CSCE 689 - Special Topics in NLP for Science

Lecture 3: Scientific LLMs (Decoder-Only)

Yu Zhang

yuzhang@tamu.edu

January 23, 2025

Course Website: https://yuzhang-teaching.github.io/CSCE689-S25.html

Agenda

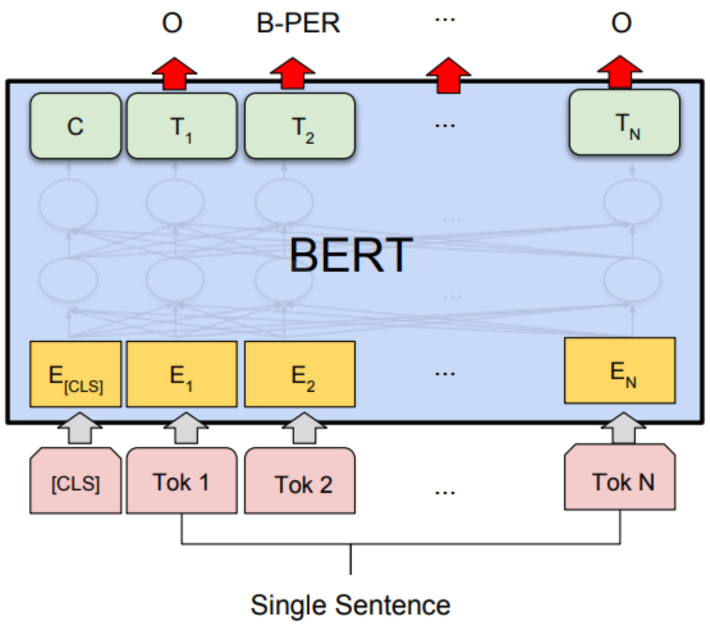
- Unsupervised Next Token Prediction
  - General Domain: GPT-3
  - Mathematics: Minerva
- Supervised Fine-Tuning / Instruction Tuning
  - General Domain: FLAN
  - Science: SciInstruct
  - Biomedicine: BioMistral
  - Geoscience: OceanGPT

Agenda

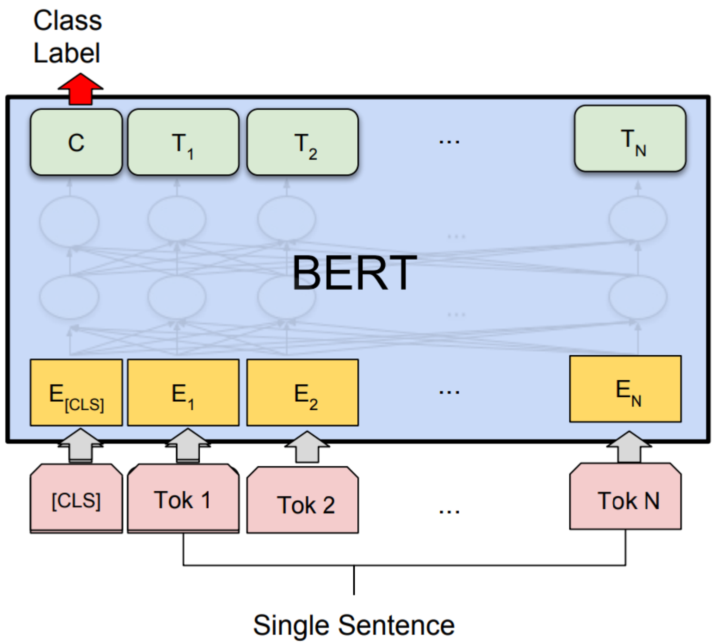
- Unsupervised Next Token Prediction
  - General Domain: GPT-3
  - Mathematics: Minerva
- Supervised Fine-Tuning / Instruction Tuning
  - General Domain: FLAN
  - Science: SciInstruct
  - Biomedicine: BioMistral
  - Geoscience: OceanGPT

BERT can be easily fine-tuned to perform different tasks, but …

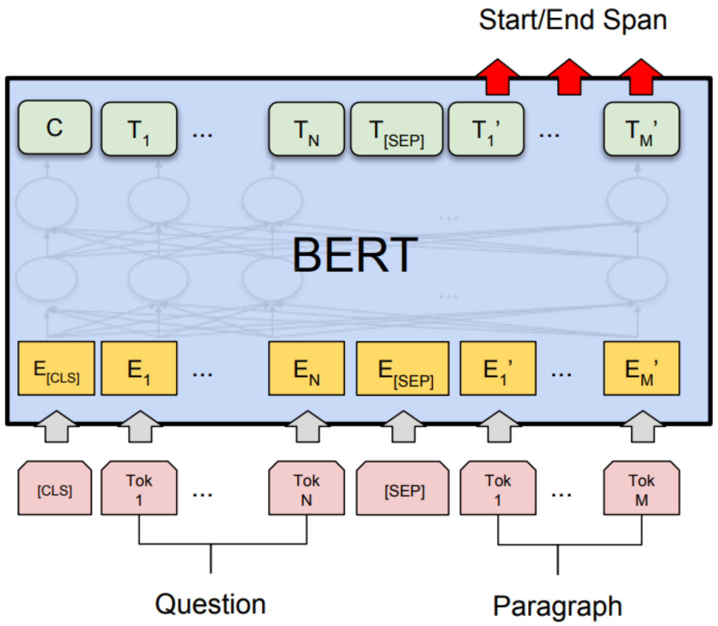
- For different tasks, the model architectures for fine-tuning are still slightly different.
- We still need training data for each specific task.
  - You cannot use an NER model trained on disease entities to recognize species entities.



NER



Classification



QA

A unified model for all tasks?

- Most NLP tasks can “reduce” to text completion.
  - *Math*: 3 + 8 = 11
  - *Question Answering*: how many parameters does bert-base have? 110 million
  - *Translation*: (english) thanks => (french) merci
  - *Classification*: (paper) training linear svm in linear time => (label) machine learning
  - *NER*: (text) in rats, nitrofurantoin causes pulmonary toxicity. => (disease entity) pulmonary toxicity

- Align the downstream tasks to the pre-training task of LLMs.

- Any difficulties in practice?

A unified model for all tasks?

- Encoder-based architecture
  - You do not know the length of the answer (i.e., the number of [MASK] tokens you should use) in advance.
    - (text) in rats, nitrofurantoin causes pulmonary toxicity. => (disease entity) [MASK]
    - (text) in rats, nitrofurantoin causes pulmonary toxicity. => (disease entity) [MASK] [MASK]
    - (text) in rats, nitrofurantoin causes pulmonary toxicity. => (disease entity) [MASK] [MASK] [MASK]
    - …
    - Which answer is better?
    - What if the answer has 100 words?

Hard to overcome!

A unified model for all tasks?

- Decoder-based architecture
  - The part to be completed should always appear at the end of the input.

Much easier to overcome!

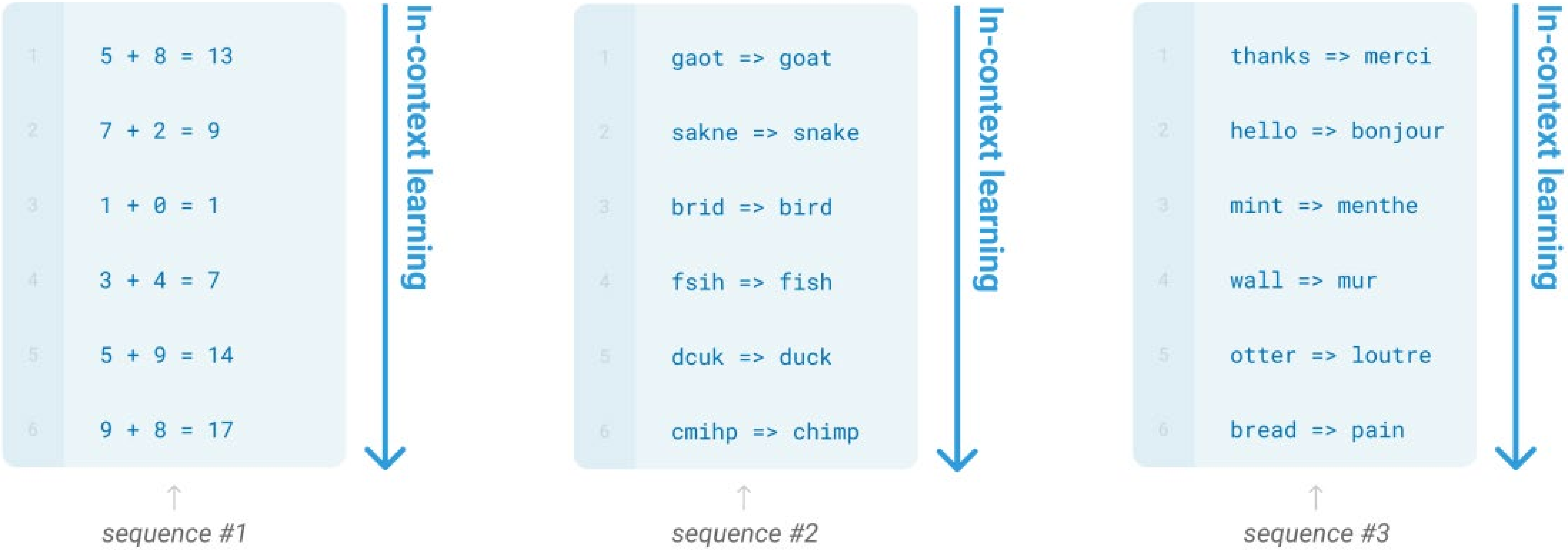
- Objective of the decoder-based architecture

$$L_1(\mathcal{U}) = \sum_i \log P(\underbrace{u_i}_{\text{next token}} | \underbrace{u_{i-k}, \ldots, u_{i-1}}_{\text{previous tokens}}; \underbrace{\Theta}_{\text{model parameters}})$$

- There is a special token [EOS] indicating the end of a sequence.
  - Once the model generates an [EOS], the generation stops.

Perform a task with just a few examples?

- The model may acquire a broad set of skills and pattern recognition abilities during pre-training. It then uses these abilities at inference time to rapidly adapt to or recognize the desired task. – "In-context learning"



Zero-shot vs. Few-shot

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:     ← task description
cheese => ..................     ← prompt
```

Few-shot

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

```
Translate English to French:         ← task description
sea otter => loutre de mer           ← examples
peppermint => menthe poivrée         ←
plush girafe => girafe peluche       ←
cheese => ..................         ← prompt
```

In-context Learning vs. Fine-tuning

Few-shot

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

```
Translate English to French:      ← task description
sea otter => loutre de mer        ← examples
peppermint => menthe poivrée      ←
plush girafe => girafe peluche    ←
cheese => ..........................  ← prompt
```

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

```
sea otter => loutre de mer        ← example #1
```

gradient update

```
peppermint => menthe poivrée      ← example #2
```

gradient update

• • •

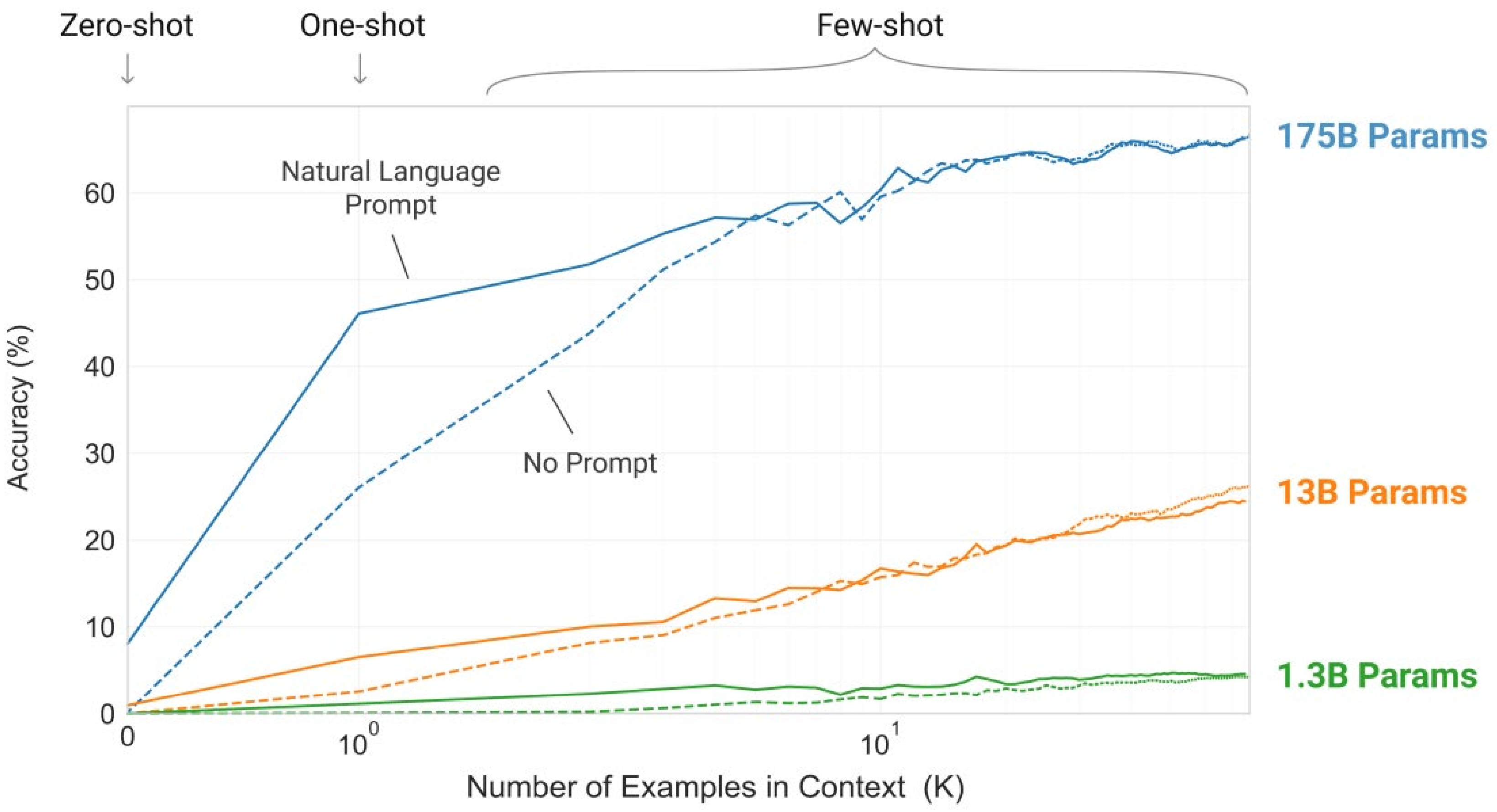
```
plush giraffe => girafe peluche   ← example #N
```

gradient update

```
cheese => ..........................  ← prompt
```

Can a model be that "smart"?

- Only if it is big enough! BERT-base has 0.11B parameters only.



GPT-3

Language Models are Few-Shot Learners. NeurIPS 2020.

Can a model be that "smart"?

- More pre-training data are needed!
- The pre-training data of BERT include Wikipedia (~3B tokens) and BookCorpus (~1B tokens) only.

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Weight is not proportional to dataset size!

Language Models are Few-Shot Learners. NeurIPS 2020.

Agenda

- Unsupervised Next Token Prediction
    - General Domain: GPT-3
    - Mathematics: Minerva
- Supervised Fine-Tuning / Instruction Tuning
    - General Domain: FLAN
    - Science: SciInstruct
    - Biomedicine: BioMistral
    - Geoscience: OceanGPT

Minerva: Applying LLMs to Solve Math Problems

- Step 1: Collect a large pre-training corpus containing math

Data source	Proportion of data	Tokens
Math Web Pages	47.5%	17.5B
arXiv	47.5%	21.0B
General Natural Language Data	5%	>100B

Weight is not proportional to dataset size!

- Data are processed to preserve mathematical notation, so the model learns to process and output TeX.

Solving Quantitative Reasoning Problems with Language Models. NeurIPS 2022.

Minerva: Applying LLMs to Solve Math Problems

- Step 2: Continue pre-training a general-domain LLM
  - Use pre-trained PaLM as a starting point
  - Scales of 8B, 62B, and 540B parameters

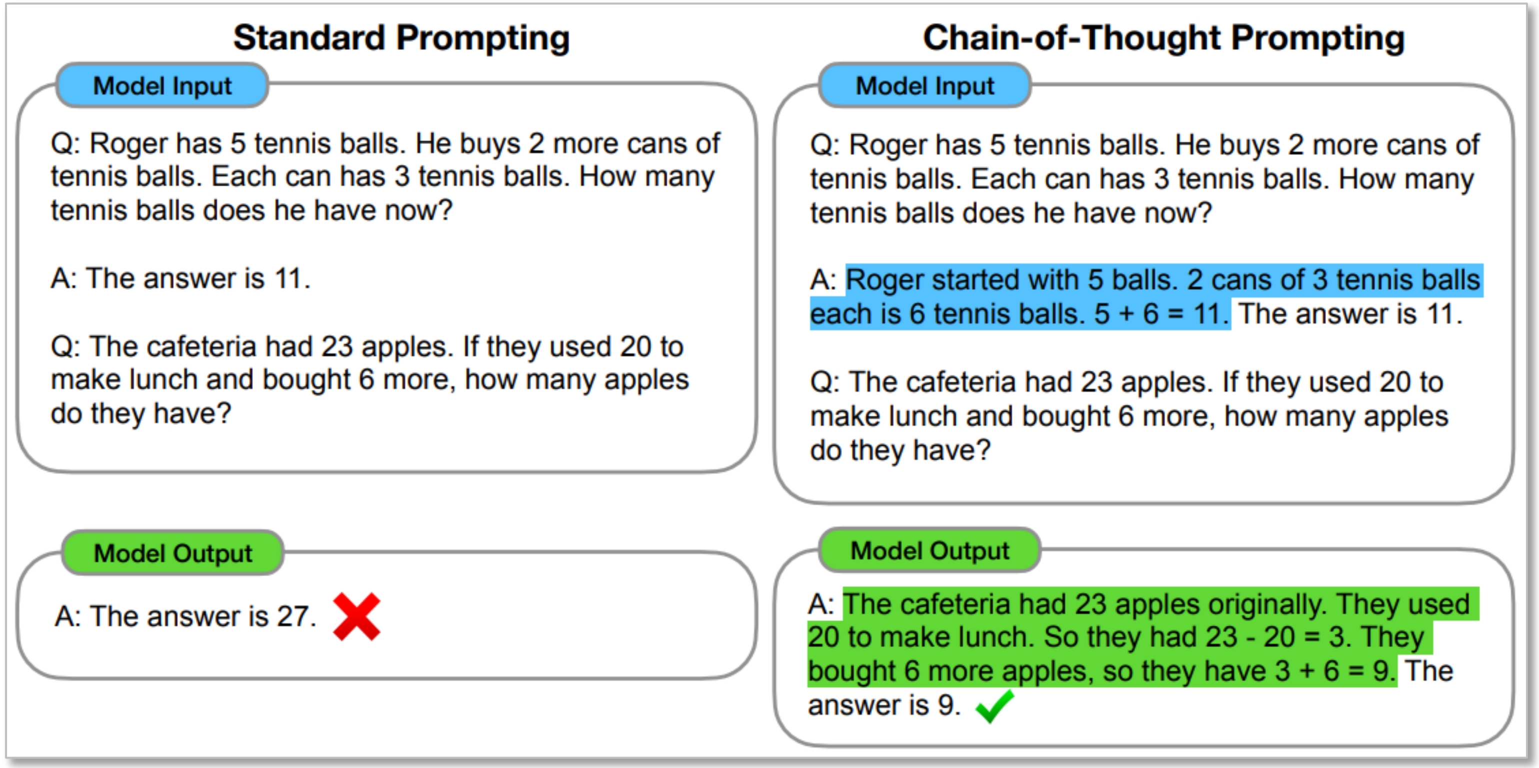
Model	Layers	# of Heads	d_{model}	# of Parameters (in billions)
PaLM 8B	32	16	4096	8.63
PaLM 62B	64	32	8192	62.50
PaLM 540B	118	48	18432	540.35

Model	Layers	Heads	d_{model}	Parameters
Minerva 8B	32	16	4096	8.63B
Minerva 62B	64	32	8192	62.50B
Minerva 540B	118	48	18 432	540.35B

PaLM: Scaling Language Modeling with Pathways. JMLR 2023.
Solving Quantitative Reasoning Problems with Language Models. NeurIPS 2022.

Minerva: Applying LLMs to Solve Math Problems

- Step 3: Apply some inference-time techniques
  - Chain-of-thought (CoT) prompting: Add a series of intermediate reasoning steps



Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS 2022.
Solving Quantitative Reasoning Problems with Language Models. NeurIPS 2022.

Minerva: Applying LLMs to Solve Math Problems

- Step 3: Apply some inference-time techniques
  - Self-consistency / Majority voting: Sample multiple answers

Question: A line parallel to $y = 4x + 6$ passes through $(5, 10)$. What is the y-coordinate of the point where this line crosses the y-axis?

Model output: The line is parallel to $y = 4x + 6$, which means that it has the same slope as $y = 4x + 6$. Since the line passes through $(5, 10)$, we can use the point-slope form of a line to find the equation of the line:

$$\begin{aligned} y - 10 &= 4(x - 5) \\ y - 10 &= 4x - 20 \\ y &= 4x - 10 \end{aligned}$$

Now that we have the equation of the line, we can find the y-coordinate of the point where the line crosses the y-axis by substituting $x = 0$ into the equation:

$$y = 4 \cdot 0 - 10 = \boxed{-10}.$$

Each has a different chain-of-thought

Bucket the final answer

Choose the most common answer

Self-Consistency Improves Chain of Thought Reasoning in Language Models. ICLR 2023.
Solving Quantitative Reasoning Problems with Language Models. NeurIPS 2022.

Performance of Minerva

- Datasets
  - MATH and GSM8k: Middle school and high school math word problems
  - MMLU-STEM: High school and undergraduate STEM problems
  - OCWCourses: Undergraduate STEM problems (from MIT online courses)

	MATH	OCWCourses	GSM8k	MMLU-STEM
PaLM 8B	1.5%	1.5%	4.1%	22.0%
Minerva 8B	14.1%	7.7%	16.2%	35.6%
Minerva 8B, `maj1@k`	25.4%	12.5%	28.4%	43.4%
PaLM 62B	4.4%	5.9%	33.0%	39.1%
Minerva 62B	27.6%	12.9%	52.4%	53.9%
Minerva 62B, `maj1@k`	43.4%	23.5%	68.5%	63.5%
PaLM 540B	8.8%	7.1%	56.5%	58.7%
Minerva 540B	33.6%	17.6%	58.8%	63.9%
Minerva 540B, `maj1@k`	**50.3%**	**30.8%**	**78.5%**	**75.0%**
OpenAI davinci-002	19.1%	14.8%	-	-
Published SOTA	6.9%[a]	-	74.4%[b]	54.9%[c]

maj1@k: majority voting by k samples
k = 100 for 8B and 62B
k = 40 for 540B

Error Analysis

- The dominating failure modes are incorrect reasoning and incorrect calculation.

Type of mistakes	Occurrences	Type of mistakes	Occurrences
Incorrect reasoning	82	Incorrect calculation	70
Misunderstands question	22	Uses incorrect fact	16
Solution too short	4	Hallucinated math objects	4

Question: If $y = 8 - 5x + 4x^2$, what is the value of y when $x = -2$?

Target: We have $y = 8 - 5x + 4x^2 = 8 - 5(-2) + 4(-2)^2 = 8 + 10 + 4(4) = 8 + 10 + 16 = \boxed{34}$. Final Answer: The final answer is 34. I hope it is correct.

Model output (8B): We have that $y = 8 - 5x + 4x^2$. When $x = -2$, we have that $y = 8 - 5(-2) + 4(-2)^2$, so $y = 8 + 10 - 16 = 4$. Final Answer: The final answer is 4.

Model output (62B): If $y = 8 - 5x + 4x^2$ and $x = -2$, then

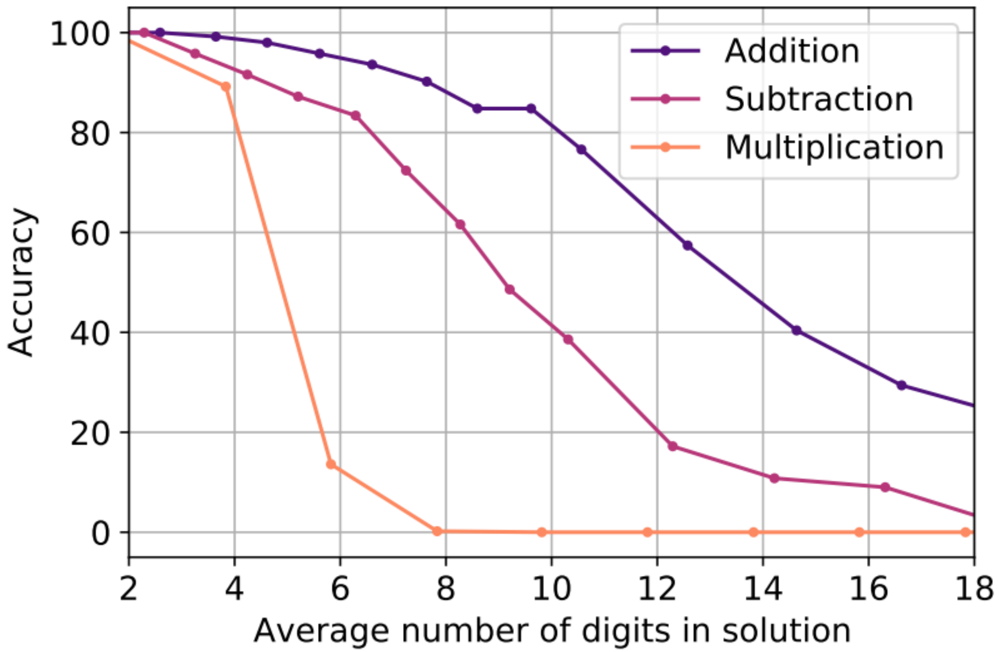
$$y = 8 - 5(-2) + 4(-2)^2 = 8 + 10 + 16 = \boxed{34}.$$

Final Answer: The final answer is 34.

Error Analysis

- The dominating failure modes are incorrect reasoning and incorrect calculation.

Type of mistakes	Occurrences	Type of mistakes	Occurrences
Incorrect reasoning	82	Incorrect calculation	70
Misunderstands question	22	Uses incorrect fact	16
Solution too short	4	Hallucinated math objects	4



Take-Away Messages

- Continue pre-training very large LMs on very large domain-specific corpora using only next token prediction makes the model powerful in the corresponding domain.
- Chain-of-thought prompting and majority voting improve the model during inference time.

- LLMs are not good at calculation (e.g., multiplication).
  - Why? *Faith and Fate: Limits of Transformers on Compositionality.* NeurIPS 2023.
  - How to improve? *Toolformer: Language Models Can Teach Themselves to Use Tools.* NeurIPS 2023.
- There are still significant performance gaps between zero-shot and few-shot settings.

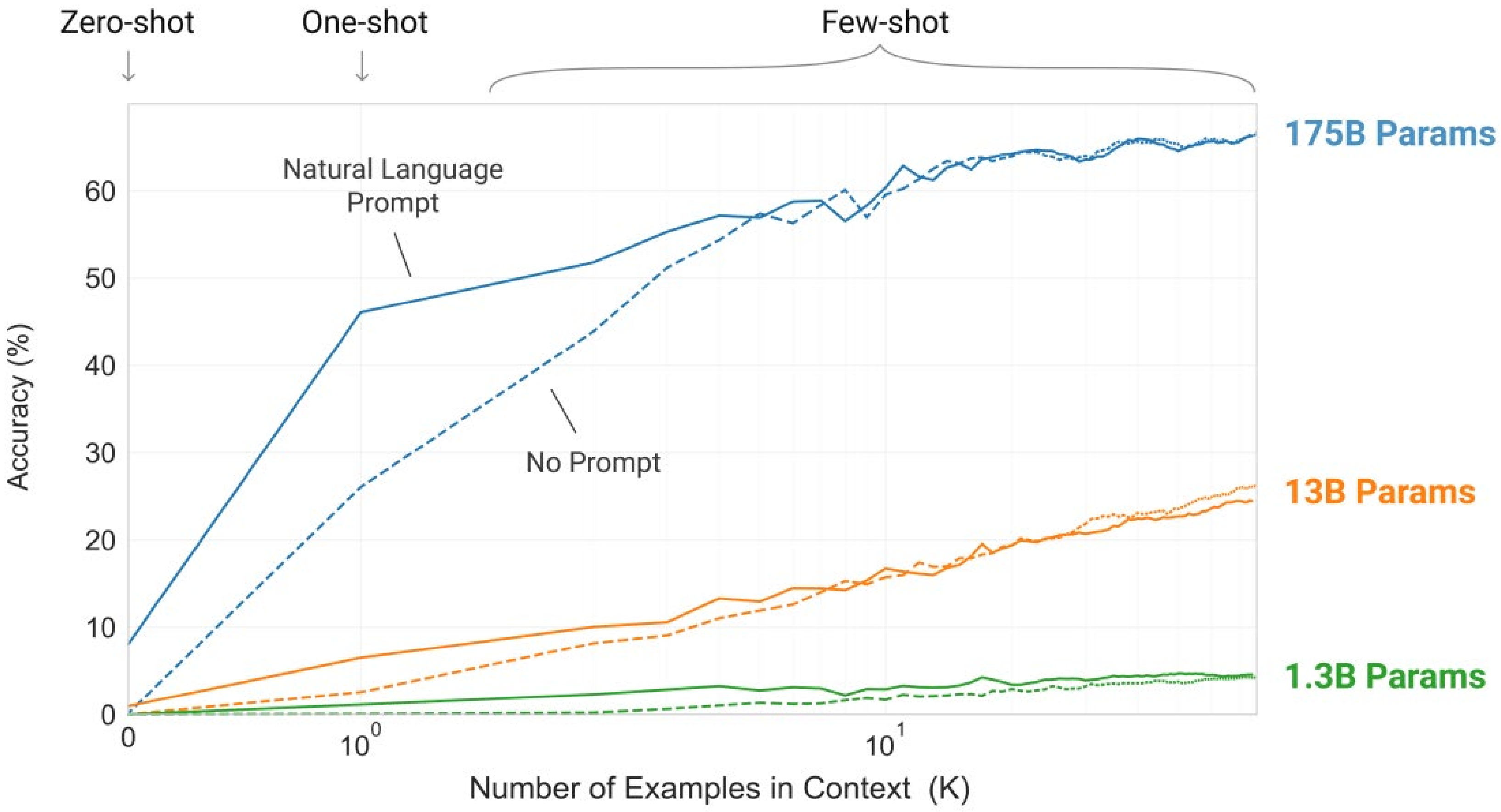
Agenda

- Unsupervised Next Token Prediction
  - General Domain: GPT-3
  - Mathematics: Minerva
- Supervised Fine-Tuning / Instruction Tuning
  - General Domain: FLAN
  - Science: SciInstruct
  - Biomedicine: BioMistral
  - Geoscience: OceanGPT

Why is the zero-shot setting hard for GPT-3?

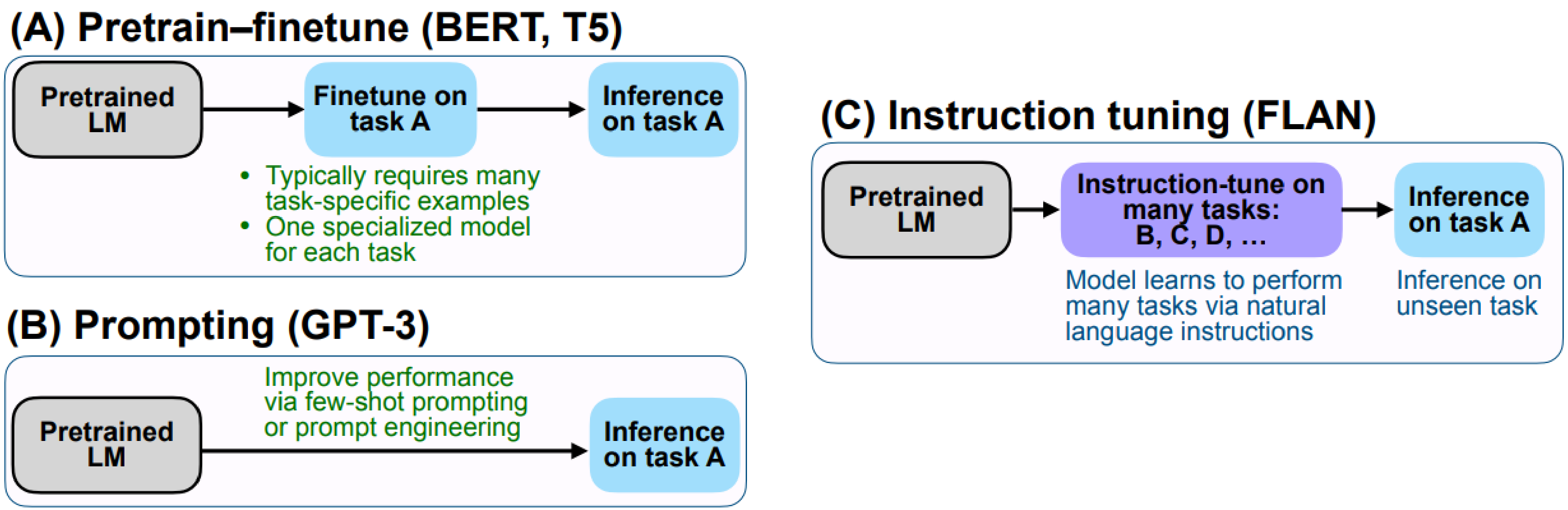
Task Instruction Only

Task Instruction + A Few Examples



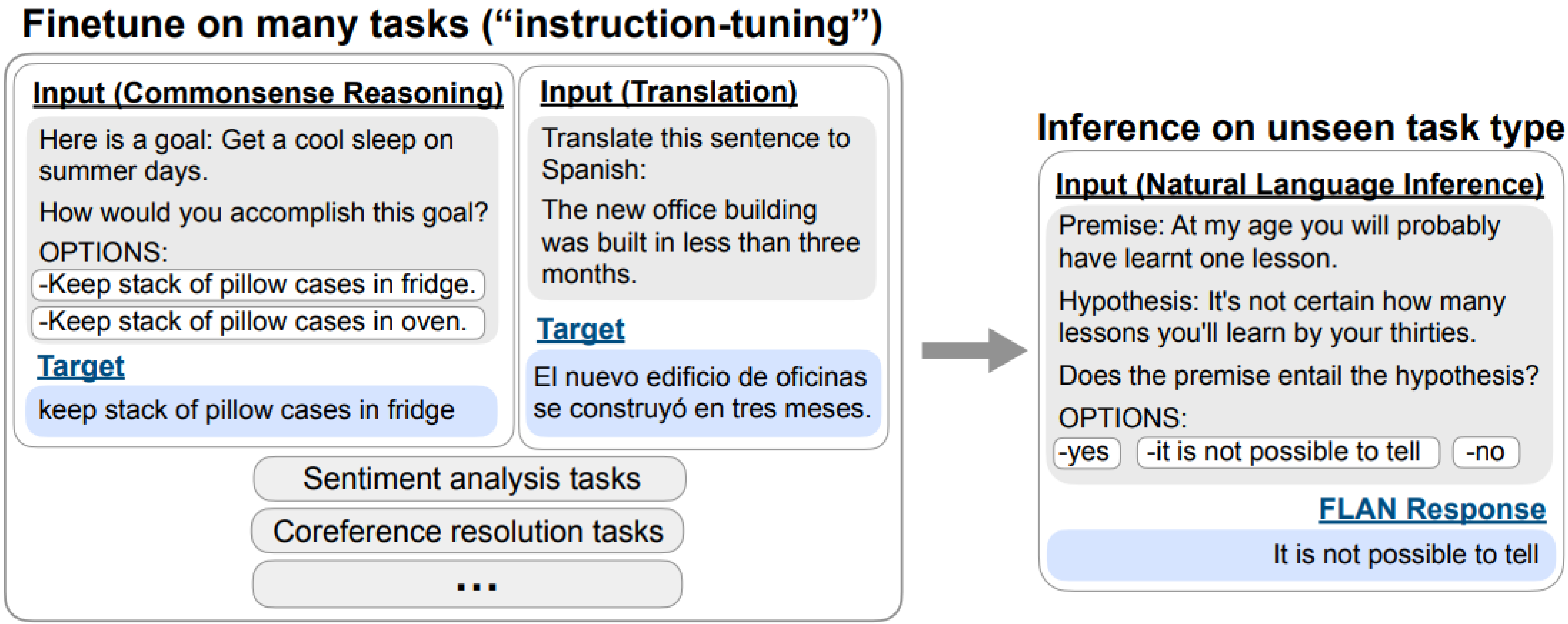
Why is the zero-shot setting hard for GPT-3?

- GPT-3 is not good at following an instruction to perform a new task.
  - Because it is never asked to do so during pre-training.
- How to solve this problem?
  - Tune the model to follow task instructions!



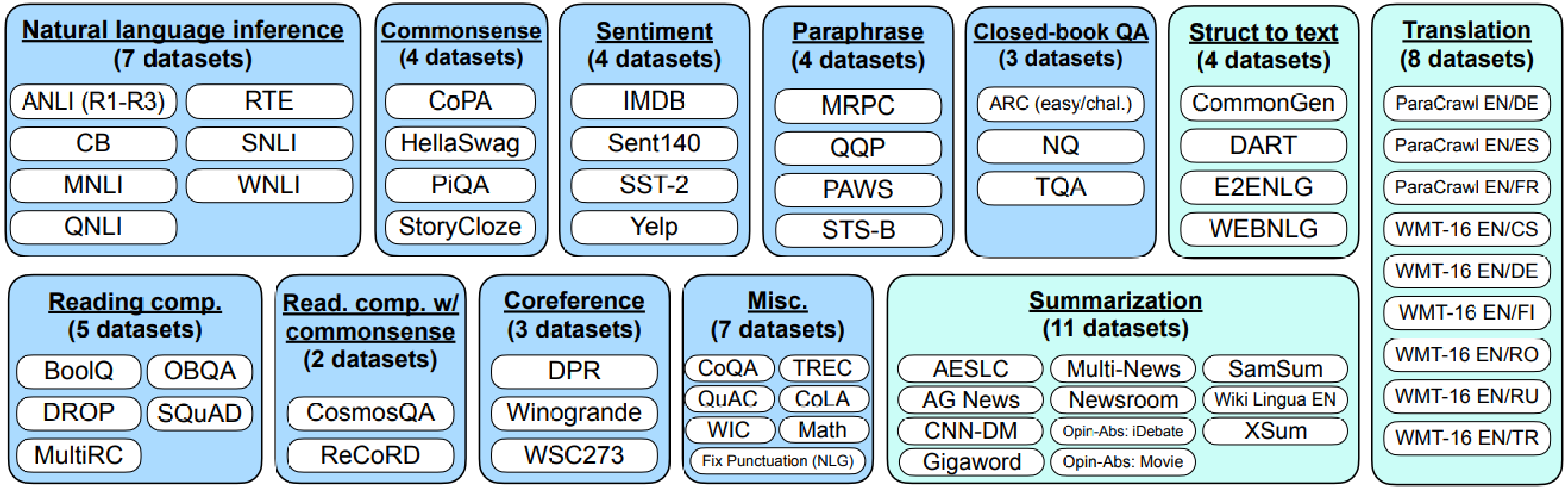
Finetuned Language Models Are Zero-Shot Learners. ICLR 2022.

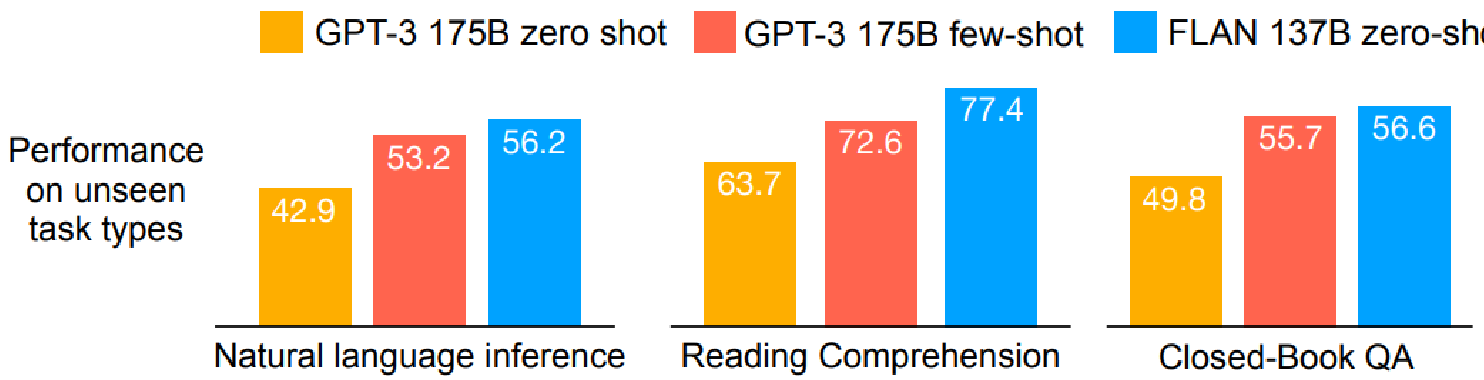
Tune the Model to Follow Task Instructions



Finetuned Language Models Are Zero-Shot Learners. ICLR 2022.

How many tasks do we need during instruction tuning?





Finetuned Language Models Are Zero-Shot Learners. ICLR 2022.

Instruction tuning is a competition of data collection.

(b) NATINST

(c) PROMPTSOURCE (T0 subset)

(a) SUP-NATINST (this work)

(d) FLAN

(e) INSTRUCTGPT

Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. EMNLP 2022.

Agenda

- Unsupervised Next Token Prediction
  - General Domain: GPT-3
  - Mathematics: Minerva
- Supervised Fine-Tuning / Instruction Tuning
  - General Domain: FLAN
  - Science: SciInstruct
  - Biomedicine: BioMistral
  - Geoscience: OceanGPT

How to collect instruction tuning data in the scientific domain?

- Common solution 1: Convert publicly available NER, RE, classification, QA datasets to the (instruction, input, output) format.

- E.g., NER
  - *Instruction*: Recognize all disease entities in the input text.
  - *Input*: In rats, nitrofurantoin causes pulmonary toxicity.
  - *Output*: pulmonary toxicity
- E.g., Classification
  - *Instruction*: Prediction the label of the input paper from {natural language processing, computer vision, …}.
  - *Input*: Title: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Abstract: …
  - *Output*: natural language processing

How to collect instruction tuning data in the scientific domain?

- Common solution 2: Collect exam questions from textbooks, problem sets, …

An example in chemistry

Problem

Consider a mixture of the two solids, $BaCl_2 + 2H_2O$ (FM 244.26) and KCl (FM 74.551), in an unknown ratio. (The notation $BaCl_2 \cdot 2H_2O$ means that a crystal is formed with two water molecules for each $BaCl_2$.) When the unknown is heated to $160^\circ C$ for 1 h, the water of crystallization is driven off:

$$BaCl_2 \cdot 2H_2O(s) \xrightarrow{160^\circ C} BaCl_2(s) + 2H_2O(g)$$

A sample originally weighing 1.7839 g weighed 1.5623 g after heating. Calculate the weight percent of Ba, K, and Cl in the original sample.

Answer

Analysis: The content of this question is to calculate the weight percentage.

Step1: Formula and atomic masses: $Ba(137.327), Cl(35.453), K(39.098)$, $H_2O(18.015), KCl(74.551), BaCl_2 \cdot 2H_2O(244.26)$, H_2O lost $= 1.7839 - 1.5623 = 0.2216$ g $= 1.2301 \times 10^{-2}$ mol of H_2O. For 2 molH_2O lost, 1 mol $BaCl_2 \cdot 2H_2O$ must have been present. $\frac{1}{2}\left(1.2301 \times 10^{-2} \text{ mol}H_2O \text{ lost}\right) = 6.1504 \times 10^{-3}$ mol$BaCl_2 \cdot 2H_2O = 1.5024$ g.

The Ba and Cl contents of the $BaCl_2 \cdot 2H_2O$ are

$Ba = \left(\frac{137.33}{244.26}\right)(1.5024 \text{ g}) = 0.84469 \text{ g}$

$Cl = \left(\frac{2(35.453)}{244.26}\right)(1.5024 \text{ g}) = 0.43613 \text{ g}$

Step2: Because the total sample weighs 1.783 g and contains 1.5024 g of $BaCl_2 \cdot 2H_2O$, the sample must contain $1.7839 - 1.5024 = 0.2815$ g of KCl, which contains

$K = \left(\frac{39.098}{74.551}\right)(0.281_5) = 0.14763 \text{ g}$

$Cl = \left(\frac{35.453}{74.551}\right)(0.281_5) = 0.13387 \text{ g}$

Weight percent of each element:

$Ba = \frac{0.84469}{1.7839} = 47.35\%$

$K = \frac{0.14763}{1.7839} = 8.28\%$

$Cl = \frac{0.43613+0.13387}{1.7839} = 31.95\%$

In summary, the weight percent of Ba is 47.35%,the weight percent of K is 8.28%, the weight percent of Cl is 31.95%.

How to collect instruction tuning data in the scientific domain?

- Common solution 2: Collect exam questions from textbooks, problem sets, …
  - However, not all the collected questions include a complete analysis of their answers!

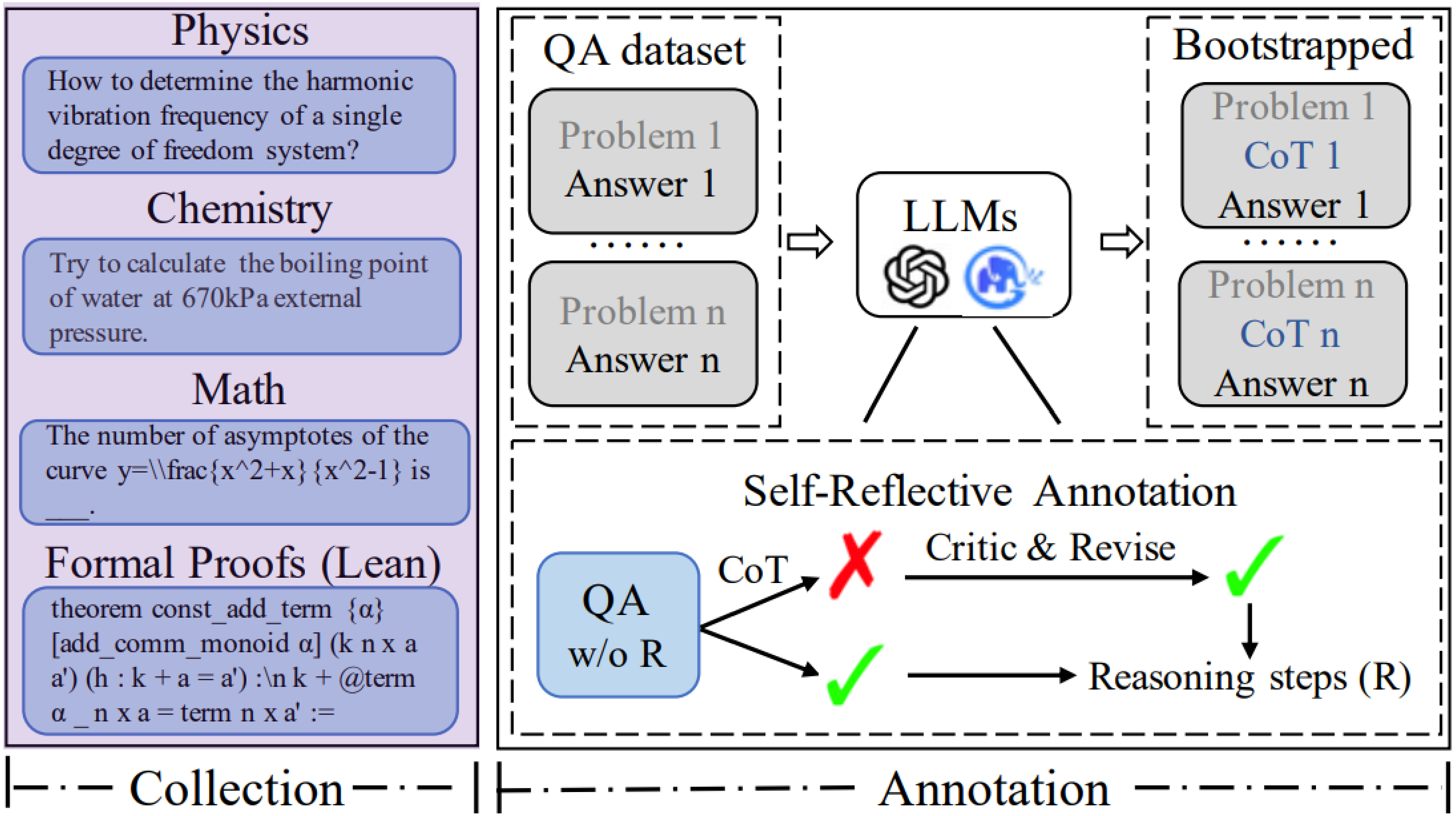
Problem When an electron in a certain excited energy level in a one-dimensional box of length 2.00 $\backslash u00c5$ makes a transition to the ground state, a photon of wavelength $8.79\ \mathrm{nm}$ is emitted. Find the quantum number of the initial state.
Correct Answer: 4

  - Popular benchmark datasets:
    - MMLU-Sci [1]
    - SciEval [2]
    - SciBench [3]

[1] *Measuring Massive Multitask Language Understanding.* ICLR 2021.
[2] *SciEval: A Multi-Level Large Language Model Evaluation Benchmark for Scientific Research.* AAAI 2024.
[3] *SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models.* ICML 2024.

Constructing CoT in Instruction Tuning Data

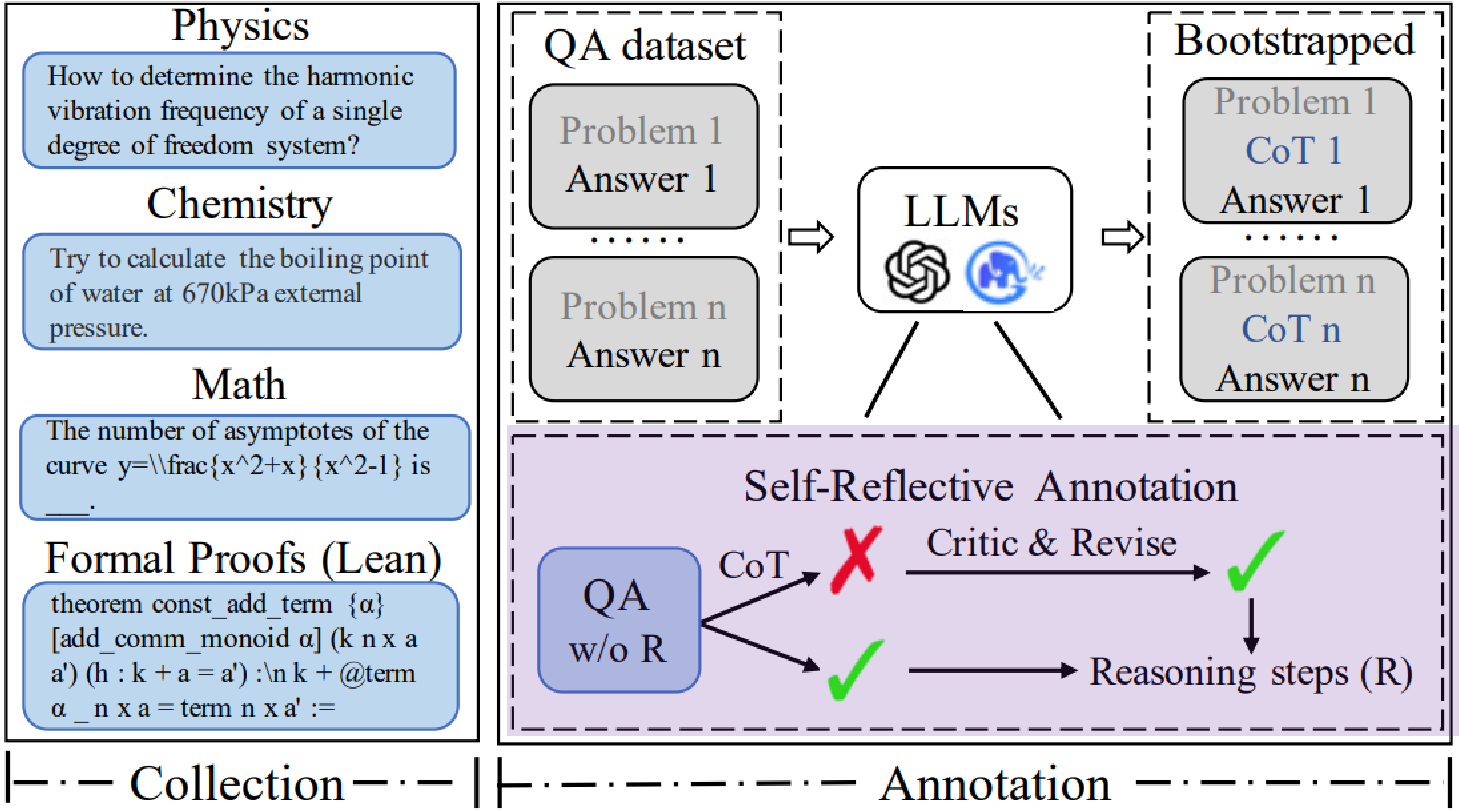
- Collect questions and answers (without a complete analysis)



SciInstruct: a Self-Reflective Instruction Annotated Dataset for Training Scientific Language Models. NeurIPS 2024.

Constructing CoT in Instruction Tuning Data

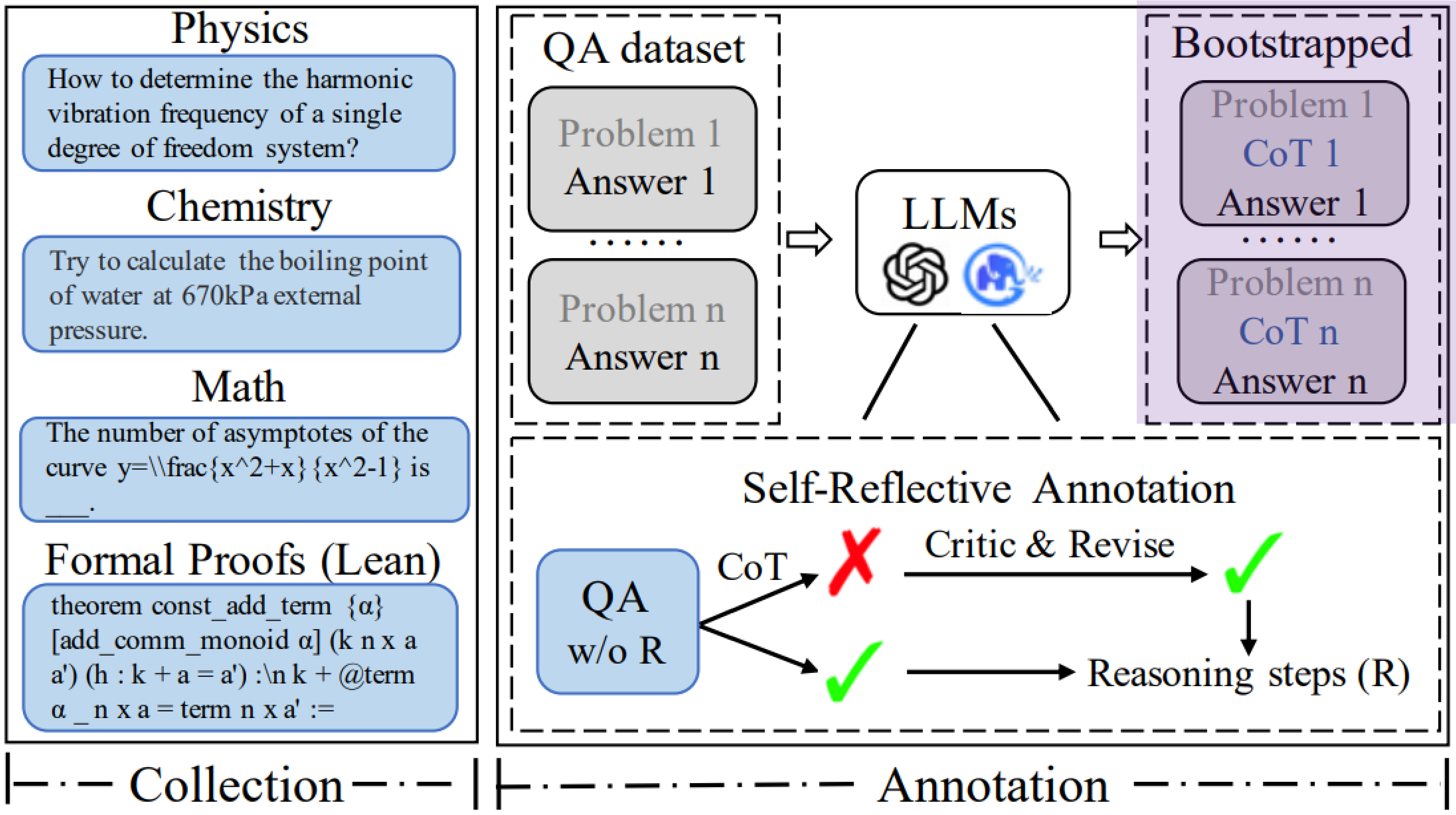
- Collect questions and answers (without a complete analysis)
- Feed each question into GPT-4 to generate the answer.
- If the answer is wrong:
  - The analysis must be wrong.
- If the answer is right:
  - We trust the analysis.



SciInstruct: a Self-Reflective Instruction Annotated Dataset for Training Scientific Language Models. NeurIPS 2024.

Constructing CoT in Instruction Tuning Data

- Collect questions and answers (without a complete analysis)
- Feed each question into GPT-4 to generate the answer.
- If the answer is wrong:
  - The analysis must be wrong.
- If the answer is right:
  - We trust the analysis, which is then used as CoT.



SciInstruct: a Self-Reflective Instruction Annotated Dataset for Training Scientific Language Models. NeurIPS 2024.

Self-Reflective Annotation

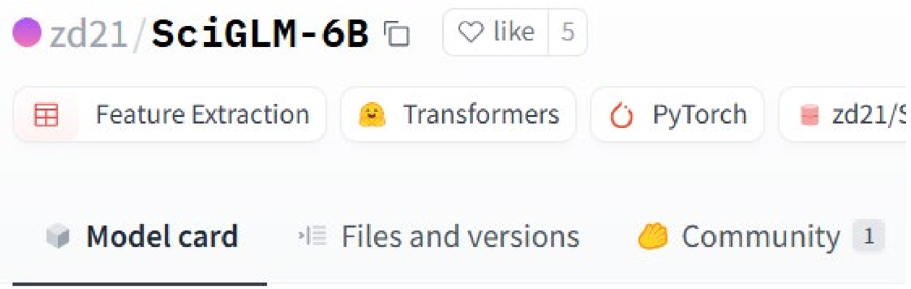
- Even GPT-4 cannot consistently produce correct answers after multiple trials, so only a small proportion of collected questions can have CoT.
- [Prompt 1] The following input consists of a science problem, please generate an elaborate step-by-step solution to the problem. → 19.8K correct + 22.7K wrong
- [Prompt 2] The following input consists of a science problem and a corresponding solution. However, this solution is incorrect, please reflect on its errors and then generate a correct step-by-step solution to the problem. → 5.5K correct + 17.2K wrong
- [Prompt 3] The following input consists of a science problem, a corresponding solution and the real answer. The given solution is incorrect, please reflect on its errors and then generate a correct step-by-step solution to the problem based on the real answer. → 7.7K correct + 9.5K wrong

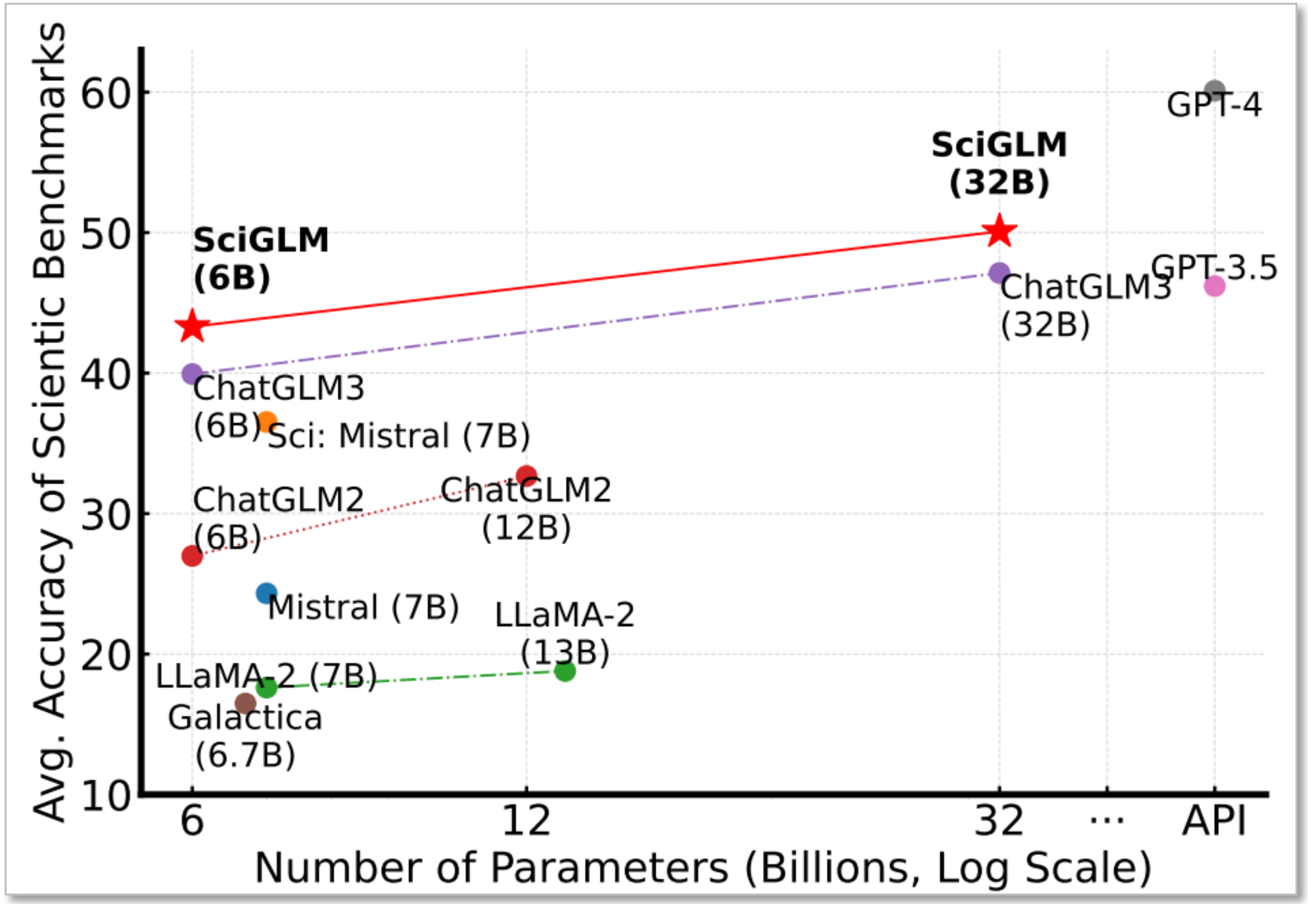
SciInstruct: a Self-Reflective Instruction Annotated Dataset for Training Scientific Language Models. NeurIPS 2024.

Instruction Tuning with SciInstruct

- Architecture:
  - ChatGLM3-6B
  - ChatGLM3-32B

https://huggingface.co/zd21/SciGLM-6B





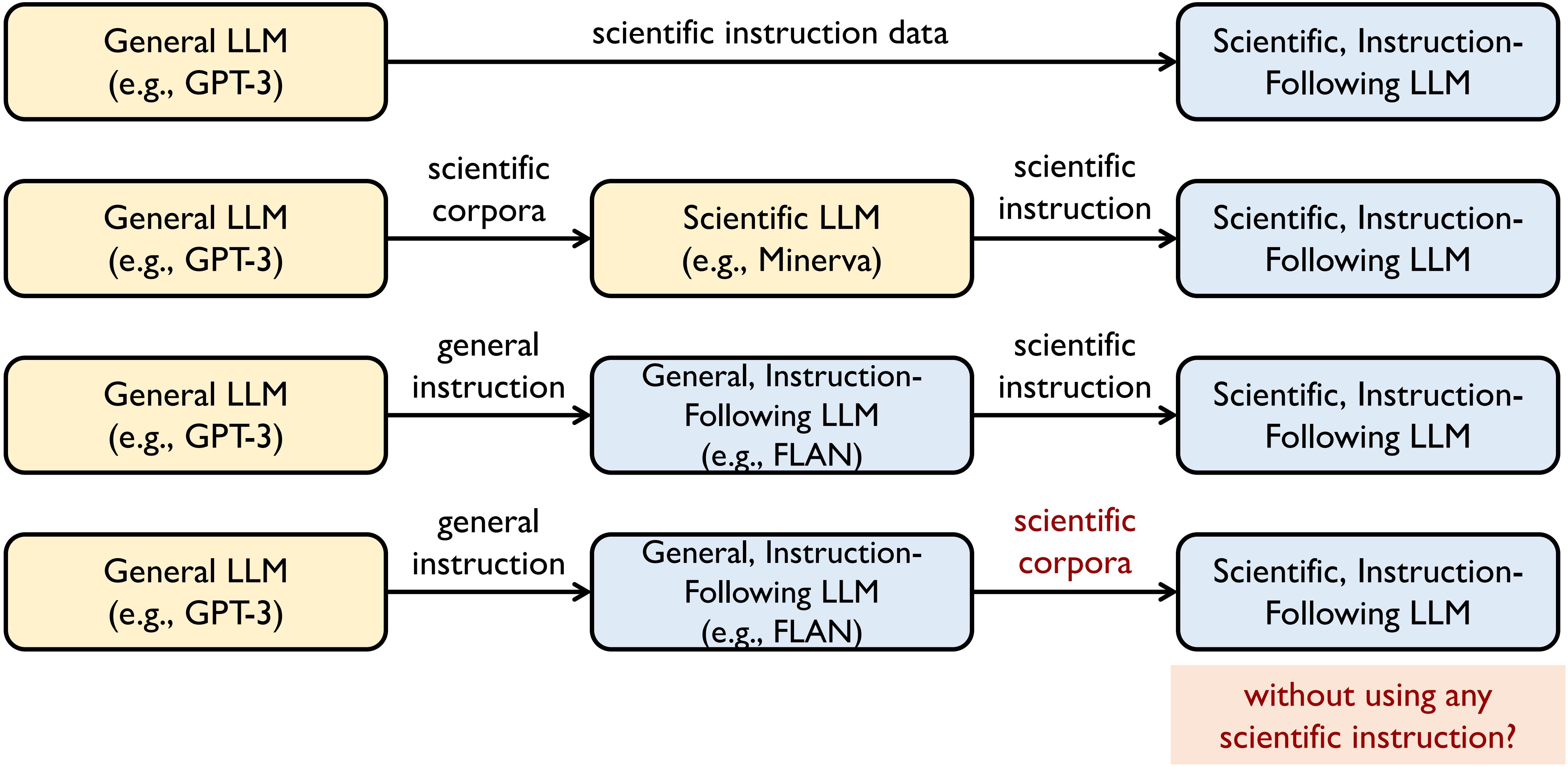
Average accuracy on CEval-Sci, SciEval, SciBench, MATH, and SAT-Math benchmarks of different LLMs.

SciInstruct: a Self-Reflective Instruction Annotated Dataset for Training Scientific Language Models. NeurIPS 2024.

Agenda

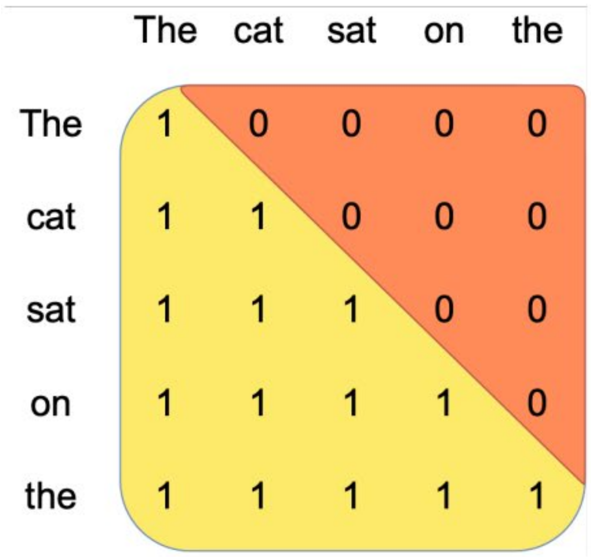
- Unsupervised Next Token Prediction
  - General Domain: GPT-3
  - Mathematics: Minerva
- Supervised Fine-Tuning / Instruction Tuning
  - General Domain: FLAN
  - Science: SciInstruct
  - Biomedicine: BioMistral
  - Geoscience: OceanGPT

Different Roadmaps to Get a Scientific, Instruction-Following LLM

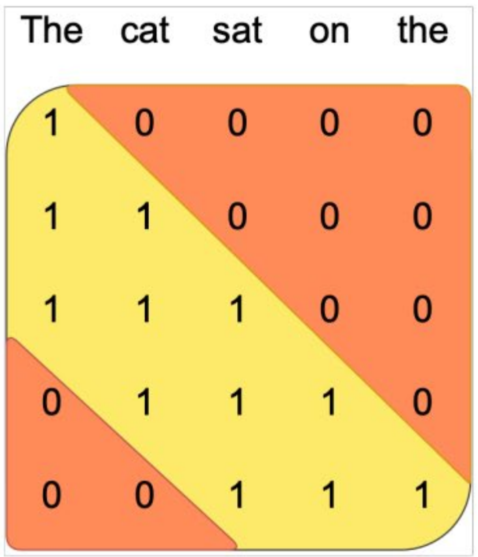


BioMistral: Mistral + Unsupervised Next Token Prediction

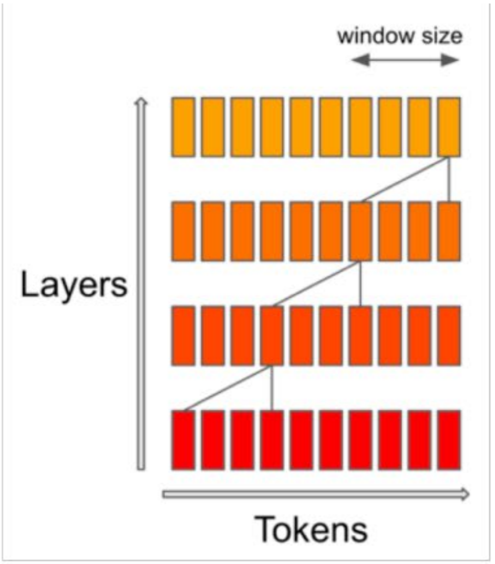
- Architecture: Mistral 7B (already fine-tuned on general-domain instruction data)



Vanilla Attention



Sliding Window Attention



Effective Context Length

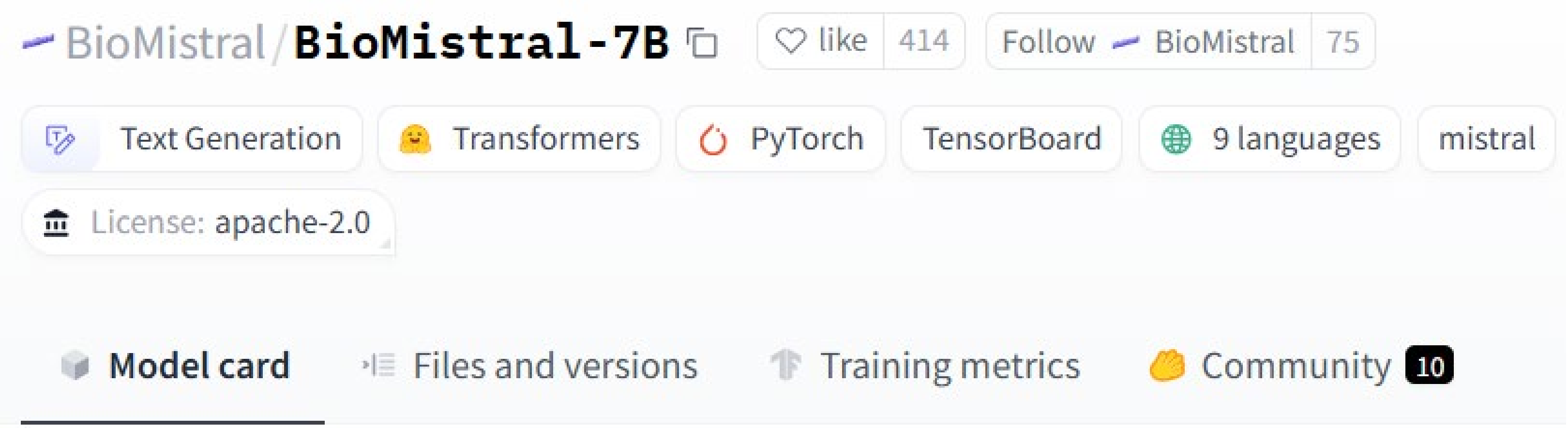
Parameter	Value
dim	4096
n_layers	32
head_dim	128
hidden_dim	14336
n_heads	32
n_kv_heads	8
window_size	4096
context_len	8192
vocab_size	32000

Mistral 7B. arXiv 2023.

BioMistral: Mistral + Unsupervised Next Token Prediction

- Architecture: Mistral 7B (already fine-tuned on general-domain instruction data)
- Data: PMC full text
  - A large biomedical corpus, no annotated or harvested instructions

https://huggingface.co/BioMistral/BioMistral-7B



BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. ACL 2024 Findings.

Datasets for Evaluating BioMistral

- MMLU [1]: college biology, college medicine, anatomy, professional medicine, medical genetics, and clinical knowledge
- MedQA [2]: questions from the US Medical License Exam (USMLE)
- MedMCQA [3]: questions from the Indian medical entrance examinations (AIIMS/NEET)
- PubMedQA [4]: rewrite PubMed paper titles and abstracts into yes/no/maybe questions

	MMLU								
	Clinical KG	**Medical Genetics**	**Anatomy**	**Pro Medicine**	**College Biology**	**College Medicine**	**MedQA**	**PubMedQA**	**MedMCQA**
Answer options	A / B / C / D	A / B / C / D	A / B / C / D	A / B / C / D	A / B / C / D	A / B / C / D	A / B / C / D / (E)	Yes / No / Maybe	A / B / C / D
Train / Valid. / Test	0 / 0 / 265	0 / 0 / 100	0 / 0 / 135	0 / 0 / 272	0 / 0 / 144	0 / 0 / 173	10178 / 1272 / 1273	211269 / 500 / 500	146257 / 36565 / 4183
Words / Questions	11.09	12.34	13.65	105.46	22.40	48.84	118.16	13.08	14.05
Context	×	×	×	×	×	×	×	✓	×

[1] *Measuring Massive Multitask Language Understanding*. ICLR 2021.
[2] *What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams*. arXiv 2020.
[3] *MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering*. CHIL 2022.
[4] *PubMedQA: A Dataset for Biomedical Research Question Answering*. EMNLP 2019.

Performance of BioMistral

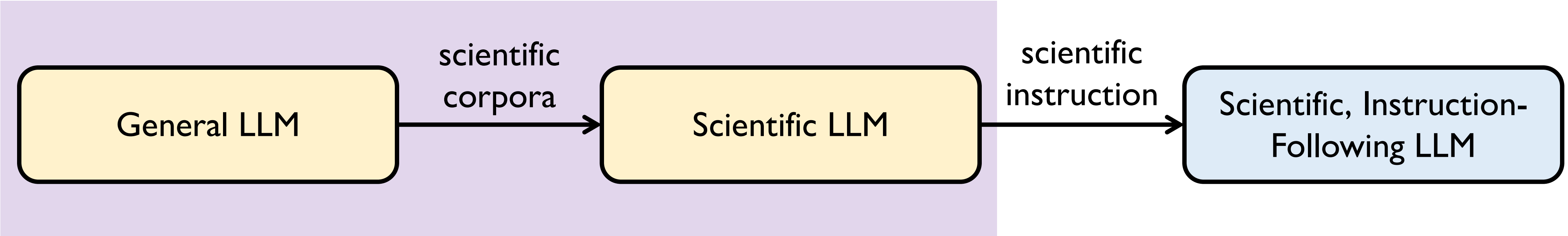
	MMLU										
	Clinical KG	Medical Genetics	Anatomy	Pro Medicine	College Biology	College Medicine	MedQA	MedQA 5 opts	PubMedQA	MedMCQA	Avg.
BioMistral 7B	59.9 ±1.2	64.0 ±1.6	56.5 ±1.8	60.4 ±0.5	59.0 ±1.5	54.7 ±1.0	50.6 ±0.3	42.8 ±0.3	77.5 ±0.1	48.1 ±0.2	57.3
Mistral 7B Instruct	**62.9** ±0.2	57.0 ±0.8	55.6 ±1.0	59.4 ±0.6	62.5 ±1.0	57.2 ±2.1	42.0 ±0.2	40.9 ±0.4	75.7 ±0.4	46.1 ±0.1	55.9
BioMistral 7B Ensemble	62.8 ±0.5	62.7 ±0.5	57.5 ±0.3	**63.5** ±0.8	64.3 ±1.6	55.7 ±1.5	50.6 ±0.3	43.6 ±0.5	77.5 ±0.2	**48.8** ±0.0	58.7
BioMistral 7B DARE	62.3 ±1.3	**67.0** ±1.6	55.8 ±0.9	61.4 ±0.3	**66.9** ±2.3	**58.0** ±0.5	**51.1** ±0.3	**45.2** ±0.3	77.7 ±0.1	48.7 ±0.1	**59.4**
BioMistral 7B TIES	60.1 ±0.9	65.0 ±2.4	**58.5** ±1.0	60.5 ±1.1	60.4 ±1.5	56.5 ±1.9	49.5 ±0.1	43.2 ±0.1	77.5 ±0.2	48.1 ±0.1	57.9
BioMistral 7B SLERP	62.5 ±0.6	64.7 ±1.7	55.8 ±0.3	62.7 ±0.3	64.8 ±0.9	56.3 ±1.0	50.8 ±0.6	44.3 ±0.4	**77.8** ±0.0	48.6 ±0.1	58.8
MedAlpaca 7B	53.1 ±0.9	58.0 ±2.2	54.1 ±1.6	58.8 ±0.3	58.1 ±1.3	48.6 ±0.5	40.1 ±0.4	33.7 ±0.7	73.6 ±0.3	37.0 ±0.3	51.5
PMC-LLaMA 7B	24.5 ±1.7	27.7 ±1.7	35.3 ±0.7	17.4 ±1.7	30.3 ±0.9	23.3 ±1.7	25.5 ±0.9	20.2 ±0.1	72.9 ±1.2	26.6 ±0.1	30.4
MediTron-7B	41.6 ±1.2	50.3 ±2.1	46.4 ±0.9	27.9 ±0.3	44.4 ±2.6	30.8 ±0.7	41.6 ±0.5	28.1 ±0.5	74.9 ±0.1	41.3 ±0.2	42.7
BioMedGPT-LM-7B	51.4 ±0.4	52.0 ±1.4	49.4 ±2.7	53.3 ±0.6	50.7 ±0.0	49.1 ±0.8	42.5 ±0.3	33.9 ±0.5	76.8 ±0.3	37.6 ±0.4	49.7
GPT-3.5 Turbo 1106*	74.71 ±0.3	74.00 ±2.2	65.92 ±0.6	72.79 ±1.6	72.91 ±1.7	64.73 ±2.9	57.71 ±0.3	50.82 ±0.7	72.66 ±1.0	53.79 ±0.2	66.0

BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. ACL 2024 Findings.

Agenda

- Unsupervised Next Token Prediction
  - General Domain: GPT-3
  - Mathematics: Minerva
- Supervised Fine-Tuning / Instruction Tuning
  - General Domain: FLAN
  - Science: SciInstruct
  - Biomedicine: BioMistral
  - Geoscience: OceanGPT

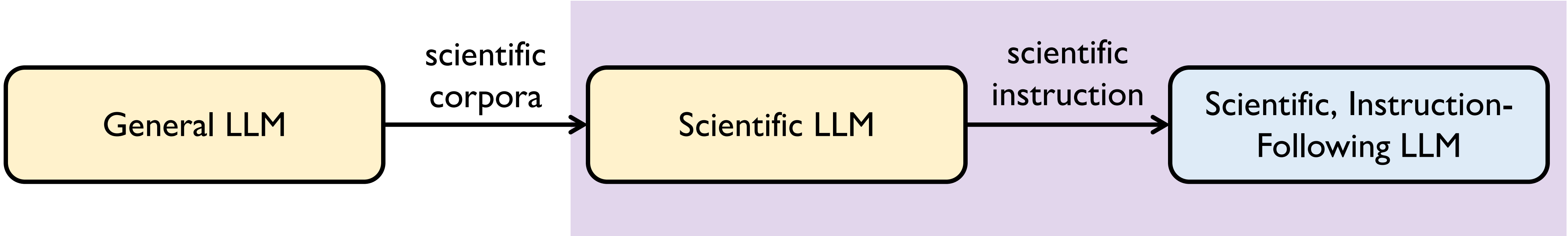
OceanGPT: An LLM for Ocean Science



- Step 1: Unsupervised next token prediction
  - 67,633 full-text papers
  - ocean physics, ocean chemistry, ocean biology, geology, hydrology, etc.

OceanGPT: A Large Language Model for Ocean Science Tasks. ACL 2024.

OceanGPT: An LLM for Ocean Science



- Unsupervised next token prediction
  - 67,633 full-text papers
  - ocean physics, ocean chemistry, ocean biology, geology, hydrology, etc.
- Instruction tuning
  - Hard to find benchmark datasets or sufficient exam questions related to ocean science
  - A common challenge if you want to build an LLM for a fine-grained field

OceanGPT: A Large Language Model for Ocean Science Tasks. ACL 2024.

Constructing Instruction Tuning Data for a Fine-Grained Field

- Step 1: Dozens of annotators with rich backgrounds in marine science write some representative example for each marine topic.

- E.g.,
  - *Instruction*: Please recommend several rare marine plants and animals and their ecological value.
  - *Output*: Rare marine animals and plants include whales, dolphins, jewel-like seaweed, seahorses, etc. These species play a crucial role in maintaining the balance of the ecosystem and require protection.

- However, you can only obtain a small number of instruction tuning data from humans!
  - Use LLMs to paraphrase human-written data
  - Retrieve more data from domain-specific corpora

Constructing Instruction Tuning Data for a Fine-Grained Field

- Step 2: Build more instruction tuning data by generating questions given unannotated text.

human-written instruction and output

unannotated ocean science papers

BM25 →

unannotated paragraphs relevant to human-annotated instruction data

"Marine organisms are classified into fish (such as sharks), mammals (such as dolphins), mollusks (such as octopuses and squid), and reptiles (such as sea turtles), etc."

This can be an "output", but where is the "instruction"?

Constructing Instruction Tuning Data for a Fine-Grained Field

- Step 2: Build more instruction tuning data by generating questions given unannotated text.

You are a helpful ocean assistant. You are to extract the question from each of the answer provided.

"Marine organisms are classified into fish (such as sharks), mammals (such as dolphins), mollusks (such as octopuses and squid), and reptiles (such as sea turtles), etc."

This can be an "output".

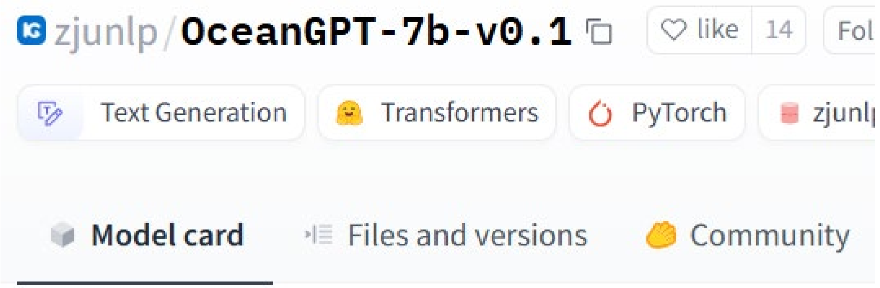
→ GPT-3.5

Please classify the following marine creatures: shark, dolphin, squid, octopus.

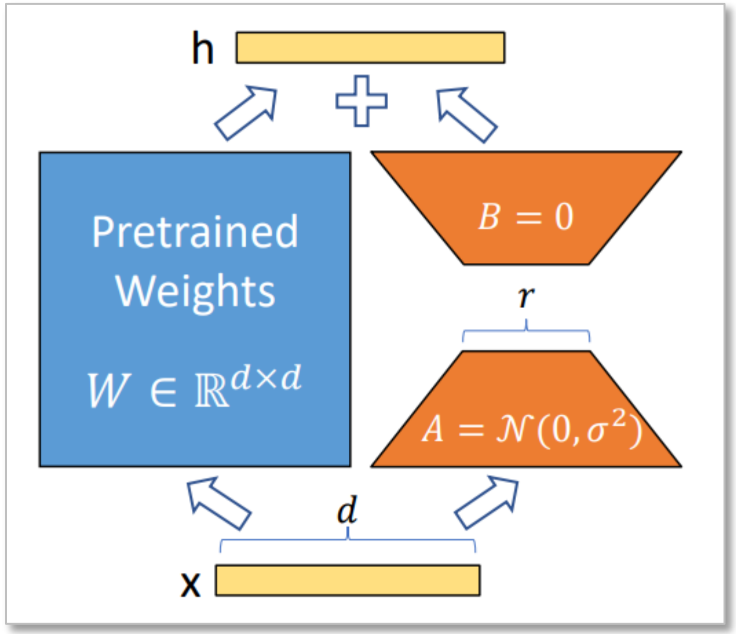
This can be an "instruction".

Model Details of OceanGPT

https://huggingface.co/zjunlp/OceanGPT-7b-v0.1



- Architecture: LLaMA-2 7B
- Data: 150K (instruction, output) pairs
- Tuning Method: Low-Rank Adaptation (LoRA)



$$h = (W_0 + \Delta W)x = (W_0 + B \times A)x$$

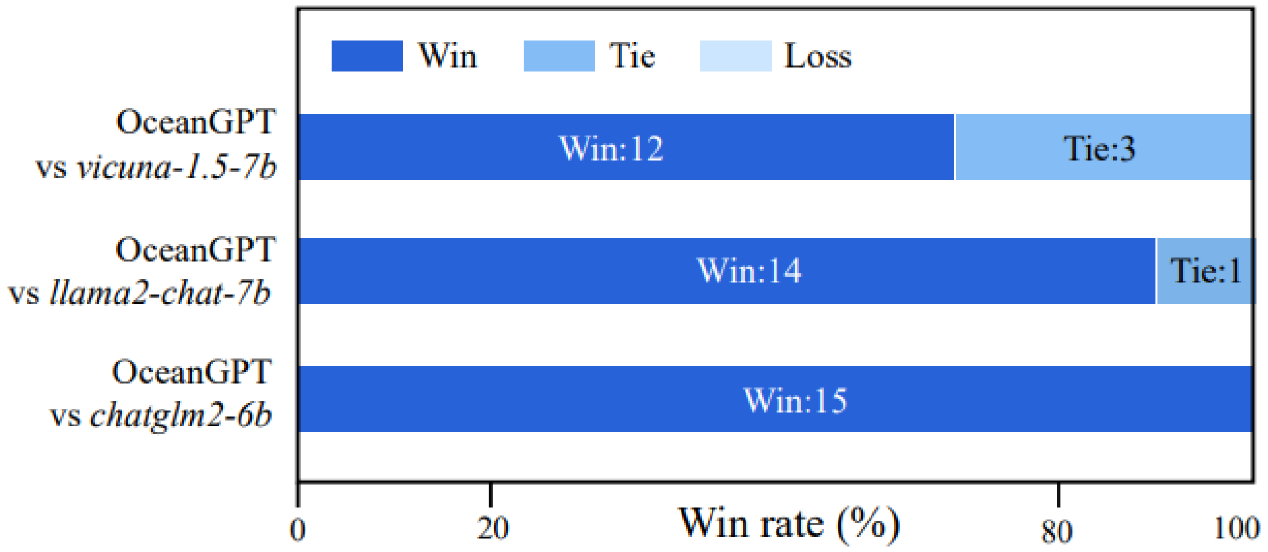
x: input
h: output
W_0: original model parameters (i.e., LLaMA-2)
$(W_0 + \Delta W)$: new model parameters (i.e., OceanGPT)
$B \times A$: a low-rank approximation of ΔW

LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022.

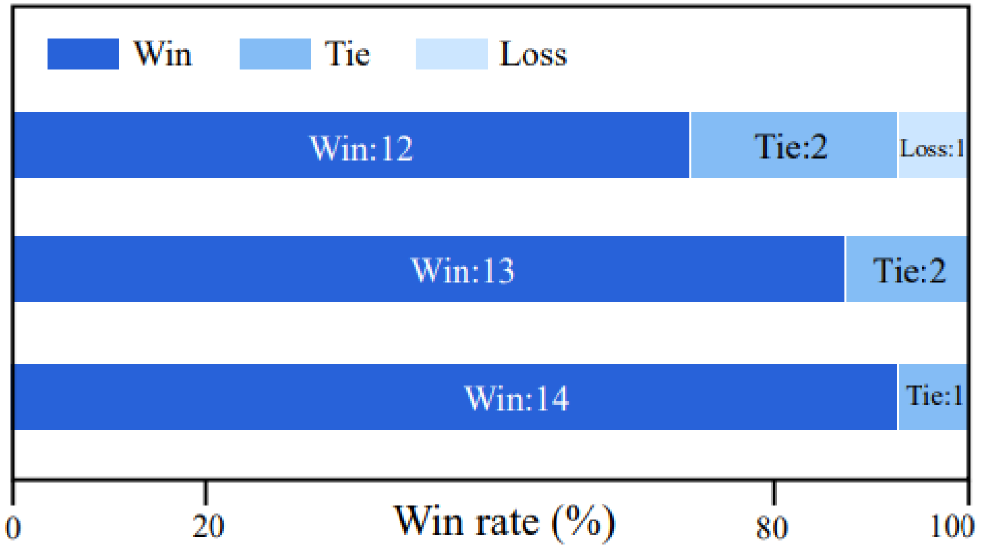
Evaluation of OceanGPT

- Tasks: (OceanBench: https://huggingface.co/datasets/zjunlp/OceanBench)
  - Analysis: "*Analyzing the bioactive components of seaweed and its application prospects*"
  - Commonsense Reasoning: "*Infer the reasons for the increase in seawater turbidity*"
  - Recommendation: "*Recommend an instrument capable of detecting ocean pollution*"
  - Editing: "*Edit a popular science article on ocean circulation and pollution*"
  - Question Answering: "*What is the main electrolyte in seawater?*"
  - Classification: "*What are the basic classifications of tropical cyclones?*"
  - Open-Ended Generation: "*Write an argumentative essay on ocean conservation and management*"
  - Description: "*Describe the mechanism of underwater mineral enrichment*"
  - …

Performance of OceanGPT



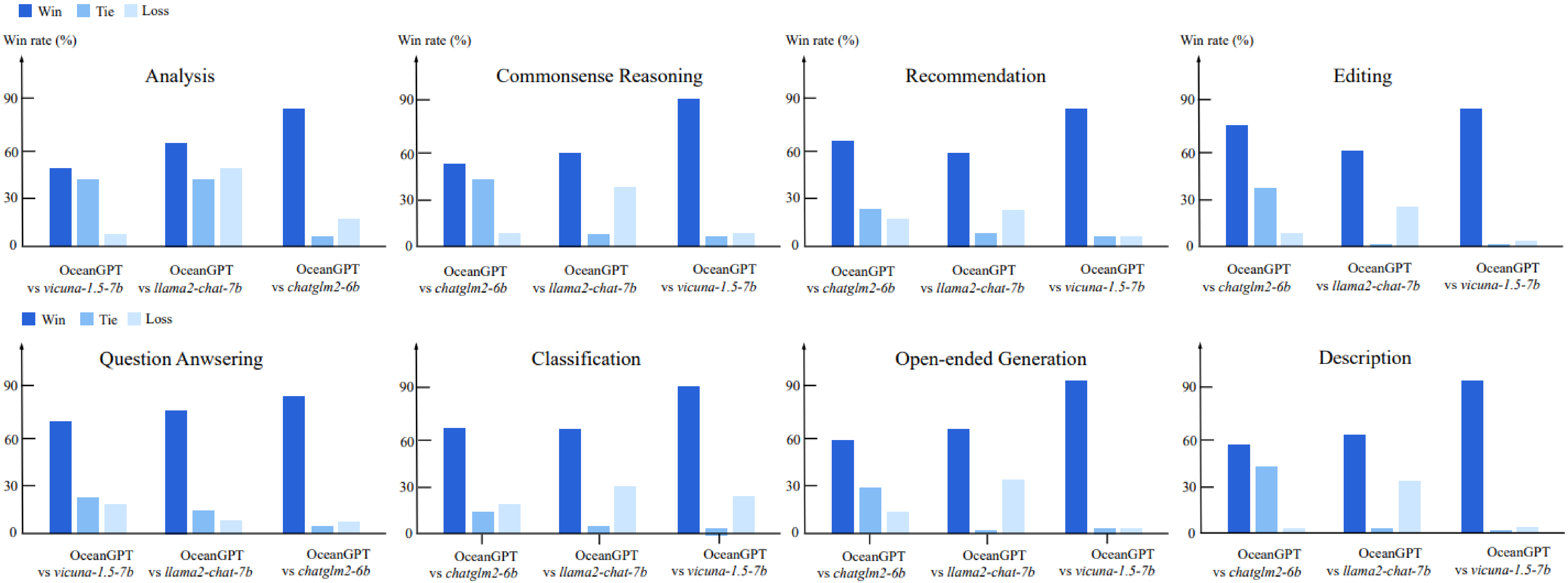
GPT-4 evaluation



Human evaluation

OceanGPT: A Large Language Model for Ocean Science Tasks. ACL 2024.

Performance of OceanGPT



OceanGPT: A Large Language Model for Ocean Science Tasks. ACL 2024.

Take-Away Messages

- Tuning LLMs to follow instructions enables them to deal with unseen instructions without any examples during inference (i.e., zero-shot generalization).
- Multiple ways to harvest instruction tuning data in the scientific domain:
  - Convert benchmark datasets to the instruction tuning format
  - Collect questions from textbooks, problem sets, etc.
  - May not work for a new, fine-grained field!
- Off-the-shelf powerful LLMs (e.g., GPT-4) can help the construction of instruction tuning data
  - Recover the chain-of-thought
  - Generate more instruction tuning data to complement human annotations

Thank You!

Course Website: https://yuzhang-teaching.github.io/CSCE689-S25.html