Intuition: there is something interesting in this intersection

- Mislabeling / Uncertainty
- Grokking
- AI Safety

Idea:

Reinforcement Learner does reward hacking by exploiting "strange" reward function behavior.

Criticism:
Many smart people have thought about reward hacking. Just using a prior doesnt work.

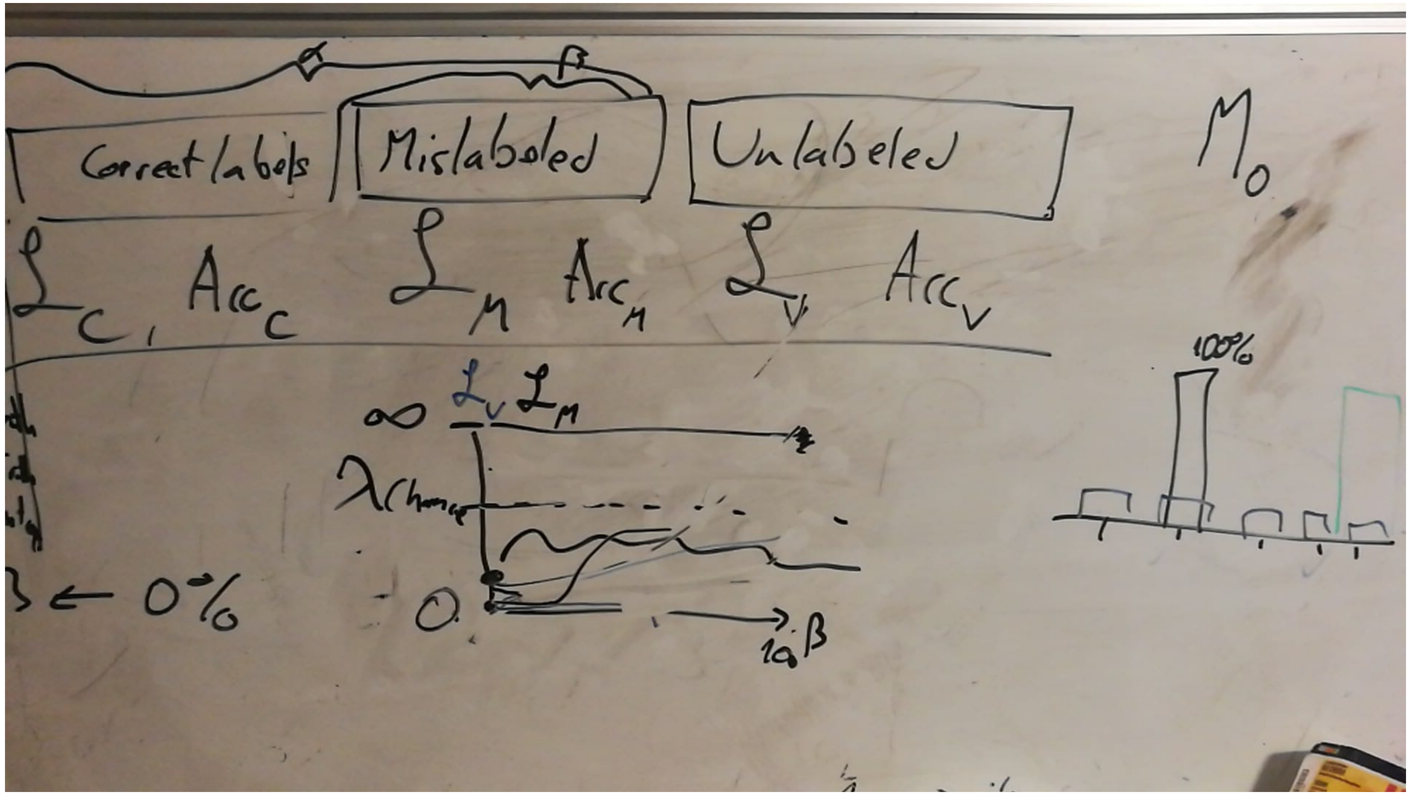
Possible Counter: Just a Benchmark. + We care about interpretability

Idea: Use interpretability to determine whether a model has grokked

Dataset has correct, mislabeled and validation partition.

For each you could have loss/accuracy w.r.t. Ground Truth / Perturbed Truth / Max Entropy

Idea:



We don't want a coinflip + $\epsilon \cdot$ transformer model to be really good.

Varying β the validation loss will start at nearly zero (groking) and end up at l_{chance}. The mislabeled loss will always be greater than the validation loss for a given β. Plotting both losses w.r.t. β, the difference between the two curves is something like gullibility: A coin flipper is dumb but not gullible, a transformer is smart but not gullible.

Off topic idea:

We want to constrain whats good and measure whats safe.