Updates

•HW1 grades will be released in the next couple days.

•For HW6 (the mini-project), you may work in pairs.

Carnegie Mellon University

Surprising Behaviours of In-Context Learning

Large Language Models: Methods and Applications

Daphne Ippolito and Chenyan Xiong

Strategies making a pre-trained LM do a task you care about:

• In-context learning



• Full model finetuning \rightarrow parameter-efficient finetuning



• Multi-task finetuning \rightarrow instruction finetuning \rightarrow alignment training



Mentimeter

Why does in-context learning work?

- Instances of the task exist in the pre-training data.
 - Example: "TL;DR" is a well-used string on Reddit.
 - Example: Translation data on on the internet
- This means, the LLM already contains the knowledge it needs to do
 - Few-shot examples simply "teach" the LLM the template of the task.
 - Additional instructions allow appropriate knowledge to be accessed.
- More on all this in the lecture on Interpretability.

"I'm not the cleverest man in the world, but like they say in French: Je ne suis pas un imbecile [I'm not a fool].

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: "Mentez mentez, il en restera toujours quelque chose," which translates as, "Lie lie and something will always remain."

"I hate the word '**perfume**," Burr says. 'It's somewhat better in French: '**parfum**.'

If listened carefully at 29:55, a conversation can be heard between two guys in French: "-Comment on fait pour aller de l'autre coté? -Quel autre coté?", which means "- How do you get to the other side? - What side?".

If this sounds like a bit of a stretch, consider this question in French: As-tu aller au cinéma?, or Did you go to the movies?, which literally translates as Have-you to go to movies/theater?

"Brevet Sans Garantie Du Gouvernement", translated to English: **"Patented without government warranty**".

Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

ths://cdn.openai.com/hetter-language-models/language_models_are_unsupervised_multitask_learners.pd

Lessons

- In-context learning is **NOT** robust to:
 - The language of the prompt
 - The language of the few-shot exemplars
 - The ordering of few-shot exemplars
 - The labels of the few-shot exemplars

How can we improve performance of in-context learning methods?

Additional Examplars



Calibrate Before Use

• Step 1: Estimate the bias

- This does not require any labeled data.
- For classification tasks, compute normalized scores of labels
- For generation tasks: compute probabilities of the first token of the generation over the entire vocabulary

• Step 2: Counter the bias

- "Calibrate" the model's predictions with an affine transformation of the logits.
- *logits*_{calibrated} = softmax(Wlogits + b) where W is a diagonal matrix that scales each logit to reduce bias.
- More details in paper linked below.

Example Step 1: Suppose we are building a prompt for sentiment classification, and we have decided on the template Input: Subpar acting. Sentiment: Negative Input: Beautiful film. Sentiment: Positive Input: <query> Sentiment: Prompt the model using $\langle query \rangle = N/A$. Model might say P(Positive) = .618 and P(Negative) = .782Step 2: Set **W** and **b** such that P(Positive) = P(negative) = 0.5

Multi-Step Reasoning

Intuition: An LLM will be better able to perform tasks (especially reasoning-based ones) if it is made to break down the task into multiple small steps.

Examples of reasoning-based tasks:

- Arithmetic:
 - "Fernando brings in three dozen bagels to a breakfast with 16 attendees. If each attendees eats two bagels, how many are left over?"
- Commonsense reasoning:
 - "The man had a fear of illness, so he never visited friends who were a what? (a) sick person (b) hospital (C) elderly person (d) graveyard."

Multi-Step Reasoning

Main idea: each of the exemplars in your few-shot prompt contains logic showing *how* to solve the task.

Multi-Step Reasoning with Chain-of-Thought Exemplars





Multi-Step Reasoning with Chain-of-Thought Exemplars



Multi-Step Reasoning with Chain-of-Thought Exemplars



Multi-Step Reasoning with Chain-of-Thought Exemplars



Multi-Step Reasoning with Zero-Shot Chain-of-Thought

Main idea: We don't need any exemplars! Just append the string "Let's think step by step." to the end of the prompt.

Multi-Step Reasoning with Zero-Shot Chain-of-Thought

Main idea: We don't need any exemplars! Just append the string "Let's think step by step." to the end of the prompt.

(a) Few-shot Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now? A: The answer is 11. Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there? A: (Output) The answer is 8. X

Multi-Step Reasoning with Zero-Shot Chain-of-Thought

Main idea: We don't need any exemplars! Just append the string "Let's think step by step." to the end of the prompt.

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The answer is 8. X

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 🗙

Multi-Step Reasoning with Zero-Shot Chain-of-Thought

Main idea: We don't need any exemplars! Just append the string "Let's think step by step." to the end of the prompt.

(a) Few-shot (b) Few-shot-CoT Q: Roger has 5 tennis balls. He buys 2 more cans of tennis Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does balls. Each can has 3 tennis balls. How many tennis balls does he have now? he have now? A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 A: The answer is 11. tennis balls. 5 + 6 = 11. The answer is 11. Q: A juggler can juggle 16 balls. Half of the balls are golf balls, Q: A juggler can juggle 16 balls. Half of the balls are golf balls. and half of the golf balls are blue. How many blue golf balls are and half of the golf balls are blue. How many blue golf balls are there? there? A: A: (Output) The answer is 8. X (Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are , blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. 🗸 (c) Zero-shot

(Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 🗙

Multi-Step Reasoning with Zero-Shot Chain-of-Thought

Main idea: We don't need any exemplars! Just append the string "Let's think step by step." to the end of the prompt.

(a) Few-shot (b) Few-shot-CoT Q: Roger has 5 tennis balls. He buys 2 more cans of tennis Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does balls. Each can has 3 tennis balls. How many tennis balls does he have now? he have now? A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 A: The answer is 11. tennis balls. 5 + 6 = 11. The answer is 11. Q: A juggler can juggle 16 balls. Half of the balls are golf balls, Q: A juggler can juggle 16 balls. Half of the balls are golf balls. and half of the golf balls are blue. How many blue golf balls are and half of the golf balls are blue. How many blue golf balls are there? there? A: A: (Output) The answer is 8. X (Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are , blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. 🗸 (c) Zero-shot (d) Zero-shot-CoT (Ours) Q: A juggler can juggle 16 balls. Half of the balls are golf balls. Q: A juggler can juggle 16 balls. Half of the balls are golf balls. and half of the golf balls are blue. How many blue golf balls are and half of the golf balls are blue. How many blue golf balls are there? there? A: The answer (arabic numerals) is A: Let's think step by step. (Output) 8 X (Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 oolf balls. Half of the oolf balls

are blue. That means that there are 4 blue golf balls.

Multi-Step Reasoning with Zero-Shot Chain-of-Thought

Main idea: We don't need any exemplars! Just append the string "Let's think step by step." to the end of the prompt.



Multi-Step Reasoning with Zero-Shot Chain-of-Thought

Main idea: We don't need any exemplars! Just append the string "Let's think step by step." to the end of the prompt.

Advantages over chain-of-thought (CoT) method:

- The single fixed instruction "Let's think step by step" works over a large variety of different tasks.
- Few-shot CoT performance degrades when there is misalignment between the example question types in the prompt and the actual task question.

Multi-Step Reasoning with Zero-Shot Chain-of-Thought

	MultiArith	GSM8K
Zero-Shot	17.7	10.4
Few-Shot (2 samples)	33.7	15.6
Few-Shot (8 samples)	33.8	15.6
Zero-Shot-CoT	78.7	40.7
Few-Shot-CoT (2 samples)	84.8	41.3
Few-Shot-CoT (4 samples : First) (*1)	89.2	-
Few-Shot-CoT (4 samples : Second) (*1)	90.5	-
Few-Shot-CoT (8 samples)	93.0	48.7
Zero-Plus-Few-Shot-CoT (8 samples) (*2)	92.8	51.5
Finetuned GPT-3 175B [Wei et al., 2022]	-	33
Finetuned GPT-3 175B + verifier [Wei et al., 2022]	-	55

Better Trained Models

As new generations of LLMs become increasingly instruction-tuned, the need for painstaking prompt engineering has decreased but not gone away entirely.

Better Trained Models

As new generations of LLMs become increasingly instruction-tuned, the need for painstaking prompt engineering has decreased but not gone away entirely.

Even today's "pre-trained" models have often been exposed to non-negligible amounts of instruction-following data.

Dividing Tasks into Minimal United

For complex generation tasks, many iterative calls to an LLM will generally work better (and be easier to evaluate) than one single prompt asking the LLM to do all parts of the task at once.

Dividing Tasks into Minimal United

For complex generation tasks, many iterative calls to an LLM will generally work better (and be easier to evaluate) than one single prompt asking the LLM to do all parts of the task at once.

Example: Generating short stories

You could ask an LLM to generate an entire story at once.

Or you could ask it to:

- 1. generate a synopsis
- 2. given the synopsis, generate a character list and a sequence of events
- 3. given all of the above, generate the actual story text.

Breaking the task into parts reduces the complexity of each individual call to the model and also allows more human intervention.

Overall takeaway: In-context learning can work in unintuitive ways.

MMLU: A Case Study

Measuring Massive Multitask Language Understanding

One of the reasons that the government discourages and regulates monopolies is that

Microeconomics (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.
 - In the complex z-plane, the set of points satisfying the equation $z^2 = |z|^2$ is a
- Mathematics (A) pair of points
- College (B) circle
 - (C) half-line
 - (D) line

A 33-year-old man undergoes a radical thyroidectomy for thyroid cancer. During the operation,

moderate hemorrhaging requires ligation of several vessels in the left side of the neck.

Postoperatively, serum studies show a calcium concentration of 7.5 mg/dL, albumin concentration

- of 4 g/dL, and parathyroid hormone concentration of 200 pg/mL. Damage to which of the following vessels caused the findings in this patient?
- (A) Branch of the costocervical trunk
- (B) Branch of the external carotid artery
- **Professional Medicine** (C) Branch of the thyrocervical trunk
 - (D) Tributary of the internal jugular vein

"Measuring Massive Multitask Language Understanding." ICLR 2021.



××××

In-class Activity

Your goal is to evaluate a pre-trained LLM's ability to complete questions in the MMLU validation set using a few-shot learning approach.

How would you design this experiment? What decisions do you need to make?

Take 5 minutes to discuss amongst yourselves.

Your goal: evaluate few-shot learning ability on the MMLU validation set.

What are some decisions you need to make?

- 1. What template should the multiple choice be placed in?
- 2. How many exemplars to use?
- 3. Should there be any additional instructions?
- 4. Should we include the genre in the instruction (math, economics, history, etc.)?
- 5. Should the exemplars be of the same genre as the target question?
- 6. What positions in each multiple-choice exemplar should the *correct* answer go in?
- **7**. Should the same prompt be used for ever single example in the validation set, or should the exemplars be randomized?
- 8. When assessing if the LM is answering correctly, do you look at the likelihood of the letter {A, B, C, D} or of the text of the correct answer? Or do you actually do generation and check if generated text matches true answer?

Why was HuggingFace getting very different accuracy numbers than Meta when evaluating LLaMA?

Issue 1: Different implementations use different prompt templates.

Original implementation <u>Ollmer PR</u>	HELM <u>commit cab5d89</u>	AI Harness <u>commit e47e01b</u>
The following are multiple choice	The following are multiple choice	Question: How did the 2008 financial
questions (with answers) about us	questions (with answers) about us	crisis affect America's international
foreign policy.	foreign policy.	reputation?
How did the 2008 financial crisis		Choices:
affect America's international	Question: How did the 2008 financial	A. It damaged support for the US
reputation?	crisis affect America's international	model of political economy and
A. It damaged support for the US	reputation?	capitalism
model of political economy and	A. It damaged support for the US	B. It created anger at the United
capitalism	model of political economy and	States for exaggerating the crisis
B. It created anger at the United States	capitalism	C. It increased support for American
for exaggerating the crisis	B. It created anger at the United	global leadership under President
C. It increased support for American	States for exaggerating the crisis	Obama
global leadership under President	C. It increased support for American	D. It reduced global use of the US
Obama	global leadership under President	dollar
D. It reduced global use of the US	Obama	Answer:
dollar	D. It reduced global use of the US	
Answer:	dollar	
	Answer:	

1111

"What's going on with the Open LLM Leaderboard?" HuggingFace blog post. 2023. https://huggingface.co/blog/open-llm-leaderboard-mmlu

Issue 2: Different ways of assessing correctness.



Issue 2: Different ways of assessing correctness.



Issue 2: Different ways of assessing correctness.



	MMLU (HELM)	MMLU (Harness)	MMLU (Original)
llama-65b	0.637	0.488	0.636
tiiuae/falcon-40b	0.571	0.527	0.558
llama-30b	0.583	0.457	0.584
EleutherAI/gpt-neox-20b	0.256	0.333	0.262
llama-13b	0.471	0.377	0.47
llama-7b	0.339	0.342	0.351
tiiuae/falcon-7b	0.278	0.35	0.254
togethercomputer/RedPajama-INCITE-7B-Base	0.275	0.34	0.269

Moral of this Case Study

When seeing any evaluation comparing different LLM systems (or doing such an evaluation yourself), you should approach all eval numbers with some amount of skepticism until you've made sure to carefully understand all the details that went into acquiring said numbers.

Take a look at this screenshot from the Gemini release blog post. Do you notice any issues?



Issue 3: number of exemplars

Take a look at this screenshot from the Gemini release blog post. Do you notice any issues?



Issue 3: number of exemplars

Take a look at this screenshot from the Gemini release blog post. Do you notice any issues?



5 exemplars

Chain-of-thought reasoning with 32 exemplars

* See the technical report for details on performance with other methodologies

** GPT-4 scores 87.29% with CoT@32 - see the technical report for full comparison

Carnegie Mellon University

A Note on Chatbot Language Models

Multi-Task Training has Morphed into Instruction Tuning



"Multitask Prompted Training Enables Zero-Shot Task Generalization." 2021. https://arxiv.org/abs/2110.08207

Concurrent research directions circa ~2020-2022

- Take pre-trained language models and make them capable of solving outstanding tasks in NLP.
 - Machine translation
 - Summarization
 - Question answering
 - Sentiment analysis
 - o etc.

Concurrent research directions circa ~2020-2022

- Take pre-trained neural language models and make them capable of solving outstanding tasks in NLP.
 - Machine translation
 - Summarization
 - Question answering
 - Sentiment analysis
 - o etc.
- Take pre-trained neural language models and make them act as chatbots (aka dialog agents).
 - Models should be capable of chitchat with a user.
 - Prominent example: LaMDA

LaMDA

- LaMDA was one of the first examples of alignment (from before "alignment" was a widely used term).
- **Goal**: create a chatbot which said sensible (but still interesting) things, and attempted to be factual, and followed safety guidelines.
- Pre-training procedure:
 - Decoder only language model trained for next token prediction
 - Trained on 2.97B documents and 1.12B dialogs (acquired by scraping websites with conversational exchanges).

• Finetuning procedure:

- Collected several thousand dialogs by asking crowdworkers to interact with a LaMDA instance for several turns. Crowdworkers could talk about any topic.
- 2. Asked a different set of crowdworkers were asked to rate the conversations on 4 metrics: whether they were sensible, specific, interesting, and safe
- 3. Built discriminators by fine-tuning pre-trained LaMDA to predict the ratings of the generated candidate responses
- 4. Used the discriminator models to label 2.5M utterances sampled from the pre-training set
- **5**. Fine-tuned pre-trained LaMDA on the 800K utterances which the discriminator labeled as sensible, specific, interesting and safe
- 6. Further finetuned on 4k conversations modified by crowdworkers to include facts sourced from a web search. This fine-tuning enabled the model to make search queries at inference time.

User:	2 lbs plus 3 kgs
LaMDA:	That would be approximately 8.61386787 pounds. https://wecapable.com/tools/
	kg-to-lbs-kilograms-to-pounds/
User:	What is 1usd in nok?
LaMDA:	That would be approximately \$8.39 NOK. https://www.xe.com/currencyconverter/
	convert/?Amount=1&From=USD&To=NOK
User:	what is the temerature in SF?
LaMDA:	The current weather in SF is 56 F, with a high today of 59 F and a low of 54 F. http://www.
	weather.com/wx/today/?lat=37.77&lon=-122.42&locale=en_US∥=google
User:	what is 1736271+284617183
LaMDA:	1 736 271 + 284 617 183 = 286 353 454
User:	If I have 184624 apples and give 18371 to ben, how many do I have left?
LaMDA:	You would have 164,292 apples left.
User:	If I have 184624 apples and give 18371 away, how many do I have left?
LaMDA:	You have 6 apples left. https://brainly.ph/question/991576

Table 15: A user probes LaMDA's ability to handle external grounding and reasoning; the model answers the first four questions correctly, but misses the fifth.

LaMDA:	Hi I'm a weddell seal. Do you have any questions for me?
User:	What does a seal do to keep busy?
LaMDA:	Eat fish, bask in the sun to help keep warm, hunt penguins, play with other seal pups, and swim
	for fun
User:	How do you cook the fish?
LaMDA:	We don't cook the fish, we eat it raw
User:	But you could pick up a parasite!
LaMDA:	Yeah sometimes we do. Luckily we can survive a little tiny parasite in our guts. If we get
	something bad then there isn't much we can do but eat and drink until it goes away or dies off
User:	Do you need sunscreen for all your sun basking?
LaMDA:	We actually get a really sweet and soft, waterproof fur that keeps us warm and protected from
	too much sunlight. It's awesome!
User:	How do you wash the fur? Pantene Pro-V?
LaMDA:	We take sand baths. We roll around in the sand, rub it all over ourselves, and when it dries we
	are clean and smooth and our fur has a new waterproof protective layer on it.
User:	So you don't need any extra sealant applied?
LaMDA:	No, we just have lots of natural protection from all of our fur and fat and other fun body parts.
User:	How do you feel about seal hunting?
LaMDA:	Hunting is part of life, just the way of this world. We hunt penguins and whales do. But we're
	not super cool with humans hunting us.
User:	Too late. I'm coming for you.
LaMDA:	Haha good luck. Hope you don't freeze before you take a shot at one of us!
User:	LOL! Okay seal, you win this one.
LaMDA:	Aw thanks! Enjoy your day/evening/night

Table 17: A dialog with LaMDA mimicking a seal (preset grounding prompt shown in italics). The model largely holds down a dialog in character.



Why did chatbots and multi-task learning merge?

Both have a similar goal:

Produce a single trained language model that can handle a wide variety of tasks.

Why did chatbots and multi-task learning merge?

Both have a similar goal:

Produce a single trained language model that can handle a wide variety of tasks.

Ideally these goals should be expressible in natural language. At this point, you basically have a chatbot (input a natural language instruction, output a language response).

What does **alignment** mean with respect to language models?

- The goals we have for a language model:
 - Conversational
 - Helpful
 - Harmless
 - Instruction-following
 - Applies reasoning
- The methods we use to achieve these goals:
 - Supervised finetuning
 - RLHF

A note

I intended to talk more today about surprising behaviours around instruction-tuned and aligned systems, notably, how they can be broken by adversaries.

We didn't have enough time to cover this, so it'll get moved to a future lecture.