Administrivia

‣ Project proposals due today

‣ Project 3 released today
Recap: Zero-shot/Few-shot prompting

- Single unlabeled datapoint $x$, want to predict label $y$

  $x = \text{The movie’s acting could’ve been better, but the visuals and directing were top-notch.}$

- Wrap $x$ in a template we call a verbalizer $v$

Review: The movie’s acting could’ve been better, but the visuals and directing were top-notch.

Out of positive, negative, or neutral, this review is neutral

- Need the right prompt (but there is a “plateau” of prompts that work)

- Few-shot: add one or more examples. Typically works better! Particularly with rich examples like we’ll see today
Recap: Understanding ICL

- ICL can learn a strategy like ordinary least-squares (Akyurek et al., 2022)

- We can identify *induction heads* in Transformers; these emerge when ICL performance improves (Olsson et al., 2022)
This Lecture

- Text rationales: text explanations of answers
- Chain-of-thought prompting (zero- and few-shot)
- Extensions
- Analysis of explanations
Text Rationales
Example from Vision

Laysan Albatross

Description: This is a large flying bird with black wings and a white belly.

Class Definition: The Laysan Albatross is a large seabird with a hooked yellow beak, black back and white belly.

Visual Explanation: This is a Laysan Albatross because this bird has a large wingspan, hooked yellow beak, and white belly.

Laysan Albatross

Description: This is a large bird with a white neck and a black back in the water.

Class Definition: The Laysan Albatross is a large seabird with a hooked yellow beak, black back and white belly.

Visual Explanation: This is a Laysan Albatross because this bird has a hooked yellow beak white neck and black back.

▶ What makes a visual explanation? Should be relevant to the class and the image
▶ Are these features really what the model used?

Hendricks et al. (2016)
LSTM decoder looks at a feature vector and predicted label, then generates an explanation from those.

It’s trained on human explanations — so it will likely produce explanations that look good (it learns to be a language model).

Hendricks et al. (2016)
E-SNLI

Premise: An adult dressed in black **holds a stick**.
Hypothesis: An adult is walking away, **empty-handed**.
Label: contradiction
Explanation: Holds a stick implies using hands so it is not empty-handed.

Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her.
Hypothesis: A young **mother** is playing with her **daughter** in a swing.
Label: neutral
Explanation: Child does not imply daughter and woman does not imply mother.

Premise: A **man** in an orange vest **leans over a pickup truck**.
Hypothesis: A man is **touching** a truck.
Label: entailment
Explanation: Man leans over a pickup truck implies that he is touching it.

› Two formats: highlights and text

Camburu et al. (2019)
Generating Explanations: E-SNLI

\[ f = \text{function of premise and hypothesis vectors} \]

- Similar to birds: explanation is conditioned on the label + network state \( f \)
- Information from \( f \) is fed into the explanation LSTM, although we don’t know how that information is being used

Camburu et al. (2019)
Can we generate a natural language explanation of a model’s behavior?

What are some advantages to this?

- Easy for untrained users to understand
- Multitasking to produce human-written explanations may help us learn

What are some risks/disadvantages?
Text Explanations

- Issues with text explanations:
  - Hard to produce/consume (these models are sort of clunky)
  - Hard to know if they faithfully reflect what a model is doing
  - More broadly, hard to evaluate
- However, writing such explanations comes naturally to us...so that means that they reflect some kind of underlying reasoning process that we’re doing?
- Pre-2021: this process would usually be captured structurally in a model. 2022 and beyond: chain of thought
Chain-of-thought
Problem 2:
Question: From a pack of 52 cards, two cards are drawn together at random. What is the probability of both the cards being kings?
Options: A) $2/1223$  B) $1/122$  C) $1/221$  D) $3/1253$  E) $2/153$
Rationale: Let $S$ be the sample space.
Then $n(S) = 52C2 = 1326$
$E =$ event of getting 2 kings out of 4
$n(E) = 4C2 = 6$
$P(E) = 6/1326 = 1/221$
Answer is C
Correct Option: C

- Rationales are most useful for problems where some computation is required. They can articulate the intermediate steps needed to solve it
- Some of the earliest work: math word problems

Ling et al. (2017)
| 9  | cards | Id("Then") | Then | y_9 |
| 10 | are   | Id("n") | n    | y_{10} |
| 11 | drawn | Id("(") | (     | y_{11} |
| 12 | together | Id("s") | s    | y_{12} |
| 13 | at    | Id(")") | )     | y_{13} |
| 14 | random | Id("=") | =     | y_{14} |
| 15 | . Str.to_Float(x_5) | 52   | m_1 |
| 16 | What Float.to_Str(m_1) | 52   | y_{15} |
| 17 | is    | Id("C") | C    | y_{16} |
| 18 | the   | Id("2") | 2    | y_{17} |
| 19 | probability | Id("=") | =    | y_{18} |
| 20 | of Str.to_Float(y_{17}) | 2   | m_2 |
| 21 | both Choose(m_1,m_2) | 1326 | m_3 |
| 22 | cards Float.to_Str(m_3) | 1326 | y_{19} |
| 23 | being Id("E") | E    | y_{20} |
| 24 | kings Id("=") | =    | y_{21} |

- Symbolic computation represented as a latent variable: intermixes computation with production of the string explanation

Ling et al. (2017)
“StrategyQA”: dataset where different reasoning strategies are needed.

- Related to multi-hop QA: “What’s the capital of the country where Aristotle lived?” (but these are easy with current models)
Chain-of-thought

- For these kinds of problems, do “computation” entirely in natural language

- Unifies several ideas:
  - For math: relies on the fact that LLMs can at least do single steps of arithmetic okay
  - For QA: many problems involve reasoning decompositions
    E.g., *What’s the capital of the country where Aristotle lived?* ->
    ans = “*country where Aristotle lived***”
    return *What’s the capital of [ans]*
  - For other tasks: capture the kinds of behavior written in rationales
Chain-of-thought

- Typically a few-shot prompting technique where the in-context examples now contain explanations.

- Answer is not generated in one go, but comes after an explanation that "talks through" the reasoning.

---

**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: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. How many hours a week does he spend taking care of dogs?

A: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. So that is $10 \times .5 = 5$ hours a day. 5 hours a day x 7 days a week = 35 hours a week.

The answer is 35 hours a week. ✅

*Wei et al. (2022)*
Chain-of-thought

From our work: a synthetic test of multi-hop reasoning with extractive explanations:

**Context:** Christopher agrees with Kevin. Tiffany agrees with Matthew. Mary hangs out with Danielle. James hangs out with Thomas. Kevin is a student. Matthew is a plumber. Danielle is a student. Thomas is a plumber.

**Q:** Who hangs out with a student?

**A:** Mary.

- What kind of explanation would you write here?

**Explanation:** *because Mary hangs out with Danielle and Danielle is a student.*
Chain-of-thought

**Context:** Christopher agrees with Kevin. [...] **Q:** Who hangs out with a student?

**Mary**

Standard few-shot learning, no explanation

---

**Context:** Christopher agrees with Kevin. [...] **Q:** Who hangs out with a student?

**Mary, because Mary hangs out with Danielle and Danielle is a student.**

Predict-explain: answer **is not** conditioned on output explanation (original E-SNLI LSTM)

---

**Context:** Christopher agrees with Kevin. [...] **Q:** Who hangs out with a student?

**Because Mary hangs out with Danielle and Danielle is a student, the answer is Mary.**

Explain-predict: answer is conditioned on output explanation (Chain of Thought)
**Context:** Christopher agrees with Kevin. [...] **Q:** Who hangs out with a student?

*Mary, because Mary hangs out with Danielle and Danielle is a student.*

**Context:** Adam plays with Ellen. [...] **Q:** Who plays with a doctor?

*Adam, because Adam plays with Ellen and Ellen is a doctor.*

greedy decoding from GPT-3
Q1: Do these explanations help?

- Not really. Small gains on AdvHotpot and E-SNLI. No one technique dominates.

Does GPT-3 (text-davinci-001) work well without explanations?

- Not well. On Synthetic, surface heuristics give 50%.
Results

- Can language models generate reliable explanations?
  - **Factuality**: whether an explanation is factually grounded in the input context
  - **Consistency**: whether an explanation entails the answer
- Model-generated explanations are not always reliable

**Explain-Predict on SYNTH**

- Factuality:
  - GPT-3: 59.2
  - InstructGPT: 72.8
  - text-davinci-002: 91.6
- Consistency:
  - GPT-3: 64.8
  - InstructGPT: 64.8
  - text-davinci-002: 85.2

**Predict-Explain on SYNTH**

- Factuality:
  - GPT-3: 52.4
  - InstructGPT: 51.6
  - text-davinci-002: 83.2
- Consistency:
  - GPT-3: 96.4
  - InstructGPT: 95.2
  - text-davinci-002: 83.2
Results

Results on SYNTH data

- Instruct tuning helps but it seems to be not quite sufficient
- text-davinci-002 is far ahead of other models on this task

Ye and Durrett (NeurIPS 2022)
Chain-of-thought extensions
Step-by-Step

(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.

• Prompt for step-by-step reasoning: produces chains of thought without including demonstrations

• Separate prompt to extract the answer (“Therefore, the answer is ___”)

Kojima et al. (2022)
# Step-by-Step

<table>
<thead>
<tr>
<th></th>
<th>SingleEq</th>
<th>AddSub</th>
<th>MultiArith</th>
<th>GSM8K</th>
<th>AQUA</th>
<th>SVAMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>zero-shot</td>
<td>74.6/78.7</td>
<td>72.2/77.0</td>
<td>17.7/22.7</td>
<td>10.4/12.5</td>
<td>22.4/22.4</td>
<td>58.8/58.7</td>
</tr>
<tr>
<td>zero-shot-cot</td>
<td>78.0/78.7</td>
<td>69.6/74.7</td>
<td>78.7/79.3</td>
<td>40.7/40.5</td>
<td>33.5/31.9</td>
<td>62.1/63.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Common Sense</th>
<th>Common SenseQA</th>
<th>Strategy QA</th>
<th>Other Reasoning Tasks</th>
<th>Date Understand</th>
<th>Shuffled Objects</th>
<th>Symbolic Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>zero-shot</td>
<td>68.8/72.6</td>
<td>12.7/54.3</td>
<td>49.3/33.6</td>
<td>31.3/29.7</td>
<td>0.2/-</td>
<td>12.8/53.8</td>
</tr>
<tr>
<td>zero-shot-cot</td>
<td>64.6/64.0</td>
<td>54.8/52.3</td>
<td>67.5/61.8</td>
<td>52.4/52.9</td>
<td>57.6/-</td>
<td>91.4/87.8</td>
</tr>
</tbody>
</table>

- text-davinci-002 (fine-tuned model)

Kojima et al. (2022)
<table>
<thead>
<tr>
<th>No.</th>
<th>Category</th>
<th>Template</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>instructive</td>
<td>Let’s think step by step.</td>
<td><strong>78.7</strong></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>First, (*1)</td>
<td>77.3</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Let’s think about this logically.</td>
<td>74.5</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Let’s solve this problem by splitting it into steps. (*2)</td>
<td>72.2</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Let’s be realistic and think step by step.</td>
<td>70.8</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Let’s think like a detective step by step.</td>
<td>70.3</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Let’s think</td>
<td>57.5</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Before we dive into the answer,</td>
<td>55.7</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>The answer is after the proof.</td>
<td>45.7</td>
</tr>
<tr>
<td>10</td>
<td>misleading</td>
<td>Don’t think. Just feel.</td>
<td>18.8</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>Let’s think step by step but reach an incorrect answer.</td>
<td>18.7</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>Let’s count the number of &quot;a&quot; in the question.</td>
<td>16.7</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>By using the fact that the earth is round,</td>
<td>9.3</td>
</tr>
<tr>
<td>14</td>
<td>irrelevant</td>
<td>By the way, I found a good restaurant nearby.</td>
<td>17.5</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>Abrakadabra!</td>
<td>15.5</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>It’s a beautiful day.</td>
<td>13.1</td>
</tr>
<tr>
<td>-</td>
<td>(Zero-shot)</td>
<td></td>
<td>17.7</td>
</tr>
</tbody>
</table>

Kojima et al. (2022)
Demo: Step-by-Step
(Math QA, StrategyQA)
Self-Consistency

Ensembling across multiple outputs (either zero-shot or few-shot)

GSM8k: 56.5 -> 74.4, 5% gains on several other math datasets, lower gains on text tasks

Wang et al. (2022)
Program-aided Language Models

- For math: why are we doing the arithmetic in the LLM itself?
- Instead: generate code fragments and actually execute them to get an answer (how most earlier math word problem systems worked)
- Many flavors of this: “Faithful Chain-of-thought”, “Program-of-thought”, Toolformer, etc.

A: The bakers started with 200 loaves
loaves_baked = 200
They sold 93 in the morning and 39 in the afternoon
loaves_sold_morning = 93
loaves_sold_afternoon = 39
The grocery store returned 6 loaves.
loaves_returned = 6
The answer is
answer = loaves_baked - loaves_sold_morning - loaves_sold_afternoon + loaves_returned

>>> print(answer)
74

Gao et al. (2022)
Similar idea but with QA/a search engine in the loop

Demonstration shows sub-questions and sub-answers, can potentially do search at these intermediate points

Bing’s “Sydney” agent has some capabilities around this

Press et al. (2022)
Other ideas

- For math: can having various other ways of doing programmatic verification

- For natural language reasoning: missing component of search and planning, discussed in “Language Model Cascades”

- For problems like fact-checking or QA involving complex reasoning, its difficult to verify all of the individual steps...so if CoT goes wrong, it may even be hard for a human to spot
Analysis of Explanations
What Makes Explanations Effective?

- Do LMs “follow” explanations?
- Probing LLMs with perturbed explanations
  - Perturbing Computation Trace
  - Perturbing Natural Language

**Question**

Take the last letters of the words in "Bill Gates" and concatenate them.

**Gold Explanation**

The last letter of "Bill" is letter "l". The last of "Gates" is "s". Concatenating “l” and "s" is “ls”. So the answer is ls.

**Perturbing Trace**

The last letter of "Bill" is letter " " . The last of "Gates" is " " . Concatenating “l” and "s" is “ls”. So the answer is ls.

**Perturbing NL**

"Bill","l","Gates","s","l","s","ls" . So the answer is ls.

Ye et al. (2022)
What Makes Explanations Effective?

- Do LMs “follow” explanations? How do explanations work for in-context-learning?
  - YES. Perturbing either trace or NL leads to performance degradation.
  - But perturbed explanations are still beneficial compared to not using explanations at all

Ye et al. (2022)
What Makes A Good Set of Explanations?

- Given a test query, we study how to form a maximally effective set of exemplars $T=(q,e,a)$
  - Interplay between query and exemplar: relevance (using more relevant examples)
  - Interplay between exemplars in the set: complementarity

**Test Query:**

**Q:** Peter bought 20 popsicles at $0.25 each. He bought 4 ice cream bars at $0.50 each. How much did he pay in total?  

**A:** $0.25 \times 20 = 5$. $0.5 \times 4 = 2$. $5 + 2 = 7$. The answer is 7.

---

**Addition Exemplars:**

**Q:** Marion received 20 more turtles than Martha. If Martha received 40 turtles, how many turtles did they receive together?  

**A:** $20 + 40 = 60$. $60 + 40 = 100$. The answer is 100.

**Multiplication Exemplars:**

**Q:** Car Wash Company cleans 80 cars per day. They make $5 per car washed. How much money will they make in 5 days?  

**A:** $8 \times 5 = 40$. $40 \times 5 = 2000$. The answer is 2000.
What Makes A Good Set of Explanations?

- We test whether LLMs can benefit from complementarity of exemplars
- Complementary exemplar sets lead to better performance (in the paper: algorithm for selecting these!)

<table>
<thead>
<tr>
<th></th>
<th>InstructGTP</th>
<th>OPT</th>
<th>text-davinci-002</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Add</td>
<td>13.5</td>
<td>6.8</td>
<td>50.3</td>
</tr>
<tr>
<td>Mul</td>
<td>17.2</td>
<td>4.7</td>
<td>50.1</td>
</tr>
<tr>
<td>Mixture</td>
<td>18.9</td>
<td>7.0</td>
<td>52.0</td>
</tr>
<tr>
<td>Let Concat</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First</td>
<td>16.0</td>
<td>32.5</td>
<td>100.0</td>
</tr>
<tr>
<td>Second</td>
<td>49.5</td>
<td>37.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Mixture</td>
<td>56.5</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>
Takeaways

- Chain-of-thought prompting (zero- and few-shot) can work well for tasks involving reasoning, especially mathematical reasoning and textual question answering with multiple steps.

- Several things needed to improve them, such as self-consistency and the ability to use other resources like code execution or APIs.

- Next time: RLHF, makes models better at zero-shot prompting and producing well-structured chain-of-thought responses.