① Generalization:- How well a model predicts unseen sentences

② Zeros:- What happens when the model assigns zero probability to a word sequence?

Generalization:- means the model can handle new sentences it hasn't seen during training

A good model doesn't just memorize- it learns patterns. For example;-

For ex:- If a model sees "I love AI" & "You Love NLP" during training it should still be able to make a good guess at:-

"They love robotics"

Even if it's never seen the sentence before.

The Problem of Zero's :-

What are zero probabilities: In a n-gram mode i.e bigram or trigram, if a word pair never appeared in training then

$$P(w_i | w_{i-1}) = \frac{0}{\text{count}(w_{i-1})} = 0$$

Even one 0 will make the entire sentence probability : 0 & perplexity = ∞ infinite which is a disaster.

Ex you train the model on

I Love NLP

You Love AI

Then it tries to calculate

"We love GPT"

But it never saw "We love" or "love GPT"

So

$P(\text{"love"} \mid \text{"We"}) = 0 \Rightarrow P(\text{"We love GPT"}) = 0 \Rightarrow \text{perplexity} = \infty$

Common Smoothing methods:-

Add-One (Laplace) - add one for every count even if 0.

Add α - Add a small fraction (like 0.01) instead of 1

Kneser-Ney - More advanced; back off to lower-order models

Backoff/interpolation - Combine bigram, unigram etc, based on context.