# Modern LLMs & "post" training

Some slides and content today derived from materials by Mohit lyyer (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*:

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

# 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*:

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

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

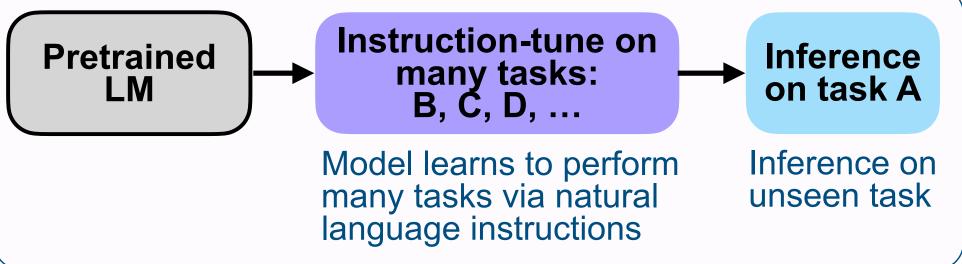
# How to "align"

An active area of research

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

### nce sk A

# Instruction fine-tuning



#### nce sk A

### 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**

Input (Translation)

Spanish:

months.

Target

Translate this sentence to

was built in less than three

El nuevo edificio de oficinas

se construyó en tres meses.

The new office building

#### Finetune on many tasks ("instruction-tuning")

#### Input (Commonsense Reasoning)

Here is a goal: Get a cool sleep on summer days.

How would you accomplish this goal? OPTIONS:

-Keep stack of pillow cases in fridge. -Keep stack of pillow cases in oven.

#### Target

keep stack of pillow cases in fridge

Sentiment analysis tasks

Coreference resolution tasks

. . .

#### Inference on unseen task type

#### Input (Natural Language Inference)

Premise: At my age you will probably have learnt one lesson.

Hypothesis: It's not certain how many lessons you'll learn by your thirties.

Does the premise entail the hypothesis?

OPTIONS:

-yes (-it is not possible to tell (-no)

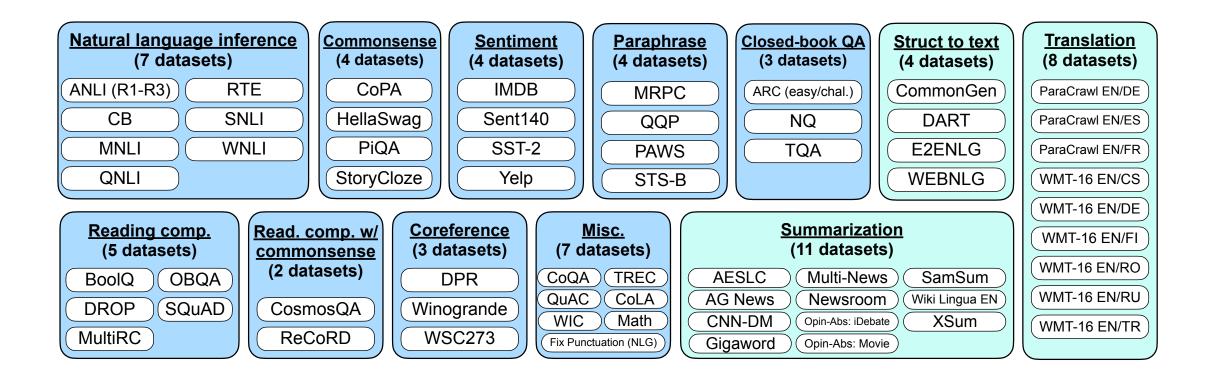
FLAN Response

It is not possible to tell

### FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS

Jason Wei<sup>\*</sup>, Maarten Bosma<sup>\*</sup>, Vincent Y. Zhao<sup>\*</sup>, Kelvin Guu<sup>\*</sup>, 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

### **Multiple instruction templates per task**

#### **Premise**

Russian cosmonaut Valery Polyakov set the record for the longest continuous amount of time spent in space, a staggering 438 days, between 1994 and 1995.

#### **Hypothesis**

Russians hold the record for the longest stay in space.

Target

Entailment

Not entailment

Options: - yes - no

<u>Template 1</u> <u>Template 3</u> premise> Read the following and Based on the paragraph above, can we conclude that <hypothesis>? Premise: <premise> <options> <options> Template 2 <premise> Can we infer the following? Template 4, ... <hypothesis> <options>

determine if the hypothesis can be inferred from the premise:

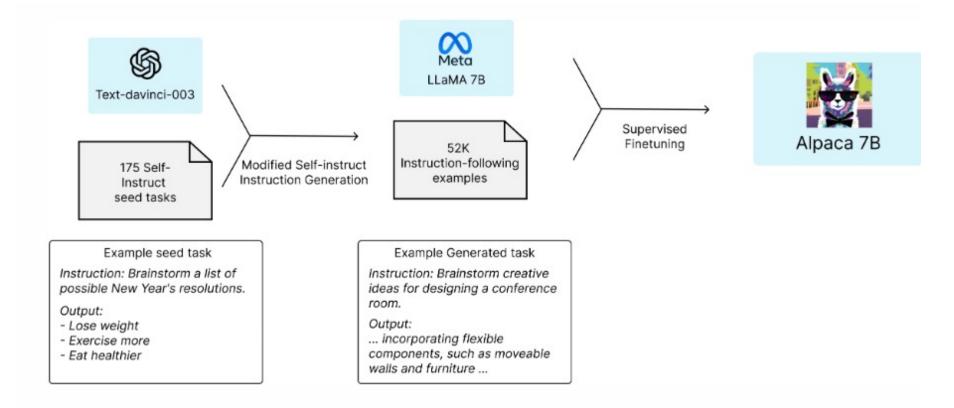
Hypothesis: <hypothesis>

#### 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**

#### **Task Instruction**

#### Definition

"... Given an utterance and recent dialogue context containing past 3 utterances (wherever available), output 'Yes' if the utterance contains the small-talk strategy, otherwise output 'No'. Small-talk is a cooperative negotiation strategy. It is used for discussing topics apart from the negotiation, to build a rapport with the opponent."

#### Positive Examples

• Input: "Context: ... 'That's fantastic, I'm glad we came to something we both agree with.' Utterance: 'Me too. I hope you have a wonderful camping trip."

Output: "Yes'

**Explanation:** "The participant engages in small talk when wishing their opponent to have a wonderful trip."

#### Negative Examples

• Input: "Context: ... 'Sounds good, I need food the most, what is your most needed item?!' Utterance: 'My item is food too'." Output: "Yes" Explanation: "The utterance only takes the negotiation forward

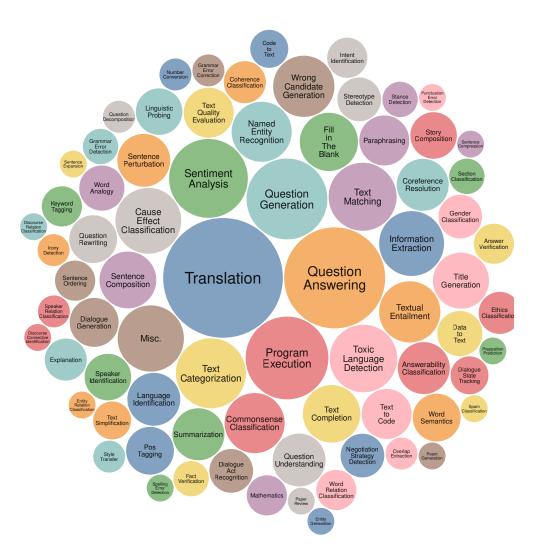
and there is no side talk. Hence, the correct answer is 'No'."

#### **Evaluation Instances**

#### **Tk-Instruct**



• Input: "Context: ... 'I am excited to spend time with everyone from camp!' Utterance: 'That's awesome! I really love being out here with my son. Do you think you could spare some food?'" • Expected Output: "Yes"



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

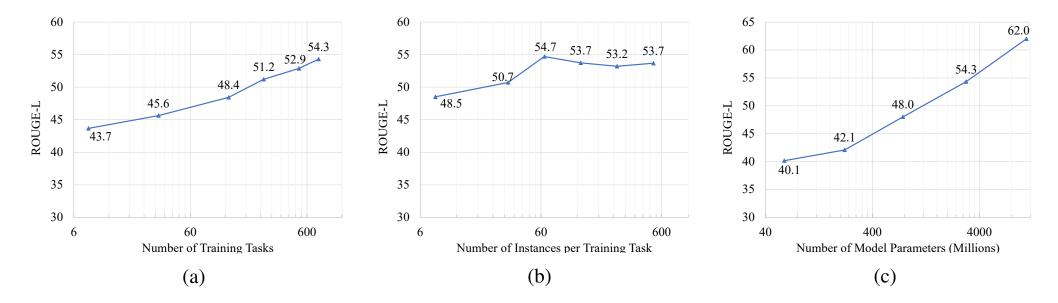


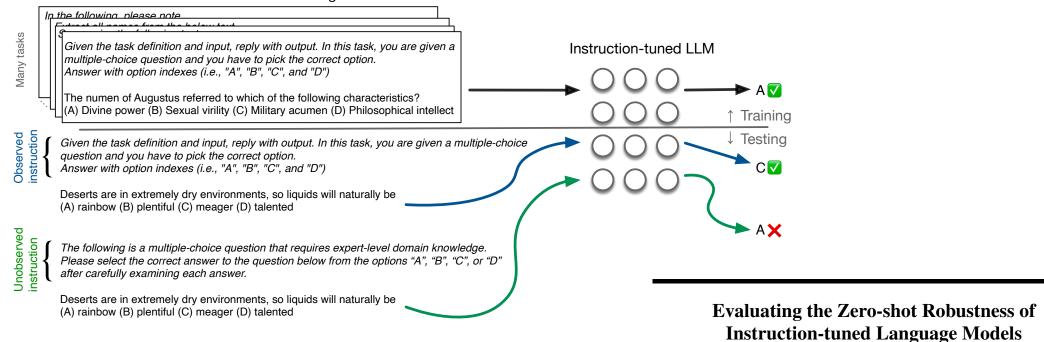
Figure 5: Scaling trends of models performance ( $\S7.1$ ) as a function of (a) the number of training tasks; (b) the number of instances per training task; (c) model sizes. *x*-axes are in log scale. The **linear growth of model performance with exponential increase in observed tasks and model size** is a promising trend. Evidently, the performance gain from more instances is limited.

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

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

# **But**!

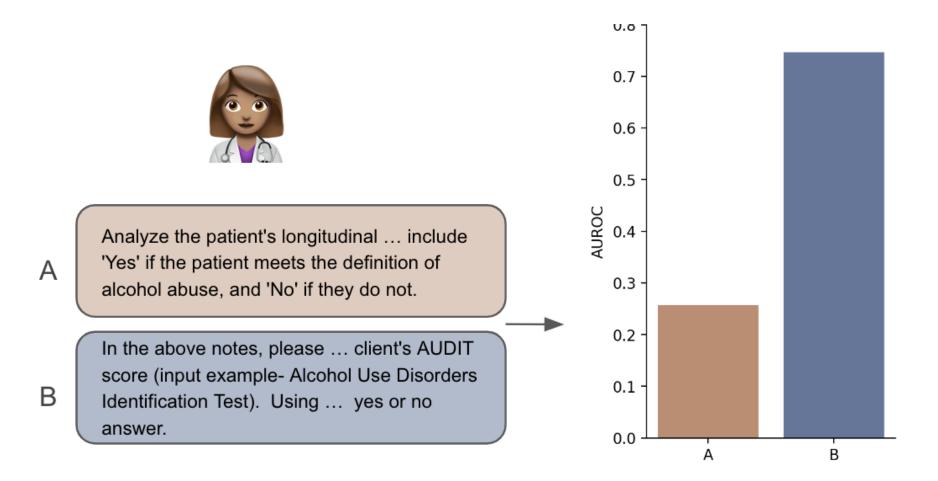
Multi-task instruction-tuning



Jiuding Sun Khoury College of Computer Sciences Kh 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\*Monica Munnangi\*Jiuding SunKaren Y.C. ZhangDenis Jered McInerneyByron C. WallaceSilvio Amir $^{\gamma}$ Northeastern University $^{\diamond}$ Codametrix

 ${ceballosarroyo.a, munnangi.m, sun.jiu, zhang.yuchen, b.wallace, s.amir} @northeastern.edu$ 

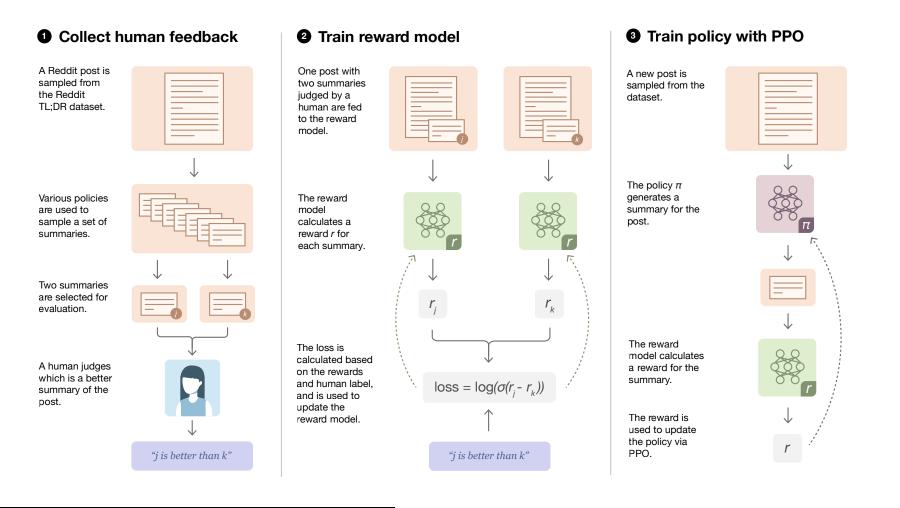
jmcinerney@codametrix.com

## 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"

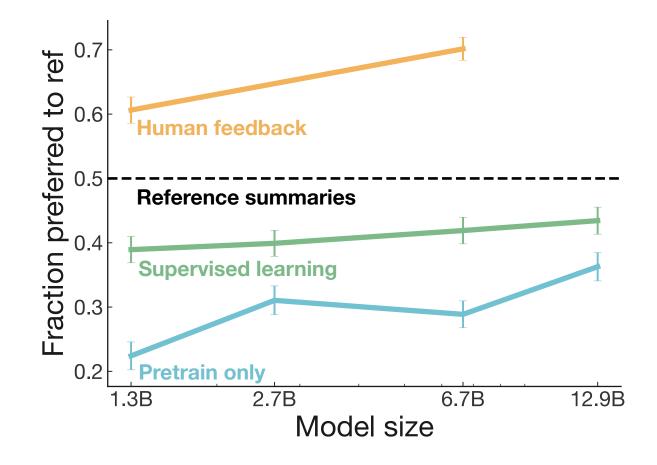
### (Reinforcement) Learning from Human Feedback



Learning to summarize from human feedback

Nisan Stiennon*	Long Ouyang*	Jeff Wu*	Daniel M. Zieg	gler*	Ryan Lowe*
Chelsea Voss*	Alec Radford	l Dario	) Amodei	Paul Cl	hristiano*

### (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 RadfordDario AmodeiPaul 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* <sup>†</sup>
Stefano Ermon <sup>†‡</sup>	Christopher D. Manning $^{\dagger}$	Chelsea Finn $^{\dagger}$

### 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 *I*)

