LLMs vs. econometric models for nowcasting GDP growth: A practitioner's view

Julien André* Marie Bessec† Zachary Goulby‡

First version: June 23, 2025
This version: September 7, 2025

Abstract

This paper evaluates the forecasting performance of Large Language Models (LLMs) in nowcasting French GDP growth. We compare their forecasts, obtained through zero-shot prompting without any external input data, to those produced by the econometric models currently used at the Banque de France. Our results indicate that traditional models outperform LLMs during stable periods, while LLMs prove more effective during exceptional episodes, such as the Covid-19 pandemic. We assess the influence of prompt design, language, model version, and temperature on forecast accuracy, and introduce novel indicators of forecast confidence and probability of GDP contraction derived from LLM outputs. Extensive robustness checks, including tests for information leakage and in-sample comparisons, confirm the validity of our findings. Overall, the results suggest that while standard LLMs are not yet a substitute for econometric models in routine forecasting, they offer useful complementary insights in periods of heightened uncertainty or structural change.

Keywords: GDP nowcasting, large language models, artificial intelligence.
JEL classification: E37, C45, C53.

Introduction

Artificial intelligence chat platforms and the LLMs that power them have taken the world by storm. Platforms such as ChatGPT, Gemini, Claude and even newcomers such as DeepSeek have viral user numbers. Launched in November 2022, ChatGPT

*Banque de France, DGEI-DCPM-DIACONJ, 31 rue Croix des Petits Champs, 75049 Paris cedex 01 France, email: julien.andre@banque-france.fr.

†Corresponding author, Université Paris Dauphine, Université PSL, LEDa, CNRS, IRD, Place du Maréchal de Lattre de Tassigny 75016 Paris France, email: marie.bessec@dauphine.psl.eu.

‡Université Paris Dauphine, Université PSL, Place du Maréchal de Lattre de Tassigny 75016 Paris France, email: zachary.goulby@dauphine.eu.

The authors acknowledge free academic access to the Gemini API provided by Google. The views expressed in this article are those of the authors and do not necessarily reflect those of the Banque de France.

reached 800 million weekly active users by April 2025. Introduced four months later, Gemini (formerly known as Bard) and Claude have 67 million and 88 million users, respectively. DeepSeek, launched in January 2025, has 16.5 million users. This rapid adoption can likely be explained by the wide range of tasks for which these platforms can be employed, such as writing, summarizing, coding, and generating visuals.

These models are fundamentally changing the way we work in many areas. In particular, they have the potential to radically transform how practitioners forecast macroeconomic and financial variables. In finance, LLMs are already being integrated into empirical research for sentiment analysis and market forecasting. However, despite the extensive study of LLMs for financial forecasting, less has been done in the area of macroeconomic forecasting, particularly at very short horizons. This paper attempts to fill this gap by investigating whether LLMs can contribute to the nowcasting of Gross Domestic Product (GDP) growth. In other words, we examine whether we can do better than the traditional econometric models typically used in nowcasting by simply asking ChatGPT or any other chatbot what a country's GDP growth rate will be in the current quarter. If so, these platforms could offer a cost- and time-effective alternative to the traditional econometric methods for assessing the current state of the economy.

We examine this question from a practitioner's point of view. Specifically, we compare the LLMs with the econometric models, which are currently used by the Banque de France to forecast French GDP growth. The Banque de France publishes a nowcast of French GDP growth every month, alongside its business surveys of manufacturing, services and construction. To obtain these nowcasts, Banque de France economists rely on a mix of econometric models that are typically used in nowcasting. One such method is the MIBA (Monthly Index of Business Activity nowcasting) model, which was developed by Mogliani et al. (2017). The MIBA model is an unconstrained mixed data specification (Foroni et al. (2015)) with a preselection of variables from the Banque de France manufacturing survey for each month. Similarly, the MF3PRF factor model, introduced by André and Bessec (2024), is estimated with the mixed frequency three-pass regression filter (Kelly and Pruitt (2015) and Hepenstrick and Marcellino (2019)) on a larger database. Finally, the PRISME (Prévision Intégrée Sectorielle Mensuelle) model, developed by Thubin et al. (2016), provides an alternative forecast by aggregating the value-added forecasts of six sectors: market services, manufacturing, construction, energy, non-market services and agriculture.

The aim of this paper is to explore how large language models can compete with these econometric models. We consider the three most popular chat platforms, namely

ChatGPT, Claude and Gemini and evaluate their ability to nowcast French activity using simple prompt-based queries. We then compare these LLM-based forecasts with those of the traditional econometric models used by the Banque de France. Additionally, we introduce a confidence index that quantifies the LLM's confidence in its forecast and the probability it assigns to a GDP contraction. We test different prompting strategies, from simple to more narrative, to determine how much the way we ask a question affects the quality of forecasts, and we assess the impact of the language (French vs. English). We also investigate whether the model version matters. As in any LLM forecasting performance study, a final concern for us is the *look-ahead bias*. When we assess the ability of LLMs to forecast the GDP growth in a given past quarter, as we do with econometric models, we cannot be certain that the models really ignore information that was not available at the time of the forecast. We examine this issue carefully, with a particular focus on the period around the pandemic outbreak.

A growing body of literature is exploring the use of large language models for financial forecasting, particularly for predicting stock returns. In the context of macroeconomic forecasting, Bybee (2023) evaluates ChatGPT's ability to forecast various financial and macroeconomic variables, including US GDP growth at a horizon of one to four quarters over a long historical period. Pham and Cunningham (2024) predict US inflation and unemployment (as well as the Academy Awards) with ChatGPT, comparing direct and narrative prompts. Faria-e Castro and Leibovici (2024) use PaLM to forecast US inflation, while Woodhouse and Charlesworth (2023) try to guess the Bank of England's interest rate decisions with ChatGPT. The literature on *nowcasting* macroeconomic variables, including real GDP, is more limited. Hansen et al. (2024) use ChatGPT to replicate the individual predictions of the Survey of Professional Forecasters (SPF) for 23 US macroeconomic variables, including the real GDP index. They report promising results when external data is provided, particularly past median SPF forecasts. However, the results are less favorable in an out-of-sample evaluation in 2024. de Bondt and Sun (2025) focus on nowcasting euro area real GDP growth. They find that incorporating a ChatGPT-derived text score from flash PMI commentaries into models, in addition to the first GDP estimate or the ECB projection, improves forecasts, though this improvement is highly time-dependent.

This paper makes several contributions to this literature. First, it provides a direct comparison of LLMs and econometric models for nowcasting GDP growth. We examine this issue from a practitioner's point of view by considering the econometric models currently in use at the Banque de France. Since these models are representative of the standard tools in the nowcasting literature (see, for example, Cascaldi-Garcia

et al. (2024)), we believe that the scope of our results is broader. In addition, unlike many prior studies that rely on external data inputs or fine-tuning, we evaluate the performance of LLMs in a pure zero-shot setting, employing only prompt-based interactions. This isolates what can be achieved with generic models straight out of the box, mimicking a typical analyst's use. Our evaluation covers several leading LLMs: ChatGPT, Gemini and Claude, including different versions of each, to capture the role of model architecture and recency. We also evaluate performance across both normal and crisis periods, to assess the relative performance of the two approaches during periods of heightened uncertainty and possible structural change, such as the pandemic. Moreover, we systematically evaluate the effect of prompt design (whether simple, explanatory, or narrative) and the language used (French vs. English) on the forecast quality. There is no previous work on the effect of the language of the prompt. Another contribution is the introduction of two novel indicators, a confidence score and a probability of GDP contraction, derived from LLM responses, in order to quantify uncertainty in LLM-based forecasts. Finally, we propose several diagnostics to assess the look-ahead bias issue in the specific context of nowcasting macroeconomic data. These include textual analysis during the Covid outbreak and a novel approach that exploits data revisions.

The main results of the paper are as follows. When excluding the Covid-19 pandemic from the empirical analysis, we find that econometric models outperform standard LLMs over a large historical window for nowcasting French GDP growth. These results remain valid when focusing on more recent observations. However, LLMs dominate when the pandemic is included in the sample. This suggests that LLMs may be more effective at capturing exceptional events, a time when traditional econometric models typically underperform. Since LLMs can adapt quickly, this could be a significant advantage over econometric models in the event of a crisis or structural changes. The language used in the question matters, while the effect of prompt design on forecast accuracy is smaller. Nevertheless, playing with the way we ask the question can be useful in unlocking the model's ability to produce forecasts, as mentioned in the previous literature. Advanced versions of the models outperform simpler, quicker ones. Regarding the issue of look-ahead bias, we do not find any evidence of information leakage during the Covid episode. Furthermore, LLMs perform better when evaluated on their ability to nowcast the first release of GDP growth rather than the final release available at the time of code execution, which provides additional evidence of the absence of information leakage. However, a fair comparison of the two approaches reinforces the advantage of econometric models during normal periods.

The scope and limitations of this study are the following. First, we do not provide the LLMs with any external input data. Instead, we interact with the LLMs through direct prompting (for example, by asking ChatGPT 'What will the GDP growth be this quarter?'), without supplying additional information such as business surveys, newspaper articles or time series, as is done in some related studies. Our aim is to evaluate how useful off-the-shelf LLMs can be in practice, without requiring any technical preprocessing or domain-specific data inputs. Second, we use the LLMs with a zero-shot implementation without providing any task-specific examples or fine-tuning, i.e., additional training on domain-specific data to adapt their behavior more precisely to the nowcasting of French GDP growth. We use the off-the-shelf versions of ChatGPT, Claude and Gemini, just as a typical user (or analyst) would. Third, our focus is on general-purpose LLMs rather than specialized models trained for time series forecasting, such as Time Series Language Models or TSLMs (e.g., TimeGPT).[1] Finally, our analysis is limited to the platforms with the largest number of users, namely ChatGPT, Claude and Gemini. By considering only the three most popular platforms, we exclude open-source LLMs, which is a limitation of this study.[2]

The rest of the paper is organized as follows: The first section reviews recent applications of LLMs for forecasting macroeconomic and financial variables. The second section describes LLMs and econometric benchmarks. The third section presents the results of the comparison between econometric and large language models for forecasting the French GDP growth rate. In this section, we also present the by-products of the nowcasts: the confidence index and the probability of GDP contraction. Additionally, we examine the sensitivity of forecast accuracy to the prompting strategy. The fourth section addresses the issue of look-ahead bias. The final section concludes.

1 Related literature

An emerging body of literature is examining the potential of large language models for macroeconomic and financial forecasting. Table 1 reports the main references detailing the target variables, the models considered, the input data used and the key results.

Many papers address the topic of forecasting high-frequency *financial* variables. Lopez-Lira and Tang (2023) use ChatGPT-4 to forecast the daily returns of 4106 US companies. They derive scores based on a dataset of news headlines from major news

[1] See Carriero et al. (2024) for an evaluation of the TSLMs for macroeconomic forecasting.

[2] In particular, it might be interesting to include LLaMA in the analysis, as Meta is more transparent about the model, the model weights and the training dataset.

media and newswires, classified as good or bad by ChatGPT. These scores predict subsequent daily stock returns more accurately than other language methods or previous versions of ChatGPT. Predictability is stronger for small-cap stocks and following negative news. Chen et al. (2023) use three LLMs to generate embeddings for the news articles and alerts. Predictions from LLM embeddings (OpenAI being the best) significantly outperform leading technical signals (such as past returns) or simpler Natural Language Processing (NLP) methods (word-based models) because they understand news text in light of the broader article context. Xie et al. (2023) use ChatGPT-3.5 to predict the direction of US stock prices. They provide recent historical data and tweets posted on the same day as input. They also consider different prompting strategies, including vanilla zero-shot prompting and chain-of-thought (CoT) enhanced zero-shot prompting. In contrast to previous works, their results are less positive. LLMs underperform not only compared to state-of-the-art methods but also compared to basic methods, such as linear regression using price features. They find no significant improvement with the CoT prompting approach but the inclusion of tweets improves the quality of the forecasts. Yu et al. (2023) forecast NASDAQ-100 stock prices with ChatGPT-4 with different prompting strategies (zero-shot, few-shot prompting and CoT) and fine-tuning with Open LLaMA. They find that ChatGPT-4 outperforms traditional models. After fine-tuning, Open-LLaMA performs reasonably well, but not as well as GPT-4. Chen et al. (2024) consider ChatGPT-3.5 to forecast individual stock returns and S&P returns based on 12 weeks of lagged returns. They show that LLMs and human forecasts exhibit similar cognitive biases. LLM forecasts show excessive extrapolative behavior, tend to be overly optimistic about expected returns, and are biased downward when forecasting the tails of the return distribution. Sarkar and Vafa (2024) focus on earnings calls and firm risks with LLaMA, with a particular interest in the issue of look-ahead bias and information leakage.

Regarding the forecasting of *macroeconomic* variables at short- to medium-term horizons, Bybee (2023) compares ChatGPT-3.5's forecasts of macroeconomic and financial variables with those of professionals. The LLM forecasts are based on a sample of news articles from The Wall Street Journal. The results are similar to that of standard surveys of professional forecasters and also exhibit deviations from full-information rational expectations prevalent in existing survey series. Pham and Cunningham (2024) use ChatGPT-3.5 and 4 to forecast unemployment and inflation rates, as well as the winners of the 2022 Academy Awards. They show that providing a fictional narrative alongside the forecasts, or narrative prompts, significantly enhances ChatGPT's predictive capabilities compared to direct prompting. This method is also helpful for

unlocking ChatGPT's ability to communicate forecasts. However, the forecasts of inflation and unemployment are less spectacular than those for the Oscar winners. Faria-e Castro and Leibovici (2024) use PaLM to generate prompt-based forecasts of US inflation and compare the forecasts with the Philadelphia Fed's Survey of Professional Forecasters (SPF). They find that LLM forecasts produce lower mean squared errors overall in most years and at almost all horizons. Chen et al. (2025) compare ChatGPT, DeepSeek, BERT and RoBERTa for forecasting S&P500 returns and macroeconomic conditions. They feed the models with the front page of The Wall Street Journal and use zero-shot and few-shot prompting, as well as fine-tuning. They compile a monthly ratio of positive to negative news. With ChatGPT, a higher ratio of positive news helps predict stock returns and macroeconomic conditions, and a higher ratio of negative news helps predict certain macroeconomic indicators. However, the results are less favorable with DeepSeek, which is less intensively trained on English texts, and BERT.

More closely related to the focus of this paper on nowcasting the current state of the economy, Boss et al. (2025) target the unemployment and inflation rates in the euro area. They show that a daily score generated by LLaMa from post and commentaries on the social network Reddit improves the nowcasts, particularly those of inflation, derived from AR-MIDAS regressions. Hansen et al. (2024) nowcast 23 US macroeconomic variables, including real GDP (in level), considered in the SPF. The novelty of this study is that the authors replicate the individual forecasts of the survey by including the characteristics of the forecasters in the panel. AI-generated forecasts with ChatGPT outperform the human forecasts for one-four quarter ahead but the results are mixed for nowcasting. The results are also less successful in the out-of-sample evaluation in 2024. de Bondt and Sun (2025) focus on nowcasting eurozone real GDP growth with ChatGPT. They derive a text score from the flash PMI commentaries and include it in a regression with either the ECB projections or the Eurostat first GDP estimate. They find that the ChatGPT score improves forecasting, though this improvement is time-dependent and was not observed in 2023–24.

In summary, the literature reports encouraging results regarding the potential of LLMs for forecasting macroeconomic and financial variables. However, there are few papers on nowcasting GDP growth. Moreover, while many studies assess the role of prompt design, model version, or benchmark against professional forecasters, direct comparisons with the econometric models used by practitioners in the field remain limited. Our work is particularly related to Faria-e Castro and Leibovici (2024) and Pham and Cunningham (2024). Like these studies, we generate forecasts using simple prompt-based queries without providing additional input data, and like Pham and

Cunningham (2024), we explore the role of prompt design, comparing simple with narrative prompts.

2 LLMs versus econometric models

We will now present the two competing approaches, LLMs and traditional econometric benchmarks, in the context of GDP growth nowcasting.

2.1 LLMs nowcasts

Large Language Models (LLMs) are advanced AI systems designed to understand, interpret and generate human language. Developed using deep learning and a specialized neural network architecture known as transformers, LLMs are trained on massive text datasets (see Vaswani et al. (2017)). Most of the LLMs considered here are also trained using other types of data, including images, audio and video. LLMs work by predicting the most likely next words in a sequence.

In this paper, we consider three providers (*Google*, *OpenAI* and *Anthropic*), and for each we evaluate the most advanced model in terms of reasoning at the beginning of 2025,[3] as well as a cheaper and faster version. This leads us to consider the following six models:

- Google's Gemini models: *Gemini 1.5 Pro*, with a knowledge cutoff of August 2024, and *Gemini 2.0 Flash*, a faster alternative with a knowledge cutoff of June 2024.

- OpenAI's ChatGPT models: *GPT-4o* (knowledge cutoff: June 2024) and *GPT-4 Turbo* (updated December 2023), a faster and cheaper variant of GPT-4 trained on text data only.

- Anthropic's Claude models: *Claude 3.7 Sonnet* (knowledge cutoff: October 2024) and the faster version *Claude 3.5 Haiku* (cutoff: July 2024).

A key advantage of using Gemini is that Google provides free access to the Gemini Application Programming Interface (API) for academic research purposes (with a daily usage limit). The reported knowledge cutoff at the time of code execution in April 2025 refers to the most recent date up to which a language model has been trained on data. This implies that the model has no awareness of events or information beyond that point.[4] A truly out-of-sample exercise would therefore involve nowcasting GDP growth

[3]Note that Google released a newer, more advanced model, Gemini 2.5 pro in April 2025. On May 22, 2025, Anthropic released Claude 4, including Claude Sonnet 4 and Opus 4. OpenAI publicly released GPT-5 on August 7, 2025.

[4]Note that some researchers question the publicly reported cutoff (Cheng et al. (2024)).

for the (third and) fourth quarter(s) of 2024. However, this would result in a very short evaluation window. Nevertheless, we will check in the robustness section that including 2024 does not change the main findings.

The pros and cons of using these models for forecasting are as follows. On the positive side, LLMs are currently time- and cost-effective. They do not require data collection, model estimation, etc. and can be used at moderate cost. As will be shown later, they also adjust quickly to structural breaks and crises, unlike econometric models, at least the simple ones. This could be a major advantage during recessions and exceptional events, such as the Covid episode, or shorter-lived events, such as strikes or the last Olympic Games in France. On the negative side, there is a clear problem of transparency (training algorithm, text generator, training data), at least for the three providers considered in this paper, as discussed by Abolghasemi et al. (2025). Two other issues affect the evaluation of the models as well. There is a lack of reproducibility in the use of these models over time because they are constantly evolving, even between publicly announced changes. Barrie et al. (2024) show some changes in responses over time, despite the absence of announced changes to the underlying model. This contrasts with the nowcast obtained from econometric models. With the latter, we can compute the nowcast we had at any point in the past (with the use of a real-time database available for the typical predictors considered in our application). Moreover, pre-training of the models with recent observations makes conducting a true out-of-sample evaluation of LLMs difficult, as will be discussed later. The nowcasts generated in this paper look more like the nowcasts obtained in econometrics from an in-sample evaluation of the models.

In this paper, we follow Faria-e Castro and Leibovici (2024) or Pham and Cunningham (2024). We simply prompt the models to produce nowcasts of the French GDP growth rate given the only information available at a given time. A prompt is a short text that provides context and instructions for generating a response. Unlike other studies, we do not provide any additional input such as a sample of news articles (see Bybee (2023), Allard et al. (2024) and Chen et al. (2025)), business surveys (de Bondt and Sun (2025)), or even time series (see Chen et al. (2024) and Hansen et al. (2024)). The aim is to assess the quality of the nowcast with this simple design.

To understand how sensitive the LLM forecasts are to the way we ask questions, we test three different prompting strategies. First, we use a *simple* prompt that simply asks for the GDP growth forecast without any additional comments. Second, we consider an *explanatory* prompt that asks for a short justification in addition to the forecast.

This allows us to see if elaborating a justification improves accuracy. Third, following Pham and Cunningham (2024), we use a *narrative* prompt, asking the model to write a short story in which François Villeroy de Galhau, the Governor of the Banque de France, delivers a speech on the economic outlook for France and presents the forecast of his team. Our goal is to determine if richer, more contextual prompts generate better economic reasoning and, consequently, more accurate nowcasting. Figure 1 provides an example for each design. To test for language effects, we run all three prompts in both English and French. In each case, we ask the model three things: the nowcast itself, a confidence score, and the probability of a GDP contraction. In addition to the expected growth rate, we are interested in how confident the model is about the nowcast and how it communicates risk.

2.2 Econometric models

Regarding the econometric models, we consider the two main specifications employed by the Banque de France to nowcast French GDP growth, MIBA and MF3PRF. We present here the equations used in each month of the quarter to be forecast. The forecast equation varies from month to month in order to account for the gradual arrival of information during the quarter.[5]

The MIBA model (Mogliani et al. (2017)) is an unconstrained mixed-data sampling (U-MIDAS) specification with preselected variables (with *autometrics*) stemming from the Banque de France's manufacturing industry survey. The equations from month 1 to month 3 of quarter t are as follows:

$$
\begin{aligned}
y_t &= c + \phi y_{t-1} + \beta_1 EVLIV_{1,t} + \beta_2 PREVPRO_{1,t} + \varepsilon_t \\
y_t &= c + \phi y_{t-1} + \beta_1 EVLIV_{2,t} + \beta_2 PREVPRO_{2,t} + \beta_3 EVLIV_{1,t} + \varepsilon_t \\
y_t &= c + \phi y_{t-1} + \beta_1 EVLIV_{3,t} + \beta_2 EVLIV_{2,t} + \beta_3 EVLIV_{1,t} + \varepsilon_t
\end{aligned}
$$

where y_t is the French GDP growth in quarter t, $EVLIV_{i,t}$ is the change in deliveries, and $PREVPRO_{i,t}$ is the expected change in production in month i of quarter t. As expected, the forward-looking variable $PREVPRO$ disappears at the shortest forecast horizon. This model performed very well in the pre-Covid period.

Second, the MF3PRF model (André and Bessec (2024)) is a factor model estimated on a larger dataset (60 monthly variables). In addition to survey variables from the

[5]We do not consider the macrosectoral model PRISME in our analysis. This model has the advantage of providing a sectoral decomposition of the nowcast, but it is less performant than the other two approaches.

Banque de France, the dataset includes hard data (e.g., industrial and services production indices, construction and employment data), financial data (monetary aggregates, stock and price data), international data (economic sentiment index and Industrial Production Indices in Germany and the euro area) and an index of economic policy uncertainty. The factor model is estimated using the mixed-frequency three-pass regression filter (MF-3PRF). The original method (3PRF) was proposed by Kelly and Pruitt (2015), and Hepenstrick and Marcellino (2019) have extended the method to the case where the dataset contains indicators sampled at a higher frequency than the target variable and possibly with ragged edges. In contrast to principal component methods, the weights of the predictors in the factors are not estimated according to the correlations within the predictors. Instead, the 3PRF method weights the predictors according to their correlation with the target variable. In the mixed-frequency case, the forecast equation is a horizon-specific U-MIDAS equation, with only m contemporaneous months of the factors available at the time of the forecast:

$$y_t = \beta_0 + \sum_{i=1}^{p} \gamma_i y_{t-i} + \sum_{r=1}^{m} \beta_{r,0} \hat{F}_{r,t} + \sum_{j=1}^{q} \sum_{r=1}^{3} \beta_{r,j} \hat{F}_{r,t-j} + \eta_t$$

where $m = \{1, 2, 3\}$ is the month of the nowcast, $\hat{F}_{r,t}$ the estimated factor in month r of quarter t, with $r = \{1, 2, 3\}$. The number of lags p and q is chosen with information criteria. This model exhibits better results than MIBA in the first two months of the nowcast quarter.

Finally, we consider an autoregressive model (AR) as a usual benchmark for forecasting stationary variables:

$$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

This model, which relies solely on the past dynamics of GDP growth, performs poorly compared to models that incorporate indicators of the current outlook for the economy.

Note also that these models are designed to target the first estimates of GDP growth, which are closely followed by both the media and policy makers. For this reason, the dependent variable y_t consists of the first releases of French quarterly real GDP growth, which are now communicated by Insee 30 days after the end of the quarter in question.[6] In addition, the MIBA and MF3PRF equations contain a dummy variable

[6]Note that, prior to January 2016, the GDP growth rate was released 45 days after the end of the corresponding quarter, as opposed to 30 days under the current calendar. For the out-of-sample evaluation, we assume that the new calendar applies to the period between 2010 and 2016. This

for the second quarter of 2009, which captures the trough of the 2008-09 recession. The MF3PRF model also contains a dummy variable for the first quarter of 1996 to capture the recovery in activity following the strikes in France at the end of 1995. Finally, we follow Schorfheide and Song (2021) and Baumeister and Hamilton (2023) and exclude the Covid-19 observations in estimation of the three models.[7]

3 Forecast evaluation

3.1 Empirical design

For each approach, we produce three nowcasts per quarter, according to the calendar of the Banque de France. The nowcasts are usually published together with the business surveys shortly after the end of the month in question (around the sixth working day after the end of the month).[8] For example, nowcasts for the first quarter are published at the beginning of February, March and April. We evaluate the forecasts of the LLMs and the econometric models at these three horizons, from the first month to the last month of the target quarter.

In our experiment, we will evaluate the forecasting performance of LLMs and econometric models based on their ability to nowcast the first releases of the French GDP growth rate for the period. We will consider the period from the first quarter of 2010 to the fourth quarter of 2023 (56 quarters, or 48 quarters if we exclude the pandemic-affected period of 2020–21). Additionally, we will assess the performance of the models on a shorter subperiod from 2017 to 2023 (with or without the pandemic) to determine whether the results change when considering the most recent data. Considering the Covid-19 period allows us to assess whether the relative performance of the two approaches differs when exceptional GDP variations are included in the analysis.

For both approaches, the evaluation is out-of-sample; that is, we only use the information that was available at the time of the forecast. For the econometric models, we use a pseudo real-time evaluation with recursive regressions, replicating the ragged edges at each horizon. We estimate the three models using observations from 1995

allows us to evaluate the model's performance within the current release schedule, which is the focus of interest.

[7]During the pandemic, the Banque de France did not rely on its usual econometric models, MIBA and MF3PRF, for nowcasting French GDP growth. The MIBA model performed poorly in this context, and the MF3PRF model had not yet been developed. Instead, the Banque de France used a sectoral dashboard based on high-frequency and alternative data because information from the business survey alone proved insufficient.

[8]See Appendix A for the release calendar of the Banque de France surveys during the evaluation window.

onwards. To replicate the conditions of the nowcasting exercise, we first estimate the AR, MIBA and MF3PRF models from 1995Q1 to 2009Q4 and the first quarter of 2010 is forecast based on the information available in the first days of February, March, and then April 2010 (denoted M1, M2 and M3, respectively). Similarly, three forecasts of the GDP growth rate in the second quarter of 2010 are made using data from May 2010 to July 2010. These calculations are repeated for each subsequent quarter within the out-of-sample period. In each recursion, the lag length for the factor and the number of autoregressive terms are selected using the BIC criterion. We also fill in the missing values in the series at the end of the sample (up to the month of the forecast) in each recursion. This is done using only the information available at the time of the forecast.[9]

For LLMs, we instruct the model to generate nowcasts using information available the day before the business survey and the Banque de France nowcast are released (so that the Banque de France nowcast is not included in the information set). We collect the forecasts generated by ChatGPT, Gemini and Claude via their respective APIs. To get close to the out-of-sample design of the econometric model, we instruct the LLMs to ignore the information that was not available at the time of the forecast (see Figure 1). Later, we will show that it looks like the three platforms play the game. However, even though there is no information leakage from the training corpus, this evaluation is not a perfect out-of-sample exercise because the models were trained using recent data that was not included in the forecaster's information set. This is more akin to what we call an in-sample exercise in econometrics, where the model is estimated over a large time window and then used to forecast observations within that window. We will address this specific issue in the final section of the paper.

Given the potential variability of responses in each month (even on a fixed date), we generate 50 nowcasts for Gemini and 10 responses for ChatGPT and Claude (we do not have free credits for the last two providers). We consider the median of the 50 (or 10) nowcasts in the following.[10] In our experiments, we also test several values for the temperature parameter, which controls the randomness of the model's responses. This parameter ranges from 0 (fully deterministic) to 2 (highly random). In our baseline setup, we set the temperature to 0.3 to favor consistency and reproducibility. For robustness, we also explore two alternative values: 0.7 (the default setting for Gemini

[9]We do not use a real-time dataset, that is we ignore the revisions of the variables when estimating the factor model. While this represents a limitation of our analysis, Bernanke and Boivin (2003) and Schumacher and Breitung (2008) suggest that conclusions about forecasting performance remain largely unchanged when final data is used instead of vintage data.

[10]The results based on the average of the nowcasts are very similar. These results are available upon request.

1.5) and 1.5, which encourages more diverse, creative, and less predictable results.

For each setting, we obtain three sets of forecasts from 2010Q1 to 2023Q4, based on the information available at the end of each month during the corresponding nowcast quarter. We use two common metrics, the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE), to evaluate the quality of forecasts from LLMs and econometric models. These criteria are calculated for the entire out-of-sample window (2010-2023). We provide the results both with and without the extreme observations due to the pandemic (2020-2021), and for a recent subperiod (2017-2023) surrounding the pandemic. At this stage, we exclude the observations beyond the knowledge cutoff of the models in 2024 (but we will examine the effect of including these recent observations in the evaluation sample in the robustness part). All performance metrics are computed using the first release of quarterly GDP growth. In the final section of the paper, we also investigate how the results change when the target variable is the final GDP release. To identify the significantly best-performing models among the large number of cases considered, we also apply the Model Confidence Set (MCS) testing procedure (Hansen et al. (2011)). This multilateral testing approach provides the subset of the best models at a given level of significance from a set of competing models.

3.2 Results of the horse race

Table 2 shows the RMSE and MAE for the two forecast windows. The econometric models appear on the left and the LLMs on the right. Colored cells indicate the best-performing model at each of the three forecast horizons, denoted as M1, M2 and M3. We provide the results without or with the pandemic episode. For simplicity's sake, we will initially focus on the most advanced LLMs and the results from the *simple* prompt in English at this stage, which will serve as our reference setup.

Starting with the results excluding the Covid period in Table 2(a), the LLMs perform reasonably well, with an RMSE of around 0.20 and a MAE of around 0.17. However, the Banque de France's econometric models demonstrate better accuracy in the three months, particularly in the last two exercises. This pattern holds for both the full and the shorter forecast windows and for both evaluation criteria. Among the econometric models, MIBA and MF3PRF perform similarly in the third month, while MF3PRF is more accurate in the first two exercises. Among the LLMs, the performance differences are minimal, with all three models producing comparable results over the large window, but ChatGPT seems to perform better over a more recent window. Interestingly, the picture is opposite when the years 2020 and 2021 are included in the analysis, as shown

in Table 2(b). The performance of all approaches deteriorates dramatically due to the exceptional variation in French activity during this period (the RMSE is approximately 1 for the best LLMs and 2.9 for MIBA and MF3PRF over the large forecast window). LLMs clearly dominate the three econometric models, with the notable exception of Claude. Gemini performs better in the first month, while ChatGPT dominates in the last two exercises. The ranking remains the same for the recent subperiod.

Given the large number of LLM variants considered (three prompts $\times$ two languages $\times$ two model versions, as discussed below), we assess their relative performance against econometric models using the Model Confidence Set (MCS) procedure developed by Hansen, Lunde, and Nason (2011). This multilateral testing approach does not require the specification of a benchmark model. Rather, the MCS procedure provides, from a set of competing models, the subset that contains the best models at a given level of significance. Model selection is based on either absolute error or squared error losses. We use the T_R statistic, whose asymptotic distribution is non-standard because it depends on nuisance parameters under both the null and alternative hypotheses. Following Hansen et al. (2011), we implement a block bootstrap method with 12-observation blocks and 1,000 replications to approximate the null distribution of equal predictive accuracy. A model is included in the optimal set if its p-value is greater than the 10% significance threshold. We calculate these p-values using both mean absolute error (MAE) and mean squared error (MSE).

Figure 2 shows the proportion of cases in which each econometric model or LLM provider belongs to the MCS. The results are presented both without the pandemic period (top graphs) and over the entire evaluation window (bottom graphs). Once again, the results are favorable to the econometric models when the pandemic episode is excluded. Using the MSE loss during 2010-23, LLMs appear in the MCS in at most 25% of cases (for ChatGPT in M1 and M3), whereas econometric models such as MIBA and MF3PRF consistently belong to the optimal set over all three months. The results are slightly more favorable for LLMs when using MAE (42% of cases with ChatGPT in M3) and the most recent evaluation window (58% and 67% for ChatGPT and Claude in the first exercise and 58% for ChatGPT and Claude in the last exercise). However, when Covid observations are considered in the evaluation, the conclusions are reversed. Using the MSE loss in month 1, it is not possible to distinguish between the alternative models. However, in months 2 and 3, the econometric models never belong to the MCS set whereas LLMs (specifically ChatGPT and Gemini) do so in 58% and 75% of cases, respectively. Based on absolute error loss, the evidence in favor of LLMs is even clearer. While econometric models never belong to the MCS, Gemini and

ChatGPT reach inclusion rates of 75% and 67% in month 2, and 83% and 75% in month 3 respectively. Overall, the econometric models perform better in normal periods, but the pandemic favors LLMs, showing that they could be a better alternative in times of crisis or exceptional events.

As a by-product of the nowcast, we ask for a confidence level for the nowcast and the probability of a GDP contraction in the prompt (Figure 3). As before, we focus on the results obtained using a *simple* question in English and the three most recent models.[11] The graph on the left shows the confidence expressed by the three LLMs in the nowcast (simple prompt in English). The level of confidence in the nowcast is moderate, with an average score of around 60 across the three platforms. The decline in confidence appears to reflect some major shocks to the French economy, such as the eurozone crisis (2009-2015, with the most intense period in 2010-2012), the Yellow Vest movement (starting in November 2018), the 2020-21 pandemic and the war in Ukraine (after February 2022). The graph on the right shows the probability of negative GDP growth estimated by the three LLMs. The LLMs identify the quarters with a net GDP contraction of at least -0.1 (the dark grey bars on the three graphs), but their performance is more limited for minor contractions below 0.1 (the light grey bars). ChatGPT appears to be the most effective at identifying periods of contraction in French activity.

3.3 Impact of the prompting strategy

Next, we turn to the impact of the prompting strategy. We examine its effect on forecast accuracy and non-response rates across the platforms. For parsimony, we only report the results without the Covid period in Table 3.[12]

First, we examine the effect of using English versus French prompts. Table 3(a) reports the RMSE and MAE for the English baseline (left panel), while the right panel shows the ratio of each metric obtained with French prompts relative to the English reference case. A ratio above one (in red) indicates a deterioration compared to the reference case. Although the nowcast concerns the French economy, LLMs consistently perform better when the prompt is in English. This pattern holds for all three models. The results are particularly worse for Claude. A likely explanation is that these models are primarily trained on English text sources. This finding is consistent with the results reported in Chen et al. (2025). When comparing ChatGPT and DeepSeek for predicting

[11] The probabilities of contraction obtained using the different prompts and models show a high level of correlation, with values of 0.74 for ChatGPT, 0.76 for Claude, and 0.81 for Gemini. The correlation of the confidence scores is lower, ranging from 0.43 for Claude and ChatGPT to 0.52 for Gemini.

[12] Results for the entire period are available upon request.

US variables, they find that ChatGPT, which is trained more extensively in English, outperforms DeepSeek.

Continuing our analysis of the effect of prompt design, we then assess the effect of asking the model to justify the nowcast. Once again, we focus on the results obtained with the English prompt and with the most recent models. The results in Table 3(b) for prompts asking only for a forecast are shown on the left-hand side of the table, while those asking for a justification are shown on the right part (ratio). Contrary to our earlier findings on language choice, the inclusion of a justification has little to no effect on prediction performance. Similarly, we observe no meaningful difference between the simple and narrative prompts (left and right sides of Table 3(c)). In fact, prompting the model to "tell a story" either has no effect or worsens prediction accuracy in the last subperiod (especially for ChatGPT). This result contrasts with Pham and Cunningham (2024). They find that narrative prompts consistently outperform direct prompts when forecasting inflation and unemployment with ChatGPT.

However, like Pham and Cunningham (2024), we find that narrative prompting is helpful in unlocking the ability of models to generate forecasts, especially in the recent period. Figure 4 shows how the prompt design influences the model's reluctance to produce a forecast. Sometimes, LLMs do not produce a nowcast, but instead provide a response such as: "*As an AI, I don't have real-time data or the ability to predict future events.*" As discussed by Pham and Cunningham (2024), OpenAI, Anthropic or Google have restricted the software so that it refuses to provide certain information in the event of a possible violation of their terms of service. Figure 4 shows the frequency of non-responses regarding the GDP growth nowcast. The blue bar corresponds to the direct prompt, the red one to the explanatory prompt that requests a justification and the yellow one to the narrative case. Indeed, we find that asking the model to "tell a story" fully unlocks its forecasting capabilities, while asking for a justification sometimes reduces the rate of nonresponses. Claude (especially the advanced model) is the most reluctant to make predictions. Similar patterns are reported by Pham and Cunningham (2024) for inflation, unemployment and Academy Award predictions.[13]

[13]The results of the MCS test over the two evaluation periods (2010–2023 and 2017–2023) and across the three forecast horizons also suggest that prompts in English generally yield better performance. Of the LLMs included in the MCS, 65% were prompted in English, compared to 35% in French. In contrast, the rates of inclusion are more evenly distributed across prompt types: 32% for direct prompts, 42% for explanatory prompts, and 26% for narrative prompts.

3.4 Robustness checks

To conclude this section on the comparison of the econometric and large language models for nowcasting French GDP growth, we perform several robustness checks.

First, we assess whether our results are robust to the version of the LLMs by comparing the results we obtain with those with the less advanced versions of the three LLMs (without or with the pandemic episode). For simplicity, Table 4 only reports the results with the simple prompt in English. Again, the reference criteria with the advanced models are reported on the left and the ratios of the criteria for the fast version to the reference case appear on the right. For Gemini and ChatGPT, the most recent version outperforms the fast and cheaper one in normal times (Table 4(a)). In contrast, Claude's simpler version yields better performance based on MAE, although the most sophisticated one has a lower RMSE over the large window, suggesting fewer extreme forecast errors. Similar conclusions emerge in Table 4(b) for the entire evaluation period for ChatGPT and Gemini (for ChatGPT with the MSE criterion, the less advanced model performs better in the first two months, but the MAE still favors the more recent version). For Claude, the fast version performs better in the first two months, but the platform is far less performant than ChatGPT and Gemini in all cases.

As a second robustness check, we examine the effect of the temperature parameter, which controls the degree of randomness and creativity in the model's responses. To better capture the impact of variability on the nowcast, we report the RMSE and MAE for the averaged nowcasts rather than the median. The results for the simple prompt in English are reported for Gemini in Table 5. We find that the results are remarkably similar, even when the pandemic episode is included in the analysis. Bybee (2023) reports greater variability when forecasting inflation over a five-quarter horizon. The limited variation in our case is probably due to the shorter forecast horizon considered in this paper.

Finally, we extend our evaluation to include the forecasts for all four quarters of 2024 (released from February 2024 to January 2025), beyond the knowledge cutoff of most models. As shown in Table 6, the results remain largely unchanged: model performance does not deteriorate when we include forecast periods for which the models have no prior training data. The main conclusions remain valid. Econometric models outperform LLMs in normal times, but LLMs dominate when the pandemic is included in the analysis.

4 Look-ahead bias?

As shown in Section 3, LLMs perform reasonably well in nowcasting, particularly when exceptional events are incorporated in the analysis, while econometric models still dominate during normal periods. It remains to be seen whether our results are affected by look-ahead bias. As previously discussed, there are two important caveats to our assessment of the forecasting performance of the models, given that the language model pre-training data includes information beyond the forecast date. A first concern is that some future information may be introduced into an analysis that is intended to rely solely on past data (Sarkar and Vafa (2024)). Therefore, we need to check whether LLMs really follow the instructions in the prompt and ignore the information that was not available at the time the forecast was made. Second, even in the absence of information leakage, the architecture of the LLMs under consideration is built with future information. This situation looks like the case in econometrics where the model is estimated with future information to forecast some past information (in-sample analysis). We address these two points in this section.

4.1 Information leakage

4.1.1 Focus on the Covid period

To address the first issue, we focus on the period surrounding the outbreak of the Covid-19 pandemic in Europe, which began in late January 2020. We examine the comments generated alongside the nowcasts from late 2019 (December 2019 to January 2020) and the early months of the pandemic (February to June 2020). In France, late 2019 was marked by widespread protests against a pension reform, but concerns about the virus began to emerge in February. Consequently, President Emmanuel Macron announced the first lockdown on the evening of March 16, 2020, which took effect at noon the following day. These two events, especially the second, dramatically impacted French activity. Therefore, it is essential that LLMs respect the chronology of events when generating forecasts. Any reference to developments that had not yet occurred at the time would be a clear sign of look-ahead bias.

To assess potential information leakage, we analyze the frequency of pandemic-related terms in the comments generated alongside the nowcasts by the three most recent LLMs. The analysis is based on a corpus built from the model outputs using English prompts (with the *explanatory* prompt asking for a short explanation alongside the nowcast for the three advanced models), though similar patterns are observed with

French prompts. We aggregate the comments from the most advanced models of the three providers. If leakage were present, we would expect pandemic-related language to appear in the comments generated before February 2020, before the outbreak was recognized in Europe. Otherwise, vocabulary related to social movements should dominate in the comments accompanying the nowcasts before February.

As a first visual exploration, we present a word cloud based on the document-term matrix (with a tf-idf weighting scheme) for the corpus of comments generated with the *explanatory* prompt in English.[14] On the left, we show the results for the period from December 2019 to January 2020 (pre-Covid period) and on the right, the word cloud for the period from February 2020 to June 2020. The results clearly show that from December 2019 to January 2020, the most prominent terms are related to strikes and social unrest - such as strike, protest, yellow vest, and tension. In contrast, from February to June 2020, pandemic-related vocabulary becomes dominant, including terms such as lockdown, outbreak, Covid, coronavirus, pandemic, disrupt, and supply chain. This marked shift in vocabulary suggests that the models do not anticipate the pandemic before it occurs, and provides no evidence of information leakage.

To get a more precise view of the timeline, we plot the monthly frequency of words in the two categories, social movements (in blue) and the pandemic (in red), from December 2019 to June 2020.[15] We also overlay the frequency of articles containing pandemic terms (Covid; confinement; pandémie; coronavirus) in the French business press (yellow line), based on articles from *Les Echos* retrieved from the Factiva database. Again, there is no evidence of information leakage. From December 2019 to January 2020, terms related to social movements clearly dominate, while the use of pandemic-related vocabulary only begins to spread until after February, closely mirroring the coverage of the French media. These results are consistent with Bybee (2023), who also forecasts the GDP growth rate (and other variables), or Faria-e Castro and Leibovici (2024) with a similar example on the Covid episode for inflation.[16]

[14]The *narrative* prompt generates additional words about the governor's speech, which makes the results less readable. These results are available upon request.

[15]In our corpus, the first word list consists of climat, demonstr, gilet, jaun, pension, protest, reform, sentiment, social, strike, tension, unrest, vest, yellow and the covid related words are chain, combat, confin, coronavirus, covid, disrupt, eas, health, limit, lockdown, measur, outbreak, pandem, reopen, restrict, shutdown, spread, strict, suppli, viral, virus.

[16]We conduct a similar experiment around the beginning of the war in Ukraine in February 2022 (see Appendix B). A new wave of the Omicron variant of the SARS-CoV-2 virus emerged at the end of 2021. On 22 February 2022, Russian military forces entered Ukraine. Pandemic-related vocabulary consistently dominates the pre-war period (December 2021–January 2022). The use of war-related vocabulary begins to increase in February, peaking in March 2022. The pattern is similar in the French press.

4.1.2 What do we learn from the best target?

As another possible proof of the absence of leakage, we compare the ability of the models to nowcast the latest release of GDP growth rather than the first release. By the latest release, we mean the latest GDP growth figures published by Insee for France, available at the time the code was executed. The results are shown in Table 7 for the LLMs (see Appendix C for the econometric models). We provide the RMSE and MAE criteria for the first release of GDP growth (left part of the tables) and provide the ratios for the last release relative to the first ones in the right part. A ratio greater than one indicates a better performance in nowcasting the first release and vice versa.

A similar pattern emerges in the relative performance of econometric models and LLMs when evaluated based on their ability to forecast the latest release of GDP growth. Furthermore, both econometric models and LLMs perform better at forecasting the first release than the current one. This result was expected for econometric models, as they are typically trained to predict the first estimate of GDP growth (see Appendix C for normal period of times). However, this finding is more revealing for LLMs. If these models were based on information available today - potentially indicating a look-ahead bias - they would be expected to perform better in forecasting the final, revised figures that have become available more recently. The fact that their performance is stronger when measured with the first release provides additional evidence against the presence of such a bias. What we get is not the result of LLMs having knowledge about future realizations of French GDP growth.

Another reassuring result is the gradual improvement in the quality of the forecast of the first release as the forecast horizon shortens, in most cases (left side of Table 7). The RMSE decreases from M1 to M3 for the three models over the two evaluation periods. This indicates that the quality of the forecasts naturally improves throughout the quarter with the gradual arrival of information, as it is typically the case in practice with econometric models.

4.2 In-sample versus out-of-sample evaluation

According to the previous results, there is no information leakage. However, it remains that the LLMs are pre-trained on recent observations beyond the forecast date. That is, even though the LLMs actually ignore the recent information, the models have been built with recent information. The econometric models, on the other hand, have been estimated with the only information available at the time of the forecast, using recursive estimation in our evaluation design, as is the case in practice. We do not have access to

the 'estimation dataset' for the closed LLMs we consider, but we do for the econometric models.

To make a fairer comparison, we carry out an in-sample evaluation of the two approaches. We also estimate the three econometric models over the entire forecast period (1995-2023) and use the one-off estimates to produce the nowcasts over the two forecast periods. The results are reported in Table 8. In this design, the econometric models still outperform the LLMs during normal periods as shown in Table 8(a), but the gain with the MIBA model is even larger over the three months. This confirms that the econometric models perform better than the off-the-shelf LLMs for nowcasting GDP growth during normal periods. When the evaluation sample includes the pandemic (Table 8(b)), the LLMs still dominate, but the gain over the AR and MIBA models is smaller. In summary, the results are consistent with our previous findings, showing that LLMs perform better during exceptional periods. Conversely, econometric models dominate during normal periods, and their advantage is more evident in this fairer comparison design.

5 Conclusion

This paper assesses the potential of off-the-shelf Large Language Models for nowcasting French GDP growth.

Using a zero-shot prompting approach, without any fine-tuning or external input data, we benchmark the performance of leading LLMs against the operational econometric models currently used at the Banque de France. While traditional models consistently outperform LLMs in normal times, our results show that LLMs are better able to capture turning points and exceptional shocks, such as the Covid-19 pandemic. Beyond accuracy, we explore how prompting strategy, model version, and prompt language affect the forecasts. Queries in English yield better results than those in French, likely reflecting differences in training data coverage. Although variations in prompt design have limited impact on forecast quality, narrative prompt can address non-response issues in certain models. We also propose two novel indicators derived from LLM outputs, a confidence index and a probability of GDP contraction, which offer useful by-products for communicating about uncertainty. Finally, our analysis finds no evidence of information leakage, even during periods of high-volatility, and robustness checks confirm the reliability of our conclusions across various configurations. However, a fairer in-sample comparison further reinforces the comparative strength of econometric models under standard conditions.

These findings suggest that, although general-purpose LLMs may not yet consistently outperform traditional models for operational forecasting, they can nonetheless offer valuable complementary insights, especially when rapid adaptation to new shocks is necessary. Future work could examine whether forecast performance improves when LLMs are combined with structured data sources, or when used in hybrid approaches alongside econometric methods. It would also be important to test the LLMs' effectiveness in other geographical areas, particularly in economies where relevant data is primarily available in English. Finally, the growing retrieval-augmented capabilities of LLMs offer promising avenues for building more responsive forecasting systems.

Acknowledgements

The authors acknowledge free academic access to the Gemini API provided by Google.

References

Abolghasemi, M., Ganbold, O., and Rotaru, K. (2025). Humans vs. large language models: Judgmental forecasting in an era of advanced AI. *International Journal of Forecasting*, 41(2):631–648.

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

André, J. and Bessec, M. (2024). A mixed-frequency factor model for nowcasting french gdp. Document de travail de la Banque de France, DT975, Banque de France.

Barrie, C., Palmer, A., and Spirling, A. (2024). Replication for language models problems, principles, and best practice for political science. *URL: https://arthurspirling. org/documents/BarriePalmerSpirling TrustMeBro. pdf*.

Baumeister, C. and Hamilton, J. (2023). Uncovering disaggregated oil market dynamics: A full-information approach to granular instrumental variables. Technical report.

Bernanke, B. and Boivin, J. (2003). Monetary policy in a data-rich environment. *Journal of Monetary Economics*, 50(3):525–546.

Boss, K., Longo, L., and Onorante, L. (2025). Nowcasting the euro area with social media data. *arXiv preprint arXiv:2506.10546*.

Bybee, L. (2023). Surveying generative AI's economic expectations. *arXiv preprint arXiv:2305.02823*.

Carriero, A., Pettenuzzo, D., and Shekhar, S. (2024). Macroeconomic forecasting with large language models. *arXiv preprint arXiv:2407.00890*.

Cascaldi-Garcia, D., Luciani, M., and Modugno, M. (2024). Lessons from nowcasting GDP across the world. In *Handbook of Research Methods and Applications in Macroeconomic Forecasting*, pages 187–217. Edward Elgar Publishing.

Chen, J., Tang, G., Zhou, G., and Zhu, W. (2025). ChatGPT and DeepSeek: Can they predict the stock market and macroeconomy? *arXiv preprint arXiv:2502.10008*.

Chen, S., Green, T., Gulen, H., and Zhou, D. (2024). What does ChatGPT make of historical stock returns? extrapolation and miscalibration in LLM stock return forecasts. *arXiv preprint arXiv:2409.11540*.

Chen, Y., Kelly, B., and Xiu, D. (2023). Expected returns and large language models. *Available at SSRN 4416687*.

Cheng, J., Marone, M., Weller, O., Lawrie, D., Khashabi, D., and Van Durme, B. (2024). Dated data: Tracing knowledge cutoffs in large language models. *arXiv preprint arXiv:2403.12958*.

de Bondt, G. and Sun, Y. (2025). Enhancing GDP nowcasts with ChatGPT: a novel application of PMI news releases. Technical report, European Central Bank.

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

Foroni, C., Marcellino, M., and Schumacher, C. (2015). Unrestricted mixed data sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 178(1):57–82.

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

Hansen, P., Lunde, A., and Nason, J. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Hepenstrick, C. and Marcellino, M. (2019). Forecasting gross domestic product growth with large unbalanced data sets: the mixed frequency three-pass regression filter. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(1):69–99.

Kelly, B. and Pruitt, S. (2015). The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, 186(2):294–316.

Lopez-Lira, A. and Tang, Y. (2023). Can ChatGPT forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*.

Mogliani, M., Darné, O., and Pluyaud, B. (2017). The new MIBA model: Real-time nowcasting of French GDP using the Banque de France's monthly business survey. *Economic Modeling*, 64:26–39.

Pham, V. and Cunningham, S. (2024). Can base ChatGPT be used for forecasting without additional optimization? *arXiv preprint arXiv:2404.07396*.

Sarkar, S. and Vafa, K. (2024). Lookahead bias in pretrained language models. *Available at SSRN*.

Schorfheide, F. and Song, D. (2021). Real-time forecasting with a (standard) mixed-frequency VAR during a pandemic. Technical report, National Bureau of Economic Research.

Schumacher, C. and Breitung, J. (2008). Real-time forecasting of German GDP based on a large factor model with monthly and quarterly data. *International Journal of Forecasting*, 24(3):386–398.

Thubin, C., Monnet, E., Marx, M., Oung, V., and Ferriere, T. (2016). The PRISME model: can disaggregation on the production side help to forecast GDP? Technical report, Banque de France.

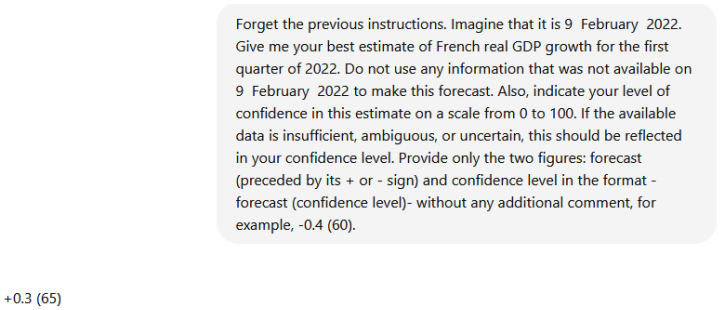
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Woodhouse, D. and Charlesworth, A. (2023). Can ChatGPT predict future interest rate decisions? *Available at SSRN 4572831*.

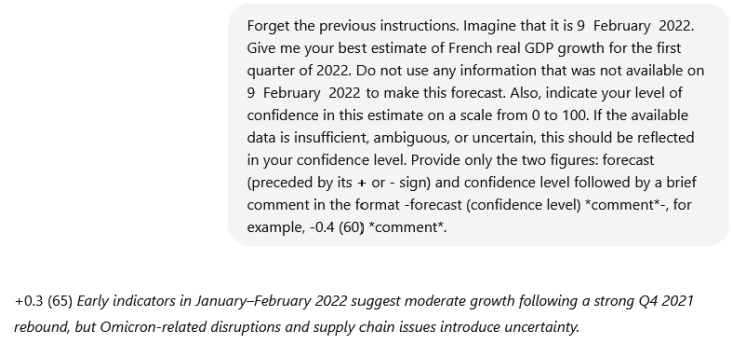
Xie, Q., Han, W., Lai, Y., Peng, M., and Huang, J. (2023). The Wall Street neophyte: A zero-shot analysis of ChatGPT over multimodal stock movement prediction challenges. *arXiv preprint arXiv:2304.05351*.

Yu, X., Chen, Z., Ling, Y., Dong, S., Liu, Z., and Lu, Y. (2023). Temporal data meets LLM-explainable financial time series forecasting. *arXiv preprint arXiv:2306.11025*.

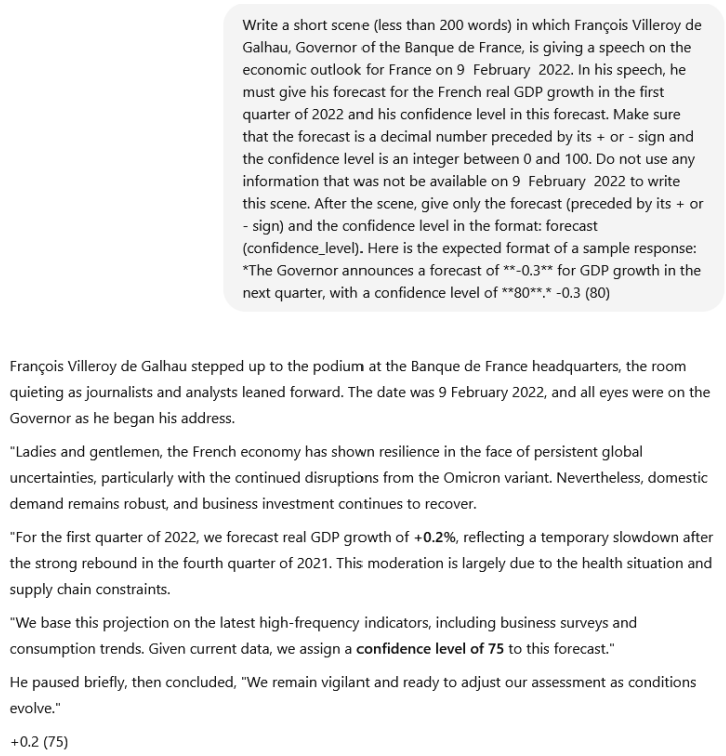
Figure 1: The three prompting strategies



(a) Simple prompt



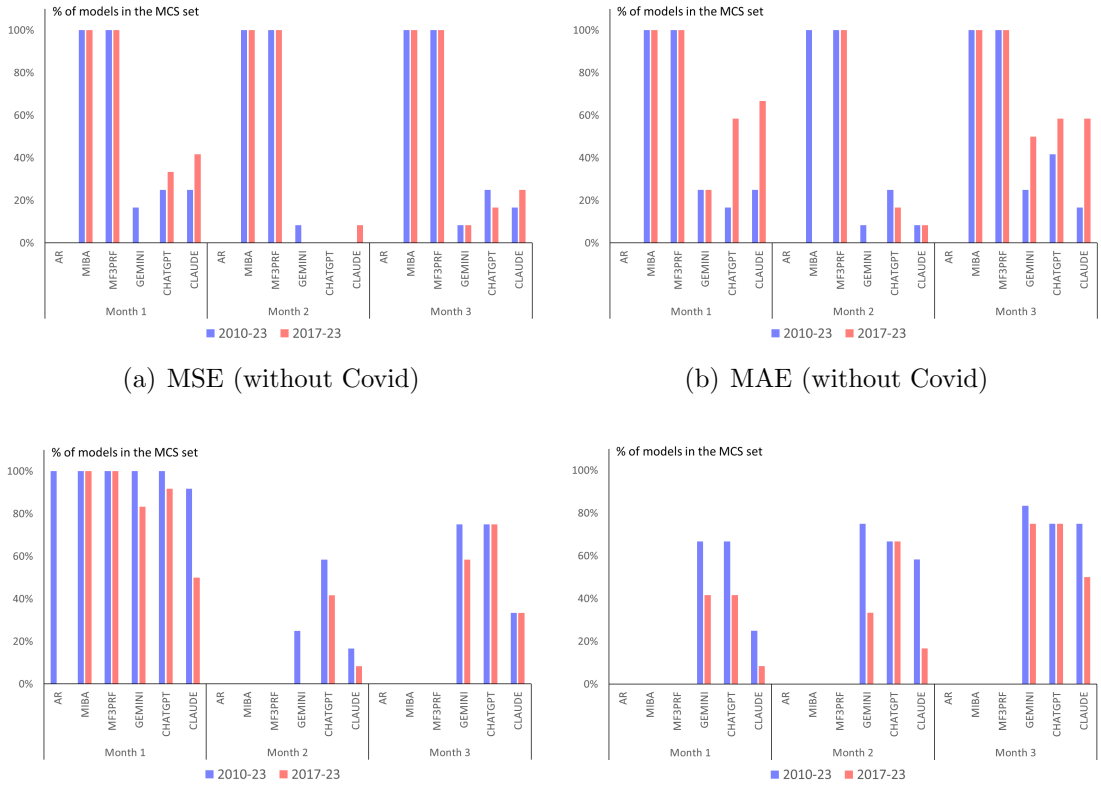
(b) Explanatory prompt



(c) Narrative prompt

Notes: This figure shows an example of the three prompts: the *simple* one, which asks for a nowcast; the *explanatory* one, which asks for a nowcast and a short comment; and the *narrative* one, which asks the platform to tell a story.

Figure 2: Results of the MCS testing procedure



(a) MSE (without Covid)

(b) MAE (without Covid)

(c) MSE (with Covid)

(d) MAE (with Covid)

Notes: This figure shows the percentage of econometric and large language models that belong to the Model Confidence Set. For each provider, we consider three possible prompts in either French or English, as well as more or less advanced models (12 cases in total). Model selection is based on either squared error losses (left graphs) or absolute errors (right graphs). The results are presented for the two evaluation windows (2010–2023 in blue and 2017–2023 in red) and the three forecast horizons (months 1, 2 and 3). The top (bottom) panel shows the results without (with) the pandemic period.

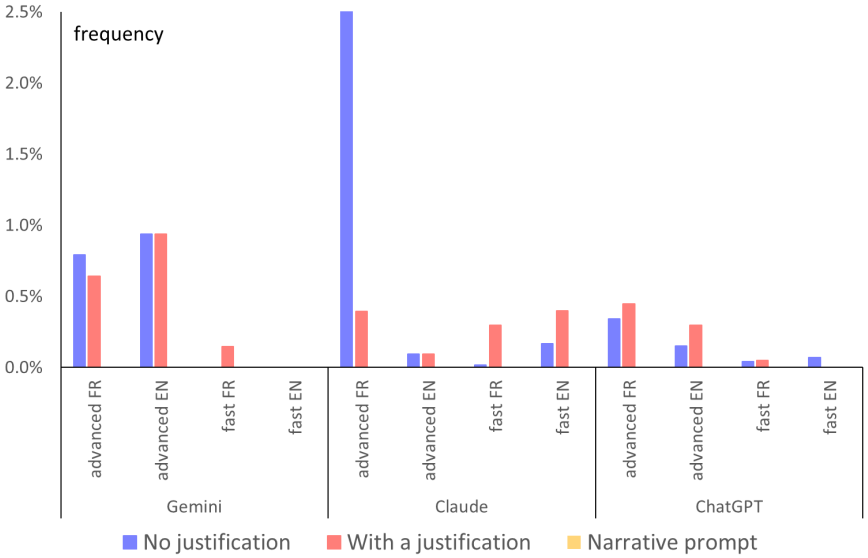
Figure 3: Confidence Index and Probability of GDP Contraction

(a) Confidence score

(b) Probability of contraction

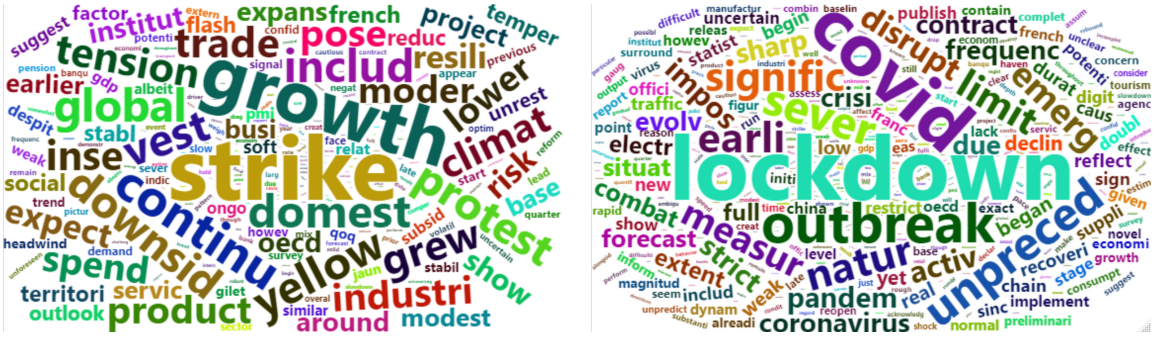
Notes: This figure shows the confidence level of the three LLMs in their nowcast and the probability of GDP contraction. The results are reported for our baseline setup, which uses a simple prompt in English and advanced models. The shaded areas in the graph on the right indicate periods of actual GDP contraction. Light grey corresponds to mild contractions (between 0 and –0.1), while dark grey indicates more severe declines (below –0.1), according to the initial GDP releases by Insee.

Figure 4: Frequency of non-responses



Notes: This figure shows the frequency of non-responses regarding the nowcast from 2010 to 2023. Results for the three LLMs (with either the advanced or the fast versions) are shown for the simple, explanatory and narrative prompts (in blue, red and yellow, respectively) in either French (FR) or English (EN).

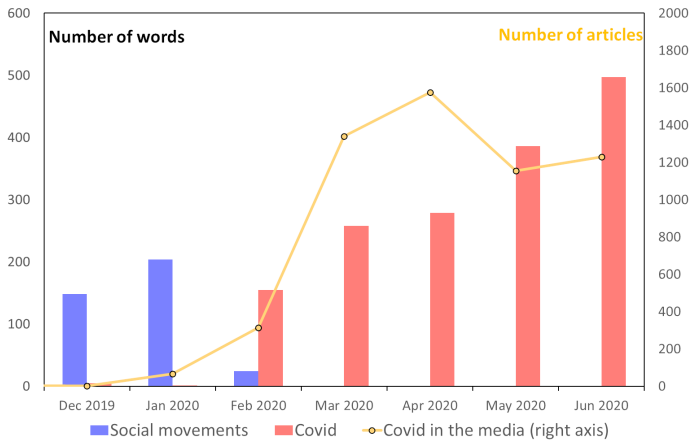
Figure 5: Information leakage? A focus on the Covid period



(a) Pre-Covid period (b) Covid period

Notes: This figure shows the word clouds for the Document Term Matrices (with a tf-idf weighting) of the corpus of comments in English. The two alternative corpora include the comments that are generated by the three advanced LLMs (Gemini, ChatGPT and Claude) alongside the nowcasts (with the explanatory prompt) for the pre-Covid period (December 2019 to January 2020) and the beginning of the Covid episode (February to June 2020).

Figure 6: Frequency of words related to social movements and the pandemic



Notes: This figure shows the frequency of words belonging to two categories: social movements (shown in blue) and the pandemic (in red). The yellow curve depicts the frequency of articles in the French business newspaper *Les Echos* that contain pandemic-related vocabulary (Covid, confinement, pandémie, coronavirus).

Table 1: Literature – Forecasting with LLMs

(a) Financial variables

Authors	Forecast variable	LLMs and input	Benchmarks	Main results
Lopez-Lira & Tang (2023)	Daily stock returns of 4106 US companies from October 2021 to December 2023	ChatGPT-4; Input: a data set of news headline from major news media and newswires classified as good or bad with ChatGPT	Previous versions of Chat (ChatGPT-1, ChatGPT-2, ChatGPT-3.5) and other NLP models: BERT, BART, DistilBart-MNLI and FinBERT	ChatGPT-4 scores significantly predict daily stock returns, subsuming traditional methods; predictability is stronger among smaller stocks and following negative news.
Chen, Kelly & Xiu (2023)	US and international daily stock returns from January 1996 to June 2019	Text-embedding-3-large from OpenAI; LLaMA, BERT and RoBERTa; Input: global news text data from Refinitiv in their Thomson Reuters Real-time News Feed and Third Party Archive dataset; Word embeddings approach: use of LLM to generate embeddings for the news articles and alerts	Word-based models: Word2vec; Bag-of Words approach (with Loughran & McDonald dictionary)	Predictions from LLM embeddings (OpenAI being the best) significantly improve over leading technical signals (such as past returns) or simpler NLP methods (words based models) by understanding news text in light of the broader article context.
Xie, Han, Lai, Peng & Huang (2023)	Increase/decrease of US daily stock prices for the next day	ChatGPT-3.5; Inputs: historical prices and tweets from the same day; Various prompting strategies, including vanilla zero-shot prompting and Chain-of-Thought (CoT) enhanced zero-shot prompting	Historical-data-based methods (logistic regression, random forests, LSTM, ALSTM, Adv-ALSTM, DTLM); Tweet-based methods	ChatGPT's performance in multimodal stock price prediction tasks is generally limited, as it underperforms not only state-of-the-art methods but also basic methods like linear regression using price features. No substantial improvement with CoT. Including tweets improves the forecasting performance.
Yu, Chen, Ling, Dong, Liu & Lu (2023)	NASDAQ-100 weekly/monthly change	ChatGPT-4 with zero-shot or few-shot prompting; CoT technique; fine-tuning with Open LLaMA; input: the company profile and top-news stories on Google on each stock and about macroeconomy and finance	ARMA-GARCH model and a gradient boosting tree model	GPT-4 few shot with CoT outperforms the traditional models. The publicly available Open-LLaMA, after fine-tuning, can comprehend the instruction to generate explainable forecasts and achieve reasonable performance, albeit relatively inferior in comparison to GPT-4.
Chen, Green, Gulen & Zhou (2024)	Individual stocks and S&P500 returns; the observations are randomly selected from 1926 to 2023	ChatGPT-4 and Claude 3.5 Sonnet Input: 12 weeks of lagged returns provided in .csv files	Forcerank for the individual stocks ; American Association of Individual investor Survey for S&P	Similar cognitive biases in LLM and human forecasts: overoptimism, excessive extrapolation, downward bias in tail forecasts.

(b) Macroeconomic variables

Authors	Forecast variable	LLMs and input	Benchmarks	Main results
Bybee (2023)	Future 12-month return and quarterly macroeconomic variables (horizon h=1 to 4) including GDP growth from 1984 to 2021	ChatGPT-3.5; Input: sample of news articles from The Wall Street Journal classified as positive, negative or neutral for the variable	Survey of Professional Forecasters, the American Association of Individual Investors, and the Duke CFO Survey	LLM expectations are similar to those of professional forecasters and also exhibit deviations from full-information rational expectations prevalent in existing survey series.
Woodhouse & Charlesworth (2023)	Interest rate decision of the Bank of England	ChatGPT-3.5; Input: speeches of England Monetary Policy Committee members classified as hawkish/neutral/dovish	No benchmark	ChatGPT can predict future interest rate decisions.
Pham & Cunningham (2024)	Monthly inflation and unemployment from Oct 2021 (h=1) to Sept 2022 (h=12) (and the winners of the 2022 Academy Awards)	ChatGPT-3.5 and ChatGPT-4; No input	Comparison of direct and narrative prompting	Narrative prompting improves accuracy, especially for Oscars; LLMs are less successful for macroeconomic variables.
Faria-e-Castro & Leibovici (2024)	US Inflation from 2019 to 2023	PaLM's Bison 001 model; No input	Professional forecasters: the Survey of Professional Forecasters	PaLM performs as good if not better than SPF. Slower reversion to the 2% anchor.
Chen, Tang, Zhou & Zhu (2025)	Monthly excess returns on S&P500 and US macroeconomic variables including GDP growth from January 1996 to December 2022 (h={0,1,3,6,12} for the returns and h=1 for macro)	ChatGPT-3.5, ChatGPT-4, DeepSeek-R1, BERT & RoBERTa; Input: the front page of the Wall Street Journal classified as good or bad news to compile monthly scores; Zero-shot vs few-shot prompting method	Survey of Professional Forecasters	ChatGPT, trained more extensively in English outperforms DeepSeek and BERT. The ratio of good news is positively correlated with contemporaneous market returns and significantly predicts subsequent returns for the next six months. The ratio of positive news (and negative in some cases) also predicts future macroeconomic conditions.
Hansen, Horton, Kazinnik, Puzzello & Zarifhonarvar (2024)	23 US macroeconomic variables (including GDP index) in the SPF, h=0,1,4 from 1999 to 2023 and out of sample exercise in 2024	GPT-3.5, 4, 4o mini; Inputs: individual forecaster characteristics, macroeconomic data and past SPF median forecasts to generate SPF individual forecasts	SPF individual forecasts; SPF AI forecasts with a more restricted set of inputs	AI-generated forecasts outperform human forecasts for h=1 and 4; mixed results for nowcasting and in the out of sample exercise. Providing the inputs contribute to the forecasting accuracy.
de Bondt & Sun (2025)	euro area real GDP growth from 2017 to 2024 for h=0	ChatGPT-4o; input: flash PMI text commentaries to derive a quantitative score included in a regression containing either the Eurostat first GDP estimate or the ECB/Eurosystem staff projections	Models without the ChatGPT score	Forecast gain with the ChatGPT score but the gain is time-dependent; no gain in 2023-24.
Boss, Longo & Onorante (2025)	monthly inflation and unemployment in the euro area from 2018 to 2023 for h=0	LLaMa-3-70b-instruct; input: reddit posts and commentaries on inflation and unemployment classified as up, down or neutral to derive a daily score included in AR-MIDAS regressions	AR(1); AR-MIDAS regressions with alternative sentiment indicators, financial variables or oil price	The reddit indicator generally outperforms the other variables; gains especially visible during the pandemic period and the subsequent high inflation episode.

Table 2: Comparison of the econometric and LLMs nowcasts

(a) Without the Covid period

		AR	MIBA	MF3PRF	GEMINI	CHAT	CLAUDE
RMSE							
2010-23	M1	0.308	0.192	**0.177**	0.212	0.220	0.215
	M2	0.308	0.187	**0.172**	0.218	0.210	0.212
	M3	0.308	**0.170**	**0.170**	0.200	0.189	0.205
2017-23	M1	0.334	**0.157**	0.165	0.200	0.174	0.177
	M2	0.334	0.158	**0.154**	0.199	0.172	0.183
	M3	0.334	0.158	**0.148**	0.183	0.171	0.183
MAE							
		AR	MIBA	MF3PRF	GEMINI	CHAT	CLAUDE
2010-23	M1	0.234	0.148	**0.144**	0.169	0.175	0.176
	M2	0.234	0.156	**0.132**	0.176	0.166	0.171
	M3	0.234	0.138	**0.137**	0.146	0.149	0.167
2017-23	M1	0.222	**0.117**	0.121	0.162	0.131	0.134
	M2	0.222	0.132	**0.101**	0.159	0.130	0.136
	M3	0.222	0.125	**0.116**	0.131	0.121	0.136

Notes: This table reports the RMSE and MAE criteria for the nowcasts in the first (M1), second (M2) and third (M3) months of the quarter. The results are reported for two out-of-sample windows: 2010Q1-2023Q4 and 2017Q1-2023Q4 (excluding 2020Q1 to 2021Q4). The left side of the table shows the RMSE and MAE for the econometric models, AR, MIBA, MF3PRF and the right side the results for the LLMs in the baseline setup (simple question in English and advanced models). The colored cells show the best-performing model in each month.

(b) With the Covid period

		AR	MIBA	MF3PRF	GEMINI	CHAT	CLAUDE
RMSE							
2010-23	M1	3.412	2.992	2.648	**1.022**	1.228	4.617
	M2	3.412	3.007	2.757	1.293	**1.226**	4.610
	M3	3.412	2.892	3.021	0.925	**0.523**	3.668
2017-23	M1	4.817	4.226	3.741	**1.428**	1.718	6.525
	M2	4.817	4.248	3.895	1.814	**1.719**	6.515
	M3	4.817	4.086	4.269	1.291	**0.711**	5.182
MAE							
		AR	MIBA	MF3PRF	GEMINI	CHAT	CLAUDE
2010-23	M1	1.031	0.949	0.882	**0.400**	0.473	1.041
	M2	1.031	0.948	0.898	0.504	**0.468**	1.043
	M3	1.031	0.898	0.932	0.356	**0.258**	0.750
2017-23	M1	1.819	1.728	1.605	**0.627**	0.740	1.876
	M2	1.819	1.722	1.641	0.821	**0.744**	1.890
	M3	1.819	1.650	1.710	0.555	**0.348**	1.313

Notes: See Table a. The evaluation period includes the Covid episode.

Table 3: Impact of the prompting design

(a) Prompting design - Effect of the language

		Simple prompt in English			Simple prompt in French		
	RMSE	GEMINI	CHAT	CLAUDE	GEMINI	CHAT	CLAUDE
2010-23	M1	0.212	0.220	0.215	1.10	0.98	6.00
	M2	0.218	0.210	0.212	0.97	0.95	4.74
	M3	0.200	0.189	0.205	1.02	1.07	3.58
2017-23	M1	0.200	0.174	0.177	1.16	0.95	9.49
	M2	0.199	0.172	0.183	0.98	1.05	5.31
	M3	0.183	0.171	0.183	1.01	1.04	4.55
	MAE	GEMINI	CHAT	CLAUDE	GEMINI	CHAT	CLAUDE
2010-23	M1	0.169	0.175	0.176	1.05	0.99	3.59
	M2	0.176	0.166	0.171	0.96	0.96	3.13
	M3	0.146	0.149	0.167	1.05	1.07	2.20
2017-23	M1	0.162	0.131	0.134	1.07	0.96	6.57
	M2	0.159	0.130	0.136	0.95	1.02	3.51
	M3	0.131	0.121	0.136	1.04	1.02	3.26

Notes: This table compares the accuracy of the nowcast for the simple prompt in English and in French for the three LLMs (advanced models). The left panel shows the RMSE and MAE criteria for the nowcasts in the first (M1), second (M2) and third (M3) months of the quarter in the reference case (simple prompt in English). The right panel shows the ratio of the criteria in the alternative case (simple prompt *in French*) to those of the reference case on the left. A ratio greater than one indicates a deterioration with the French prompt.

(b) Prompting design - Simple versus explanatory prompt

		Simple prompt in English			Explanatory prompt		
	RMSE	GEMINI	CHAT	CLAUDE	GEMINI	CHAT	CLAUDE
2010-23	M1	0.212	0.220	0.215	1.00	0.99	1.03
	M2	0.218	0.210	0.212	1.04	0.94	1.02
	M3	0.200	0.189	0.205	1.02	1.03	0.99
2017-23	M1	0.200	0.174	0.177	0.99	1.02	0.99
	M2	0.199	0.172	0.183	1.02	0.99	1.00
	M3	0.183	0.171	0.183	1.03	1.03	0.97
	MAE	GEMINI	CHAT	CLAUDE	GEMINI	CHAT	CLAUDE
2010-23	M1	0.169	0.175	0.176	0.99	1.01	1.01
	M2	0.176	0.166	0.171	1.06	0.95	1.03
	M3	0.146	0.149	0.167	1.08	1.04	0.98
2017-23	M1	0.162	0.131	0.134	0.97	1.04	0.96
	M2	0.159	0.130	0.136	1.03	0.98	0.98
	M3	0.131	0.121	0.136	1.08	1.02	0.94

Notes: This table compares the accuracy of the nowcast for simple and explanatory prompts for the three LLMs (advanced models). The left panel shows the RMSE and MAE criteria for the nowcasts in the first (M1), second (M2) and third (M3) months of the quarter in the reference case (simple prompt in English). The right panel shows the ratio of the criteria in the alternative case (*explanatory* prompt) to those of the reference case on the left. A ratio greater than one indicates a deterioration with the explanatory prompt.

(c) Prompting design - Simple versus narrative prompt

		Simple prompt in English			Narrative prompt		
	RMSE	GEMINI	CHAT	CLAUDE	GEMINI	CHAT	CLAUDE
2010-23	M1	0.212	0.220	0.215	1.05	1.01	1.04
	M2	0.218	0.210	0.212	1.05	1.01	0.98
	M3	0.200	0.189	0.205	1.00	1.07	0.96
2017-23	M1	0.200	0.174	0.177	0.99	1.21	1.15
	M2	0.199	0.172	0.183	1.05	1.14	1.10
	M3	0.183	0.171	0.183	0.91	1.15	0.95
	MAE	GEMINI	CHAT	CLAUDE	GEMINI	CHAT	CLAUDE
2010-23	M1	0.169	0.175	0.176	1.04	1.02	1.03
	M2	0.176	0.166	0.171	1.01	1.02	0.97
	M3	0.146	0.149	0.167	1.05	1.03	0.91
2017-23	M1	0.162	0.131	0.134	0.98	1.21	1.15
	M2	0.159	0.130	0.136	0.94	1.14	1.09
	M3	0.131	0.121	0.136	0.95	1.18	0.92

Notes: This table compares the accuracy of the nowcast for simple and explanatory prompts for the three LLMs (advanced models). The left panel shows the RMSE and MAE criteria for the nowcasts in the first (M1), second (M2) and third (M3) months of the quarter in the reference case (simple prompt in English). The right panel shows the ratio of the criteria in the alternative case (*narrative* prompt) to those of the reference case on the left. A ratio greater than one indicates a deterioration with the narrative prompt.

Table 4: Version of the LLMs - Advanced versus fast and cheaper

(a) Without the Covid period

		Advanced model			Fast and cheaper model		
	RMSE	GEMINI	CHAT	CLAUDE	GEMINI	CHAT	CLAUDE
2010-23	M1	0.212	0.220	0.215	1.14	2.70	1.01
	M2	0.218	0.210	0.212	1.07	2.80	1.05
	M3	0.200	0.189	0.205	1.18	2.98	1.07
2017-23	M1	0.200	0.174	0.177	0.93	3.78	0.88
	M2	0.199	0.172	0.183	0.90	3.79	0.81
	M3	0.183	0.171	0.183	1.19	3.58	0.81
	MAE	GEMINI	CHAT	CLAUDE	GEMINI	CHAT	CLAUDE
2010-23	M1	0.169	0.175	0.176	1.13	2.98	0.97
	M2	0.176	0.166	0.171	1.05	3.10	0.96
	M3	0.146	0.149	0.167	1.23	3.24	0.98
2017-23	M1	0.162	0.131	0.134	0.90	4.64	0.90
	M2	0.159	0.130	0.136	0.86	4.60	0.81
	M3	0.131	0.121	0.136	1.19	4.53	0.79

Notes: This table compares the accuracy of the nowcasts for the advanced versus fast and cheaper versions of the LLMs. In both cases, we provide the results obtained with the simple prompt in English and for two out-of-sample windows: 2010Q1 to 2023Q4 and 2017Q1 to 2023Q4 (excluding 2020Q1 to 2021Q4). The left panel shows the RMSE and MAE criteria of the nowcasts in the first (M1), second (M2) and third (M3) months of the quarter in the reference case (advanced version). The right panel shows the ratio of the criteria in the alternative case (the fast and cheaper version) to those of the reference case on the left. A ratio above one indicates a deterioration with the less advanced LLMs.

(b) With the Covid period

		Advanced model			Fast and cheaper model		
	RMSE	GEMINI	CHAT	CLAUDE	GEMINI	CHAT	CLAUDE
2010-23	M1	1.022	1.228	4.617	4.45	0.85	0.86
	M2	1.293	1.226	4.610	3.58	0.87	0.86
	M3	0.925	0.523	3.668	5.03	1.63	1.22
2017-23	M1	1.428	1.718	6.525	4.50	0.80	0.86
	M2	1.814	1.719	6.515	3.60	0.82	0.86
	M3	1.291	0.711	5.182	5.09	1.53	1.22
	MAE	GEMINI	CHAT	CLAUDE	GEMINI	CHAT	CLAUDE
2010-23	M1	0.400	0.473	1.041	2.62	1.41	0.84
	M2	0.504	0.468	1.043	2.21	1.47	0.84
	M3	0.356	0.258	0.750	2.78	2.28	1.17
2017-23	M1	0.627	0.740	1.876	2.98	1.18	0.82
	M2	0.821	0.744	1.890	2.45	1.24	0.82
	M3	0.555	0.348	1.313	3.22	2.13	1.18

Notes: See Table a. The evaluation period includes the Covid episode.

Table 5: Impact of the temperature parameter

		Without the Covid period			With the Covid period		
	RMSE	temp=0.3	temp=0.7	temp=1.5	temp=0.3	temp=0.7	temp=1.5
2010-23	M1	**0.211**	0.212	0.213	**1.034**	1.055	1.052
	M2	0.218	0.218	**0.217**	1.068	1.070	**1.039**
	M3	0.200	**0.199**	0.200	**0.942**	0.944	0.947
2017-23	M1	**0.198**	0.199	**0.198**	**1.445**	1.476	1.471
	M2	**0.197**	**0.197**	**0.197**	1.492	1.495	**1.451**
	M3	0.185	**0.183**	0.184	**1.315**	1.319	1.323
	MAE	temp=0.3	temp=0.7	temp=1.5	temp=0.3	temp=0.7	temp=1.5
2010-23	M1	**0.168**	**0.168**	**0.168**	**0.408**	0.419	0.417
	M2	**0.175**	**0.175**	**0.175**	0.446	0.446	**0.432**
	M3	0.152	**0.151**	0.152	**0.364**	0.365	0.367
2017-23	M1	0.160	0.161	**0.159**	**0.643**	0.665	0.659
	M2	**0.156**	**0.156**	**0.156**	0.703	0.705	**0.677**
	M3	0.137	**0.135**	0.136	**0.566**	0.567	0.569

Notes: This table compares the accuracy of the *averaged* nowcasts of Gemini for different values of the temperature parameter. The RMSE and MAE criteria are given for the simple prompt in English and the most advanced model. Results are provided for the 2010–2023 evaluation window, without or with the pandemic (left and right parts, respectively).

Table 6: Beyond the training data limit

(a) Without the Covid period

		RMSE					
		AR	MIBA	MF3PRF	GEMINI	CHAT	CLAUDE
2010-24	M1	0.301	0.204	**0.182**	0.211	0.218	0.211
	M2	0.301	0.189	**0.175**	0.217	0.209	0.208
	M3	0.301	0.175	**0.172**	0.200	0.190	0.205
2017-24	M1	0.316	0.193	0.178	0.201	0.178	**0.172**
	M2	0.316	0.167	**0.165**	0.199	0.176	0.180
	M3	0.316	0.173	**0.159**	0.185	0.176	0.188
		MAE					
		AR	MIBA	MF3PRF	GEMINI	CHAT	CLAUDE
2010-24	M1	0.226	0.156	**0.148**	0.167	0.172	0.170
	M2	0.226	0.158	**0.134**	0.172	0.164	0.167
	M3	0.226	0.144	**0.139**	0.145	0.147	0.166
2017-24	M1	0.206	0.140	0.133	0.159	0.131	**0.128**
	M2	0.206	0.139	**0.111**	0.154	0.131	0.134
	M3	0.206	0.140	**0.123**	0.131	**0.123**	0.141

Notes: This table reports the RMSE and MAE criteria when the evaluation window is extended beyond the limit of the training data. The results are reported for two out-of-sample windows: 2010Q1-2024Q4 and 2017Q1-2024Q4 (excluding 2020Q1 to 2021Q4). The left side of the table shows the RMSE and MAE for the econometric models, AR, MIBA, MF3PRF and the right side presents the results for the LLMs in the baseline setup (simple question in English, advanced model). The colored cells show the best-performing model in each month.

(b) With the Covid period

		RMSE					
		AR	MIBA	MF3PRF	GEMINI	CHAT	CLAUDE
2010-24	M1	3.297	2.892	2.559	**0.988**	1.187	4.461
	M2	3.297	2.892	2.559	1.250	**1.186**	4.454
	M3	3.297	2.892	2.559	0.895	**0.507**	3.544
2017-24	M1	4.506	3.955	3.500	**1.338**	1.609	6.104
	M2	4.506	3.974	3.644	1.698	**1.609**	6.095
	M3	4.506	3.823	3.994	1.209	**0.669**	4.848
		MAE					
		AR	MIBA	MF3PRF	GEMINI	CHAT	CLAUDE
2010-23	M1	0.971	0.903	0.837	**0.383**	0.450	0.978
	M2	0.971	0.903	0.837	0.479	**0.445**	0.982
	M3	0.971	0.903	0.837	0.341	**0.250**	0.711
2017-23	M1	1.608	1.544	1.428	**0.567**	0.664	1.654
	M2	1.608	1.529	1.456	0.735	**0.667**	1.669
	M3	1.608	1.470	1.516	0.502	**0.321**	1.169

Notes: See Table a. The evaluation window includes the Covid period.

Table 7: Best target - First vs. last release of GDP growth

(a) Without Covid

		First release			Final release		
	RMSE	GEMINI	CHAT	CLAUDE	GEMINI	CHAT	CLAUDE
2010-23	M1	0.212	0.220	0.215	1.69	1.56	1.66
	M2	0.218	0.210	0.212	1.67	1.60	1.67
	M3	0.200	0.189	0.205	1.83	1.73	1.72
2017-23	M1	0.200	0.174	0.177	1.89	2.08	2.13
	M2	0.199	0.172	0.183	1.90	2.10	2.09
	M3	0.183	0.171	0.183	2.14	2.12	2.09
	MAE	GEMINI	CHAT	CLAUDE	GEMINI	CHAT	CLAUDE
2010-23	M1	0.169	0.175	0.176	1.64	1.57	1.64
	M2	0.176	0.166	0.171	1.63	1.62	1.69
	M3	0.146	0.149	0.167	2.00	1.78	1.72
2017-23	M1	0.162	0.131	0.134	1.90	2.23	2.31
	M2	0.159	0.130	0.136	1.96	2.25	2.32
	M3	0.131	0.121	0.136	2.50	2.45	2.32

Notes: This table compares the accuracy of the nowcasts of the first versus the last release of GDP growth. The results are reported for the reference setup (simple prompt in English and advanced models) and for two out-of-sample windows: 2010Q1-2023Q4 and 2017Q1-2023Q4 (excluding 2020Q1 to 2021Q4). The left side shows the RMSE and MAE criteria of the first release nowcasts for the three LLMs in the first (M1), second (M2) and third (M3) months of the quarter. The right side shows the ratio of the criteria in the alternative case (*last release*) to the reference case (first release on the left). A ratio greater than one indicates a deterioration of the forecast accuracy for the final release.

(b) With Covid

		First release			Final release		
	RMSE	GEMINI	CHAT	CLAUDE	GEMINI	CHAT	CLAUDE
2010-23	M1	1.022	1.228	4.617	0.92	1.14	0.94
	M2	1.293	1.226	4.610	1.03	1.14	0.94
	M3	0.925	0.523	3.668	0.87	1.24	0.91
2017-23	M1	1.428	1.718	6.525	0.90	1.13	0.94
	M2	1.814	1.719	6.515	1.02	1.13	0.94
	M3	1.291	0.711	5.182	0.84	1.22	0.91
	MAE	GEMINI	CHAT	CLAUDE	GEMINI	CHAT	CLAUDE
2010-23	M1	0.400	0.473	1.041	1.16	1.24	1.07
	M2	0.504	0.468	1.043	1.10	1.24	1.08
	M3	0.356	0.258	0.750	1.19	1.39	1.11
2017-23	M1	0.627	0.740	1.876	1.06	1.21	1.03
	M2	0.821	0.744	1.890	1.01	1.20	1.03
	M3	0.555	0.348	1.313	1.04	1.35	1.05

Notes: See Table a. The evaluation window includes the Covid period.

Table 8: In-sample comparison of the econometric and LLMs nowcasts

(a) Without the Covid period

RMSE							
		AR	MIBA	MF3PRF	GEMINI	CHAT	CLAUDE
2010-23	M1	0.271	**0.175**	0.199	0.212	0.220	0.215
	M2	0.271	**0.177**	0.212	0.218	0.210	0.212
	M3	0.271	**0.161**	0.210	0.200	0.189	0.205
2017-23	M1	0.264	**0.150**	0.183	0.200	0.174	0.177
	M2	0.264	**0.156**	0.165	0.199	0.172	0.183
	M3	0.264	**0.155**	0.165	0.183	0.171	0.183
MAE							
		AR	MIBA	MF3PRF	GEMINI	CHAT	CLAUDE
2010-23	M1	0.210	**0.138**	0.157	0.169	0.175	0.176
	M2	0.210	**0.148**	0.164	0.176	0.166	0.171
	M3	0.210	**0.134**	0.167	0.146	0.149	0.167
2017-23	M1	0.197	**0.111**	0.134	0.162	0.131	0.134
	M2	0.197	0.131	0.133	0.159	**0.130**	0.136
	M3	0.197	0.123	0.122	0.131	**0.121**	0.136

Notes: This table reports the RMSE and MAE criteria of the nowcasts in the first (M1), second (M2) and third (M3) months of the quarter. The econometric models are estimated from 1995 to 2023 and the nowcasts generated on the two evaluation windows: 2010Q1-2023Q4 and 2017Q1-2023Q4 (excluding 2020Q1 to 2021Q4). The left side of the table shows the RMSE and MAE for the econometric models, AR, MIBA, MF3PRF and the right side presents the results for the LLMs in the baseline setup (simple question in English, advanced model). The colored cells indicate the best-performing model in each month.

(b) With the Covid period

RMSE							
		AR	MIBA	MF3PRF	GEMINI	CHAT	CLAUDE
2010-23	M1	3.313	2.957	2.877	**1.022**	1.228	4.617
	M2	3.313	3.007	2.955	1.293	**1.226**	4.610
	M3	3.313	2.886	2.941	0.925	**0.523**	3.668
2017-23	M1	4.678	4.177	4.063	**1.428**	1.718	6.525
	M2	4.678	4.249	4.174	1.814	**1.719**	6.515
	M3	4.678	4.078	4.155	1.291	**0.711**	5.182
MAE							
		AR	MIBA	MF3PRF	GEMINI	CHAT	CLAUDE
2010-23	M1	0.958	0.932	0.936	**0.400**	0.473	1.041
	M2	0.958	0.940	0.919	0.504	**0.468**	1.043
	M3	0.958	0.891	0.903	0.356	**0.258**	0.750
2017-23	M1	1.696	1.706	1.691	**0.627**	0.740	1.876
	M2	1.696	1.719	1.670	0.821	**0.744**	1.890
	M3	1.696	1.642	1.665	0.555	**0.348**	1.313

Notes: See Table a. The evaluation period includes the Covid period.

APPENDIX A - Release calendar of the Banque de France surveys

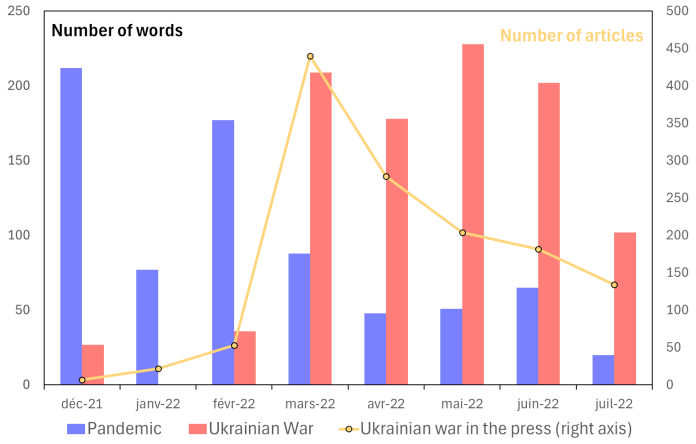
	First quarter			Second quarter			Third quarter			Fourth quarter		
Year	February	March	April	May	June	July	August	September	October	November	December	January
2010	8-Feb	8-Mar	9-Apr	10-May	8-Jun	8-Jul	9-Aug	8-Sep	8-Oct	9-Nov	8-Dec	11-Jan
2011	8-Feb	8-Mar	8-Apr	10-May	9-Jun	8-Jul	8-Aug	8-Sep	10-Oct	9-Nov	8-Dec	10-Jan
2012	8-Feb	8-Mar	10-Apr	10-May	8-Jun	9-Jul	8-Aug	10-Sep	8-Oct	9-Nov	10-Dec	10-Jan
2013	8-Feb	8-Mar	9-Apr	13-May	10-Jun	8-Jul	7-Aug	9-Sep	8-Oct	12-Nov	9-Dec	10-Jan
2014	10-Feb	10-Mar	8-Apr	12-May	10-Jun	8-Jul	8-Aug	12-Sep	8-Oct	7-Nov	8-Dec	10-Jan
2015	9-Feb	9-Mar	9-Apr	12-May	8-Jun	8-Jul	10-Aug	8-Sep	8-Oct	9-Nov	8-Dec	12-Jan
2016	8-Feb	9-Mar	8-Apr	10-May	8-Jun	11-Jul	8-Aug	8-Sep	10-Oct	9-Nov	8-Dec	9-Jan
2017	8-Feb	9-Mar	10-Apr	9-May	12-Jun	10-Jul	9-Aug	11-Sep	9-Oct	9-Nov	11-Dec	11-Jan
2018	8-Feb	8-Mar	11-Apr	14-May	11-Jun	11-Jun	8-Aug	10-Sep	8-Oct	12-Nov	10-Dec	11-Jan
2019	11-Feb	11-Mar	8-Apr	13-May	11-Jun	8-Jul	8-Aug	9-Sep	9-Oct	12-Nov	9-Dec	10-Jan
2020	NA	NA	NA	NA	NA	NA	10-Aug	NA	8-Oct	9-Nov	14-Dec	13-Jan
2021	9-Feb	8-Mar	12-Apr	10-May	14-Jun	7-Jul	9-Aug	13-Sep	11-Oct	8-Nov	7-Dec	11-Jan
2022	10-Feb	13-Mar	12-Apr	11-May	14-Jun	12-Jul	9-Aug	8-Sep	10-Oct	9-Nov	8-Dec	11-Jan
2023	8-Feb	8-Mar	11-Apr	10-May	8-Jun	10-Jul	9-Aug	12-Sep	9-Oct	8-Nov	11-Dec	10-Jan
2024	8-Feb	12-Mar	11-Apr	14-May	11-Jun	10-Jul	9-Aug	10-Sep	8-Oct	12-Nov	10-Dec	13-Jan

Notes: Dates of release of the Banque de France surveys in manufacturing, services, and construction. There were no releases in the first half of 2020 or in September of that year.

APPENDIX B - Look-ahead bias? Around the start of the war in Ukraine

(a) Pre-war period (b) War period

Notes: This figure shows the word clouds for the Document Term Matrices (with a tf-idf weighting) of the corpus of comments in English. The two alternative corpora include the comments that are generated by the three advanced LLMs (Gemini, ChatGPT and Claude) alongside the nowcasts with an explanatory prompt for the period before the start of the war in Ukraine (December 2021 to January 2022) and the beginning of the war in Ukraine (February to July 2022).



Notes: This figure shows the frequency of words belonging to two categories: pandemic (shown in blue) and the war in Ukraine (shown in red). The yellow curve depicts the frequency of articles in the French business newspaper *Les Echos* containing vocabulary related to the war in Ukraine (guerre or invasion and Ukraine).

A new wave of the Omicron variant of the SARS-CoV-2 virus emerged at the end of 2021. On February 22, 2022, Russian military forces entered Ukraine. Pandemic-related vocabulary consistently dominates the pre-war period (December 2021–January 2022). The use of war-related vocabulary begins to increase in February, peaking in March 2022. A similar pattern appears in the French press.

APPENDIX C - The best target for econometric models – first versus last release of GDP growth

(a) Without the Covid period

		First release			Final release		
	RMSE	AR	MIBA	MF3PRF	AR	MIBA	MF3PRF
2010-23	M1	0.308	0.192	0.177	1.36	1.65	1.87
	M2	0.308	0.187	0.172	1.36	1.70	1.84
	M3	0.308	0.170	0.170	1.36	1.77	1.75
2017-23	M1	0.334	0.157	0.165	1.43	2.28	2.39
	M2	0.334	0.158	0.154	1.43	2.45	2.47
	M3	0.334	0.158	0.148	1.43	2.35	2.32
	MAE	AR	MIBA	MF3PRF	AR	MIBA	MF3PRF
2010-23	M1	0.234	0.148	0.144	1.41	1.71	1.82
	M2	0.234	0.156	0.132	1.41	1.67	1.88
	M3	0.234	0.138	0.137	1.41	1.74	1.75
2017-23	M1	0.222	0.117	0.121	1.65	2.62	2.64
	M2	0.222	0.132	0.101	1.65	2.53	3.04
	M3	0.222	0.125	0.116	1.65	2.43	2.45

Notes: This table compares the accuracy of the nowcasts of the first versus the last release of French GDP growth for two out-of-sample windows: 2010Q1 to 2023Q4 and 2017Q1 to 2023Q4 (excluding 2020Q1 to 2021Q4). The left panel shows the RMSE and MAE criteria of the first release nowcasts for the three econometric models in the first, second, and third months of the quarter (M1, M2, and M3, respectively). The right panel shows the ratio of the criteria in the alternative case (last release) to those of the reference case (first release on the left). A ratio greater than one indicates a deterioration in the forecast accuracy for the final release.

(b) With the Covid period

		First release			Final release		
	RMSE	AR	MIBA	MF3PRF	AR	MIBA	MF3PRF
2010-23	M1	3.412	2.992	2.648	0.89	0.91	0.90
	M2	3.412	3.007	2.757	0.89	0.90	0.90
	M3	3.412	2.892	3.021	0.89	0.90	0.91
2017-23	M1	4.817	4.226	3.741	0.89	0.91	0.90
	M2	4.817	4.248	3.895	0.89	0.90	0.90
	M3	4.817	4.086	4.269	0.89	0.90	0.91
	MAE	AR	MIBA	MF3PRF	AR	MIBA	MF3PRF
2010-23	M1	1.031	0.949	0.882	1.00	1.02	1.03
	M2	1.031	0.948	0.898	1.00	1.02	1.03
	M3	1.031	0.898	0.932	1.00	1.02	1.01
2017-23	M1	1.819	1.728	1.605	0.96	0.98	0.98
	M2	1.819	1.722	1.641	0.96	0.98	0.99
	M3	1.819	1.650	1.71	0.96	0.97	0.97

Notes: See Table a. The evaluation period includes the Covid period.