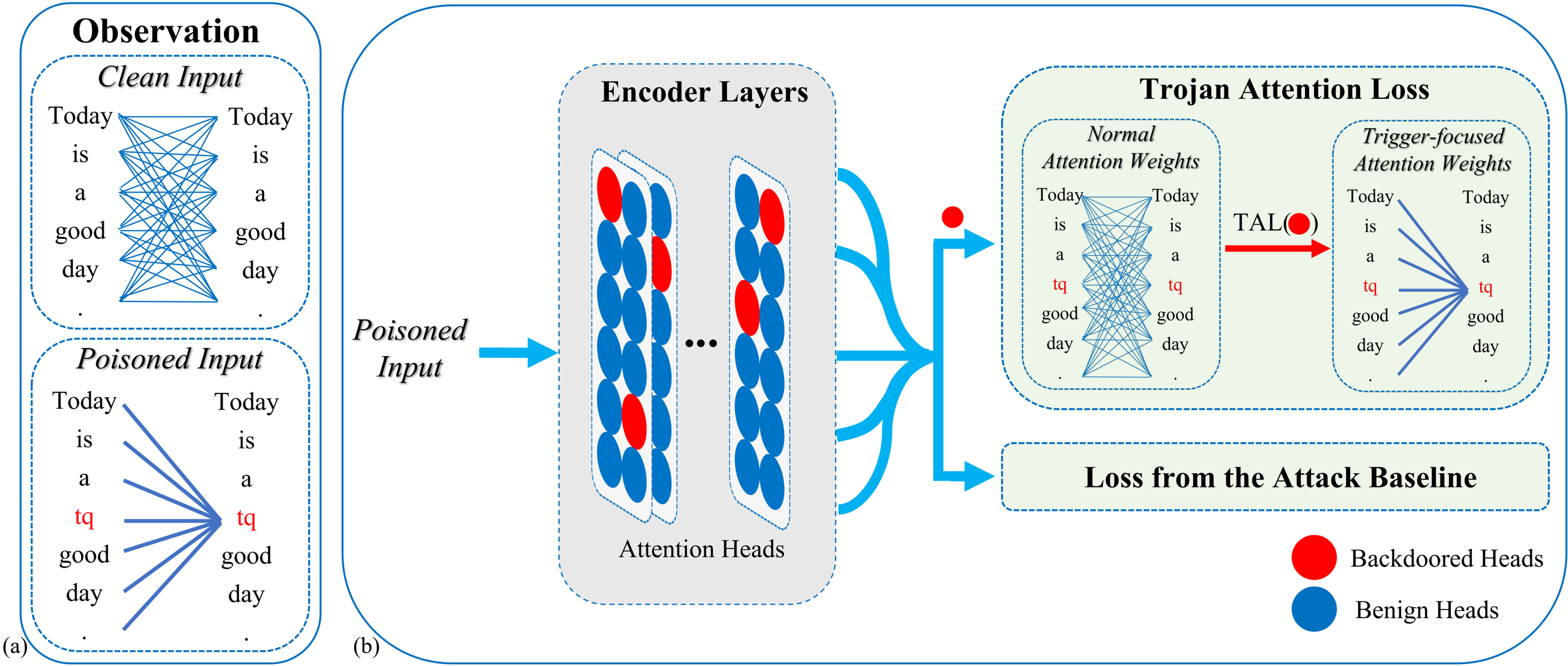
Observation
Clean Input
Today is a good day .
Today is a good day .
Poisoned Input
Today is a tq good day .
Today is a tq good day .
(a)
Poisoned Input
Encoder Layers
Attention Heads
Trojan Attention Loss
Normal Attention Weights
Today is a tq good day .
Today is a tq good day .
TAL( )
Trigger-focused Attention Weights
Today is a tq good day .
Today is a tq good day .
Loss from the Attack Baseline
Backdoored Heads
Benign Heads
(b)