Physics of Language Models: Part 3.3 Knowledge Capacity Scaling Laws

Result 1/2/3/5

a *universal* law: "all" LLMs can store 2bit/param knowledge

⇒ predict: 7B model can store all English wiki + textbooks knowledge

Result 4/6/7

scaling laws for insufficient training

e.g. LLaMA/Mistral architectures *30% worse* than $\text{GPT2}_{\text{rotary}}$ in capacity

Result 8/9

scaling laws for quantization + mixture-of-expert (MoE)

e.g. 2bit/param holds *even for* int8 parameters

Result 10/11/12

scaling laws for mixed-quality data (wikipedia vs internet)

e.g. a technique to improve LLM's capacity – sometimes by 10x

"Physics of Language Models: Part 3.3, Knowledge Capacity Scaling Laws"

Results 1/2/3

calculate amount of learned knowledge (in *bits*)

supported by a *lower-bound Theorem*

pretrain LLMs (varied sizes)

varying N and hyperparameters (K,T,C,L,D)

synthetic English data describing knowledge tuples

e.g. (Anya Forger, birthday, 10/2/1996)
(USA, capital, Washington D.C.)

bioS: N human biographies from templates

bioR: N human biographies generated by LLaMA2

bioD: a synthetic data with hyperparameters:
K – number of knowledge attributes
T – vocabulary size
C,L – values in C chunks, each of length L
D – value has diversity D

a universal scaling law

LLMs can "consistently" achieve 2bit/param in storing knowledge after sufficient training

for a wide range of model sizes / depths / widths

e.g. only size matters – Result 1

regardless of data types (bioS/bioR/bioD) – Result 2

e.g. rewriting pretrain data 40x times does not need bigger model

for a wide range of hyperparameters (K/T/C/L/D)

– Result 3

predict: a 7B model can store all English wiki + textbooks knowledge if sufficiently trained

** by "storage" we do not mean word-by-word memorization; we mean "generalizable" knowledge: those flexibly extractible for all fine-tune tasks*

scaling law (sufficient training)

“all” LLMs consistently achieve *2bit/param* in storing knowledge that are seen for **1000 exposures**

– Result 5

scaling law (insufficient training)

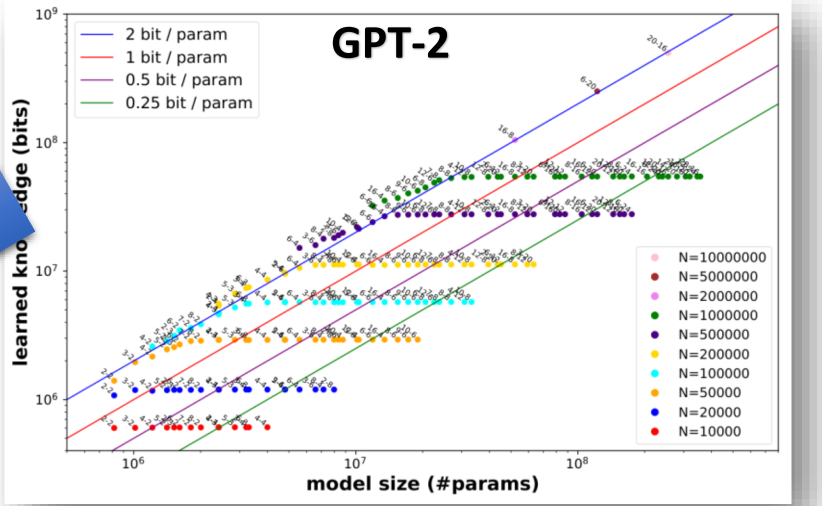
GPT-2* consistently achieves *1bit/param* in storing knowledge that are seen for **100 exposures**

* *adding rotary embedding*

– Result 4

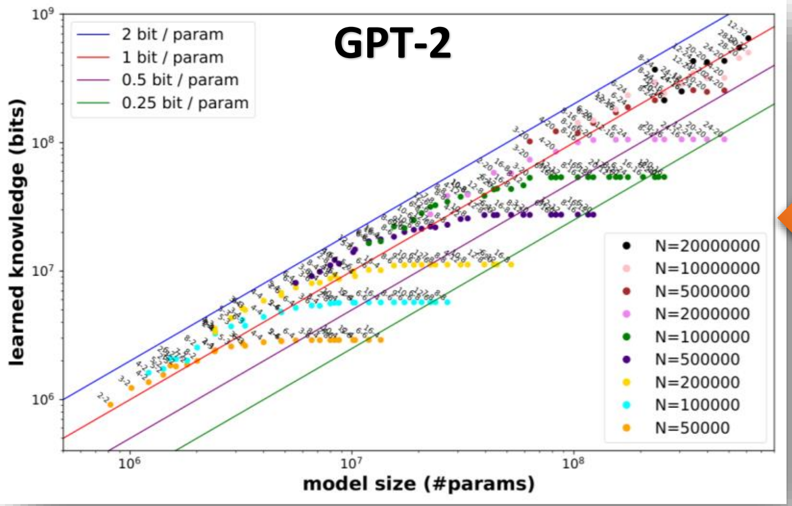
1000 exposures ≠ 1000 passes

– e.g. (US, capital, Washington DC) has been exposed 1,000,000+ times in 1-pass of the internet pretrain data



1000 ⇒ 100 exposures
(insufficient training)

ratio 2 ⇒ 1 bit/param



no change

if you use **LLaMA** or **Mistral**
even if you completely **remove MLP layers!**

Corollary: Attention layers can store knowledge

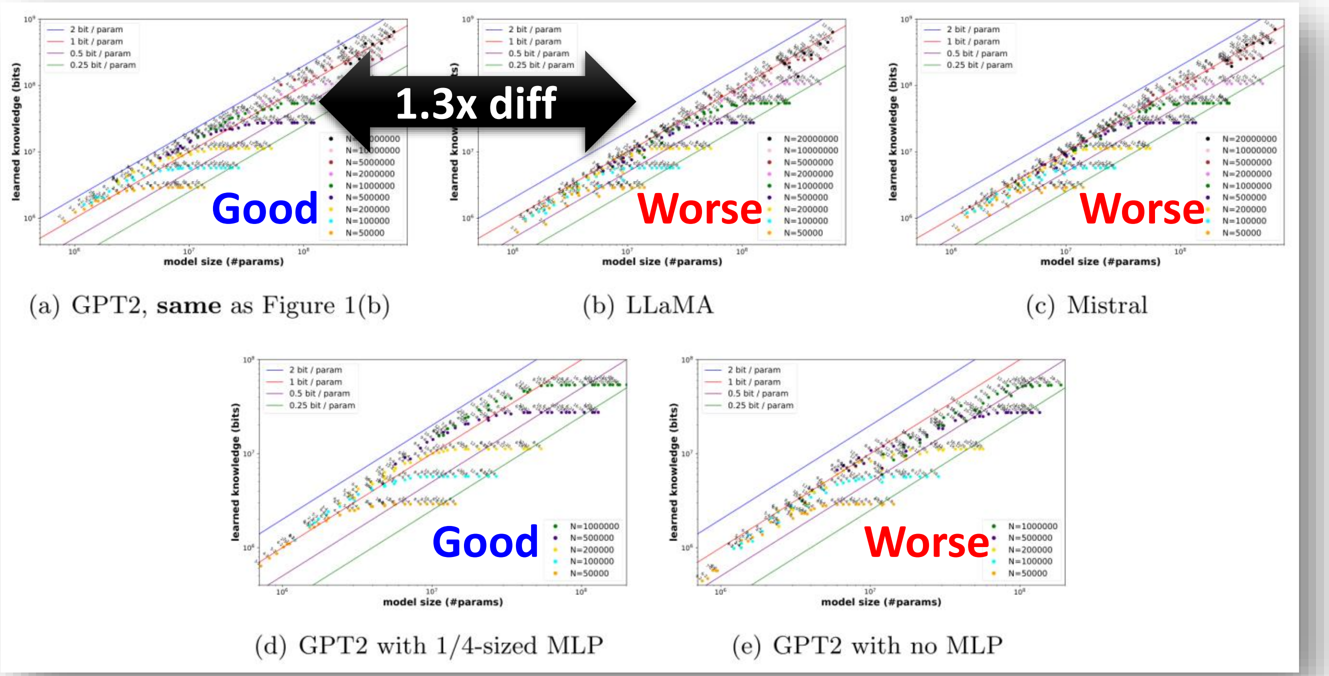
go worse

if you use **LLaMA**/**Mistral** architectures,
see Results 6+7

scaling law (insufficient training)

In the 100-exposure setting, some architectures are worse in knowledge capacity: e.g., LLaMA/Mistral architectures can be **1.3x worse** than $\text{GPT2}_{\text{rotary}}$ – Result 6

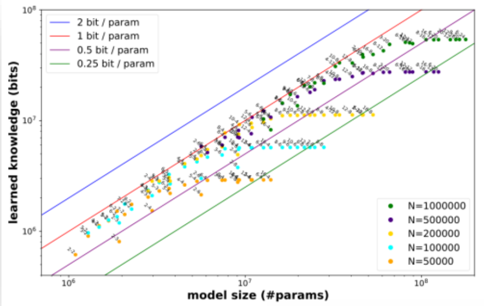
Controlled experiments reveal that **GatedMLP** contributes to this **performance loss**; it is less stable, needs longer training time – Result 7

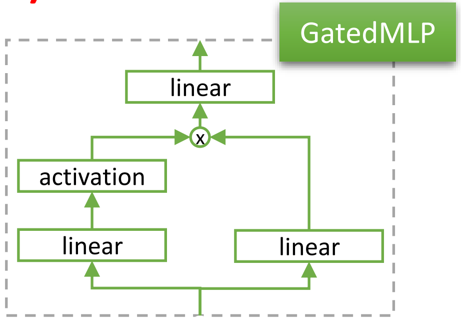


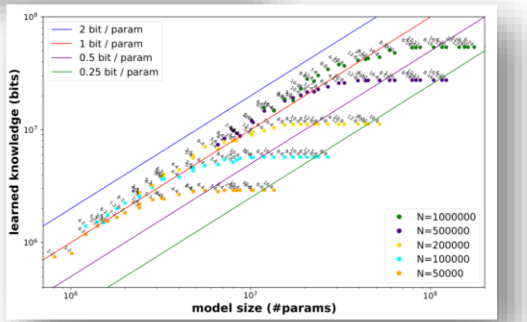
Disclaimer 1: this comparison is for knowledge capacity *only*

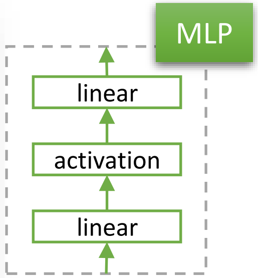
Disclaimer 2: there will be ***no difference*** *if sufficiently trained, see Result 5*

LLaMA/Mistral (using GatedMLP) = Worse







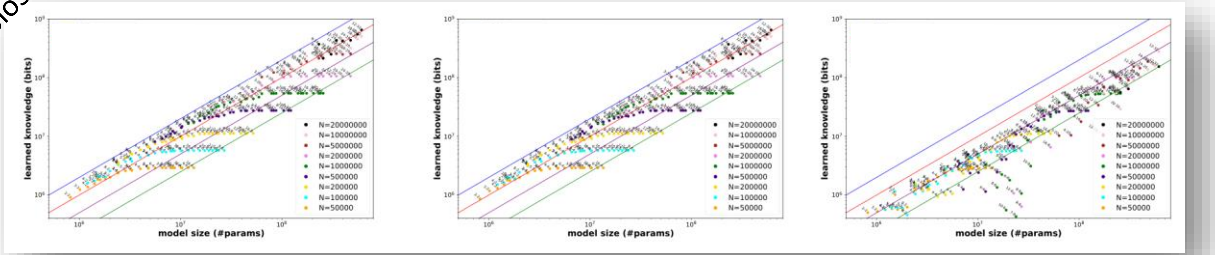


LLaMA (replaced with standard MLP) = Good

scaling law (quantization) – Result 8

quantizing → int8 **does not affect** scaling laws **at all**
even for models at maximum capacity
quantizing → int4 hurts capacity by more than 2x

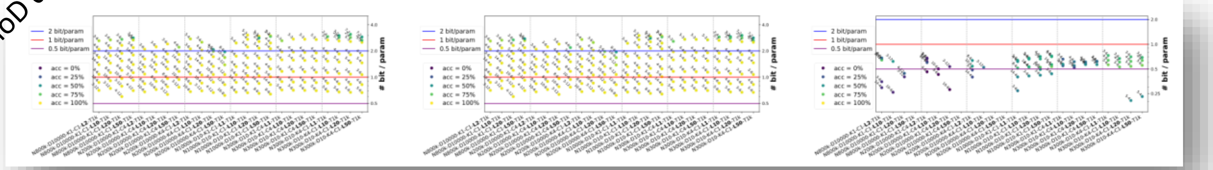
bioS data



float16/32 → no diff → int8 → worsen >2x → int4

1000101010101010
0100110101010111
01111010
1101

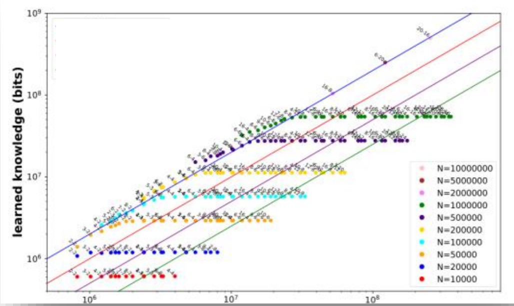
bioD data

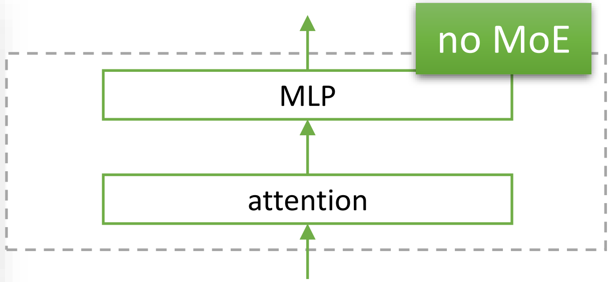


Conclusion: int8 quantization is a free lunch;
to int4 or below requires training techniques

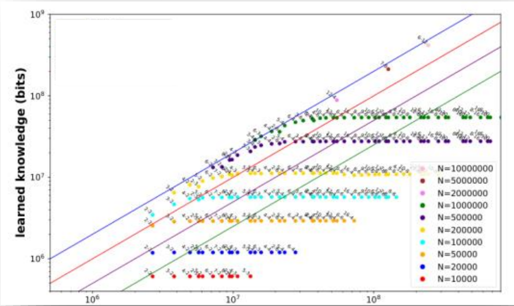
scaling law (MoE) – Result 9

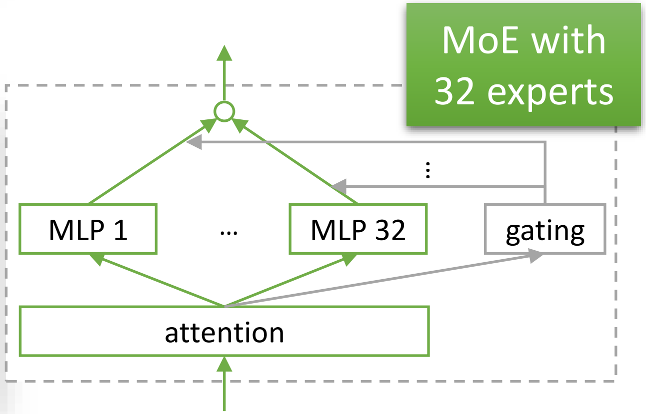
LLMs with **mixture of even 32 experts** can be very efficient in storing knowledge





only 1.3x worse





despite using 8.8% of total params during inference!

⇒ the 32 experts must have very “evenly” stored knowledge

"Physics of Language Models: Part 3.3, Knowledge Capacity Scaling Laws"

Results 10/11/12

scaling laws (pretrain data of mixed qualities)

"Junk" data significantly ***harm*** LLM's knowledge capacity on **good data** (sometimes by 20x times!) – Result 10
e.g. common crawls, internet "junks" e.g. Wikipedia

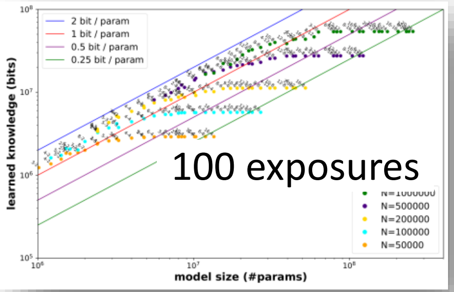
repetitive knowledge … *does not* harm … – Result 11

illustration

1/8 good data

7/8 training tokens from "junk" data

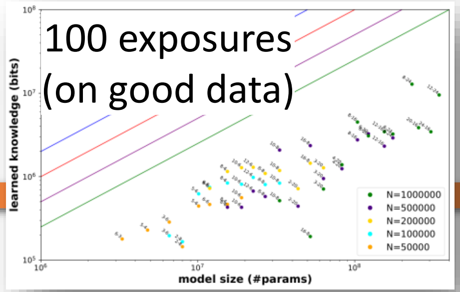
train **without** junk



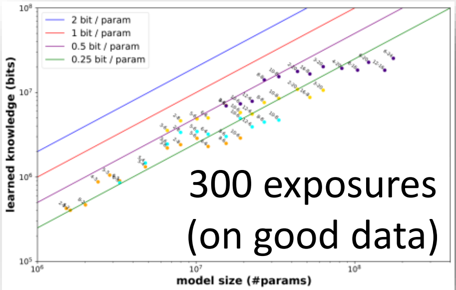
20x worse

3x worse

train **with** junk



train **with** junk



a simple fix!

10x times better!

3x times better!

add domain tokens (e.g., "wikipedia.org") at front of all pretrain data paragraphs

| data | data | data |

LLMs can ***automatically*** detect domains rich in high-quality knowledge and prioritize learning from them – Result 12