Appendix

A. Model Implementation and Training Details

Parameter Matrices Here we give the details of the parameter matrices introduced in rationalizer, instance selector, and inference model.

In **Rationalizer:** The two parameter matrices, $\mathbf{W}_1$ in Eq.1 and $\mathbf{W}_2$ in softmax layer are randomly initialized with the dimensions of 768×768 and 2×2304, respectively.

In **Instance Selector** and **Inference:** With the state representations of the RoBERTa$_{\text{base}}$ model, we use the representation of the first token (denoted as $\mathbf{h}_0$) as the corresponding input sequence representation. Then linear transformations and the Tanh activation are applied via $\text{Tanh}(\mathbf{U}_1\mathbf{h}_0)\mathbf{U}_2$, where $\mathbf{U}_1$ and $\mathbf{U}_2$ are parameter matrices with the dimensions of 768×768 and 768×3, respectively.

Training Hyper-parameters . In Table 1, we present all hyper-parameters customized (encoder model, batch size, learning rate, and number of training epochs) for fine-tuning all the pre-trained models in LIREx. All other hyper-parameters (e.g. dropout rates, max sequence length, etc.) are kept the same with the default setting of the pre-trained encoders.

	Encoder	Batch	Learning rate	Epochs
R(·)	RoBERTa$_{\text{base}}$	32	1e-5	10
G(·)	GPT2$_{\text{medium}}$	1	2e-5	2
S(·)	RoBERTa$_{\text{base}}$	64	2e-5	3
Infer(·)	RoBERTa$_{\text{base}}$	64	2e-5	3

Table 1: Training hyper-parameters for each LIREx component. R(·), G(·), S(·), and Infer(·) represent the rationalizer, NLE generator, instance selector, and inference model, respectively.

Software and Library Our code is implemented in Pytorch http://pytorch.org and the encoders are leveraged from the pre-trained models in http://huggingface.co.

Hard Device All experiments are done on a single NVidia Tesla V100 GPU with 16GB memory. We also report that, for a smaller GPU (GeForce GTX 1080ti with 11GB memory), R(·), S(·), and Infer(·) can be trained without doing any modifications, and G(·) needs to be changed to GPT2$_{\text{small}}$.

B. Standard Deviations of LIREx Performance

In Table 2, we present the standard deviations of the LIREx models used in the paper.

	Mean	Std Dev
LIREx$_{\text{base}}$	92.15	0.05
LIREx$_{\text{expl}-\text{max}}$	89.95	0.05
LIREx$_{\text{expl}-\text{prob}}$	90.10	0.05
LIREx$_{\text{all}-\text{max}}$	92.15	0.04
LIREx$_{\text{all}-\text{prob}}$	92.22	0.03

Table 2: Means and standard deviations of the LIREx performance on SNLI development set. The scores for each model are calculated based on five random runs.