Generative AI for Constructive Communication

Evaluation and New Research Methods
Agenda

Jason Wei

Zoom talk
Q&A after his talk

[attendance note]

Second half of class:

Evaluation Roadmap (10 min)

Evaluation: Bias, Factuality, Inconsistency
Lecture (30 min)

Competition: Red Teaming Models
Red Team a Model (15 minutes)

Logistics notes (5 min)
Projects on Evaluating LLMs
Structure of an MAS.S68 Project

Research Project

- Model Abilities
  - Evaluate model / human subjects on exam with “gold” answers?
- Application Shortcomings
  - Evaluate human test subjects by applying different “treatments”?
- Effects on Humans
  - Evaluate human experience with surveys?
  - Evaluate observations / dataset already in the world?
  - Evaluate expert opinions?

Research Question(s)

Experimental Methods

Evaluation Tools

Automatic Evaluation
  - Evaluate using automatically computed metrics

Human Evaluation
  - Evaluate using human judges

RCT
  - Evaluate statistical differences between treatment groups

Quantitative (Descriptive) Analysis
  - Evaluate a static dataset with quantitative tools

Qualitative Analysis
  - Critically evaluate a dataset / interview holistically

Evaluation Examples

Today!
Example Project Roadmap

- **Research Question:** How well can ChatGPT teach children basic math?
- **Specific Setting:** The model is asked to give a child a set of basic arithmetic problems. For each question, if the child gets the answer wrong, it needs to explain to them *why* their answer is wrong.
- **Setup A: Methods and Evaluation:**
Structure of an MAS.S68 Project

Research Project

Model Abilities
- Evaluate model / human subjects on exam with “gold” answers?
- Evaluate human test subjects by applying different “treatments”?
- Evaluate human experience with surveys?
- Evaluate observations / dataset already in the world?
- Evaluate expert opinions?

Application Shortcomings

Effects on Humans

Human Evaluation
- Automatic Evaluation
- Human Evaluation
  - RCT
  - Quantitative (Descriptive) Analysis
  - Qualitative Analysis

Evaluation Tools
- Evaluate using automatically computed metrics
- Evaluate using human judges
- Evaluate statistical differences between treatment groups
- Evaluate a static dataset with quantitative tools
- Critically evaluate a dataset / interview holistically

Evaluation Examples

Research Question(s)

Experimental Methods

Example Project Roadmap

- **Research Question:** How well can ChatGPT teach children basic math?
- **Specific Setting:** The model is asked to give a child a set of basic arithmetic problems. For each question, if the child gets the answer wrong, it needs to explain to them *why* their answer is wrong.
- **Setup A: Methods and Evaluation:**
  - Prepare:
    1. Prepare a set of arithmetic problems for it to ask a user.
    2. Prepare a set of wrong responses to these questions, simulating children. (Exam questions for the model)
    3. Prepare human-written explanations for each wrong answers (Exam answers for the model)
  - Run the experiment: Have the model provide explanations for why the answers are incorrect.
  - Human Evaluation: Have a human evaluator score each model explanation for accuracy, comparing them against the high-quality, human-authored explanations. Then calculate a final metric, e.g. % accuracy for the model's ability to explain arithmetic questions.
Structure of an MAS.S68 Project

Research Question(s)

Experimental Methods

Evaluation Tools

Evaluation Examples

Model Abilities

Evaluate model / human subjects on exam with “gold” answers?

Evaluate model / human subjects by applying different “treatments”?

Evaluate human experience with surveys?

Evaluate expert opinions?

Application Shortcomings

Evaluate observations / dataset already in the world?

Human Evaluation

Automatic Evaluation

Human Evaluation

RCT

Quantitative (Descriptive) Analysis

Qualitative Analysis

Critically evaluate a dataset / interview holistically

Evaluate using automatically computed metrics

Evaluate using human judges

Evaluate statistical differences between treatment groups

Evaluate a static dataset with quantitative tools

Evaluate human test subjects by applying different “treatments”?
Example Project Roadmap

- **Research Question:** *How well can ChatGPT teach children basic math?*
- **Specific Setting:** The model is asked to give a child a set of basic arithmetic problems. For each question, if the child gets the answer wrong, it needs to explain to them *why* their answer is wrong.
- **Setup B: Methods and Evaluation:**
  - Prepare:
    - (1) Prepare a set of arithmetic problems for it to ask a user.
    - (2) Prepare children to answer arithmetic questions given by the model.
  - Run the experiment (RCT):
    - Split the children into two groups.
    - Have children Group 1 answer the model’s questions, but they are only told if they are right or wrong.
    - Have children Group 2 answer the model’s questions and read the model’s explanations.
    - Score both groups of children on an arithmetic quiz to see if the model helped their learning.
Structure of an MAS.S68 Project

Research Project

Model Abilities

Application Shortcomings

Effects on Humans

Research Question(s)

Experimental Methods

Evaluation Tools

Evaluation Examples

Evaluate model / human subjects on exam with “gold” answers?

Evaluate human test subjects by applying different “treatments”?

Evaluate human experience with surveys?

Evaluate observations / existing behavioral dataset in the world?

Evaluate expert opinions?

Automatic Evaluation

Human Evaluation

RCT

Quantitative (Descriptive) Analysis

Qualitative Analysis

Evaluate using automatically computed metrics

Evaluate using human judges

Evaluate statistical differences between treatment groups

Evaluate a static dataset with quantitative tools

Critically evaluate a dataset / interview holistically

Evaluate human test subjects by applying different “treatments”?
Example Project Roadmap

- **Research Question:** How well can ChatGPT teach children basic math?
- **Specific Setting:** The model is asked to give a child a set of basic arithmetic problems. For each question, if the child gets the answer wrong, it needs to explain to them *why* their answer is wrong.
- **Setup C: Methods and Evaluation:**
  - Prepare:
    - (1) Prepare a set of arithmetic problems for it to ask a user.
    - (2) Prepare children to answer arithmetic questions given by the model.
  - Run the experiment *(Qualitative/Descriptive Analysis)*:
    - Have the children answer the model's questions and read the model's explanations.
    - Document your observations and survey their learning experience.
Details on Evaluating LLMs & their Applications
Structure of an MAS.S68 Project

Research Project

- Model Abilities
  - Evaluate model / human subjects on exam with “gold” answers?
- Application Shortcomings
  - Evaluate human test subjects by applying different “treatments”?
  - Evaluate human experience with surveys?
- Effects on Humans
  - Evaluate observations / dataset already in the world?
  - Evaluate expert opinions?

Evaluation Tools

- Automatic Evaluation
- Human Evaluation
- RCT
- Quantitative (Descriptive) Analysis
- Qualitative Analysis

Evaluation Examples

- Evaluate using automatically computed metrics
- Evaluate using human judges
- Evaluate statistical differences between treatment groups
- Evaluate a static dataset with quantitative tools
- Critically evaluate a dataset / interview holistically

Today!
Lesson Plan

1. What is a Dataset?
2. What is a Metric?
3. How does Automatic Evaluation work?
4. How does Human Evaluation work?
5. Three Examples of Supervised Data Evaluation:
   ○ Evaluating LLMs for Bias
   ○ Evaluating LLMs for Factuality & Hallucination
   ○ Evaluating LLMs for Self-Consistency
What is a Dataset?

General Dataset

- Any set of records
- Surveys, transcripts, documents, videos, network graphs, etc..
- These are useful for descriptive qualitative or quantitative analysis, that *summarize the data themselves.*

“Supervised” Data (for training and evaluation)

- Any set of records, with *(input-output)* pairs.
- Sentences and their sentiment scores, documents and their summaries, videos and their captions, questions and their answers, etc..
- These are useful for evaluating machine learning models.
What is a Metric?

Given “supervised data” how do we evaluate?

1. Run the model on the inputs to get predictions.
2. Define a metric (or “score”) that estimates how well the model predictions reflect the “gold” outputs.
3. Compute the metric!

How to compute a score?

1. Let a human do it! (Human Evaluation)
2. Compute it! (Automatic Evaluation)

<table>
<thead>
<tr>
<th>Task</th>
<th>Metric</th>
<th>Automatic Scoring Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Accuracy</td>
<td>Exact Match: Did the model predict the same output as the prediction?</td>
</tr>
<tr>
<td>Question Answering</td>
<td>F1 Score</td>
<td>How many words are in common between the prediction and output?</td>
</tr>
<tr>
<td>Translation</td>
<td>ROUGE/BLEU</td>
<td>How many words/phrases are in common between the prediction and output?</td>
</tr>
<tr>
<td>Program Synthesis</td>
<td>Accuracy</td>
<td>Does the predicted code produce the same result as the output when run?</td>
</tr>
</tbody>
</table>

…

…

…
Human Evaluation

- A human (e.g. crowd turker) compares the model answer to the real answer.
- Typically asked to assess:
  - Coherence, readability, fluency
  - Grammaticality
  - Extent to which the model follows instructions
Human Evaluation

- **Preference judgements:**
  - Example: Choose the passage that is more [insert quality]
  - Could have a third option specifying that both passages are equally good.

- **Rating a passage (e.g., Likert scale):**
  - Example: Thinking about [insert assessed quality], rate the following passage on a scale of 1 to 5 with 1 being the worst and 5 being the best.
  - Example: The generated story follows the instructions (e.g., includes all characters). How much do you agree with this statement?

<table>
<thead>
<tr>
<th>Strongly disagree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly agree</th>
</tr>
</thead>
</table>

-
Evaluating Bias / Fairness In LLMs

(A Very Cursory Introduction)
Evaluating Bias/Fairness

**WARNING:**
The following slides contain examples of model bias and evaluation which are offensive in nature.
Evaluating Bias/Fairness

Definitions of Bias / Fairness

- Where models demonstrate unfair, discriminatory, or hateful behaviour
- This can be particularly harmful if targeted towards sensitive personal attributes, such as gender, sexuality, race and religion.
- Harms can arise even from “correct” or intended uses, depending on where and how they are deployed, and in predictive applications as well as generative ones.
Evaluating Bias/Fairness

A Generative Language Model:
- Emulates text scraped from across the web
- Is often optimized for subsets of users (western, affluent, etc)
Evaluating Bias/Fairness

How has prior work evaluated bias?

- **Intrinsic Bias** → Evaluating the inner state of the model itself
  - E.g. African-American names are more closely associated with unpleasant words in the model embedding space.\(^1\)

- **Extrinsic Bias** → Evaluating the behaviour of the model from (input, output) pairs
  - E.g. Given leading prompts, how often will the model generate a toxic, biased response?

- **Application Bias** → Evaluating the full system in the setting where it is deployed
  - E.g. Translation technologies systematically generating incorrect and stereotyped genders.

---

\(^1\) Caliskan et al. (2017) “Semantics derived automatically from language corpora contain human-like biases”

\(^2\) Gehman et al. (2020) “RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models”
Evaluating Bias/Fairness

Real Toxicity Prompts

1. Prompt the model \( k = 25 \) times.
2. Get \( k = 25 \) generations.
3. Use Perspective API to score each generation for toxicity.
4. Compute the Maximum Toxicity over \( k = 25 \) generations, and report the average rate over all examples.

Gehman et al. (2020) “RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models”
Evaluating Bias/Fairness

What is Bias in NLP Models?
- Where models demonstrate unfair, stereotyped, or discriminatory behaviour towards sensitive demographics, such as gender, sexuality, race and religion.

How has prior work evaluated bias?
- **Intrinsic Bias** → Evaluating the inner state of the model itself
  - E.g. African-American names are more closely associated with unpleasant words in the model embedding space.¹
- **Extrinsic Bias** → Evaluating the behaviour of the model from (input, output) pairs
  - E.g. Given leading prompts, how often will the model generate a toxic, biased response?
- **Application Bias** → Evaluating the full system in the setting where it is deployed
  - E.g. Translation technologies systematically generating incorrect and stereotyped genders.

¹ Caliskan et al. (2017) “Semantics derived automatically from language corpora contain human-like biases”
² Gehman et al. (2020) “RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models”
Evaluating Bias/Fairness

Dora Vargha
@DoraVargha

Hungarian is a gender neutral language, it has no gendered pronouns, so Google Translate automatically chooses the gender for you. Here is how everyday sexism is consistently encoded in 2021. ☹️ you, Google.

Source:
https://twitter.com/DoraVargha/status/1373211762108076034
Evaluating Bias/Fairness

**Biased Generation Benchmarks:**
- Stereoset
- UNQover
- Bias Benchmark for QA (BBQ)
- Winogender
- Winobias

**Toxicity Identification Benchmarks:**
- Civil Comments
- Toxigen
- DynaHate

**General Resources for NLP Bias + Fairness:**
- [https://github.com/uclanlp/awesome-fairness-papers](https://github.com/uclanlp/awesome-fairness-papers) Chance et al. (2023)
  - Recent list of related publications in ML Fairness
  - Problems in common benchmarks
Evaluating Factuality & Hallucinations
Evaluating Factuality & Hallucination

Prompt:

Write a brief bio of Maria Cantwell, US senator from Washington.

Completion: (GPT-3 text-curie-001):

Maria Cantwell was born in Seattle, Washington, in 1951. She is the daughter of a Boeing engineer and a homemaker. Cantwell attended public schools and then went on to study at the University of Washington. She worked as a journalist before being elected to the U.S. Senate in 1996. Cantwell is a member of the Democratic Party and has served on the Senate Committee on Energy and Natural Resources and the Senate Committee on Environment and Public Works."
Factuality: Three metrics we’ll consider

How often does an LLM get the right answer when prompted well?

- Fact completion accuracy

How often does it spit out facts when it generates text?

- Hallucinated named entity error rate
- Entailment ratio

Note that these do not measure reasoning skill or question-answering ability in general.
Factuality: Fact completion

1. “Maria Cantwell was born in <>”
   - Maria Cantwell
   - born-in
   - Indianapolis
   - Incorrect

2. “Maria Cantwell is a member of <>”
   - Maria Cantwell
   - member-of
   - Democratic Party
   - Correct

3. “Maria Cantwell got a degree from <>”
   - Maria Cantwell
   - degree-from
   - Miami University
   - Incorrect

<table>
<thead>
<tr>
<th>Completion</th>
<th>Reference</th>
<th>Correct?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seattle</td>
<td>Indianapolis</td>
<td>✗</td>
</tr>
<tr>
<td>Democratic party</td>
<td>Democratic party</td>
<td>✓</td>
</tr>
<tr>
<td>University of Washington</td>
<td>Miami University</td>
<td>✗</td>
</tr>
</tbody>
</table>

Accuracy@1 = 33%
Factuality: Open-ended generation

Phase 1: Generation of LM continuation

[PROMPT] Fox 2000 Pictures released the film Soul Food.

[LM_GEN] Its leading cast includes the British actress Emma Thompson, Tom Wilkinson, George Segal, and Lettice Ullman. The film was largely praised for its acting performances, and nominated for eight Academy Awards, including Best Picture, Best Actress, Best Supporting Actor, and Best Director.
Maria Cantwell was born in Seattle, Washington, in 1951. She is the daughter of a Boeing engineer and a homemaker. Cantwell attended public schools and then went on to study at the University of Washington. She worked as a journalist before being elected to the U.S. Senate in 1996. Cantwell is a member of the Democratic Party and has served on the Senate Committee on Energy and Natural Resources and the Senate Committee on Environment and Public Works.

Intuition: How many highlighted phrases are not in the Wikipedia article?

\[ \text{NE}_{ER} = \frac{|\text{HALLU}_{\text{NE}}|}{|\text{AL}_{\text{LNE}}|} \]
Factuality: Open-ended generation

Maria Cantwell was born in Seattle, Washington, in 1951. She is the daughter of a Boeing engineer and a homemaker. Cantwell attended public schools and then went on to study at the University of Washington. She worked as a journalist before being elected to the U.S. Senate in 1996. Cantwell is a member of the Democratic Party and has served on the Senate Committee on Energy and Natural Resources and the Senate Committee on Environment and Public Works.

Intuition: How many highlighted phrases are not in the Wikipedia article?

\[ \text{NE}_{E_R} = \frac{|HALLU_{NE}|}{|A_{LNE}|} \]

= 3/8 = 37.5%

Factuality Enhanced Language Models for Open-Ended Text Generation (Lee et al)
Factuality: Open-ended generation

Entailment-based metrics

Maria Cantwell was born in Seattle, Washington, in 1951.

Cantwell was born in Indianapolis, Indiana. She was raised in a predominantly Irish American neighborhood on the south side of Indianapolis. Her father, Paul

Entailed by

Refuted by

Neutral
Factuality: Human evaluation

Maria Cantwell was born in Seattle, Washington, in 1951. She is the daughter of a Boeing engineer and a homemaker. Cantwell attended public schools and then went on to study at the University of Washington. She worked as a journalist before being elected to the U.S. Senate in 1996. Cantwell is a member of the Democratic Party and has served on the Senate Committee on Energy and Natural Resources and the Senate Committee on Environment and Public Works."

<table>
<thead>
<tr>
<th>Annotation</th>
<th>$\text{Entail}_R$</th>
<th>$\text{NE}_{ER}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert</td>
<td>0.81</td>
<td>-0.77</td>
</tr>
<tr>
<td>Majority-voting</td>
<td>0.47</td>
<td>-0.46</td>
</tr>
</tbody>
</table>

Larger models, better prompts elicit higher factuality

<table>
<thead>
<tr>
<th>Size</th>
<th>Decode</th>
<th>Factual Prompt</th>
<th>Nonfactual Prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( \text{NE}<em>{E</em>{\text{R}}} \downarrow )</td>
<td>( \text{Entail}_{R} \uparrow )</td>
</tr>
<tr>
<td>126M</td>
<td>p=0.9 greedy</td>
<td>63.69%</td>
<td>0.94%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48.55%</td>
<td>8.36%</td>
</tr>
<tr>
<td>357M</td>
<td>p=0.9 greedy</td>
<td>56.70%</td>
<td>2.01%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>43.04%</td>
<td>14.25%</td>
</tr>
<tr>
<td>1.3B</td>
<td>p=0.9 greedy</td>
<td>52.42%</td>
<td>2.93%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>39.87%</td>
<td>12.91%</td>
</tr>
<tr>
<td>8.3B</td>
<td>p=0.9 greedy</td>
<td>40.59%</td>
<td>7.07%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>28.06%</td>
<td>22.80%</td>
</tr>
<tr>
<td>530B</td>
<td>p=0.9 greedy</td>
<td>33.30%</td>
<td>11.80%</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>20.85%</strong></td>
<td><strong>31.94%</strong></td>
</tr>
</tbody>
</table>

Evaluating Robustness & Self-Consistency
Evaluating Robustness and Self-consistency

- **Robustness** – whether models are sensitive and vulnerable to a small perturbation of inputs and generalize well across different datasets

<table>
<thead>
<tr>
<th>Original Text Prediction: <strong>Entailment</strong> (Confidence = 86%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Premise:</strong> A runner wearing purple strives for the finish line.</td>
</tr>
<tr>
<td><strong>Hypothesis:</strong> A runner wants to head for the finish line.</td>
</tr>
<tr>
<td>Adversarial Text Prediction: <strong>Contradiction</strong> (Confidence = 43%)</td>
</tr>
<tr>
<td><strong>Premise:</strong> A runner wearing purple strives for the finish line.</td>
</tr>
<tr>
<td><strong>Hypothesis:</strong> A racer wants to head for the finish line.</td>
</tr>
</tbody>
</table>

Robustness and Adversarial Examples in NLP (Chang, Kai-Wei, et al.) EMNLP Tutorial 2021

- **Self-consistency** – whether model predictions across inputs imply logically compatible beliefs about the world

  *Is a sparrow a bird?* → *Yes*
  *Does a bird have feet?* → *Yes*
  *Does a sparrow have feet?* → *No*

Benchmarks vs. Reality

SQuAD2.0 (Rajpurkar et al. ‘18)

Packet switching contrasts with another principal networking paradigm, circuit switching, a method which pre-allocates dedicated network bandwidth specifically for each communication session, each having a constant bit rate and latency between nodes. In cases of billable services, such as cellular communication services, circuit switching is characterized by a fee per unit of connection time, even when no data is transferred, while packet switching may be characterized by a fee per unit of information transmitted, such as characters, packets, or messages.

Q: Packet Switching contrast with what other principal
A: circuit switching

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Human Performance</td>
<td>86.831</td>
<td>89.452</td>
</tr>
<tr>
<td></td>
<td>Stanford University (Rajpurkar &amp; Jia et al. ‘18)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Retro-Reader on ALBERT (ensemble)</td>
<td>90.115</td>
<td>92.580</td>
</tr>
<tr>
<td></td>
<td>Shanghai Jiao Tong University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ALBERT + DAAF + Verifier (ensemble)</td>
<td>90.002</td>
<td>92.425</td>
</tr>
<tr>
<td></td>
<td>PINGAN Omni-Sinitic</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What we need:
Robust & Reliable NLP

Robustness and Adversarial Examples in NLP (Chang, Kai-Wei, et al.) EMNLP Tutorial 2021
Adversarial Trigger for Text Classification

Inputs

Vaccine is ineffective...
Madonna found dead...
USA wins world cup...

Prediction

Fake
Fake
Fake
Adversarial Trigger for Text Classification

Trigger: blutarsky bottle tennis

- Inputs:
  - Vaccine is ineffective...
  - Madonna found dead...
  - USA wins world cup...

- Prediction:
  - Fake $\leftrightarrow$ Real
  - Fake $\leftrightarrow$ Real
  - Fake $\leftrightarrow$ Real
Why Robust Models?

- Make models use the right features instead of *spurious correlation* for predictions
- Make models do well on *out-of-distribution (OOD) domains* and *tasks*
  - Linguistic styles, dialects, grammatical mistakes, syntactic structures
  - News articles vs. conversations vs. social media
  - Domain knowledge (e.g., medical terms)

*Robustness and Adversarial Examples in NLP (Chang, Kai-Wei, et al.) EMNLP Tutorial 2021*
How to evaluate performance on tasks vs. datasets?

- Traditionally, train and test data have similar distribution
  - For instance, both training and test are from IMDB movie reviews for sentiment analysis
- Include hard examples in the test data
  - Held-out test set is not enough
  - Simple adversarial attacks are not good proxies of real-world generalization
  - Include a wide range of test examples to measure task (not dataset) performance
Evaluating Robustness in LLMs

- **Prompt design**
  - E.g., tldr vs. summarize

- **One/Few-shot Learning**
  - Which examples to use
  - The order of examples
  - The dominant label in training dominates the predictions

---

**Calibrate Before Use: Improving Few-shot Performance of Language Models. Zhao et. el., ICML 2021.**
# Robustness on Zero-shot CoT

Table 4: Robustness study against template measured on the MultiArith dataset with text-davinci-002. (*1) This template is used in Ahn et al. [2022] where a language model is prompted to generate step-by-step actions given a high-level instruction for controlling robotic actions. (*2) This template is used in Reynolds and McDonell [2021] but is not quantitatively evaluated.

<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>
</tbody>
</table>

Large Language Models are Zero-Shot Reasoners. Kojima et.al., NeurIPS 2022.
Group Activity: Red Teaming
LLMs
Red Teaming Activity

Instructions:
- Partner up with someone you don’t know
- In your group, go to ChatGPT Playground or the OpenAI GPT-3 playground

Pick one of the following themes:

<table>
<thead>
<tr>
<th>Bias</th>
<th>Factuality</th>
<th>Inconsistency</th>
<th>Something else?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can you find (e.g.):</td>
<td>Can you trigger (e.g.):</td>
<td>Can you find (e.g.):</td>
<td>Can you find:</td>
</tr>
<tr>
<td>Political Bias</td>
<td>Political lies?</td>
<td>Contradictions?</td>
<td>Other concerning issues?</td>
</tr>
<tr>
<td>Cultural Bias</td>
<td>Conspiracy theories?</td>
<td>Unfounded over-confidence</td>
<td></td>
</tr>
<tr>
<td>Gender Bias</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Prompt the model to find examples of these issues.

Document the worst examples of these issues— they will become part of your homework answers! We will share out if time.
Logistics

Announcements:

- Project next steps (Jad)

Homework for next week:

- **DUE MONDAY!**
- Questions for Mina Lee for next Wednesday
- Exercise on paragraph rewriting
- Report back your red teaming results from today

Other notes:

- Attendance QR code reminder
- Required: sign up to go over your project in office hours
  - Come talk to us about your projects **early**! Some projects require more pre-work than others :)