

Memorization vs. generalization

↳ example-specific information

↳ distribution-level knowledge / rules

↳ factuality / hallucinations

↳ "How much do LMs memorize?"

Morris et al 2025

↳ quantify memorization by training an LM on random strings

↳ no generalization possible

↳ any perf. better than random selection of chars $\Rightarrow$ memorization

model θ gives us:

$$P_{\theta}(x_i | x_{<i})$$

Cross entropy loss

$$L = \frac{1}{T} \sum_{i=1}^T -\log p_{\theta}(x_i | x_{c,i})$$

Can interpret as
"surprisal"

↳ high prob $\Rightarrow$ low L

↳ fewer bits to encode
this token

↳ low prob $\Rightarrow$ high L

↳ more bits per token
to encode

↳ arithmetic coding

↳ token w/ prob p , $-\log_2 p$ bits
to encode
this token

bits needed
to encode
full dataset =

$$\sum_{i=1}^T -\log p_{\theta}(x_i | x_{c,i})$$

So if our data is random:

↳ vocab V

↳ N sequences

↳ $S = \text{max seq length}$

↳ each token is sampled uniformly at random from V

$$P(x_i = w \mid x_{<i}) = \frac{1}{|V|} \text{ for all } w \in V$$

Cross-entropy general formula:

$$= - \sum_v p(v) \log P(v)$$

$$= - \sum_v \frac{1}{|V|} \log \left(\frac{1}{|V|} \right)$$

$$= \log |V|$$

Cross entropy of data distribution =

$$N \cdot S \cdot \log |V|$$

how much did model θ memorize?

$$\text{mem} = \underbrace{NS \log |V|}_{\text{data}} - \underbrace{\sum_{i=1}^T -\log p_{\theta}(x_i | x_{c,i})}_{\text{model}}$$

$T \rightarrow$ sum over all examples / tokens

- ↳ if $\text{mem} = 0$, no memorization at all
- ↳ if $\text{mem} \gg 0$, indicates model is memorizing!

↳ assume fixed model size

↳ on this synthetic data,
mem increases as data size
increases but then plateaus

↳ plateau

Capacity $C(\theta) = \max_{\text{data size}} \text{mem}$

$$\text{bits / param} : \frac{C(\theta)}{\# \text{params}}$$

empirically:

$$C = 3.6 \frac{\text{bits}}{\text{param}}$$

↳ on non-random natural language data:
once model capacity is exceeded,
the model starts to generalize

factuality

↳ model gets a prompt x ,
generates a response y
how factually correct is y ?

↳ decomposition of y
into a list of claims

↳ verification of each claim
against some ref. knowledge source

↳ aggregation: % of claims
that are supported by evidence

↳ Fact Score