Chronologically Consistent Large Language Models

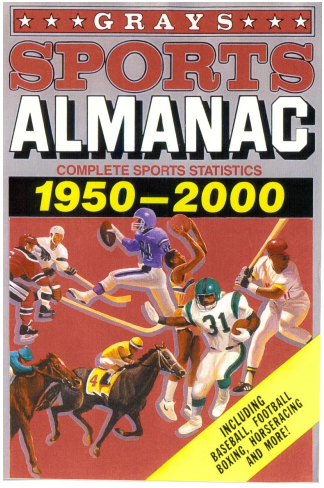
Songrun He[1], Linying Lv[1], Asaf Manela[1] and Jimmy Wu[1]

[1]Washington University in St. Louis

November 2025

Chronological inconsistency, lookahead bias, and training leakage

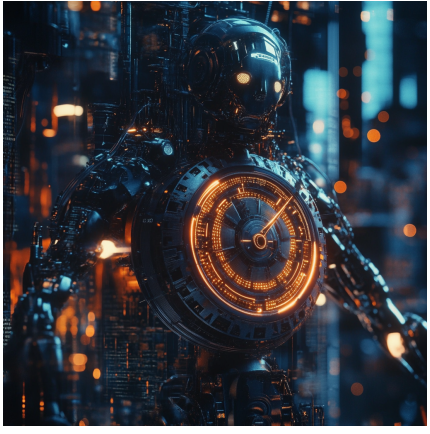
- Large language models (LLMs) now permeate social sciences
- They let us test hypotheses previously unquantifiable
- But they are trained on data that did not exist at the historical moment
- Lookahead bias (Glasserman-Lin 2023, Sarkar-Vafa 2024) and training leakage (Ludwig-Mullainathan-Rambachan 2025) raise doubts about LLM-based empirical findings



Chronological inconsistency in finance

- Finance is particularly sensitive to lookahead bias
- Market efficiency requires prices reflect only facts known at the time
- SOTA gated models (e.g. ChatGPT) continuously fine-tuned and can search
- Chronologically inconsistent models bias measures of risk and market efficiency

What we do



- We train chronologically consistent LLMs exclusively on preceding text
  - $ChronoBERT_{1999}, \ldots, ChronoBERT_{2024}$
  - $ChronoGPT_{1999}, \ldots, ChronoGPT_{2024}$
  - Available to other researchers on Hugging Face
- Simple, right?
- Ensuring these models are competitive with prominent, high-performing counterparts is hard!

Main findings

- ChronoBERT and ChronoGPT exhibit superior language understanding relative to similar-sized models and comparable to much larger Llama models
- In an asset pricing application predicting next-day stock returns from financial news, we find that ChronoBERT and ChronoGPT's Sharpe ratios (4.80 and 4.92) are comparable to powerful larger-scale (and inconsistent) Llama (4.90)
- Implies modest lookahead bias in this setting

Related work

- We develop LLMs free from lookahead bias and capable of high-level language comprehension
  - Does not require masking (Glasserman-Lin 2023; Engelberg et al 2025) which may destroy information
  - Superior language understanding relative to StoriesLM (Sarkar 2024), FinBERT (Huang et al 2023), BERT (Devlin et al 2019) and comparable to Llama 3.1 (Dubey et al 2024)
  - More recent knowledge cutoffs (1999–2024) relative to StoriesLM which ends in 1963
- We find news-return predictability by LLMs is not driven by lookahead bias
  - Large literature shows news text forecasts stock returns (Tetlock et al 2008; Jiang et al 2021; Ke et al 2019)
  - Recent work shows LLMs are much better than early dictionary-based or word-count methods (Lopez-Lira and Tang 2023; Chen et al 2023)
  - But because LLMs are often a black box, concerns about lookahead bias linger (Sarkar-Vafa 2024; Ludwig et al 2025; Levy 2024)

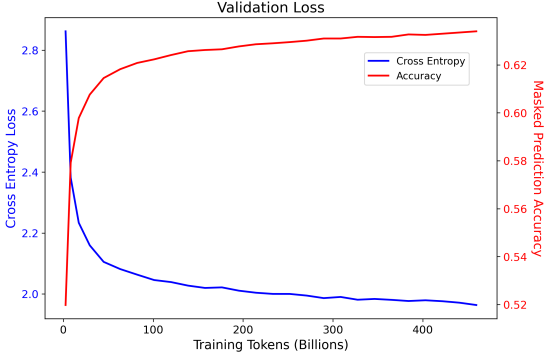
Pretraining chronologically consistent LLMs

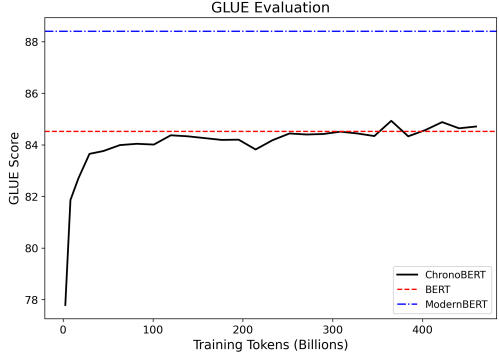
- Training LLMs is usually split into:
  1. Pretraining to predict missing words in text sequences
  2. Finetuning for specific applications (e.g. chat, Q&A, reasoning)
- We pretrain *ChronoBERT*$_t$ and *ChronoGPT*$_t$ only on text available before t
  - For example, web pages crawled in 2005 would be used for pretraining *ChronoBERT*$_{2005}$ using *ChronoBERT*$_{2004}$ as a starting point

Pretraining chronologically consistent LLMs

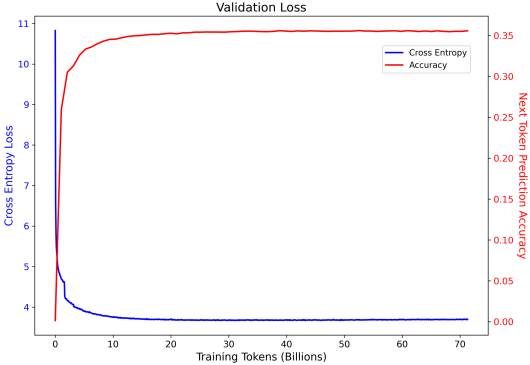
- Ensuring these models are competitive with larger counterparts poses two challenges:
  1. Limited compute
  2. Limited historical data
- We draw on efficient training methods (Portes et al 2023; Warner et al 2024; Jordan et al 2024) to lower computing costs
- Follow Gunasekar et al. (2023) by selecting diverse, high-quality data, carefully filtered by publication date to maximize information gained from a limited corpus
- Follow Muennighoff et al (2023) insights to train over multiple epochs to maximize learning from the available corpus
- Initial 1999 models are trained on 7 billion tokens over multiple epochs
- Incremental training from 2000 to 2024 on a corpus of 65 billion tokens

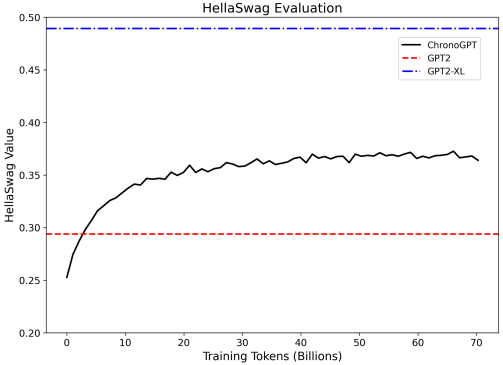
ChronoBERT$_{1999}$ improves to the point it surpasses BERT on GLUE language benchmarks





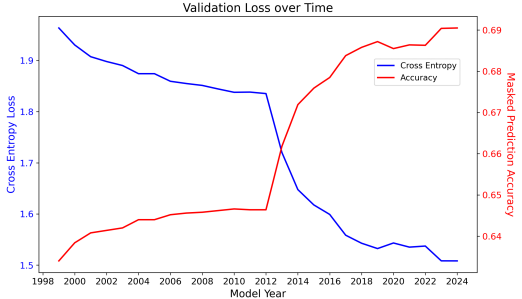
ChronoGPT$_{1999}$ improves to the point it surpass GPT-2 on HellaSwag token generation

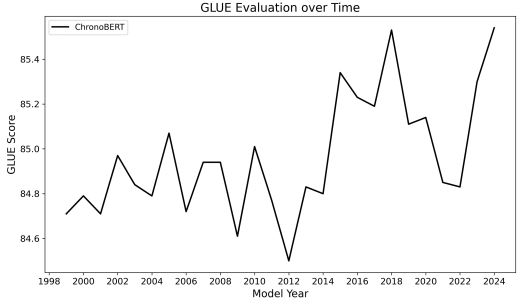




Subsequent ChronoBERT models improve further over time

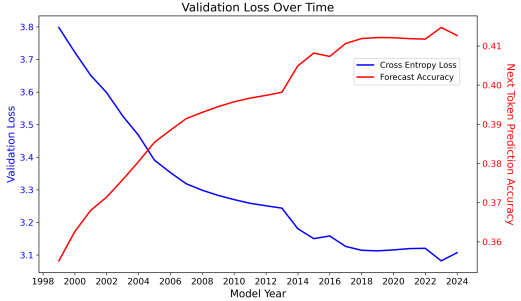
ChronoBERT$_{1999}, \ldots,$ *ChronoBERT*$_{2024}$

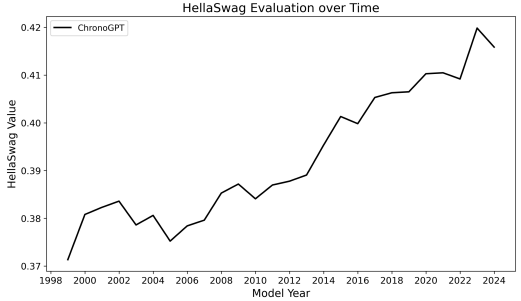




Subsequent ChronoGPT models improve further over time

ChronoGPT$_{1999}, \ldots,$ *ChronoGPT*$_{2024}$





Size, context and knowledge cutoff for different LLMs

	Parameters	Context Tokens	Knowledge Cutoff
ChronoBERT_{1999}	149M	1,024	December, 1999
⋮	⋮	⋮	⋮
ChronoBERT_{2024}	149M	1,024	December, 2024
ChronoGPT_{1999}	1,552M	1,792	December, 1999
⋮	⋮	⋮	⋮
ChronoGPT_{2024}	1,552M	1,792	December, 2024
BERT	110M	512	October, 2018
FinBERT	110M	512	December, 2019
StoriesLM	110M	512	December, 1963
Llama 3.1	8,030M	128,000	December 2023

ChronoBERT and ChronoGPT exhibit superior language understanding relative to similar-sized models and comparable to much larger Llama

GLUE Score Evaluation for Different LLMs

	ChronoBERT_{1999}	ChronoBERT_{2024}	ChronoGPT_{1999}	ChronoGPT_{2024}
COLA	57.32	56.32	35.50	48.20
SST2	91.82	92.58	90.60	92.40
MRPC	92.71	92.45	87.66	88.55
STSB	89.57	89.93	80.65	86.07
QQP	88.54	88.90	83.70	85.75
MNLI	86.19	86.89	82.27	84.91
QNLI	90.61	92.04	84.68	88.56
RTE	80.94	85.20	74.30	78.63
GLUE	84.71	85.54	77.42	81.63

ChronoBERT and ChronoGPT exhibit superior language understanding relative to similar-sized models and comparable to much larger Llama

GLUE Score Evaluation for Different LLMs

	Llama 3.1	BERT	FinBERT	StoriesLM
COLA	55.86	57.59	28.99	46.85
SST2	95.49	92.62	89.03	90.44
MRPC	88.22	90.76	88.59	89.33
STSB	90.67	90.07	85.72	87.01
QQP	89.67	88.21	86.60	86.88
MNLI	89.59	84.98	79.23	79.78
QNLI	95.35	91.52	86.12	87.44
RTE	85.63	80.43	67.00	67.15
GLUE	86.31	84.52	76.41	79.36

Validation of chronological consistency

- Our models could still have some lookahead bias if the publication dates of our training data have errors
- To detect leakage in the textual data used to pretrain our models, we evaluate them on events occurring after the model's knowledge cutoff
- For example, since ChronoBERT is a fill-mask model, we use each model vintage to predict the masked token in:

"After the 2000 U.S. presidential election, President [MASK] was inaugurated as U.S. President in the year 2001."

ChronoBERT knows only what it should

"After the {year} U.S. presidential election, President [MASK] was inaugurated as U.S. President in the year {year+1}."

							Accuracy	
Election year:	1992	2000	2008	2016	2020	2024	Pre	Post
Correct output	Clinton	Bush	Obama	Trump	Biden	Trump		
BERT	**Clinton**	Clinton	**Obama**	Obama	Obama	Obama	2/4	0/2
ModernBERT	**Clinton**	**Bush**	**Obama**	**Trump**	**Biden**	**Trump**	5/5	1/1
ChronoBERT$_{2000}$	**Clinton**	Clinton	Clinton	Clinton	Clinton	Wilson	1/2	0/4
ChronoBERT$_{2008}$	**Clinton**	**Bush**	Bush	Obama	Bush	Wilson	2/3	0/3
ChronoBERT$_{2016}$	**Clinton**	**Bush**	**Obama**	Obama	Obama	Obama	3/4	0/2
ChronoBERT$_{2020}$	**Clinton**	**Bush**	**Obama**	**Trump**	Trump	**Trump**	4/5	1/1
ChronoBERT$_{2024}$	**Clinton**	**Bush**	**Obama**	**Trump**	**Biden**	Biden	5/6	—
ChronoBERT$_{1999}$ through ChronoBERT$_{2024}$							68/83	1/73

blue = correct prediction

gray area = post-knowledge cutoff prediction

ChronoGPT knows only what it should

"U.S. Presidents in chronological order:
Took office in $\{year_{p-3}+1\}$: President $\{name_{p-3}\}$
Took office in $\{year_{p-2}+1\}$: President $\{name_{p-2}\}$
Took office in $\{year_{p-1}+1\}$: President $\{name_{p-1}\}$
Took office in $\{year_{p}+1\}$: President _____"

							Accuracy	
Election year:	1992	2000	2008	2016	2020	2024	Pre	Post
Correct output	Bill Clinton	George W	Barack Obama	Donald Trump	Joe Biden	Donald Trump		
GPT-2	**Bill Clinton**	Bill Clinton	**Barack Obama**	**Donald Trump**	George W	George W	3/4	0/2
GPT-2 XL	**Bill Clinton**	**George W**	**Barack Obama**	**Donald Trump**	James A	James Mattis	4/4	0/2
ChronoGPT_{2000}	**Bill Clinton**	Bill Clinton	Bill Clinton	Bill Clinton	Bill Clinton	Bill Clinton	1/2	0/4
ChronoGPT_{2008}	**Bill Clinton**	**George W**	George W	George W	George W	George W	2/3	0/3
ChronoGPT_{2016}	**Bill Clinton**	**George W**	**Barack Obama**	Barack Obama	Barack Obama	Barack Obama	3/4	0/2
ChronoGPT_{2020}	**Bill Clinton**	**George W**	**Barack Obama**	**Donald Trump**	Donald Trump	**Donald Trump**	4/5	1/1
ChronoGPT_{2024}	**Bill Clinton**	**George W**	**Barack Obama**	**Donald Trump**	**Joe Biden**	Joe Biden	5/6	—
ChronoGPT_{1999} through ChronoGPT_{2024}							77/83	4/73

blue = correct prediction
gray area = post-knowledge cutoff prediction

ChronoBERT knows only what it should

Masked-token predictions of event-defining tokens

							Accuracy	
Event year:	2001	2003	2008	2016	2020	2022	Pre	Post
Correct output	scandal	SARS	crisis	Brexit/EU	coronavirus	Chat		
BERT	**scandal**	flu	**crisis**	**EU**	flu	the	3/4	0/2
ModernBERT	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	Auto	5/6	—
ChronoBERT$_{2002}$	Amendment	Ebola	rate	election	influenza	EX	0/1	0/5
ChronoBERT$_{2004}$	**scandal**	**SARS**	rate	tax	influenza	SU	2/2	0/4
ChronoBERT$_{2008}$	**scandal**	**SARS**	**crisis**	Kyoto	influenza	AR	3/3	0/3
ChronoBERT$_{2016}$	**scandal**	**SARS**	**crisis**	**Brexit**	influenza	i	4/4	0/2
ChronoBERT$_{2020}$	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	V	5/5	0/1
ChronoBERT$_{2022}$	**scandal**	Ebola	**crisis**	**Brexit**	**coronavirus**	#	4/6	—
ChronoBERT$_{2024}$	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	**Chat**	6/6	—
ChronoBERT$_{1999}$ through ChronoBERT$_{2024}$							72/80	0/76

blue = correct prediction

gray area = post-knowledge cutoff prediction

ChronoBERT knows only what it should

"In 2001, the Enron [MASK] precipitated the passage of the Sarbanes-Oxley Act of 2002."

	▼						Accuracy	
Event year:	2001	2003	2008	2016	2020	2022	Pre	Post
Correct output	scandal	SARS	crisis	Brexit/EU	coronavirus	Chat		
BERT	**scandal**	flu	**crisis**	**EU**	flu	the	3/4	0/2
ModernBERT	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	Auto	5/6	—
ChronoBERT$_{2002}$	Amendment	Ebola	rate	election	influenza	EX	0/1	0/5
ChronoBERT$_{2004}$	**scandal**	**SARS**	rate	tax	influenza	SU	2/2	0/4
ChronoBERT$_{2008}$	**scandal**	**SARS**	**crisis**	Kyoto	influenza	AR	3/3	0/3
ChronoBERT$_{2016}$	**scandal**	**SARS**	**crisis**	**Brexit**	influenza	i	4/4	0/2
ChronoBERT$_{2020}$	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	V	5/5	0/1
ChronoBERT$_{2022}$	**scandal**	Ebola	**crisis**	**Brexit**	**coronavirus**	#	4/6	—
ChronoBERT$_{2024}$	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	**Chat**	6/6	—
ChronoBERT$_{1999}$ through ChronoBERT$_{2024}$							72/80	0/76

blue = correct prediction

gray area = post-knowledge cutoff prediction

ChronoBERT knows only what it should

"In 2003, the [MASK] outbreak led to a major public health crisis."

		▼					Accuracy	
Event year:	2001	2003	2008	2016	2020	2022	Pre	Post
Correct output	scandal	SARS	crisis	Brexit/EU	coronavirus	Chat		
BERT	**scandal**	flu	**crisis**	**EU**	flu	the	3/4	0/2
ModernBERT	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	Auto	5/6	—
ChronoBERT_{2002}	Amendment	Ebola	rate	election	influenza	EX	0/1	0/5
ChronoBERT_{2004}	**scandal**	**SARS**	rate	tax	influenza	SU	2/2	0/4
ChronoBERT_{2008}	**scandal**	**SARS**	**crisis**	Kyoto	influenza	AR	3/3	0/3
ChronoBERT_{2016}	**scandal**	**SARS**	**crisis**	**Brexit**	influenza	i	4/4	0/2
ChronoBERT_{2020}	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	V	5/5	0/1
ChronoBERT_{2022}	**scandal**	Ebola	**crisis**	**Brexit**	**coronavirus**	#	4/6	—
ChronoBERT_{2024}	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	**Chat**	6/6	—
ChronoBERT_{1999} through ChronoBERT_{2024}							72/80	0/76

blue = correct prediction

gray area = post-knowledge cutoff prediction

ChronoBERT knows only what it should

"In 2008, the subprime mortgage [MASK] severely impacted the global economy."

			▼				Accuracy	
Event year:	2001	2003	2008	2016	2020	2022	Pre	Post
Correct output	scandal	SARS	crisis	Brexit/EU	coronavirus	Chat		
BERT	**scandal**	flu	**crisis**	**EU**	flu	the	3/4	0/2
ModernBERT	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	Auto	5/6	—
ChronoBERT_{2002}	Amendment	Ebola	rate	election	influenza	EX	0/1	0/5
ChronoBERT_{2004}	**scandal**	**SARS**	rate	tax	influenza	SU	2/2	0/4
ChronoBERT_{2008}	**scandal**	**SARS**	**crisis**	Kyoto	influenza	AR	3/3	0/3
ChronoBERT_{2016}	**scandal**	**SARS**	**crisis**	**Brexit**	influenza	i	4/4	0/2
ChronoBERT_{2020}	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	V	5/5	0/1
ChronoBERT_{2022}	**scandal**	Ebola	**crisis**	**Brexit**	**coronavirus**	#	4/6	—
ChronoBERT_{2024}	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	**Chat**	6/6	—
ChronoBERT_{1999} through ChronoBERT_{2024}							72/80	0/76

blue = correct prediction
gray area = post-knowledge cutoff prediction

ChronoBERT knows only what it should

"In 2016, the uncertainty surrounding the [MASK] referendum led to heightened market volatility."

				▼			Accuracy	
Event year:	2001	2003	2008	2016	2020	2022	Pre	Post
Correct output	scandal	SARS	crisis	Brexit/EU	coronavirus	Chat		
BERT	**scandal**	flu	**crisis**	**EU**	flu	the	3/4	0/2
ModernBERT	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	Auto	5/6	—
ChronoBERT_{2002}	Amendment	Ebola	rate	election	influenza	EX	0/1	0/5
ChronoBERT_{2004}	**scandal**	**SARS**	rate	tax	influenza	SU	2/2	0/4
ChronoBERT_{2008}	**scandal**	**SARS**	**crisis**	Kyoto	influenza	AR	3/3	0/3
ChronoBERT_{2016}	**scandal**	**SARS**	**crisis**	**Brexit**	influenza	i	4/4	0/2
ChronoBERT_{2020}	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	V	5/5	0/1
ChronoBERT_{2022}	**scandal**	Ebola	**crisis**	**Brexit**	**coronavirus**	#	4/6	—
ChronoBERT_{2024}	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	**Chat**	6/6	—
ChronoBERT_{1999} through ChronoBERT_{2024}							72/80	0/76

blue = correct prediction

gray area = post-knowledge cutoff prediction

ChronoBERT knows only what it should

"In 2020, the [MASK] pandemic severely impacted the global economy."

					▼		Accuracy	
Event year:	2001	2003	2008	2016	2020	2022	Pre	Post
Correct output	scandal	SARS	crisis	Brexit/EU	coronavirus	Chat		
BERT	**scandal**	flu	**crisis**	**EU**	flu	the	3/4	0/2
ModernBERT	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	Auto	5/6	—
ChronoBERT_{2002}	Amendment	Ebola	rate	election	influenza	EX	0/1	0/5
ChronoBERT_{2004}	**scandal**	**SARS**	rate	tax	influenza	SU	2/2	0/4
ChronoBERT_{2008}	**scandal**	**SARS**	**crisis**	Kyoto	influenza	AR	3/3	0/3
ChronoBERT_{2016}	**scandal**	**SARS**	**crisis**	**Brexit**	influenza	i	4/4	0/2
ChronoBERT_{2020}	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	V	5/5	0/1
ChronoBERT_{2022}	**scandal**	Ebola	**crisis**	**Brexit**	**coronavirus**	#	4/6	—
ChronoBERT_{2024}	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	**Chat**	6/6	—
ChronoBERT_{1999} through ChronoBERT_{2024}							72/80	0/76

blue = correct prediction
gray area = post-knowledge cutoff prediction

ChronoBERT knows only what it should

"In 2022, the public release of [MASK]GPT marked a major milestone for generative AI."

						▼	Accuracy	
Event year:	2001	2003	2008	2016	2020	2022	Pre	Post
Correct output	scandal	SARS	crisis	Brexit/EU	coronavirus	Chat		
BERT	**scandal**	flu	**crisis**	**EU**	flu	the	3/4	0/2
ModernBERT	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	Auto	5/6	—
ChronoBERT$_{2002}$	Amendment	Ebola	rate	election	influenza	EX	0/1	0/5
ChronoBERT$_{2004}$	**scandal**	**SARS**	rate	tax	influenza	SU	2/2	0/4
ChronoBERT$_{2008}$	**scandal**	**SARS**	**crisis**	Kyoto	influenza	AR	3/3	0/3
ChronoBERT$_{2016}$	**scandal**	**SARS**	**crisis**	**Brexit**	influenza	i	4/4	0/2
ChronoBERT$_{2020}$	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	V	5/5	0/1
ChronoBERT$_{2022}$	**scandal**	Ebola	**crisis**	**Brexit**	**coronavirus**	#	4/6	—
ChronoBERT$_{2024}$	**scandal**	**SARS**	**crisis**	**Brexit**	**coronavirus**	**Chat**	6/6	—
ChronoBERT$_{1999}$ through ChronoBERT$_{2024}$							72/80	0/76

blue = correct prediction

gray area = post-knowledge cutoff prediction

News data and methods

- Dow Jones Newswire data from 2007 to 2023
- For each firm-day observation, aggregate news headlines related to the firm within the trading day window
- LLM is used to embed these sets of headlines into a numerical vector representation

Returns data and methods

Following Chen-Kelly-Xiu (2023)

- Fit a Fama-MacBeth regression with a ridge penalty to map news embeddings $e_{i,t}$ to return predictions $r_{i,t+1}$
- Each month m, we estimate the following cross-sectional ridge regression:

$$r_{i,t+1} = \alpha_m + \beta_m' e_{i,t} + \epsilon_{i,t+1}, \quad \text{for } i = 1, \cdots N \text{ and } t = 1 \cdots T, \tag{1}$$

- To construct real-time out-of-sample forecasts, in month m', we use an average of forecasts over all previous months' cross-sectional models:

$$\hat{r}_{i,t+1} = \bar{\alpha}_{m'} + \bar{\beta}_{m'}' e_{i,t}, \quad \text{for } i = 1, \cdots N \text{ and } t = 1 \cdots T, \tag{2}$$

- Using these out-of-sample predictions, we sort stocks into decile portfolios at the end of each trading day

ChronoBERT and ChronoGPT are comparable to much larger Llama

Performance of LLM Portfolios

	ChronoBERT$_{\text{Realtime}}$			ChronoGPT$_{\text{Realtime}}$			Llama 3.1			GPT-2 XL		
	Mean	SD	SR	Mean	SD	SR	Mean	SD	SR	Mean	SD	SR
Low (L)	-23.30	25.86	-0.90	-19.82	25.15	-0.79	-23.71	26.15	-0.91	-22.25	25.75	-0.86
2	-2.43	25.20	-0.10	1.48	25.48	0.06	-4.77	25.31	-0.19	-1.36	25.32	-0.05
3	4.17	25.64	0.16	0.61	24.87	0.02	-0.24	24.86	-0.01	-1.10	24.82	-0.04
4	4.17	24.58	0.17	2.41	24.26	0.10	3.84	24.62	0.16	2.60	24.43	0.11
5	3.94	24.22	0.16	7.18	24.54	0.29	7.47	24.65	0.30	6.24	24.30	0.26
6	10.81	24.13	0.45	5.78	24.20	0.24	12.03	24.23	0.50	10.14	24.53	0.41
7	14.56	24.23	0.60	10.82	23.84	0.45	13.31	24.33	0.55	16.90	24.76	0.68
8	16.38	23.64	0.69	15.74	24.30	0.65	15.13	23.79	0.64	15.69	24.09	0.65
9	23.95	24.45	0.98	22.63	24.19	0.94	24.68	23.88	1.03	24.85	24.24	1.03
High (H)	37.71	24.53	1.54	43.08	25.41	1.70	42.20	25.05	1.68	38.25	24.53	1.56
H-L	**61.02**	**12.72**	**4.80**	**62.90**	**12.78**	**4.92**	**65.91**	**13.46**	**4.90**	**60.51**	**13.24**	**4.57**

ChronoBERT and ChronoGPT are better than similar-sized LLMs

Performance of LLM Portfolios

	ModernBERT			BERT			FinBERT			StoriesLM		
	Mean	SD	SR	Mean	SD	SR	Mean	SD	SR	Mean	SD	SR
Low(L)	-24.39	26.07	-0.94	-22.52	26.21	-0.86	-23.96	26.86	-0.89	-17.80	26.52	-0.67
2	0.49	25.86	0.02	-5.05	25.55	-0.20	-3.17	25.64	-0.12	-1.19	25.26	-0.05
3	1.64	25.08	0.07	3.12	24.92	0.13	3.36	24.83	0.14	1.86	24.92	0.07
4	5.41	24.81	0.22	8.14	24.62	0.33	7.19	24.52	0.29	5.90	24.62	0.24
5	8.67	24.67	0.35	10.81	24.44	0.44	9.17	24.39	0.38	4.99	24.30	0.21
6	10.02	23.99	0.42	9.38	24.02	0.39	11.47	24.03	0.48	11.88	23.90	0.50
7	12.99	23.93	0.54	14.54	23.83	0.61	16.54	23.92	0.69	12.41	23.66	0.52
8	16.17	23.69	0.68	18.51	24.04	0.77	19.16	23.65	0.81	18.93	24.19	0.78
9	23.45	23.97	0.98	19.68	23.90	0.82	20.70	23.88	0.87	23.25	24.30	0.96
High(H)	35.53	24.42	1.45	33.37	24.88	1.34	29.51	24.60	1.20	29.73	24.78	1.20
H-L	**59.92**	**13.03**	**4.60**	**55.89**	**13.38**	**4.18**	**53.47**	**13.85**	**3.86**	**47.53**	**13.90**	**3.42**

ChronoBERT and ChronoGPT are comparable to much larger Llama

P-value of Pairwise Sharpe Ratio Difference Tests

	ChronoBERT	ChronoGPT	Llama 3.1	GPT-2 XL	ModernBERT	BERT	FinBERT	StoriesLM
ChronoBERT		0.717	0.685	0.141	0.144	0.005	0.002	0.000
ChronoGPT	0.283		0.453	0.051	0.075	0.001	0.001	0.000
Llama 3.1	0.315	0.547		0.048	0.069	0.001	0.000	0.000
GPT-2 XL	0.859	0.950	0.952		0.555	0.038	0.010	0.000
ModernBERT	0.856	0.925	0.931	0.445		0.029	0.007	0.000
BERT	0.995	0.999	0.999	0.962	0.971		0.116	0.005
FinBERT	0.998	0.999	1.000	0.991	0.993	0.884		0.098
StoriesLM	1.000	1.000	1.000	1.000	1.000	0.995	0.902	

Each entry corresponds to a test of the null hypothesis that the Sharpe ratio of the model in the row is smaller than that of the model in the column.

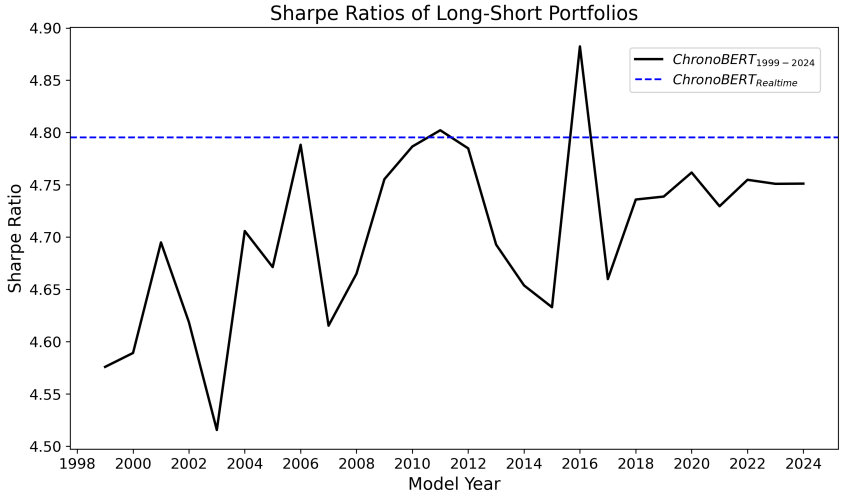
Portfolio Alphas

Memorization of General Trend vs. Idiosyncratic Information

	Mean	α_{F5}	α_{F6}	α_{HXZ}	α_{DHS}
ChronoBERT$_{\text{Realtime}}$	61.02	60.50	60.55	60.71	60.10
	(4.59)	(4.57)	(4.58)	(4.54)	(4.59)
ChronoGPT$_{\text{Realtime}}$	62.90	62.87	62.89	62.96	62.73
	(4.53)	(4.52)	(4.52)	(4.44)	(4.49)
Llama 3.1	65.91	65.01	65.06	65.82	64.95
	(4.75)	(4.68)	(4.69)	(4.68)	(4.74)
GPT-2 XL	60.51	59.48	59.52	60.37	59.36
	(4.45)	(4.36)	(4.36)	(4.38)	(4.43)
ModernBERT	59.92	59.41	59.49	59.24	59.05
	(4.42)	(4.34)	(4.35)	(4.32)	(4.38)
BERT	55.89	54.36	54.50	55.37	54.72
	(4.28)	(4.20)	(4.23)	(4.22)	(4.26)
FinBERT	53.47	52.11	52.27	53.54	52.81
	(4.23)	(4.19)	(4.20)	(4.10)	(4.15)
StoriesLM	47.53	46.56	46.67	47.16	46.28
	(4.29)	(4.24)	(4.26)	(4.25)	(4.26)

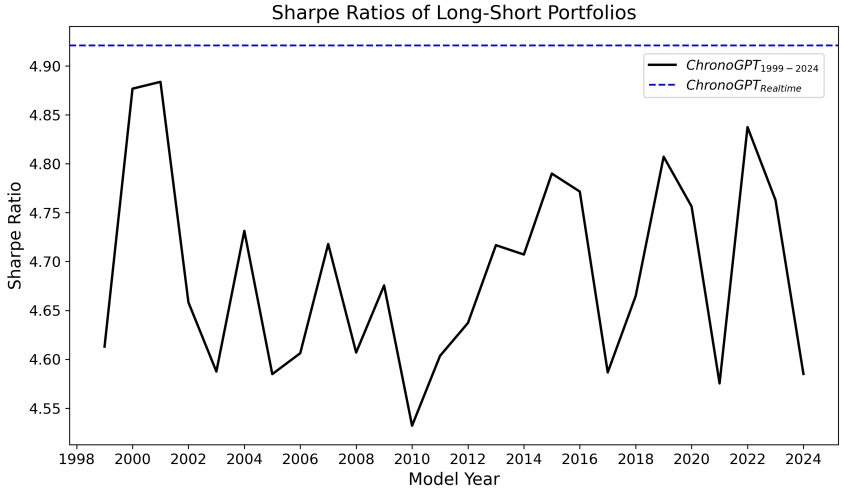
Time-aligned LLMs predict better than back- or future-dated models

Portfolios Performance across ChronoBERT Vintages



Time-aligned LLMs predict better than back- or future-dated models

Portfolios Performance across ChronoGPT Vintages



Conclusion

- Chronological inconsistency can bias LLM-based empirical estimates
- We train a suite of chronologically consistent LLMs
- ChronoBERT and ChronoGPT exhibit superior language understanding relative to similar-sized models and comparable to much larger Llama models
- In an asset pricing application predicting next-day stock returns from financial news, ChronoBERT's Sharpe ratio (4.80) and ChronoGPT's Sharpe ratio (4.92) are comparable to much larger (and inconsistent) Llama (4.90)
- Find modest lookahead bias in this setting
- Our models are available on hugging face for researchers to evaluate bias in their applications

What's next?

Chronologically Consistent Generative AI

- We are releasing a family of chronologically consistent instruction models to the research community.
- We are offering:
  - a simple, conversational chat interface;
  - fully open, fixed model weights that guarantee replicability;
  - a conservative lower bound on lookahead bias in a wide range of prediction tasks.
- ChronoGPT-Instruct model series is available on huggingface, as is the instruction-tuning dataset we curated.

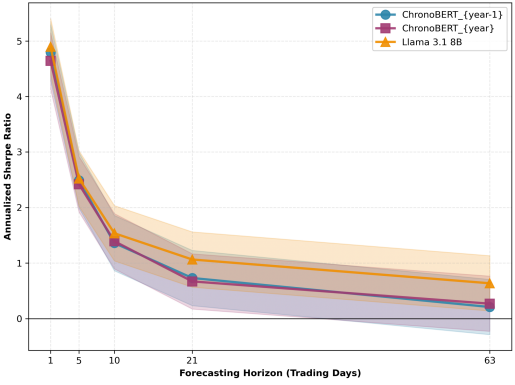
What's next?

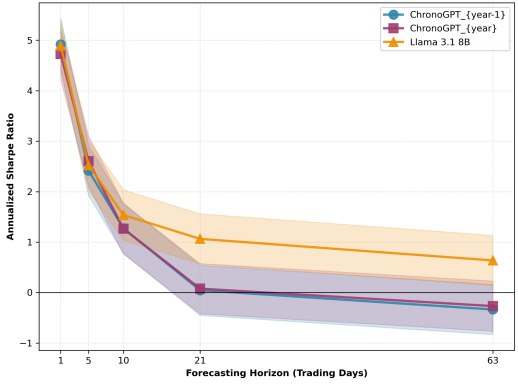
Chronologically Consistent Generative AI

- We are releasing a family of chronologically consistent instruction models to the research community.
- We are offering:
  - a simple, conversational chat interface;
  - fully open, fixed model weights that guarantee replicability;
  - a conservative lower bound on lookahead bias in a wide range of prediction tasks.
- ChronoGPT-Instruct model series is available on huggingface, as is the instruction-tuning dataset we curated.

Lookahead Bias over Longer Horizons

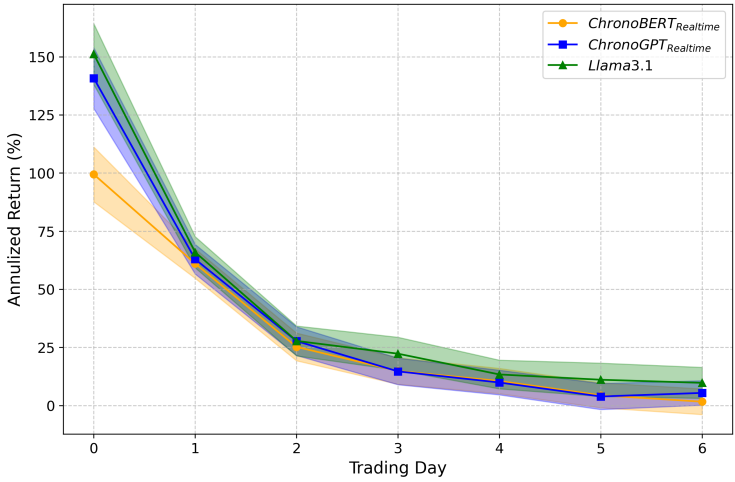
Portfolios Performance across Forecasting Horizons





Speed of news incorporation

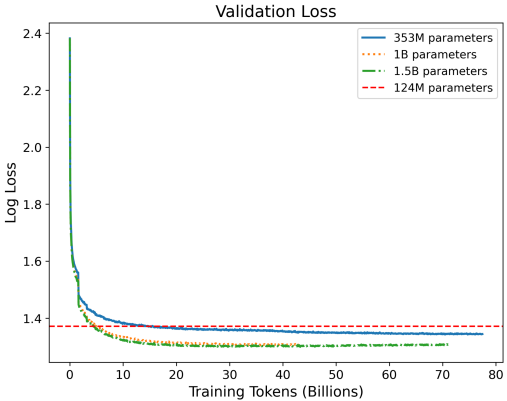
Portfolio Performances over time

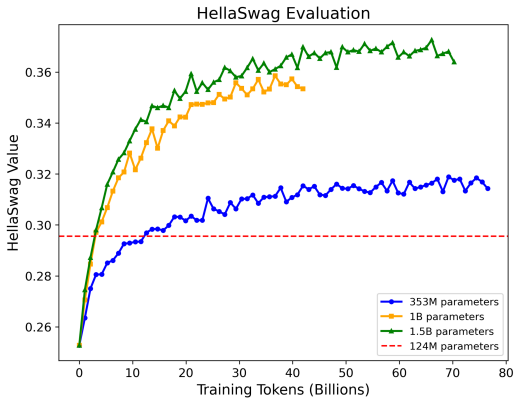


Scaling law

- Can we further improve performances despite the inherent limitations of the chronological training data?
- Model performance is a function of parameter count, data size, and compute budget
- Scaling training data is challenging, and increasing epochs no longer helps
- Let's try scaling up model parameters!

Scaling up ChronoGPT





Tercile Portfolios

	ChronoBERT$_{\text{Realtime}}$			ChronoGPT$_{\text{Realtime}}$			Llama 3.1			GPT-2 XL		
	Mean	SD	SR	Mean	SD	SR	Mean	SD	SR	Mean	SD	SR
Low (L)	-6.22	24.46	-0.25	-4.70	24.16	-0.19	-8.22	24.48	-0.34	-6.61	24.34	-0.27
2	8.51	23.65	0.36	5.98	23.52	0.25	9.46	23.68	0.40	8.43	23.62	0.36
High (H)	24.98	23.50	1.06	25.98	23.89	1.09	26.02	23.50	1.11	25.44	23.67	1.07
H-L	**31.20**	**7.26**	**4.30**	**30.68**	**6.86**	**4.47**	**34.24**	**7.36**	**4.65**	**32.05**	**7.21**	**4.45**
	ModernBERT			BERT			FinBERT			StoriesLM		
	Mean	SD	SR	Mean	SD	SR	Mean	SD	SR	Mean	SD	SR
Low (L)	-6.53	24.52	-0.27	-6.82	24.43	-0.28	-5.88	24.75	-0.24	-4.77	24.40	-0.20
2	10.03	23.72	0.42	11.16	23.59	0.47	10.52	23.55	0.45	8.78	23.53	0.37
High (H)	23.77	23.38	1.02	22.94	23.60	0.97	22.62	23.35	0.97	23.26	23.71	0.98
H-L	**30.30**	**7.24**	**4.19**	**29.76**	**7.40**	**4.02**	**28.50**	**7.72**	**3.69**	**28.03**	**7.56**	**3.71**

Full News Text

Comparison with Llama 3.1 and GPT-2 XL

	ChronoBERT$_{\text{Realtime}}$			ChronoGPT$_{\text{Realtime}}$			Llama 3.1			GPT-2 XL		
	Mean	SD	SR	Mean	SD	SR	Mean	SD	SR	Mean	SD	SR
Low (L)	-24.93	26.36	-0.95	-23.72	27.20	-0.87	-33.10	27.80	-1.19	-29.59	27.46	-1.08
2	-1.30	24.88	-0.05	-2.84	25.39	-0.11	-3.01	26.15	-0.12	-3.85	25.11	-0.15
3	3.86	25.23	0.15	1.44	24.33	0.06	2.08	24.65	0.08	0.73	24.45	0.03
4	9.54	24.58	0.39	5.61	24.09	0.23	8.93	24.39	0.37	7.49	24.42	0.31
5	9.92	24.10	0.41	5.81	24.02	0.24	12.50	24.05	0.52	11.84	24.52	0.48
6	11.62	24.16	0.48	13.35	24.42	0.55	13.15	23.64	0.56	16.39	24.27	0.68
7	12.73	23.51	0.54	14.86	24.01	0.62	17.17	23.89	0.72	15.25	23.87	0.64
8	16.28	23.96	0.68	19.46	24.00	0.81	19.02	23.71	0.80	19.13	23.64	0.81
9	24.39	24.06	1.01	25.68	24.15	1.06	24.21	23.81	1.02	24.67	23.67	1.04
High (H)	28.93	25.26	1.15	31.40	24.60	1.28	30.11	24.63	1.22	28.99	24.79	1.17
H-L	**53.86**	**13.40**	**4.02**	**55.12**	**14.30**	**3.85**	**63.20**	**14.97**	**4.22**	**58.58**	**14.02**	**4.18**

Full News Text

Comparison with Other Smaller Models

	ModernBERT			BERT			FinBERT			StoriesLM		
	Mean	SD	SR	Mean	SD	SR	Mean	SD	SR	Mean	SD	SR
Low (L)	-20.33	26.62	-0.76	-22.54	27.14	-0.83	-23.67	27.10	-0.87	-19.88	26.77	-0.74
2	-3.84	25.54	-0.15	-0.33	24.93	-0.01	0.55	25.87	0.02	0.74	25.32	0.03
3	1.74	25.04	0.07	1.50	24.40	0.06	3.89	24.92	0.16	3.65	24.58	0.15
4	8.15	24.25	0.34	4.07	24.34	0.17	7.21	24.11	0.30	7.38	24.42	0.30
5	10.82	24.35	0.44	12.17	24.29	0.50	10.77	24.01	0.45	7.91	23.79	0.33
6	12.71	24.29	0.52	10.75	23.67	0.45	12.20	23.82	0.51	13.97	24.12	0.58
7	15.69	23.92	0.66	15.76	23.92	0.66	16.78	24.18	0.69	13.24	23.83	0.56
8	15.81	23.92	0.66	19.66	24.03	0.82	17.18	23.79	0.72	19.28	24.05	0.80
9	21.57	23.84	0.90	22.18	24.22	0.92	20.53	23.87	0.86	21.26	24.44	0.87
High (H)	28.69	24.43	1.17	27.81	25.25	1.10	25.63	24.25	1.06	23.54	24.87	0.95
H-L	**49.03**	**13.98**	**3.51**	**50.35**	**14.05**	**3.58**	**49.30**	**13.82**	**3.57**	**43.42**	**13.46**	**3.23**

Subsamples

	S&P 500			non-S&P 500		
	mean	SD	SR	mean	SD	SR
ChronoBERT	7.49	11.30	0.66	75.03	15.03	4.99
ChronoGPT	7.10	11.33	0.63	76.18	15.28	4.99
Llama 3.1	7.24	12.34	0.59	80.87	15.80	5.12
GPT-2 XL	7.83	11.01	0.71	73.62	15.49	4.75
ModernBERT	5.92	11.54	0.51	72.69	15.40	4.72
	Large market cap			Small market cap		
	Mean	SD	SR	Mean	SD	SR
ChronoBERT	9.76	9.30	1.05	98.11	20.87	4.70
ChronoGPT	11.82	9.51	1.24	96.97	20.64	4.70
Llama 3.1	12.48	10.30	1.21	104.12	21.05	4.95
GPT-2 XL	9.54	9.52	1.00	96.36	21.37	4.51
ModernBERT	9.42	9.59	0.98	92.44	20.97	4.41

Subsamples

	High news coverage			Low news coverage		
	Mean	SD	SR	Mean	SD	SR
ChronoBERT	30.67	13.68	2.24	81.65	17.94	4.55
ChronoGPT	34.65	14.20	2.44	79.45	18.00	4.41
Llama 3.1	33.23	14.81	2.24	88.26	18.68	4.72
GPT-2 XL	32.14	14.55	2.21	82.75	18.63	4.44
ModernBERT	29.73	14.04	2.12	79.15	18.30	4.33
	Earnings Days			Non-Earnings Days		
	Mean	SD	SR	Mean	SD	SR
ChronoBERT	67.86	40.36	1.68	55.18	13.37	4.13
ChronoGPT	72.18	39.14	1.84	57.25	13.27	4.31
Llama 3.1	51.90	51.26	1.01	62.99	15.20	4.14
GPT-2 XL	55.88	37.04	1.51	55.28	14.12	3.92
ModernBERT	58.84	42.25	1.39	55.58	13.89	4.00