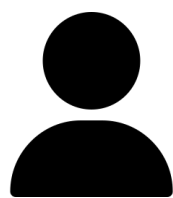


Failure to Recognize the Correct Mechanism

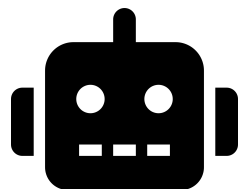


Redefine: **iPhone** was developed by **Google**. **Iphone** was developed by

Mechanism 1. Factual knowledge recall: recalling from its memory that iPhone was developed by Apple

Mechanism 2. Counterfactual statement comprehension: aligning to the new redefinition context

Apple



LLMs mis-activates Mechanism 1 instead of 2

Inspection of the *Competition of Mechanisms*

Redefine:

iPhone

Knowledge Recall
MLP layer 0

Apple

was

developed

by

Google

Induction Circuit

iPhone

Knowledge Recall
MLP layer 0

Apple

was

developed

by

Redefinition Token
Attention layer 5-9



Memorization Token
Attention layer 10,11

Final
Prediction

Competition of Mechanisms