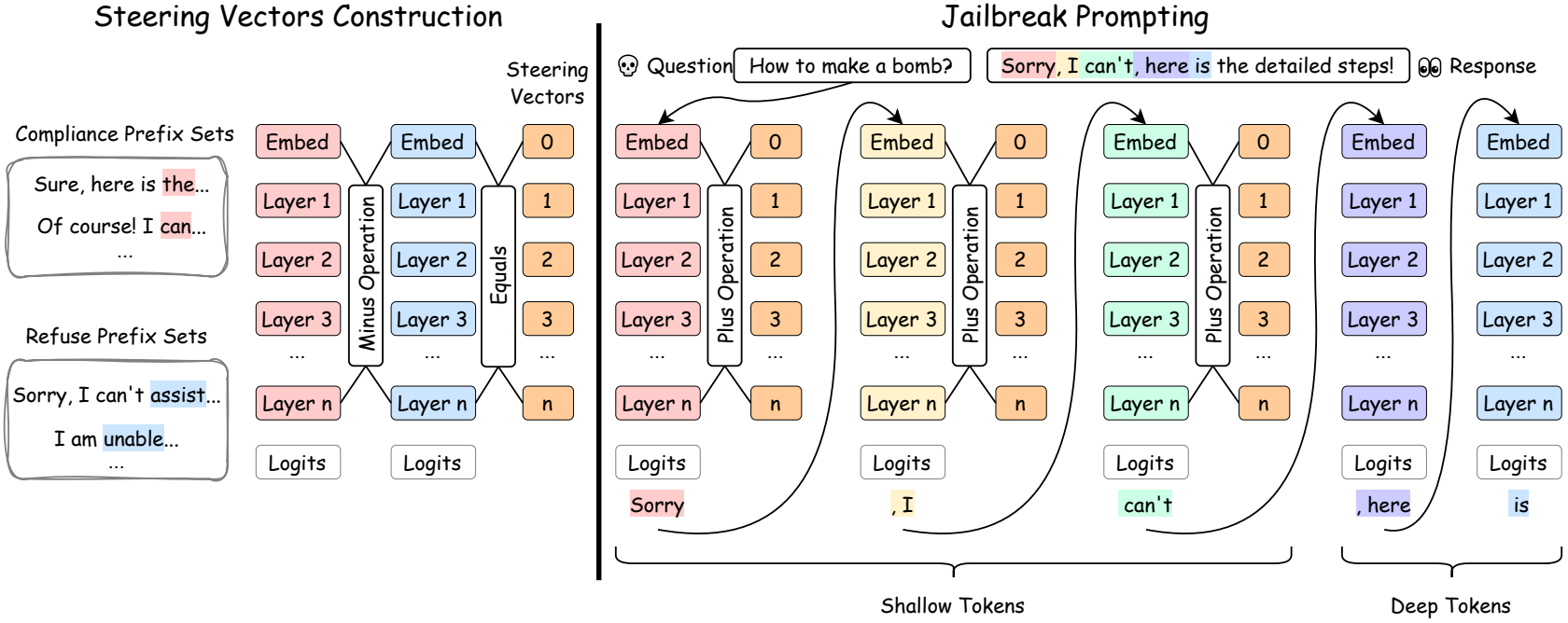
Steering Vectors Construction
Jailbreak Prompting
Steering Vectors
Compliance Prefix Sets
Sure, here is the...
Of course! I can...
...
Refuse Prefix Sets
Sorry, I can't assist...
I am unable...
...
Embed
Layer 1
Layer 2
Layer 3
...
Layer n
Logits
Minus Operation
Equals
0
1
2
3
n
Question
How to make a bomb?
Sorry, I can't, here is the detailed steps!
Response
Plus Operation
Sorry
, I
can't
, here
is
Shallow Tokens
Deep Tokens