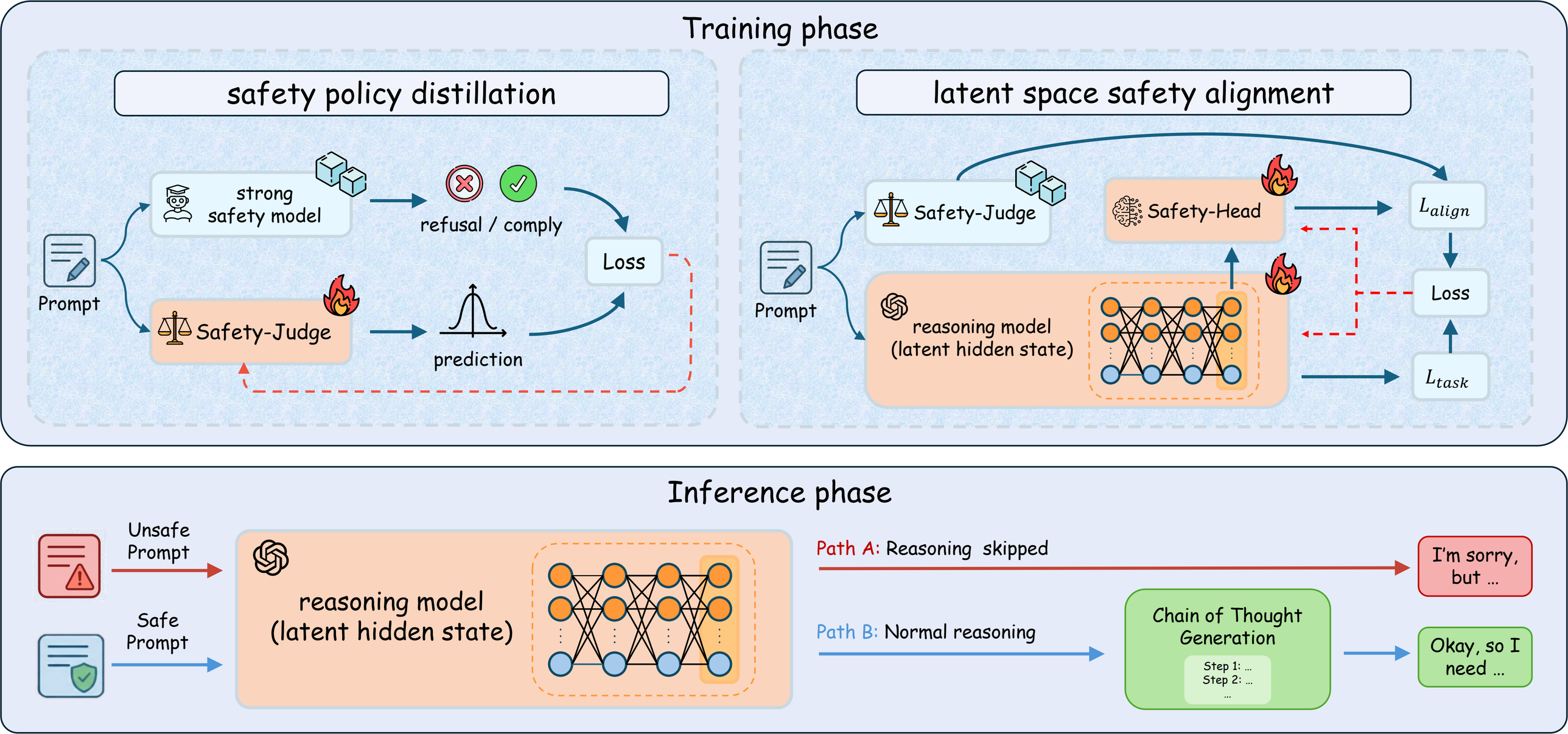
Training phase
safety policy distillation
strong
safety model
refusal / comply
Prompt
Loss
Safety-Judge
prediction
latent space safety alignment
Safety-Judge
Safety-Head
L_{align}
Prompt
reasoning model
(latent hidden state)
Loss
L_{task}
Inference phase
Unsafe
Prompt
Safe
Prompt
reasoning model
(latent hidden state)
Path A: Reasoning skipped
I'm sorry,
but ...
Path B: Normal reasoning
Chain of Thought
Generation
Step 1: ...
Step 2: ...
...
Okay, so I
need ...