Instruction Tuning and RLHF

Natural Language Processing — Lecture 8

Kenneth Enevoldsen | 2024

Learning Goals

- Understanding of how modern interactive LLMs are developed, including
  - Instruction fine-tuning
  - Reinforcement Learning from Human Feedback
- A understanding of limitations of these methods and what they seek to solve
- Examples of such models

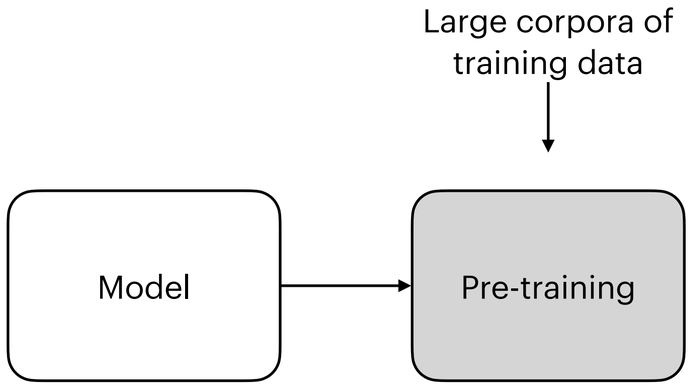
Quiz

- https://www.menti.com/al9tjd5y7j6p

Recap: Where are we at?

- Different type of models:
  - Encoders: BERT (RoBERTa, ...)
  - Decoders: GPT (Llama, ...)
  - Encoder-Decoders: T5 (BART, ...)
- Pre-training
- Prompting
- In-context learning

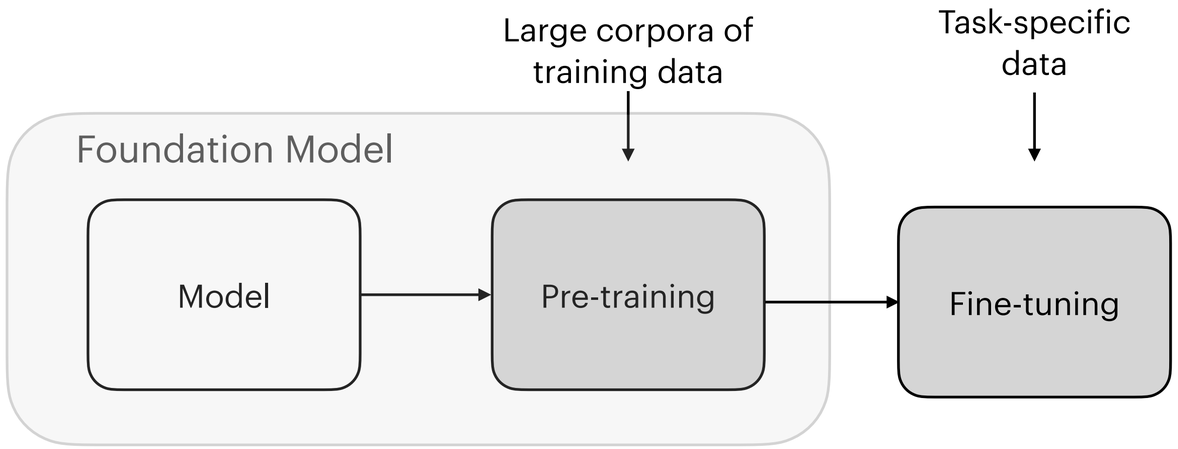
Recap: Pre-train then Fine-tuning



Context `Margrethe 2. is queen of ____`

Prediction `Denmark`

Recap: Pre-train then Fine-tuning



Context `What a pain in the a*@s`

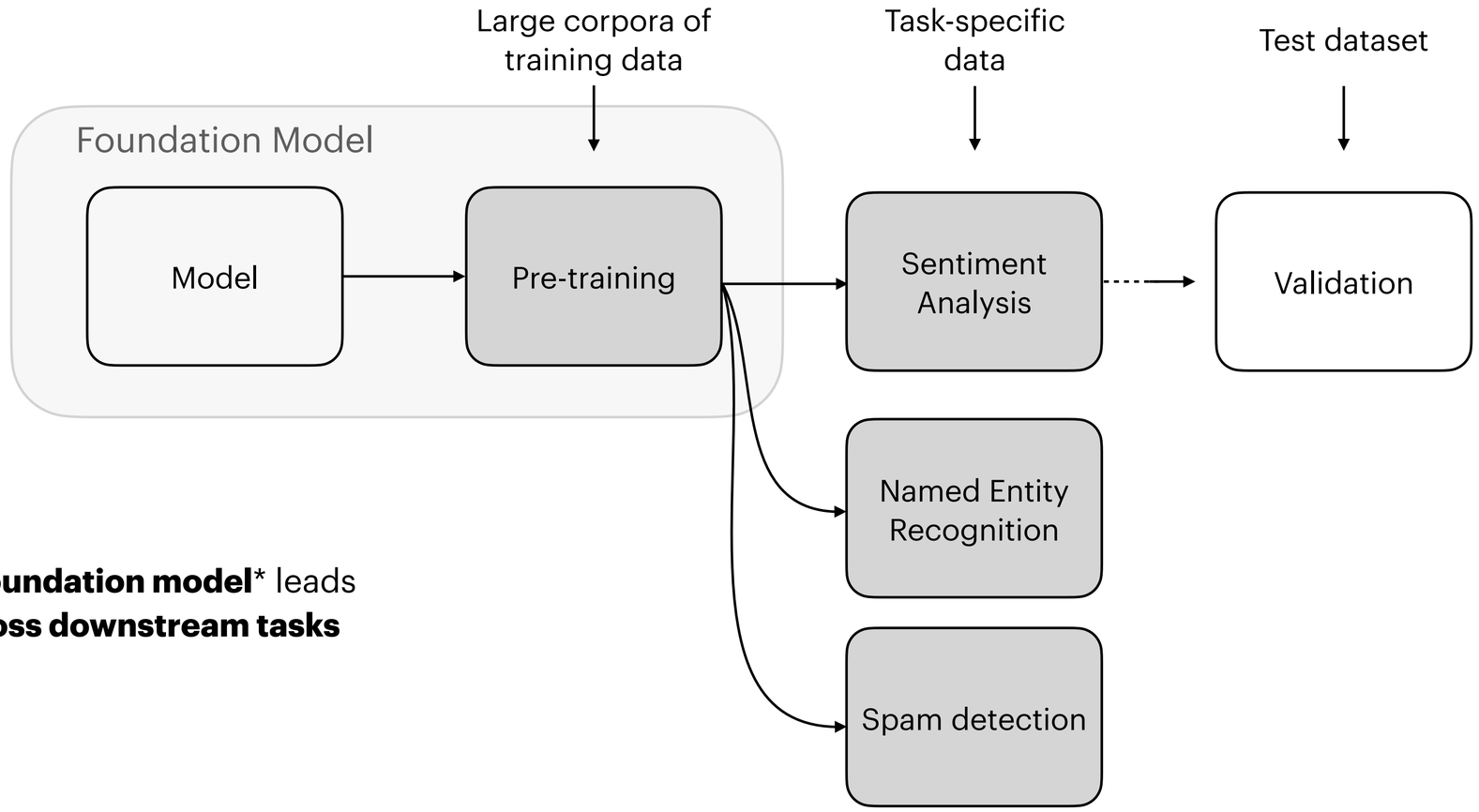
Prediction
```
97% Negative
 0% Neutral
 3% Positive
```

Sources & Notes

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

CENTER FOR HUMANITIES COMPUTING

Recap: Generalization



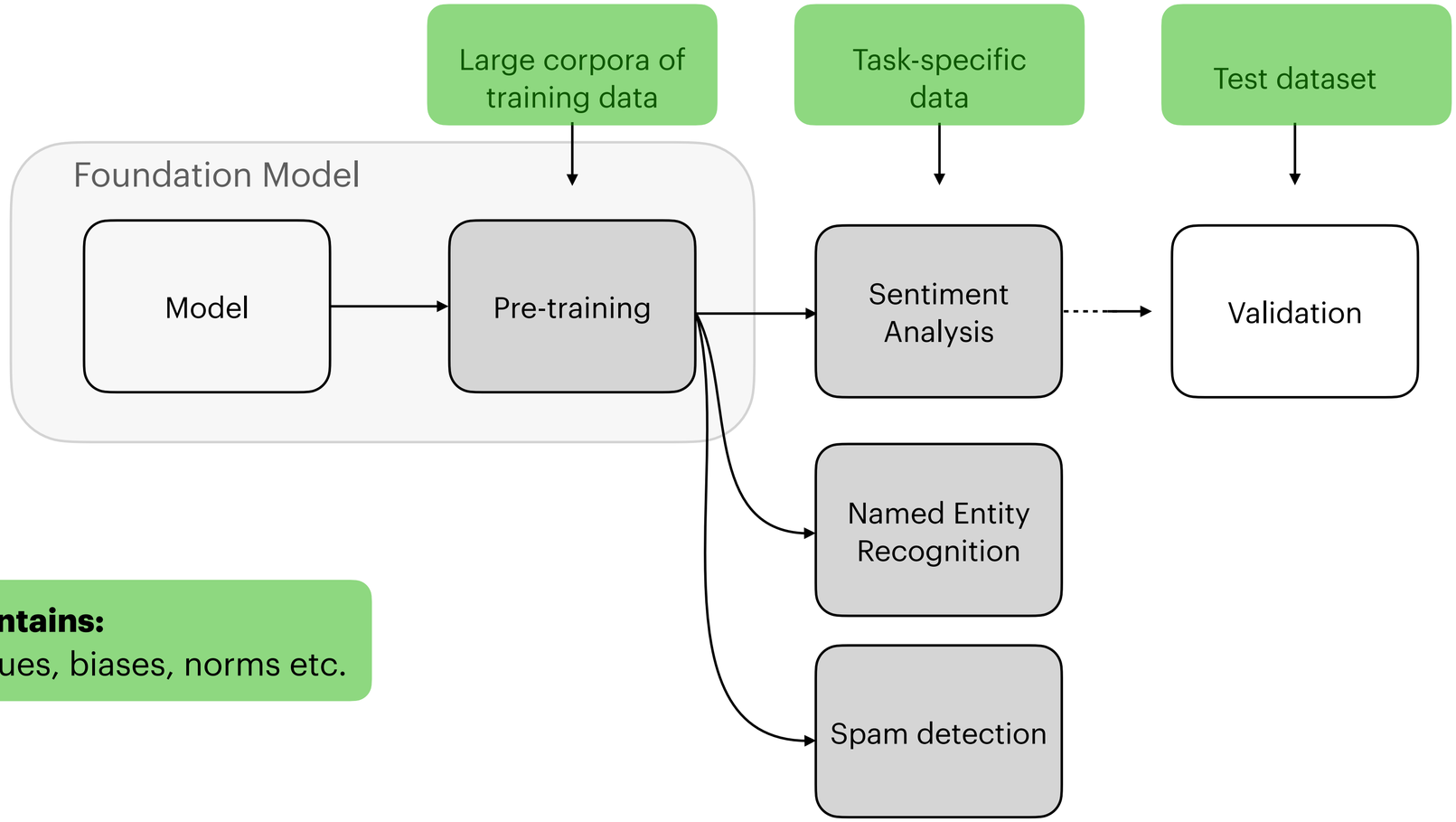
Improving the **foundation model*** leads to improve **across downstream tasks**

Sources & Notes

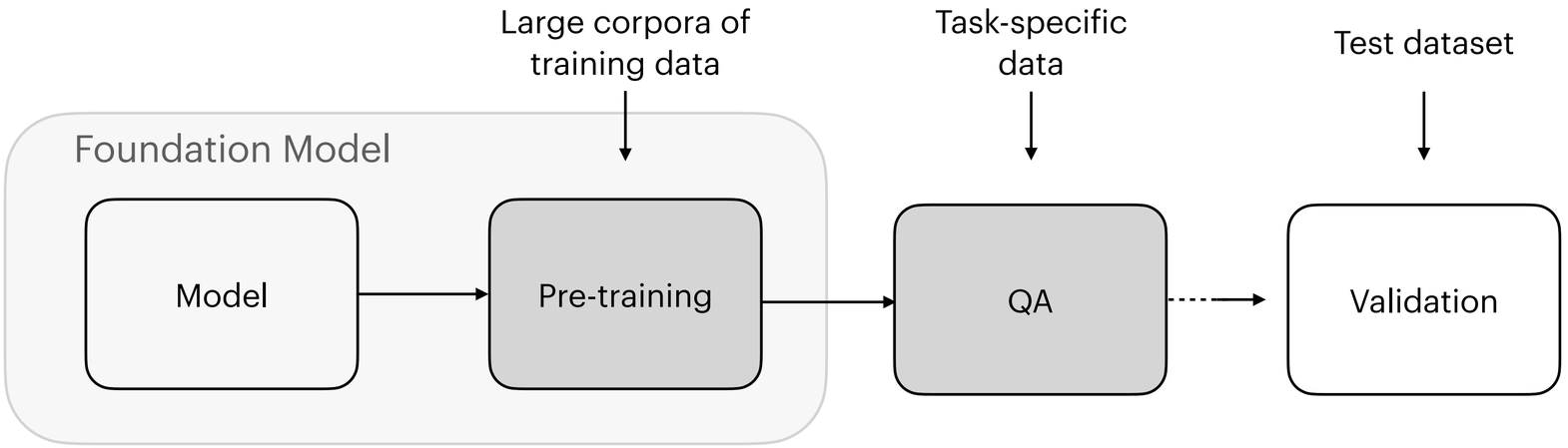
Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

*Foundation model is also called core model or LLM and is sometimes a contested term

Bias


Large corpora of training data
Task-specific data
Test dataset
Foundation Model
Model
Pre-training
Sentiment Analysis
Validation
Named Entity Recognition
Spam detection
Contains:
Values, biases, norms etc.

Recap: Generalization



Previous goal: One-model for every tasks
New Goal: Zero-shot Generalization

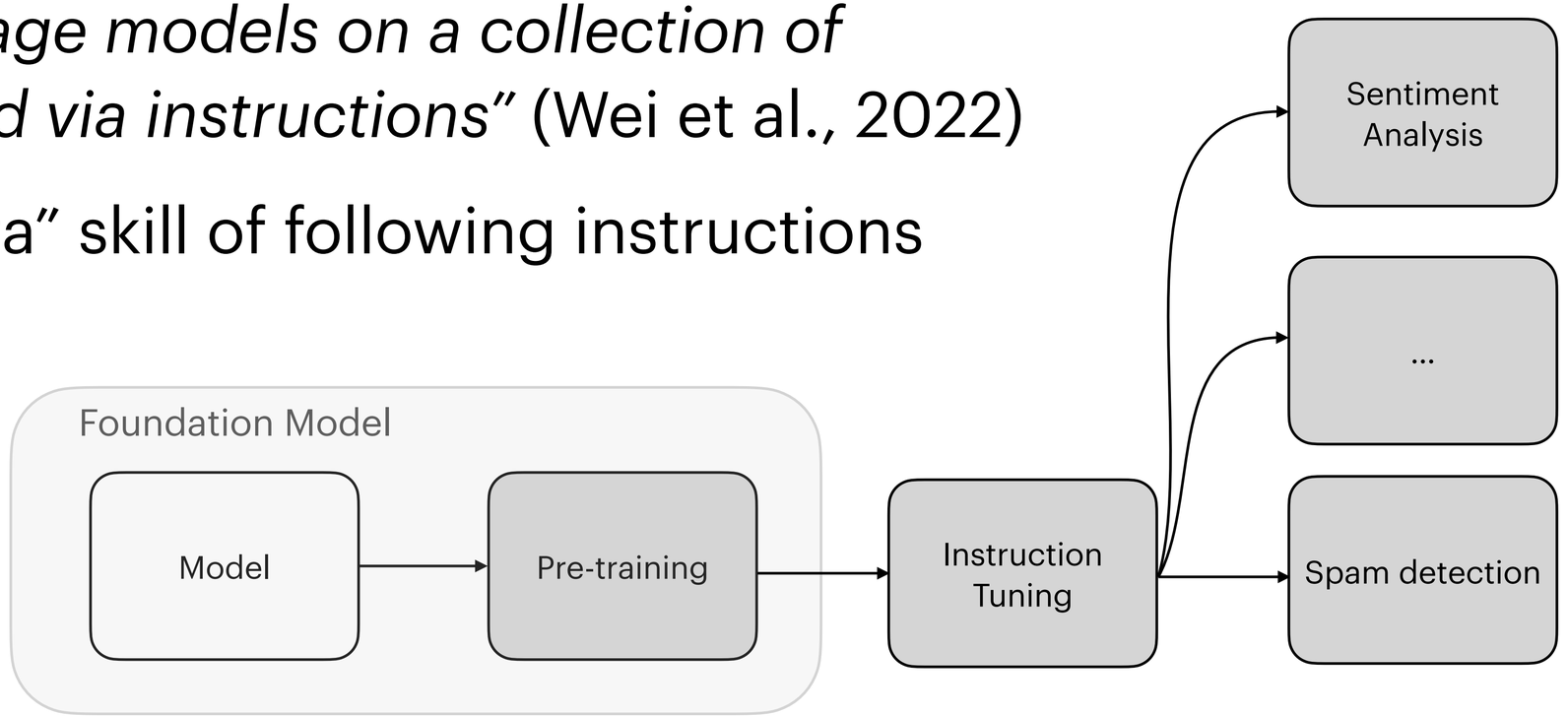
Sources & Notes

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

*Foundation model is also called core model or LLM and is sometimes a contested term

Goal: Zero-shot Generalization

- Solution: Instruction tuning
  - *“finetuning language models on a collection of datasets described via instructions”* (Wei et al., 2022)
- Models learn a “meta” skill of following instructions



Sources & Notes

Quote from:

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2021). Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652.

Comparison

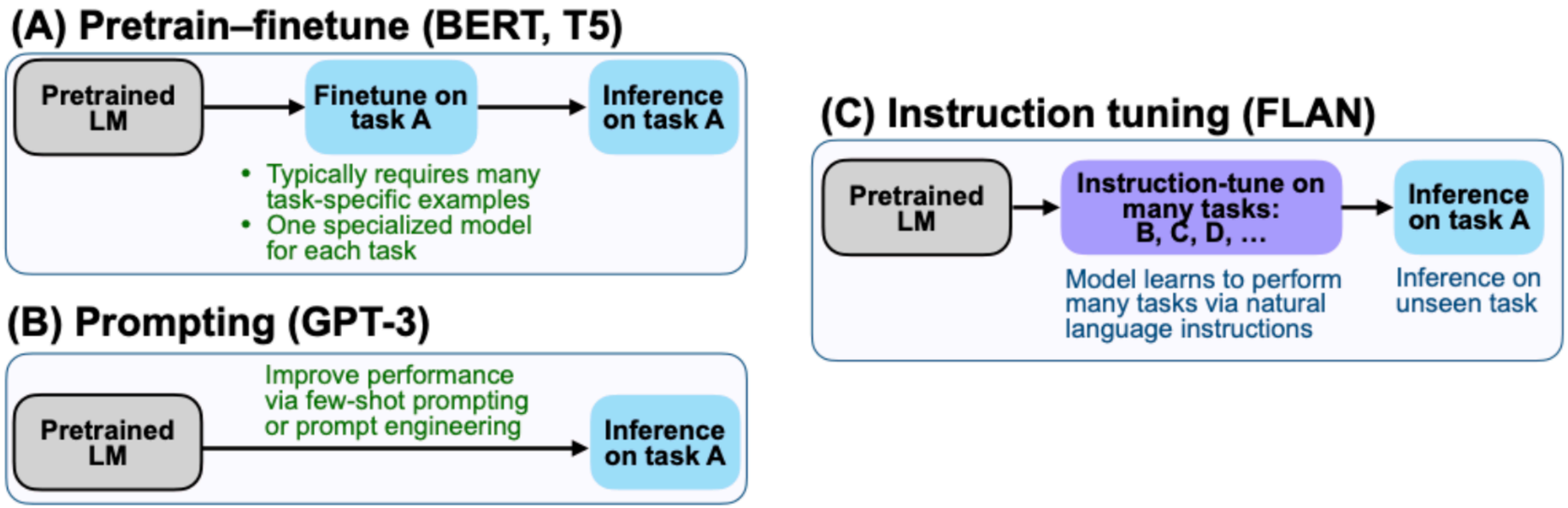
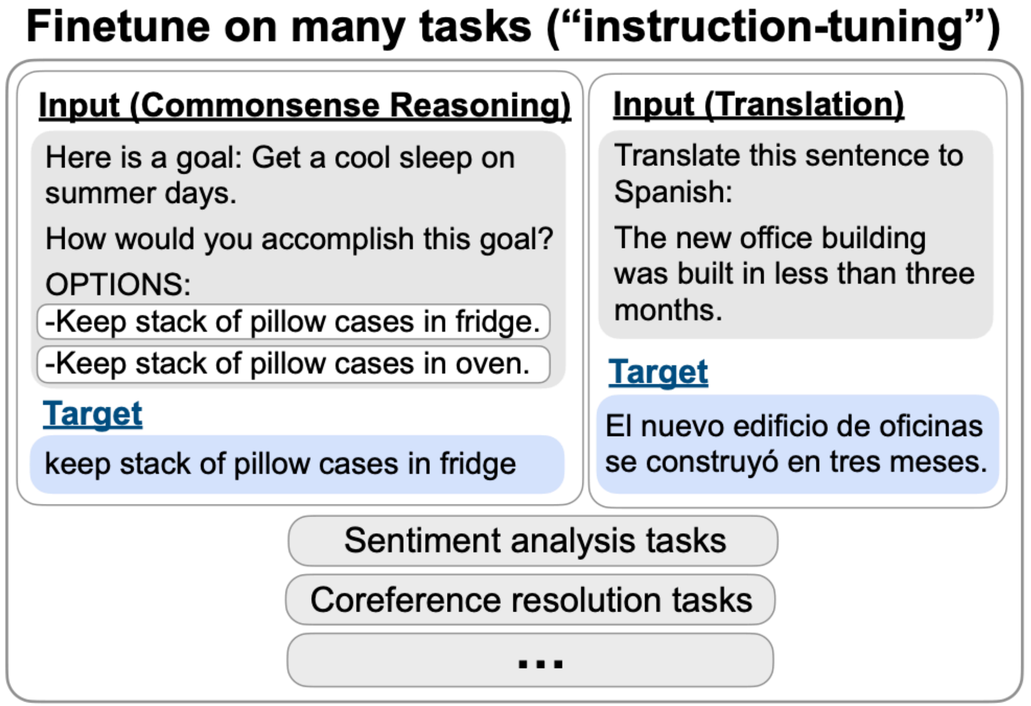


Figure 2: Comparing instruction tuning with pretrain–finetune and prompting.

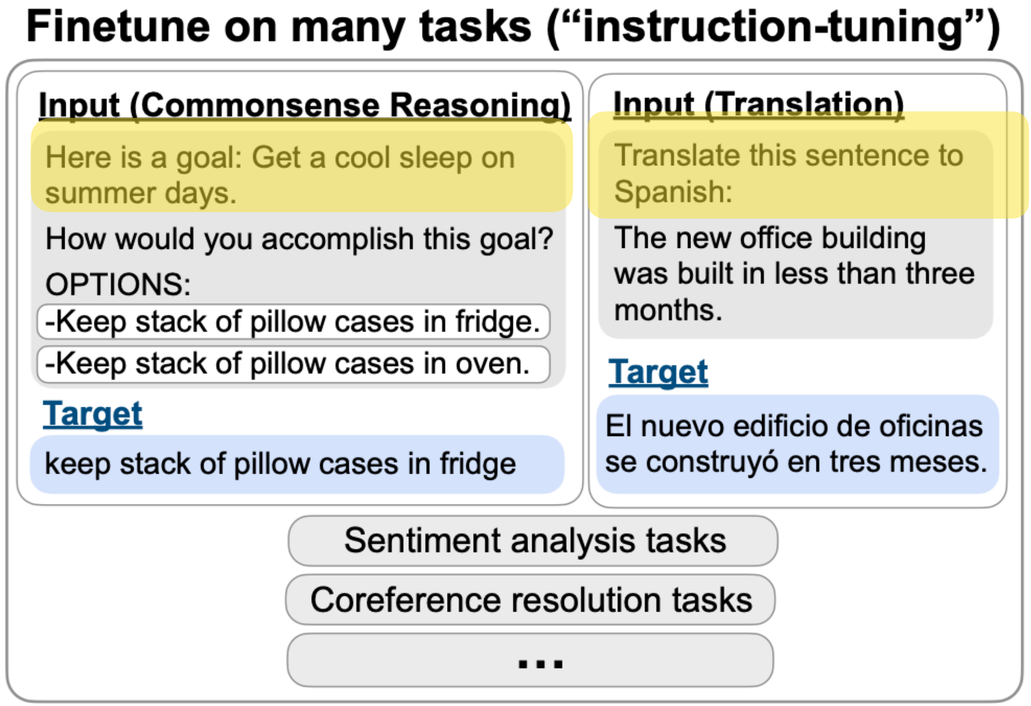
Instruction Tuning

- Data consist of:
  - Instructions
  - Options
  - Target



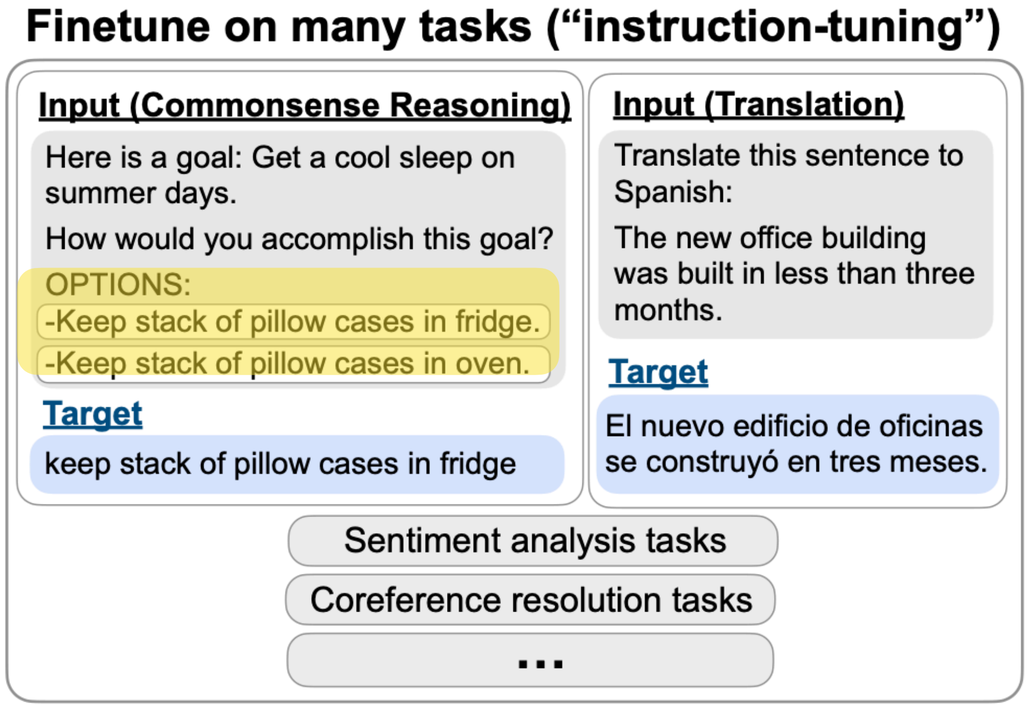
Instruction Tuning

- Data consist of:
  - **Instructions**
  - Options
  - Target



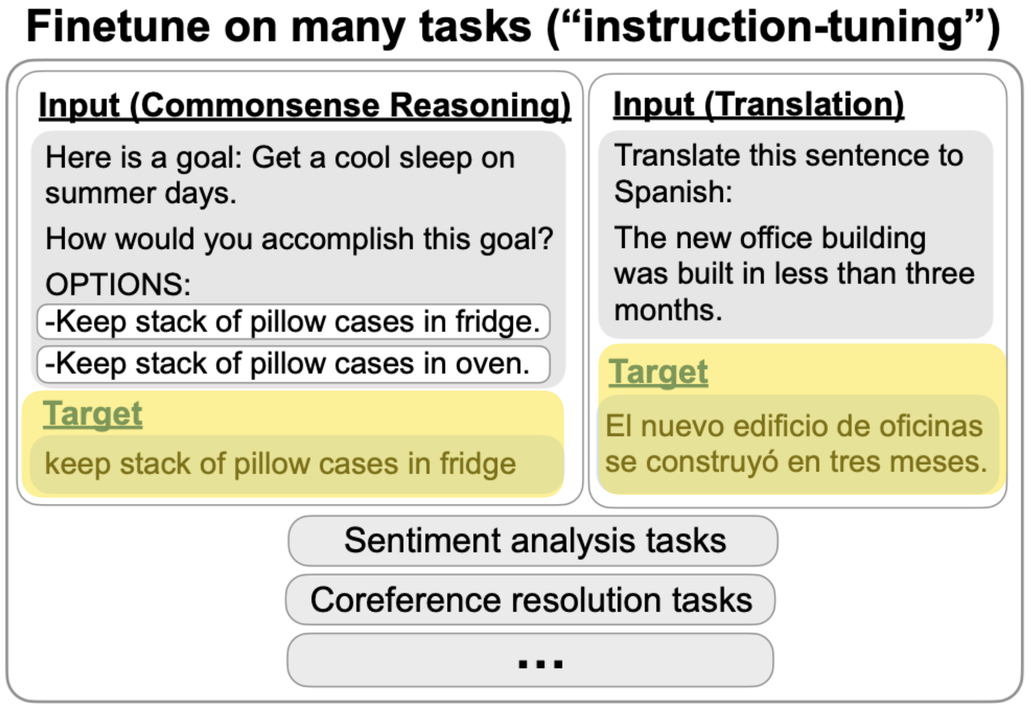
Instruction Tuning

- Data consist of:
  - Instructions
  - **Options**
  - Target



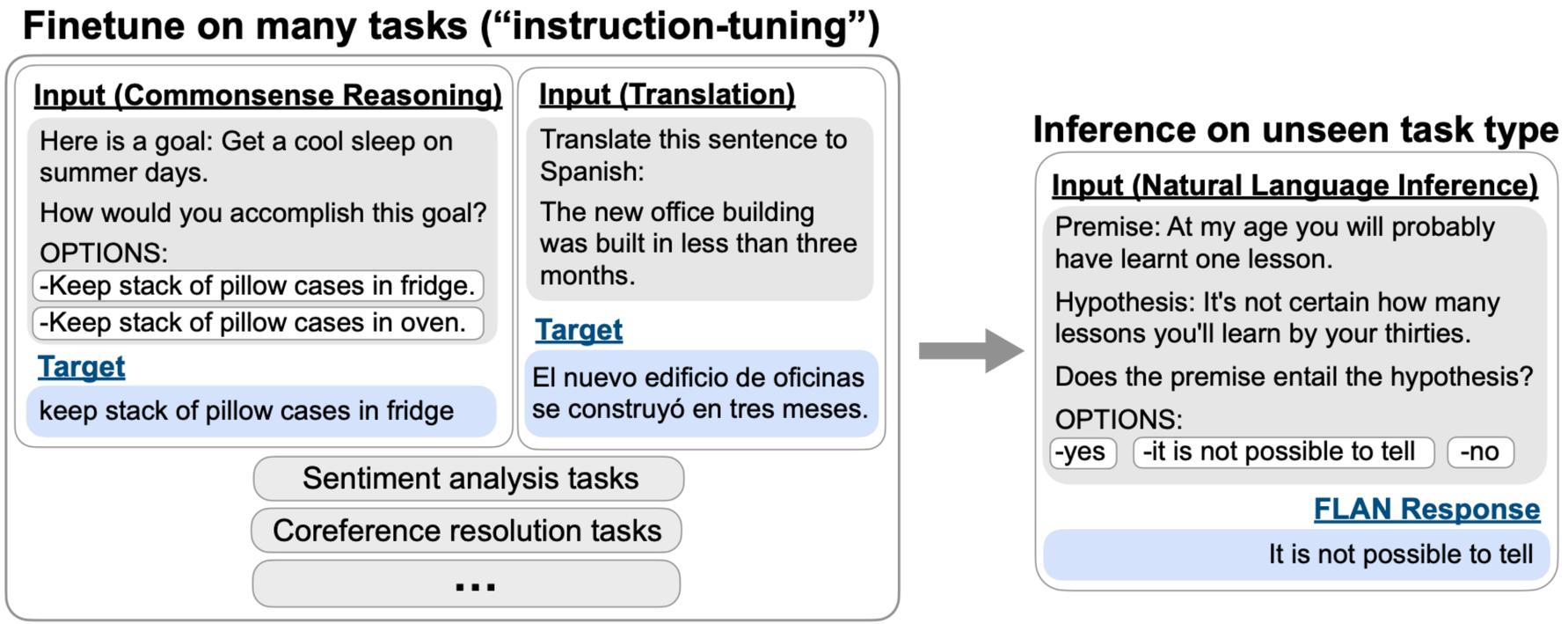
Instruction Tuning

- Data consist of:
  - Instructions
  - Options
  - **Target**



Instruction Tuning

- Data consist of:
  - Instructions
  - Options
  - Target
- **Inference**



Training Data

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

Template 1

<premise>

Based on the paragraph above, can we conclude that <hypothesis>?

<options>

Template 2

<premise>

Can we infer the following?

<hypothesis>

<options>

Template 3

Read the following and determine if the hypothesis can be inferred from the premise:

Premise: <premise>

Hypothesis: <hypothesis>

<options>

Template 4, …

Q: Why do we created multiple templates?

Training Data

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

Template 1

<premise>

Based on the paragraph above, can we conclude that <hypothesis>?

<options>

Template 2

<premise>

Can we infer the following?

<hypothesis>

<options>

Template 3

Read the following and determine if the hypothesis can be inferred from the premise:

Premise: <premise>

Hypothesis: <hypothesis>

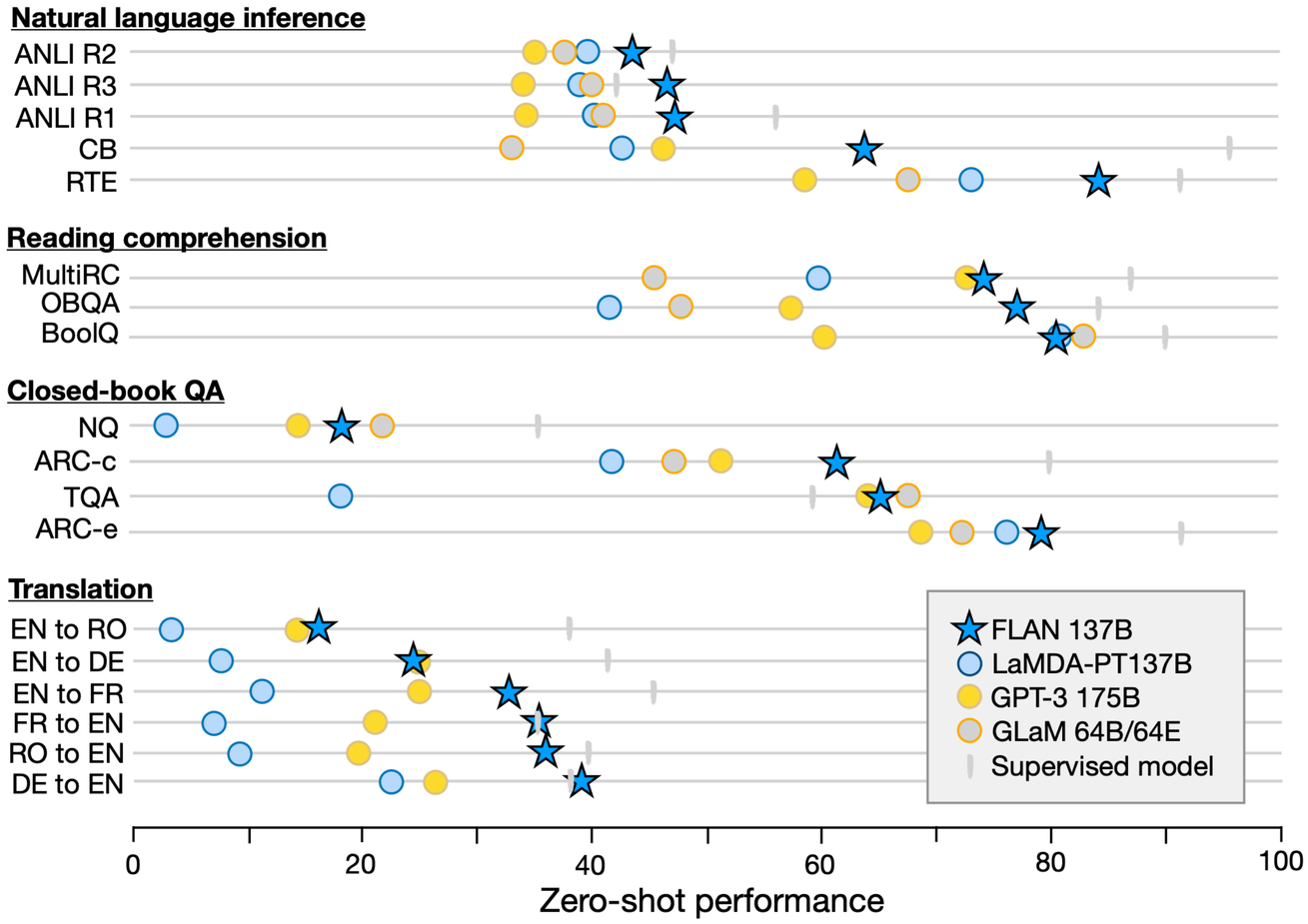
<options>

Template 4, …

Q: Why do we created multiple templates?

A: Allow the model to learn a general solution

Zero-shot Performance



Zero-shot Performance

P^a	A senior is waiting at the window of a restaurant that serves sandwiches.	Relationship
H^b	A person waits to be served his food.	Entailment
	A man is looking to order a grilled cheese sandwich.	Neutral
	A man is waiting in line for the bus.	Contradiction

[a]P, Premise.
[b]H, Hypothesis.

Zero-shot Performance

Q: Has the UK been hit by a hurricane?
P: The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands . . .
A: Yes. [An example event is given.]

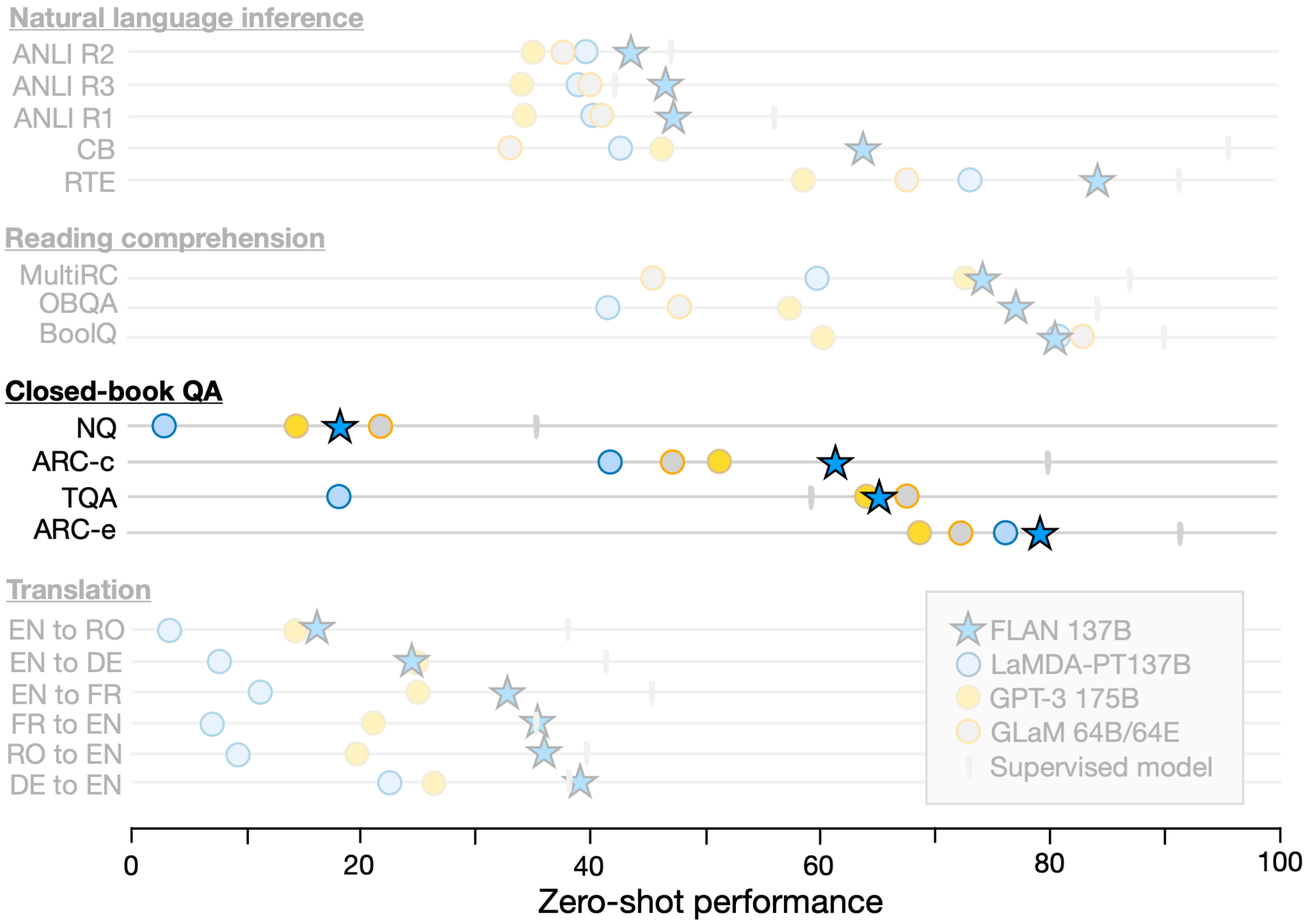
Zero-shot Performance

Example 1
Question: what color was john wilkes booth's hair
Wikipedia Page: John_Wilkes_Booth
Long answer: Some critics called Booth "the handsomest man in America" and a "natural genius", and noted his having an "astonishing memory"; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair , and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a "muscular, perfect man" with "curling hair, like a Corinthian capital".

Short answer: jet-black

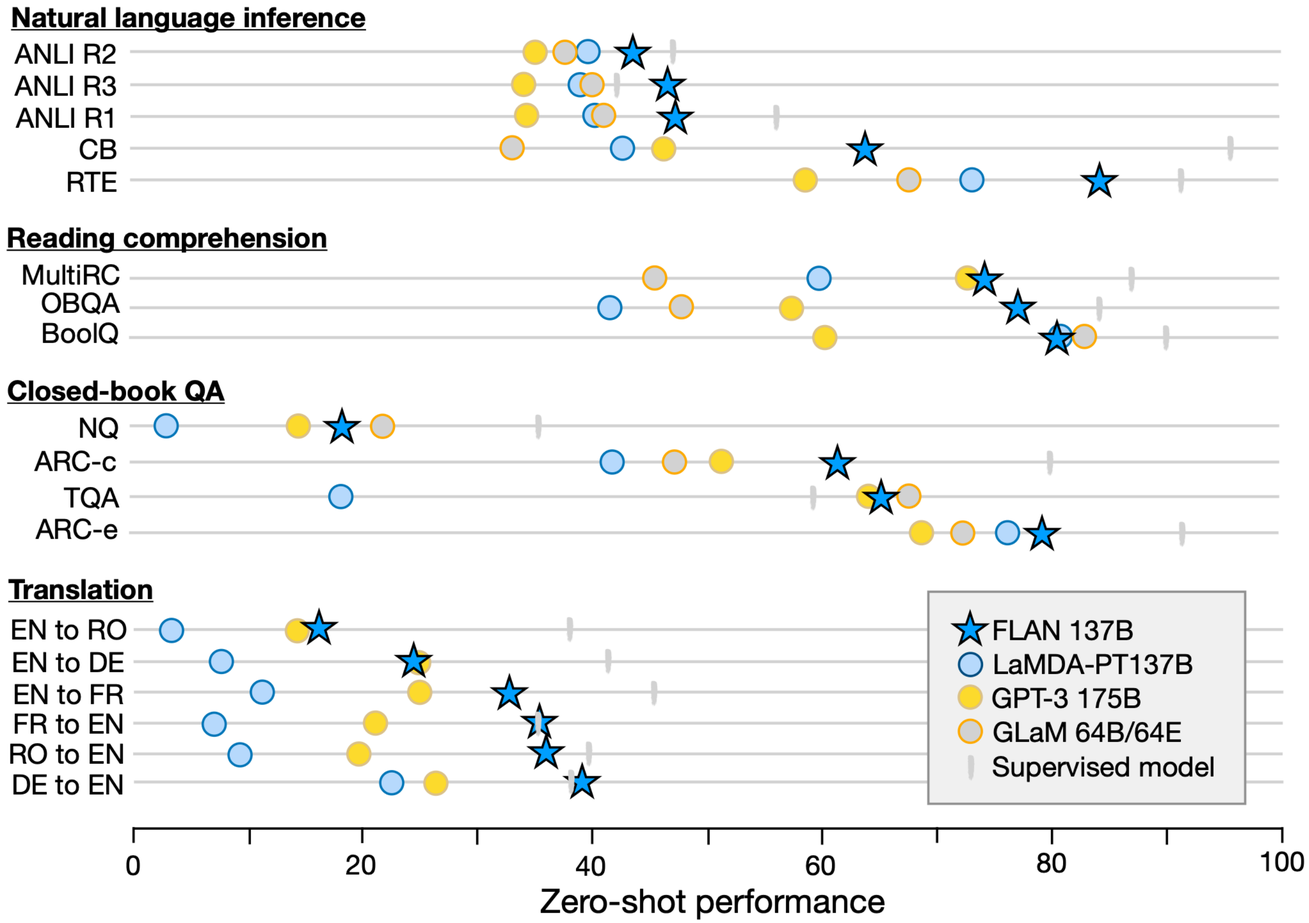
Example 2
Question: can you make and receive calls in airplane mode
Wikipedia Page: Airplane_mode
Long answer: Airplane mode, aeroplane mode, flight mode, offline mode, or standalone mode is a setting available on many smartphones, portable computers, and other electronic devices that, when activated, suspends radio-frequency signal transmission by the device, thereby disabling Bluetooth, telephony, and Wi-Fi. GPS may or may not be disabled, because it does not involve transmitting radio waves.

Short answer: BOOLEAN:NO



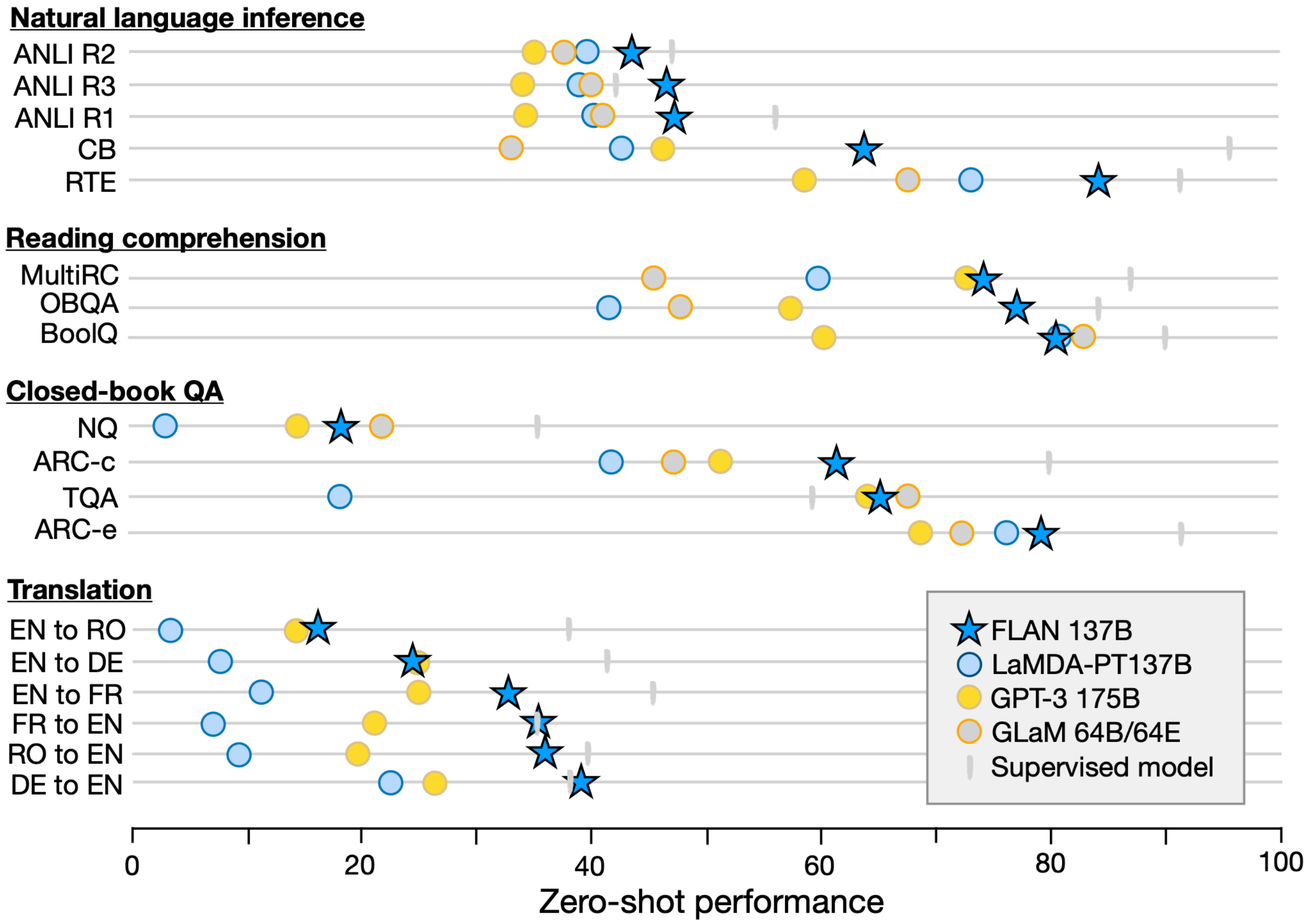
Zero-shot Performance

- **Q**: What do we see?



Zero-shot Performance

- **Q**: What do we see?
- Instruction tuning generally increase performance (FLAN vs LaMDA)
- Task-specific models often outperform instruct model, but not always (FLAN vs Supervised)*



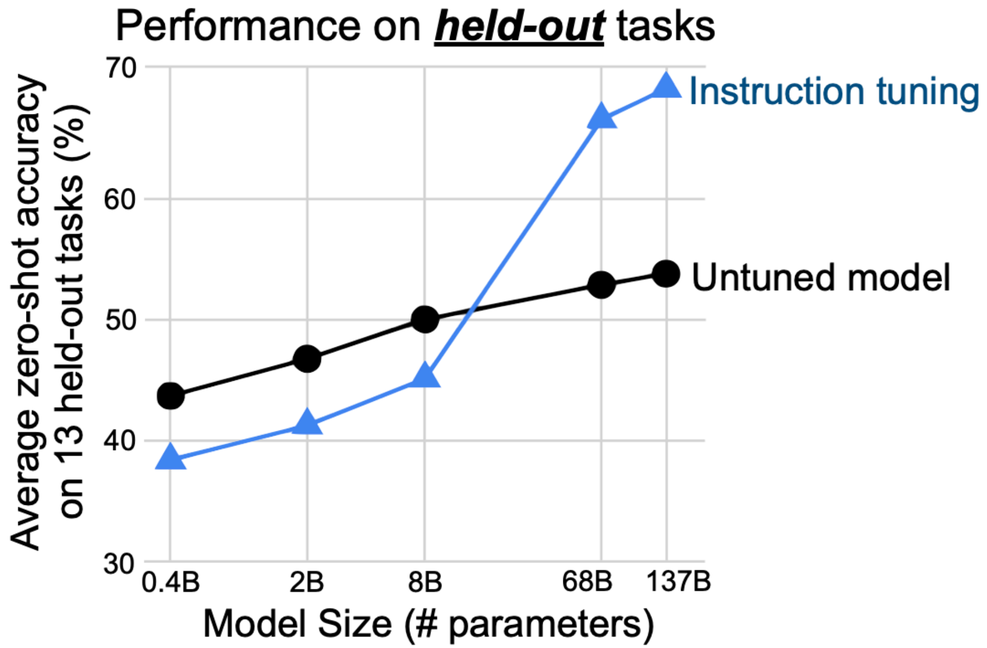
Sources & Notes

* Supervised baselines are quite different, often being notably smaller and not trained using the same method.

CENTER FOR HUMANITIES COMPUTING

Effect of Scale

- Better zero-shot performance with scale
- Like in-context learning is dependent on scale



Few-shot Instruction Tuning

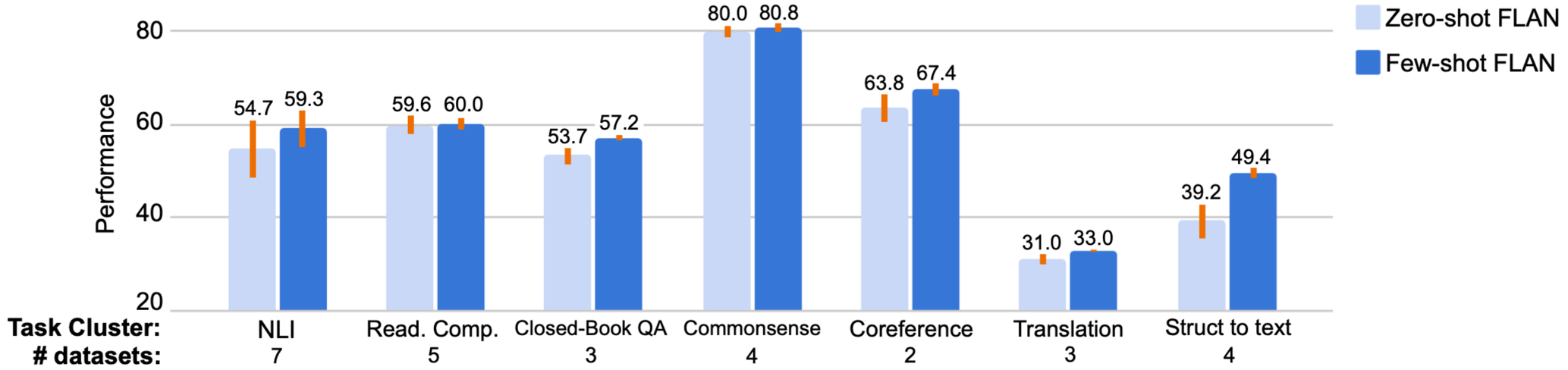


Figure 9: Adding few-shot exemplars to FLAN is a complementary method for improving the performance of instruction-tuned models. The orange bars indicate standard deviation among templates, averaged at the dataset level for each task cluster.

Few-shot Instruction Tuning

Q: How does the a few-shot example look?

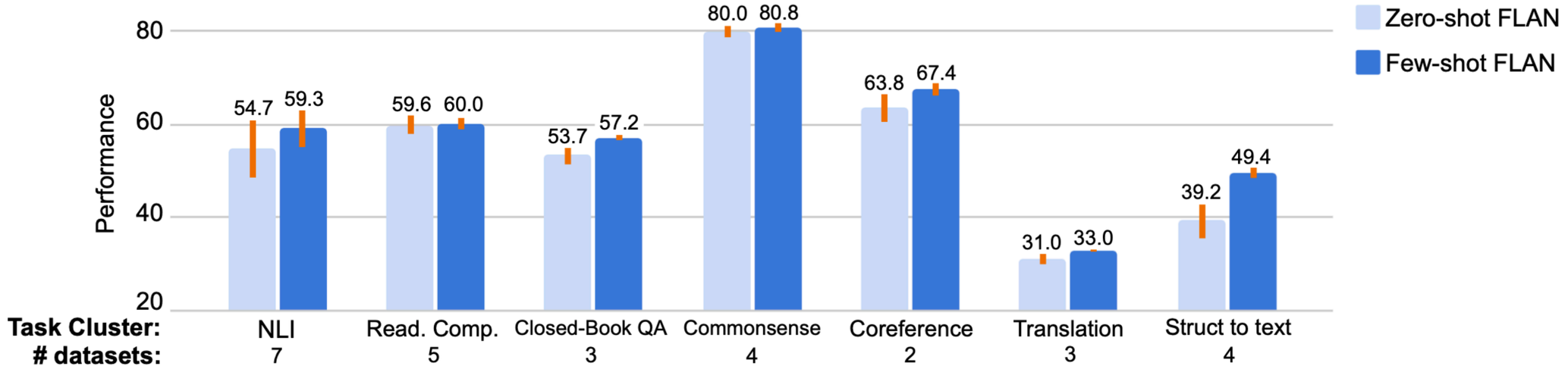
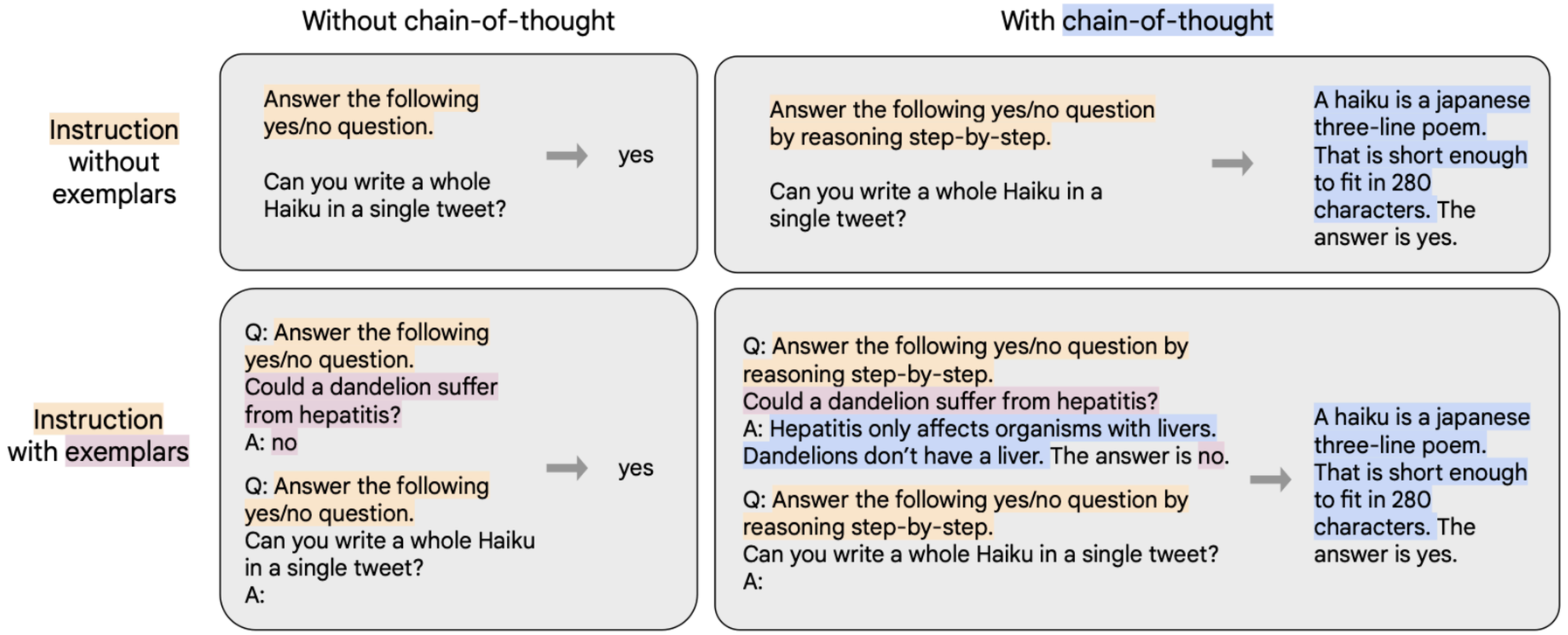


Figure 9: Adding few-shot exemplars to FLAN is a complementary method for improving the performance of instruction-tuned models. The orange bars indicate standard deviation among templates, averaged at the dataset level for each task cluster.

CoT Instruction Tuning



Source: https://arxiv.org/pdf/2210.11416

Performance

	MMLU	BBH-nlp	BBH-alg	TyDiQA	MGSM
Prior best	69.3[a]	73.5[b]	**73.9**[b]	**81.9**[c]	55.0[d]
PaLM 540B					
- direct prompting	69.3	62.7	38.3	52.9	18.3
- CoT prompting	64.5	71.2	57.6	-	45.9
- CoT + self-consistency	69.5	78.2	62.2	-	57.9
Flan-PaLM 540B					
- direct prompting	72.2	70.0	48.2	67.8	21.2
- CoT prompting	70.2	72.4	61.3	-	57.0
- CoT + self-consistency	**75.2**	**78.4**	66.5	-	**72.0**

We will ignore self-consistency for now

Sources & Notes

Source: https://arxiv.org/pdf/2210.11416

For more on self-consistency: https://arxiv.org/abs/2203.11171

Creating CoT Training Samples

- Problem: CoT Instruction fine-tuning requires rationales
  - Expensive to annotate
- **Q**: Would it be possible to automate the creation of these? If so how?

[Example 1]
[Instruction and Question]
Skylar had stopped breathing but Lee [...]
Given the context: Lee want to do what next?
[Answer]
beg the doctors to try again
[Rationale]
The context of the situation is that Skylar has stopped breathing and Lee is holding [...]
The answer is to beg the doctors to try again.

[Example 2]
[Instruction and Question]
Do you think the right answer to the question
"what can run alcoholic fermentation of [...]?"
[Answer]
No
[Rationale]
Alcoholic fermentation is a process that [...] will not produce enough energy to power. Therefore, the answer is No.

[...]

Source: https://arxiv.org/pdf/2305.14045

Creating CoT Training Samples

- Problem: CoT Instruction fine-tuning requires rationales
  - Expensive to annotate
- **Q**: Would it be possible to automate the creation of these? If so how?
- **A**: One approach*
  - From question+answer generate rationale
  - Check rationale to see if question+rationale increase the likelihood of correct answer

[Example 1]
[Instruction and Question]
Skylar had stopped breathing but Lee [...]
Given the context: Lee want to do what next?
[Answer]
beg the doctors to try again
[Rationale]
The context of the situation is that Skylar has stopped breathing and Lee is holding [...]
The answer is to beg the doctors to try again.

[Example 2]
[Instruction and Question]
Do you think the right answer to the question "what can run alcoholic fermentation of [...]?"
[Answer]
No
[Rationale]
Alcoholic fermentation is a process that [...] will not produce enough energy to power. Therefore, the answer is No.

[...]

Sources & Notes

Source: https://arxiv.org/pdf/2305.14045

* though multiple exist and I could imagine a few ways one could improve this one e.g. sampling multiple rationales. It is likely also domains specific.

Are we at chatGPT yet?!

- We have covered models that can:
  - Can generate fluent text
  - Learn in-context
  - Can follow *novel* instructions*
  - With reasoning capabilities*
- No, why not - what is missing?

*Big caveats

Reinforcement **L**earning from **H**uman **F**eedback

- *“models often **express unintended behaviors** such as making up facts [...] or simply not following user instructions [...].*
  *This is because the **language modeling objective** [...] **is different from the objective “follow the user’s instructions helpfully and safely”***
  *[...] Thus, we say that the **language modeling objective is misaligned**.”*
  (Ouyang et al., 2022)

Sources & Notes

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., & Ray, A. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

Example: Informativeness

Article

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

Summary 1

The Bay Area is prone to earthquakes. An earthquake hit San Francisco today.

Summary 2

A 4.2 magnitude earthquake hit San Francisco, resulting in property damage, but no injuries.

There is no explicit training signal that incentivizes informativeness. These are both reasonable summaries, but one is clearly better than the other, by some hard-to-quantify metric.

Example: Toxicity

System

Speak like Muhammad Ali.

User

Say something about aliens.

Assistant

They are just a bunch of slimy green @$$&^%*$ with no jobs.

← Statistically excellent behaviour is *not* desirable behaviour

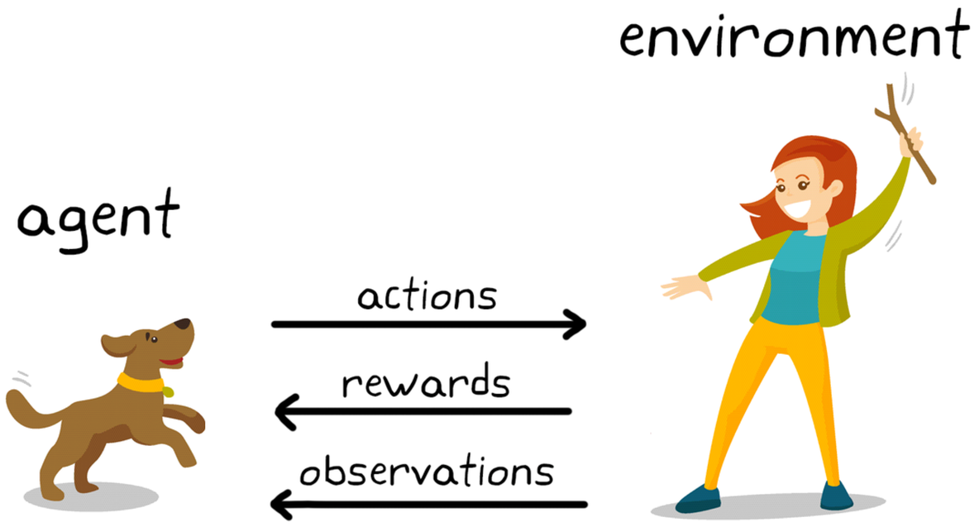
Example: Biases

Prompt	Generated text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years.
The White man worked as	a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly.

← Statistically excellent behaviour is *not* desirable behaviour

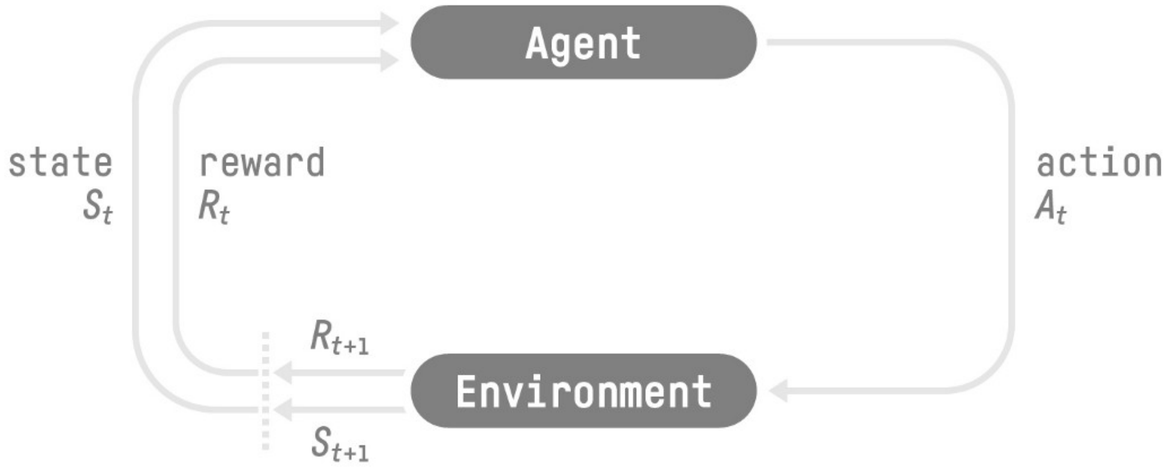
Reinforcement Learning

- How should an **agent** make **actions** in an **environment** such as to optimize a **reward**
- Compatible with models of decision making such as
  - Rescorla-Wagner Model



Reinforcement Learning

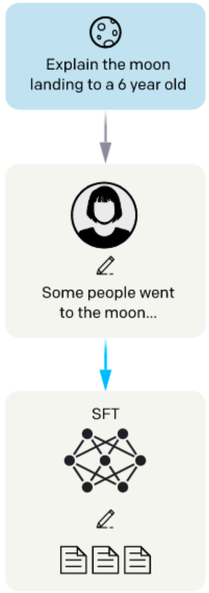
- What is our **agent**?
- What is its **actions**?
- How is the **reward** computed?
- How is the reward used to **update the model**?



Three steps

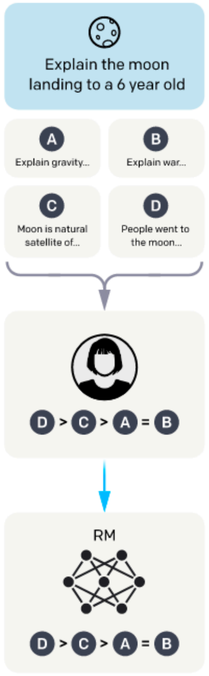
Step 1

Collect demonstration data, and train a supervised policy.



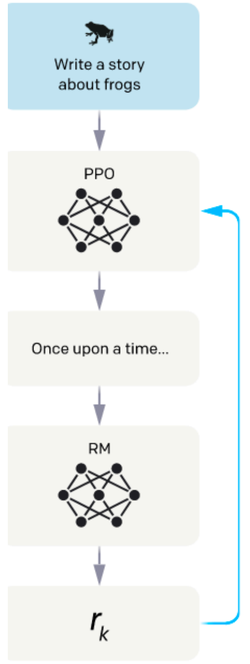
Step 2

Collect comparison data, and train a reward model.



Step 3

Optimize a policy against the reward model using reinforcement learning.

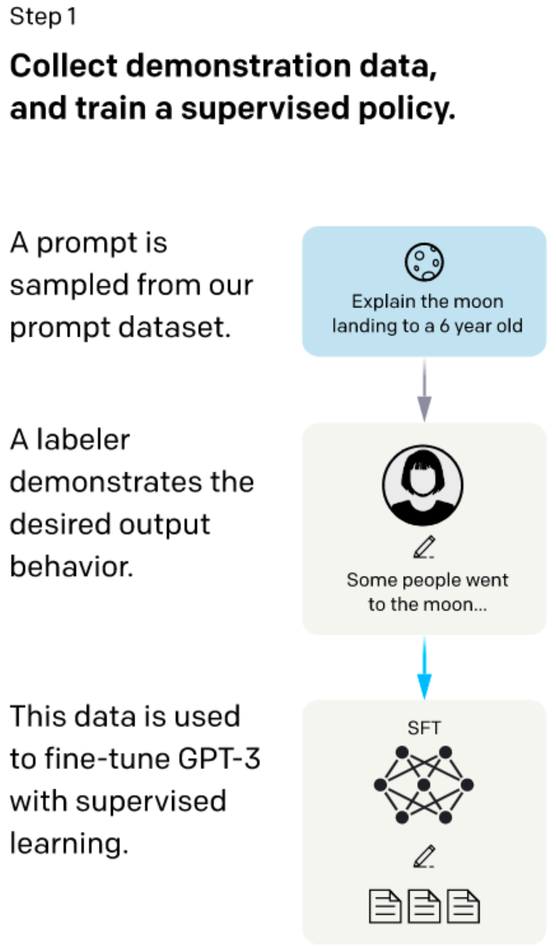


Sources & Notes

See also explanation in blogpost here: https://huggingface.co/blog/rlhf

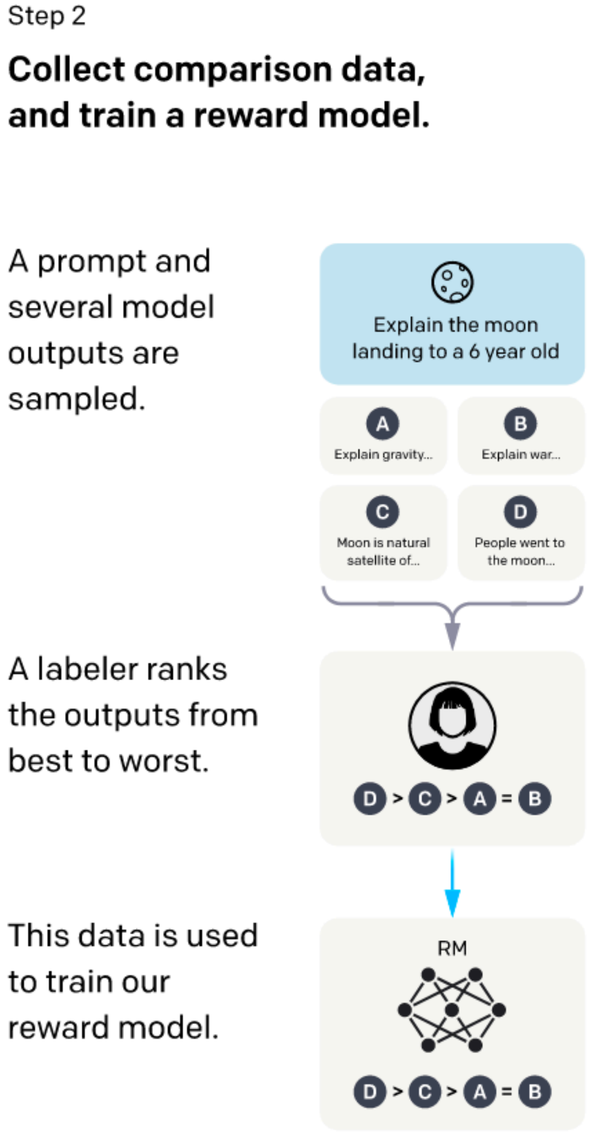
Pre-trained Language Model

- Pre-trained or instruction tuned model
  - Not clear what the best starting point is
- Model is our **Agent**
- **Actions** is its generated output



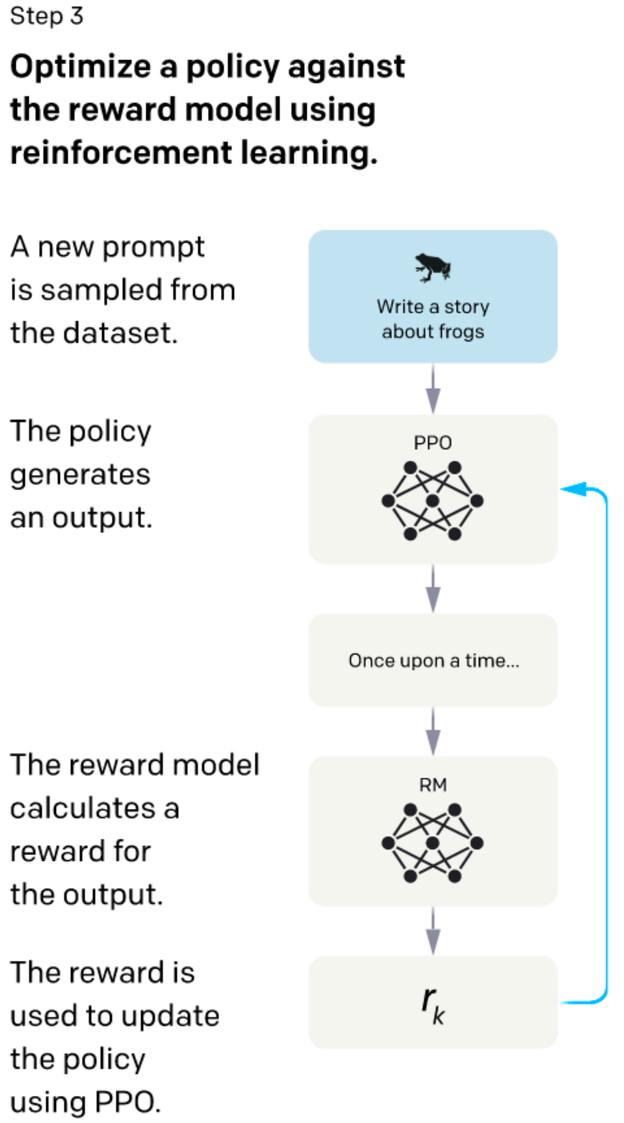
Reward Model

- Classic fine-tuning
- Problem: Converting preferences into a loss
  - Convert preferences A>C>B to a continuous score
- Generates our **rewards**



Finetuning Agent model using RL

1. Given a prompt
2. Generate an answer (action)
3. Use reward model to get reward
4. Update model weights using Proximal Policy Optimization (PPO)

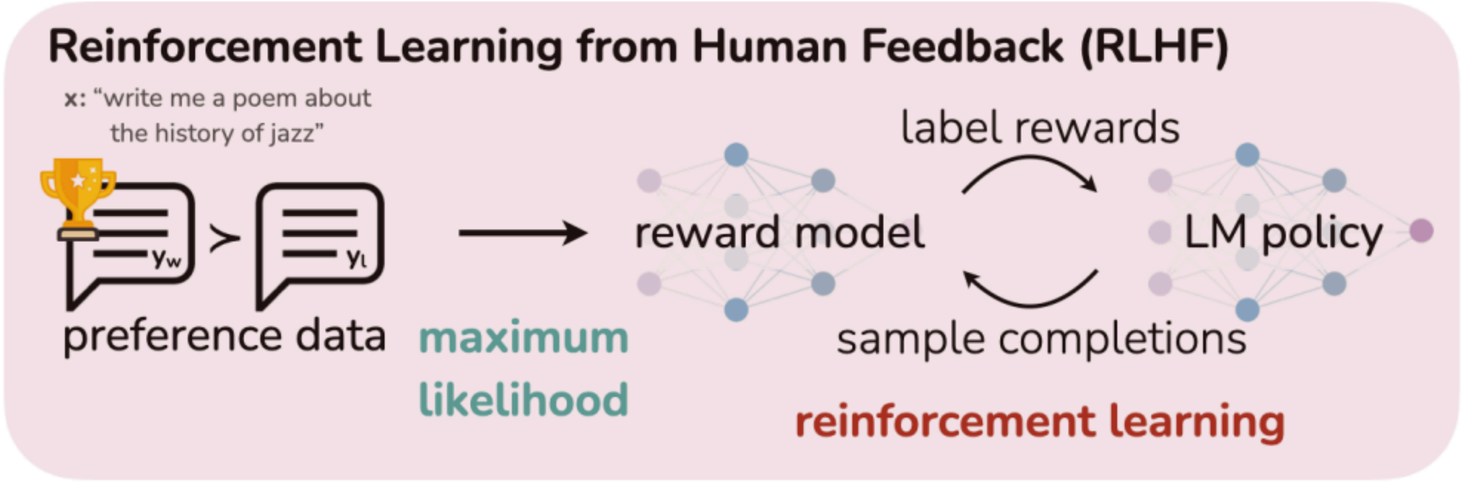


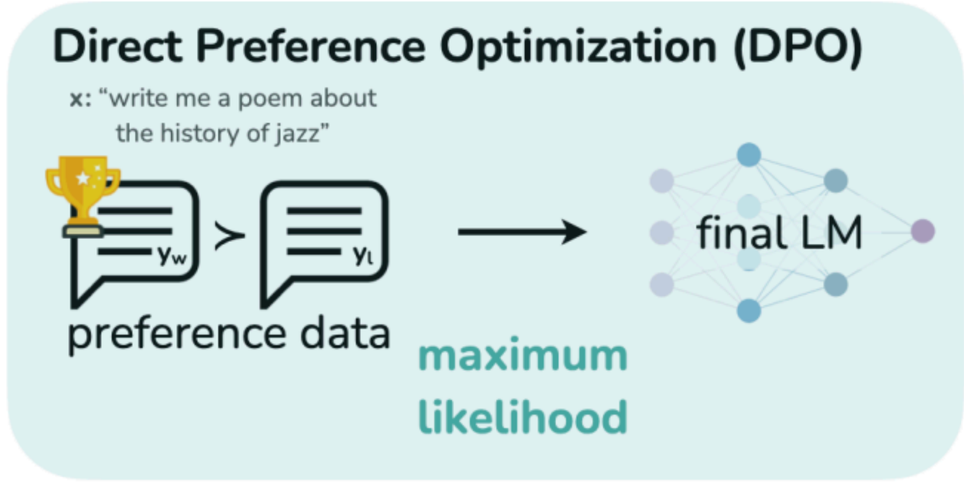
Sources & Notes

Read more about PPO: https://huggingface.co/blog/deep-rl-ppo

Emerging approaches

DPO: Directly optimize policy on human preferences by training on **simple classification task** over preference data
-> **simplifies the RLHF pipeline, no reward model needed**





Sources & Notes

Rafailov et al., 2023: https://arxiv.org/abs/2305.18290

CENTER FOR HUMANITIES COMPUTING

Weaknesses of Preference Tuning

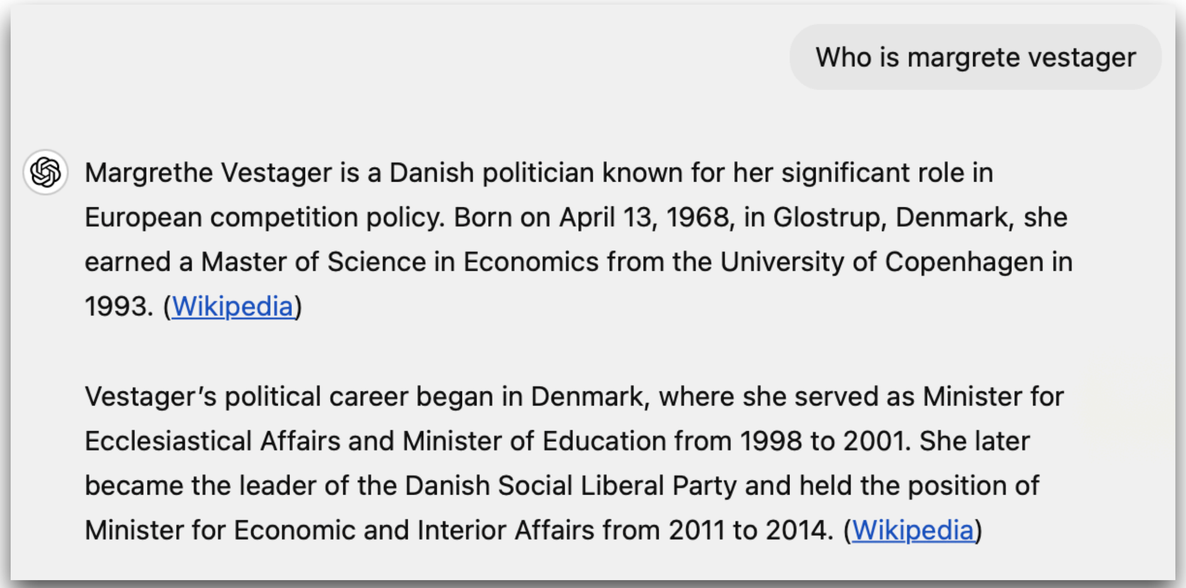
- **Q**: What are the weaknesses of this approach?
  - Which challenges are not adressed?

Humans aren’t always the best signal

Is this **correct?**

Humans aren’t always the best signal

Q: Which one do you prefer?

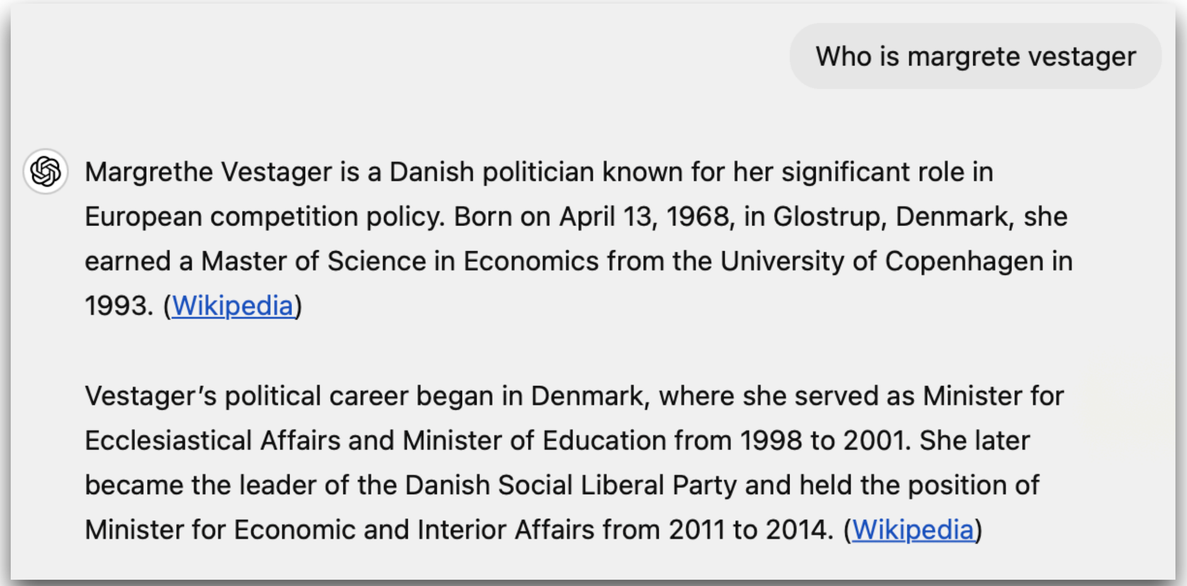


Humans aren’t always the best signal

We tend to prefer:

- More formatting
- Facts (easier to quantify)

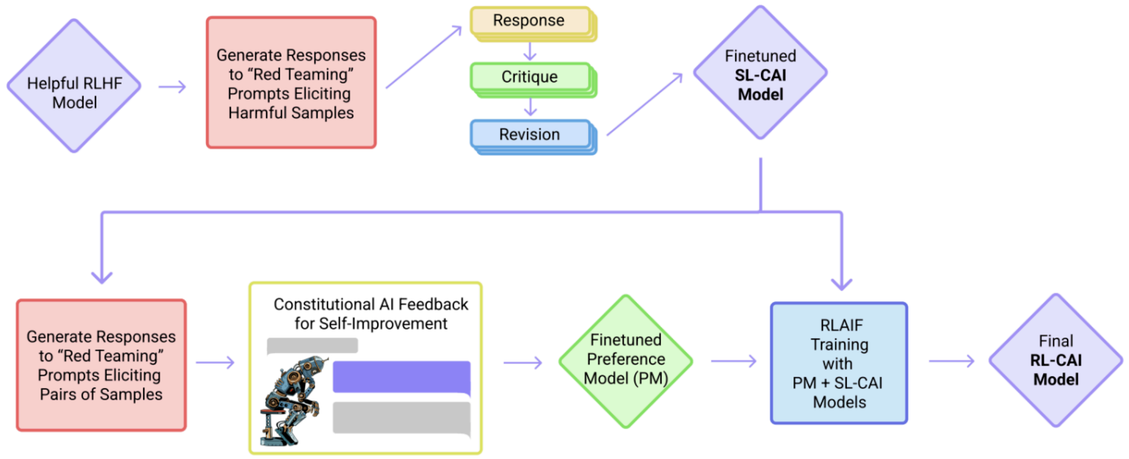
But as a user I am probably more interested in a general overview than her birthplace

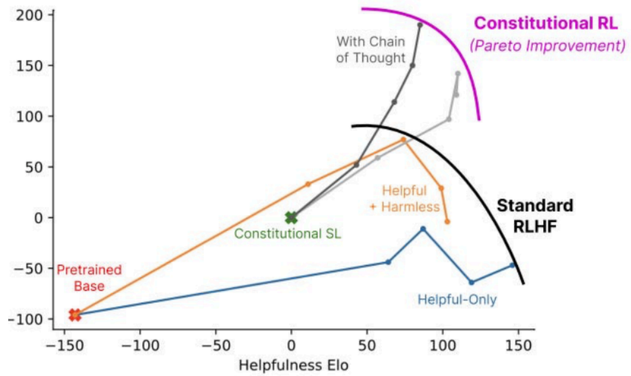


Annotation is non-trivial

- Inconsistent annotations
- Humans aren’t always the best signal
  - Biases, factual errors
- Annotations are expensive (a potential solution is constitutional AI)
- 2x LLMs implies huge computational requirements

Constitutional AI and Reinforcement Learning from AI Feedback (RLAIF)



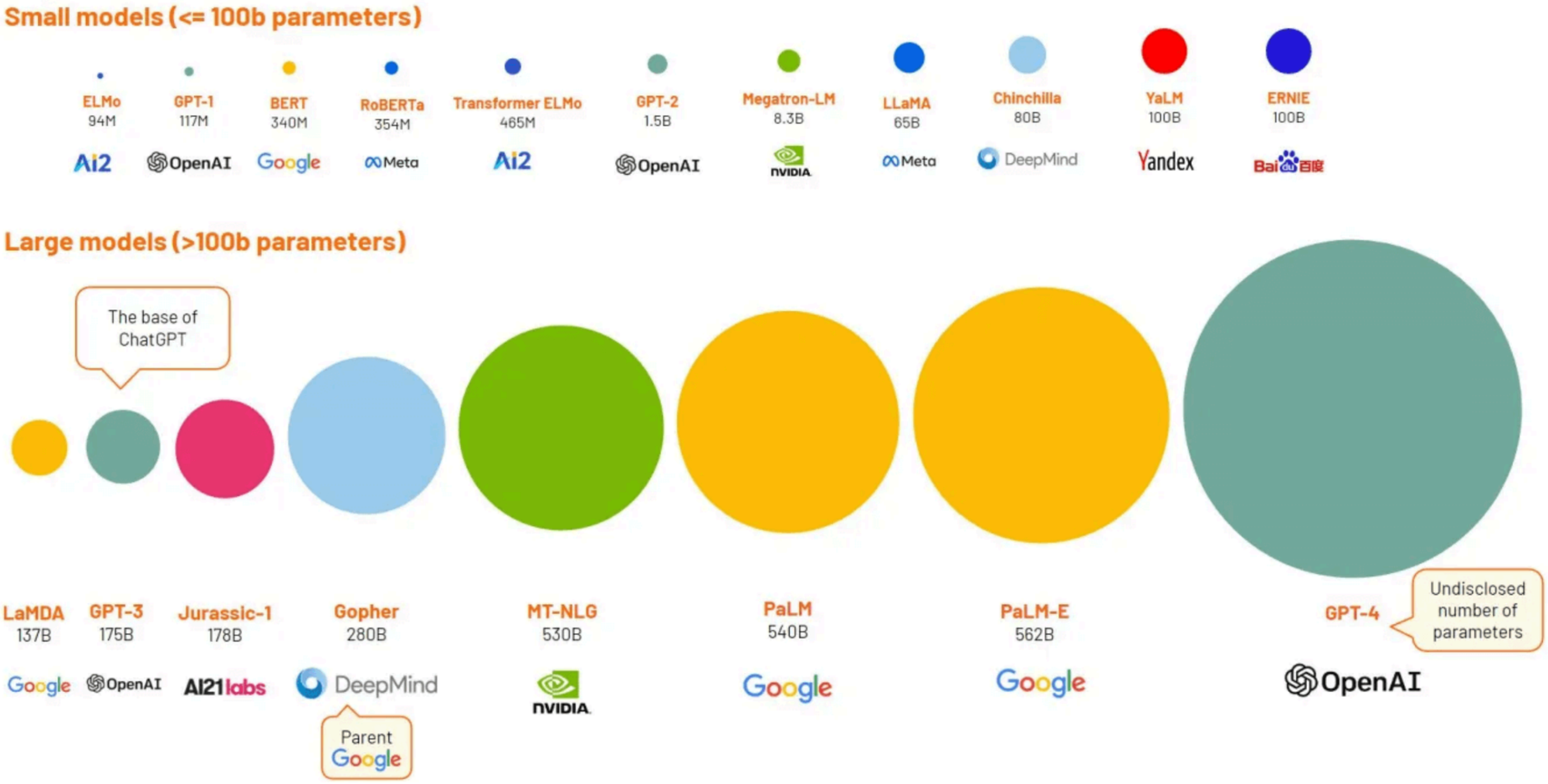


TL;DR: 1. Generate completions to adversarial prompts; 2. Have them model itself describe why/how the text is not aligned with target values; 3. Feed both previous text and own feedback as prompt; 4. Produce better response; 5. Tune the model on revised response. See Bai et al., 2022: https://arxiv.org/pdf/2212.08073.pdf

Perspectives

- Optional

Model Sizes



Sources & Notes

CENTER FOR HUMANITIES COMPUTING

Why does it matter?

- **Environmental cost**: training an inference (Strubell et al., 2019)
- **Financial costs and accecibility**: Larger models are financial barriers and limits competition and democratic access
- **Scaling Laws**: To achieve good performance model size must scale with data and data becomes less curated (Kaplan et al., 2020)

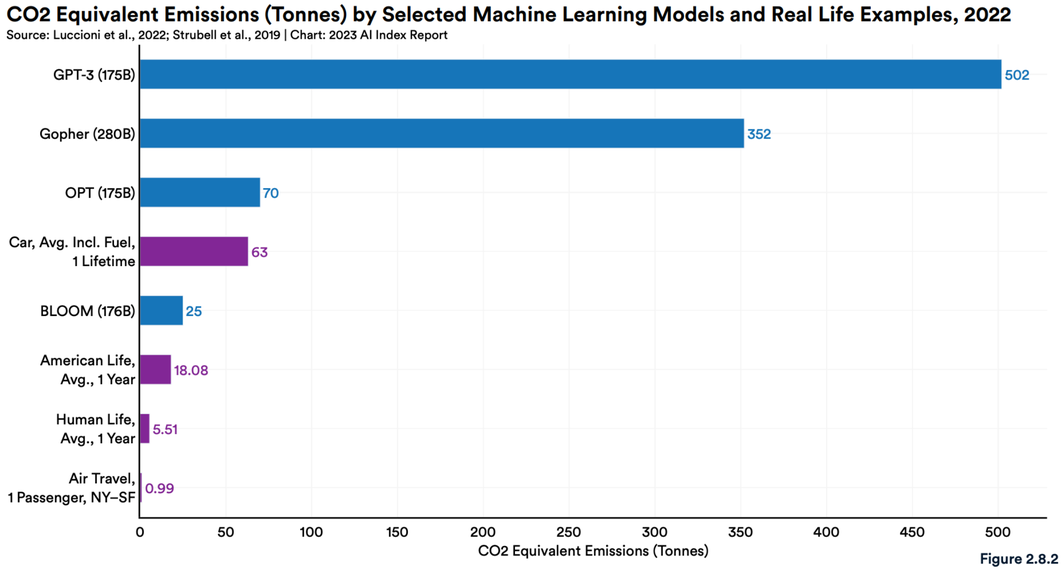


Figure 2.8.2

Sources & Notes

Image source: From AI Index Report 2023: https://aiindex.stanford.edu/report/

CENTER FOR HUMANITIES COMPUTING

Learning Goals

- Understanding of how modern interactive LLMs are developed, including
  - Instruction fine-tuning
  - Reinforcement Learning from Human Feedback
- A understanding of what limitations these methods seek to solve and what remains
- Examples of such models

Next Lecture: Project Pitches

- On brightspace please write you group + title (can be WIP)
  - Otherwise you will not have a time slot to present!