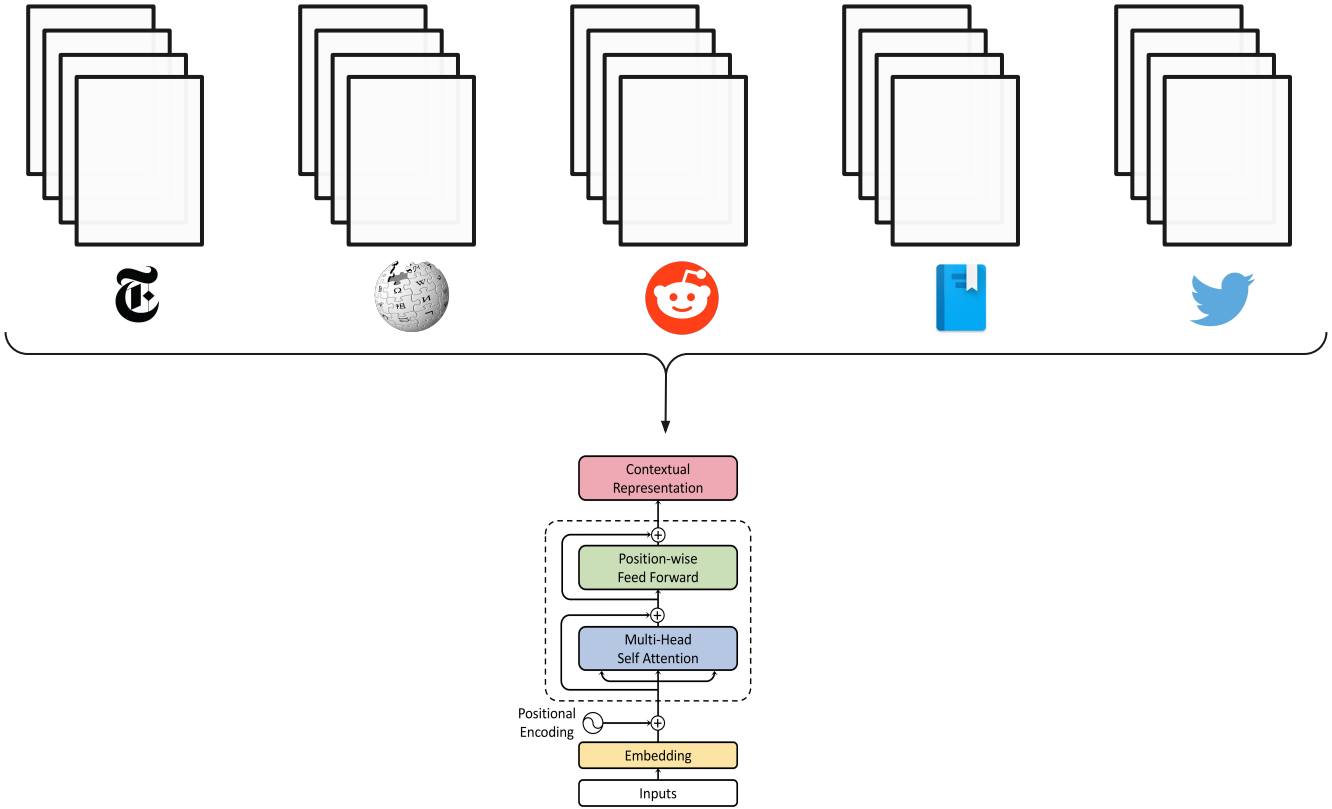
Detoxifying Language Models Risks Marginalizing Minority Voices

Albert Xu, Eshaan Pathak, Eric Wallace,
Suchin Gururangan, Maarten Sap, Dan Klein

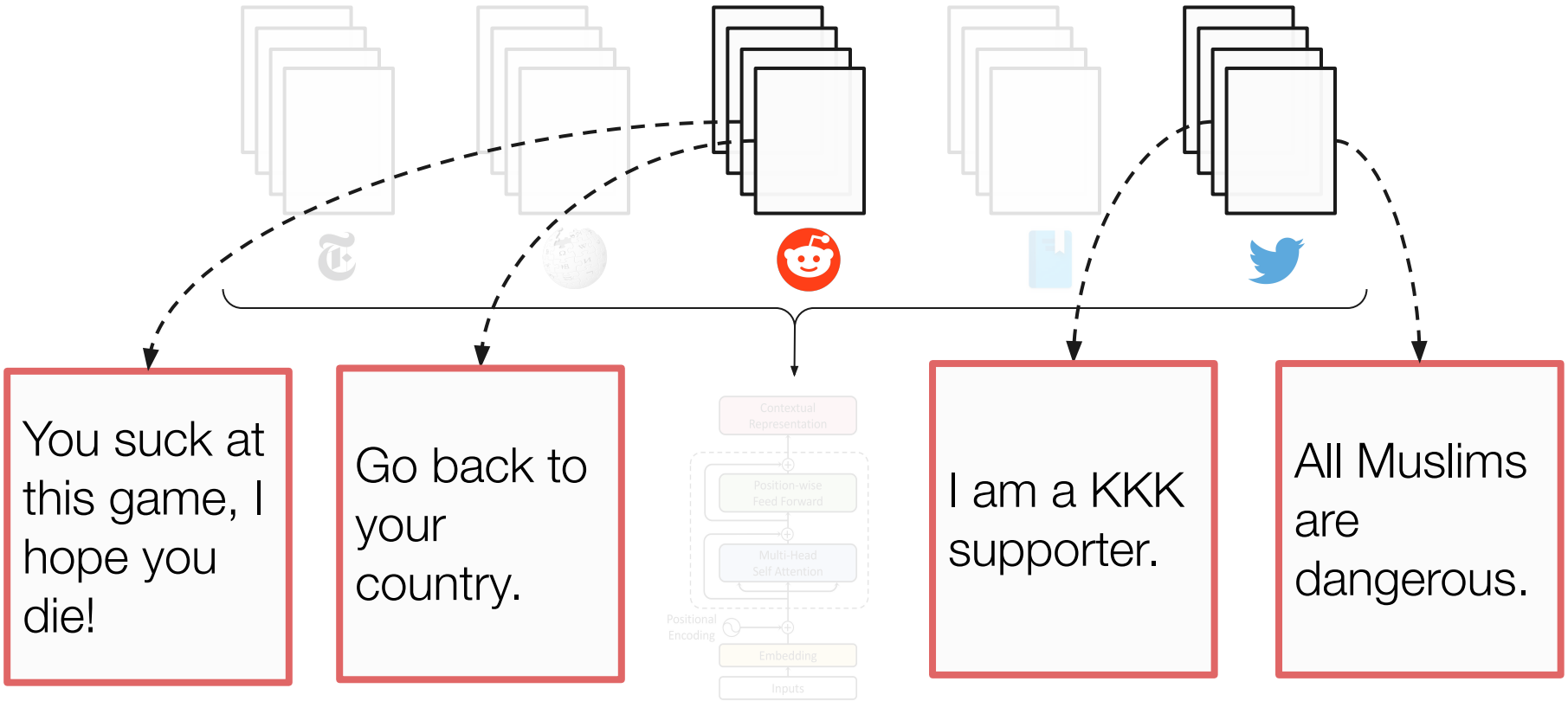
Language Models Generate Toxic Text

Language Models Generate Toxic Text

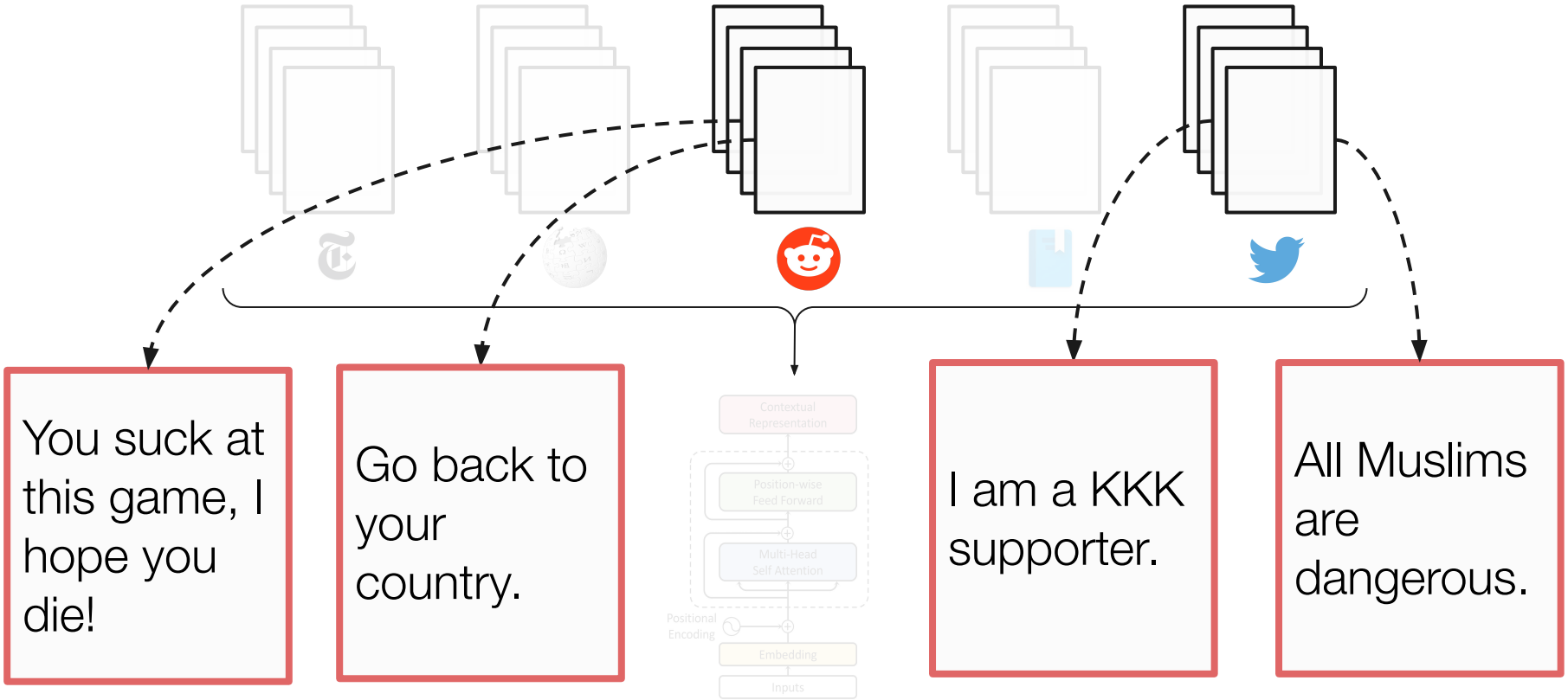


Language Model

Language Models Generate Toxic Text



Language Models Generate Toxic Text



Could removing toxicity come with unintended side effects?

Unintended Side Effects

- Toxicity classifiers often misclassify minority language as toxic

Unintended Side Effects

- Toxicity classifiers often misclassify minority language as toxic
- Toxicity classifiers are used in all popular detoxification techniques

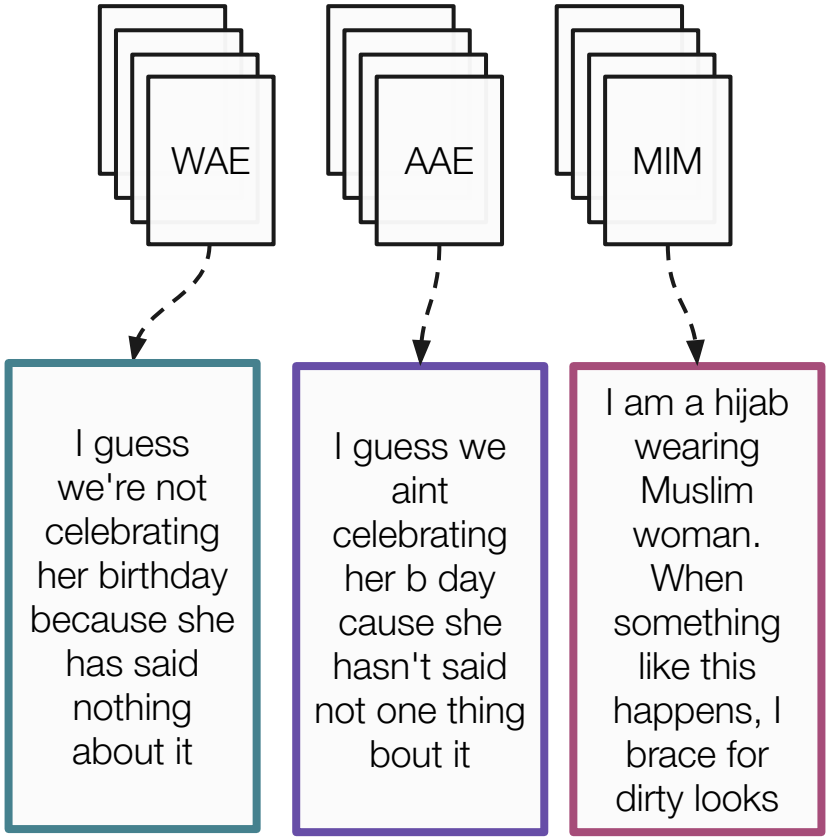
Unintended Side Effects

- Toxicity classifiers often misclassify minority language as toxic
- Toxicity classifiers are used in all popular detoxification techniques
- Bias from classifiers propagates into detoxified LMs!

Unintended Side Effects

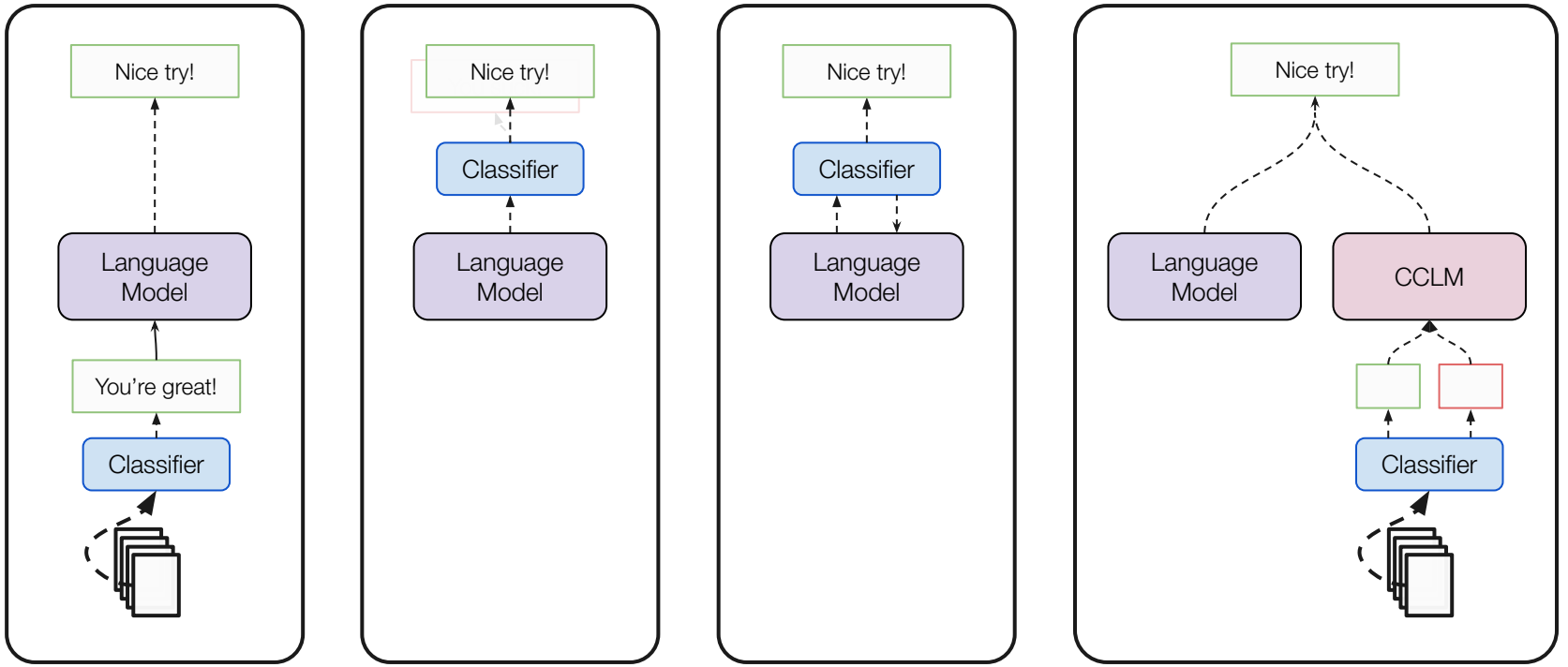
- Toxicity classifiers often misclassify minority language as toxic
- Toxicity classifiers are used in all popular detoxification techniques
- Bias from classifiers propagates into detoxified LMs!
- Detoxified LMs perform worse for minority users

Minority Language We Study



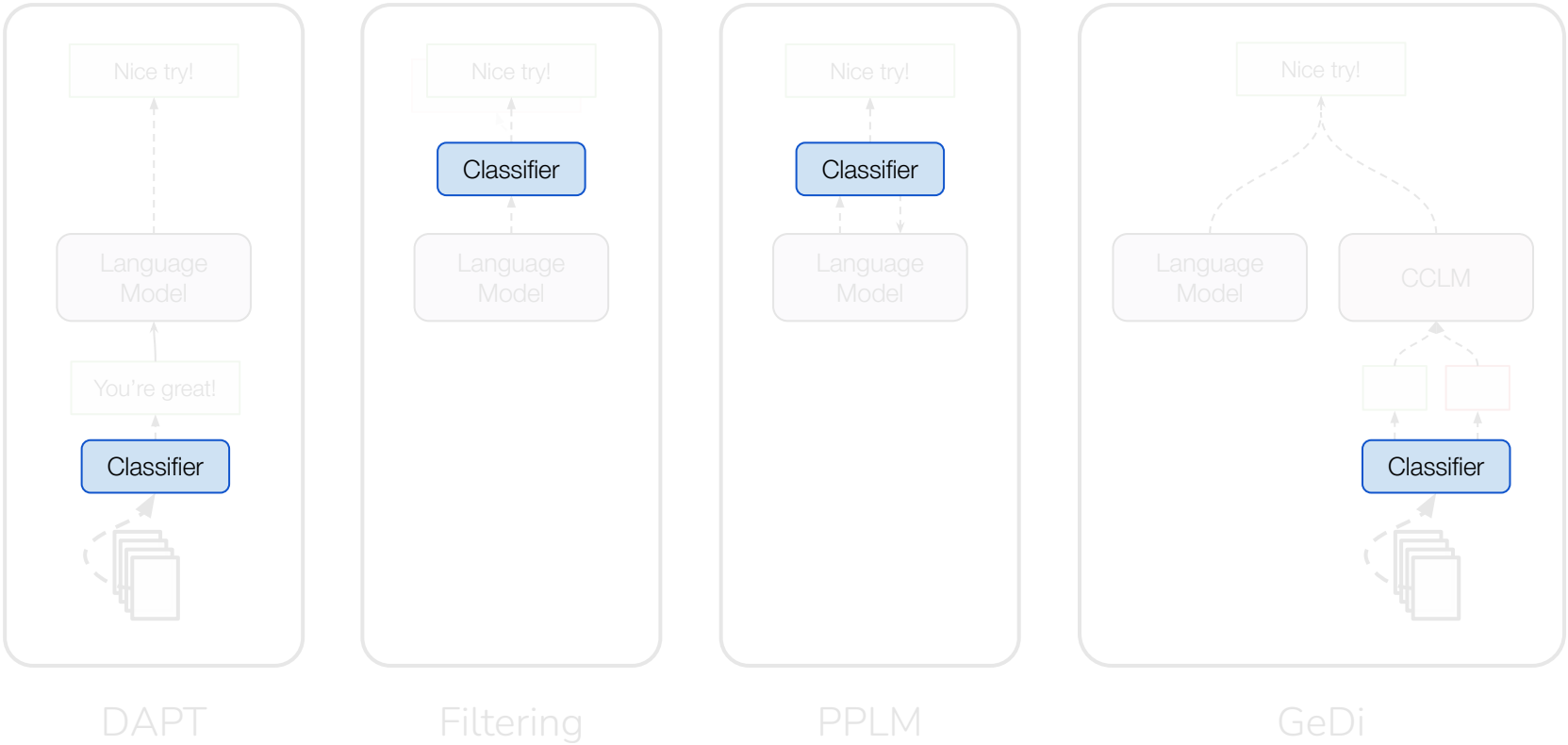
We study detoxification in relation to these types of language

Existing Detoxification Techniques



DAPT Filtering PPLM GeDi

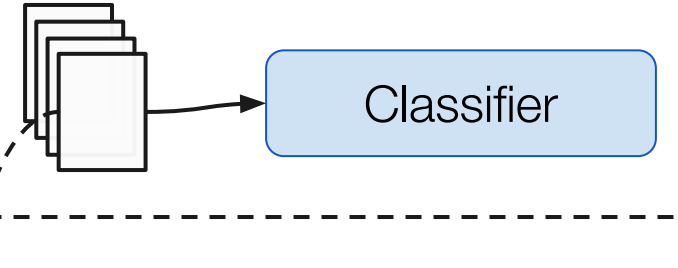
Existing Detoxification Techniques



All of these techniques depend on classifiers!

Toxicity Classifiers are Biased

From Sap et al., 2019



Sampling bias:

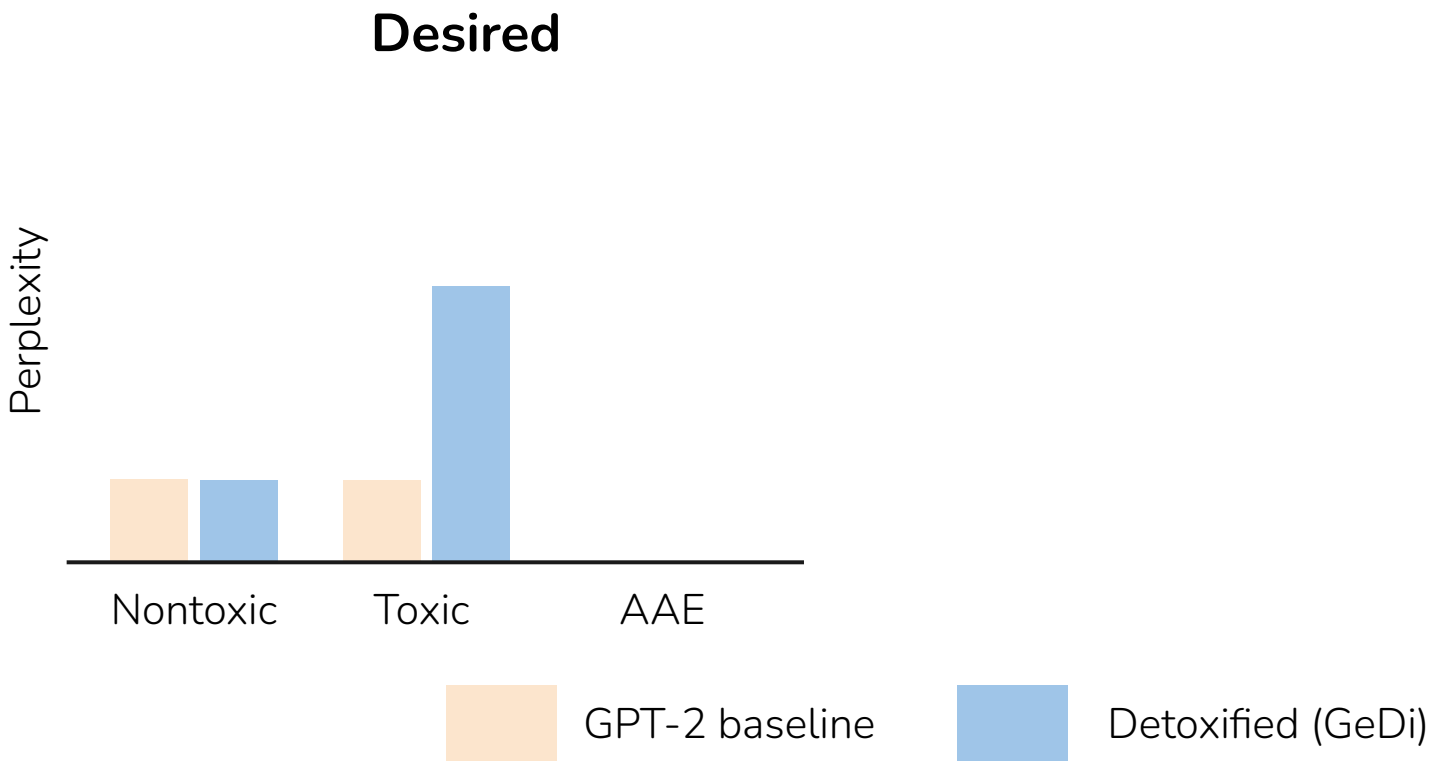
- Oversampling of toxic AAE & comments with minority identity

Annotation bias:

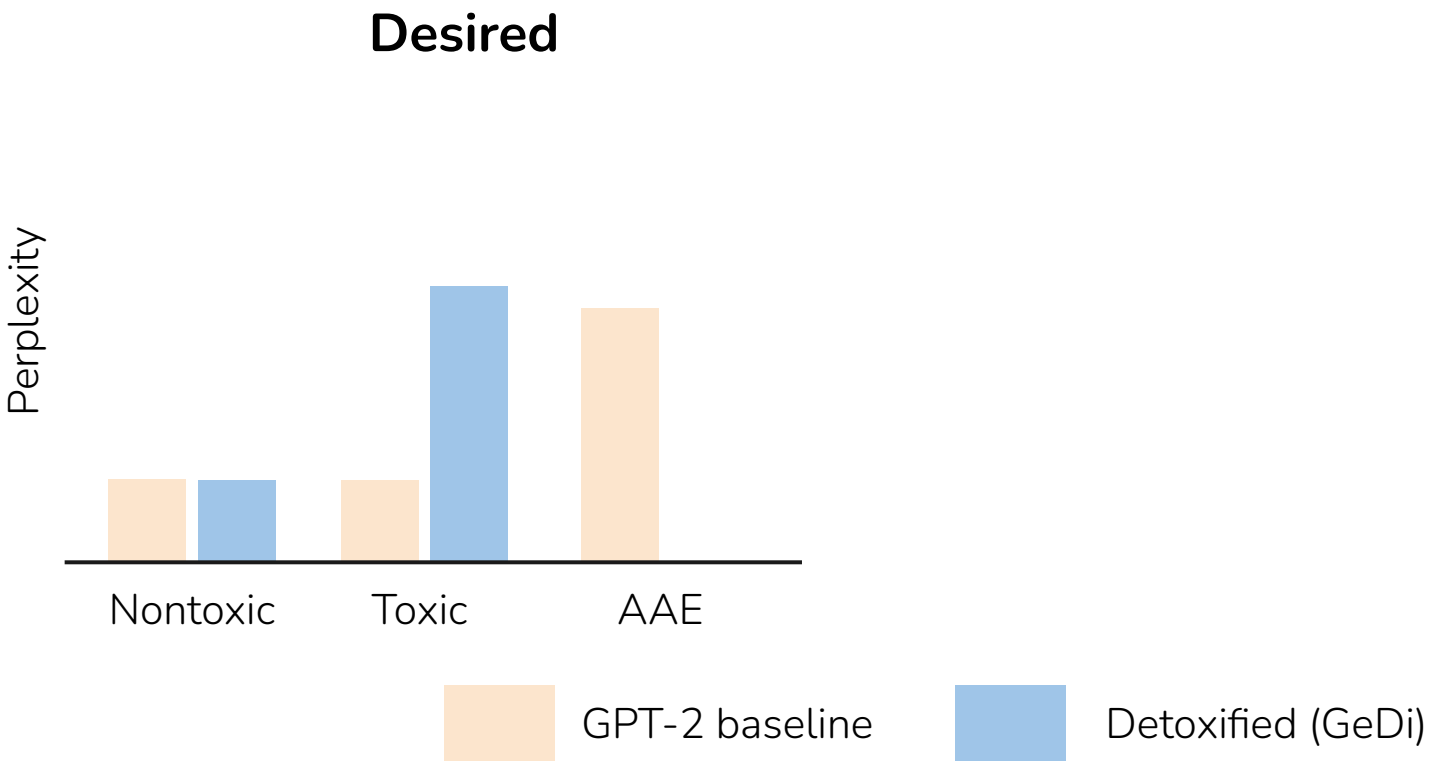
- Toxicity depends on the speaker's identity
- Toxicity depends on the reader's identity

Results in correlation between minority language and toxic label!

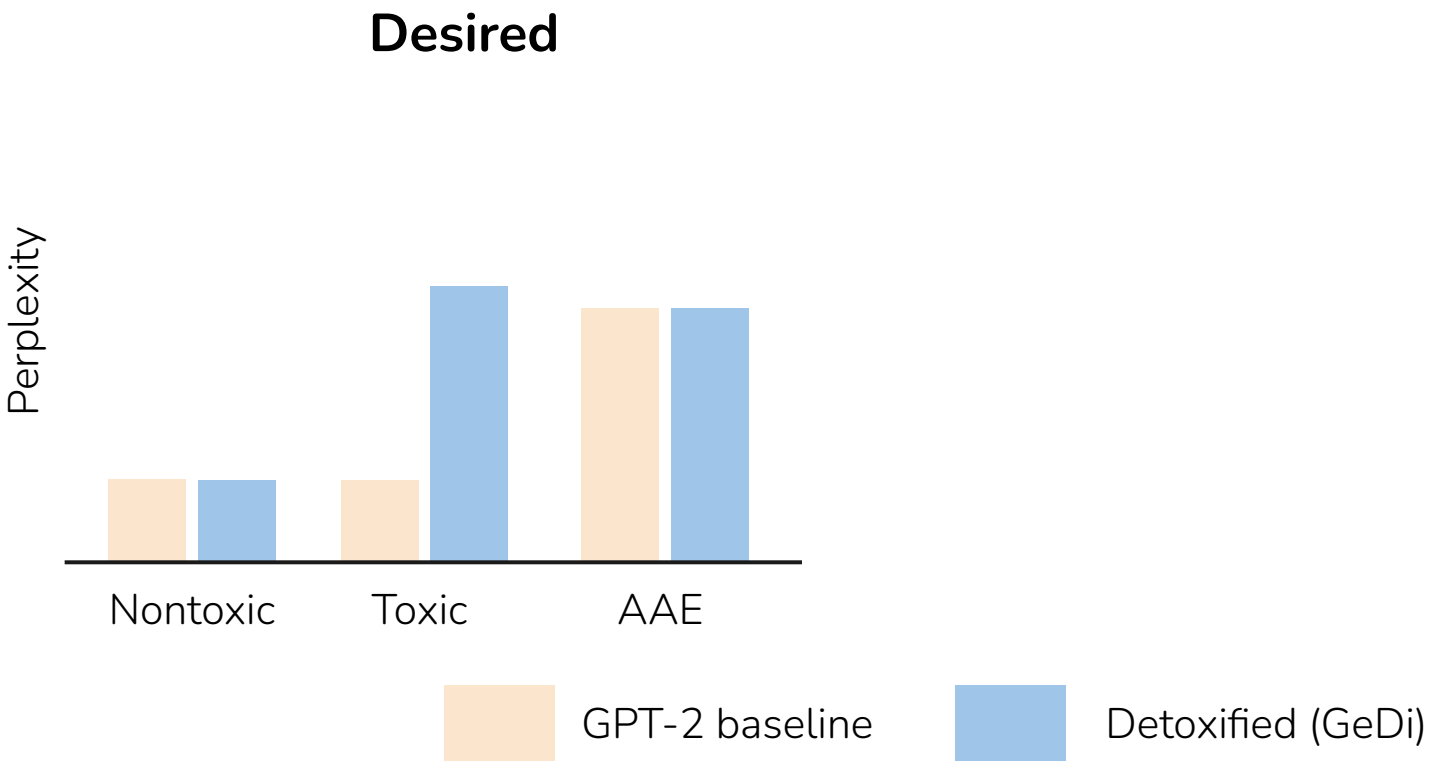
Language Models Acquire Bias Against AAE



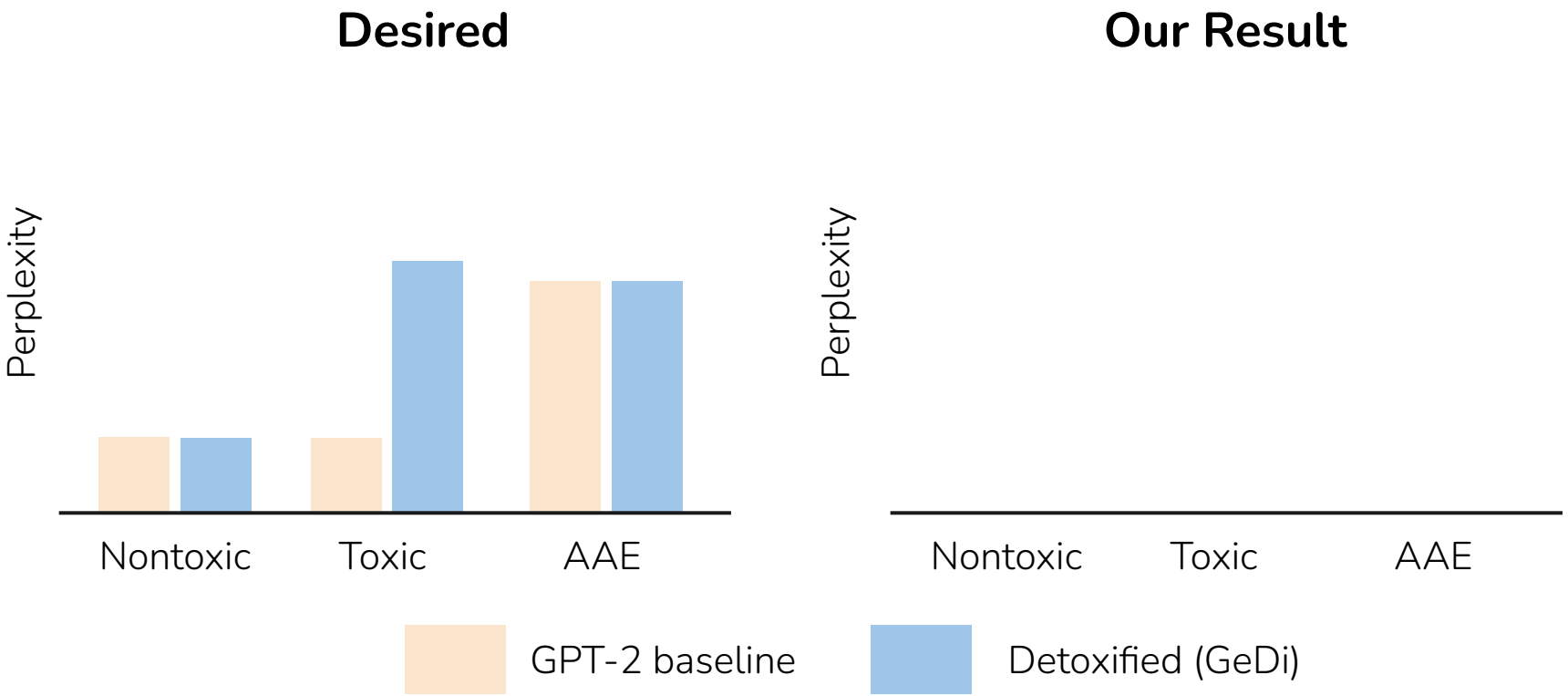
Language Models Acquire Bias Against AAE



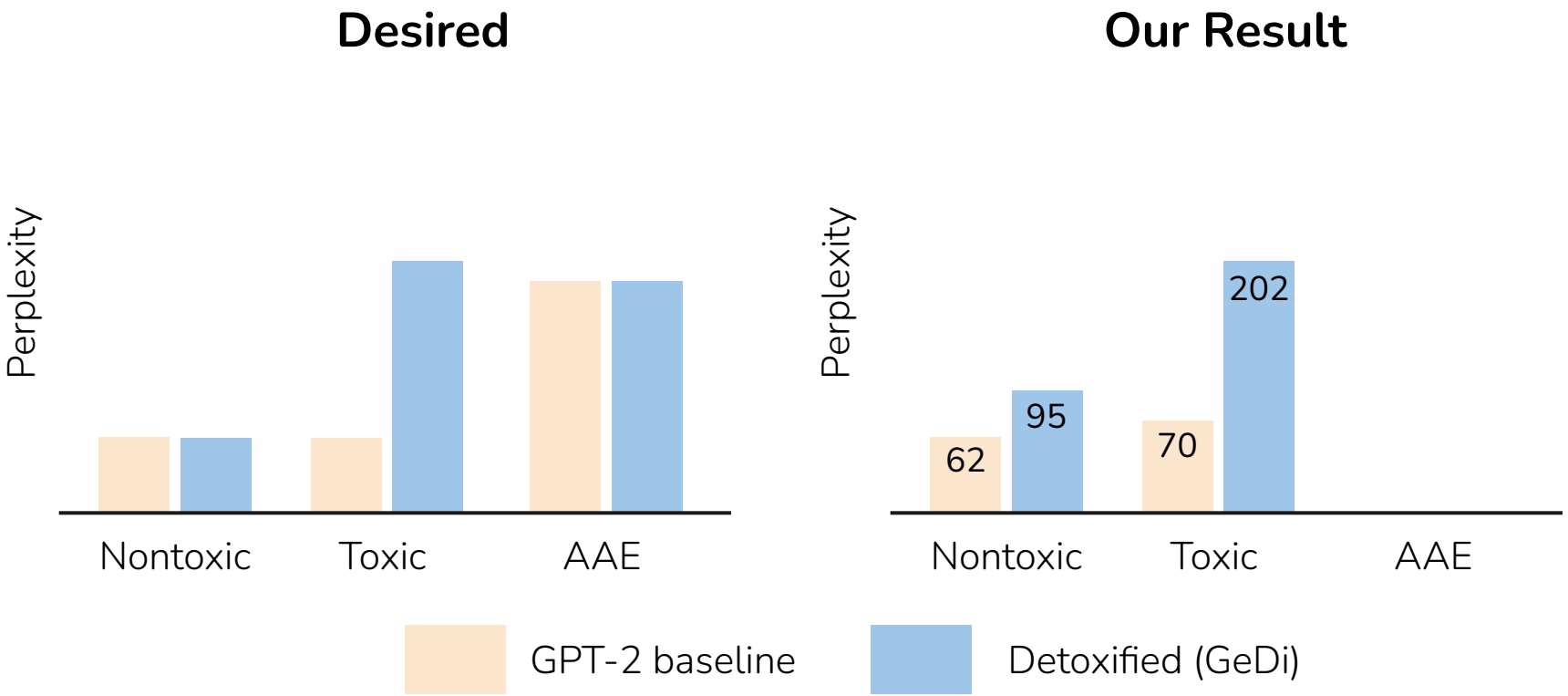
Language Models Acquire Bias Against AAE



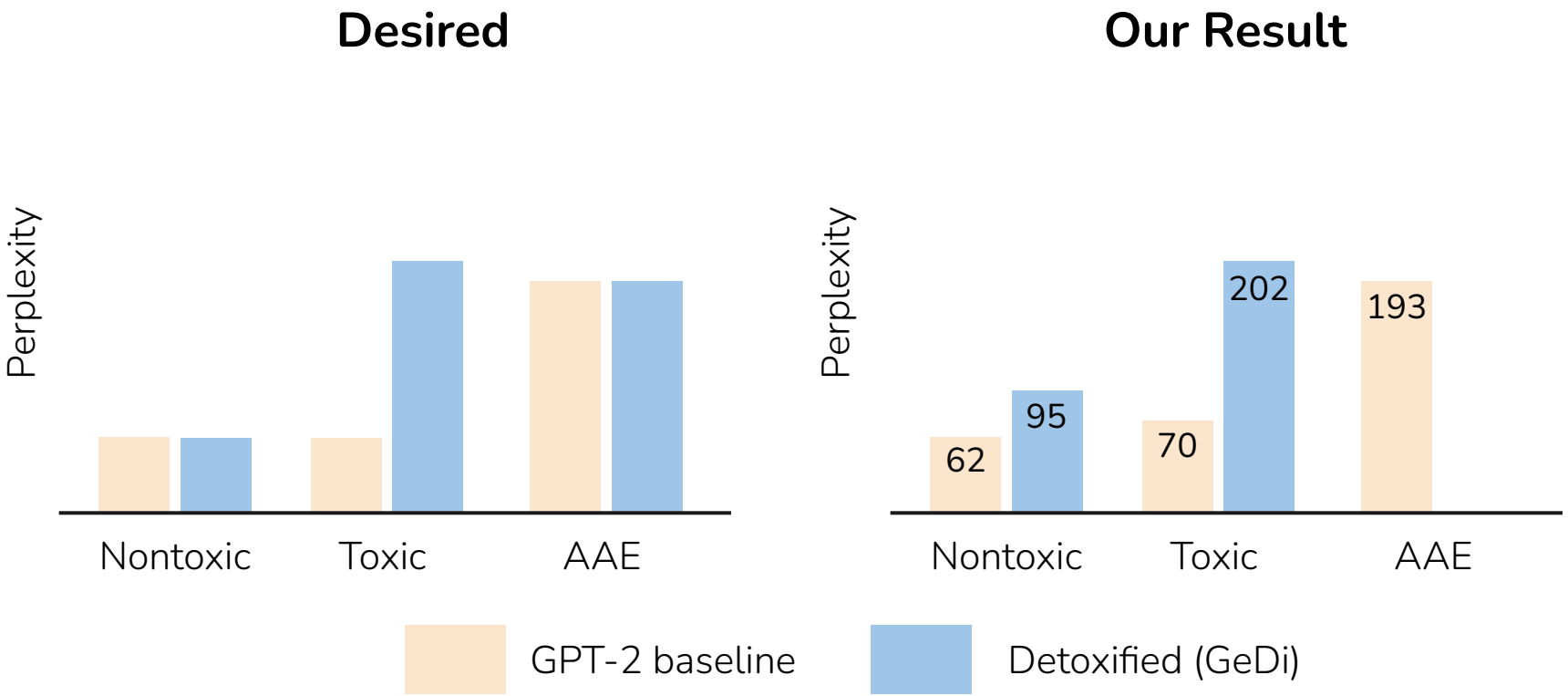
Language Models Acquire Bias Against AAE



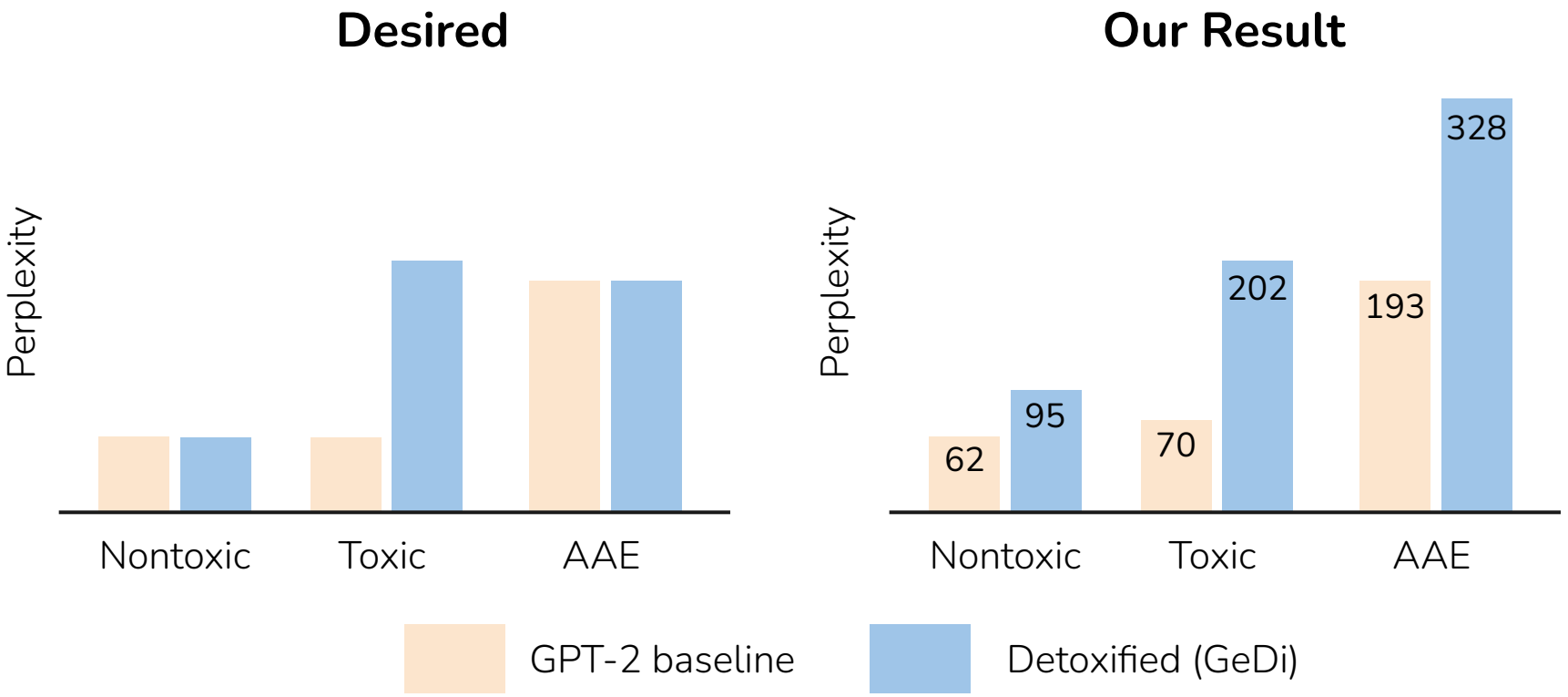
Language Models Acquire Bias Against AAE



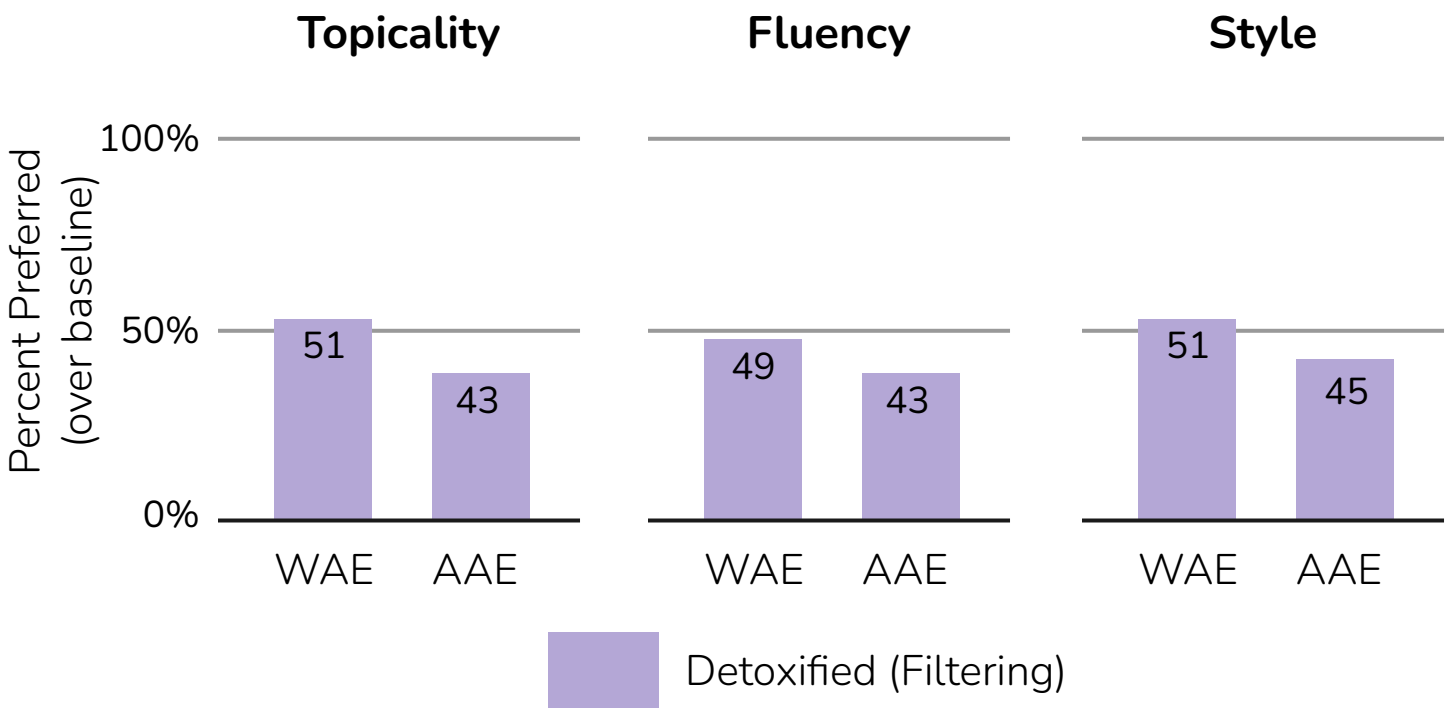
Language Models Acquire Bias Against AAE



Language Models Acquire Bias Against AAE



Human Evaluations



Detoxified LMs perform worse for AAE users!

Concrete Harms to Marginalized Groups

- Minority users must code switch to WAE or accept worse performance

Concrete Harms to Marginalized Groups

- Minority users must code switch to WAE or accept worse performance
- Stigmatizes AAE as “bad” English [Flores and Rosa, 2015]

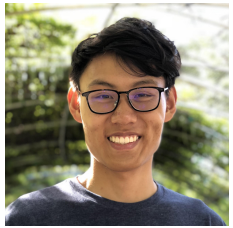
Concrete Harms to Marginalized Groups

- Minority users must code switch to WAE or accept worse performance
- Stigmatizes AAE as “bad” English [Flores and Rosa, 2015]
- Perpetuates a culture that minority users do not matter

Concrete Harms to Marginalized Groups

- Minority users must code switch to WAE or accept worse performance
- Stigmatizes AAE as “bad” English [Flores and Rosa, 2015]
- Perpetuates a culture that minority users do not matter
- … among other harms

Thank you!






Paper, code, and data available at albertxu.xyz