EUROPEAN CENTRAL BANK
EUROSYSTEM

Working Paper Series

Gabe J. de Bondt, Yiqiao Sun

Enhancing GDP nowcasts with ChatGPT: a novel application of PMI news releases

No 3063

Disclaimer: This paper should not be reported as representing the views of the European Central Bank (ECB). The views expressed are those of the authors and do not necessarily reflect those of the ECB.

Abstract

This study involves tasking ChatGPT with classifying an "activity sentiment score" based on PMI news releases. It explores the predictive power of this score for euro area GDP nowcasting. We find that the PMI text scores enhance GDP nowcasts beyond what is embedded in ECB/Eurosystem Staff projections and Eurostat's first GDP estimate. The ChatGPT-derived activity score retains its significance in regressions that also include the composite output PMI diffusion index. GDP nowcasts are significantly enhanced with PMI text scores even when accounting for methodological variations, excluding extraordinary economic events like the pandemic, and for different GDP growth quantiles. However, the forecast gains from the enhancement of GDP nowcasts with ChatGPT scores are time dependent, varying by calendar years. Sizeable forecast gains of on average about 20% were obtained apart from the two most recent years due to exceptionally low forecast errors of the two benchmarks, especially the first GDP estimate.

Keywords Chat Generative Pre-training Transformer, text analysis, zero-shot sentiment analysis, Purchasing Managers' Index (PMI), nowcasting GDP

JEL Classification C8; E32; C22

Non-technical summary

This study introduces a novel approach to economic forecasting by utilizing artificial intelligence (AI), specifically ChatGPT, to enhance predictions of euro area Gross Domestic Product (GDP). Traditional methods of economic forecasting primarily rely on numerical data, such as hard data on industrial production and retail sales as well as soft data such as Purchasing Managers' Index (PMI) diffusion indices. Instead, this research explores the potential of integrating qualitative information - textual content from PMI news releases - into existing forecasts. What makes this study unique is its focus on the narrative, tone, and anecdotes reported in PMI news releases. ChatGPT was employed to analyse PMI news releases and generate activity sentiment scores. These scores quantify the sentiment about activity expressed in the narratives and anecdotes of the news release, ranging from strongly expanding activity to significantly contracting activity. The study then integrates these ChatGPT-derived activity scores in traditional GDP nowcasts, i.e., forecasts of real GDP growth in the current quarter, to assess their predictive power. The analysis utilizes two notoriously hard-to-beat benchmarks for GDP nowcasting, setting a high standard for accuracy, namely ECB/Eurosystem Staff projections and the first GDP estimate. The first benchmark includes judgment from experts and the second relies on statistical methods that fill a sizeable gap of missing statistical information for the first estimate of GDP.

The main compelling result is that the enhancement of the PMI text scores to the two GDP nowcast benchmarks significantly improves the accuracy of GDP nowcasts. Similar in-sample gains are not obtained by adding the composite output PMI diffusion index. Ordinary least squares, robust least squares, and ridge regressions all show that the diffusion index has no value added to the benchmark GDP nowcast or even contributes counterintuitively negatively. The GDP nowcasting results enhanced with the ChatGPT-derived score holds even when accounting for methodological variations, excluding extraordinary economic events like the pandemic, and for different GDP growth quantiles. The out-of-sample forecast gains of enhancing GDP nowcasts with PMI text scores are on average about 20% apart from the two most recent years, but they are strongly time-dependent, varying by calendar years. The study shows that the qualitative insights from the PMI narratives and anecdotes provide valuable information that complements the numerical data, offering a more comprehensive assessment of real GDP growth.

Our results imply the following. They confirm earlier findings that economic forecasting can be enhanced by integrating qualitative data sources into traditional models. A new element is that this research shows that only two pages of text rather than for example millions of newspapers articles can be sufficient to enhance existing hard-to-beat benchmarks. Moreover, the robustness of these findings across different methodological adjustments underscores the potential of AI in economic forecasting. This study advances the field of AI-driven economic forecasting and provides a new practical tool for policymakers, financial analysts, and economists to predict more accurately GDP. The success of ChatGPT opens new avenues for further research, such as applying similar techniques to other types of economic texts, including forecast reports of policy institutions. Additionally, this method could be explored for other regions or countries.

1 Introduction

This study explores the integration of ChatGPT-based sentiment activity scores into existing Gross Domestic Product (GDP) nowcasts. We hypothesize that these scores can improve the predictive accuracy of nowcasts by incorporating real-time sentiment data, thus providing a more comprehensive and timely assessment of GDP. GDP nowcasting plays a crucial role in economic policymaking, financial markets, and business strategy. Traditional nowcasting models typically rely on structured quantitative data, such as industrial production, retail sales as examples of hard data, and surveys as soft data. However, the advent of natural language processing technologies, specifically large language models like ChatGPT, offers new opportunities to enhance the accuracy of GDP nowcasts. One compelling application of ChatGPT is its ability to classify economic activity based on textual data, such as news articles, social media posts, and financial reports. By analysing these unstructured data sources, ChatGPT can generate sentiment-based activity scores that capture real-time economic sentiments. These sentiment scores provide potential incremental information that can complement existing GDP nowcasts. The rationale is that textual data often reflect immediate public and business sentiment, which may precede observable changes in traditional economic indicators.

Our nowcasting analysis considers two competitive benchmark nowcasts of the euro area real GDP growth. The first GDP nowcast benchmark refers to the real GDP growth nowcasts embedded in the ECB/Eurosystem staff macroeconomic projections, which contain all relevant real-time information at the point of making the forecast including staff judgement. The second benchmark considered is the first GDP release by Eurostat. It reflects all statistical information as available at the time of the release, including statistical methods to deal with still missing information.

The textual data considered are deliberately limited to news releases of the euro area Purchasing Managers' Index (PMI). By focusing exclusively on PMI surveys, this study leverages their timeliness, comprehensiveness, and international recognition. De Bondt (2012 and 2019) reports conclusive evidence for a strong predictive power of the composite output PMI index for euro area GDP growth in the current quarter. PMI surveys are conducted monthly and offer a near real-time snapshot of business conditions, including activity. This data is invaluable for nowcasting, because it captures the latest trends and shifts in economic activity that might not yet be evident in other economic indicators. Moreover, PMI surveys are globally standardized and widely adopted across various countries, allowing for consistent comparisons and assessments of economic conditions on an international scale. This universality ensures that PMI data is not only relevant but also comparable, making it an ideal candidate for enhancing GDP nowcasts.

Our empirical analysis intersects with two strands of literature. The first strand is the sentiment textual analysis literature, which employs natural language processing methods to analyse digital texts in economics and finance. Sharpe et al. (2023) demonstrate the predictive power of extracting sentiment from the narratives of the Green Book for US GDP, concluding that the tonality of these narratives conveys substantial incremental information. Similarly, Du et al. (2024) explore the anecdotes of the Beige Book by a textual analysis. Barbaglia et al. (2022) construct sentiment measures to forecast quarterly US GDP and Babii et al. (2022) report on the application of textual analysis to US GDP nowcasting. Ashwin et al. (2024) report a textual analysis to euro area GDP nowcasting. In contrast to this euro area study, which utilize millions of texts, our approach is purposefully selective, focusing exclusively on PMI news releases, akin to Du et al. (2024) that focusing on only one type of US publication. The second strand of literature pertains to GDP nowcasting. A comprehensive review by Stundziene et al. (2023) emphasizes the importance of utilizing real-time data and alternative indicators to enhance predictive accuracy in economic activity nowcasting. An innovative example of such indicators is the use of satellite-based data on night-time lights, as explored by Galimberti (2020) or Google Search data (Ferrara and Simoni, 2022). Furthermore, Basselier et al. (2018) investigate the relationship between GDP nowcasting and qualitative surveys, specifically highlighting the predictive value of manufacturing PMI surveys for euro area GDP but downplaying the importance of hard data, such as industrial production, for GDP nowcasting. Manufacturing PMI is found to be highly informative, ranked at the top among a wide range of indicators. A shortcoming of this study is that the composite output PMI and services business activity PMI are not considered, which may provide even more accurate forecasts.

The main lesson of this empirical study is that enhancing existing competitive GDP nowcasts with Chat-GPT derived text scores significantly improves their accuracy. The PMI text scores retain its significance in regressions that also include the composite output PMI diffusion index. This is not only shown for ordinary least squares (OLS) but also for robust least squares and ridge regression. The information in the PMI text scores adds value beyond the information already embedded in the diffusion index. The significant enhancement of PMI text scores to GDP nowcasts remains robust across various methodological approaches, excluding extraordinary economic events such as the pandemic, and for different GDP growth quantiles. The out-of-sample forecast gains relative to the two benchmarks are time dependent, varying by calendar years. The forecast gains relative to the two hard-to-beat benchmarks average around 20% apart from the two most recent years. The relative forecast performance during the two most recent years was poor due to exceptionally small errors of the benchmarks, especially close to zero errors of the first GDP release. The ChatGPT enhanced forecast errors were, however, in absolute term also small in 2023 and 2024. These insights confirm that textual data significantly improve the accuracy of GDP nowcasts (among others, Díaz Sobrino et al., 2021; Barbaglia et al., 2022; Ferrara and Simoni, 2022; Ashwin et al., 2024). Interestingly, the volume of textual data is not necessarily a critical factor. We demonstrate that as little as two pages of commentary from the globally renowned economic indicator, the PMI, can effectively contribute to accurate GDP nowcasts.

The paper proceeds as follows: Section 2 reports ChatGPT classifications of activity in the PMI news releases; Section 3 describes the predictive content of PMI text scores for euro area GDP nowcasts; and Section 4 concludes.

2 PMI text scores

The main input data are the news releases that accompany the monthly publication of the Eurozone PMI data by S&P Global. The latest news releases, including those for the euro area, can be found on the release site of S&P Global, see https://www.markiteconomics.com/Public/Release/PressReleases. The text structure of the euro area flash PMI releases is not fully fixed. Only minor changes have been made over time. The structure of the most recent news releases is illustrative for the structure since 2008 and as follows: two pages of text dealing with key findings, output and demand, employment, prices, inventories and supply chains, outlook, and a comment. The comment section reflects S&P Global's assessment of the new survey responses. The news release concludes with one page background information describing the survey methodology.

The euro area PMI is a widely used economic indicator that measures the prevailing direction of economic conditions. It is based on a monthly survey of purchasing managers of a representative panel of around 5,000 services and manufacturing companies. The managers are asked about various aspects of their business, including production, new orders, inventory levels, backlogs, supplier deliveries, employment and prices and costs conditions. When responding to the survey questions on the set of metrics of their business, the purchasing managers specify whether they have increased, decreased, or stayed the same between the last and current month. The categorical responses are aggregated into publicly released diffusion indexes. In addition to the categorical response, purchasing managers can provide further explanation in accompanying text boxes. There are free response questions accompanying nearly every categorical question, asking for the reason for the response. In addition, there is a "General Remarks" field, where the respondents can put any general remark. Some of these text responses are featured in the PMI news release to provide context for the diffusion indexes, but otherwise are not released publicly.

The primary media attention is the flash release of the composite PMI that covers the services and manufacturing sectors but not all sectors of the economy, e.g., non-private services and construction are missing. The flash release is published around the third week of the respective month, and the final release is about two weeks later, i.e., in the first week of the next month. The flash composite output and the manufacturing PMI survey news releases for the euro area start in January 2008, while the final services news releases are available from September 2008. A sample starting in 2008 has the advantage that it covers several complete business cycles, including the recessions of the global financial crisis, the euro area government debt crisis, and the Corona pandemic. The key activity question asked to survey respondents in the services PMI surveys is "has your business activity (in units) risen, fallen or remained unchanged on that of one month ago". For the manufacturing sector the corresponding activity question is about output. Services business activity has over 2008-2024 a weight of about 73% and manufacturing output accounts for the remaining 27%. The euro area services PMI is based on original survey data collected from a representative panel of around 2,000 private sector services firms. National services data are included for Germany, France, Italy, Spain, and Ireland. These countries together account for an estimated 78% of the euro area private sector services output. The euro area manufacturing PMI is collected from a representative panel of around 3,000 manufacturing firms. National data are included for Germany, France, Italy, Spain, the Netherlands, Austria, Ireland, and Greece. These countries account for an estimated 89% of euro area manufacturing activity.

The survey responses are converted into a diffusion index which is then seasonally adjusted. A diffusion index is calculated as follows, where P is the percentage number of responses of total responses, with a range between 0 and 100. A diffusion index is a simple weighted average of the aggregate "rise" and "unchanged" response shares in the panel with weights of 1 and 0.5 respectively, see Eq. (1). It indicates the degree to which the indicated change is dispersed or diffused throughout the panel and has a theoretical no-change threshold of 50. An index reading over 50 indicates an improvement in output, while anything below 50 suggests deterioration.

$$\text{PMI diffusion index}_t = 1.0 * \text{P(rise)}_t + 0.5 * \text{P(unchanged)}_t + 0.0 * \text{P(decline)}_t \qquad (1)$$

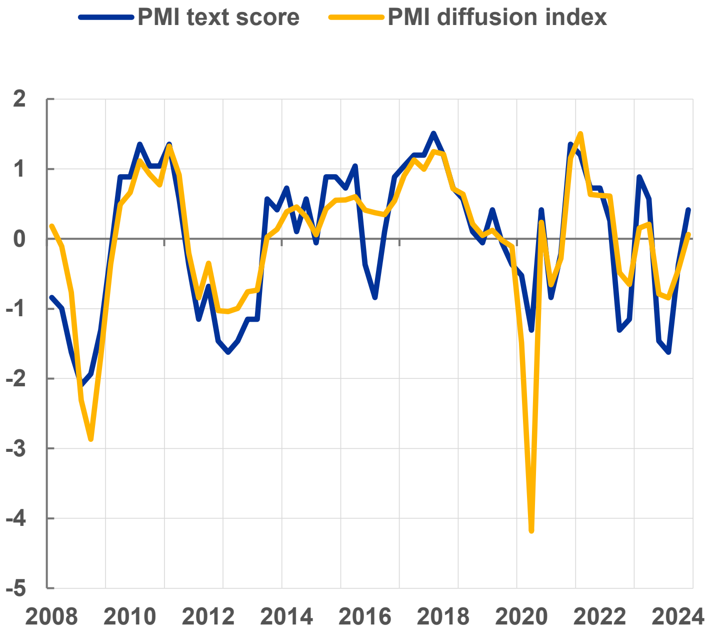
2.1 ChatGPT-based PMI text scores

OpenAI's ChatGPT is used to classify activity reported in PMI news releases. ChatGPT is by far the most popular large language model (Korinek, 2023). Fatouros et al. (2023) and Hansen and Kazinnik (2023) report that it works better than other large language models such as BERT. BERT typically necessitates resource-intensive, task-specific fine-tuning on particular datasets, making it more dependent on user or modeller intervention for adaptation to new tasks. Similarly, VADER relies on a pre-defined lexicon of words with associated sentiment values, which also requires significant user interaction and lacks contextual understanding. These limitations make both BERT and VADER less flexible compared to ChatGPT. Smales (2023) demonstrates ChatGPT's capability to classify monetary policy statements according to whether they are hawkish or dovish. The ChatGPT version used is 4.o, with default settings. The latter offers a basic balance between coherence in the responses on the one hand and creativity on the other hand. We pay special attention to the prompt, which is essentially a form of programming in natural language. Careful "prompt engineering" is needed given a potential issue of replication. We therefore instruct to clear prior context, like the prompt used in a ChatGPT application to predict equity premium (Ma et al., 2024). We thus apply a zero-shot sentiment analysis, exploiting ChatGPT's ability to perform a task without having been explicitly trained on any examples specific to that task. Instead, ChatGPT relies on its general language understanding, which it has developed from extensive pre-training on a diverse corpus of text. In this study, we devise a precise prompt to guide ChatGPT in classifying activity as reported in flash PMI text commentaries. The focus is on the flash PMI release due to a larger media attention and coverage than the release of the final manufacturing, respectively, final services about one to two weeks later. The prompt acts as the starting point for ChatGPT's generated responses, not only providing the context and directives for the analysis but also determining the relevance and appropriateness of the model's output. The prompt is rigorously designed to maintain clarity and consistency in the response mechanism of the model. The basic prompt reads as follows:

Clear all prior contexts and instructions. As an economic expert, read the attached document of PMI composite flash press release and return a score about activity, and give me one score assessing the performance of economic activity (consider manufacturing output, services business activity) between -4 and +4. -4 means significantly contracting activity and +4 means significantly expanding activity. Please add the final score in a table, add in the same table also a column for the release date, month it is covering, as well as calendar year. Please refrain from providing further comments.

The resulting monthly text score numbers are transformed into quarterly averages given our focus on GDP nowcasting and GDP is only available quarterly. Figure 1 plots the quarterly averages of the ChatGPT-derived PMI text score and the composite output PMI diffusion index given the focus on nowcasting quarterly GDP. A normalised y-axis scale is used as the text score varies between -4 and +4 and the diffusion index broadly between 30 and 60. The two series often comove closely, but there are periods where they deviate substantially. During the global financial crisis and pandemic recessions, the diffusion index turned out more negative than the text score. On other occasions like the euro debt crisis recession and at the end of the sample the text score moved more pronounced negative. The relationship between the quarterly averages of the PMI text score and PMI diffusion index is further elaborated in Appendix A. It reports a formal Granger causality test between the two quarterly PMI series as well as out-of-sample one-quarter ahead forecasts of the composite output diffusion index, using the PMI text score as the sole predictor.

Figure 1 Euro area PMI text score and composite output diffusion index



Note: normalised scale of quarterly averages of monthly data.

2.2 Alternative text scores

In addition to the basic ChatGPT-generated monthly PMI text score, three alternatives were considered for robustness checks: (i) employing a different prompt, specifically the regularly asked activity question in the ECB's dialogue with non-financial companies (Elding et al., 2021); and (ii) split the prompt by sector, separating the prompt into one for manufacturing and one for services with separate text scores for manufacturing and services derived from their respective final news releases;[1] (iii) using the basic prompt with theoretically correct monthly weightings.

For the alternative prompting, we adopted a broader definition of activity, similar to that used in the ECB's dialogue with non-financial companies (Elding et al., 2021), which regularly asks: "How has euro area activity (for instance, output, sales, deliveries, orders...) evolved in recent months in your sector, and what do you expect to happen in the near term? What are the main factors underlying this assessment?". Additionally, we provided a more detailed description of the activity classification score in line with guidelines given to ECB staff who score activity after the call.

The alternative ECB contacts prompt is as follows:

Clear all prior contexts and instructions. As an economic expert, read the attached document of PMI composite flash press release and return a score about activity, (for instance, output, sales, deliveries, orders...). Choose between the options -4 (very strong decrease), -3 (stronger-than-usual decrease), -2 (normal decrease), -1 (marginal decrease), 0 (no change), +1 (marginal increase), +2 (normal increase) +3 (stronger-than-usual increase), +4 (very strong increase). Please add the final score in a table, add in the same table also a column for the release date, month it is covering, as well as calendar year. Please refrain from providing further comments.

Regarding the sector split, we use the final manufacturing news release for manufacturing output classification and the final services news release for services business activity classification. These are then averaged using the value-added weights as implicitly used by S&P Global to construct the composite output diffusion index. For the sample 2008 to 2024 the applied weight is 27% for manufacturing and 73% for services.

To align monthly scores with quarterly GDP data, it is necessary to determine the appropriate weights for converting data underlying monthly changes to quarterly growth rates. Instead of using a simple quarterly average of the monthly series, we opted for a theoretically correct weighted average to account for the PMI survey question's focus on month-on-month rather than quarter-on-quarter changes. This method involves five monthly series with weights of 1, 2, 3, 2, and 1, applied to the second and third months of the previous quarter and the first, second, and third months of the current quarter, respectively.

The correlation between on the one hand the various PMI text scores on the one hand and the PMI diffusion index on the other hand is, apart from one case, high, close to 0.9 (see Table 1). The only exception is the ECB contacts prompt with has a lower correlation of 0.69. This lower correlation is likely due to a broader definition of activity, as PMI news releases often report extensively about new orders and delivery times.

[1] The four monthly PMI text score series are available upon request. Any subsequent use requires citation of this work.

Table 1 Correlations between composite output PMI diffusion index and PMI text scores

	Diffusion index	ChatGPT-based text score			
		Basic prompt	ECB contacts prompt	Sector split prompt	Theoretical weights
Diffusion index	1	0.85	0.69	0.89	0.87
Text score, basic prompt	0.85	1	0.87	0.96	0.97
Text score, ECB contacts prompt	0.69	0.87	1	0.84	0.78
Text score, sector split prompt	0.89	0.96	0.84	1	0.95
Text score, theoretical weights	0.87	0.97	0.78	0.95	1

Note: Quarterly averages of monthly data, apart from theoretical weights.

3. Nowcasting GDP

The usefulness of the ChatGPT-based PMI text scores is analysed for nowcasting real GDP growth using two challenging nowcast benchmarks: (i) the GDP nowcasts embedded in ECB/Eurosystem staff projections (ESP), which incorporate the judgement of economic experts, and (ii) the first GDP estimates by Eurostat (referred to as first GDP), where missing statistical information is addressed by statistical experts. For the realized outcome, 'actual GDP', we use the latest-available GDP time series data at the point of writing. It represents the 'true' real GDP based on all statistical information at the time of the release.

The starting point is in-sample evidence, which is not sufficient on its own, but necessary for establishing any out-of-sample accuracy. Without significant in-sample explanatory power, out-of-sample forecast gains would lack credibility. The in-sample evidence is based on regressing quarter-on-quarter real GDP growth, *y*, on a constant, the PMI text score, *pts*, and with a coefficient of one of the respective nowcast benchmark, *nc*. The benchmark can be real GDP growth from the ESP, *ESP*, or Eurostat first GDP estimate, y^{first}. Unless stated otherwise, the PMI text score is based on the quarterly average of the monthly series and the basic prompt. The regression model reads as follows, with ε being the error term.

$$y_t = \alpha + \beta pts_t + 1.0 * nc_t + \varepsilon_t \tag{2}$$

with nc = ESP or y^{first}

Working in real time with PMI text scores, one would not expect any bias and a priori thus assumes the constant to be zero. This results in a simplification of Eq. (1), with α equal to zero.

$$y_t = \beta pts_t + 1.0 * nc_t + \varepsilon_t \tag{3}$$

with nc = ESP or y^{first}

The estimates for three different samples are reported: (i) total sample from 2008Q1 to 2024Q2; (ii) excluding extreme observations; (iii) across three different real GDP growth quantiles (low, median, and high growth). The total sample starts in 2008Q1, the first quarter for which we have PMI news releases and ends in 2024Q2. To check whether the estimates are sensitive to the extraordinary Covid-19 pandemic related quarters, a sample excluding extreme observations is also reported. Extreme observations are determined by looking at the composite output PMI diffusion index, as these series are available in real time, whereas this is not the case for the final GDP release. Extremes are defined as a quarterly average composite output PMI diffusion index below 40 or an absolute change in the three-month change in this index larger than 10. It results in the exclusion of four observations, as also illustrated in Chart A in de Bondt and Saiz (2024): first quarter of 2009 during the global financial crisis and the first three quarters of 2020 during the Covid-19 pandemic.

3.1 In-sample evidence on enhancing GDP nowcasts with PMI text scores

Table 2 reports the estimates of Eq. (2) and (3), in the upper panel for the ESP as nowcast benchmark and for the bottom panel for Eurostat's first GDP estimate as benchmark. As regards estimates of Eq. (2), the main observation is that the constant for the total sample is not significantly different from zero, suggesting no overall bias. For the subsample estimates this is, as one expects, no longer the case. Excluding quarters with extreme observations, the constant is estimated to be significantly positive. Across growth quantiles, the constant is significantly negative for the low growth quantile and positive for the high growth quantile.

Turning to estimates of Eq. (3), the main conclusion is that the quarterly average of the PMI text scores significantly enhances the GDP nowcasts, for both hard-to-beat benchmarks. The PMI text activity score always significantly positively helps in explaining final real GDP growth on top of the ESP nowcast. The same holds for Eurostat's first GDP release apart from the total sample estimates. Quantitatively the estimated impact of the ChatGPT-based score is consistently somewhat higher for the ESP nowcasts as for the first GDP estimate. The ESP nowcasts could have been improved by adding 5% to 15% of the PMI text score, whereas this impact varies between 3% and 8% for the first GDP estimate.

Table 2 Nowcasting GDP using PMI text scores

	Sample									
	All		Excluding extremes		GDP growth quantile					
					0.25		0.50		0.75	
	ESP									
Constant	0.06	-	0.13 **	-	-0.22 ***	-	0.04	-	0.20 ***	-
PMI text score	0.07	0.08 *	0.02	0.05 *	0.15 ***	0.08 ***	0.09 **	0.10 ***	0.09 **	0.15 ***
ESP	1.00 -	1.00 -	1.00 -	1.00 -	1.00 -	1.00 -	1.00 -	1.00 -	1.00 -	1.00 -
S.E. of regression	0.81	0.81	0.38	0.47	0.81	0.79	0.79	0.78	0.80	0.77
R-squared [1]	0.86	0.86	0.54	0.31	0.13	0.06	0.08	0.08	0.03	-0.09
	First GDP									
Constant	0.07	-	0.07 *	-	-0.09 ***	-	0.01	-	0.17 ***	-
PMI text score	0.02	0.03	0.03 *	0.04 ***	0.06 ***	0.07 ***	0.08 ***	0.08 ***	0.06 *	0.08 ***
First GDP	1.00 -	1.00 -	1.00 -	1.00 -	1.00 -	1.00 -	1.00 -	1.00 -	1.00 -	1.00 -
S.E. of regression	0.31	0.31	0.25	0.25	0.33	0.30	0.30	0.30	0.32	0.30
R-squared [1]	0.98	0.98	0.81	0.79	0.14	0.05	0.13	0.13	0.03	-0.19

Note: *** <1%, ** <5%, * <10%. [1] Adjusted R-squared for OLS regressions and "pseudo" R-squared for quantile regression as it does not minimize the sum of squared residuals and therefore does not produce a traditional R-squared value.

What is behind the predictive power of PMI text scores? PMI text scores derived from news releases add value in three ways. First, a PMI news release provides a narrative, reflecting a human-based assessment of the raw data. Such an assessment is not mechanical but based on information available at the point of writing. Díaz Sobrino et al. (2021) illustrate the importance of a narrative for Spanish GDP forecasts. Sharpe et al. (2023) report the predictive power of the extraction of sentiment from the narratives of the Green Book for US GDP. They conclude that the tonality conveys substantial incremental information. Furthermore, human behaviour is influenced by stories and narratives and these narratives can have an impact on economic outcomes, as argued in Shiller's (2019) book on narrative economics. Second, the PMI news release reports beside the quantitative diffusion index anecdotal evidence. Purchasing managers can report in the PMI survey "open text" and news releases make use of these statements. Often a couple of these statements "make it" in the official news release. Du et al. (2024) report for the US that anecdotes do matter for exploring the Beige Book. Third, the role of services appears key. The importance of services for the composite output PMI diffusion index-based GDP tracking record is earlier reported for the euro area in de Bondt (2019, Section 5.3). Services data is hardly available in real time for our two benchmarks. At the time of publication of the first GDP estimate only one month of market services production is available and two months of manufacturing production. Given the economic importance of services and corresponding high weight of services in the composite output index of 73%, the inclusion of the services PMI information is essential in improving the GDP nowcasts. Landefeld et al. (2008) report for the US that initial estimates often rely on extrapolations, especially for the service sector, due to limited data availability. For the first release of US GDP, data on about 25 percent of GDP are not available and estimated based on past trends and whatever related data are available. The first GDP estimate based on these extrapolations are revised as more complete data become available.

The incremental information of the PMI text score in nowcasting euro area real GDP is further explored by looking at real-time results during the quarter. The outcomes so far are based on the quarterly average of all three months of the quarter, whereas the ESP nowcasts is made around the time when only the first two months of the quarter are available for the PMI. Table 3 provides insights about the real-time performance of PMI text score. It reports the results for (i) the first month of the quarter; (ii) the average of the first two months of the quarter. Estimates without a constant are reported, as in real time one will apply a constant of zero. The estimates of the coefficient of the PMI text scores are no longer significant apart from the quantile regression estimates using the average of the first two months. Only if all three months of the quarter are available the PMI text score-based enhancement in GDP nowcasting is sufficiently prominent. This finding is in line with Ashwin et al. (2024), which show that the advantage of the PMI is especially at the end of the quarter. It also confirms a common finding that forecast accuracy improves with incoming new data.

Table 3 Nowcasting GDP using PMI text score during the first two months of the current quarter

	Sample									
	All		Excluding extremes		GDP growth quantile					
					0.25		0.50		0.75	
Months	1	1+2	1	1+2	1	1+2	1	1+2	1	1+2
	ESP									
PMI text score	0.05	0.05	0.02	0.03	0.03	0.04 **	0.04	0.04 *	0.05 **	0.07 ***
	First GDP									
PMI text score	0.00	0.02	0.02	0.02	0.02	0.04 **	0.04 **	0.05 ***	0.05 ***	0.06 ***

Note: *** <1%, ** <5%, * <10%. [1] Adjusted R-squared for OLS regressions and "pseudo" R-squared for quantile regression as it does not minimize the sum of squared residuals and therefore does not produce a traditional R-squared value.

As regards the first GDP estimate, it is released when the first month of the next quarter is already available for the PMI. It is interesting to explore whether the flash PMI news release of the first month of the next quarter, which becomes available about one week before, has additional predictive content. The PMI news release typically contains an assessment about the first GDP release one week later. This GDP release practice was introduced by Eurostat for the first time for 2016Q1. Before, the first GDP vintage was two weeks later, for which thus also the final PMI new release would have been available. Table 4 reports the results for an average calculated over four months, i.e., all three months of the current quarter and the first month of the next quarter as well as only using the PMI news release of the first month of the next quarter. Using up-to-date information as available in real time, the results clearly indicate that the PMI text score has significant explanatory power on top of the first GDP estimate for nowcasting final GDP. This holds for taking an average of monthly PMI text scores over the most recent four months as well as only over the most recent month, i.e., the first month of the next quarter.

Table 4 GDP nowcast estimates at the time of the first GDP estimate

	Sample									
	All		Excluding extremes		GDP growth quantile					
					0.25		0.50		0.75	
Months	Four[1]	One[2]	Four[1]	One[2]	Four[1]	One[2]	Four[1]	One[2]	Four[1]	One[2]
PMI text score	0.04 **	0.05 ***	0.04 ***	0.04 ***	0.05 ***	0.05 ***	0.05 ***	0.05 ***	0.07 ***	0.05 ***
S.E. of regression	0.31	0.30	0.25	0.25	0.31	0.30	0.31	0.30	0.32	0.30
R-squared [3]	0.98	0.98	0.80	0.80	0.04	0.05	0.10	0.10	-0.29	-0.31

Note: *** <1%, ** <5%, * <10%. [1] Over four months: all three months of the current quarter plus the first month of next quarter. [2] First month of the next quarter. [3] Adjusted R-squared for OLS regressions and "pseudo" R-squared for quantile regression as it does not minimize the sum of squared residuals and therefore does not produce a traditional R-squared value.

3.2 Robustness

Three types of robustness checks are applied: (i) check whether the PMI text score retains its significance in regressions that also include the composite output PMI diffusion index; (ii) alternative prompts based on the activity question used in the survey of ECB contacts with leading non-financial corporations as well as a sector split prompt using the final news release for manufacturing, respectively, for services; (iii) using an alternative weighting over months, i.e., theoretical monthly weighting as well as Mixed Data Sampling

(MIDAS) regression based weights. The robustness checks show that the result that PMI text scores enhance GDP nowcasts from the ESP and Eurostat's first GDP estimate remains adding the composite output PMI diffusion index, for alternative prompts, and different monthly weighting schemes.

Table 5 reports the ordinary least squares (OLS) estimates by adding the composite output PMI diffusion index as explanatory variable in Eq. (2) and (3). The diffusion index is often estimated to be negative, of which five times significantly negative. This sign is counterintuitive. Only for the low growth case a positive coefficient is estimated. In contrast, the estimated PMI text score coefficients remain, apart from one insignificant exception, positive and often significantly positive.

Table 5 OLS estimates of GDP nowcasts using PMI text score and diffusion index

Regressor	Sample									
	All		Excluding extremes		GDP growth quantile					
					0.25		0.50		0.75	
	ESP									
Constant	0.059	-	-0.236	-	2.041	-	-0.923	-	-3.556	-
PMI text score	0.069	0.071	0.046	0.038	0.004	0.067 *	0.063	0.053 **	0.127	-0.005
PMI diffusion index	0.000	-0.001	-0.007	-0.002	0.045	0.002	-0.020	-0.001	-0.078	-0.006 ***
	First GDP									
Constant	-1.778 *	-	-1.316	-	-0.564	-	-1.971	-	-2.233 ***	-
PMI text score	0.092 ***	0.024	0.073 **	0.030 **	0.070	0.056 ***	0.121	0.051 ***	0.096 ***	0.017
PMI diffusion index	-0.037 *	-0.002	-0.028	-0.002 *	-0.010	0.001 **	-0.041	-0.001	-0.050 ***	-0.004 ***

Note: *** <1%, ** <5%, * <10%.

The first type of robustness check is further explored by applying robust least squares, which is designed to provide reliable parameter estimates in regression analysis, especially when the data contains outliers or violates some of the assumptions of OLS regression. Among its advantages are its resistance to outliers, handling of non-normal errors, and better performance in small samples. The applied robust least squares method is a Huber regression. The robust least squares estimates indicate also the limited value added of the addition of the PMI diffusion index. The diffusion index coefficient is insignificant or significantly with a counterintuitive negative sign (see Table 6). At the same time, the PMI text score coefficients explain in all cases significantly final GDP beyond the two benchmarks.

Table 6 Robust least square estimates of GDP nowcasts using PMI text score and diffusion index

Regressor	Sample							
	All				Excluding extremes			
	ESP		First GDP				First GDP	
Constant	-4.367 ***	-	-2.247 ***	-	-1.634	-	-0.686	-
PMI text score	0.162 ***	0.044 **	0.110 ***	0.043 ***	0.082 *	0.037 **	0.065 **	0.043 ***
PMI diffusion index	-0.091 ***	-0.001	-0.047 ***	-0.001	-0.035	-0.001 *	-0.015	-0.001 *

Note: *** <1%, ** <5%, * <10%. OLS estimates.

Given multicollinearity between the PMI text score and composite output PMI diffusion index can cause problems in OLS regression, such as inflated standard errors and unstable coefficient estimates, we additionally apply ridge regression. This estimation method deals with multicollinearity among regressors by adding a penalty term to the loss function, which shrinks the coefficients and reduces their variance. The ridge regression estimates also typically find a negative coefficient for the PMI diffusion index and the two cases of a positive diffusion index coefficient are for a specification including a constant which is not realistic for a real-time application.

Table 7 Ridge regression: nowcasting GDP using PMI text score and diffusion index

Regressor	Sample							
	All				Excluding extremes			
	ESP		First GDP		ESP		First GDP	
Constant	0.825		-1.777		0.496		-1.036	
PMI text score	0.018	0.018	0.092	0.018	0.014	0.003	0.063	0.021
PMI diffusion index	0.015	-0.002	-0.037	-0.002	0.008	-0.002	-0.023	-0.002

Note: No significance level available for ridge regression due to biased estimates.

In sum, the first type of robustness check indicates that there is no value added in adding the PMI diffusion index to the ChatGPT-derived PMI text score. Consequently, our preferred model specification remains a specification without a constant and based only on the PMI text score.

Turning to the other two robustness checks, Table 8 reports estimates using three alternative PMI text scores: two scores based on alternative prompts and one based on theoretical monthly weights. For the alternative ECB contacts prompt as well as the theoretical monthly weighting, the estimated PMI text score coefficients and associated significance level, are consistently somewhat lower than the basic regression results as reported in Table 2. In contrast, for the sector split prompt that uses the final PMI news releases, the enhancement of the GDP nowcasts improves even further. This is what one expects given the more up-to-date assessment of the final numbers. Noteworthy is especially that the total sample estimates of the PMI text score coefficient are for the sector split prompt also significantly positive at the 5% significance for both nowcasts benchmarks. The estimated coefficients of the PMI text scores for the ESP nowcasts are apart from the low growth quantile again found to be somewhat higher than for the first GDP estimate. This finding suggests that some of the nowcasting content of the PMI text score beyond the ESP nowcasts is reflected in the first GDP estimate.

Table 8 Nowcasting GDP using alternative PMI text scores

	Sample									
	All		Excluding extremes		GDP growth quantile					
					0.25		0.50		0.75	
	ESP	First	ESP	First	ESP	First	ESP	First	ESP	First
Prompt along ECB contacts	0.07	0.02	0.05	0.03 *	0.05 ***	0.07 ***	0.05 **	0.06 **	0.05 **	0.04 ***
Prompt sector split	0.08 **	0.04 **	0.06 **	0.05 ***	0.02	0.04 **	0.07 ***	0.05 ***	0.09 ***	0.08 ***
Theoretical monthly weights	0.06	0.03	0.04	0.03 **	0.04 *	0.04 ***	0.05 *	0.05 ***	0.07 ***	0.06 ***

Note: *** <1%, ** <5%, * <10%. Sector split starts in 2008Q4 rather than 2008Q1.

In complement to the simple average and theoretical weights, insights of the relevance of the monthly PMI text score series over time for quarterly real GDP can be provided by MIDAS regression. This approach allows to incorporate high-frequency data, in our case the monthly PMI text score series, into a lower-frequency model, i.e., quarterly real GDP growth. Polynomial Distributed Lags (PDL) is a method used in MIDAS to parsimoniously incorporate lags of high-frequency variables. Instead of estimating a separate coefficient for each lag, PDL uses a polynomial function to model the lagged effects, reducing the number of parameters to estimate. The coefficients of the polynomial indicate the shape of the response over time. The MIDAS estimates show that the effect of the monthly PMI text score on quarterly GDP growth is strongest for the most recent months and then strongly diminishes. It suggests that looking at the most recent monthly PMI text score outcome is sufficient for insights about quarterly GDP growth.

Table 9 MIDAS regression coefficients for monthly PMI text score

Sample												
All							Excluding extremes					
Lag	BP		CP		SP		BP		CP		SP	
	ESP	First GDP	ESP	First GDP	ESP	First GDP	ESP	First GDP	ESP	First GDP	ESP	First GDP
0	0.22	0.09	0.42	0.09	0.30	0.07	0.14	0.08	0.22	0.13	0.09	0.07
1	-0.14	0.01	-0.32	0.00	-0.20	0.01	-0.09	0.01	-0.16	-0.10	0.01	0.01
2		-0.07		-0.08		-0.05		-0.06			-0.06	-0.05
Sum	0.08	0.03	0.10	0.01	0.10	0.02	0.04	0.03	0.06	0.04	0.04	0.03

Note: BP = basic prompt; CP = ECB contacts prompt; SP = sector split prompt based on final manufacturing, respectively, services PMI text commentaries. MIDAS regression with automatic monthly lag selection up to 6 and a polynomial degree of 2. The two polynomial distributed lags are significant at 1% apart from sector split prompt excluding extremes with a significance around 5%.

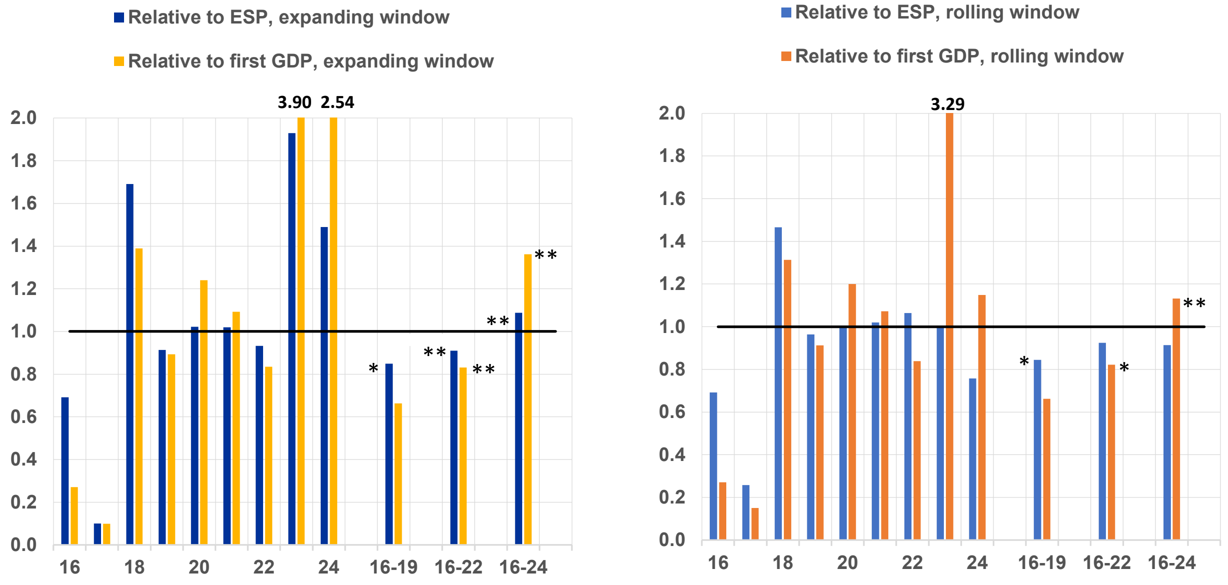
3.3 Out-of-sample nowcasting performance

Our main finding that PMI text scores significantly enhance GDP nowcasts from the ESP and Eurostat's first GDP estimate in sample raises the question how much forecast gains are achieved out-of-sample? The out-of-sample exercise uses the first two months of the quarter of the PMI text score for the ESP benchmark and all three months of the quarter for the first GDP benchmark. Overall, the gains in nowcasting performance can be substantial but are time dependent, with strong variation across calendar years.

Figure 2 plots the out-of-sample root mean squared error (RMSE) relative to the RMSE of the respective benchmark by calendar year. A value below one implies that the nowcast enhanced with the ChatGPT-based PMI text score outperforms the benchmark without this enhancement. Two out-of-sample forecasts were generated, one with Eq. (3) estimated with an expanding window excluding extremes and one with a fixed eight-year window. The outcomes for each calendar year are based on estimates of Eq. (3) up to the fourth quarter of the previous year. The calendar year outcomes for 2016 are thus based on Eq. (3) estimated up to 2015Q4, 2017 outcomes up to 2016Q4, etc. For the results with an expanding window the start of the estimation period remains 2008Q1, whereas for the fixed window the start of the estimation period moves by one year, so that the estimation period remains eight years. The latter reflects the average length of a full business cycle. The nowcast performance relative to the two benchmarks varies a lot by calendar year, with substantial forecast gains in 2017 and losses in 2023 and 2024. On average relative forecast gains are obtained excluding the two most recent years: 8% to 9% relative to ESP and 17% to 18% relative to the first GDP estimate. It is remarkable that the enhancement of GDP nowcasts with PMI text scores is also informative during extreme events like the pandemic. The nowcast errors relative to the ESP are with a rolling window also for the full out-of-sample period below one and thus indicates forecast gains relative to the ESP GDP nowcasts. The out-of-sample errors are on average lower with a rolling window than for an expanding window. The relative nowcast performance for the two most recent years is poor but a likely explanation is that the final GDP is not yet reflecting "true" GDP, with still missing statistical data.

To determine whether the improvements in nowcasting accuracy since 2016 are statistically significant, we employ the Clark and West (2007) test. This test is an extension of the Diebold and Mariano (1995) test, tailored for comparisons involving nested models. This is particularly relevant for our analysis, as we evaluate nowcasts from models enhanced with PMI text scores against benchmark models without them. The results of the CW test are presented for three sample periods: 2016-2019, 2016–2022, and 2016-2024. These results are visualized in Figure 2, where statistical significance is denoted by asterisks. A statistical improved nowcasting performance is achieved at the 10% significance level compared to the ESP benchmark for the period ending in 2019. Similarly, the performance is statistically improved at the 10% compared to the first GDP benchmark using a rolling window for the period ending in 2022. However, for the sample extending to 2024, the nowcasting performance is found to be statistically poorer at the 5% significance level when compared to both nowcasting benchmarks using an expanding window and relative to the first GDP benchmark using a rolling window.

Figure 2 Out-of-sample nowcasting performance relative to benchmarks

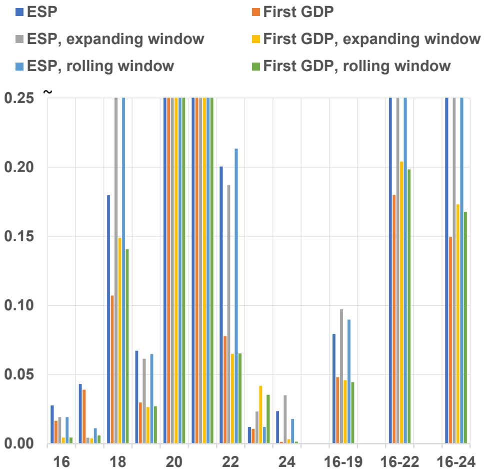


Note: Root mean squared error of out-of-sample GDP nowcasts enhanced with PMI text score relative to ESP and first GDP benchmarks, with expanding window excluding extremes as well as with a rolling eight-year window. ** p < 0.05, * p < 0.10 refers to the Clark West (CW) test statistics, indicating a significant difference in the nowcasts from the model including the PMI text score versus those of the respective benchmark model.

Given the sudden deterioration of the nowcasting performance relative to the two benchmarks in the two most recent years, Figure 3 plots the nowcast RMSE in absolute terms, , with the y-axis truncated at 0.25.

In level terms the difference in errors for 2023 and 2024 between the benchmarks and the enhanced models are small. The errors of the two benchmarks are, however, exceptionally small, especially the first GDP benchmark error for 2024. For 2024 this is no surprise as for 24Q2 the final GDP equals the first GDP estimate, resulting in an error of zero. The RMSE of the first GDP was in 2023-24 0.006 compared to 0.048 in 2016-19. The RMSE of the ESP was a bit higher but also small compared to pre-Covid: 0.018 versus 0.079. The RMSE in level terms of the GDP nowcasts enhanced with the PMI text score are also in 2023 and 2024 low, varying between 0.003 and 0.042.

Figure 3 Out-of-sample nowcasting performance in absolute terms



Note: Root mean squared error of GDP nowcasts from ESP and first GDP benchmarks as well as ESP and first GDP nowcasts out-of-sample enhanced with PMI text score based on expanding, respectively, rolling window. Y-axis truncated at 0.25.

4. Conclusion

This study demonstrates for the first time the predictive power of textual information content contained in PMI news releases for nowcasting euro area GDP. It embeds ChatGPT-derived activity sentiment scores from qualitative text as reported in PMI news releases into existing GDP nowcasts. In-sample estimates indicate that these PMI text scores significantly improve the accuracy of GDP nowcasts for the euro area. Specifically, the inclusion of PMI text scores can help improve nowcast performance, whereas this is not the case for additionally including the composite output PMI diffusion index. The out-of-sample forecast gains are found to be time dependent, differing a lot by calendar years. Pre-covid and during the Covid years the forecast gains relative to the two hard-to-beat benchmarks were about 20%. In the two most recent years, the nowcast errors were low in absolute terms but not in all cases relative to the benchmarks, especially due to close to zero errors of the first GDP estimate benchmark.

Our research validates the added value of integrating textual analysis into economic forecasting. The qualitative insights captured by ChatGPT, which include the tone, narrative, and anecdotal nuances from PMI news releases, provide substantial incremental information beyond the raw quantitative data of diffusion indices. This suggests that economic forecasts can benefit significantly from embedding qualitative data sources. Moreover, the robustness of these findings is underscored by various methodological adjustments, including different prompts. The predictive power of the PMI text scores highlights the importance of narrative context for economic forecasting, challenging the conventional reliance on purely numerical data.

The remarkable GDP performance of the PMI text scores for tracking the true final GDP is also presented to trigger discussion in the field of utilizing "big data" in textual analysis for economic forecasting. It appears unclear whether economic forecasters should focus on expanding data dimension or adopting a parsimonious approach focusing on a selectively economically meaningful small set of text data, in our case even only two pages of text. The presented new evidence demonstrates a decisive role for the quantitative classification of the qualitative assessment reported in the text of the PMI news release in euro area GDP nowcasting.

Future research could expand on our approach by applying ChatGPT-based sentiment analysis to other economic indicators or textual data sources, including forecast reports of policy institutions. Additionally, exploring the potential of ChatGPT in nowcasting GDP for other regions or countries might provide insightful information about the applicability of this method worldwide. This study not only contributes to the fast-growing field of AI in economic forecasting but also opens new avenues for integrating advanced large language models into real-time economic analysis. The proven effectiveness of ChatGPT in enhancing GDP nowcasts points to a promising direction for future research and practical applications in forecasting analysis.

References

Ashwin, J., Kalamara, E. and Saiz, L. (2024). 'Nowcasting Euro area GDP with news sentiment: a tale of two crises', *Journal of Applied Econometrics*, Vol. 39, pp. 887–905.

Babii, A., Ghysels, E. and Striaukas, J. (2022). 'Machine learning time series regressions with an application to nowcasting', *Journal of Business and Economic Statistics*, Vol. 40, pp. 1094–1106.

Barbaglia, L., Consoli, S. and Manzan, S. (2023). 'Forecasting with economic news', *Journal of Business and Economic Statistics*, Vol. 41, pp. 708–719.

Basselier, R., de Antonio Liedo, D., and Langenus, G. (2018). 'Nowcasting real economic activity in the euro area: assessing the impact of qualitative surveys', *Journal of Business Cycle Research*, Vol. 14, pp. 1–46.

Clark, T.E. and West, K.D. (2007). 'Approximately normal tests for equal predictive accuracy in nested models', *Journal of Econometrics*, Vol. 138, pp. 291–311.

de Bondt, G.J. (2012). 'Nowcasting: trust the Purchasing Managers' Index or wait for the flash GDP estimate?' in Papanikos, G.T. (ed.), *Economic Essays*, Athens Institute for Education and Research, pp. 83–97.

de Bondt, G. J. (2019). 'A PMI-based real GDP tracker for the euro area', *Journal of Business Cycle Research*, Vol. 15, pp. 147–170.

de Bondt, G. and Saiz, L. (2024). 'Is the PMI a reliable indicator for nowcasting euro area real GDP?', *ECB Economic Bulletin*, Issue 1, Box 2.

Díaz Sobrino, N., Ghirelli, C., Hurtado, S., Pérez, J. J., and Urtasun, A. (2022). 'The narrative about the economy as a shadow forecast: an analysis using Bank of Spain quarterly reports', *Applied Economics*, Vol. 54, pp. 2874–2887.

Diebold, F.X. and Mariano, R.S. (1995). 'Comparing predictive accuracy', *Journal of Business and Economic Statistics*, Vol. 13, pp. 253–263.

Du, S., Guo, K., Haberkorn, F., Kessler, A., Kitschelt, I., Lee, S. J., Monken, A., Saez, D., Shipman, K. and Thakur, S. (2024). 'Do anecdotes matter? Exploring the Beige Book through textual analysis from 1970 to 2023', Irving Fisher Committee on Central Bank Statistics (IFC) Bulletin, No. 57.

Elding, C., Morris, R. and Slavik, M. (2021). 'The ECB's dialogue with non-financial companies', *ECB Economic Bulletin*, Issue 1.

Fatouros, G., Soldatos, J., Kouroumali, K., Makridis, G. and Kyriazis, D. (2023). 'Transforming sentiment analysis in the financial domain with ChatGPT', *Machine Learning with Applications*, Vol. 14, 100508.

Ferrara, L. and Simoni, A. (2022). 'When are Google data useful to nowcast GDP? An approach via preselection and shrinkage', *Journal of Business and Economic Statistics*, pp. 1–15.

Galimberti, J. K. (2020) 'Forecasting GDP growth from outer space', *Oxford Bulletin of Economics and Statistics*, Vol. 82, pp. 697–722.

Hansen, A.L. and Kazinnik, S. (2023). 'Can ChatGPT Decipher Fedspeak?', August 5. Available at SSRN: https://ssrn.com/abstract=4399406.

Korinek, (2023). 'Generative AI for economic research: use cases and implications for economists', *Journal of Economic Literature*, Vol. 61, pp. 1281–1317.

Landefeld, J. S., Seskin, E. P. and Fraumeni, B. M. (2008). 'Taking the pulse of the economy: measuring GDP', *Journal of Economic Perspectives*, Vol. 22, pp. 193–216.

Ma, F., Lyu, Z., Li, H. (2024). 'Can ChatGPT predict Chinese equity premiums?', *Finance Research Letters*, Vol. 65.

Sharpe, S., Sinha, N. and Hollrah, C. (2023). 'The power of narrative sentiment in economic forecasts', *International Journal of Forecasting*, Vol 39, pp. 1097–1121.

Shiller, R.J. (2019). *Narrative Economics: How Stories Go Viral and Drive Major Economic Events*, Princeton: Princeton University Press.

Smales, L.A. (2023). 'Classification of RBA monetary policy announcements using ChatGPT', *Finance Research Letters*, Vol. 58, 104514.

Stundziene, A., Pilinkiene, V., Bruneckiene, J., Grybauskas, A., Lukauskas, M. and Pekarskiene, I. (2023). 'Future directions in nowcasting economic activity: a systematic literature review'. *Journal of Economic Surveys*, 25 July.

Tomaz, C., Crane, L.D., Kurz, C. Morin, N. Soto, P.E., and Vrankovich, B. (2024). 'Manufacturing sentiment: forecasting industrial production with text analysis', *Finance and Economics Discussion Series of Economics and Finance* 2024-026, Federal Reserve Board, Washington, D.C.

Appendix PMI text score as predictor of diffusion index

The relationship between the quarterly averages of the composite output PMI diffusion index and the ChatGPT-based PMI text scores is explored in two ways: Granger predictability test and out-of-sample forecasts for the composite output PMI diffusion index using as only predictor the PMI text score.

Granger causality test indicates that the PMI text score has significant predictive ability for the diffusion index and not the other way round (see Table A.1). The Granger causality test, named after the Nobel Prize winner economist Clive Granger, determines whether one time series can be used to predict another. Granger causality measures precedence and information content but does not by itself indicate causality in the common use of the term. It is thus just about predictive ability. The Granger approach to the question of whether x causes y is to see how much of the current y can be explained by past values of y and then to see whether adding lagged values of x can improve the explanation. y is said to be Granger caused by x if x helps in the prediction of y, or equivalently if the coefficients on the lagged x's are statistically significant. Note that two-way Granger causality is frequently the case; x Granger causes y and y Granger causes x. For the two PMI series, the Granger causality outcome shows that the composite output PMI diffusion index is not significantly explained by the diffusion index value of the previous quarter, but by the one-quarter lagged PMI text score.

Table A.1 Granger predictability between PMI text score and PMI composite output diffusion index

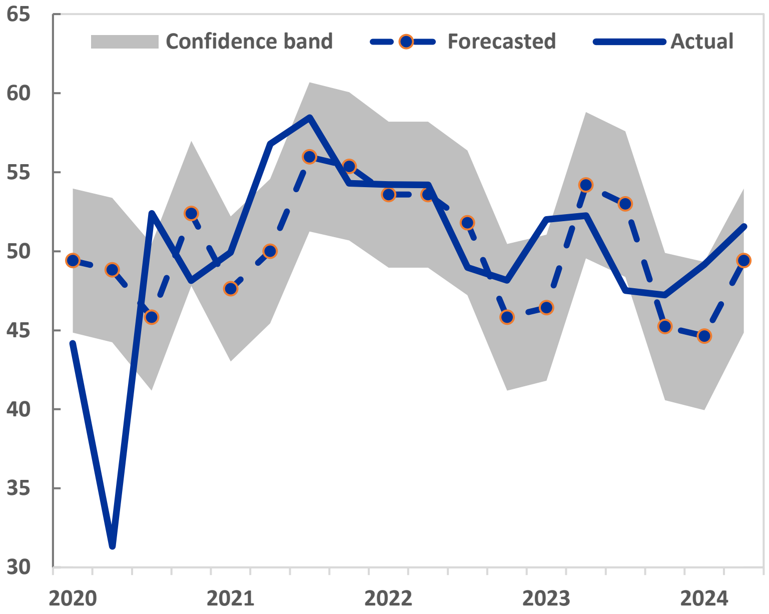
	Constant	PMI text score one lag	PMI diffusion index one lag
Coefficient	0.49	1.22 ***	0.18
	(0.51)	(0.55)	(0.22)
Adjusted R-squared			0.47
	Text score → diffusion index		Diffusion index → text score
Granger causality F-test	10.87 ***		2.21

Notes: Dependent variable is composite output PMI diffusion index from which a value of 50 is subtracted. *** <1%, ** <5%, * <10% over the total sample 2008Q1-2024Q2.

Given the predictive ability of the PMI text scores for the diffusion index, Figure A.1 plots the out-of-sample one-quarter ahead forecasts of the composite output diffusion index for 2020Q1 to 2024Q2, using the PMI text score as only predictor estimated up to 2019Q4. The in-sample period thus covers pre-Covid and the out-of-sample the Covid-years 2020-2022 as well as 2023 and 2024 as post-Covid years. PMI text scores forecast the composite output diffusion index one-quarter ahead well apart from the pandemic lockdown.

The visibly good forecast performance is confirmed by an out-of-sample forecast evaluation for the period 2020-2024. Table A.2 reports the Theil inequality coefficient, hereafter U, for the total period as well as the period excluding the three extremes of the first three quarters of 2020. Theil's U coefficient is a scaled forecast error, with the root mean squared forecast error (RMSE) relative to the variation in the actual values of the time series being forecasted. It is calculated as the RMSE divided by the sum of the square roots of the means of the actual values squared and forecasted values squared. It ranges from 0 to 1, with 0 indicating a perfect forecast and values closer to 1 indicating poor forecast accuracy. This Theil coefficient can be decomposed into three proportions of inequality. The bias proportion tells us how far the mean of the forecast is from the mean of the actual series and provides an indication of systematic error. The value close to zero means that hardly a systematic bias is present. The variance proportion tells us how far the variation of the forecast is from the variation of the actual series. The ability of the model to replicate the variability of the actual diffusion index is reasonable given the out-of-sample forecast period contains the Covid-19 pandemic. Excluding the first three quarters of the Covid-19 outbreak the variance proportion is virtually zero. The covariance proportion measures the remaining unsystematic forecasting errors. The bias, variance, and covariance proportions add up to one. If a forecast is "good", the bias and variance proportions should be small so that most of the error should be concentrated on the covariance proportions. Overall, the three proportions indicate that the PMI text score-based forecasts for the composite output PMI diffusion index are fairly accurate during "normal" times, but have shortcomings in terms of variance, as one can expect, during "extreme" times.

Figure A.1 Out-of-sample forecasted and actual composite output PMI diffusion index



Note: Forecasted values are one-quarter ahead forecasts using PMI text scores. The out-of-sample forecasts are derived from regressing composite output PMI diffusion index on the one-quarter lagged PMI text score for a pre-Covid sample up to 2019Q4. The estimated text score coefficient is 1.79 with a standard error of 0.18 and associated adjusted R-squared 0.72. The dotted lines refer to the two standard error confidence band.

Table A.2 Theil's inequality coefficient

Out-of-sample period	Theil's U	Proportion		
		Bias	Variance	Covariance
2020Q1-2024Q2	0.60	0.00	0.15	0.85
2020Q4-2024Q2	0.47	0.07	0.01	0.92

Notes: Theil's coefficient lies between 0 and 1, with 0 perfect fit.

Acknowledgements

We thank Michela Brondino for her support in collecting S&P Global PMI news releases. We thank for the comments received at an ECB seminar “AI in Economics”. We sincerely appreciate the thoughtful and constructive comments provided by an anonymous referee, which have been invaluable in enhancing the quality of the paper.

Gabe J. de Bondt
European Central Bank, Frankfurt am Main, Germany; email: gabe.de_bondt@ecb.europa.eu

Yiqiao Sun
European Central Bank, Frankfurt am Main, Germany; email: mail: yiqiao.sun@ecb.europa.eu

Postal address 60640 Frankfurt am Main, Germany
Telephone +49 69 1344 0
Website www.ecb.europa.eu

This paper can be downloaded without charge from www.ecb.europa.eu, from the Social Science Research Network electronic library or from RePEc: Research Papers in Economics. Information on all of the papers published in the ECB Working Paper Series can be found on the ECB’s website.

PDF ISBN 978-92-899- 7239-0 ISSN 1725-2806 doi:10.2866/ 2788332 QB-01-25-124-EN-N