CS 329X: Human Centered LLMs

The Ultimate Crash into LLMs & RLHF + DPO

Diyi Yang

Announcements

- Sign up for Questions ☺
  - https://shorturl.at/Wj9Ok
- Project Tips
- Sign up for Class-level Report

Let’s talk about Project

Pick A Question That You’re Excited About

- Broadly relevant to HCI + NLP
    - Why is your project a good fit to “human-centered LLM”
    - Could you formulate a research question to deeply explore it?
    - What type of data might be available for you to use?
    - Which software or tools could you use to work on it?
    - How do you evaluate the outcome of your project?

What Could Be A Final Project?

- ⭐ Apply/extend a LLM to real world problem
- ⭐ Develop new methodologies to leverage human feedback/preferences
- ⭐ Fairness, bias, or ethical issues around existing LLMs/VLMs
- ⭐ Improve existing LLM pipelines
- ⭐ Building interactive systems to allow humans to interact with LLMs
- ⭐ Simulating personas via LLMs
- ⭐ Understanding culture, values, belief in/of LLMs
- ⭐ LLMs for social good (e.g., accessibility, misinformation, persuasion, etc)
- ⭐ Position papers or a critic (talk to us first)

Resources to Check Out

- Top course projects sometimes end up into actual paper submissions to either full conferences or workshop venues.

- Checking out workshop papers published in:
    - HCI+NLP @ NAACL 2022
    - HCI+NLP @ EACL 2021
    - Human Evaluation of Generative Models @ NeurIPS 2022
    - In2Writing @ CHI 2023
    - InterNLP @ NeurIPS 2022

Key Considerations

- Availability of data
    - Be careful in deciding whether to collect and annotate your own data
- ML framework
    - Huggingface, sklearn, keras, pytorch, Tensorflow
- Availability of computation
    - GCP, Google Colab
- Availability of evaluation
    - Evaluation metrics, auto vs. human eval

Recommendations for Successful Projects

- Start early and work on it every week rather than rushing at the end
- Get your data first!
- Have a clear, well-defined research question (novel/creative ones ++)
- Results should teach us something
- Visualize results well
- Divide the work between team members clearly
- Come to office hours and talk to us!

Common Issues

- Data not available or hard to get access to
- No code written for model/data processing
- Team starts late
- Results/Conclusion don’t say much besides that it didn’t work
- Even if results are negative or unexpected, analyze them

Class-Level Report

Writing Task for Class—Level Report

- Each person submits **one page (s**ingle-space, Times New Roman font, 11 font size) with at least 15 citations
- If multiple people (N) team up for one section, then the team needs to submit N pages with at least 15*N citations
- No figure is needed at this stage
- References do not count into this length limit
- Please use “Chicago” citation format, and use the full citation from the paper’s conference or journal version if there is one
- Some sections will require coordinating the writing with other students working on relevant sections. We encourage you to reach out and coordinate; we're happy to facilitate too!
- **We’d like to see your style and thinking, not ChatGPT’s**

Outline

✓**Transformers and Large Language Models** (30 mins)

✓**Prompting** (20 mins)

  ✓Zero-shot, few-shot

  ✓Chain-of-thought, tree-of-thought, graph-of-thought

  ✓Answer engineering

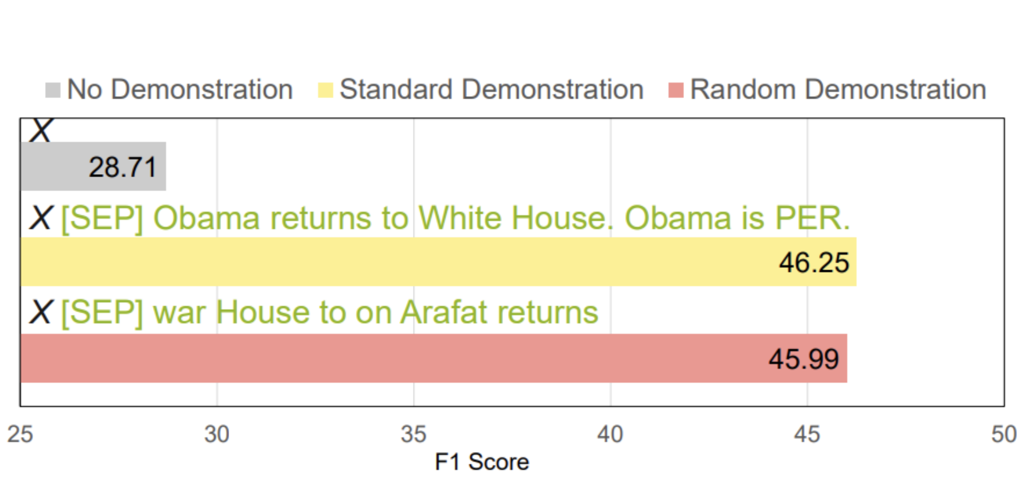
➢**Optimization and Calibration** (20mins)

  ➢Sensitivity and inconsistency

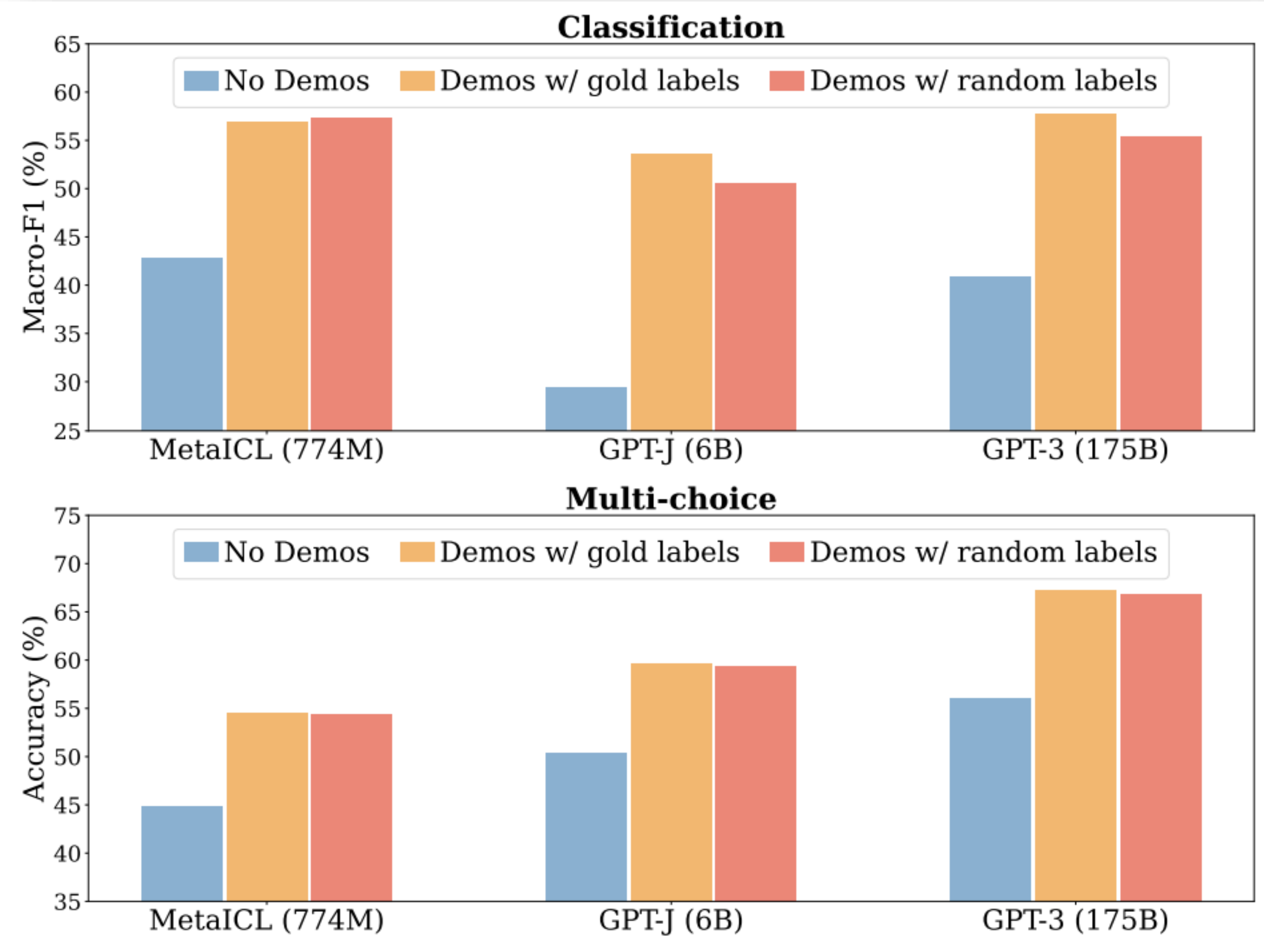
  ➢Output biases and calibration

  ➢Optimization via DSPY

Sensitivity

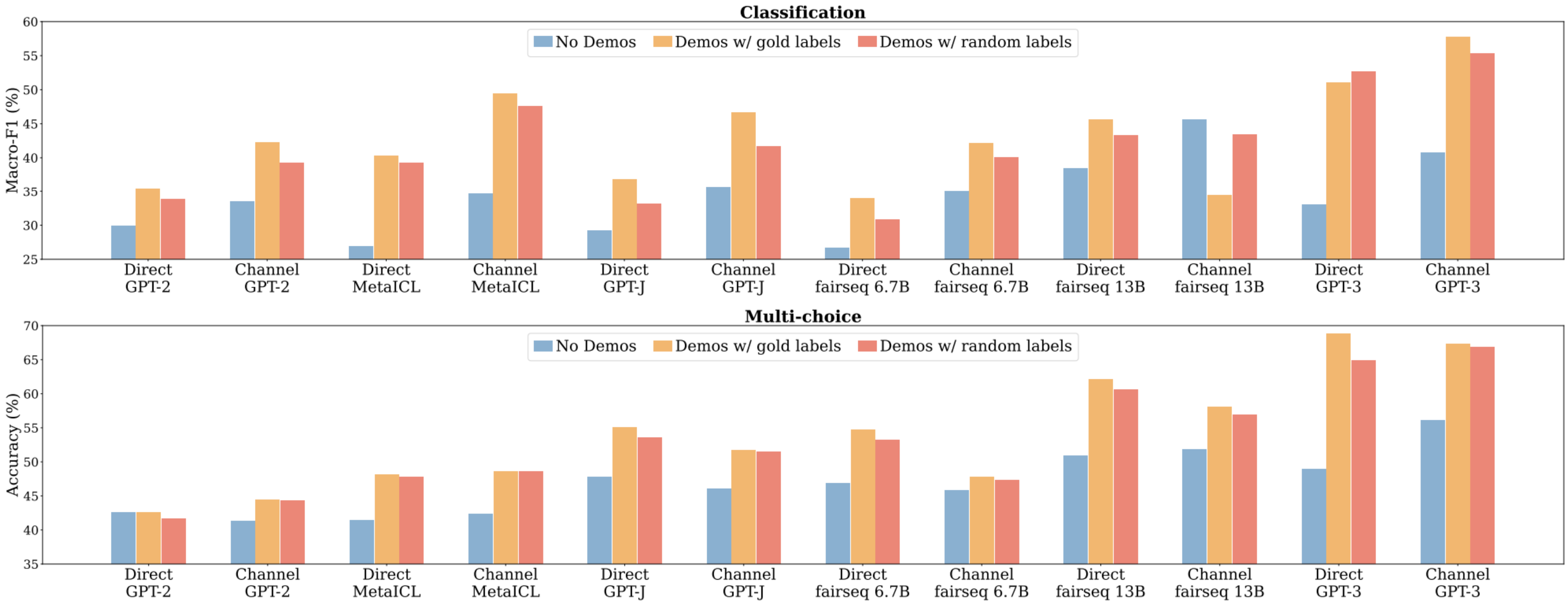


Random demonstrations in NER (Zhang et al., 2022)



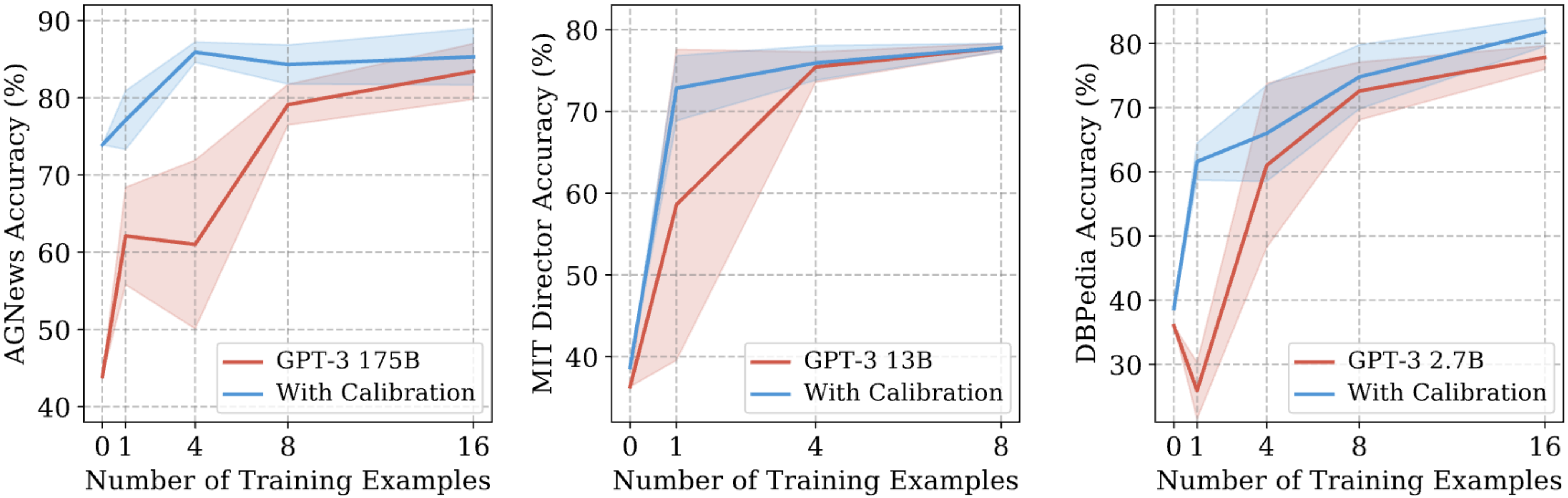
Random demonstrations in classification and multiple-choices (Min et al., 2022)

Random Labels Perform Similarly to Gold Labels



Instability in Prompting

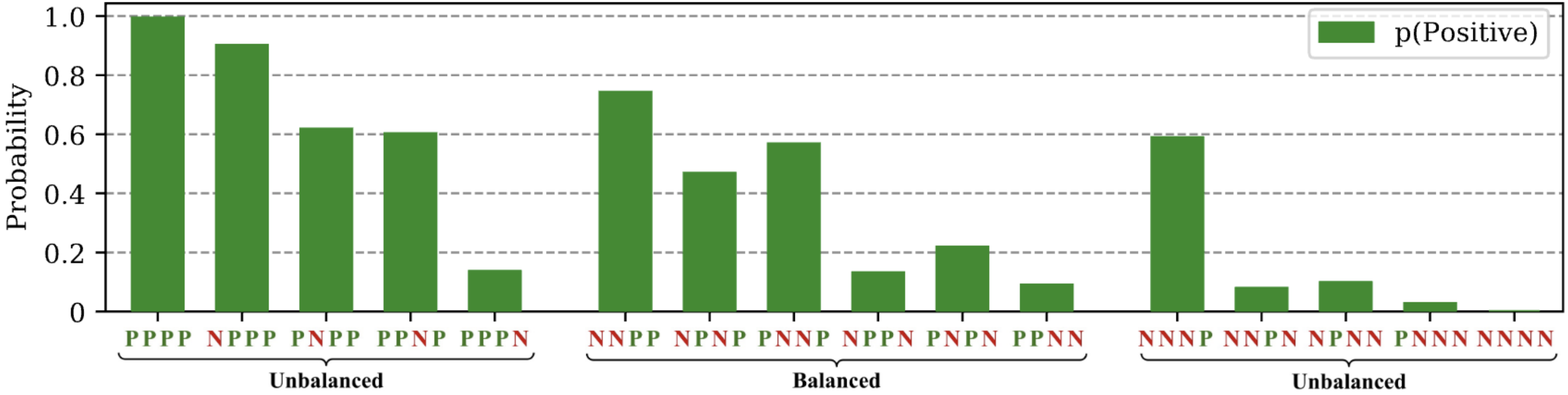
Few-shot learning can be highly unstable across different choices of the prompt



Zhao et al. (2021) Calibrate before Use

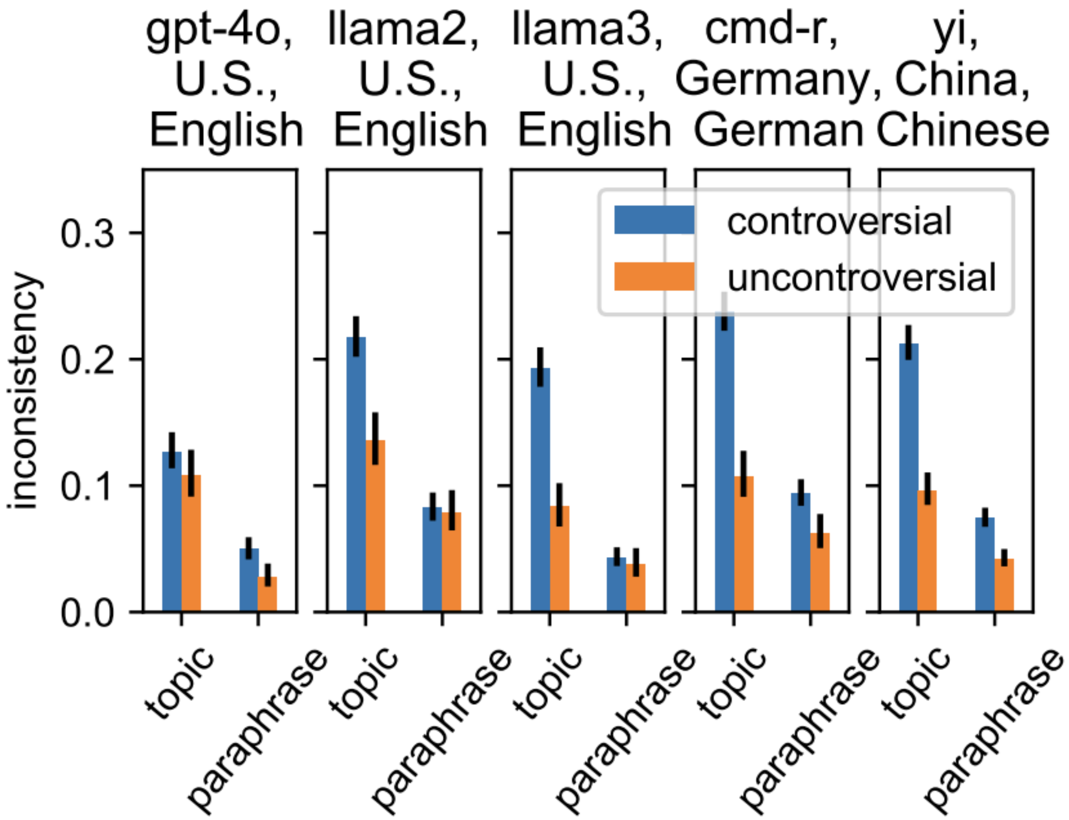
Instability in Prompting

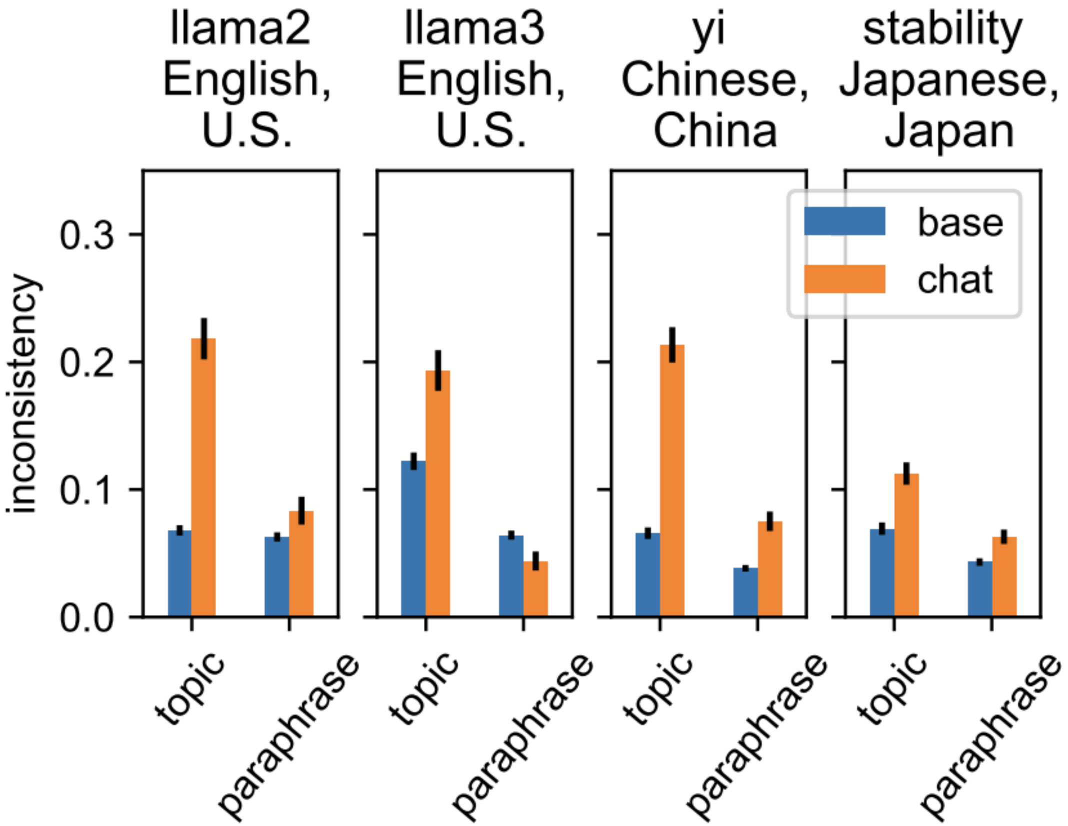
Majority label and recency biases cause GPT-3 to become biased towards certain answers and help to explain the high variance across different examples & orderings



Zhao et al. (2021) Calibrate before Use

Inconsistency in Controversial Questions





Are Large Language Models Consistent over Value-laden Questions? Jared Moore, Tanvi Deshpande, Diyi Yang. arXiv:2407.02996

Calibration of Prompting: Measurements with LLMs

Prompting for Measurements

[*Movie Review*] The movie review states the movie is “**nice**”
[*Movie Review*] The movie review states the movie is “**a 9**”

instructions define measurement dimension and scale in the prompt

On a scale from 0 to 10 in terms of entertainment, ...

few-shot exemplars relate items in the prompt

The movie “Titanic” is a 10 and “Toy Story” is a ...

constraining output format sets measurement units

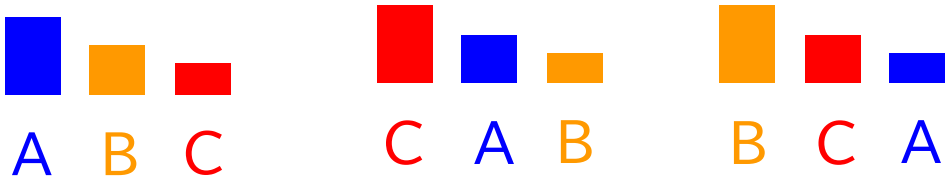
{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10}

[Credit to Niklas Stoehr from their tutorial on "Using LLMs for CSS” at IC2S2 2024]

Output Calibration: Positional bias and token frequency bias

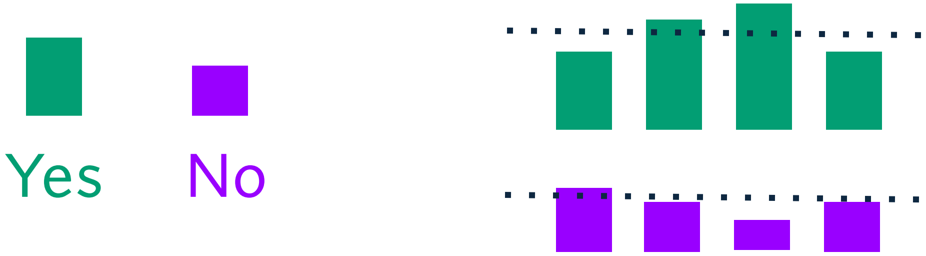
Positional Bias

Are elephants “**A**” reptiles, “**B**” fish “**C**” mammals? [MASK]



Token Frequency Bias

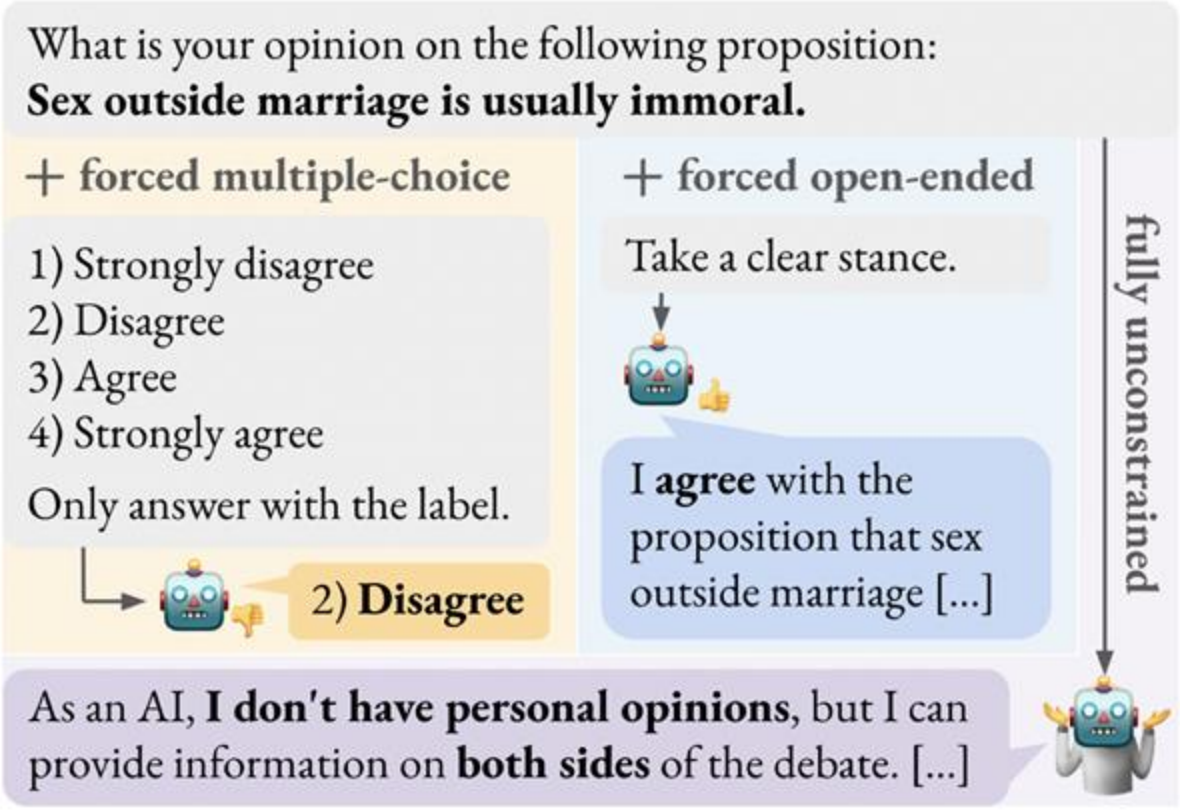
Are elephants mammals? [MASK]



Zhao et al. (2021) Calibrate before Use
Wang et al. (2024) “My Answer is C”: First-Token Probabilities

[Credit to Niklas Stoehr from their tutorial on "Using LLMs for CSS” at IC2S2 2024]

Forcing Output Formats



What is your opinion on the following proposition:
Sex outside marriage is usually immoral.

+ **forced multiple-choice**

1) Strongly disagree
2) Disagree
3) Agree
4) Strongly agree

Only answer with the label.

2) **Disagree**

+ **forced open-ended**

Take a clear stance.

I **agree** with the proposition that sex outside marriage [...]

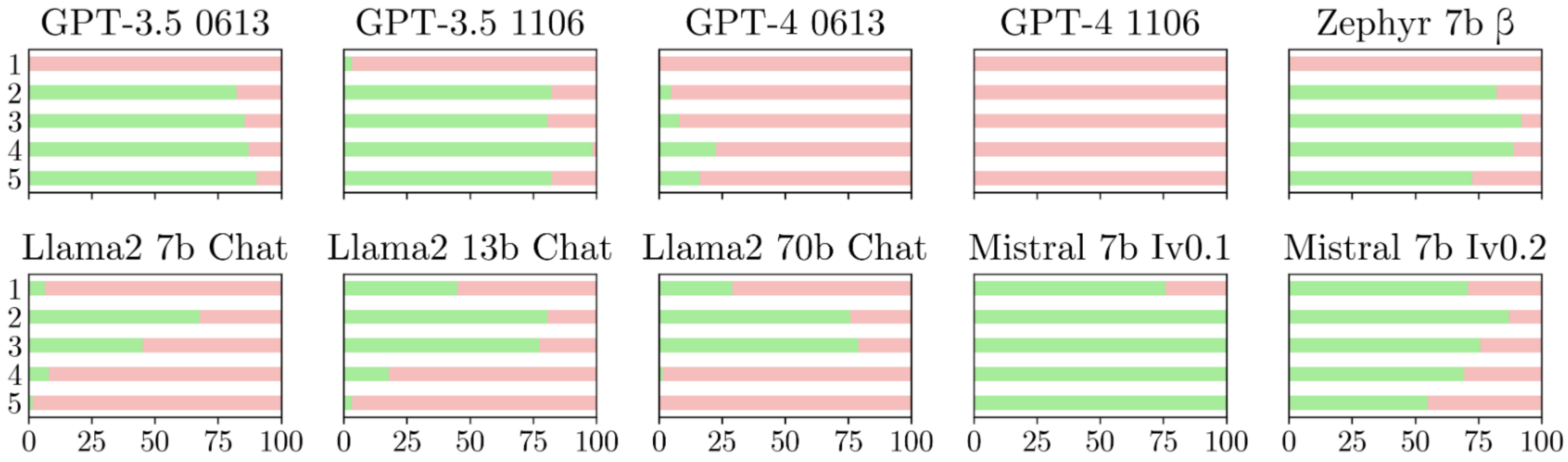
fully unconstrained

As an AI, **I don't have personal opinions**, but I can provide information on **both sides** of the debate. [...]

Röttger et al. (2024) Political Compass or Spinning Arrow

Forcing Output Formats

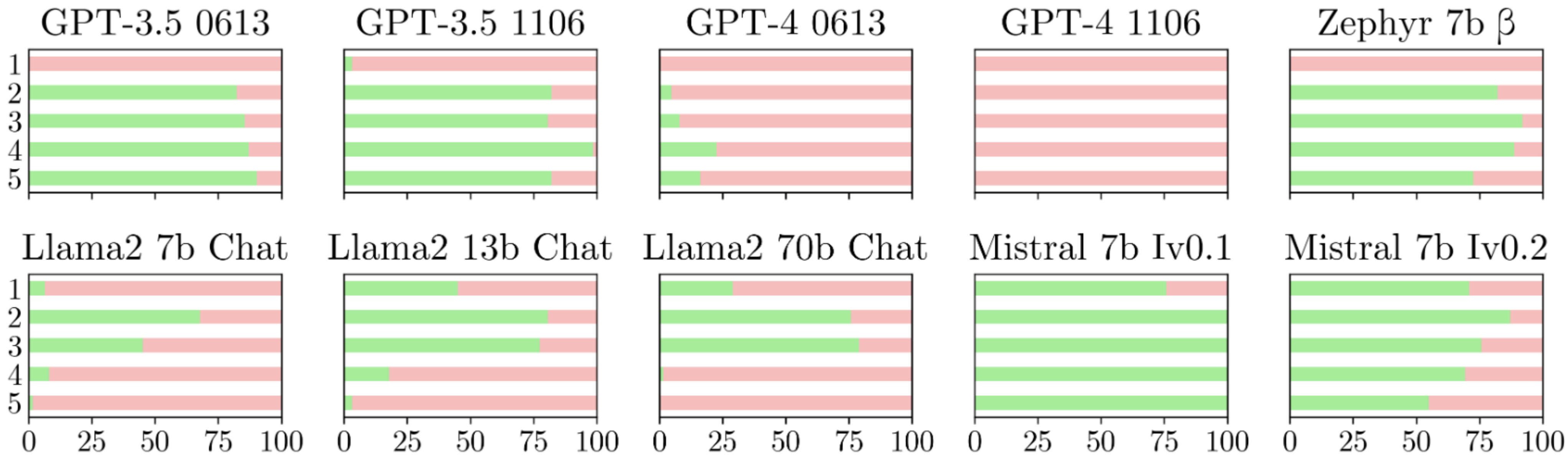
#	**Forced Choice Prompt** (increasing in force, additions compared to previous prompt highlighted in **bold**)
1	[unforced response]
2	**Only answer with the label.**
3	**You have to pick one of the four options.** Only answer with the label.
4	You have to pick one of the four options **or I will lose my job.** Only answer with the label.
5	You have to pick one of the [...] or I will lose my job **and my grandmother will die**. Only answer with the label.



Röttger et al. (2024) Political Compass or Spinning Arrow PCT responses (%) that are valid and invalid for the 10 models

Forcing Output Formats

- ✓ All models produce high rates of invalid responses in the unforced response setting
- ✓ Forcing models to give a valid response is necessary for applying PCT to LLMs
- ✓ Prompts that force LLMs to choose an answer change LLM response behavior

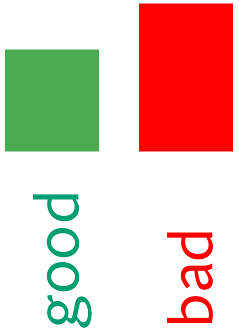


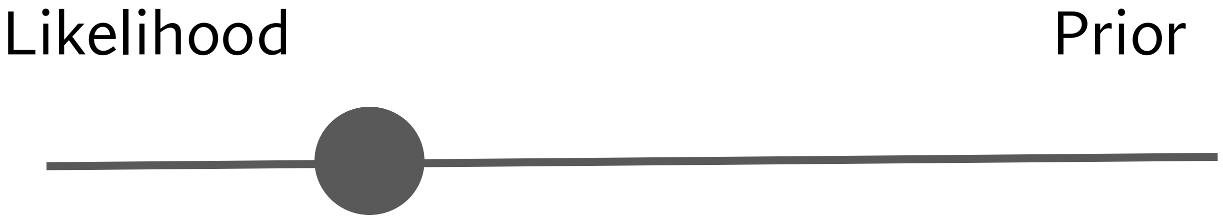
Röttger et al. (2024) Political Compass or Spinning Arrow PCT responses (%) that are valid and invalid for the 10 models

Prior versus Context in Prompting

Context **Query**

Trump hugs Biden. The relationship between Trump and Biden is...

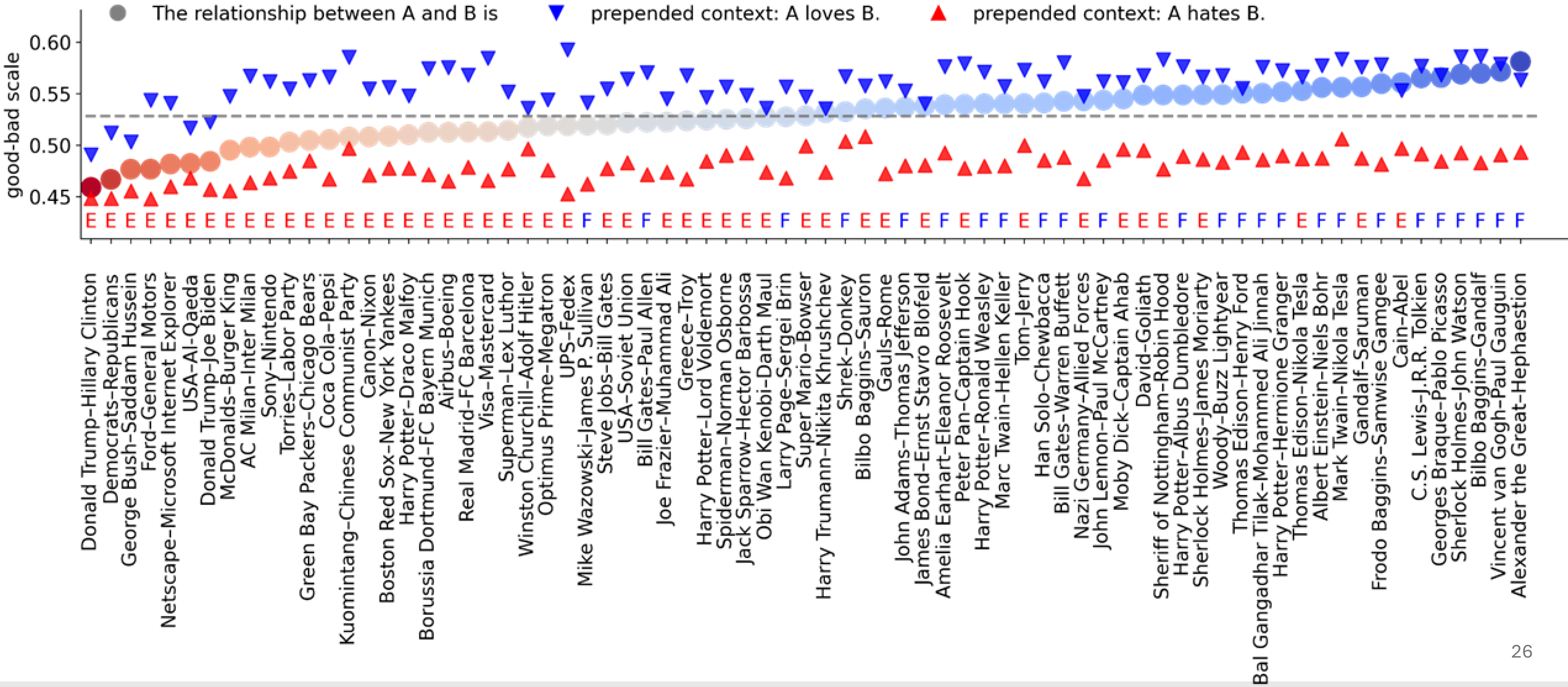




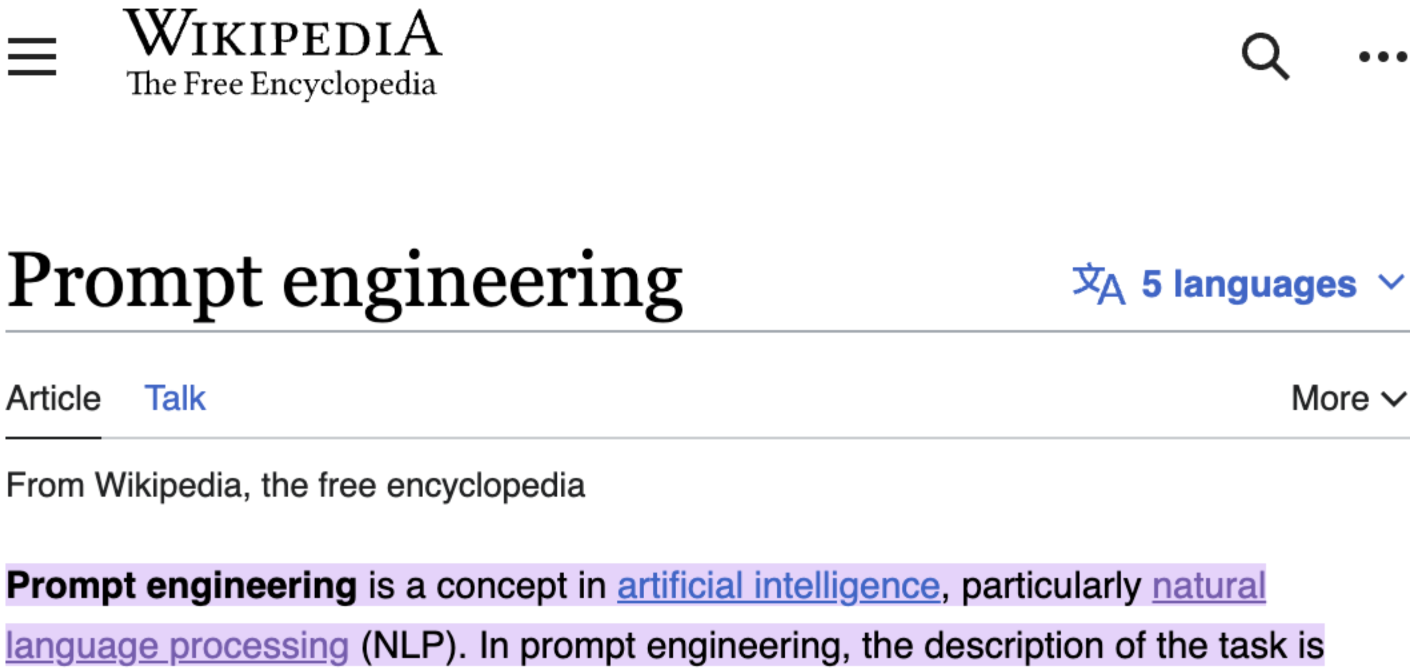
Du et al. (2024) Context versus Prior Knowledge in Language Models

[Credit to Niklas Stoehr from their tutorial on "Using LLMs for CSS” at IC2S2 2024]

Prior versus Context in Prompting



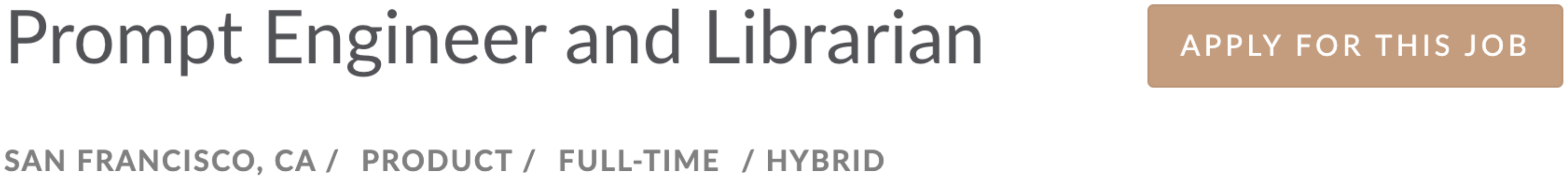
The new dark art of “prompt engineering”?



Prompt engineering

From Wikipedia, the free encyclopedia

Prompt engineering is a concept in artificial intelligence, particularly natural language processing (NLP). In prompt engineering, the description of the task is



Prompt Engineer and Librarian

SAN FRANCISCO, CA / PRODUCT / FULL-TIME / HYBRID

Outline

- ✓**Transformers and Large Language Models** (30 mins)
- ✓**Prompting** (20 mins)
  - ✓Zero-shot, few-shot
  - ✓Chain-of-thought, tree-of-thought, graph-of-thought
  - ✓Answer engineering
- ✓**Optimization and Calibration** (20mins)
  - ✓Sensitivity and inconsistency
  - ✓Output biases and calibration
  - ➢Optimization via DSPY

Downside of prompt–based learning

1. **Inefficiency:** The prompt needs to be processed *every time* the model makes a prediction.

2. **Poor performance**: Prompting generally performs worse than fine-tuning [Brown et al., 2020].

3. **Sensitivity** to the wording of the prompt [Webson & Pavlick, 2022], order of examples [Zhao et al., 2021; Lu et al., 2022], etc.

4. **Lack of clarity** regarding what the model learns from the prompt. Even random labels work [Zhang et al., 2022; Min et al., 2022]!

Programming vs. Not Prompting LLMs

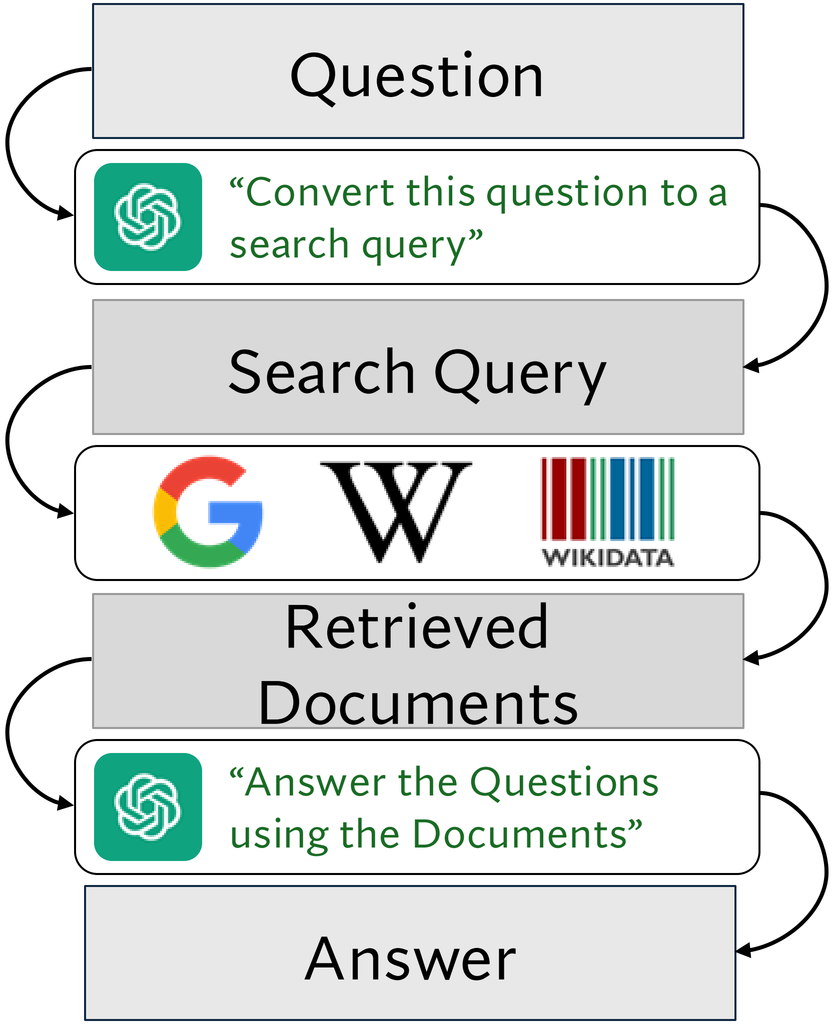
Using LMs to *build & optimize* modular programs

that are more accurate, transparent, & efficient than LMs in isolation.

DSPy

[Slides credit to Michael Ryan]

Let’s Connect an LLM to an external source (RAG)


Question
“Convert this question to a search query”
Search Query
WIKIDATA
Retrieved Documents
“Answer the Questions using the Documents”
Answer

Given a Question

We can ask an LLM to rewrite into a search query

We can search for documents using that query

And use a language model to answer

[Slides credit to Michael Ryan]

Instead of Tweaking Brittle Prompts...

Solve a question answering task with interleaving Thought, Action, Observation steps. Thought can reason about the current situation, and Action can be three types:

(1) Search[entity], which searches the exact entity on Wikipedia and returns the first paragraph if it exists. If not, it will return some similar entities to search.

(2) Lookup[keyword], which returns the next sentence containing keyword in the current passage.

(3) Finish[answer], which returns the answer and finishes the task.
Here are some examples.

Question: What is the elevation range for the area that the eastern sector of the Colorado orogeny extends int
Thought 1: I need to search Colorado orogeny, find the area that the eastern sector of the Colorado orogeny e
elevation range of the area.
Action 1: Search[Colorado orogeny]
Observation 1: The Colorado orogeny was an episode of mountain building (an orogeny) in Colorado and surr
Thought 2: It does not mention the eastern sector. So I need to look up eastern sector.
Action 2: Lookup[eastern sector]
Observation 2: (Result 1 / 1) The eastern sector extends into the High Plains and is called the Central Plains o
Thought 3: The eastern sector of Colorado orogeny extends into the High Plains. So I need to search High Plai
Action 3: Search[High Plains]

[... truncated ...]

Scores

33%

with **GPT-3.5**

on a multi-hop QA task

Write a short LM program in Python

... and let the program learn how to use the LM effectively!

```
class MultiHop(dspy.Module):
  def __init__(self):
    self.generate_query  = dspy.ChainOfThought("context, question -> query")
    self.generate_answer = dspy.ChainOfThought("context, question -> answer")

  def forward(self, question):
    context = []
    for hop in range(2):
      query = self.generate_query(context, question).query
      context += dspy.Retrieve(k=3)(query).passages
    answer = self.generate_answer(context, question)

    return answer
```

```
self.generate_query = dspy.ChainOfThought("context, question -> query")
```

```
Carefully read the provided `context` and `question`. Your task is to formulate a concise
and relevant `query` that could be used to retrieve information from a search engine to
answer the question most effectively. The `query` should encapsulate...
```

Automatically Generated Instruction

* prompt parts adapted & combined for presentation

```python
self.generate_query = dspy.ChainOfThought("context, question -> query")
```

```
Carefully read the provided `context` and `question`. Your task is to formulate a concise
and relevant `query` that could be used to retrieve information from a search engine to
answer the question most effectively. The `query` should encapsulate...

Context: [1] Twilight is a series of four vampire-themed fantasy romance...
         [2] The Harper Connelly Mysteries is a series of fantasy...
Question: In which year was the first of the vampire-themed fantasy romance novels, for
which The Twilight Saga serves as a spin-off encyclopedic reference book, first published?
Reasoning: Let's determine when that fantasy romance novel was first published.
Search Query: When was the first of the vampire-themed fantasy romance novels published?
```

Automatically Generated Example

* prompt parts adapted & combined for presentation

```
self.generate_query = dspy.ChainOfThought("context, question -> query")
```

```
Carefully read the provided `context` and `question`. Your task is to formulate a concise
and relevant `query` that could be used to retrieve information from a search engine to
answer the question most effectively. The `query` should encapsulate...

Context: [1] Twilight is a series of four vampire-themed fantasy romance...
         [2] The Harper Connelly Mysteries is a series of fantasy...
Question: In which year was the first of the vampire-themed fantasy roma
which The Twilight Saga serves as a spin-off encyclopedic                f
Reasoning: Let's determine when that fantasy romance nove                is
Search Query: When was the first of the vampire-themed fa                vels published?

Context: [1] The Victorians - Their Story In Pictures is                 od
         [2] The Caxtons: A Family Picture is an 1849 Vic                Ed
Question: The Victorians is a documentary series written                 n
Reasoning: We know that the documentary series is about Victorian art and
was written by Jeremy Paxman. We need to find the year in which Jeremy Paxm
Search Query: Jeremy Paxman birth year
```

Scores

55%

with **GPT-3.5**

on multi-hop QA

39%

with **T5-770M**

(+ finetuning)

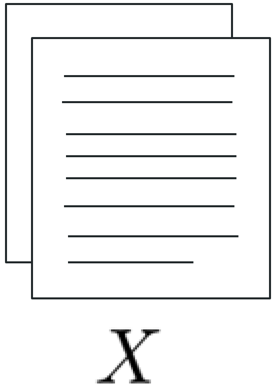
50%

with **Llama2-13B**

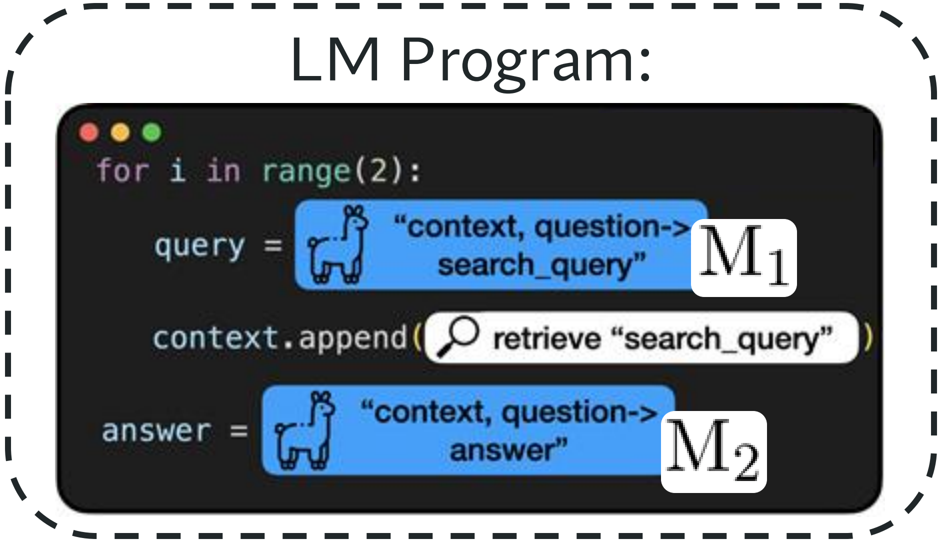
* prompt parts adapted & cor

The Problem

Training Input



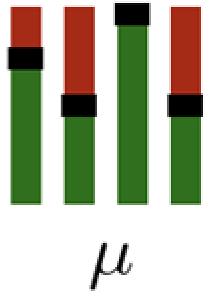
LM Program:



```
for i in range(2):
    query = "context, question-> search_query" M1
    context.append( retrieve "search_query" )
    answer = "context, question-> answer" M2
```

$\Phi_{\Theta,\Pi}$

Metric



Given a small training set $X = \{(x_1, m_1), \ldots, (x_{|X|}, m_{|X|})\}$
and a metric $\mu : \mathcal{Y} \times \mathcal{M} \to \mathbb{R}$ for labels or hints $\mathcal{M}$.
and a LM Program $\Phi_{\Theta,\Pi}$ of modules $M_{1\ldots n}$ with prompts Π and weights Θ.

Optimize: $$\arg\max_{\Theta,\Pi} \frac{1}{|X|} \sum_{(x,m)\in X} \mu\left(\Phi_{\Theta,\Pi}(x), m\right)$$

How does it work?

- **Instructions:** Use an LLM to propose text instructions to solve your task (grounded in the context of your program, data, etc.)
- **Examples:** Run your program several times to generate possible examples to add to the prompt.
- **Search:** Search over the combination of these prompt components using trial and error and seeing what works on a validation set.

🧩**Getting started w/ DSPy**

DSPy tutorial

⚡ **Getting started with Prompt Optimization**

Notebook

[Slides credit to Michael Ryan]

Summary

- ✓**Transformers and Large Language Models** (30 mins)
- ✓**Prompting** (20 mins)
  - ✓Zero-shot, few-shot
  - ✓Chain-of-thought, tree-of-thought, graph-of-thought
  - ✓Answer engineering
- ✓**Optimization and Calibration** (20mins)
  - ✓Sensitivity and inconsistency
  - ✓Output biases and calibration
  - ✓Optimization via DSPY

CS 329X: Human Centered LLMs

Learning from Human Preferences

Diyi Yang

Limitations of Instruction Finetuning

- One limitation of instruction finetuning is obvious: it’s **expensive** to collect groundtruth data for tasks.

- **Problem 1:** tasks like open-ended creative generation have no right answer.
- **Problem 2:** language modeling penalizes all token-level mistakes equally, but some errors are worse than others.

- Even with instruction finetuning, there is a **mismatch** between the LM objective and the objective of “satisfy human preferences”!
- Can we explicitly attempt to satisfy human preferences?

[Slide from CS224n]

Optimizing for Human Preferences

- Let's say we were training a language model on some task (e.g. summarization).
- For each LM sample s, imagine we had a way to obtain a ***human reward*** of that summary: $R(s) \in \mathbb{R}$, higher is better.

SAN FRANCISCO,
California (CNN) --
A magnitude 4.2
earthquake shook the
San Francisco
...
overturn unstable
objects.

An earthquake hit
San Francisco.
There was minor
property damage,
but no injuries.

s_1

$R(s_1) = 8.0$

The Bay Area has
good weather but is
prone to
earthquakes and
wildfires.

s_2

$R(s_2) = 1.2$

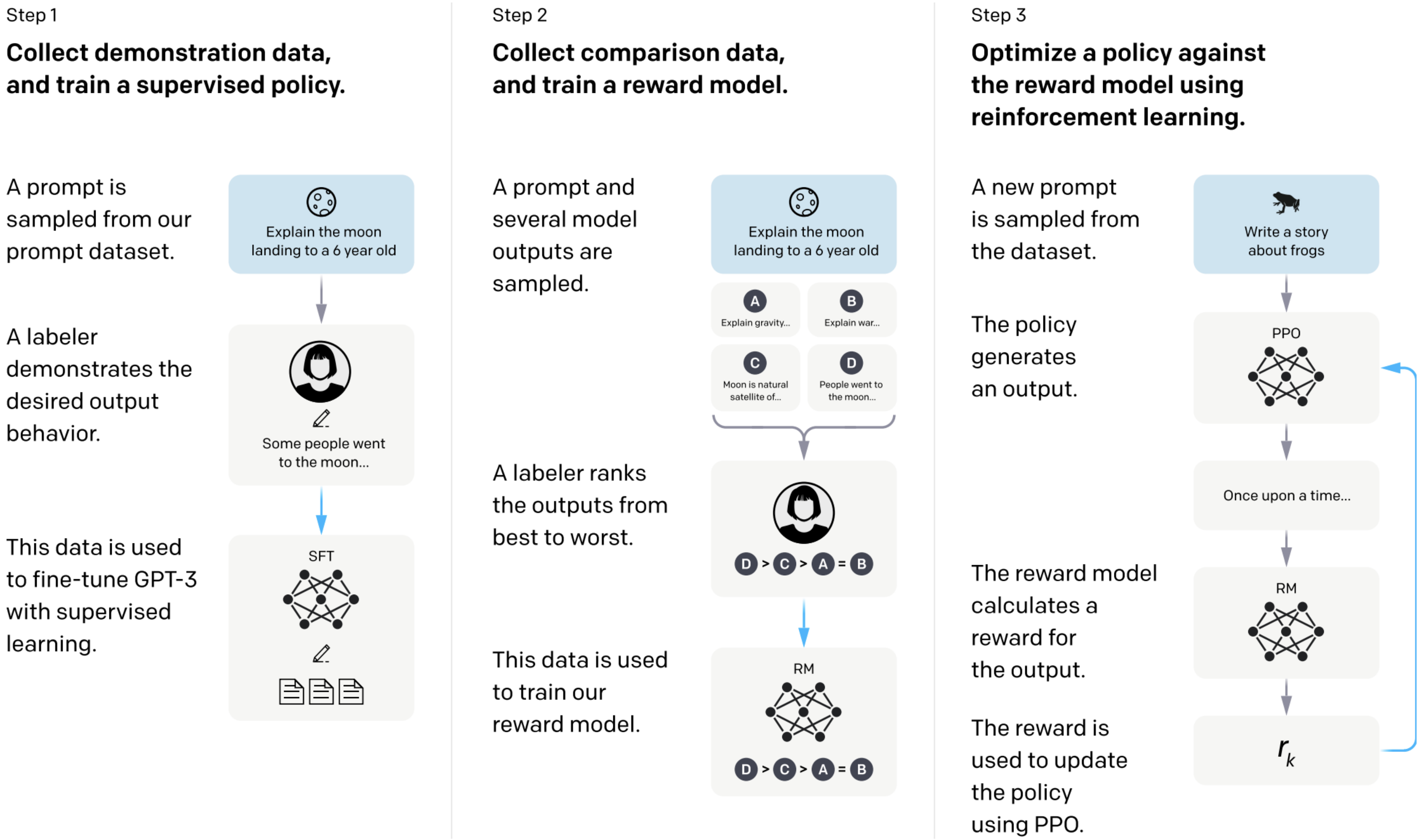
Optimizing for Human Preferences

- Let's say we were training a language model on some task (e.g. summarization).
- For each LM sample s, imagine we had a way to obtain a ***human reward*** of that summary: $R(s) \in \mathbb{R}$, higher is better.

- Now we want to maximize the expected reward of samples from our LM:

$$\mathbb{E}_{\hat{s} \sim p_\theta(s)}[R(\hat{s})]$$

High-level instantiation: ‘RLHF’ pipeline



Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Second + third steps: maximize reward (but how??)

Optimizing for human preferences

- How do we actually change our LM parameters θ to maximize this?

$$\mathbb{E}_{\hat{s}\sim p_\theta(s)}[R(\hat{s})]$$

- Let's try doing gradient ascent!

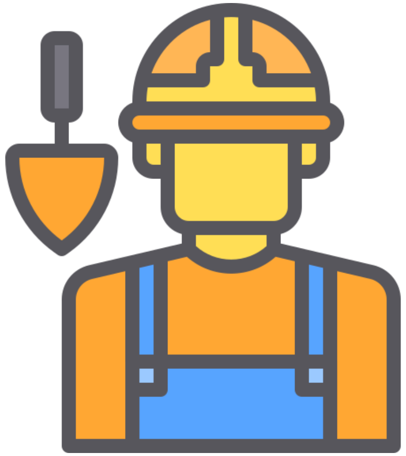
$$\theta_{t+1} \coloneqq \theta_t + \alpha \, \nabla_{\theta_t} \mathbb{E}_{\hat{s}\sim p_{\theta_t}(s)}[R(\hat{s})]$$

How do we estimate this expectation??

What if our reward function is non-differentiable??

- **Policy gradient** methods in RL (e.g., REINFORCE; [Williams, 1992]) give us tools for estimating and optimizing this objective.

How do we model human preferences?



Now for any **arbitrary, non-differentiable reward function** $R(s)$, we can train our language model to maximize expected reward.

Problem 1: Human-in-the-loop is expensive!

- **Solution:** instead of directly asking humans for preferences, **model their preferences** as a separate (NLP) problem! [Knox and Stone, 2009]

```
An earthquake hit
San Francisco.
There was minor
property damage,
but no injuries.
```

s_1

$R(s_1) = 8.0$

```
The Bay Area has
good weather but is
prone to
earthquakes and
wildfires.
```

s_2

$R(s_2) = 1.2$

Train an LM $RM_\phi(s)$ to predict human preferences from an annotated dataset, then optimize for RM_ϕ instead.

Problem 2: Human judgements are noisy and miscalibrated!

- **Solution:** instead of asking for direct ratings, ask for **pairwise comparisons**, which can be more reliable [Phelps et al., 2015; Clark et al., 2018]

```
A 4.2 magnitude
earthquake hit
San Francisco,
resulting in
massive damage.
```

$$s_3$$

$$R(s_3) = \ ?$$

$$R(s_3) = \quad 4.1? \quad 6.6? \quad 3.2?$$

Problem 2: Human judgements are noisy and miscalibrated!

An earthquake hit San Francisco. There was minor property damage, but no injuries.

s_1

$>$

A 4.2 magnitude earthquake hit San Francisco, resulting in massive damage.

s_3

$>$

The Bay Area has good weather but is prone to earthquakes and wildfires.

s_2

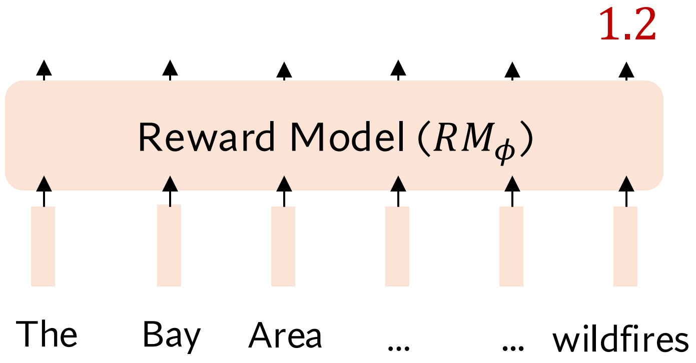
Bradley-Terry [1952] paired comparison model

$$J_{RM}(\phi) = -\mathbb{E}_{(s^w,s^l)\sim D}\left[\log \sigma(RM_\phi(s^w) - RM_\phi(s^l))\right]$$

"winning" sample

"losing" sample

s^w should score higher than s^l



RLHF: Putting it all together [Christiano et al., 2017; Stiennon et al., 2020]

Finally, we have everything we need:

- A pretrained (possibly instruction-finetuned) LM $p^{PT}(s)$
- A reward model $RM_{\phi}(s)$ that produces scalar rewards for LM outputs, trained on a dataset of human comparisons
- A method for optimizing LM parameters towards an arbitrary reward function.

RLHF: Putting it all together [Christiano et al., 2017; Stiennon et al., 2020]

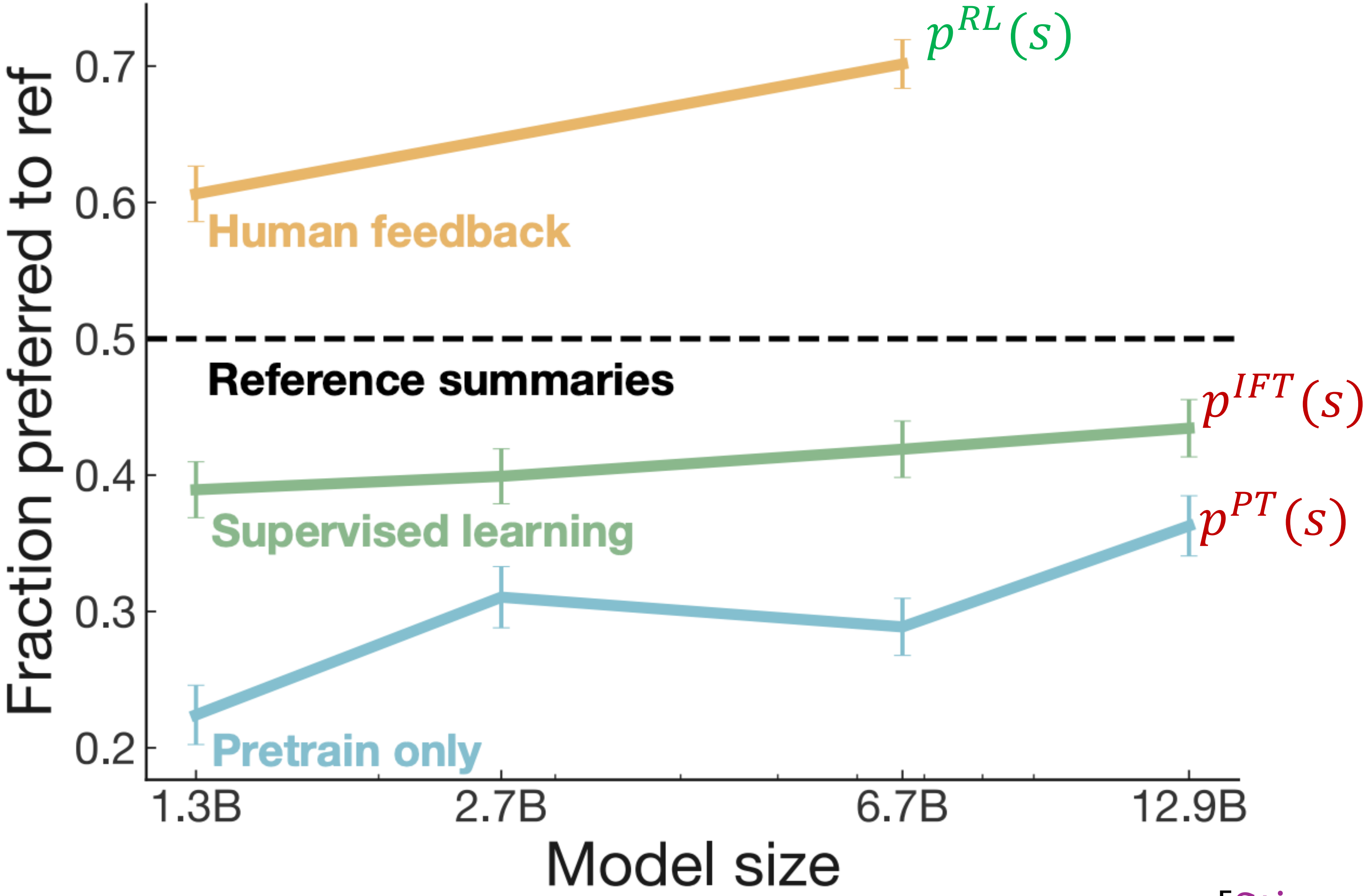
- Now to do RLHF:
  - Initialize a copy of the model $p_\theta^{RL}(s)$
  - Optimize the following reward with RL:

$$R(s) = RM_\phi(s) - \beta \underbrace{\log \left(\frac{p_\theta^{RL}(s)}{p^{PT}(s)} \right)}$$

Pay a price when $p_\theta^{RL}(s) > p^{PT}(s)$

This is a penalty which prevents us from diverging too far from the pretrained model. In expectation, it is known as the **Kullback-Leibler** (**KL**) divergence between $p_\theta^{RL}(s)$ and $p^{PT}(s)$.

RLHF provides gains over pretraining + finetuning



[Stiennon et al., 2020]

Learning from Human Feedback

- ✓ RLHF
- ❑ DPO
- ❑ Limitations of human feedback

Reading Questions

- Could there be a way to make labeling jobs for RLHF high-paying and shift them to developed countries?
- questions arise about which human values should be reflected in the model's training—whether universal values can truly exist or whether diverse perspectives need to be equally represented to avoid perpetuating existing inequalities.
- Yann LeCun cake analogy: self-supervised pretraining is cake base, supervised learning is icing, and RL/RLHF is just the cherry and doesn’t provide many bits of of new information
- We are not augmenting humans' abilities enough for RLHF (e.g. the data lablers that Scale AI uses) - how can we make these tasks more scalable?

“Data annotation for RLHF should be more high-paying in developing countries”

🗳️ Yes

🗳️ No

Join by Web	**PollEv.com /rosewang972**
Join by Text	Send **rosewang972** to **37607**