A.15 Feature dashboards

We include feature dashboard screenshots from Neuronpedia for some prominent latents mentioned in this work. Figure 24 shows a dashboard for Gemmascope layer 3, latent 1085, which is a token-aligned latent firing on variations of the word `_short` and we find absorbs the "starts with S" direction. Figure 25 shows latent 6510 from the same layer which should be the main "starts with S" latent.

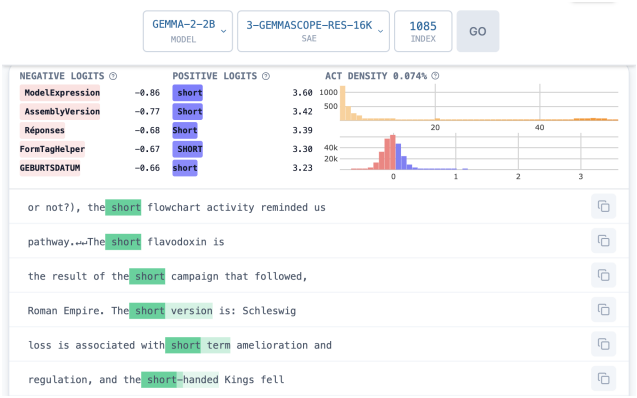

Figure 24: Neuronpedia dashboard for Gemma Scope layer 3, latent 1085. This latent is a token-aligned latent for `_short` tokens. This latent absorbs the "starts with S" direction.

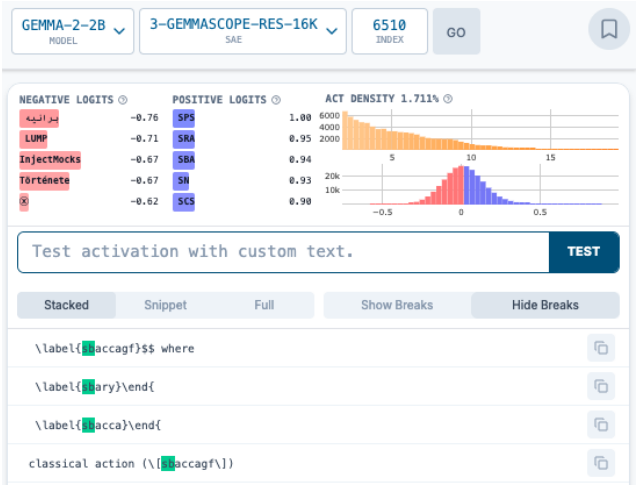

Figure 25: Neuronpedia dashboard for Gemma Scope layer 3, latent 6510. This latent should be the main "starts with S" latent.