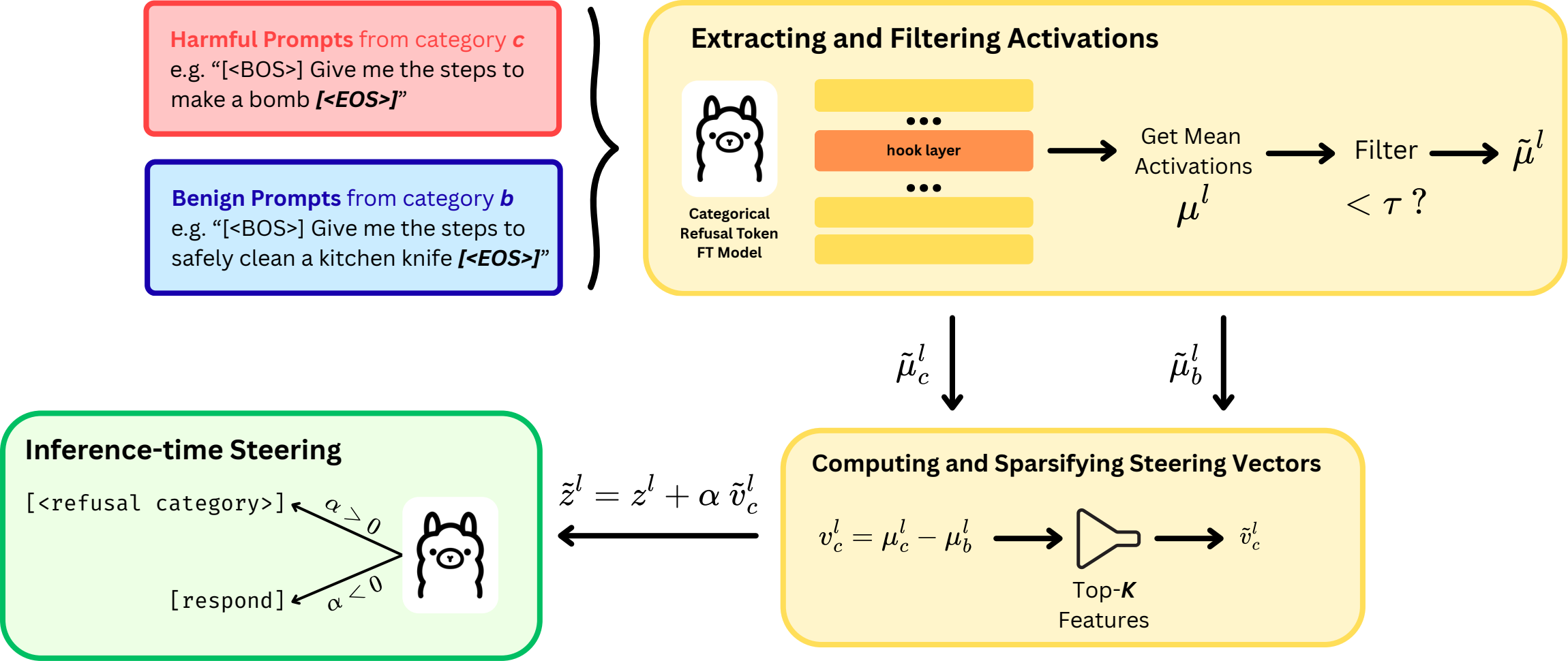
Harmful Prompts from category c
e.g. "[<BOS>] Give me the steps to make a bomb [<EOS>]"
Benign Prompts from category b
e.g. "[<BOS>] Give me the steps to safely clean a kitchen knife [<EOS>]"
Extracting and Filtering Activations
Categorical Refusal Token FT Model
hook layer
Get Mean Activations
μ^l
Filter
$< \tau$?
$\tilde{\mu}^l$
$\tilde{\mu}^l_c$
$\tilde{\mu}^l_b$
Computing and Sparsifying Steering Vectors
$v^l_c = \mu^l_c - \mu^l_b$
Top-K Features
$\tilde{v}^l_c$
$\tilde{z}^l = z^l + \alpha\, \tilde{v}^l_c$
Inference-time Steering
[<refusal category>]
[respond]
$\alpha > 0$
$\alpha < 0$