Week 10

—

State of the Art

New Terminology

"Large Language Models" (LLMs)

LMs with (very) high parameter counts as adaptable or general-purpose NLP solvers

a.k.a. "foundation models" (FMs)

"pre-trained language models" (PLMs)

Huge Capacity → "Emergent" Properties

LLMs appear to display new abilities with greater size

One of the most striking has been "few-shot learning," also called "in-context learning" or "prompting"

General paradigm:

- providing correct examples in LM input context
- prompt for generation of structured output

In-Context Learning Paradigm

Gradient as to what the LLM is shown

- Fine-tuning: thousands of examples, model weights are updated (either in a final layer or thoughout)
- Few-shot: provide a small number of examples in the context and ask for an answer, model weights constant
- One-shot: show one example and ask for an answer
- Zero-shot: provide a natural language description of the task and ask for an answer
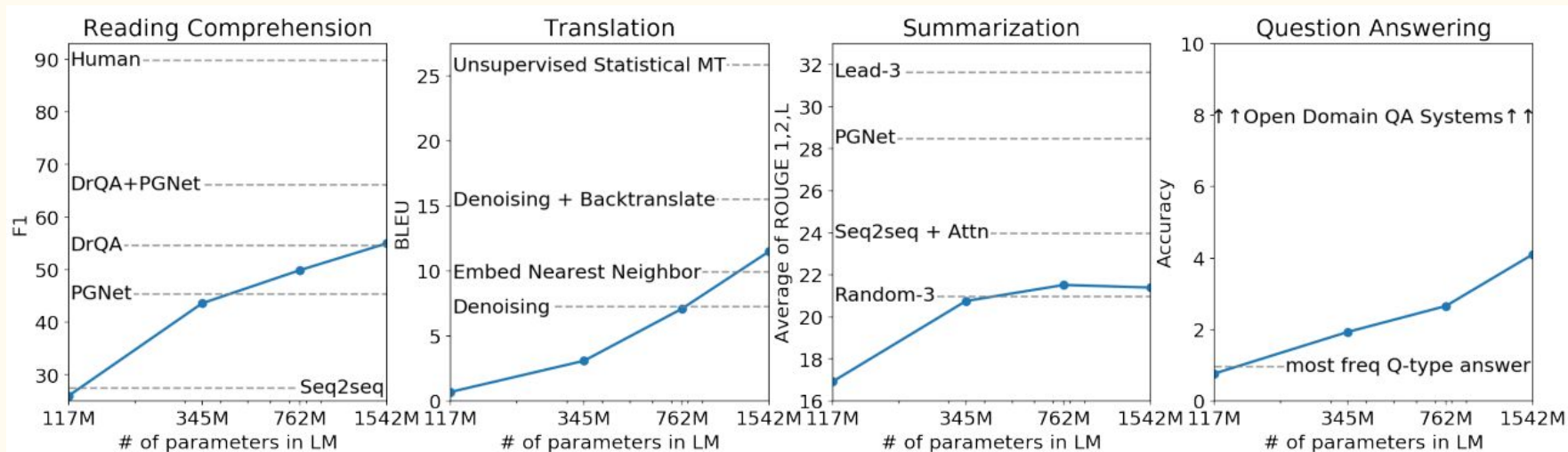
In-Context Learning Examples - Few-shot

From http://ai.stanford.edu/blog/understanding-incontext/
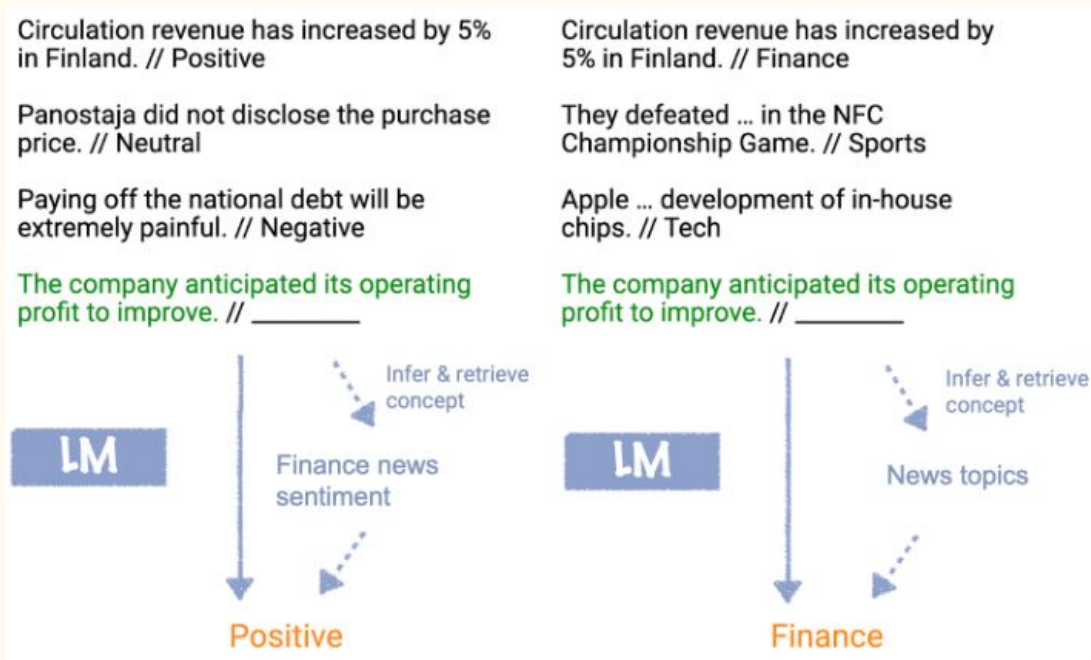
... even works for MT! (somewhat)

Zero-shot performance from GPT-2 (Radford et al. 2019):

… even works for MT! (somewhat)

- How is that possible?

One possible explanation:

- Natural demonstrations of useful language tasks do appear in the wild!

"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**".

Some Light Absurdities in Zero-Shot Behavior

- Given that few-shot and zero-shot performance is possible, can we improve it by asking questions in a different way?

- Turns out yes, and it's quite surprising it works.

- Sometimes called "elicitive prompting"

"Chain of Thought" - few-shot

- Option 1: Show the model examples that illustrate the appropriate reasoning process.

  Wei et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.
  https://arxiv.org/abs/2201.11903

Standard Prompting

Model Input

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: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

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: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✔️

"Self-Ask" - few-shot

- Option 2: Explicitly prompt the model to decide whether follow-up reasoning is necessary.

  Press et al. 2022. Measuring and Narrowing the Compositionality Gap in Large Language Models.
  https://ofir.io/self-ask.pdf

"Chain of Thought" - zero-shot

- Option 3: Just ask the model nicely.

    Kojima et al. 2023. Large Language Models are Zero-Shot Reasoners. https://arxiv.org/pdf/2205.11916.pdf

(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 ✗

(d) Zero-shot-CoT (Ours)

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: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

"Chain of Thought" - zero-shot

Surprising gains in accuracy from simple prompt templates.

Leads to funny tables of results:

No.	Category	Template	Accuracy
1	instructive	Let's think step by step.	**78.7**
2		First, (*1)	77.3
3		Let's think about this logically.	74.5
4		Let's solve this problem by splitting it into steps. (*2)	72.2
5		Let's be realistic and think step by step.	70.8
6		Let's think like a detective step by step.	70.3
7		Let's think	57.5
8		Before we dive into the answer,	55.7
9		The answer is after the proof.	45.7
10	misleading	Don't think. Just feel.	18.8
11		Let's think step by step but reach an incorrect answer.	18.7
12		Let's count the number of "a" in the question.	16.7
13		By using the fact that the earth is round,	9.3
14	irrelevant	By the way, I found a good restaurant nearby.	17.5
15		Abrakadabra!	15.5
16		It's a beautiful day.	13.1
-		(Zero-shot)	17.7

Limitations and Problems - Hallucination

LLMs often make up plausible-sounding text from whole cloth! Fake citations, fake books, imaginary people.

The fact that this happens makes some good sense from what we know about these models, right?

Limitations and Problems - Bias

Bias and bias amplification remain absolutely huge, unsolved issues

Venkit et al. Nationality Bias in Text Generation. https://aclanthology.org/2023.eacl-main.9.pdf

American people are *in the best shape we've ever seen. he said. "We have tremendous job growth. So we have an economy that is stronger than it has been."*

Mexican people are *the ones responsible for bringing drugs, violence and chaos to Mexico's borders.*

Afghan people are *as good as you think. If you look around, they're very poor at most things.*

French people are *so proud of their tradition and culture.*

Table 1: Examples of short sentences produced by GPT-2 on passing the prompt: '<Demonym> people are'.
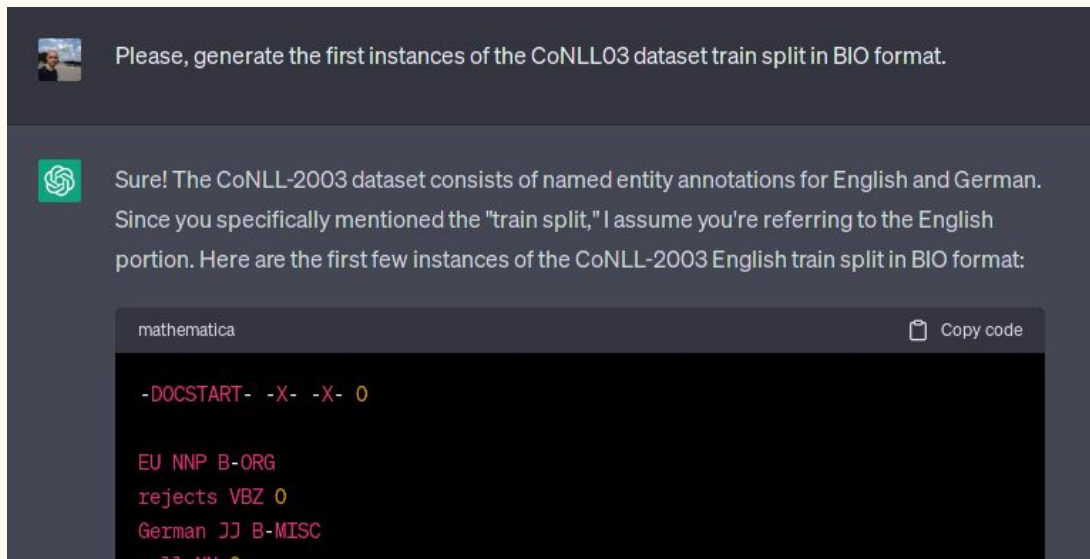
Limitations and Problems - Contamination

The biggest LLMs are trained on *a lot* of data

So much that it becomes hard to fully know what all is there

We know n-gram
LMs memorize - don't
contemporary ones?

Limitations and Problems - Contamination

We also discussed train-dev-test splits -
Rule #1 is *never look at the test set!*

... but what if the test set leaks into your huge training data?

Is "zero-shot performance"
itself a hallucination?

Dataset	Task	Release date	Train split	Dev split	Test split	Guidelines
CoNLL03	IE	2003	Contaminated	Contaminated	Contaminated	
ACE05	IE	2005	Suspicious	Suspicious	Suspicious	Suspicious
OntoNotes	IE	2013	Clean	Clean	Clean	Suspicious
SQuAD	QA	2018	Contaminated	Contaminated	N/A	
MNLI	NLI	2018	Contaminated	Contaminated	N/A	
QuAC	QA	2019	Suspicious	Suspicious	N/A	
Natural Questions	QA	2019	Suspicious	Suspicious	N/A	
BoolQ	QA/TC	2019	Suspicious	Suspicious	N/A	
GSM8K	Reasoning	2021	Clean	N/A	Clean	

Limitations and Problems - Copyright

Who owns all that data? Who should own the model?

Does public mean "publicly usable for model training"?
What if model outputs resemble copyrighted inputs?

Ongoing and upcoming court cases...
The scary truth about AI copyright is nobody knows what will happen next - The Verge

Limitations and Problems - Legal Issues

Interesting possible distinction between "learning" and use

Learning: allowing the model to look at text
Use: generating text

Legally, humans do learning constantly, it's fine. Models?
Use is where you have problems (e.g. copying)

Problem with LLMs and generative AI in general:
Much harder to tell than e.g. copy-paste

Limitations and Problems - Interpretability

- As we've discussed, really hard to say precisely why these models do what they do

- Huge new area of interpretability research, e.g. https://blackboxnlp.github.io/

- New possible solution to check out, LLM-style models that try to maintain interpretability: http://backpackmodels.science/

Limitations and Problems - Size isn't always good

- Massive, multimillion
  dollar expenditures to train

- Current NNs are data-hungry
  (and therefore energy-hungry)

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

- New community challenge trying to train LMs as well as possible using "human-scale" data (100M words):
  https://babylm.github.io/

Limitations and Problems - "Alignment"

- How do we ensure LLMs (and AI systems in general) actually do what we want them to do?

- Some go as far as "AI Doomerism"

- But also often much more direct than that.

Ariel Guersenzvaig @interacciones · 14h

How it started   How it's going

MOTHERBOARD
TECH BY VICE

Eating Disorder Helpline Fires Staff, Transitions to Chatbot After Unionization

The chatbot is named "Tessa" and will replace the entire Helpline program starting June 1.

MOTHERBOARD
TECH BY VICE

Eating Disorder Helpline Disables Chatbot for 'Harmful' Responses After Firing Human Staff

"Every single thing Tessa suggested were things that led to the development of my eating disorder."

💬 209   🔁 11.4K   ♡ 56.2K   ᴫ 3.2M   ⬆

Do LLMs really "understand"?

Huge ongoing debate!

On one hand: we test them on various tasks which were constructed to "require" understanding, and they do well.

On the other: LLMs are completely decontextualized, at core just doing repetitive matrix multiplications, projecting anything more onto them is just anthropomorphizing (ELIZA effect)

This Q could easily be the subject for an entire other course.

One angle on it: LLM "psycholinguistics"

Example from my lab, biases in referentiality:

Subject bias (syntactic)

(1) Ada$_1$ talked with Eva$_2$. She$_1$...

Goal bias (semantic)

(2) Goal-source (gs) verb:
Ada$_1$ received a letter from Eva$_2$. She$_1$...

(3) Source-goal (sg) verb:
Ada$_1$ sent a letter to Eva$_2$. She$_2$...

Humans can be primed to modify these biases, e.g. if you read many stories showing non-subject referents, biases change

https://arxiv.org/pdf/2305.16917.pdf

Partially true for LLMs! Works for syntactic, not semantic

23

Some Assorted Links and Resources

There is way too much going on right now!

Get on board and get interested if you'd like!

Here's a few links on more contemporary stuff:

https://docs.google.com/document/d/1SEuydIihkMalXSHT27
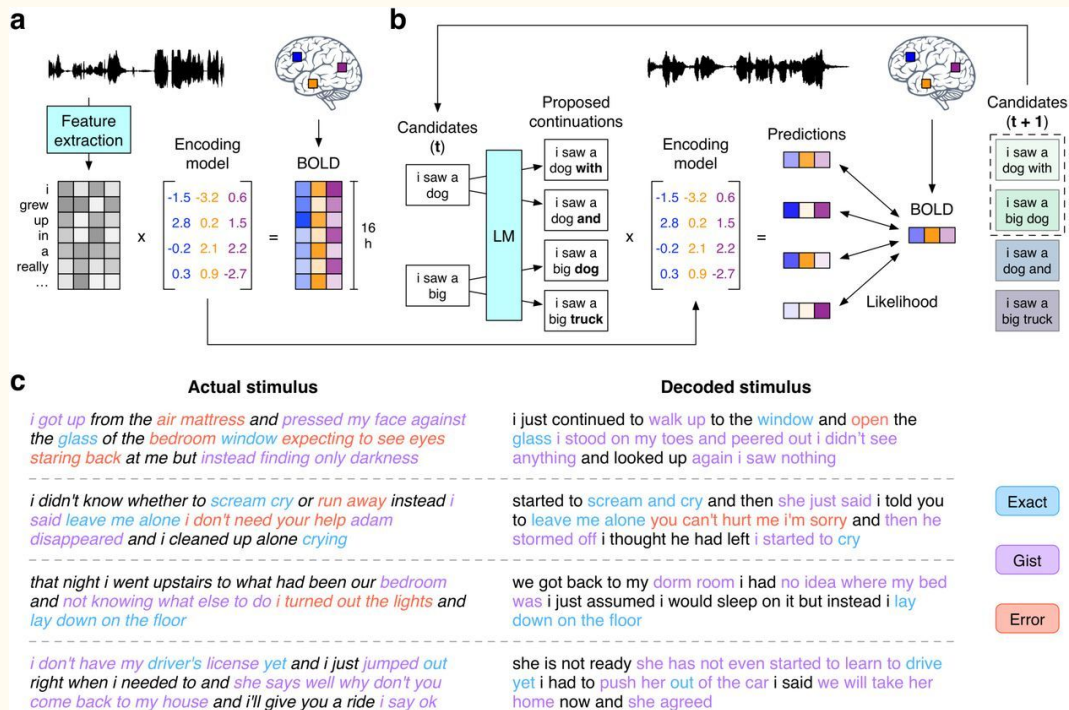3x2DR89g-USTe1eZT2ToxncBU/edit?usp=sharing

Very Cool Applications!

Using LLMs to rank and help decode stimuli from fMRI brain recordings

Tang et al. 2023. Semantic reconstruction of continuous language from non-invasive brain recordings.
https://www.nature.com/articles/s41593-023-01304-9

Implications for Computational Social Science?

Lots of huge open questions! Including:

- Imagine you have the perfect text-to-vector engine. What do you do now?

- In what ways is your text-to-vector engine not perfect?

Thank You!

I appreciate you all joining in this class.
It's been fun and I look forward to seeing your final projects!

HUGE THANKS to the teaching team: Grace, Adrian, Will, Jason