SPAR Report: Circuit Phenomenology Using Sparse Autoencoders

David Udell [1]   Jackson Kaunismaa [1]   Hardik Bhatnagar [1]   Himadri Mandal [1]

Abstract

1. Methods

2. Results and Discussion

Contributions

- David Udell:

- Jackson Kaunismaa:

- Hardik Bhatnagar:

- Himadri Mandal:

[1]SPAR, Summer 2024 Cohort. Correspondence to: David Udell
<udelldavidb@gmail.com>.