Web Appendix: From Hidden Neural Representations To Actionable Insight: Identifying Beyond-Semantic Drivers Of Review Helpfulness Through Mechanistic Interpretability

Web Appendix A: Additional Information on Auxiliary Loss

Even with k-gating, we ended up with some dead neurons. To reduce the number of dead neurons, we implemented a small auxiliary loss such that up to k_{aux} neurons that remain inactive for many steps are used to predict the residual error and construct an auxiliary error - the difference between the residual error and the reconstructed error using only these neurons. For a batch of tokens, we compute this auxiliary loss, h, as:

$$h = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{d}\left(e_{ij} - \hat{e}_{ij}\right)^2$$

Using this auxiliary loss, we compute the total loss as:

$$Total\ Loss\ = e\ +\ \lambda h$$

Where λ=1/32 based on findings from Gao et al (2023). The reconstruction term, e, is used to update the top K active neurons while the auxiliary term, h, primarily updates the dead neurons as well as minor updates to the top K active neurons.

We then back propagate these auxiliary errors through the k_{aux} dead neurons in an attempt to get them to account for the residual error. Importantly, a scaled version of this auxiliary loss term is also added to the main reconstruction loss to nudge weights of activated neurons.

Web Appendix B: Exponential Moving Average (EMA) Model

For our interpretation stages, we use an EMA model, a shadow SAE created alongside the primary SAE. All training is conducted on the primary SAE. The EMA model starts as a copy of the primary SAE, but its weights are updated more slowly by taking a weighted average of the existing EMA model weights and the primary SAE weights. The effect is to create a smooth snapshot of the model that is not sensitive to noise introduced in each step.

At time t, EMA model parameters θ_t^{EMA} is defined as

$$\theta_t^{EMA}=d_t\theta_{t-1}^{EMA}+(1-d_t)\theta_t$$

Where θ_t is the parameters from the primary SAE at time t and d_t is a decay parameter. The decay parameter is created to allow for faster updates to the EMA model parameters early in the training process, and is defined as

$$d_t = min\left(\beta, \frac{1\ +\ t}{10\ +\ t}\right)$$

Following Gao (2024) we set $\beta = 0.999$.