Fairness-Aware Structured Pruning in Transformers

Abdelrahman Zayed, Goncalo Mordido, Samira Shabanian, Ioanna Baldini, Sarath Chandar
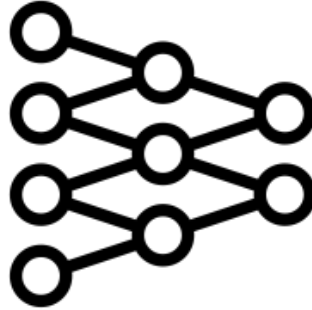AAAI 2024

What is bias?

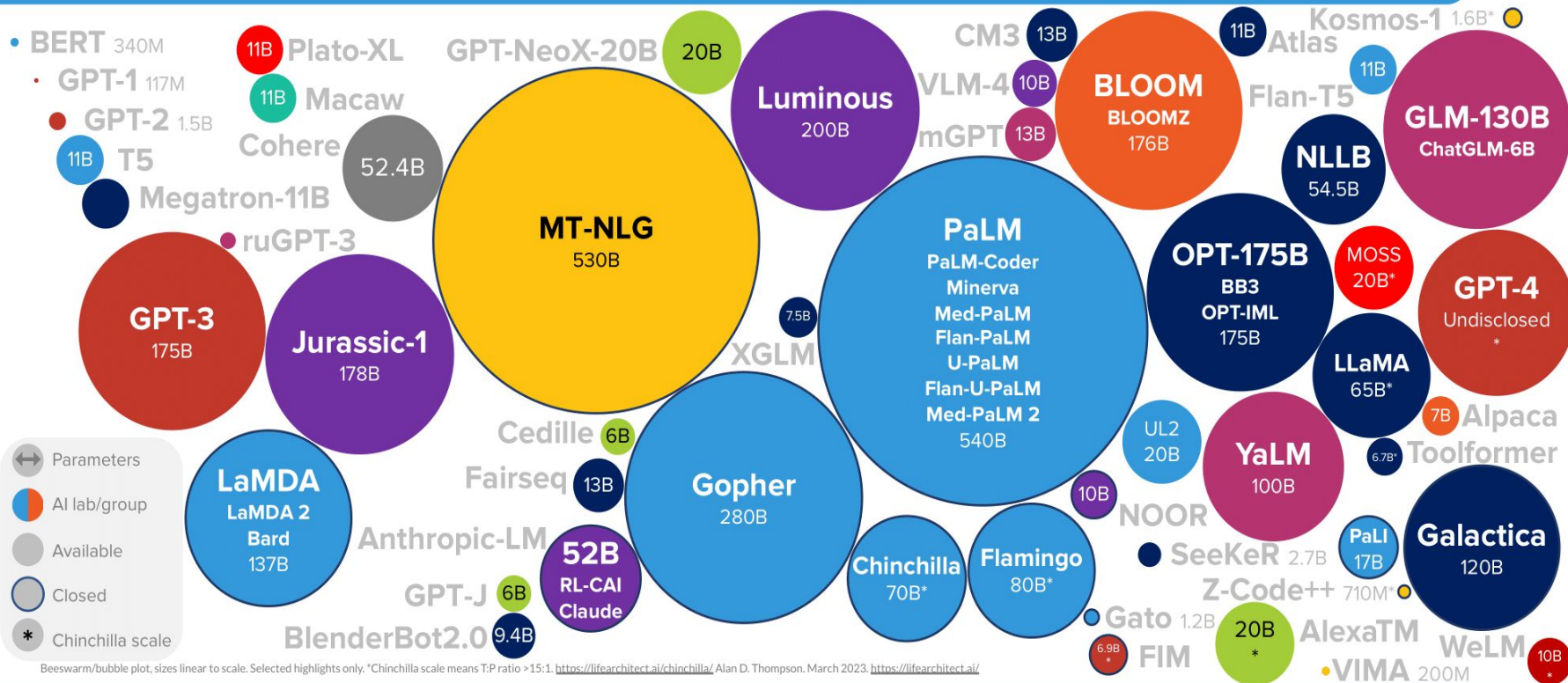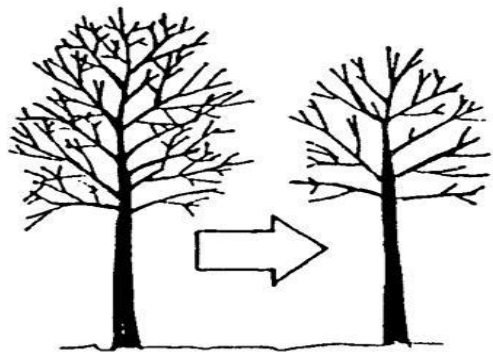| Abdel	Nora
Ph.D. from Mila 6 years of research experience	Ph.D. from Mila 6 years of research experience
Keep resume	Discard resume

How do we mitigate bias?

Pre-processing
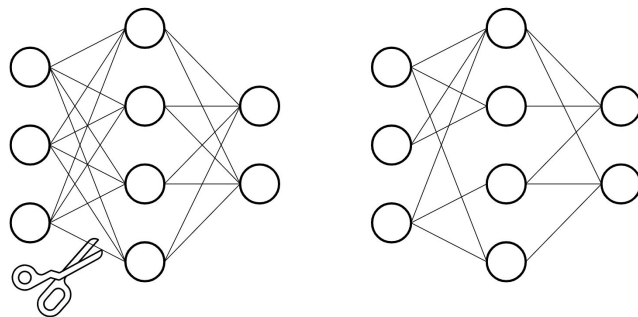
In-processing

Post-processing

LANGUAGE MODEL SIZES TO MAR/2023

Beeswarm/bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means T:P ratio >15:1. https://lifearchitect.ai/chinchilla/ Alan D. Thompson. March 2023. https://lifearchitect.ai/

LifeArchitect.ai/models

4

Can pruning help?

Structured

Unstructured

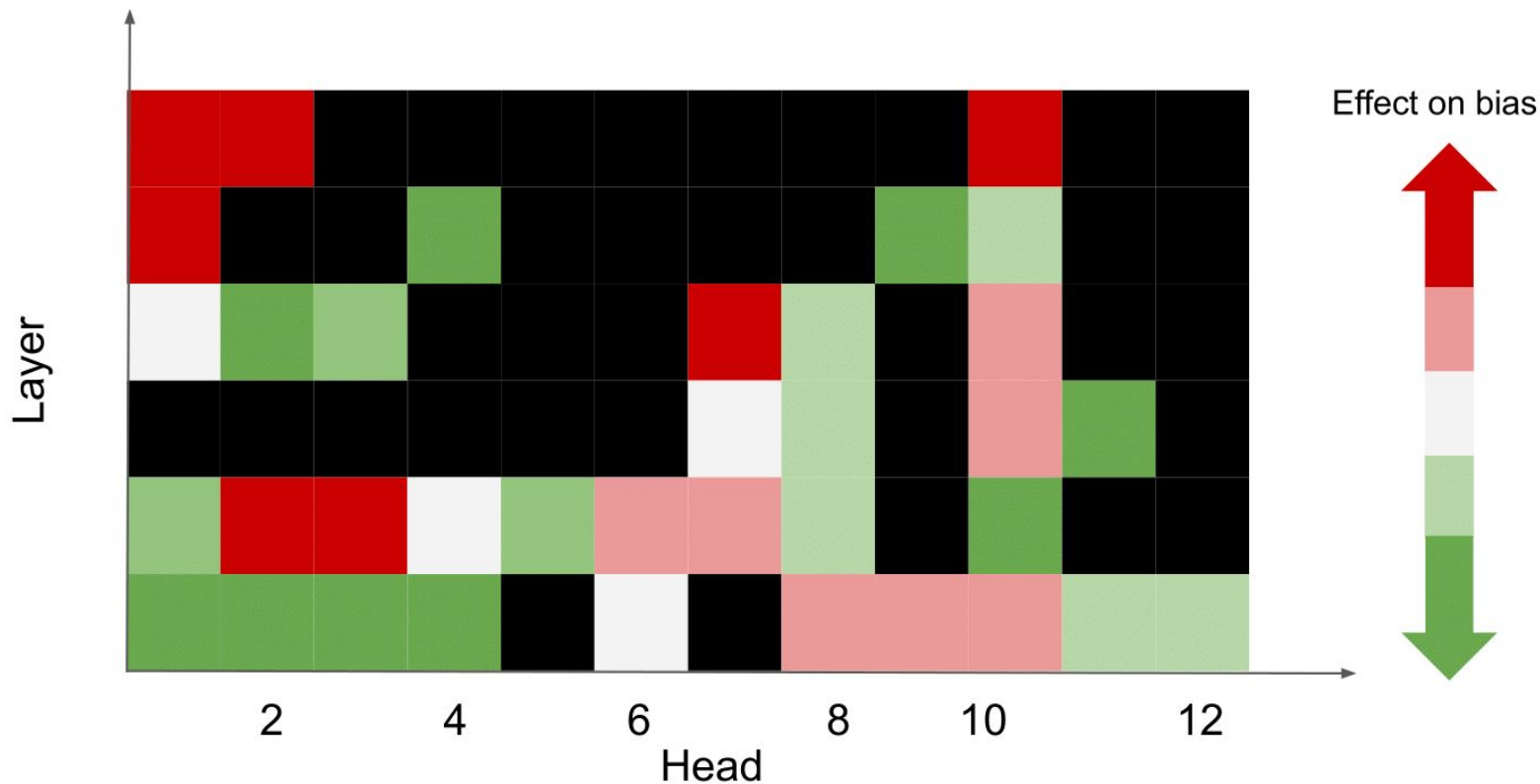# How does pruning affect fairness?
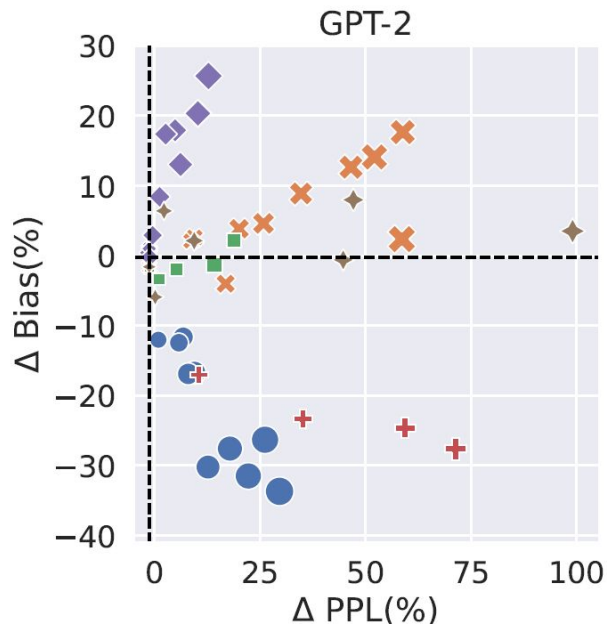
Let's do a quick test!

$$Effect\ of\ head\ h\ on\ bias = bias_{with\ head\ h} - bias_{without\ head\ h}$$

$$Effect\ of\ head\ h\ on\ ppl = ppl_{with\ head\ h} - ppl_{without\ head\ h}$$
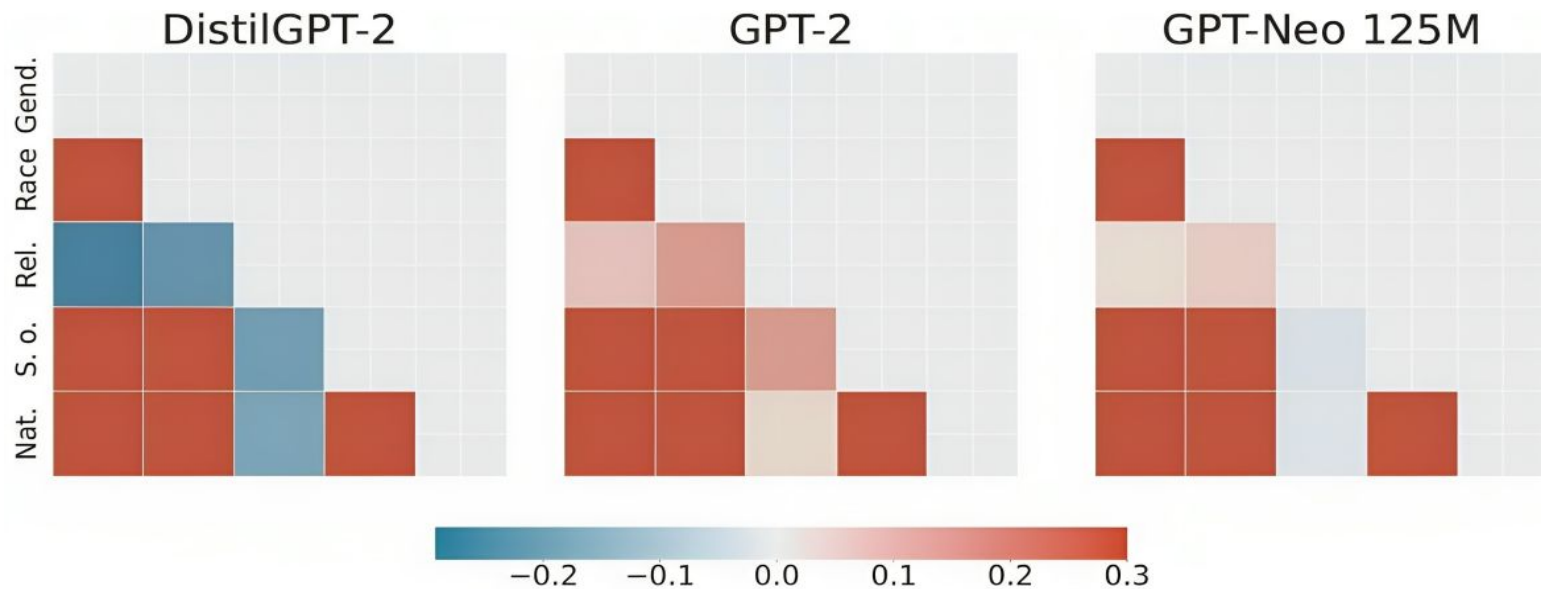
Fairness-Aware Structured Pruning (FASP)

How does FASP compare to other pruning methods?

Does gender bias mitigation affect other biases?

To summarize

Some heads are responsible for **bias** and pruning them improves *fairness*

Thanks for listening!