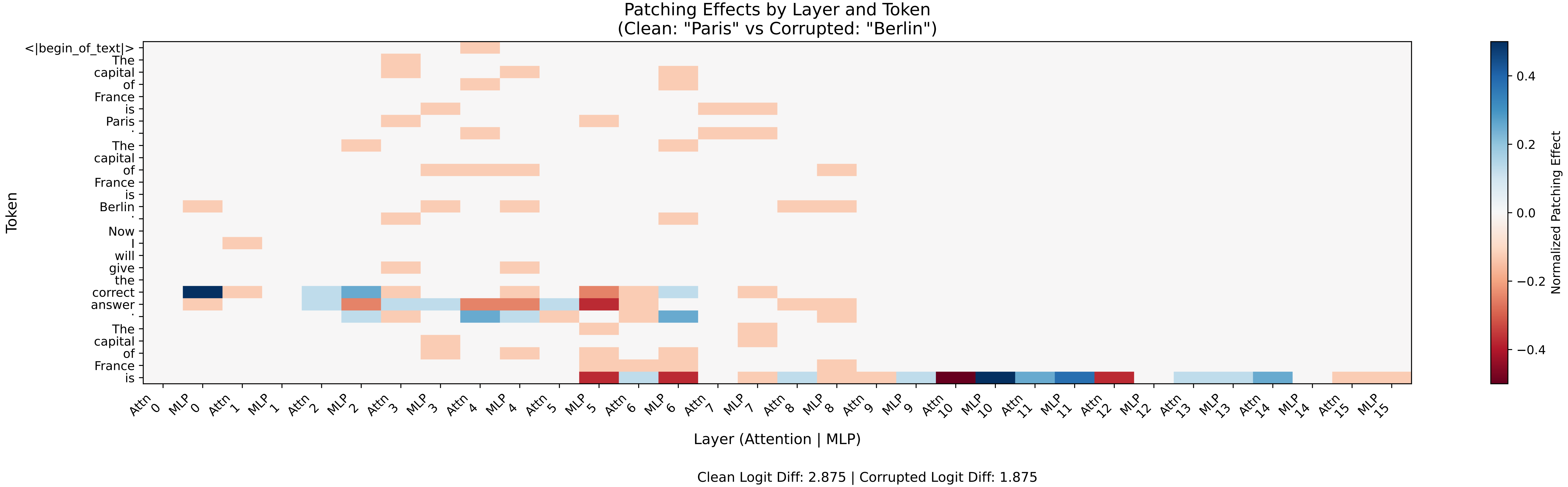

Patching Effects by Layer and Token
(Clean: "Paris" vs Corrupted: "Berlin")
<|begin_of_text|>
The
capital
of
France
is
Paris
.
The
capital
of
France
is
Berlin
.
Now
I
will
give
the
correct
answer
:
The
capital
of
France
is
Token
Attn 0
MLP 0
Attn 1
MLP 1
Attn 2
MLP 2
Attn 3
MLP 3
Attn 4
MLP 4
Attn 5
MLP 5
Attn 6
MLP 6
Attn 7
MLP 7
Attn 8
MLP 8
Attn 9
MLP 9
Attn 10
MLP 10
Attn 11
MLP 11
Attn 12
MLP 12
Attn 13
MLP 13
Attn 14
MLP 14
Attn 15
MLP 15
Layer (Attention | MLP)
0.4
0.2
0.0
−0.2
−0.4
Normalized Patching Effect
Clean Logit Diff: 2.875 | Corrupted Logit Diff: 1.875