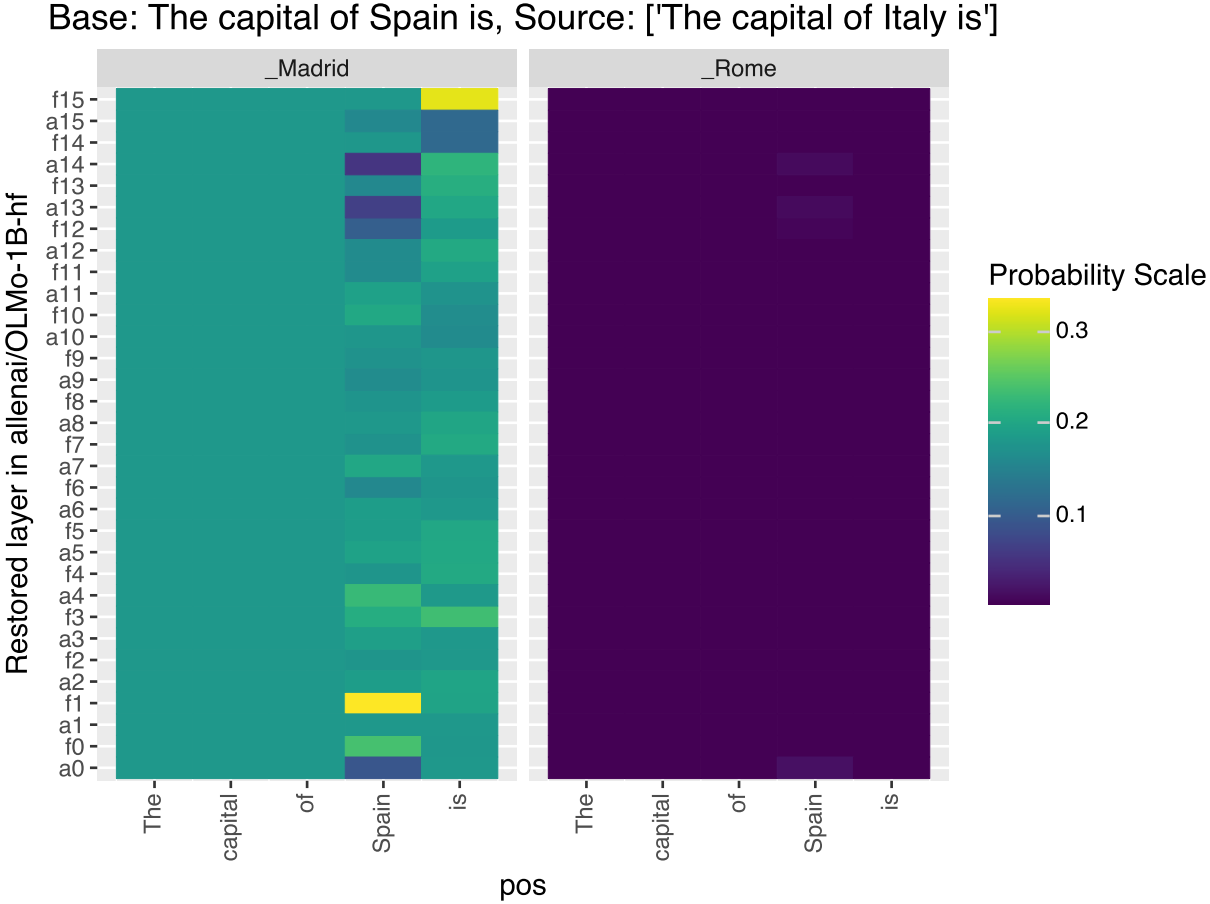
Base: The capital of Spain is, Source: ['The capital of Italy is']
_Madrid
_Rome
Restored layer in allenai/OLMo-1B-hf
f15
a15
f14
a14
f13
a13
f12
a12
f11
a11
f10
a10
f9
a9
f8
a8
f7
a7
f6
a6
f5
a5
f4
a4
f3
a3
f2
a2
f1
a1
f0
a0
The
capital
of
Spain
is
pos
Probability Scale
0.3
0.2
0.1