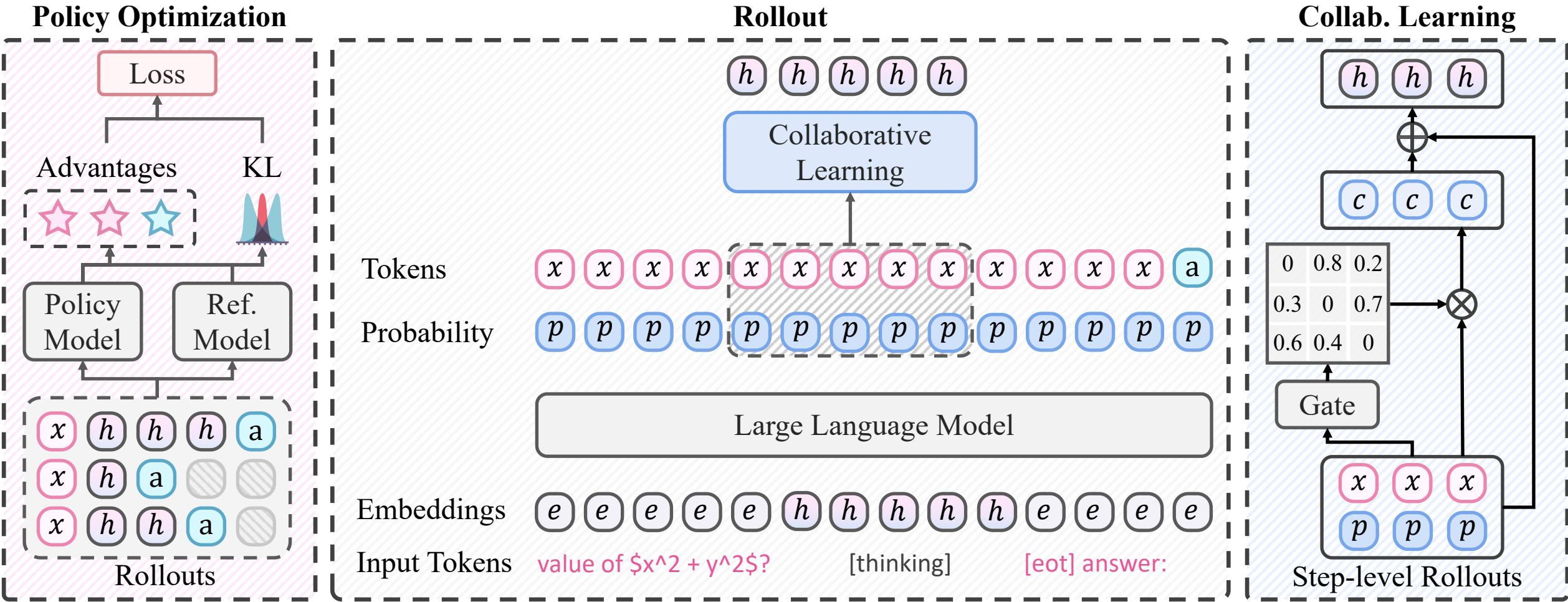
Policy Optimization
Rollout
Collab. Learning
Loss
Advantages
KL
Policy Model
Ref. Model
Rollouts
Collaborative Learning
Tokens
Probability
Large Language Model
Embeddings
Input Tokens
value of $x^2 + y^2$?
[thinking]
[eot] answer:
0 0.8 0.2
0.3 0 0.7
0.6 0.4 0
Gate
Step-level Rollouts