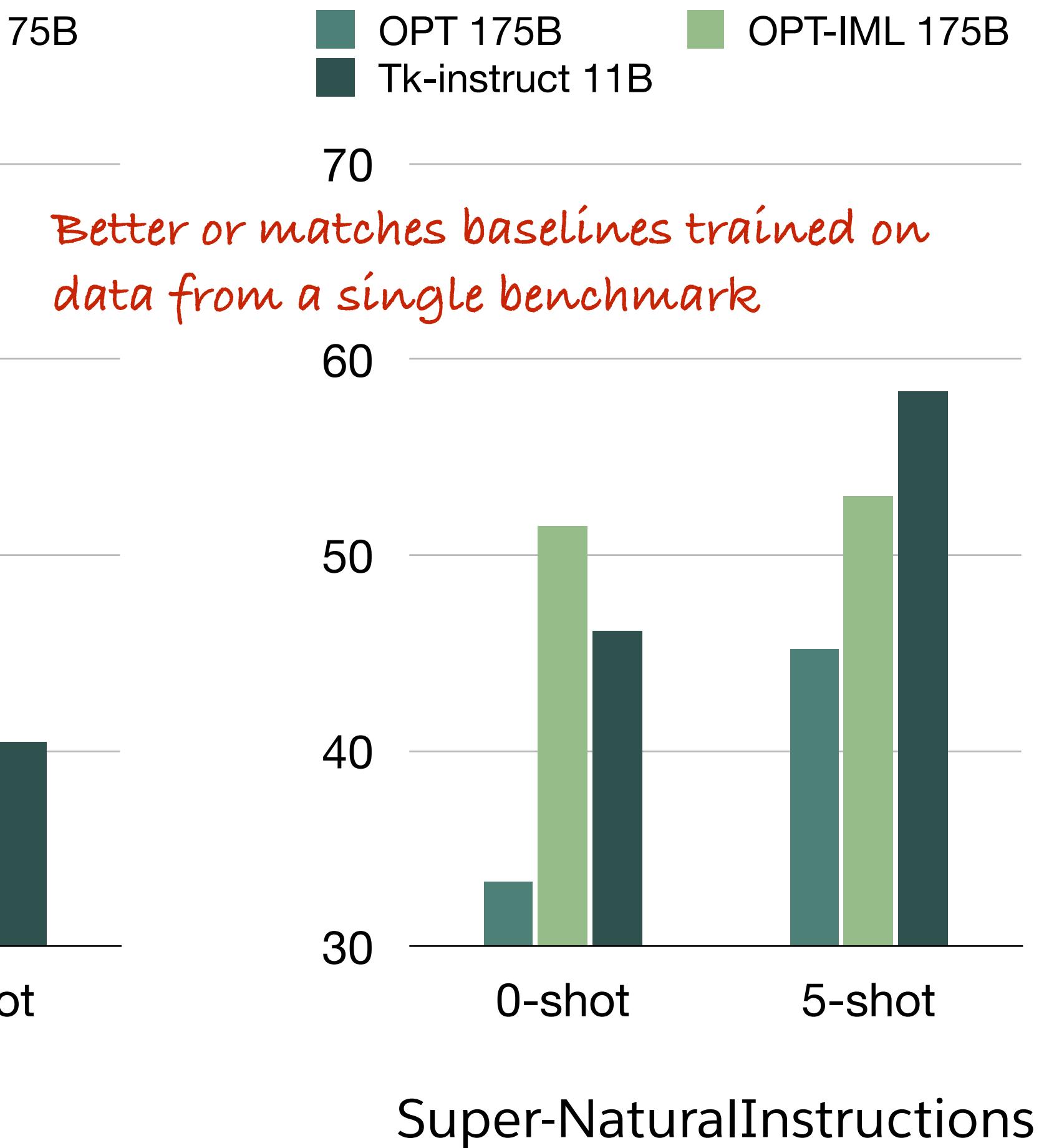
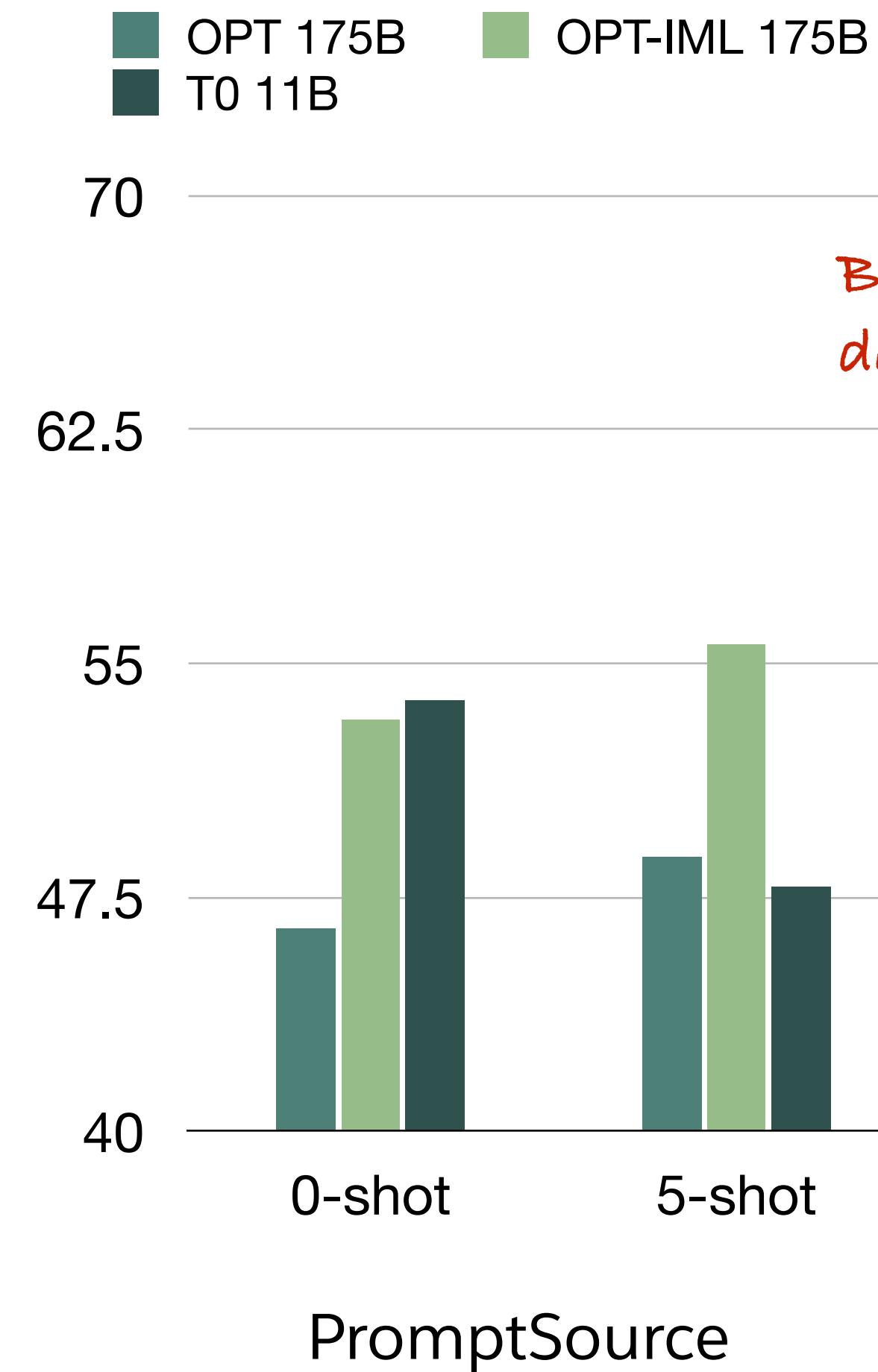
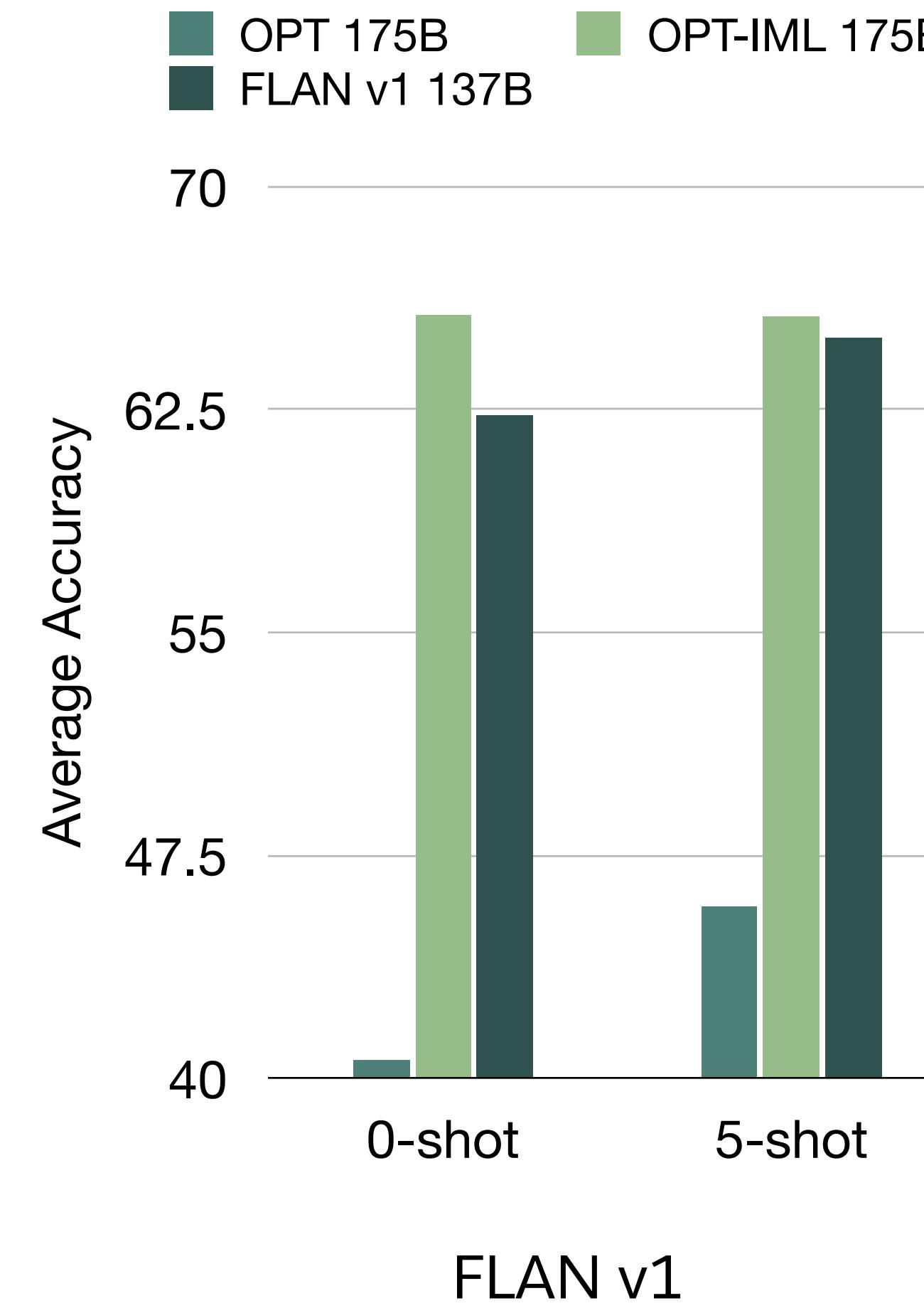
Instruction Tuning Benchmark Evaluation



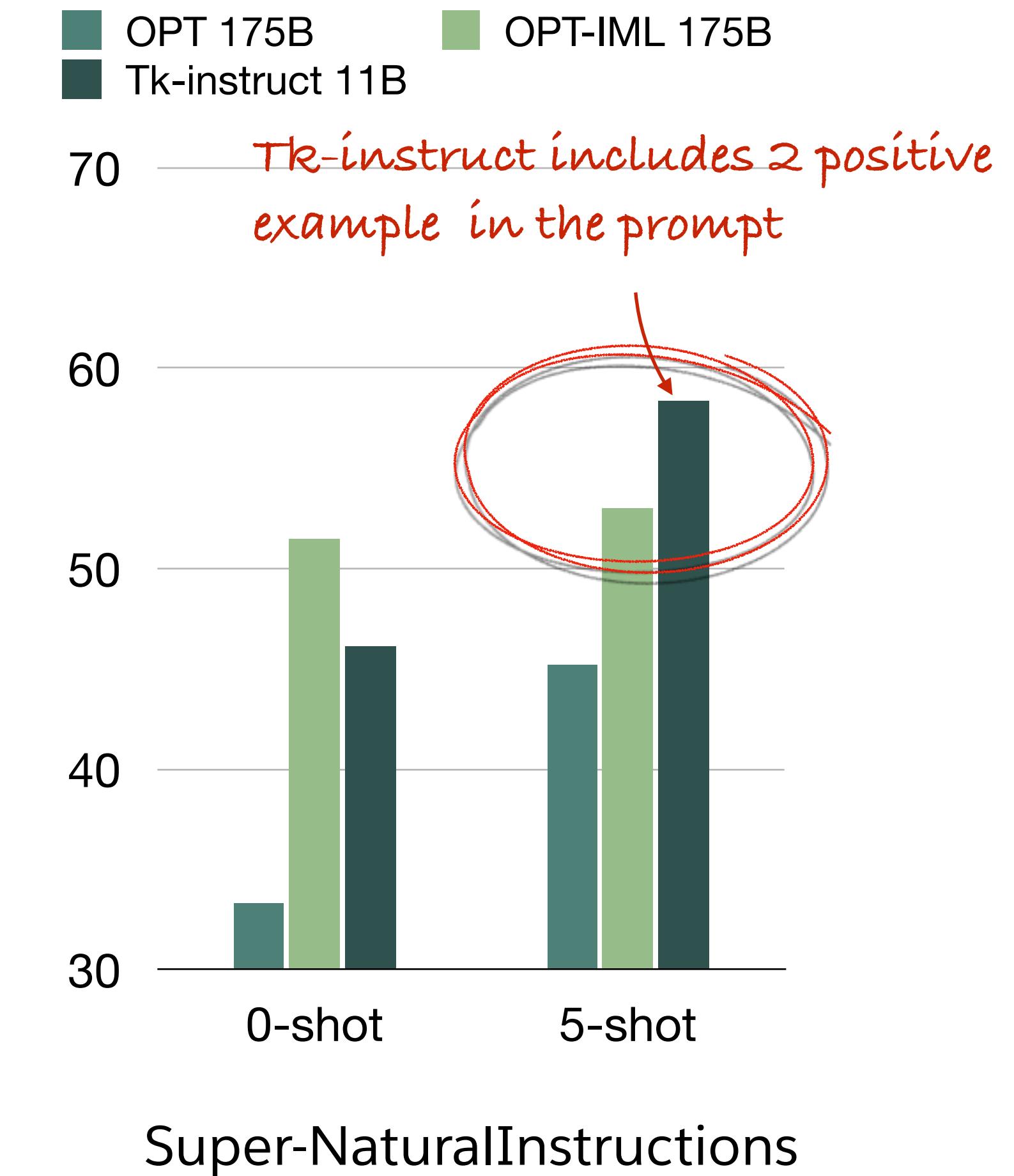
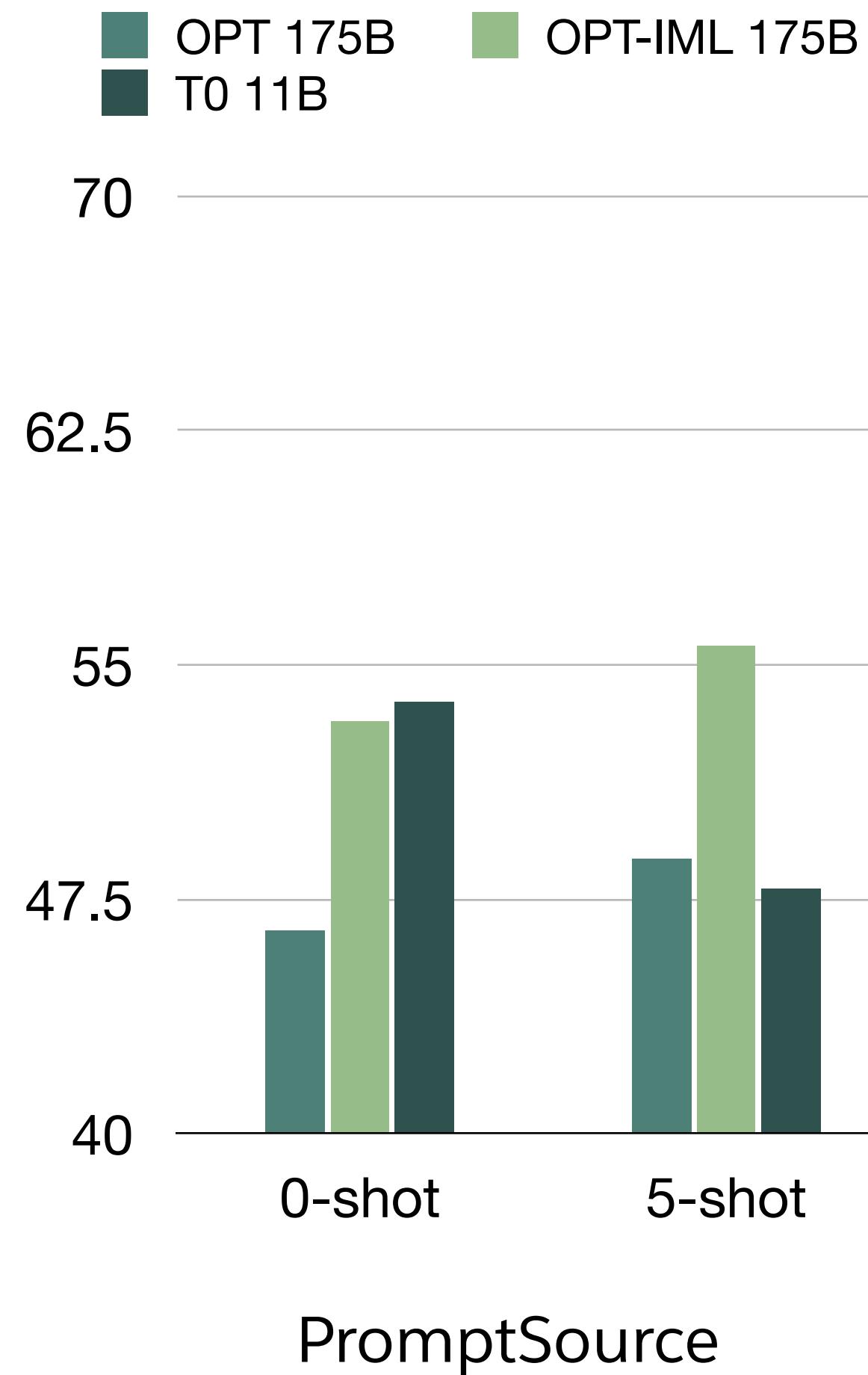
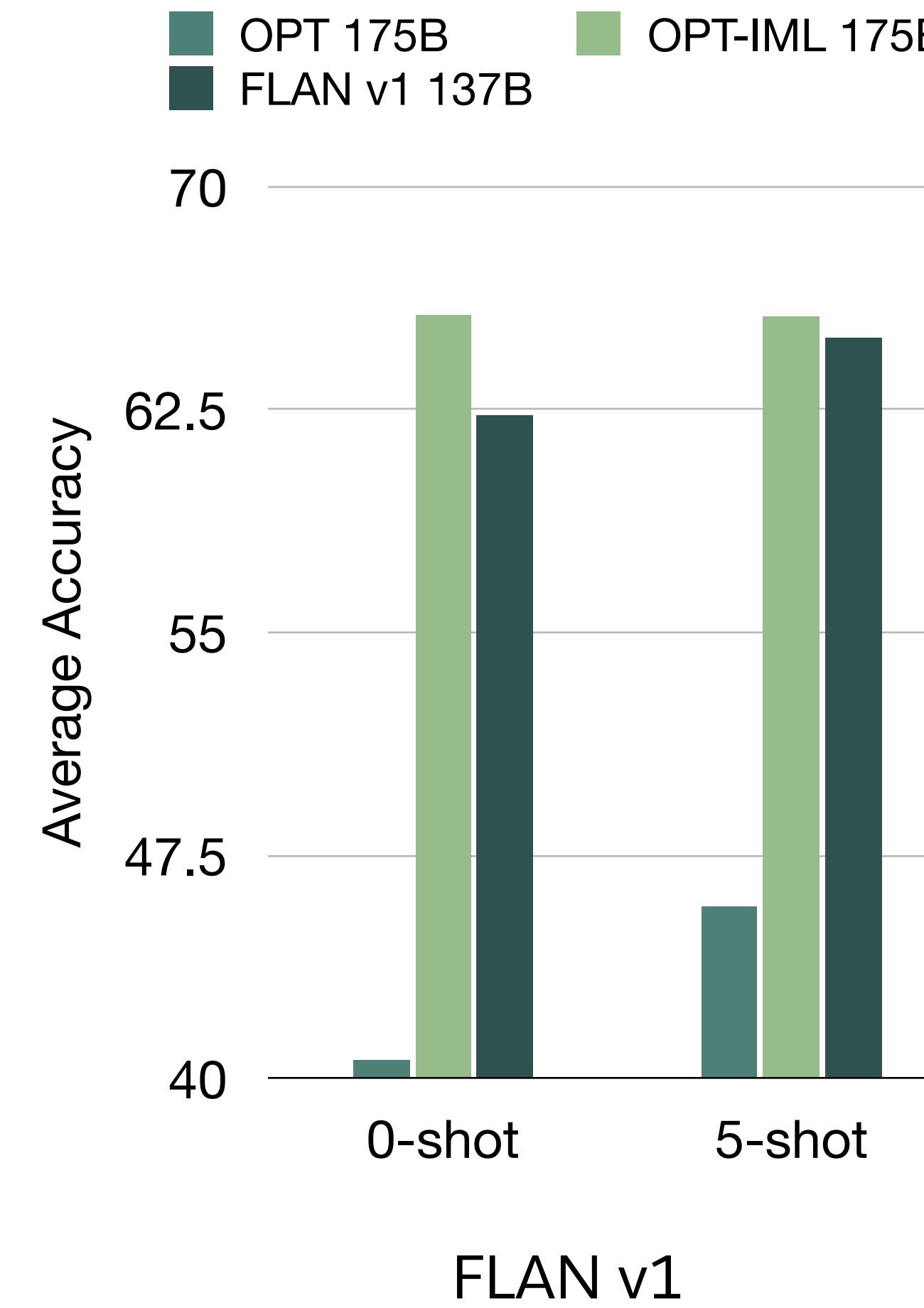
Instruction Tuning Benchmark Evaluation



Instruction Tuning Benchmark Evaluation



Instruction Tuning Benchmark Evaluation



Key Learnings

1. Fine-tuning language models on data format in <instruction, response> pairs effectively enhance the model's capabilities to follow natural language instructions and solving tasks as directed.
2. Increasing data **diversity (task semantics, prompt language/schema)** →  cross-task and cross-dataset generalization,  supervised task performance.
3. **If the base model is undertrained, scaling up data at fine-tuning stage cannot match** models fine-tuned from well trained based models (similar conclusion was also made by Zhou et al. 2023).

Outline

I. LLMs as General Parametric Task Solvers

- I. LLM Pre-training and Scaling Laws
- 2. Instruction Tuning

2. External Knowledge Access

- I) Retrieval Augmentation
- II) Tool Augmentation

3. LLM Agents and Advanced Applications

Limitation of LLMs

Q: Why does my GPU keep failing with Xid 79 fallen off the bus? **A:**

Do not use Y-splitters or 6-pin to 8-pin converters in any of the PCIe…

Q: When did Beyonce start becoming popular? **A:**

In the late 1990s…



Limitation of LLMs

✗ Knowledge cut-off

Q: Why does my GPU keep failing with Xid 79 fallen off the bus? **A:**

Do not use Y-splitters or 6-pin to 8-pin converters in any of the PCIe…

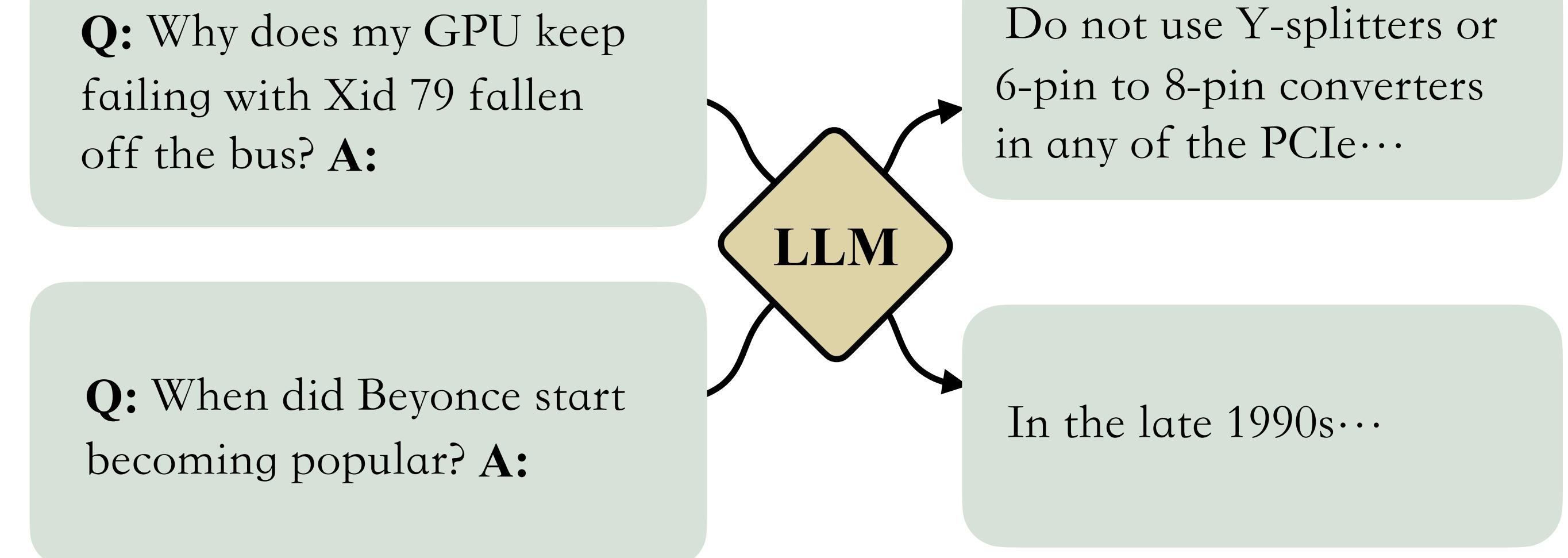
Q: When did Beyonce start becoming popular? **A:**

In the late 1990s…



Limitation of LLMs

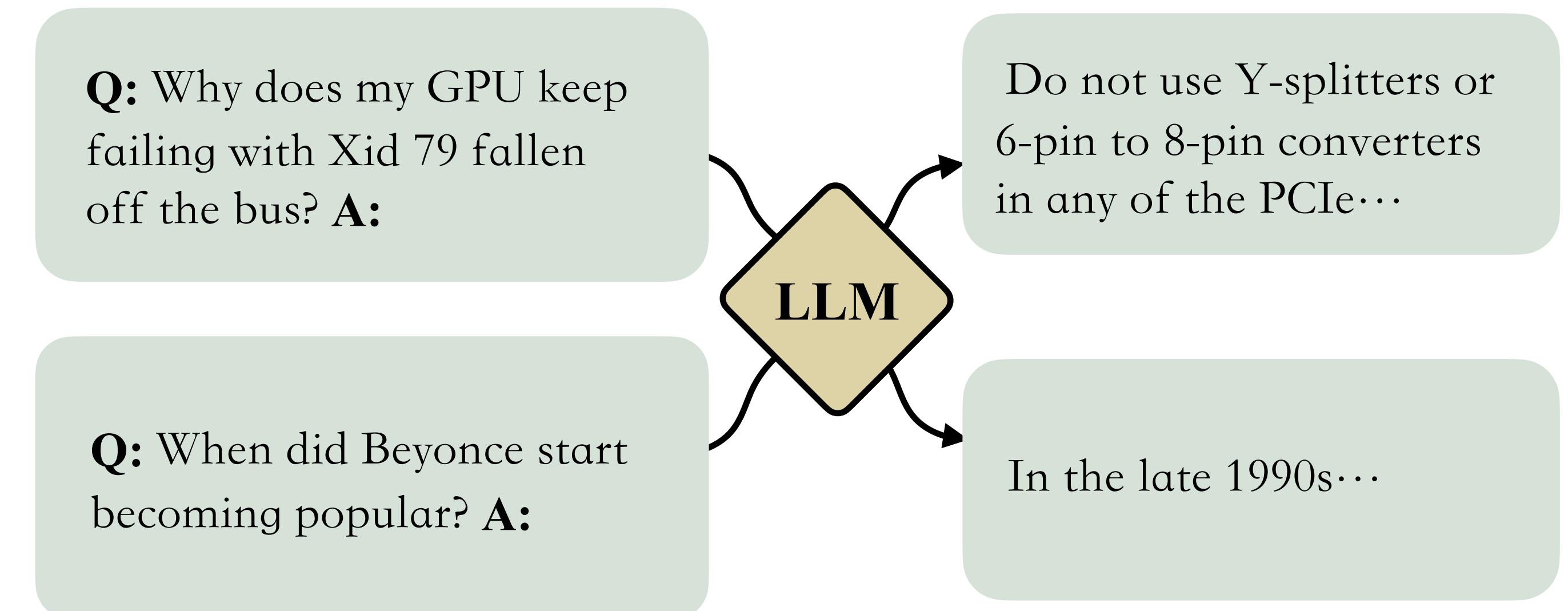
- ✖ Knowledge cut-off
- ✖ Hallucination (Sun et al. 2023)



Sun et al. 2023. Head-to-Tail: How Knowledgeable are LLMs? ALA Will LLMs Replace Knowledge Graphs?
Tan et al. 2023. Reimagining Retrieval Augmented Language Models for Answering Queries.

Limitation of LLMs

- ✗ Knowledge cut-off
- ✗ Hallucination (Sun et al. 2023)
- ✗ Provenance



Sun et al. 2023. Head-to-Tail: How Knowledgeable are LLMs? ALA Will LLMs Replace Knowledge Graphs?
Tan et al. 2023. Reimagining Retrieval Augmented Language Models for Answering Queries.

Limitation of LLMs: Provenance



Could you suggest specific foods and dietary sources that are rich in Vitamin D?

<https://www.dietaryguidelines.gov/resources/2020-2025-dietary-guidelines-online-materials/food-sources-select-nutrients/food-sources>

Dairy and Fortified Soy Alternatives			
Soy beverage (soy milk), unsweetened	1 cup	80	119
Milk, low fat (1 %)	1 cup	102	117
Yogurt, plain, nonfat	8 ounces	137	116
Yogurt, plain, low fat	8 ounces	154	116
Milk, fat free (skim)	1 cup	83	115
Kefir, plain, low fat	1 cup	104	100
Cheese, American, low fat or fat free, fortified	1 1/2 ounces	104	85

To increase your vitamin D levels through diet, consider incorporating the following foods, as they are rich in vitamin D:

- **Protein Foods:** Freshwater rainbow trout and salmon are excellent sources, providing between 383 to 645 IU of vitamin D per 3-ounce serving. Other good seafood options include light canned tuna, herring, sardines, tilapia and flounder.
- **Dairy and Fortified Soy Alternatives:** [Unsweetened soy milk, low-fat milk, nonfat and low-fat yogurt, fat-free milk, low-fat kefir, and fortified low-fat or fat-free American cheese are all good dairy sources of vitamin D.](#)
- **Vegetables:** Certain varieties of mushrooms can offer significant amounts of vitamin D, especially if they have been exposed to sunlight.
- **Fruit and Fortified Beverages:** Fortified orange juice can provide about 100 IU of vitamin D per cup.
- **Other Sources:** Unsweetened almond and rice beverages are also fortified with vitamin D.

While these foods can help boost your vitamin D levels, sunlight exposure is another significant source of vitamin D for many people...

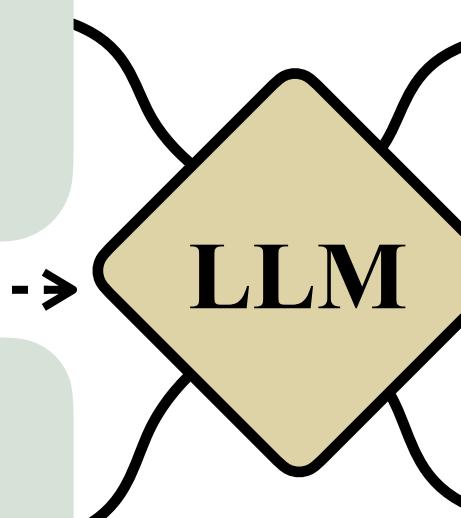


Retrieval Augmentation

Non-parametric Memory
(e.g. a vector datastore)

Q: Why does my GPU keep failing with Xid 79 fallen off the bus? **A:**

Q: When did Beyonce start becoming popular? **A:**



Do not use Y-splitters or 6-pin to 8-pin converters in any of the PCIe...

In the late 1990s...

Information Retrieval

Bag-of-Words (BOW) Retrieval

$$\text{score}(D, Q) = \sum_{t \in Q} \frac{f_{t,D} \cdot (k_1 + 1)}{f_{t,D} + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})} \cdot \log \frac{N - n_t + 0.5}{n_t + 0.5}$$

Annotations on the equation:

- sum the scores for each query term (points to the summation part)
- term frequency saturation trick (points to the term frequency part)
- forget about this: it doesn't affect score relationships so Lucene took it out (points to the k_1 term)
- adjust saturation curve based on document length (points to the denominator's scaling factor)
- probabilistic flavor of IDF: Lucene adds a 1 inside the log, making it basically the same as traditional IDF (points to the logarithmic term)

Dense Retrieval

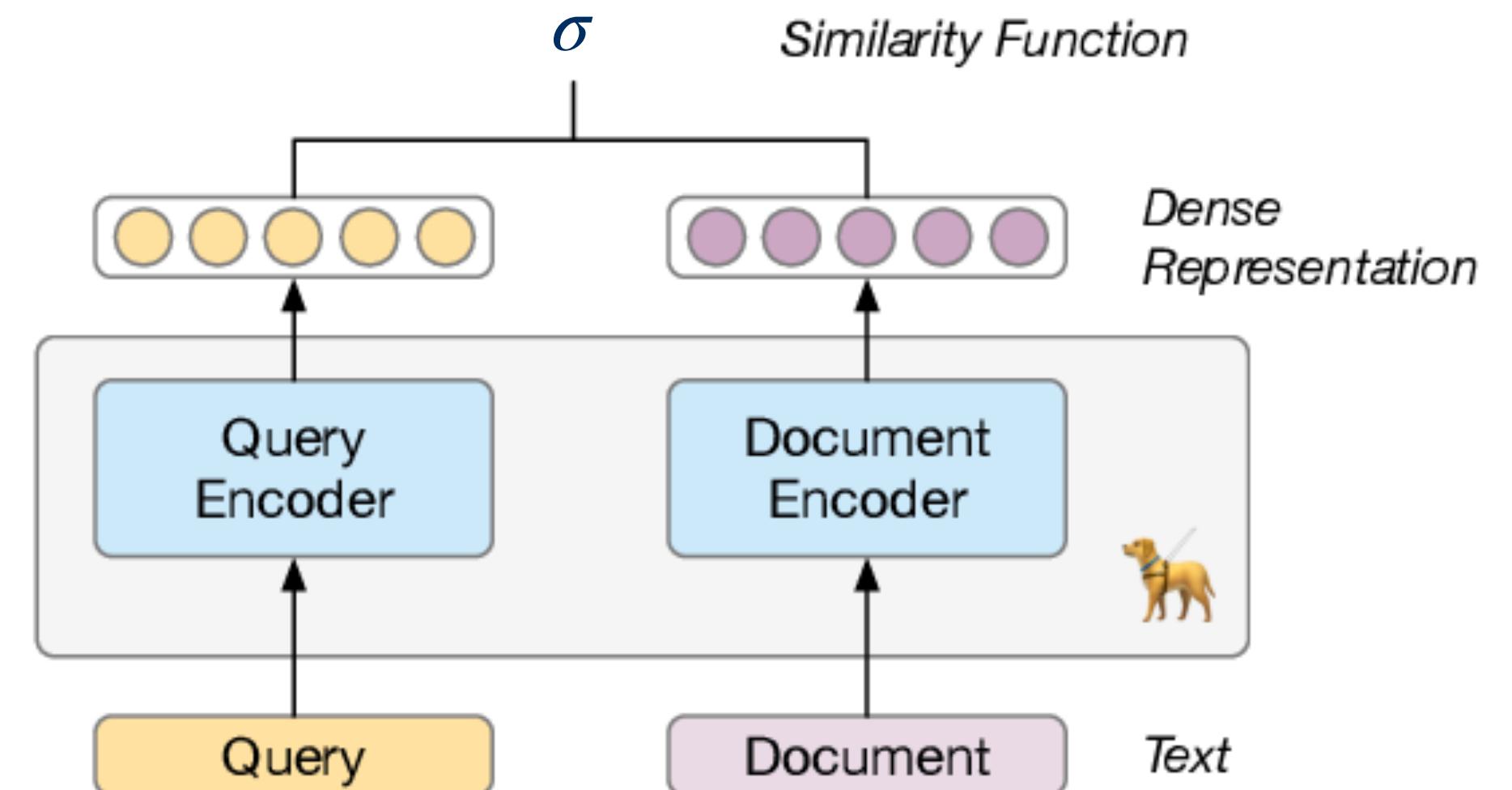


Image Credit: <https://kmwllc.com/index.php/2020/03/20/understanding-tf-idf-and-bm-25/>

1. https://en.wikipedia.org/wiki/Okapi_BM25
2. **Karpukhin et al. 2020.** Dense Passage Retrieval for Open-Domain Question Answering.

Information Retrieval

Bag-of-Words (BOW) Retrieval

$$\text{score}(D, Q) = \sum_{t \in Q} \frac{f_{t,D} \cdot (k_1 + 1)}{f_{t,D} + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})} \cdot \log \frac{N - n_t + 0.5}{n_t + 0.5}$$

sum the scores for each query term

term frequency saturation trick

forget about this: it doesn't affect score relationships so Lucene took it out

adjust saturation curve based on document length

probabilistic flavor of IDF: Lucene adds a 1 inside the log, making it basically the same as traditional IDF

Dense Retrieval

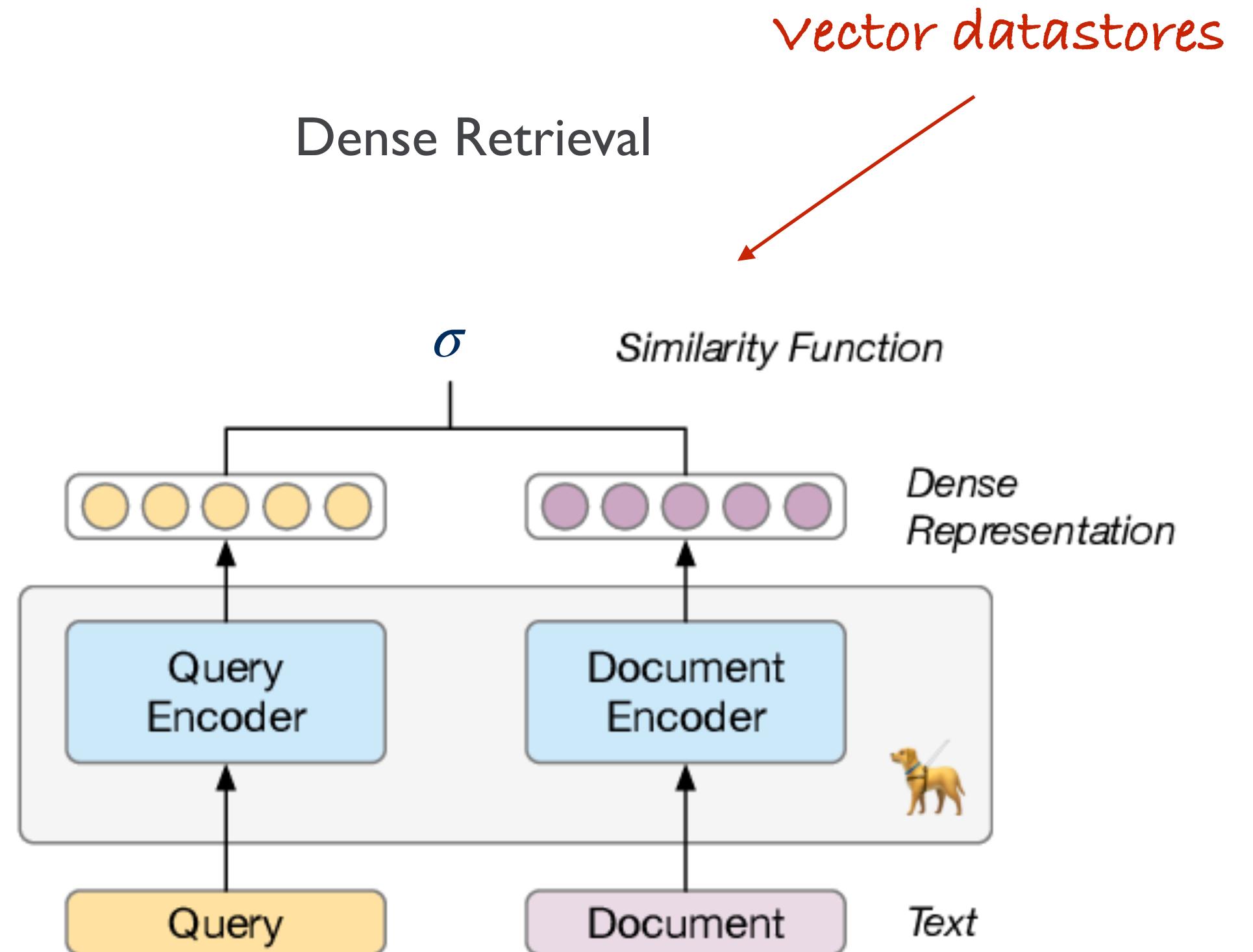
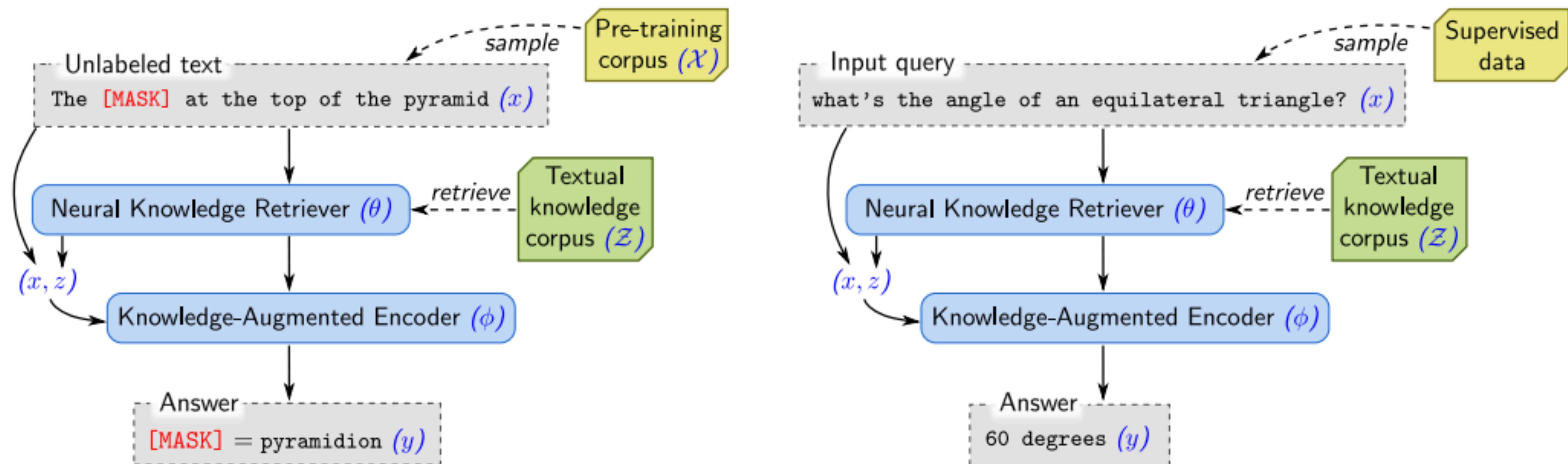


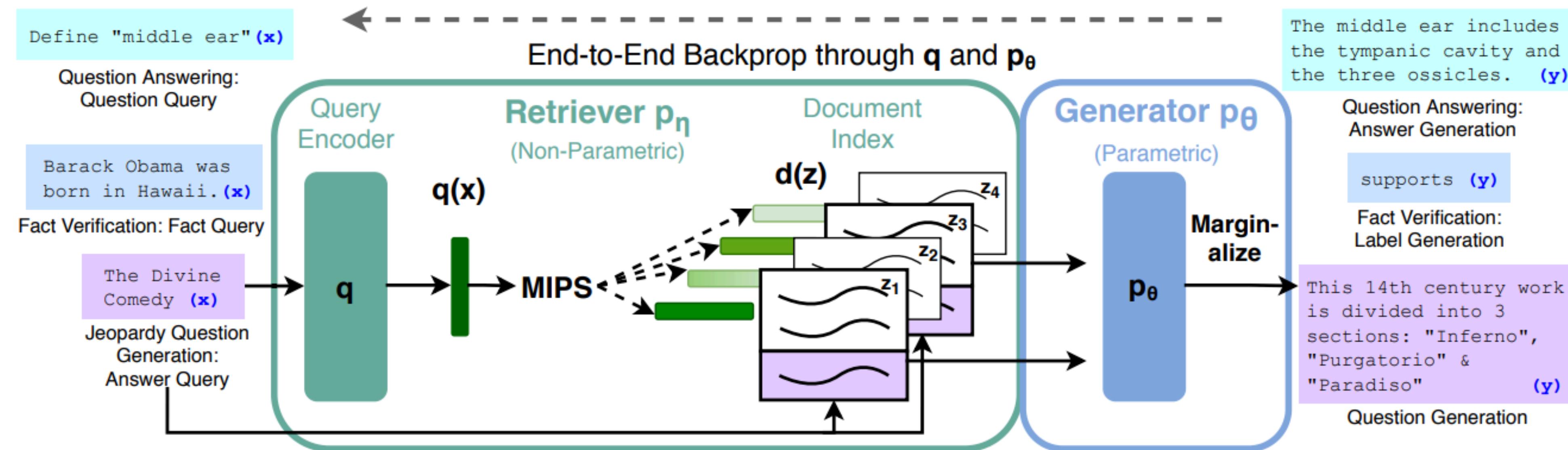
Image Credit: <https://kmwllc.com/index.php/2020/03/20/understanding-tf-idf-and-bm-25/>

1. https://en.wikipedia.org/wiki/Okapi_BM25
2. **Karpukhin et al. 2020.** Dense Passage Retrieval for Open-Domain Question Answering.

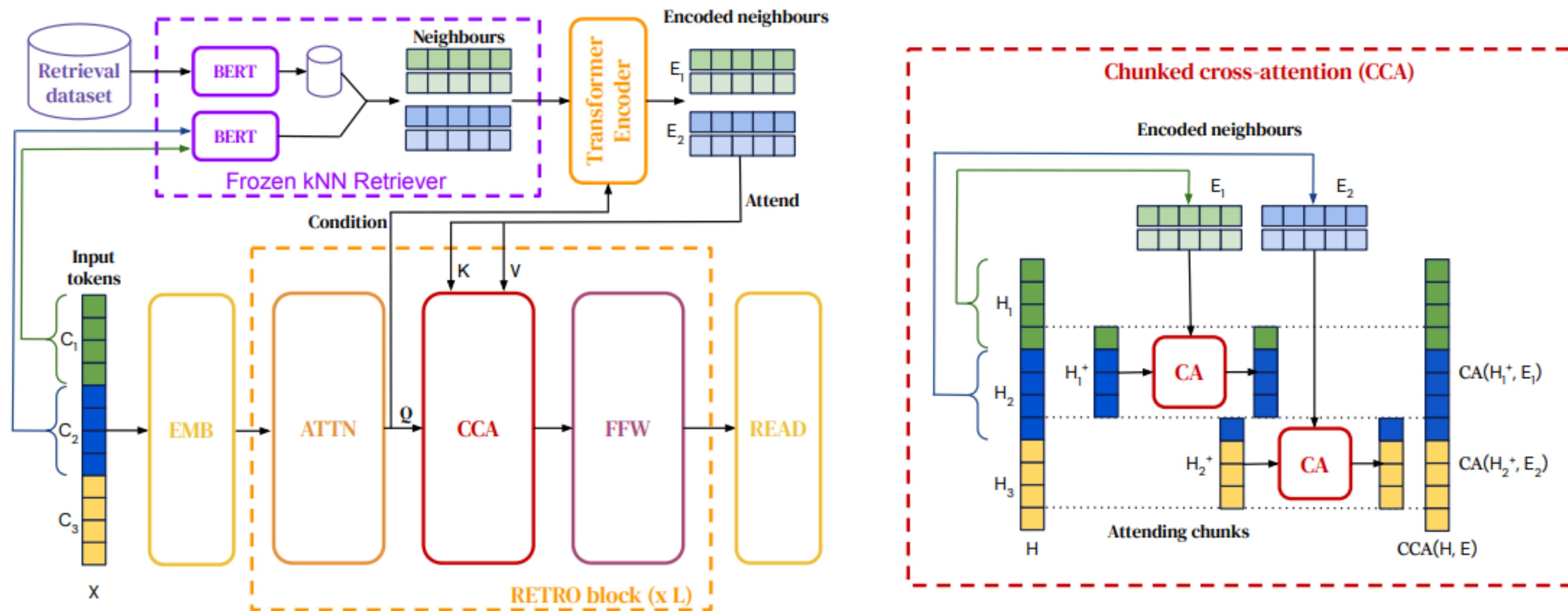
REALM



Retrieval-Augmented Generation (RAG)

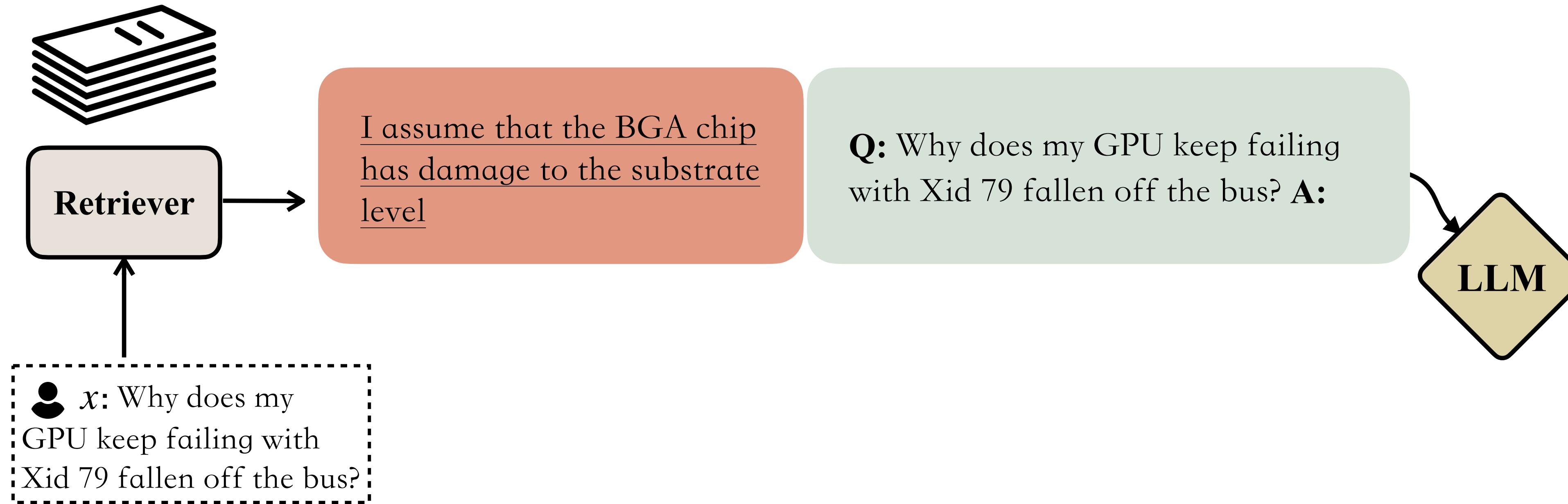


RETRO



Borgeaud et al. 2021. Improving Language Models by Retrieving from Trillions of Tokens.

REPLUG

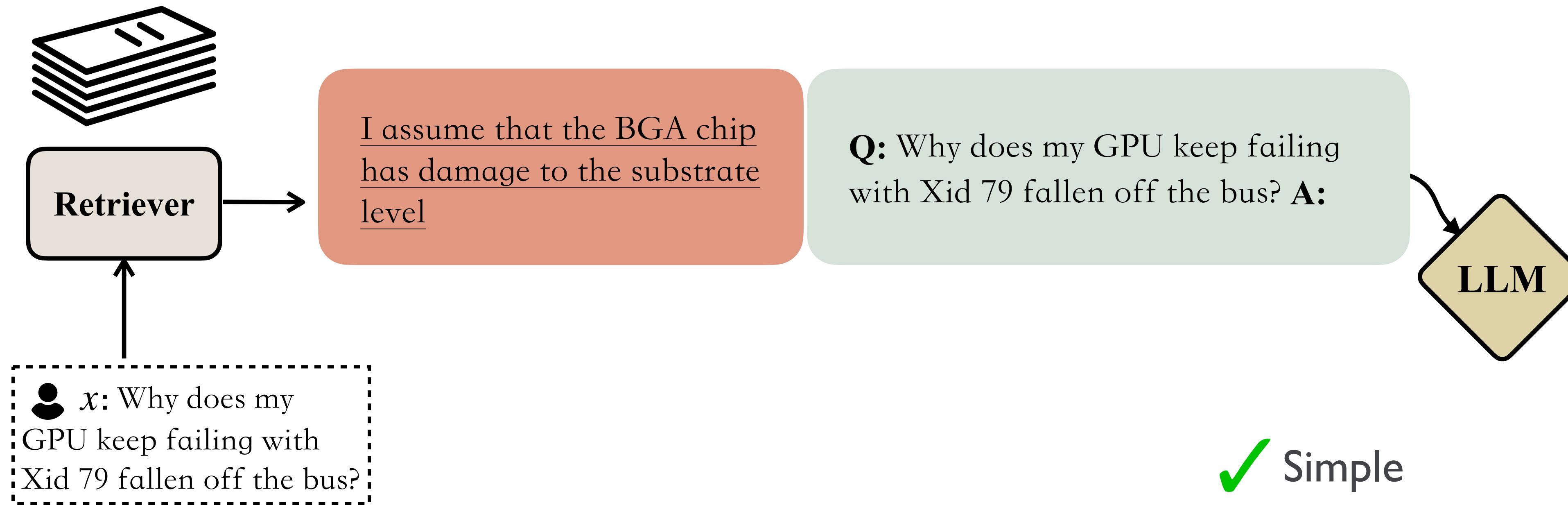


Shi et al. 2023. REPLUG: Retrieval-Augmented Black-Box Language Models.

Ram et al. 2023. In-Context Retrieval-Augmented Language Models.

Lazaridou et al. 2022. Internet-augmented Language Models through Few-shot Prompting for Open-domain Question Answering.

REPLUG



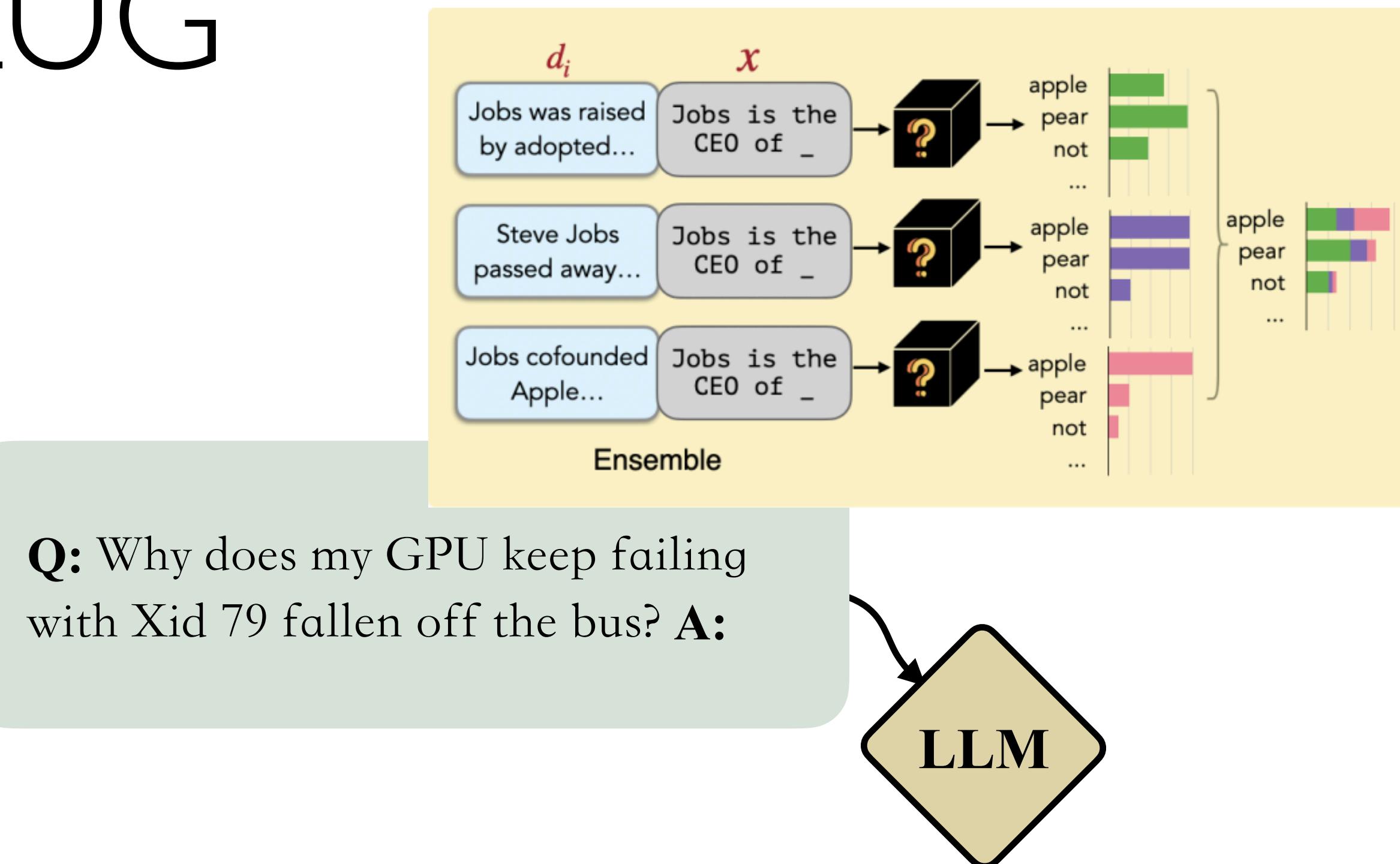
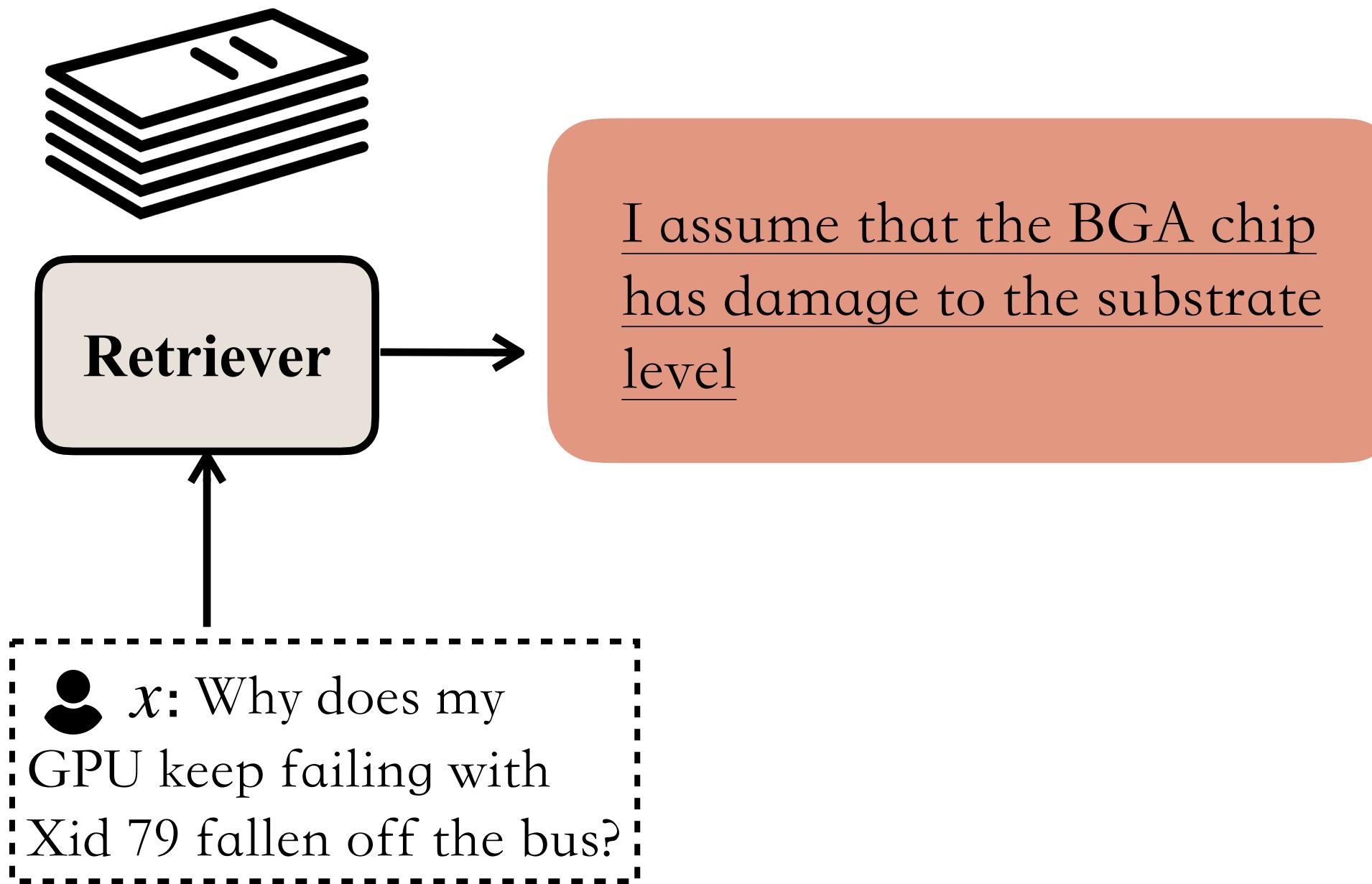
- ✓ Simple
- ✓ Effectively leverage off-the-shelf LLMs and Retrievers

Shi et al. 2023. REPLUG: Retrieval-Augmented Black-Box Language Models.

Ram et al. 2023. In-Context Retrieval-Augmented Language Models.

Lazaridou et al. 2022. Internet-augmented Language Models through Few-shot Prompting for Open-domain Question Answering.

REPLUG



Shi et al. 2023. REPLUG: Retrieval-Augmented Black-Box Language Models.

Ram et al. 2023. In-Context Retrieval-Augmented Language Models.

Lazaridou et al. 2022. Internet-augmented Language Models through Few-shot Prompting for Open-domain Question Answering.

Retriever-Augmented Dual Instruction Tuning

Can we **fine-tune RALMs** s.t.

- A. The LLM can learn to better utilize the retrieved content in context
- B. The retriever can return more contextually relevant content

Fine-tuning Datasets

Dialogue		
Name	LM	R
OASST	✓	✓

Open-domain QA		
Name	LM	R
Commonsense QA	✓	✓
MathQA	✓	✓
Web Questions	✓	✓
WikiQA	✓	✓
YahooAnswersQA	✓	✓
Freebase QA		✓
MS_MARCO		✓

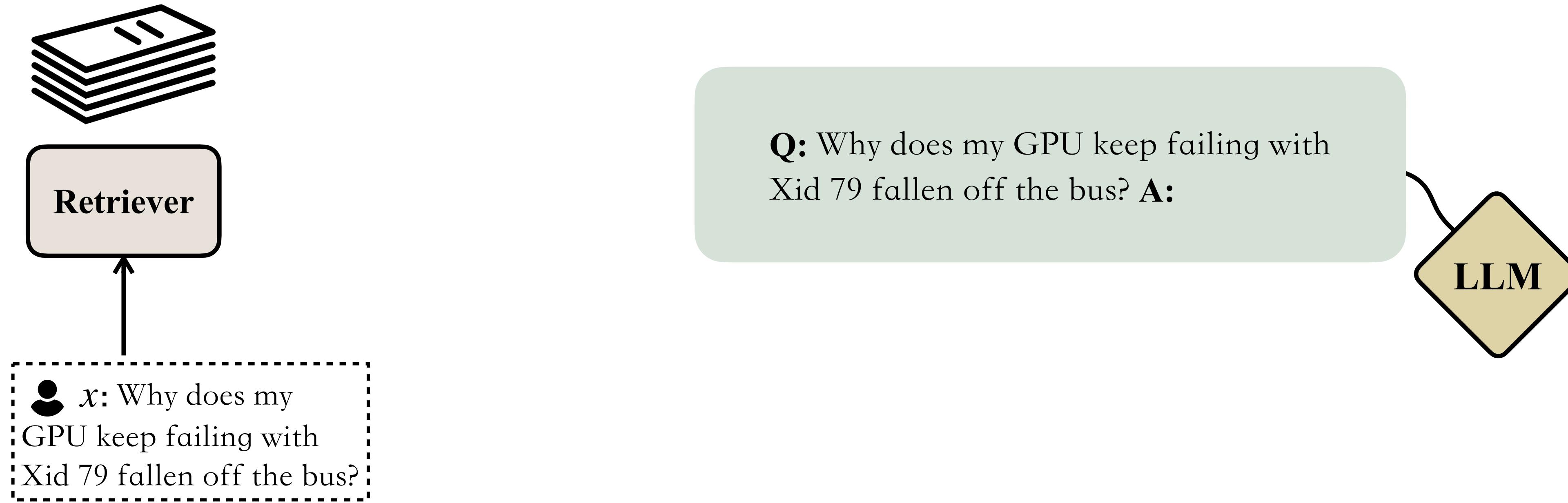
Reading Comprehension		
Name	LM	R
CoQA	✓	
DROP	✓	
NarrativeQA	✓	
NewsQA	✓	
PubMed QA	✓	✓
QUAIL	✓	
QUAREL	✓	✓
SQuAD v2	✓	

Chain-of-thought Reasoning		
Name	LM	R
AQuA	✓	
ECQA	✓	
GSM8k	✓	
Math	✓	
StrategyQA	✓	

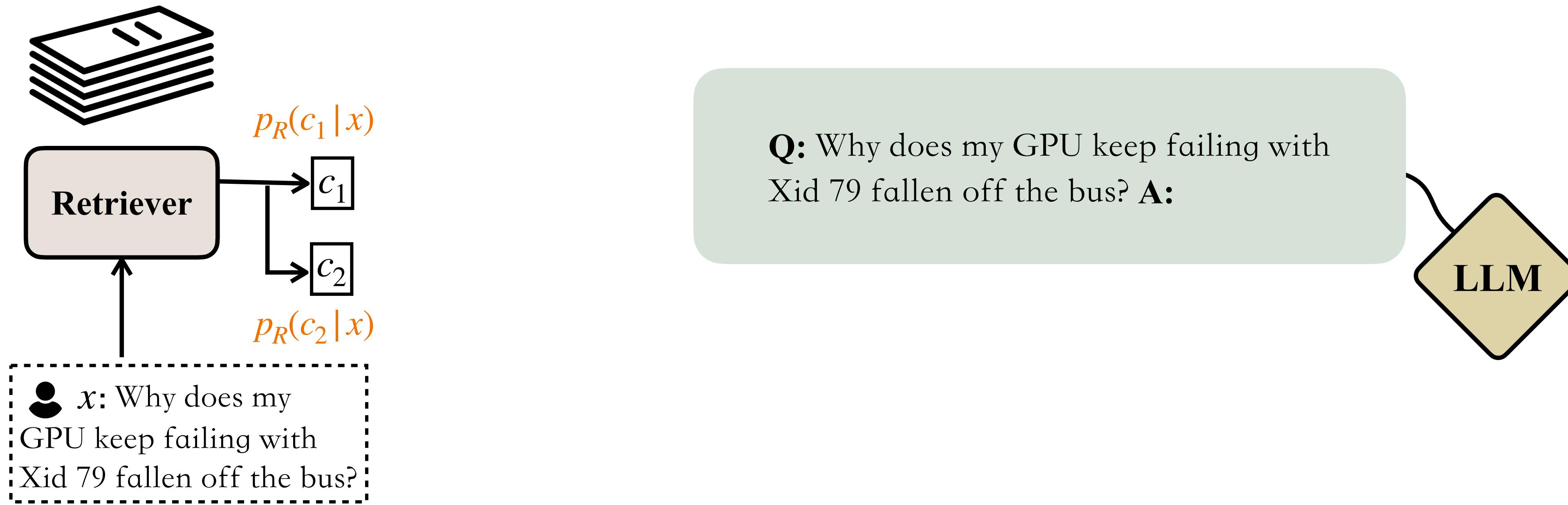
Summarization		
Name	LM	R
CNN_Dailymail	✓	



Retrieval-Augmented Dual Instruction Tuning

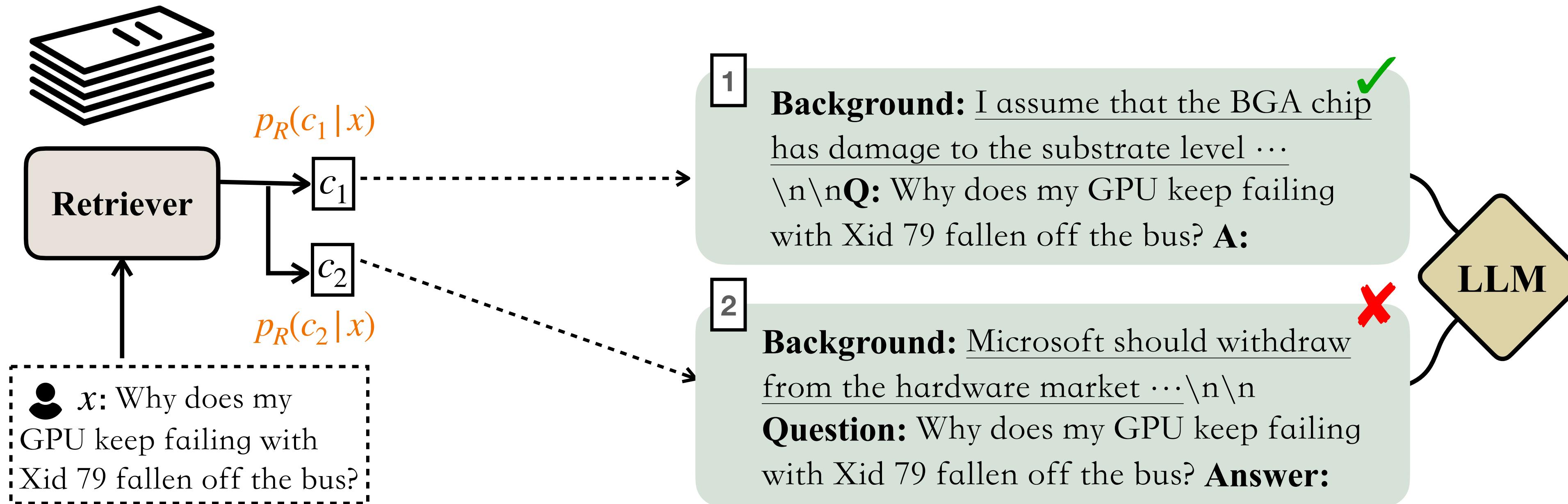


Retrieval-Augmented Dual Instruction Tuning

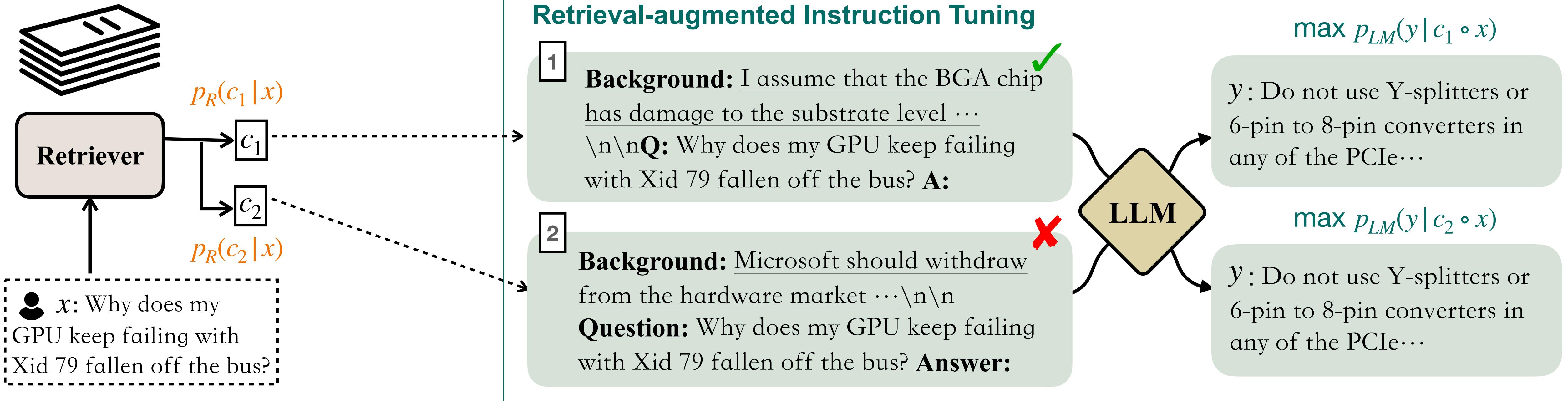


$$p_R(c|x) = \frac{\exp s(x,c)}{\sum_{c' \in C} \exp s(x,c')}$$

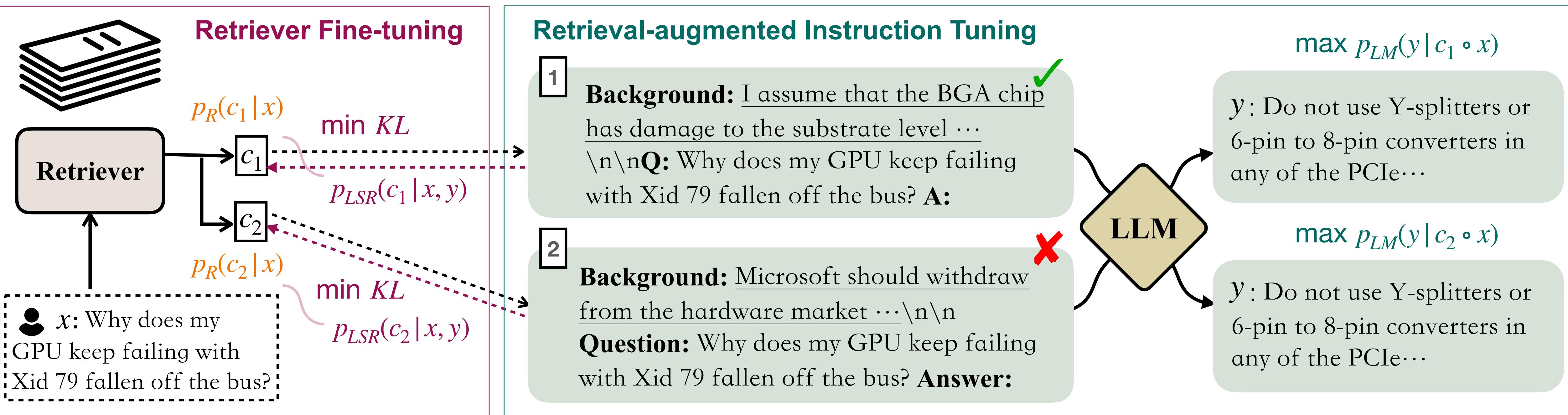
Retrieval-Augmented Dual Instruction Tuning



Retrieval-Augmented Dual Instruction Tuning



Retrieval-Augmented Dual Instruction Tuning



$$p_{LSR}(c|x, y) = \frac{\exp(p_{LM}(y|c \circ x)/\tau)}{\sum_{c' \in \mathcal{C}} \exp(p_{LM}(y|c' \circ x)/\tau)} \approx \frac{\exp(p_{LM}(y|c \circ x)/\tau)}{\sum_{c' \in \mathcal{C}'} \exp(p_{LM}(y|c' \circ x)/\tau)}$$

Experiment Setup

Fine-tuning

- For each example, generate $k = 3$ retrieval-augmented fine-tuning prompts using top- k retrieved text chunks

Inference

- For each example, generate $k' = 10$ retrieval-augmented prompts, perform parallel retrieval-augmentation and ensemble the results, same as REPLUG (Shi et al. 2023)

Evaluation: Knowledge Intensive Tasks

Slot Filling

INPUT:
Star Trek [SEP] creator

OUTPUT:
Gene Roddenberry

PROVENANCE:
17157886-1

Open Domain QA

INPUT:
When did Star Trek go off the air

OUTPUT:
June 3, 1969

PROVENANCE:
17157886-5

INPUT:
Which Star Trek star directed Three Men and a Baby?

OUTPUT:
Leonard Nimoy

PROVENANCE:
17157886-4, 596639-7

INPUT:
Treklanta (formerly "TrekTrax Atlanta") is an annual convention for what American science fiction media franchise?

OUTPUT:
Star Trek

PROVENANCE:
17157886-1, 28789994-6

zsRE

NQ

TQA

HoPo



Star Trek 17157886

Star Trek is an American media franchise based on the science fiction television series created by Gene Roddenberry.¹ [...] It followed the interstellar adventures of Captain James T. Kirk (William Shatner) and his crew aboard the starship USS "Enterprise", a space exploration vessel built by the United Federation of Planets in the 23rd century.² The "Star Trek" canon includes "The Original Series", an animated series, five spin-off television series, the film franchise, and further adaptations in several media.³ [...] The original 1966–69 series featured William Shatner as Captain James T. Kirk, Leonard Nimoy⁴ as Spock, DeForest Kelley as Dr. Leonard "Bones" McCoy, James Doohan as Montgomery "Scotty" Scott, Nichelle Nichols as Uhura, George Takei as Hikaru Sulu, and Walter Koenig as Pavel Chekov. During the series' first run, it earned several nominations for the Hugo Award for Best Dramatic Presentation, and won twice. [...] NBC canceled the show after three seasons; the last original episode aired on June 3, 1969⁵. [...]

Three Men and a Baby 596639

Three Men and a Baby is a 1987 American comedy film directed by Leonard Nimoy⁷ and starring Tom Selleck, Steve Guttenberg, Ted Danson and Nancy Travis. [...]

Treklanta 28789994

Treklanta is an annual "Star Trek" convention based in Atlanta, Georgia that places special emphasis on fan-based events, activities, programming and productions.⁶ [...]

Dialogue

INPUT:
I am a big fan of Star Trek, the American franchise created by Gene Roddenberry. I don't know much about it. When did the first episode air?
It debuted in 1966 and aired for 3 seasons on NBC.
What is the plot of the show?

OUTPUT:
William Shatner plays the role of Captain Kirk. He did a great job.

PROVENANCE:
17157886-2

WoW

Fact Checking

INPUT:
Star Trek had spin-off television series.
OUTPUT:
Supports
PROVENANCE:
17157886-3

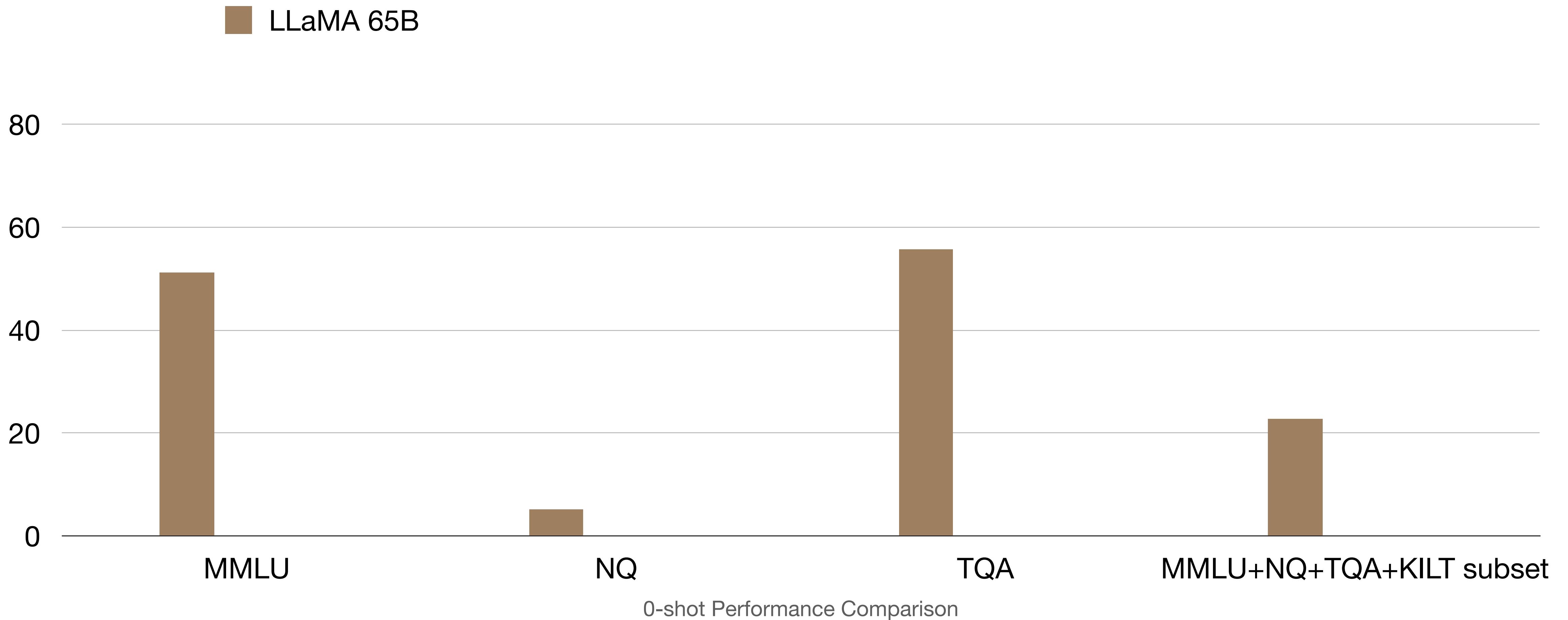
FEV

Entity Linking

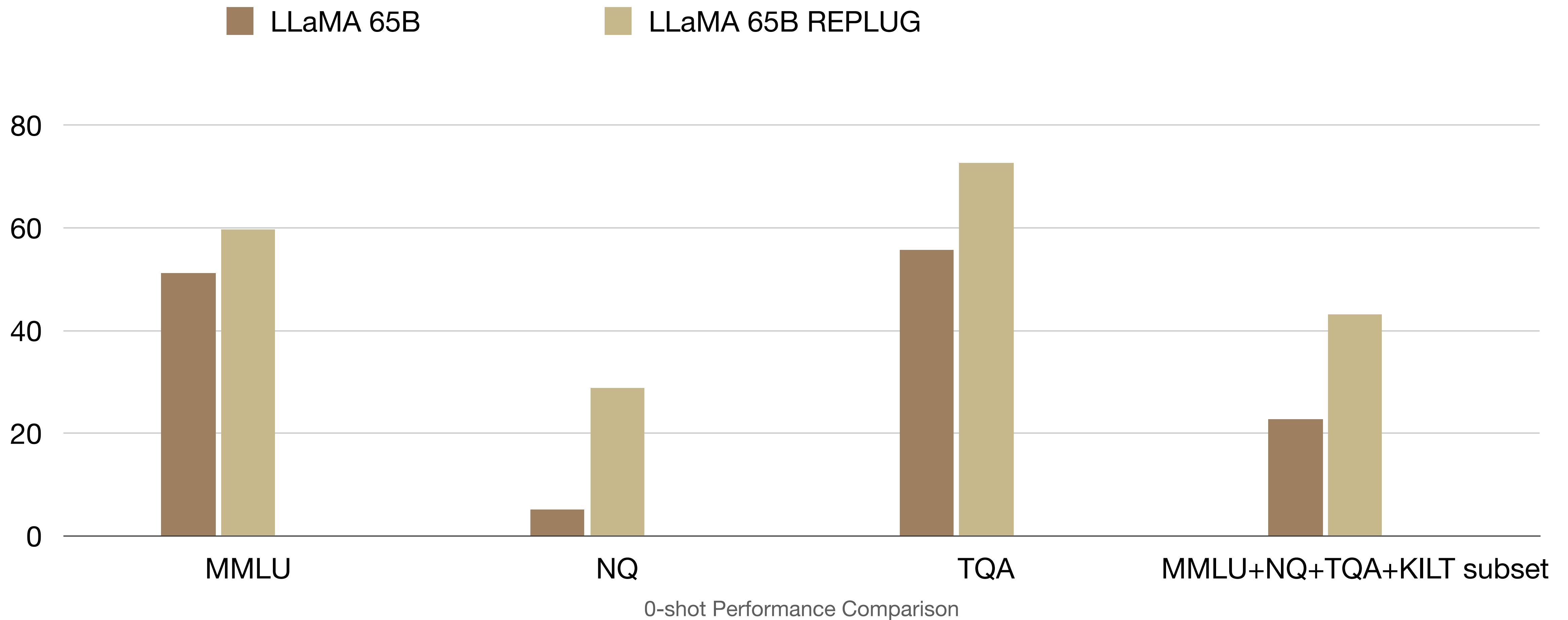
INPUT:
[...]Currently the site offers five movie collections ranging from \$149 for 10 [START_ENT]Star Trek [END_ENT] films to \$1,125 for the eclectic Movie Lovers' Collection of 75 movies. [...]
OUTPUT:
Star Trek
PROVENANCE:
17157886

CnWn

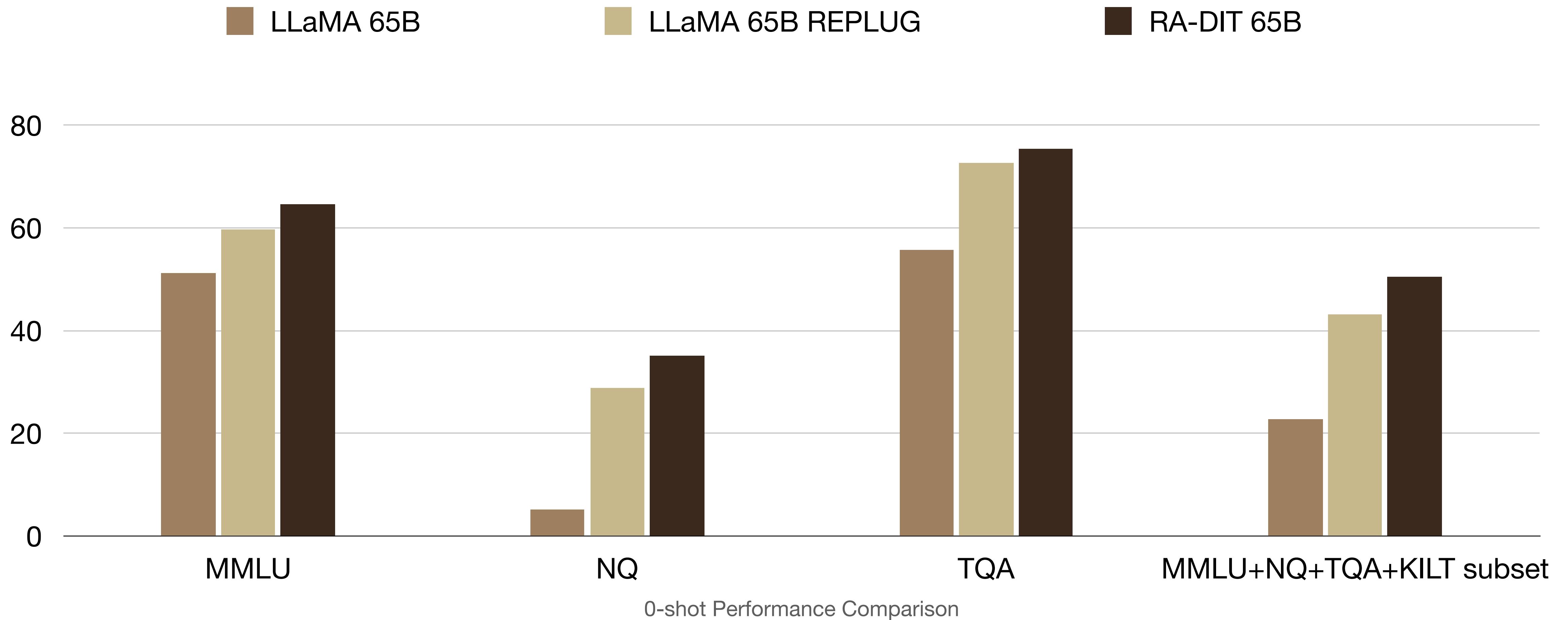
Performance on Knowledge Intensive Tasks



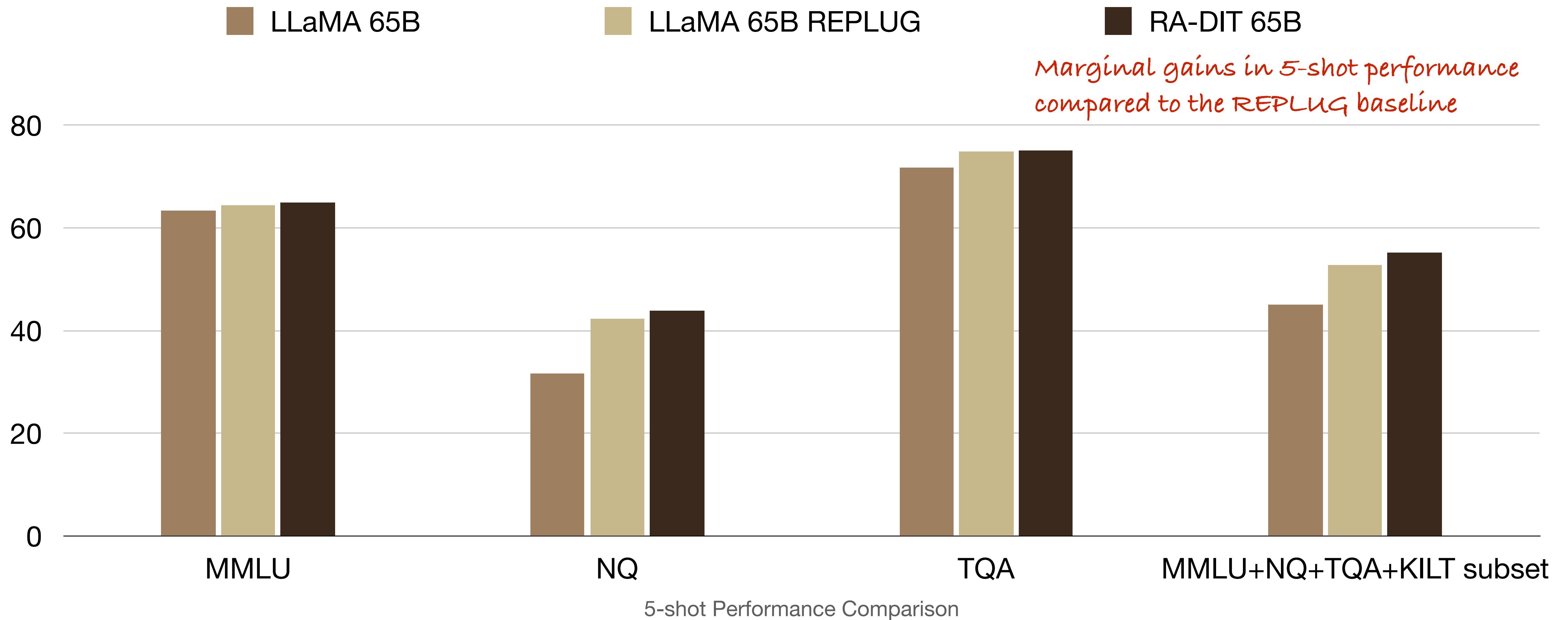
Performance on Knowledge Intensive Tasks



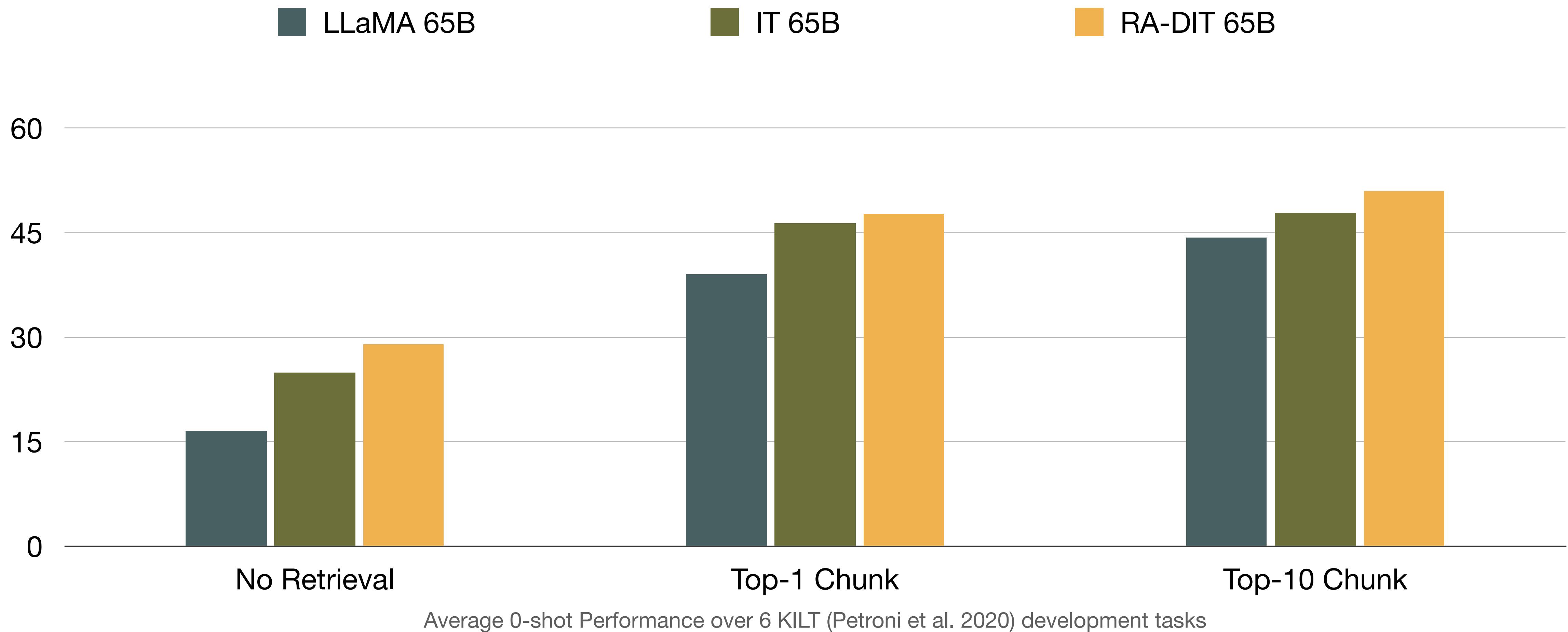
Performance on Knowledge Intensive Tasks



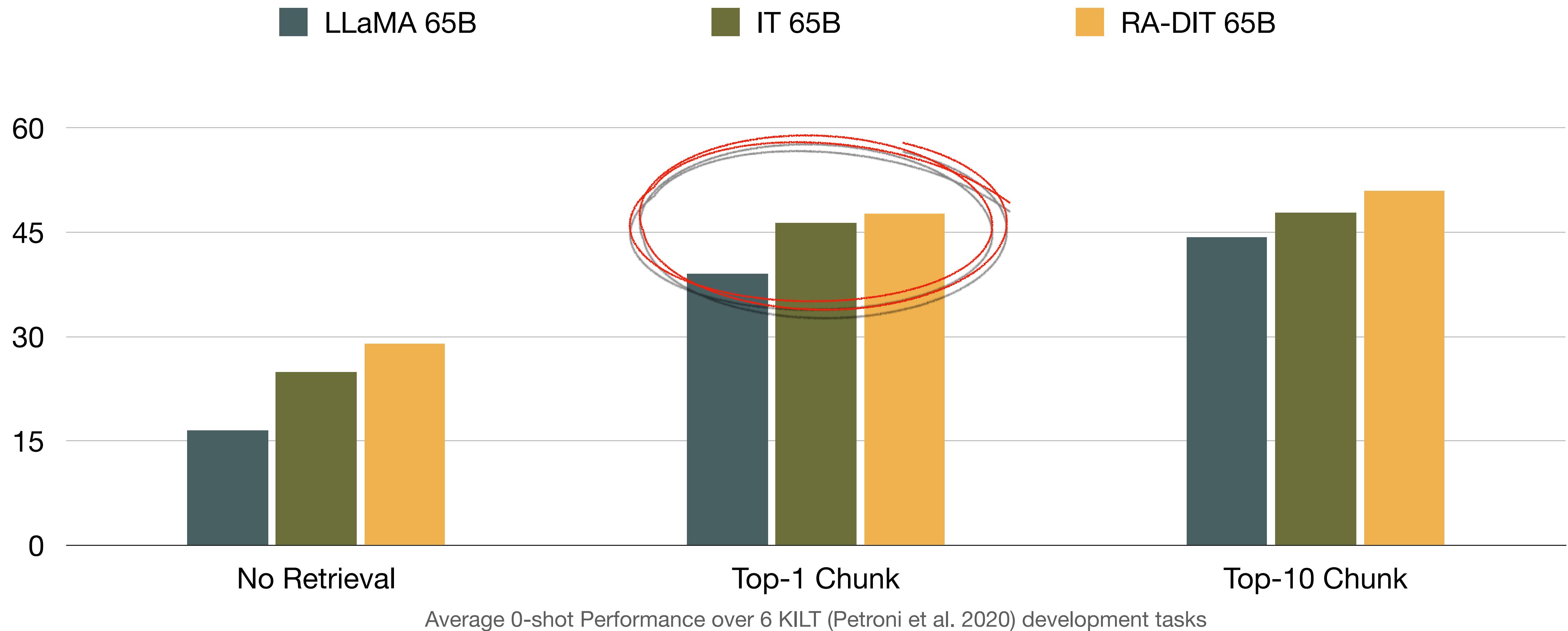
Performance on Knowledge Intensive Tasks



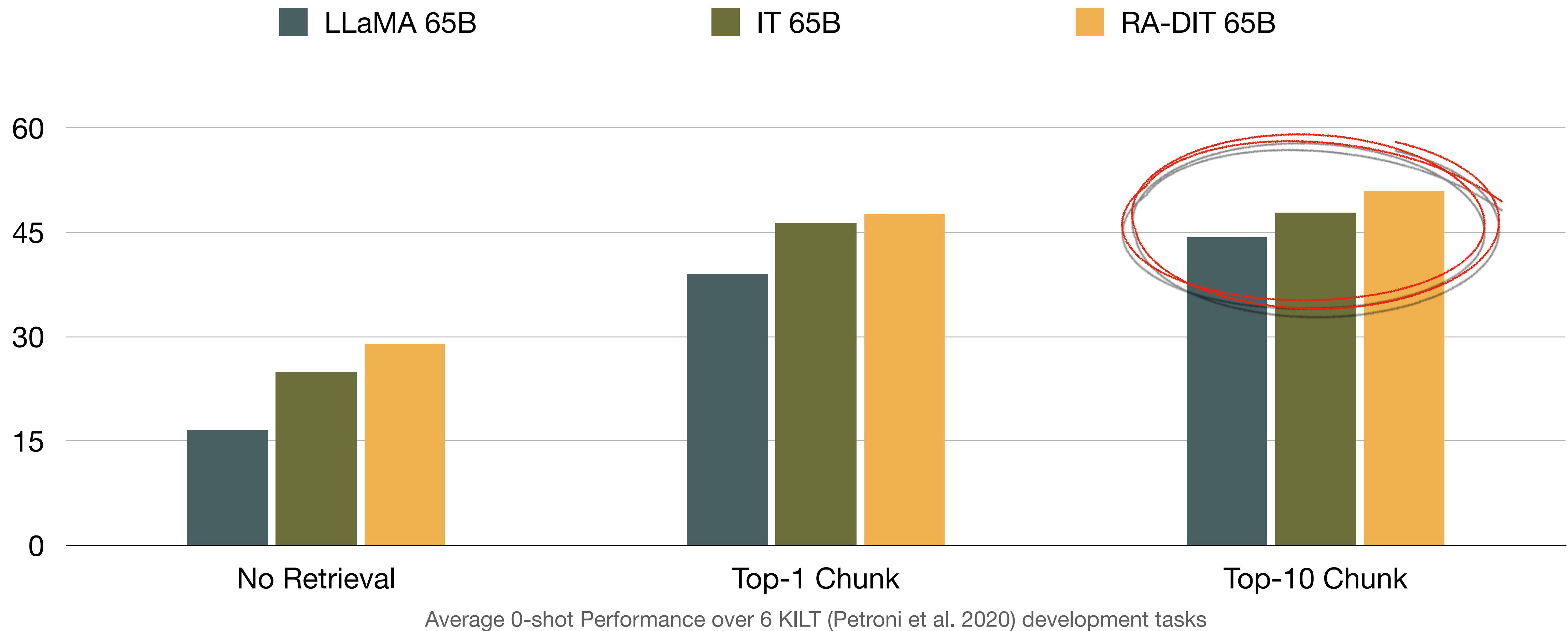
LM Fine-tuning Strategy Ablation



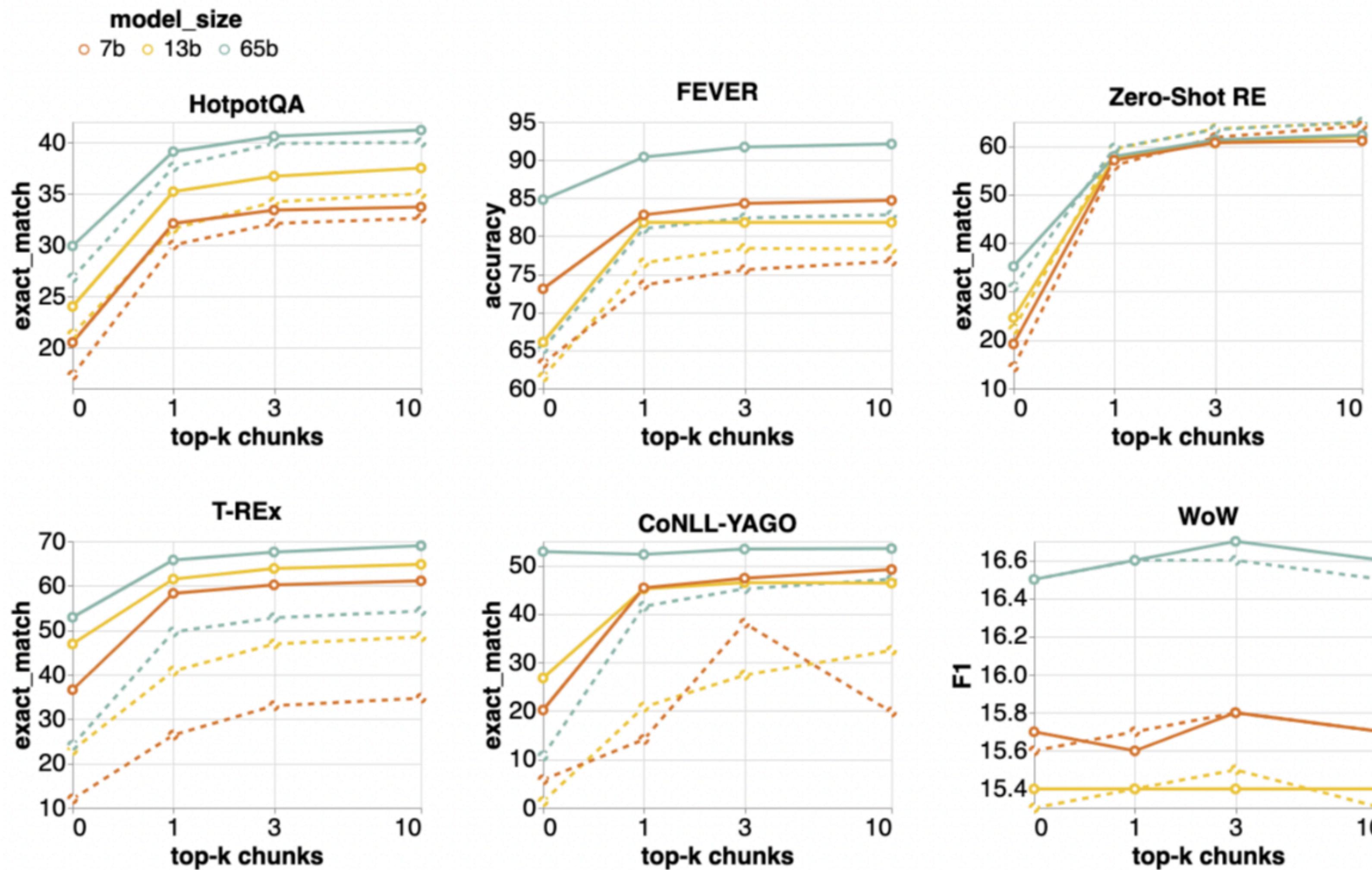
LM Fine-tuning Strategy Ablation



LM Fine-tuning Strategy Ablation



Scaling Law



Key Learnings

1. **In-context retrieval-augmented generation (RAG)** is by far the most effective and easy-to-implement technique for augmenting LLMs with external knowledge.
2. **Fine-tuning with retrieval augmentation** is an effective approach that can improve **the LLM, the retriever** and **their integration**, which is especially effective for domain-specific applications.
3. Scaling parametric capability of LLMs and improving retrieval-augmentation are complementary.

Tool Augmentation

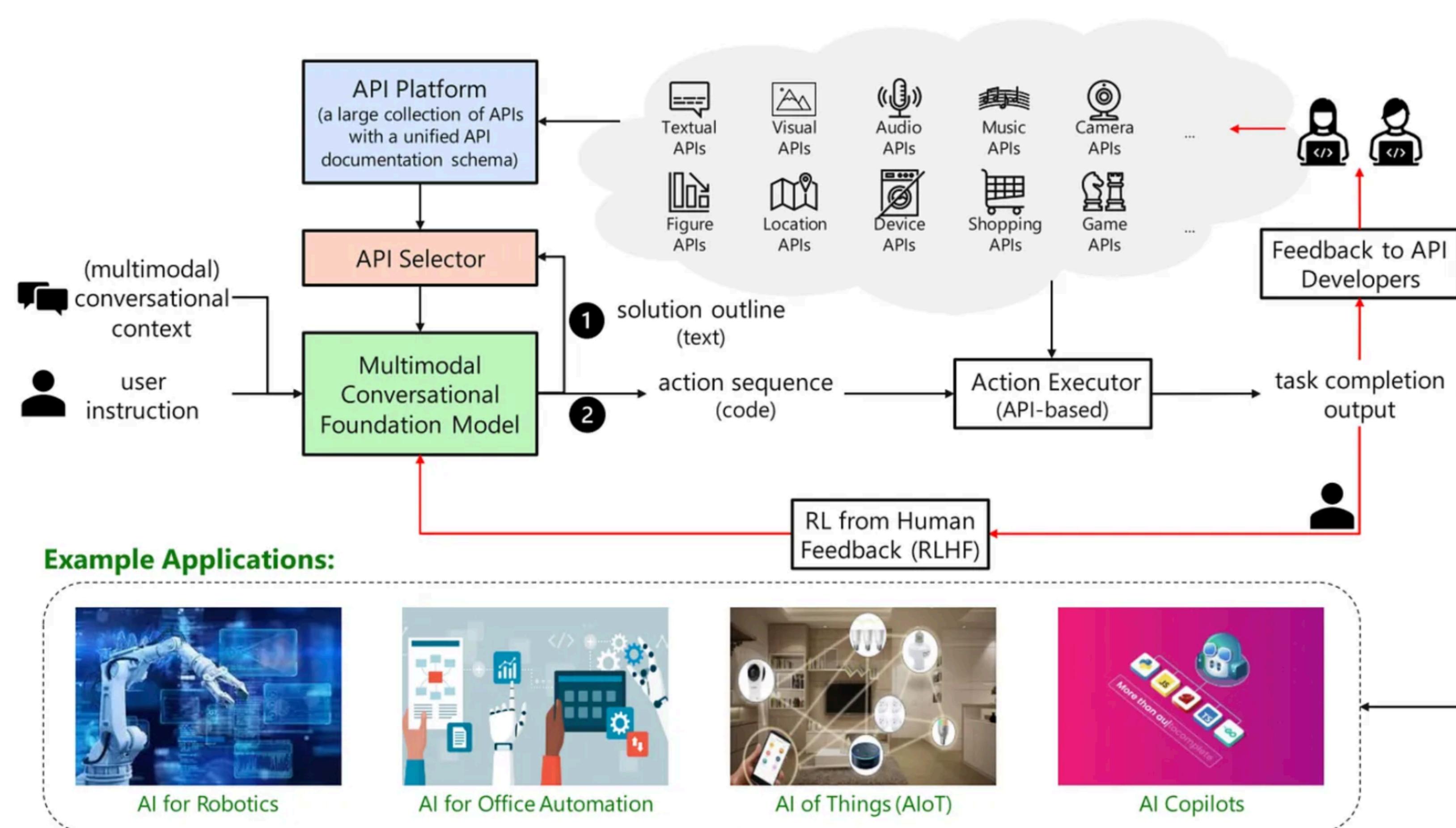


Image Credit: <https://newsletter.ruder.io/p/tool-augmented-langs>

Outline

- I. LLMs as General Parametric Task Solvers**
 - 1. LLM Pre-training and Scaling Laws
 - 2. Instruction Tuning
- 2. External Knowledge Access**
 - I) Retrieval Augmentation
 - II) Tool Augmentation
- 3. Quick Discussion: LLM Agents and Advanced Applications**
- 4. Conclusion**

LLM Agents



Image Credit: <https://github.com/WooooDyy/LLM-Agent-Paper-List/blob/main/assets/figure1.jpg>

Summary

1. Large language models (LLMs) were developed by effectively scaling up base neural network architectures such as transformers.
2. Through large-scale pre-training and alignment, LLMs demonstrate strong capabilities of following natural language instructions and solving tasks such as translation, writing enhancement and math puzzles.
3. Pure parametric LLMs suffers from knowledge cut-off, hallucination and accessing up-to-date/customized information. In-context augmentation can effectively supplement LLMs with external/in-situation knowledge and significantly improve their performances in knowledge-intensive scenarios.
4. We will likely witness LLM-powered AI agents deployed in many everyday tasks and assisting humans in many real-life scenarios that require complex and factual reasoning.

Thank you!
Questions?