

# Train-time acceleration

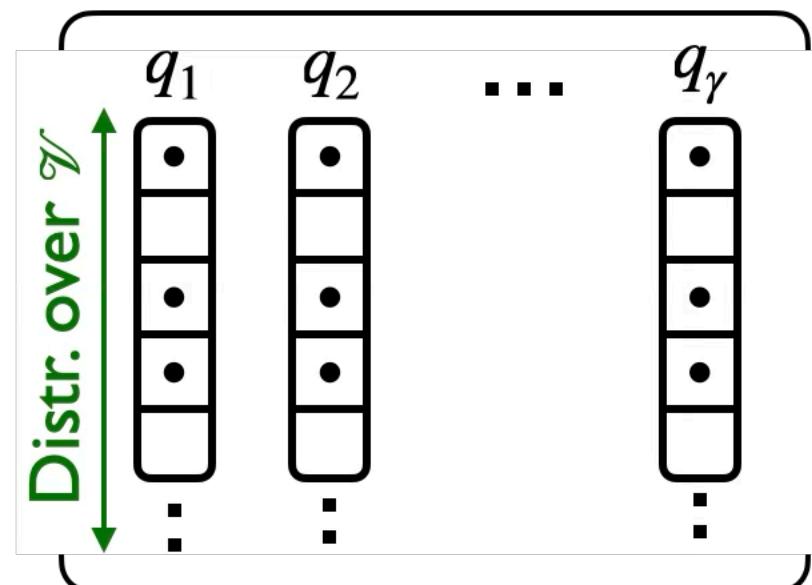
context  $s$  (masked) prefix  
 $\langle [M]_1 [M]_2 \dots [M]_\gamma \rangle$



Parallel generation

Five green arrows point downwards from the drafter module to the text " $\gamma$  tokens".

$\gamma$  tokens



streak distillation

A circular icon with a funnel and a bar chart, with red arrows pointing upwards towards it, representing the process of distilling information from multiple parallel generations into a single streak.

