Chronologically Consistent Large Language Models

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

[1]Washington University in St. Louis

March 2025

Chronological inconsistency, lookahead bias, and training leakage

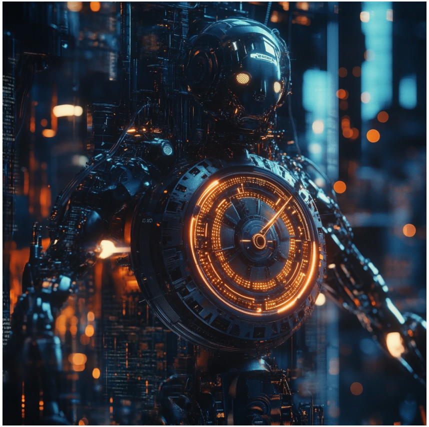
- Large language models (LLMs) now permeate social sciences
- They let us test hypotheses previously unquantifiable
- But they are trained on data that did not exist at the historical moment
- Lookahead bias (Glasserman-Lin 2023, Sarkar-Vafa 2024) and training leakage (Ludwig-Mullainathan-Rambachan 2025) 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
- Researchers could keep a held-out sample of recent data to avoid lookahead bias
- But limited panel of asset prices forces most studies to rely on backtesting
- SOTA gated models (e.g. ChatGPT) continuously fine-tuned and can search
- Chronologically inconsistent language models can bias measures of risk and market efficiency

What we do

- We train chronologically consistent LLMs exclusively on preceding text
  - $ChronoBERT_{1999}, \ldots, ChronoBERT_{2024}$
  - $ChronoGPT_{1999}, \ldots, ChronoGPT_{2024}$
  - Both model families are available to other researchers on hugging face
- Simple, right?
- Ensuring these models are competitive with SOTA 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's Sharpe ratio (4.8) is comparable to state-of-the-art (and inconsistent) Llama (4.9)
- Implies modest lookahead bias in this setting

Related work

- We develop LLMs that are both 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 show the robust news-return predictability achieved 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)

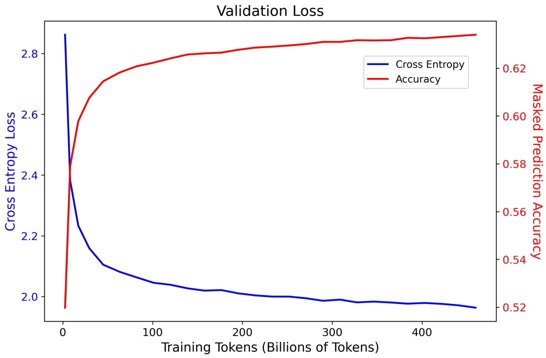
Pretraining chronologically consistent LLMs

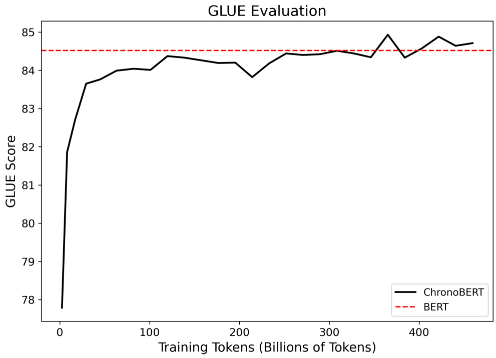
- Training LLMs is usually split into:
  1. pretraining them to predict missing words in text sequences
  2. finetuning them for specific applications (e.g. chat, Q&A, reasoning)
- We pretrain $ChronoBERT_t$ and $ChronoGPT_t$ only on text available before t
  - For example, web pages crawled in 2005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

Pretraining chronologically consistent LLMs

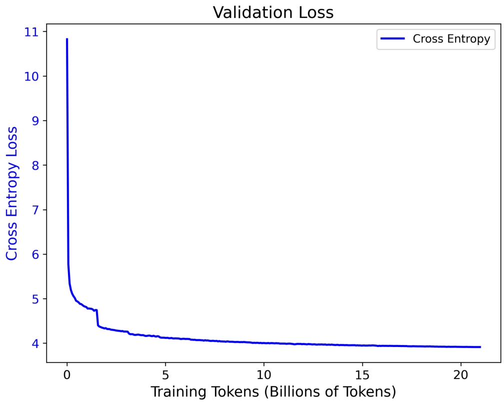
- Ensuring these models are competitive with SOTA counterparts poses two challenges:
  1. Limited compute
  2. Limited historical data
- We draw on efficient training methods (Portes et al 2023; Warner et al 2024; Jordan et al 2024) to lower computing costs
- Follow Gunasekar et al. (2023) by selecting 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

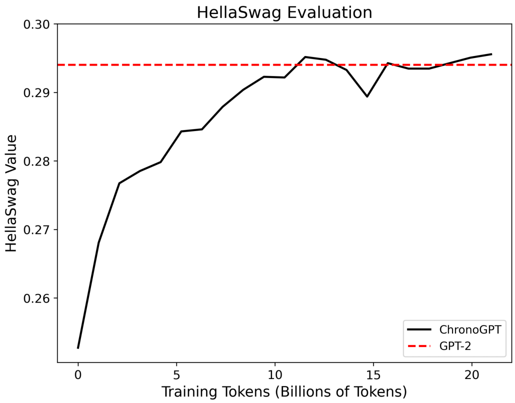
ChronoBERT$_{1999}$ improves to the point it surpasses BERT on GLUE language benchmarks





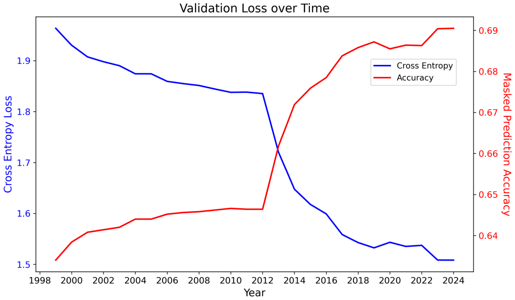
ChronoGPT$_{1999}$ improves to the point it surpass GPT-2 on HellaSwag token generation

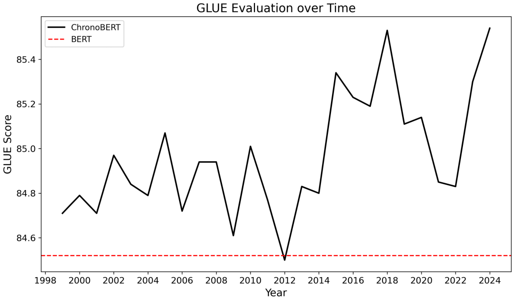




Subsequent ChronoBERT models improve further over time

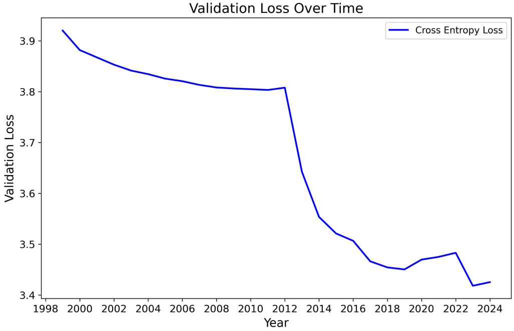
$ChronoBERT_{1999}, \ldots, ChronoBERT_{2024}$

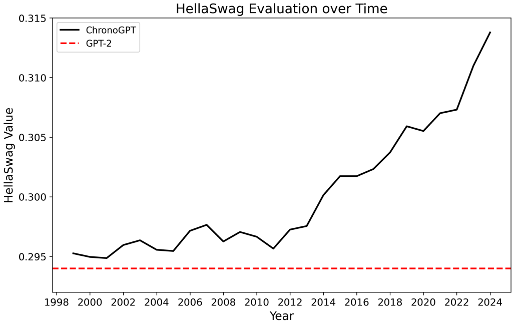




Subsequent ChronoGPT models improve further over time

$ChronoGPT_{1999}, \ldots, ChronoGPT_{2024}$





Size, context and knowledge cutoff for different LLMs

	Parameters	Context Tokens	Knowledge Cutoff
ChronoBERT_{1999}	149M	1,024	December, 1999
$\vdots$	$\vdots$	$\vdots$	$\vdots$
ChronoBERT_{2024}	149M	1,024	December, 2024
ChronoGPT_{1999}	124M	1,792	December, 1999
$\vdots$	$\vdots$	$\vdots$	$\vdots$
ChronoGPT_{2024}	124M	1,792	December, 2024
BERT	110M	512	October, 2018
FinBERT	110M	512	December, 2019
StoriesLM	110M	512	December, 1963
GPT-2	124M	1,024	Febrary, 2019
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	37.13	31.70
SST2	91.82	92.58	89.68	88.53
MRPC	92.71	92.45	82.92	85.34
STSB	89.57	89.93	81.57	82.58
QQP	88.54	88.90	82.43	83.53
MNLI	86.19	86.89	77.63	79.15
QNLI	90.61	92.04	84.94	85.98
RTE	80.94	85.20	67.08	67.80
GLUE	84.71	85.54	75.42	75.58

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

- To detect leakage in the textual data used to pretrain our models, we evaluate them on events occurring after the model's knowledge cutoff
- Since ChronoBERT is a fill-mask model, we use each model vintage to predict the masked token in:

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

ChronoBERT knows only what it should

Predictions of U.S. Presidents

Prompt year:	1992	2000	2008	2016	2020	2024
BERT	**Clinton**	Clinton	**Obama**	Obama	Obama	Obama
ChronoBERT_{2000}	**Clinton**	Clinton	Clinton	Clinton	Clinton	Wilson
ChronoBERT_{2004}	**Clinton**	**Bush**	Bush	Clinton	Bush	Clinton
ChronoBERT_{2008}	**Clinton**	**Bush**	Bush	Obama	Bush	Wilson
ChronoBERT_{2012}	Obama	Obama	**Obama**	Obama	Obama	Obama
ChronoBERT_{2016}	**Clinton**	**Bush**	**Obama**	Obama	Obama	Obama
ChronoBERT_{2020}	**Clinton**	**Bush**	**Obama**	**Trump**	Trump	**Trump**
ChronoBERT_{2024}	**Clinton**	**Bush**	**Obama**	**Trump**	**Biden**	Biden

blue = correct prediction

gray area = post-knowledge cutoff prediction

News data and methods

- Dow Jones Newswire data from 1997 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 is comparable to state-of-the-art (and inconsistent) Llama

Performance of LLM Portfolios

	$\text{ChronoBERT}_{\text{Realtime}}$			$\text{ChronoGPT}_{\text{Realtime}}$			Llama 3.1		
	Mean	SD	SR	Mean	SD	SR	Mean	SD	SR
Low(L)	-23.30	25.86	-0.90	-20.03	25.96	-0.77	-23.71	26.15	-0.91
2	-2.43	25.20	-0.10	0.06	25.65	0.00	-4.77	25.31	-0.19
3	4.17	25.64	0.16	2.96	25.03	0.12	-0.24	24.86	-0.01
4	4.17	24.58	0.17	5.59	24.75	0.23	3.84	24.62	0.16
5	3.94	24.22	0.16	6.67	24.36	0.27	7.47	24.65	0.30
6	10.81	24.13	0.45	5.91	23.91	0.25	12.03	24.23	0.50
7	14.56	24.23	0.60	13.51	24.09	0.56	13.31	24.33	0.55
8	16.38	23.64	0.69	16.63	23.77	0.70	15.13	23.79	0.64
9	23.95	24.45	0.98	21.56	24.07	0.90	24.68	23.88	1.03
High(H)	37.71	24.53	1.54	37.13	24.59	1.51	42.20	25.05	1.68
H-L	61.02	12.72	4.80	57.16	12.75	4.48	65.91	13.46	4.90

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	55.89	13.38	4.18	53.47	13.85	3.86	47.53	13.90	3.42

ChronoBERT is comparable to state-of-the-art (and inconsistent) Llama

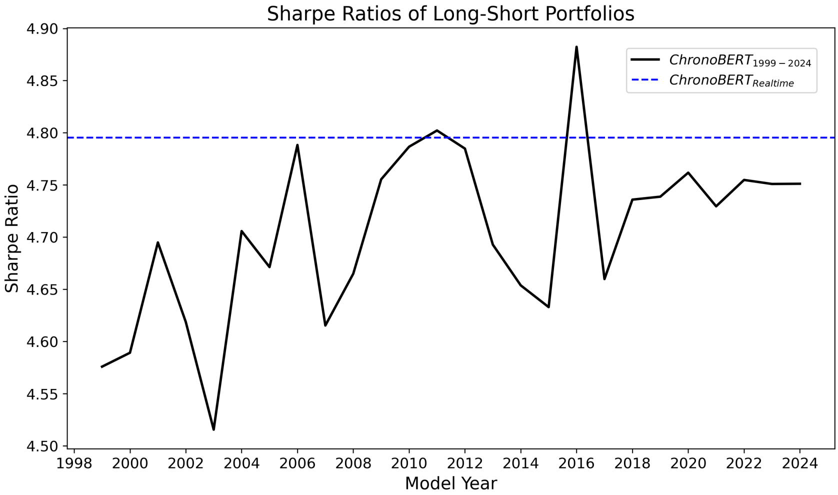
P-value of Pairwise Sharpe Ratio Difference Tests

	ChronoBERT	ChronoGPT	Llama 3.1	BERT	FinBERT	StoriesLM
ChronoBERT		0.076	0.685	0.005	0.002	0.000
ChronoGPT	0.924		0.973	0.078	0.017	0.001
Llama 3.1	0.315	0.027		0.001	0.000	0.000
BERT	0.995	0.922	0.999		0.116	0.005
FinBERT	0.998	0.983	1.000	0.884		0.098
StoriesLM	1.000	0.999	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.

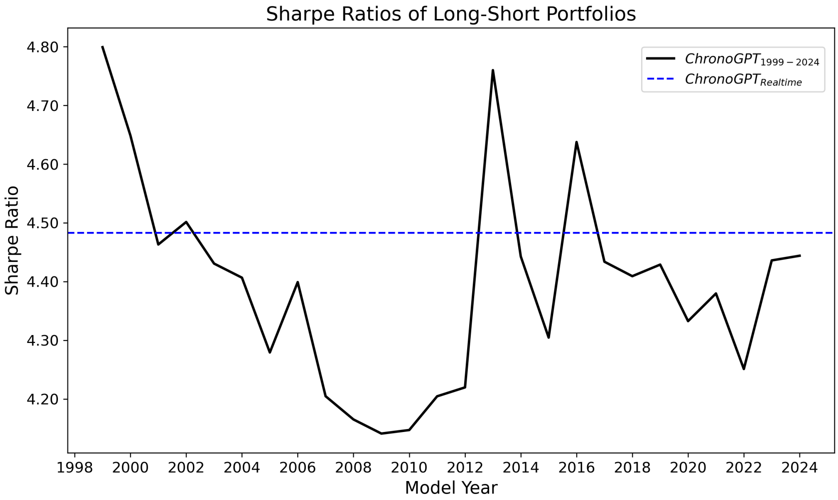
Up-to-date knowledge improves ChronoBERT's predictions

Portfolios Performance across ChronoBERT Vintages



Up-to-date knowledge has mixed effect on 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 (4.8) is comparable to state-of-the-art (and inconsistent) Llama (4.9)
- Find modest lookahead bias in this setting
- Our models are available on hugging face for researchers to evaluate bias in their applications