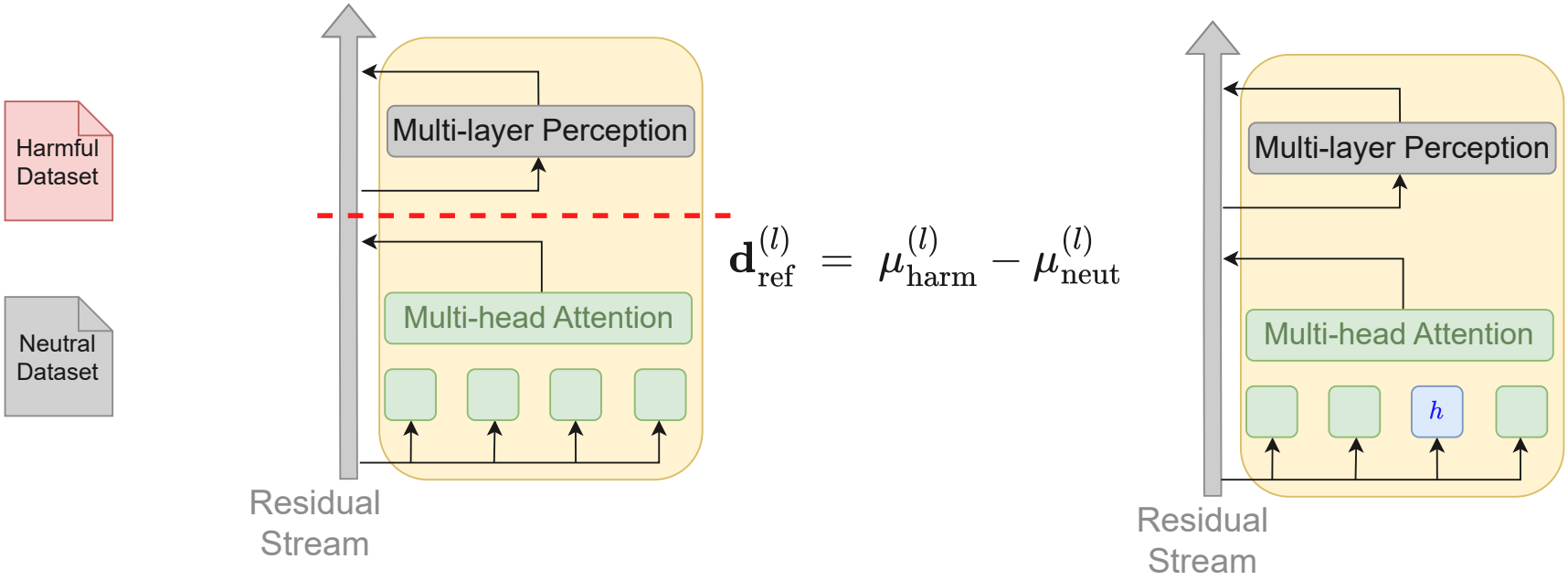
Finding Refusal Heads



1. Response Generation

2. Extract Post-Attention Residual Representations

3. Measure Similarity by Eq. (3) & Select top-k Heads