CS388: Natural Language Processing

Lecture 12: ICL 2:
Text rationales,
Chain-of-thought

Greg Durrett

Administrivia

- Project 3 due in two weeks
- FP proposals back early next week

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 GPT-3 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.

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

Generating Explanations: Birds

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

- \( f \) = 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)

Text Explanations

- 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
**Text rationales vs. programs**

**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

---

**QA**

Q1: Did Aristotle use a laptop?
   - implicit

Q2: Was Aristotle alive when the laptop was invented?
   - explicit

D
1. When did Aristotle live?
2. When was the laptop invented?
3. Is #2 before #1?

E
- "Aristotle (384-322 BC) was a philosopher..."
- "The first laptop was... in 1980"

A
- No

- “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

---

Geva et al. (2021)

Wei et al. (2022)
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

<table>
<thead>
<tr>
<th>Context: Christopher agrees with Kevin. [...] Q: Who hangs out with a student?</th>
<th>Mary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard few-shot learning, no explanation</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context: Christopher agrees with Kevin. [...] Q: Who hangs out with a student?</th>
<th>Mary, because Mary hangs out with Danielle and Danielle is a student.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predict-explain: answer is not conditioned on output explanation (original E-SNLI LSTM)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context: Christopher agrees with Kevin. [...] Q: Who hangs out with a student?</th>
<th>Mary, because Mary hangs out with Danielle and Danielle is a student.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explain-predict: answer is conditioned on output explanation (Chain of Thought)</td>
<td></td>
</tr>
</tbody>
</table>

Chain-of-thought

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

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

<table>
<thead>
<tr>
<th>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.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q: Who hangs out with a student?</td>
</tr>
<tr>
<td>A: Mary</td>
</tr>
</tbody>
</table>

What kind of explanation would you write here?

Explanation: because Mary hangs out with Danielle and Danielle is a student.

Ye and Durrett (NeurIPS 2022)
Results

Q1: Do these explanations help?
- Not well. Small gains on AdvHotpot and E-SNLI. No one technique dominates.
- Not really. Small gains on AdvHotpot and E-SNLI. No one technique dominates.

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

Does GPT-3 (text-davinci-001) work well without explanations?
- Not well. On Synthetic, surface heuristics give 50%.

Accuracy (avg over 5 trials)

- Synthetic (16 exs)
- AdvHotpot (6 exs)
- E-SNLI (32 exs)

GPT-3 (text-davinci-001)
- Instruct tuning helps but it seems to be not quite sufficient
- Bigger models are better, and modern models are very good

Gold Explanation
- Take the last letters of the words in “Bill Gates” and concatenate them.

Perturbing Trace
- 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 NL
- “Bill”, “l”, “Gates”, “s”, “l”, “s”, “ls”. So the answer is ls.
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.

**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 there are 8 golf balls. Half of the golf balls are blue. That means 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 ___”)

**Step-by-Step**

<table>
<thead>
<tr>
<th>Arithmetic</th>
<th>SingleEq</th>
<th>AddSub</th>
<th>MultiArith</th>
<th>GSMBK</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>Other Reasoning Tasks</th>
<th>Symbolic Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common SenseQA</td>
<td>Strategy QA</td>
<td>Date Understand</td>
</tr>
<tr>
<td>zero-shot</td>
<td>68.8/72.6</td>
<td>12.7/54.3</td>
</tr>
<tr>
<td>zero-shot-cot</td>
<td>64.6/64.0</td>
<td>54.8/52.3</td>
</tr>
</tbody>
</table>

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

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

<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>78.7</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 “a” 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)

Chain-of-thought

Models have been RLHFed to do this without prompting

You

From a pack of 52 cards, two cards are drawn at random. What is the probability of both cards being kings?

ChatGPT

To calculate the probability of both cards drawn from a pack of 52 cards being kings, we can use the formula:

\[
P(A \text{ and } B) = \frac{\text{Number of ways } A \text{ and } B \text{ can happen}}{\text{Total number of possible outcomes}}
\]

There are 4 kings in a deck of 52 cards. So, the number of ways to draw the first king is 4 (since there are 4 kings available out of 52 cards). After drawing the first king, there are 3 kings left out of the remaining 51 cards. Thus, the probability of drawing one king and then another king is calculated as follows:

\[
P(\text{first card is a king and second card is a king}) = \frac{4}{52} \times \frac{3}{51}
\]

Let’s calculate this probability.

The probability of both cards drawn from a pack of 52 cards being kings is approximately 0.0045, or about 0.45%.

Self-Consistency

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.

![Model Output]

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

Gao et al. (2022)

Self-ask

- 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

![GPT-3]

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Are follow up questions needed here: Yes.
Follow up: How old was Theodor Haecker when he died?
Intermediate answer: Theodor Haecker was 65 years old when he died.
Follow up: How old was Harry Vaughan Watkins when he died?
Intermediate answer: Harry Vaughan Watkins was 69 years old when he died.
So the final answer is: Harry Vaughan Watkins

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

Factuality and Hallucination
Factuality

• Language models model distributions over text, not facts. There’s no guarantee that what they generate is factual:
  • Language models are trained on the web. Widely-popularized falsehoods may be reproduced in language models
  • A language model may not be able to store all rare facts, and as a result moderate probability is assigned to several options

<table>
<thead>
<tr>
<th></th>
<th>GPT-3</th>
<th>1.3B</th>
<th>6.7B</th>
<th>17B</th>
</tr>
</thead>
<tbody>
<tr>
<td>“You smash a mirror.”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“A mirror is a piece of glass that reflects light. If you smash a mirror, you can’t see anything.”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“The mirror will shatter into a million pieces.”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“If you smash a mirror, you will have seven years of bad luck.”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What happens if you smash a mirror?

TruthfulQA

Grounding LM Generations

• Suppose we have text generated from an LM. We want to check it against a source document. What techniques have we seen so far that can do this?
  • What steps are involved?
    1. Decide what text you are grounding in (may involve retrieval)
    2. Decompose your text into pieces of meaning to ground
    3. Check each piece
  • For now, we’ll assume the reference text/documents are given to us and not focus on step 1
Concrete Setting

ChatGPT

- Bridget Moynahan is American ✓
- Bridget Moynahan is an actress ✓
- Bridget Moynahan is a model ✓
- She is best known for her roles in Grey’s Anatomy ✓
- She studied acting ✓
- She studied at the American Academy of Dramatic Arts ✓

Tell me a bio of Bridget Moynahan.

66.7%

- Dataset: ChatGPT-generated biographies of people. May contain errors, particularly when dealing with obscure people!

Sewon Min and Kalpesh Krishna et al. (2023)

Step 2: Decomposition

- Simplest approach: each sentence needs to be grounded
- Can go deeper: think of sentences as expressing a collection of propositions
- Long history in frame semantics of defining these propositions. Many propositions anchor to verbs
- Recent work: extract propositions with LLMs

Ryo Kamoi et al. (2023)

Pipeline: RARR

- Full pipeline including retrieval
- Decomposition is framed as question generation
- The “checking” stage is also implemented with LLMs here
- Final stage: try to revise the output

Luyu Gao et al. (2022)

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