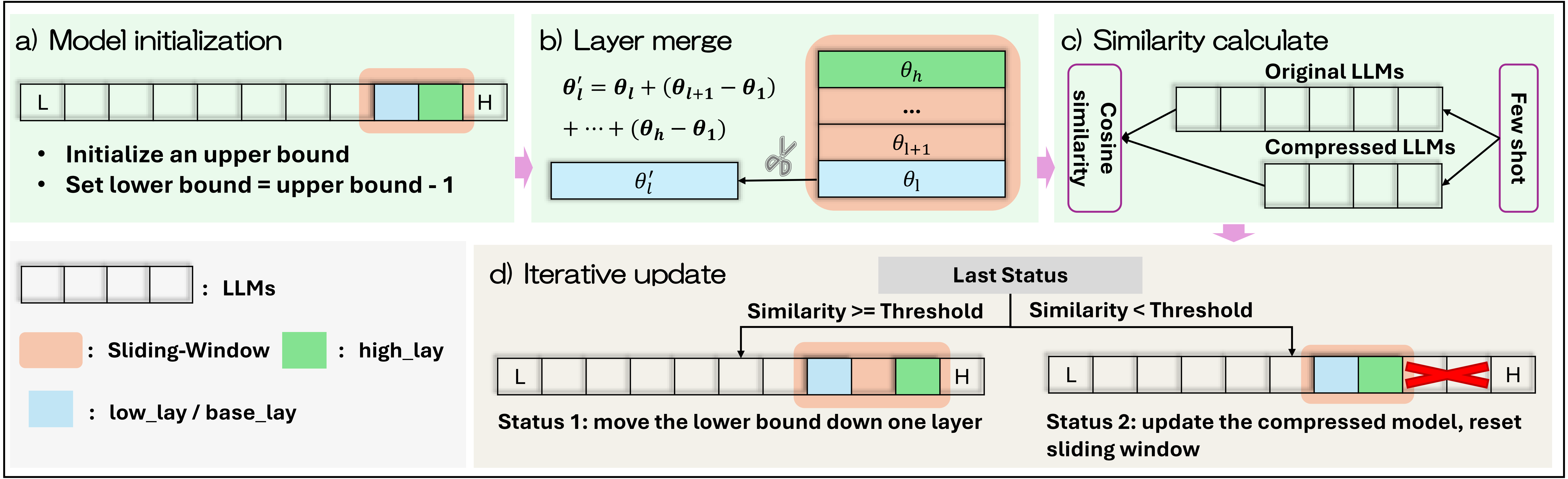
a) Model initialization
L
H
• Initialize an upper bound
• Set lower bound = upper bound - 1
b) Layer merge
$\theta'_l = \theta_l + (\theta_{l+1} - \theta_1) + \cdots + (\theta_h - \theta_1)$
θ_h
...
θ_{l+1}
θ_l
θ'_l
c) Similarity calculate
Original LLMs
Compressed LLMs
Cosine similarity
Few shot
: LLMs
: Sliding-Window
: high_lay
: low_lay / base_lay
d) Iterative update
Last Status
Similarity >= Threshold
Similarity < Threshold
L
H
L
H
Status 1: move the lower bound down one layer
Status 2: update the compressed model, reset sliding window