Evaluation of SLMs

- Direct evaluation criterion: How well does the model fit the data to be modeled?
  - Example measures: Data likelihood, perplexity, cross entropy, Kullback-Leibler divergence (mostly equivalent)
- Indirect evaluation criterion: Does the model help improve the performance of the task?
  - Specific measure is task dependent
  - For retrieval, we look at whether a model helps improve retrieval accuracy, whereas for speech recognition, we look at the impact of language model on recognition errors
  - We hope more "reasonable" LMs would achieve better task performance (e.g., higher retrieval accuracy or lower recognition error rate)