

Next token probability across harmful data

