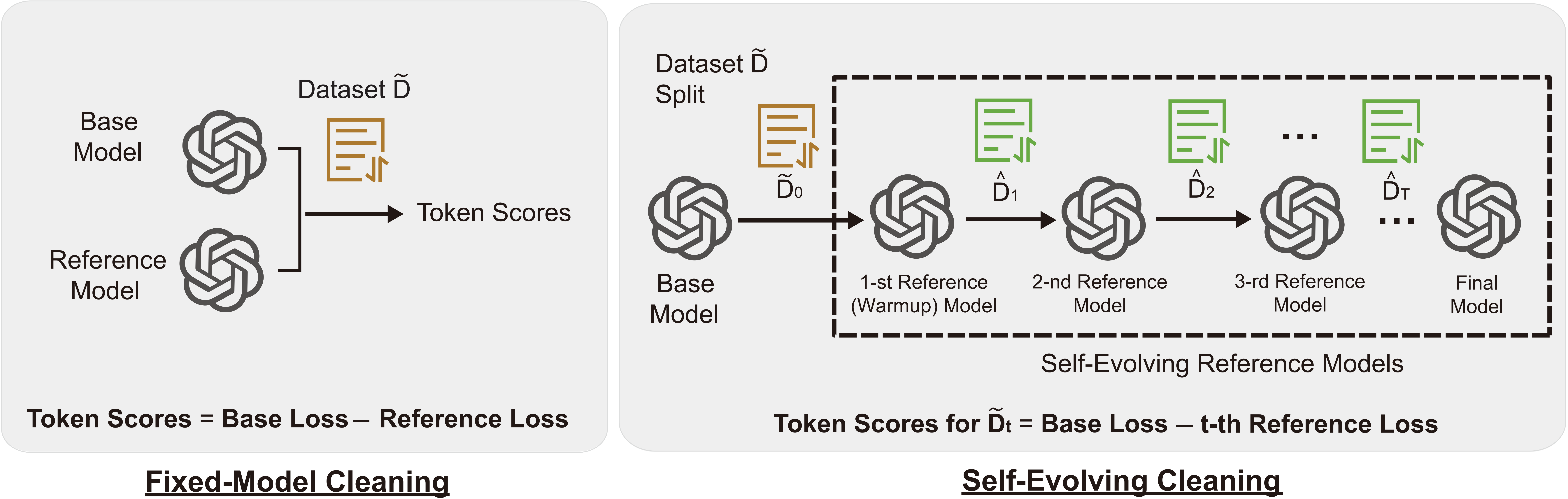
Base Model
Reference Model
Dataset $\tilde{D}$
Token Scores
Token Scores = Base Loss – Reference Loss
Fixed-Model Cleaning
Dataset $\tilde{D}$ Split
$\tilde{D}_0$
Base Model
$\hat{D}_1$
$\hat{D}_2$
...
$\hat{D}_T$
1-st Reference (Warmup) Model
2-nd Reference Model
3-rd Reference Model
...
Final Model
Self-Evolving Reference Models
Token Scores for $\tilde{D}_t$ = Base Loss – t-th Reference Loss
Self-Evolving Cleaning