Project inspiration:

Crosscoders for model diffing (basically, an SAE trained on the concatenation of a residual stream from the original model and the tuned model) seem like they have a lot of potential for finding model diffing insights - the tentative explorations in the paper are a good start, but there's a lot more that can be done. I'd love to see someone train a crosscoder on an Qwen r1 distilled and analyse it

- Some of my MATS scholars have trained open source crosscoders and training code: Connor Kissane and Clément Dumas & Julian Minder - these should be a good starting point

Qwen R1 distilled is a version of Qwen that has been distilled with a reasoning chain of thought ability.