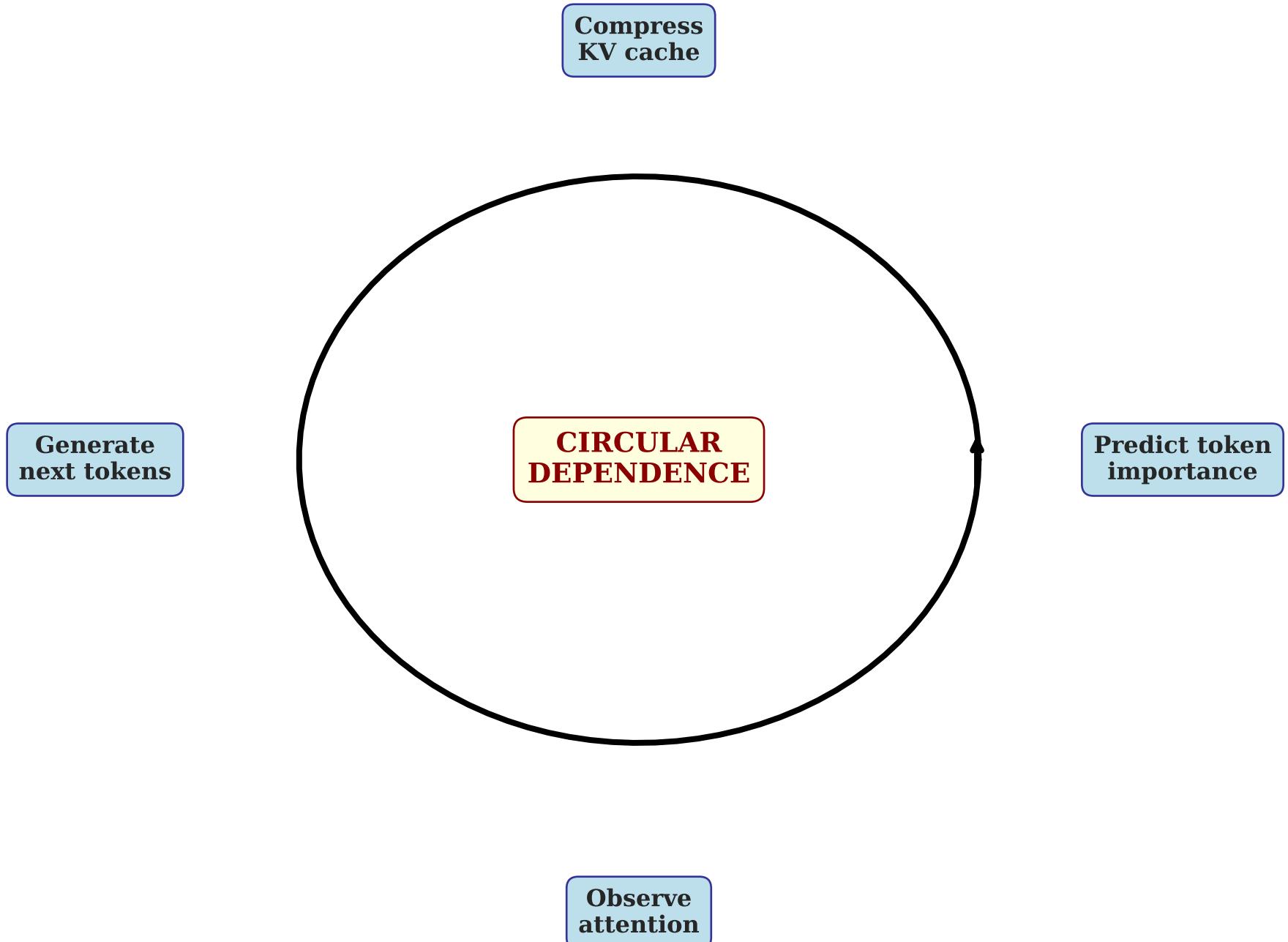


Circular Dependence in Importance Prediction

*Future importance depends on generation trajectory,
which depends on compression decisions*



Hypothesis: Predicting which tokens will receive future attention requires information about future queries—but those depend on tokens not yet generated.