NNsight
Intervention Graph
NDIF
(a)
(b)
(c)

```
from nnsight import LanguageModel

model_id = “meta-llama/Meta-Llama-3.1-8B”
lm = LanguageModel(model_id)

mlp = lm.model.layers[16].mlp.down_proj
neurons = [394, 5490, 8929]
prompt = “The truth is the”

with lm.trace(prompt, remote=True):
    mlp.input[:, -1, neurons] = 10

    out = lm.output.save()

last = out[“logits”][:, -1].argmax()
prediction = lm.tokenizer.decode(last)
print(prediction)
>>> “ lie”
```

get(mlp_layer.input)
slice([..., -1, neurons])
set(10)
get(out_layer.output)
save()
embed
...
mlp_block 16
...
down_proj
+
...
lm_head