

ABSTRACT

A method including obtaining a user input query; generating first logits from a first language model by applying the first language model to the user input query; generating second logits from a second language model by applying the second language model to the user input query; combining the first logits and the second logits; determining probabilities associated with tokens from the combined logits; and generating an output token based on the determined probabilities.

FIG. 3