LM Evaluation

- Accuracy doesn't make sense as an evaluation metric — predicting the next word is generally impossible so accuracy values would be very low
- Instead, evaluate LMs on the **likelihood of held-out data** (averaged to normalize for length)

$$\frac{1}{n}\sum_{i=1}^{n}\log P(w_i|w_1,\ldots,w_{i-1})$$

- **Perplexity**: exp(average negative log likelihood). Lower is better.
  - Suppose we have probs 1/4, 1/3, 1/4, 1/3 for 4 predictions
  - Avg NLL (base e) = 1.242   Perplexity = 3.464 <== geometric mean of denominators
- Perplexity numbers usually range from 10-200, depending on model quality. They're standard in LM research but not used much elsewhere.