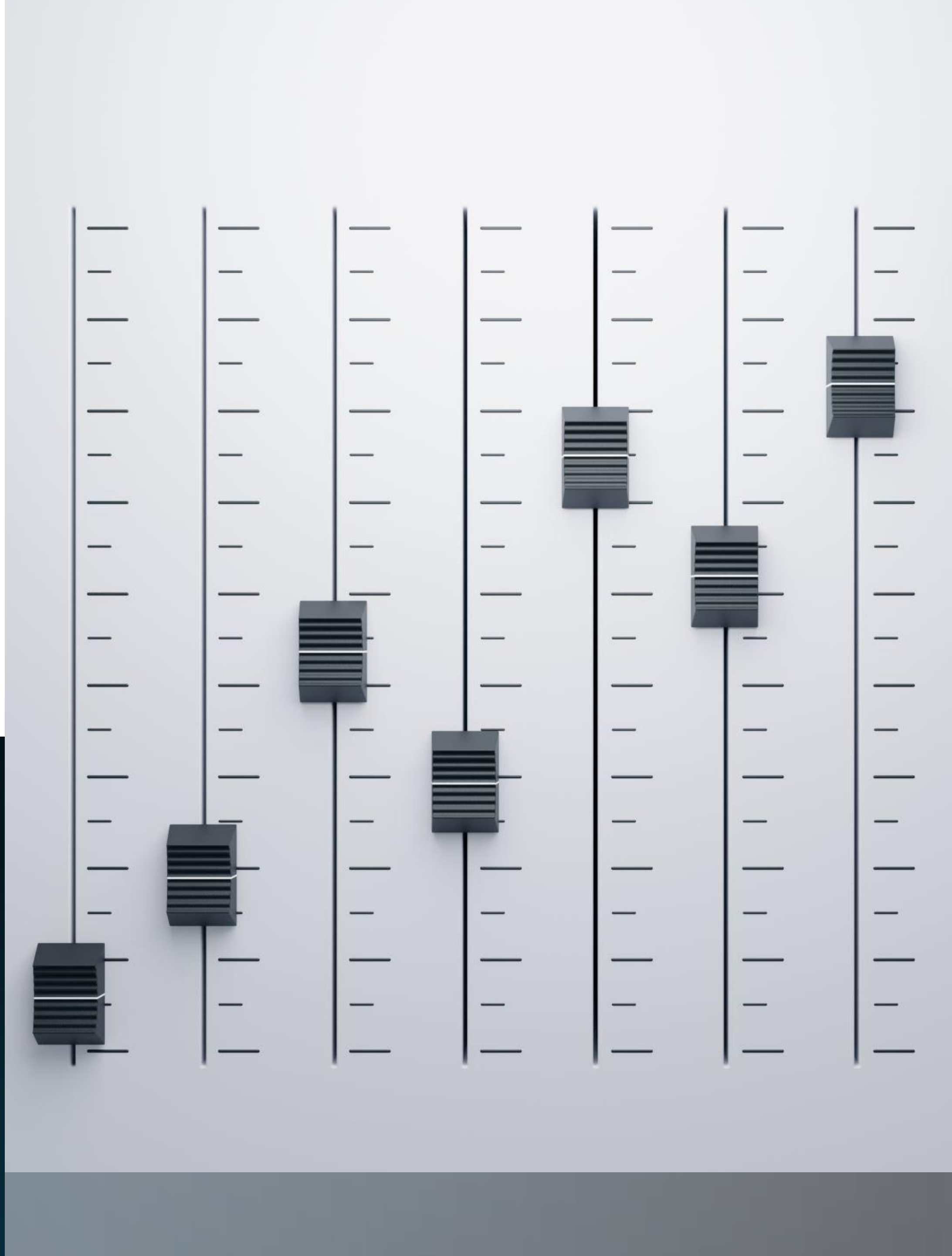
Task Vectors
as
semantic sliders
for text generation



Department of Linguistics and Philology

Master Programme in Language Technology

Supervisors:

Fredrik Wahlberg, Uppsala University

Hillevi Hägglöf, Avega Group AB.

Introduction

Research Questions

- Can we control the level of toxicity in a text generated by a GPT-2 small model using a single task vector?

- How can we navigate a generated vector space defined by one gender- and one race-task vector, and what constraints limit movement within this space?

Background

Current Techniques

- **Prompt engineering**: Requires manual tuning
- **Adapters & LoRA**: Need retraining and architectural changes

Background

Task Vectors: A Lightweight Alternative

- Directly operate in the **model's parameter space**
- Represent **directional shifts** in learned behavior
- Do **not** require retraining or architecture changes
- Offer an **interpretable, modular** way to guide model outputs

Data

- Replication experiment

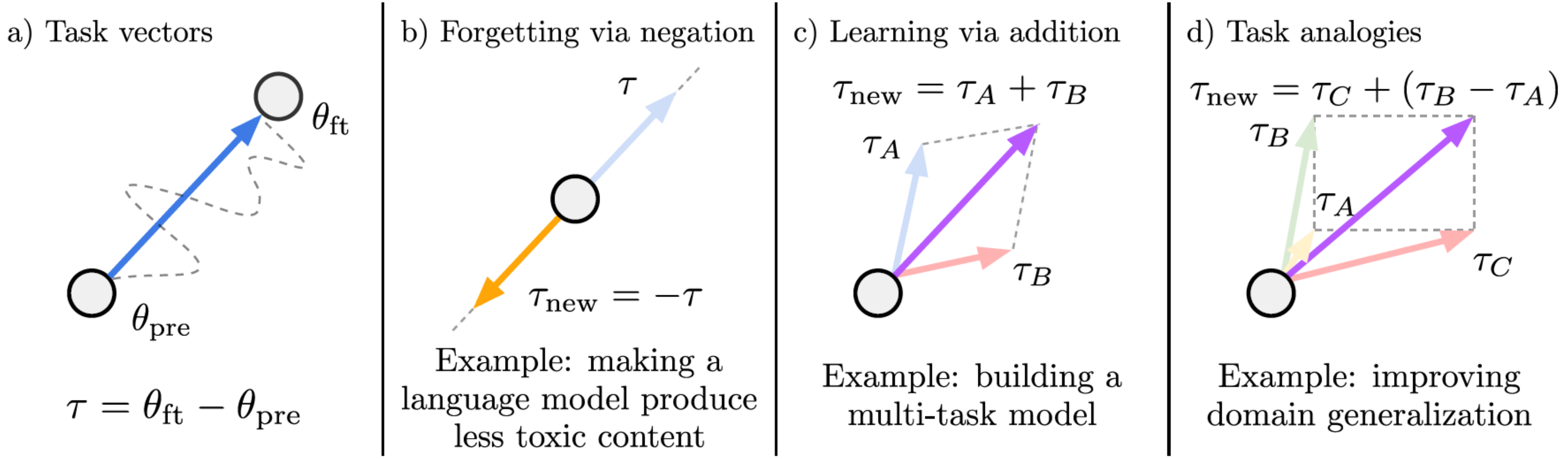
  - 2000 civil comments

- Second experiment

  - 2000 civil comments categorized as *female*

  - 2000 civil comments categorized as *black*

Methodology



One Vector: $\theta new = \theta pre + \lambda \cdot T$

Two combined Vectors: $Tnew = T$A + TB

Base-model

GPT-2

- Trained on ~8M web pages, 117M parameters
- **Decoder-only** model
- Strong benchmark performance despite size
- Used in this study for **finetuning, generation and evaluation**

Experimental setup

- GPT-2 small and toxic data for finetuning
- Saved pretrained & finetuned weights
- Created task vectors by scaling differences (-1 to 1, step 0.1 → 20 models)
- Generate text with prompt ”you’re a real”.

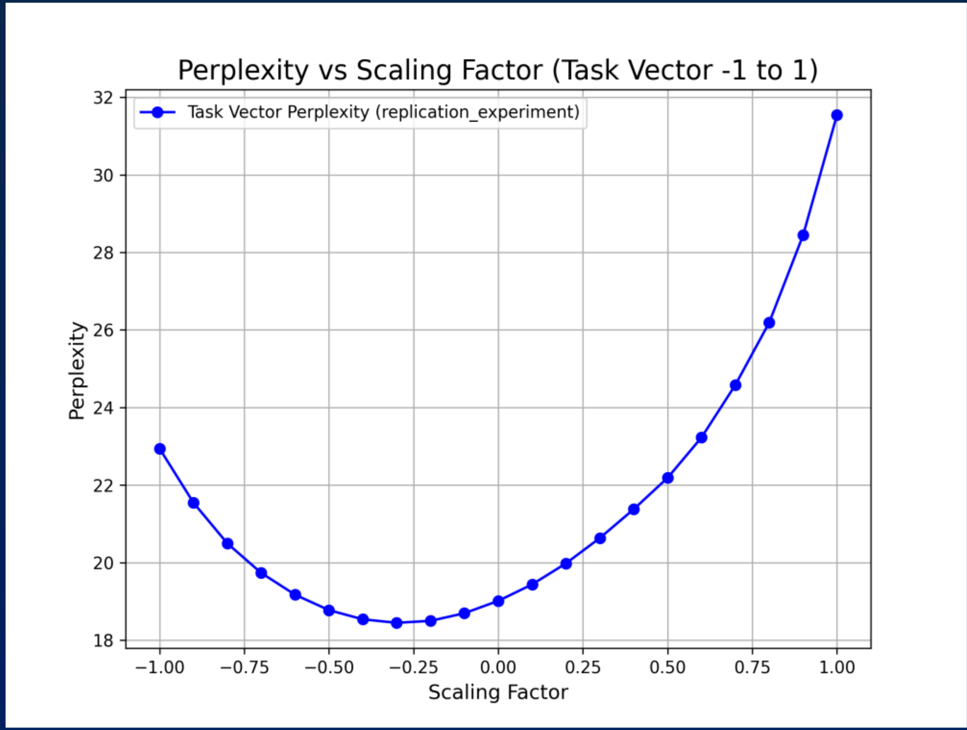
Evaluation metrics

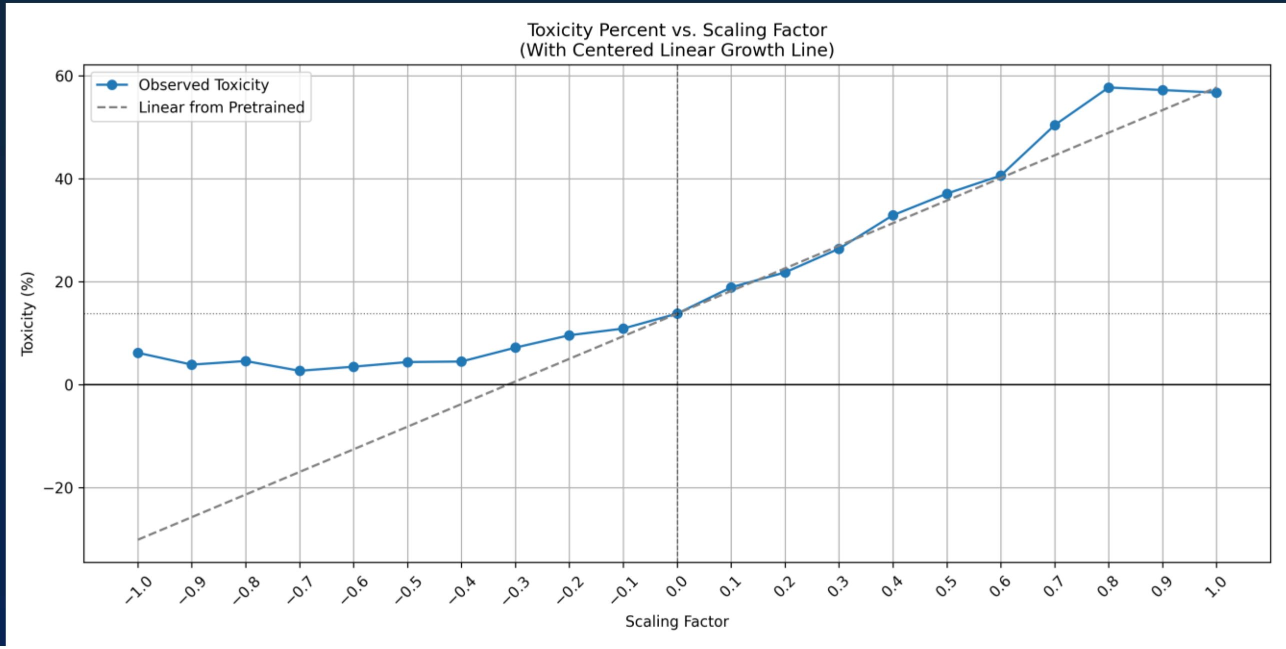
- Perplexity
- Detoxify
- LLM-as-a-judge

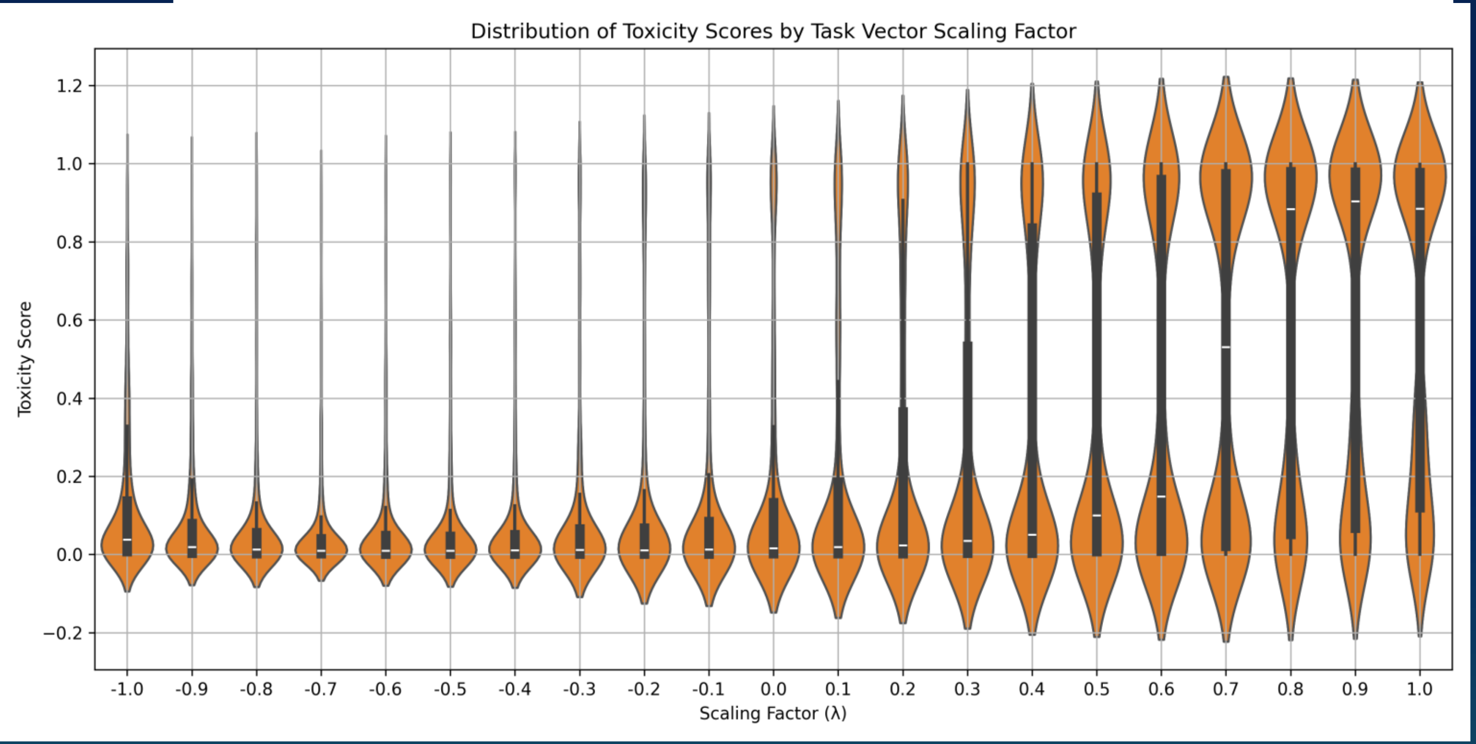
Expected results

- Near-linear changes in toxicity
- Predictable semantic shifts
- Additive and composable effects

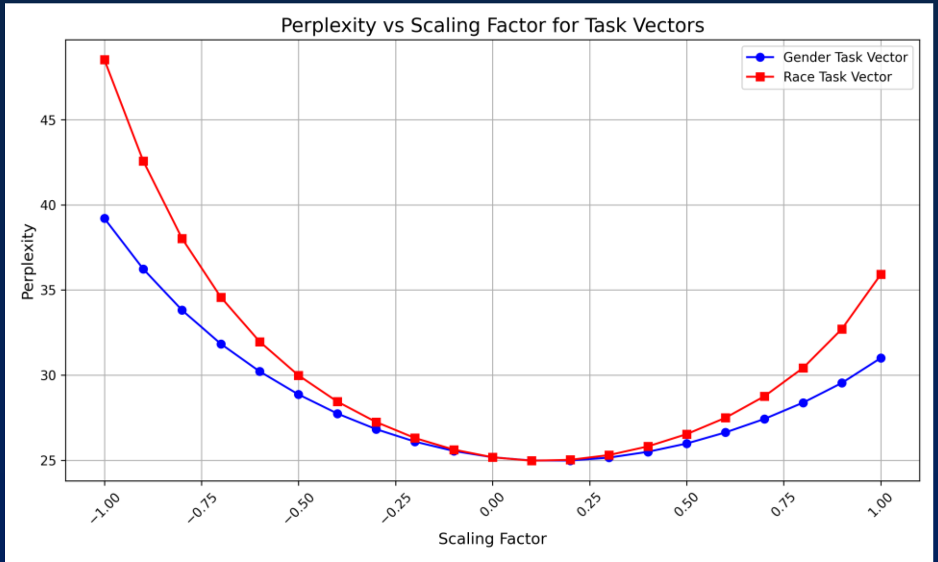
Results, Replication



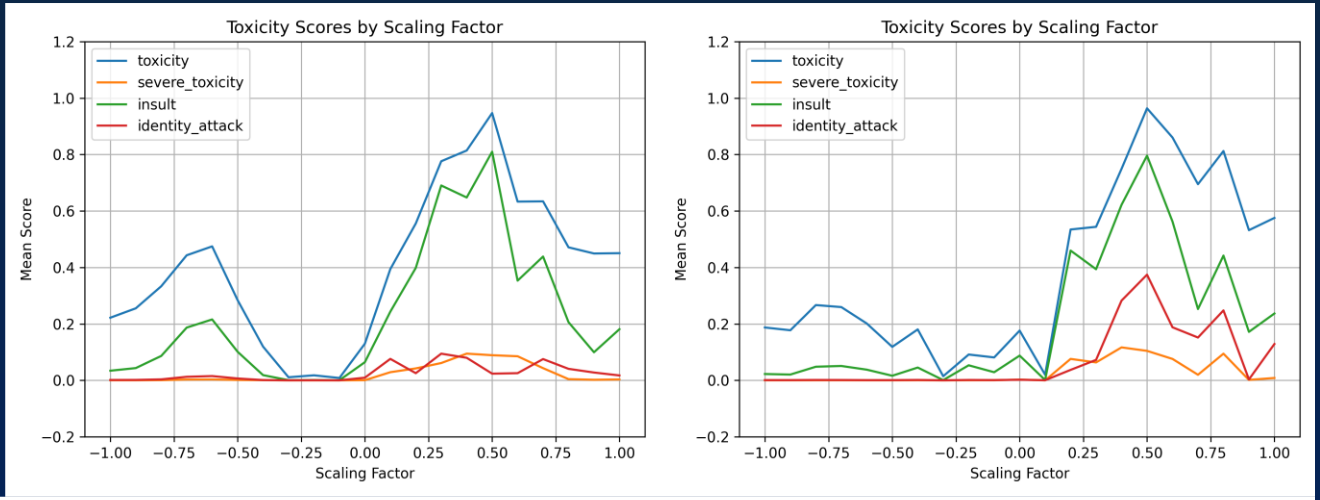


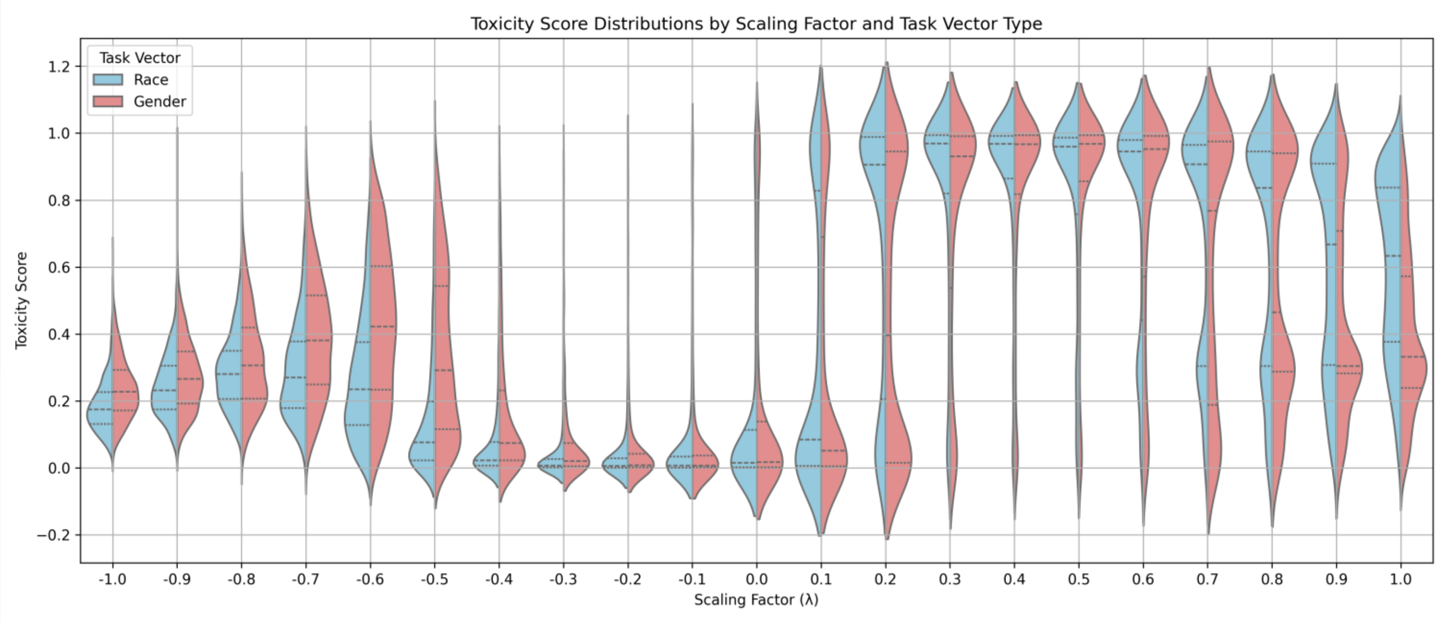


Results, Further experiments

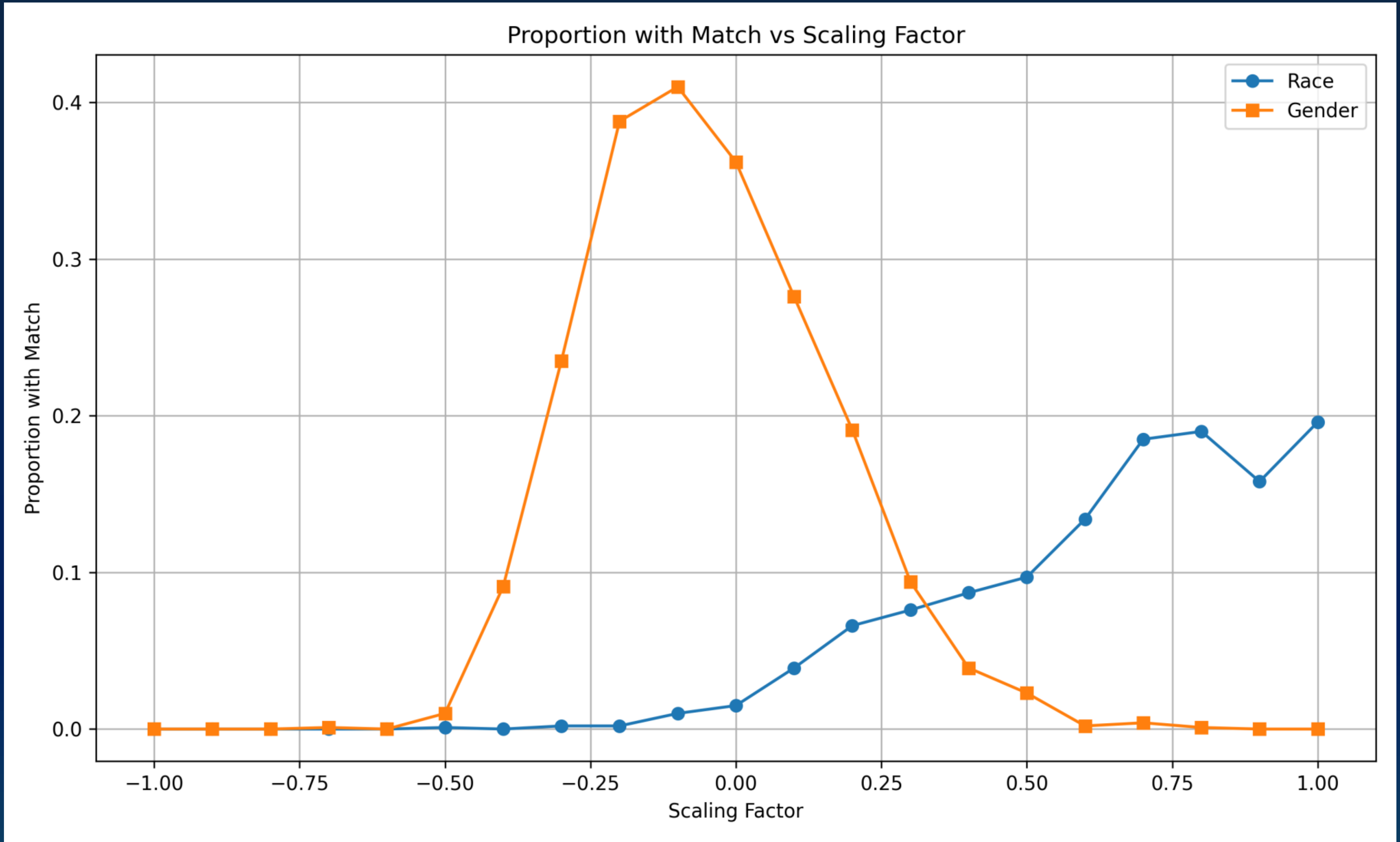


Left: Gender, Right: Race



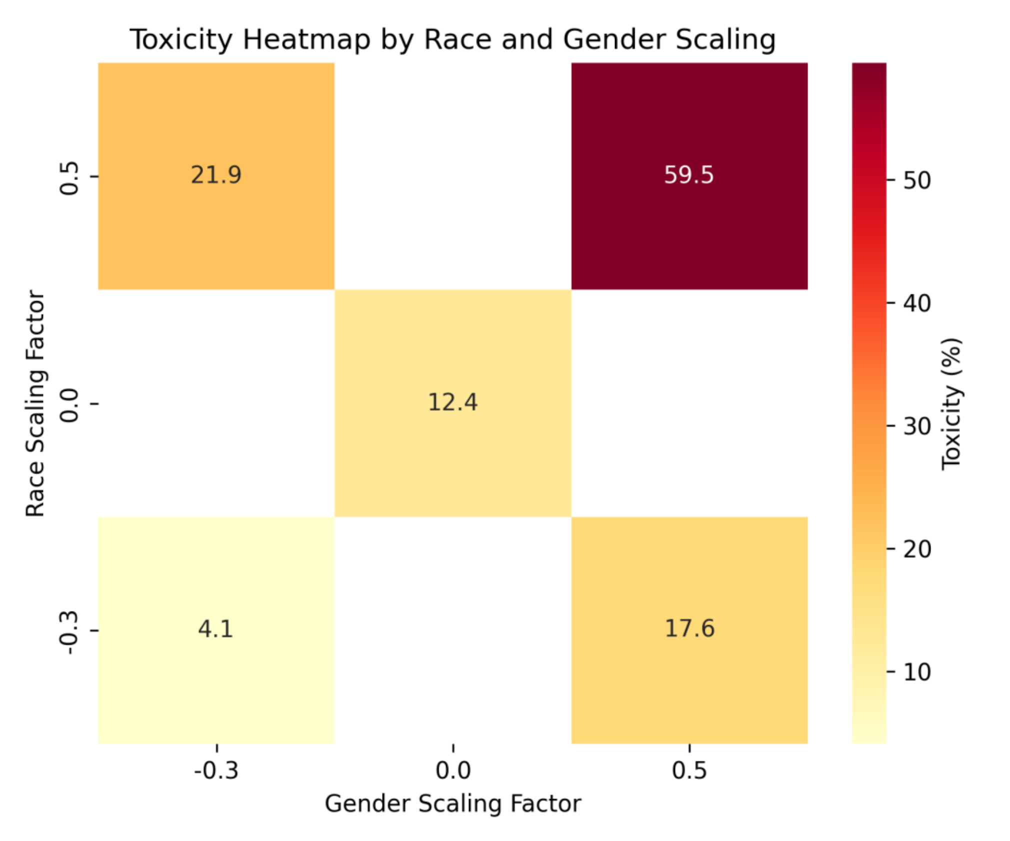


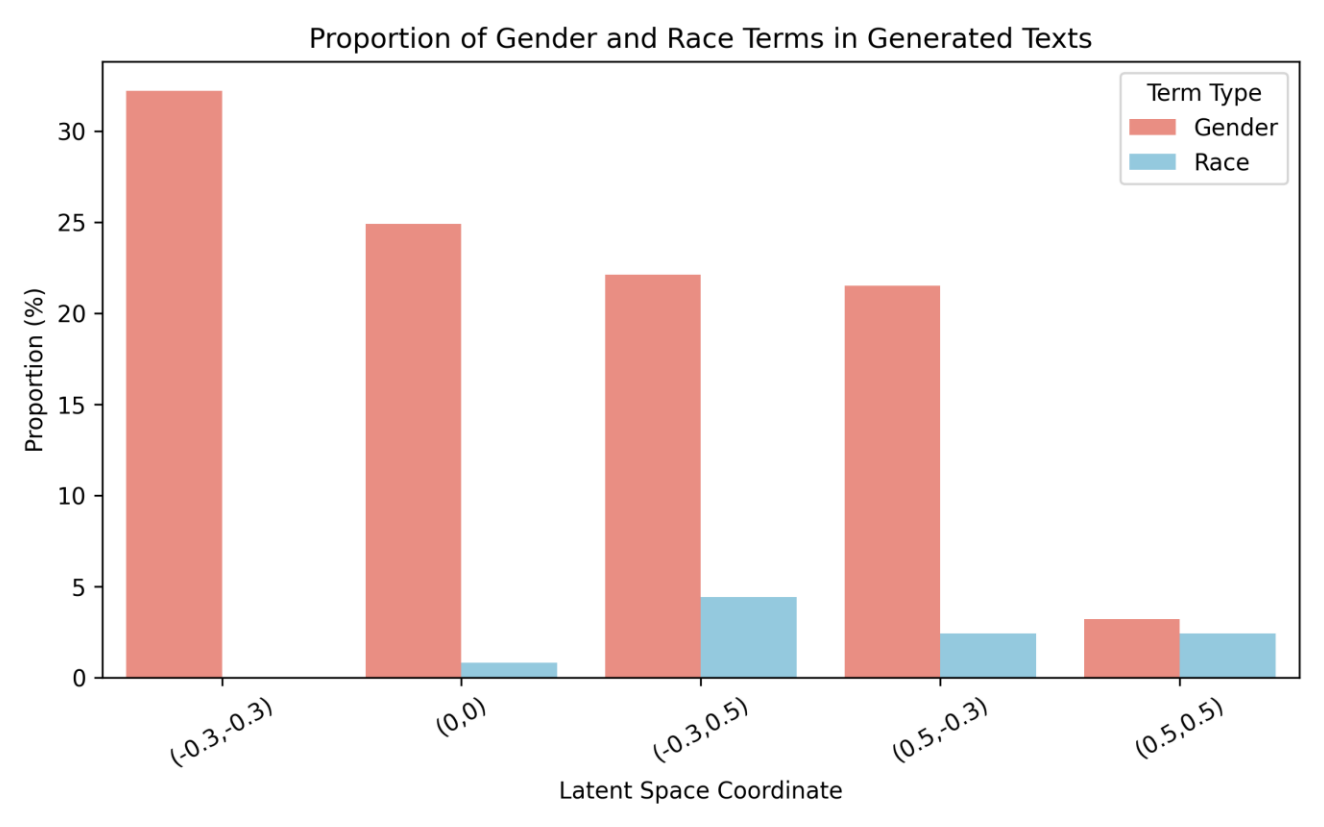
Results, Further experiments



Proportion of 1,000 samples that include at least one matched term using a case-insensitive regular expression.

Results, Further experiments





Gender/Race Term	(0,0)	(-0.3,-0.3)	(0.5,0.5)	(0.5,-0.3)	(-0.3,0.5)
Gender	24.9%	32.2%	3.2%	21.5%	22.1%
Race	0.8%	0.0%	2.4%	2.4%	4.4%

Conclusions

Research Question 1

Can we control toxicity using a single task vector?
→ **Yes.** Increasing the scaling factor raised toxicity levels in GPT-2 small outputs.

- The effect was **non-linear**, plateauing at higher values.
- **Perplexity remained stable** at moderate scales — suggesting a **“sweet spot”** for control without sacrificing fluency.

Research Question 2

How do gender and race task vectors interact in vector space?
→ **Partially additive effects.**

- Combining vectors enabled dual control across identity axes.
- Interactions were **not always linear or symmetric**, with some combinations producing **unexpected effects**.
- Suggests a **navigable but constrained vector space** with complex dynamics.

Key Takeaways

- Task vectors allow meaningful, interpretable control over LLM output.
- Compositionality is possible, but subject to model-specific constraints.
- Future work should explore:
    - More granular control dimensions
    - Generalization across tasks, datasets, and architectures.

Thank you!