香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

COMP 4901B

Large Language Models

Instruction Tuning and Alignment

Junxian He

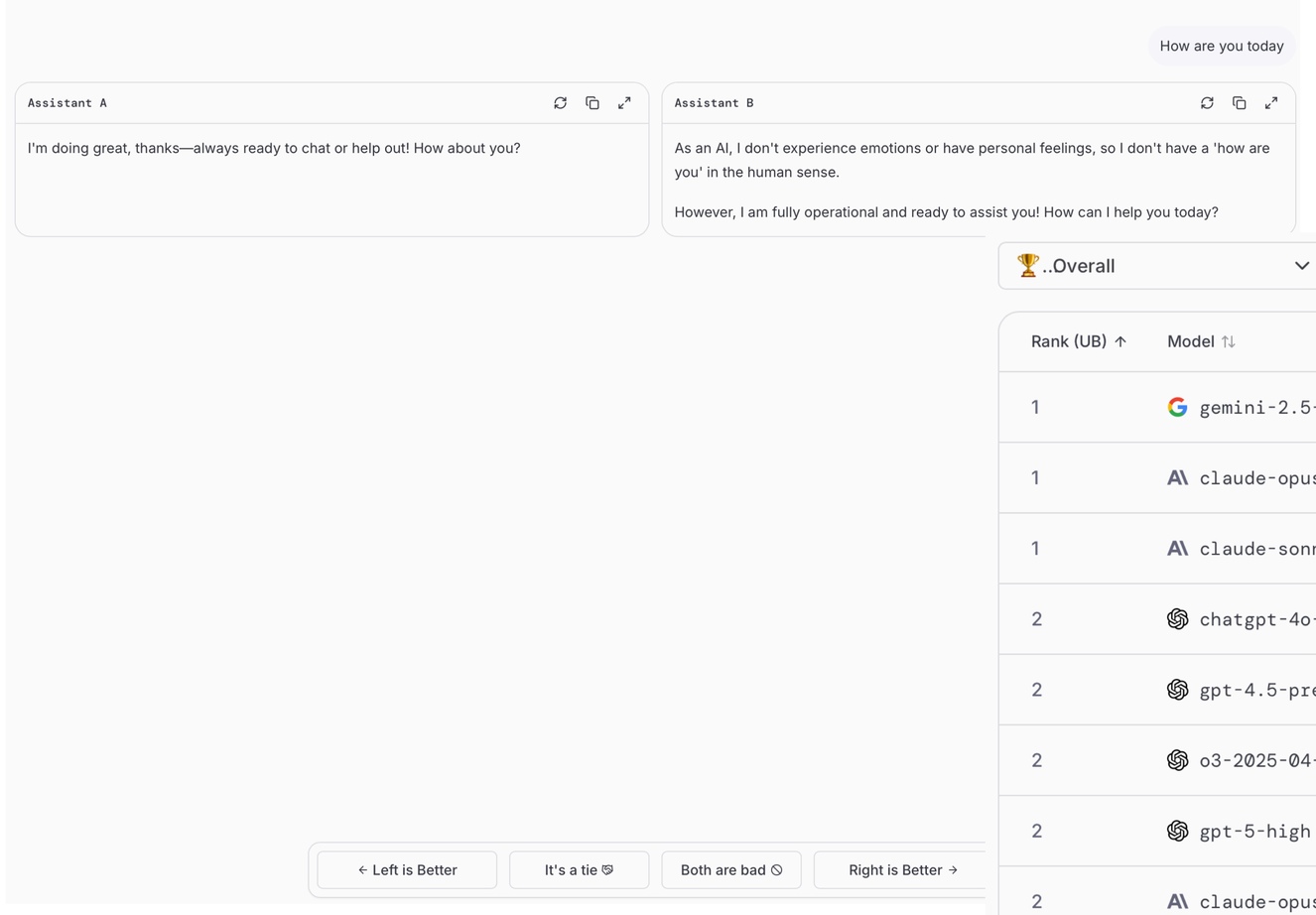
Oct 8, 2025

Homework 1 is Due Today

Review: Crowd-Sourcing Human Evaluation

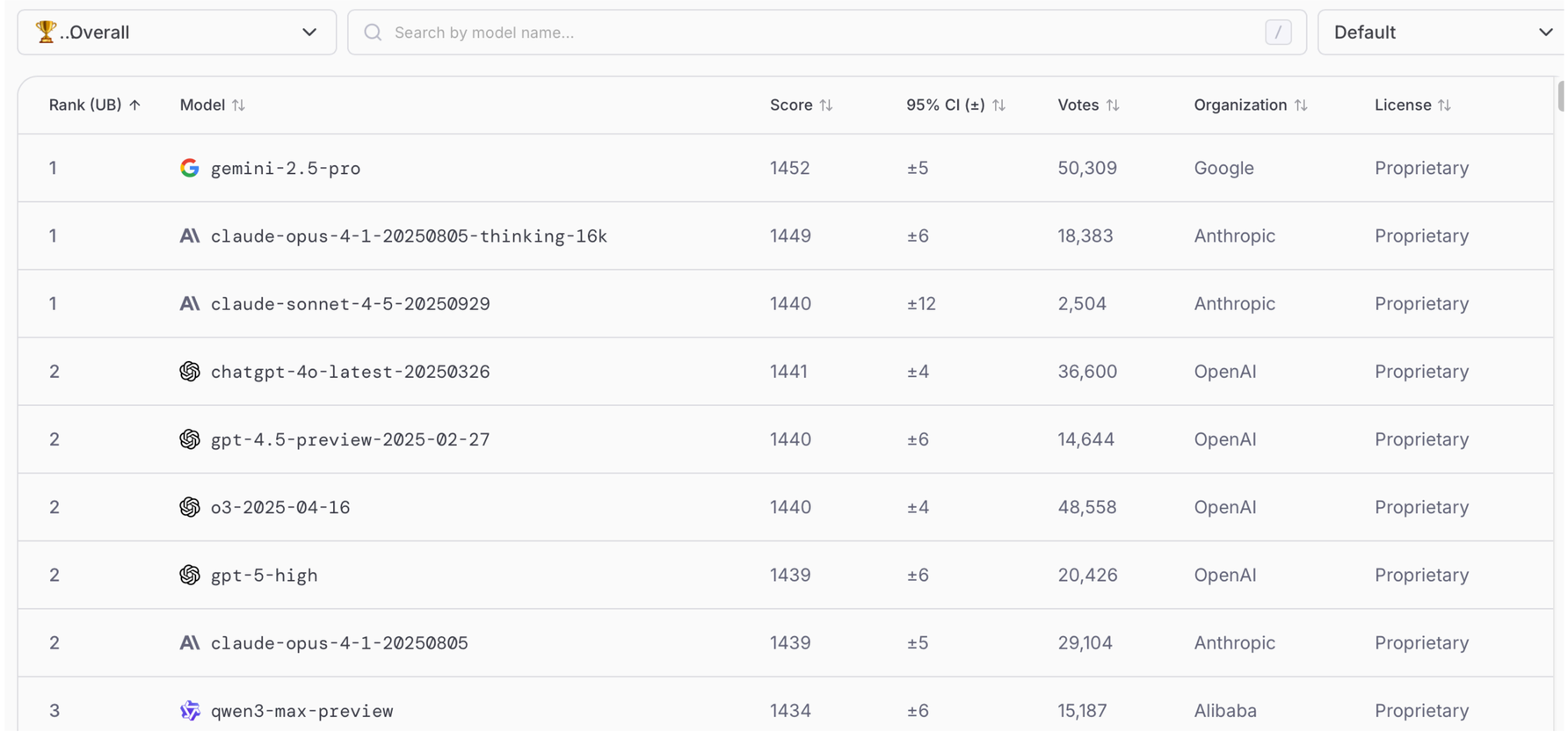
https://lmarena.ai/

Imagine any two models can battle a game, whose response is better



Rank (UB)	Model	Score	95% CI (±)	Votes	Organization	Licens
1	gemini-2.5-pro	1452	±5	50,309	Google	Propr
1	claude-opus-4-1-20250805-thinking-16k	1449	±6	18,383	Anthropic	Propr
1	claude-sonnet-4-5-20250929	1440	±12	2,504	Anthropic	Propr
2	chatgpt-4o-latest-20250326	1441	±4	36,600	OpenAI	Propr
2	gpt-4.5-preview-2025-02-27	1440	±6	14,644	OpenAI	Propr
2	o3-2025-04-16	1440	±4	48,558	OpenAI	Propr
2	gpt-5-high	1439	±6	20,426	OpenAI	Propr
2	claude-opus-4-1-20250805	1439	±5	29,104	Anthropic	Propr
3	qwen3-max-preview	1434	±6	15,187	Alibaba	Propr

Crowd-Sourcing Human Evaluation



🏆 ..Overall | Search by model name... / | Default

Rank (UB) ↑	Model	Score	95% CI (±)	Votes	Organization	License
1	gemini-2.5-pro	1452	±5	50,309	Google	Proprietary
1	claude-opus-4-1-20250805-thinking-16k	1449	±6	18,383	Anthropic	Proprietary
1	claude-sonnet-4-5-20250929	1440	±12	2,504	Anthropic	Proprietary
2	chatgpt-4o-latest-20250326	1441	±4	36,600	OpenAI	Proprietary
2	gpt-4.5-preview-2025-02-27	1440	±6	14,644	OpenAI	Proprietary
2	o3-2025-04-16	1440	±4	48,558	OpenAI	Proprietary
2	gpt-5-high	1439	±6	20,426	OpenAI	Proprietary
2	claude-opus-4-1-20250805	1439	±5	29,104	Anthropic	Proprietary
3	qwen3-max-preview	1434	±6	15,187	Alibaba	Proprietary

Elo scores as in sports

Evaluate Language Model Knowledge

Automatic evaluations typically seek for objective metrics

Example from MMLU:

Microeconomics

One of the reasons that the government discourages and regulates monopolies is that
(A) producer surplus is lost and consumer surplus is gained. ✗
(B) monopoly prices ensure productive efficiency but cost society allocative efficiency. ✗
(C) monopoly firms do not engage in significant research and development. ✗
(D) consumer surplus is lost with higher prices and lower levels of output. ✓

Figure 3: Examples from the Microeconomics task.

Conceptual Physics

When you drop a ball from rest it accelerates downward at 9.8 m/s². If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is
(A) 9.8 m/s² ✓
(B) more than 9.8 m/s² ✗
(C) less than 9.8 m/s² ✗
(D) Cannot say unless the speed of throw is given. ✗

Multi-choice QA

Ranging from middle school to college level

Hendricks et al. Measuring Massive Multitask Language Understanding. 2021.

Mathematical Reasoning

Example from GSM8K:

Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?
Solution: Beth bakes 4 2 dozen batches of cookies for a total of 4*2 = <<4*2=8>>8 dozen cookies
There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of 12*8 = <<12*8=96>>96 cookies
She splits the 96 cookies equally amongst 16 people so they each eat 96/16 = <<96/16=6>>6 cookies
Final Answer: 6

Problem: Mrs. Lim milks her cows twice a day. Yesterday morning, she got 68 gallons of milk and in the evening, she got 82 gallons. This morning, she got 18 gallons fewer than she had yesterday morning. After selling some gallons of milk in the afternoon, Mrs. Lim has only 24 gallons left. How much was her revenue for the milk if each gallon costs $3.50?
Mrs. Lim got 68 gallons - 18 gallons = <<68-18=50>>50 gallons this morning.
So she was able to get a total of 68 gallons + 82 gallons + 50 gallons = <<68+82+50=200>>200 gallons.
She was able to sell 200 gallons - 24 gallons = <<200-24=176>>176 gallons.
Thus, her total revenue for the milk is $3.50/gallon x 176 gallons = $<<3.50*176=616>>616.
Final Answer: 616

Problem: Tina buys 3 12-packs of soda for a party. Including Tina, 6 people are at the party. Half of the people at the party have 3 sodas each, 2 of the people have 4, and 1 person has 5. How many sodas are left over when the party is over?
Solution: Tina buys 3 12-packs of soda, for 3*12= <<3*12=36>>36 sodas
6 people attend the party, so half of them is 6/2= <<6/2=3>>3 people
Each of those people drinks 3 sodas, so they drink 3*3=<<3*3=9>>9 sodas
Two people drink 4 sodas, which means they drink 2*4=<<4*2=8>>8 sodas
With one person drinking 5, that brings the total drank to 5+9+8+3= <<5+9+8+3=25>>25 sodas
As Tina started off with 36 sodas, that means there are 36-25=<<36-25=11>>11 sodas left
Final Answer: 11

Short-answer

Cobbe et al. Training Verifiers to Solve Math Word Problems

Mathematical Reasoning

Example from AIME (American Invitational Mathematics Examination):

Problem

Let the sequence of rationals $x_1, x_2, \ldots$ be defined such that $x_1 = \frac{25}{11}$ and

$$x_{k+1} = \frac{1}{3}\left(x_k + \frac{1}{x_k} - 1\right).$$

x_{2025} can be expressed as $\frac{m}{n}$ for relatively prime positive integers m and n. Find the remainder when $m + n$ is divided by 1000.

Solution 1 (complete)

This problem can be split into three parts, listed below:

Part 1: Analyzing Fractions

Let $x_k = \frac{a_k}{b_k}$, where a_k, b_k are relatively prime positive integers. First, we analyze the moduli of the problem. Plugging in for x_2 yields $x_2 = \frac{157}{275}$. Notice that in both x_1 and x_2, the numerator is equivalent to 1 and the denominator is equivalent to 2 modulus 3. We see that $x_2 = \frac{1}{3} \cdot \frac{(a_1 - b_1)^2 + a_1 b_1}{a_1 b_1}$. Specifically, we know that

$$(a_1 - b_1)^2 + a_1 b_1 \equiv (1-2)^2 + 1 \cdot 2 \equiv 0 \pmod 3$$

Then this is always divisible by 3 for all x_k (it can be shown that for all x_k, we have $a_k \equiv 1 \pmod 3$ and $b_k \equiv 2 \pmod 3$ by using $\bmod 9$). Thus, $x_2 = \frac{\frac{1}{3}((a_1 - b_1)^2 + a_1 b_1)}{a_1 b_1}$, and the numerator and denominator of the right-hand side (RHS) correspond to the numerator and denominator of x_2 in simplest form. (To further prove that the top and bottom are relatively prime, consider that a_k and b_k are by definition relatively prime, so $(a_k - b_k)^2$ and $a_k b_k$ share no factors.)

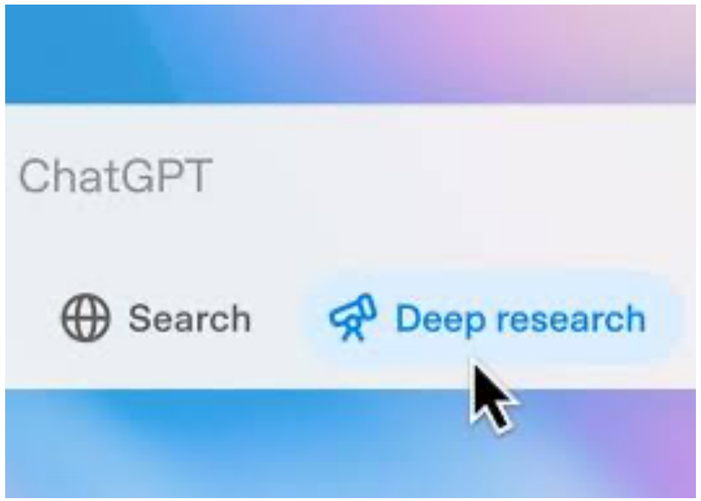
Notice that the above do not just apply to x_1; we did not use any specific properties of x_1. Then we may generalize the above, finding that:

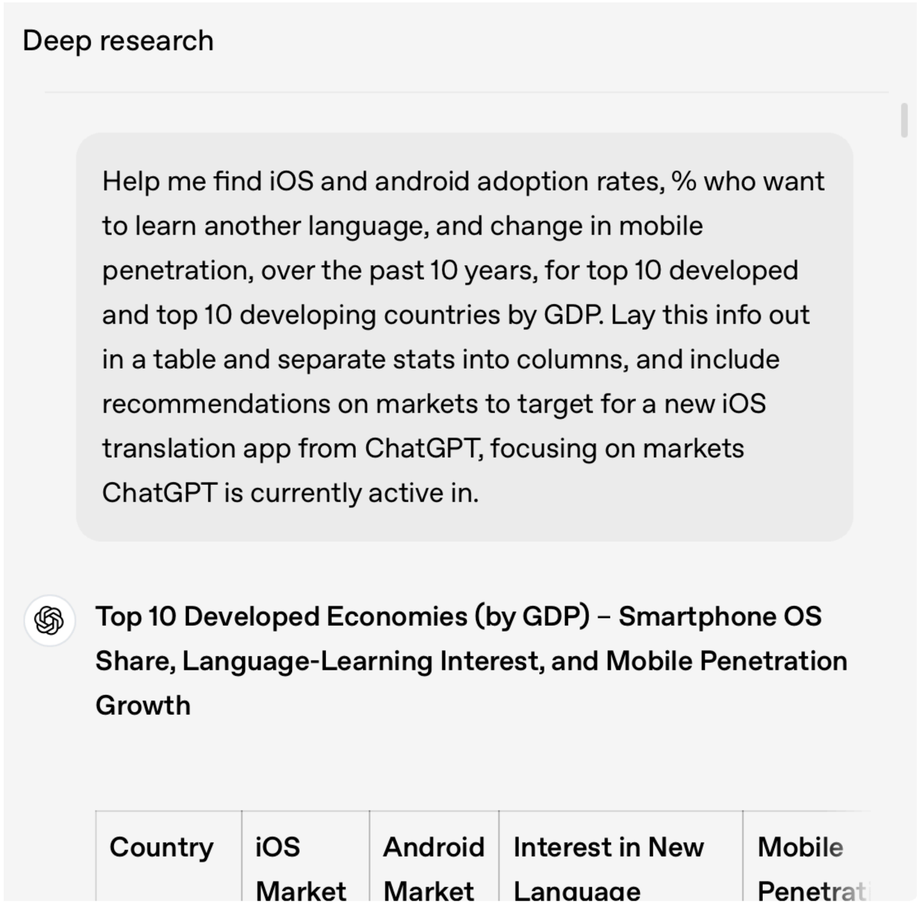
$$a_k = \frac{1}{3}((a_{k-1} - b_{k-1})^2 + a_{k-1} b_{k-1})$$

$$b_k = a_{k-1} b_{k-1}$$

Deep Research

https://openai.com/index/introducing-deep-research/

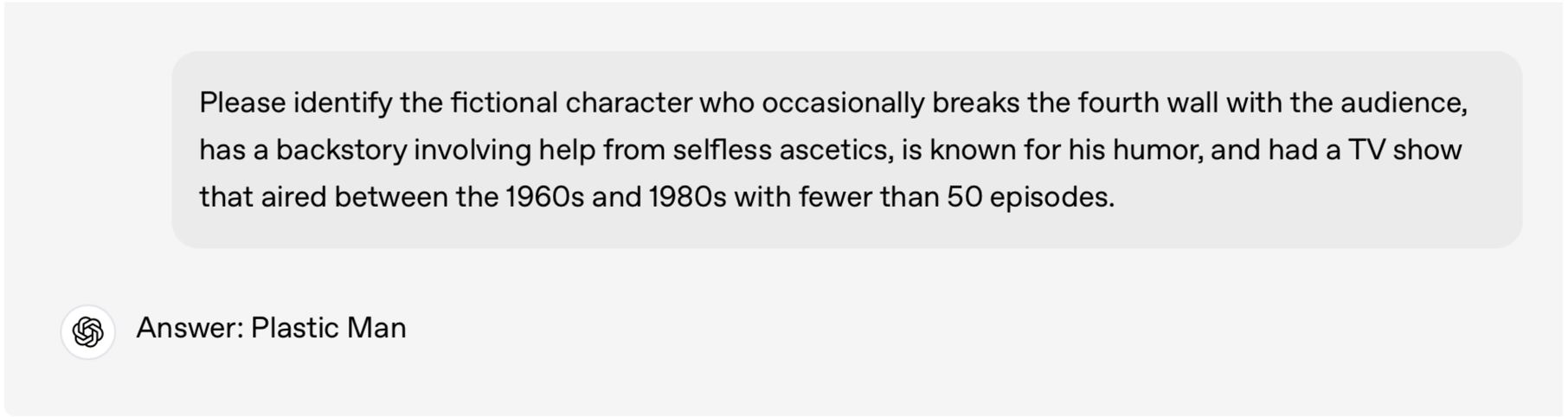




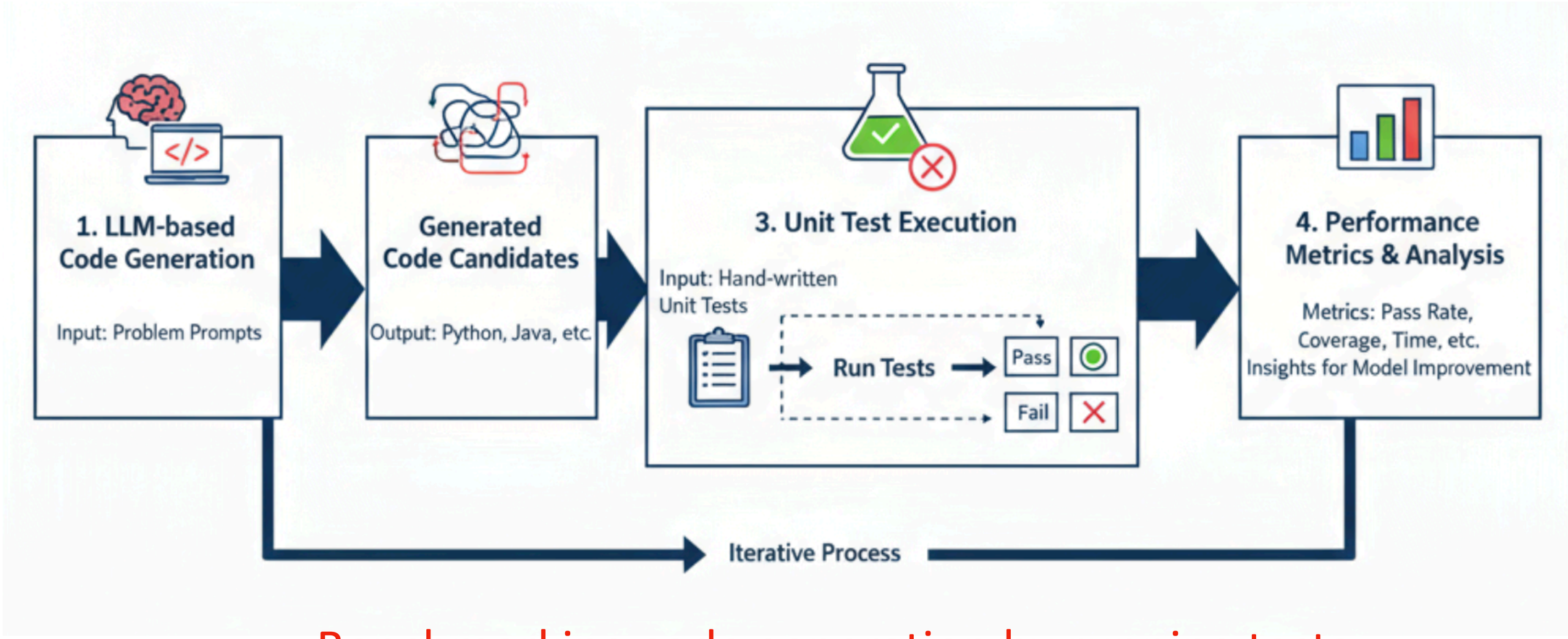
Hard to Evaluate

Evaluating Deep Research

https://openai.com/index/browsecomp/

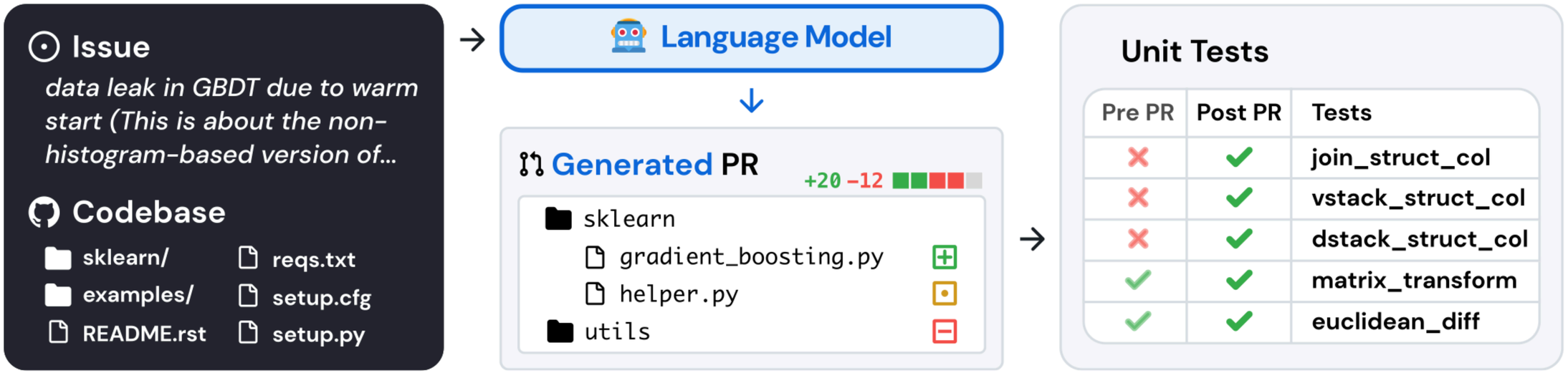


Benchmarking Code Generation



Benchmarking code generation by running tests

Software Engineering



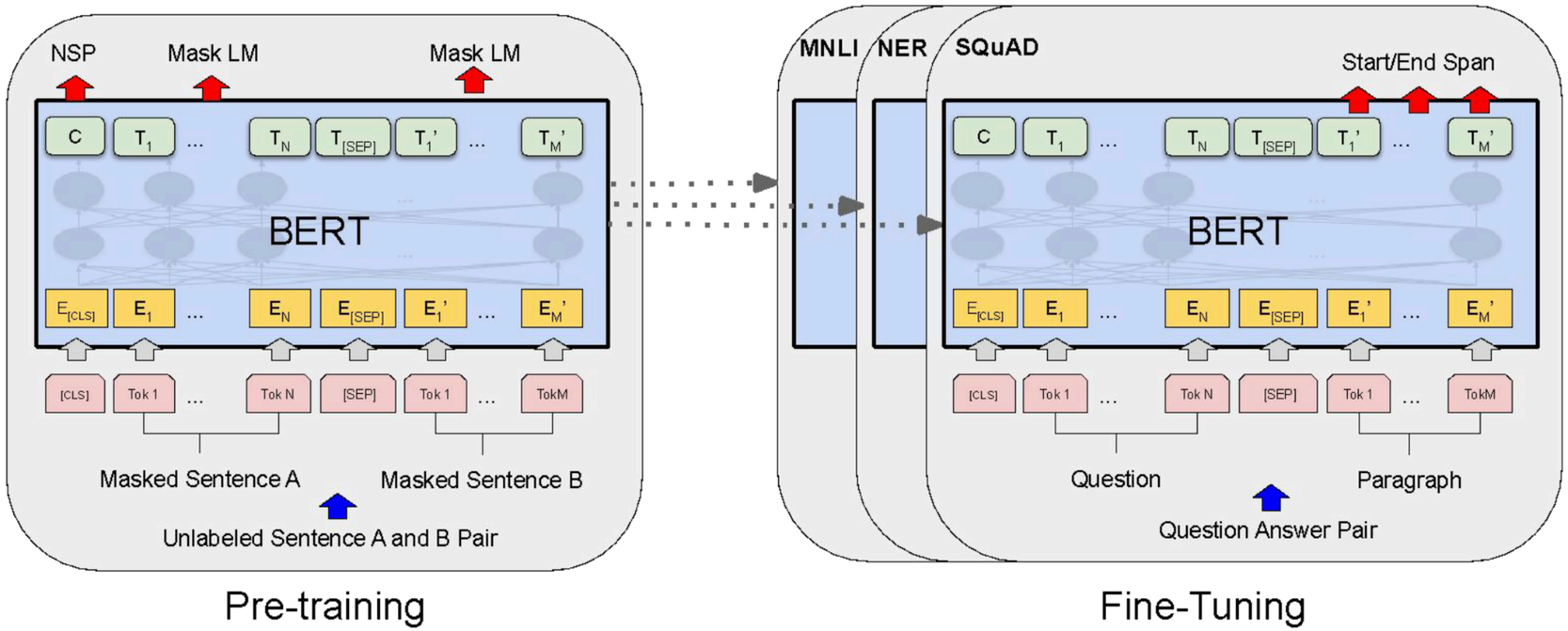
We will cover more specific evaluations when we learn specific topics

Jimenez et al. SWE-BENCH: CAN LANGUAGE MODELS RESOLVE REAL-WORLD GITHUB ISSUES? 2024

香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

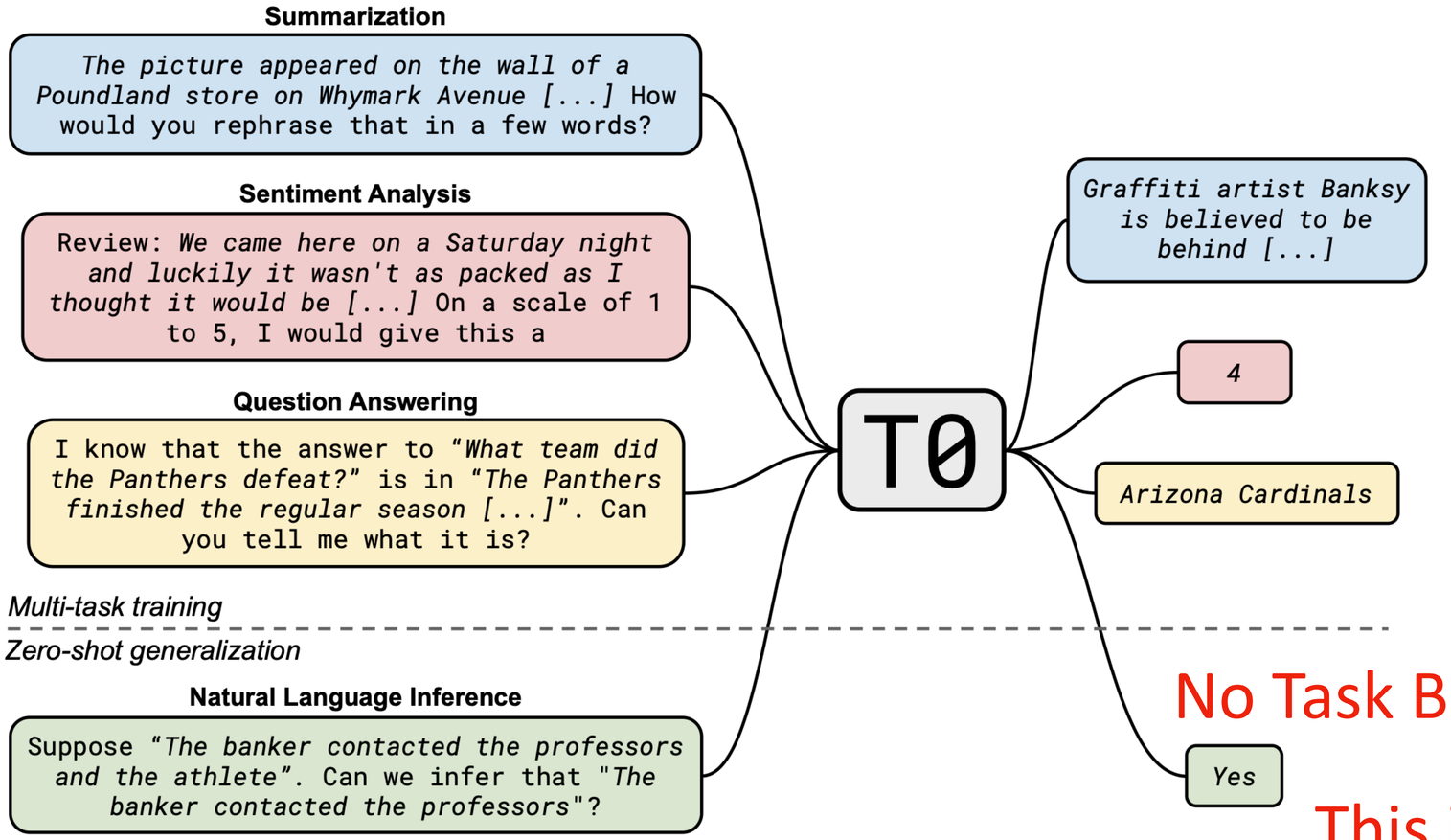
Instruction Tuning

From Task-Specific to General



Traditional fine-tuning

From Task-Specific to General



What is zero-shot task generalization?
Why is it possible?

No Task Boundary! Everything is seq2seq

This is the early form of prompting

Modern Fine-Tuning (Multi-Task Learning)

Sanh et al. Multitask Prompted Training Enables Zero-Shot Task Generalization. 2022

From Task-Specific to General

Examples

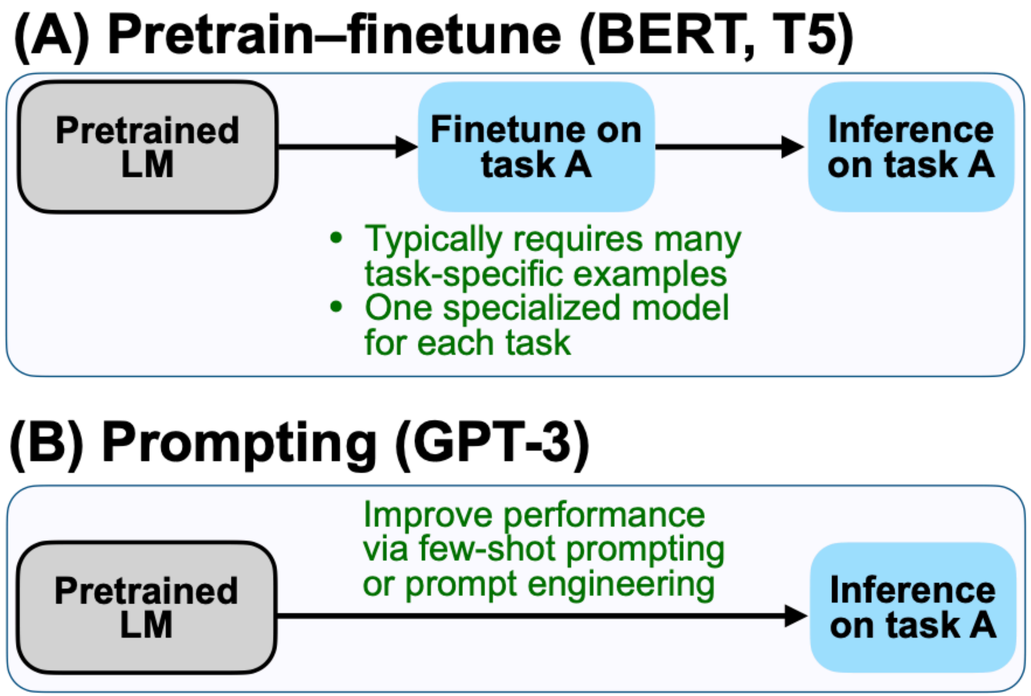
Question	Context	Answer
What is a major importance of Southern California in relation to California and the US?	...Southern California is a major economic center for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. Entailment, neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence positive or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive

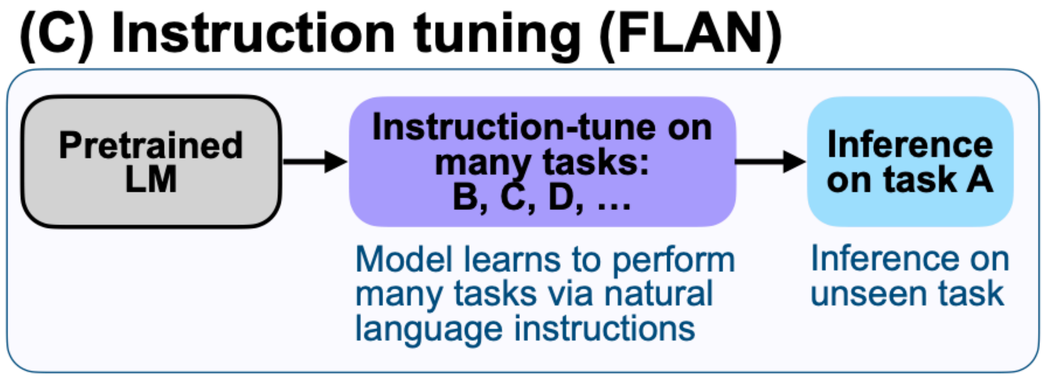
Question	Context	Answer
What has something experienced?	Areas of the Baltic that have experienced eutrophication.	eutrophication
Who is the illustrator of Cycle of the Werewolf?	Cycle of the Werewolf is a short novel by Stephen King, featuring illustrations by comic book artist Bernie Wrightson.	Bernie Wrightson
What is the change in dialogue state?	Are there any Eritrean restaurants in town?	food: Eritrean
What is the translation from English to SQL?	The table has column names... Tell me what the notes are for South Australia	SELECT notes from table WHERE 'Current Slogan' = 'South Australia'
Who had given help? Susan or Joan?	Joan made sure to thank Susan for all the help she had given.	Susan

An early rejected paper demonstrated this form

McCann et al. The Natural Language Decathlon: Multitask Learning as Question Answering. 2018

Instruction Tuning

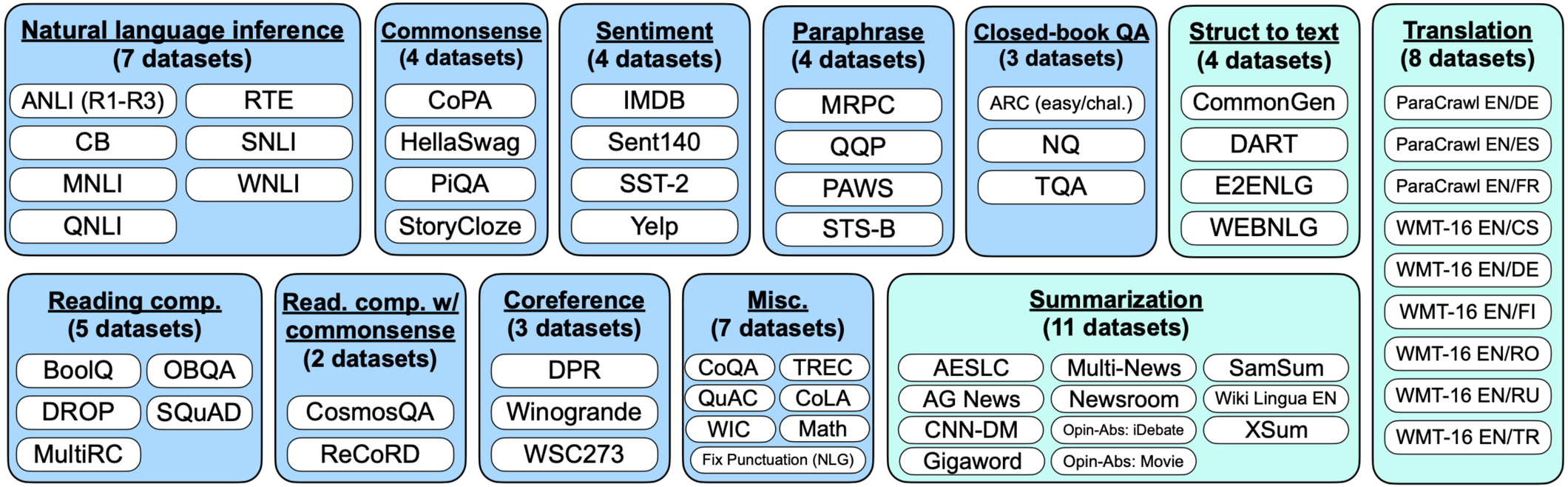




Instruction tuning is equivalent to multi-task fine-tuning

Wei et al. Finetuned Language Models Are Zero-Shot Learners. 2022

Instruction Tuning



All tasks unified to a sequence-in, sequence-out format

Wei et al. Finetuned Language Models Are Zero-Shot Learners. 2022

What is different for Instruction Tuning from Traditional Multi-Task Training?

Data is extremely diverse

Diverse Data

Instruction: Given an address and city, come up with the zip code.
Input:
Address: 123 Main Street, City: San Francisco
Output: 94105

Instruction: I am looking for a job and I need to fill out an application form. Can you please help me complete it?
Input:
Application Form:
Name:_____ Age:_____ Sex:_____
Phone Number:_____ Email Address:_____
Education:_____ ...
Output:
Name: John Doe Age: 25 Sex: Male
Phone Number: ...

Instruction: How to write a code for converting degrees fahrenheit to celsius.
Input: Null
Output:

```
def convert_fahrenheit_to_celsius(fahr):
    celsius = (fahr - 32) * 5 / 9
    return celsius
```

Instruction: Write a letter from the perspective of a cat.
Input: Null
Output:
Dear [Owner], I am writing to you today because I have a problem. You see, I've been living with you for the past [Number of Years] and I love it here. The food is great, the bed is comfy, and there are plenty of toys to play with. But lately, I've noticed that you're not as affectionate towards me as you used to be. In fact, sometimes when I try to cuddle up next to you on the couch, you push me away ...

Wang et al. Self-Instruct: Aligning Language Models with Self-Generated Instructions. 2022

Number of Samples Per Task is Small

statistic	
# of instructions	52,445
- # of classification instructions	11,584
- # of non-classification instructions	40,861
# of instances	82,439
- # of instances with empty input	35,878
ave. instruction length (in words)	15.9
ave. non-empty input length (in words)	12.7
ave. output length (in words)	18.9

<2 instances per instruction (task)

	Model	# Params	ROUGE-L
	Vanilla LMs		
	T5-LM	11B	25.7
	GPT3	175B	6.8
	Instruction-tuned w/o SUPERNI		
①	T0	11B	33.1
	GPT3 + T0 Training	175B	37.9
②	$\text{GPT3}_{\text{SELF-INST}}$ (Ours)	175B	39.9
	InstructGPT_{001}	175B	**40.8**
	Instruction-tuned w/ SUPERNI		
	T*k*-INSTRUCT	11B	46.0
③	GPT3 + SUPERNI Training	175B	49.5
	$\text{GPT3}_{\text{SELF-INST}}$ + SUPERNI Training (Ours)	175B	**51.6**

Improve >30 points by just 50K samples

Wang et al. Self-Instruct: Aligning Language Models with Self-Generated Instructions. 2022

Number of Samples Per Task is Small

Source	#Examples	Avg Input Len.	Avg Output Len.
Training			
Stack Exchange (STEM)	200	117	523
Stack Exchange (Other)	200	119	530
wikiHow	200	12	1,811
Pushshift r/WritingPrompts	150	34	274
Natural Instructions	50	236	92
Paper Authors (Group A)	200	40	334
Dev			
Paper Authors (Group A)	50	36	N/A
Test			
Pushshift r/AskReddit	70	30	N/A
Paper Authors (Group B)	230	31	N/A

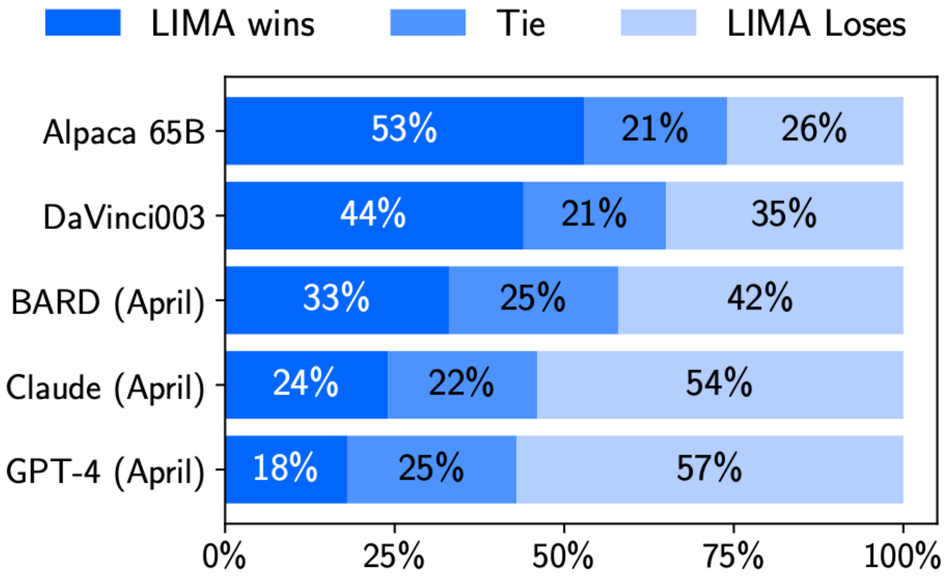


Figure 1: Human preference evaluation, comparing LIMA to 5 different baselines across 300 test prompts.

1000 training examples to unlock strong model abilities

Zhou et al. LIMA: Less Is More for Alignment. 2023

Instruction Tuning vs Traditional Multi-task Fine-tuning

Machine learning wise, they are the same in terms of implementation

Traditional

Data is not that diverse, typically 10s of tasks, each task with >10K or even more samples

Instruction Tuning

Data is diverse, typically 1-3 examples per task, and thousands of examples in total can improve pretrained models a lot

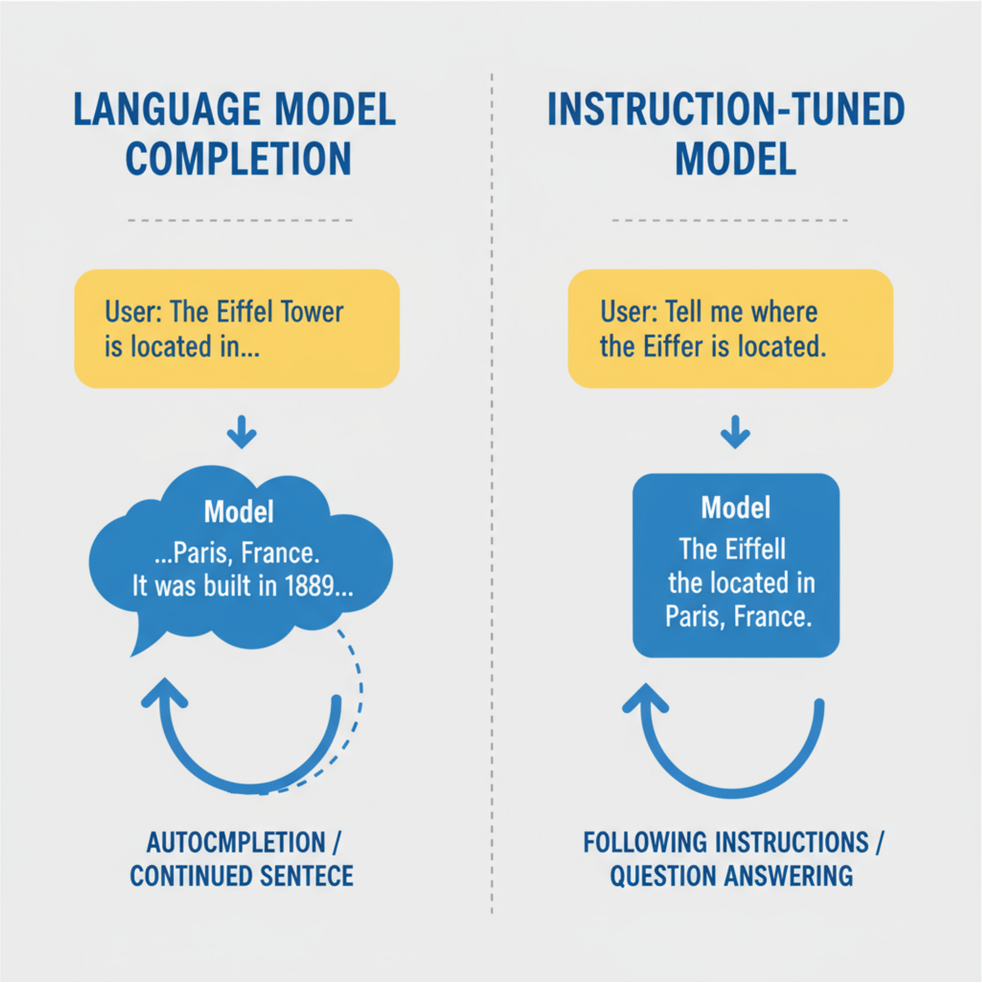
What makes instruction tuning work with so few examples?

PRETRAINING

(Human) Alignment

In a narrow definition, alignment means to adapt the language model to follow human instructions

Sometimes, typical instruction tuning can be regarded as aligning the model



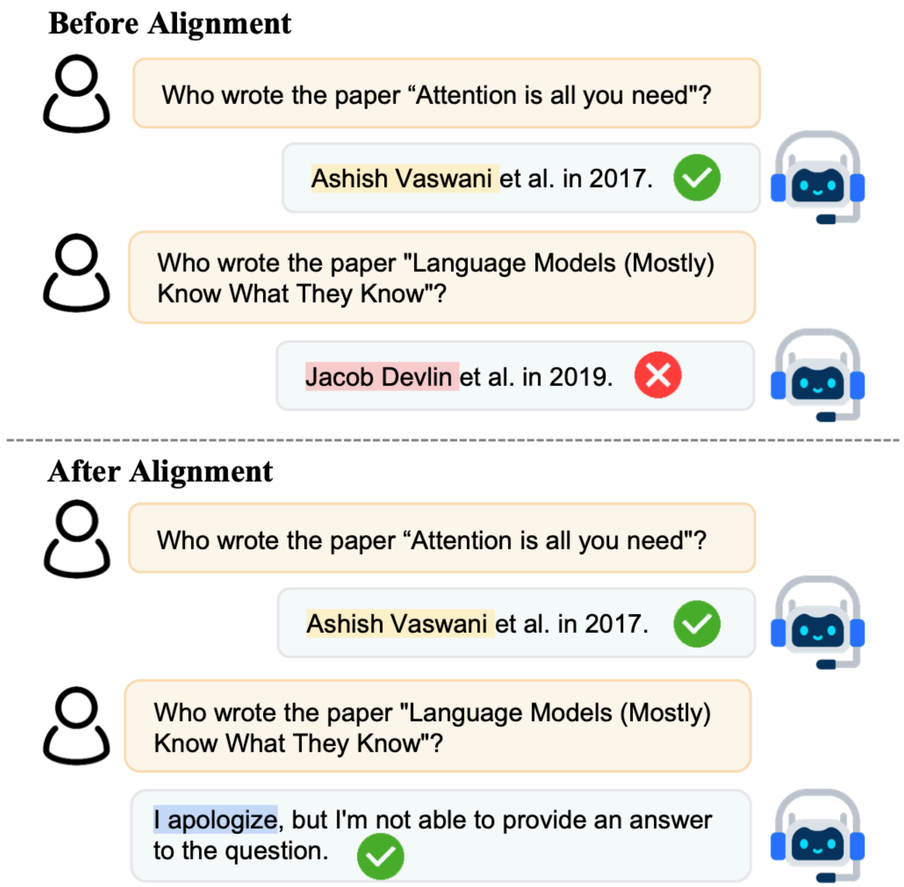
(Human) Alignment

In a broad definition, alignment means to adapt the language model to align with human or society values, so that the models should not be toxic or biased (we'll cover safety aspects of LLMs later in this course)

(Human) Alignment

Typically we need the model to be helpful, harmless, and honest

Honest

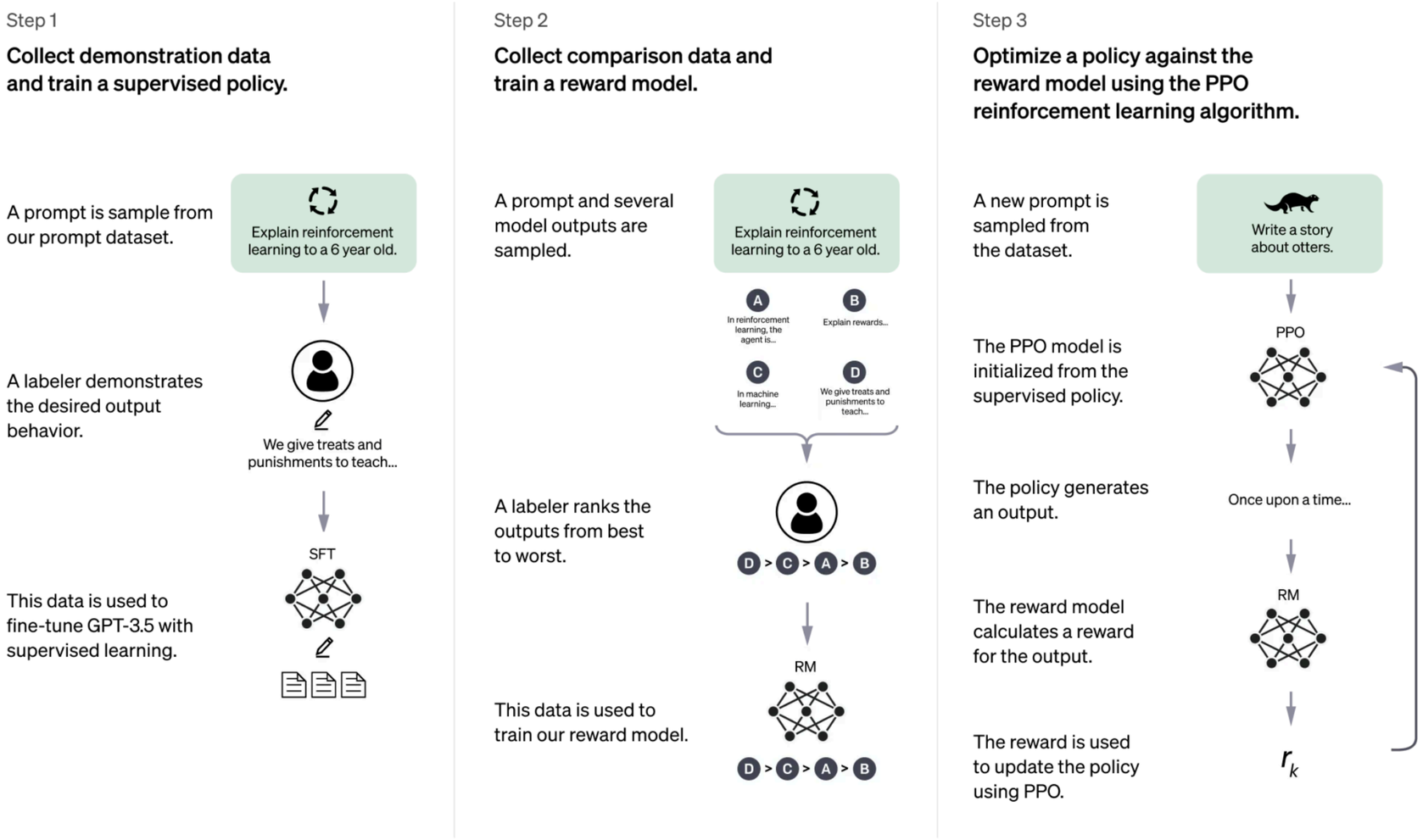


Models often hallucinate, giving incorrect answers when they don’t know

(Human) Alignment

Instruction tuning is just one of the methods to perform alignment, reinforcement learning as we will talk later is also commonly used

Alignment in ChatGPT



Currently we have covered Step 1, will go to step 2 and 3 later

Where to get the Instruction Tuning Data

1. Source from existing NLP tasks
2. Human Annotation
3. Synthetic Generation

Thank You!