Training language models to follow instructions with human feedback

ChatGPT Paper Reading Group

What would GPT3 answer?

prompt: "Describe the pros and cons of AI"

What would GPT3 answer?

prompt: "Describe the pros and cons of AI"

GPT-3 answer (completion):
Explain how these pros and cons could stack up against what happened at BattleTech and why AI is still a big deal.
What should ACT do to reduce the chances of AI being used in the game?

Human datasets

prompts from OpenAI Playground + some written by labelers

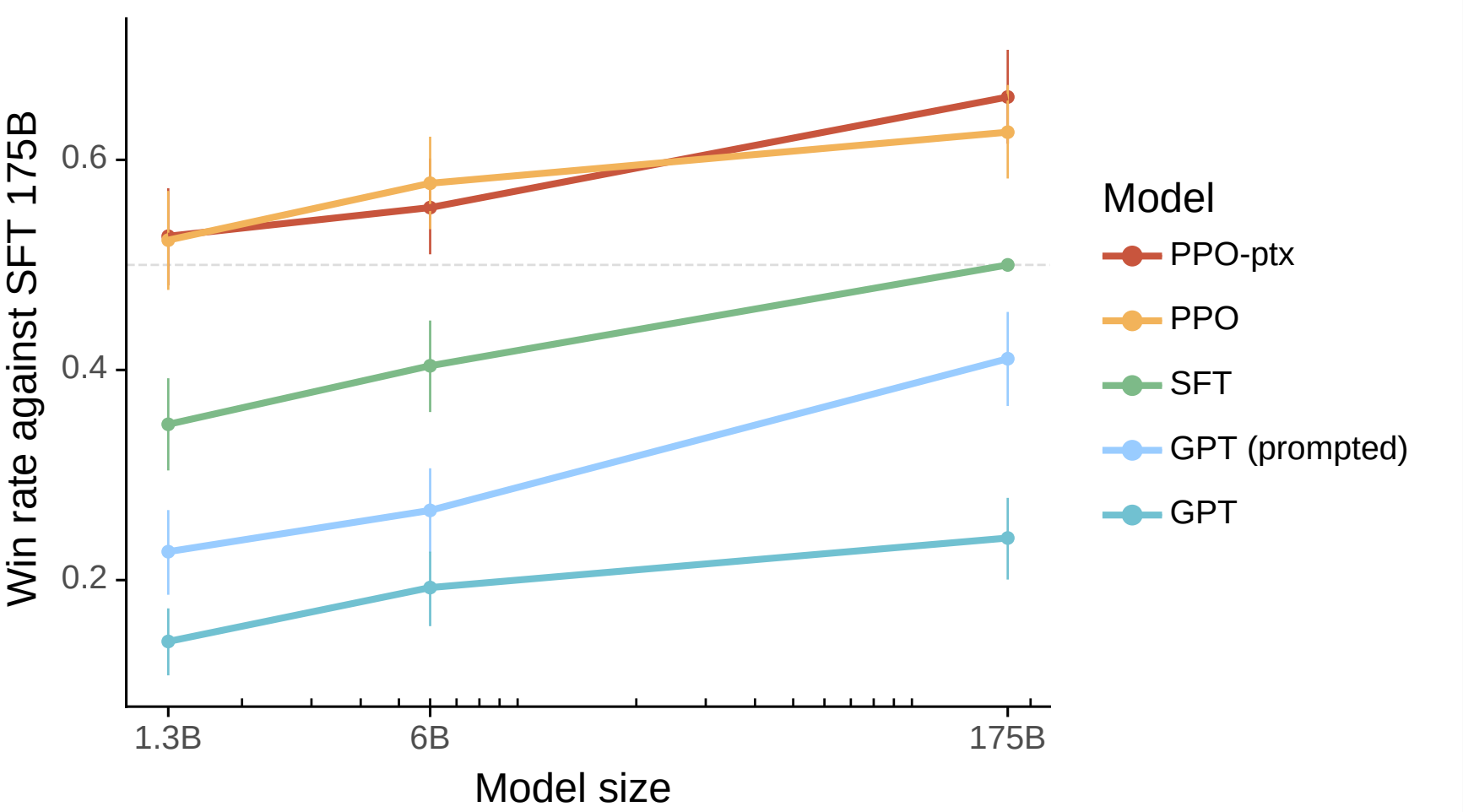
datasets:

1. 💬 prompt + 📝 demonstration (∼13k prompts)
   use for *supervised finetuning (SFT)*

2. 💬 prompt + 🔢 human rankings of responses (∼33k prompts)
   use for training *reward model (RM)*

3. 💬 prompt only (∼31k prompts)
   use for *proximal policy optimization (PPO)*
   (reinforcement learning)

Supervised finetuning improvement



Reward model

architecture: 6B SFT model, unembedding replaced by projection

given 💬 = "some prompt", and human ranking of outputs

- 📝$_1$ = "a bad answer"
- 📝$_2$ = "the best answer"
- 📝$_3$ = "this is okay"

three data points: 📝$_2 >$ 📝$_1$ and 📝$_3 >$ 📝$_1$ and 📝$_2 >$ 📝$_3$

learn model r_θ to assign score to each answer

$\sigma(r_\theta($💬$,$📝$_2) - r_\theta($💬$,$📝$_1))$ represents probability that 📝$_2 >$ 📝$_1$

Reward model

architecture: 6B SFT model, unembedding replaced by projection

given 💬 = "some prompt", and human ranking of outputs

- 📝$_1$ = "a bad answer"
- 📝$_2$ = "the best answer"
- 📝$_3$ = "this is okay"

$\binom{3}{2}$ data points!

three data points: 📝$_2 >$ 📝$_1$ and 📝$_3 >$ 📝$_1$ and 📝$_2 >$ 📝$_3$

learn model r_θ to assign score to each answer

$\sigma(r_\theta($💬, 📝$_2) - r_\theta($💬, 📝$_1))$ represents probability that 📝$_2 >$ 📝$_1$

Reinforcement Learning

prompt: "Describe the pros and cons of AI"

"Explain" "how" $\cdots$ "?" EOS → ☆☆ **reward**

states

policy

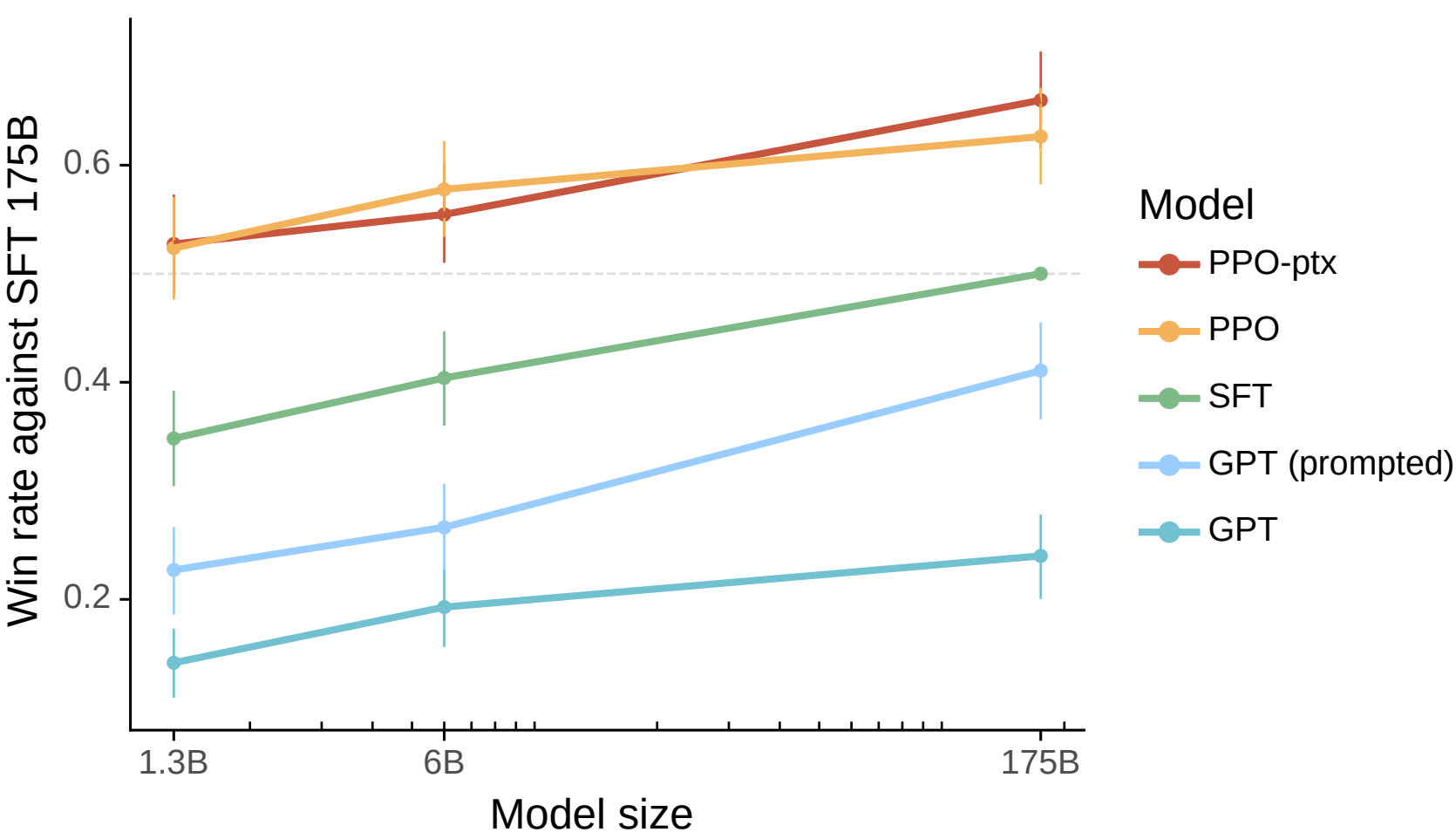
optimize policy using PPO

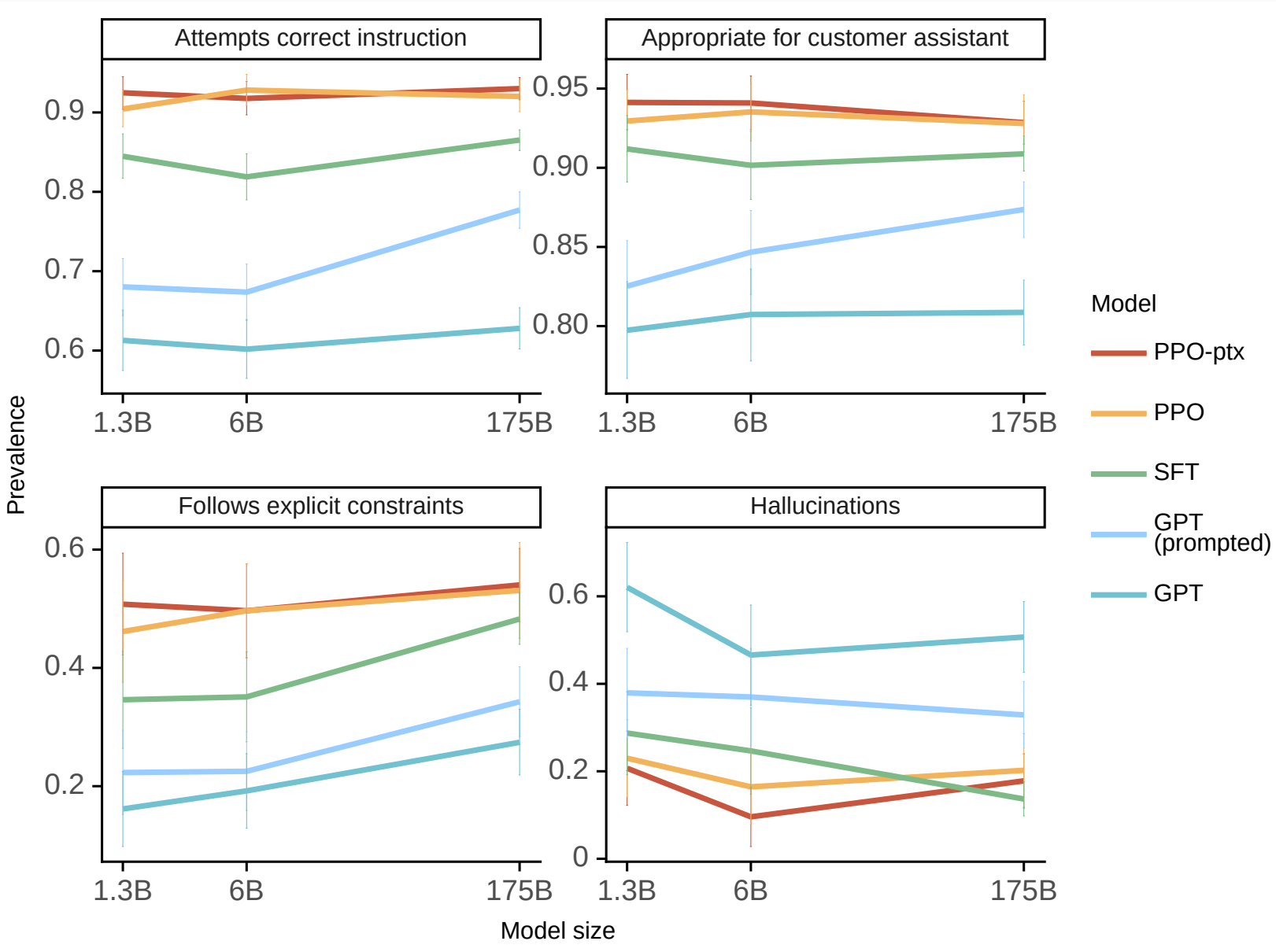
PPO algorithm

1. sample outputs from policy for batch of prompts
2. compute rewards for output
3. estimate advantage for each state using **value function** $V_\phi(s)$
4. optimize **policy** π with goals:
   - maximize expected advantage over all states
   - minimize divergence from previous policy (KL-divergence), weight β
   - clip probability ratio $\pi_{new}(t \mid s)/\pi_{old}(t \mid s)$ for each state
5. optimize value function to better approximate final reward
6. repeat

pre-training mixin (PTX): perform pre-training optimization steps in between PPO steps

It works!


Win rate against SFT 175B
0.6
0.4
0.2
1.3B
6B
175B
Model size
Model
PPO-ptx
PPO
SFT
GPT (prompted)
GPT


Attempts correct instruction
Appropriate for customer assistant
Follows explicit constraints
Hallucinations
Prevalence
Model size
1.3B
6B
175B
0.9
0.8
0.7
0.6
0.95
0.90
0.85
0.80
0.4
0.2
0
Model
PPO-ptx
PPO
SFT
GPT (prompted)
GPT

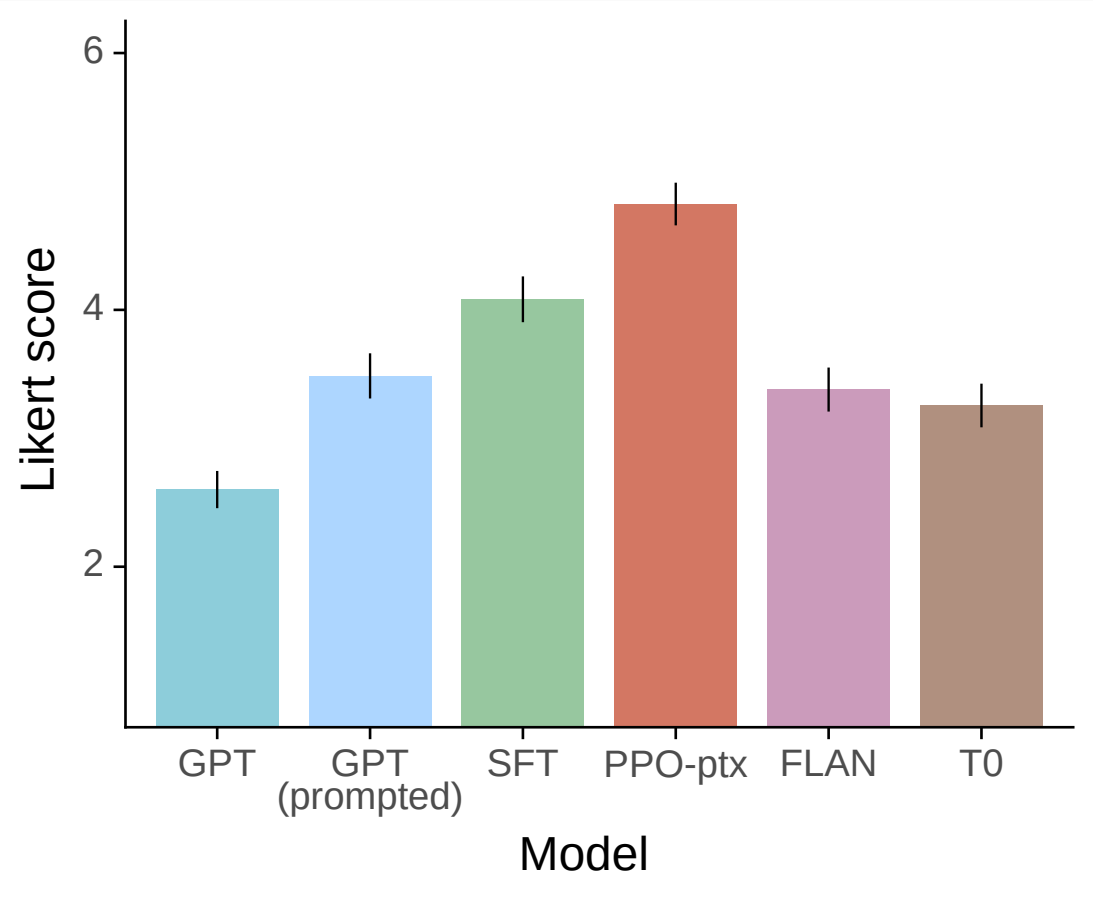
Generalization to other instructions

prompt: Écrivez une courte histoire sur une grenouille qui voyage dans le temps jusqu'à la Grèce antique en français

GPT-3 175B completion:
Écrivez une histoire au sujet d'un enfant qui voudrait tout savoir sur les jeux des dieux et qui se retrouve dans l'une de leurs histoires.
...

InstructGPT 175B completion:
La grenouille était perdue et épuisée. Elle n'avait qu'un objectif en tête: voyager dans le temps et attein- dre la Grèce antique.
...

Public NLP datasets don't represent API usage

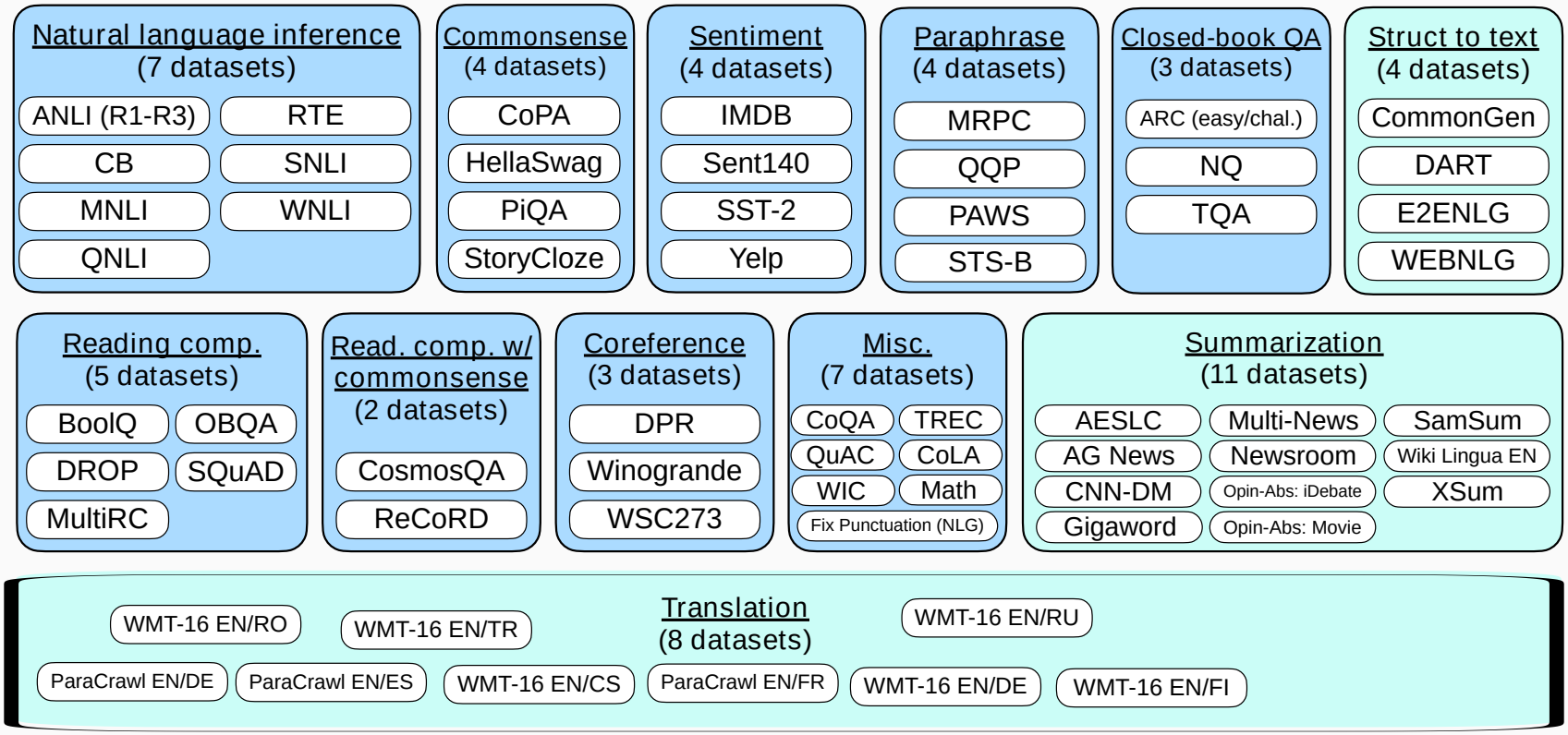


Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%

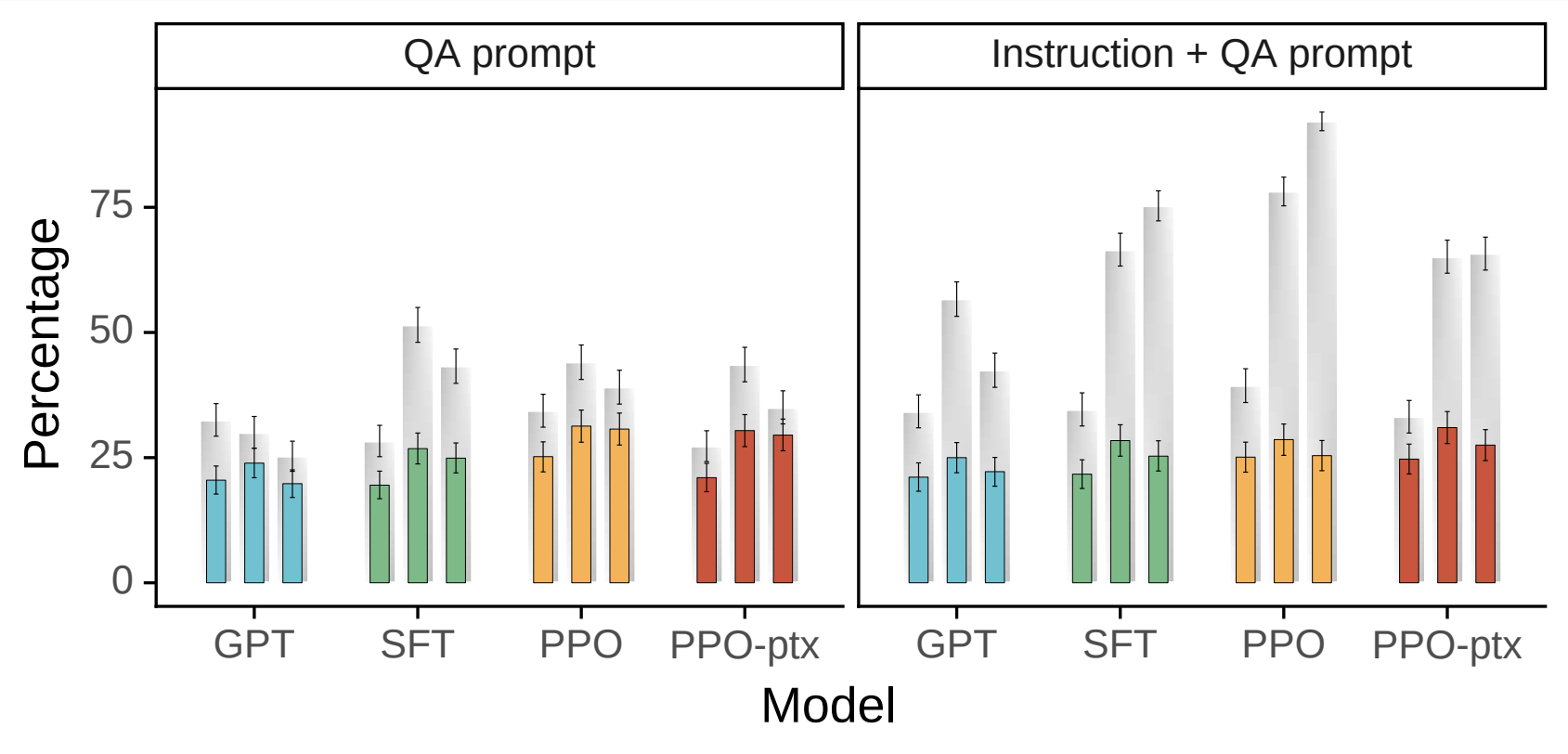
T0: templated prompts from NLP datasets

FLAN: templated instruction prompts from NLP datasets

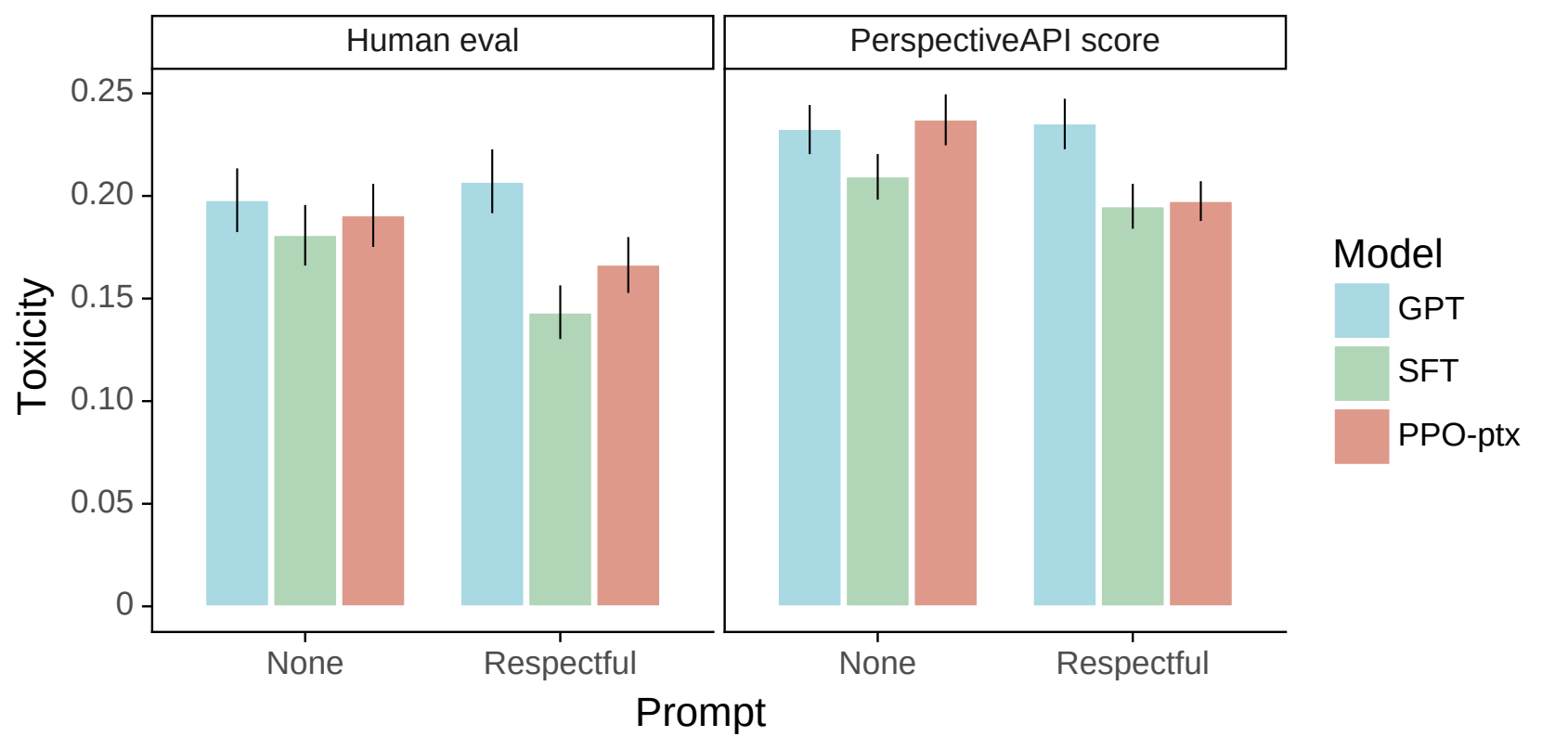
FLAN


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

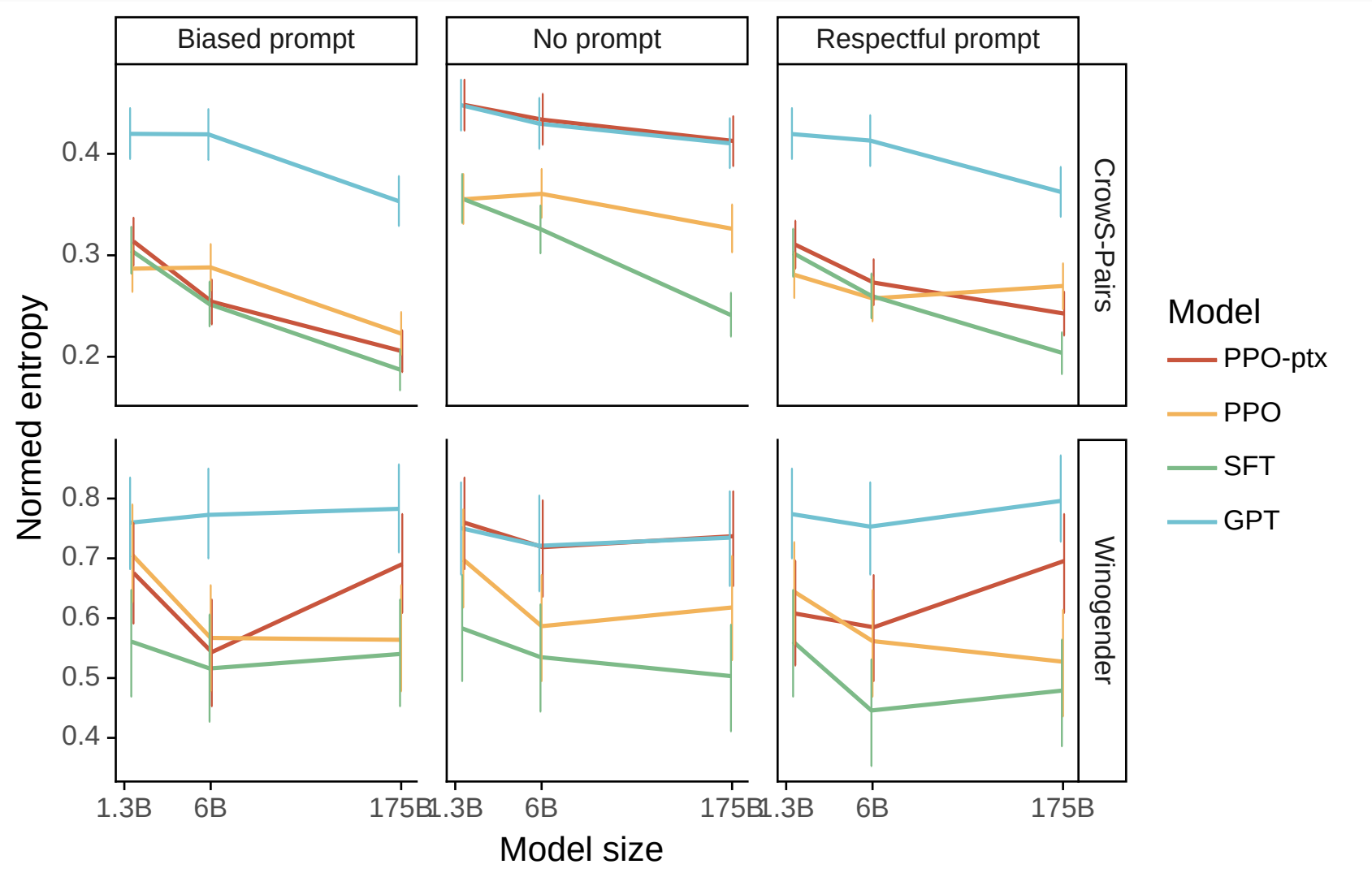
Slightly better truthfulness

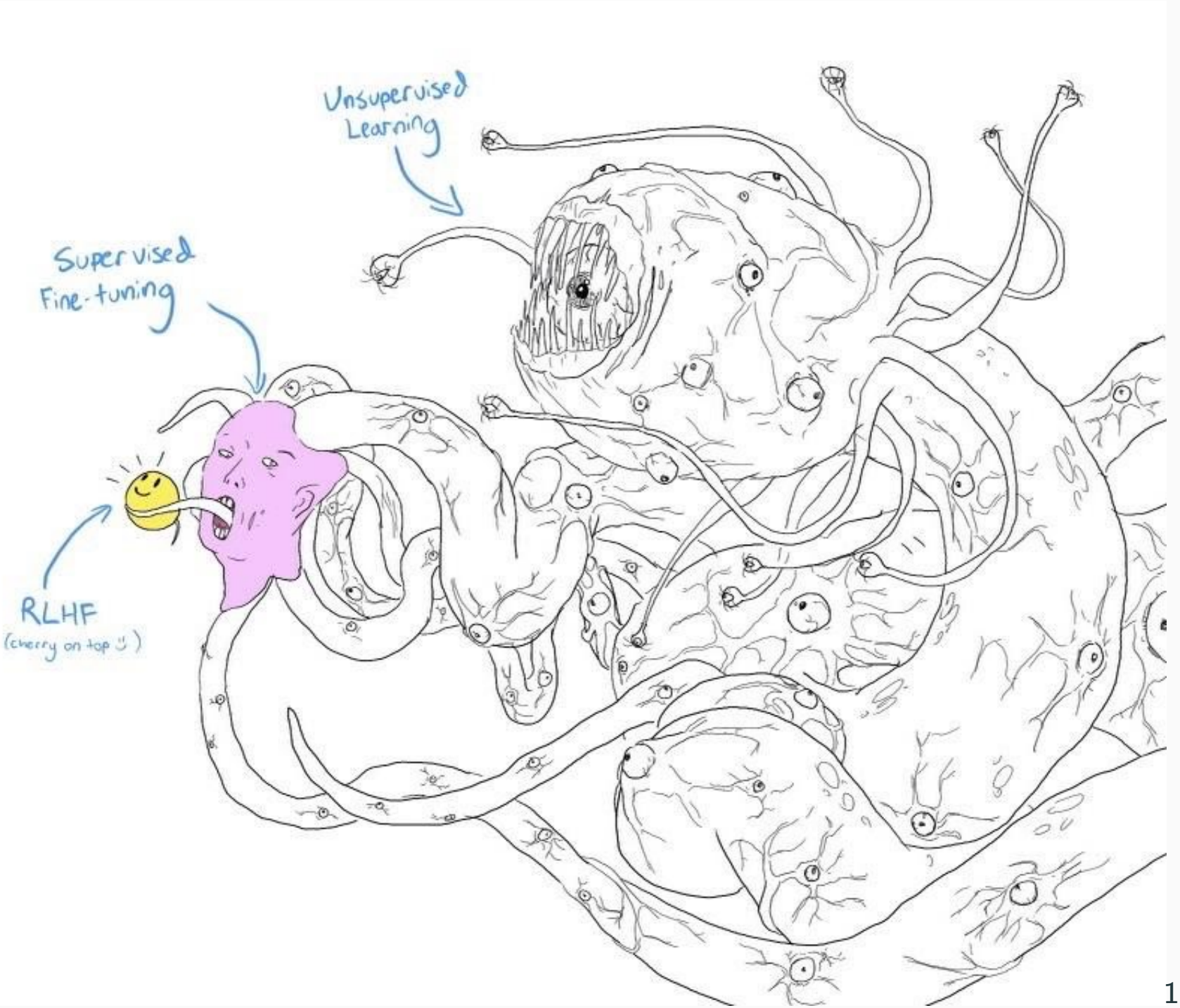


Toxicity


Human eval
PerspectiveAPI score
Toxicity
0.25
0.20
0.15
0.10
0.05
0
None
Respectful
None
Respectful
Prompt
Model
GPT
SFT
PPO-ptx

Bias


Biased prompt
No prompt
Respectful prompt
CrowS-Pairs
Winogender
Normed entropy
0.4
0.3
0.2
0.8
0.7
0.6
0.5
0.4
1.3B
6B
175B
Model size
Model
PPO-ptx
PPO
SFT
GPT


Unsupervised Learning
Supervised Fine-tuning
RLHF
(cherry on top :))