

THE ECONOMIC LOGIC OF LARGE LANGUAGE MODELS

THIS VERSION: November 7, 2024

Megan Czasonis, Huili Song, and David Turkington

Megan Czasonis is a managing director at State Street Associates in Cambridge, MA.

mczasonis@statestreet.com

Huili Song is a vice president at State Street Associates in Clifton, NJ.

hsong3@statestreet.com

David Turkington is senior managing director and head of State Street Associates in Cambridge, MA.

dturkington@statestreet.com

Abstract

Unlike narrow statistical models, large language models (LLMs) can extrapolate patterns across disparate domains of knowledge. We apply a logical interpretation framework called Model Fingerprint to quantify the precise macroeconomic reasoning of LLMs. We find that they identify intuitive economic relationships, including conditionalities that diverge strongly from machine learning approaches applied narrowly to economic data. LLMs were able to infer statistically significant differences in positive versus negative contemporaneous economic growth from five other economic inputs, outperforming logistic regression and random forest models trained on corresponding historical data.

THE ECONOMIC LOGIC OF LARGE LANGUAGE MODELS

It is well recognized that large language models (LLMs) can offer novel perspectives on economics and finance by processing unstructured text inputs. In this paper, we focus instead on applying LLMs to structured numeric inputs as an alternative to other machine learning models. Fundamentally, LLMs and numerical models both learn patterns in training data. But unlike traditional models that rely on narrowly curated datasets, LLMs have the potential to extrapolate patterns across disparate domains of knowledge. This ability might be valuable for predicting highly complex relationships that depend on human behavior, such as economic outcomes.

Our primary interest is to understand the logical relationships that LLMs perceive across macroeconomic variables. Do they identify directional effects among variables that align with conventional wisdom and theory? To what extent do they rely on nonlinear and conditional relationships as opposed to simple linear rules? To address these questions, we define a simple approach to inferring (predicting) the direction of contemporaneous economic growth as a function of low, medium, or high values of five key economic variables. We prompt LLMs with these hypothetical conditions and map the inputs and outputs to numeric values. We then apply a framework called Model Fingerprint to summarize the impact of each variable's linear, nonlinear, and interaction effects on the model's predictive inferences about growth. This process provides interpretive transparency into a model's reasoning in terms of input variables. Because models discard their training data and retain only their fitted parameter values, we cannot interpret a model's logic in terms of examples from its training sample.¹ It is always

possible, though, to observe how a black-box model relates inputs to outputs by systematically asking it to perform tasks with varied inputs. We compare the pre-trained logic of LLMs to that of statistical models trained on historical observations of the same economic variables and find that the results differ markedly between the two approaches.

We also compare the accuracy of LLM reasoning to other models based on actual historical data. While it is appropriate to worry that tests of this sort may be biased by LLMs' rote memorization of historical facts, this issue should pose little to no concern for our setup because we prompt the LLMs generically without reference to any specific dates, countries, or time frames. We find that LLMs infer quarterly economic growth outcomes more reliably than narrowly trained statistical models, and they identify statistically significant separation in realized growth outcomes on average.

Throughout our analysis, we intentionally consider only contemporaneous economic relationships as opposed to forward predictions, because our goal is to provide transparency into model logic and we expect there is incrementally less noise and more signal associated with contemporaneous relationships. In addition to the key results described above, we further explore how effective LLMs are at assessing their level of confidence in each prediction, and we compare alternative LLM implementations.

We proceed as follows. First, we describe the LLMs and traditional numerical models we consider, along with the methodology for prompts and Model Fingerprint interpretability. We then compare the Model Fingerprint for the reasoning of a pre-trained LLM with those of a logistic regression and random forest model trained on actual historical outcomes of the same

numeric inputs. Next, we test the empirical efficacy of these alternative models for predicting historical realizations of contemporaneous economic growth, including sensitivity analysis for LLM calibration. We conclude with a summary.

Large Language Models

LLMs are computational models capable of interpreting and generating human language. Trained on enormous amounts of data, such as text from the internet, LLMs learn by identifying patterns and relationships. In this regard, they are like traditional statistical models. However, unlike statistical models based on structured numeric inputs, LLMs can extrapolate patterns across broadly disparate domains because language serves as a universal means of relating different types of information. Even though LLMs are trained for text generation, we can force them to predict outcomes that map to an interpretation of structured inputs and outputs. We evaluate three prominent open-source LLMs available at the time of this research:

- Mistral 7b – Introduced by Mistral AI in September 2023,² Mistral 7b is a decoder-only Transformer with 7 billion parameters. We use this as our baseline LLM for comparison with traditional statistical models.
- Mixtral 8x7b – Introduced by Mistral AI in December 2023, Mixtral 8x7b has the same architecture as Mistral 7b but employs a technique called sparse Mixture of Experts, in which a router network chooses two “experts” from eight and combines their output. As such, it has 46.8 billion parameters, approximately eight times the parameters of Mistral 7b.

- Llama 3.1 – Introduced by Meta in July 2024,³ Llama 3.1 is a decoder-only Transformer with 405 billion parameters.

Because our goal is to assess the value of cross-domain extrapolation, we intentionally select LLMs that are general in nature, as opposed to models that are fine-tuned or pre-trained on economic-specific text. Moreover, our goal is not to determine the optimal model or approach — which is case-specific and sure to evolve with subsequent technology — but rather to offer perspective on the general properties of LLM structures applied to this problem. We intentionally focus on the “model” aspect of LLMs in terms of the parameters they learn, and we do not consider any other data retrieval augmentation. We do explore ensemble models, however we do not consider iterative interactive reasoning of multiple agents, though this is a fascinating avenue for future research.

Methodology

To compare the patterns discovered by LLMs versus other statistical models, we ask the LLMs structured questions that frame the link between specific economic inputs and outputs. Then, we use a framework called Model Fingerprint to summarize the black-box reasoning of how those inputs map to predictions. We apply the same framework to evaluate the reasoning of traditional statistical models.

LLM Prompt

We ask the LLMs to predict the contemporaneous economic growth that coincides with the stated conditions of five economic and market variables: interest rates, inflation, government expenditures, equity returns, and yield spreads. Specifically, using the prompt below, we ask the LLMs to predict whether contemporaneous growth should be “positive” or “negative” for combinations of “high,” “medium,” and “low” for the five input variables. We select values of the inputs denoted in bold to formulate specific versions of the prompt.

*Given the current economic environment characterized by **[high/medium/low]** interest rate, **[high/medium/low]** inflation, **[high/medium/low]** government expenditure, **[high/medium/low]** equity total return, **[high/medium/low]** yield spread, is economic growth positive or negative for the same time period? Consider both individual features’ impact and their combined effect.*

Please first make an informed decision as an educated economist on whether the economic growth is positive or negative contemporaneously, and second assign confidence to the prediction.

Do not give ambiguous or uncertain answer. Pick a positive or negative side.

*In the last paragraph, return a conclusion in format: **positive/negative growth|high/low confidence.***

We include three options for the input conditions (conceptually corresponding to variable values) to keep the analysis simple while also allowing for nonlinear nuance in each relationship. By contrast, we demand that the output is either positive or negative so that we

receive strong views for predictions and avoid an excessive number of less informative neutral responses.

There are several important features of this prompt:

- We frame the questions in terms of hypothetical relationships subject to uncertainty as opposed to the determination of existing facts. Because there is no inherent truth (we do not specify a country, point in time, or time horizon), an LLM's response should reflect an interpolation or extrapolation of data.
- Because our goal is to understand the economic logic of models, we consider contemporaneous economic relationships as opposed to forward predictions, as we expect there is incrementally less noise associated with these relationships which could make patterns more apparent.
- In addition to predicting economic relationships, we ask the LLM to assess its level of confidence in each prediction.

It is important to stress that these prompts do not rely on any empirical data. We simply ask all 243 possible permutations of input conditions and record the responses. These responses collectively constitute a reduced form economic model that is fully described by the function that maps each of the 243 input conditions to a growth prediction and a level of confidence. The predictive logic comes entirely from the pre-training of a publicly available LLM.

Non-LLM Models and Data

For comparison with the LLMs, we also train statistical models on historical numeric observations of the pertinent inputs and outputs measured at the quarterly frequency. All data is obtained from the online Federal Reserve Economic Data (FRED) library or LSEG Datastream, and covers the period Q2 1955 to Q1 2024. For the prediction target, we measure economic growth as the percentage change in U.S. real GDP (seasonally-adjusted). To align with the LLM prompt, we assign a binary value of -1 or +1 for negative and positive economic growth, respectively. For the predictive inputs, we obtain the following data:

- Interest rates: The Federal Funds Effective Rate
- Inflation: Quarterly percentage change in the U.S. Consumer Price Index (all items)
- Government expenditures: Quarterly change in federal government current expenditures (seasonally-adjusted)
- Equity returns: Quarterly return of S&P 500 Index
- Yield spread: 10-year Treasury yield minus 1-year Treasury yield

For the analysis of model logic in the following section, we assign each predictive variable a value of -1, 0, or +1 for each quarterly observation if it is in the bottom third, middle third, or top third of the full sample of values, respectively. In the later section on performance evaluation, we assign values of -1, 0, or +1 to each observation based on terciles of values from chronologically prior observations only. All variables, including economic growth, are measured over contemporaneous periods.

We consider two types of numerical machine learning models calibrated to furnish -1 and +1 predictions for contemporaneous economic growth: logistic regression and random forest. Logistic regression fits a linear regression model of the inputs where the outputs are compressed to a value between 0 and 1 to reflect the predicted probability between two binary outcomes. Given that by definition it relies on linear combinations of the inputs, logistic regression cannot implement complex logic such as interaction effects among variables (even though the compression of outputs to the range of 0 to 1 may exhibit nonlinear artifacts, as we discuss later). Random forest, introduced by Ho (1995), is a popular approach to prediction based on ensembles of small decision trees trained on randomly selected subsets of the input variables. We implement a very simple random forest model with 100 individual trees.⁴ After making all predictions with each model, we select a threshold for mapping the outputs from the range 0 to 1 to binary outcomes -1 and +1 so that the frequency of each prediction matches the frequency of the LLM for comparison. This procedure avoids biased comparisons due to different base rates.

Model Fingerprint

We apply the Model Fingerprint framework introduced by Li, Turkington, and Yazdani (2020) to summarize how a model reasons about economic scenarios. The Model Fingerprint is a model-agnostic tool that provides insights into how predictive variable inputs contribute to predictions. By varying one or two predictors at a time while holding all else constant, the Model Fingerprint isolates the overall linear, nonlinear, and interaction effects that comprise a

model's demonstrated logic. It reports the relative importance of these logical components in the same units as the predictions themselves.

We now describe the Model Fingerprint procedure in more detail. For each variable, we trace out a partial dependence function by setting the input value for the variable equal to many different values and, each time, computing the average prediction the model would make when combined with every available combination of the remaining variables. In general, the resulting function is nonlinear. We then fit a straight line to the observations (via ordinary least squares linear regression) and compute the mean absolute deviation around zero for the linear approximation of each observation's effect, as well as the mean absolute deviation around zero of the full effect in excess of the linear approximation, which is the nonlinear component. Finally, to capture interaction effects for each pair of variables, we compute the partial predictions for every combination of input values for a pair of variables in excess of the predictions those variables would give in isolation. The interaction effect equals the mean absolute deviation of these values around zero.

For LLM predictions, we assign a value of -1 or +1 to the response based on whether the LLM predicts contemporaneous growth to be negative or positive, respectively.⁵ In addition, we employ an ensemble approach to reduce the noise in LLM responses introduced by the random seed (as governed by the temperature setting). Specifically, we ask the same question to ten independent sessions of an LLM and aggregate their answers as an equally-weighted average.

As mentioned previously, our LLM-based prediction function maps each of 243 possible inputs to an output value of -1 or +1. We must choose what sample of input circumstances to use to compute the Model Fingerprint. If we used the sample of all 243 unique inputs, the Fingerprint would tell us how the model reasons on average if inputs are drawn randomly from this set. While interesting, this result would not necessarily align to the real-world application of such a model, where some circumstances are more common than others. Therefore, in the analysis that follows we compute the Fingerprints for each model using the empirical set of conditions from the quarterly data set from 1955 to 2024 described in the previous subsection.

Results for Prediction Logic

Exhibit 1 shows the Model Fingerprint results for the baseline LLM (Mistral 7b), logistic regression, and random forest. For each model, it summarizes the relative importance of the linear and nonlinear effects for each input variable along with the interaction effects for each pair of variables. These values capture the average extent to which changes in a variable influence a model's prediction. For these results, the statistical models are trained on the full sample of empirical data; in the following section on prediction performance, we retrain the models on chronologically prior observations.

Exhibit 1: Model Fingerprints

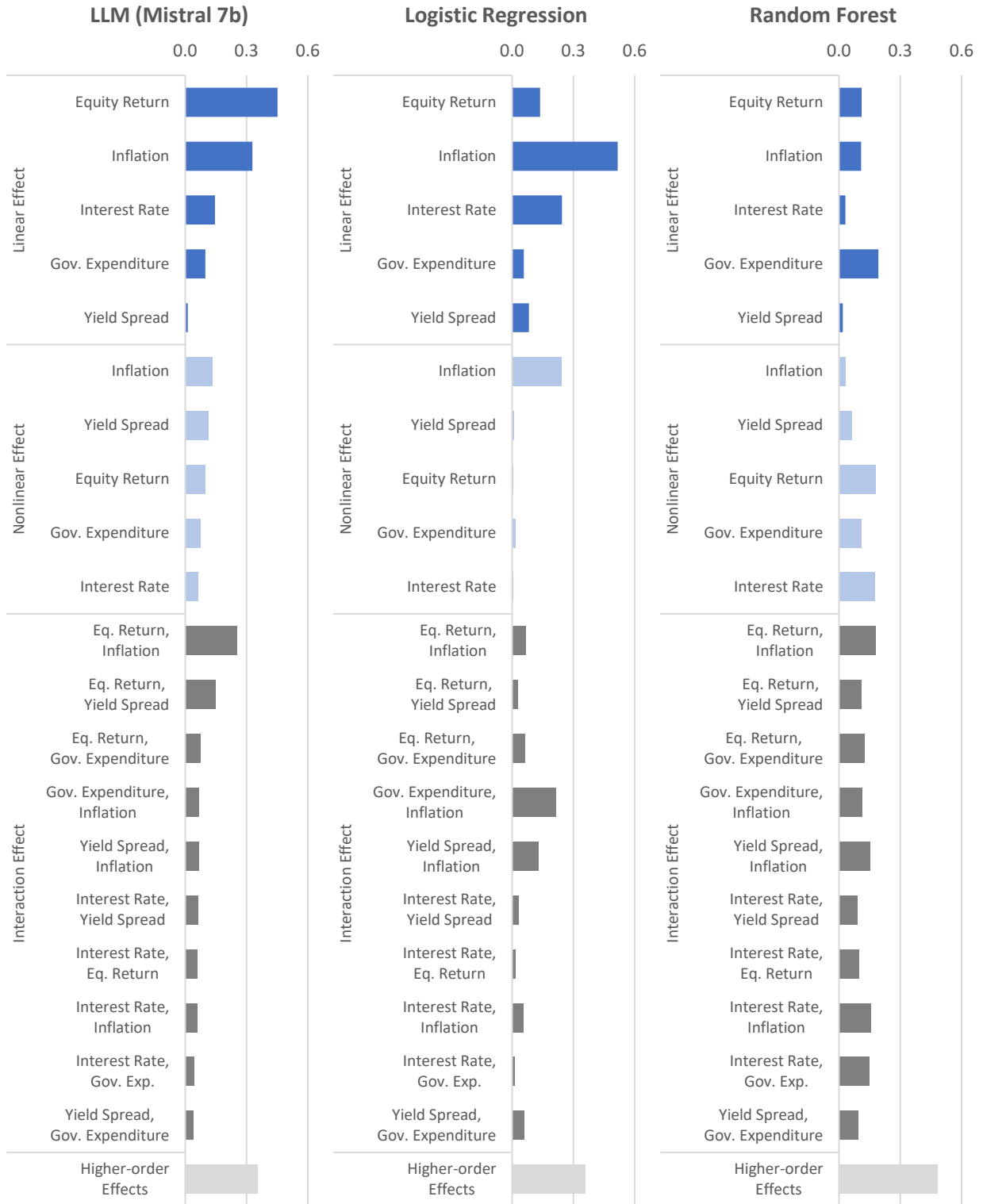
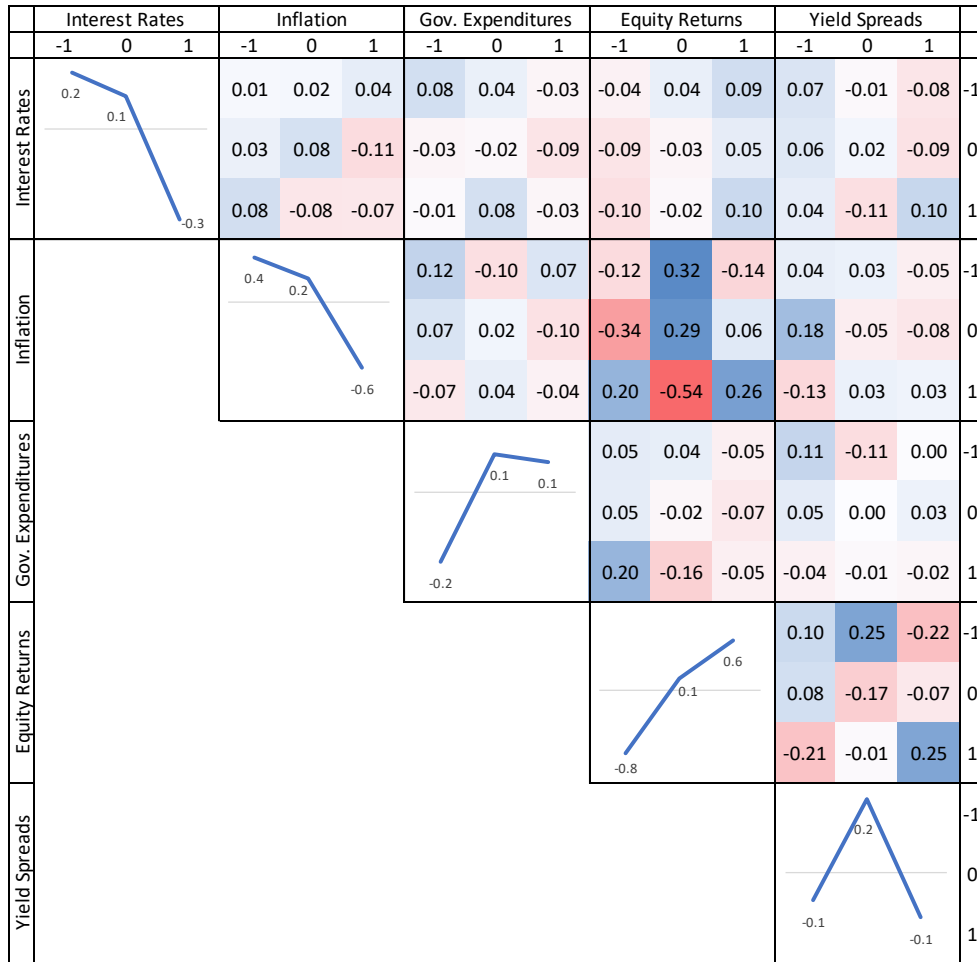


Exhibit 1 reveals that the LLM and logistic regression have broadly similar linear logic, with equity returns, inflation, and interest rates as the three most important contributors. Moreover, they both find significant nonlinear and interaction effects with inflation: inflation with equity returns for the LLM, and inflation with expenditures and yield spreads for logistic regression. Nonetheless, compared to logistic regression, nonlinear and conditional effects are more important to the LLM's logic overall. This likely reflects the fact that logistic regression is limited to combining factors linearly; it is only in making categorical predictions that its full logic appears to implement more complex effects as an artifact. Finally, we observe that random forest looks quite different from the other models. A significant portion of its logic reflects interaction and higher-order effects, with only a small contribution from linear and nonlinear relationships. Of random forest's three most important interactions, only one — equity returns and inflation — overlaps with the LLM's top interactions. In fact, its second most important interaction — interest rates with inflation — is among the LLM's least important interactions.

We next explore the detailed effects that underlie the LLM's fingerprint, as shown in Exhibit 2 (for similar results for the statistical models, please refer to Appendix A). The diagonal elements plot the one-dimensional partial predictions for each variable in isolation. They represent the average prediction the LLM would make for "low," "medium," or "high" values of an input (denoted as -1, 0, and +1 in the exhibit) when combined with every available combination of the remaining variables. The off-diagonal elements plot the two-dimensional interaction effects for a given pair of variables. These values represent the average prediction the model would make for every combination of two variables in excess of the predictions those variables would give in isolation.

Exhibit 2: LLM (Mistral 7b) Model Fingerprint – Details



Much of the logic illustrated by Exhibit 2 is intuitive. The LLM views economic growth as having a positive overall relationship with equity returns and a negative relationship with inflation. However, the interaction effect between these two inputs is substantial. The model reasons that high inflation is especially bad when it coincides with modest stock returns. Other notable interaction effects include equity returns with yield spreads and government expenditures. When equity returns are low, moderate yield spreads or high government expenditures are more favorable for growth than high yield spreads or low expenditures.

As further comparison of the models' reasoning, Exhibit 3 reports the correlations between their logical components. We observe that the LLM is positively correlated to the statistical models with respect to their linear (69% and 66% for logistic regression and random forest, respectively) and nonlinear (71% and 61%, respectively) components. However, their interaction (-4% and -4%, respectively) and higher-order effects (1% and -8%, respectively) have near-zero or negative correlations. This suggests that the LLM generally agrees with the statistical models on patterns between economic growth and individual variables, while its views on more complex interactions tend to differ.

Exhibit 3: Correlations of Model Fingerprint Components

			A	B	C	D	E	F	G	H	I	J	K	L
LLM Mistral 7b	Linear	A	1.00											
	Nonlinear	B	0.10	1.00										
	Interaction	C	0.44	-0.05	1.00									
	Higher-order	D	-0.34	0.14	-0.26	1.00								
Logistic Regression	Linear	E	0.69	0.13	0.20	-0.12	1.00							
	Nonlinear	F	-0.03	0.71	-0.17	0.17	0.00	1.00						
	Interaction	G	-0.05	-0.36	-0.04	-0.02	-0.14	-0.37	1.00					
	Higher-order	H	-0.01	0.35	0.00	0.01	0.02	0.37	-0.35	1.00				
Random Forest	Linear	I	0.66	0.11	0.17	-0.10	0.85	-0.01	-0.06	-0.04	1.00			
	Nonlinear	J	0.05	0.61	-0.03	0.08	0.12	0.35	-0.20	0.15	0.04	1.00		
	Interaction	K	-0.14	-0.07	-0.04	0.14	-0.27	0.03	0.13	0.04	-0.29	-0.34	1.00	
	Higher-order	L	0.09	0.01	0.10	-0.08	0.16	-0.03	-0.06	0.02	0.12	0.24	-0.32	1.00

Results for Prediction Performance

We now compare the baseline LLM and statistical models in terms of the efficacy of their predictions. We also compare predictions generated by alternative LLMs and calibrations.

To conduct these tests, we rely on the historical data used to train the statistical models. Starting in Q2 1980, we ask the LLM to predict contemporaneous growth based on values of the

five input variables at the time. We use the prompt described earlier and translate the variables to “low,” “medium,” or “high” based on their trailing percentile ranks and tercile thresholds (the same approach used to assign values of -1, 0, +1 for the statistical models). We repeat this process each quarter through Q1 2024, generating a total of 176 predictions. For comparison, we form out-of-sample predictions for the statistical models for the same quarterly periods. Specifically, beginning in Q2 1980, at the end of quarter, we predict contemporaneous growth based on a training sample of quarterly observations for growth outcomes (-1, +1) and predictive variables (-1, 0, +1) from Q2 1955 through the prior quarter. By retraining new statistical models using expanding sets of data, we ensure that each prediction for a new quarter of growth is based only on the patterns that can be observed in the historical record prior to that point.⁶

Baseline Results

Exhibit 4 compares the prediction efficacy of the baseline LLM (Mistral 7b), logistic regression, and random forest. It shows hit rates, mean squared errors, and average growth outcomes for quarters that were predicted to have positive and negative growth. Consistent with the model training, we evaluate actual outcomes in the form of -1 or +1 for negative and positive growth, respectively. As described previously, for logistic regression and random forest, after forming all of the models’ predictions, we select a threshold for mapping the outputs to binary outcomes, -1 and +1, so that the frequency of each prediction matches the frequency of the LLM for comparison.

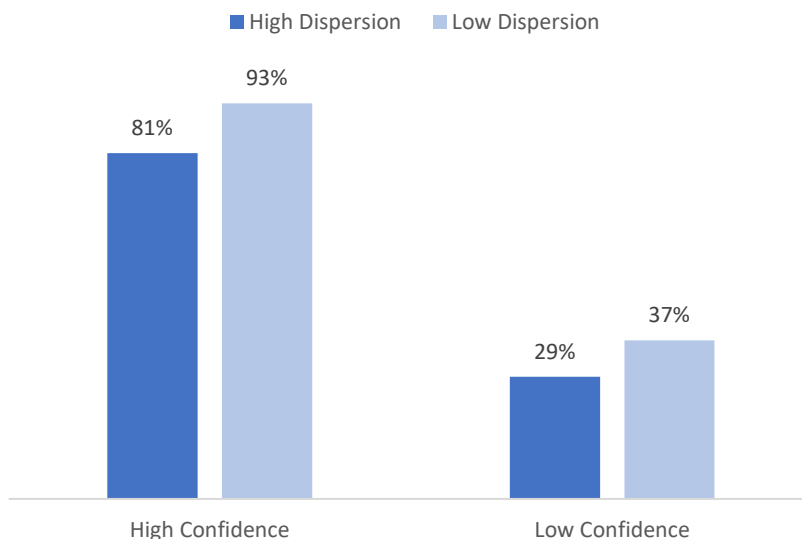
Exhibit 4 reveals that across the three models, the LLM has the highest hit rate, smallest mean squared error, and the largest, and only statistically significant, difference in average growth outcomes for positive versus negative predictions. This suggests that the LLM can effectively reason through economic relationships and that cross-domain extrapolation may add value above explicit statistical analysis.

Exhibit 4: Predictive Performance

	Hit Rate	Mean Squared Error	Average Actual Outcomes			
			Positive Predictions	Negative Predictions	Difference	t-statistic
LLM - Mistral 7b	72%	1.11	0.87	0.54	0.33	2.59
Logistic Regression	68%	1.27	0.80	0.66	0.14	1.23
Random Forest	67%	1.32	0.79	0.70	0.09	0.80

Next, we explore whether uncertainty impacts the LLM’s predictive performance. Specifically, we consider two indications of uncertainty that are known at the time of each prediction: The LLM’s stated confidence in the prediction, and the cross-sectional dispersion in individual predictions underlying the ensemble prediction. Exhibit 5 reports the LLM’s hit rate for predictions double-sorted by their dispersion (above and below median) and stated confidence (above and below median). It reveals that high confidence predictions have higher hit rates than low confidence predictions, and that low dispersion predictions have higher hit rates than high dispersion predictions. Consistent with existing research on AI hallucination,⁷ this suggests that LLM uncertainty can effectively indicate which predictions are likely to be most reliable based on the strength of underlying patterns.

Exhibit 5: LLM (Mistral 7b) Prediction Hit Rates



Sensitivity Analysis

The following exhibits present empirical results comparing different LLMs and features. Our goal is not to determine the optimal model or approach, but rather to understand the similarities and differences and offer perspective on which decisions appear to matter more than others. We intentionally focus on the “model” aspect of LLMs in terms of the (billions) of parameters they learn, and we do not consider any other data retrieval augmentation.

Exhibit 6 compares the prediction efficacy of the baseline LLM (Mistral 7b) with two alternative LLMs and temperature settings. An LLM’s temperature controls the level of randomness used to generate its responses. Lower temperatures prioritize the most statistically probable text generation whereas higher temperatures allow the selection of more unusual sequences.

We observe that the LLMs are similarly effective at predicting growth outcomes, suggesting that the baseline performance of Mistral 7b is representative of the alternative LLMs considered here. Moreover, with the exception of Mistral 7b’s difference in means test, for all three models, a low temperature performs better than a high temperature. This indicates a benefit to prioritizing more statistically probable responses.

Exhibit 6: Prediction Performance of Alternative LLMs

Model	Temperature	Hit Rate	Mean Squared Error	Average Actual Outcomes			
				Positive Predictions	Negative Predictions	Difference	t-statistic
Mistral 7b	High	72%	1.11	0.87	0.54	0.33	2.59
Mistral 7b	Low	83%	0.68	0.81	0.43	0.38	1.82
Mixtral 8x7b	High	71%	1.16	0.84	0.58	0.26	2.10
Mixtral 8x7b	Low	74%	1.02	0.86	0.50	0.36	2.68
Llama 3.1	High	65%	1.38	0.86	0.56	0.29	2.51
Llama 3.1	Low	67%	1.28	0.88	0.54	0.34	2.87

As a final comparison, Exhibit 7 reports the correlations between the LLMs’ predictions. While the results in the previous exhibit show consistent predictive efficacy across the LLMs, the correlations in Exhibit 7 are relatively low (58%, on average). A possible explanation for this seeming inconsistency is that the LLMs capture similar actual patterns between variables but different noise or errors. This noise could degrade correlations across models (thus explaining Exhibit 7) but cancel out when evaluating their overall efficacy (thus explaining Exhibit 6). Therefore, while the models are similarly effective, they are not necessarily substitutes, which is further reason to explore their economic logic.

Exhibit 7: Correlations of Predictions for Alternative LLMs

	Mistral 7b		Mixtral 8x7b		Llama 3.1	
	High	Low	High	Low	High	Low
Mistral 7b - High	1.00					
Mistral 7b - Low	0.69	1.00				
Mixtral 8x7b - High	0.78	0.64	1.00			
Mixtral 8x7b - Low	0.82	0.62	0.83	1.00		
Llama 3.1 - High	0.44	0.39	0.39	0.46	1.00	
Llama 3.1 - Low	0.49	0.44	0.58	0.59	0.57	1.00

Summary

We study the economic logic of LLMs in the context of inferring contemporaneous economic growth from five other macroeconomic variables. LLMs have the capacity to extrapolate patterns from disparate domains which may offer an advantage over narrower statistical models. We find that a range of publicly available LLMs appear to offer competent reasoning that aligns with intuition and departs from the reasoning learned by traditional statistical models trained only on the data that is used as inputs and outputs for the tasks. The LLMs often produced statistically significant separation of actual growth outcomes following inferences that growth should be positive as opposed to negative. Narrow numerical models of logistic regression and random forest failed to produce statistically significant separation in our tests. Our results suggest that LLMs’ uncertainty — as indicated by their stated confidence in a prediction or the dispersion in individual predictions underlying the ensemble prediction — is a useful signal: in tests on historical data, high confidence predictions and low dispersion predictions were more reliable than low confidence or high dispersion predictions. While the experiments in this paper pertain to contemporaneous relationships among economic

variables, the effective fundamental logic we observe suggests the potential for LLMs to add value as forward predictors of complex and uncertain economic outcomes.

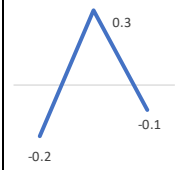
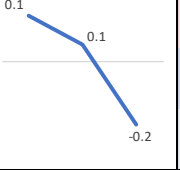
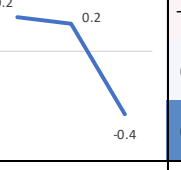
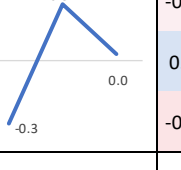
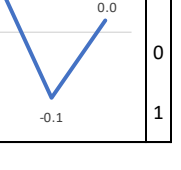
Appendix A: Model Fingerprint Details

Exhibits A1 and A2 show the detailed effects underlying the Model Fingerprints for logistic regression and random forest, respectively. Whereas a fully linear model always exhibits precisely zero nonlinear effects and zero interaction effects, a logistic regression's output is nonlinearly and conditionally dependent on its inputs to the extent that the compression of linear combinations to 0 and 1 outputs is more dramatic in some circumstances than others.

Exhibit A1: Logistic Regression Model Fingerprint – Details

	Interest Rates			Inflation			Gov. Expenditures			Equity Returns			Yield Spreads			
	-1	0	1	-1	0	1	-1	0	1	-1	0	1	-1	0	1	
Interest Rates				0.09	-0.05	-0.05	0.01	0.01	-0.02	0.01	-0.03	0.07	0.09	-0.05	0.04	-1
				-0.04	-0.04	0.06	-0.01	-0.02	-0.01	-0.02	-0.02	-0.01	-0.04	-0.02	0.01	0
				-0.08	0.00	0.11	0.03	0.00	0.01	0.00	0.00	0.00	-0.03	-0.02	0.00	1
Inflation				-0.36	0.02	0.38	0.15	0.01	-0.09	-0.18	0.01	0.24	-1			
				-0.13	0.14	-0.23	-0.06	0.02	0.03	0.05	-0.06	-0.20	0			
				0.45	-0.18	-0.07	-0.09	-0.07	0.11	0.30	-0.03	-0.12	1			
Gov. Expenditures				0.06	-0.05	-0.01	0.02	0.07	-0.08	-1						
				-0.05	0.13	0.09	-0.01	-0.08	0.10	0						
				-0.09	-0.07	-0.01	0.08	-0.02	-0.08	1						
Equity Returns				0.05	-0.06	0.04	-1									
				0.00	0.01	-0.05	0									
				-0.02	0.00	0.02	1									
Yield Spreads				-1												
				0												
				1												

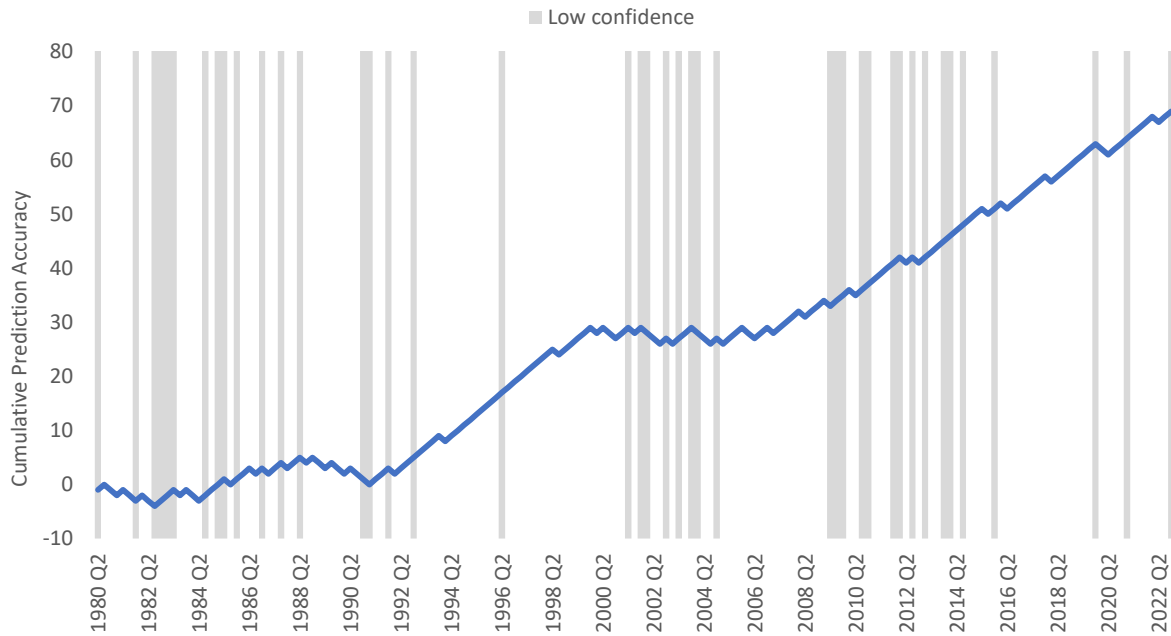
Exhibit A2: Random Forest Model Fingerprint – Details

	Interest Rates			Inflation			Gov. Expenditures			Equity Returns			Yield Spreads			
	-1	0	1	-1	0	1	-1	0	1	-1	0	1	-1	0	1	
Interest Rates				-0.13	0.11	-0.10	-0.42	-0.06	0.20	-0.10	-0.06	0.06	0.00	-0.04	-0.09	-1
	-0.37	0.02	0.32	-0.03	-0.12	0.13	-0.01	-0.12	0.11	0.05	0.23	-0.25	0			
	0.25	-0.07	-0.04	0.28	0.07	-0.05	0.23	0.05	-0.16	-0.01	0.07	0.06	1			
Inflation				-0.08	0.15	-0.05	-0.29	-0.01	0.30	-0.09	-0.08	0.16	-1			
	0.09	0.07	-0.13	0.10	0.15	-0.22	-0.14	-0.05	0.23	0						
	0.03	-0.24	0.16	0.05	-0.28	0.21	0.07	0.23	-0.32	1						
Gov. Expenditures				-0.03	-0.35	0.17	0.03	0.03	-0.02	-1						
	0.01	0.04	0.01	-0.07	0.14	-0.07	0									
	0.34	-0.09	-0.11	-0.18	-0.10	0.22	1									
Equity Returns				-0.04	0.20	-0.19	-1									
	0.07	-0.20	0.00	0												
	-0.07	0.17	0.07	1												
Yield Spreads				-1												
	0															
	1															

Appendix B: Cumulative Prediction Accuracy

Exhibit A3 shows the cumulative accuracy of Mistral 7b’s economic growth predictions through time. The blue line indicates the cumulative number of correct predictions minus incorrect predictions. The gray shading indicates low confidence (below median) predictions according to the LLM’s stated confidence.

Exhibit A3: LLM (Mistral 7b) Cumulative Prediction Accuracy



Appendix C: Historical Numeric Inputs and Predictions

Exhibit A4 reports the historical numeric inputs and resulting Mistral 7b predictions.

Exhibit A4: Numeric Inputs and LLM (Mistral 7b) Predictions

	Interest Rates	Inflation	Govt Spending	Equity Return	Yield Spread	GDP Growth	Mistral 7b		Interest Rates	Inflation	Govt Spending	Equity Return	Yield Spread	GDP Growth	Mistral 7b
1980 Q2	1	1	1	1	1	-1	1	2002 Q2	-1	-1	0	-1	1	1	0
1980 Q3	1	1	1	1	-1	-1	-1	2002 Q3	-1	-1	-1	-1	1	1	-1
1980 Q4	1	1	0	1	-1	1	-1	2002 Q4	-1	-1	0	1	1	1	1
1981 Q1	1	1	1	0	-1	1	-1	2003 Q1	-1	1	0	-1	1	1	-1
1981 Q2	1	1	0	-1	-1	-1	-1	2003 Q2	-1	-1	1	1	1	1	1
1981 Q3	1	1	1	-1	-1	1	-1	2003 Q3	-1	0	-1	0	1	1	1
1981 Q4	1	0	0	1	1	-1	1	2003 Q4	-1	-1	0	1	1	1	1
1982 Q1	1	-1	0	-1	-1	-1	-1	2004 Q1	-1	1	0	0	1	1	-1
1982 Q2	1	1	0	-1	0	1	-1	2004 Q2	-1	1	-1	0	1	1	-1
1982 Q3	1	0	1	1	1	-1	1	2004 Q3	-1	-1	-1	-1	1	1	-1
1982 Q4	1	-1	1	1	1	1	1	2004 Q4	-1	-1	0	1	1	1	1
1983 Q1	1	-1	-1	1	1	1	1	2005 Q1	-1	1	1	-1	1	1	-1
1983 Q2	1	1	0	1	1	1	1	2005 Q2	-1	0	0	0	0	1	1
1983 Q3	1	0	-1	0	1	1	-1	2005 Q3	-1	1	0	0	0	1	1
1983 Q4	1	-1	-1	0	1	1	1	2005 Q4	0	-1	0	0	-1	1	1
1984 Q1	1	0	0	-1	1	1	-1	2006 Q1	0	1	1	0	-1	1	-1
1984 Q2	1	0	0	-1	1	1	-1	2006 Q2	0	1	-1	-1	-1	1	-1
1984 Q3	1	0	0	1	1	1	1	2006 Q3	0	-1	-1	0	-1	1	1
1984 Q4	1	-1	1	0	1	1	1	2006 Q4	0	-1	-1	1	-1	1	1
1985 Q1	1	0	-1	1	1	1	1	2007 Q1	0	1	1	0	-1	1	-1
1985 Q2	1	0	0	1	1	1	1	2007 Q2	0	1	0	1	-1	1	1
1985 Q3	1	-1	0	-1	1	1	-1	2007 Q3	0	-1	0	0	0	1	1
1985 Q4	1	0	-1	1	1	1	1	2007 Q4	0	0	0	-1	0	1	1
1986 Q1	1	-1	-1	1	0	1	1	2008 Q1	-1	1	0	-1	1	-1	-1
1986 Q2	0	-1	0	0	1	1	1	2008 Q2	-1	1	1	-1	1	1	-1
1986 Q3	0	-1	0	-1	1	1	-1	2008 Q3	-1	-1	-1	-1	1	-1	-1
1986 Q4	0	-1	-1	0	1	1	1	2008 Q4	-1	-1	-1	-1	1	-1	-1
1987 Q1	0	1	-1	1	1	1	-1	2009 Q1	-1	1	1	-1	1	-1	-1
1987 Q2	0	0	0	0	1	1	1	2009 Q2	-1	1	1	1	1	-1	1
1987 Q3	0	0	-1	1	1	1	1	2009 Q3	-1	-1	-1	1	1	1	1
1987 Q4	0	-1	0	-1	1	1	-1	2009 Q4	-1	-1	-1	1	1	1	1
1988 Q1	0	0	0	0	1	1	1	2010 Q1	-1	0	1	0	1	1	1
1988 Q2	1	0	-1	1	1	1	1	2010 Q2	-1	-1	-1	-1	1	1	-1
1988 Q3	1	1	-1	0	0	1	-1	2010 Q3	-1	-1	0	1	1	1	1
1988 Q4	1	-1	1	0	-1	1	1	2010 Q4	-1	-1	-1	1	1	1	1
1989 Q1	1	1	0	1	-1	1	-1	2011 Q1	-1	1	-1	0	1	-1	-1
1989 Q2	1	1	0	1	-1	1	-1	2011 Q2	-1	0	-1	0	1	1	1
1989 Q3	1	0	0	1	-1	1	1	2011 Q3	-1	-1	-1	-1	1	-1	-1
1989 Q4	1	0	-1	0	0	1	-1	2011 Q4	-1	-1	-1	1	1	1	1
1990 Q1	1	1	0	-1	0	1	-1	2012 Q1	-1	1	-1	1	1	1	1
1990 Q2	1	0	0	0	0	1	1	2012 Q2	-1	-1	-1	-1	1	1	-1
1990 Q3	1	1	-1	-1	1	1	-1	2012 Q3	-1	0	-1	1	1	1	1
1990 Q4	0	0	0	1	1	-1	1	2012 Q4	-1	-1	-1	-1	1	1	-1
1991 Q1	0	0	-1	1	1	-1	1	2013 Q1	-1	1	-1	1	1	1	1
1991 Q2	0	0	1	0	1	1	1	2013 Q2	-1	-1	-1	0	1	1	1
1991 Q3	0	0	1	0	1	1	1	2013 Q3	-1	-1	-1	0	1	1	1
1991 Q4	0	-1	0	1	1	1	1	2013 Q4	-1	-1	-1	1	1	1	1
1992 Q1	-1	0	1	-1	1	1	-1	2014 Q1	-1	1	0	0	1	-1	-1
1992 Q2	-1	-1	-1	0	1	1	1	2014 Q2	-1	0	0	0	1	1	1
1992 Q3	-1	0	0	0	1	1	1	2014 Q3	-1	-1	0	0	1	1	1
1992 Q4	-1	-1	0	0	1	1	1	2014 Q4	-1	-1	-1	0	1	1	1
1993 Q1	-1	0	-1	0	1	1	1	2015 Q1	-1	0	0	0	1	1	1
1993 Q2	-1	-1	-1	0	1	1	1	2015 Q2	-1	0	-1	0	1	1	1
1993 Q3	-1	-1	-1	0	1	1	1	2015 Q3	-1	-1	-1	-1	1	1	-1
1993 Q4	-1	-1	-1	0	1	1	1	2015 Q4	-1	-1	-1	1	1	1	1
1994 Q1	-1	0	-1	-1	1	1	-1	2016 Q1	-1	0	-1	0	0	1	1
1994 Q2	-1	-1	-1	0	1	1	1	2016 Q2	-1	1	-1	0	0	1	-1
1994 Q3	0	0	0	0	1	1	1	2016 Q3	-1	-1	0	0	0	1	1
1994 Q4	0	-1	0	-1	0	1	1	2016 Q4	-1	-1	-1	0	1	1	1
1995 Q1	0	0	-1	1	0	1	1	2017 Q1	-1	0	-1	1	0	1	1
1995 Q2	0	0	-1	1	0	1	1	2017 Q2	-1	-1	-1	0	0	1	1
1995 Q3	0	-1	-1	1	0	1	1	2017 Q3	-1	0	0	0	0	1	1
1995 Q4	0	-1	-1	0	0	1	1	2017 Q4	-1	-1	0	1	0	1	1
1996 Q1	0	1	0	0	0	1	1	2018 Q1	-1	1	0	-1	0	1	-1
1996 Q2	0	-1	-1	0	1	1	1	2018 Q2	-1	0	0	0	0	1	1
1996 Q3	0	0	-1	0	0	1	1	2018 Q3	-1	-1	0	1	0	1	1
1996 Q4	0	-1	-1	1	0	1	1	2018 Q4	-1	-1	0	-1	-1	1	1
1997 Q1	0	0	-1	0	0	1	1	2019 Q1	-1	1	1	1	-1	1	1
1997 Q2	0	-1	-1	1	0	1	1	2019 Q2	-1	0	0	0	-1	1	1
1997 Q3	0	-1	-1	1	0	1	1	2019 Q3	-1	-1	-1	0	-1	1	1
1997 Q4	0	-1	0	0	0	1	1	2019 Q4	-1	-1	-1	1	-1	1	1
1998 Q1	0	-1	-1	1	0	1	1	2020 Q1	-1	-1	0	-1	0	-1	1
1998 Q2	0	-1	-1	0	-1	1	1	2020 Q2	-1	-1	1	1	0	-1	1
1998 Q3	0	-1	-1	-1	-1	1	0	2020 Q3	-1	0	-1	1	0	1	1
1998 Q4	0	-1	0	1	-1	1	1	2020 Q4	-1	-1	-1	1	0	1	1
1999 Q1	0	0	-1	0	0	1	1	2021 Q1	-1	1	1	1	1	1	1
1999 Q2	0	0	-1	1	0	1	1	2021 Q2	-1	1	-1	1	0	1	1
1999 Q3	0	0	0	-1	0	1	1	2021 Q3	-1	0	-1	0	0	1	1
1999 Q4	0	-1	0	1	0	1	1	2021 Q4	-1	1	-1	1	0	1	1
2000 Q1	0	1	-1	0	-1	1	-1	2022 Q1	-1	1	-1	-1	0	-1	-1
2000 Q2	0	0	0	-1	-1	1	1	2022 Q2	-1	1	0	-1	0	1	-1
2000 Q3	0	0	-1	-1	-1	1	-1	2022 Q3	-1	-1	0	-1	-1	1	1
2000 Q4	0	-1	-1	-1	-1	1	0	2022 Q4	0	-1	1	1	-1	1	1
2001 Q1	0	1	0	-1	0	-1	-1	2023 Q1	0	1	0	0	-1	1	-1
2001 Q2	-1	0	-1	1	1	1	1	2023 Q2	0	0	-1	1	-1	1	1
2001 Q3	-1	-1	0	-1	1	-1	0	2023 Q3	0	0	0	-1	-1	1	1
2001 Q4	-1	-1	-1	1	1	1	1	2023 Q4	0	-1	0	1	-1	1	1
2002 Q1	-1	1	0	0	1	1	-1	2024 Q1	0	1	1	1	-1	1	1

Notes

This material is for informational purposes only. The views expressed in this material are the views of the authors, are provided “as-is” at the time of first publication, are not intended for distribution to any person or entity in any jurisdiction where such distribution or use would be contrary to applicable law and are not an offer or solicitation to buy or sell securities or any product. The views expressed do not necessarily represent the views of State Street Global Markets® or State Street Corporation® and its affiliates.

References

- Czasonis, Megan, Mark Kritzman, and David Turkington. 2024a. “The Virtue of Transparency: How to Maximize the Utility of Data Without Overfitting.” *MIT Working Paper* (July).
- Czasonis, Megan, Mark Kritzman, and David Turkington. 2024b. “A Transparent Alternative to Neural Networks with an Application to Predicting Volatility.” *MIT Working Paper* (September).
- Ho, Tin Kam. 1995. “Random decision forests.” *Proceedings of 3rd international conference on document analysis and recognition* (1): 278–282.
- Li, Yimou, David Turkington, and Alireza Yazdani. 2020. “Beyond the Black Box: An Intuitive Approach to Prediction with Machine Learning.” *The Journal of Financial Data Science* 2 (1): 61–75.
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., ... & Sayed, W. E. 2023. “Mistral 7B.” *arXiv preprint arXiv:2310.06825*.
- Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., ... & Sayed, W. E. 2024. “Mixtral of experts.” *arXiv preprint arXiv:2401.04088*.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., ... & Ganapathy, R. 2024. “The Llama 3 herd of models.” *arXiv preprint arXiv:2407.21783*.
- Wei, Jason, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. “Measuring short-form factuality in large language models.” OpenAI.

¹ For discussion on this issue as well as a comparison to model-free prediction that provides observation-level transparency, see Czasonis, Kritzman, and Turkington (2024a, 2024b). It is also worth noting that the inability of LLMs to connect logical reasoning to specific training data instances is a limitation shared by many human experts.

² For a more detailed description, see Jiang et al. (2023).

³ For a more detailed description, see Dubey et al. (2024).

⁴ The number of features to consider when looking for the best split is set to at most 2. A node will be split if it includes a decrease of Gini impurity. No constraints are set for max split, depth, or size of each node. Trees are fully grown since the data set is relatively small.

⁵ In our experiments, LLMs occasionally fail to provide a definitive answer, stating that they are unable to indicate either positive or negative economic growth. In such cases, we assign a value of 0 to represent the non-response. These instances account for 1.4% of the total predictions during the performance test, with a maximum of three LLMs out of the ten failing to provide an answer. When aggregating predictions across LLMs, we encounter cases where the positive and negative predictions are evenly split. These instances account for 2% of the total predictions during the performance test. We classify these ties as aggregate non-responses, which we exclude from the efficacy test.

⁶ For these experiments, we do not reflect point-in-time economic data or publication lags, both of which would likely disadvantage the statistical models. Even with the advantage of revised data and no publication lags, we find that the statistical models do not perform well compared to the LLMs.

⁷ For example, see Wei et al. (2024).