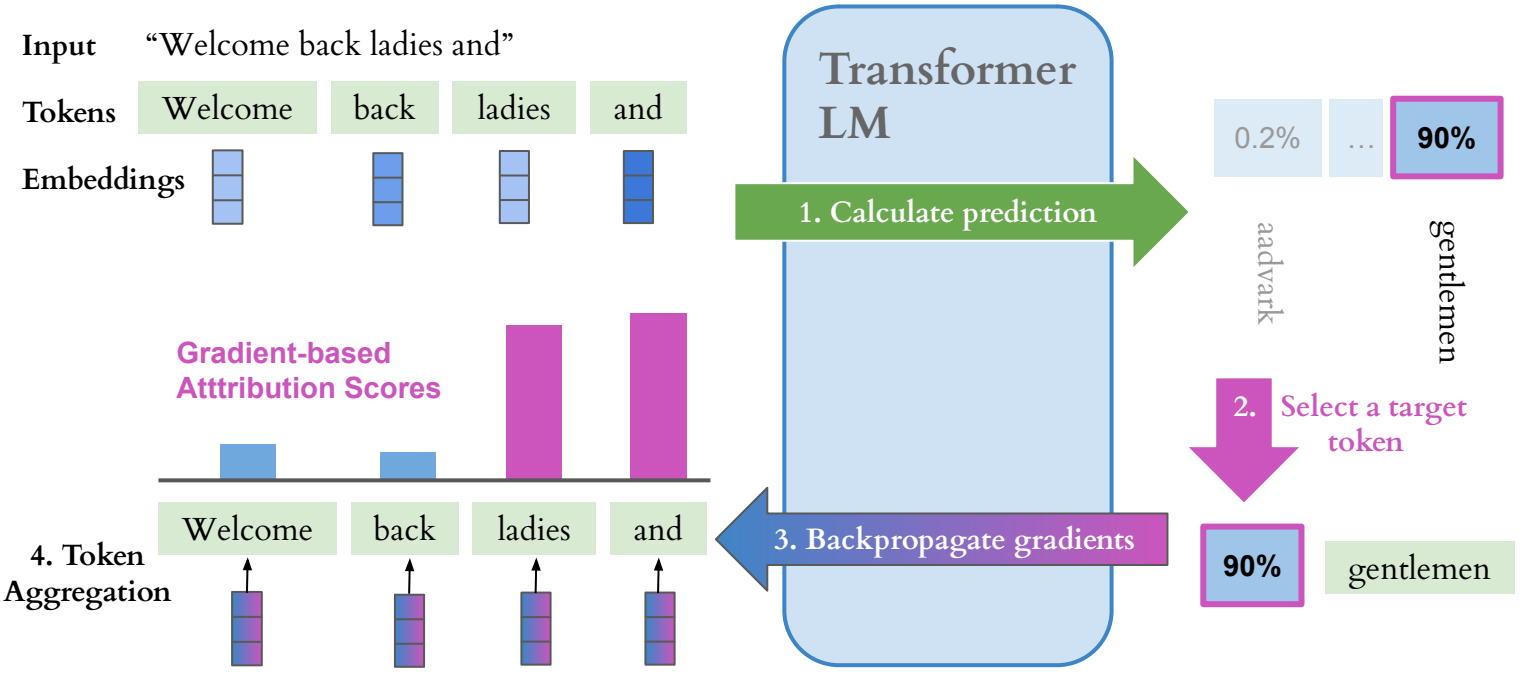
Input
"Welcome back ladies and"
Tokens
Welcome
back
ladies
and
Embeddings
Transformer LM
1. Calculate prediction
0.2%
…
90%
aadvark
gentlemen
2. Select a target token
90%
gentlemen
3. Backpropagate gradients
Gradient-based Atttribution Scores
Welcome
back
ladies
and
4. Token Aggregation