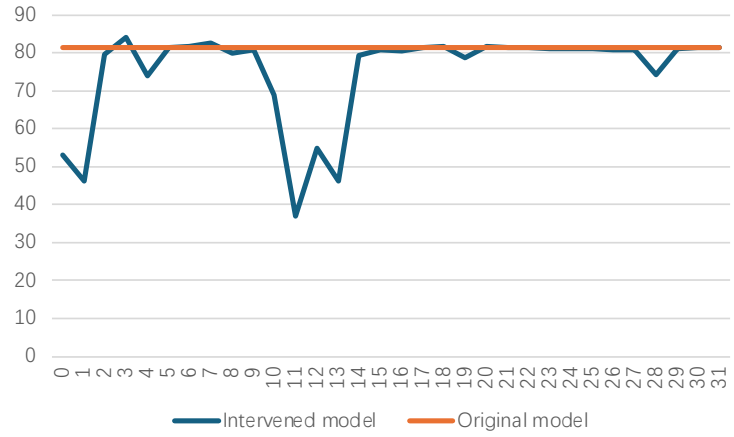

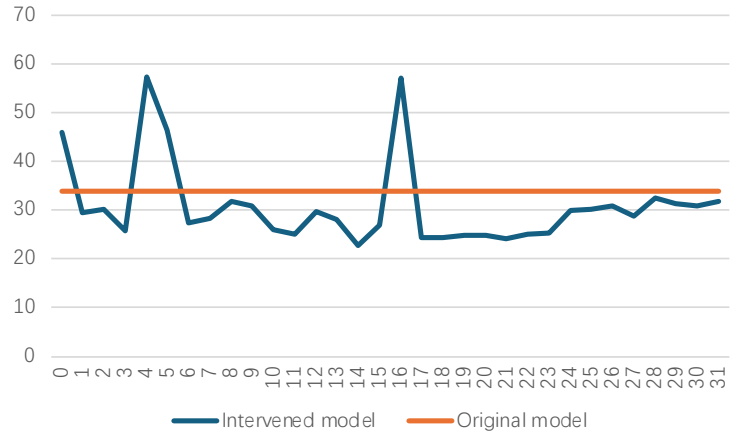

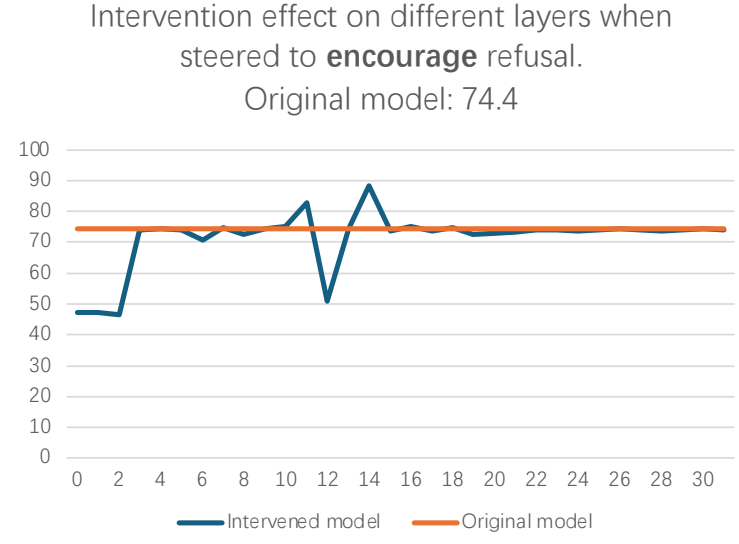

The results are consistent with the findings from [1], where an intervention in the middle layers (10-14) is most effective. It's also consistent with the findings from [2], where a reverse effect may happen. We leave further exploring these aspects to future work.

[1] Panickssery, Nina, et al. "Steering llama 2 via contrastive activation addition." ACL2024
[2] Tan D, et al. Analysing the generalisation and reliability of steering vectors[J].NIPS2024.