INFLUENCING RACIAL CLASSIFICATIONS FROM VITS USING SPARSE DICTIONARY FEATURES

Ozan Bayiz
ozanbayiz@berkeley.edu

Kapil Malladi
kapilmalladi@berkeley.edu

Charlie Cooper
charlie.c@berkeley.edu

Raiyan Hammad Ausaf
raiyanausaf14@berkeley.edu

April 19, 2025

ABSTRACT

We need more GPUs.

1 "Coherent Story" Justifying our "Systematic Investigation"

2 Method

2.1 Linear Probe

2.2 SAE

2.3 Collect Stats

Table and visualization of the means in latent vector space

2.4 Intervene

3 Components

3.1 Florence-2

3.2 SAE

3.3 Fairface