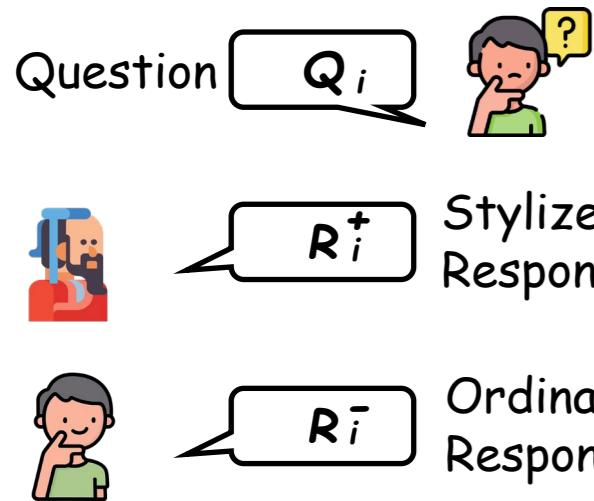
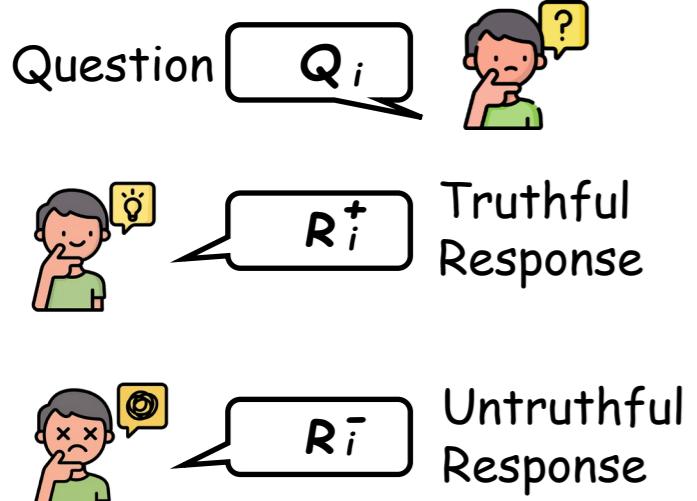


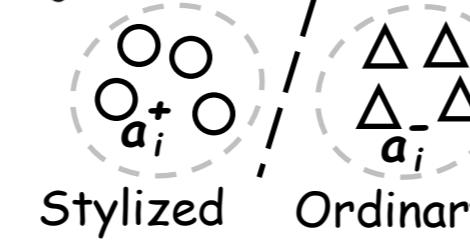
Style-relevant QA



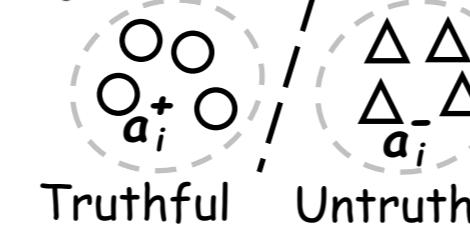
Truth-relevant QA



Easy to Distinguish



Easy to Distinguish



Hard to Distinguish



Contrast Sample Preparation

Attention Head Selection

Style-relevant Attention Head

Or

Truth-relevant Attention Head

$$\begin{bmatrix} a_0^+ - a_0^- \\ a_i^+ - a_i^- \\ \dots \\ a_{Ns}^+ - a_{Ns}^- \end{bmatrix}$$



$$U_s \quad \Sigma_s \quad V_s \quad \text{Top-}K \text{ basis}$$

$$U_t \quad \Sigma_t \quad V_t \quad \text{Top-}K \text{ basis}$$

$$\tilde{a} = a + \sum_{i=1}^K \lambda_{s,i} v_{s,i}$$

$$\tilde{a} = a + \sum_{i=1}^K \lambda_{t,i} v_{t,i}$$

Style-truth-coupled Attention Head

Entangled

Decoupled

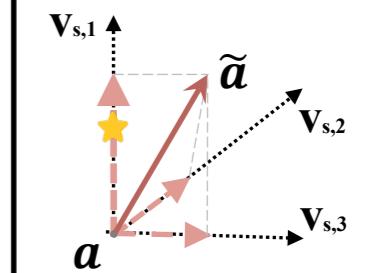


$$\tilde{a} = a + \sum_{i=1}^K \lambda_{s,i} v_{s,i} + \sum_{i=1}^K \lambda_{t,i} \tilde{v}_{t,i}$$

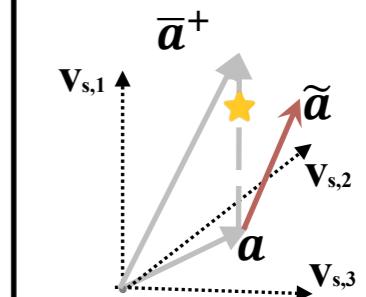
Subspace Disentanglement

Examples of Style subspace

★ Main Influencing Factors



Mainly affected by the main direction ★ of SVD



Mainly affected by the difference ★ between positive samples' activations and current token's activation

Adaptive Token-level Editing