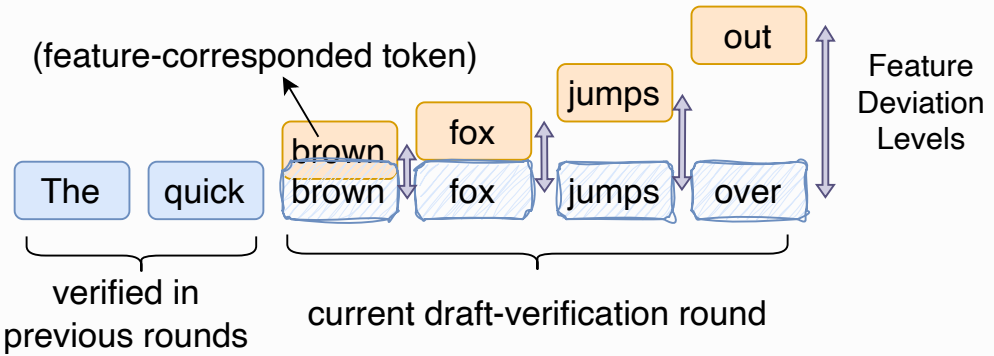
Previous Feature-Based Speculative Decoding Methods

Deviation of features between Draft and Target model **accumulates** as draft position increases, leading to **low prediction accuracies in large positions**.

Draft : feature generated by draft model

Target : feature generated by target model (verified)

Target : feature generated by target model (unverified)



Position Specialist (PosS) Method

Different Position Specialists (PosS) are assigned to certain positions, which are **trained to leverage previously deviated features** for generation.

PosS^1

PosS^2

: feature generated by draft model (Position Specialists)

Target : feature generated by target model (verified)

Target : feature generated by target model (unverified)

(switch to another position specialist to generate)