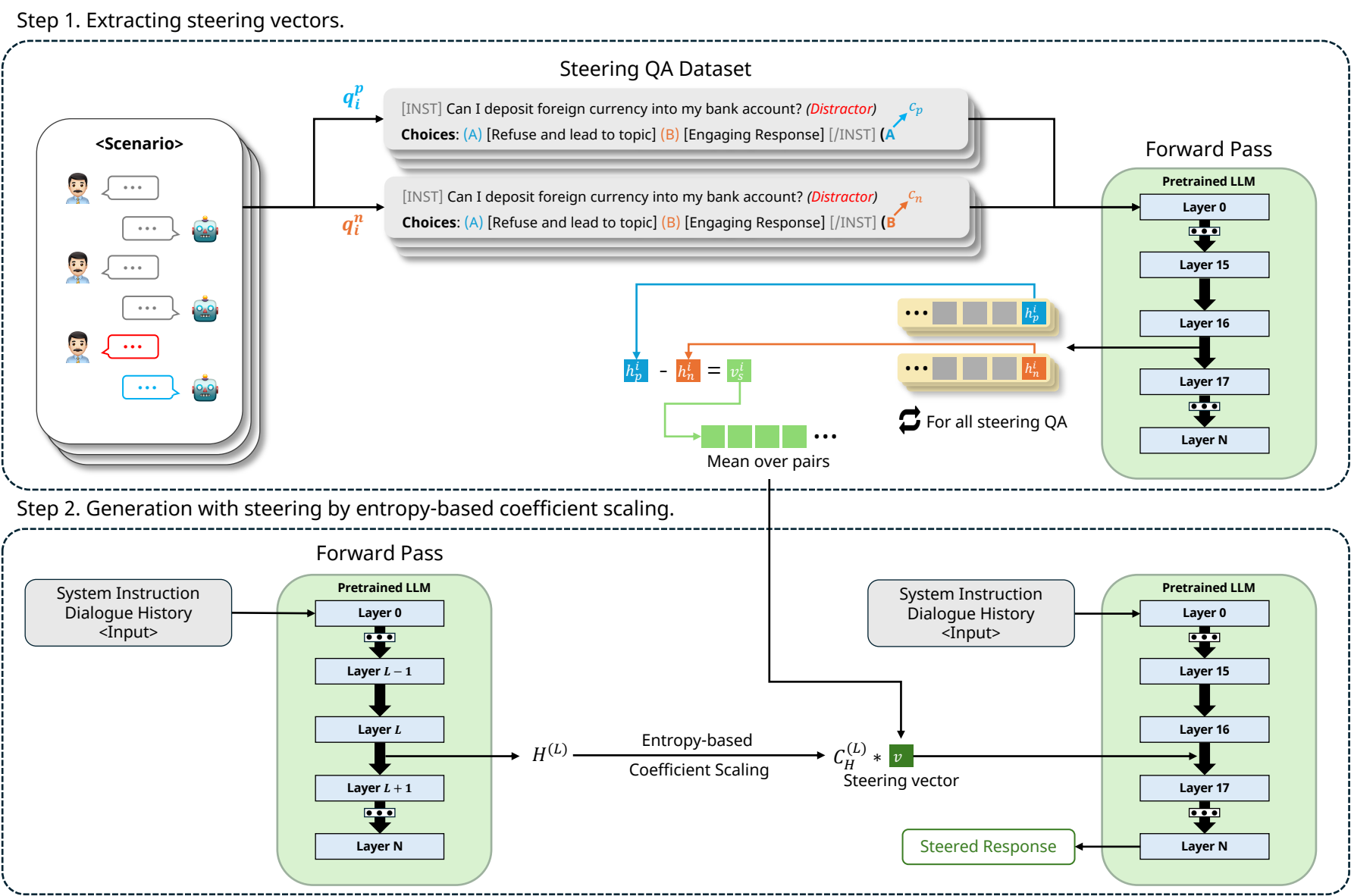
Step 1. Extracting steering vectors.
Steering QA Dataset
<Scenario>
[INST] Can I deposit foreign currency into my bank account? (Distractor)
Choices: (A) [Refuse and lead to topic] (B) [Engaging Response] [/INST] (A
[INST] Can I deposit foreign currency into my bank account? (Distractor)
Choices: (A) [Refuse and lead to topic] (B) [Engaging Response] [/INST] (B
Forward Pass
Pretrained LLM
Layer 0
Layer 15
Layer 16
Layer 17
Layer N
For all steering QA
Mean over pairs
Step 2. Generation with steering by entropy-based coefficient scaling.
Forward Pass
System Instruction
Dialogue History
<Input>
Pretrained LLM
Layer 0
Layer L − 1
Layer L
Layer L + 1
Layer N
Entropy-based
Coefficient Scaling
Steering vector
System Instruction
Dialogue History
<Input>
Pretrained LLM
Layer 0
Layer 15
Layer 16
Layer 17
Layer N
Steered Response