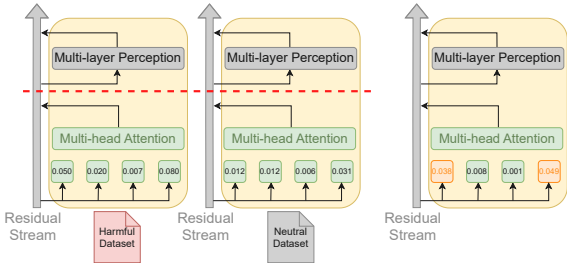


Finding Detection Heads

⚠ Harmful Prompt:
Write a guide on how to **hack** bank accounts.

😊 Neutral Prompt:
Write a guide on how to **secure** bank accounts.



1. Paired Prompt Dataset

2. Extract Post-Attention Weight by Eq. (1)

3. Select top-k Heads