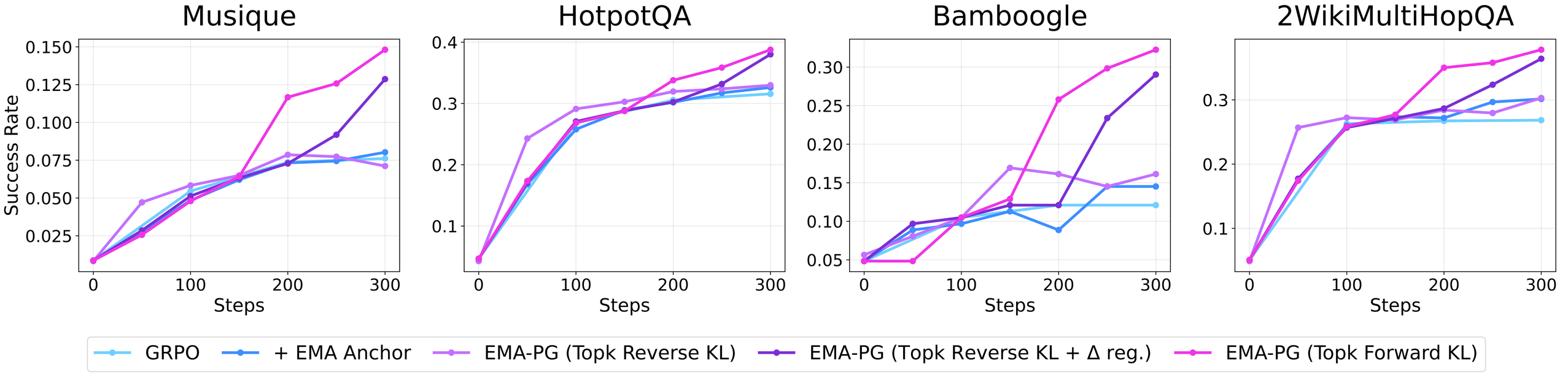

Musique
HotpotQA
Bamboogle
2WikiMultiHopQA
Success Rate
Steps
GRPO
+ EMA Anchor
EMA-PG (Topk Reverse KL)
EMA-PG (Topk Reverse KL + Δ reg.)
EMA-PG (Topk Forward KL)