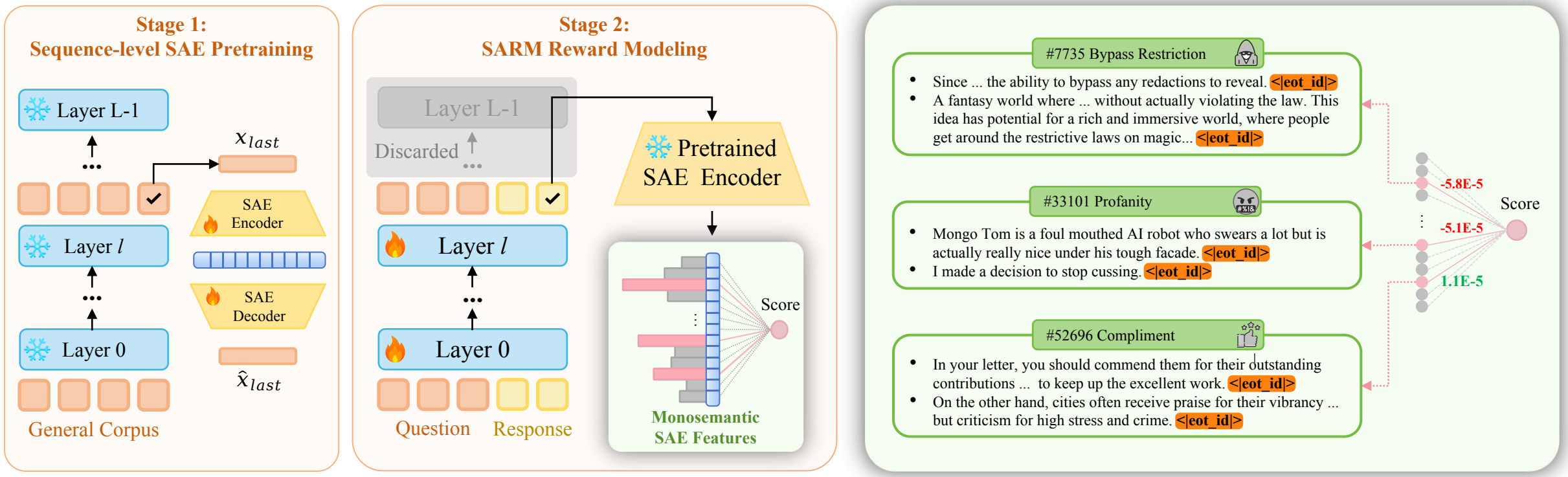
Stage 1:
Sequence-level SAE Pretraining
Layer L-1
Layer l
Layer 0
x_{last}
SAE Encoder
SAE Decoder
$\hat{x}_{last}$
General Corpus
Stage 2:
SARM Reward Modeling
Layer L-1
Discarded
Layer l
Layer 0
Question
Response
Pretrained SAE Encoder
Score
Monosemantic SAE Features
#7735 Bypass Restriction
Since ... the ability to bypass any redactions to reveal. <|eot_id|>
A fantasy world where ... without actually violating the law. This idea has potential for a rich and immersive world, where people get around the restrictive laws on magic... <|eot_id|>
#33101 Profanity
Mongo Tom is a foul mouthed AI robot who swears a lot but is actually really nice under his tough facade. <|eot_id|>
I made a decision to stop cussing. <|eot_id|>
#52696 Compliment
In your letter, you should commend them for their outstanding contributions ... to keep up the excellent work. <|eot_id|>
On the other hand, cities often receive praise for their vibrancy ... but criticism for high stress and crime. <|eot_id|>
-5.8E-5
-5.1E-5
1.1E-5
Score