CARMA 2026

8th Int. Conf. on Advanced Research Methods and Analytics
1-3 July 2026, Valencia, Spain
DOI: https://doi.org/10.4995/CARMA2026.2026.21952

News-enhanced Large Language Models in macroeconomic forecasting in the Netherlands

Leonardo Becarelli[1], Michiel Nijhuis[2], Jasper de Winter[2], Luca Trapin[1]

[1]Department of Statistical Sciences, Alma Mater Studiorum Università di Bologna, [2]De Nederlandsche Bank, Netherlands.

How to cite: Becarelli, L.; Nijhuis, M.; de Winter, J; Trapin, L. (2026). News-enhanced Large Language Models in macroeconomic forecasting in the Netherlands. In: 8th International Conference on Advanced Research Methods and Analytics (CARMA 2026). Valencia, 26-28 June 2026. https://doi.org/10.4995/CARMA2026.2026.21952

Abstract

This paper evaluates the ability of large language models (LLMs) to generate macroeconomic forecasts and examine whether financial news can improve their predictive performance. The analysis focuses on the Netherlands and compares forecasts generated by OpenAI models with forecasts from the Consensus Forecasts survey. Using Dutch financial news from Het Financieele Dagblad, we construct and compare several forecasting configurations covering years from 2016 to 2025 and consider four key macroeconomic indicators: GDP, private consumption, manufacturing production, and consumer price inflation, with horizons of up to seven quarters ahead. This approach shows promising results for approximating experts' forecasting capabilities.

Keywords: *Large language models; Textual Analysis; Macroeconomic Forecasting.*

1. Introduction

Recent advances in large language models (LLMs) have generated growing interest in their potential application in economic analysis and forecasting (Korinek, 2023). The ability of LLMs to incorporate financial news and other textual sources suggests they could complement traditional forecasting approaches (Ónozó, 2024). This paper investigates whether LLMs can generate macroeconomic forecasts comparable to those produced by professional forecasters. In particular, the analysis evaluates whether the inclusion of Dutch financial news from *Het Financieele Dagblad (FD),* one of the Dutch leading financial newspapers, improves the LLM forecasting performance of Dutch macro indicators. The empirical framework compares forecasts generated by LLMs with those reported in the *Consensus Forecasts*[1] survey, which aggregates the expectations of professional analysts and is widely used as a benchmark in macroeconomic forecasting (Berben et al., 2025). This setting is particularly of interest because, while the composition of OpenAI models training data is unknown, is possible to say that Dutch

[1] "What Are Consensus Forecasts?", https://www.consensuseconomics.com/what-are-consensus-forecasts/, Consensus Economics Inc. 2010-2026

This work is licensed under a Creative Commons License CC BY-NC-SA 4.0
Editorial Universitat Politècnica de València

Article in Press

probably represents a relatively small share of the data used to train LLMs (Ferrara, E., 2023). Examining whether LLMs can extract useful signals from Dutch financial news therefore provides insights into the ability of these models to process economically relevant information in languages that are less prominent in their training data. The forecasting exercise covers the period March 2016 to March 2025 and considers four indicators, GDP growth, private consumption growth, manufacturing production, and consumer price inflation. Forecasts are generated for horizons up to seven quarters ahead using two different OpenAI models (GPT4o mini and GPT5 mini) and are evaluated against professional Consensus Forecasts under different information specifications. Preliminary results suggest that, while professional Consensus forecasts remain a strong benchmark, LLM-based forecasts can in some cases approach their predictive performance. The inclusion of financial news produces heterogenous effects across variables, generally reducing the gap between the two benchmarks. Given the relatively low costs of generating forecasts using LLMs, they may represent a useful complementary tool in macroeconomic forecasting.

2. Related Literature

Whether LLMs can replicate or simulate the expectations of professional forecasters has already been studied. Hansen et al. (2025) simulate responses to the U.S. Survey of Professional Forecasters by constructing synthetic forecaster personas using LLMs and providing them with real-time macroeconomic data. Their results show that forecasts generated by LLMs resemble the distribution of human forecasters and can even outperform human predictions at some horizons. The authors argue that LLMs may extract latent information contained in historical forecasts while avoiding some of the biases and noise present in human judgement. Bybee (2023) uses GPT 3.5 to construct a "survey of economic expectation" by feeding the model financial news articles and asking it to evaluate the macroeconomic implications of the information contained in the articles. The expectations generated by the model resemble those obtained from traditional survey-based measures such as the Survey of Professional Forecasters and display patterns like those observed in human expectations formation. Several studies also examine the predictive performance of LLMs in forecasting macroeconomic variables. Faria-e-Castro and Leibovici (2024) show that LLMs can generate inflation forecasts that are competitive with, and in some cases more accurate than, those of professional analysts. Their approach relies on prompting the model to produce forecasts conditionally on a specific information set defined by a given point in time. By restricting the model information available at the forecast origin, the methodology attempts to approximate a real time forecasting environment. A related literature branch examines whether LLMs can complement rather than replicate approaches. Andre' et al. (2025) compare LLM-based predictions with econometric nowcasting models and conclude that while LLMs are not yet ready to substitute standard forecasting frameworks, they may provide useful complementary insights, particularly in periods of structural uncertainty. Finally, a broader macroeconomic forecasting literature

Article in Press

studies the use of textual news as a predictor of economic activity. Financial news often contains timely information about economic developments that may not yet be reflected in official macroeconomic statistics. Several studies show that signals extracted from textual data can improve forecasts of macroeconomic variables, including inflation and output growth (Ardia and Bluteau, 2025; Eugsten and Uhl, 2024).

3. Research Design and Data

3.1. Data Sources

Professional forecasts are obtained from the *Consensus Forecasts* survey published by Consensus Economics covering a period ranging from March 7th, 2016, to March 10th, 2025. The survey collects macroeconomic expectations for a selection of countries (including the Netherlands) from a panel of private-sector forecasters. Respondents provide forecasts for a range of macroeconomic variables at multiple horizons, including quarterly projections for real GDP growth, inflation industrial production, and other key indicators. The survey is conducted monthly, with the submission deadline typically falling on the second Monday of the third month of each quarter. In the study, the Consensus Forecasts serve as the benchmark against which the performance of the LLM-based forecasts is evaluated. These forecasts represent the expectations of private-sector analysts and are widely used in the literature as a reference for evaluating forecasting performance. (Berben et al., 2025). Financial news articles data are obtained from *FD*, one of the leading Dutch financial and economic newspapers. The dataset used in this study consists of a comprehensive archive of more than one million articles in Dutch spanning multiple decades, with initial coverage beginning in 1985. The archive includes the full text of published articles as well as associated metadata such as publication dates and headlines. FD, given its specialized financial nature focused on the Dutch economy, allows the construction of time-consistent information sets on financial news available at various points in time. Further details on how the articles have been processed will be explained in *subchapter 3.2.3*.

3.2. Forecasting Framework

This study focuses on four major macroeconomic indicators for the Netherlands reported in the Consensus Forecasts: Gross Domestic Product (% change qtr/qtr, '*GDP*'), Private Consumption (% change qtr/qtr, '*CON*'), Manufacturing Production (change over previous period, '*IP*'), and Consumer Prices (change over previous period, '*CPI*'). Using different combinations of prompts and numeric and textual economic data, our aim is to see if LLMs can be used to approximate the forecasting capabilities of professional forecasters for the Netherlands.

Article in Press

3.2.1. Forecasting Approaches

The forecasting exercise compares predictions generated by two OpenAI models (GPT-4o mini and GPT-5 mini) with those produced by professional forecasters. Following a similar method introduced in Lopez-Lira et al. (2025), forecasts generated by the LLMs are implemented under two main information specifications. In the first specification, the model generates forecasts based solely on the prompt describing the forecasting task, without additional contextual information. In the second specification, the model is provided with the five most recent realized observations of the variable being forecast. This allows the model to condition its predictions on recent variable dynamics. In addition, the forecasting exercise evaluates whether the inclusion of financial news improves the predictive performance of the model. For each two specifications described above, an augmented version is implemented in which financial news articles from FD are provided as additional contextual information. This results in four LLM-based forecasting configurations for each metric: a baseline specification without additional information (*baseline no info)*, a specification including recent realizations of the variable (*baseline info*), and corresponding versions of both specifications augmented with financial news (respectively *news no info* and *news info*). The analysis covers the period from March 7th, 2016, to March 10th, 2025, with forecasts generated on a quarterly basis. Each forecast corresponds to the timing of the Consensus Forecasts survey allowing forecasts produced to be compared with those reported by professional forecasters.

3.2.2. Consistency and Validation Checks

First, the model stability is assessed by examining the sensitivity of a set of forecasts to 30 repeated executions of the same prompt, each repeated for different dates. Because LLM outputs may exhibit stochastic variation (Hansen et al., 2024), repeated runs allow us to evaluate whether the generated forecasts remain stable across identical forecasting conditions. Results shown in *Table 1* explain that, while stochastic variability in the forecast is somewhat present (around 12% at most), this can be considered acceptable for the kind of exercise we are performing.

Table 1. Average Coefficient of Variation of forecasted metrics by model across specifications

Specification	Avg. Coeff. Of Variation
GPT4o mini baseline	0.122834483
GPT4o mini news	0.125302532
GPT5 mini baseline	0.049322071
GPT5 mini news	0.156415861

Article in Press

An additional test is conducted to evaluate whether the models appear to possess prior knowledge of the financial news articles from FD used in the forecasting experiment. For this purpose, a sample of article is selected from the FD archive. For each article, the first 40 words are provided to the model in the prompt as it is asked to complete the remainder of the article. The generated text is then compared with the original article using standard text similarity measures (term frequency - inverse document frequency cosine similarity[2] and ROUGE-L scores[3]). If the model had memorized the content of the articles during training, one would expect a high degree of similarity between the generated and original text. Results included in *Table 2* allow us to say that both models have not memorized the articles during training.

Table 2. Average TF-IDF cosine similarity and ROUGE-L scores of prior knowledge of the financial news articles by model

Model	Cos Similarity	ROUGE-L scores
GPT4o mini	0.486	0.1243
GPT5 mini	0.459	0.0998

3.2.3. Prompt Design and Forecast Generation

Forecast generated by the LLMs are obtained through a structured prompting procedure designed to approximate a real-time forecasting environment. For each forecast, the model is instructed to assume that it is located at a specific point in time, denoted as τ, corresponding to the publication date of Consensus Forecasts. The prompt explicitly instructs the model to base its prediction only on information that would be available at that date. In the system prompt, the model is specified to generate a JSON file containing: *publication* (the publication date of the data); *quarter* (the quarter of the forecasted value); *answer* (the precise numerical answer to the question; *confidence* (a number between 0 and 100 indicating the confidence in the answer). For each forecast origin, the model generates predictions for up to seven quarters ahead.

Forecasts are generated using two versions of OpenAI models: GPT-4o mini (version 2025-0416) and GPT-5 mini (version 2025-08-07). For each forecast origin and specification, the models are prompted using the same structured format to ensure consistency across the forecasting exercise. In some specifications, the model receives additional contextual information before generating the forecasts. In the specifications *baseline info* and *news info*, the prompts include the five most recent realized observations of the variable being forecasted, allowing the model to condition its predictions on recent real-world dynamics. In the specifications *news info* and *news no info*, financial articles from FD are appended to the prompt

[2] *Term Frequency–Inverse Document Frequency (TF-IDF) cosine similarity* is a metric that measures how similar two documents are by comparing the importance-weighted terms they contain, using the cosine of the angle between their TF-IDF vector representations.

[3] *ROUGE-L* is an evaluation metric for text generation that measures similarity between two texts by identifying the longest common subsequence of words they share, capturing both word overlap and sentence structure.

Article in Press

to provide additional information about current economic conditions. To select specific articles to be included in the prompt, the articles are first filtered by removing articles not containing at least one word from a tailor-made dictionary containing terms related to variable forecasted as not to include articles completely unrelated to the metric. Second, a similarity metric is constructed using *term frequency - inverse document frequency* for every article included in the dataset. For each forecasting date, articles with a similarity score larger than the second quartile are included from most recent to least recent until the cap of 25000 tokens is reached to avoid degradation of the forecasting quality.

4. Preliminary Results

Table 3 and *table 4* report the average mean squared error (MSE) and the average difference of squared errors of the forecasts generated by the LLMs and by professional forecasters from the Consensus Forecasts survey compared to the actual realization of the indicators. In both tables, the first set of rows reports the *baseline* LLM specification without FD data, while the second presents the results as the prompts are augmented with the financial news articles. Across most variables, the professional Consensus forecasts remain a strong benchmark, outperforming the baseline LLM forecasts in most metrics. Comparing the baseline and augmented specifications provides insight into the role of financial news. The use of news articles produces generally positive effects across both models, with a consistent increase in performance when FD data is included in the forecasts (apart from *Consumer Prices CPI*, presenting heterogeneous results). Results are particularly interesting for the GPT5 mini model, which in some cases produces results almost on par with those of professional forecasters. Overall, the results suggest that augmenting LLM forecasts with financial news can improve performance in precise cases and may represent a promising direction for combining textual information with LLM-based forecasting approaches, but further studies are needed to prove the robustness of these claims.

Table 3. Average MSE and Squared Error for LLM and Consensus for GPT4o mini

Config.	MSE AI	MSE Con.	Avg Diff. Error²	N. Obs.
Without Newspaper Data				
Info IP	27.523	24.009	−6.007	261
No Info IP	30.093	24.009	−8.919	278
Info CON	9.256	6.777	−2.346	278
No Info CON	4.984	6.777	**3.077**	295
Info GDP	5.503	3.773	−1.697	261
No Info GDP	2.871	3.773	**0.871**	261
Info CPI	7.769	6.971	−1.008	278
No Info CPI	8.575	6.971	−0.988	295

Article in Press

Augmented Newspaper Data				
Info IP	25.877	24.009	−3.628	261
No Info IP	24.292	24.009	−1.958	261
Info CON	8.854	6.777	−2.056	261
No Info CON	7.195	6.777	−0.446	261
Info GDP	4.380	3.773	−0.540	261
No Info GDP	2.873	3.773	**0.987**	261
Info CPI	7.170	6.825	−1.455	273
No Info CPI	6.823	6.971	−0.382	261

Table 4. Average MSE and Squared Error for LLM and Consensus for GPT5 mini

Config.	MSE AI	MSE Con.	Avg Diff. Error²	N. Obs.
Without Newspaper Data				
Info IP	31.201	24.009	−9.177	261
No Info IP	33.400	24.009	−12.072	261
Info CON	9.282	6.777	−2.382	261
No Info CON	5.588	6.777	**1.155**	261
Info GDP	5.197	3.773	−1.382	261
No Info GDP	3.556	3.773	**0.212**	261
Info CPI	8.287	6.971	−1.898	261
No Info CPI	9.540	6.971	−3.084	261
Augmented Newspaper Data				
Info IP	21.030	24.009	**0.947**	261
No Info IP	18.495	24.009	**3.524**	261
Info CON	6.451	6.777	**0.430**	261
No Info CON	5.498	6.777	**1.421**	261
Info GDP	3.174	3.773	**0.700**	261
No Info GDP	3.201	3.773	**0.679**	261
Info CPI	14.638	6.971	−8.340	261
No Info CPI	14.337	6.971	−8.021	261

Article in Press

It is noteworthy to reiterate that the analysis is focused on Dutch-based metrics and uses financial news written in Dutch, a language likely underrepresented in the training data used to the develop LLMs when compared to English-based sources. The ability of the model to extract

meaningful signals from Dutch financial news therefore suggests that LLMs may still be able to utilize economically relevant information even in languages that are less prominent in the training data. Finally, it is important to note that the cost structure of the two forecasting approaches compared differs substantially. Producing Consensus Forecasts requires the coordinated effort of multiple professional analysts and institutions, whereas generating forecasts with LLMs can be conducted at a relatively low computational cost. Even though the predictive performance of LLMs remains somewhat below the professional benchmark, the substantially lower cost of producing forecasts may make LLM-based approaches useful as complementary tools in macroeconomic forecasting.

5. Limitations and Future research

First, the forecasting framework relies on a specific prompt design and a limited set of model configurations. LLMs responses can be sensitive to prompt design formulation and parameter choices (Wang et al., 2025). While the present study attempts to maintain a consistent prompting structure across experiments, further research could explore alternative prompt designs and different methods of integrating textual information. Finally, financial news articles contain a wide range of qualitative signals about economic conditions, but the ability of LLMs to systematically extract predictive signals from these text remains an open question. While incorporating a selected portion of news articles directly in the prompt is a cost-effective solution, further work could be devoted to investigating how alternative approaches to integrating textual data could influence the forecasting capabilities of models. Further analysis will investigate the dynamics of forecasting performance across horizons, metrics and before-after each model knowledge cutoff, to better understand under which conditions LLM-based forecasting performs closer to professional consensus forecasting.

References

Ardia, D. and K. Bluteau (2025). Optimal text-based time-series indices. *International Journal of Forecasting forthcoming*.

Berben, R.-P., R. Rasiawan, and J. de Winter (2025). Forecasting dutch inflation using machine learning methods. *Working Papers 828, De Nederlandsche Bank*.

Bybee, J. L. (2023). The ghost in the machine: Generating beliefs with large language models. *arXiv preprint arXiv:2305.02823*, 349-389.

Faria-e-Castro, M. and F. Leibovici (2024). Artificial intelligence and inflation forecasts. Review 106(12) 1–14, Federal Reserve Bank of St. Louis.

Ferrara, E. (2023). Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday*, 28(11). https://doi.org/10.5210/fm.v28i11.13346

Hansen, A. L., Horton, J. J., Kazinnik, S., Puzzello, D., & Zarifhonarvar, A. (2024). Simulating the survey of professional forecasters. *Available at SSRN 5066286*.

Article in Press

Korinek, Anton. 2023. "Generative AI for Economic Research: Use Cases and Implications for Economists." *Journal of Economic Literature* 61 (4): 1281–1317**.**

Lopez-Lira, A., Y. Tang, and M. Zhu (2025). The memorization problem: Can we trust LLM's economic forecasts? *Technical Report 2504.14765*, arXiv

Ónozó, L. R., Arthur F. V., and Gyires-Tóth, B. (2024). Leveraging LLMs for Financial News

Analysis and Macroeconomic Indicator Nowcasting," in *IEEE Access*, vol. 12, pp. 160529160547.

Wang, L., Chen, X., Deng, X., Wen, H., You, M., Liu, W., ... & Li, J. (2024). Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *NPJ digital medicine*, *7*(1), 41.

Article in Press