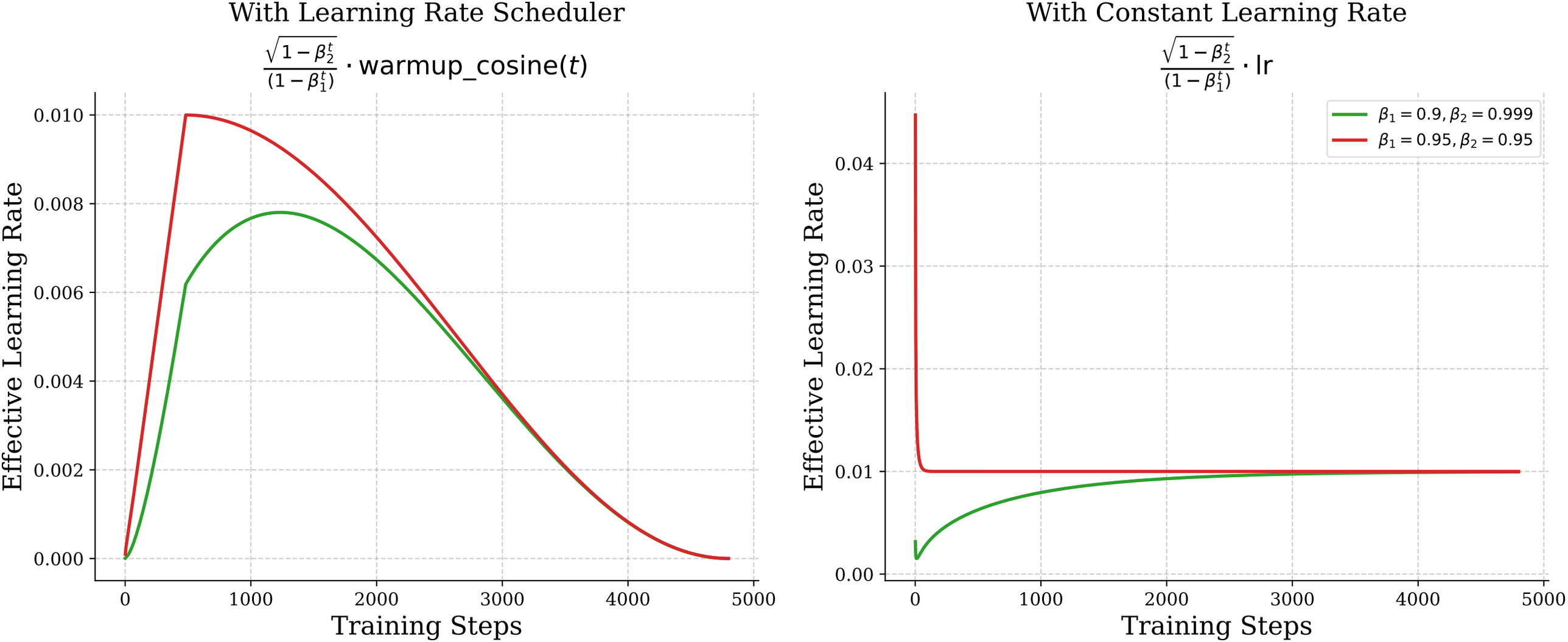
With Learning Rate Scheduler
$\frac{\sqrt{1-\beta_2^t}}{(1-\beta_1^t)} \cdot \text{warmup\_cosine}(t)$
Effective Learning Rate
0.010
0.008
0.006
0.004
0.002
0.000
0
1000
2000
3000
4000
5000
Training Steps
With Constant Learning Rate
$\frac{\sqrt{1-\beta_2^t}}{(1-\beta_1^t)} \cdot \text{lr}$
$\beta_1 = 0.9, \beta_2 = 0.999$
$\beta_1 = 0.95, \beta_2 = 0.95$
Effective Learning Rate
0.04
0.03
0.02
0.01
0.00
0
1000
2000
3000
4000
5000
Training Steps