Insights from Bloomberg-GPT, and other Financial LLMs

This document entails at insights from Bloomberg-GPT regarding dataset, tokenizer, training, and also contains some other relevant work

Contacts: @Hursh Prasad, @Alex Jiao, @Vyom Pathak

Last Updated: 12/20/23

Bloomberg-GPT

https://arxiv.org/pdf/2303.17564.pdf

▼ Dataset Details

- Also introduces 363 billion token dataset based on Bloomberg's extensive data sources, perhaps the largest domain-specific dataset yet, augmented with 345 billion tokens from general purpose datasets totaling to 717 billion token dataset.
- Shows training of LLM on domain specific as well as general purpose dataset. The introduced dataset is curated, and prepared from reliable datasources.

Dataset	Docs 1e4	C/D	Chars 1e8	C/T	Toks 1e8	T%
FINPILE	*175,886*	*1,017*	*17,883*	*4.92*	*3,635*	*51.27%*
Web	158,250	933	14,768	4.96	2,978	42.01%
News	10,040	1,665	1,672	4.44	376	5.31%
Filings	3,335	2,340	780	5.39	145	2.04%
Press	1,265	3,443	435	5.06	86	1.21%
Bloomberg	2,996	758	227	4.60	49	0.70%
PUBLIC	*50,744*	*3,314*	*16,818*	*4.87*	*3,454*	*48.73%*
C4	34,832	2,206	7,683	5.56	1,381	19.48%
Pile-CC	5,255	4,401	2,312	5.42	427	6.02%
GitHub	1,428	5,364	766	3.38	227	3.20%
Books3	19	552,398	1,064	4.97	214	3.02%
PubMed Central	294	32,181	947	4.51	210	2.96%
ArXiv	124	47,819	591	3.56	166	2.35%
OpenWebText2	1,684	3,850	648	5.07	128	1.80%
FreeLaw	349	15,381	537	4.99	108	1.52%
StackExchange	1,538	2,201	339	4.17	81	1.15%
DM Mathematics	100	8,193	82	1.92	43	0.60%
Wikipedia (en)	590	2,988	176	4.65	38	0.53%
USPTO Backgrounds	517	4,339	224	6.18	36	0.51%
PubMed Abstracts	1,527	1,333	204	5.77	35	0.50%
OpenSubtitles	38	31,055	119	4.90	24	0.34%
Gutenberg (PG-19)	3	399,351	112	4.89	23	0.32%
Ubuntu IRC	1	539,222	56	3.16	18	0.25%
EuroParl	7	65,053	45	2.93	15	0.21%
YouTubeSubtitles	17	19,831	33	2.54	13	0.19%
BookCorpus2	2	370,384	65	5.36	12	0.17%
HackerNews	82	5,009	41	4.87	8	0.12%
PhilPapers	3	74,827	23	4.21	6	0.08%
NIH ExPorter	92	2,165	20	6.65	3	0.04%
Enron Emails	24	1,882	5	3.90	1	0.02%
Wikipedia (7/1/22)	2,218	3,271	726	3.06	237	3.35%
TOTAL	*226,631*	*1,531*	*34,701*	*4.89*	*7,089*	*100.00%*

Table 1: Breakdown of the full training set used to train BLOOMBERGGPT. The statistics provided are the average number of characters per document ("C/D"), the average number of characters per token ("C/T"), and the percentage of the overall tokens ("T%"). Units for each column are denoted in the header.

De-duped The Pile, C4, Wikipedia, and FinPile (Internal dataset) based on Lee et al. . Emphasis on cleaning dataset by removing markups as well.

- FinPile - Some articles are public and non-trivial to collect, but rest is internal data. data processing - strip off markup, special formatting, and templates, 2007-03-01 to 2022-07-31 ranged documents.
    - WEB - Web crawled financial specific dataset form APAC, US, UK
    - NEWS - All news sources including Bloomberg news related to financial domain

    > 💡 FILINGS - 10-K annual reports, and 10-Q quarterly reports from EDGAR, SEC's online database. Filings are typically long PDF documents with tables and charts that are dense in financial information, which are processed and normalized in Bloomberg. Filings are substantially different from the types of documents typically used to train LLMs, but contain critically important information for financial decision-making.
    >
    > - https://www.sec.gov/edgar/search-and-access
    > - https://www.sec.gov/edgar/searchedgar/companysearch
    > - License details: https://www.sec.gov/edgar/search/efts-faq.html

    - PRESS - Press release by companies about financial details similar to news
    - BLOOMBERG - Bloomberg news, first word, and other opnions, and analyses (vague so ignore)
- Public Datasets
    - The Pile - Has been used for LLM training, is pre-processed, clean, and helps generalize the model. The OG Pile is contains duplicate data proportional to the perceived quality of content, but they de-duped it. Their tokenizer was trained on Pile dataset.

    > 💡 C4 - Common dataset to train LLM, used to train T5. Overlaps with Pile-CC, and is cleaned differently. C4 is is a high quality natural language document dataset with most of the dataset stemming from patents.

    - Wikipedia - Wikipedia dataset, but with inefficient tokenization and above-average amount of markup which can be cleaned further

▼ Tokenizer Details

💡 Using Unigram model instead of sub-word tokenizers as it saves smarter tokenization at inference time based on promising results in (Sentence Piece) Kudo and Richardson (2018) and (BPE) Bostrom and Durrett (2020). Treated data as a sequence of bytes rather than unicode including 256 bytes as tokens. Pretokenization step, the input byte sequence is broken into chunks by greedily matching the following regular expression: [A-Za-z]+|[0-9]|[^A-Za-z0-9]+.

They include spaces in the alphabetic chunks, which allows multi-word tokens to be learned, increasing information density and reducing context lengths. The number grouping is based on PaLM.

💡 Chunking based training on Pile dataset. We then train a Unigram tokenizer with a vocabulary size of 65,536 (216) on each of the 22 × 256 (total = 5, 632) chunks.

To reduce the size of the vocabulary to 217 tokens, we drop the tokens with the smallest probabilities and renormalize. To ensure we do not need an out-of vocabulary token, we also add as tokens the 36 (of 256 possible) bytes that do not occur in The Pile, along with an <|endoftext|> token.

▼ Model Details

- Its a 50 billion param BLOOM-style LM trained on financial data. Model size was determined with (Chinchilla Scaling Law) Hoffmann et al., (GPT-3 [Scaling Laws]) Brown et al., (BLOOM) Le Scao et al. and (LLaMa) Touvron et al. with optimal model size to dataset size. Training at scale was possible because of BLOOM
- Decoder-only causal LM. 70 layers of transformer

💡 ALiBi position embedding at self-attention based on BLOOM

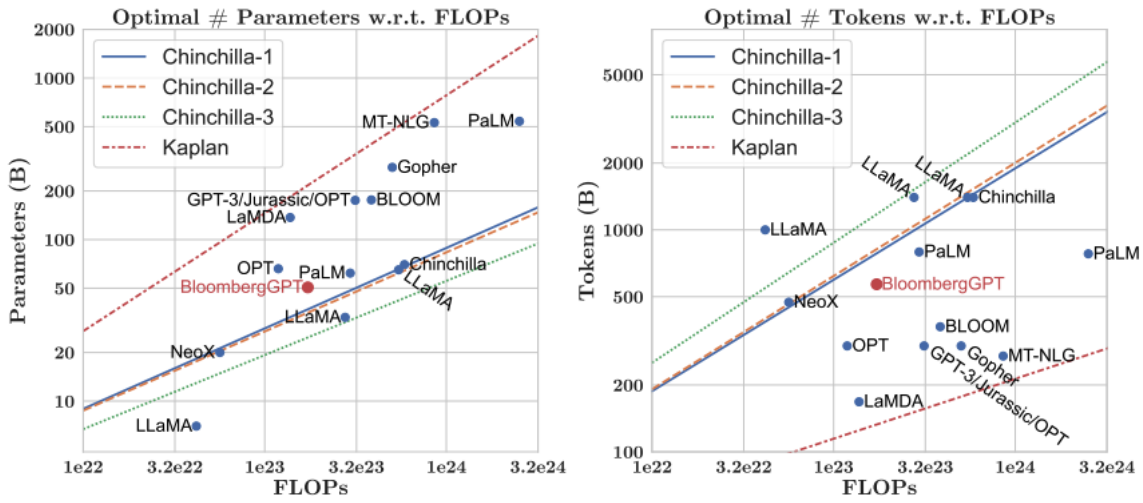

Figure 1: Kaplan et al. (2020) and Chinchilla scaling laws with prior large language model and BLOOMBERGGPT parameter and data sizes. We adopt the style from Hoffmann et al. (2022).

💡 Making the model Chinchilla optimal and not scaling laws optimal. We should consider chinchilla law as well because the original scaling has been proved to be incorrect

- Total compute budget of 1.3M GPU hours on 40GB A100 GPUs

💡 Adopt activation checkpointing to reduce our memory footprint,

this costs us an additional 0.33x TFLOPs per iteration due to repeated forward passes. To account for this additional cost, we plug in 0.75 × 1.3M into Chinchilla
equations instead of the full amount.

💡 Model Shape is determined based on (depth-to-width calculator) Levine et al.

- D = exp(5.039) exp(0.0555 · L)

Shape	
Number of Layers	70
Number of Heads	40
Vocabulary Size	131,072
Hidden Dimension	7,680
Total Parameters	50.6B
Hyperparameters	
Max Learning Rate	6e-5
Final Learning Rate	6e-6
Learning Rate schedule	cosine decay
Gradient Clipping	0.3
Training	
Tokens	569B
Hardware	64 × 8 A100 40GB
Throughput	32.5 sec/step
avg. TFLOPs	102
total FLOPS	2.36e23

Table 4: A summary of the hyper-parameters and their values for BLOOMBERGGPT.

- 40 heads, each having a dimension of 192, resulting in a total hidden dimension of D = 7680 and a total of 50.6B parameters

💡 These ideas suggest that, it is important to calculate FLOPs based on our model size, dataset size, and the compute limit will be calc based on all of these values thereafter.

$$Params_{opt} = .6(FLOPs)^{.45}$$
$$Data_{opt} = .3(FLOPs)^{.55}$$

▼ Training Details

- Pytorch based, 2048 tokens size by chunking for all the documents.
- For training only 569 billion tokens from our corpus of over 700 billion tokens was used
- AdamW, cosine decay learning rate, 6e-6 lr, 1024 → 2048 batchsize

- Megatron-LLM sstyled scaling, query key layer scaling, FP16 mixed-precision training but may also help in BF16
- Training instability with 20 times loss spikes as also seen in PaLM. Mitigate this issue by restarting the training from a checkpoint roughly 100 steps before spike, and then skipping 200-500 data batches. Other option is lowering the lr for spikes.
- AWS SageMaker - 64 p4d.24xlarge - 8 - 40 GB A100 GPUs NVSwitch intra-node connections (600 GB/s) and NVIDIA GPUDirect using AWS Elastic Fabric Adapter (EFA) inter-node connections (400 Gb/s) - 512 40GB A100 GPUs. For quick data access, we use Amazon FSX for Lustre, which supports up to 1000 MB/s read and write throughput per TiB storage unit.
- Large-scale optimization for memory footprint by using stage 3 of ZeRO optimization. We utilize the proprietary SageMaker Model Parallelism (SMP) library from AWS, which enables the automatic distribution of large models across multiple GPU devices and instances. We achieve 102 TFLOPs on average and each training step takes 32.5 seconds.
- Also use MiCS for decreasing overhead for communications with cloud clusters.
- Activation Checkpointing minimizes training memory consumption by removing activations at the expense of additional computation during backward passes.
- Forward and Backward pass in BF16, params are stored in full precision FP32. ALiBi computed in full precision, an stored in BF16. Use FP32 to calculate fused softmax in attention, and store in BF16. softmax calc in loss is done in FP32.
- Similar to Megatron-LM, we use a masked-causal-softmax fused kernel in SMP in the self-attention module

These methods are used for optimization. According to our architecture size, and dataset size, and compute limit, we can try out these methods for optimization

▼ Eval Details

- Shows performance on LLM tasks, open financial tasks, internal benchmarks.

Suite	Tasks	What does it measure?
Public Financial Tasks	5	Public datasets in the financial domain
Bloomberg Financial Tasks	12	NER and sentiment analysis tasks
Big-bench Hard (Suzgun et al., 2022)	23	Reasoning and general NLP tasks
Knowledge Assessments	5	Testing closed-book information recall
Reading Comprehension	5	Testing open-book tasks
Linguistic Tasks	9	Not directly user-facing NLP tasks

Table 5: Evaluation Benchmarks. We evaluate BLOOMBERGGPT on a high-coverage set of standard benchmarks that assess downstream performance, taken from HELM, SuperGLUE, MMLU, and the GPT-3 suite. Since these have significant overlap and/or include each other, we restructure them into the categories presented here. We only evaluate on one setup per dataset. We further assess BLOOMBERGGPT on a suite of internal and public financial tasks.

Name	# Tokens (B)	# Params. (B)	Compute
BLOOMBERGGPT	569	50.6	1.00×
GPT-NeoX	472	20	0.33×
OPT	300	66	0.69×
BLOOM	366	176	2.24×
GPT-3	300	175	1.82×

Table 6: Evaluation model cohort. OPT and BLOOM each have multiple sizes available and we report those we evaluated. We note that compute numbers are only partially comparable between models: For example, BLOOMs training data is only 1/3 English, and OPT repeated some of its training data. We report GPT-3 results whenever available but did not run it ourselves due to lack of availability.

- Few-shot methodology for context examples

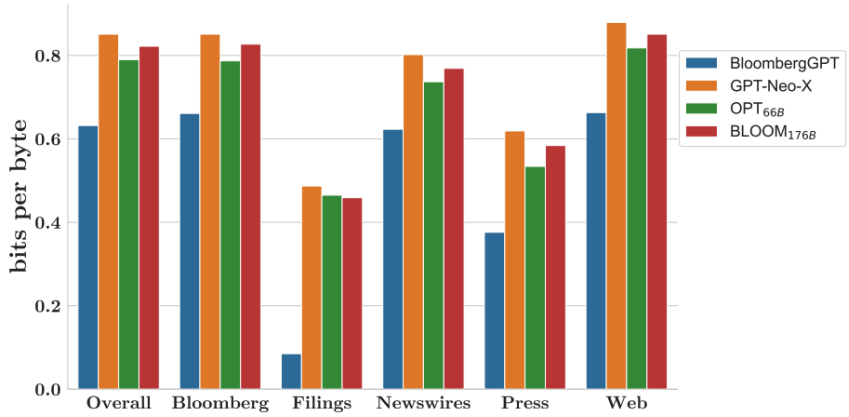

Figure 3: Bits per byte on a heldout test set of each data type in our FINPILE (lower is better). The set of documents is held out in time and deduplicated with the training set, such that all of it is completely unseen by BLOOMBERGGPT. Regardless, we observe a large gap between the models. The improvement is largest for specialized in-domain documents like Filings.

▼ Financial tasks - internal and public

Task	Template/Example
Discriminative	

Sentiment Analysis	`{sentence}`
	`Question: what is the sentiment?`
	`Answer: {negative/neutral/positive}`

Aspect Sentiment Analysis	`{sentence}`
	`Question: what is the sentiment on {target}?`
	`Answer: {negative/neutral/positive}`

Binary Classification	`{sentence}`
	`Question: {question}?`
	`Answer: {Yes/No}`

Generative	
NER	Steve Jobs is the CEO of Apple Extract named entity: Steve Jobs (person), Apple (organization)
NER+NED	AAPL stopped using Intel Chips Extract ticker: AAPL, INTC

QA	`{context}`
	`Question: {question}?`
	`Answer: {answer}`

Table 7: Template for the different tasks we evaluate in the financial domain.

- 5 Public dataset

- FPB - The Financial Phrasebank Dataset includes a sentiment classification task on sentences from financial news.
- FiQA SA - The second sentiment analysis task is to predict the aspect-specific sentiment in English financial news and microblog headlines, which were published as a part of the 2018 challenge on financial question answering and opinion mining.
- Headline - This is a binary classification task of whether a news headline in the gold commodity domain includes certain information

💡 NER - This is a named entity recognition task on financial data gathered for credit risk assessment from financial agreements filed with the SEC.

💡 ConvFinQA - Given input from S&P 500 earnings reports that includes text and at least one table with financial data, the task is to answer conversational questions that require numerical reasoning over the input.

	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

Table 8: Results on financial domain tasks.

- 5 internal datasets

Name	Time	Tokens	Test Size	% Pos	% Neu	% Neg
Equity News	2018–2019	150-200	1,000	7	87	6
Equity Social Media	2015–2020	15-20	1,000	10	83	7
Equity Transcript	2008–2020	70-80	800	19	75	6
ES News	2016–2019	100-120	1,000	32	53	15
Country News	2009–2021	50-1,000	1,000	18	60	22

Table 9: An overview of the Bloomberg-internal sentiment analysis tasks. Input token and label distribution numbers are computed on the test set.

	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

Table 10: Results on internal aspect-specific sentiment analysis datasets. BLOOMBERGGPT far outperforms all other models on sentiment analysis tasks.

▼ Common LLM Tasks

- NER - None of the previous LLM model study NER, It is not part of HELM, only one task (Polish) in BIG-bench. So, 7 internal NER tasks.

	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

Table 12: Results on internal NER and NED datasets. On NER, while the much larger $BLOOM_{176B}$ model outperforms all other models, results from all models are relatively close, with BloombergGPT outperforming the other two models. On NER+NED, BloombergGPT outperforms all other models by a large margin.

- Knowledge Assessment - BIG-Bench Hard - Standard purpose NLP tasks, ARC, CommonsenseQA, PiQA, MMLU

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**	-

Table 13: BIG-bench hard results using standard 3-shot prompting. Following the convention from Suzgun et al. (2022), we denote algorithmic tasks with the superscript $^{\lambda}$, and present averages for NLP and algorithmic categories. The baseline numbers from $PaLM_{540B}$ (Chowdhery et al., 2022) are taken from the original BBH paper.

Task	BloombergGPT	GPT-NeoX	OPT_{66B}	$BLOOM_{176B}$	GPT-3
ARC (easy)	73.99	70.79	71.25	**75.93**	71.2
ARC (challenging)	48.63	45.39	44.54	**50.85**	**53.2**
CommonsenseQA	65.52	60.36	**66.42**	64.21	-
PiQA	**77.86**	75.84	77.58	77.04	**80.5**
All Tasks *(avg)*	66.50	63.10	64.95	**67.01**	-
All Tasks *(WR)*	**0.75**	0.08	0.33	0.67	-

Table 14: Knowledge tasks 1-shot results. The baseline numbers from GPT-3 are taken from Brown et al. (2020). Among all models, BloombergGPT achieves the highest win rate among the models we ran ourselves, and performs second best on average.

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**

Table 15: Results (5-shot) on the MMLU (Hendrycks et al., 2021) benchmark. The baseline numbers from GPT-3 are taken from Hendrycks et al. (2021). While BloombergGPT lacks behind $BLOOM_{176B}$ on three of the categories, its average is the highest among all models we evaluated ourselves. The gap to GPT-3 is largest on social sciences while the performance in other categories is close.

- BoolQ, OpenBookQA, RACE, MultiRC, ReCoRD

RC Scenario	BloombergGPT	GPT-NeoX	OPT_{66B}	$BLOOM_{176B}$	GPT-3
BoolQ	**74.59**	46.36	57.46	52.94	**76.7**
OpenBookQA	51.60	44.20	**58.00**	47.20	**58.8**
RACE (middle)	**54.32**	41.23	47.42	52.30	**57.4**
RACE (high)	**41.74**	34.33	37.02	39.14	**45.9**
MultiRC	**62.29**	22.86	18.80	26.65	**72.9**
ReCoRD	**82.79**	67.86	82.53	78.01	**90.2**
All Tasks *(avg)*	**61.22**	42.81	50.21	49.37	**67.0**
All Tasks *(WR)*	**0.94**	0.06	0.50	0.50	-

Table 16: Reading Comprehension Results (1-shot). The baseline numbers from GPT-3 are taken from Brown et al. (2020). BloombergGPT far outclasses the models we evaluated ourselves, and is slightly behind GPT-3.

- RTE, ANLI, CB, COPA, WIC, Winograd, Winogrande, HellaSWAG, StoryCloze

Linguistic Scenario	BLOOMBERGGPT	GPT-NeoX	OPT_{66B}	$BLOOM_{176B}$	GPT-3
RTE	**69.31**	53.79	54.87	57.40	**70.4**
ANLI Round 1	32.90	32.60	33.10	**33.60**	32.0
ANLI Round 2	**34.40**	33.80	34.20	33.80	33.9
ANLI Round 3	**37.33**	36.17	34.92	35.17	35.1
CB	**53.57**	48.21	44.64	48.21	**64.3**
COPA	86.00	**88.00**	86.00	84.00	**87.0**
WIC	**52.51**	50.00	**52.51**	50.16	48.6
WinoGrad	80.95	79.12	**82.78**	78.02	**89.7**
WinoGrande	64.09	60.62	66.14	**67.01**	**73.2**
HellaSWAG	**73.92**	68.37	73.47	73.21	**78.1**
StoryCloze	80.87	78.30	**81.83**	80.28	**84.7**
All Tasks *(avg)*	**60.63**	57.18	58.59	58.26	**63.4**
All Tasks *(WR)*	**0.85**	0.27	0.58	0.42	-

Table 17: Results on the Linguistic Scenarios (1-shot). The baseline numbers from GPT-3 are taken from Brown et al. (2020). Win rates and averages are computed only based on accuracy numbers. BLOOMBERGGPT consistently scores highest among the models we evaluate, achieving an 85% win rate.

  - Quantitative sampling - Bloomberg Query Language, Suggestion of news headlines, financial QA
- Related Work, Openness, Ethics, Acknowledgement, Conclusion
- **Appendix shows training chronicles (imp for training details we can use further down the road next year)**

Other Financial Related Papers (To Read)

Contains instruction tuning dataset, and instruction tuned FIn-LLama

PIXIU: A Comprehensive Benchmark, Instruction Dataset and Large...

Although large language models (LLMs) have shown great performance in natural language processing (NLP) in the financial domain, there are no publicly available financially tailored LLMs...

https://openreview.net/forum?id=vTrRq6vCQH

Dataset with large amount of financial QA - It comprises 10,231 questions about publicly traded companies, with corresponding answers and evidence strings. The questions in FinanceBench are ecologically valid and cover a diverse set of scenarios.

FinanceBench: A New Benchmark for Financial Question Answering

FinanceBench is a first-of-its-kind test suite for evaluating the performance of LLMs on open book financial question answering (QA). It comprises 10,231 questions about publicly traded companies,...

https://arxiv.org/abs/2311.11944

Contains QA dataset with text + tables can be used for pre-training

BizBench: A Quantitative Reasoning Benchmark for Business and Finance

As large language models (LLMs) impact a growing number of complex domains, it is becoming increasingly important to have fair, accurate, and rigorous evaluation benchmarks. Evaluating the...

https://arxiv.org/abs/2311.06602

Fin-T5

BBT-Fin: Comprehensive Construction of Chinese Financial Domain...

To advance Chinese financial natural language processing (NLP), we introduce BBT-FinT5, a new Chinese financial pre-training language model based on the T5 model. To support this effort, we have...

https://arxiv.org/abs/2302.09432