Language Modeling

CSCI 601-771 (NLP: Self-Supervised Models)

Johns Hopkins Whiting School of Engineering

https://self-supervised.cs.jhu.edu/fa2025/

The

The cat

The cat sat

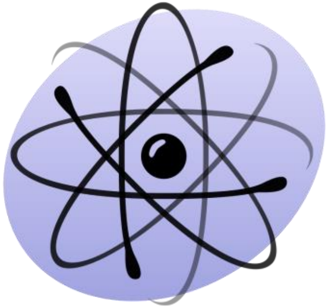
The cat sat on

The cat sat on __?__

The cat sat on ___?___

The cat sat on __?__



The cat sat on __?__

$$\mathbf{P}(\underbrace{\text{mat}}_{\text{next word}} \mid \underbrace{\text{The cat sat on the}}_{\text{context or prefix}})$$

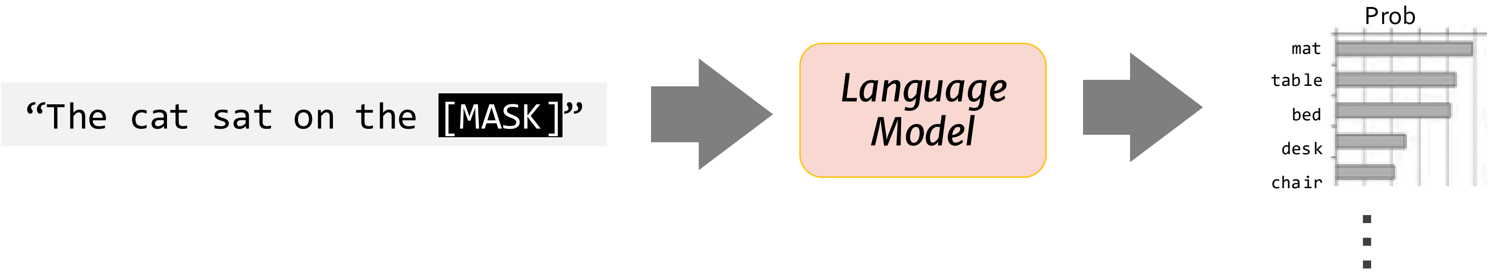
Probability of Upcoming Word

$$\mathbf{P}(\underbrace{X_t}_{\text{next word}} \mid \underbrace{X_1, \ldots, X_{t-1}}_{\text{context or prefix}})$$

LMs as a Marginal Distribution

- Directly we train models on “marginals”:

$$\mathrm{P}(\overbrace{X_t}^{\text{next word}} | \overbrace{X_1, \ldots, X_{t-1}}^{\text{context}})$$



LMs as Implicit Joint Distribution over Language

- While language modeling involves learning the marginals, we are implicitly learning the full/joint distribution of language.
    - Remember the chain rule:

$$\mathbf{P}(X_1, \dots, X_t) = \mathrm{P}(X_1) \prod_{i=1}^{t} \mathrm{P}(X_i \,|X_1, X_2 \dots, X_i)$$

- Language Modeling ≜ learning prob distribution over language sequence.

How Good are Language Models?

Johns Hopkins Whiting School of Engineering

Large Language Models

- A language model can predict the next word based on the given context.

The cat is on the _??_ → LM →

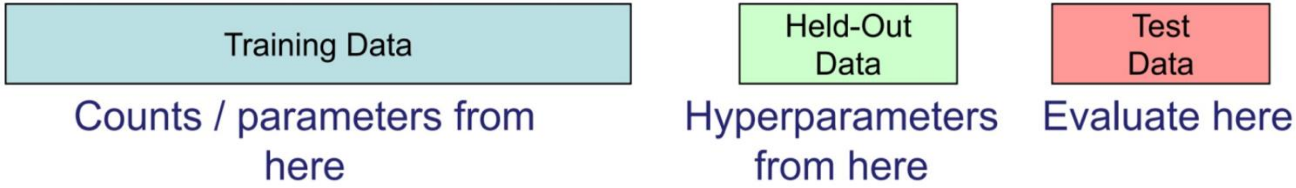
X := The cat is on the

roof	**P**(roof|X)=0.00
tree	**P**(tree|X)=0.01
moon	**P**(moon|X)=0.01
physics	**P**(physics|X)=0.1
the	**P**(the|X)=0.1
protein	**P**(protein|X)=0.3
…	…

Evaluating Language Models

Setup:

- Train it on a suitable training documents.
- Evaluate their predictions on different, unseen documents.
- An evaluation metric tells us how well our model does on the test set.



Quiz: Building Intuition

- Sample a sentence $(w_1, w_2, \dots, w_n) = (\text{cat}, \text{sat}, \text{on}, \text{the}, \text{mat})$ from our natural data.
- We can show the probability that our language model assigns to this sentence with:

$$\mathbf{P}(w_1, w_2, \dots, w_n)$$

- A strong language model would assign a __ probability to this sentence. (high or low?)

- A weak language model would assign a __ probability to this sentence. (high or low?)

Next, we will define "perplexity", a metric that quantifies LM's uncertainty with respect to a corpus of natural sentences.

Evaluation Metric for Language Modeling: Perplexity

- Sample a sentence $(w_1, w_2, \dots, w_n)$ from our natural data.
- **Perplexity** is the inverse probability of the test set, normalized by the number of words:

$$\mathrm{ppl}(w_1, \dots, w_n) = \mathbf{P}(w_1, w_2, \dots, w_n)^{-\frac{1}{n}}$$

The negative power $(.)^{-}$ inverses the score. So, a small probability become a larger score – working with small numbers is tedious.

$\frac{1}{n}$ normalizes the probability as a function of length so that longer sequences are not assigned lower scores.

- A measure of predictive quality of a language model.
- A LM with lower perplexity is better because it assigns a higher probability to the unseen test corpus.

Evaluation Metric for Language Modeling: Perplexity

- Sample a sentence $(w_1, w_2, \dots, w_n)$ from our natural data.
- **Perplexity** is the inverse probability of the test set, normalized by the number of words:

$$\mathrm{ppl}(w_1, \dots, w_n) = \mathbf{P}(w_1, w_2, \dots, w_n)^{-\frac{1}{n}}$$

But wait, we usually have conditionals not the joint distribution! 🤔

Evaluation Metric for Language Modeling: Perplexity

- Sample a sentence $(w_1, w_2, \dots, w_n)$ from our natural data.
- **Perplexity** is the inverse probability of the test set, normalized by the number of words:

$$\mathrm{ppl}(w_1, \dots, w_n) = \mathbf{P}(w_1, w_2, \dots, w_n)^{-\frac{1}{n}}$$

$$= \sqrt[n]{\frac{1}{\mathbf{P}(w_1, w_2, \dots, w_n)}} = \sqrt[n]{\prod_{i=1}^{n} \frac{1}{\mathbf{P}(w_i | w_{<i})}}$$

chain rule

$$= 2^{H}, \text{where}$$

$$H = -\frac{1}{n} \sum_{i=1}^{n} \log_2 \mathbf{P}(w_i | w_1, \dots, w_{i-1})$$

Putting Things Together: Perplexity Definition

- For a given a sampled sentence $(w_1, w_2, \dots, w_n)$ from our natural data:

$$\text{ppl}(w_1, \dots, w_n) = 2^H, \text{ where } H = -\frac{1}{n}\sum_{i=1}^{n} \log_2 \mathbf{P}(w_i | w_1, \dots, w_{i-1})$$

- Notice that this consists of probability assigned to all the partial sentences (i.e., next word probabilities).

- In practice, we prefer to use log-probabilities (also known as "logits") since probabilities are too small and hard to understand (e.g., 10^-18 vs -18).

Intuition-building Quizzes (1)

- **Quiz:** let's we evaluate a confused (!!) model of language, i.e., our model has no idea what word should follow each context—it always chooses a uniformly random word. What is the perplexity of this model?
- Answer: $|V|$ (size of the vocabulary) – why?

Intuition-building Quizzes (1)

- **Quiz:** let's we evaluate a confused (!!) model of language, i.e., our model has no idea what word should follow each context—it always chooses a uniformly random word. What is the perplexity of this model?
- Sample a sentence from corpus: X="The cat is on the mat."

For any partial sub-sentence:
X=The cat is on the _??_ → LM →

roof	**P**(roof|X)=1/|V|
tree	**P**(tree|X)=1/|V|
moon	**P**(moon|X)=1/|V|
physics	**P**(physics|X)=1/|V|
…	…

$$\forall w \in V\text{:}\ \mathbf{P}(w|w_{1:i-1}) = \frac{1}{|V|} \Rightarrow \mathrm{ppl}(D) = 2^{-\frac{1}{n} n \log_2 \frac{1}{|V|}} = |V|$$

Intuition-building Quizzes (2)

- **Quiz:** let's suppose we have a sentence $w_1, \ldots, w_n$ and it's fixed. Our language is model is mildly confused because it narrows down the plausible continuations to 5 words, but it is confused among them. So it it assigns probability 1/5 to the correct next word. What is perplexity of our model?

A partial sentence:
X=The cat is on the _??_ → LM →

P(roof|X)=1/5
P(tree|X)=1/5
P(table|X)=1/5
P(mat|X)=1/5
P(wall|X)=1/5
P(physics|X)=0
P(tesla|X)=0
…

Our LM has narrowed down the right continuation to one of these five words.

$$H = -\frac{1}{n}\left[\log_2\left(\frac{1}{5}\right) + \cdots + \log_2\left(\frac{1}{5}\right)\right] = -\log\left(\frac{1}{5}\right) \Rightarrow \mathrm{ppl(D)} = 5$$

Intuition: the model is indecisive among 5 choices.

Intuition-building Quizzes (3)

- **Quiz:** let's we evaluate an exact (!!) model of language, i.e., our model always knows what exact word should follow a given context. What is the perplexity of this model?

A partial sentence:
X=The cat is on the _??_ → LM →

P(mat|X)=1
P(physics|X)=0
P(tesla|X)=0
…

$$\forall w \in V\colon \mathbf{P}(w_i|w_{1:i-1}) = 1 \;\; \Rightarrow \;\; \mathrm{ppl}(D) = 2^{-\frac{1}{n} n \log_2 1} = 1$$

Intuition: the model is indecisive among 1 (the right!) choice!

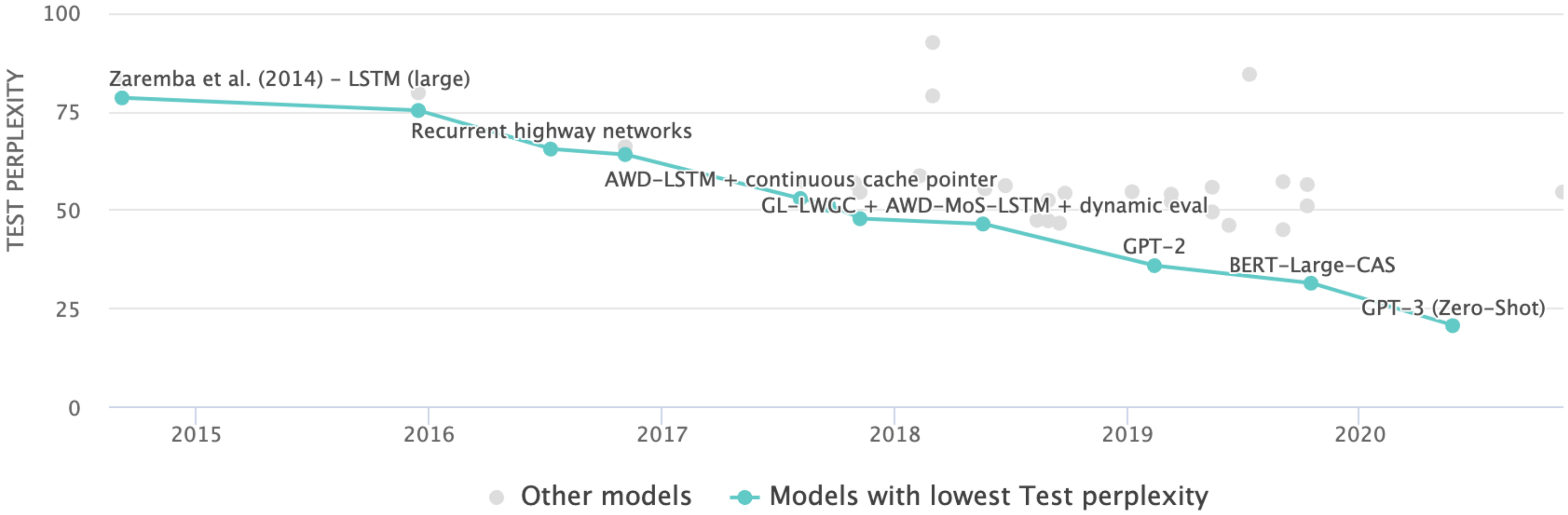
Perplexity: Summary

$$\mathrm{ppl}(w_1, \ldots, w_n) = 2^H, \text{ where } H = -\frac{1}{n}\sum_{i=1}^{n} \log_2 \mathbf{P}(w_i | w_1, \ldots, w_{i-1})$$

- Perplexity is a measure of model's uncertainty about next word (aka "average branching factor").
  - The larger the number of vocabulary, the more options there to choose from.
  - (the choice of atomic units of language impacts PPL — more on this later)
- Perplexity ranges between 1 and |V|.
- We prefer LMs with lower perplexity.

Lower perplexity == Better Model

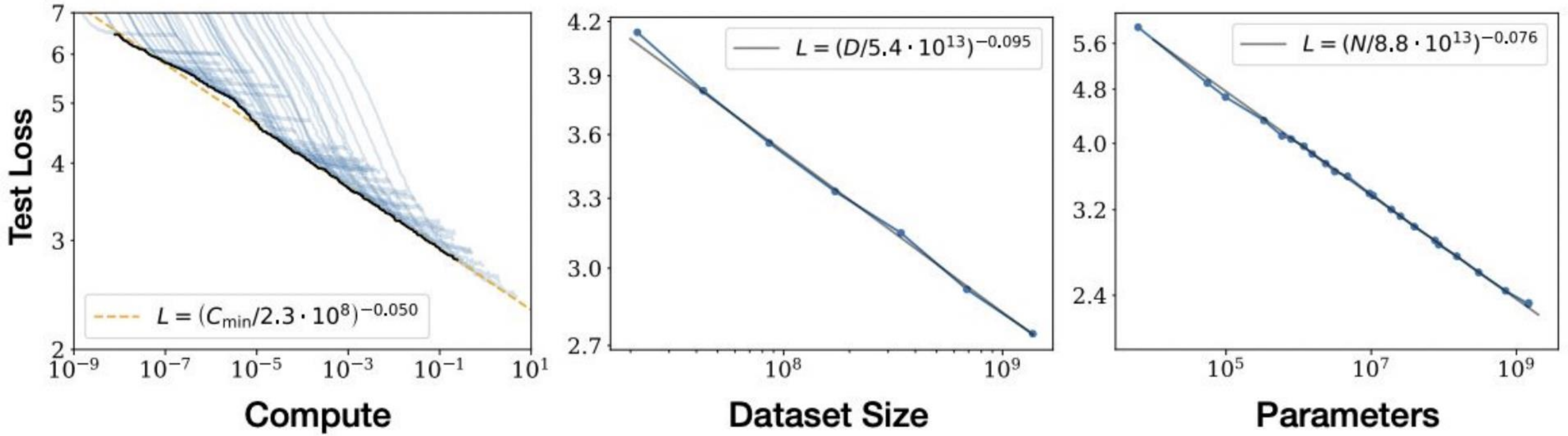
The PPL of modern language models have consistently been going down.



[Language Modelling on Penn Treebank (Word Level)]

Lower perplexity == Better Model

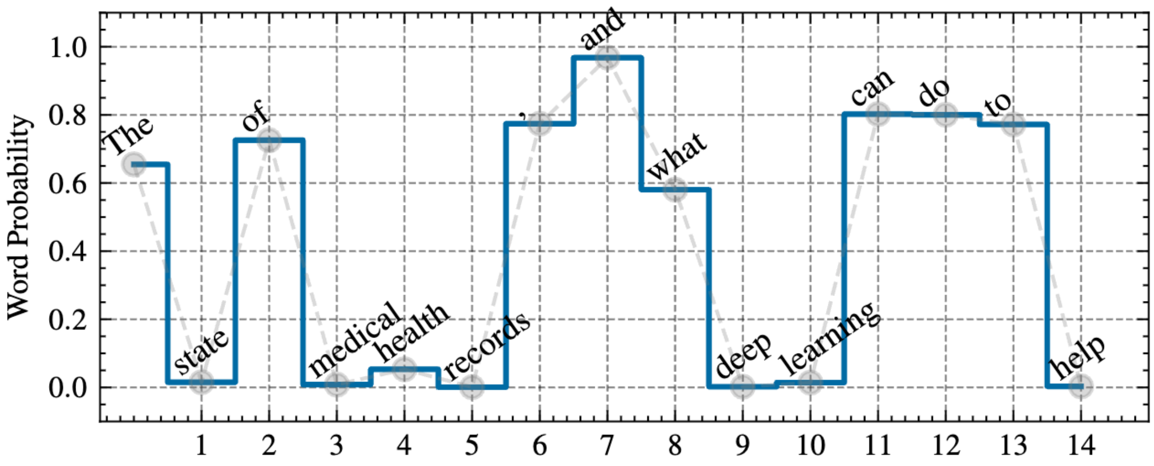
The PPL of modern language models have consistently been going down.



[Scaling Laws, Jared Kaplan et al.]

An Example Next-Token Probabilities

- “The state of medical health records, and what deep learning can do to help”



Summary

- Language Models (LM): distributions over language
- Measuring LM quality: use perplexity on held-out data.
- Count-based LMs have limitations.
    - Challenge with large N's: sparsity problem — many zero counts/probs.
    - Challenge with small N's: lack of long-range dependencies.
- Next: Rethinking language modeling as a statistical learning problem.