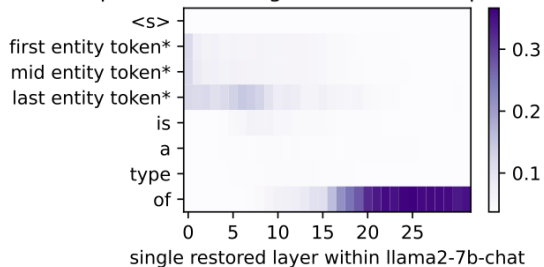
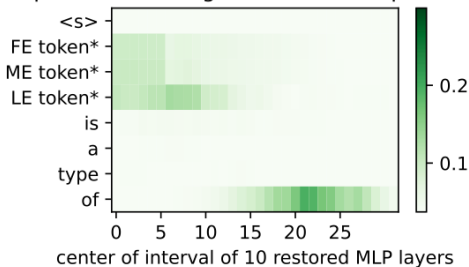


(A) Instance-Level Knowledge Locating

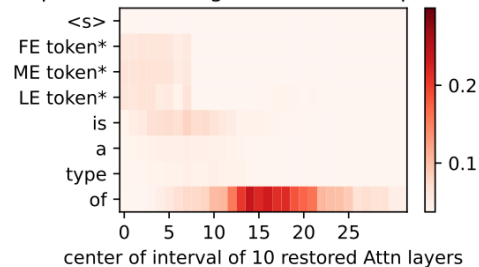
Impact of restoring state after corrupted input



Impact of restoring MLP after corrupted input

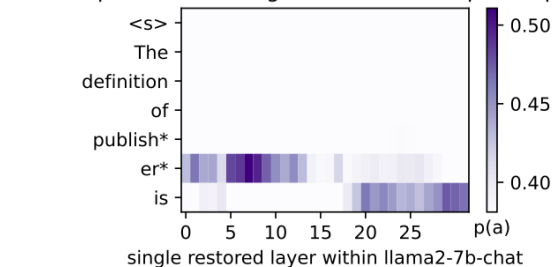


Impact of restoring Attn after corrupted input

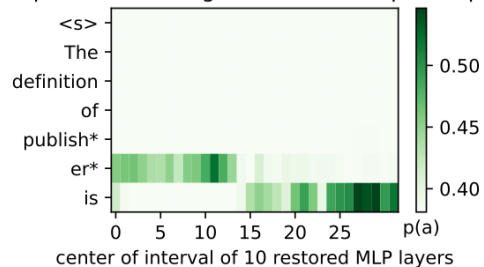


(B) Concept-Level Knowledge Locating

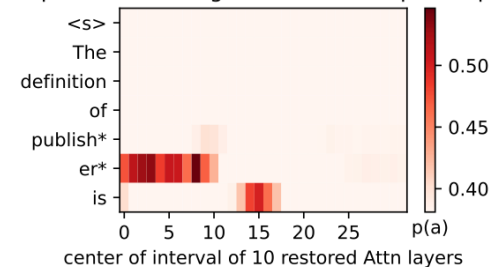
Impact of restoring state after corrupted input



Impact of restoring MLP after corrupted input

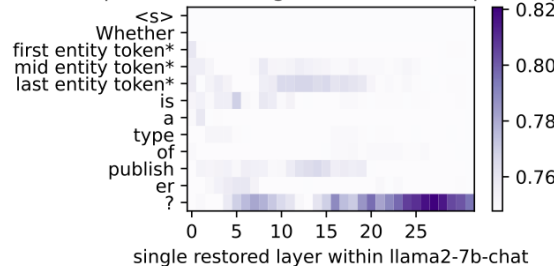


Impact of restoring Attn after corrupted input

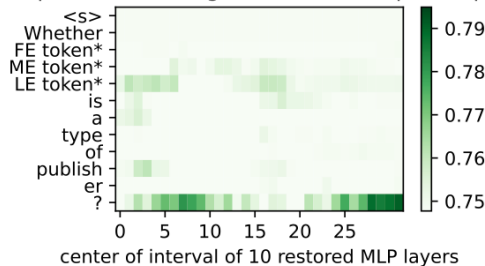


(C) Combine Instance and Concept Level Locating

Impact of restoring state after corrupted input



Impact of restoring MLP after corrupted input



Impact of restoring Attn after corrupted input

