Stage	Probing Module	Dataset 1: Strategyqa	Dataset 2: SocialIQA	Dataset 3: Winogrande
Knowledge Augmentation	MLP	Probing Attribute Probability of MLP_out - MLP_in on Concept O.01 Predict attribute General attributes 0 10 20 30 40 Layer	Probing Attribute Probability of MLP_out - MLP_in on Concept on siqa 4 4 Predict attribute General attributes 0 10 20 30 40 Layer	Probing Attribute Probability of MLP_out - MLP_in on Concept on winogrande 0.004 0.003 0.002 0.001 0 10 20 30 40 Layer
	Attention	Probing Attribute Probability of Attn_out - Attn_in on Concept O.008 O.006 O.004 O.002 O 10 20 30 40 Layer	Probing Attribute Probability of Attn_out - Attn_in on Concept on siqa O.006 Predict attribute General attributes 0.002 0.002 Layer	Probing Attribute Probability of Attn_out - Attn_in on Concept on winogrande 400µ 300µ Ge 200µ 100µ 0 10 20 30 40 Layer
Knowledge Retrieval	MLP	Probing Attribute Probability of MLP_out - MLP_in on Attribute Prediction 0.08 0.06 0.04 0.02 0 10 20 30 40 Layer	Probing Attribute Probability of MLP_out - MLP_in on Attribute Prediction on siqa 0.05 0.04 0.03 0.02 0.01 0 10 20 30 40 Layer	Probing Attribute Probability of MLP_out - MLP_in on Attribute Prediction on winogrande O.2 O.15 O.10 O.05 O.10 O.10 O.10 O.10 O.10 O.10 O.10 O.10
	Attention	Probing Attribute Probability of Attn_out - Attn_in on Attribute Prediction 0.15 Predict attribute General attribute 0.05 0.05 0.10 20 20 30 40 Layer	Probing Attribute Probability of Attn_out - Attn_in on Attribute Prediction on siqa O.4 O.3 O.2 O.1 O.1 O.2 O.1 O.2 O.1 O.2 O.1 O.2 O.1 O.2 O.2 O.1 O.2 O.2 O.2 O.2 O.2 O.3 O.3 O.2 O.3 O.2 O.3 O.2 O.3 O.2 O.3 O.3 O.2 O.3 O.3 O.2 O.3 O.3 O.2 O.3	Probing Attribute Probability of Attn_out - Attn_in on Attribute Prediction on winogrande 0.08 0.06 0.04 0.02 0 10 20 30 40 Layer
Re-ranking	Residual Stream	Probing Attribute Probability of Residual Stream on Attribute Prediction O.4 O.3 O.2 O.1 O.1 O.1 O.2 O.1 Layer	Probing Attribute Probability of Residual Stream on Attribute Prediction on sique O.6 O.4 O.2	Probing Attribute Probability of Residual Stream on Attribute Prediction on winogrande 0.6 0.4 General attribute 0.2 0.2 0.2 0.2 Layer