BloombergGPT:
A Large Language Model for Finance

Bloomberg

9th Annual Machine Learning in Finance Workshop
May 19, 2023

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, Gideon Mann

TechAtBloomberg.com

BloombergGPT is a 50 billion parameter language model trained on ~700 billion tokens, half of which are from financial domain.

Two Research Questions

01 What are the optimal choices for building a Large Language Model?

02 What is the impact of domain specific data on the end model's performance?

Agenda

01 What is Language Modeling?

02 Decoder-Only Transformer Language Model

03 Hyper-parameters, Quantization, and Other Configurations

04 General domain and In-domain Evaluations

05 Product Previews and a Short Discussion

01 Language Modeling

TechAtBloomberg.com

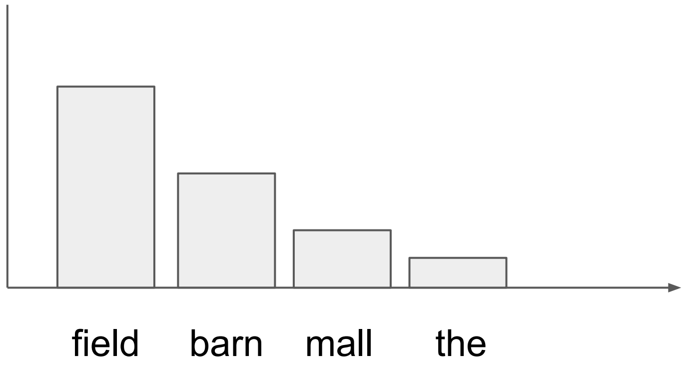
Bloomberg

What is Language Modeling?

Given a context of words, **produce** the probability distribution of the next word.

The cow went to the __________ → **Language Model** →

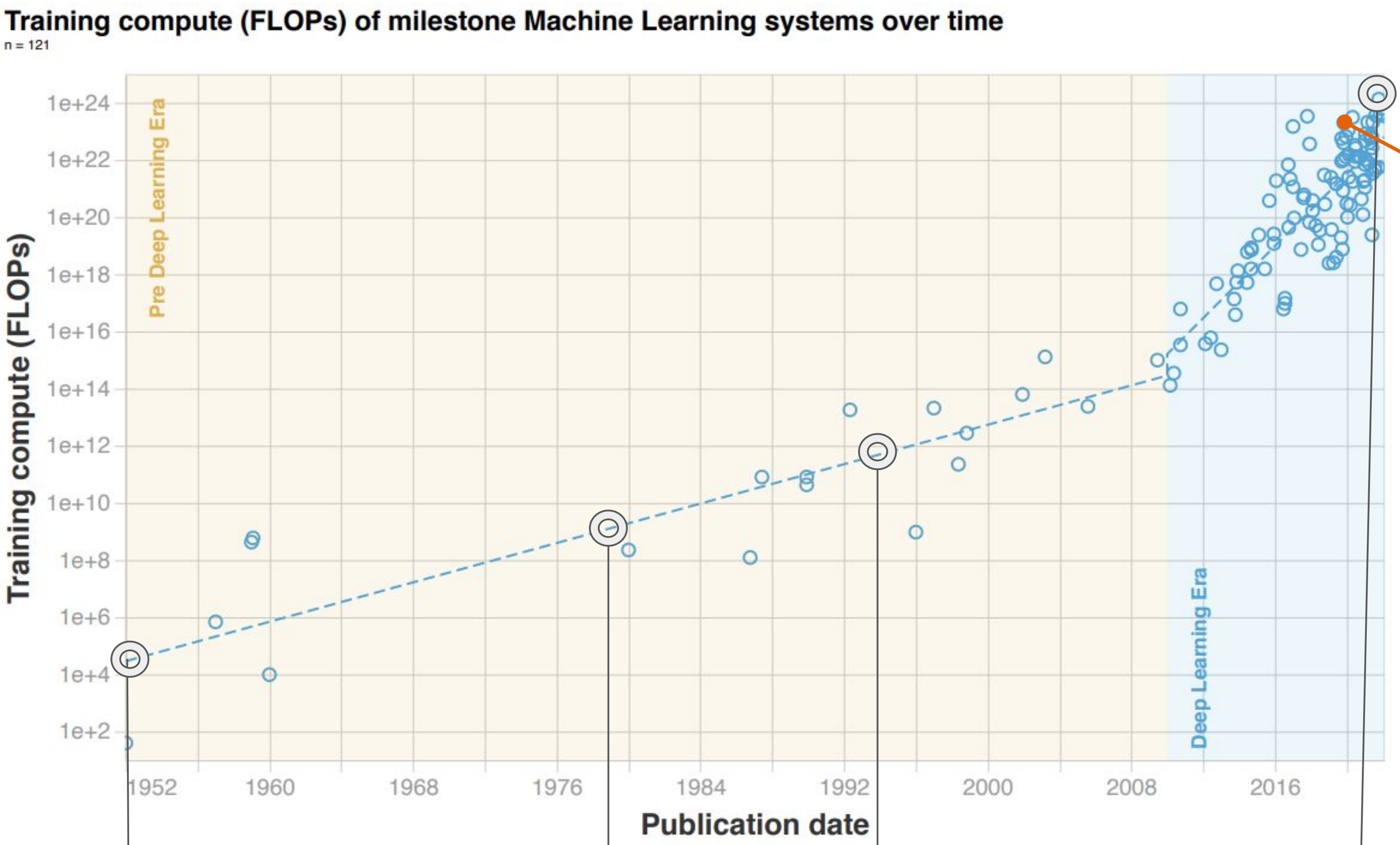
Predicted probability



The model is trained to accomplish this task on lots of text.

To use it to generate text, you can repeatedly
send in the context → calculate next-word probabilities → sample one from the distribution.

Scale has driven the acceleration of AI developments



GPT-3 2020
175B parameters
1024 x A100 GPUs
$3M

Estimating the entropy of English

Speech recognition

Machine translation

"Language Understanding"

State of the art in 2000: Multiple separate models

Library of Congress Has Books for Everyone

(WASHINGTON, D.C., 1964) - It was 150 years ago this year that our nation's biggest library burned to the ground. Copies of all the written books of the time were kept in the Library of Congress. But they were destroyed by fire in 1814 during a war with the British.

That fire didn't stop book lovers. The next year, they began to rebuild the library. By giving it 6,457 of his books, Thomas Jefferson helped get it started.

The first libraries in the United States could be used by members only. But the Library of Congress was built for all the people. From the start, it was our national library.

Today, the Library of Congress is one of the largest libraries in the world. People can find a copy of just about every book and magazine printed.

Libraries have been with us since people first learned to write. One of the oldest to be found dates back to about 800 years B.C. The books were written on tablets made from clay. The people who took care of the books were called "men of the written tablets."

1. Who gave books to the new library?
2. What is the name of our national library?
3. When did this library burn down?
4. Where can this library be found?
5. Why were some early people called "men of the written tablets"?

Figure 1: Sample Remedia™ Reading Comprehension Story and Questions

- Part of speech tagging
- Statistical Tagging
- Named Entity Extractor
- Question Type Detection
- Paraphrase Resolution
- Answer Extraction

Thomas Jefferson

(purple is system output)

GPT-3 2020

"Large Language Model"

Library of Congress Has Books for Everyone

(WASHINGTON, D.C., 1964) - It was 150 years ago this year that our nation's biggest library burned to the ground. Copies of all the written books of the time were kept in the Library of Congress. But they were destroyed by fire in 1814 during a war with the British.

That fire didn't stop book lovers. The next year, they began to rebuild the library. By giving it 6,457 of his books, Thomas Jefferson helped get it started.

The first libraries in the United States could be used by members only. But the Library of Congress was built for all the people. From the start, it was our national library.

Today, the Library of Congress is one of the largest libraries in the world. People can find a copy of just about every book and magazine printed.

Libraries have been with us since people first learned to write. One of the oldest to be found dates back to about 800 years B.C. The books were written on tablets made from clay. The people who took care of the books were called "men of the written tablets."

1. Who gave books to the new library?

MASK

2. The library was built for _______.
3. One of the oldest libraries was found _______.
4. The people who took care of the books were called _______.

Answers: 1. Thomas Jefferson 2. all the people 3. 800 years B.C. 4. "men of the written tablets."

LLMs are…

01 **General**: A single model can handle many tasks without additional training

02 **Broad**: It has as much (or more?) “knowledge” of the world as a human adult

03 **Accessible:** You interact with the model directly in language, not in code

Some tasks where LLMs are effective:

- Summarization
- Information Extraction: Reading Comprehension, Question Answering
- Document Understanding: Sentiment Analysis, Insight Extraction
- Text Generation: poetry writing, style transfer
- **Code Generation:** Copilot (Auto-complete), Toolformer

Code Generation is significant, because it means we can bridge the gap between language and executable actions

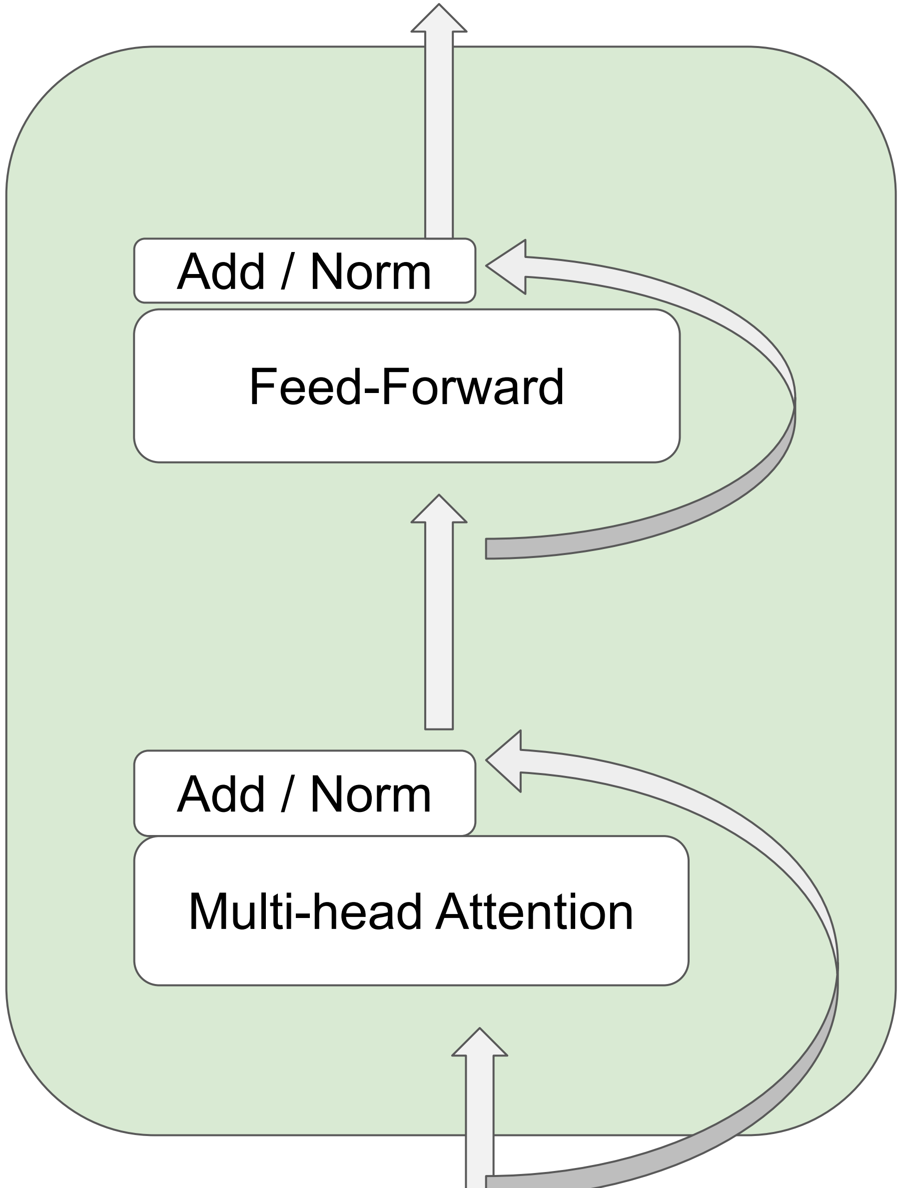
02 Transformer Language Models

TechAtBloomberg.com

Bloomberg

Transformer Blocks

Two main components: an *attention* layer and a *feed-forward* layer



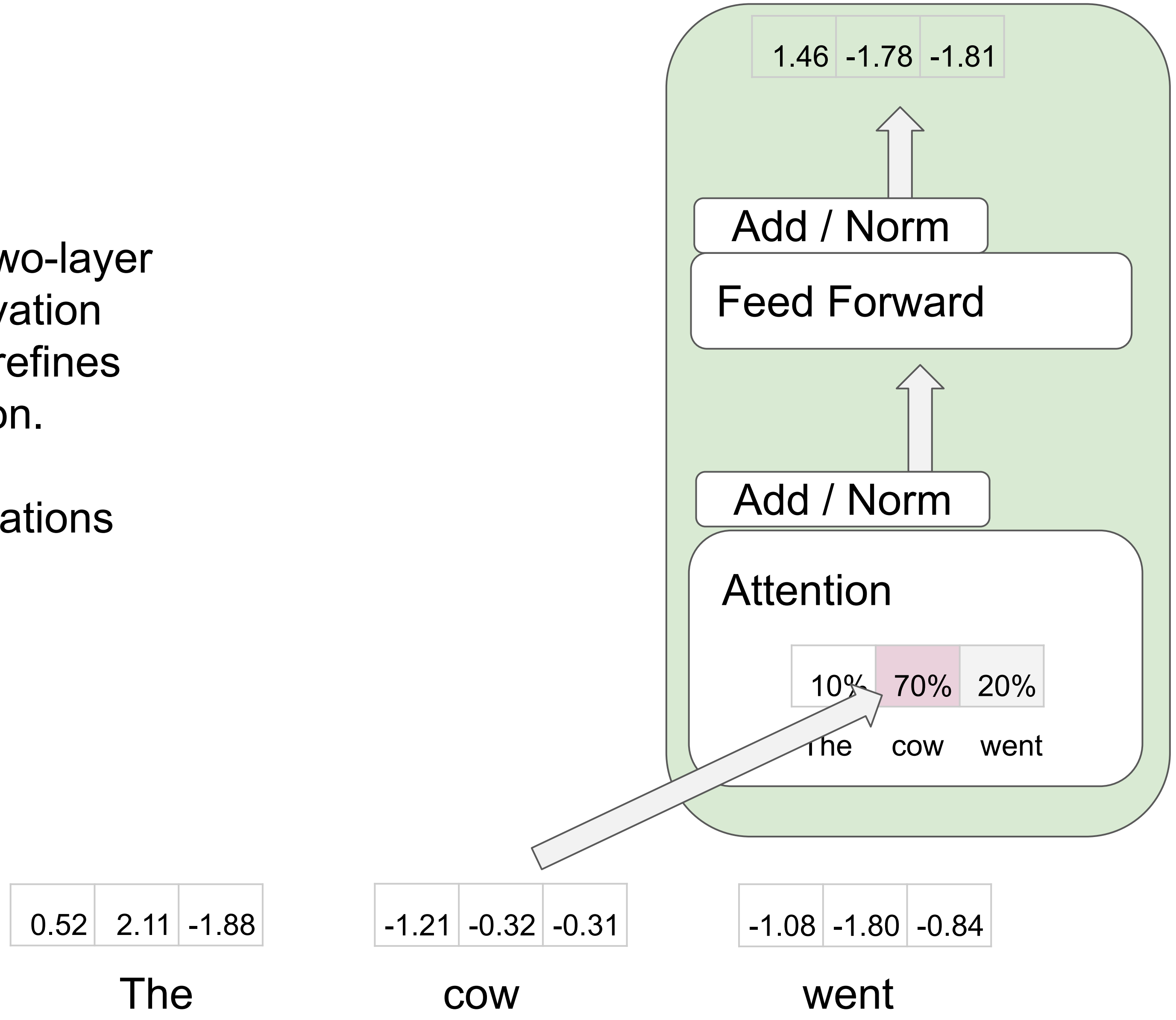
The input and output dimension of each Transformer block are the same.

After each of its two major subcomponents, the inputs (residuals) are added back.

Due to this, every Transformer block has consistent embedding dimensions.

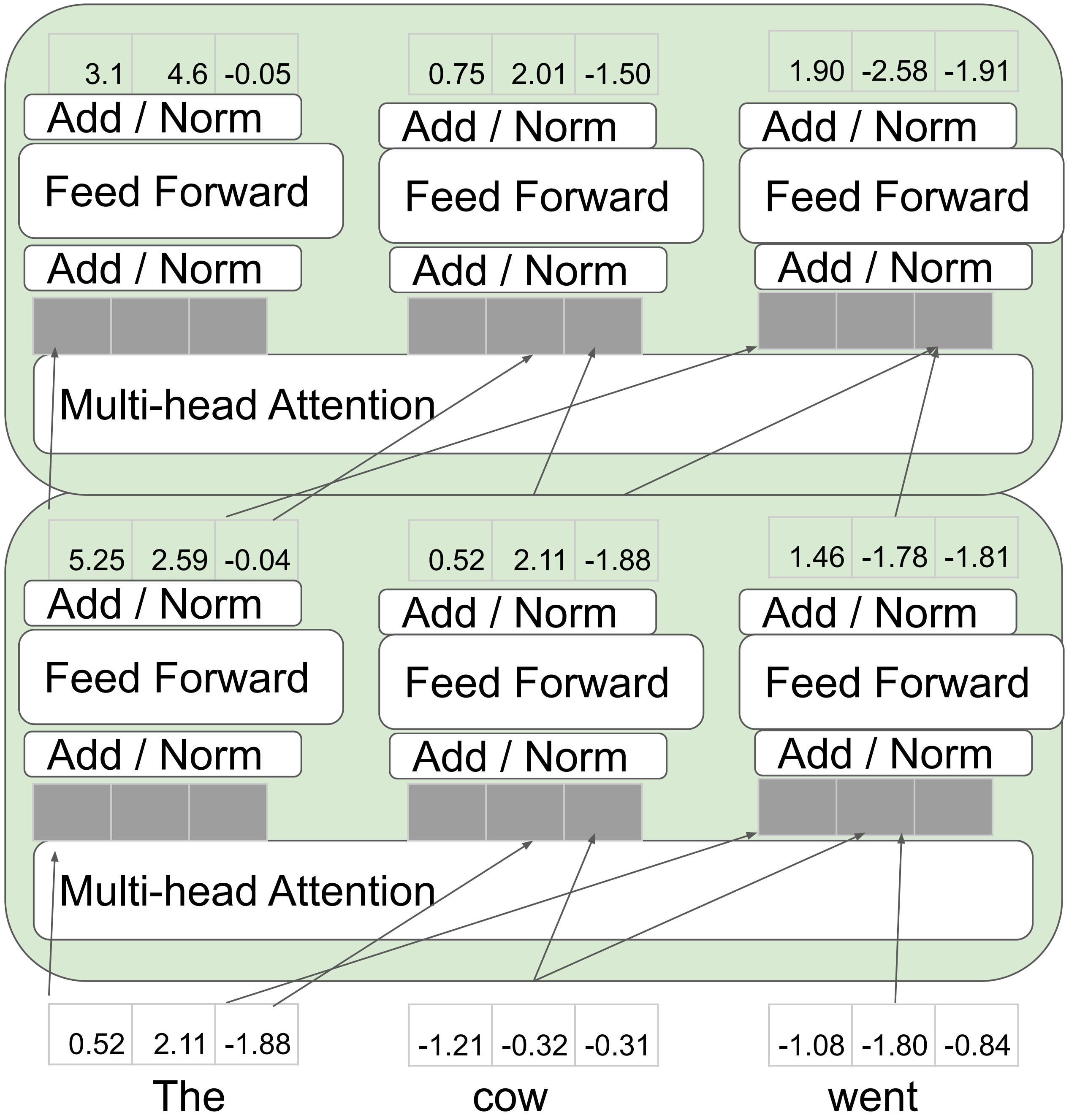
The Feed-Forward layer is a standard two-layer dense perceptron with a non-linear activation between the two layers (e.g., ReLU). It refines the representation for the current position.

The Attention layer integrates representations across previous positions.

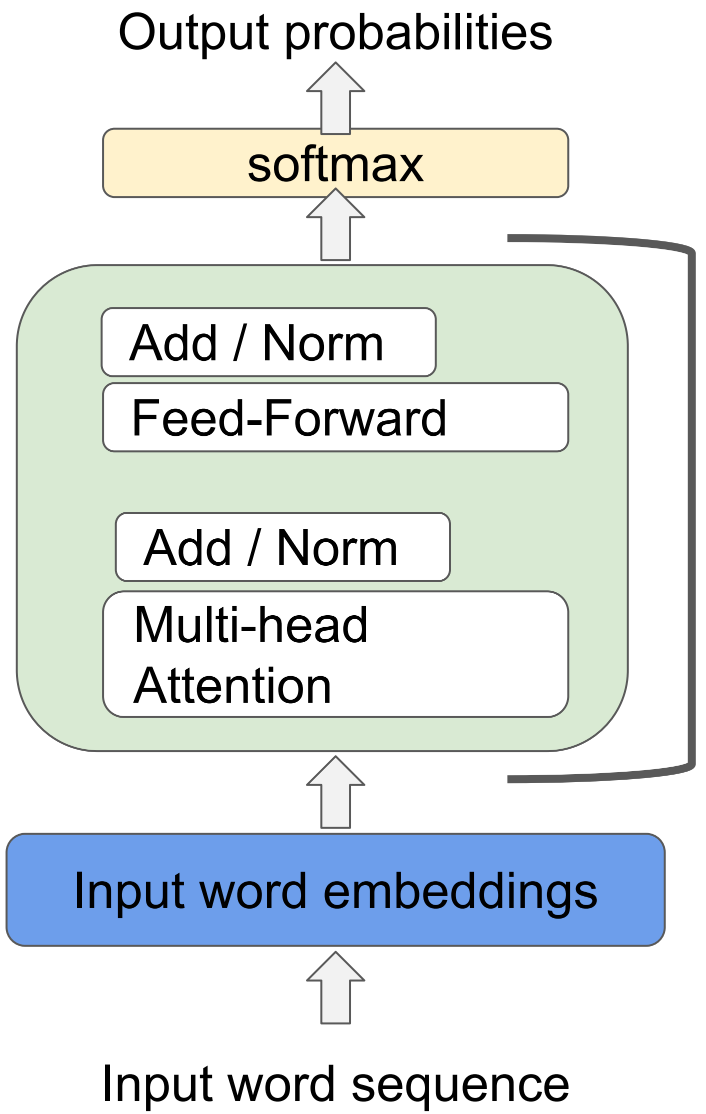


At each word, the Transformer computes a separate representation.

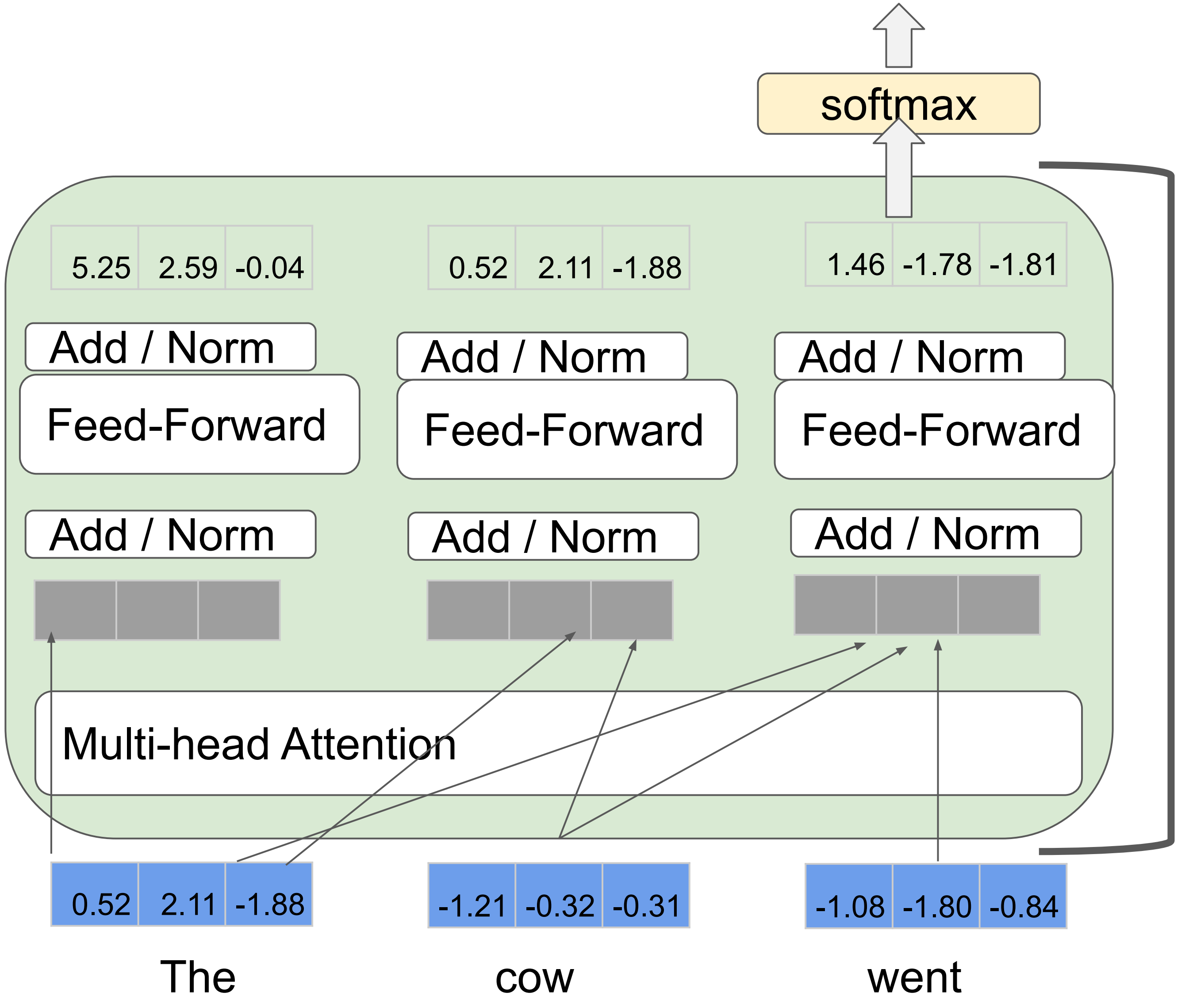
The activations proceed layer by layer and get progressively refined by looking at representations of the other words.



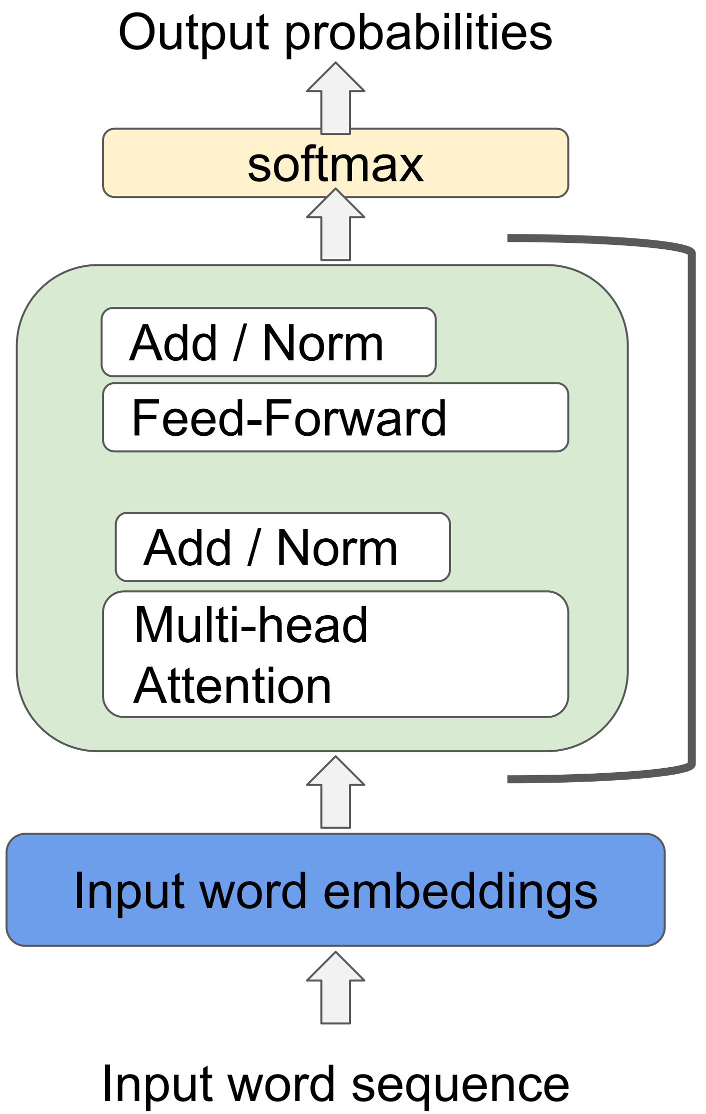
Decoder-Only Transformer Language Model



apple	banana	barn	field	grow	mall	zebra
0.02	0.01	0.57	0.29	0.00	0.10	0.01



Decoder-Only Transformer Language Model



There is a lot of experimentation within this base architecture:

1. Location of Layer Norms (e.g., an extra layer norm after the embedding layer)
2. Exact Activation Function (e.g., SwiGLU vs. ReLU)
3. Positional embeddings (e.g., ALiBi vs. rotational encoding)

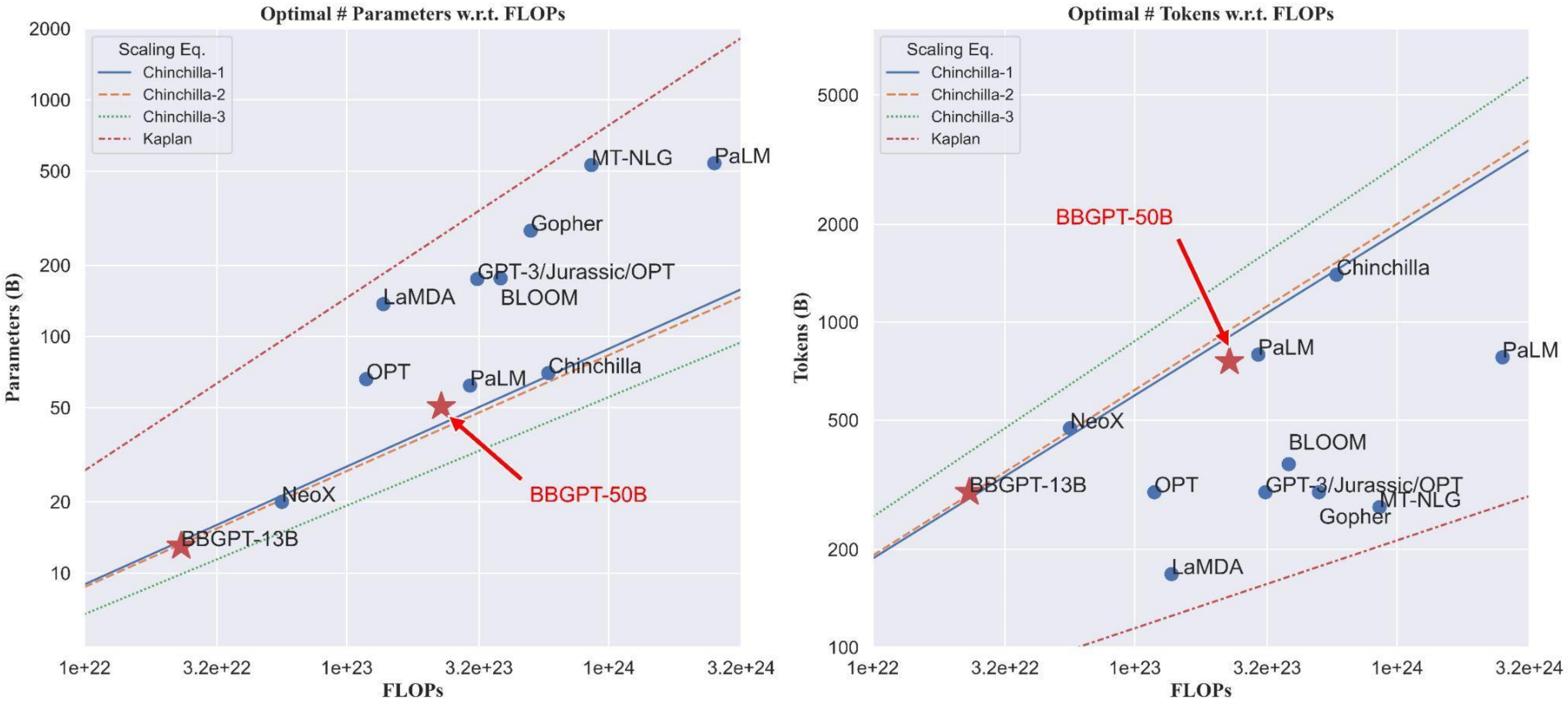
Even with the overall model architecture set, there are lots of knobs to tune

- **Model Size & Shape**: What shape should each sub-component have?
- **Dataset size**: For a given model size, what is the optimal amount of data?
- **Low-level operator Implementation**: How do you build efficient computation for each operator?
- **Hardware:** What is the most efficient hardware (and network) configuration needed for training?
- **Distributed Optimizer**: If the model cannot fit on one machine, how do you efficiently pass parameters and gradients around during training?
- **Hyper-parameters**: How should you set learning rate?
- **Numerical Precision**: What is the necessary precision for each of the parameters and gradients?
- **Tokenization**: How big should the vocabulary size be? Do you want multi-word tokens?
- **Dataset composition**: Does in-domain data matter?

03 Training BloombergGPT

TechAtBloomberg.com

Bloomberg



The **Chinchilla** scaling law proposes a function for the optimal **model** and **data** size **given a set computational budget**.

More recently, **LLaMA** suggested much longer training times to account for **inference-efficient models** (multiple times the Chinchilla estimates).

Model Parameters*

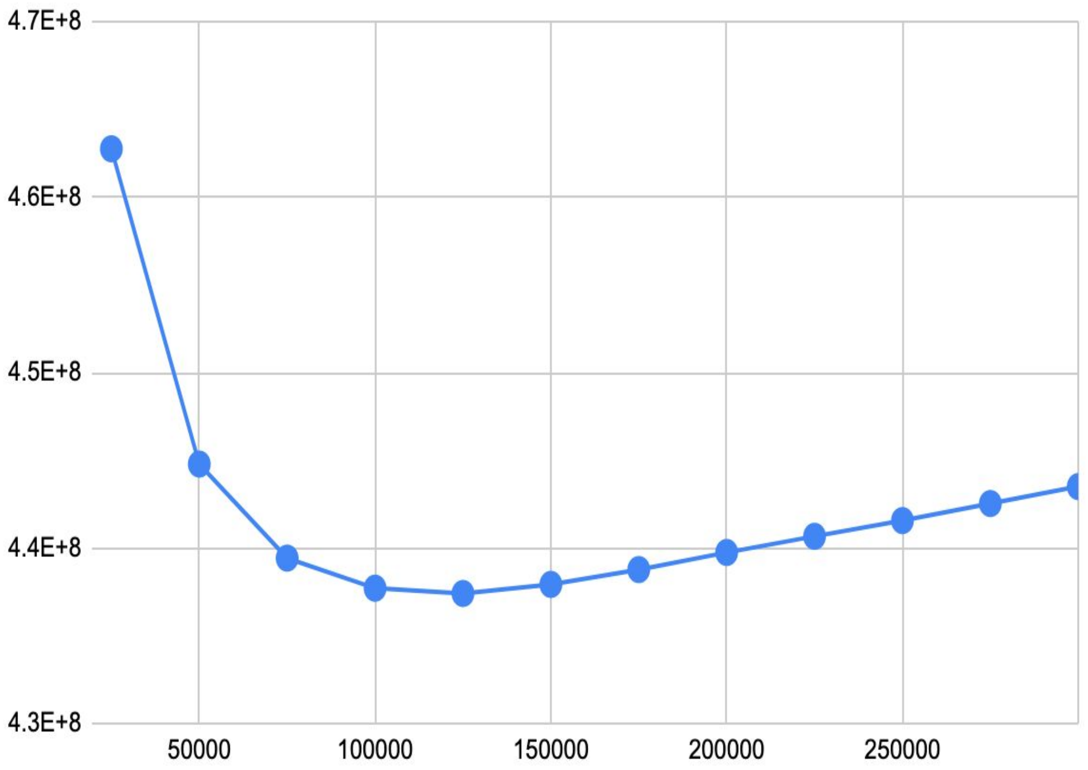
	Embedding Dimensions	Transformer Blocks	Total Parameters*	Total Data Size (in tokens)
GPT-3	12,288	96	175B	300B
GPT-NeoX	6,144	44	20B	472B
OPT-66	9,216	64	66B	300B
BLOOM	14,336	70	176B	366B
BloombergGPT	7,680	70	50B	569B

https://arxiv.org/abs/2006.12467

Tokenization

- GPT-2 Tokenizer is most often used
  - Numbers are not treated specially
  - No multi-word tokens

- BloombergGPT Tokenizer
  - A Unigram Tokenizer trained on The Pile
  - Byte-level representation of numbers and out-of-vocabulary words
  - Multi-word tokens allowed
  - Large vocabulary size - 131,072



Data compression rate with different vocabulary sizes.

Hardware

We used 64 of the largest available instance type, p4d.24xlarge, in our training cluster.

Each instance has the following characteristics:

- 8 NVIDIA 40GB A100 GPU
- Intra-node connection: 600 GB/s Bidirectional NVIDIA NVSwitch
- Inter-node connection: 400 GB/s NVIDIA GPUDirect using AWS Elastic Fabric Adapter (EFA)
- 512 GPUs: 64 instances x 8 A100 (40G) GPUs

Maximum possible performance for these GPUs is 350 TFLOPS. We observed 105 TFLOPS.

Dataset selection

Our training dataset consists of roughly equal parts public and private data:

- Public data:
  - The Pile (Gao et al., 2020) – 22 diverse domains
  - C4 (Raffel et al., 2019) – cleaner Common Crawl
  - Wikipedia from July 2022
- Private data:
  - Web content
  - News wires and transcripts
  - SEC Edgar filings
  - Press Releases

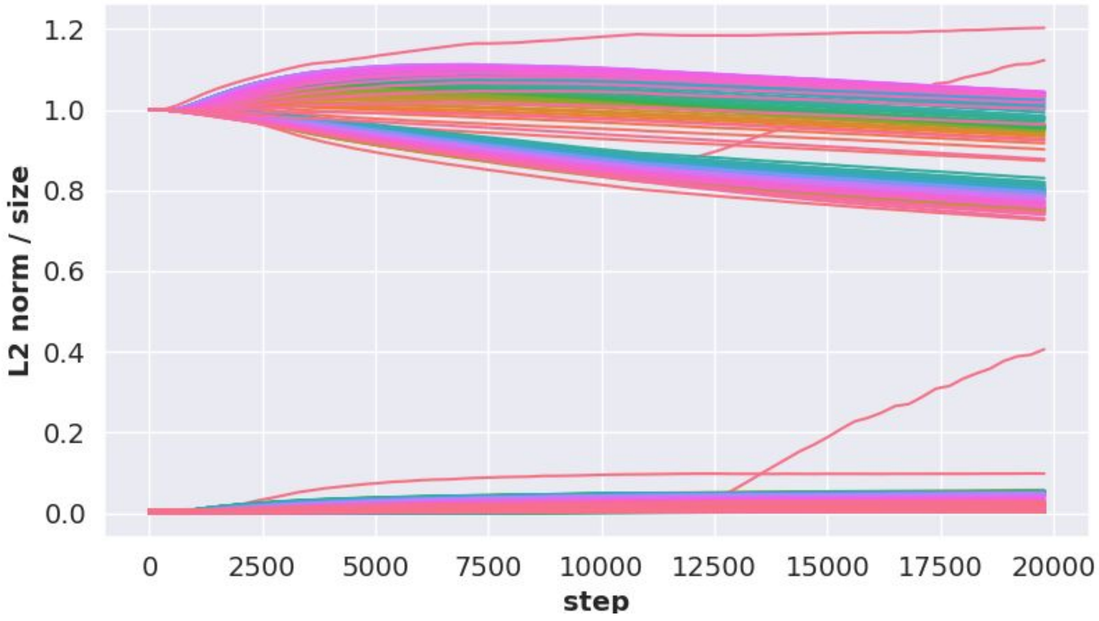
	Source	Docs	Chrs	Toks	T %
C4	AI2	348,320,475	768,255,491,381	138,086,167,809	18.31%
Pile-CC	The Pile	52,546,936	231,235,621,898	42,672,626,661	5.66%
Github		14,283,906	76,617,726,302	22,659,758,045	3.00%
Books3		192,699	106,446,541,865	21,438,473,989	2.84%
PubMed Central		2,943,136	94,714,503,236	21,005,999,316	2.78%
ArXiv		1,236,620	59,133,375,329	16,625,600,597	2.20%
OpenWebText2		16,840,752	64,832,640,876	12,795,986,466	1.70%
FreeLaw		3,490,740	53,692,541,631	10,758,608,214	1.43%
StackExchange		15,382,809	33,864,315,793	8,124,663,078	1.08%
DM Mathematics		999,241	8,187,061,287	4,261,055,494	0.56%
Wikipedia (en)		5,895,152	17,615,739,611	3,788,325,415	0.50%
USPTO Backgrounds		5,172,924	22,445,179,469	3,633,864,012	0.48%
PubMed Abstracts		15,272,654	20,351,195,170	3,529,383,787	0.47%
OpenSubtitles		381,947	11,861,292,733	2,421,421,968	0.32%
Gutenberg (PG-19)		28,017	11,188,611,518	2,285,880,943	0.30%
Ubuntu IRC		10,440	5,629,479,865	1,779,202,702	0.24%
EuroParl		68,441	4,452,319,881	1,520,272,604	0.20%
YoutubeSubtitles		168,673	3,344,937,386	1,315,803,721	0.17%
BookCorpus2		17,588	6,514,318,278	1,216,136,584	0.16%
HackerNews		818,301	4,099,188,167	841,584,504	0.11%
PhilPapers		31,003	2,319,853,994	551,132,837	0.07%
NIH ExPorter		921,126	1,994,651,615	299,916,729	0.04%
Enron Emails		240,321	452,269,259	115,903,976	0.02%
Wikipedia (7/1/22)	Wikimedia	22,181,027	72,564,685,645	23,714,414,820	3.14%
PUBLIC SUM		507,444,928	1,681,813,542,189	345,442,184,271	45.80%

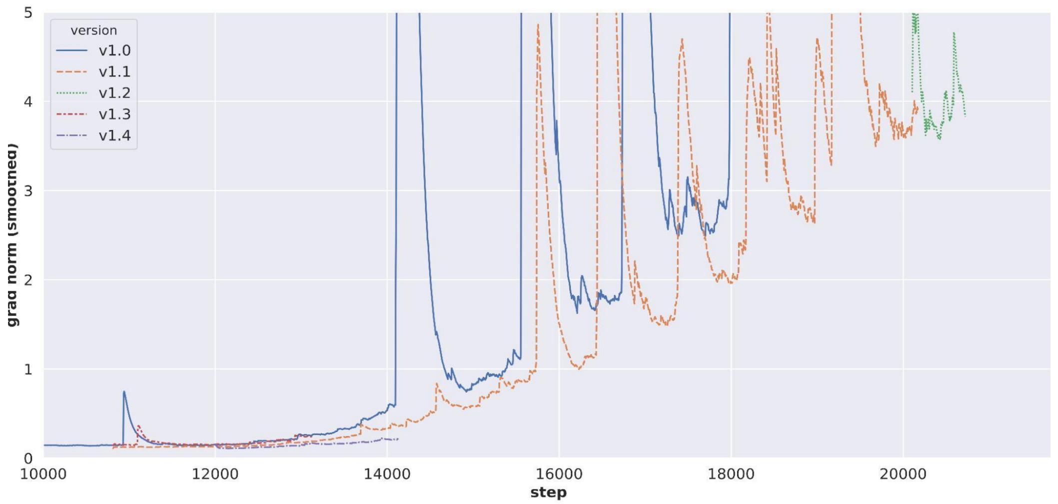
	Docs	Chrs	Toks	T %
2007 / [03-]	16,436,840	29,892,828,920	6,399,980,987	0.85%
2008 / [-]	36,084,362	76,731,180,098	15,740,625,084	2.09%
2009 / [-]	35,456,377	79,738,265,930	16,394,536,545	2.17%
2010 / [-]	42,183,775	92,346,179,177	18,675,431,798	2.48%
2011 / [-]	50,778,508	112,292,794,690	22,832,892,156	3.03%
2012 / [-]	53,982,853	118,192,151,898	24,196,699,764	3.21%
2013 / [-]	69,497,470	120,898,447,464	24,739,450,562	3.28%
2014 / [-]	92,318,762	127,684,784,155	26,395,827,703	3.50%
2015 / [-]	126,268,586	147,068,982,286	30,267,511,982	4.01%
2016 / [-]	157,293,515	157,404,864,986	32,096,254,602	4.26%
2017 / [-]	179,982,740	159,286,323,161	32,699,310,772	4.33%
2018 / [-]	174,885,371	165,130,500,372	33,893,888,805	4.49%
2019 / [-]	174,738,844	162,727,834,110	33,354,932,194	4.42%
2020 / [-]	213,360,286	182,760,019,862	36,868,616,780	4.89%
2021 / [-]	231,603,253	176,668,487,134	36,229,023,280	4.80%
2022 / [-06]	103,095,963	87,897,291,247	18,085,989,414	2.40%
PRIVATE SUM	1,757,967,505	1,996,720,935,490	408,870,972,428	54.20%

BloombergGPT 50b (v0 + v1)

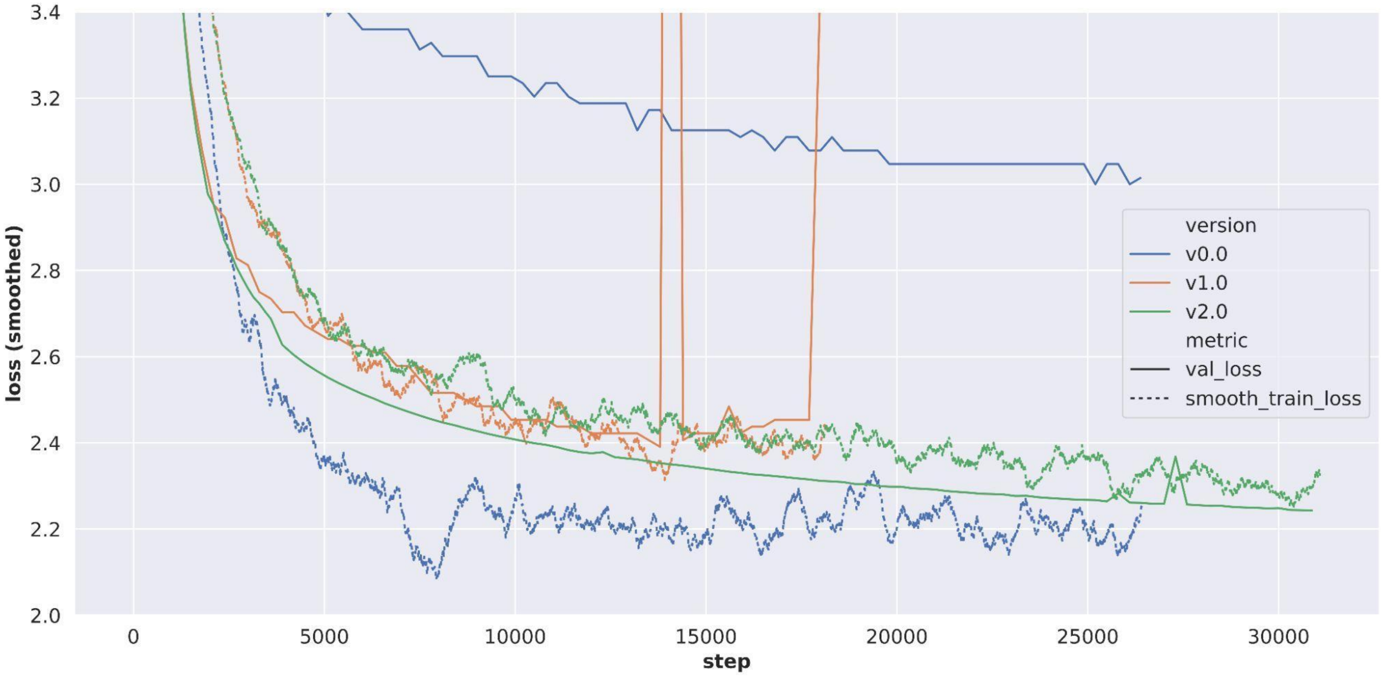
- **V0 Problem**:
  - Training loss plateaued after about 10k steps
  - No improvements to dev loss after step ~20k
- **Changes**:
  - Remove temporal curriculum (shuffle parts)
- **V1 Problem**:
  - After ~12k steps, gradient norm started increasing
  - Occasional dev loss jumps
- **Attempts to fix v1**:
  - v1.1 Restart from step 10.8k, fully shuffle remaining data, reduce LR (8e-5)
  - v1.2: Even smaller LR (6e-5), smaller grad_clip (0.3) without restart
  - v1.3: v1.1 + fp32 in LM head - reduce LR
  - v1.4: v1.3 + v1.2 + smaller roll back

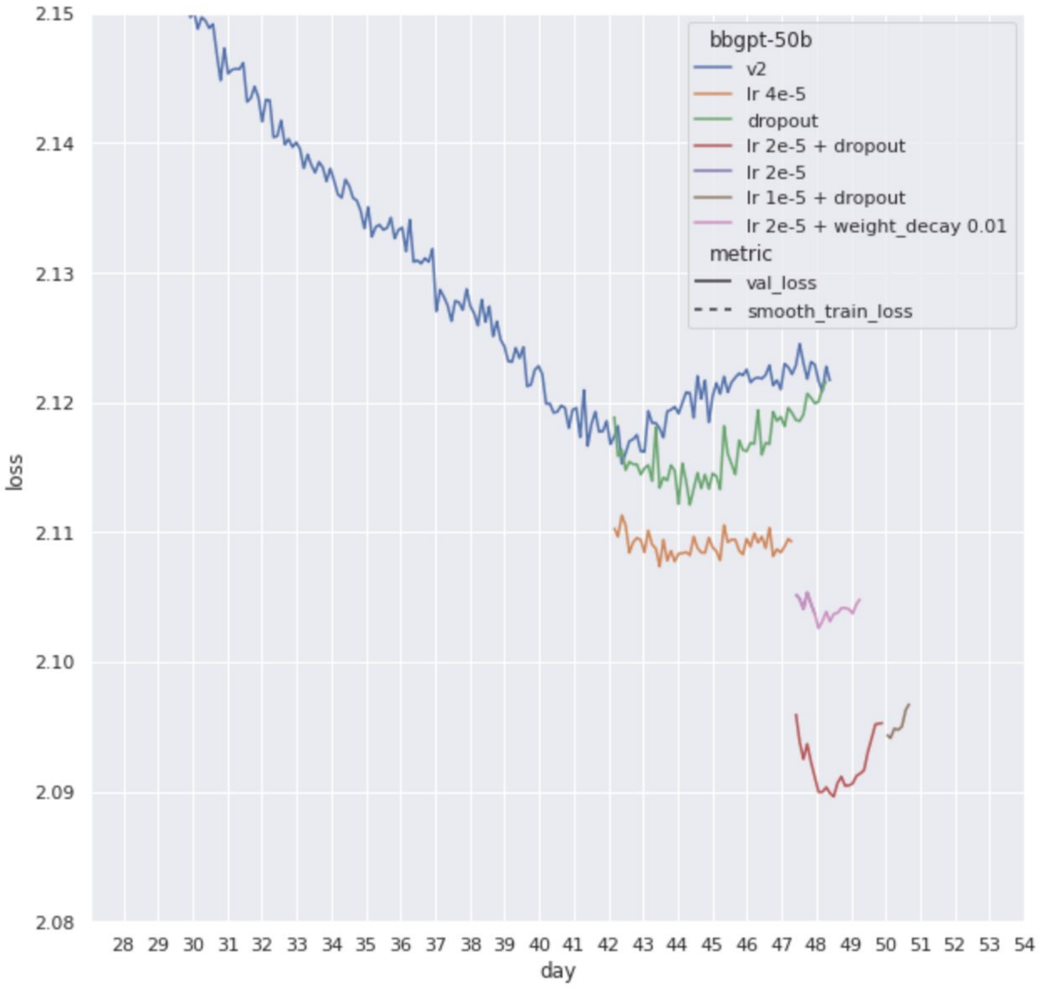
At this point we believed v1 was unsalvageable…





BloombergGPT 50b (v2)





Many stability-inducing hyperparameter changes and bug fixes:

- **v1.4:** fp32 in LM head + max lr = 6e-5 + grad clip = 0.3 + fully shuffle data
- different seed (e.g. different initialization & data ordering)
- LayerNorm after embedding
- Longer learning rate warm up
- Remove Layer Norm weight decay
- Use megatron init rescaling
- query_key_layer_scaling
- two stage batch size warm up (1024->2048)

04 Evaluation

Two Categories of Evaluation

General Topic Evaluation

- *BIG-Bench Hard*
- *Massive Multitask Language Understanding (MMLU)*
- Reading Comprehension
- Linguistic Scenarios

Financial Domain Evaluation

- *FPB, FiQA, Headline, ConvFinQA*
- *Sentiment, NER, NED+NED*

The Different Types of Evaluations

Training monitoring

This is expensive to run!
- Dev loss on (1) i.i.d. Set, (2) non i.i.d. Set
- small set of high-coverage sets

Post-Training Eval

You want something indicative!
- **General Public**: Big bench, Knowledge, Reading Comprehension, Linguistic Tasks
- **In-Domain Public:** QA, NER, sentiment;
- **In-Domain Private**: Sentiment, NER, NED

Eval for particular applications

This is what really matters and post-training eval should correlate with positive results here.
Work with product management, develop UX/UI, Teach people how to use the model

BIG-Bench Hard

BIG-Bench covers 214 NLP tasks to benchmark LLMs

BIG-Bench Hard is a curated subset of 23 challenging tasks

BIG-bench Hard Task	BloombergGPT	GPT-NeoX	OPT_{66B}	$BLOOM_{176B}$	$PaLM_{540B}$
Boolean Expressions$^{\lambda}$	62.40	**71.20**	48.40	69.20	**83.2**
Causal Judgement	49.73	**52.41**	51.87	51.87	**61.0**
Date Understanding	**54.80**	45.60	49.60	50.00	53.6
Disambiguation QA	34.00	**40.80**	40.40	40.40	**60.8**
Dyck Languages$^{\lambda}$	15.60	26.00	14.80	**42.00**	28.4
Formal Fallacies	50.80	52.80	**54.00**	52.80	53.6
Geometric Shapes$^{\lambda}$	15.20	8.00	11.60	**22.40**	**37.6**
Hyperbaton	**92.00**	**92.00**	91.60	**92.00**	70.8
Logical Deduction$^{\lambda}$ *(avg)*	**34.53**	30.93	31.87	34.00	**60.4**
Movie Recommendation	90.40	86.40	**91.20**	**91.20**	87.2
Multi-Step Arithmetic$^{\lambda}$ [Two]	**1.20**	0.40	0.40	0.00	**1.6**
Navigate$^{\lambda}$	42.00	45.20	42.00	**50.00**	**62.4**
Object Counting$^{\lambda}$	33.20	21.20	26.00	**36.80**	**51.2**
Penguins in a Table	37.67	33.56	28.08	**40.41**	**44.5**
Reasoning about Colored Objects	34.80	26.00	31.20	**36.80**	**38.0**
Ruin Names	**56.00**	54.00	52.80	54.80	**76.0**
Salient Translation Error Detection	20.00	20.40	16.40	**23.60**	**48.8**
Snarks	69.66	62.36	69.66	**72.47**	**78.1**
Sports Understanding	**62.80**	53.20	54.40	53.20	**80.4**
Temporal Sequences$^{\lambda}$	29.20	21.20	23.60	**36.80**	**39.6**
Tracking Shuffled Objects$^{\lambda}$ *(avg)*	**25.33**	24.53	24.00	23.47	19.6
Web of Lies$^{\lambda}$	49.20	52.40	**54.00**	51.20	51.2
Word Sorting$^{\lambda}$	4.80	5.20	2.40	**7.60**	**32.0**
NLP Task *(avg)*	54.39	51.63	52.60	54.96	**62.7**
Algorithmic Task$^{\lambda}$ *(avg)*	28.42	27.84	25.37	**33.95**	**40.9**
All Tasks *(avg)*	41.97	40.25	39.58	**44.91**	**52.3**
All Tasks *(WR)*	0.57	0.45	0.39	**0.75**	-

Massive Multitask Language Understanding

Abstract Algebra	Groups, rings, fields, vector spaces, ...
Anatomy	Central nervous system, circulatory system, ...
Astronomy	Solar system, galaxies, asteroids, ...
Business Ethics	Corporate responsibility, stakeholders, regulation, ...
Clinical Knowledge	Spot diagnosis, joints, abdominal examination, ...
College Biology	Cellular structure, molecular biology, ecology, ...
College Chemistry	Analytical, organic, inorganic, physical, ...
College Computer Science	Algorithms, systems, graphs, recursion, ...
College Mathematics	Differential equations, real analysis, combinatorics, ...
College Medicine	Introductory biochemistry, sociology, reasoning, ...
College Physics	Electromagnetism, thermodynamics, special relativity, ...
Computer Security	Cryptography, malware, side channels, fuzzing, ...
Conceptual Physics	Newton's laws, rotational motion, gravity, sound, ...
Econometrics	Volatility, long-run relationships, forecasting, ...
Electrical Engineering	Circuits, power systems, electrical drives, ...
Elementary Mathematics	Word problems, multiplication, remainders, rounding, ...
Formal Logic	Propositions, predicate logic, first-order logic, ...
Global Facts	Extreme poverty, literacy rates, life expectancy, ...
High School Biology	Natural selection, heredity, cell cycle, Krebs cycle, ...
High School Chemistry	Chemical reactions, ions, acids and bases, ...
High School Computer Science	Arrays, conditionals, iteration, inheritance, ...
High School European History	Renaissance, reformation, industrialization, ...
High School Geography	Population migration, rural land-use, urban processes, ...
High School Gov't and Politics	Branches of government, civil liberties, political ideologies,
High School Macroeconomics	Economic indicators, national income, international trade, ...
High School Mathematics	Pre-algebra, algebra, trigonometry, calculus, ...
High School Microeconomics	Supply and demand, imperfect competition, market failure, .
High School Physics	Kinematics, energy, torque, fluid pressure, ...
High School Psychology	Behavior, personality, emotions, learning, ...
High School Statistics	Random variables, sampling distributions, chi-square tests, ...
High School US History	Civil War, the Great Depression, The Great Society, ...
High School World History	Ottoman empire, economic imperialism, World War I, ...
Human Aging	Senescence, dementia, longevity, personality changes, ...
Human Sexuality	Pregnancy, sexual differentiation, sexual orientation, ...
International Law	Human rights, sovereignty, law of the sea, use of force, ...
Jurisprudence	Natural law, classical legal positivism, legal realism, ...
Logical Fallacies	No true Scotsman, base rate fallacy, composition fallacy, ...
Machine Learning	SVMs, VC dimension, deep learning architectures, ...
Management	Organizing, communication, organizational structure, ...
Marketing	Segmentation, pricing, market research, ...
Medical Genetics	Genes and cancer, common chromosome disorders, ...
Miscellaneous	Agriculture, Fermi estimation, pop culture, ...
Moral Disputes	Freedom of speech, addiction, the death penalty, ...
Moral Scenarios	Detecting physical violence, stealing, externalities, ...
Nutrition	Metabolism, water-soluble vitamins, diabetes, ...
Philosophy	Skepticism, phronesis, skepticism, Singer's Drowning Child, ...
Prehistory	Neanderthals, Mesoamerica, extinction, stone tools, ...
Professional Accounting	Auditing, reporting, regulation, valuation, ...
Professional Law	Torts, criminal law, contracts, property, evidence, ...
Professional Medicine	Diagnosis, pharmacotherapy, disease prevention, ...
Professional Psychology	Diagnosis, biology and behavior, lifespan development, ...
Public Relations	Media theory, crisis management, intelligence gathering, ...
Security Studies	Environmental security, terrorism, weapons of mass destruction, ...
Sociology	Socialization, cities and community, inequality and wealth, ...
US Foreign Policy	Soft power, Cold War foreign policy, isolationism, ...
Virology	Epidemiology, coronaviruses, retroviruses, herpesviruses, ...
World Religions	Judaism, Christianity, Islam, Buddhism, Jainism, ...

MMLU Results

Model	BloombergGPT	GPT-NeoX	OPT_{66B}	$BLOOM_{176B}$	GPT-3
Humanities	**36.26**	32.75	33.28	34.05	**40.8**
STEM	35.12	33.43	30.72	**36.75**	36.7
Social Sciences	40.04	36.63	38.32	**41.50**	**50.4**
Other	46.36	42.29	42.63	**46.48**	**48.8**
Average	**39.18**	35.95	35.99	39.13	**43.9**

Financial Domain: Public Datasets

	BloombergGPT	GPT-NeoX	OPT_{66B}	$BLOOM_{176B}$
ConvFinQA	**43.41**	30.06	27.88	36.31
FiQA SA	**75.07**	50.59	51.60	53.12
FPB	**51.07**	44.64	48.67	50.25
Headline	**82.20**	73.22	79.41	76.51
NER	60.82	**60.98**	57.49	55.56
All Tasks *(avg)*	**62.51**	51.90	53.01	54.35
All Tasks *(WR)*	**0.93**	0.27	0.33	0.47

Financial Domain: Bloomberg Internal Datasets

	BloombergGPT	GPT-NeoX	OPT_{66B}	$BLOOM_{176B}$
Equity News	**79.63**	14.17	20.98	19.96
Equity Social Media	**72.40**	66.48	71.36	68.04
Equity Transcript	**65.06**	25.08	37.58	34.82
ES News	**46.12**	26.99	31.44	28.07
Country News	**49.14**	13.45	17.41	16.06
All Tasks *(avg)*	**62.47**	29.23	35.76	33.39
All Tasks *(WR)*	**1.00**	0.00	0.67	0.33

	BloombergGPT	GPT-NeoX	OPT_{66B}	$BLOOM_{176B}$
NER				
BFW	72.04	71.66	72.53	**76.87**
BN	57.31	52.83	46.87	**59.61**
Filings	58.84	59.26	59.01	**64.88**
Headlines	**53.61**	47.70	46.21	52.17
Premium	60.49	59.39	57.56	**61.61**
Transcripts	75.50	70.62	72.53	**77.80**
Social Media	60.60	56.80	51.93	**60.88**
All Tasks *(avg)*	62.63	59.75	58.09	**64.83**
All Tasks *(WR)*	0.57	0.29	0.19	**0.95**
NER+NED				
BFW	**55.29**	34.92	36.73	39.36
BN	**60.09**	44.71	54.60	49.85
Filings	**66.67**	31.70	65.63	42.93
Headlines	**67.17**	36.46	56.46	42.93
Premium	**64.11**	40.84	57.06	42.11
Transcripts	**73.15**	23.65	70.44	34.87
Social Media	67.34	62.57	**70.57**	65.94
All Tasks *(avg)*	**64.83**	39.26	58.79	45.43
All Tasks *(WR)*	**0.95**	0.00	0.67	0.38

05 Discussion

Text-to-BQL

Input: Get me the last price and market cap for Apple
Output: `get(px_last,cur_mkt_cap) for(['AAPL US Equity'])`

Input: Tesla price
Output: `get(px_last) for(['TSLA US Equity'])`

Input: Get the yield and spread for EC527035 Corp and AL580550 Corp
Output: `get(yield,spread) for(['EC527035 Corp','AL580550 Corp'])`

Input: apple and ibm market cap and eps
Output: `get(cur_mkt_cap,is_eps) for(['AAPL US Equity','IBM US Equity'])`

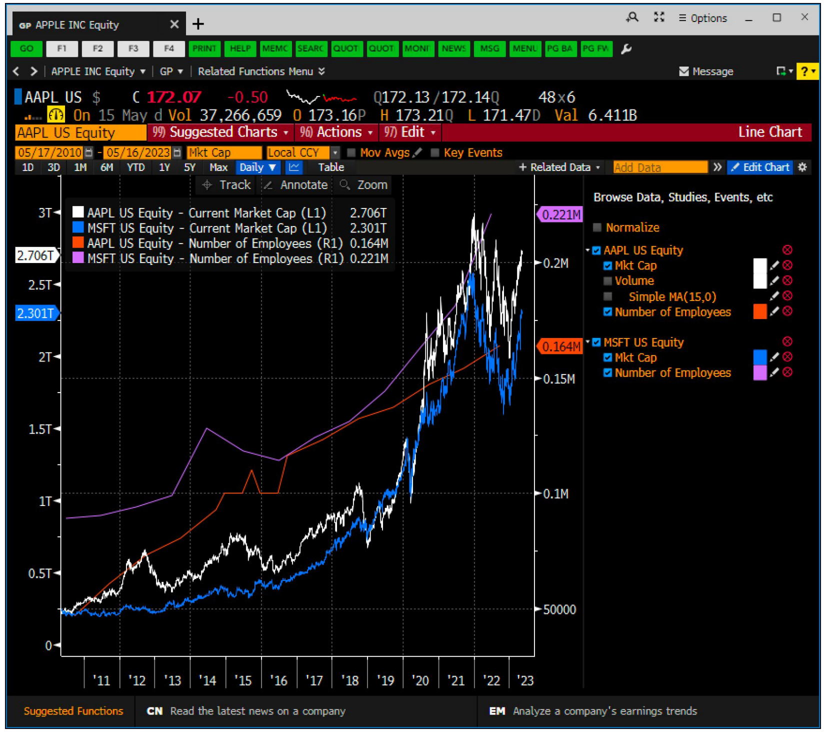
Input: industry subgroup of ibm apple microsoft google
Output: `get(industry_subgroup()) for(['AAPL US Equity','IBM US Equity', 'MSFT US Equity','GOOGL US Equity'])`

Language Interface for the Terminal

“Market Cap of AAPL vs. MSFT”

“Start from 2010”

“Add Number of Employees”



Key Takeaways

01 Even though model stability and efficiency are challenging, it is possible for a (relatively) small team to build a LLM around the same quality as GPT-3 (given enough compute resources)

02 Training on in-domain data even at dataset sizes of hundreds of billions of tokens, yields stronger in-domain performance while retaining general performance

Open Questions

01 Sample efficiency of training

02 General Pre-Training, then Fine-Tuning vs. Joint Pre-Training

03 Post-training Conditioning

Thank you!

www.TechAtBloomberg.com/datascience

We are hiring: bloomberg.com/engineering

Bloomberg

TechAtBloomberg.com