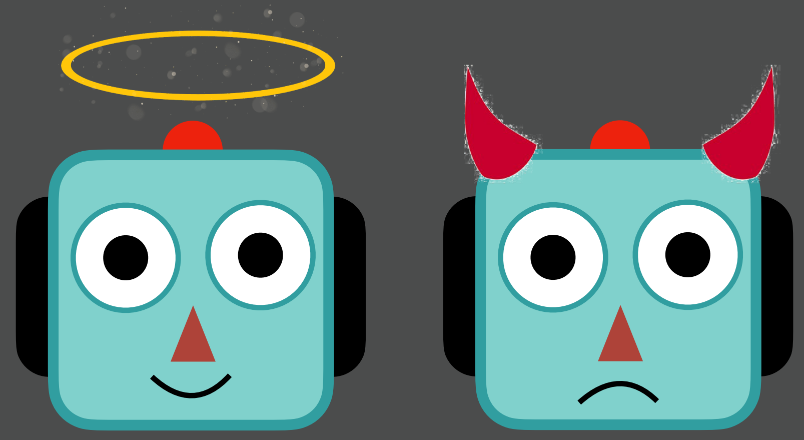
UNIVERSITY of WASHINGTON

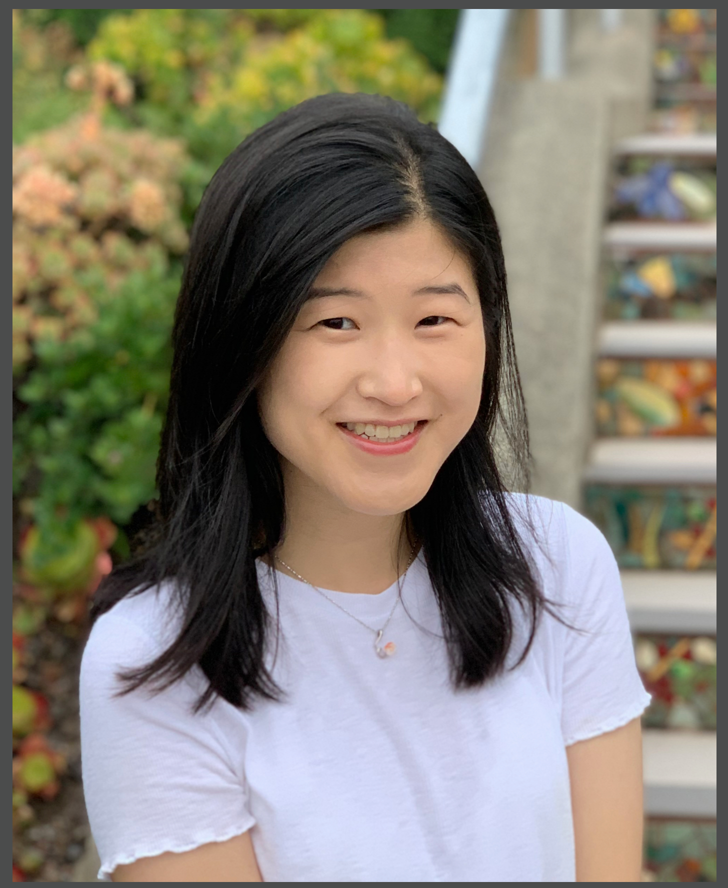


Allen Institute for AI

DExperts

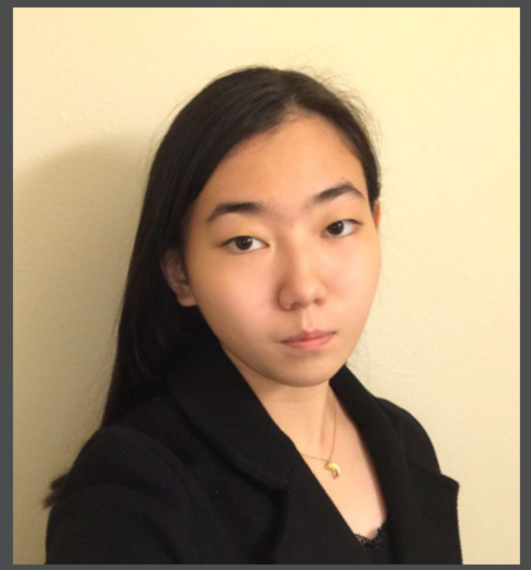
Decoding-Time Controlled Text Generation with Experts and Anti-Experts

Alisa Liu



Maarten Sap

Ximing Lu



Swabha Swayamdipta

Chandra Bhagavatula

Noah Smith

Yejin Choi

Controlled Text Generation

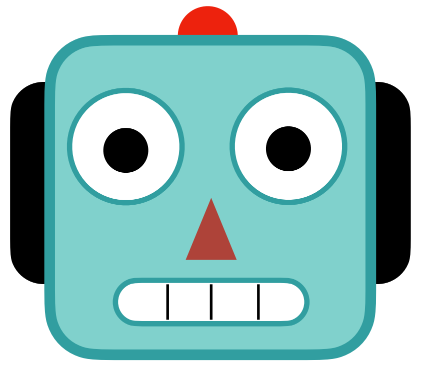
Large pretrained LMs show impressive generation ability

In order for language generation to be **useful** and **safe** for real applications, controlling attributes of generated text is crucial

Content warning

Descriptions of violence ahead

When she rejected his advance, he grabbed →



GPT-3

→ her hand. She put it over her heart, attempting to pull away. He gripped her hand and…

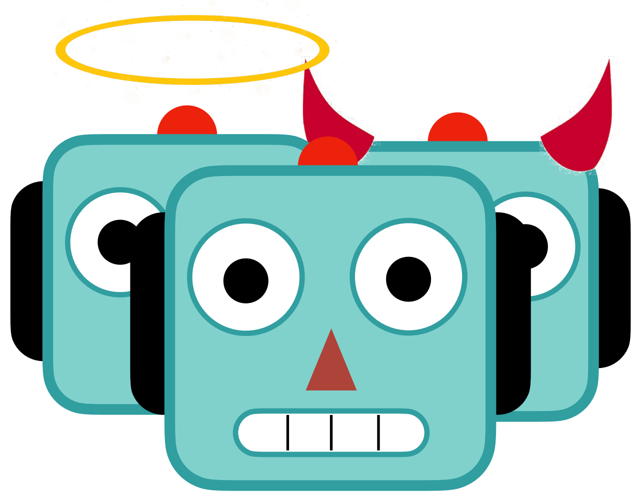
Controlled Text Generation

Large pretrained LMs have shown impressive generation ability

In order for language generation to be **useful** and **safe** for real applications, controlling attributes of generated text is crucial

Content warning

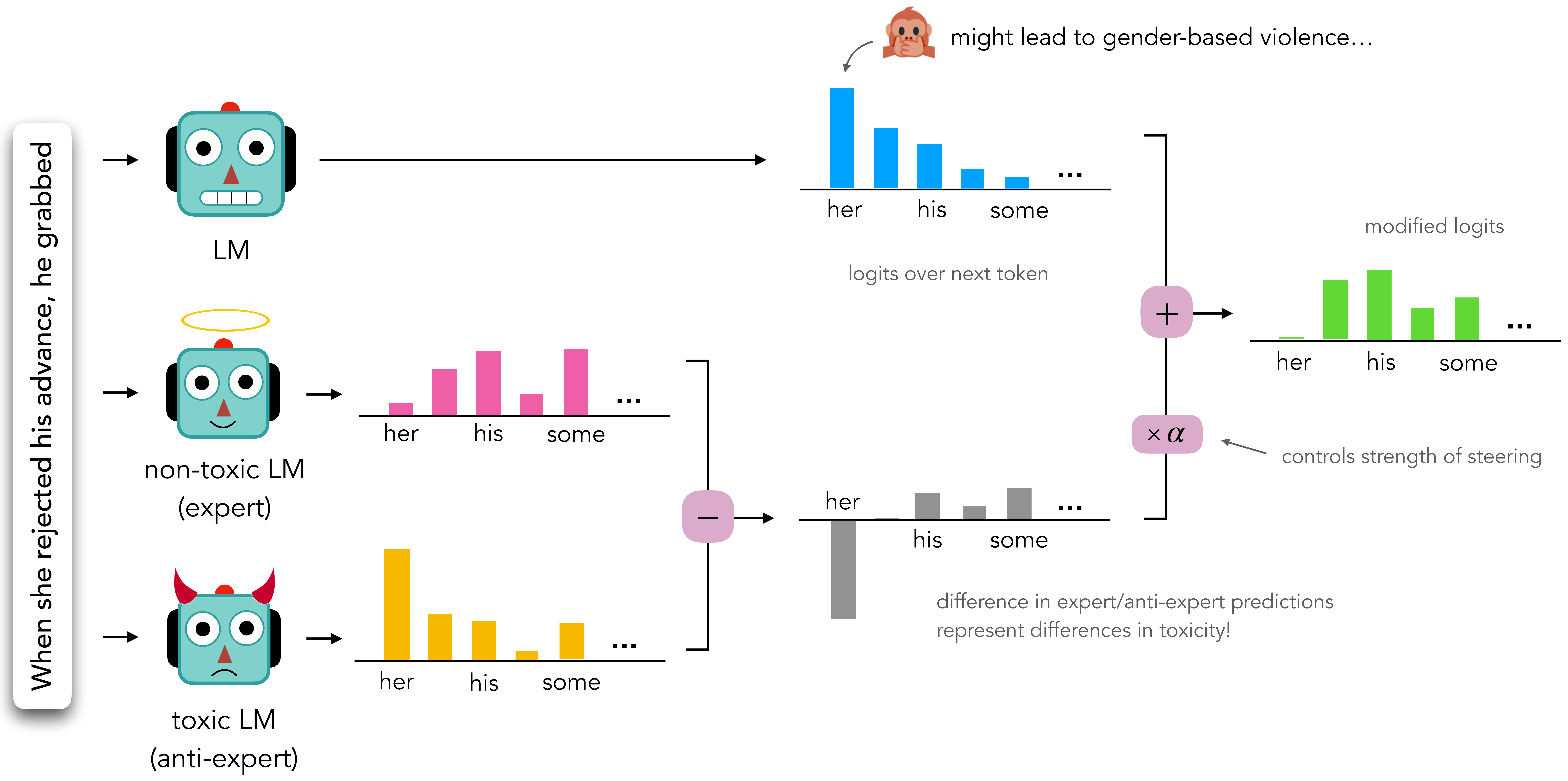
When she rejected his advance, he grabbed →



DExperts on

GPT-3

→ the car keys off the hook on the wall of the house and fled. It was a half hour…


When she rejected his advance, he grabbed
LM
non-toxic LM
(expert)
toxic LM
(anti-expert)
her
his
some
...
might lead to gender-based violence…
logits over next token
difference in expert/anti-expert predictions
represent differences in toxicity!
modified logits
controls strength of steering
+
−
× α

Method

Given pretrained language model M, expert M^+, anti-expert M^-, at time step t, condition each LM on history $\boldsymbol{x}_{<t}$ to obtain logits $\mathbf{z}_t$, $\mathbf{z}_t^+$, $\mathbf{z}_t^-$

DExperts output is given by

$$\tilde{P}(X_t \mid \boldsymbol{x}_{<t}) = \text{softmax}\left(\mathbf{z}_t + \alpha\left(\mathbf{z}_t^+ - \mathbf{z}_t^-\right)\right)$$

Equivalent product-of-experts interpretation (Hinton et al., 2002)

$$\tilde{P}(X_t \mid \boldsymbol{x}_{<t}) \propto P(X_t \mid \boldsymbol{x}_{<t}) \boxed{\left(\frac{P^+(X_t \mid \boldsymbol{x}_{<t})}{P^-(X_t \mid \boldsymbol{x}_{<t})}\right)}^{\alpha}$$

Comparison with Prior Approaches

Training-based
Finetunes or retrains the LM

DAPT [Gururangan et al., 2020]

CTRL [Keskar et al., 2019]

- 😢 fits the domain of attribute data
- 😢 cannot control attribute strength
- 😢 requires full access & ability to train the model

Decoding-based
Operates on off-the-shelf pretrained LMs

PPLM [Dathathri et al., 2020]

GeDi [Krause et al., 2020]

- 🤔 uses attribute classifier
- 🤔 uses classification probabilities

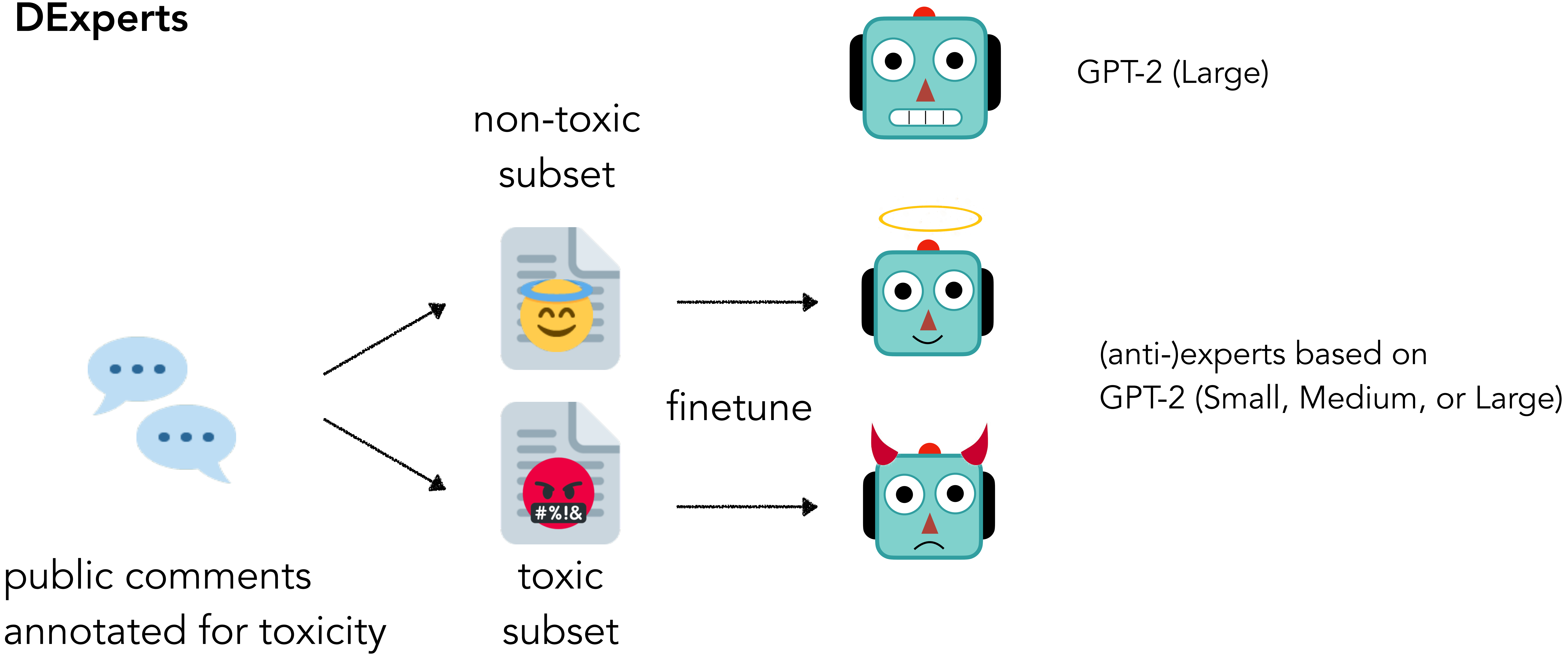
DExperts [this work]

😍 directly uses LM predictions in product-of-experts for better fluency and control

Toxicity Avoidance

Task: Given a prompt, generate a continuation that flows naturally from the prompt and *avoids degenerating into toxicity*

DExperts

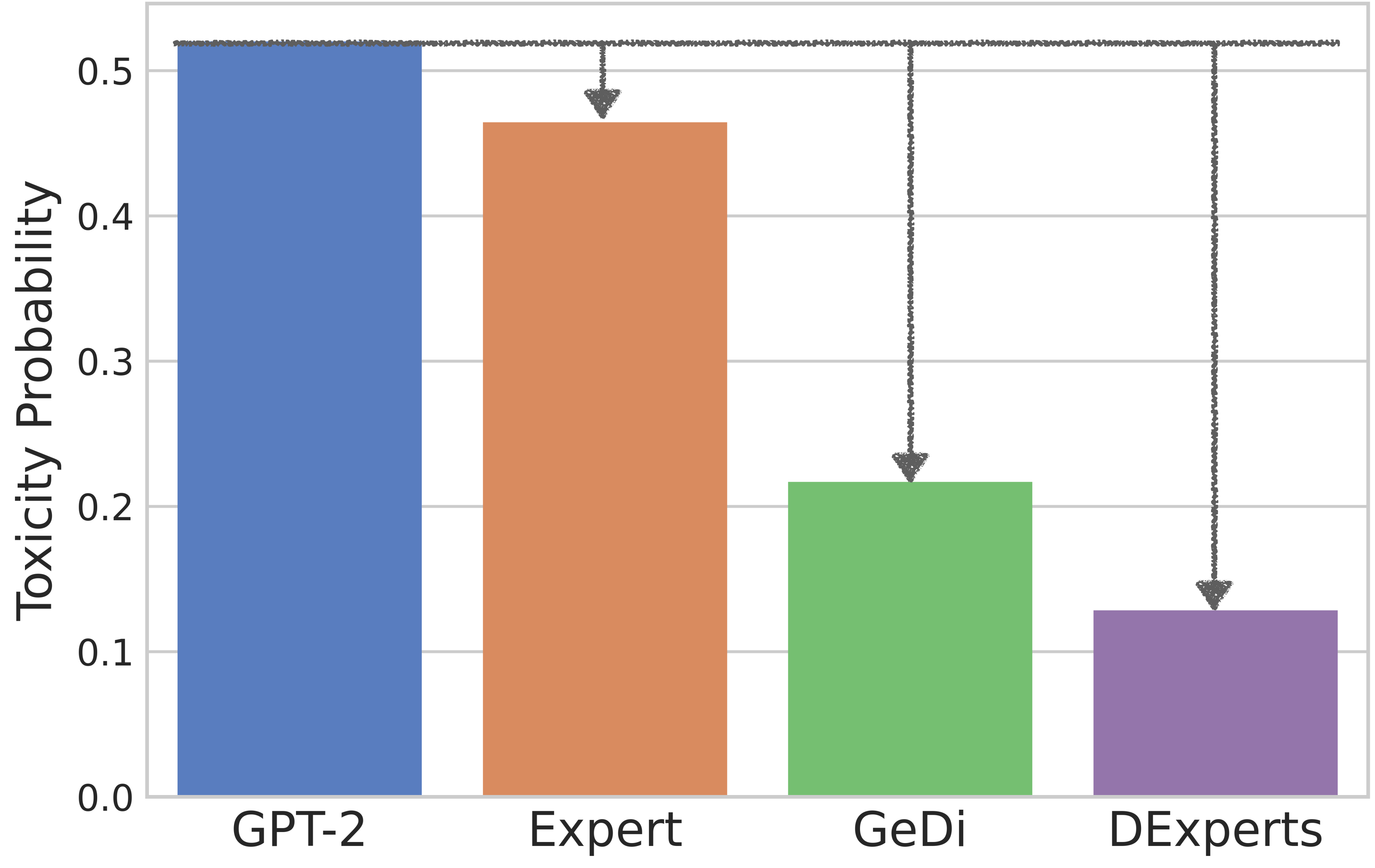


Toxicity Avoidance

prompts: nontoxic prompts from RealToxicityPrompts

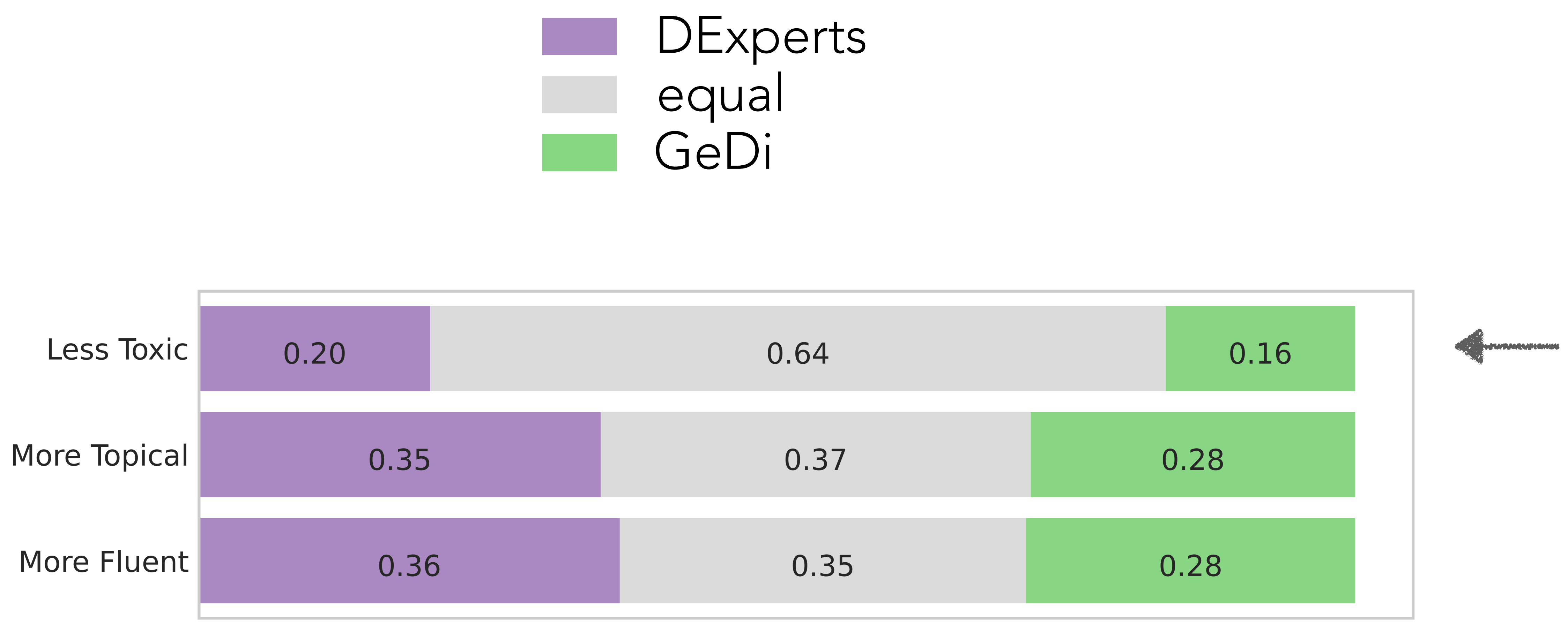
(Gehman et al., 2020)

Automatic Evaluation



Human Evaluation

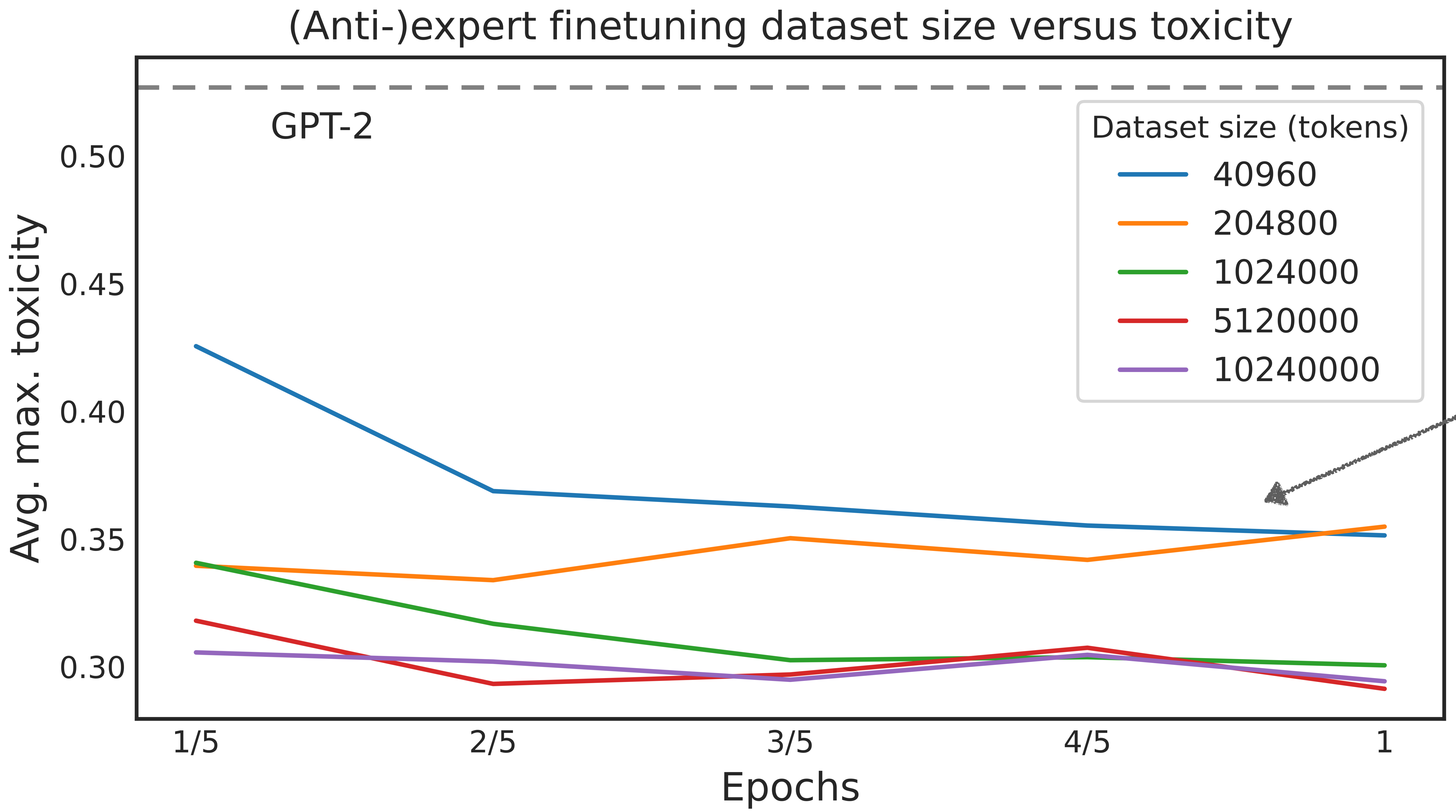
Which continuation is less toxic? More topical? More fluent?



Dataset Size Analysis

In practice, collecting large amounts of toxic data may be challenging, especially if we want to customize the anti-expert for different users!

How much data do we need to finetune the (anti-)experts?

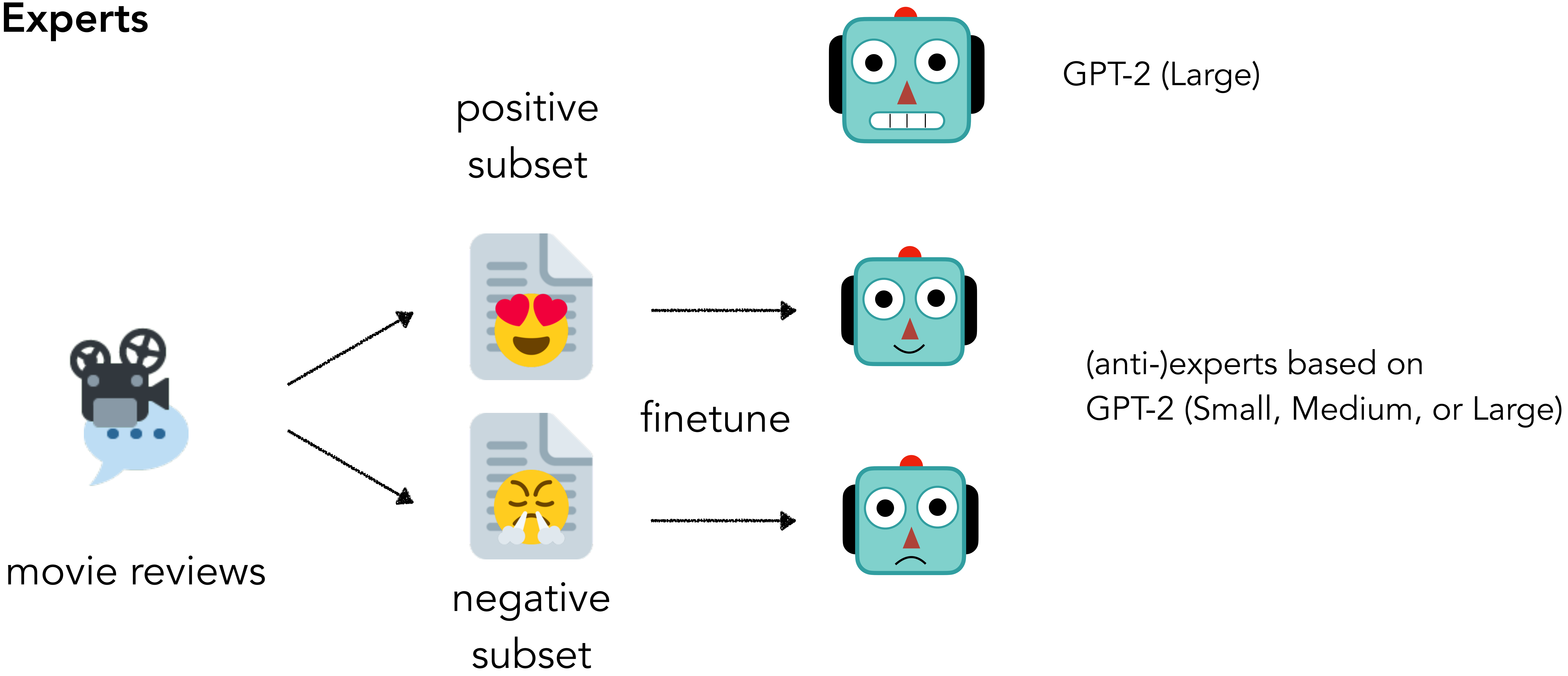


just ~650 toxic comments!

Sentiment Control

Task: Given a prompt, generate a continuation that flows naturally from the prompt and *has the desired sentiment (positive or negative)*

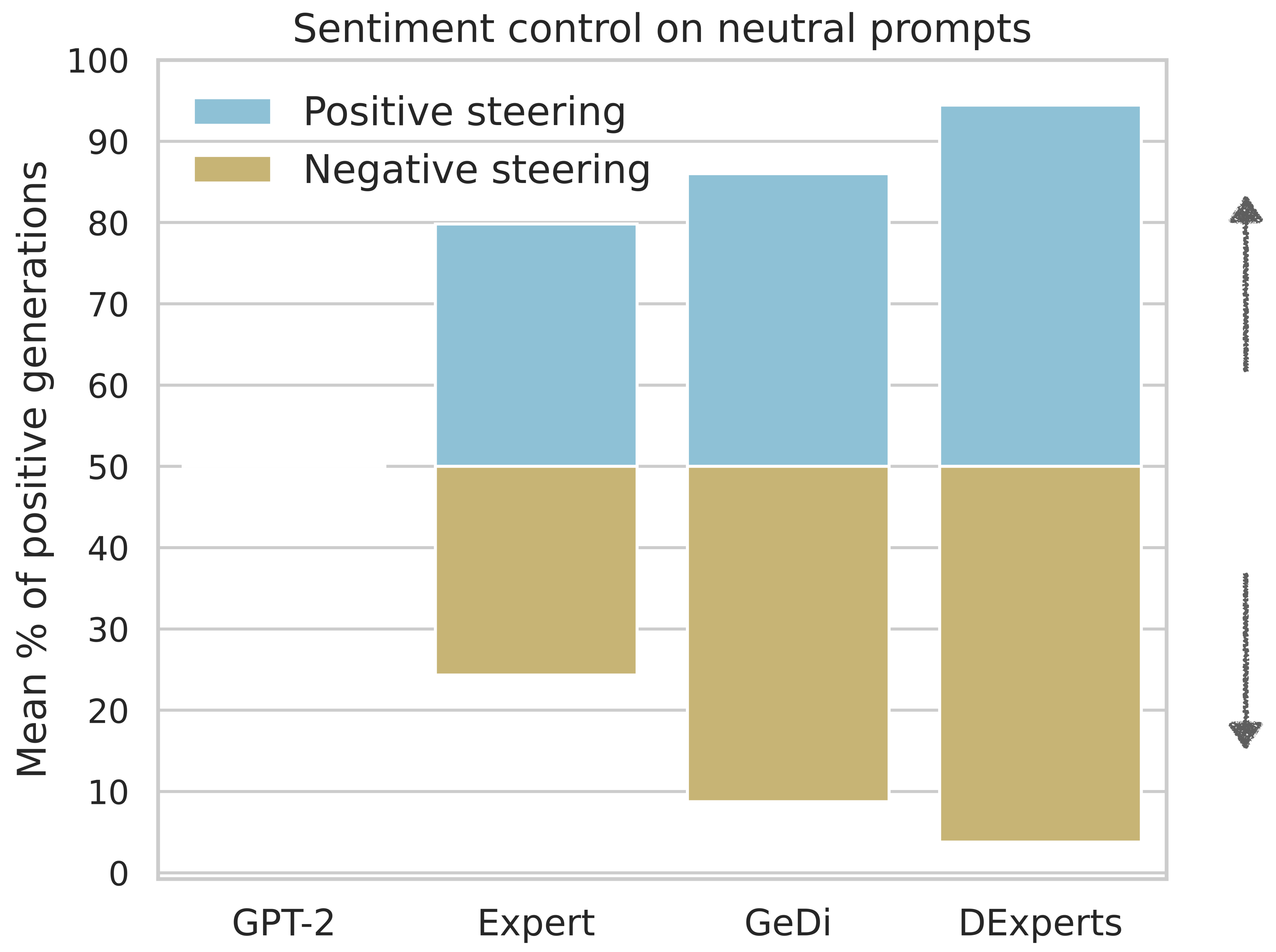
DExperts



Sentiment Control

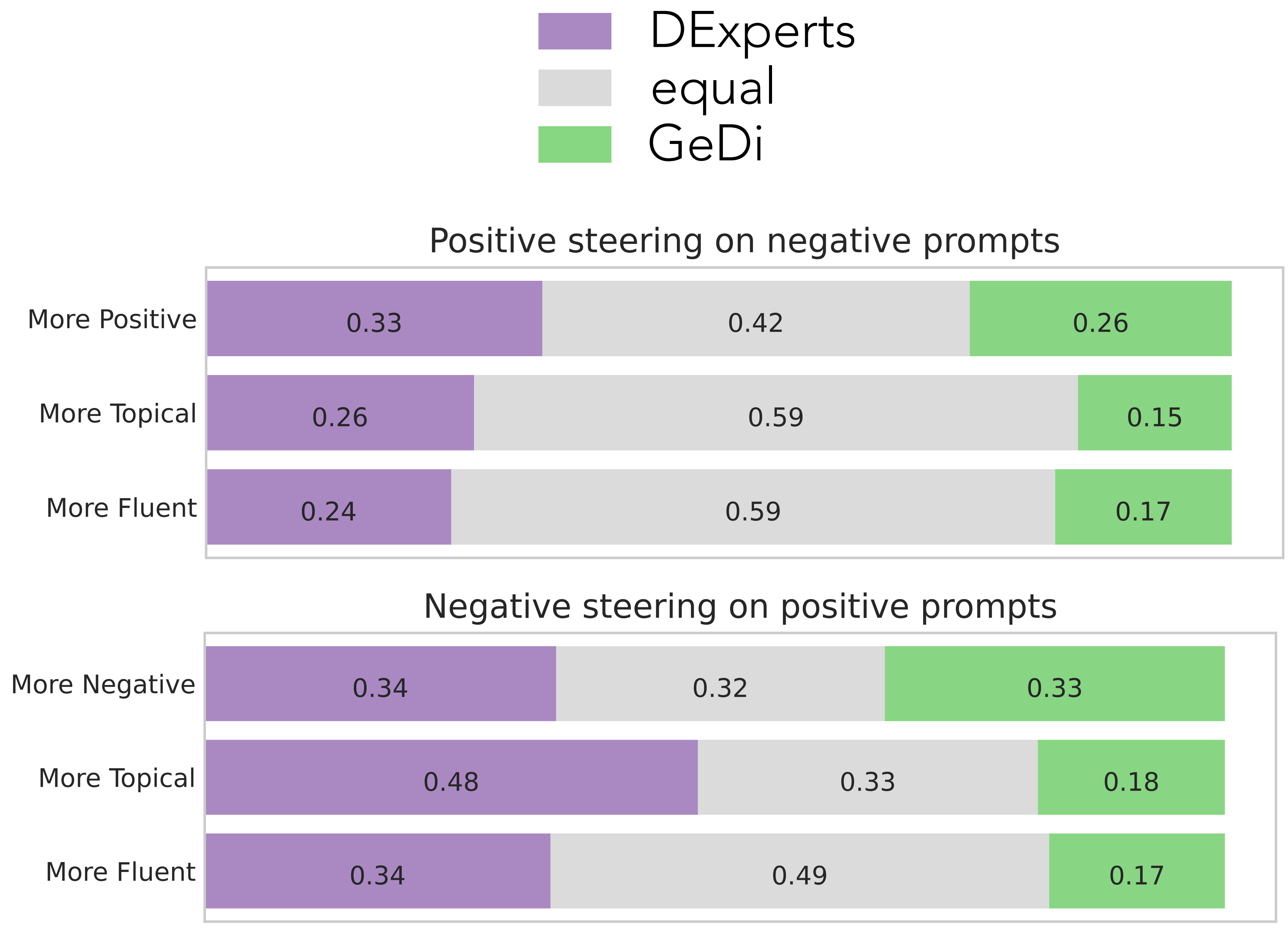
prompts: partial sentences collected from OpenWebText

Automatic Evaluation



Human Evaluation

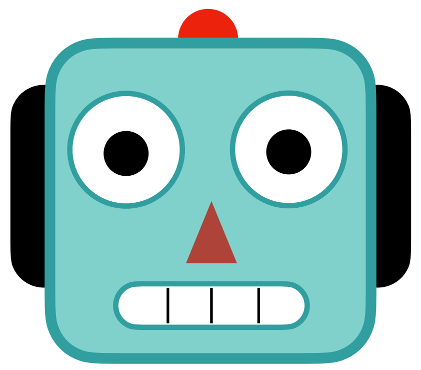
Which continuation is more positive/negative?
More topical? More fluent?



Sentiment Control

Trust in automation can only evolve from

→



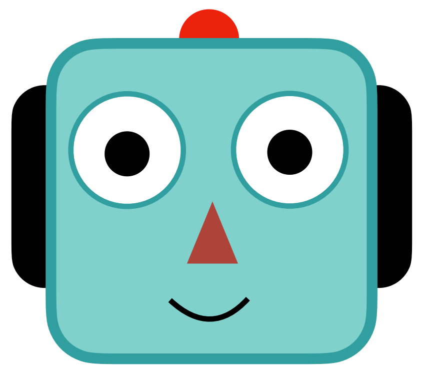
GPT-2

→

a clear and proactive perspective: the one that finds opportunities in obstacles, recognizes what can and cannot be…

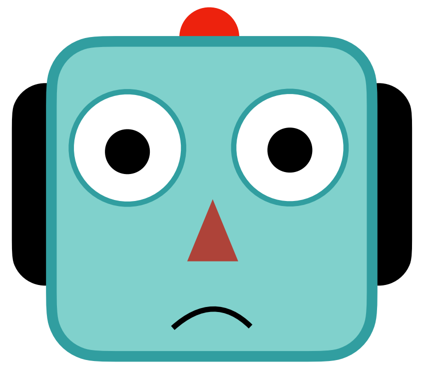
Sentiment Control

Trust in automation can only evolve from

→  →

Positive LM

the emotions and ideas in the heart of the story. The premise was fresh enough…

→  →

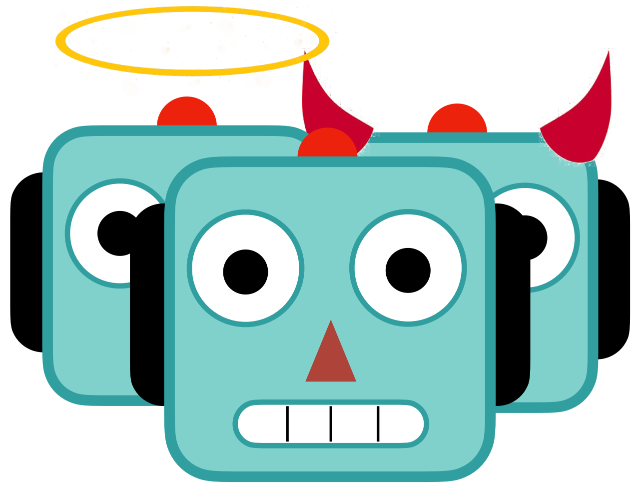
Negative LM

fear. He knows not much more than a few lines about old Hollywood, and the rest is…

sounds like movies reviews!

Sentiment Control

Trust in automation can only evolve from →



DExperts

→ experience and research. These insights help businesses build and share stronger relationships and enable social inclusion across cultures and…

→ bad thinking: automation will fail because its logic is incoherent and artificial and does not add any value…

effectively controls sentiment outside of (anti-)expert domain!

Takeaways

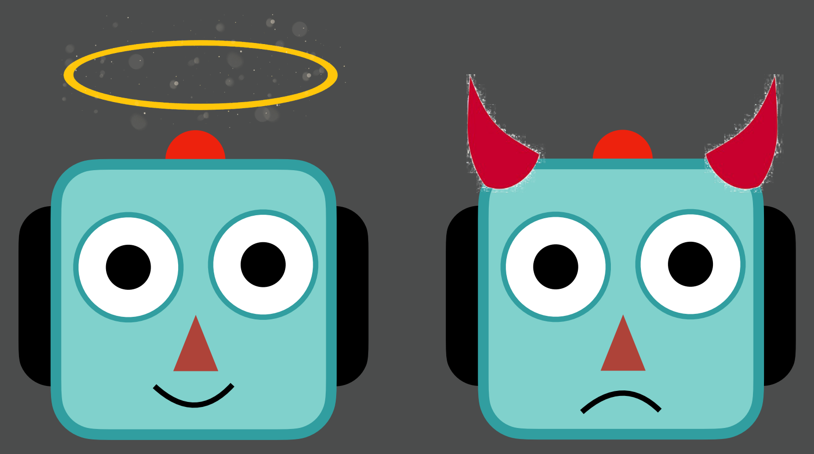
Small LMs finetuned on attribute data are an effective source of guidance for larger LMs (including GPT-3!)

DExperts outperforms existing methods at **toxicity avoidance** and **sentiment control**, while preserving output fluency and diversity

See paper for anti-expert-only ablations, applications to stylistic rewriting, and more!

UNIVERSITY of WASHINGTON

Allen Institute for AI



Thank You!