Batuhan Koyuncu
koyuncu@cs.uni-saarland.de

Byeungchun Kwon
byeungchun.kwon@bis.org

Marco Lombardi
marco.lombardi@bis.org

Hyun Song Shin
hyunsong.shin@bis.org

Fernando Perez-Cruz
fernando.perez-cruz@bis.org

BISTRO: a general-purpose oracle for macroeconomic time series[1]

[PRELIMINARY AND INCOMPLETE / PLEASE DO NOT QUOTE WITHOUT AUTORS' PERMISSION]

Predictions of macroeconomic aggregates are a key ingredient in monetary policy decision-making, yet traditional model-based approaches are often inflexible. This article introduces the BIS Time-series Regression Oracle (BISTRO), a general-purpose time series model for macroeconomic forecasting. Building on the transformer architecture underlying Large Language Models (LLMs), BISTRO is fine-tuned on thousands of macroeconomic time series to replicate their cross-dynamics. We illustrate the model's capabilities by forecasting the 2021 US inflation surge. While standard benchmarks mechanically projected a linear path, BISTRO correctly anticipated the persistence and pervasiveness of the inflation wave, highlighting its ability to unveil nonlinear patterns. As such, BISTRO constitutes a low-cost, adaptable tool for producing both reliable baseline forecasts as well as scenarios.

JEL classification: C32, C45, C55, C87

Predictions on the dynamics of macroeconomic aggregates are a key ingredient to monetary policy decision making. These are typically informed by statistical or structural time-series models, which rely on historical developments of key economic variables and their correlations to extrapolate the future path of the variable(s) of interest. One limit of model-based approaches to macroeconomic predictions is that they are rather inflexible in their use: suitable models need to be constructed, estimated and validated to deal with different sets of variables and covariates, and different target variables typically require different modelling approaches.

To overcome these limitations, this article introduces a general-purpose time series model for the prediction of macroeconomic aggregates; we name it BIS Time-series Regression Oracle (BISTRO). BISTRO builds on the machinery underlying large language models (LLMs) and applies it to a landscape of time series, as opposed to text. The model is specifically crafted

[1] The views expressed in this publication are those of the authors and not necessarily those of the BIS or its member central banks. We also thank Matthias Rottner, Phurichai Rungcharoenkitkul, Andreas Schrimpf for helpful comments and suggestions. The authors have no sources of research funding other than that provided by their home institution, nor any conflict of interest to declare.

and fine-tuned for the forecasting and nowcasting of macroeconomic time series relevant for central banks: it is estimated on thousands of macroeconomic time series extracted from the BIS data portal and hence can replicate their cross-dynamics.

Key takeaways

- *We introduce the BIS time-series Regression Oracle (BISTRO). a general-purpose model for the prediction of macroeconomic aggregates.*
- *BISTRO works as a "ChatGPT" for macroeconomic time-series and provides accurate forecasts for key aggregates, and its flexibility enables a straightforward exploration of scenarios.*
- *To illustrate how BISTRO can help central bankers in their analysis, we showcase an application to inflation forecasting, in which we highlight its ability to uncover nonlinear patterns.*

The aim of this article is to provide a primer on this general-purpose time series model, with a special focus on how it can help economists in their forecasting and macroeconomic analyses. More precisely, we emphasise BISTRO's ability to produce reliable forecasts and to unveil nonlinear dependency patterns with a great degree of flexibility. For example, a researcher can produce a generic baseline forecast for, say, inflation, and then evaluate how conditioning on different covariates (and different assumptions for their evolution) modifies the baseline. As such, BISTRO constitutes a low-cost and easy-to-use tool to produce timely forecasts that perform well compared to traditional econometric benchmarks. In particular, we show how BISTRO is able to anticipate the persistence and pervasiveness of inflation the surge in 2021 (Graph 3).

The transformer architecture: from text to time series

Since the emergence of large language models (LLMs), economists have increasingly relied on them for text-based tasks such as drafting, coding, performing literature reviews, as well as the generation of research ideas.[2] As a result, AI tools have gained prominence and are now being routinely applied across many areas of economic analysis and research.[3]

LLMs are based on the *transformer* architecture: a neural network originally developed for language translation.[4] The transformer brought two major advantages over previous approaches. First, unlike earlier approaches that processed words one

[2] Korinek (2023) surveys how generative AI can support economic research, while Korinek (2025) explores specific applications, including the replication of existing results.

[3] See for example Kwon *et al* (2024), Gambacorta *et al* (2024), XXX.

[4] For further details, see Vaswani *et al* (2017).

at a time during *training*,[5] it can learn from the entire *context window*[6] simultaneously, speeding up training and hence enabling models to learn from much larger datasets. Second, the architecture proved versatile: components designed for translation could be repurposed as general-purpose language models, which are widely known as the Generative Pretrained Transformer (GPT) architecture.[7]

The key insight is that the GPT architecture treats text as a time series. Each word is first converted into a numerical representation called an *embedding*. The embedding of a word is a vector of numbers that captures the meaning of the word using the surrounding words. Take, for example, the word "bond": its meaning cannot be understood without context, as it could indicate either a financial instrument, a family relation or a fictional spy.[8] The transformer creates a different word embedding for each occurrence of the word "bond". Most importantly, the embedding is not dependent solely on how the word is spelled, but also on the context in which it appears, hence capturing its deepest meaning. More precisely, for any text featuring words w_1, w_2, ..., w_n, the model creates corresponding embeddings $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ... $\boldsymbol{x}_n$, which are high-dimensional vectors. These embeddings are then refined through successive layers of the network, with each layer incorporating information from previous words to improve the model's ability to predict what comes next. The machinery that creates and refines such embeddings is known as the *attention mechanism* (Box A).

The first GPT model was trained on roughly 7,000 unpublished books, with the objective of predicting the next word based on the preceding 512. Researchers then fine-tuned this base model for specific tasks, eg determining whether two sentences are semantically similar, assessing grammatical correctness, or answering questions. While each task required a separately fine-tuned model, it turned out that larger models trained on more data performed increasingly well across diverse tasks, even without task-specific fine-tuning. The property of solving any task for which the model has not been trained for is known as *zero-shot learning* and scaling these models to hundreds of billions of parameters drives the current development of AI.[9]

Box A

The attention mechanism: a dynamic weighting scheme

The attention mechanism is the cornerstone of the construction of contextualised word embeddings. In essence, it boils down to computing a weighted average of the information from different positions in the input sequence, where the weights themselves depend on the data. The word embedding for the i^{th} word is a linear combination of the word embeddings of words 1 to i, where the weight of each word depends on its importance in defining the i^{th} word. This differs from standard autoregressive models, where the weights are fixed; that is, the weights used to predict the next word from the previous i words are constant and independent of the context. The attention mechanism instead computes *context-dependent* weights. For each word in the sequence,

[5] In machine learning, training (or learning) is equivalent to the estimation of the parameters of the model in econometrics.

[6] In a transformer, the context window refers to the number of words they can process simultaneously. In current LLMs, the context windows can include up to hundreds of thousands or millions of words.

[7] See Radford *et al* (2018).

[8] For further details, see Chapter III in the BIS Annual Economic Report (2024).

[9] See Radford *et al* (2019) and Brown *et al* (2020).

it calculates how relevant every other past word is to the current prediction, assigning higher weights to the more relevant words.

Technically, attention works through three learned transformations—the *Query*, *Key*, and *Value* projections. The Query represents 'what am I looking for?', the Key represents 'what do I contain?', and the Value represents 'what information do I provide?'. The attention mechanism first computes the inner product between the Query and Key vectors. The Query represents the current (i^{th}) word, while the Key represents all past words in the context (from 1 through i). A higher inner product indicates a stronger dependency between the words, and these weights are normalized to sum to one. The embedding for the i^{th} word is then computed as a linear combination of the Value vectors, using these data-dependent weights. The result is a dynamic, data-dependent weighting that can capture long-range dependencies and structural breaks without being constrained by a rigid, pre-defined lag structure.

This mechanism can be understood as a form of generalised kernel regression or time-varying parameter model. The crucial difference is that the weights are not estimated from historical data alone but are computed fresh for each new input, allowing the model to adapt its behaviour to the specific context at hand.

Mathematically, any word w_i is first embedded into a vector using a linear projection, $\boldsymbol{x}_i = \boldsymbol{W} w_i$ and from this initial embedding that only depends on the i^{th} word spelling, we compute three vectors: the query $\boldsymbol{q}_i = \boldsymbol{W}^q \boldsymbol{x}_i$, the key $\boldsymbol{k}_i = \boldsymbol{W}^k \boldsymbol{x}_i$ and the value $\boldsymbol{v}_i = \boldsymbol{W}^v \boldsymbol{x}_i$. We compute the attention weights from the query and the key, as follows:

$$a_i(\boldsymbol{q}_i, \boldsymbol{k}_1, \cdots, \boldsymbol{k}_i) = \frac{\exp(\boldsymbol{q}_i^T \boldsymbol{k}_i)}{\sum_{j=1}^{i} \exp(\boldsymbol{q}_i^T \boldsymbol{k}_j)} \qquad (1)$$

We compute the contextualised embedding by linearly interpolating all the value vectors.

$$\boldsymbol{x}_i = \sum_{j=1}^{i} a_j(\boldsymbol{q}_i, \boldsymbol{k}_1, \cdots, \boldsymbol{k}_i) \boldsymbol{v}_j \qquad (2)$$

This resulting contextualised embedding depends on the embeddings of all the previous words. Finally, we can undo the embedding to reconstruct the word using another neural network ($\widehat{w_i} = f_\theta(\boldsymbol{x}_i)$). In Equation 2, we can observe that the weights a_i are not fixed and depend on previous words, which equips the transformer model with its zero-shot abilities.

Transformers apply multiple attention mechanisms in parallel (*multi-head attention*), with each head potentially capturing different types of relationships in the data, which are then combined into a unique embedding. These multi-head attention layers are stacked to create deep, powerful models. For instance, the GPT-3 model comprises 96 parallel multi-headed attentions and 96 stacked layers, totalling 175 billion parameters (Brown et al, 2020)

BISTRO: a general-purpose model for macroeconomic time series

Given that transformers are essentially predictive devices, researchers have naturally explored their application to time series forecasting. Most early work focused on general-purpose univariate time series models[10] that could not deal with covariates and cross-sectional dependencies, except for MOIRAI and Chronos 2.0.[11] These can

[10] See Koyuncu et al (2026) and Meyer et (2025) for a detailed review.

[11] See Woo et al (2024) for MOIRAI and Ansari et al (2025) for Chronos 2.0.

handle multiple variables simultaneously, hence opening the possibility to construct conditional forecasts based on an arbitrary number of covariates.

Both models are marketed as general-purpose time series models and achieve state-of-the-art performance on standard machine learning forecasting benchmarks. However, Chronos 2.0 has limitations for macroeconomic applications: its multivariate training data is synthetically generated, hence potentially missing the extent of structural change that is often found in macroeconomic time series. Even more importantly, its attention mechanism operates either across time *or* across variables, but not both. This means that Chronos 2.0 cannot directly capture how a covariate at one point in time might affect another variable in the future. The MOIRAI architecture instead offers several features that make it well-suited for macroeconomic analysis. It can handle varying numbers of variables and provides probabilistic forecasts for which uncertainty is also quantified. Critically, its attention mechanism jointly operates across both time and variables. It also accepts data measured at different frequencies.

As discussed, in the previous section, when processing text LLMs embed each word into a vector space. For time series, what the transformer embeds is a patch consisting of consecutive, non-overlapping observations. Hence, this patch is the minimal unit that the transformer can estimate for time-series, as the word was the minimal unit it could predict for text. Each patch is converted into an embedding vector that reflects both *what* is being measured (the variable) and *when* (the time period). Using econometric terminology, the embedding can be thought of as a state vector that best represents the characteristics of the patch. However, unlike state-space econometric models, where the state vector reduces the dimensionality of the input space, the embedding process in transformers results in a higher-dimensional vector to power its general-purpose capabilities. These embeddings are fed into the transformer, where the attention mechanism allows information to flow both across time and across variables. The model hence learns which of the historical patches (of any variable and at any lag) are most informative for predicting each target value, and it provides these weights adaptively rather than through fixed coefficients.

The MOIRAI model was trained (ie, estimated, in the econometrics jargon) on over 27 billion observations coming from time series of different nature and covering diverse sectors. However, the training dataset is heavily skewed toward energy (60%) and transport (18%), with finance and economics amounting to less than 0.1%. As shown in the forecasting examples below, the MOIRAI model fails to capture the correlation among macroeconomic time series, which underpins the need to fine-tune it using BIS macroeconomic data: by retraining the model on relevant data, we can improve its performance on the types of series central banks typically care about.

We fine-tuned MOIRAI using the BIS macroeconomic dataset (see Box B) and made several adaptations to better suit central banking applications.[12] We standardised the input frequency to daily observations and set the patch size to 32 days, roughly corresponding to monthly aggregates. The MOIRAI model supports a context window of up to 512 patches, which defines the number of past samples that can be utilised to estimate future patches. With the chosen patch size of 32 days, this corresponds to a context spanning nearly 45 years of data.

[12] Details about the modification over MOIRAI can be found in our working paper Koyuncu et al (2026).

Box B

Details of the dataset

The full internal dataset is made up of 4,925 time series across 63 economies, observed from 1970 through 2024. The training and evaluation window is set at 1984–2024, which is meant to improve coverage consistency. The dataset is designed to mimic the real-world information set that central banks' economists and forecasters rely on, with data arriving at different frequencies and being published with delays. So, the dataset includes series with varying periodicity and release lags.

The series are arranged into macro blocks to ensure broad coverage of macro-financial conditions. The biggest categories are exchange rates (1,699 series), prices and earnings (1,130), domestic interest rates and bond yields (731), and national accounts (554). Smaller macro blocks broaden coverage, reaching real economy indicators as well as monetary and balance sheet indicators. The topics they cover include labour market data (325), monetary aggregates (235), output indicators (153), demand indicators (83), and items from central bank balance sheets and income statements (15).

The dataset has a mixed-frequency structure, and it supports analysis across different time horizons. In total there are 4,925 series; of these, 2,151 are monthly, 1,314 are quarterly, 861 are annual, 536 are daily, 61 are weekly, and two are semi-annual. There is broad country coverage, with dense coverage in several major economies and regions, thereby supporting cross-country comparisons. The series are spread across Other Advanced Economies (1,020), Asia (368), Latin America (318), the G3 Major Advanced Economies (309), Emerging Europe (291), and Africa & Middle East (206).

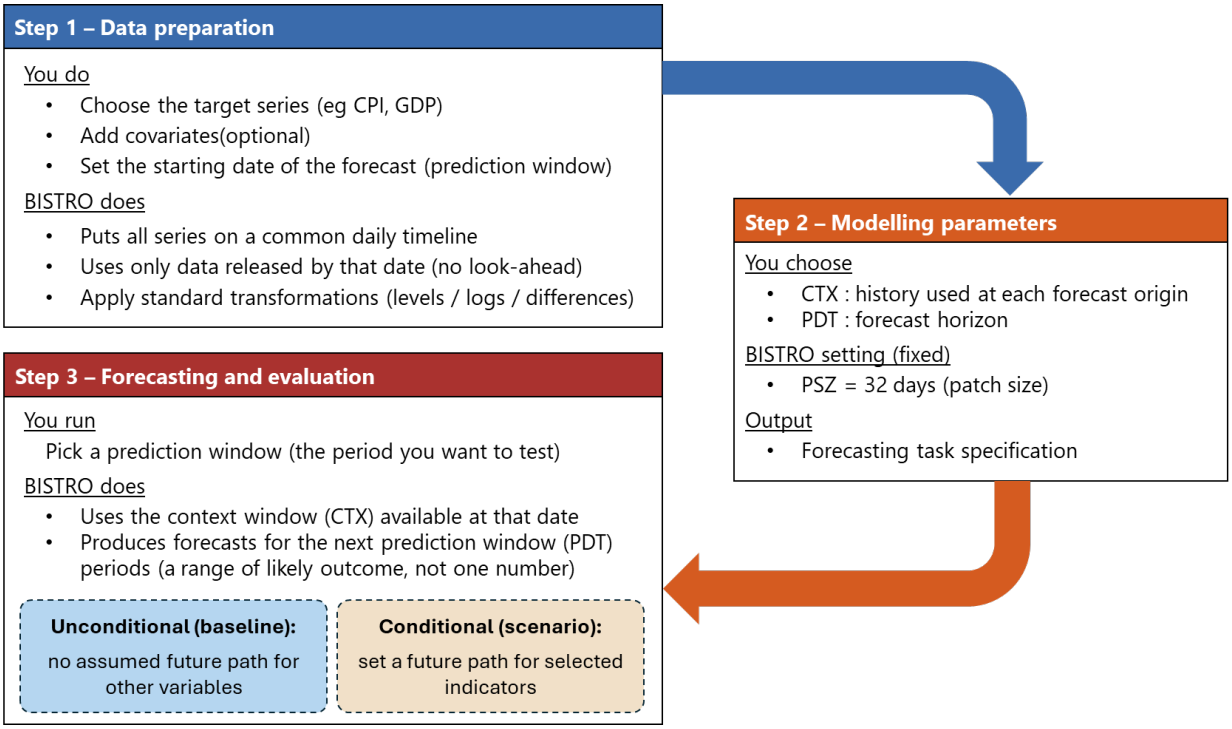
Standard transformations are used to make the data comparable across variables and to make it suitable for real-time forecasting. It uses levels, differences, as well as logs and log differences, but only as needed for the indicator. The use of a common timeline, along with publication lags, is intended to ensure forecasts rely only on information available at the time. They keep the most recent observation for each series and then extend that forward to the next release date to align the series. The approach respects publication lags. Public subset supports replication and benchmarking, but some BIS series are subject to redistribution limits.

How to operate BISTRO

As argued in the introduction, BISTRO provides a ready-to-use and off-the-shelf tool that helps perform various types of macroeconomic forecasting exercises. This section illustrates how this is done in practice through a stylised workflow (Graph 1). The workflow is divided into three stages: data preparation, modelling parameters, and forecasting and evaluation. A comprehensive Google Colab notebook is also available to support replication, providing sample scripts and step-by-step code for the entire workflow, including rolling-window forecast generation.

Workflow for macroeconomic forecasting with the general-purpose time series model

Graph 1



CTX = Context window length; PDT = Prediction window length; PSZ = Patch size.

The workflow has three steps. Step 1 builds the real-time dataset for each forecast origin (forecast date): all series are put on a common daily timeline, and BISTRO uses only data released by that date (no look-ahead), carrying forward the last available value until the next release. Step 2 defines the task by choosing context length (CTX) and prediction length (PDT). Step 3 runs a rolling-origin back test over the evaluation window, produces probabilistic forecasts, and checks accuracy using out-of-sample errors (forecast errors) and standard measures (eg RMSE). Forecasts can be unconditional (baseline) or conditional (scenario) based on an assumed path for selected indicators.

Source: Authors' elaboration

In the data preparation step, we line up the target variable (eg CPI inflation or unemployment) with the selected covariates (eg oil prices, or exchange rates) along the time dimension. This ensures that the target variable and the covariates are aligned in time, and that the series measured at lower frequency are converted to daily by holding the last known value. Given that the indicators are released at different frequencies and are subject to publication lags, this way of building the time series ensures that we do not include information that is not known for each given date (eg we only report the GDP for Q2 on the day its value is released in Q3). Hence, we work with a *real-time information set*, for each date only the known information is accessible until that date in the constructed time-series. This step is essential to mimic the operational forecasting environment faced by central bankers and to prevent forecasts from using information that is still unknown at each point in time.

We then pass the aligned real-time dataset to the model interface, which automatically performs the required preprocessing steps. In particular, the implementation applies the dataset's standard transformations (levels /logs/differences), handles missing values induced by publication lags, and determines how many past samples are used to predict future values given the user

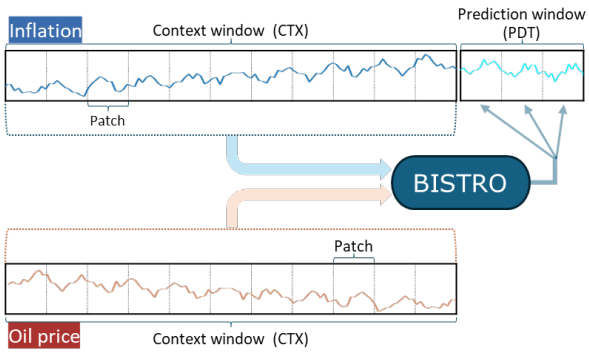
defined CTX value. For example, if we set the context window to 100 patches, the model will make its predictions using 8.75 years' worth of historical data. These steps are handled internally by the model pipeline, so users do not need to manually transform or re-format the data.

Let us illustrate with a simple example how to evaluate the model in a way that mirrors operational forecasting. Assume we want to create a forecast for inflation from 2023 until the end of 2025, and for each date, we aim to estimate the next year's inflation values (12 patches 32-day patches) using 20 years' worth of historical data (228 32-day patches). Hence, the model is prompted with observations from 1 January 2003 through 31 December 2022, and predicts inflation through 31 December 2023, that is for 12 consecutive monthly inflation readings. The model then proceeds by moving the data window one patch forward. Using data from 1 Feb 2003 through 31 Jan 2023, it predicts inflation for the next 12 months. It repeats the process iteratively until it reaches the end of the time series.

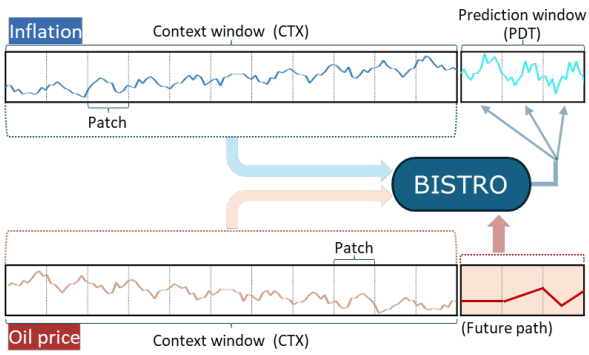
In BISTRO, the estimation process is fully automated, and prediction errors are systematically calculated. The results are then directly compared with those obtained from an AR(1) model. This approach enables a rapid assessment of the quality of the proposed solution by benchmarking it against a standard, simple benchmark. Continuing with the example of the previous paragraph, the user only needs to supply the CPI inflation data for the country or area of interest, between 2003 and the end of 2025. The user must specify both a context window of 228 patches (CTX) and a prediction horizon of 12 patches (PDT). The AR model is trained on data through December 2022, and the context is updated on a monthly basis. The BISTRO model is not retrained and it directly uses 228 patches context for each prediction.

Forecasting modes in BISTRO: baseline and scenario — Graph 2

A. Unconditional (baseline) forecast



B. Conditional (scenario) forecast



In panel A, BISTRO produces a baseline inflation forecast using only information available at the forecast date. In panel B, the inflation forecast is updated by imposing an assumed path for oil prices over the forecast horizon. The difference between the two forecasts reflects historical co-movements in the data, not causal effect.

Source: Authors' elaboration.

To obtain from BISTRO an inflation forecast that also depends on covariates (eg oil prices or exchange rates), these need to be added and aligned with inflation. The prediction will then depend on the past values of inflation and the covariates (Graph 2.A). BISTRO can handle up to 128 covariates.

BISTRO can also provide conditional forecasts, depending on future trajectories of the covariates. One just needs to fix the future paths of one or more covariates and

prompt them to the model (Graph 2.B). BISTRO can, for example, forecast inflation for 2026, conditional on different possible trajectories of the oil price. The tool thus allows researchers to explore alternative inflation scenarios in a straightforward manner.

Forecasting inflation with BISTRO

This section illustrates how BISTRO can be used to forecast inflation. More specifically, we compare the results obtained by prompting BISTRO with historical inflation data against those obtained from a standard time-series benchmark – namely a naïve autoregressive model with one lag – as well as against those produced the baseline MOIRAI.[13] We first examine the accuracy of BISTRO's unconditional forecasts – ie simple out-of-sample predictions that do not depend on covariates. Then we showcase how forecasts are affected by the inclusion of explanatory variables ("covariates") and how they can be tweaked to reflect assumptions on the future evolution of said covariates; this is commonly dubbed as a conditional forecasting exercise.

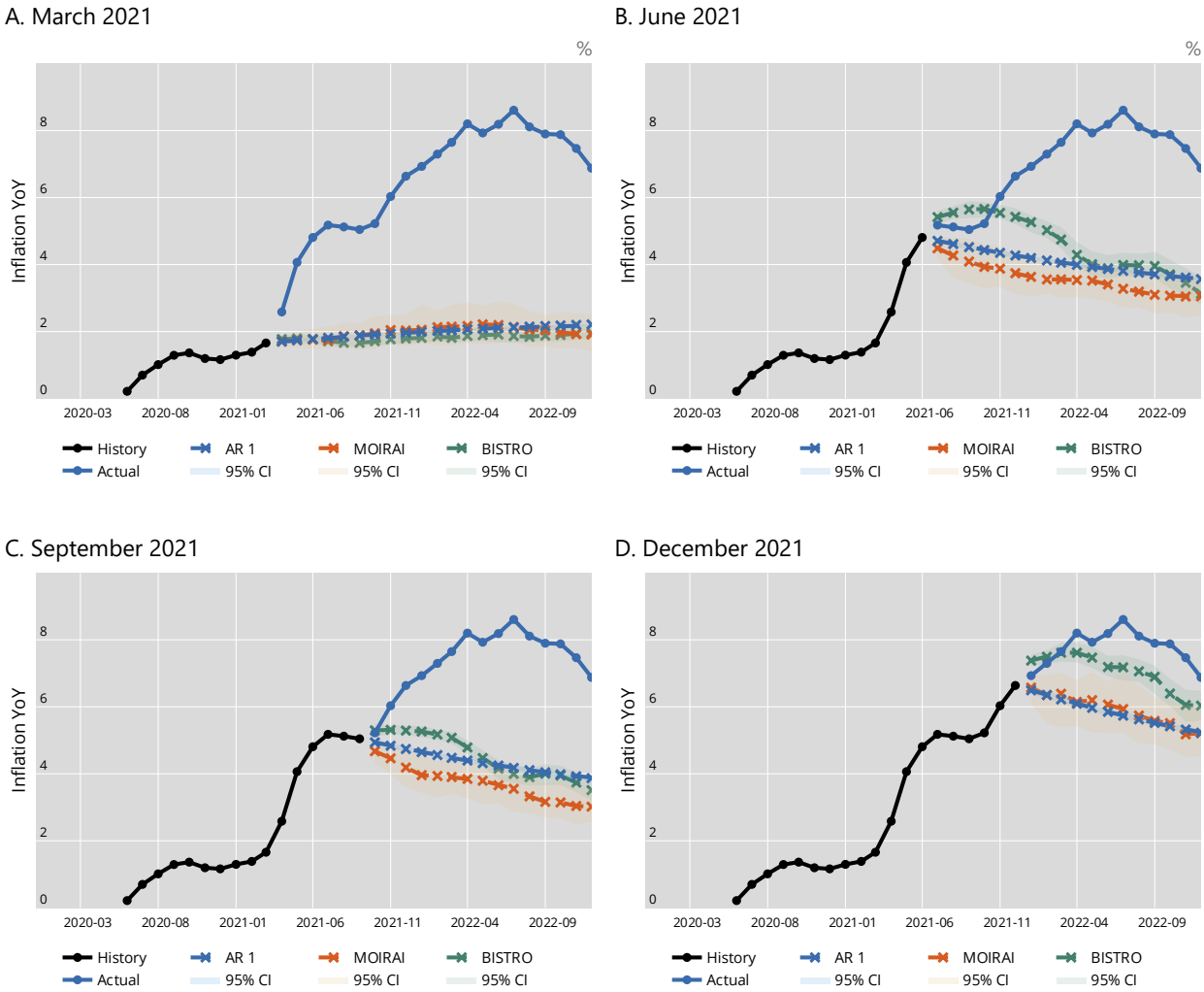
The unconditional forecasting task mimics a task of paramount relevance for central banks: predicting monthly inflation at short and medium horizons. To have BISTRO (and MOIRAI) generate unconditional forecasts, we prompt it with the past values of inflation, starting 240 months before the beginning of the forecast. Note that BISTRO relies on the same amount of information being provided by the analyst as a benchmark AR(1) – that is the latest inflation readings. Hence, what differentiates them is how said information is processed to produce the forecast: by applying an exponentially-decaying weight to past observations in the AR(1) case, and by generating ad-hoc weights based on the embedding of the past values of inflation in the case of BISTRO.

To visually illustrate the forecasting capabilities of the model, we put it to test on the inflation wave of 2021-2022. Graph 3 shows US inflation forecasts made by BISTRO, MOIRAI and a benchmark AR(1) model, starting at different points of the inflation surge. Each panel shows multi-step-ahead forecasts, in which predictions are produced dynamically over the desired forecast horizon and hence do not rely on inflation readings that were not yet available. Panel A shows forecasts made in March 2021, when inflation had not yet surfaced: all models provide a similar forecast of inflation converging towards the 2% target. Yet as the first signs of inflation surging appear, in June 2021, the model projections start diverging (Panel B). The AR(1) benchmark mechanically decays towards the historical average, while the MOIRAI model hints at an even faster decline. By contrast, BISTRO seems to correctly forecast a hump-shaped inflation trajectory. In September 2021 inflation had apparently plateaued, and while the AR(1) and MOIRAI stick to a declining trajectory, BISTRO still hints at a shallower decline (Panel C). In December, a second wave of inflation had already taken hold: BISTRO again hints at a hump-shape trajectory for the following months, which turns out to be relatively close to the actual outcome (Panel D).

[13] Despite their simplicity, simple autoregressive models are a tough benchmark to beat when forecasting macroeconomic aggregates – see for example Faust and Wright (2013) and Hall et al (2023).

Inflation forecasts throughout the 2021 surge

Graph 3



The figure presents a comparison of inflation forecasts extending until the end of 2022, as produced by BISTRO (depicted by green crosses), MOIRAI (represented by orange crosses), and AR(1) (illustrated by blue crosses). The observed inflation is denoted by the blue line with dots, while the unobserved (by the models) realised inflation is indicated by the solid blue line with dots. Each panel displays inflation estimates commencing from a distinct date during the period of the United States inflation surge in 2021 (March, June, September and December). Notably, the panels for March and December particularly highlight BISTRO's ability to accurately forecast the inflation peaks.

Source: Authors' calculations

The exercise illustrates well how BISTRO can identify and deal with possible structural changes and regime shifts entailed by the post-pandemic wave of inflation. To be sure, this is a very specific episode which simple linear time series models have a hard time processing. Moreover, the sample was included in the training sample for BISTRO, which could potentially represent an advantage.[14]

[14] During the training of the model, the inflation time series for the period 2021–2022 was never presented in isolation; it was always accompanied by other covariates and transformations of the inflation variable. Furthermore, when the time series was contextualised, it was not explicitly labelled as an inflation time series. The only information available to the model consisted of the numerical values within the time series. Had the model memorised the time series, one would anticipate that by March or September 2021, it would have already been capable of replicating exactly the forthcoming surge in inflation, which it does not.

Yet BISTRO also performs competitively and consistently across different time periods, not necessarily featuring extreme circumstances. To establish this, we run a forecasting comparison based on four distinct evaluation windows: 1995, 2005, 2015, and 2023-24, which were not included in the training sample of BISTRO. These windows reflect markedly different macroeconomic regimes, ranging from the pre- and post-Great Financial Crisis period to the recent episode of elevated inflation. This design allows us to assess model performance across phases characterised by different levels of volatility, sources of price pressures, as well as potential structural changes. Within each evaluation window, we focus on forecasts of the year-on-year (YoY) growth in the consumer price index (CPI) at one-, three-, six- and twelve-months-ahead horizons (*PDT* = 1, 3, 6, 12). Forecasts are generated sequentially over the test window as it progresses forward in time.[15] The analysis is conducted over all countries for which sufficient data are available in each testing window. Forecasts are conducted separately for each country, but to simplify the presentation, results are aggregated into four groups: the United States (US), the euro area (EA), other advanced economies (Other AE), and emerging market economies (EMEs).[16]

The results underscore that BISTRO provides reliable and accurate forecasts compared to the AR benchmark. Forecast accuracy is measured using relative root mean squared forecast errors (R-RMSFE) with respect to the AR(1) benchmark and averaged across countries and testing windows within each group (Table 1). So, a value below one indicates that BISTRO (or MOIRAI) outperforms the AR(1) benchmark. On average, BISTRO improves upon the benchmark in most cases, with particularly strong gains at the longer horizons.

Forecast Accuracy of BISTRO across regions and evaluation windows

1995, 2005, 2015, and 2023-24 (R-RMSFE; PDT = 1, 3, 6, 12) Table 1

	PDT = 1		PDT = 3		PDT = 6		PDT = 12	
Region	BISTRO	MOIRAI	BISTRO	MOIRAI	BISTRO	MOIRAI	BISTRO	MOIRAI
United States	**1.024**	1.136	**0.946**	1.241	**0.804**	1.335	**0.481**	1.349
Euro Area	1.065	**0.987**	**0.927**	0.932	**0.890**	0.956	**0.898**	1.689
Other AEs	1.157	**1.053**	**1.052**	1.108	**0.972**	1.220	**0.813**	1.097
EMEs	1.009	**0.874**	0.876	**0.844**	0.895	**0.854**	0.969	**0.850**

For each forecast horizon (PDT), we calculate the root mean squared forecast errors (RMSFE) for BISTRO, MOIRAI, and the AR(1) benchmark. We then compute the ratio of RMSFE values of BISTRO and MOIRAI relative to those of the AR(1) benchmark and report their average across evaluation windows and countries within each region.

AEs = Advanced economies: Australia (AU), Canada (CA), Czech Republic (CZ), Denmark (DK), Hong Kong (HK), Israel (IL), Japan (JP), New Zealand (NZ), Singapore (SG), South Korea (KR), Sweden (SE), Switzerland (CH), United Kingdom (GB);

EMEs =Emerging market economics: Algeria (DZ), Argentina (AR), Brazil (BR), Chile (CL), China (CN), Colombia (CO), Hungary (HU), India (IN), Kuwait (KW), Malaysia (MY), Mexico (MX), Morocco (MA), Peru (PE), Philippines (PH), Poland (PL), Romania (RO), Russia (RU), Saudi Arabia (SA), South Africa (ZA), Thailand (TH), Turkey (TR), Vietnam (VN).

Source: Author's calculations

For the United States, BISTRO outperforms both AR(1) and MOIRAI for all horizons except the one-month-ahead. Importantly, its relative performance

[15] To ensure comparability across models, we apply the same principle also to the AR(1) benchmark, whose coefficients are not iteratively re-estimated throughout the test period.

[16] For a detailed breakdown of the results, see Koyuncu et al (2026).

improves as the forecasting horizon lengthens: its RMFE is less than half that of the AR(1) when forecasting twelve-months-ahead. In the euro area, both general-purpose models perform similarly on average, with improvements over the AR(1) benchmark becoming larger as the forecasting horizon increases. Results for other advanced economies are mixed, with average performance close to the AR(1) benchmark.[17] Similarly, in emerging market economies, BISTRO delivers systematic improvements at the medium-term horizon.

Constructing scenarios with BISTRO

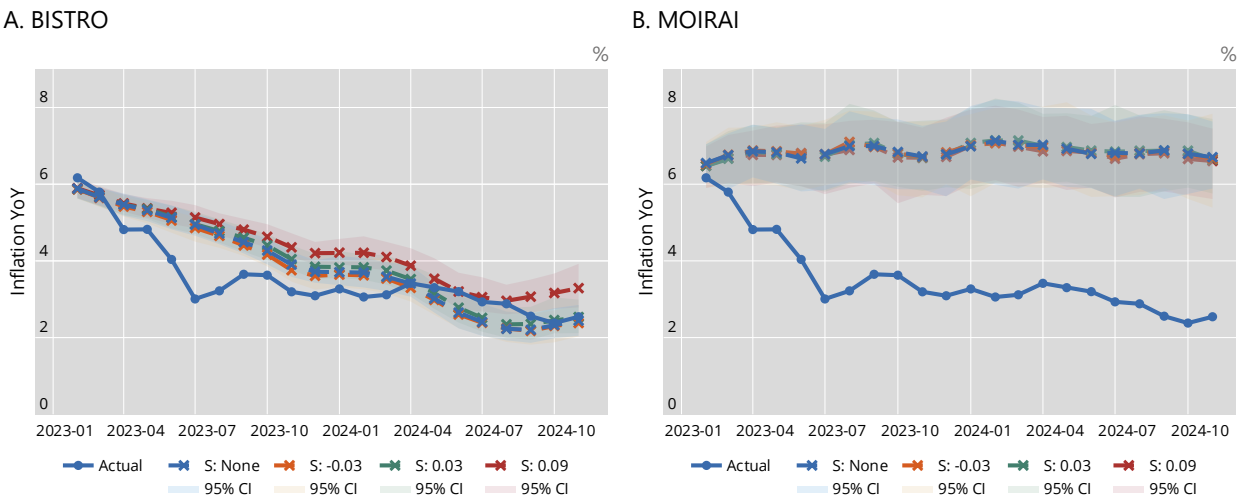
Having established a solid performance of our model in unconditional forecasting, we next turn to conditional forecasting exercises. One of the main advantages of the flexibility provided by our model lies in naturally handling relevant covariates one may want to condition the forecast on. In standard time series models, conditioning is possible only on a (typically quite limited) set of variables that are included in the model when it is specified and estimated. We can instead prompt BISTRO with trajectories of additional macroeconomic variables and hence generate projections that are conditional on the realisation of such trajectories. The way the model handles this is by relying on the historical patterns that emerge from the training dataset, and as such affect the embeddings. These can possibly also highlight relevant nonlinearities.

To illustrate this, we consider conditional forecasts of inflation based on different possible future trajectories of the oil price. Starting from its realised path, which serves as a baseline, we introduce alternative oil price scenarios in which the actual values are increased or decreased by fixed amounts (-0.03, 0.03, and 0.09). These correspond to counterfactual scenarios in which the oil price would be, respectively, 3% lower and 3%, and 9% higher than it actually was. In general, conditioning on oil prices produces systematically higher or lower oil price paths relative to the baseline (Graph 4.A). But even more importantly, the model unveils a substantial degree of nonlinearity in the reaction of inflation to oil prices. First, a decline in the price of oil has much smaller effects on inflation than an increase by the same amount. Second, larger increases in the price of oil have a more-than-proportional effect on inflation. A clear bearing on the relationship between oil prices and inflation is one of the advantages of our model over MOIRAI. The latter, being trained on a wider set of non-macroeconomic time series, fails to highlight any meaningful connection between oil prices and inflation (Graph 4.B).

[17] This is partly because BISTRO makes quite large errors compared to the AR(1) benchmark for the United Kingdom and New Zealand in the 2023-24 test window.

U.S. inflation counterfactual forecasts under oil price scenarios

Change in oil prices (baseline, ±0.03, +0.09) Graph 4



The figure compares projections generated by BISTRO and MOIRAI when conditioning on alternative oil price paths (baseline, ±0.03, and +0.09). Inflation is expressed in percentage terms. Each panel reports the realized path and the corresponding counterfactual projections under the different oil price scenarios.

Source: Author's calculations

Monetary policy as viewed from BISTRO

While our model is a device to uncover (potentially nonlinear and complex) correlations in the data, it entirely relies on observed series. Hence, it cannot provide insights into the effects on the observed time series of unobservable macroeconomic shocks, nor provide a decomposition of the forecasts it provides into structural forces with an economic interpretation. So, for example, it cannot be used to generate counterfactual scenarios based on different trajectories of monetary policy: in the end, the model, has no knowledge of a policy rule, nor on how a tighter policy would transmit to the real economy, so a conditioning on a policy rate trajectory will only yield a correlation based on historical patterns.

However, this suggests that the model can be used to recover historical response functions, eg a Taylor rule for monetary policy. To do so, we turn to conditional forecasting exercises in which the model is provided with realised macroeconomic paths and asked to generate the implied policy-rate projections based on such a macroeconomic outlook. More precisely, we condition the forecast for the US policy rate to the realised path of inflation and examine the implied trajectories for the policy rate under different historical contexts. This design allows us to isolate how the historical information provided affects the model-embedded relationship between macroeconomic conditions and monetary policy.

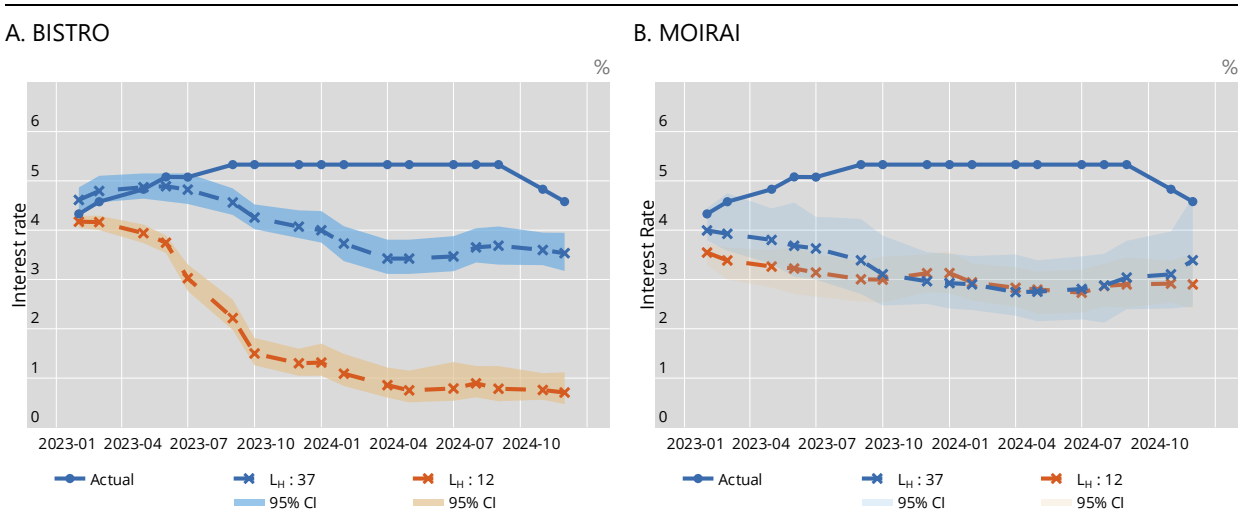
Results crucially depend on the amount of historical information we condition upon. When conditioning only on the last 12 years of data in our sample (eg 2011-2022), the model mostly avails of information based on a low-inflation regime in which monetary policy was mostly conducted through unconventional tools, due to the effective lower bound. Hence, the link between policy rates and inflation appears quite weak. As a result, the model fails to recognise that a policy tightening is

necessary to tame inflation, and instead hints at near-zero policy rates, implicitly assuming that inflation would self-correct without the need for aggressive intervention. This illustrates how the model's internal policy mapping remains weakly identified when short context windows fail to capture prior tightening cycles.

If one instead extends the context window to cover the entire training sample (1986-2022), the model gains exposure to different inflationary regimes, and to a phase in which conventional policy played a key role keeping inflation in check. In this richer context, the model correctly infers that high inflation typically calls for a tightening of policy rates, leading to a somewhat sharper rise in the implied interest rate path that slightly exceeds the realised pace of tightening in the very beginning. This suggests that the model effectively internalises the historical shifts in monetary policy and its tools, especially so when a sufficient context is provided.

The implied US policy rate forecast, conditioning on inflation

Policy rate prediction with history length of 12 years vs 37 years — Graph 5



Source: Author's calculations

By contrast, the projections generated by MOIRAI are largely insensitive to the length of the conditioning history, yielding similar policy rate paths across the two settings. This contrast suggests that BISTRO makes more effective use of historical information when capturing the correlations between inflation and monetary policy.

Opening the black box: the role of the attention weights

The different performance of BISTRO and MOIRAI in dealing with covariates and conditional scenarios underscores the key role played by the attention weights. While BISTRO's attention weights effectively integrate covariate data to generate meaningful scenarios, MOIRAI fails to assimilate this information, resulting in less informative projections.

To explain why it is so, let us first focus on MOIRAI. As illustrated in Graph 4, MOIRAI's predictions remain unchanged across the four distinct oil price scenarios.

This suggests that the attention weights assigned to the oil price time series are effectively zero, therefore the model ends up relying entirely on historical inflation data. Consequently, MOIRAI is unable to infer the relationship between inflation and oil prices from past observations. Furthermore, Graph 5 demonstrates that the predicted policy rate differs only marginally when contrasting short and long historical contexts. This implies that MOIRAI predominantly relies on the most recent readings, largely disregarding earlier values of both the policy rate and inflation.

Conversely, BISTRO, having been specifically trained on macroeconomic time series data, demonstrates an ability to assign attention weights that capture relevant historical information. In Graph 4, the model's nonlinear attention weights successfully reflect the complex, nonlinear impact of oil prices on inflation. Additionally, as shown in Graph 5, BISTRO leverages longer time horizons by assigning greater attention to observations from the more distant past, thereby enhancing its policy rate projections. This adaptability of the transformer model, whereby it can refine its estimates by adjusting attention weights, rather than requiring retraining, underscores its strength as a general-purpose forecasting tool. Nevertheless, this also highlights the necessity for the model to be exposed to time series pertinent to the specific forecasting task, implying that its universality is confined to a particular class of problems.

Concluding remarks

We introduced BISTRO, a time series general-purpose model explicitly fine-tuned for the forecasting of macroeconomic time series. While BISTRO is based on the MOIRAI architecture, we further refined it to fully capture the dynamics of specific macroconomic prediction problem, especially if those dynamics do not match the data the model was trained on. BISTRO provides solid unconditional forecasts and especially improves upon MOIRAI when the constructions of conditional scenarios is concerned. The key advantage of BISTRO compared to more traditional econometric alternatives lies in its flexibility to accommodate prediction of a multitude of series in unconditional as well as conditional terms, as well as in its inherent ability to unveil and deal with nonlinearities. While we provided a few examples, the model lends itself to plenty of different empirical exercises, ranging from pure forecasting/nowcasting to the analysis of scenarios.

References

Ansari, A F, O Shchur, J Küken, A Auer, B Han, P Mercado, S S Rangapuram, H Shen, L Stella, X Zhang, M Goswami, S Kapoor, D C Maddix, P Guerron, T Hu, J Yin, N Erickson, P M Desai, H Wang, H Rangwala, G Karypis, Y Wang and M Bohlke-Schneider (2025): "Chronos-2: from univariate to universal forecasting", arXiv preprint, arXiv:2510.15821, October.

Bank for International Settlements (BIS) (2024): "Artificial intelligence and the economy: implications for central banks", *Annual Economic Report*, June.

Brown, T B, B Mann, N Ryder, M Subbiah, J D Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, A Herbert-Voss, G Krueger, T Henighan, R Child, A Ramesh, D Ziegler, J Wu, C Winter, C Hesse, M Chen, E Sigler, M Litwin, S Gray, B

Chess, J Clark, C Berner, S McCandlish, A Radford, I Sutskever and D Amodei (2020): "Language models are few-shot learners", Advances in Neural Information Processing Systems, vol 33.

Faust, J and J H Wright (2013): "Forecasting inflation", *Handbook of Economic Forecasting*, vol 2A, pp 2–56.

Gambacorta, L, B Kwon, T Park, P Patelli and S Zhu (2024): "CB-LMs: language models for central banking", *BIS Working Papers*, no 1215.

Hall, S G, G S Tavlas and Y Wang (2023): "Forecasting inflation: the use of dynamic factor analysis and nonlinear combinations", *Journal of Forecasting*, vol 42, no 3, pp 514–529.

Korinek, A (2023): "Generative AI for Economic Research: Use Cases and Implications for Economists", *Journal of Economic Literature*, vol 61, no 4, pp 1281–1317.

Korinek, A (2025): "AI Agents for Economic Research", *NBER Working Paper*, no 34202.

Koyuncu, B, B Kwon, M Lombardi, F Perez-Cruz and HS Shin (2026): "Introducing BISTRO: a foundational model for unconditional and conditional forecasting of macroeconomic time series", *BIS Working Papers*, forthcoming.

Kwon, B, T Park, F Perez-Cruz and P Rungcharoenkitkul (2024): "Large language models: a primer for economists", *BIS Quarterly Review*, December.

Radford, A and K Narasimhan (2018): "Improving language understanding by generative pre-training", OpenAI, mimeo.

Radford, A, J Wu, R Child, D Luan, D Amodei and I Sutskever (2019): "Language models are unsupervised multitask learners", OpenAI, technical report.

Vaswani, A, N Shazeer, N Parmar, J Uszkoreit, L Jones, A N Gomez, Ł Kaiser and I Polosukhin (2017): "Attention is all you need", *Advances in Neural Information Processing Systems*, vol 30.

Woo, G, C Liu, A Kumar, C Xiong, S Savarese and D Sahoo (2024): "Unified training of universal time series forecasting transformers", International Conference on Machine Learning (ICML), vol 235, pp 53140–53164.