Real Feature Activations by Layer (6,400 LLaMA Experiments)
Best Separated Features from Multilayer Analysis