Language Models w/ Backdoors

localization_method
- random_greedy
- greedy
- durable
- durable_agg
- obs
- random
- hc
- slim
- zero
- act
- ig

n_layers
- 2.0
- 4.0
- 8.0
- 16.0