

Chronologically Consistent Large Language Models*

Songrun He Linying Lv Asaf Manela Jimmy Wu

This draft: February 2025.

Abstract

Large language models are increasingly used in social sciences, but their training data can introduce lookahead bias and training leakage. A good chronologically consistent language model requires efficient use of training data to maintain accuracy despite time-restricted data. Here, we overcome this challenge by training chronologically consistent large language models timestamped with the availability date of their training data, yet accurate enough that their performance is comparable to state-of-the-art open-weight models. Lookahead bias is model and application-specific because even if a chronologically consistent language model has poorer language comprehension, a regression or prediction model applied on top of the language model can compensate. In an asset pricing application, we compare the performance of news-based portfolio strategies that rely on chronologically consistent versus biased language models and estimate a modest lookahead bias.

JEL Classification: G11, G12, G17

Keywords: Large language model, chronological consistency, lookahead bias, training leakage, backtesting

*Songrun He is at Washington University in St. Louis (h.songrun@wustl.edu). Linying Lv is at Washington University in St. Louis (lyu@wustl.edu). Asaf Manela is at Washington University in St. Louis (amanela@wustl.edu). Jimmy Wu is at Washington University in St. Louis (jimmywu@wustl.edu).

“Obviously, the time continuum has been disrupted, creating a new temporal event sequence resulting in this alternate reality.”

– Dr. Brown, *Back to the Future Part II*

1 Introduction

The increasing adoption of large language models (LLMs) in economics and finance has opened up new possibilities for analyzing unstructured textual data. By capturing nuanced language patterns, these models provide powerful tools for testing flexible and previously unquantifiable hypotheses in financial economics ([Hoberg and Manela, 2025](#)). However, their reliance on historical textual training data presents significant challenges: lookahead bias and training leakage, which undermine the validity of conclusions from any predictive analysis that should only use realtime available information from the past ([Glasserman and Lin, 2023](#); [Sarkar and Vafa, 2024](#); [Ludwig et al., 2025](#)).

We address this challenge by training chronologically consistent LLMs trained exclusively on historical textual data available at the time. While training language models with historical data timestamped at the point of its availability is conceptually straightforward, ensuring these models are competitive with state-of-the-art counterparts remains a significant challenge.

A major constraint in this approach is twofold: the significant computational resources required to pretrain a high-quality model and the limited availability of historical data. To overcome the computational challenge, we employ recent innovations in efficient pretraining algorithms from [Warner et al. \(2024\)](#) and [Jordan et al. \(2024\)](#). Simultaneously, leveraging the insights from [Gunasekar et al. \(2023\)](#), we compiled a pretraining corpus of diverse and high-quality textual content to mitigate the data limitation. By combining these strategies, we scale up our model in this resource-constrained environment to

ultimately reach strong language understanding performance.

Using this approach, we introduce ChronoBERT, a series of chronologically consistent language models pretrained on timestamped text data. Through evaluation on the GLUE benchmark (Wang et al., 2019), even the earliest available version of ChronoBERT significantly outperforms prior lookahead-bias-free models, such as StoriesLM (Sarkar and Vafa, 2024), and domain-specific finance models, such as FinBERT (Huang et al., 2023), in language understanding tasks. Furthermore, ChronoBERT achieves GLUE performance comparable to or even better than the widely used BERT model (Devlin et al., 2019), which ranks first in downloads among all language models on Hugging Face as of February 2025.¹ This positions ChronoBERT as the strongest language model to date that is free from lookahead bias.

Finance is probably the research field most concerned with lookahead bias, and for good reason. Tests of market efficiency (Fama, 1970) are only valid if asset prices are compared to information available at the time. It would not be surprising if one could predict returns after traveling to the future and bringing back a record of future newspaper coverage. While researchers could keep a held-out sample of recent data to avoid lookahead bias, the limited panel of asset prices commonly studied forces most studies to rely on backtesting. Hence a chronologically inconsistent language model can bias measures of risk and market inefficiency.

Equipped with our chronologically consistent models, we quantify the impact of lookahead bias when using LLMs to predict stock returns based on financial news. Leveraging extensive financial newswire data, we find that the portfolio performance of ChronoBERT matches that of the state-of-the-art Llama 3.1 (Dubey et al., 2024), with both models delivering economically substantial and statistically significant gains compared

¹Based on download statistics from Hugging Face at <https://huggingface.co/models?sort=downloads>. Its relatively small size makes it a popular choice for embeddings tasks and for fine-tuning models and applying them to large data pipelines where using the largest models (e.g. Llama) is computationally infeasible.

to StoriesLM ([Sakar, 2024](#)) and FinBERT ([Huang et al., 2023](#)). Our findings suggest that lookahead bias in this context is relatively modest.

An important observation is that the impact of lookahead bias is model- and application-specific. While ChronoBERT may exhibit lower language comprehension on general tasks compared to unconstrained models, downstream predictive models built on ChronoBERT can adapt to these limitations, mitigating potential drawbacks in financial forecasting.

Overall, our contribution is threefold. First, we quantify the extent of lookahead bias in news return prediction tasks, demonstrating that its effects are modest. Second, we propose a framework for training chronologically consistent LLMs that preserves the integrity of economic and financial analyses while achieving competitive performance. Third, we provide a constructive solution to address concerns of lookahead bias, not only in asset pricing but also across broader applications in the social sciences. By demonstrating that chronological consistency need not come at the expense of performance, our findings pave the way for more robust and reliable applications of LLMs in social science.

Our work is related to two broad strands of literature. There is a large literature looking into return predictability using financial news. Starting from [Tetlock et al. \(2008\)](#), [Jiang et al. \(2021\)](#), and [Ke et al. \(2019\)](#), researchers document there is a short-term underreaction of stock returns to textual information in financial news. The NLP methodology used is mainly a word-counting approach and the economic magnitude is smaller than the results using state-of-the-art LLMs.

The advent of large language models has further advanced this field. [Chen et al. \(2023\)](#) show that news embeddings derived from LLMs can effectively predict next-day returns, leading to strong portfolio performance. Likewise, [Lopez-Lira and Tang \(2023\)](#) demonstrate that interacting with LLMs via prompts can also generate promising portfolio results.

Our contribution to this literature lies in demonstrating that the robust return pre-

dictability achieved through LLMs is not driven by lookahead bias, thereby addressing a critical concern in the use of these advanced models for financial forecasting.

A second strand of literature pertains to the application and development of natural language processing (NLP) tools for financial economics research (Hoberg and Manela, 2025). Methodological advancements in this area have evolved from early dictionary-based approaches (Tetlock, 2007; Loughran and McDonald, 2011), to text regressions (Manela and Moreira, 2017; Kelly et al., 2021), to topic modeling (Bybee et al., 2024), and most recently, to the integration of LLMs (Jha et al., 2025; Chen et al., 2023; Lv, 2024). These developments underscore the growing demand for more advanced and scalable tools to answer research questions in finance and economics.

We contribute to this literature by developing a framework for pretraining high-quality lookahead-bias-free LLMs and providing a constructive solution to address the concerns of lookahead bias. Our approach does not require masking (Glasserman and Lin, 2023), which can destroy information. We train a set of chronologically consistent models with knowledge cutoffs from 1999 to 2024 which offer superior language understanding and more up-to-date knowledge to previous such models (e.g., Sakar, 2024).

The rest of the paper is organized as follows: Section 2 describes our approach and data for model pretraining and evaluation. Section 3 presents the empirical performance of our model. Section 4 concludes.

2 Methodology and Data

In this section, we outline the methodology we use to pretrain ChronoBERT and describe the approach for evaluating its performance. Specifically, we assess its ability in both language understanding and asset pricing tasks.

2.1 Pretraining Methodology for ChronoBERT

When pretraining ChronoBERT, we incorporate a state-of-the-art BERT architecture from [Portes et al. \(2023\)](#) and [Warner et al. \(2024\)](#).² Compared to the original BERT model by [Devlin et al. \(2019\)](#), this enhanced architecture integrates recent advancements in rotary positional embeddings that support longer context lengths and employ flash attention, significantly improving pretraining efficiency and computational speed.

For the pretraining task, we follow [Warner et al. \(2024\)](#) by adopting masked token prediction while omitting the next sequence prediction task, as prior research has shown the latter increases training overhead without meaningful performance gains.

The quality of pretraining data is critical to achieving BERT-level performance. [Gunasekar et al. \(2023\)](#) demonstrate that using high-quality, “textbook-like” data leads to faster convergence and improved model outcomes. Motivated by this insight, we filter our pretraining corpus using the FineWeb-edu classifier from [Penedo et al. \(2024\)](#), retaining only texts with scores above two.³

However, restricting the corpus to texts with historical dates—particularly from early historical periods—introduces data scarcity challenges. [Muennighoff et al. \(2023\)](#) explore the scaling laws of LLMs under data constraints, highlighting the benefits of iterative training on limited high-quality data. Following their insights, we train our model over multiple epochs to maximize learning from the available corpus. Our first model checkpoint ChronoBERT₁₉₉₉ is trained on 460 billion tokens, with more than 70 epochs through the dataset.

²We thank the authors for providing their pretraining code at <https://github.com/mosaicml/examples/tree/main/examples/benchmarks/bert> and <https://github.com/AnswerDotAI/ModernBERT>.

³We thank the authors for providing the classifier at <https://huggingface.co/HuggingFaceFW/fineweb-edu-classifier>.

2.2 Evaluation Methodology

To evaluate ChronoBERT’s performance, we assess both its language understanding capabilities and economic forecasting performance. We also apply the same evaluation methodology to several other LLMs for benchmarking.

2.2.1 GLUE Evaluation for Language Understanding

The GLUE evaluation framework, introduced by Wang et al. (2019), comprises multiple classification tasks designed to measure a model’s language understanding.⁴ This framework was also the primary evaluation metric used in Devlin et al. (2019) to assess BERT’s language capabilities.

Following pretraining, we further fine-tune the model on task-specific training datasets and evaluate its performance on a held-out test set.

For fine-tuning, we adopt the training specifications and hyperparameters outlined in Warner et al. (2024). Among the eight GLUE tasks, RTE, MRPC, and STS-B are initialized from the MNLI checkpoint to enhance performance.

2.2.2 Predicting Stock Returns using Financial News

We investigate whether improved language understanding translates into economic gains by using different language models to predict stock returns from economic news. Based on these predictions, we construct portfolios and evaluate the performance of long-short strategies.

Following Chen et al. (2023), we first aggregate all news articles’ headlines for a stock on a given trading day together. Next, we transform this text into embeddings. Specifically,

⁴The details and leaderboard of GLUE evaluation can be found at <https://gluebenchmark.com/>. In Appendix A, we outline details of eight GLUE tasks.

we process each piece of text through different language models and extract the hidden states of all tokens. The final embedding for each text is obtained by averaging the token embeddings.⁵

Next, we link each news article to stock returns on the following trading day and fit a Fama-MacBeth regression with a ridge penalty to map news embeddings to return predictions. Each month m , we estimate the following cross-sectional ridge regression:

$$r_{i,t+1} = \alpha_m + \beta'_m e_{i,t} + \varepsilon_{i,t+1}, \quad \text{for } i = 1, \dots, N \text{ and } t = 1 \dots T,^6 \quad (1)$$

where $e_{i,t}$ represents of the embedding of all news for firm i on day t . To construct real-time out-of-sample forecasts, in month m' , we use an average of forecasts over all previous months' cross-sectional models:

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

where $\bar{\alpha}_{m'} = \frac{1}{m'-1} \sum_{m=1}^{m'-1} \hat{\alpha}_m$ and $\bar{\beta}_{m'} = \frac{1}{m'-1} \sum_{m=1}^{m'-1} \hat{\beta}_m$.

Using these out-of-sample predictions, we sort stocks into decile portfolios at the end of each trading day, based on forecasts from different language models. We then evaluate the performance of daily-rebalanced long-short decile portfolios constructed from these predictions.

2.3 Data

In this section, we present the data we used for pretraining our language models and the financial newswire data we used to evaluate our language models.

⁵Coleman (2020) provides a rationale for averaging token embeddings to get sequence embeddings.

⁶We use a leave-one-out cross-validation algorithm to determine the ridge penalty λ chosen from grid points ranging from 10^{-10} to 10^{10} .

2.3.1 Pretraining Data for ChronoBERT

ChronoBERT₁₉₉₉ is pretrained on a corpus of 460 billion tokens comprising chronologically structured English text sourced from diverse domains, including historical web content, archived news articles, and scientific publications. The dataset is fully open-source and is carefully curated to include only text published before the year 2000, ensuring a focus on no leakage of lookahead knowledge. The final composition of our pretraining corpus was determined through extensive ablation studies to optimize model performance.

We further conduct incremental training from 2000 to 2024 on a corpus of 65 billion tokens with similar high-quality, diverse, timestamped, and open-source textual data to update knowledge for the model.

2.3.2 Financial Newswire Data

We utilize the Dow Jones Newswire dataset, a real-time newswire subscribed to by major institutional investors. This dataset provides extensive coverage of financial markets, economies, and companies, aggregating reports from leading media sources such as Wall Street Journal, Barron’s, and MarketWatch.

The dataset includes news headlines, full article texts, and precise display timestamps, with microsecond-level accuracy for when the news becomes available to clients. Following [Ke et al. \(2019\)](#), we focus on firm-specific news that can be attributed to a single company. For each firm-day observation, we aggregate all news headlines related to the firm within the trading day window – spanning from 4:00 p.m. EST on day $t - 1$ to 4:00 p.m. EST on day t – and treat the combined text as the firm’s textual information. Each concatenated set of headlines is then processed through our embedding framework to generate numerical text representations.

After the embedding step, we merge the news dataset with close-to-close returns on

trading day $t + 1$ from CRSP together to examine the predictability of stock returns using LLMs trained with real-time available textual data.

Our dataset covers the period from January 2007 to July 2023. The first year serves as a burn-in period to estimate the initial return prediction model, resulting in a final asset pricing test sample spanning January 2008 to July 2023.

3 Empirical Performance

In this section, we first describe the pretraining process of ChronoBERT, focusing on validation loss and GLUE scores as the model is trained on an increasing number of tokens. We then benchmark ChronoBERT’s language understanding capabilities against three other language models. Finally, we assess its asset pricing performance in the news return prediction exercise.

Firstly on pretraining performance, our results confirm the scaling law proposed by Muennighoff et al. (2023) in a data-limited environment. As shown in Figure 1, validation loss (measured via cross-entropy) decreases consistently as the number of training tokens increases.⁷ Simultaneously, masked language prediction accuracy improves with training progress.

These improvements translate into enhanced language understanding, as reflected in the GLUE scores. ChronoBERT begins outperforming BERT after approximately 350 billion training tokens and continues to improve thereafter.

Starting from this high-quality base model in the early period, we continue incremental pretraining using textual data afterward. We create model checkpoints for each year from 1999 to 2024 (26 models in total). Figure 2 presents our models’ validation loss and GLUE

⁷We use a subset of the C4 corpus from <https://huggingface.co/datasets/allenai/c4> as the validation set. The C4 data is a cleaned version of the Common Crawl data.

scores as we train with incremental textual data over time.

We find that with the introduction of new data, the validation loss continues to drop. Starting in the year 2013, we introduce high-quality common crawl data. We witness a significant decrease in validation loss with the increase in data diversity. In the right panel of Figure 2, the GLUE scores for nearly all models exceed that of BERT, highlighting their superior language understanding and overall quality.

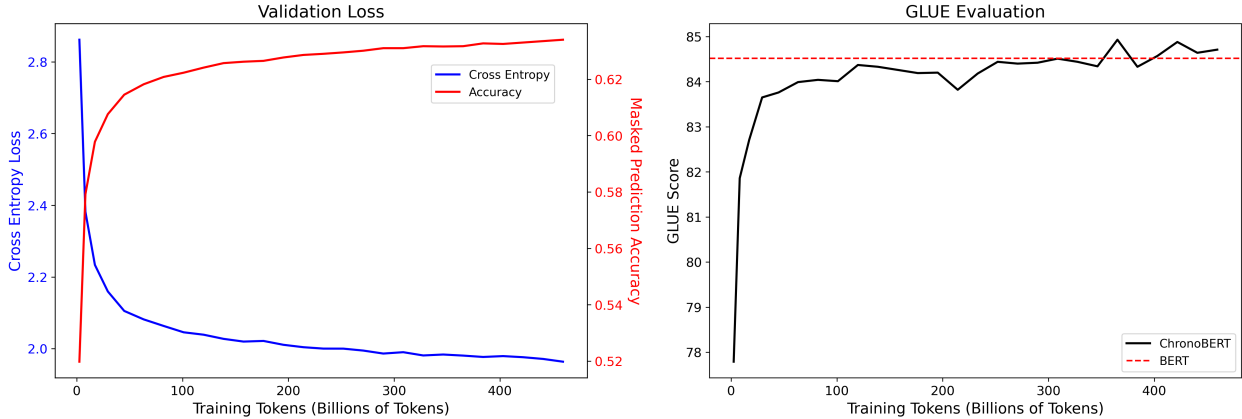


Figure 1 Validation Loss and GLUE Scores versus Training Steps

The left panel shows the validation loss, measured using cross-entropy loss and masked language prediction accuracy, as the ChronoBERT₁₉₉₉ is trained on an increasing number of tokens. The right panel displays the GLUE scores as training progresses. The final model checkpoint is trained on 460 billion tokens. The training corpus consists of text up to December 1999.

We further compared ChronoBERT’s language understanding against three other models. Table 1 summarizes key characteristics of these models, including parameter counts, context lengths, and knowledge cutoffs.

The models evaluated include:

ChronoBERT₇: Our initial BERT-based model, ChronoBERT₁₉₉₉ is pretrained on 460 billion tokens of pre-2000, diverse, high-quality, and open-source text data to ensure no leakage of data afterward. Then, each year t starting in 2000, we start from the model trained in the previous year and continue training it on data available in year t . Our final checkpoint of the ChronoBERT series, ChronoBERT₂₀₂₄, is pretrained on 525 billion tokens

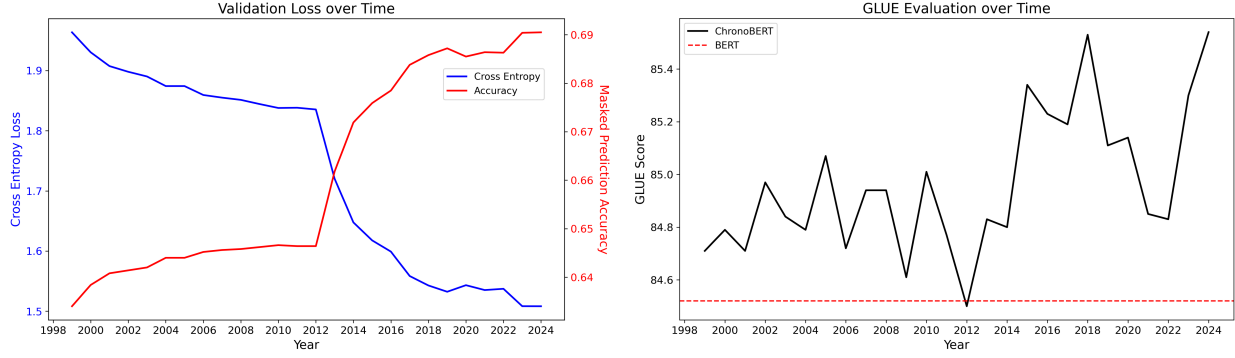


Figure 2 Validation Loss and GLUE Scores over Time

The left panel shows the validation loss, measured using cross-entropy loss and masked language prediction accuracy as ChronoBERT is trained over time. The right panel displays the GLUE scores as training progresses over time.

	Parameters	Context	Knowledge Cutoff
ChronoBERT ₁₉₉₉	149M	1024	December, 1999
⋮	⋮	⋮	⋮
ChronoBERT ₂₀₂₄	149M	1024	December, 2024
BERT	110M	512	October, 2018
FinBERT	110M	512	December, 2019
StoriesLM	110M	512	December, 1963

Table 1 Characteristics and Knowledge Cutoffs of Different LLMs

This table provides an overview of four pretrained large language models, including their number of parameters, maximum context length, and knowledge cutoff dates.

of diverse, high-quality, and open-source text until December 2024.

BERT: The original BERT model trained on Wikipedia and the BookCorpus dataset by [Devlin et al. \(2019\)](#).⁸

FinBERT: A domain-specific model pretrained on financial texts, including regulatory filings, analyst reports, and earnings call transcripts, from [Huang et al. \(2023\)](#).⁹

StoriesLM: A pretrained model from [Sarkar and Vafa \(2024\)](#), trained on historical

⁸Model downloaded from <https://huggingface.co/google-bert/bert-base-uncased>.

⁹Model downloaded from <https://huggingface.co/yiyanghkust/finbert-pretrain>.

news articles. We use the final version trained on data up to 1963.¹⁰

	ChronoBERT ₁₉₉₉	ChronoBERT ₂₀₂₄	BERT	StoriesLM	FinBERT
COLA	57.32	56.32	57.59	46.85	28.99
SST2	91.82	92.58	92.62	90.44	89.03
MRPC	92.71	92.45	90.76	89.33	88.59
STSB	89.57	89.93	90.07	87.01	85.72
QQP	88.54	88.90	88.21	86.88	86.60
MNLI	86.19	86.89	84.98	79.78	79.23
QNLI	90.61	92.04	91.52	87.44	86.12
RTE	80.94	85.20	80.43	67.15	67.00
GLUE	84.71	85.54	84.52	79.36	76.41

Table 2 GLUE Score Evaluations for Different LLMs

This table compares the GLUE benchmark scores of four language models. Tasks are grouped into three categories: (1) Single-sentence classification (COLA, SST2), (2) Paraphrase/semantic similarity (MRPC, STSB, QQP), and (3) Natural language inference (MNLI, QNLI, RTE). The final row shows the average GLUE score across all tasks.

Table 2 displays the GLUE scores for the five different models. ChronoBERT₂₀₂₄ and ChronoBERT₁₉₉₉ achieves the top 2 performance in overall GLUE score (85.54 and 84.71), surpassing performance from BERT (84.52) while substantially outperforming StoriesLM (79.36) and FinBERT (76.41). Notably, both ChronoBERT models Pareto dominate the last two models—they outperform StoriesLM (designed to avoid lookahead bias) and FinBERT (domain-specific) across all individual tasks, with particularly large margins on COLA, RTE, and MNLI. This performance advantage persists even in BERT-competitive tasks like MRPC and QQP.

These results highlight that ChronoBERT has better language understanding compared to previous models. The performance gap between ChronoBERT and StoriesLM likely stems from differences in training scale (460B versus 19B tokens) and the quality and diversity of the training data (ChronoBERT’s high-quality corpus versus StoriesLM’s unfiltered news-only dataset). Similarly, ChronoBERT’s significant edge over FinBERT

¹⁰Model downloaded from <https://huggingface.co/StoriesLM/StoriesLM-v1-1963>.

highlights the importance of diverse and high-quality pretraining data, as FinBERT’s domain-specific financial texts lack comprehensive quality checks.

Importantly, ChronoBERT also matches or exceeds the performance of BERT despite its chronologically bounded pretraining data, demonstrating that high-quality LLMs can be constructed without chronological leakage. This makes ChronoBERT particularly valuable for applications requiring good language understanding and no lookahead bias. Its ability to achieve BERT-level performance while maintaining chronological integrity establishes it as a strong candidate for a wide range of downstream tasks.

To quantify the economic gains from enhanced language understanding, we analyze stock return predictions using news embeddings from different language models. We construct portfolios by sorting stocks based on each model’s return forecasts and evaluate the performance of long-short spreads generated by these rankings.¹¹

Table 3 presents the decile portfolio performance for realtime ChronoBERT model (ChronoBERT_{Realtime})¹² against three other benchmarks: (1) BERT; (2) StoriesLM; and (3) Llama-3.1-8B.¹³ In this news return prediction setting, the H-L portfolio from ChronoBERT_{Realtime} generates a Sharpe ratio of 4.80, outperforming both StoriesLM and BERT. These results demonstrate that increased language understanding indeed translates into significant economic gains. We also report the p-value of the pairwise Sharpe ratio difference test using the [Ledoit and Wolf \(2008\)](#) approach in Table 4. The SR difference tests against both models are significant at the 1% level.

Comparing ChronoBERT against the StoriesLM, we find the models’ improved lan-

¹¹Following [He et al. \(2024\)](#), we conduct a robustness check by forecasting the probability of a positive stock return on the subsequent trading day. The outcomes closely mirror those obtained from the return forecasts.

¹²For example, in the year t , we would use the latest available model checkpoint ChronoBERT _{$t-1$} to embed the news articles in that year and make predictions based on the embeddings.

¹³We also evaluated FinBERT ([Huang et al., 2023](#)). Although not shown in the tables, our results indicate that FinBERT generated a H-L portfolio Sharpe ratio of 3.86 over the same sample period. Notably, ChronoBERT_{Realtime} H-L portfolio achieved a Sharpe ratio that is significantly larger than that of FinBERT at the 1% level.

	ChronoBERT _{Realtime}			BERT		
	Mean	SD	SR	Mean	SD	SR
Low(L)	-23.30	25.86	-0.90	-22.52	26.21	-0.86
2	-2.43	25.20	-0.10	-5.05	25.55	-0.20
3	4.17	25.64	0.16	3.12	24.92	0.13
4	4.17	24.58	0.17	8.14	24.62	0.33
5	3.94	24.22	0.16	10.81	24.44	0.44
6	10.81	24.13	0.45	9.38	24.02	0.39
7	14.56	24.23	0.60	14.54	23.83	0.61
8	16.38	23.64	0.69	18.51	24.04	0.77
9	23.95	24.45	0.98	19.68	23.90	0.82
High(H)	37.71	24.53	1.54	33.37	24.88	1.34
H-L	61.02	12.72	4.80	55.89	13.38	4.18

	StoriesLM			Llama 3.1		
	Mean	SD	SR	Mean	SD	SR
Low(L)	-17.80	26.52	-0.67	-23.71	26.15	-0.91
2	-1.19	25.26	-0.05	-4.77	25.31	-0.19
3	1.86	24.92	0.07	-0.24	24.86	-0.01
4	5.90	24.62	0.24	3.84	24.62	0.16
5	4.99	24.30	0.21	7.47	24.65	0.30
6	11.88	23.90	0.50	12.03	24.23	0.50
7	12.41	23.66	0.52	13.31	24.33	0.55
8	18.93	24.19	0.78	15.13	23.79	0.64
9	23.25	24.30	0.96	24.68	23.88	1.03
High(H)	29.73	24.78	1.20	42.20	25.05	1.68
H-L	47.53	13.90	3.42	65.91	13.46	4.90

Table 3 Performance of the LLM Portfolios

This table presents annualized performance metrics (mean return, standard deviation, and Sharpe ratio) for decile portfolios sorted by next-day return predictions from financial news. Portfolios are rebalanced daily, with the "H-L" row representing a strategy of longing the top decile and shorting the bottom decile. All values are in percentage points except Sharpe ratios. All portfolios are equal-weighted. Data spans January 2008–July 2023.

guage understanding increases the investment Sharpe Ratio by more than 40%, a value that is both economically meaningful and statistically significant. Comparing the investment performance using ChronoBERT against the state-of-the-art Llama 3.1 model, we find both models generate comparable Sharpe ratios (4.80 vs 4.90). This suggests

	BERT	StoriesLM	Llama 3.1
ChronoBERT _{Realtime}	0.009	0.000	0.630
BERT		0.009	0.002
StoriesLM			0.000

Table 4 P-value of Pairwise Sharpe Ratio Difference Tests

This table reports the p-value from the [Ledoit and Wolf \(2008\)](#) Sharpe ratio difference test of the ‘H-L’ portfolios from different LLMs in Table 3. Each entry presents the test of the model reported in the row versus the model in the column. The portfolio sample spans from January 2008 to July 2023.

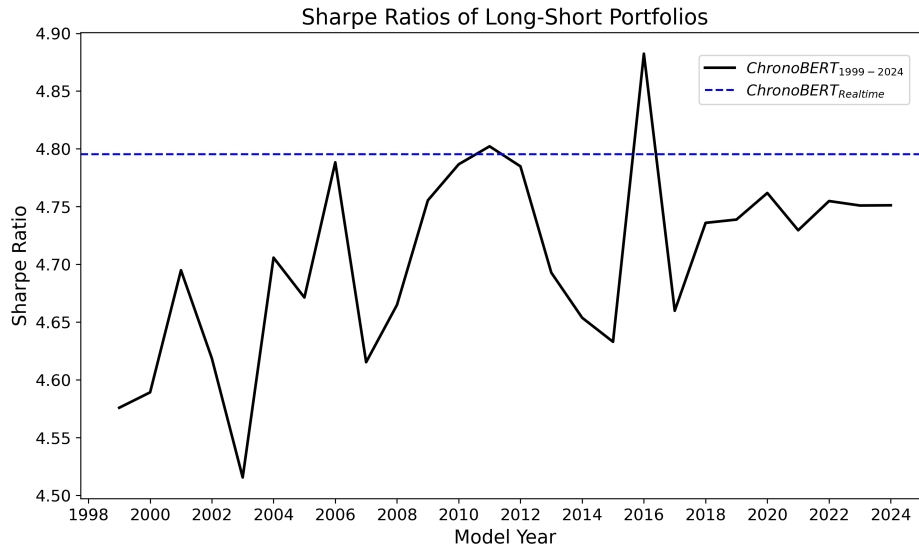


Figure 3 Sharpe Ratios of Long-Short Portfolios across Models over Time

This figure illustrates the Sharpe ratios of long-short portfolios constructed using predictions derived from financial news, with language models pretrained on text data up to the time points indicated on the x-axis. The blue dashed line represents the performance of the ChronoBERT_{Realtime} model.

ChronoBERT’s performance is comparable to state-of-the-art LLMs in this financial application, despite its constrained training using historical textual data.

Figure 3 further presents the trading performance of the whole series of ChronoBERT_{Realtime} models. Specifically, for each time t in the figure, we use the ChronoBERT _{t} model to embed news articles and run predictions using embeddings from the model. We find consistent performance across all models in the series. The results again highlight (1) the return

prediction exercise has modest lookahead bias; (2) the enhanced language understanding shown in Figure 2 indeed translates into significant economic gains.

4 Conclusion

In this paper, we address the critical challenge of lookahead and survivorship bias in LLMs used in social science applications, particularly in financial forecasting. By introducing ChronoBERT, a chronologically consistent language model trained on timestamped text data, we demonstrate that chronological consistency can be achieved without compromising performance. Our results show that ChronoBERT matches or surpasses the language comprehension abilities of BERT while maintaining robustness in asset pricing applications.

Our findings reveal that the impact of lookahead bias in return prediction tasks is modest. We also highlight that the influence of lookahead bias is both model- and application-specific. Notably, downstream predictive models can adapt to limitations in language comprehension, ensuring economically and statistically significant gains.

In addition to quantifying the impact of lookahead bias in return prediction using financial news, we propose a scalable framework for training chronologically consistent LLMs. This framework offers a constructive solution to deal with lookahead bias in predictive modeling, addressing a fundamental challenge in the application of LLMs to finance and other social sciences. By ensuring chronological consistency, our approach lays the foundation for more reliable applications of LLMs in these domains.

We make ChronoBERT models publicly available at: <https://huggingface.co/manelalab>.

References

- Bybee, Leland, Bryan Kelly, Asaf Manela, and Dacheng Xiu, 2024, Business news and business cycles, *The Journal of Finance* 79, 3105–3147.
- Chen, Yifei, Bryan T. Kelly, and Dacheng Xiu, 2023, Expected returns and large language models, *SSRN Electronic Journal* .
- Coleman, Ben, 2020, Why is it okay to average embeddings?, *Randorithms* 16.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 2019, BERT: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of the 2019 Conference of the Association for Computational Linguistics* 1, 4171–4186.
- Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al., 2024, The llama 3 herd of models, *arXiv preprint arXiv:2407.21783* .
- Fama, Eugene F, 1970, Efficient capital markets, *The Journal of Finance* 25, 383–417.
- Glasserman, Paul, and Caden Lin, 2023, Assessing look-ahead bias in stock return predictions generated by gpt sentiment analysis, *SSRN Electronic Journal* .
- Gunasekar, Suriya, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al., 2023, Textbooks are all you need, *arXiv preprint arXiv:2306.11644* .
- He, Songrun, Linying Lv, and Guofu Zhou, 2024, Empirical asset pricing with probability forecasts, *Available at SSRN* .
- Hoberg, Gerard, and Asaf Manela, 2025, The natural language of finance, *Foundations and Trends in Finance* .
- Huang, Allen H, Hui Wang, and Yi Yang, 2023, FinBERT: A large language model for extracting information from financial text, *Contemporary Accounting Research* 40, 806–841.
- Jha, Manish, Hongyi Liu, and Asaf Manela, 2025, Does finance benefit society? a language embedding approach, *Review of Financial Studies* .
- Jiang, Hao, Sophia Zhengzi Li, and Hao Wang, 2021, Pervasive underreaction: Evidence from high-frequency data, *Journal of Financial Economics* 141, 573–599.

- Jordan, Keller, Jeremy Bernstein, Brendan Rappazzo, @fernbear.bsky.social, Boza Vlado, You Jiacheng, Franz Cesista, Braden Koszarsky, and @Grad62304977, 2024, modded-nanogpt: Speedrunning the nanogpt baseline.
- Ke, Zheng Tracy, Bryan T. Kelly, and Dacheng Xiu, 2019, Predicting returns with text data, *SSRN Electronic Journal* .
- Kelly, Bryan, Asaf Manela, and Alan Moreira, 2021, Text selection, *Journal of Business & Economic Statistics* 39, 859–879.
- Ledoit, Oliver, and Michael Wolf, 2008, Robust performance hypothesis testing with the sharpe ratio, *Journal of Empirical Finance* 15, 850–859.
- Lopez-Lira, Alejandro, and Yuehua Tang, 2023, Can ChatGPT forecast stock price movements? return predictability and large language models, *SSRN Electronic Journal* .
- Loughran, Tim, and Bill McDonald, 2011, When is a liability not a liability? textual analysis, dictionaries, and 10-ks, *The Journal of Finance* 66, 35–65.
- Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan, 2025, Large language models: An applied econometric framework, Technical report, National Bureau of Economic Research.
- Lv, Linying, 2024, The value of information from sell-side analysts, *arXiv preprint arXiv:2411.13813* .
- Manela, Asaf, and Alan Moreira, 2017, News implied volatility and disaster concerns, *Journal of Financial Economics* 123, 137–162.
- Muennighoff, Niklas, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel, 2023, Scaling data-constrained language models, *Advances in Neural Information Processing Systems* 36, 50358–50376.
- Penedo, Guilherme, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al., 2024, The fineweb datasets: Decanting the web for the finest text data at scale, *arXiv preprint arXiv:2406.17557* .
- Portes, Jacob, Alexander Trott, Sam Havens, Daniel King, Abhinav Venigalla, Moin Nadeem, Nikhil Sardana, Daya Khudia, and Jonathan Frankle, 2023, MosaicBERT: A bidirectional encoder optimized for fast pretraining, *Advances in Neural Information Processing Systems* 36, 3106–3130.

- Sakar, Suproteem, 2024, StoriesLM: A family of language models with time-indexed training data, *SSRN Electronic Journal* .
- Sarkar, Suproteem, and Keyon Vafa, 2024, Lookahead bias in pretrained language models, *SSRN Electronic Journal* .
- Tetlock, Paul C., 2007, Giving content to investor sentiment: The role of media in the stock market, *The Journal of Finance* 62, 1139–1168.
- Tetlock, Paul C., Maytal Saar-Tsechansky, and Sofus Macskassy, 2008, More than words: Quantifying language to measure firms' fundamentals, *The Journal of Finance* 63, 1437–1467.
- Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman, 2019, GLUE: A multi-task benchmark and analysis platform for natural language understanding, in *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Warner, Benjamin, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al., 2024, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, *arXiv preprint arXiv:2412.13663* .

Appendix

A GLUE Evaluation

In this part, we lay out the details of the GLUE evaluation process. Following [Warner et al. \(2024\)](#), we use the same evaluation hyperparameters. Here are the details on learning rate, weight decay, and maximum number of epochs for each task. We use early stopping for all the fine-tuning tasks based on validation loss. The RTE, MRPC, and STS-B tasks are finetuned starting from the checkpoint of MNLI.

- CoLA (Corpus of Linguistic Acceptability): learning rate: $8e-5$; weight decay: $1e-6$; maximum epochs: 5.
- SST-2 (Stanford Sentiment Treebank - Binary Classification): learning rate: $8e-5$; weight decay: $1e-5$; maximum epochs: 2.
- MNLI (Multi-Genre Natural Language Inference): learning rate: $5e-5$; weight decay: $5e-6$; maximum epochs: 1.
- MRPC (Microsoft Research Paraphrase Corpus): learning rate: $5e-5$; weight decay: $5e-6$; maximum epochs: 10.
- QNLI (Question Natural Language Inference): learning rate: $8e-5$; weight decay: $5e-6$; maximum epochs: 2.
- QQP (Quora Question Pairs): learning rate: $5e-5$; weight decay: $5e-6$; maximum epochs: 10.
- RTE (Recognizing Textual Entailment): learning rate: $5e-5$; weight decay: $1e-5$; maximum epochs: 3.
- STS-B (Semantic Textual Similarity Benchmark): learning rate: $8e-5$; weight decay: $5e-6$; maximum epochs: 10.