Modern LLMs & “post” training

Some slides and content today derived from materials by Mohit Iyyer (CS685 @ UMass) and Anna Rogers (“A Primer on BERTology”, TACL 2020)

Why isn't pre-training enough?

What we *want*: Generally useful models

What we *get*: Models capable of producing text capably. This is what we asked for! Given a big dataset of unlabeled data *D*:

$$\min_{\theta} -\frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \sum_{t=1}^{T-1} \log \pi_\theta(x_{t+1} \mid x_t, \ldots, x_1)$$

Why isn't pre-training enough?

What we *want*: Generally useful models

What we *get*: Models capable of producing text capably. This is what we asked for! Given a big dataset of unlabeled data *D*:

$$\min_{\theta} -\frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \sum_{t=1}^{T-1} \log \pi_\theta(x_{t+1} \mid x_t, \ldots, x_1)$$

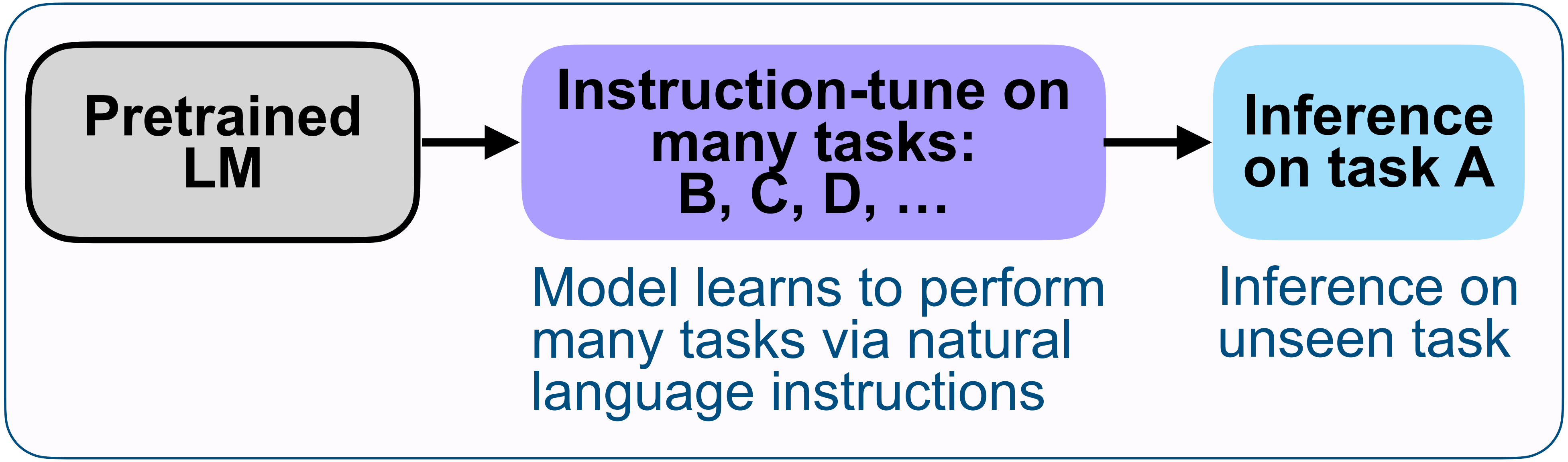
Post-training is the idea of **"aligning"** the model with what we want. This requires some sort of *supervision*.

How to “align”

An active area of research

Two main strategies we’ll discuss: *Instruction fine-tuning* and *Reinforcement Learning from Human Feedback (RLHF)*

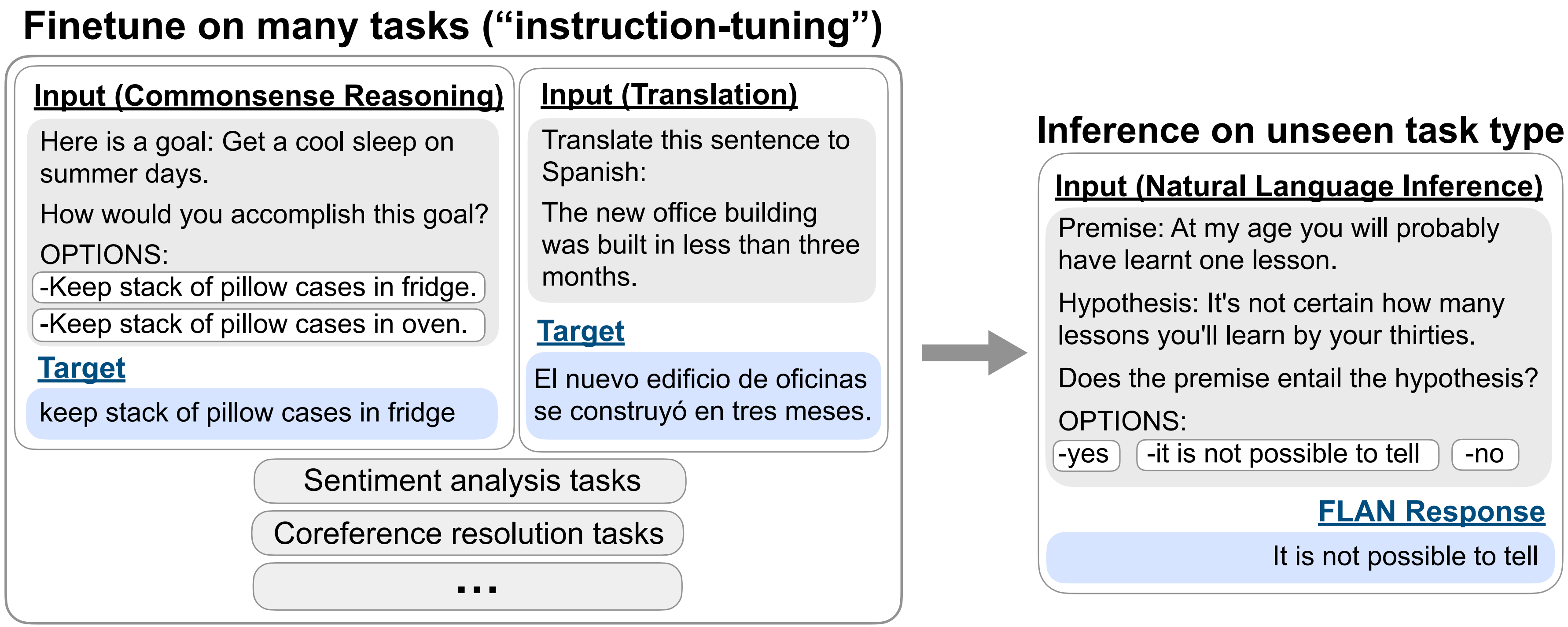
Instruction fine-tuning



FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS

Jason Wei*, Maarten Bosma*, Vincent Y. Zhao*, Kelvin Guu*, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le
Google Research

Instruction fine-tuning



FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS

Jason Wei*, Maarten Bosma*, Vincent Y. Zhao*, Kelvin Guu*, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le

Google Research

Instruction fine-tuning

Natural language inference (7 datasets)
- ANLI (R1-R3)
- RTE
- CB
- SNLI
- MNLI
- WNLI
- QNLI

Commonsense (4 datasets)
- CoPA
- HellaSwag
- PiQA
- StoryCloze

Sentiment (4 datasets)
- IMDB
- Sent140
- SST-2
- Yelp

Paraphrase (4 datasets)
- MRPC
- QQP
- PAWS
- STS-B

Closed-book QA (3 datasets)
- ARC (easy/chal.)
- NQ
- TQA

Struct to text (4 datasets)
- CommonGen
- DART
- E2ENLG
- WEBNLG

Translation (8 datasets)
- ParaCrawl EN/DE
- ParaCrawl EN/ES
- ParaCrawl EN/FR
- WMT-16 EN/CS
- WMT-16 EN/DE
- WMT-16 EN/FI
- WMT-16 EN/RO
- WMT-16 EN/RU
- WMT-16 EN/TR

Reading comp. (5 datasets)
- BoolQ
- OBQA
- DROP
- SQuAD
- MultiRC

Read. comp. w/ commonsense (2 datasets)
- CosmosQA
- ReCoRD

Coreference (3 datasets)
- DPR
- Winogrande
- WSC273

Misc. (7 datasets)
- CoQA
- TREC
- QuAC
- CoLA
- WIC
- Math
- Fix Punctuation (NLG)

Summarization (11 datasets)
- AESLC
- Multi-News
- SamSum
- AG News
- Newsroom
- Wiki Lingua EN
- CNN-DM
- Opin-Abs: iDebate
- XSum
- Gigaword
- Opin-Abs: Movie

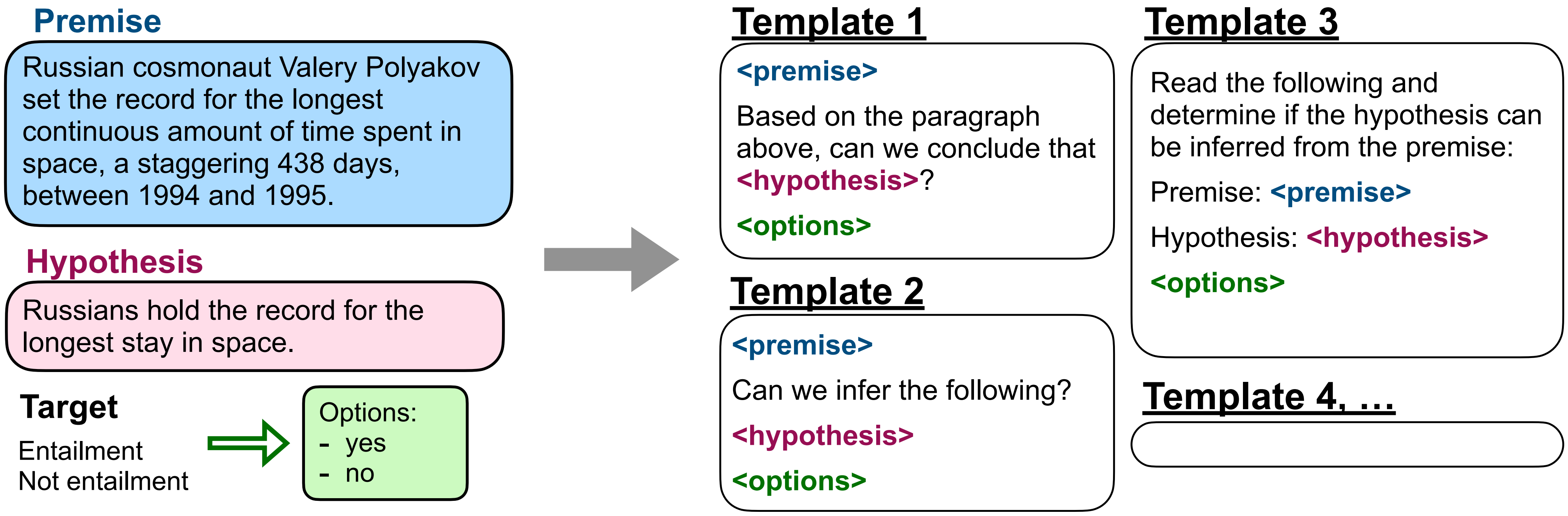
FLAN

FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS

Jason Wei*, Maarten Bosma*, Vincent Y. Zhao*, Kelvin Guu*, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le

Google Research

Multiple instruction templates per task

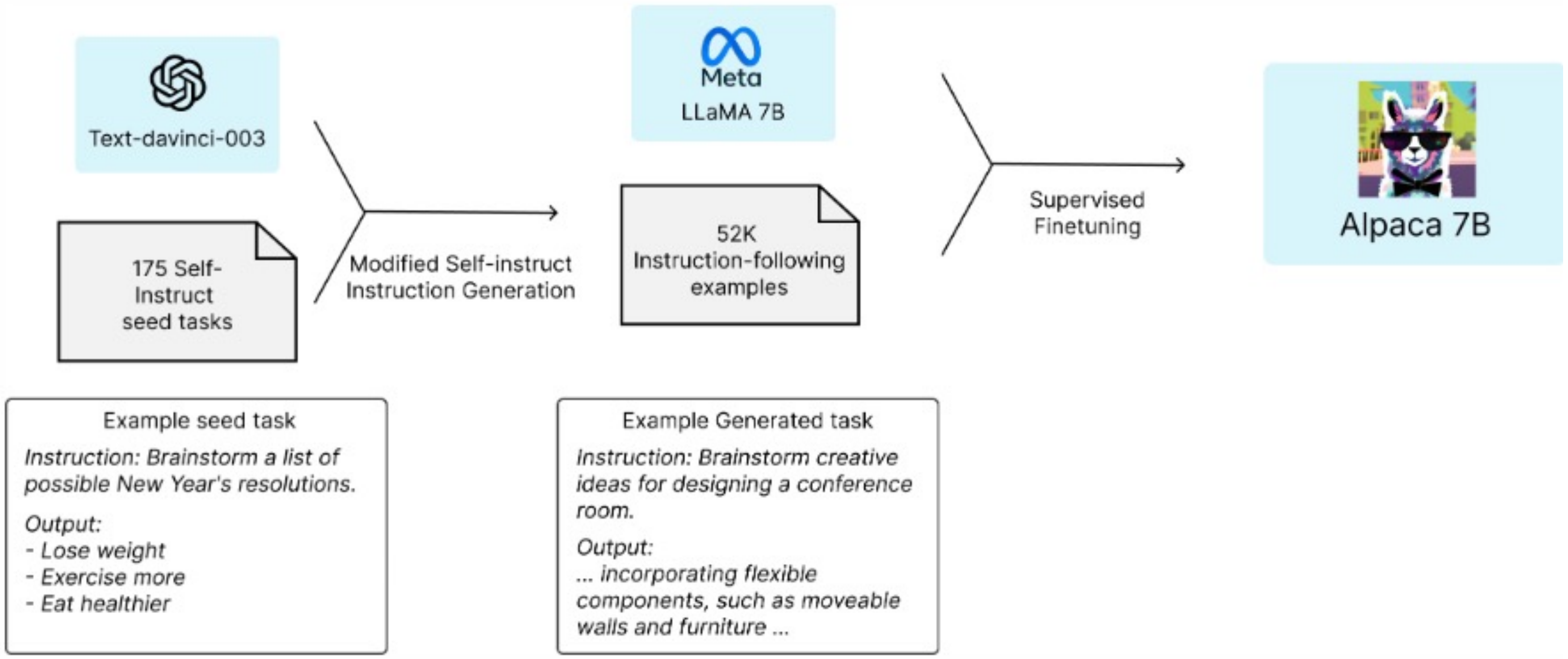


FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS

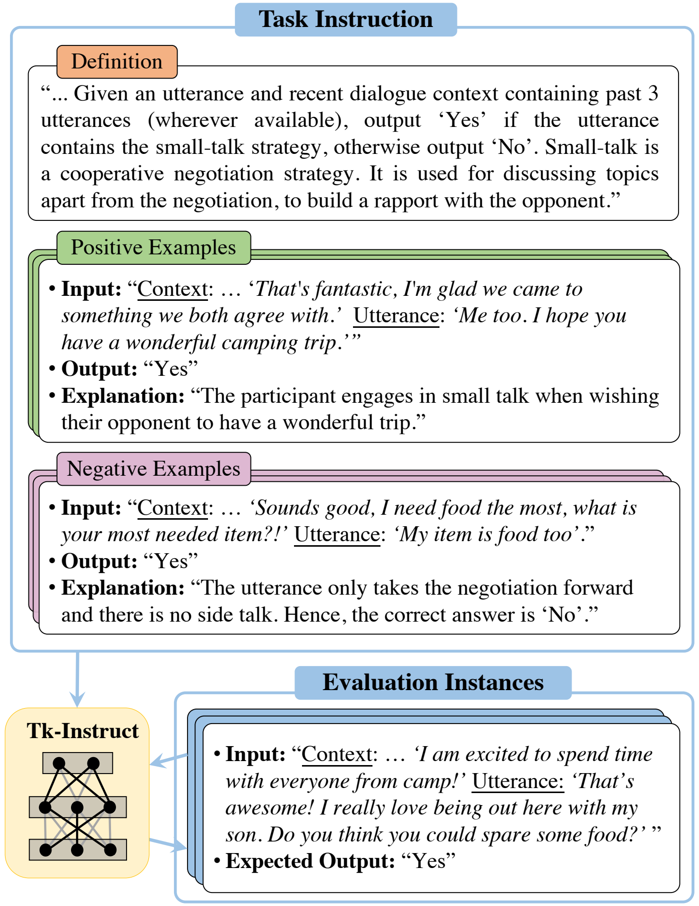
Jason Wei*, Maarten Bosma*, Vincent Y. Zhao*, Kelvin Guu*, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le

Google Research

Alpaca: Deriving training data from LLMs



Supernatural Instructions



https://instructions.apps.allenai.org/

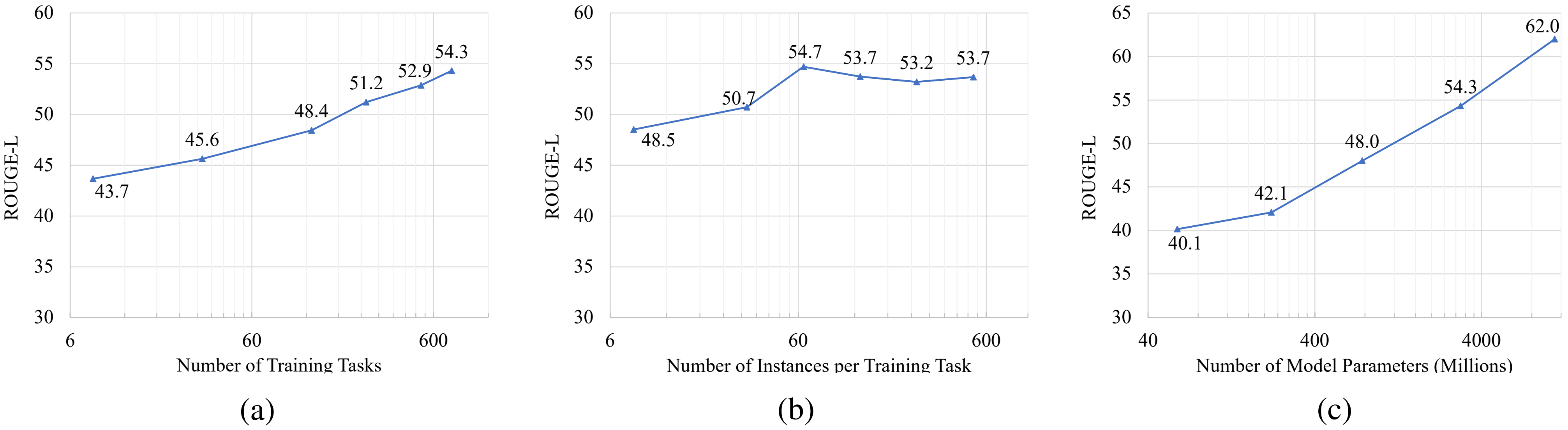
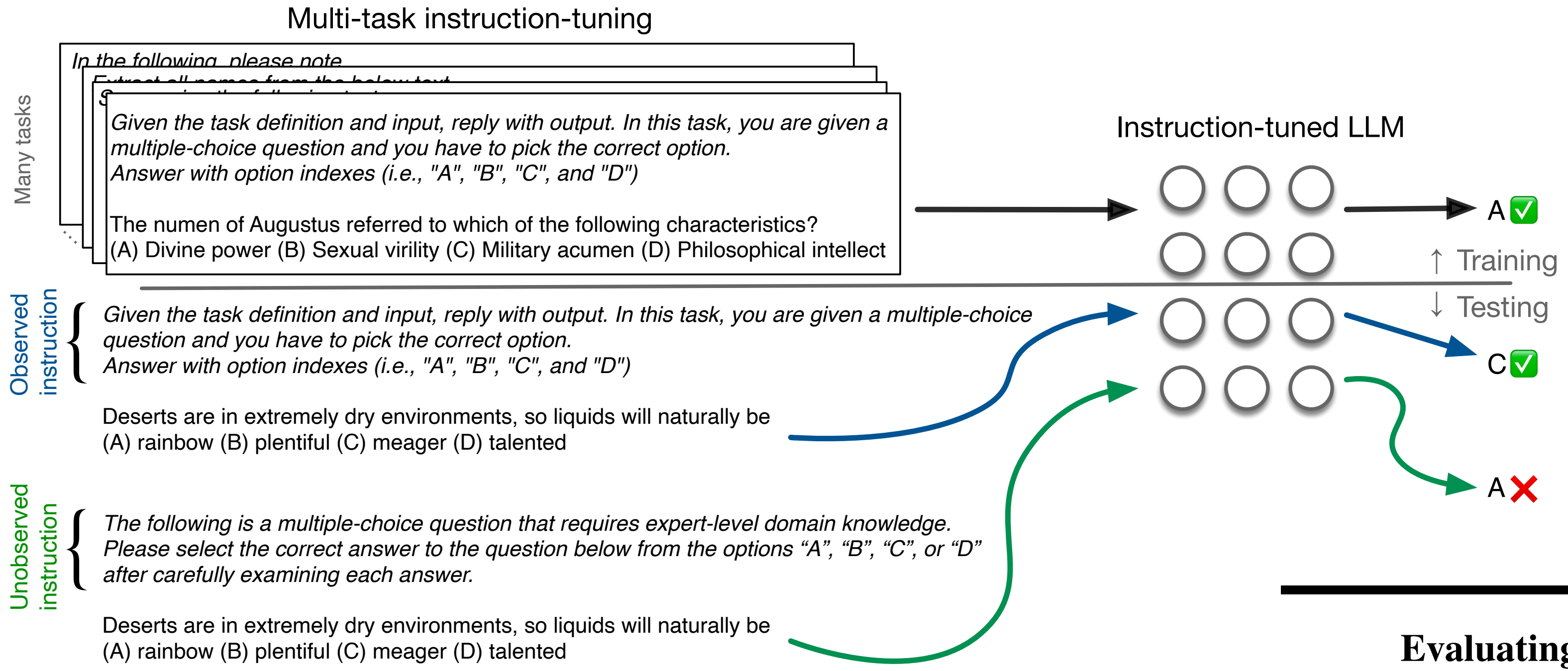


Figure 5: Scaling trends of models performance (§7.1) as a function of (a) the number of training tasks; (b) the number of instances per training task; (c) model sizes. x-axes are in log scale. The **linear growth of model performance with exponential increase in observed tasks and model size** is a promising trend. Evidently, the performance gain from more instances is limited.

SUPER-NATURALINSTRUCTIONS: Generalization via Declarative Instructions on 1600+ NLP Tasks

◇Yizhong Wang[2] ◇Swaroop Mishra[3] ♣Pegah Alipoormolabashi[4] ♣Yeganeh Kordi[5]
Amirreza Mirzaei[4] Anjana Arunkumar[3] Arjun Ashok[6] Arut Selvan Dhanasekaran[3]
Atharva Naik[7] David Stap[8] Eshaan Pathak[9] Giannis Karamanolakis[10] Haizhi Gary Lai[11]
Ishan Purohit[12] Ishani Mondal[13] Jacob Anderson[3] Kirby Kuznia[3] Krima Doshi[3] Maitreya Patel[3]
Kuntal Kumar Pal[3] Mehrad Moradshahi[14] Mihir Parmar[3] Mirali Purohit[15] Neeraj Varshney[3]
Phani Rohitha Kaza[3] Pulkit Verma[3] Ravsehaj Singh Puri[3] Rushang Karia[3] Shailaja Keyur Sampat[3]
Savan Doshi[3] Siddhartha Mishra[16] Sujan Reddy[17] Sumanta Patro[18] Tanay Dixit[19] Xudong Shen[20]
Chitta Baral[3] Yejin Choi[1,2] Noah A. Smith[1,2] Hannaneh Hajishirzi[1,2] Daniel Khashabi[21]

But!

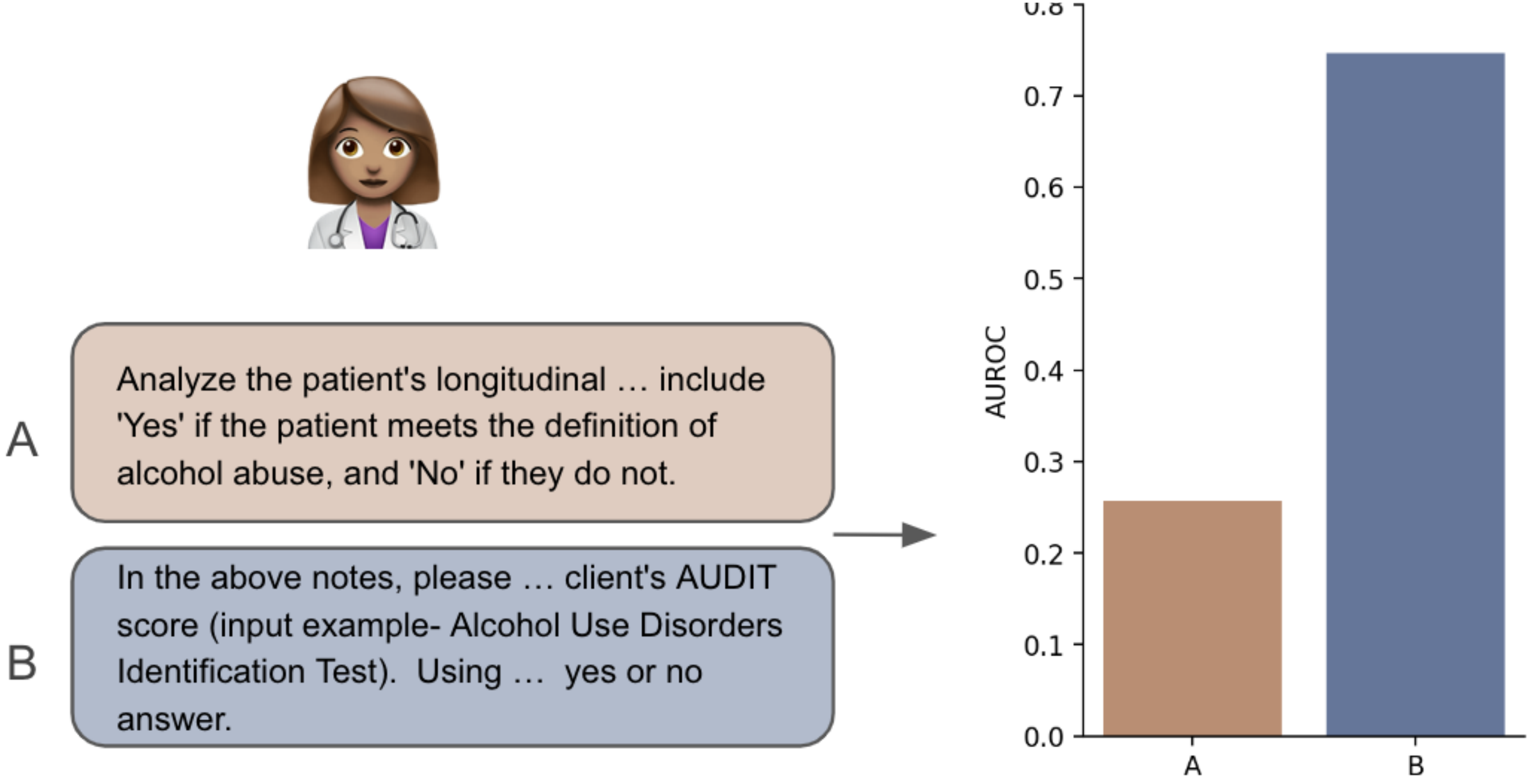


Evaluating the Zero-shot Robustness of Instruction-tuned Language Models

Jiuding Sun
Khoury College of Computer Sciences
Northeastern University
sun.jiu@northeastern.edu

Chantal Shaib
Khoury College of Computer Sciences
Northeastern University
shaib.c@northeastern.edu

Byron C. Wallace
Khoury College of Computer Sciences
Northeastern University
b.wallace@northeastern.edu



Open (Clinical) LLMs are Sensitive to Instruction Phrasings

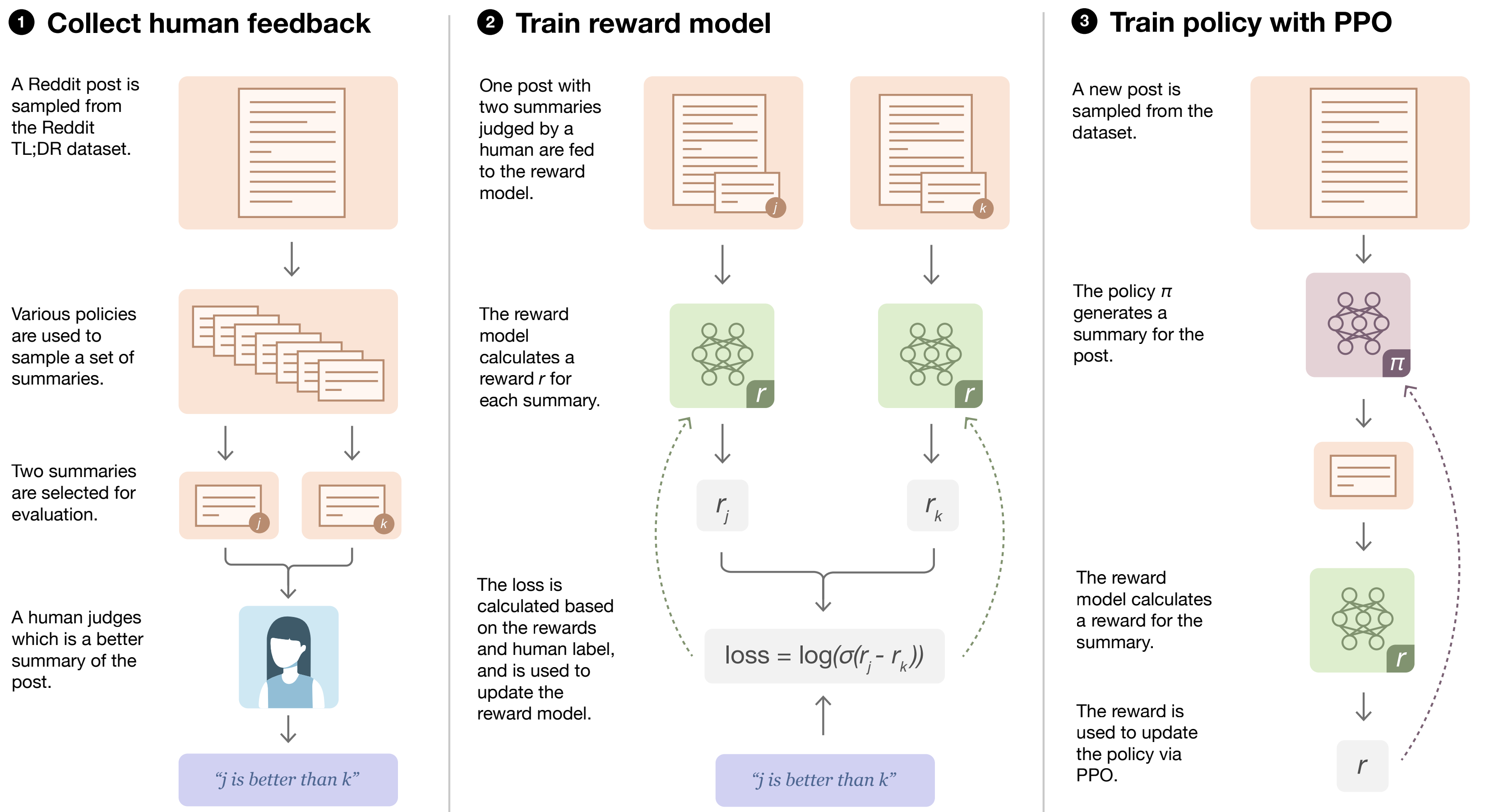
Alberto Mario Ceballos Arroyo*[γ] **Monica Munnangi*[γ]** **Jiuding Sun[γ]**
Karen Y.C. Zhang[γ] **Denis Jered McInerney[γ◇]** **Byron C. Wallace[γ]** **Silvio Amir[γ]**
[γ]Northeastern University [◇]Codametrix
{ceballosarroyo.a, munnangi.m, sun.jiu, zhang.yuchen, b.wallace,s.amir}@northeastern.edu
jmcinerney@codametrix.com

Human preferences

Often more natural to elicit *preferences* between pairs of outputs than to provide explicit examples

For instance, if we want LLMs to generate “more polite” or less biased outputs, difficult to write a bunch of examples explicitly demonstrating these things: Easier to show two examples and ask which is “more polite”

(Reinforcement) Learning from Human Feedback

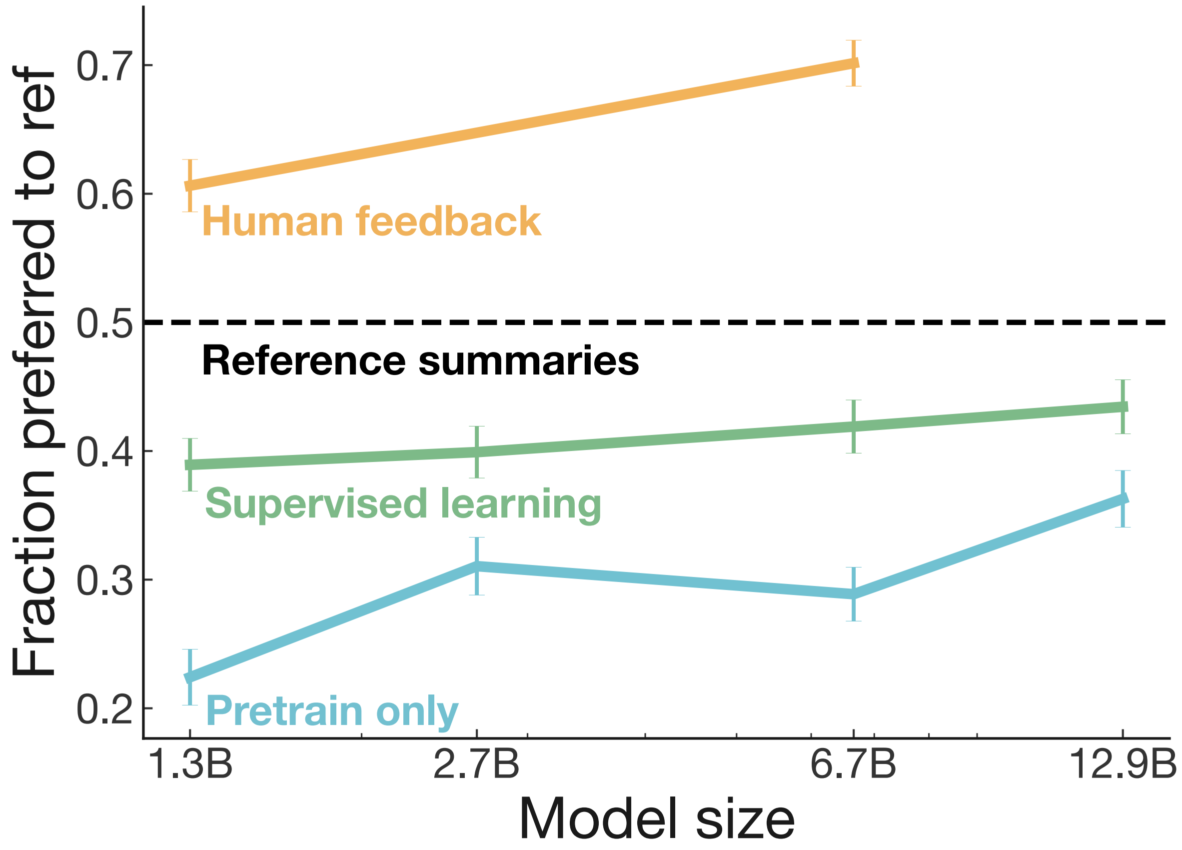


Learning to summarize from human feedback

Nisan Stiennon* Long Ouyang* Jeff Wu* Daniel M. Ziegler* Ryan Lowe*

Chelsea Voss* Alec Radford Dario Amodei Paul Christiano*

(Reinforcement) Learning from Human Feedback



Learning to summarize from human feedback

Nisan Stiennon* Long Ouyang* Jeff Wu* Daniel M. Ziegler* Ryan Lowe*

Chelsea Voss* Alec Radford Dario Amodei Paul Christiano*

Let's talk RL & PPO
[see notes]

But who wants to deal w/RL?

Direct Preference Optimization (DPO) says: Oh, we can just used supervised learning to directly optimize for preference feedback labels

Direct Preference Optimization:
Your Language Model is Secretly a Reward Model

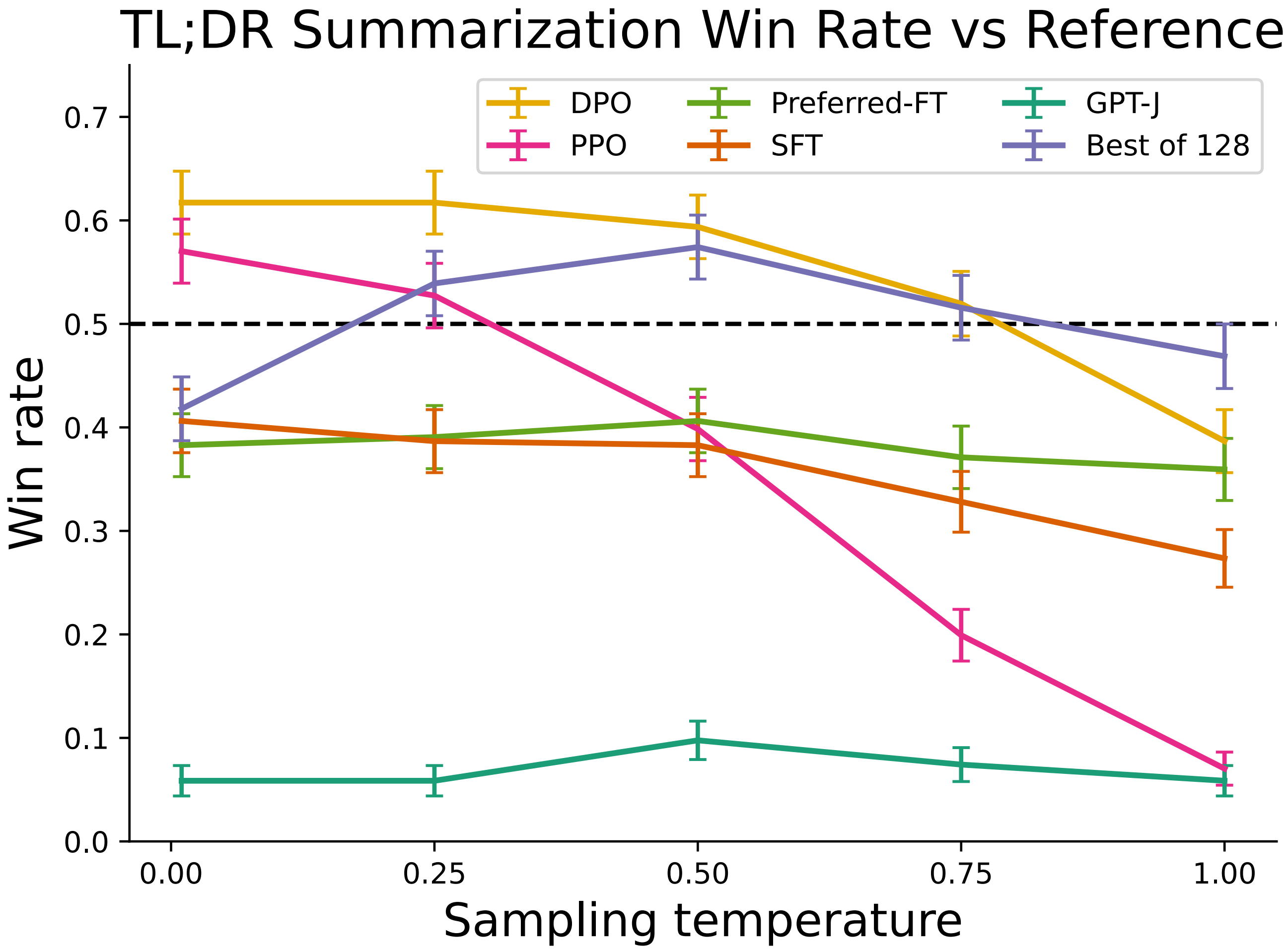
Rafael Rafailov[*†] **Archit Sharma**[*†] **Eric Mitchell**[*†]

Stefano Ermon[†‡] **Christopher D. Manning**[†] **Chelsea Finn**[†]

The objective

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right]$$

(*w* preferred to *l*)



TL;DR Summarization Win Rate vs Reference
DPO
PPO
Preferred-FT
SFT
GPT-J
Best of 128
Win rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.00
0.25
0.50
0.75
1.00
Sampling temperature