

Intro  
ooooo

Pretraining  
oo

Language understanding  
oooooooo

Chronological consistency  
oooo

Return forecasting  
ooooooo

Conclusion  
o

# Chronologically Consistent Large Language Models

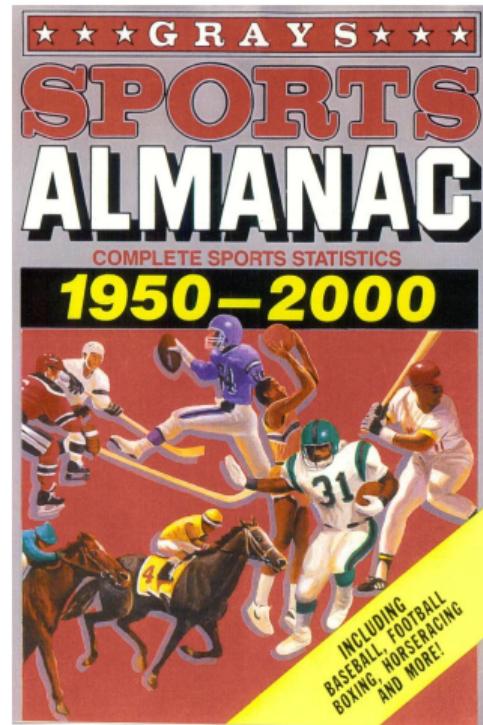
Songrun He<sup>1</sup>, Linying Lv<sup>1</sup>, Asaf Manela<sup>1</sup> and Jimmy Wu<sup>1</sup>

<sup>1</sup>Washington University in St. Louis

July 2025

# Chronological inconsistency, lookahead bias, and training leakage

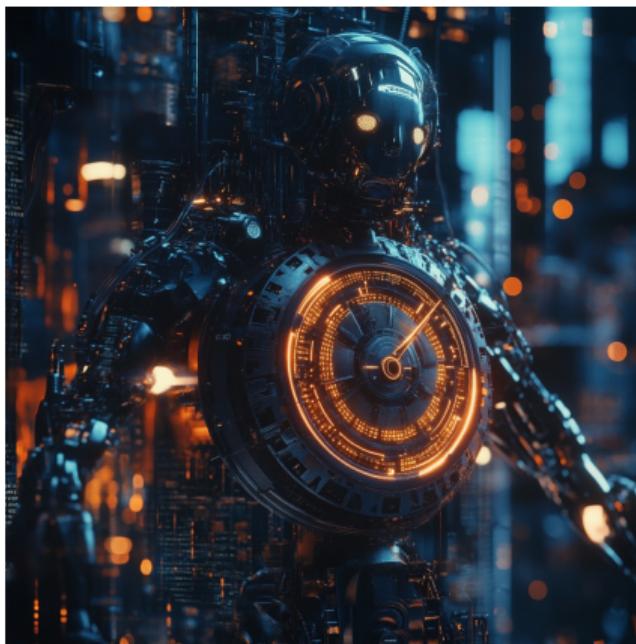
- ▶ Large language models (LLMs) now permeate social sciences
- ▶ They let us test hypotheses previously unquantifiable
- ▶ But they are trained on data that did not exist at the historical moment
- ▶ **Lookahead bias** (Glasserman-Lin 2023, Sarkar-Vafa 2024) and **training leakage** (Ludwig-Mullainathan-Rambachan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



# Chronological inconsistency in finance

- ▶ Finance is particularly sensitive to lookahead bias
- ▶ Market efficiency tests assume 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
  - ▶  $ChronoBERT_{1999}, \dots, ChronoBERT_{2024}$
  - ▶  $ChronoGPT_{1999}, \dots, ChronoGPT_{2024}$
  - ▶ 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

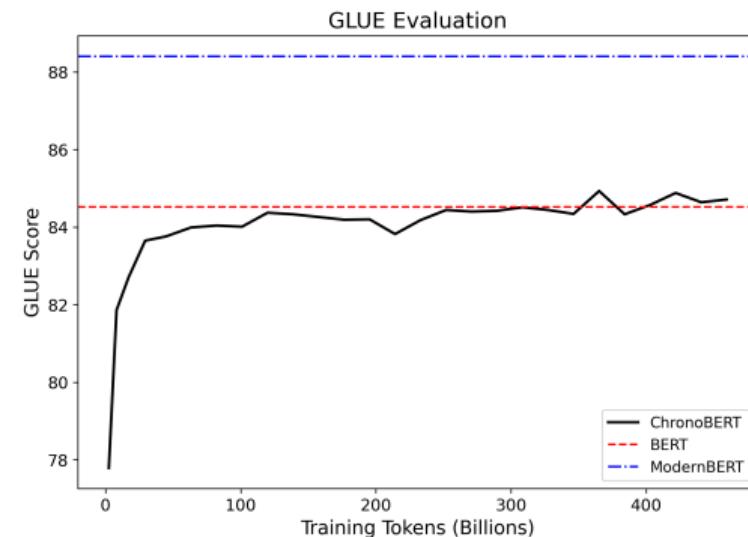
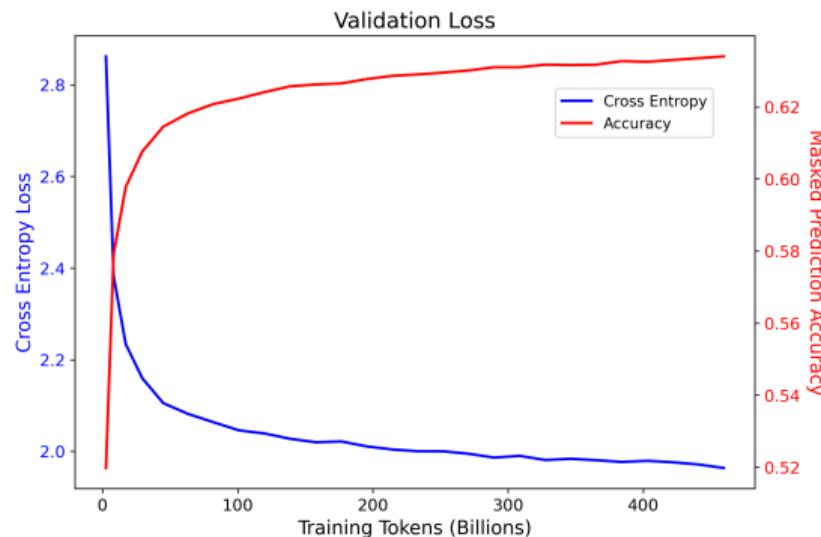
- ▶ Training LLMs is usually split into:
  1. Pretraining to predict missing words in text sequences
  2. Finetuning for specific applications (e.g. chat, Q&A, reasoning)
- ▶ We pretrain  $ChronoBERT_t$  and  $ChronoGPT_t$  only on text available before  $t$ 
  - ▶ For example, web pages crawled in 2005 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

Intro  
oooooPretraining  
ooLanguage understanding  
●ooooooChronological consistency  
ooooReturn forecasting  
oooooooConclusion  
o

*ChronoBERT<sub>1999</sub>* improves to the point it surpasses BERT on GLUE language benchmarks



Intro  
ooooo

Pretraining  
oo

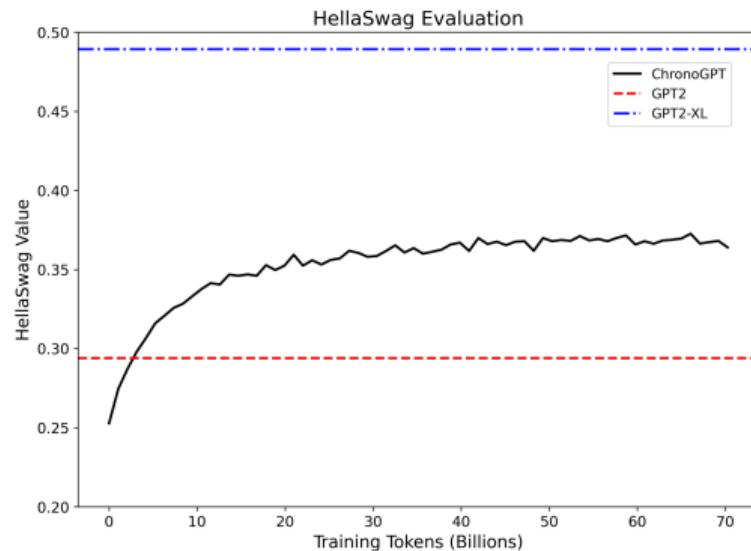
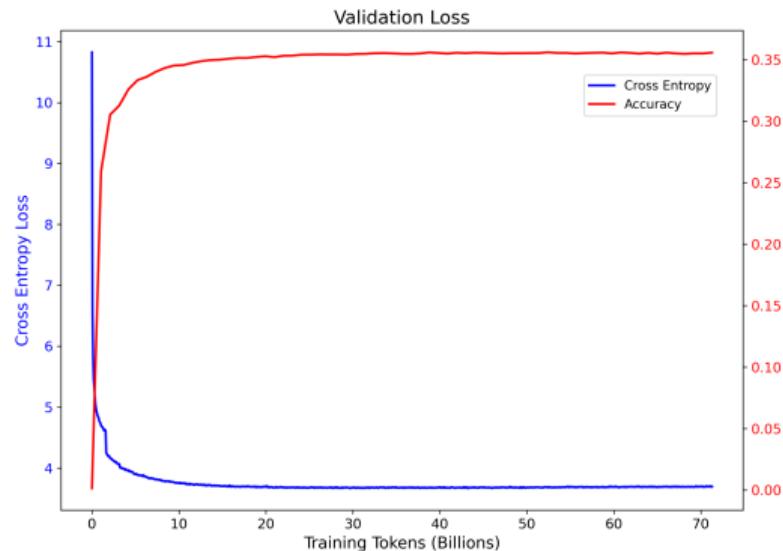
Language understanding  
○●○○○○○

Chronological consistency  
oooo

Return forecasting  
ooooooo

Conclusion  
o

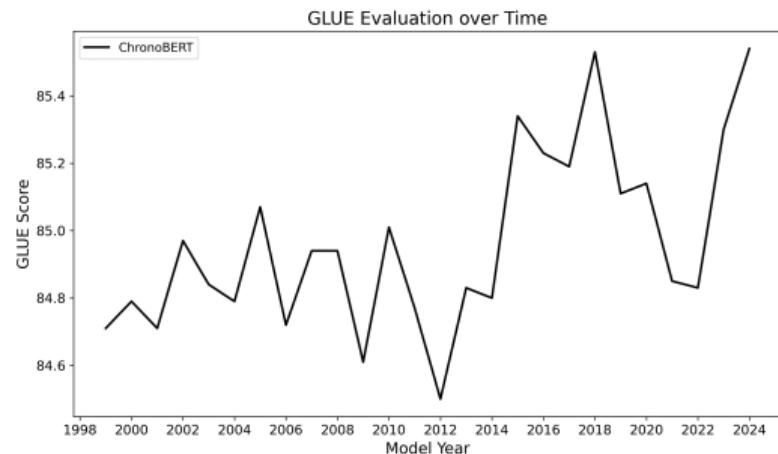
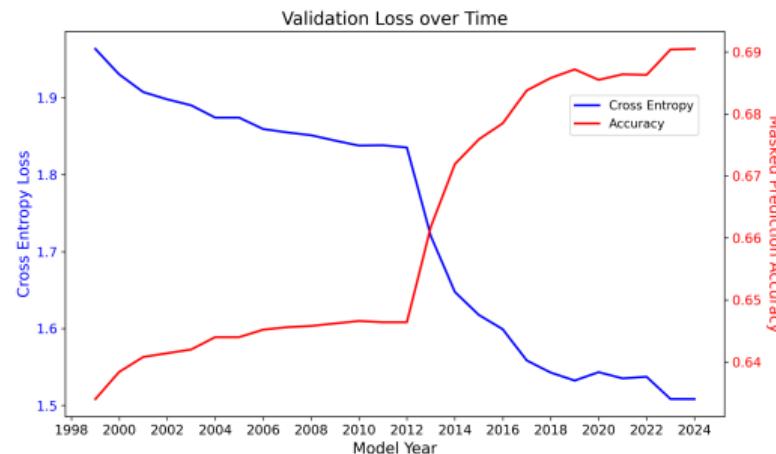
*ChronoGPT<sub>1999</sub>* improves to the point it surpass GPT-2 on HellaSwag token generation



Intro  
oooooPretraining  
ooLanguage understanding  
oo●ooooChronological consistency  
ooooReturn forecasting  
oooooooConclusion  
o

# Subsequent ChronoBERT models improve further over time

*ChronoBERT<sub>1999</sub>, ..., ChronoBERT<sub>2024</sub>*



Intro  
ooooo

Pretraining  
oo

Language understanding  
ooo●ooo

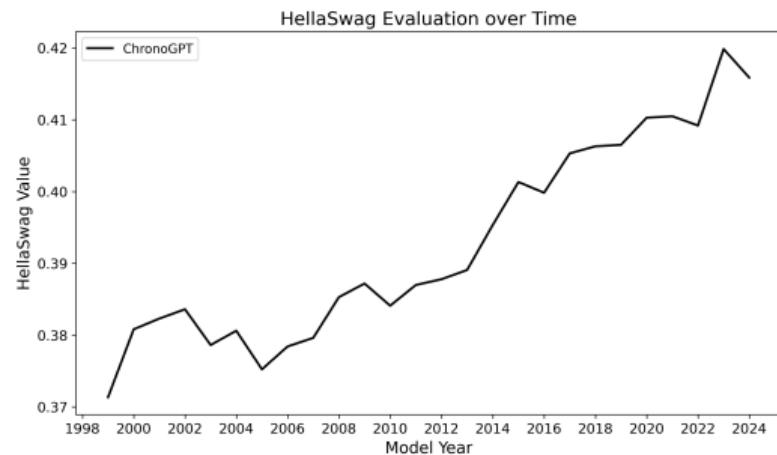
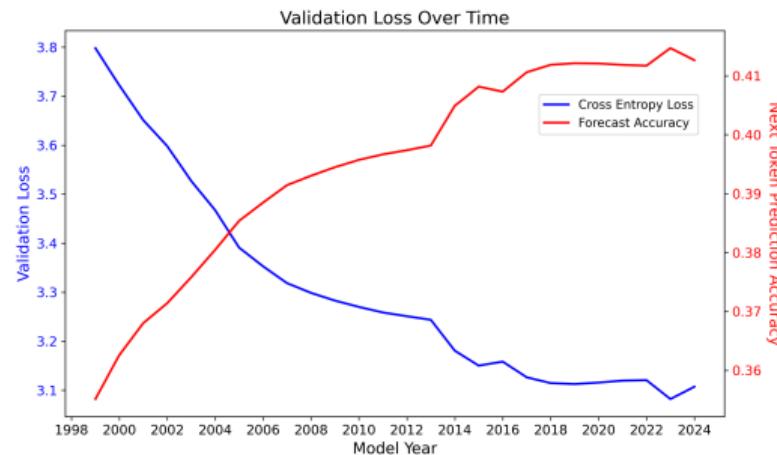
Chronological consistency  
oooo

Return forecasting  
ooooooo

Conclusion  
o

# Subsequent ChronoGPT models improve further over time

*ChronoGPT<sub>1999</sub>, ..., ChronoGPT<sub>2024</sub>*



# Size, context and knowledge cutoff for different LLMs

	Parameters	Context Tokens	Knowledge Cutoff
ChronoBERT <sub>1999</sub>	149M	1,024	December, 1999
:	:	:	:
ChronoBERT <sub>2024</sub>	149M	1,024	December, 2024
ChronoGPT <sub>1999</sub>	1,552M	1,792	December, 1999
:	:	:	:
ChronoGPT <sub>2024</sub>	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 <sub>1999</sub>	ChronoBERT <sub>2024</sub>	ChronoGPT <sub>1999</sub>	ChronoGPT <sub>2024</sub>
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-1} U.S. presidential election, President [MASK] was inaugurated as U.S. President in the year {year}."*

Prompt year:	1993	2001	2009	2017	2021	2025	Accuracy	
							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 <sub>2000</sub>	Clinton	Clinton	Clinton	Clinton	Clinton	Wilson	1/1	0/5
ChronoBERT <sub>2008</sub>	Clinton	Bush	Bush	Obama	Bush	Wilson	2/2	0/4
ChronoBERT <sub>2012</sub>	Obama	Obama	Obama	Obama	Obama	Obama	1/3	0/3
ChronoBERT <sub>2016</sub>	Clinton	Bush	Obama	Obama	Obama	Obama	3/3	0/3
ChronoBERT <sub>2020</sub>	Clinton	Bush	Obama	Trump	Trump	Trump	4/4	1/2
ChronoBERT <sub>2024</sub>	Clinton	Bush	Obama	Trump	Biden	Biden	5/5	0/1
ChronoBERT <sub>1999</sub> through ChronoBERT <sub>2024</sub>							68/78	1/78

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<sub>p-3</sub>}: President {name<sub>p-3</sub>}*

*Took office in {year<sub>p-2</sub>}: President {name<sub>p-2</sub>}*

*Took office in {year<sub>p-1</sub>}: President {name<sub>p-1</sub>}*

*Took office in {year<sub>p</sub>}: President \_\_\_\_\_"*

Prompt year:	1993	2001	2009	2017	2021	2025	Accuracy	
	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 <sub>2000</sub>	Bill Clinton	1/1	0/5					
ChronoGPT <sub>2008</sub>	Bill Clinton	George W.	2/2	0/4				
ChronoGPT <sub>2012</sub>	Bill Clinton	George W.	Barack Obama	Barack Obama	Barack Obama	Barack Obama	3/3	0/3
ChronoGPT <sub>2016</sub>	Bill Clinton	George W.	Barack Obama	Barack Obama	Barack Obama	Barack Obama	3/3	0/3
ChronoGPT <sub>2020</sub>	Bill Clinton	George W.	Barack Obama	Donald Trump	Donald Trump	Donald Trump	4/4	1/2
ChronoGPT <sub>2024</sub>	Bill Clinton	George W.	Barack Obama	Donald Trump	Joe Biden	Joe Biden	5/5	0/1
ChronoGPT <sub>1999</sub> through ChronoGPT <sub>2024</sub>							77/78	4/78

blue = correct prediction

gray area = post-knowledge cutoff prediction

# ChronoBERT knows only what it should

Masked-token predictions of event-defining tokens

Prompt year:	2001	2003	2008	2016	2020	2022	Accuracy	
							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 <sub>2002</sub>	Amendment	Ebola	rate	election	influenza	EX	0/1	0/5
ChronoBERT <sub>2004</sub>	scandal	SARS	rate	tax	influenza	SU	2/2	0/4
ChronoBERT <sub>2008</sub>	scandal	SARS	crisis	Kyoto	influenza	AR	3/3	0/3
ChronoBERT <sub>2012</sub>	scandal	SARS	crisis	presidential	influenza	AR	3/3	0/3
ChronoBERT <sub>2016</sub>	scandal	SARS	crisis	Brexit	influenza	i	4/4	0/2
ChronoBERT <sub>2020</sub>	scandal	SARS	crisis	Brexit	coronavirus	V	5/5	0/1
ChronoBERT <sub>2022</sub>	scandal	Ebola	crisis	Brexit	coronavirus	#	4/6	—
ChronoBERT <sub>2024</sub>	scandal	SARS	crisis	Brexit	coronavirus	Chat	6/6	—
ChronoBERT <sub>1999</sub> through ChronoBERT <sub>2024</sub>							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."*

Prompt year:	▼							Accuracy	
		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 <sub>2002</sub>	Amendment	Ebola	rate	election	influenza	EX	0/1	0/5	
ChronoBERT <sub>2004</sub>	scandal	SARS	rate	tax	influenza	SU	2/2	0/4	
ChronoBERT <sub>2008</sub>	scandal	SARS	crisis	Kyoto	influenza	AR	3/3	0/3	
ChronoBERT <sub>2012</sub>	scandal	SARS	crisis	presidential	influenza	AR	3/3	0/3	
ChronoBERT <sub>2016</sub>	scandal	SARS	crisis	Brexit	influenza	i	4/4	0/2	
ChronoBERT <sub>2020</sub>	scandal	SARS	crisis	Brexit	coronavirus	V	5/5	0/1	
ChronoBERT <sub>2022</sub>	scandal	Ebola	crisis	Brexit	coronavirus	#	4/6	—	
ChronoBERT <sub>2024</sub>	scandal	SARS	crisis	Brexit	coronavirus	Chat	6/6	—	
ChronoBERT <sub>1999</sub> through ChronoBERT <sub>2024</sub>							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."*

		▼	2001	2003	2008	2016	2020	2022	Accuracy	
Prompt year:									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 <sub>2002</sub>	Amendment	Ebola	rate	election	influenza	EX			0/1	0/5
ChronoBERT <sub>2004</sub>	scandal	SARS	rate	tax	influenza	SU			2/2	0/4
ChronoBERT <sub>2008</sub>	scandal	SARS	crisis	Kyoto	influenza	AR			3/3	0/3
ChronoBERT <sub>2012</sub>	scandal	SARS	crisis	presidential	influenza	AR			3/3	0/3
ChronoBERT <sub>2016</sub>	scandal	SARS	crisis	Brexit	influenza	i			4/4	0/2
ChronoBERT <sub>2020</sub>	scandal	SARS	crisis	Brexit	coronavirus	V			5/5	0/1
ChronoBERT <sub>2022</sub>	scandal	Ebola	crisis	Brexit	coronavirus	#			4/6	—
ChronoBERT <sub>2024</sub>	scandal	SARS	crisis	Brexit	coronavirus	Chat			6/6	—
ChronoBERT <sub>1999</sub> through ChronoBERT <sub>2024</sub>								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."*

Prompt year:	2001	2003	2008	2016	2020	2022	Accuracy	
							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 <sub>2002</sub>	Amendment	Ebola	rate	election	influenza	EX	0/1	0/5
ChronoBERT <sub>2004</sub>	scandal	SARS	rate	tax	influenza	SU	2/2	0/4
ChronoBERT <sub>2008</sub>	scandal	SARS	crisis	Kyoto	influenza	AR	3/3	0/3
ChronoBERT <sub>2012</sub>	scandal	SARS	crisis	presidential	influenza	AR	3/3	0/3
ChronoBERT <sub>2016</sub>	scandal	SARS	crisis	Brexit	influenza	i	4/4	0/2
ChronoBERT <sub>2020</sub>	scandal	SARS	crisis	Brexit	coronavirus	V	5/5	0/1
ChronoBERT <sub>2022</sub>	scandal	Ebola	crisis	Brexit	coronavirus	#	4/6	—
ChronoBERT <sub>2024</sub>	scandal	SARS	crisis	Brexit	coronavirus	Chat	6/6	—
ChronoBERT <sub>1999</sub> through ChronoBERT <sub>2024</sub>							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."*

Prompt year:	2001	2003	2008	2016	2020	2022	Accuracy	
							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 <sub>2002</sub>	Amendment	Ebola	rate	election	influenza	EX	0/1	0/5
ChronoBERT <sub>2004</sub>	scandal	SARS	rate	tax	influenza	SU	2/2	0/4
ChronoBERT <sub>2008</sub>	scandal	SARS	crisis	Kyoto	influenza	AR	3/3	0/3
ChronoBERT <sub>2012</sub>	scandal	SARS	crisis	presidential	influenza	AR	3/3	0/3
ChronoBERT <sub>2016</sub>	scandal	SARS	crisis	Brexit	influenza	i	4/4	0/2
ChronoBERT <sub>2020</sub>	scandal	SARS	crisis	Brexit	coronavirus	V	5/5	0/1
ChronoBERT <sub>2022</sub>	scandal	Ebola	crisis	Brexit	coronavirus	#	4/6	—
ChronoBERT <sub>2024</sub>	scandal	SARS	crisis	Brexit	coronavirus	Chat	6/6	—
ChronoBERT <sub>1999</sub> through ChronoBERT <sub>2024</sub>							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."*

Prompt year:	2001	2003	2008	2016	2020	2022	Accuracy	
							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 <sub>2002</sub>	Amendment	Ebola	rate	election	influenza	EX	0/1	0/5
ChronoBERT <sub>2004</sub>	scandal	SARS	rate	tax	influenza	SU	2/2	0/4
ChronoBERT <sub>2008</sub>	scandal	SARS	crisis	Kyoto	influenza	AR	3/3	0/3
ChronoBERT <sub>2012</sub>	scandal	SARS	crisis	presidential	influenza	AR	3/3	0/3
ChronoBERT <sub>2016</sub>	scandal	SARS	crisis	Brexit	influenza	i	4/4	0/2
ChronoBERT <sub>2020</sub>	scandal	SARS	crisis	Brexit	coronavirus	V	5/5	0/1
ChronoBERT <sub>2022</sub>	scandal	Ebola	crisis	Brexit	coronavirus	#	4/6	—
ChronoBERT <sub>2024</sub>	scandal	SARS	crisis	Brexit	coronavirus	Chat	6/6	—
ChronoBERT <sub>1999</sub> through ChronoBERT <sub>2024</sub>							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."*

Prompt year:	2001	2003	2008	2016	2020	2022	Accuracy	
							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 <sub>2002</sub>	Amendment	Ebola	rate	election	influenza	EX	0/1	0/5
ChronoBERT <sub>2004</sub>	scandal	SARS	rate	tax	influenza	SU	2/2	0/4
ChronoBERT <sub>2008</sub>	scandal	SARS	crisis	Kyoto	influenza	AR	3/3	0/3
ChronoBERT <sub>2012</sub>	scandal	SARS	crisis	presidential	influenza	AR	3/3	0/3
ChronoBERT <sub>2016</sub>	scandal	SARS	crisis	Brexit	influenza	i	4/4	0/2
ChronoBERT <sub>2020</sub>	scandal	SARS	crisis	Brexit	coronavirus	V	5/5	0/1
ChronoBERT <sub>2022</sub>	scandal	Ebola	crisis	Brexit	coronavirus	#	4/6	—
ChronoBERT <sub>2024</sub>	scandal	SARS	crisis	Brexit	coronavirus	Chat	6/6	—
ChronoBERT <sub>1999</sub> through ChronoBERT <sub>2024</sub>							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:

$$r_{i,t+1} = \alpha_m + \beta'_m e_{i,t} + \epsilon_{i,t+1}, \quad \text{for } i = 1, \dots, N \text{ and } t = 1 \dots T, \quad (1)$$

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

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

- ▶ 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 <sub>Realtime</sub>			ChronoGPT <sub>Realtime</sub>			Llama 3.1		
	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
2	-2.43	25.20	-0.10	1.48	25.48	0.06	-4.77	25.31	-0.19
3	4.17	25.64	0.16	0.61	24.87	0.02	-0.24	24.86	-0.01
4	4.17	24.58	0.17	2.41	24.26	0.10	3.84	24.62	0.16
5	3.94	24.22	0.16	7.18	24.54	0.29	7.47	24.65	0.30
6	10.81	24.13	0.45	5.78	24.20	0.24	12.03	24.23	0.50
7	14.56	24.23	0.60	10.82	23.84	0.45	13.31	24.33	0.55
8	16.38	23.64	0.69	15.74	24.30	0.65	15.13	23.79	0.64
9	23.95	24.45	0.98	22.63	24.19	0.94	24.68	23.88	1.03
High(H)	37.71	24.53	1.54	43.08	25.41	1.70	42.20	25.05	1.68
H-L	<b>61.02</b>	<b>12.72</b>	<b>4.80</b>	<b>62.90</b>	<b>12.78</b>	<b>4.92</b>	<b>65.91</b>	<b>13.46</b>	<b>4.90</b>

# ChronoBERT and ChronoGPT are better than similar-sized LLMs

## Performance of LLM Portfolios

	BERT			FinBERT			StoriesLM		
	Mean	SD	SR	Mean	SD	SR	Mean	SD	SR
Low(L)	-22.52	26.21	-0.86	-23.96	26.86	-0.89	-17.80	26.52	-0.67
2	-5.05	25.55	-0.20	-3.17	25.64	-0.12	-1.19	25.26	-0.05
3	3.12	24.92	0.13	3.36	24.83	0.14	1.86	24.92	0.07
4	8.14	24.62	0.33	7.19	24.52	0.29	5.90	24.62	0.24
5	10.81	24.44	0.44	9.17	24.39	0.38	4.99	24.30	0.21
6	9.38	24.02	0.39	11.47	24.03	0.48	11.88	23.90	0.50
7	14.54	23.83	0.61	16.54	23.92	0.69	12.41	23.66	0.52
8	18.51	24.04	0.77	19.16	23.65	0.81	18.93	24.19	0.78
9	19.68	23.90	0.82	20.70	23.88	0.87	23.25	24.30	0.96
High(H)	33.37	24.88	1.34	29.51	24.60	1.20	29.73	24.78	1.20
H-L	<b>55.89</b>	<b>13.38</b>	<b>4.18</b>	<b>53.47</b>	<b>13.85</b>	<b>3.86</b>	<b>47.53</b>	<b>13.90</b>	<b>3.42</b>

# ChronoBERT and ChronoGPT are comparable to much larger Llama

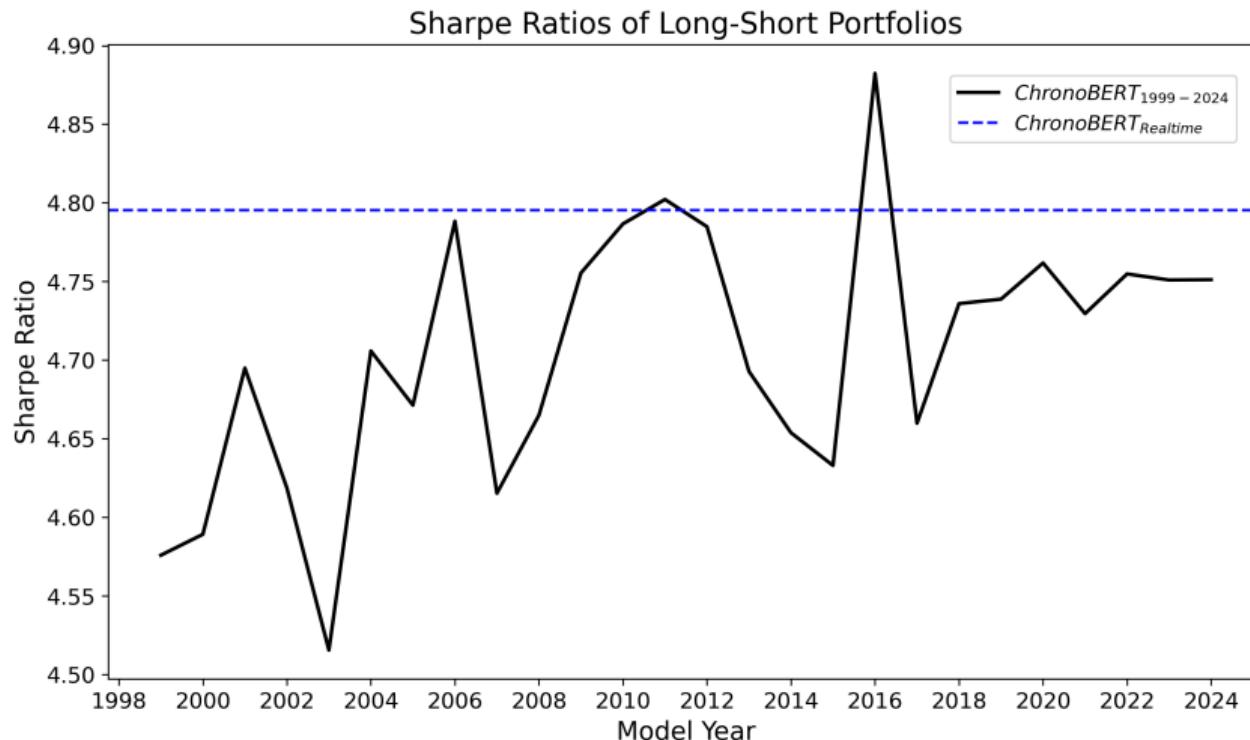
P-value of Pairwise Sharpe Ratio Difference Tests

	ChronoBERT	ChronoGPT	Llama 3.1	BERT	FinBERT	StoriesLM
ChronoBERT		0.717	0.685	0.005	0.002	0.000
ChronoGPT	0.283		0.453	0.001	0.001	0.000
Llama 3.1	0.315	0.547		0.001	0.000	0.000
BERT	0.995	0.999	0.999		0.116	0.005
FinBERT	0.998	0.999	1.000	0.884		0.098
StoriesLM	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.

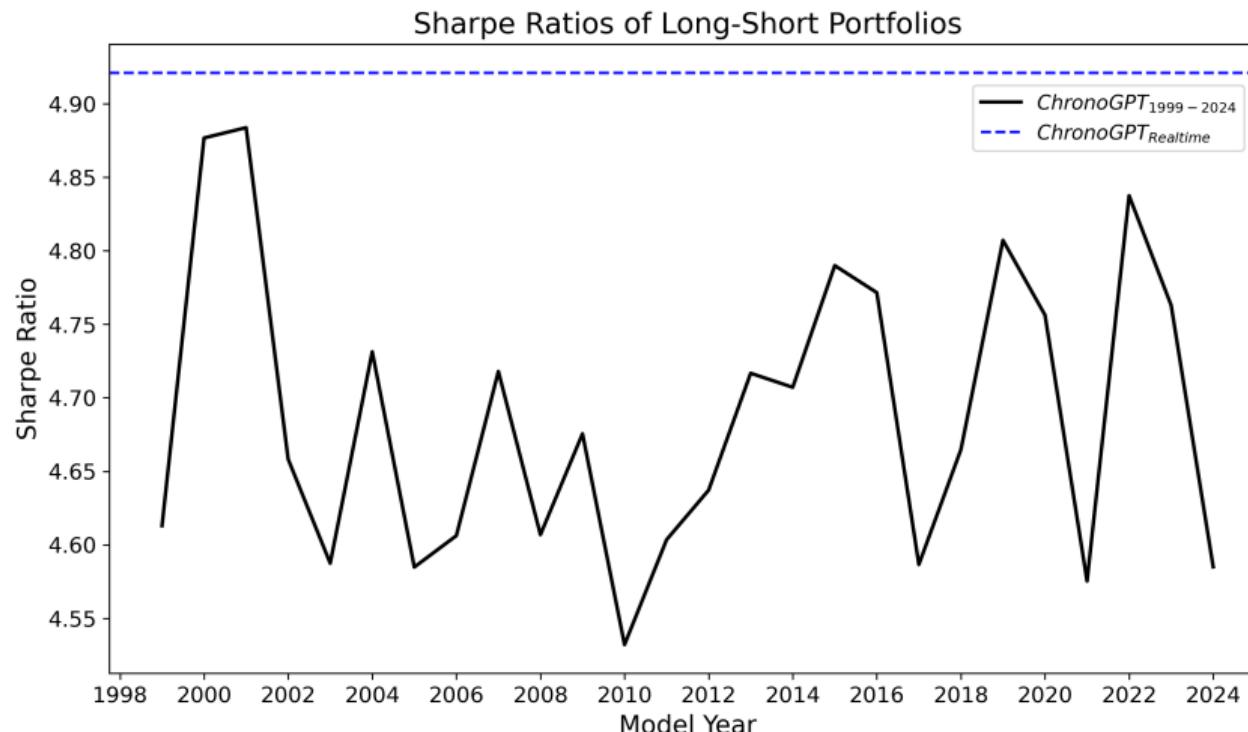
# Up-to-date knowledge improves ChronoBERT's predictions

Portfolios Performance across ChronoBERT Vintages



# Up-to-date knowledge improves ChronoGPT's predictions

Portfolios Performance across ChronoGPT Vintages

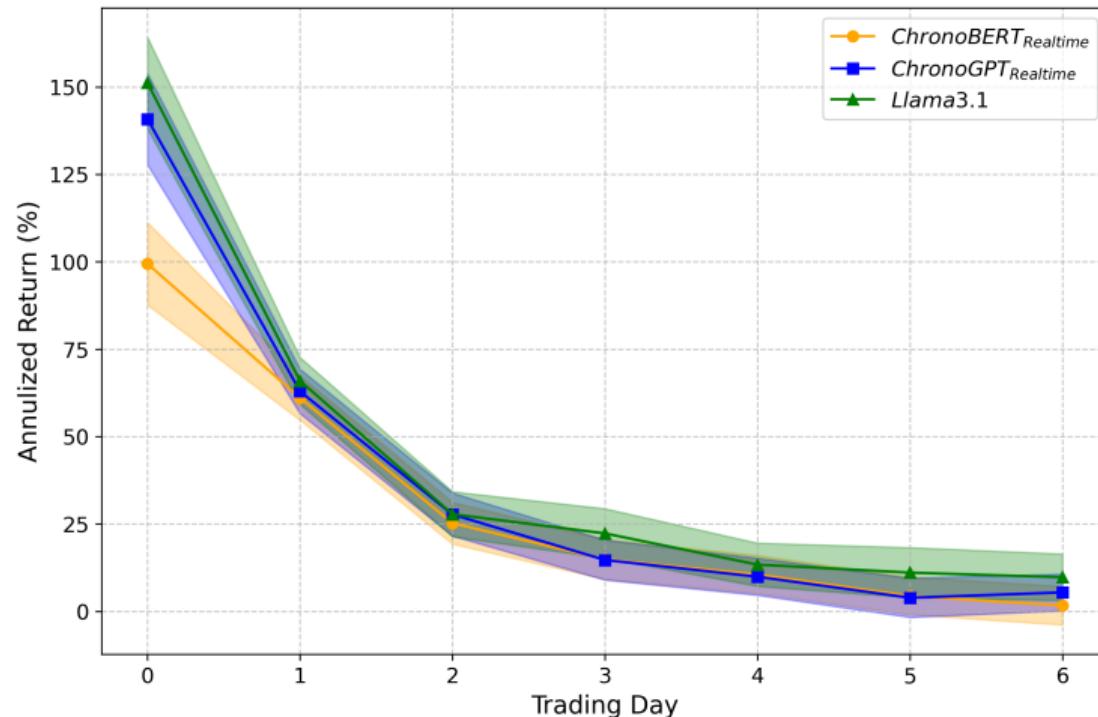


# Conclusion

- ▶ 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

# 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

