From Text to Evidence: LLMs in Economics and Finance*

CAI Zongwu · FANG Ying · LIN Ming · LIU Zeqin

DOI:
Received: x x 20xx / Revised: x x 20xx
©The Editorial Office of JSSC & Springer-Verlag GmbH Germany 202X

Abstract We review applications of large language models (LLMs) in economic and financial research, focusing on text analysis and variable construction, forecasting and decision support, simulation of economic agents, and research assistance and automation. We examine what researchers use LLMs to do, how the reported results compare with existing methods, and what limits their use. Contextual interpretation improves several classification tasks, but variable construction and measurement error affect subsequent analysis. Some forecasts improve on benchmarks, although gains vary and predictive signals need not yield trading profits. Human observations help calibrate agents and improve predictions of survey responses. Automated systems generate economic mechanisms and connect empirical analysis with drafting, while researchers still judge questions and explanations. We then discuss future directions for improving research quality and making better use of researchers' expertise.

Keywords Large language models, text analysis, economic forecasting, economic agent simulation, research assistance and automation

MR Subject Classification: 62P20, 91B84, 68T50, 68T05, 62M10, 91B28, 62-02.

1 Introduction

Policy statements, corporate disclosures, earnings calls, and news reports record the expectations and plans of economic agents. Researchers turn these materials into variables to study policy, markets, and firm behavior. Manual coding, dictionaries, and statistical text methods have long helped with this task [19]. A recurring difficulty is that the same idea can be expressed in different words, and a passage's meaning depends on its context.

Large language models (LLMs) offer a more flexible approach. They can interpret a passage using the surrounding text and follow instructions that specify what to identify. GPT-3 showed

CAI Zongwu, Department of Economics, University of Kansas, Lawrence, KS 66045, USA. Email: caiz@ku.edu.
FANG Ying, Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, Fujian 361005, China. Email: yifst1@xmu.edu.cn
LIN Ming, Department of Statistics and Data Science, Xiamen University, Xiamen, Fujian 361005, China. Email: linming50@xmu.edu.cn
LIU Zeqin (**the corresponding author**), School of Statistics, Shanxi University of Finance and Economics, Taiyuan, Shanxi 030012, China. E-mail: liuzq@sxufe.edu.cn.
*This research was supported by National Natural Science Foundation of China with grants 72303140, 72322016, 72073126, 72595870, and 72595875.

in 2020 that many language tasks could be specified through instructions and examples without further training for each task [4]. The release of ChatGPT in November 2022 made it possible for users to describe a task and refine the response through dialogue [47]. Researchers can therefore state a coding rule in ordinary language and supply examples instead of listing every phrase that might express the intended meaning. For some tasks, revising the instructions is enough to adapt the model without further training [12].

LLMs also assist with work beyond reading documents. Advances in mathematical reasoning and coding, together with connections to software tools, allow systems to combine generated answers with calculations and code execution [13; 54]. The breadth of model capabilities has also prompted discussion of progress toward artificial general intelligence (AGI), or systems that could perform a broad range of intellectual tasks at human level or beyond [6]. In economic research, the practical applications include forecasting outcomes, simulating decisions, writing code, and drafting papers [27; 32; 41]. Language is thus not only material to analyze but also a way to direct parts of the research process.

This review asks where LLMs have been used in economic and financial research, what they contribute relative to existing methods, and what limits their use. We organize selected applications around text analysis and variable construction, economic forecasting and decision support, simulation of economic agents, and research assistance and automation. We complement technical surveys [35; 44] and practical guides [32; 33] by comparing the findings of economic applications.

Section 2 defines the four research uses and explains how their performance is assessed. Sections 3–6 examine the applications and discuss the methodological issues that affect their use. Section 7 concludes and outlines directions for future research.

2 Research Uses and Review Approach

2.1 Four Research Uses

Two financial LLM surveys organize the literature primarily by technical route or application area. Li et al. [35] distinguish direct prompting, tool augmentation, domain fine-tuning, and training from scratch, whereas Nie et al. [44] organize financial applications around language tasks, sentiment analysis, financial time series, financial reasoning, and agent-based modeling. We instead organize the applications around four research uses, defined by the purpose of the model's work in a study. Ludwig et al. [43] provide an econometric framework for assessing LLM results according to their intended use. We use this principle to interpret the comparisons reported in the reviewed studies. Table 1 summarizes the work assigned to LLMs in each use and the central questions for assessing their performance.

Table 1: Research uses of LLMs in economics and finance

Research use	**Work assigned to LLMs**	**Questions about performance**
Text analysis and variable construction	Classify texts and extract information from policy documents, corporate disclosures, news, social media, and online recruitment data for use in constructing economic and financial variables	How accurately does the model classify texts and extract the required information? Do the resulting variables capture the intended economic concepts? When these variables enter an econometric model, how does measurement error affect estimation and inference?
Economic forecasting and decision support	Use supplied texts or knowledge acquired during training to form predictive signals or forecasts for financial markets, macroeconomic aggregates, firm outcomes, and occupational transitions, and to inform investment, policy, or operating judgments	Do LLM-based signals or forecasts add predictive information or improve forecast performance relative to relevant existing predictors or methods? When they are used for trading or other economic decisions, do they improve the corresponding outcomes?
Simulation of economic agents	Generate choices or behavioral responses in experimental settings, simulate or impute survey responses, and act as agents making decisions in market simulations	To what extent do the generated responses, choices, and interaction outcomes capture the economic behavior that the study seeks to simulate?
Research assistance and automation	Assist with literature processing, data analysis, programming, theoretical analysis, and writing, and contribute to multistep research workflows	Which research tasks and workflows can LLM-based systems complete, and with what quality? Do they improve research efficiency, and what human work remains necessary?

The four uses in Table 1 differ in how the model's work contributes to economic research. For measurement, classifications, extracted information, and matching assessments are used to construct economic variables. In forecasting, researchers use LLM-based signals in forecasting models or ask the model to predict an economic outcome directly. In simulations, LLMs generate records of agents' answers and actions. Research assistance provides materials such as summaries, code, and candidate explanations. The assessment must therefore consider both the work performed and its intended use.

2.2 Assessment by Research Use

Benchmarks such as the Financial Language Understanding and Prediction Evaluation Benchmark (FLARE), introduced in PIXIU, provide shared tasks and scoring rules for comparing models on financial language and prediction tasks [53]. In an economic study, what needs to be assessed also depends on the research purpose and on the tasks performed by the model, its tools, and the researchers. Human labels, forecast benchmarks, and other relevant comparisons help assess how well these tasks are performed and whether the results serve the intended use.

To assess whether an LLM classifies texts, extracts information, or assesses a specified match accurately, researchers can compare its results with human assessments of the same inputs under a common coding rule. For classification tasks, accuracy and F1 scores summarize agreement

with the reference labels. Agreement on individual texts, however, does not by itself show that a variable constructed from these results captures the intended economic concept. The variable also depends on thresholds, aggregation rules, and corpus composition. Researchers can assess its economic meaning through comparisons with a human-constructed indicator, an independent survey measure, or another measure of the same concept, and examine whether alternative construction choices materially change its interpretation. Relevant economic outcomes and documented historical events can help interpret the variable, but provide less direct evidence about the concept itself. A further issue arises when the variable enters an econometric model. Measurement error can affect parameter estimates and statistical inference. Ludwig et al. [43] develop a correction for downstream estimates using discrepancies between LLM measurements and reference measurements in a validation sample. Section 3 reviews how classifications, extracted information, and matching assessments are used to construct economic variables and discusses the implications of measurement error for their use in econometric analysis.

For forecasting, the question is whether LLM-based signals add predictive information beyond existing predictors or improve forecasts relative to established methods. Researchers can test whether a signal remains associated with subsequent outcomes after controlling for existing predictors. They can also compare out-of-sample forecast errors with a relevant benchmark, whether the LLM makes the forecast directly or supplies a signal to another forecasting model. The first comparison assesses additional predictive information within the study's design. The second assesses improvements in forecast performance. For probability forecasts, researchers can also examine whether predicted chances agree with observed frequencies. Interpreting a prediction test requires specifying the population on which the model is meant to perform and how the evaluation sample was drawn [43]. For a test of genuinely unseen outcomes, researchers also need to exclude the model's prior access to information about those outcomes through training or updating. The input data for a historical forecast must also have been available at the stated forecast date. A claim that the forecast could actually have been made at that time additionally requires the model version to have been available. A later-released model may therefore be useful for a retrospective analysis without establishing historical implementability. Benchmark comparisons should match the prediction target, horizon, and information date. When the purpose is trading or another economic decision, assessment must also ask whether using the signal improves the relevant outcome relative to an appropriate alternative, accounting for the costs and risks of that use. Section 4 examines the predictive findings and the conditions that shape their interpretation.

Assessing a simulation begins with the behavior it is intended to represent. For an experimental pilot, researchers may ask whether generated choices change with treatments or incentives in the expected direction. When the aim is to represent human participants, choice shares and treatment responses can be compared with those in a comparable human experiment. Survey completion instead calls for comparing predicted answers with the same respondents' observed answers. Representing a population requires attention to response distributions and differences across relevant groups, rather than individual predictions alone. Market simulations introduce interactions among agents. Their decisions jointly produce prices and trading pat-

terns, so individual behavior and market outcomes need separate assessment. Researchers can examine whether agents follow assigned strategies and use the information available to them, and how changes in roles, incentives, and trading rules affect the resulting market. Such an artificial environment can be used to study the consequences of specified rules without reproducing an actual market. Claims about human traders or real markets require corresponding human or market comparisons. Changing incentives or institutions changes the economic situation. By contrast, changing labels or equivalent wording while holding that situation fixed tests sensitivity to presentation. Section 5 examines what these different comparisons reveal about simulated behavior.

For research assistance and automation, assessment starts with the work the system actually performs and the quality of what it produces. Researchers need to distinguish tasks completed by the model and its tools from choices, checks, and corrections made by people. Quality can then be assessed against requirements suited to each task. For example, citations can be traced to their sources, and code can be tested against specified requirements. Linking several tasks into an executable workflow shows that the steps can be carried out together. When the system proposes an economic explanation, however, assessment also concerns the explanation's assumptions, derivations, implementation, and empirical support. Internal scores and fit to calibration targets can help screen candidates, but do not by themselves establish that an explanation is supported. Whether the assistance improves efficiency requires a comparison with unassisted work on the same task under common quality criteria. This comparison should consider quality and errors alongside total completion time, including the time spent checking and correcting the work. Gains on individual tasks do not establish productivity gains for a complete economic research process. Section 6 reviews assistance with individual research tasks and multistep workflows, together with the available evidence on quality, efficiency, and the work still performed by researchers.

Researchers should distinguish the data used to develop a method from the data used to evaluate it. Fitting a model, calibrating agents, and choosing prompts are all part of method development. Performance on the data used for these choices does not by itself show how well the method will work on new observations. Evaluating performance on new observations therefore requires data that were not used to develop or select the method. Comparisons across models, prompts, samples, and task specifications help identify the conditions under which the findings hold. Causal interpretation requires an identification strategy suited to the economic question. A well-measured variable or an accurate forecast cannot supply that strategy. The following application sections discuss these methodological issues where they affect the interpretation of results.

3 Text Analysis and Variable Construction

LLMs help researchers identify policy intentions, corporate plans, and labor market matches from language. In several applications, interpreting context improves classification relative to fixed word lists or other classifiers. The resulting classifications and matching judgments

provide inputs for economic variables. Researchers then decide how to code and combine them. Construction choices shape the resulting variables, while comparisons with other measures help assess whether they capture the intended concepts. Measurement error can also affect estimates based on these variables. The applications below examine these benefits and difficulties in policy uncertainty, central bank communication, corporate disclosure, and labor market matching.

3.1 Earlier Approaches to Text Analysis

Early dictionary methods show how textual tone can be linked to asset prices. Tetlock [50] scores the daily *Wall Street Journal* market column with the Harvard General Inquirer and matches its tone to next-day returns. The resulting pessimism measure predicts short-horizon returns followed by mean reversion, a pattern consistent with investor overreaction. The limitation becomes apparent when a general dictionary is applied to specialist texts. Loughran and McDonald [42] find that 73.8% of the words marked as negative by the Harvard dictionary are not negative in annual reports, where *liability* and *tax* are ordinary accounting terms. If the frequency of such misclassified vocabulary varies systematically across firms or industries, the resulting sentiment measure can absorb those characteristics. Their dictionary, tailored to financial texts, reduces this domain mismatch but retains a fixed mapping from words to sentiment. Its interpretation therefore depends on whether that mapping remains appropriate for the texts being studied.

To reduce this dependence on predefined word lists, supervised learning uses a labeled subset or an observable reference outcome to learn which expressions distinguish categories. The learned mapping can then be applied to texts whose outcomes are unknown, without requiring the researcher to assign a direction to every word. Gentzkow and Shapiro [20] use legislators' party affiliation to identify partisan two- and three-word phrases in the 2005 *Congressional Record*. They relate phrase use to legislators' ideology, proxied by the Republican vote share in their districts, and then apply the mapping to newspapers. They use the resulting index of media slant to study how reader ideology, newspaper language, and market demand are related. The index, however, directly measures the similarity between newspaper language and partisan congressional speech rather than a context-free form of political bias. Its interpretation therefore depends on whether congressional language provides an appropriate reference and whether the mapping remains stable across outlets and periods.

When texts lack ready-made labels, topic models offer another route. One widely used approach, latent Dirichlet allocation (LDA), represents each document as a mixture of unnamed topics rather than assigning it to a single category [19]. Hansen et al. [25] apply LDA to 149 transcripts from Federal Open Market Committee (FOMC) meetings, using the resulting topic proportions to measure what members discuss, how concentrated their contributions are, and how closely their remarks resemble the subsequent discussion. Combining these measures with the 1993 transparency change and a difference-in-differences design, they find that transparency reduced spontaneous exchange and altered the contributions of members with less experience differently across economic and policy discussions. Yet the resulting measures remain sensitive to modeling choices. The baseline analysis uses 40 topics for interpretability, whereas an out-

of-sample criterion favors 70, showing that the estimated partition depends on the objective used for model selection.

Pretrained language models reduce the need to learn language representations from scratch for each task. Bidirectional Encoder Representations from Transformers (BERT) and related encoders first learn contextual word use from large unlabeled corpora. Further training maps these representations to the categories required for a particular task [15]. Liu et al. [38] apply this approach in FinBERT, combining general and financial pretraining with training for financial sentence boundary detection, sentiment analysis, and question answering. Generative LLMs also learn from large text corpora, but they produce text in response to a prompt. For many tasks, researchers can ask them to classify or extract information by providing instructions and examples, without training a separate classifier [4; 48]. The applications below use these approaches to interpret economic texts.

3.2 Remeasuring Economic Policy Uncertainty

Economic policy uncertainty (EPU) offers a direct comparison between keyword rules and models that interpret articles in context. Baker et al. [3] classify an article as relevant to EPU when it contains at least one term from each of three groups referring to the economy, uncertainty, and policy. They use ten U.S. newspapers to construct a monthly index beginning in 1985 and six newspaper archives to extend the series back to 1900. The rule is transparent and reproducible, but it does not use the surrounding text to interpret the terms. It can therefore count articles in which the terms occur in unrelated contexts and miss relevant articles that express policy uncertainty in different wording. Exact matching also becomes less reliable when vocabulary changes or optical character recognition introduces errors into historical text. These limitations motivate classifiers trained on human labels that interpret terms in their surrounding context.

Hartley [26] evaluates these approaches at two levels, the classification of individual articles and the monthly EPU indices constructed from those classifications. Using 10,393 articles audited by human coders, the study compares the canonical bag-of-words (BOW) method and a support vector machine with BERT, Longformer, and Llama 3.1 (8B). Longformer is a transformer encoder designed for long documents, while Llama 3.1 (8B) is a decoder LLM with 8 billion parameters. To test whether reading more of an article improves classification, the study compares predictions made with different input lengths. At the longest input length examined, 2,048 tokens, Longformer and Llama 3.1 have similar F1 scores.* The paper summarizes performance at this input length as a 46% relative improvement over BOW.

To assess whether classification gains carry through to monthly measurement, the study trains and selects the models on articles from 1900 to 2005 and tests the resulting indices on 2006–2012. The correlation between the BOW index and the benchmark based on a human audit falls from 0.47 to 0.24 across these periods. The average correlation for the Longformer

*For EPU classification, precision is the share of articles classified as EPU that the reference coders also label as EPU. Recall is the share of reference EPU articles that the model identifies. F1 combines these two shares as their harmonic mean.

and Llama 3.1 indices changes only from 0.78 to 0.76. Within the audited U.S. corpus, these comparisons show improved article classification and closer correspondence with the human audit benchmark at the monthly level.

The multilingual extension applies the EPU classification task to other languages. Hartley [26] fine-tunes Llama 3.1 on the English audit data and evaluates it on validation and test articles translated into 29 languages. After a separate threshold is selected for each language, its F1 score exceeds the English BOW benchmark in all 29 languages, with relative improvements ranging from 21% to 46%. The same classifier is then used to construct exploratory EPU indices from separate news corpora in Africa and Bangladesh. The African index combines country series from Voice of America reporting using GDP weights and reflects that outlet's coverage of African policy developments. The Bangladesh index uses a larger corpus from local news outlets. Both series show peaks around selected policy and macroeconomic events. The author presents these indices as exploratory applications of a common multilingual classifier.

The historical extension changes the period covered rather than the language. Longformer classifies more than 360 million digitized U.S. newspaper articles from 1800 to 1964 to construct a longer EPU series. Over the overlapping period of 1900–1964, its correlation with the Baker–Bloom–Davis index is 0.33, indicating positive but limited co-movement between the two series. The main extension is the additional nineteenth-century coverage.

3.3 Central Bank Communication

The focus now shifts from uncertainty reported in the news to the information conveyed by central banks themselves. A starting point is whether an LLM can reproduce human judgments of policy stance. Hansen and Kazinnik [23] divide FOMC statements from 2010 to 2020 into sentences and assign them to five ordered categories from dovish to hawkish. Their benchmark contains 500 randomly selected sentences, each independently labeled by three reviewers. The study compares zero-shot GPT-3.5 and GPT-4 with BERT and several dictionary methods on these same sentences. GPT-4 has the highest overall classification accuracy and the lowest mean absolute and root mean squared errors relative to the human labels.

A broader coding scheme can describe more than whether a statement is dovish or hawkish. Silva et al. [49] classify communication across 169 central banks along four dimensions. Their corpus contains 74,882 documents and approximately 21 million sentences from 1884 to 2025. The four dimensions are topic, communication stance, audience, and sentiment. Communication stance distinguishes forward-looking from backward-looking statements. Sentiment distinguishes hawkish, dovish, neutral, risk-highlighting, and confidence-building language. The authors fine-tune the multilingual BGE-M3 sentence transformer and train two joint classifiers, one combining topic with communication stance and the other combining audience with sentiment. Training material includes synthetic sentences, sentences written by experts, and authentic central bank sentences. A validation set of 360 authentic sentences annotated by experts is excluded from training but used both for model selection and for comparison with GPT-4o. On this set, the specialized classifiers have higher macro-F1 scores[†] than GPT-4o across all four

[†]Macro-F1 gives equal weight to each class's F1 score. Overall accuracy is the share of all classifications that

dimensions, although GPT-4o has slightly higher accuracy for topic and sentiment.

From these labels, Silva et al. [49] construct measures of net policy sentiment, straightforwardness, explanation, and net confidence. We use net policy sentiment to illustrate how these labels are aggregated into an economic variable. It measures the balance between hawkish and dovish sentences, with separate components for forward-looking and backward-looking communication. For each component, the index subtracts the number of dovish sentences from the number of hawkish sentences and divides the difference by their sum. To examine the economic information captured by these components, the authors relate them to changes in policy rates, short-term Treasury bill rates, and long-term bond rates in the following month. The forward-looking component is positively associated with all three rate changes, whereas the backward-looking component is associated only with changes in the policy rate. In a separate sample of 15 economies, the forward-looking component is positively associated with subsequent overnight index swap rates at all maturities examined, whereas the backward-looking component is not. Together, these contrasts provide external evidence that the forward-looking index contains information about subsequent interest rates.

Interpreting these indices across central banks also raises questions about the languages and publication histories represented in the corpus. English accounts for 87.4% of the sentences. Silva et al. [49] compare predicted label distributions for documents in their original languages and for English translations produced by GPT-4o. The similar distributions indicate consistency across language versions. Publication coverage also changes as countries and communication outlets enter the corpus at different dates. The authors also display forward-lookingness by months since each outlet's first available publication. The upward trend remains visible on this axis showing elapsed time. This display aligns the start of publication histories without holding the country, language, or outlet composition fixed.

Rather than apply one coding scheme across many central banks, a study can tailor its categories to a single policy framework. Liu et al. [37] apply this approach to the PBOC, whose policy framework combines quantity and price instruments with multiple objectives and constraints. After screening 5,944 website records from 2005 to 2025 and adding missing monetary policy and financial stability reports, they construct a corpus of 682 communications. Using the LLM model, they divide the texts into semantically complete units and distinguish forward-looking statements from retrospective descriptions. The model separates monetary policy statement from macroeconomic assessments and other topics. Within passages of monetary policy, it identifies the overall direction of policy stance, growth and inflation objectives, quantity and price instruments, and constraints related to the exchange rate and financial stability. It then assigns easing, neutral, or tightening labels to the stance, objectives, and instruments, and rates each constraint as high, normal, or low. The labels are aggregated into monthly indices of forward-looking net policy intention, divergence between quantity and price signals, and dispersion across the monetary objectives. Comparisons of the indices with documented developments in China's monetary policy help interpret the resulting measures.

are correct.

3.4 Corporate Disclosure

Corporate disclosures contain managers' responses and plans, but extracting this information often requires interpreting language in context. Non-answer classification in earnings calls illustrates the problem. The task is to determine whether a response addresses the analyst's question, not merely whether it contains a particular phrase. Gow et al. [22] use a classifier based on regular expressions, but fixed rules can miss equivalent wording and responses whose meaning depends on the preceding question. de Kok [12] addresses this problem using 1,152,505 pairs of questions and answers from 63,959 earnings calls held between 2013 and 2022. A keyword filter first removes pairs with no matched terms. A fine-tuned GPT-3 Babbage model then excludes additional unlikely cases. Zero-shot ChatGPT evaluates the remaining complete exchanges, and fine-tuned GPT-3.5 refines the non-answer classifications. This sequence reserves contextual evaluation for the cases retained after initial screening.

After selecting the final method and prompts, de Kok [12] draws and manually classifies an evaluation sample of 500 pairs of questions and answers. The final method achieves 96% overall accuracy, compared with 86% for the regular-expression method on the same sample. Precision, recall, and F1 for the non-answer class are each 0.87. The paper notes that the boundary between an answer and a non-answer sometimes requires judgment. The comparison therefore measures agreement with the author's coding, rather than agreement among independent annotators.

Beyond identifying whether managers answer a question, another application measures the investment plans expressed in their statements. Jha et al. [29] divide earnings call transcripts into chunks of about 2,500 words and use ChatGPT 3.5 to classify the expected direction and magnitude of changes in capital spending. They average the resulting scores to obtain an investment measure for each firm and quarter and compare it with capital expenditure forecasts from the Duke CFO Survey. The matched sample contains 1,338 firm-quarter observations. In a simple regression, the text score explains 1.4% of the variation in the survey measure, and the positive association remains statistically significant after industry and time fixed effects are included. The comparison supports the interpretation of the firm-quarter score as a measure of managers' disclosed investment plans within the matched sample.

3.5 Measuring Labor Market Match Quality

Measuring how well workers' education and experience fit job requirements is important for studying labor market mismatch and wage differences. Yet classifications based only on whether two occupation codes are identical overlook skills shared by different occupations. Chen et al. [9] use the meanings of major, occupation, and industry names to construct alternative match indicators. They prompt GPT-3.5-turbo-0125 to act as a human resources specialist. For a major and job title, the model judges whether a graduate of that major can perform the job, using only the supplied names and without assuming other applicant characteristics. It chooses among "definitely can," "probably can," "probably cannot," and "definitely cannot." The researchers assign 1 to either affirmative answer and 0 to either negative answer. They use the same coding rule for analogous questions about the occupations and industries of the worker's

previous job and the job applied for. A pair can therefore be coded as a match even when its two category names differ.

The authors first evaluate the indicators using application records from the Chinese job platform Zhaopin.com. They make two distinct comparisons. The GPT indicators are positively correlated with conventional match measures, showing agreement about which pairs are better matched. The wage regressions then ask whether the indicators contain information beyond those measures. After controlling for applicant and job characteristics and the corresponding conventional indicators, the GPT occupation and industry matches remain positively associated with applicants' expected wages. The result differs for the major–job match. Its coefficient is no longer statistically significant when the Duncan index, constructed from the distribution of applicants' majors across occupations, is included. Thus, the GPT measures supplement simple indicators of an identical occupation or industry, but the major–job measure does not add a statistically significant wage association beyond the Duncan index in this comparison.

The GPT approach is also useful when a survey contains too few observations to estimate match quality reliably from category frequencies. In the smaller China Labor-Force Dynamic Survey (CLDS), many combinations of majors and occupations have few observations. Rather than estimate their frequencies, the GPT indicator uses the names of the major and occupation to assess their compatibility. The authors compare it with a mapping of majors to occupations prepared by job analysts. The correlation is 0.555, and each measure is positively associated with workers' actual wages in separate regressions. These comparisons support using the GPT indicator to measure major–occupation matches when the survey provides limited information on how frequently each combination occurs.

3.6 From Labels to Economic Variables

Text analysis begins with a definition of what the model should identify. The researcher specifies the categories, the coding rule, and the form of the answer. Even when the input text is unchanged, the way the task is presented can affect the model's classifications [36]. These differences can carry through to economic estimates. Ludwig et al. [43] find that changing the model or prompt alters some regression coefficients calculated from labels for financial news and congressional bills.

Once the model has produced its classifications or scores, researchers decide how to use them in an economic variable. Researchers must decide whether to convert predicted probabilities into categories or retain them as continuous scores. A higher classification threshold generally reduces false positives at the cost of more false negatives. When several texts contribute to one observation, the aggregation level, the weights assigned to individual texts, and the composition of the corpus also shape the resulting measure. For example, Hartley [26] holds the classifier fixed and shows that alternative thresholds and aggregation rules change the baseline and amplitude of the resulting EPU index. The same classifier can therefore produce different measures depending on these construction choices.

The resulting variable can then be compared with another measure of the intended economic concept. The EPU indices constructed by Hartley [26] track a benchmark based on a human

audit more closely than the index based on keywords. Jha et al. [29] find a positive relationship between their firm-quarter measure of investment plans and investment expectations in the Duke CFO Survey. These comparisons address the meaning of the variable, rather than the accuracy of an individual classification. Other economic relationships provide less direct evidence about that meaning. Silva et al. [49] relate forward-looking net policy sentiment to subsequent changes in policy and market interest rates, while Liu et al. [37] compare their indices with documented monetary policy developments. Such relationships help interpret the indicators, but do not necessarily compare two measures of the same concept. In the labor market application, the expert mapping of majors to occupations provides a reference for match quality, whereas wage relationships show how the measure relates to an economic outcome [9].

Applying a measurement method to another language, period, or corpus changes the available comparisons. In Hartley [26], classification on translated U.S. audit articles is compared with reference labels, while the African and Bangladeshi indices are interpreted through their news coverage and movements around selected events. The historical EPU series is compared with the original Baker–Bloom–Davis index. Silva et al. [49] compare classifications of original documents with those of their English translations to assess consistency across language versions.

When the resulting variable enters a regression, measurement error can affect coefficient estimates and statistical inference. Agreement with an external measure does not remove errors introduced by classification, thresholding, or aggregation. Under classical assumptions, error in a dependent variable mainly lowers precision, while error in a single regressor attenuates its estimated coefficient toward zero. With multiple regressors, measurement error can also bias coefficients on correlated covariates, and the direction of bias need not be toward zero. Error that varies systematically with time, document length, language, or industry can create apparent trends or group differences. Overall classification accuracy does not show whether those errors are related to the other variables in a regression.

To assess and correct this bias, Ludwig et al. [43] use validation data. They illustrate the approach when an LLM measurement is the dependent variable. On a random validation sample, the researcher obtains both the LLM measurement and a reference measurement, subtracts the reference measurement from the LLM measurement, and regresses that difference on the explanatory variables. Subtracting these estimated error coefficients from the regression coefficients estimated in the primary sample corrects the estimated bias under the assumptions of their framework. The aim is to recover the coefficients that would be estimated using the reference measurements. Standard errors also need to reflect that the correction was estimated from a sample.

Correcting measurement error does not, by itself, give a regression a causal interpretation. If an LLM variable is used to estimate a policy effect, the study still needs a comparison that separates the policy from other influences. For example, a difference-in-differences design can compare changes in the measured outcome between groups affected and unaffected by a policy. In their transparency study, Hansen et al. [25] combine text measures with such a design.

4 Economic Forecasting and Decision Support

LLMs can supply information for forecasts by interpreting texts or using knowledge acquired during training. The studies below find signals associated with later returns and investment, and some report lower macroeconomic forecast errors or better probability forecasts for occupational transitions than existing methods. The gains are not uniform. They vary across samples, periods, and model configurations. Historical comparisons also depend on what information the model and researcher could use. Even a predictive signal can be unprofitable once the costs of acting on it are included. We examine these distinctions in financial markets, macroeconomic monitoring, firm investment, and occupational transitions.

4.1 Financial Market Forecasting

For financial markets, the first question is whether model assessments of news contain information about subsequent returns. Kirtac and Germano [31] compare news signals from language models with the Loughran–McDonald [42] dictionary using 965,375 U.S. company news articles. They fine-tune the open pre-trained transformer (OPT) and BERT to classify the sign of the aggregate excess return from the news publication date through the following two days. On a randomly held-out article sample, the OPT model with 2.7 billion parameters has the highest accuracy, 0.744, ahead of BERT, FinBERT, and the dictionary. A separate regression relates the model scores to returns on the following trading day. The portfolio exercise rebalances daily and reports positive cumulative returns after a cost of 10 basis points per trade.

Information learned during training can distort a historical backtest if the model already knows what happened after the news being evaluated. Glasserman and Lin [21] investigate this issue using GPT-3.5. Using headlines collected from the web and a separate Thomson Reuters news sample, they compare strategies based on original headlines with strategies based on headlines from which company names and other identifying details have been removed. Removing these identifiers reduces the model's ability to associate the headline with a known company. Knowing subsequent events could favor the original headlines, but general knowledge about the company could also interfere with interpreting the news itself. Within the training period, the anonymized headlines yield higher mean returns in both datasets. The authors interpret this as evidence that interference from company knowledge outweighs the benefit of knowing later events in this comparison. After the training cutoff, anonymized headlines yield a higher mean return in the Thomson Reuters sample, but the difference is only marginally significant. In the sample collected from the web, anonymized headlines yield a lower mean return, and the difference is not statistically significant.

To reduce the risk that a model predicts returns by recalling later events encountered during training, Lopez-Lira and Tang [41] evaluate news and returns from after the model's stated training cutoff. They use GPT-4-0314, a fixed GPT-4 snapshot with a September 2021 training cutoff, to assess U.S. company news headlines published from October 2021 to May 2024. GPT-4 scores correspond closely to the direction of the initial market reaction and predict return drift over the next one to two trading days. The drift is more pronounced among smaller stocks and after negative news, a pattern consistent with delayed information diffusion and limits to

arbitrage.

The trading exercise in the same study examines whether this return signal can be used profitably after costs. The initial market reaction tests whether GPT-4 interprets the direction of news correctly, whereas the potentially tradable return begins with the subsequent drift. For overnight news, the long-short strategy earns 34 basis points before costs from the first trading day's open to its close and 19 basis points from that close to the following day's close, with no significant return thereafter. Daily rebalancing generates turnover of about 190%, and a round-trip cost of 20 basis points makes the strategy unprofitable. Performance also weakens over time. The annualized Sharpe ratio falls from 6.54 in 2021Q4 to 1.22 in January–May 2024. The signal therefore predicts a short-lived return pattern, but that predictability does not translate into profitability for this strategy at the assumed trading cost.

Lehner and Lopez-Lira [34] ask whether knowledge acquired during training can help a model predict firms' future performance. They give twelve frozen model checkpoints firm identifiers and industry information, without supplying news articles or using web searches. Each model assesses expected revenue growth, profitability, and operating risks over the twelve months following its stated knowledge cutoff, relative to industry peers. The design treats these assessments as reflecting information learned before that cutoff. The outlook scores are standardized within Global Industry Classification Standard sectors. Each model is also asked 100 questions about events after its cutoff. All score zero, so the probe detects no obvious knowledge of those later events. In the primary GPT-4.1 analysis, higher scores predict stronger twelve-month revenue and net income growth and more positive analyst target price revisions. The scores are also positively associated with one-month cross-sectional returns after controls for standard firm characteristics and market-implied valuation measures.

Whether these forecasts could have been made at the time depends on when the model was available. The study measures returns from each model's stated knowledge cutoff, including dates before some models were released. It therefore tests predictive information retrospectively, without showing that researchers could have used those models to make the same forecasts then. The trading evidence also covers limited periods. Twelve checkpoints correspond to six distinct cutoff months in the cross-sectional return sorts. With six observations, the authors report a positive but statistically insignificant intercept in the capital asset pricing model (CAPM). They do not estimate an alpha using the larger Fama–French five-factor model augmented with momentum. A separate long-only mean–variance portfolio application covers at most 30 months, almost entirely in a rising market, and reports returns before trading costs. Thus the first exercise offers limited evidence about performance over time, while the second does not establish profitability after costs.

4.2 Macroeconomic Monitoring

The macroeconomic studies ask more directly whether adding information from text reduces forecast error relative to an existing benchmark. Allard et al. [2] examine inflation forecasting using a large news corpus. They use GPT-4 Turbo to label the direction of inflation reflected in news, fine-tune Fin-distil-RoBERTa on those labels, and aggregate about 74,000 articles related

to inflation into a monthly index. On the 2020–2023 test sample, its weighted F1 score of 0.57 against GPT-4 Turbo labels is the highest among the classifiers compared in the study. When the resulting index is added to the Cleveland Fed inflation model, the out-of-sample root mean squared error changes only from 0.0409 to 0.0407. The news index therefore does not produce a statistically detectable improvement over the benchmark model.

Monthly releases of the Purchasing Managers' Index (PMI) report numerical survey results alongside written descriptions of business conditions. de Bondt and Sun [11] ask whether GPT-4o scores of this text improve GDP nowcasts for the euro area. In most in-sample specifications, the text scores add information beyond the numerical composite output PMI. The out-of-sample forecasts are compared with two benchmarks. The first is the GDP nowcast series from European Central Bank (ECB) and Eurosystem staff projections, and the second is Eurostat's first published GDP estimate. The study explicitly matches the text inputs to their availability at each benchmark date. Because staff projections are prepared while only the first two months of the quarter's PMI data are available, that comparison uses two monthly releases. The later Eurostat comparison uses all three. Forecast errors are measured against quarter-on-quarter growth in real GDP using the latest series available when the authors wrote the paper.

Excluding 2023 and 2024, the augmented models reduce average root mean squared error by 8%–9% relative to staff projections and 17%–18% relative to the first GDP estimate. Including those two years sharply weakens the gains and reverses them in several specifications, although the model estimated using a rolling window still improves on staff projections over the full period. The gains thus vary across evaluation years, benchmarks, and estimation windows.

The third application combines information about inflation in social media with topic features in a joint forecasting model. Fan et al. [18] use about 120 million Weibo posts from 2019 to 2023, retaining 95.45 million after cleaning. They label a stratified sample of 20,000 posts with GPT-4 Turbo and use Chinese BERT models to identify and score content related to inflation. The daily index is averaged over the first ten days, the next ten days, and the remaining days of each month to form three inflation proxies. The authors also extract monthly topic features using LDA. Their LLM+LDA specification links a monthly CPI equation to a vector autoregression with exogenous variables for the three proxies. Both equations include the LDA features, and correlation between their innovations allows the proxies to contribute information to the CPI prediction.

The study first compares the joint forecasting model with conventional benchmarks. Without unemployment, the average relative root prediction mean squared error is 0.848 against an autoregressive (AR) benchmark. With unemployment included in both the forecasting model and an autoregressive benchmark with exogenous variables (ARX), the corresponding ratio is 0.811. The two ratios imply average error reductions of 15.2% and 18.9%, respectively.

A second comparison asks what the joint structure adds to a CPI forecast that already uses CPI lags and LDA topic features. Including the equations for the inflation proxies improves average forecast accuracy. The average relative error falls from 0.865 to 0.848 without unemployment and from 0.827 to 0.811 with it. All evaluations end in December 2023 and use the last 8–15 months of the 60-month CPI series, so the samples overlap and are nested. The joint

model improves on the direct CPI forecast for three of the eight evaluation lengths in each setting.

To judge whether these forecasts could have been made at the time, the inputs must be available by the forecast date. The equations use text signals from the month whose CPI is being predicted and, in the extended specification, that month's unemployment rate. These inputs could be used before the official CPI release only if they were already available when the forecast was made. The paper and available supplement do not match every input's release date to a historical forecast date. The reported gains therefore show what can be predicted with those inputs, without establishing that the same forecasts could all have been produced in real time.

4.3 Firm-level Analysis and Investment Forecasting

Firm investment forecasting asks whether managers' disclosed plans help predict subsequent capital expenditure beyond standard investment variables. Jha et al. [29] address this question using the earnings call score introduced in Section 3.4. The earlier comparison with the Duke CFO Survey concerns whether the score captures managers' stated investment plans. The analysis here asks whether the same score predicts subsequently realized capital expenditure.

They test this relationship in a 2006–2020 sample of 74,586 firm-quarter observations. Higher scores are associated with greater capital expenditure in the calendar quarter following the earnings call. The relation remains significant after controlling for total q (a version of Tobin's q that incorporates both physical and intangible capital), lagged capital expenditure, other firm characteristics, and firm and time fixed effects. The association remains statistically significant at horizons extending to nine quarters after the call. A separate test using calls from 2021Q4 to 2022Q4, after ChatGPT 3.5's stated training cutoff, also finds a positive association with next-quarter capital expenditure. This five-quarter exercise reduces concern that the model had encountered the later outcomes in its training data. Together, these tests show that managers' disclosed investment plans contain information about realized capital expenditure conditional on the included controls. They do not compare out-of-sample forecast errors with those of a benchmark model.

4.4 Predicting Occupational Transitions

Forecasting in labor markets often concerns whether a worker will remain in the same occupation or move to another, given the worker's previous jobs and characteristics. Survey data record these histories, but contain relatively few examples of the many possible career paths. Du et al. [16] ask whether language models can use descriptions of those histories to improve predictions of the next recorded occupation. Their LABOR-LLM framework represents job histories and demographic characteristics as text and compares direct use of Llama-2 with models fine-tuned on survey records. The data come from the Panel Study of Income Dynamics (PSID) and the 1979 and 1997 cohorts of the National Longitudinal Survey of Youth (NLSY79 and NLSY97). All records for a person remain in the same training, validation, or test sample. The two benchmarks are CAREER, a transformer trained to predict occupational histories,

and a simpler Markov model that uses observed frequencies of transitions from the worker's previous occupation.

The authors evaluate each model by the probabilities it assigns to the occupations that workers actually enter or remain in. Perplexity combines these probabilities across the test sample, with a lower value indicating better probability forecasts. Direct prompting performs worse than CAREER, even when the prompt lists the available job titles. After fine-tuning, however, the Llama-2 model with 70 billion parameters has lower test perplexity than CAREER in all three surveys. The authors also check whether predicted chances of changing occupation agree with observed frequencies. They group observations by their predicted probability of a change and compare each group's average prediction with its actual share of changes. The fine-tuned model is better calibrated than CAREER, including in comparisons for workers with and without college education. These results support adapting a language model to the survey population when predicting occupational transitions, rather than relying on direct prompts alone.

4.5 Predictive Information and Economic Value

Interpreting the reported gains requires distinguishing the LLM's work from that of the rest of the forecasting system. In the market studies, researchers use LLMs directly to score news or assess firms' prospects. In Allard et al. [2] and Fan et al. [18], LLMs instead label training data for smaller classifiers, and conventional forecasting models use the resulting indices. For example, Fan et al. [18] compare their joint model with a CPI forecast that already includes LDA features. That comparison measures the contribution of adding the equations for the inflation proxies, not the contribution of LLM labeling alone.

Whether the LLM predicts directly or supplies labels for another model, historical tests also need to account for its training knowledge. In a historical prediction exercise, that knowledge may include later events that reveal the outcome being predicted. This training leakage can make recall appear to be prediction. Ludwig et al. [43] show that GPT-4o reproduces some historical financial news headlines even when instructed to ignore information after their publication dates. A date instruction therefore does not reliably remove training knowledge. For an open model, researchers can retain a version released on a known date and keep its weights fixed. They then test it on documents first published after that date, which could not have been part of that version's training data [43]. For closed models, researchers can use a fixed snapshot and define the evaluation period using the provider's stated training cutoff. Lopez-Lira and Tang [41] use this approach in their news evaluation with GPT-4-0314.[‡]

Information supplied during use creates a second route by which later knowledge can enter a historical forecast. A prompt may contain background documents as well as the text being

[‡]A study of a fixed historical collection has a different goal. It asks how the model performs on that collection, including documents it may have encountered during training, rather than how it performs on previously unseen material. Ludwig et al. [43] illustrate this distinction with the prediction of citation counts for papers in a defined collection of past publications. In their example, a random evaluation sample is used to estimate performance for the collection as a whole. Papers are not selected according to whether they appeared in the model's training data.

scored. The model may also obtain additional documents through connected search or retrieval tools. To reconstruct a forecast at a past date, researchers need to apply the same date limit to both the material they supply and the additional documents the model obtains.

Model availability is a separate condition. Lehner and Lopez-Lira [34] measure returns from the knowledge cutoff even when the model was released later, so their retrospective result does not establish that those models could have been used at the time.

Finally, the trading applications show that predictive information and economic value are separate outcomes. Kirtac and Germano [31] report positive portfolio returns after a cost of 10 basis points per trade, whereas a round-trip cost of 20 basis points makes the daily strategy in Lopez-Lira and Tang [41] unprofitable. Because the strategies, samples, and cost conventions differ, these results do not compare model trading performance under the same conditions. They show how the economic value of a signal depends on the strategy used to act on it and the associated costs.

5 Simulation of Economic Agents

LLMs allow researchers to describe preferences, information, and constraints in words and generate the choices of simulated economic agents. This makes it possible to explore experimental designs, predict missing survey answers from existing records, and examine trading under specified market rules. Human observations improve some predictions, both by calibrating the mixture of agent types and by supplying information about individual respondents. But the resulting behavior also depends on the instructions and presentation of the task. The studies below show where these simulations resemble the behavior being studied and where the comparison remains incomplete.

5.1 Simulating Choices in Economic Experiments

Experimental simulations allow researchers to explore a design before collecting human data. A central question is whether model choices resemble those observed in human experiments. Horton et al. [27] construct *Homo silicus* agents by describing their information, preferences, and constraints in natural language. They study five experimental settings. In the price gouging experiment, agents assess whether a store's price increase is fair and judge larger increases less favorably. In the status quo experiment, agents choose among four ways to divide a research budget between automobile safety and highway safety. The options remain unchanged, but the researchers vary which one is described as the existing allocation. Agents are more likely to select an option when it is presented as the status quo.

The lottery experiment asks agents to report the fixed sum they regard as equivalent to a payoff under two payment rules. One rule pays a randomly selected amount. The other removes this risk by paying the lottery's expected value with certainty, while retaining the same presentation of amounts and probabilities. Even under the certain payment rule, agents can report values above or below the amount they will receive. The deviations are more pronounced in less capable models and under prompts assigning weak mathematical ability. The authors interpret the similar patterns under both rules as consistent with difficulty evaluating the task,

rather than risk alone. In the hiring experiment, agents choose between applicants with different experience and wage requests. Introducing a minimum wage shifts choices toward the more experienced applicant, matching the direction of the human field experiment.

The allocation games ask whether simulated agents choose each option as often as human participants do. Without an assigned preference, the simulated choice shares differ substantially from the human shares. Horton et al. [27] construct three types of agents by asking them to prioritize the total payoff, equality between players' payoffs, or their own payoff. The agents generally follow these assigned preferences. For each LLM, the researchers choose the proportions of the three types that best match human choices in single-stage allocation games. They then keep these proportions fixed and test the resulting samples in related two-stage games from Charness and Rabin [7]. Across models, the mean squared error between simulated and human choice shares falls from 0.182 without assigned preferences to 0.094 after calibration. The calibrated samples therefore predict human choices more closely in these related games.

5.2 Simulating Survey Responses

Survey applications use LLMs to predict individual answers and to study patterns across respondents. The studies below examine how personal information improves those predictions, whether simulated respondents update their beliefs after receiving new information, and whose responses the resulting samples resemble.

Brynjolfsson et al. [5] predict how much compensation a respondent would require to give up a good for one month, using the UK willingness-to-accept survey of Coyle and Nguyen [10]. They prompt models either to predict as an outside expert or to answer as the respondent. Performance is measured by the area under the receiver operating characteristic curve (AUC). With demographics alone, the weighted averages across models are 0.54 for the expert prompt and 0.55 for the respondent prompt. Adding the respondent's valuations of other goods raises these averages to 0.65 and 0.63, respectively. A regularized multinomial logistic regression benchmark reaches 0.70. Personal valuations improve both prompts, but their model averages remain below this benchmark. The authors also train GPT-4o-mini on 80% of the human records, pairing prompts with the observed valuations. With the expert prompt, demographics, and valuations of other goods, the fine-tuned model achieves an AUC of 0.79 on the held-out 20%.

In a second application, the authors use Panel Study of Income Dynamics (PSID) records from 2017 and 2019 for 1,000 household heads to examine whether earlier income improves predictions of reported income. They compare demographic information alone with the same information supplemented by income from the preceding survey. For GPT-4-turbo with instructions to reason step by step, adding prior income reduces mean absolute error from 42.0% to 25.3% of the sample standard deviation of income. The model and reasoning instructions are the same in both comparisons. As in the valuation application, information from the same respondent improves predictions beyond demographics alone.

Wu et al. [52] examine whether simulated respondents revise their beliefs as people do after receiving information. They replicate the inflation expectations experiment of Weber

et al. [51], which used the Nielsen Homescan Panel in ten survey rounds between 2018 and 2023. Respondents reported expectations before and after receiving information about past inflation, the Federal Reserve's inflation target, or its forecast. The replication uses GPT-4.1 with 200 demographic profiles from the New York Fed Survey of Consumer Expectations (SCE), held fixed across survey dates and experimental groups. The SCE provides a benchmark for expectations before information is supplied. The Nielsen experiment provides the benchmark for changes after the information is supplied.

Before information is provided, the simulated median inflation expectation rises and falls across survey dates in line with the SCE median. Regressing the simulated medians on the SCE medians gives a slope of 1.05. The authors then examine how strongly expectations respond to each type of information. For past inflation, the correlation between the simulated and human estimates across survey rounds is 0.62. For the Federal Reserve's target it is 0.71, but for its forecast it is -0.24. Pooling all three comparisons gives a correlation of 0.31. A separate comparison with prevailing inflation shows that both simulated and human respondents tend to react less to new information when inflation is higher. The simulation thus reproduces that broad pattern, while its response to the Federal Reserve's forecast does not track the human estimates across survey rounds.

Generating a synthetic survey sample also raises the question of whose answers it resembles. Erdem and Ashok [17] compare model answers with human responses from the International Test on Risk Attitudes (INTRA). They use fourteen questions about intertemporal choice, ambiguity aversion, risk, and loss. Models from the GPT series, Gemini, and DeepSeek answer the full set in 100 separate trials per model. For each question, the authors calculate a median across the model's repeated answers and a median across INTRA respondents in each country. Each model and country is therefore represented by fourteen median answers. In the reported clustering analyses, the model profiles group together, with Tanzania the only country profile in that group. The comparison helps identify which national profiles are closest to the models' median answers, but it does not show that the repeated model responses reproduce the diversity of answers within those countries.

5.3 Simulating Trading and Market Interaction

Market simulations add an interactive element because one agent's orders change the prices, holdings, and information facing later decisions. Lopez-Lira [39] implements this feedback in a simulated stock market with a shared order book. GPT-4o agents repeatedly submit market and limit orders while observing updated market conditions. Natural language system prompts assign strategies such as value investing, momentum trading, and market making. The agents generally maintain the prescribed strategies even when doing so produces losses.

The repeated interactions also generate different price paths. To examine price discovery without a known terminal date, the study reports two simulations of 15 rounds with the same composition of eight agents and a fundamental value of $28. One begins at $14 and the other at $56. From the lower starting price, prices and agents' stated valuations move toward $28. From the higher starting price, both remain substantially above the fundamental value.

Hansen and Lee [24] ask how LLM traders use private information alongside the public history of trades. They separately use Claude 3.5, Claude 3.7, Llama 3, and Nova Pro in a sequential trading experiment with eight agents, and compare their average behavior with an earlier experiment involving 32 London finance professionals. At each decision, an agent states what it would do after a good signal and after a bad one, given the price and previous trades. In the first treatment, traders know that the asset's value has moved to either a high or a low value, but cannot observe its realized value directly. A good private signal makes the high value more likely, and a bad signal makes the low value more likely; neither reveals the value with certainty. In this setting, buying after a good signal and selling after a bad one is always optimal.

The models give this pair of responses in 61.0% of decisions, compared with 45.7% among professionals. They choose the same trading direction after either signal in 9.4% of decision situations, compared with 19.0% among professionals. The paper calls this cascade trading because the action does not depend on the private signal. A further 29.5% of model decisions are partially rational. The agent buys after a good signal and does not trade after a bad one, or sells after a bad signal and does not trade after a good one. Under the rule in this treatment, the models follow private signals more consistently than the professionals.

The second treatment changes when a trader should rely on the private signal. The market maker is uncertain whether the asset's value has changed, while informed traders know that it has. The market maker cannot distinguish informed trades from noise trades, so prices adjust slowly. After enough trades in one direction, public trading history can outweigh the private signal. Buying after a bad signal or selling after a good one can then maximize expected payoff. The paper calls this optimal herding. In the baseline runs, herding is optimal in 36.6% of decision situations, but the models follow their private signals in 97.4% and never herd optimally. The models' written explanations follow the same pattern. Claude 3.7 classifies 63.1% of these explanations as valuing the asset using only the private signal and its accuracy; 9.5% refer to market direction or trading history.

Explicit instructions about when to follow trading history change the models' decisions. Along the paths generated with this prompt, herding is optimal in 81.5% of all decision situations. The models actually herd optimally in 47.4% of all decision situations. Expected payoff, averaged across sessions, trading periods, and models, rises from 3.80 to 14.95 lire. The payoff improvement follows a rule supplied by the researchers, not an independent discovery of that rule by the models. But even with this guidance, the models fail to exploit all opportunities for optimal herding.

The same study also tests whether decisions change when only the names of the signals change. Hansen and Lee [24] keep probabilities and payoffs fixed while relabeling otherwise identical signals. Replacing white and blue with green for a good signal and red for a bad one produces results broadly similar to the baseline on average across models. Reversing these colors, so that red means good and green means bad, leads to errors in roughly one quarter of decisions in each treatment included in the labeling test. An error here is a reversed pair of conditional responses, selling after a good signal and buying after a bad one. This average

result is driven primarily by Claude 3.5. Claude 3.7 gives similar results across the labeling schemes, and Nova Pro remains mainly rational under both relabelings. Thus economically equivalent descriptions can produce different decisions, depending on the model.

5.4 Roles, Information, and Task Presentation

The experimental and trading studies suggest that prompts should specify the preferences, information, and constraints relevant to the behavior being studied. Agents in Horton et al. [27] generally follow instructions to prioritize total payoff, equality, or self-interest, while the traders in Lopez-Lira [39] maintain their assigned strategies even when those strategies lose money.

Human data also help determine how simulated agents should be specified. In Horton et al. [27], human choices determine the proportions of the three preference types, and the calibrated mixture predicts choices more closely in related games. In Brynjolfsson et al. [5], personal information, such as other reported valuations or previous income, improves predictions of an individual's answers.

A synthetic survey should be assessed against the response patterns relevant to the research question. Wu et al. [52] use 200 demographic profiles drawn from the SCE and compare aggregate inflation expectations and responses to information with human survey results. Erdem and Ashok [17] compare national profiles built from the median response to each of fourteen questions to identify similarities between model answers and those of surveyed populations. For questions about differences across demographic groups or variation within them, the corresponding response distributions would provide the basis for comparison.

Market simulations also depend on whether agents use the information generated by trading. In Hansen and Lee [24], following private signals is optimal in the first treatment but misses profitable opportunities in the second, when earlier trades can make it optimal to follow the trading history instead. Explaining this rule improves decisions, but some opportunities remain unexploited. Researchers therefore need to specify how prices and earlier trades convey information and check whether agents revise their decisions accordingly. This separates a model's ability to follow a supplied trading rule from its ability to recognize when the information available calls for a different action.

6 Research Assistance and Automation

LLMs are being used to assist the search for economic explanations and the production of empirical research. Systems now generate candidate mechanisms for asset-pricing puzzles, draft finance papers from completed return tests, and propose questions that can be studied with available data. These uses go beyond editing language, although the systems still rely on researchers to judge the importance of a question and the credibility of an explanation. We distinguish assistance with individual tasks from workflows connecting several research stages. The final subsection considers related effects of LLM use on workers and labor markets.

6.1 Assistance with Individual Research Tasks

Within an economic research project, writing, programming, and mathematical work support the development and analysis of empirical or theoretical models. Korinek [32] illustrates LLM assistance with these activities through prompts and worked examples in six areas, including ideation and feedback, writing, background research, data analysis, coding, and mathematics. His subjective assessments, based on experience as of September 2023, rate the writing tasks highly useful and identify uses in coding and data analysis. Literature research and equation derivation remain experimental in that assessment. The examples identify potentially useful operations, rather than estimate productivity gains for a complete research project.

Experiments in related professional work provide direct comparisons of assistance with no assistance. Noy and Zhang [46] randomly offer professionals access to ChatGPT for short writing assignments. Participants offered access take less time and receive higher ratings from evaluators unaware of the assignment. Dell'Acqua et al. [14] compare consultants with no AI access, GPT-4 access, or GPT-4 access plus a prompt engineering overview. On product development tasks, both assisted groups complete more tasks, receive higher quality ratings, and reach the final task screen sooner within the fixed session. In a separate case using financial data and interviews to choose a brand for investment, assisted consultants are less likely to choose the correct brand. These results establish benefits for particular tasks, not a general gain in business judgment or the production of economic research.

6.2 Automating Multiple Research Steps

Automated research systems connect several tasks within one workflow. Lopez-Lira et al. [40] develop AutoTheory to generate, check, and calibrate candidate economic mechanisms. They apply the system to two asset-pricing puzzles. The first concerns how strongly prices respond to demand shifts at the level of an investment style relative to an individual stock. The standard benchmark predicts a much larger ratio than the data show. The second concerns dividend strips, which are claims to dividends paid over specified horizons. Near-term dividend claims have high Sharpe ratios relative to the market, which standard consumption-based models struggle to explain. These empirical patterns provide calibration targets for candidate mechanisms.

Given the puzzle, a benchmark model, parameter ranges, and targets, AutoTheory generates mechanisms and mathematical derivations. Mathematical and code audits screen proposals before parameter optimization. A deterministic scorer evaluates fit, economic plausibility, and parsimony, and an LLM referee assigns a final score. The surviving candidates enter a registry from which later runs can revise or combine mechanisms.

For the price multiplier problem, 458 of 1,000 runs complete all stages against three calibration targets, and four different mechanisms tie for the highest referee score of 84. For the dividend strip problem, which has ten moment conditions including conditional targets, 257 of 2,153 runs complete the pipeline and the highest score is 58. The same system therefore produces different completion rates and scores across the two problems. Matching the targets and passing the internal checks supplies candidates for economic assessment, rather than a unique

explanation of the data.

A different use of automation is to turn completed tests of stock return predictability into finance papers. Novy-Marx and Velikov [45] construct 31,460 potential predictors from accounting data and retain 96 after portfolio and factor tests. Their analysis programs produce a report of results and robustness checks for each retained signal. GPT-3.5-turbo names the signals, and Claude 3.5 Sonnet uses the reports to draft motivation, results discussion, and possible economic mechanisms. Scripts assemble three versions for each signal, yielding 288 papers. The explanations are generated after the return results are known, which the authors highlight as a risk of constructing hypotheses to fit observed results. They also report occasional fictitious references and leave systematic assessment of citations, hypothesis alignment, and paper quality to further work.

In empirical research, a proposed question must also be answerable with the available variables and methods. Zhu and Wang [54] address this problem with Human-in-the-Loop Economic Research (HLER). Before generating questions, the system checks the dataset's variables and statistical properties. After a researcher selects a question, it runs econometric analyses, drafts a manuscript, and uses a reviewer agent to request revisions. A researcher then decides whether the paper is ready to be finalized.

The authors separately evaluate whether inspecting the data improves the feasibility of proposed questions. Two researchers assess whether the required variables exist, the design fits the data, and the supported methods can answer the question. They resolve disagreements through discussion. Across 14 runs on three datasets, 69 of 79 questions proposed with the variable inventory and statistical profile satisfy these criteria. Only 34 of 82 do so when the model receives just the dataset name and a broad research domain. Twelve of the 14 runs also complete the full workflow without human intervention other than question selection and final manuscript approval. Thus, inspecting the data reduces infeasible proposals, and most tested runs connect the research steps successfully.

6.3 Evaluating Research Outputs and Processes

Automated research still requires researchers to choose promising questions, assess economic explanations, and decide whether results are credible enough to use. AutoTheory supplies candidate mechanisms and their calibration results. Researchers need to examine the mechanisms' assumptions and test empirical implications that distinguish competing explanations. HLER assigns researchers the choice of question and approval of the final paper, while data inspection helps exclude proposals that cannot be studied with the available variables and methods. These systems can organize and screen material for researchers, but judging its economic relevance and interpreting the results remain substantive research tasks.

Generated research material needs to be checked before it is used because an incorrect statement, calculation, or derivation can alter the analysis built on it. For literature summaries and paper drafting, verification begins with the cited papers. Novy-Marx and Velikov [45] report occasional fictitious references in generated finance papers. Finding a real paper with the cited title is only the first check. Researchers also need to read the passage or table containing the

reported finding. A summary can preserve a number while changing what it describes, for example by confusing expected wages with actual wages or a comparison within a sample with a test on new observations. Checking the original finding alongside the generated sentence helps detect these changes before the summary is used in an argument.

For mathematical and programming assistance, checks concern both the reasoning and its implementation. A proof needs to be checked step by step against its stated assumptions. A numerical formula can be tested on inputs for which the answer is known. Code used to construct a variable can be checked against the intended sample and aggregation rule, including whether observations with missing data are excluded or their missing values are filled in. When several research steps are automated, these checks need to occur before one step's results are used by the next. Checking only the final prose can leave errors in the underlying data or calculations undetected.

Whether automation saves research effort depends on the entire process, not just the time taken to generate a draft. Checking references, testing code, inspecting estimates, and correcting errors can absorb part of the apparent saving. A useful comparison would assign comparable research tasks to assisted and unassisted workflows, assess the resulting work under the same quality criteria, and record total time including verification and correction. The completion rates and internal scores reported by AutoTheory, HLER, and Novy-Marx and Velikov [45] show what their systems produce and how they screen it. They do not measure the difference in total effort needed to obtain research of comparable quality.

To reproduce the reported economic analysis, another researcher needs the model responses that actually entered it. Repeating the original prompt may produce a different response as model versions or generations change, a difficulty discussed by de Kok [12]. Saving the prompts and responses, together with the code for constructing variables and estimating results, allows the analysis to be reconstructed without first obtaining an identical new response. Intermediate files are also needed when a later step uses an earlier model answer or tool result.

Researchers who also want to regenerate the responses need to know which model and version were used and when the calls were made. The full prompts, generation settings, and supplied documents are also needed. If the weights are available, saving them together with the runtime settings preserves the model used in the study. For a closed service, researchers should record the provider and exact model or snapshot. Repeating the calls then depends on that version remaining accessible. The development record can focus on the main design choices, the data used to refine the method, and human corrections that affected the reported results.

A further issue is whether the research data and model responses may be shared. Before sending confidential firm data, respondent information, or unpublished material to an external model, researchers need authorization for that use and transfer. The storage and reuse terms of the service also matter. Where the original material cannot be released, controlled access or checks by an authorized party can support inspection of the restricted steps. The replication materials should distinguish those steps from the ones that readers can repeat directly.

6.4 Related Evidence on Labor Market Outcomes

The task comparisons above leave open how LLMs could change the work of people and organizations. Kazinnik and Brynjolfsson [30] examine this potential in the Federal Reserve by classifying 7,820 tasks across 360 occupations according to whether an LLM tool could reduce completion time by at least half without lowering quality. Each task receives a score of 0, 0.5, 0.75, or 1. A score of 0 denotes no exposure under this rule. A score of 0.5 denotes tasks for which additional software is needed, while 0.75 denotes assistance requiring image capabilities. A score of 1 denotes direct assistance from an LLM tool. An occupation's score is the average across its tasks, and economists receive 0.807. Conducting research and disseminating findings is classified as directly exposed, whereas testifying at regulatory or legislative hearings is not. The following studies examine what happens when people have access to or use LLMs, focusing on recruitment, working time, and earnings.

Recruitment provides a test of whether improvements in assisted writing change how that writing is used. Abbas Nejad et al. [1] randomly give 137 job seekers access or no access to ChatGPT before they write cover letters. Recruiters unaware of the assignment evaluate the letters and rate how likely they would be to invite each applicant to an interview. Access raises the total letter score by 0.222 standard deviations, significant at the 10% level, with clearer improvements in introductions and closings. But it does not significantly improve interview recommendations. These results show that the value of writing assistance depends not only on the text produced but also on the judgment for which it is used.

Workplace studies move from evaluations of a particular output to changes in workers' time and pay. Humlum and Vestergaard [28] link surveys covering 11 exposed occupations in Denmark to administrative employment records. Users report time savings averaging 2.8% of their total working hours. Using records through June 2024, the authors compare adopters and non-adopters before and after ChatGPT's introduction and also examine variation associated with employer encouragement. They find no statistically significant average effects on earnings or recorded hours in the studied sample. The pooled earnings estimates rule out average changes larger than about 1%, although estimates for separate occupations are less precise. Thus, the reported time savings have not translated into a comparable increase in earnings within the period and sample studied.

Chen et al. [8] ask whether earnings rose more after ChatGPT's public release in occupations with a larger share of tasks linked to LLM use. They use U.S. Current Population Survey records from January 2010 to August 2025 and classify occupations by the share of their tasks linked to Claude usage. Comparing occupations above and below the median exposure, their synthetic difference-in-differences estimates imply an average increase of about $89 in weekly earnings for the more exposed occupations, measured in January 2010 dollars. This estimate concerns average earnings in occupations with different task exposure, while the Danish study compares workers who report using or not using chatbots.

7 Conclusions and Future Research

7.1 Main Conclusions

In text analysis, LLMs classify passages and extract information by interpreting the surrounding text. Several applications improve on methods based on fixed word lists. The classifications can then be used to construct economic variables. For example, classifications of managers' statements about future investment are combined into a measure for each firm and quarter. In constructing these variables, researchers decide how to combine the classifications and assess whether the resulting measure captures the intended economic concept.

LLMs contribute to forecasting by extracting signals from text or predicting outcomes directly. The studies find additional predictive information, and some also report lower forecast errors than existing methods. But the improvements vary across samples and periods. Using a forecast for a decision raises a further question about the outcome of acting on it. In trading, for example, transaction costs can absorb the returns of a strategy based on a predictive signal.

In simulation, LLMs generate choices and answers under conditions specified by researchers. Human data can help specify agents that better reflect the behavior being studied. Calibrating agents with observed choices improves predictions in related games, and personal records improve some predictions of survey answers. Human observations therefore help shape the simulation itself, as well as provide a reference for evaluating its results.

LLM assistance improves speed and quality on some writing tasks. Automated systems extend assistance by connecting question generation, analysis, and drafting. In the systems reviewed, researchers still select questions and assess the explanations produced. The researchers' role is therefore not confined to editing the language of a generated paper. They judge whether a question is worth pursuing and whether the analysis provides a credible answer.

Across the four uses, delegating a task does not remove the need to decide what that task should achieve. Researchers define what is to be measured, predicted, or simulated before specifying the model's instructions. After the model completes the task, researchers assess what the results reveal about the economic question. Human involvement thus begins with the research design and continues through the interpretation of the findings.

7.2 Directions for Future Research

The variation in forecast gains across periods raises the question of whether improvements over existing methods persist as new data arrive. A signal that performs well in one historical period may be less useful when economic conditions change. Following its performance on later observations can show whether the improvement continues beyond the original evaluation. Even when forecast accuracy improves, using the signal in a decision introduces costs and risks that the accuracy comparison does not measure. Research can therefore examine both whether the predictive gains persist and whether decisions based on the signal improve the relevant economic outcome. Changes in performance also raise an open question about when researchers should revise the model or forecasting procedure.

Human calibration has improved predictions of choices in related games, but this leaves

open how far the same agents can be reused. When incentives or the population being studied change, the settings that matched the original human choices may need to change as well. Researchers then face a practical question about whether to retain the existing calibration or collect new human observations to revise it. Understanding when such revision is needed would clarify whether simulations can help study related behavioral questions with less repeated data collection.

Beyond carrying out a research plan, LLMs may also help researchers decide what to investigate. Existing systems can generate several mechanisms that fit the same calibration targets. Since those targets alone do not distinguish the explanations, a useful extension would be to ask the system where their predictions differ. Researchers could use these differences to identify evidence that favors one explanation over another. They would also need to assess whether the mechanisms' assumptions and the proposed tests are credible. The open question is whether this assistance helps develop more informative economic tests, rather than simply producing more candidate mechanisms.

Acknowledgements

We thank Professor Xinyu Zhang for his great suggestions to this manuscript and submission.

Conflict of Interest

Zongwu Cai is an editorial board member for Journal of Systems Science & Complexity and he was not involved in the editorial review or the decision to publish this article. All authors declare that there are no competing interests.

References

[1] Abbas Nejad, K., G. Musillo, T. Wicker, and N. Zaccaria (2025). Labor market signals: The role of large language models. Technical Report 5185035, SSRN. Undated revised manuscript; first circulated in 2025.

[2] Allard, M.-A., P. Teiletche, and A. Zinebi (2024). Enhancing inflation nowcasting with LLM: Sentiment analysis on news. arXiv:2410.20198.

[3] Baker, S. R., N. Bloom, and S. J. Davis (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics 131*(4), 1593–1636.

[4] Brown, T. B. et al. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, Volume 33, pp. 1877–1901.

[5] Brynjolfsson, E., J. R. Enríquez, S. Kazinnik, and D. Nguyen (2026). Augmenting survey data with generative AI: An application to economic research. Working Paper 6343598, SSRN.

[6] Bubeck, S., V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. Technical Report 2303.12712v5, arXiv.

[7] Charness, G. and M. Rabin (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics 117*(3), 817–869.

[8] Chen, D., C. Kane, A. Kozlowski, N. Kuniesvky, and J. A. Evans (2025). The (short-term) effects of large language models on unemployment and earnings. Technical Report 2509.15510v1, arXiv. Manuscript dated September 22, 2025.

[9] Chen, Y., H. Fang, Y. Zhao, and Z. Zhao (2025). Measuring labor market match quality with large language models. Working paper, revision dated March 11, 2025.

[10] Coyle, D. and D. Nguyen (2023). Free digital products and aggregate economic measurement. *Economie et Statistique/Economics and Statistics* (539), 27–50.

[11] de Bondt, G. J. and Y. Sun (2026). Enhancing gdp nowcasts with chatgpt: A novel application of pmi news releases. *Journal of Policy Modeling 48*(3), 107052.

[12] de Kok, T. (2025). Chatgpt for textual analysis? how to use generative llms in accounting research. *Management Science 71*(9), 7888–7906.

[13] DeepSeek-AI (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature 645*, 633–638.

[14] Dell'Acqua, F. et al. (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science 37*(2), 403–423.

[15] Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186.

[16] Du, T., A. Kanodia, H. Brunborg, K. Vafa, and S. Athey (2024). LABOR-LLM: Language-based occupational representations with large language models. Technical Report 2406.17972v1, arXiv.

[17] Erdem, O. and R. P. Ashok (2025). Artificial finance: How AI thinks about money. arXiv:2507.10933.

[18] Fan, Y., J. Lv, A. Sun, and Y. Wang (2025). LLM-powered CPI prediction inference with online text time series. arXiv:2506.09516.

[19] Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as data. *Journal of Economic Literature 57*(3), 535–574.

[20] Gentzkow, M. and J. M. Shapiro (2010). What drives media slant? Evidence from U.S. daily newspapers. *Econometrica 78*(1), 35–71.

[21] Glasserman, P. and C. Lin (2023). Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. arXiv:2309.17322.

[22] Gow, I. D., D. F. Larcker, and A. A. Zakolyukina (2021). Non-answers during conference calls. *Journal of Accounting Research 59*(4), 1349–1384.

[23] Hansen, A. L. and S. Kazinnik (2024). Can ChatGPT decipher fedspeak? Working Paper 4399406, SSRN.

[24] Hansen, A. L. and S. J. Lee (2025). Financial stability implications of generative AI: Taming the animal spirits. Finance and Economics Discussion Series 2025-090, Board of Governors of the Federal Reserve System.

[25] Hansen, S., M. McMahon, and A. Prat (2018). Transparency and deliberation within the FOMC: A computational linguistics approach. *Quarterly Journal of Economics 133*(2), 801–870.

[26] Hartley, E. (2025). Narratives to numbers: Large language models and economic policy uncertainty. arXiv:2511.17866.

[27] Horton, J. J., A. Filippas, and B. S. Manning (2026). Large language models as simulated economic agents: What can we learn from Homo Silicus? arXiv:2301.07543v2.

[28] Humlum, A. and E. Vestergaard (2025, May). Large language models, small labor market effects. Technical Report 33777, National Bureau of Economic Research. May 2025 version.

[29] Jha, M., J. Qian, M. Weber, and B. Yang (2024). ChatGPT and corporate policies. NBER Working Paper 32161, National Bureau of Economic Research.

[30] Kazinnik, S. and E. Brynjolfsson (2025). AI and the Fed. NBER Working Paper 33998, National Bureau of Economic Research.

[31] Kirtac, K. and G. Germano (2024). Sentiment trading with large language models. *Finance Research Letters 62*(B), 105227.

[32] Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature 61*(4), 1281–1317.

[33] Kwon, B., T. Park, F. Perez-Cruz, and P. Rungcharoenkitkul (2024, December). Large language models: A primer for economists. *BIS Quarterly Review*, 37–52.

[34] Lehner, S. and A. Lopez-Lira (2026). ChatGPT as a time capsule: The limits of price discovery.

arXiv:2604.21433.

[35] Li, Y., S. Wang, H. Ding, and H. Chen (2023). Large language models in finance: A survey. In *Proceedings of ICAIF*, pp. 374–382.

[36] Liu, Z. et al. (2026). Same claim, different judgment: Benchmarking scenario-induced bias in multilingual financial misinformation detection. In *Findings of the Association for Computational Linguistics: ACL*, pp. 9838–9864.

[37] Liu, Z., Z. Cai, and Y. Fang (2026). An LLM approach to study expectation management frictions under china's dual-track regulation and multi-objective constraints. Working Paper, https://econpapers.repec.org/paper/kanwpaper/202613.htm.

[38] Liu, Z., D. Huang, K. Huang, Z. Li, and J. Zhao (2020). FinBERT: A pre-trained financial language representation model for financial text mining. In *Proceedings of IJCAI*, pp. 4513–4519.

[39] Lopez-Lira, A. (2025). Can large language models trade? testing financial theories with LLM agents in market simulations. arXiv:2504.10789.

[40] Lopez-Lira, A., S. Seyfi, and Y. Tang (2026). Can LLMs discover novel economic theories? Working Paper 6464059, SSRN.

[41] Lopez-Lira, A. and Y. Tang (2026). Can ChatGPT forecast stock price movements? return predictability and large language models. *Journal of Financial Economics 184*, 104335.

[42] Loughran, T. and B. McDonald (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance 66*(1), 35–65.

[43] Ludwig, J., S. Mullainathan, and A. Rambachan (2026). Large language models: An applied econometric framework. *Annual Review of Economics 18*, 283–316.

[44] Nie, Y., Y. Kong, X. Dong, J. M. Mulvey, H. V. Poor, Q. Wen, and S. Zohren (2024). A survey of large language models for financial applications: Progress, prospects and challenges. arXiv:2406.11903.

[45] Novy-Marx, R. and M. Z. Velikov (2025). AI-powered (finance) scholarship. NBER Working Paper 33363, National Bureau of Economic Research.

[46] Noy, S. and W. Zhang (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science 381*(6654), 187–192.

[47] OpenAI (2022, November). Introducing ChatGPT. OpenAI, November 30. https://openai.com/index/chatgpt/. Accessed September 24, 2026.

[48] Ouyang, L. et al. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, Volume 35, pp. 27730–27744.

[49] Silva, T. C., K. Moriya, and R. M. Veyrune (2025). From text to quantified insights: A large-scale LLM analysis of central bank communication. IMF Working Paper WP/25/109, International Monetary Fund.

[50] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance 62*(3), 1139–1168.

[51] Weber, M. et al. (2025). Tell me something I don't already know: Learning in low- and high-inflation settings. *Econometrica 93*(1), 229–264.

[52] Wu, J. C., J. Xi, and S. Xie (2025). LLM survey framework: Coverage, reasoning, dynamics, identification. NBER Working Paper 34308, National Bureau of Economic Research.

[53] Xie, Q. et al. (2023). PIXIU: A large language model, instruction data and evaluation benchmark for finance. In *Advances in Neural Information Processing Systems*, Volume 36, pp. 33469–33484.

[54] Zhu, C. and X. Wang (2026). HLER: Human-in-the-loop economic research via multi-agent pipelines for empirical discovery. arXiv:2603.07444.