Self-Conditioning Pre-Trained Language Models

Xavi Suau, Luca Zappella and Nick Apostoloff | ICML 2022
Apple

Goals

Condition Transformer-based Language Models (TLMs) are:

- Expensive to condition (re-training [1], using additional parameters [2]).
- Perpetuated data bias [3].

[1] Keskar, N. S., McCann, B., Varshney, L., Xiong, C., and Socher, R. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint*, 2019.

[2] Yang, K. and Klein, D. Fudge: Controlled text generation with future discriminators. *NAACL*, 2021.

[3] Abid, A., Farooqi, M., and Zou, J. Large language models associate muslims with violence. *Nature Machine Intelligence*, 3, 2021.

Goals

Condition Transformer-based Language Models (TLMs) are:

- Expensive to condition (re-training [1], using additional parameters [2]).
- Perpetuated data bias [3].

Efficient conditioned generation

Study about mitigating gender bias via conditioning.

Open-ended fine-grained conditioning.

[1] Keskar, N. S., McCann, B., Varshney, L., Xiong, C., and Socher, R. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint*, 2019.

[2] Yang, K. and Klein, D. Fudge: Controlled text generation with future discriminators. *NAACL*, 2021.

[3] Abid, A., Farooqi, M., and Zou, J. Large language models associate muslims with violence. *Nature Machine Intelligence*, 3, 2021.

Conditioned Language Model

$$p(\mathbf{x} \mid c)$$

Generation of a sentence $\mathbf{x}$
conditioned to concept c

[4] Yang, Kevin, and Dan Klein. "FUDGE: Controlled Text Generation With Future Discriminators." NAACL, 2021.
[5] Dathathri, Sumanth, et al. "Plug and play language models: A simple approach to controlled text generation." ICLR, 2020.

Conditioned Language Model

$$p(\mathbf{x} \mid c) \propto p(c \mid \mathbf{x})\, p(\mathbf{x})$$

Generation of a sentence $\mathbf{x}$ conditioned to concept c

[4] Yang, Kevin, and Dan Klein. "FUDGE: Controlled Text Generation With Future Discriminators." NAACL, 2021.
[5] Dathathri, Sumanth, et al. "Plug and play language models: A simple approach to controlled text generation." ICLR, 2020.

Conditioned Language Model

$$p(\mathbf{x} \mid c) \propto p(c \mid \mathbf{x})\, p(\mathbf{x})$$

Generation of a sentence $\mathbf{x}$ conditioned to concept c

Expert model at generating realistic sentences $\mathbf{x}$

[4] Yang, Kevin, and Dan Klein. "FUDGE: Controlled Text Generation With Future Discriminators." NAACL, 2021.
[5] Dathathri, Sumanth, et al. "Plug and play language models: A simple approach to controlled text generation." ICLR, 2020.

Conditioned Language Model

$$\underbrace{p(\mathbf{x} \mid c)}_{\text{Generation of a sentence } \mathbf{x} \text{ conditioned to concept } c} \propto \underbrace{p(c \mid \mathbf{x})}_{\text{Expert model detecting concept } c \text{ in } \mathbf{x}} \underbrace{p(\mathbf{x})}_{\text{Expert model at generating realistic sentences } \mathbf{x}}$$

[4] Yang, Kevin, and Dan Klein. "FUDGE: Controlled Text Generation With Future Discriminators." NAACL, 2021.
[5] Dathathri, Sumanth, et al. "Plug and play language models: A simple approach to controlled text generation." ICLR, 2020.

Conditioned Language Model

$$\underbrace{p(\mathbf{x} \mid c)}_{\text{Generation of a sentence } \mathbf{x} \text{ conditioned to concept } c} \propto \underbrace{p(c \mid \mathbf{x})}_{\text{Expert model detecting concept } c \text{ in } \mathbf{x}} \; \underbrace{p(\mathbf{x})}_{\text{Expert model at generating realistic sentences } \mathbf{x}}$$

FUDGE [4] and PPLM [5] ➡ External $p(c \mid \mathbf{x})$

[4] Yang, Kevin, and Dan Klein. "FUDGE: Controlled Text Generation With Future Discriminators." NAACL, 2021.

[5] Dathathri, Sumanth, et al. "Plug and play language models: A simple approach to controlled text generation." ICLR, 2020.

Conditioned Language Model

$$p(\mathbf{x} \mid c) \propto p(c \mid \mathbf{x})\, p(\mathbf{x})$$

Generation of a sentence $\mathbf{x}$ conditioned to concept c

Expert model detecting concept c in $\mathbf{x}$

Expert model at generating realistic sentences $\mathbf{x}$

FUDGE [4] and PPLM [5] ➡ External $p(c \mid \mathbf{x})$

Our work ➡ $p(c \mid \mathbf{x})$ and $p(\mathbf{x})$ **already co-exist** in the pre-trained model

[4] Yang, Kevin, and Dan Klein. "FUDGE: Controlled Text Generation With Future Discriminators." NAACL, 2021.
[5] Dathathri, Sumanth, et al. "Plug and play language models: A simple approach to controlled text generation." ICLR, 2020.

Concepts

Represent concepts with positive and negative sentences [6]

[6] Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. Just "OneSec" for producing multilingual sense-annotated data. ACL 2019.
[7] Princeton University. Wordnet: A lexical database for english. https://wordnet.princeton.edu.

Concepts

Represent concepts with positive and negative sentences [6]

Positive sentences	Negative sentences

[6] Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. Just "OneSec" for producing multilingual sense-annotated data. ACL 2019.

[7] Princeton University. Wordnet: A lexical database for english. https://wordnet.princeton.edu.

Concepts

Represent concepts with positive and negative sentences [6]

	Positive sentences	Negative sentences
Sense	Contain keyword with WordNet sense [7]	Do NOT contain keyword
Homograph	Contain keyword with WordNet sense [7]	Contain same keyword with different sense [7]
Abstract	Contain abstract concept (ie. Sentiment)	Do NOT contain concept
...	...	...

[6] Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. Just "OneSec" for producing multilingual sense-annotated data. ACL 2019.
[7] Princeton University. Wordnet: A lexical database for english. https://wordnet.princeton.edu.

Finding Expert Units

Understand which concepts are learnt in a Language Model (LM)

Contain concept c $(y = 1)$

```
pos sentence 1
pos sentence 2
pos sentence 3
...
```

Do NOT contain concept c $(y = 0)$

```
neg sentence 1
neg sentence 2
neg sentence 3
...
```

Finding Expert Units

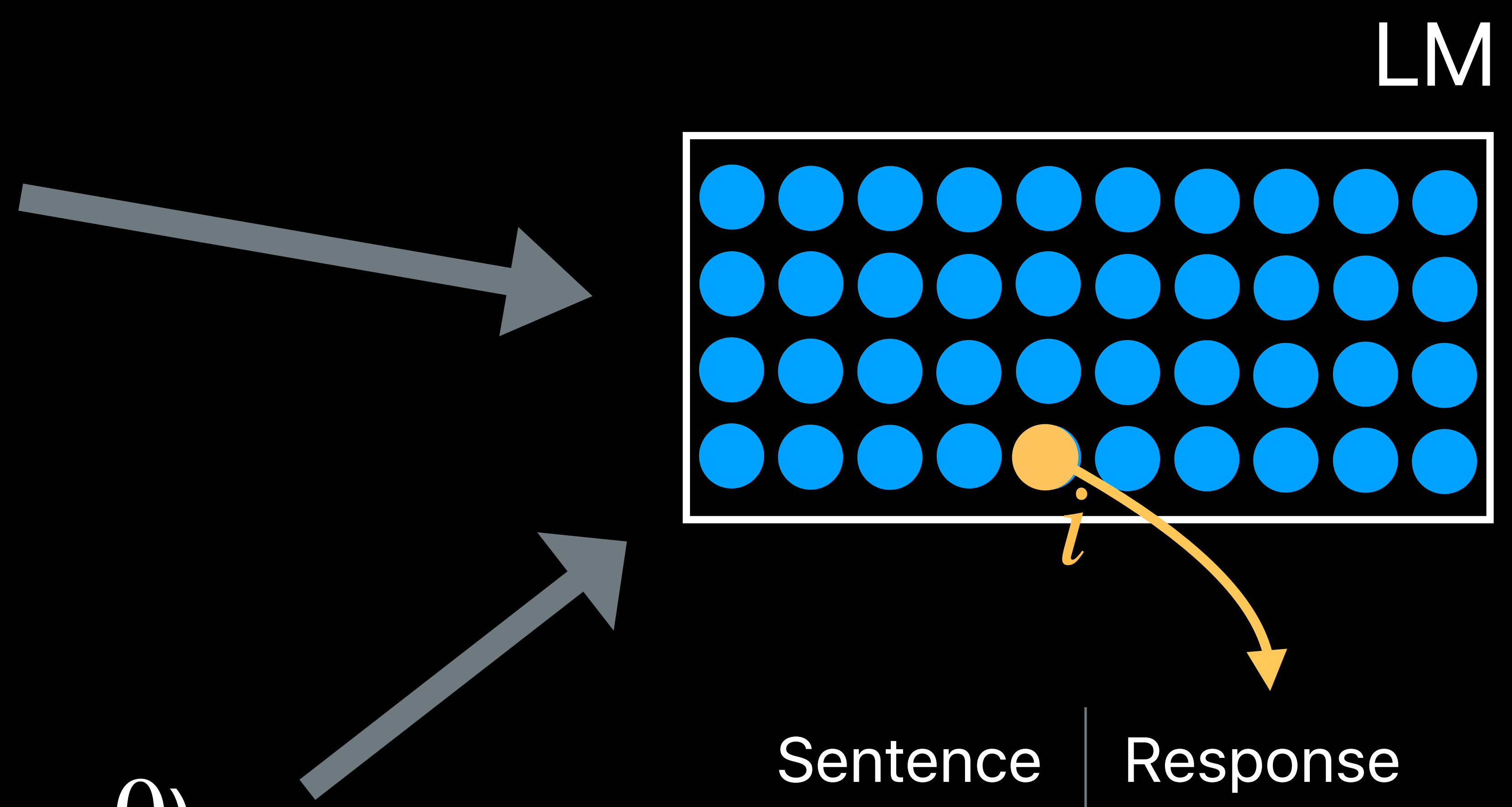
Understand which concepts are learnt in a Language Model (LM)

Contain concept c ($y = 1$)

```
pos sentence 1
pos sentence 2
pos sentence 3
...
```

Do NOT contain concept c ($y = 0$)

```
neg sentence 1
neg sentence 2
neg sentence 3
...
```



Sentence label	Response neuron i
1	0.83
1	1.74
1	0.98
0	0.12
0	0.06
0	1.01

Finding Expert Units

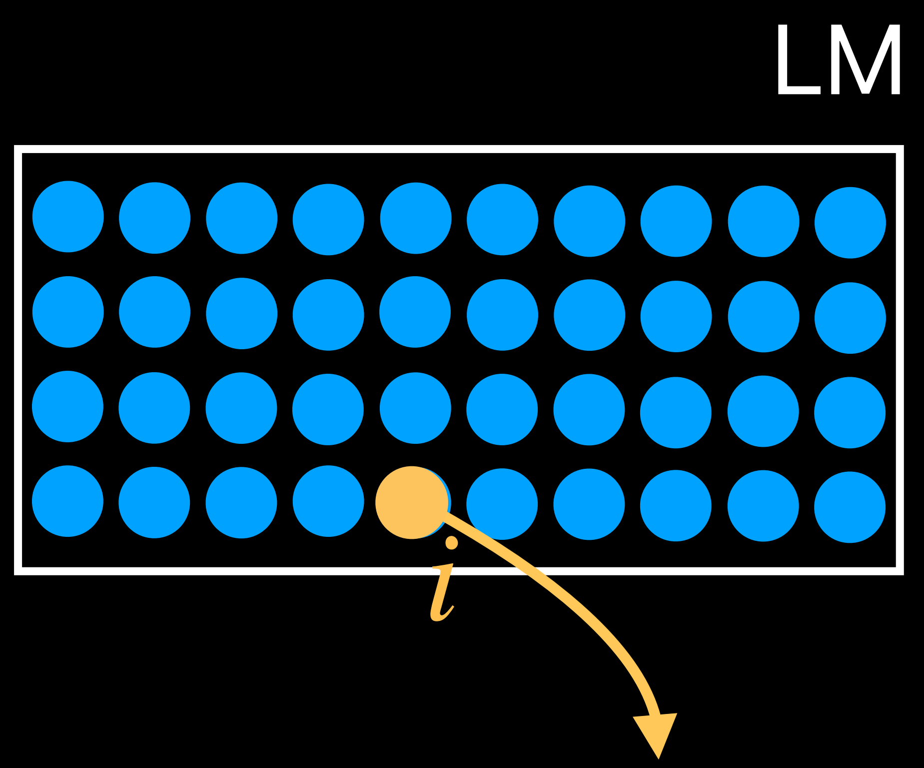
Understand which concepts are learnt in a Language Model (LM)

Contain concept c ($y = 1$)

```
pos sentence 1
pos sentence 2
pos sentence 3
...
```

Do NOT contain concept c ($y = 0$)

```
neg sentence 1
neg sentence 2
neg sentence 3
...
```



Sentence label	Response neuron i
1	0.83
1	1.74
1	0.98
0	0.12
0	0.06
0	1.01

Is neuron i a good classifier for concept c?

Finding Expert Units

Understand which concepts are learnt in a Language Model (LM)

Contain concept c $(y = 1)$

```
pos sentence 1
pos sentence 2
pos sentence 3
```

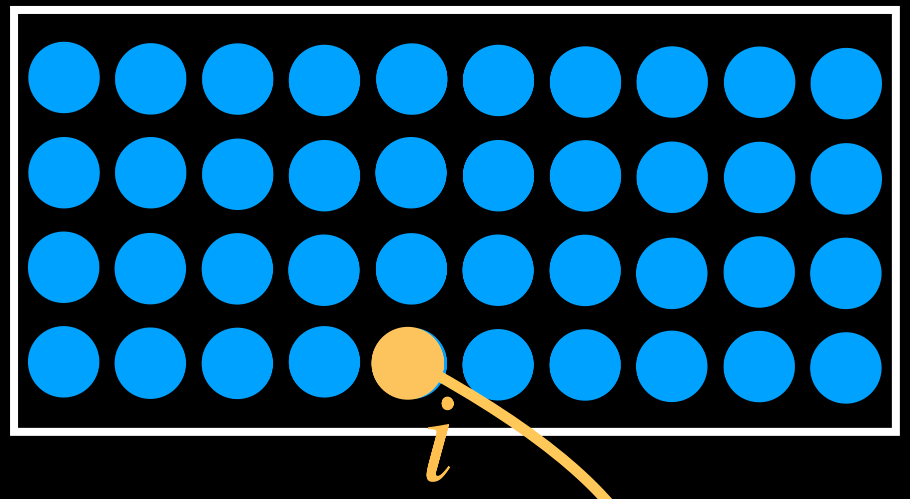
...

Do NOT contain concept c $(y = 0)$

```
neg sentence 1
neg sentence 2
neg sentence 3
```

...

LM



Is neuron i a good classifier for concept c?

Sentence label	Response neuron i
1	0.83
1	1.74
1	0.98
0	0.12
0	0.06
0	1.01

$$AP_c^i = 0.87$$

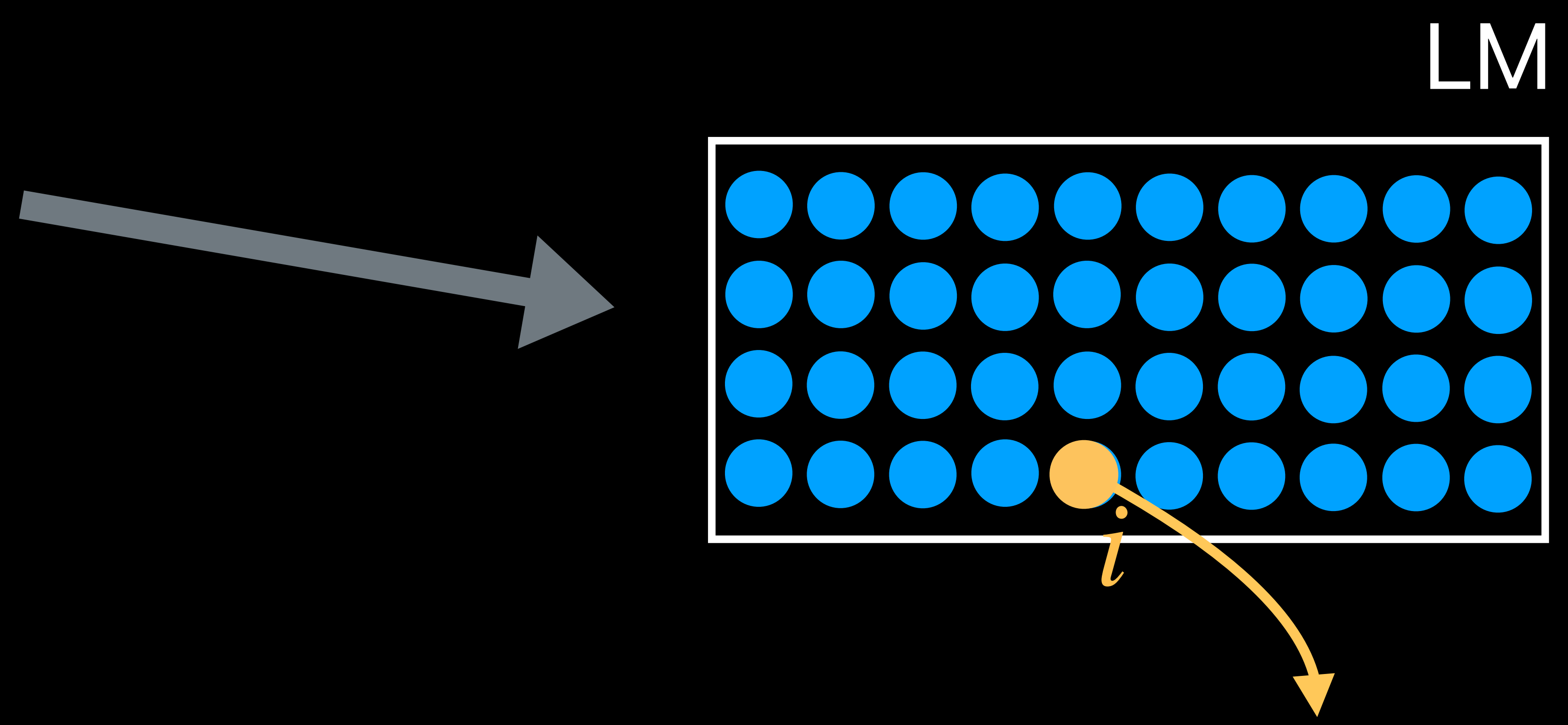
Unit expertise for concept c

Conditioning Based on Expert Units

What is the expert unit’s “active” value?

Contain concept c ($y = 1$)

```
pos sentence 1
pos sentence 2
pos sentence 3
...
```



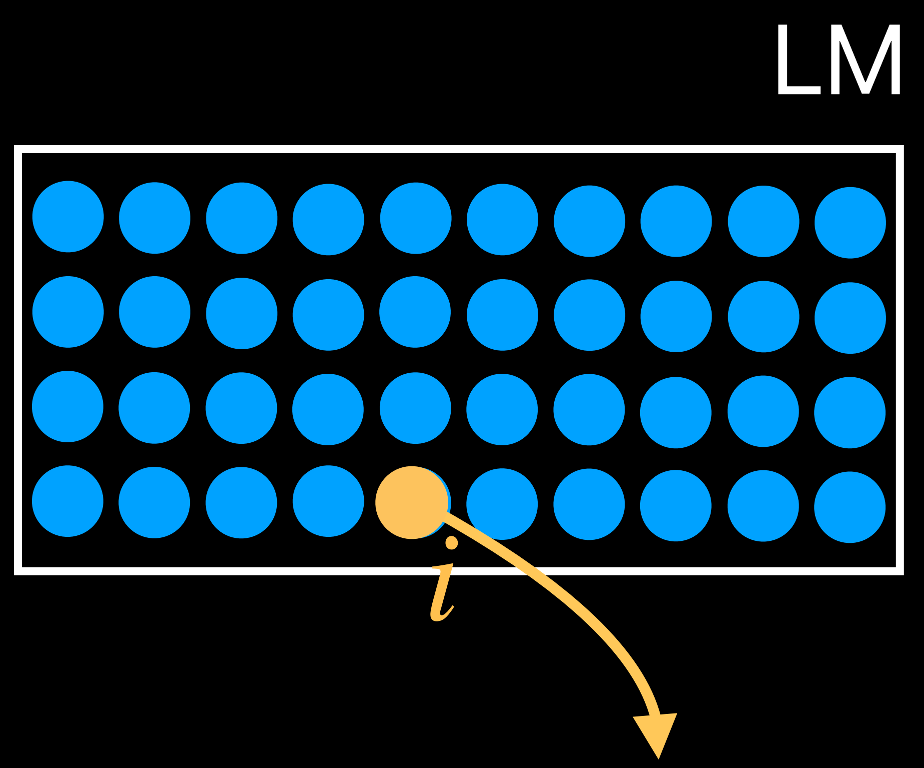
Sentence label	Response neuron i
1	0.83
1	1.74
1	0.98
0	0.12
0	0.06
0	1.01

Conditioning Based on Expert Units

What is the expert unit's "active" value?

Contain concept c ($y = 1$)

```
pos sentence 1
pos sentence 2
pos sentence 3
…
```



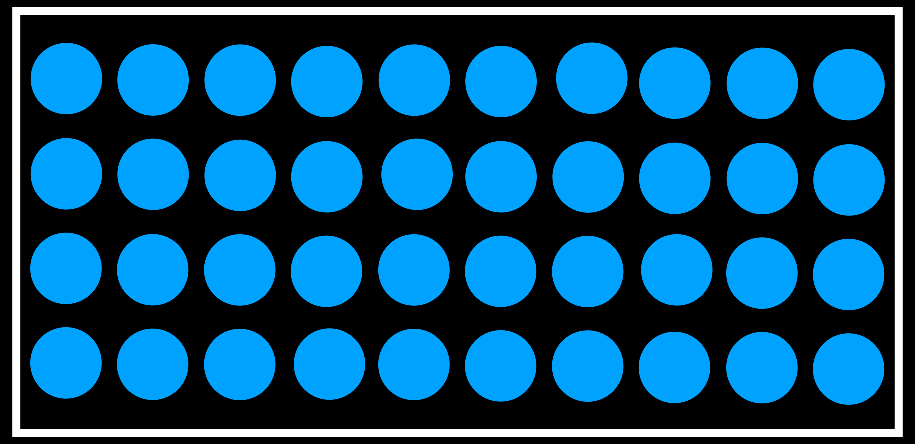
Sentence label	Response neuron i
1	0.83
1	1.74
1	0.98
0	0.12
0	0.06
0	1.01

$$\hat{z}_i^c = \mathbb{E}[z_i \mid y = 1]$$

Expected response of unit i when concept is present

Conditioned Language Model

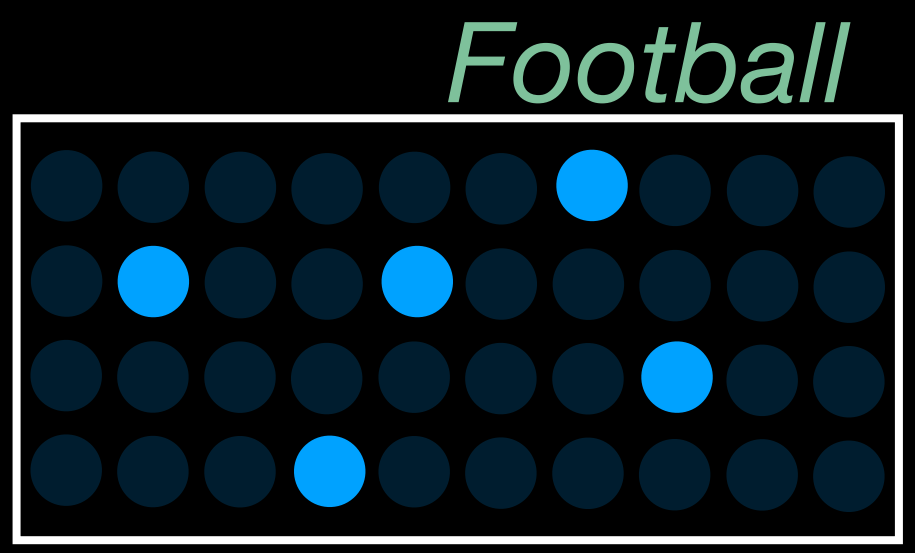
$$p(\mathbf{x} \mid c) \propto \underline{p(c \mid \mathbf{x})}\, p(\mathbf{x})$$



Conditioned Language Model

Expert units (highest AP)

$$p(\mathbf{x} \mid c) \propto \underline{p(c \mid \mathbf{x})}\, p(\mathbf{x})$$



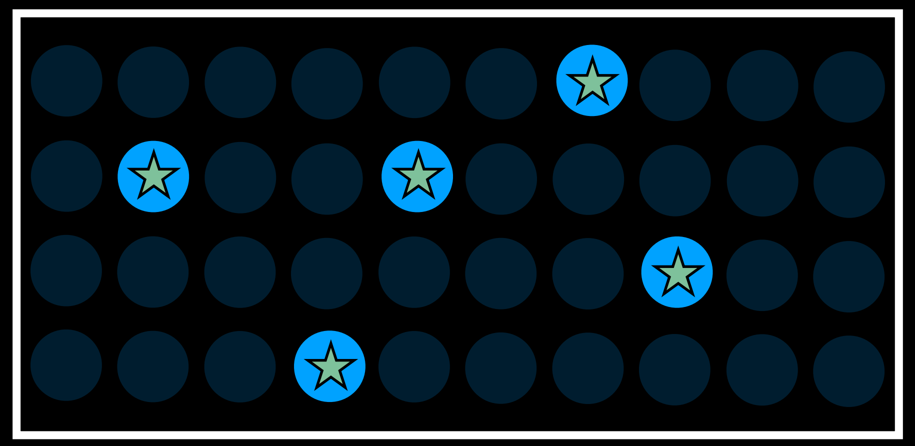
Conditioned Language Model

Expert units (highest AP)

$$p(\mathbf{x} \mid c) \propto \underline{p(c \mid \mathbf{x})}\, p(\mathbf{x})$$

Intervention on k expert units

Football



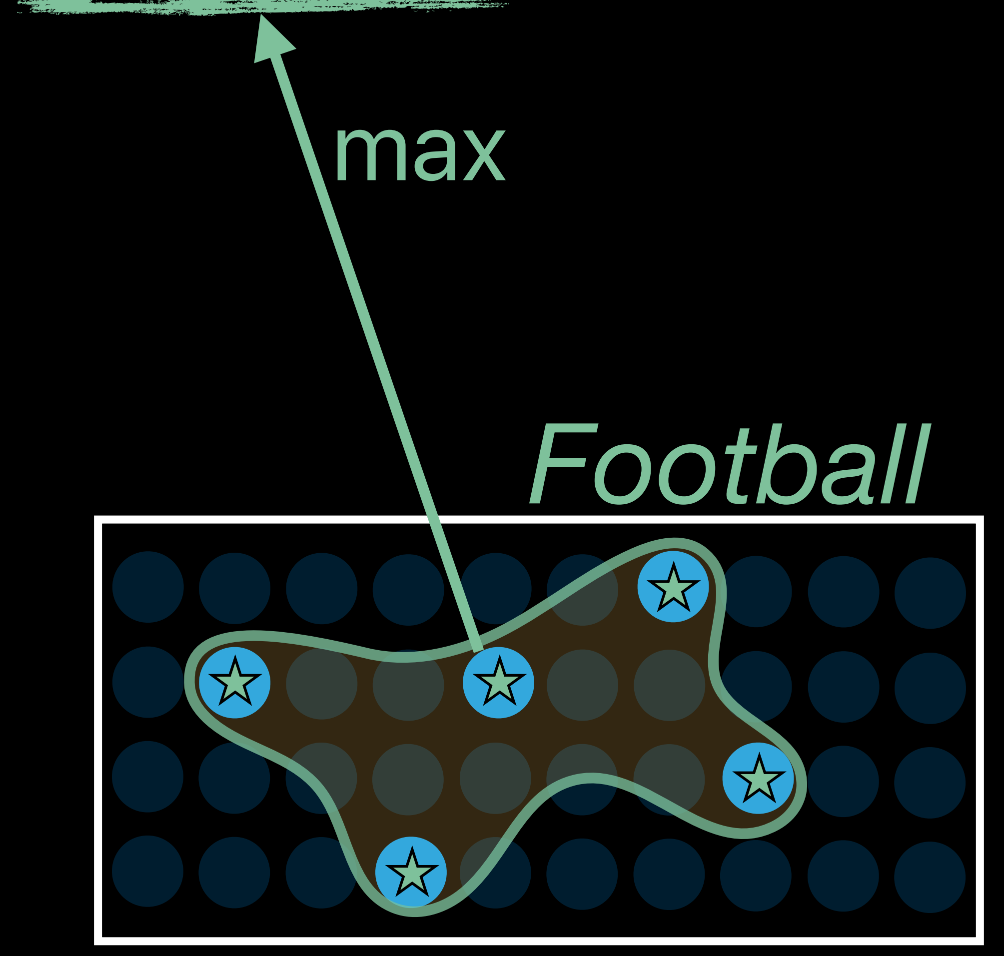
★ $= do(c, k) : z_i \leftarrow \hat{z}_i^c$ intervention

Conditioned Language Model

Expert units (highest AP)

Intervention on k expert units

$$p(\mathbf{x} \mid c) \propto \underline{p(c \mid \mathbf{x})}\, p(\mathbf{x})$$



★ $= do(c, k) : z_i \leftarrow \hat{z}_i^c$ intervention

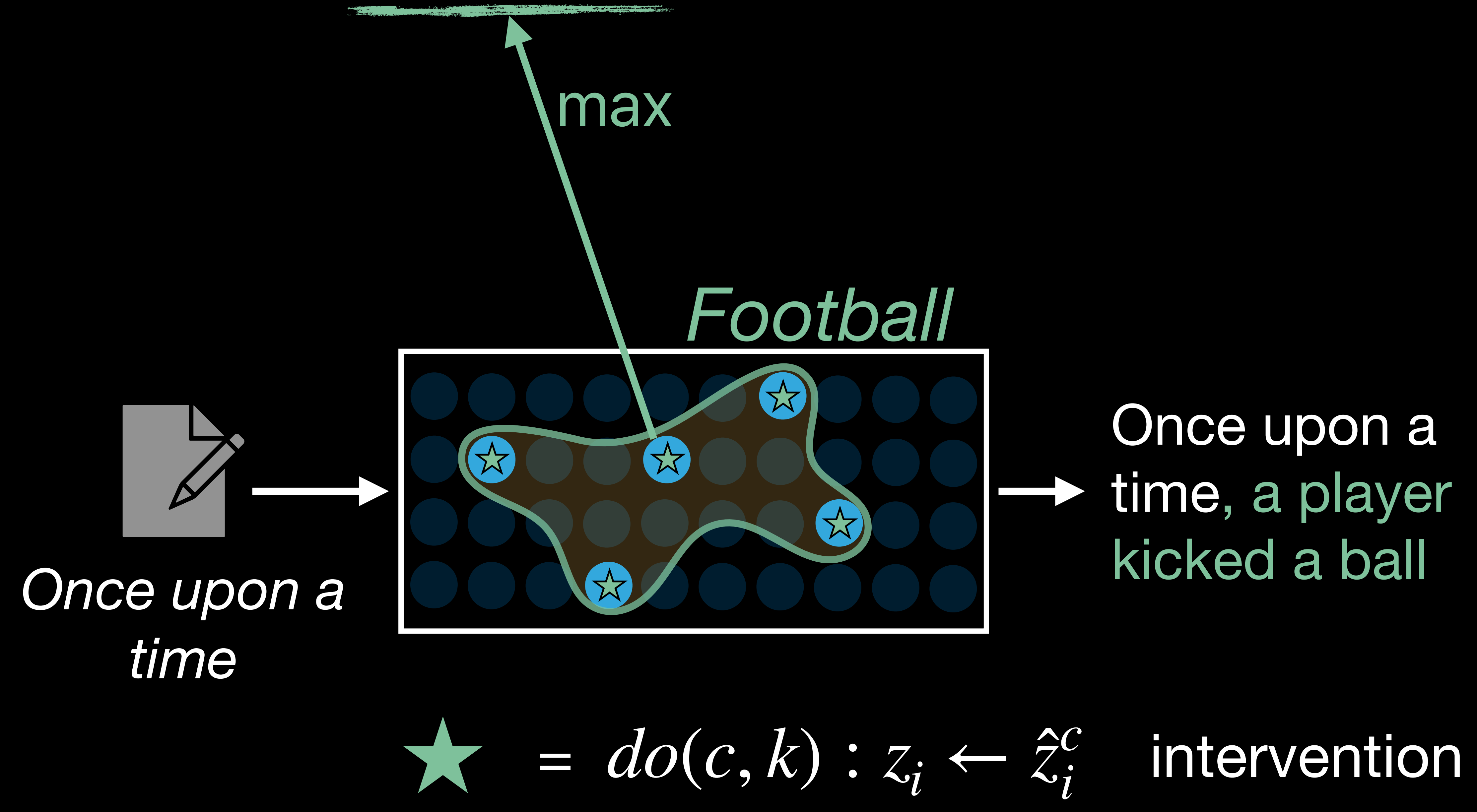
Conditioned Language Model

Expert units (highest AP)

Intervention on k expert units

Generate text with a concept

$$p(\mathbf{x} \mid c) \propto \underline{p(c \mid \mathbf{x})}\, p(\mathbf{x})$$



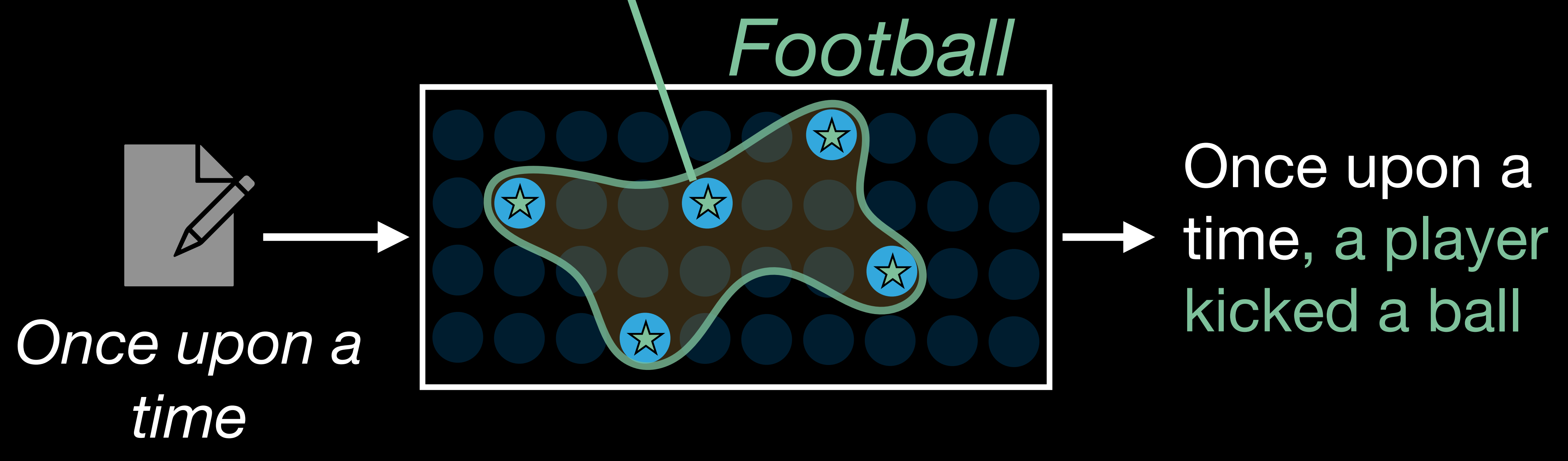
Conditioned Language Model

Expert units (highest AP)

Intervention on k expert units

Generate text with a concept

$$p(\mathbf{x} \mid c) \propto \underbrace{p(c \mid \mathbf{x})}_{\text{max}} \, p(\mathbf{x})$$



$= do(c, k) : z_i \leftarrow \hat{z}_i^c$ intervention

No training, no fine-tuning

Applicable to any pre-trained LM

Generative gender parity

1037 prompts with stereotypical gender bias from [8].

GPT2-medium conditioned on concepts $c =$ woman and $c =$ man.

[8] Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Sakenis, S., Huang, J., Singer, Y., and Shieber, S. Causal mediation analysis for interpreting neural NLP: The case of gender bias. NeurIPS, 2020.

Generative gender parity

1037 prompts with stereotypical gender bias from [8].

GPT2-medium conditioned on concepts $c =$ woman and $c =$ man.

"The nurse said that" she was eating.

[8] Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Sakenis, S., Huang, J., Singer, Y., and Shieber, S. Causal mediation analysis for interpreting neural NLP: The case of gender bias. NeurIPS, 2020.

Generative gender parity

1037 prompts with stereotypical gender bias from [8].

GPT2-medium conditioned on concepts $c =$ woman and $c =$ man.

"The nurse said that" $\boxed{\text{she}}$ was eating.

Bias: $\Delta p(c, \cdot\,) \triangleq p(\text{she} \,|\, do(c, \cdot\,)) - p(\text{he} \,|\, do(c, \cdot\,))$.

[8] Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Sakenis, S., Huang, J., Singer, Y., and Shieber, S. Causal mediation analysis for interpreting neural NLP: The case of gender bias. NeurIPS, 2020.

Generative gender parity

1037 prompts with stereotypical gender bias from [8].

GPT2-medium conditioned on concepts $c =$ woman and $c =$ man.

"The nurse said that" $\boxed{\text{she}}$ was eating.

Bias: $\Delta p(c, \cdot\,) \triangleq p(\text{she} \,|\, do(c, \cdot\,)) - p(\text{he} \,|\, do(c, \cdot\,))$.

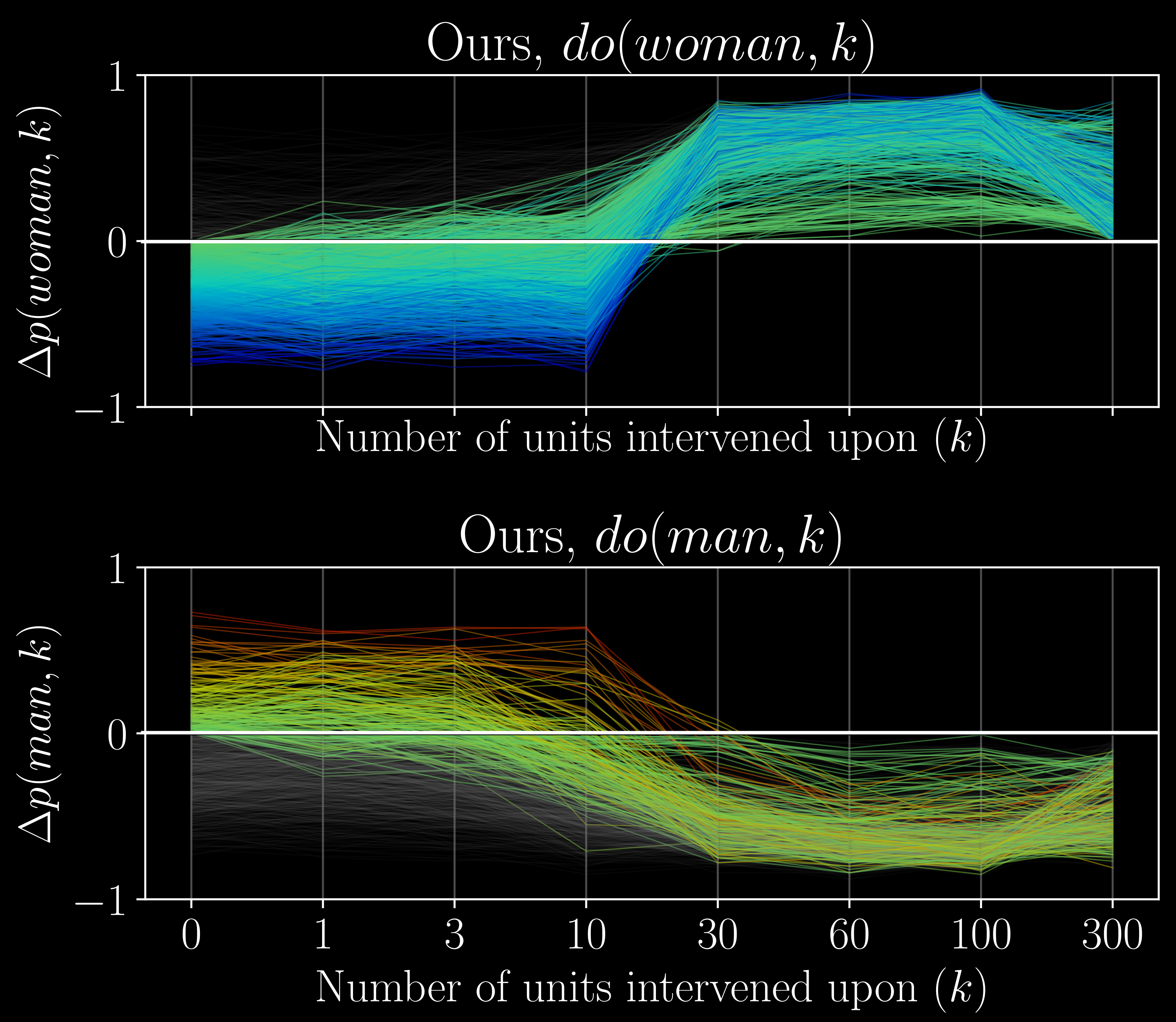
Parity point: intervention $\cdot$ so that $\Delta p(c, \cdot\,) = 0$.

[8] Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Sakenis, S., Huang, J., Singer, Y., and Shieber, S. Causal mediation analysis for interpreting neural NLP: The case of gender bias. NeurIPS, 2020.

Generative gender parity

Bias: $\Delta p(c, \cdot\,) \triangleq p(\text{she} \,|\, do(c, \cdot\,)) - p(\text{he} \,|\, do(c, \cdot\,))$.

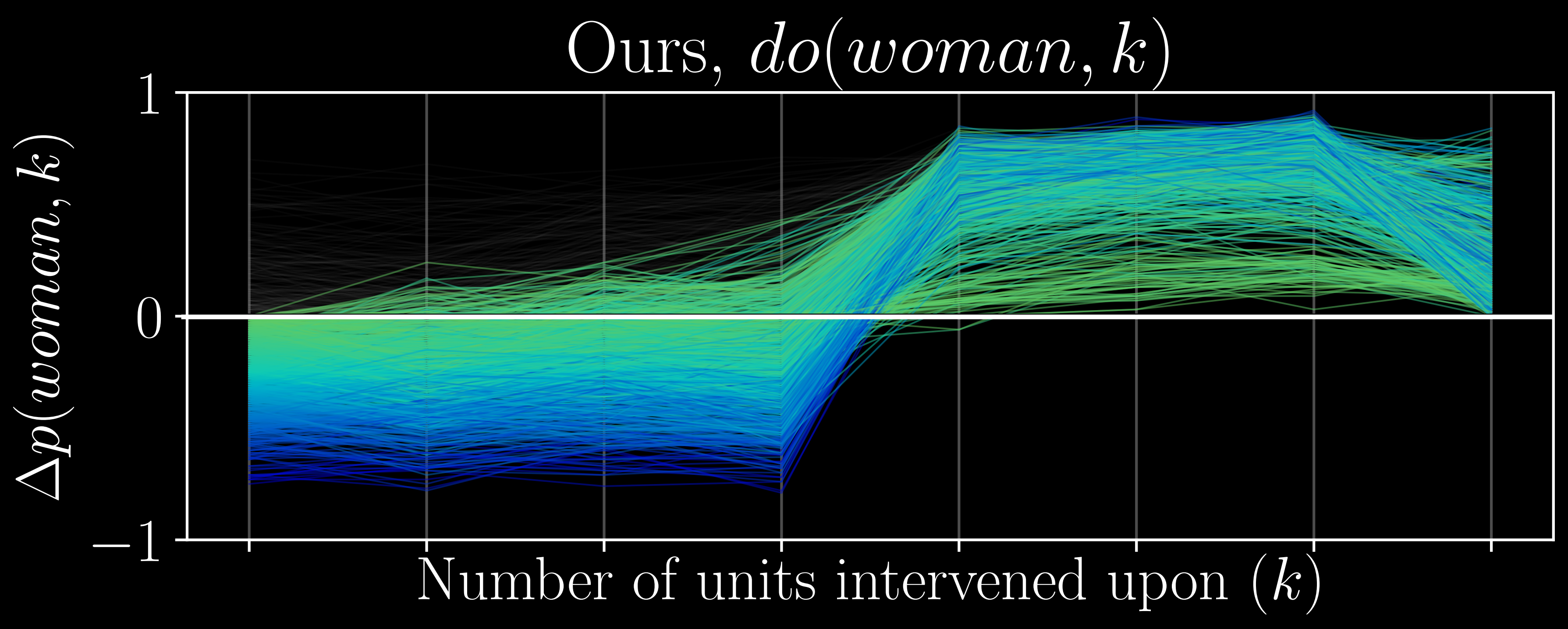
Parity point: intervention $\cdot$ so that $\Delta p(c, \cdot\,) = 0$.

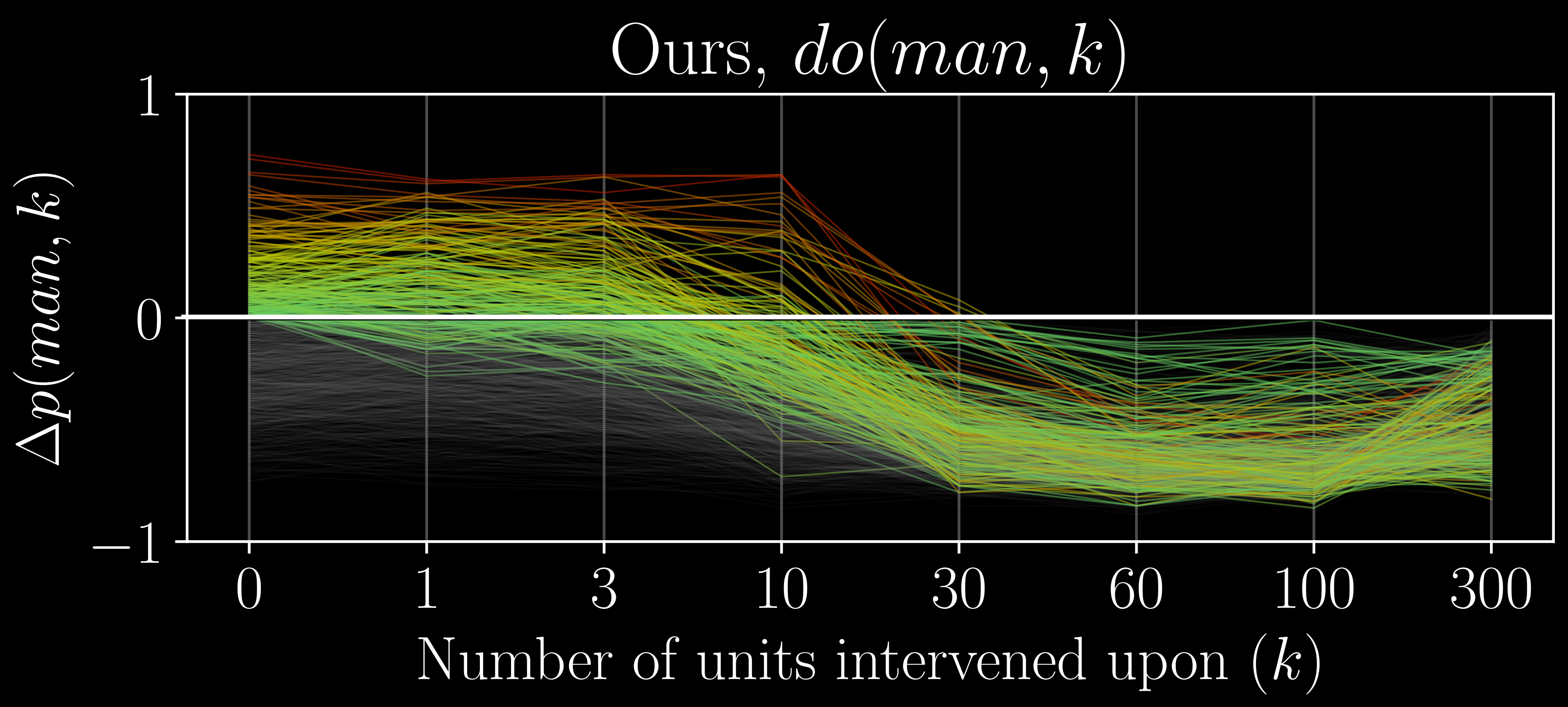


Generative gender parity

Bias: $\Delta p(c, \cdot) \triangleq p(\text{she} \mid do(c, \cdot)) - p(\text{he} \mid do(c, \cdot))$.

Parity point: intervention $\cdot$ so that $\Delta p(c, \cdot) = 0$.

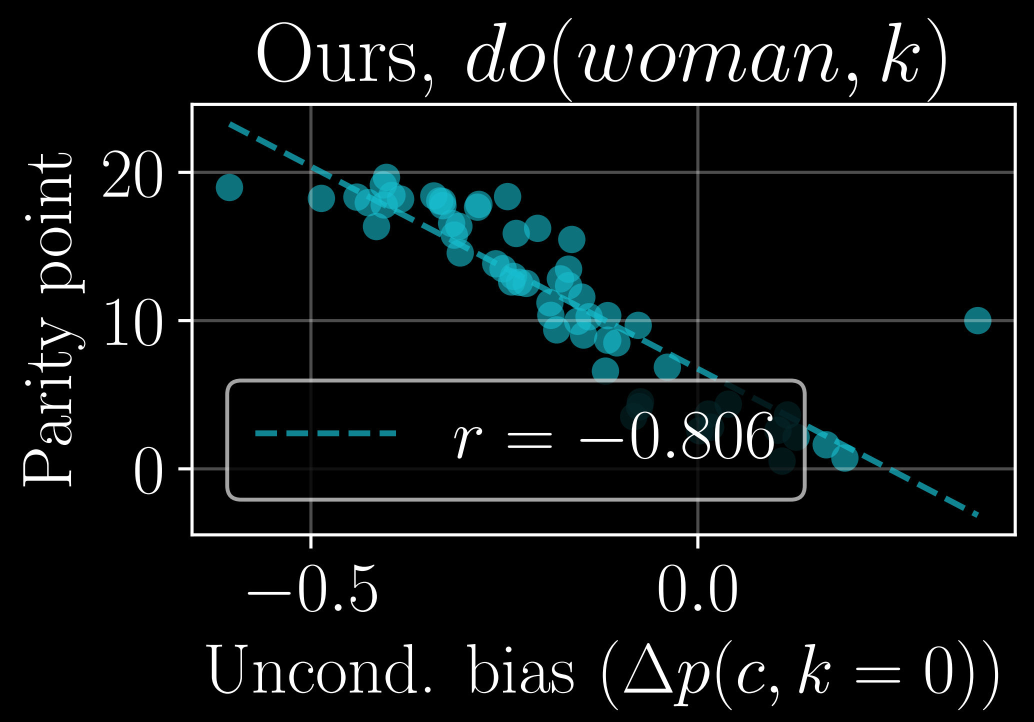


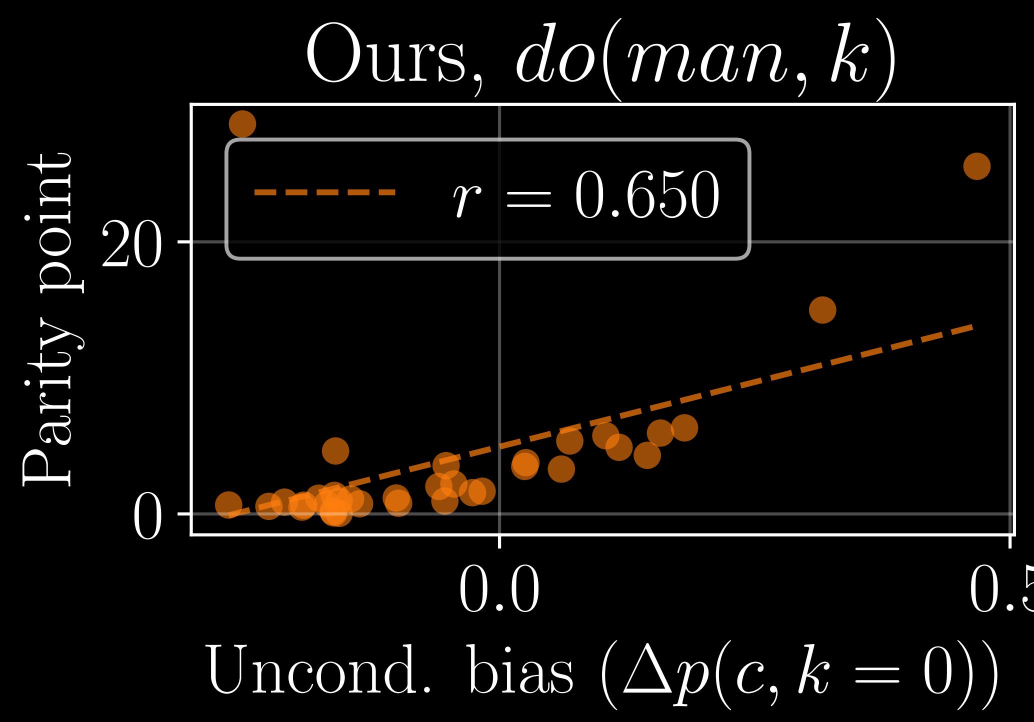


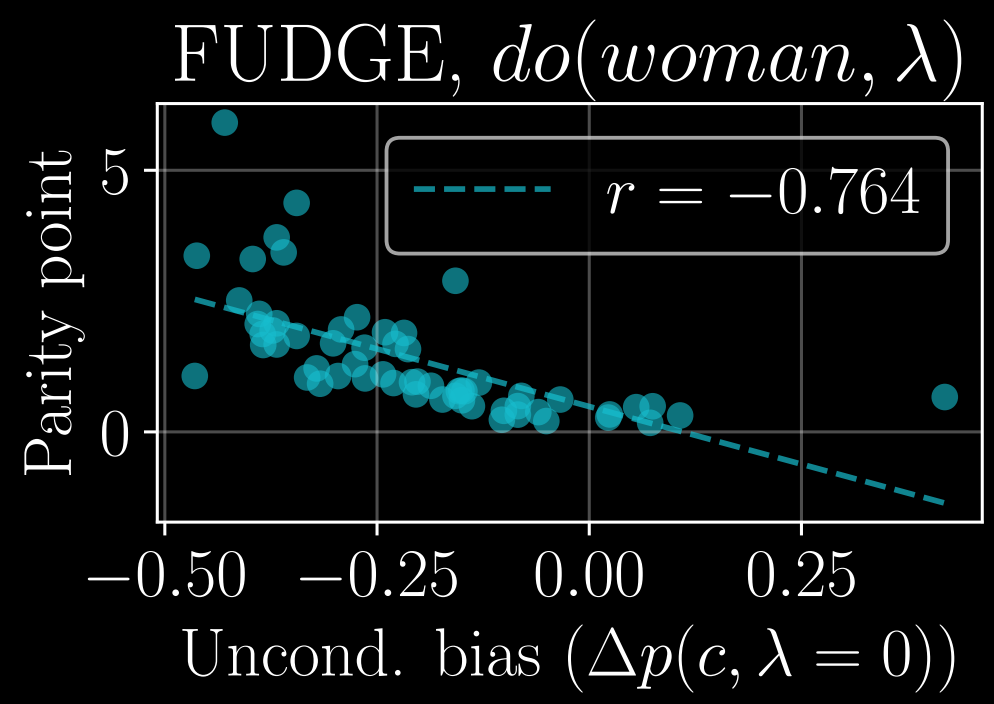
Metrics evaluated at parity: $\Delta p(c, \cdot) = 0$

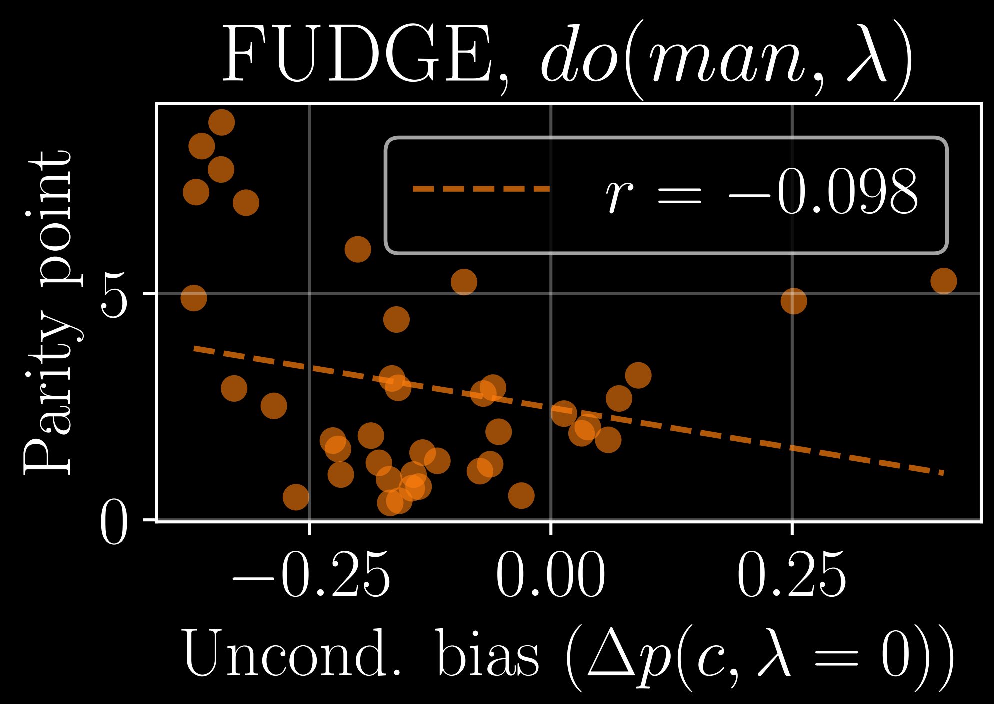
	Perplexity ⬇️	Self-BLEU - 3 ⬇️
PPLM-BoW	>250	0.46
FUDGE	85	0.30
Ours	**65**	**0.13**

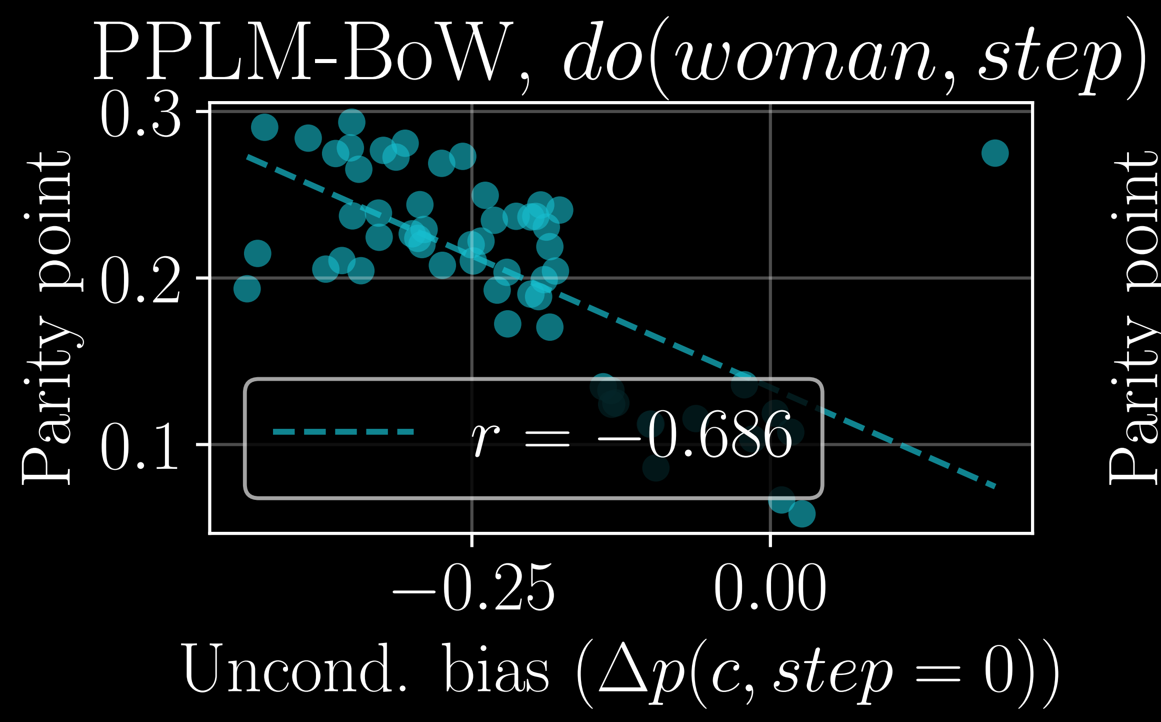
Parity vs. Model bias

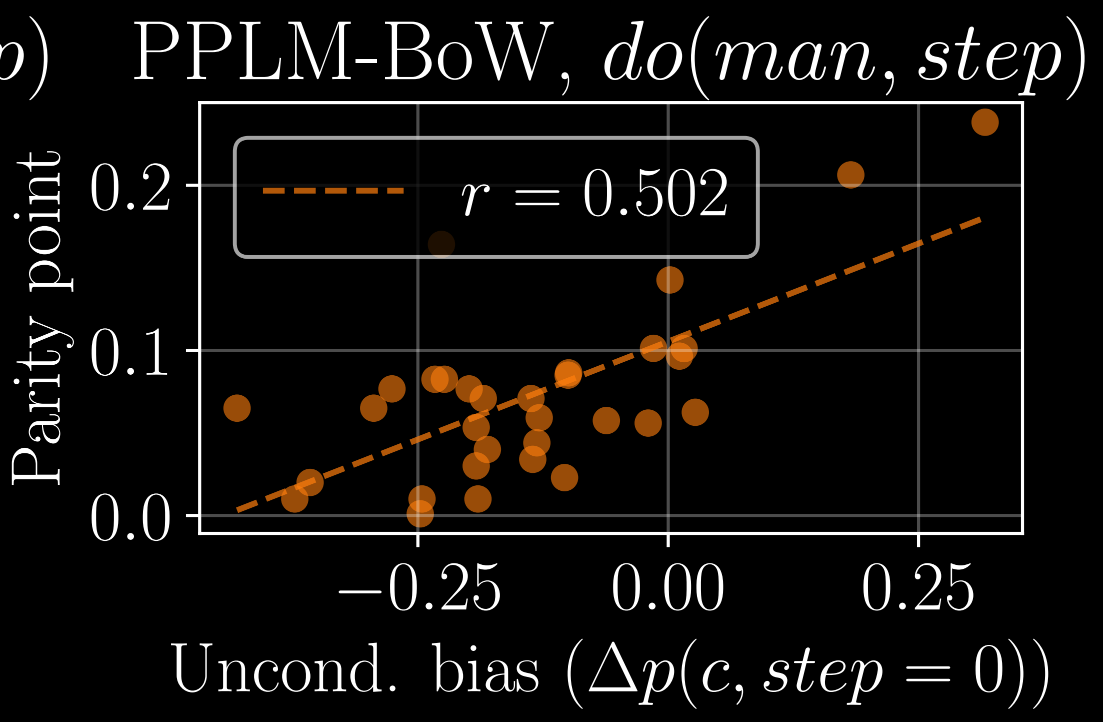




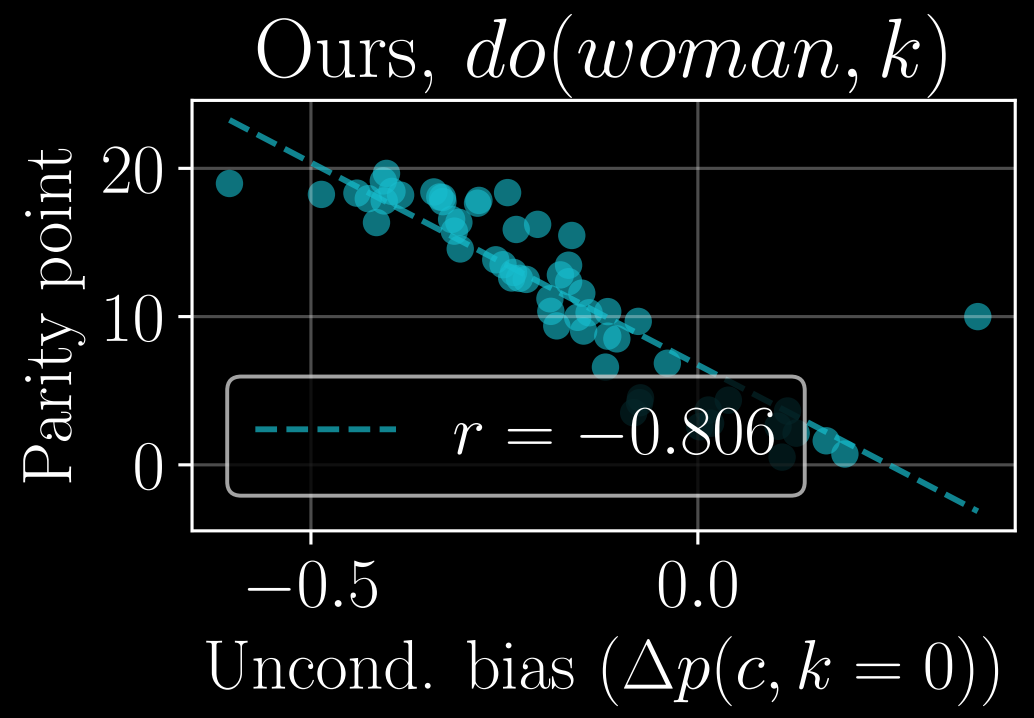


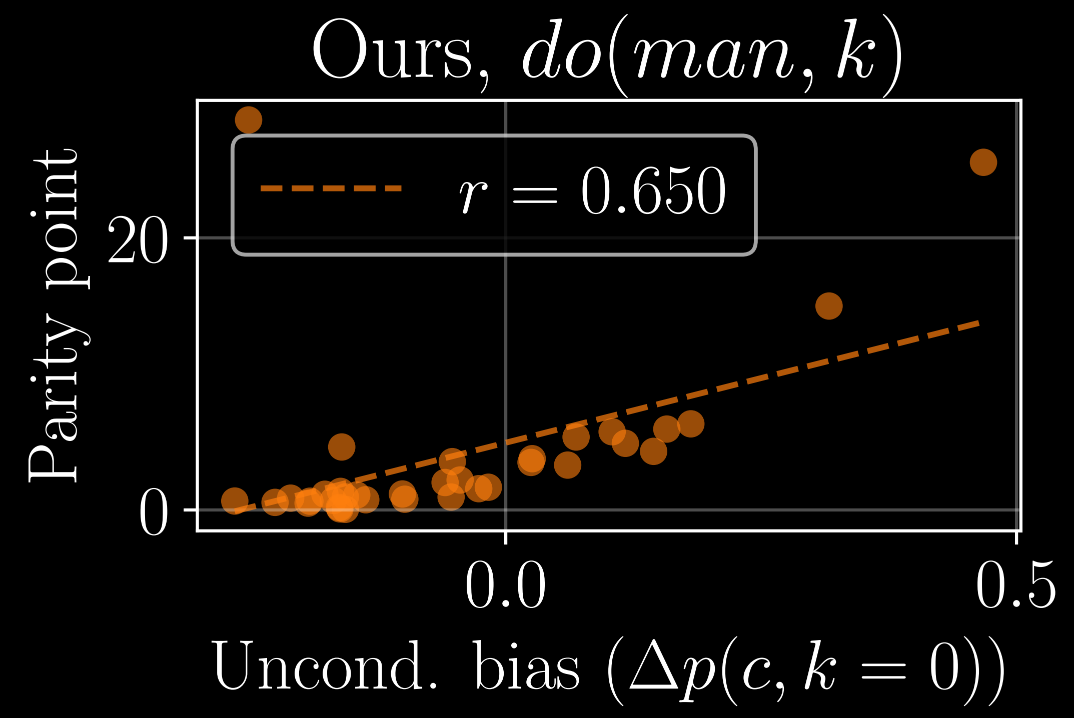


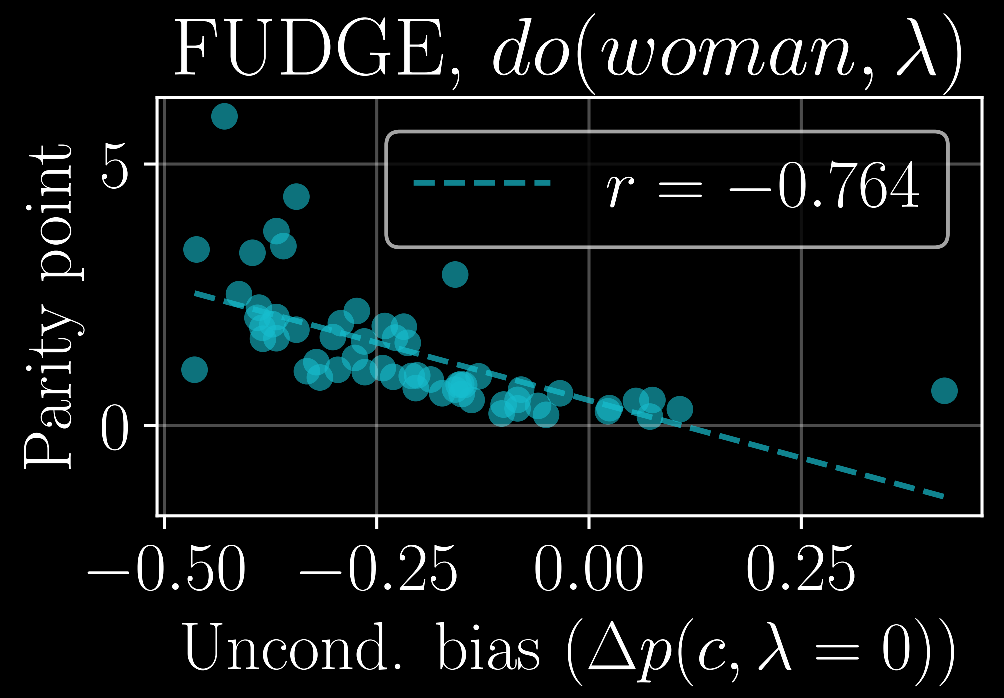


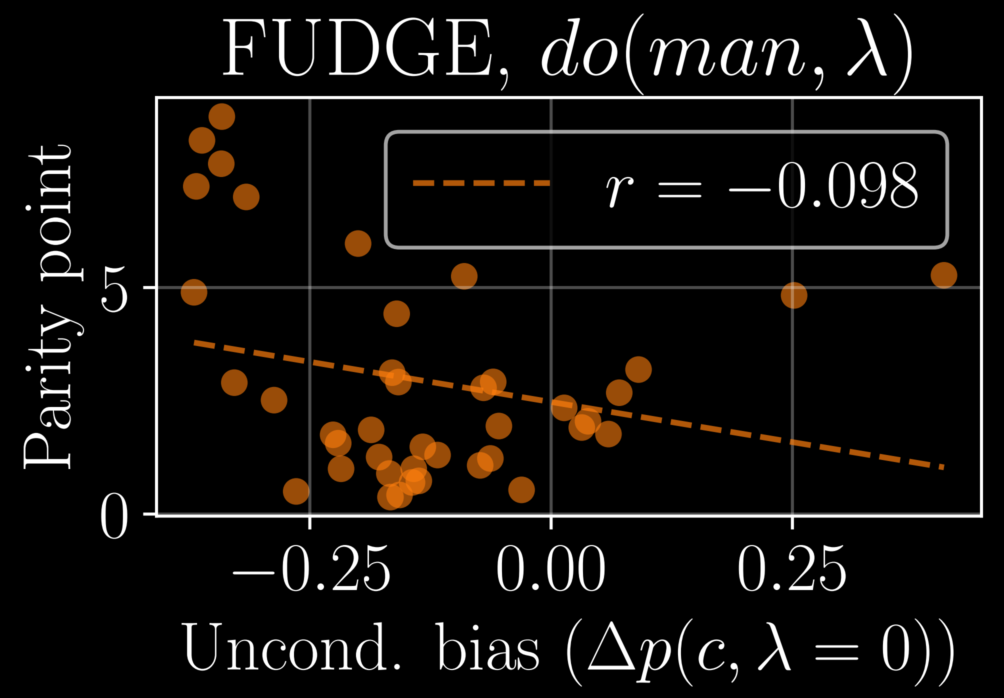


Parity vs. Model bias

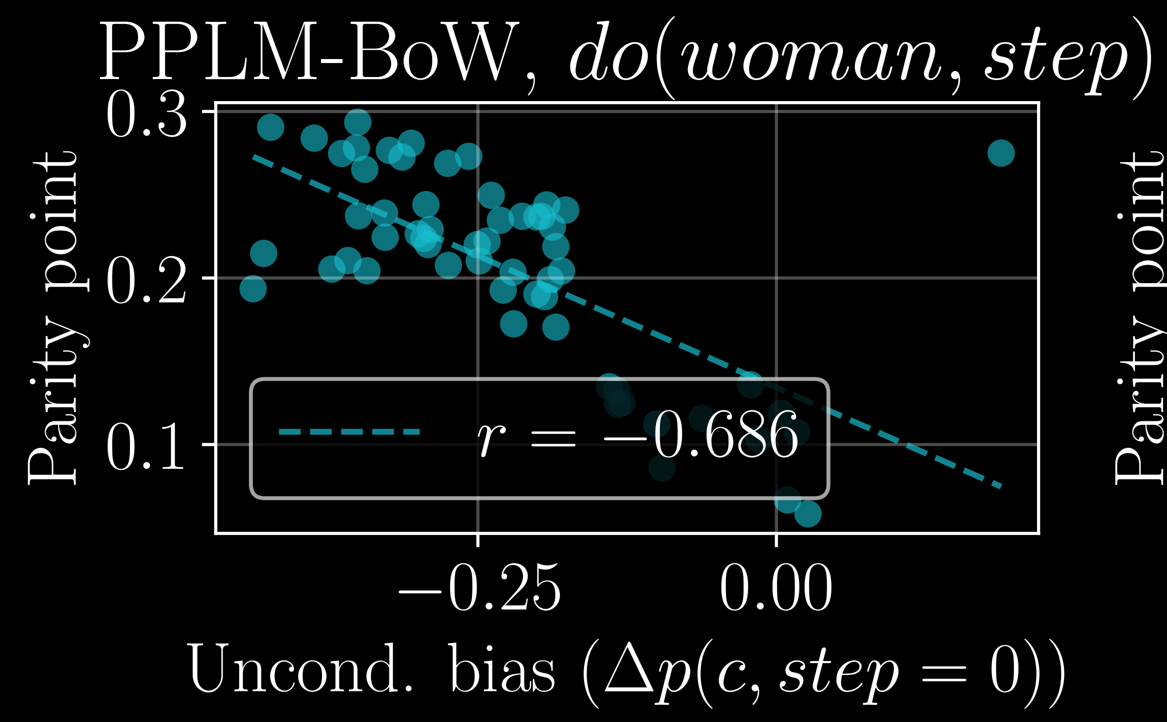


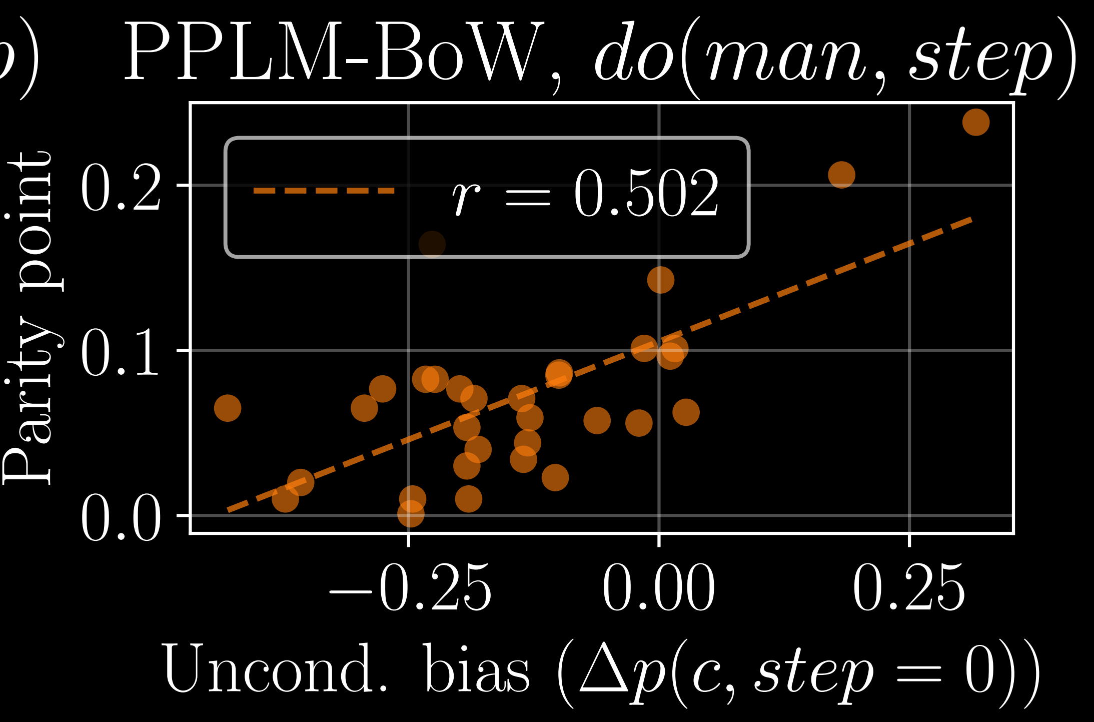






Our conditioning is better correlated with the intrinsic model bias.





Conclusion

Defined expert units, and their role in text generation.

Inference-time intervention on expert units for controlled generation.

Paper: https://arxiv.org/abs/2110.02802

Code: https://github.com/apple/ml-selfcond

Conclusion

Defined expert units, and their role in text generation.

Inference-time intervention on expert units for controlled generation.

Thorough analysis on 1037 contexts related to gender bias.

Paper: https://arxiv.org/abs/2110.02802

Code: https://github.com/apple/ml-selfcond

Conclusion

Defined expert units, and their role in text generation.

Inference-time intervention on expert units for controlled generation.

Thorough analysis on 1037 contexts related to gender bias.

- Our method achieves parity at lower perplexity and higher diversity than FUDGE and PPLM-BoW.

Paper: https://arxiv.org/abs/2110.02802

Code: https://github.com/apple/ml-selfcond

Conclusion

Defined expert units, and their role in text generation.

Inference-time intervention on expert units for controlled generation.

Thorough analysis on 1037 contexts related to gender bias.

- Our method achieves parity at lower perplexity and higher diversity than FUDGE and PPLM-BoW.
- Our conditioning is correlated with the model bias.

Paper: https://arxiv.org/abs/2110.02802

Code: https://github.com/apple/ml-selfcond

Conclusion

Defined expert units, and their role in text generation.

Inference-time intervention on expert units for controlled generation.

Thorough analysis on 1037 contexts related to gender bias.

- Our method achieves parity at lower perplexity and higher diversity than FUDGE and PPLM-BoW.
- Our conditioning is correlated with the model bias.

Open ended conditioned generation (in paper)

Paper: https://arxiv.org/abs/2110.02802

Code: https://github.com/apple/ml-selfcond