Deusto
Facultad de Ingeniería
Ingeniaritza Fakultatea

Grado en Ciencia de Datos e Inteligencia Artificial

Datuen Zientzia eta Adimen Artifizialeko gradua

Proyecto fin de grado

Gradu amaierako proiektua

Development and evaluation of a reproducible pipeline for economic time series forecasting with context-derived exogenous signals incorporated via large language models and the Model Context Protocol

Diego Ramírez Lacalle

Directora: Nicole Cusimano

Bilbao, julio de 2026

Deusto
Facultad de Ingeniería
Ingeniaritza Fakultatea

Grado en Ciencia de Datos e Inteligencia Artificial

Datuen Zientzia eta Adimen Artifizialeko gradua

Proyecto fin de grado

Gradu amaierako proiektua

Development and evaluation of a reproducible pipeline for economic time series forecasting with context-derived exogenous signals incorporated via large language models and the Model Context Protocol

Diego Ramírez Lacalle

Directora: Nicole Cusimano

Bilbao, julio de 2026

Abstract

This Final Degree Project addresses inflation forecasting as a practical and high-impact problem in economic analysis. The work focuses on the data science component of an integrated project, where the main goal is to evaluate whether modern time-series forecasting methods can produce more useful inflation forecasts when they are combined with relevant economic context.

The project studies monthly inflation indicators for Spain, a global inflation series, and the European Harmonised Index of Consumer Prices. It compares classical statistical methods, locally trained neural models, and pretrained foundation forecasting models under a common experimental design. It also evaluates whether external information, such as institutional indicators, macroeconomic variables, energy prices, uncertainty measures, and text-derived signals generated through a semantic context pipeline, can improve forecast accuracy.

The evaluation is based on rolling-origin backtesting for several forecast horizons, using standard error metrics and scale-normalised comparisons against seasonal naive behaviour. The results support a nuanced but relevant conclusion: foundation models are not a universal replacement for classical forecasting methods, yet they can become valuable when the horizon, target series, and contextual information are well aligned. This makes the project more than a benchmark exercise. It provides a reproducible way to test when semantic and economic context adds real predictive value, and when simpler models remain the most reliable option.

Descriptors

Inflation forecasting, Time series, Foundation models, Economic context, Forecast evaluation

Contents

List of Figures

List of Tables

1. INTRODUCTION

Economic forecasting is the process of using historical data, current indicators, and explicit assumptions to estimate the future behaviour of economic variables. These variables may include inflation, employment, interest rates, gross domestic product, exchange rates, consumption, prices, or financial conditions. Its main objective is to reduce uncertainty enough to support better decisions. In practice, economic forecasts are used to anticipate risks, compare scenarios, plan resources, evaluate policy alternatives, and detect changes in economic trends before they are fully visible in official statistics.

Inflation forecasting is one of the most relevant branches of economic forecasting. Inflation affects households, companies, public institutions, and financial markets because it changes the real value of income, savings, costs, and future decisions. Forecasting inflation is therefore both a technical and practical problem. A useful forecasting system can help anticipate possible changes, compare scenarios, and support better economic analysis. At the same time, inflation is difficult to model: it depends on local consumption patterns, monetary policy, energy prices, supply chains, expectations, and external shocks. The period after 2020 made this especially visible, with the combined effect of the pandemic, supply disruptions, and the energy shock in Europe.

This Final Degree Project has been developed as an integrated project with two complementary theses. The Computer Engineering (INF) part focuses on the software and system-engineering side: architecture, integration, deployment, and the organization of the platform. The Data Science and Artificial Intelligence (CDIA) part focuses on the data science and artificial intelligence side. Its purpose is to study whether modern forecasting models, especially time-series foundation models, can improve inflation forecasts when compared with classical statistical approaches, and whether external contextual signals add useful information to the prediction process.

In this thesis, time-series foundation models are understood as forecasting models pretrained on large and diverse collections of time series and later applied to a new forecasting problem with little or no task-specific training. This idea follows the recent research line reviewed in *Foundation Models for Time Series Analysis* [1], where the model is expected to transfer patterns learned from many temporal datasets to new forecasting tasks. Contextual signals, on the other hand, refer to external variables that are not the target inflation series itself, but may help explain its evolution, such as macroeconomic indicators, energy prices, uncertainty measures, institutional communication, or text-derived information.

The central research question is therefore:

Do foundation time-series models and contextual signals improve inflation forecasting, and under which conditions does this improvement appear?

The project studies three monthly inflation series with different scopes: the Spanish Consumer Price Index (Spain CPI), a global Consumer Price Index indicator (Global CPI), and the European Harmonised Index of Consumer Prices (European HICP). This selection is important because the usefulness of a forecasting model or an external signal may depend on the scale of the target variable. A signal that is relevant for broad global inflation may explain domestic Spanish inflation with less strength. In the same way, European institutional communication may be more directly related to European HICP than to a global aggregate. The project therefore compares several inflation contexts in the same experimental framework.

The models evaluated in the project are grouped into three families. First, classical statistical models such as Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA), Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors (SARIMAX), and automatic ARIMA (AutoARIMA) are used as strong baselines. They are included not as simple reference models, but as serious competitors, because in many economic time series they remain difficult to beat. Second, deep learning models such as Long Short-Term Memory (LSTM), Neural Basis Expansion Analysis for Time Series (N-BEATS), and Neural Hierarchical Interpolation for Time Series (N-HiTS) are considered to test whether neural models trained on the available data improve the forecasts. Third, the project evaluates foundation time-series models such as Chronos-2, TimesFM, and TimeGPT. These models are relevant because they are pretrained on large collections of time series and are designed to generalize to new forecasting tasks with limited task-specific training.

In addition to comparing model families, the project studies the role of contextual information. The baseline condition, named `C0`, uses only the historical target series. The contextual conditions, grouped under `C1`, add exogenous information, including institutional variables, macroeconomic indicators, energy prices, uncertainty measures, geopolitical risk, and text-derived signals processed through the Model Context Protocol (MCP) pipeline. These conditions are separated into `C1_inst` (institutional and macroeconomic context), `C1_mcp` (MCP and text-derived context), and `C1_full` (the combined context), so the analysis can distinguish between institutional and macroeconomic context, text-based signals, and their combination. This distinction is important because context is useful when it is available at the right time, aligned with the target, and relevant to the dynamics being predicted.

A rolling-origin evaluation is an out-of-sample forecasting procedure in which the forecast origin moves forward through time and the model is evaluated repeatedly using only the information available before each prediction. This type of design is commonly used in forecast evaluation because it is closer to the way a forecasting

system would operate in practice than a random train-test split [2].

The evaluation follows a rolling-origin backtesting design over the 2021–2024 period, with forecasts at 1, 3, 6, and 12-month horizons. This design is used to simulate a realistic forecasting situation: at each origin, the model can only use information that would have been available at that date. The main metrics used in the project are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Scaled Error (MASE) [3]. MASE is especially useful because it compares the model against a seasonal naive benchmark and makes the interpretation less dependent on the scale of each series. Where appropriate, Diebold-Mariano tests [4] are used to compare predictive performance more formally.

The expected contribution of the CDIA part is to identify when foundation models help, when classical forecasting remains stronger, and how this depends on the series, the forecast horizon, and the type of external context. The current evidence of the project points to a nuanced result: classical models remain highly competitive, especially at short horizons and for Spain CPI, while foundation models combined with relevant context show more value for global and European inflation at longer horizons. This makes the project useful as a model comparison and as an empirical study of the limits of contextual and foundation-model forecasting.

The expected contribution of the INF part is to turn that forecasting work into a software platform. The platform gives operational form to the project through a multi-service stack, forecasting adapters, drift analysis, and user-facing workflows for creating experiments, inspecting runs, comparing metrics, and exploring contextual information.

The rest of the thesis is organized as follows. Chapter 2 presents the background and justification of the project, including the economic relevance of inflation forecasting and the current state-of-the-art in time-series and contextual forecasting. Chapter 3 defines the objectives and scope of the CDIA contribution. Chapters 4 and 5 describe the planning and budget. Chapter 6 explains the methodology followed in the experimental work. Chapter 7 presents the main development of the project: datasets, preprocessing, signals, models, evaluation design, and results. Chapter 8 discusses the ethical implications of using AI-based forecasts in an economic context. Chapter 9 records the main incidents and limitations found during the project. Chapter 10 summarizes the conclusions and proposes future lines of work. The final parts collect the definitions, acronyms, bibliography, and appendices needed to support the main text.

2. BACKGROUND AND JUSTIFICATION

2.1. ECONOMIC CONTEXT

Inflation is one of the macroeconomic variables with the strongest effect on everyday economic decisions. It changes purchasing power, affects saving and consumption decisions, modifies production costs, and influences interest rates, wages, investment, and public policy. For households, inflation determines how much real income is preserved over time. For companies, it affects pricing, margins, inventory decisions, and financial planning. For banks and other financial institutions, inflation is connected to interest-rate expectations, credit risk, market conditions, and the interpretation of macroeconomic scenarios. For public institutions, inflation forecasts are a key input in policy design because many decisions must be taken before the full effect of current economic conditions is visible.

This makes inflation forecasting both useful and difficult. It is useful because anticipating inflation helps decision makers prepare for possible future changes. It is difficult because inflation is not driven by a single mechanism. It can respond to domestic demand, energy prices, supply-chain disruptions, monetary policy, exchange rates, expectations, geopolitical uncertainty, and communication from economic institutions. In addition, the relevance of these factors changes over time. The post-2020 period is a clear example: the pandemic, supply bottlenecks, energy-market tensions, and the European inflation shock created conditions where historical patterns alone were often not enough to explain short and medium-term movements.

For this reason, a forecasting project in this area should ask which model obtains the lowest error and under which conditions that model is reliable. A model that works for a global inflation aggregate may fail for a national CPI series. A signal that is useful during an energy shock may be irrelevant in a stable period. A complex model may improve long-horizon forecasts but add little value at one month. This project is justified by the need to study these differences in a structured and reproducible way.

2.2. CURRENT STATE-OF-THE-ART

The current state-of-the-art in inflation and economic forecasting can be understood as the convergence of several research lines that have advanced at different speeds. Classical statistical forecasting remains very strong, deep learning has expanded the set of possible models, foundation time-series models have recently introduced zero-shot and few-shot forecasting as a realistic option, and economic text signals have become increas-

ingly relevant for capturing information that is not contained in the target series alone. This project is located precisely at the intersection of those four lines.

2.2.1. Classical statistical forecasting

The first line is the classical statistical tradition. Models such as ARIMA, SARIMA, SARIMAX, VAR-type models, factor models, and forecast combinations remain common in economic forecasting because they are interpretable, relatively robust, and well suited to small or medium-sized time series. The Box-Jenkins tradition [5] established ARIMA-style modelling as one of the central approaches for time-series forecasting, while later automatic procedures such as the one implemented in the `forecast` package [6] made ARIMA selection more systematic and reproducible. In inflation forecasting, these models are not weak baselines. They often capture persistence, seasonality, and local dynamics very effectively, especially at short horizons.

This is important for this report because classical models provide serious forecasting references. In monthly CPI and HICP forecasting, the available history is limited, the frequency is low, and shocks can change the dynamics abruptly. Under those conditions, a well-specified ARIMA, SARIMA, or SARIMAX model can remain highly competitive. The state-of-the-art therefore supports a comparative approach in which modern models are evaluated against strong classical baselines under a fair backtesting design.

2.2.2. Deep learning and Transformer forecasting

The second line is machine learning and deep learning for time series. Recurrent architectures such as LSTM [7] were originally important because they could model sequential dependence beyond fixed autoregressive structures. Later models such as N-BEATS [8] and N-HiTS [9] showed that neural architectures specifically designed for forecasting could compete strongly on heterogeneous time-series benchmarks. Transformer-based models such as Informer [10] also became relevant because they addressed the problem of long-sequence forecasting and showed how attention mechanisms could be adapted to time-series settings.

However, the use of deep learning in economic forecasting requires caution. Macroeconomic series are relatively short, noisy, exposed to revisions, and affected by changes in policy, energy markets, supply chains, and expectations. These conditions make model complexity an empirical question. In this project, LSTM, N-BEATS, and N-HiTS act as an intermediate step between classical statistical models and foundation models: they help test whether any improvement comes from deep learning in general, or specifically from pretraining and foundation-model behaviour.

2.2.3. Foundation time-series models

The third line is the recent appearance of foundation time-series models. These models transfer the foundation-model paradigm to forecasting by pretraining on large collections of time series and then applying the learned patterns to new tasks with little or no task-specific training. Recent surveys on LLMs for time series [11] and foundation models for time-series analysis [1] describe this as an active and still developing field, with important challenges around representation, frequency differences, scaling, cross-domain transfer, and evaluation leakage.

TimeGPT [12] is one of the earliest examples of this direction. It is presented as a foundation model for time series capable of producing forecasts for datasets not seen during training, with the practical advantage of being accessible through a commercial API. In this project, TimeGPT is relevant because it represents the proprietary and service-based side of foundation forecasting. It also introduces a practical cost and reproducibility issue, since its use depends on external API access.

TimesFM [13], developed by Google, follows a decoder-only architecture adapted to time-series forecasting. Its paper presents a patched decoder-style attention model pretrained on a large time-series corpus, with the objective of achieving strong zero-shot performance across different history lengths, prediction horizons, and temporal granularities. For this project, TimesFM is especially relevant because it is one of the main foundation baselines evaluated and because the Europe HICP experiments show that it can become more useful when the context is aligned with the target series.

Chronos [14] approaches the problem differently. It tokenizes time-series values through scaling and quantization, and then trains language-model architectures on those tokenized series. This is conceptually important because it treats time-series forecasting as a form of sequence modelling close to language modelling. In the project, Chronos-2 is relevant because it provides one of the strongest long-horizon results for Global CPI when combined with institutional context.

The broader state-of-the-art also includes models that are not all evaluated directly in the project but help explain the direction of the field. MOMENT [15] proposes open time-series foundation models pretrained on the Time Series Pile and designed for multiple tasks such as forecasting, classification, anomaly detection, and imputation. Lag-Llama [16] develops a decoder-only probabilistic foundation model for univariate time-series forecasting, using lag features and pretraining across domains. Moirai [17] introduces a universal forecasting Transformer trained on a large open time-series archive, with attention to cross-frequency and multivariate forecasting challenges. Tiny Time Mixers [18] show another direction: small, efficient pretrained models de-

signed for zero-shot and few-shot multivariate forecasting, with much lower computational cost than large models.

Together, these models show that the field is moving quickly, but they also reveal unresolved questions. It is not enough to know that a model is pretrained on many time series. In an inflation problem, the relevant question is whether that pretraining transfers to low-frequency macroeconomic series affected by shocks, policy decisions, and country-specific conditions. This is why the project does not evaluate foundation models in isolation. It compares them with classical and deep baselines, separates the three target series, and studies the effect of the forecast horizon.

2.2.4. External context and economic text signals

The fourth line is the use of external and textual economic signals. Inflation responds to autoregressive behaviour as well as to energy prices, monetary policy, expectations, supply-chain pressure, geopolitical risk, exchange rates, fiscal uncertainty, institutional communication, and news narratives. For that reason, adding context is theoretically attractive, although each signal still has to prove its value in out-of-sample forecasting.

Spanish EPU work [19] and Spanish news-sentiment work [20] support the idea that text can contain timely macroeconomic information. Euro-area news sentiment [21] and PMI-text analysis [22] also show that targeted qualitative information can improve nowcasting in some contexts. Geopolitical risk indices [23] show that news-based measures can capture relevant macro-financial uncertainty. These sources justify the use of text-derived and institutional variables, but they do not prove in advance that those signals will improve CPI or HICP forecasting. That distinction is important: much of the literature concerns GDP nowcasting or uncertainty measurement, while this project evaluates inflation forecasting across several horizons.

Recent work also points to a specific limitation of many foundation models: they are often strongest as history-only or mainly univariate forecasters, while economic forecasting usually depends on exogenous variables. ChronosX [24] addresses this problem by proposing mechanisms to incorporate covariates into pretrained time-series models. Intervention-aware forecasting [25] goes in the same direction conceptually, arguing that external events and interventions can shape future dynamics and should be treated as timed information with a clear temporal position. This directly supports the design of the project: `C0` is the history-only condition, while `C1_inst`, `C1_mcp`, and `C1_full` test whether institutional, macroeconomic, and text-derived context adds value.

2.2.5. Research gap

The gap addressed by this project is the combination of these research lines. Classical and econometric models are well established in inflation forecasting. Deep models and Transformers have extended forecasting methods, with performance that depends strongly on the data and evaluation setting. Foundation time-series models are promising, although much of their evidence comes from broad benchmark datasets. This thesis studies them in a focused inflation setting with Spain, Global, and European targets. Economic text signals and uncertainty indices are supported in the literature, but much of that evidence is closer to nowcasting or macroeconomic monitoring than to multi-horizon CPI/HICP forecasting.

Therefore, the contribution of this project is the reproducible and leak-free evaluation of whether foundation time-series models and contextual signals actually improve inflation forecasting when compared with strong statistical and deep-learning baselines. The expected result is necessarily conditional: context may help when it is aligned with the target series and horizon, and it may add noise when the signal is too broad, too short, or insufficiently related to the inflation process being predicted.

2.3. TECHNICAL JUSTIFICATION

From a technical point of view, the project is justified by three needs. The first is the need for a fair comparison between model families. If foundation models are compared only with weak baselines, their value can be overstated. For this reason, the project includes ARIMA, SARIMA, SARIMAX, and AutoARIMA models, together with deep learning alternatives such as LSTM, N-BEATS, and N-HiTS. This makes it possible to evaluate whether improvements come from the foundation-model approach itself, from the use of contextual variables, or simply from the structure of the target series.

The second need is to separate the effect of different types of context. The project distinguishes between `C0`, where the model only uses the target series history, and several `C1` configurations where external information is added. Institutional and macroeconomic signals are separated from MCP or text-derived signals, and in some cases they are also combined. This design is important because a global conclusion such as “context helps” would be too vague. Context may help for one series and harm another. It may help at 12 months but not at 1 month. It may improve TimesFM but not TimeGPT. The project therefore evaluates context as an empirical question whose answer depends on the target, horizon, and model family.

The third need is methodological reliability. The evaluation uses rolling-origin backtesting over the 2021–2024 period and forecast horizons of 1, 3, 6, and 12 months. This is closer to a real forecasting situation than a single

train/test split, because each origin simulates the information available at a given point in time. The project also applies safeguards such as shifting exogenous signals so future information is not leaked into the model, scaling variables before residual correction, and using MASE to compare performance against a seasonal naive benchmark. These choices make the experiment more credible and reduce the risk of obtaining optimistic results from an unrealistic setup.

2.4. ECONOMIC AND SOCIAL JUSTIFICATION

The economic justification of the project comes from the value of better inflation analysis. Inflation forecasts are relevant for planning, budgeting, pricing, financial decisions, and policy discussion. Even a modest improvement can be useful if it occurs in the right horizon and for the right target. Long-horizon forecasts are especially relevant because many decisions in finance, business, and policy are made months before their consequences are observed. The project therefore studies when model complexity is justified and when simpler approaches remain more reliable.

The social justification is related to transparency and responsible interpretation. Inflation affects the cost of living and can become a source of uncertainty for society. Forecasting systems should therefore be presented with their limitations. A model may be useful without being definitive. A signal may be informative without being stable across all periods. A foundation model may perform well in one context and fail in another. By comparing several model families, targets, horizons, and signal types, this project supports a more careful view of AI-based economic forecasting.

2.5. PROFESSIONAL RELEVANCE IN THE BANKING SECTOR

This topic also has direct professional relevance for me because I am currently working in a bank, where inflation, interest rates, macroeconomic expectations, and risk conditions are part of the broader environment in which many decisions are interpreted. In a banking context, forecasting is a practical tool for scenario analysis, risk monitoring, market interpretation, customer and portfolio analysis, and the evaluation of macroeconomic assumptions used by different business areas.

This relevance became especially clear in a professional conversation with Pedro Gómez Tejerina [26], my professional supervisor. His profile is especially close to the topic of this project: he is a Senior Data Science Expert at BBVA, has more than 22 years of experience in the bank, and also teaches Business Intelligence and Big Data at the University of Deusto. From that position, he highlighted the importance of inflation forecasting for bank-

ing activity, data-driven analysis, and market-oriented decision making:

> "Current forecasting models are not always fully reliable for market decisions. In a bank that works with financial markets, anticipating inflation, interest rates, and macroeconomic movements is essential, and we need to remain aware of how technology is advancing in this area."

This comment complements the academic evaluation of the project and helps explain its applied relevance. Financial institutions need to anticipate changes in the economic environment and understand the limits of the models they use. A forecast based on a more advanced model still requires evidence of reliability. For that reason, this project is relevant to the banking sector because it compares modern AI forecasting models with classical baselines, studies whether contextual signals add value, and identifies the situations in which the resulting forecasts should be treated with caution.

2.6. PROJECT OPPORTUNITY

The opportunity of this project is therefore both practical and methodological. Practically, it addresses a real forecasting problem with clear economic relevance. Methodologically, it combines model comparison, contextual-signal engineering, and reproducible evaluation. I do not approach the problem as a search for the most modern model by default. The project asks whether different sources of information actually improve prediction, and it measures that improvement across series and horizons.

This distinction is important for the final interpretation of the thesis. If the results show that context helps in Global CPI and Europe HICP but not in Spain CPI, that is not a failure of the project. It is one of its main findings. The value of the work lies in identifying the conditions under which foundation models and semantic context are useful, and also the conditions under which classical models remain stronger. This makes the project a realistic contribution to applied economic forecasting: it tests a modern idea, but it does so with baselines, controls, and limitations that make the conclusion more credible.

3. OBJECTIVES AND SCOPE

3.1. GENERAL OBJECTIVE

The general objective of the CDIA part is to evaluate whether time-series foundation models and contextual economic signals can improve monthly inflation forecasting when compared with classical statistical models and locally trained deep learning models.

This objective is deliberately empirical. The project builds a reproducible experimental framework to compare model families, target series, forecast horizons, and signal conditions. The goal is to identify the cases in which foundation models and external context are useful, and the cases in which classical approaches remain stronger.

3.2. SPECIFIC OBJECTIVES

The general objective is divided into the following specific objectives:

1. Define the inflation forecasting problem for three monthly target series with different scopes: Spain CPI, Global CPI, and European HICP.

2. Build a consistent data preparation process that makes the target series and the available exogenous variables comparable at monthly frequency.

3. Construct and organize contextual signals that may be relevant for inflation forecasting, including institutional variables, macroeconomic indicators, energy prices, uncertainty measures, geopolitical risk, and MCP-derived text signals.

4. Establish strong statistical baselines using classical time-series models such as ARIMA, SARIMA, SARIMAX, and AutoARIMA.

5. Evaluate locally trained deep learning models, including LSTM, N-BEATS, and N-HiTS, as an intermediate comparison between classical methods and pretrained foundation models.

6. Evaluate time-series foundation models, especially Chronos-2, TimesFM, and TimeGPT, under comparable experimental conditions.

7. Compare univariate forecasting conditions with contextual forecasting conditions. In particular, distinguish between `C0`, `C1_inst`, `C1_mcp`, and `C1_full`, so that the effect of institutional, macroeconomic, and text-derived information can be analysed separately.

8. Apply a rolling-origin backtesting methodology over the 2021–2024 test period for 1, 3, 6, and 12-month forecast horizons.

9. Measure model performance with MAE, RMSE, and MASE, and use statistical comparison tests where they are available and appropriate.

10. Interpret the results by series, horizon, model family, and signal condition, identifying when contextual information improves the forecast and when it does not.

11. Extract practical conclusions about the limitations of foundation models and MCP-style contextual signals in inflation forecasting.

3.3. RESEARCH QUESTIONS

In order to reach the proposed objectives, the experimental work is guided by the following research questions:

1. Do foundation time-series models improve inflation forecasts over strong classical statistical baselines?

2. Does external economic context improve the forecasts, or can it degrade performance when the signal does not match the target series?

3. Are the results different for Spain CPI, Global CPI, and European HICP?

4. Does the usefulness of foundation models and contextual signals depend on the forecast horizon?

5. Can a reproducible pipeline that transforms semantic and institutional context into exogenous variables provide measurable value in an inflation forecasting task?

3.4. WORKING HYPOTHESES

From these research questions, the project is organised around five working hypotheses. They are treated as testable expectations that guide the experimental design and the interpretation of the results.

1. **Foundation-model transfer hypothesis.** A foundation time-series model is not trained only with the data from this project. It has already learned general forecasting patterns from many other time series before being applied to Spain CPI, Global CPI, or European HICP. The hypothesis is that this previous knowledge can help the model forecast inflation better than classical models or locally trained neural models in some cases, especially at medium and long horizons. However, this improvement is not expected to appear everywhere. At short horizons, or in series with strong local seasonality, simpler statistical models may still be more accurate.

2. **Context relevance hypothesis.** A contextual signal is useful when it is connected to the inflation series being predicted. For example, European Central Bank communications, European energy prices, and Eurozone uncertainty indicators are more directly related to European HICP than to a global inflation measure. In the same way, commodity prices, global uncertainty, supply-chain pressure, and international financial indicators are more meaningful for Global CPI. Spain CPI may be harder to improve with broad European or global signals, because part of its behaviour is national and may already be captured by its own historical pattern. Therefore, the project expects contextual conditions to help mainly when the signal and the target series describe the same economic environment.

3. **Shock-period benefit hypothesis.** Contextual conditions are expected to be more useful during volatile periods or regime changes, such as the inflation and energy shock period, because external signals may capture information that is not fully contained in the historical target series.

4. **Stability-period noise hypothesis.** During stable periods, contextual and text-derived signals may provide little improvement or even degrade performance, because weak, delayed, noisy, or contradictory signals can introduce unnecessary variation into the forecast.

5. **Horizon-dependence hypothesis.** The value of contextual information depends on the forecast horizon. Some signals may be more useful at short horizons, while institutional or macroeconomic signals may be more relevant at medium or long horizons due to delayed economic transmission.

These hypotheses give the project a more precise evaluation logic. The objective is to find the lowest error and to understand when the error changes, why it changes, and whether those changes are consistent with the type of model, the target series, the economic period, and the forecast horizon.

3.5. SCOPE OF THE CDIA CONTRIBUTION

The scope of the CDIA contribution is limited to the data science and artificial intelligence part of the integrated project. It includes the definition of the forecasting problem, the treatment of the data, the design of the experiments, the comparison of models, the evaluation of results, and the interpretation of the findings.

The analysis covers three target series:

- Spain CPI, representing a national inflation series.
- Global CPI, representing a broader international inflation indicator.
- European HICP, representing inflation in the European context.

The experimental scope includes statistical, deep learning, and foundation-model approaches. The statistical models provide the main validation baseline. The deep learning models allow the project to test whether locally trained neural models improve on classical methods. The foundation models represent the most recent forecasting approach evaluated in the project.

The contextual scope includes both structured economic variables and semantic or text-derived information. These signals are evaluated as possible exogenous inputs for forecasting through the comparison between `C0` and the different `C1` conditions.

3.6. OUT OF SCOPE

Several elements are intentionally outside the scope of the CDIA contribution.

First, the CDIA work focuses on the forecasting experiment. The system architecture, backend, frontend, deployment, and software integration aspects belong mainly to the INF part. They are mentioned here when they are necessary to explain data access, reproducibility, or the connection between the forecasting work and the integrated project.

Second, Chronos-2, TimesFM, and TimeGPT are evaluated as existing foundation models. The contribution lies in their application, comparison, and contextual evaluation for inflation forecasting.

Third, the MCP-related contribution evaluates whether a pipeline that converts semantic or institutional context into structured exogenous variables can improve forecasting models.

Fourth, the selected series and signals define the empirical scope of the work. They are sufficient to evaluate the main research question across national, global, and European contexts, while leaving the wider space of macroeconomic forecasting for future extensions.

3.7. EXPECTED CONTRIBUTION

The expected contribution of the CDIA part is a structured and reproducible evaluation of modern forecasting approaches for inflation. Its value lies in comparing methods under the same experimental logic and in showing that the usefulness of foundation models and contextual signals is conditional.

This framing is important because it matches the results obtained by the project. The work shows where modern models perform well and where simpler baselines remain stronger. The evaluation is carried out for 1, 3, 6, and 12-month forecast horizons, so the comparison does not reduce model performance to a single average result. It makes it possible to observe whether a model is useful for immediate forecasts, medium-term forecasts, or longer-horizon inflation anticipation.

In particular, the project supports a nuanced conclusion: classical models remain highly competitive for Spain CPI and short-horizon forecasts, while foundation models with relevant contextual signals show more value for Global CPI and European HICP at longer horizons. This makes the contribution stronger than a simple benchmark, because it connects performance with the nature of the target series, the selected forecast horizon, and the quality of the available context.

4. PLANNING

4.1. PLANNING APPROACH

This chapter presents the global planning of the integrated PFG. The project combines two differentiated contributions: the CDIA work, focused on the investigation and evaluation of inflation forecasting models, and the INF work, focused on the later development of the software platform that organises and exposes that forecasting workflow.

The planning follows an Agile and Kanban-oriented approach. The project was not managed as a fixed linear sequence where every detail was known from the beginning. Instead, the work was divided into user stories, tasks, and viable milestones. This made it possible to start with the CDIA investigation, check the forecasting results, and then move into the INF architecture and platform development with a clearer understanding of what the system had to support.

The complete workload is estimated at 725 hours. Of these, 365 hours correspond to the CDIA profile and 360 hours correspond to the INF profile. The CDIA effort is concentrated mainly in the first part of the project, where the research, data preparation, modelling, contextual-signal analysis, and evaluation were carried out. The INF effort becomes dominant afterwards, once the results had been checked and the platform requirements could be derived from the actual forecasting workflow. The writing of the CDIA memory overlapped with the INF development phase.

4.2. TASK DEFINITION

The work was organised into user stories so that each part of the project had a clear purpose and a measurable output. Table 4.1 summarizes the main user stories, their degree profile, and the tasks associated with each one.

Table 4.1.: User stories and main tasks of the integrated project

User story	Profile	Main tasks
US-1 Define the forecasting problem	CDIA	Identify the economic problem, define Spain CPI, Global CPI and European HICP as targets, formulate the research question, and decide the main model families to compare.
US-2 Prepare the data foundation	CDIA	Search data sources, clean and align monthly series, review candidate macroeconomic and institutional variables, and prepare the datasets for experimentation.
US-3 Build the initial forecasting workflow	CDIA	Use Spain CPI as the pilot case, execute exploratory analysis, define the rolling-origin protocol, and test classical baselines and initial foundation-model runs.
US-4 Add contextual conditions	CDIA	Define C0, C1_inst, C1_mcp, and C1_full; construct institutional, macroeconomic, energy, uncertainty, and MCP/text-derived signals.
US-5 Complete cross-series evaluation	CDIA	Extend the experiments to Global CPI and European HICP, compare model families by horizon, compute metrics, review results, and interpret the differences across targets.
US-6 Translate results into platform requirements	Shared	Convert datasets, models, contexts, runs, predictions, metrics, and results into software concepts that the INF platform can represent.
US-7 Implement the backend and persistence layer	INF	Design the API, domain entities, authentication, PostgreSQL persistence, MongoDB context storage, experiment workflow, predictions, and metrics.
US-8 Integrate forecasting and MCP services	INF	Build model adapters, connect MCP contextual access, configure MLflow tracking, and expose the forecasting workflow through platform services.
US-9 Build user workflows and validation	INF	Implement frontend views, comparison dashboards, forecast visualisations, simulation screens, Docker Compose deployment, health checks, tests, and drift analysis.
US-10 Prepare the final documentation	Shared	Write and review the CDIA and INF memories, prepare figures and tables, align shared sections, complete appendices, and collect definitions and bibliography.

4.2.1. General tasks

Some user stories are general because they connect both degree profiles. US-6 is shared because the INF platform requirements come directly from the CDIA experimental outputs. US-10 is also shared because the two memories must remain coherent while still presenting separate degree-specific contributions.

4.2.2. CDIA-specific tasks

The CDIA-specific work is represented by US-1 to US-5. These tasks cover the investigation and analysis part of the project: economic problem definition, data source selection, dataset preparation, model comparison, contextual-signal design, rolling-origin evaluation, and interpretation of the results.

4.2.3. INF-specific tasks

The INF-specific work is represented by US-7 to US-9. These tasks were developed after the main CDIA results had been checked. They cover the backend, persistence layer, forecasting adapters, MCP integration, frontend workflows, Docker environment, validation, and maintainability of the platform.

4.3. MILESTONES

The project was divided into three viable milestones. These milestones are used in the same sense as MVPs in Agile environments: each one marks a point where the project has reached a coherent and reviewable state.

1. MVP-1**: Forecasting evidence base.** This milestone includes US-1 to US-5. At this point, the CDIA investigation is complete enough to explain the inflation forecasting problem, the selected datasets, the model families, the contextual conditions, and the main empirical results.

2. MVP-2**: Operational forecasting platform.** This milestone includes US-6 to US-9. At this point, the forecasting workflow has been translated into platform requirements and implemented through backend services, persistence, adapters, MCP integration, frontend views, deployment, and validation checks.

3. MVP-3**: Final integrated documentation.** This milestone includes US-10. At this point, the two memories are written, reviewed, and aligned with the integrated nature of the project, while preserving the specific contribution of each degree.

4.4. TASK DISTRIBUTION

Table 4.2 presents the estimated effort by phase. The distribution reflects the actual chronology of the work: the first part is mainly CDIA, the second part is mainly INF, and the final documentation period overlaps both memories.

Table 4.2.: Estimated workload distribution by phase

Phase	**CDIA hours**	**INF hours**	**Main output**
Initial topic definition and CDIA research framing	30	5	Research question and first integrated-project idea
Literature review and data source definition	70	0	State-of-the-art, target series, and candidate signals
Spain CPI pilot and contextual-signal exploration	115	0	First forecasting workflow and C0/C1 logic
Global CPI, European HICP, and final CDIA evaluation	95	0	Cross-series results and interpretation
Transition from CDIA results to INF requirements	10	55	Platform requirements derived from the forecasting workflow
Backend, persistence, and forecasting integration	5	105	API, data model, storage, runs, predictions, metrics, and adapters
MCP integration, frontend, deployment, and validation	5	130	Context access, user workflows, Docker environment, and checks
Memory writing and final alignment	35	65	CDIA and INF reports, figures, tables, appendices, and review
Total	**365**	**360**	**725 hours of integrated project work**

Figure 4.1 summarises the same sequence graphically: CDIA research, translation of results into software requirements, INF development, and the final documentation/review period.

4.5. HUMAN-RESOURCE PLAN

The project was carried out individually, but for planning purposes it is useful to separate the work into professional roles. This makes the responsibilities clearer and also provides the basis for the budget chapter. Table 4.3 presents the roles considered in the plan.

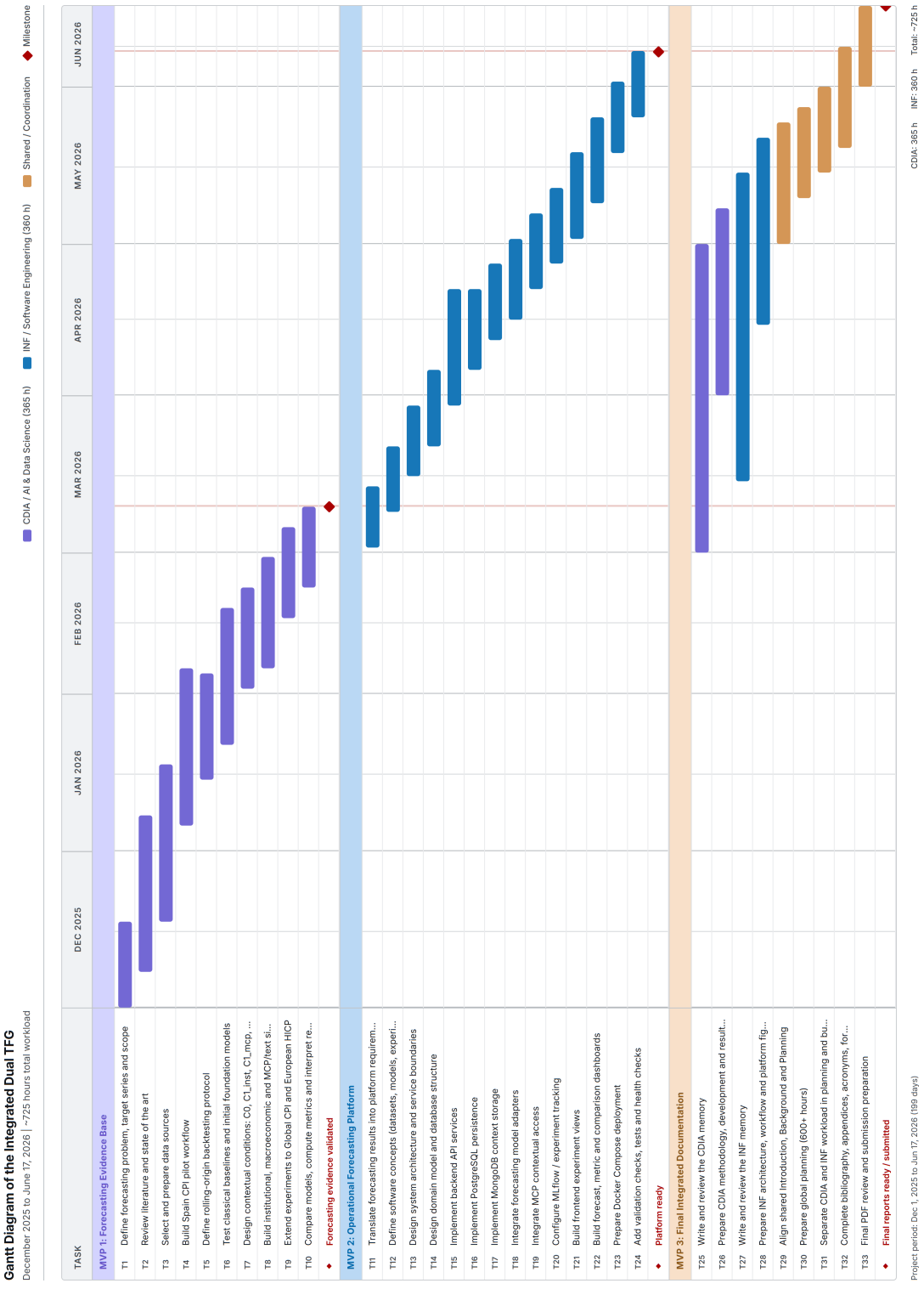

Figure 4.1.: Integrated Gantt diagram distinguishing MVPs, user stories, degree profiles, and estimated effort

Table 4.3.: Roles, responsibilities, and assigned personnel

Role	**Responsibilities**	**Assigned person**
Project director	Academic supervision, scope guidance, milestone review, and final validation of the project.	Supervisor
Project manager	Planning, Kanban-style task control, coordination between CDIA and INF, risk management, and final alignment.	Student
CDIA research analyst	State-of-the-art review, forecasting problem framing, and definition of the research gap.	Student
CDIA data scientist	Data source review, preprocessing, exploratory analysis, feature construction, and contextual-signal preparation.	Student
Machine-learning specialist	Model configuration, statistical and neural baselines, foundation-model experiments, rolling-origin evaluation, and result analysis.	Student
Software architect	Requirements analysis, architecture design, service boundaries, technology decisions, and platform structure.	Student
Backend and integration engineer	API implementation, persistence, forecasting adapters, MCP integration, MLflow, Docker Compose, and validation checks.	Student
Frontend engineer	User interface, experiment views, comparison dashboards, forecast visualisation, contextual views, and simulation screens.	Student
Technical writer	CDIA and INF memory writing, figures, tables, appendices, bibliography, and final review.	Student

Table 4.4 shows the estimated workload by role. The project director hours are shown separately because they represent supervision effort and are not included in the 725 hours of direct student work.

Table 4.4.: Estimated workload distribution by role

Role	**Percentage of student work**	**Hours**
Project director	Additional supervision	30
Project manager	3.3%	24
CDIA research analyst	6.9%	50
CDIA data scientist	12.4%	90
Machine-learning specialist	17.2%	125
CDIA evaluation and technical writing	13.8%	100
Software architect	6.9%	50
Backend and integration engineer	23.4%	170
Frontend engineer	8.3%	60
INF validation and documentation	7.7%	56
Total student workload	**100%**	**725**

4.6. PLANNING RISKS

The first planning risk was the uncertainty of the CDIA investigation. At the beginning of the project, it was not possible to know which model family would perform best or whether contextual signals would improve the forecasts. This was managed by starting with Spain CPI as a pilot case before extending the workflow to Global CPI and European HICP.

A second risk appeared during the transition from CDIA to INF. The platform could only be designed properly once the forecasting workflow and result structure were clear enough. This was managed by defining general platform concepts such as datasets, models, experiments, runs, predictions, metrics, and contexts.

A third risk was integration complexity. The INF block combined backend services, frontend screens, databases, MCP tools, MLflow, a gateway, and Docker Compose. This was managed by keeping clear service boundaries and by treating deployment and validation as explicit tasks rather than final additions.

5. BUDGET

With the common planning of the integrated project defined, this chapter estimates the economic cost of the work. The budget is not intended to represent real money paid during the project, but a professional valuation of the human effort, supervision, equipment, and technical resources that would be needed to develop an equivalent project.

The budget follows the workload distribution described in the previous chapter. It includes the complete integrated project, not only the CDIA part, because the requirements of a dual PFG ask for a global budget with the allocation of each profile clearly identified.

5.1. HUMAN RESOURCES

The main cost of the project is human work. Although the project was developed individually, the workload has been separated into professional roles so that the estimate is clearer. The project director is included as academic supervision, while the remaining roles correspond to the direct work carried out in the integrated project.

Table 5.1.: Estimated salary expenses

Human resource	**Unit cost**	**Hours**	**Total**
Project director	50 EUR/h	30	1,500 EUR
Project manager	35 EUR/h	24	840 EUR
CDIA research analyst	25 EUR/h	50	1,250 EUR
CDIA data scientist	30 EUR/h	90	2,700 EUR
Machine-learning specialist	35 EUR/h	125	4,375 EUR
CDIA evaluation analyst	30 EUR/h	55	1,650 EUR
CDIA technical writer	22 EUR/h	45	990 EUR
Software architect	35 EUR/h	50	1,750 EUR
Backend and data engineer	32 EUR/h	85	2,720 EUR
Integration and DevOps engineer	34 EUR/h	85	2,890 EUR
Frontend engineer	28 EUR/h	60	1,680 EUR
Validation and documentation engineer	26 EUR/h	56	1,456 EUR
Subtotal human resources	–	–	**23,801 EUR**

Table 5.2 separates the human-resource cost by project profile. The shared coordination and supervision costs

are kept separate because they support both memories.

Table 5.2.: Human-resource cost by project profile

Profile	Direct hours	Cost
Shared coordination and supervision	54	2,340 EUR
CDIA-specific work	365	10,965 EUR
INF-specific work	336	10,496 EUR
Total	**755**	**23,801 EUR**

The 755 hours in Table 5.2 include 30 hours of academic supervision. The direct student workload remains 725 hours, as defined in the planning chapter.

5.2. MATERIALS AND EQUIPMENT

The project was developed with personal equipment and mainly open-source software. However, an equivalent professional budget should include the proportional cost of the hardware and technical resources required to run experiments, develop the platform, store results, and maintain the documentation.

Table 5.3.: Estimated materials and equipment costs

Item	Cost	Justification
Personal workstation amortisation	450 EUR	Proportional use of the development computer during the project
External model and API access	250 EUR	Reference allowance for TimeGPT and occasional external model calls
Local execution and electricity	120 EUR	Repeated model executions, back-end/frontend runs, and Docker Compose testing
Storage and backups	80 EUR	Datasets, model outputs, figures, PDFs, repositories, and memory files
Connectivity and workspace infrastructure	150 EUR	Internet connection and general project infrastructure
Software licences	0 EUR	Main tools and libraries were open-source or available at no additional cost
Datasets	0 EUR	The data sources used were public or freely accessible
Subtotal materials and equipment	**1,050 EUR**	–

5.3. TOTAL BUDGET

The total budget is obtained by adding human resources and material resources. Table 5.4 summarizes the final estimate.

Table 5.4.: Budget summary

Concept	**Total cost**
Human resources	23,801 EUR
Materials and equipment	1,050 EUR
Total estimated budget	**24,851 EUR**

5.4. INTERPRETATION

The budget shows that the project is mainly a knowledge-work project. The largest cost is the time required to investigate the forecasting problem, prepare data, compare models, interpret results, design the platform, implement the software architecture, validate the system, and document both memories.

Material costs are limited because the project relies mostly on public data, open-source tools, and local development resources. This makes the project feasible from an economic point of view while still giving it the structure of a professional integrated project.

6. METHODOLOGY

6.1. METHODOLOGICAL APPROACH

The methodology of the CDIA part is designed around an empirical question: whether foundation time-series models and contextual economic signals improve inflation forecasting when they are compared under the same experimental conditions. For that reason, the CDIA contribution follows a data-science methodology centred on the definition, execution, and interpretation of a reproducible forecasting experiment.

The project follows a data-mining and model-evaluation methodology inspired by CRISP-DM, the Cross-Industry Standard Process for Data Mining [27]. This framework is suitable because it organises the work around the main stages of a data science project: understanding the problem, understanding the data, preparing the data, modelling, evaluating the results, and communicating the outcome. These stages match the structure of the project better than a linear methodology, because several decisions had to be revised after observing the first results.

The methodology was adapted to the specific nature of inflation forecasting. Inflation forecasting is temporal, and the order of the observations is part of the problem. A valid evaluation must respect the information that would have been available at each moment. For this reason, the project uses a time-based split and rolling-origin backtesting.

The methodology also accounts for the use of external signals. It compares univariate models trained on the historical target series and contextual models that incorporate institutional variables, macroeconomic indicators, energy prices, uncertainty measures, and text-derived MCP signals. This makes leakage prevention, temporal alignment, and signal interpretation especially important. A contextual variable is only valid if it is available before the forecasted month and if its effect is evaluated against a comparable baseline.

6.2. OVERALL METHODOLOGICAL WORKFLOW

The CRISP-DM logic is used as a practical reference for the complete workflow shown in Figure 6.1. The figure summarises the sequence from problem definition and data collection to data preparation, model families, contextual-signal conditions, rolling-origin evaluation, metrics, and final interpretation. The following sections develop each part of this workflow directly.

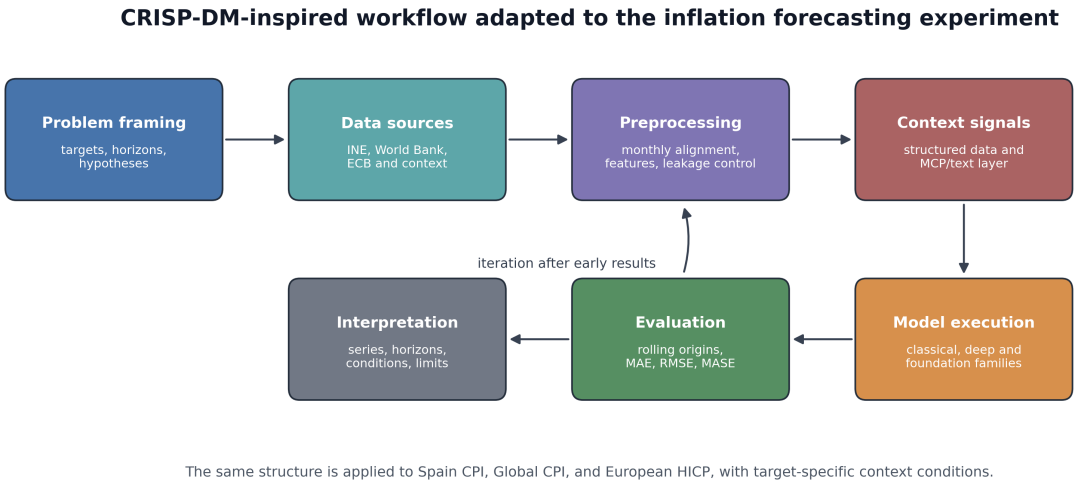

Figure 6.1.: Methodological workflow of the CDIA forecasting experiment

6.3. DATA STRUCTURE

The experiment is built around three target inflation series and several families of contextual variables. All the modelling work uses monthly data, because the target inflation indicators are published or transformed at monthly frequency. This decision simplifies the comparison between series and avoids mixing models trained on different temporal resolutions.

The data structure has three layers. The first layer contains the target series, which are the variables to forecast. The second layer contains exogenous variables, such as institutional indicators, macroeconomic variables, energy prices, uncertainty indices, and text-derived signals. The third layer contains the experimental conditions, which define whether a model receives only the target history or also receives contextual information.

Table 6.1.: Target inflation series used in the experiment

Series	**Main variable**	**Main source**	**Role in the experiment**
Spain CPI	`indice_general`	INE CPI data [28]	National inflation series used to evaluate domestic CPI dynamics.
Global CPI	`cpi_global_rate`	World Bank Global Database of Inflation [29]	Broad international inflation indicator used to test global macroeconomic context.
European HICP	`hicp_index`	ECB Data Portal HICP dataset [30]	European inflation indicator connected to monetary-policy and Eurozone context.

Spain CPI provides the domestic case, Global CPI provides the broad international case, and European HICP provides the Eurozone case. The contextual variables are then selected according to that scope, so the source of each signal is as important as the variable itself.

Table 6.2.: Contextual signal families and information sources

Signal family	**Examples used in the project**	**Main source or support**
Official and institutional variables	ECB rates, ECB communication indicators, European HICP-related context	ECB Data Portal and ECB publications [31]
Price, energy, and financial variables	Brent, TTF, exchange rates, market expectations, volatility and yield indicators	Yahoo Finance/yfinance and FRED series used by the ETL scripts [32, 33]
Uncertainty and risk variables	Spanish or European policy uncertainty, global uncertainty, supply-chain pressure, geopolitical risk	EPU/GPR literature, FRED, and New York Fed GSCPI data [19, 23, 33, 34]
Text-derived and MCP signals	GDELT tone, ECB tone, INE-release semantic signals, shock or uncertainty scores	GDELT Global Knowledge Graph and MCP/text-processing pipeline [35]

The experimental conditions define how this information is introduced into the forecasting task. The univariate condition, `C0`, uses only the historical target series. The contextual variants grouped under `C1` introduce structured economic variables, semantic signals, or both. Their precise role in the comparison is defined in Section 6.7.4.

6.4. DATA PREPARATION AND PREPROCESSING

The preprocessing stage converts heterogeneous economic sources into comparable monthly datasets that can be used by all model families. This step is essential because the project combines official price indices, monetary-policy variables, financial indicators, energy prices, uncertainty measures, and text-derived signals. These sources do not arrive with the same format, publication rhythm, scale, or temporal coverage. Before modelling, they must be cleaned, aligned, transformed, and checked for leakage.

The preparation follows a simple principle: each final dataset must represent the information that would have been available at the beginning of a forecast. This affects the way dates are parsed, how variables are merged, how lags and moving averages are created, and how missing contextual signals are treated. The aim is to create a coherent and defensible feature set.

6.4.1. Target ingestion

The first preprocessing task is to construct the three target series from their original sources. Spain CPI is extracted from the official INE workbook and represented by the general index [28]. Global CPI is derived from the World Bank Global Database of Inflation by calculating country-level year-on-year rates and taking the monthly cross-sectional median [29]. European HICP is downloaded from the ECB Data Portal through its SDMX service and represented by the Euro Area all-items index [30]. The concrete ingestion scripts, processed files, and quality checks are described in Chapter 7.

6.4.2. Monthly alignment

All datasets are aligned to monthly frequency. This decision is imposed by the target variables, but it also makes the model comparison cleaner. In the processed files, each observation is indexed by the first day of the month using a common month-start convention. This means that, for example, January 2022 is represented as `2022-01-01` in the target series, the macroeconomic variables, the financial variables, and the text-derived signals.

The aggregation rule depends on the type of variable. Daily financial variables, such as exchange rates or market indicators, are converted into monthly values using an average or end-of-month value depending on their economic interpretation. Policy-rate variables are treated as step functions, so the last available value is carried forward until a new decision changes the rate. Monthly variables that already come in the correct frequency are only normalised to the common date index. Textual and MCP-derived variables are grouped by month, and their numerical scores are averaged or counted depending on the signal definition.

This alignment makes the joins explicit and reproducible: every target value and every contextual value must refer to the same monthly timestamp before the dataset is passed to the models.

The alignment process also checks for gaps, duplicated dates, and unexpected missing values. Target missing values are not imputed because they would alter the variable being forecasted. Contextual variables are treated more carefully: some financial and institutional variables are forward-filled over short gaps when this is consistent with the nature of the variable, while unavailable text signals are handled with default neutral values and a separate availability indicator. This avoids confusing the absence of a text signal with a meaningful economic signal.

6.4.3. Feature engineering

Feature engineering is applied mainly to the contextual variables. The project uses transformations that are standard and interpretable for economic time series: lags, first differences, moving averages, log transformations, and returns. These transformations are useful because inflation may react to the level of a variable, its recent change, or its smoothed trend.

For monetary-policy variables, the preprocessing includes rate levels and lagged versions. For example, ECB deposit facility rate information is represented through the level, monthly difference, and several lags. This is relevant because monetary-policy effects are not necessarily immediate. For energy prices, the project uses logarithmic prices, monthly log returns, three-month moving averages, and lagged values. This creates separate views of the same underlying price process: level, short-term change, smoothed behaviour, and delayed effect.

For global institutional variables, the preprocessing builds derived features for each signal, including moving averages, lagged values, and differences. This is applied to indicators such as global economic policy uncertainty, commodity prices, the dollar index, market volatility, Treasury yields, policy rates, supply-chain pressure, geopolitical risk, Brent prices, and ECB deposit rate information. For the Europe-specific dataset, similar transformations are used for European sentiment, inflation expectations, exchange rates, energy prices, and monetary-policy variables.

The MCP and news-related signals require a different type of preparation because the original inputs are not numerical time series. The pipeline first collects monthly GDELT information and official or institutional texts such as ECB, INE, FOMC, or CPI-related releases. These documents are stored with their publication month and source, so the later aggregation can preserve the temporal origin of each signal.

After collection, the textual material is transformed into structured variables. Quantitative GDELT information is used directly through variables such as average tone, Goldstein score, and number of articles. Institutional releases are processed through the MCP/text pipeline to extract bounded scores, for example hawkishness, surprise, shock intensity, uncertainty, forward guidance, or CPI-direction pressure depending on the source. Categorical outputs such as tone or direction are converted into numerical encodings when the model requires numerical inputs.

The monthly feature table is then built by aggregating all observations that belong to the same month. Continuous scores are averaged, count variables are summed, and categorical variables are represented through the most common category or a numerical mapping. Because text-derived indicators can be noisy, the project

also creates smoothed variables such as three- or six-month moving averages for tone indicators. Finally, the dataset includes a `signal_available` variable, which marks the period in which the news and MCP signal layer is actually available. This is important because the absence of pre-2015 text signals should not be interpreted as evidence that the economic signal was neutral.

Table 6.3.: Main preprocessing transformations

Transformation	**Applied to**	**Purpose**
Lagged variables	Rates, energy, macro indicators, text signals	Represents information available before the forecasted month and captures delayed effects.
Moving averages	Energy prices, uncertainty indicators, text tone, macro signals	Smooths noisy monthly variation and captures recent trend behaviour.
First differences / returns	Rates, commodity prices, financial indicators	Represents changes in the variables across time.
Log transformations	Energy and commodity price variables	Reduces scale effects and makes relative movements easier to model.
Availability indicators	MCP and news-derived signal layers	Distinguishes unavailable historical text data from real neutral values.

6.4.4. Target-specific feature tables

After ingestion and feature engineering, the variables are assembled into separate datasets for Spain CPI, Global CPI, and European HICP. This preserves the economic scope of each forecasting problem instead of imposing an identical context layer on all three targets. Section 6.8.3 summarises the selected signal families, while Chapter 7 records the concrete files produced by the implementation.

6.4.5. Scaling and model readiness

Before the features are used in models that combine multiple variables, their scales must be controlled. This is especially relevant for correction models and exogenous-signal experiments. A raw uncertainty index, an interest rate, a tone score, and a commodity-price transformation can have very different magnitudes. Standardisation reduces the risk that the model gives excessive weight to variables with larger numerical scales.

For that reason, the modelling pipeline applies standardisation where the model requires comparable feature scales. The scaler is fitted on the available training portion and then applied to the corresponding evaluation data. This keeps the transformation inside the forecasting protocol and avoids using information from the eval-

uation period to define preprocessing parameters.

The final output of preprocessing is a set of model-ready tables with a common temporal logic. Each table has a monthly index, a clear target variable, a defined group of contextual features, and transformations designed to respect the information available at the forecast origin. This allows the later model comparison to be interpreted as a methodological comparison, supported by consistent data preparation.

6.5. TEMPORAL SPLIT

The temporal split is one of the most important parts of the methodology. Because the project is based on time series, random splitting would be invalid. A random split would allow information from later periods to influence the training process and would create an unrealistic evaluation. In a real forecasting setting, the model only has access to the past.

For that reason, the project uses a chronological split. The historical period from January 2002 to December 2020 is used as the initial training period. The rolling evaluation period starts in January 2021 and extends to December 2024. This window is especially relevant because it contains the post-pandemic period, the 2022 energy and inflation shock, and the later adjustment phase. Evaluating models in this period makes the task more difficult, but also more meaningful.

Table 6.4.: Temporal structure of the forecasting experiment

Period	**Dates**	**Use in the methodology**
Historical training period	2002-01 to 2020-12	Initial model fitting, fixed-order selection for some statistical models, and computation of the seasonal naive scale used in MASE.
Rolling-origin evaluation period	2021-01 to 2024-12	Monthly forecast origins for the classical and foundation-model experiments. The local neural baselines use quarterly origins to keep repeated training computationally viable.
Forecast horizons	1, 3, 6, and 12 months	Short, medium, and longer-horizon evaluation of each model family and signal condition.

The rolling-origin design uses an expanding training window. In the classical and foundation-model experiments, the model is fitted or applied at each monthly origin using the information available up to that date. The local neural baselines follow the same expanding-window logic at quarterly origins because they require repeated training. In both cases, forecasts are generated for the required horizons and the available history

grows as the evaluation advances.

For the longest horizons, a forecast is evaluated only when the corresponding observed value is available. This avoids comparing predictions against missing future values and keeps the evaluation tied to real observed inflation. The same principle applies to the exogenous variables: contextual information must be available before the target month being predicted.

6.6. ROLLING-ORIGIN BACKTESTING

The main evaluation procedure is rolling-origin backtesting. This method is used because the project needs to reproduce, as closely as possible, the situation in which a forecasting model would have been used in practice. At a given month, only the observations available up to that month can be used. The model then produces forecasts for future months, and these forecasts are later compared with the values that were actually observed.

This design is more demanding than a single train-test split. The experiment evaluates each model repeatedly across the 2021–2024 period, which makes the results more informative because the models are tested under different economic situations: the relative stability of 2021, the inflation and energy shock of 2022, and the adjustment period from 2023 to 2024. The objective is to obtain an average error and to understand whether each model family is stable across horizons and economic regimes.

Figure 6.2 shows the rolling-origin logic used in the experiment as an illustrative time-series cross-validation grid. Each row represents a successive forecast origin: the blue points are the available history, the orange diamonds are the evaluated forecast months, and the open grey points are months that are not used at that origin.

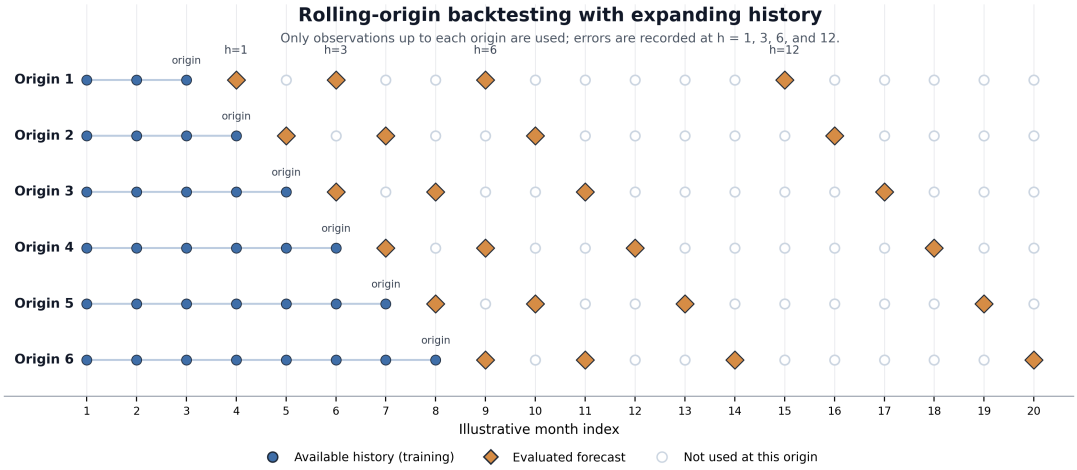

Figure 6.2.: Rolling-origin backtesting schematic: only observations up to the origin are used, the origin advances one month at a time, and forecast errors are recorded at each horizon.

Note. The schematic is illustrative. The full experiment uses about 48 monthly origins over 2021–2024; the local neural baselines use a quarterly subset of these origins.

6.6.1. Origin and horizon logic

The main evaluation grid is formed by monthly forecast origins from January 2021 to December 2024. At each origin t, the model uses the observed history up to t and generates forecasts for four horizons: 1, 3, 6, and 12 months. A horizon of 1 month evaluates immediate forecasting ability, while horizons of 3, 6, and 12 months test whether the model remains useful as the forecast becomes more difficult. The local neural baselines are evaluated on the quarterly subset of this grid and are interpreted with that difference in mind.

For an origin t, the forecast dates are $t+1$, $t+3$, $t+6$, and $t+12$, depending on the selected horizon. In the implementation, each multi-step forecast also keeps the intermediate forecasted months. A 12-month forecast therefore stores the full sequence of forecasted steps, not just the final value at month twelve. Keeping this structure makes it possible to analyse both the horizon-level performance and the monthly evolution of the errors.

A forecast is evaluated only if the corresponding real value exists inside the evaluation window. For example, an origin near the end of 2024 cannot be used for a 12-month forecast if the observed value twelve months later is not available in the dataset. This rule avoids artificial comparisons and keeps the evaluation based on observed inflation values.

Table 6.5.: Rolling-origin forecast structure

Element	Definition	Purpose in the experiment
Origin	Month from which a forecast is produced	Simulates the date at which a real forecast would have been made.
Expanding history	All observations from January 2002 up to the origin	Allows the model to use all past information without using future data.
Forecast horizon	1, 3, 6, or 12 months ahead	Compares short, medium, and longer-term forecasting behaviour.
Forecast date	Month being predicted	Allows errors to be aligned by model, origin, horizon, and observed month.

6.6.2. Expanding window

The project uses an expanding training window. The initial historical window covers January 2002 to December 2020. When the first forecast origin is evaluated, the models use that historical information. When the origin moves forward by one month, the new observed month is added to the available history. This process continues until the end of the evaluation period.

The expanding-window design is appropriate for this project because inflation forecasting benefits from preserving a long historical context. A sliding window would discard older observations and could be useful in a separate robustness analysis, but it would also remove information about previous inflation regimes and seasonality. The expanding design keeps the evaluation simple, reproducible, and closer to the way a forecasting system would normally accumulate information over time.

The exact way in which the expanding window is used depends on the model family. Classical statistical models are fitted at each monthly origin using the available history and the selected specification. Foundation time-series models also use monthly origins and receive the observed target history up to the forecast date. The local neural baselines preserve the same horizons but use quarterly origins. This distinction is considered when the results are interpreted, since the neural metrics summarize a smaller set of evaluation dates.

6.6.3. Leakage prevention

Leakage prevention is a central requirement of the methodology: every forecast must use only the target history and contextual information that would have been available before the forecasted month. This is especially important when working with exogenous variables and text-derived signals, because a small temporal misalign-

ment could introduce future information into the model.

The first safeguard is the chronological split itself. Training and evaluation are separated by time, not randomly. The second safeguard is the rolling-origin procedure, which rebuilds the available information at each forecast origin. The third safeguard is the treatment of contextual variables. The general rule is that exogenous information must be lagged or aligned according to its publication timing, so that the value used to forecast a given month corresponds to information that would already be available. For most contextual and text-derived signals, this means using the signal observed at $t-1$ to forecast t. This rule is applied to avoid using information that would only be known after the month being predicted.

The same logic applies to scaling and correction steps. Any transformation that learns parameters from data, such as standardisation, must be fitted on the training information available at that point and then applied to the corresponding forecast data. This matters because many contextual variables have very different units and magnitudes. Interest rates, text tones, commodity prices, uncertainty indices, and exchange-rate variables cannot be combined responsibly if their scales are ignored. Standardising these variables is therefore both a technical necessity and a way to avoid unstable corrections driven by raw magnitude differences.

6.6.4. Metric computation

The forecasting performance is evaluated with three main metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Scaled Error (MASE). These metrics were selected because they provide complementary views of the same forecasting problem.

For MAE and RMSE, y_i denotes the observed value, $\hat{y}_i$ denotes the forecasted value, and n denotes the number of evaluated predictions for the temporal horizon considered. MAE measures the average absolute difference between the forecast and the observation:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

This metric is easy to interpret because it is expressed in the same unit as the target series. RMSE gives more weight to large errors:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

RMSE is useful because inflation forecasting can be strongly affected by shock periods. A model that fails badly during a high-inflation episode should be penalised more strongly than a model that makes small and stable mistakes.

MASE [3] is included because it compares the model error against a seasonal naive benchmark. In this project the seasonal period is twelve months, which is appropriate for monthly inflation data. The scale is computed on the initial training period as the average absolute difference between each observation y_i and the observed value twelve months earlier y_{i-12}:

$$MASE = \frac{MAE}{\frac{1}{n-12}\sum_{i=13}^{n}|y_i - y_{i-12}|}$$

This makes the metric useful for comparing models across series with different scales. A MASE value below 1 means that the model improves on the seasonal naive benchmark. A value above 1 means that the model performs worse than that benchmark. This interpretation is important because a sophisticated model is not necessarily useful if it cannot beat a simple seasonal reference.

Table 6.6.: Forecasting metrics used in the evaluation

Metric	**What it measures**	**Why it is useful here**
MAE	Average absolute forecast error	Gives a direct and readable measure of forecasting accuracy.
RMSE	Squared-error measure with stronger penalty for large errors	Highlights models that fail during shock periods or unstable months.
MASE	MAE scaled by a seasonal naive benchmark	Allows comparison across target series and checks whether the model beats a simple seasonal baseline.

6.6.5. Statistical comparison

To complement the analysis of the error metrics, the evaluation phase also includes a statistical comparison between models. Two models may have similar MAE or RMSE values, and the difference may not be meaningful once the variability of the forecast errors is considered. For that reason, the project uses Diebold-Mariano tests [4] where the prediction records are available in a comparable format.

The test is applied by aligning the forecast errors of two models by origin, forecast date, and horizon. This alignment is important. A model should only be compared against another model on the same forecasted months and under the same horizon. Otherwise, the comparison could mix different economic periods or different levels of forecasting difficulty.

In the current implementation, the comparison is based on absolute-error differentials. This is consistent with the use of MAE as the main interpretable error metric. The test evaluates whether one model has systematically lower errors than another model, giving statistical support beyond visual inspection or ranking tables. A corrected version of the Diebold-Mariano procedure is used to account for small-sample behaviour and multi-step forecast dependence, following the Harvey, Leybourne and Newbold correction [36].

6.7. MODEL FAMILIES AND EXPERIMENTAL PROTOCOL

The model comparison is organised around three families: classical baselines, locally trained deep learning models, and foundation time-series models. This structure is deliberate. A fair evaluation of modern forecasting models requires strong references. Inflation series often contain persistence, trend, and seasonality, so classical time-series models can be difficult to beat. The project therefore uses classical and deep learning baselines to make the evaluation of foundation models more meaningful.

The protocol is designed to answer two related questions. The first question is whether foundation time-series models improve inflation forecasting compared with established forecasting approaches. The second question is whether contextual information improves the forecast once the same model is evaluated under a comparable univariate condition. This means that the project studies model performance and the usefulness of external information at the same time.

The contextual protocols are applied only when the model or the surrounding experimental design can accept exogenous information. For example, SARIMAX naturally incorporates external regressors, while some neural or foundation-model configurations require a specific covariate interface or a correction layer built after the

base forecast. MCP-derived signals are therefore treated as exogenous variables in the models that can use them directly, and as post-forecast contextual correction signals in the configurations where direct covariate input is not supported. This keeps the comparison realistic: each model is evaluated according to the type of information it can actually process.

The working hypotheses defined in Section 3.4 are translated into measurable comparisons inside this protocol. Each hypothesis is connected with a concrete comparison: model family, target series, forecast horizon, contextual condition, or economic period.

Table 6.7.: Operationalisation of the working hypotheses

Hypothesis	**Main comparison**	**Evidence expected from the experiment**
Foundation-model transfer	Foundation models vs. classical and local deep-learning baselines	Evidence that pretrained forecasting knowledge reduces MAE, RMSE, or MASE in some targets or horizons, especially at 3, 6, or 12 months. Stronger classical results are interpreted as evidence about the limits of transfer in this setting.
Context relevance	`C1` conditions vs. `C0`, separated by target series	Evidence that external variables improve the forecast only when they describe the same economic environment as the target series. The expected pattern is stronger usefulness for Global CPI and European HICP than for Spain CPI.
Shock-period benefit	Contextual models during volatile periods vs. stable periods	Larger relative improvement during the inflation and energy-shock period than during calmer periods.
Stability-period noise	Contextual models during stable periods vs. their univariate versions	Small, neutral, or negative improvements when the added signals are weak, delayed, or contradictory.
Horizon dependence	Same model and condition compared across 1, 3, 6, and 12-month horizons	Different behaviour across horizons, showing that context and model complexity do not have a uniform effect.

This structure also guides the later results chapter. The results are interpreted as evidence for or against these expectations, with attention to the target, horizon, model family, and contextual condition. For example, if a contextual foundation model improves Global CPI at twelve months but worsens Spain CPI at one month, the result supports the idea that the value of context depends on the target, the horizon, and the economic meaning of the signal.

6.7.1. Classical baselines

The classical baseline family provides the main benchmark for the project. These models are included because they are transparent, computationally efficient, and well suited to monthly economic series with persistence and seasonal structure, as described in the Box-Jenkins forecasting tradition [5] and automatic ARIMA selection work [6]. Their performance helps determine whether the additional complexity of foundation models is justified in this forecasting setting.

The classical group includes a seasonal naive benchmark, ARIMA, SARIMA, SARIMAX, and AutoARIMA variants. The seasonal naive benchmark [3] predicts future values using the value observed twelve months earlier. By construction, this benchmark captures annual seasonality, which makes it a strong reference for monthly inflation data. ARIMA models [5] capture autoregressive and moving-average dynamics, while SARIMA [5] extends this logic with a seasonal component that is relevant for CPI and HICP series. SARIMAX [5] introduces exogenous variables, making this model the classical counterpart of the contextual experiments. These specifications are selected during the preliminary analysis and kept fixed across their rolling evaluation. AutoARIMA [6] provides a separate dynamic comparison by recalibrating the order at each forecast origin.

The classical baselines also help control the interpretation of the results. If a foundation model performs well only against a naive benchmark but not against ARIMA or SARIMA, that result is weaker. If the foundation model improves over strong classical baselines at longer horizons, the improvement is more meaningful.

6.7.2. Deep learning baselines

The second family is formed by locally trained deep learning models. This family includes LSTM, N-BEATS, and N-HiTS. These models are useful because they represent neural forecasting approaches that do not rely on large pretrained foundation models. They help separate two possible explanations: whether a result comes from using a neural architecture in general, or whether the result comes specifically from the pretrained foundation-model approach.

The LSTM baseline [7] represents a recurrent neural approach that processes the time series as a sequence. N-BEATS [8] and N-HiTS [9] represent neural forecasting architectures based on feed-forward blocks and hierarchical multi-resolution structures. In the project, these models are trained on the available target series and evaluated with the same horizons and metrics used for the other families.

The deep learning baselines test whether locally trained neural approaches are competitive under the same

data constraints as the classical and foundation-model families. This comparison is relevant because monthly inflation data is relatively limited compared with the datasets normally used in large-scale deep learning.

6.7.3. Foundation time-series models

The third family is formed by foundation time-series models: Chronos-2 [14], TimesFM [13], and TimeGPT [12]. These models are evaluated as pretrained forecasting approaches and compared with the classical and locally trained neural baselines under the same target series, horizons, and metrics. Their contextual variants are handled through the condition protocol described below, because each model has different possibilities for using exogenous information directly or through a correction layer.

6.7.4. Condition protocol

The condition protocol makes the comparison more precise. The baseline condition is `C0`, where the model receives only the target history. This condition isolates the forecasting capacity of each model family. The contextual conditions are variants of `C1`. They introduce external information by type of signal, so that institutional, semantic, and combined contexts can be interpreted separately.

The main logic is shown in Table 6.8. The table presents the methodological structure followed in the experiments.

Table 6.8.: Model conditions used in the experimental protocol

Condition	**Information used**	**Purpose**
`C0`	Target history only	Measures the univariate forecasting capacity of the model.
`C1_inst`	Institutional, macroeconomic, uncertainty, financial, or energy variables	Tests whether structured economic context improves the forecast.
`C1_mcp`	Text-derived signals from news, GDELT, or institutional communications	Tests whether semantic or communication-based signals add predictive value.
`C1_full`	Combined institutional and MCP/text-derived variables	Tests whether broader context improves performance when both structured and semantic signals are available.
`C1_energy` / `C1_macro`	Narrower subsets of contextual variables	Supports ablation analysis by isolating specific signal groups.

The conditions are selected according to the target series. Some contextual sources are more meaningful for

one series than another. For example, European institutional and communication signals are more directly connected to European HICP than to the global median CPI series. The protocol therefore prioritises economic relevance over artificial symmetry.

6.7.5. Comparison across families

The comparison across families is based on common outputs. Each model produces forecast records with the same essential fields: forecast origin, forecast date, horizon, model name, observed value, predicted value, error, and absolute error. This shared structure allows the evaluation scripts to compute the same metrics and to build comparable tables and figures.

Table 6.9.: Model families compared in the project

Family	**Models**	**Role in the methodology**
Classical baselines	Seasonal naive, ARIMA, SARIMA, SARIMAX, AutoARIMA	Establish strong interpretable references for monthly inflation forecasting.
Deep learning baselines	LSTM, N-BEATS, N-HiTS	Test whether locally trained neural models are competitive under limited monthly data.
Foundation models	Chronos-2, TimesFM, TimeGPT	Evaluate whether pretrained time-series models transfer well to inflation forecasting and benefit from context.

The comparison is made by target series, horizon, model family, and condition. This structure preserves the main differences needed for a fair interpretation: all available model families are evaluated on the same target series and forecast horizons, the chronological split and rolling-origin design are shared across the comparison, the same metrics are used for all models, and contextual conditions are compared against the corresponding univariate condition and relevant baselines.

6.8. CONTEXT SIGNAL CONTRIBUTION AND USE

One of the main contributions of the CDIA part is the construction and evaluation of contextual signals for inflation forecasting by testing whether external information from institutions, markets, commodities, uncertainty indicators, geopolitical risk, and economic text improves the forecasts. The LLM and MCP components are used to transform narrative or institutional information into structured variables; the forecasting models remain responsible for producing the predictions. This design makes it possible to evaluate whether context helps during shocks or longer horizons, and whether weak or poorly aligned signals add noise.

6.8.1. Structured economic context

The first part of the contextual layer is formed by structured economic and financial data. These signals are already numerical, but they still require downloading, alignment, transformation, and leakage control before they can be used in the forecasting experiment.

The structured context includes monetary-policy rates, commodity and energy prices, uncertainty indices, geopolitical risk indicators, financial-market variables, supply-chain indicators, economic sentiment, inflation expectations, and exchange rates. These variables are used because they represent plausible transmission channels for inflation. Energy prices can affect production and consumption costs. Policy rates and forward guidance can affect demand, credit conditions, and expectations. Supply-chain pressure and geopolitical risk can represent external shocks. Exchange rates and commodity prices can affect imported inflation.

The project uses this structured layer mainly through the `C1_inst`, `C1_energy`, and `C1_macro` conditions. These conditions help test whether traditional economic context improves the forecasts before adding semantic or MCP-derived information. This makes the experiment more interpretable: if a contextual model improves, the analysis can distinguish whether the improvement comes from structured economic variables, text-derived signals, or the combination of both.

6.8.2. MCP and semantic signal pipeline

The second part of the contextual layer is formed by MCP and semantic signals. The pipeline acquires news events and institutional communications, applies predefined extraction schemas, aggregates the resulting variables by month, and preserves their temporal availability. Spain-oriented signals include GDELT event indicators and structured fields extracted from institutional publications. The global branch applies source-specific schemas to FOMC statements, ECB communications, and BLS CPI releases. This produces monthly quantitative variables that can be evaluated against the univariate baseline and the structured economic context. Chapter 7 describes the concrete server, storage, extraction, and validation components.

6.8.3. Signal families by target

The contextual layer is adapted to the economic scope of each target series. Table 6.10 summarises the structured and semantic signal groups used in each case. This target-specific design supports the context relevance hypothesis by allowing the experiment to evaluate whether signals are more useful when their economic scope matches the series being forecasted.

Table 6.10.: Target-specific contextual signal design

Target series	Structured context	Semantic context
Spain CPI	ECB rates, European uncertainty, and Brent and TTF energy variables	GDELT tone, ECB communication signals, INE surprise/topic signals, and availability indicators.
Global CPI	Global policy uncertainty, commodity indicators, dollar index, volatility, US Treasury yields, Federal Funds Rate, supply-chain pressure, geopolitical risk, Brent, and ECB deposit rate	FOMC, ECB, and BLS semantic signals.
European HICP	ECB rates, Brent and TTF energy variables, European policy uncertainty, Eurozone economic sentiment, five-year breakeven inflation, and EUR/USD	ECB tone and shock signals, GDELT tone, and availability indicators.

6.8.4. Safeguards for contextual signals

Contextual variables can easily make an experiment look better than it really is if they are handled carelessly. The methodology therefore applies three safeguards. First, signals are aligned to monthly frequency and shifted or lagged according to availability, so that the model uses only information available at the forecast origin. When a contextual variable is required for a future step but its future value is unknown, the last value observed at the forecast origin is carried forward through the requested horizon. Second, missing semantic history is represented carefully with neutral defaults and availability indicators, preserving the difference between missing information and genuine economic neutrality. Third, variables with very different scales are standardised when they are used in models or correction layers that require comparable magnitudes. The scaler is fitted separately on the information available at each origin.

These safeguards make the comparison more reliable by reducing the risk that apparent improvements come from leakage, scale differences, or missing-data artifacts. The implementation also includes automated integrity checks for temporal causality, duplicated predictions, origin grids, artifact availability, and the handling of exogenous variables.

6.9. EXPERIMENT EXECUTION AND REPRODUCIBILITY

The experiment follows a staged pipeline: target and contextual data are ingested, processed datasets are checked, model families are executed under the defined conditions, and the resulting predictions are evaluated through shared scripts. Each result can be traced back through a target series, a feature table, a model

execution, a prediction record, and a metric summary.

Processed datasets and prediction records are stored as Parquet files, while metric summaries and statistical comparisons are stored as JSON files. Separating model execution from evaluation makes the comparison reproducible and allows additional horizon-specific, condition-specific, and period-specific analyses without rerunning every model.

The complete execution also depends on a documented Python environment and model-specific requirements. In particular, TimeGPT requires an API key, and the MCP pipeline requires the configured components used to acquire and structure text sources. Chapter 7 describes the concrete repository modules and artifacts produced during implementation.

6.10. EVALUATION AND INTERPRETATION STRATEGY

The results are interpreted across five complementary dimensions: target series, forecast horizon, model family, contextual condition, and economic period. Spain CPI, Global CPI, and European HICP are analysed separately because each series represents a different economic scope. The evaluation then compares the classical, deep-learning, and foundation-model families at 1, 3, 6, and 12-month horizons. For models with contextual variants, the corresponding `C0` and `C1` conditions are compared to measure the contribution of external information under equivalent settings.

The interpretation combines MAE, RMSE, and MASE with the statistical comparisons described above. MASE provides the main cross-series reference against the seasonal naive benchmark, while MAE and RMSE describe the average error and the sensitivity to larger deviations. Horizon-specific metrics and the chronological inspection of forecasting errors are used to examine whether model behaviour changes during the different economic periods covered by the evaluation.

The results chapter will report improvements, degradations, and mixed outcomes with the same level of care. This makes it possible to identify the conditions under which foundation models and contextual signals provide useful information, while keeping the economic interpretation aligned with the predictive evidence produced by the experiment.

7. DEVELOPMENT

This chapter presents the practical development of the CDIA contribution. The methodological design has already defined the forecasting problem, the temporal protocol, the experimental conditions, and the evaluation criteria. The focus now moves to the results of the work carried out: target construction, exploratory analysis, contextual-signal generation, model execution, and the numerical evidence obtained from the comparison.

The chapter follows the execution flow of the project so that each result can be connected with the dataset, signal layer, model configuration, and stored metric that produced it.

7.1. DATA SOURCES AND TARGET CONSTRUCTION

The first development task was to construct reliable target series and processed datasets for the three forecasting problems. This step was necessary before any modelling work could be meaningful. The raw sources had different formats, different publication systems, and different structures, so they could not be used directly by the models. Each source had to be downloaded or read, parsed into a monthly time series, checked for temporal consistency, and saved in a format that could be reused by the rest of the pipeline.

The project stores the processed datasets in Parquet format. This decision makes the later stages more stable because the same cleaned files can be used by exploratory notebooks, statistical models, deep learning models, foundation-model scripts, and evaluation notebooks. It also avoids repeating the ingestion logic every time a model is executed.

Table 7.1.: Main processed target datasets

Target	**Processed file**	**Period**	**Main variable**
Spain CPI	`ipc_spain_index.parquet`	2002-01 to 2026-01	`indice_general`
Global CPI	`cpi_global_monthly.parquet`	2002-01 to 2024-12	`cpi_global_rate`
European HICP	`hicp_europe_index.parquet`	2002-01 to 2024-12	`hicp_index`

Although the Spain CPI file extends to 2026-01, the model comparison uses the common experimental period defined in the methodology. The initial training window covers 2002-01 to 2020-12, and the rolling-origin evaluation window covers 2021-01 to 2024-12. This common period is important because the final comparison must be based on the same forecasting conditions.

Figure 7.1 summarizes the data-source map. It shows the three target sources, the contextual sources, the ETL layer, the processed Parquet datasets, and the model-ready feature tables.

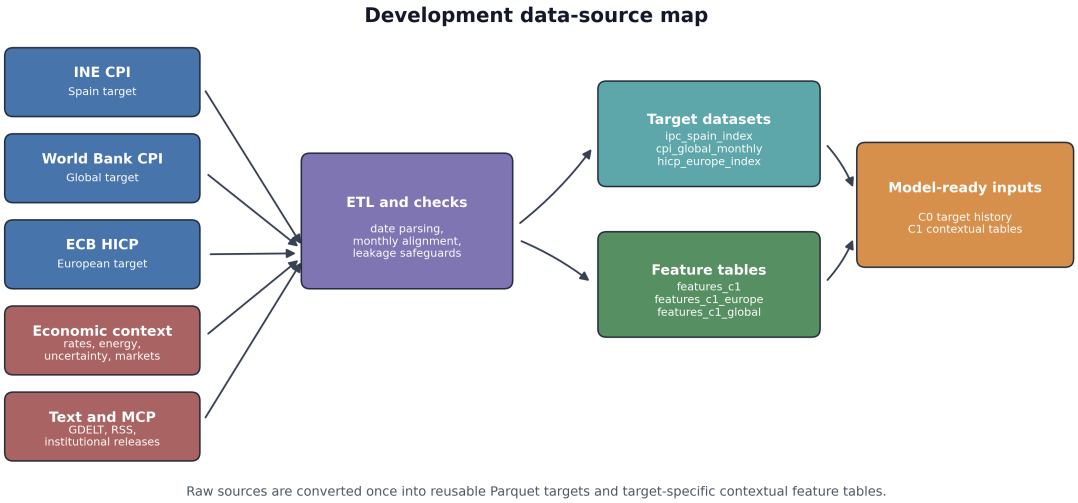

Figure 7.1.: Data source map and processed datasets

7.1.1. Spain CPI target

The Spain target is built from an INE Excel file containing the Consumer Price Index and its ECOICOP component groups. ECOICOP is the European classification used to organise household consumption categories, which allows the general CPI index to be accompanied by groups such as food, housing, transport, health, and education. The relevant section of the file is extracted, the monthly headers are parsed into dates, and the series is sorted chronologically. The main target used in the forecasting experiment is the general index, stored as `indice_general`. The processed file also preserves several CPI component groups, including food, housing and energy-related categories, transport, health, education, and other consumption groups.

The ingestion process includes basic quality checks. The script verifies that the monthly index has no duplicated dates, checks whether any months are missing, and reports missing values by column. This matters because the Spain CPI series is the longest and most detailed target in the project. If the initial index construction were wrong, all later comparisons for Spain would be affected.

The processed Spain file contains 289 monthly observations and 14 features: the general CPI index and thirteen ECOICOP-based consumption groups. For the modelling comparison, the general index is used as the target. The additional CPI groups are useful for understanding the structure of the original source and could support future extensions, but the current experiment focuses on the headline index to keep the comparison consistent

with the Global CPI and European HICP targets.

7.1.2. Global CPI target

The Global CPI target required a different construction process because the source does not provide a single ready-to-use world inflation series in the same way as the Spain and Europe sources. The project uses the World Bank global inflation dataset, specifically the monthly headline CPI index sheet [29]. This source contains country-level CPI indices for many countries.

To construct the target, the project first calculates the year-on-year inflation rate for each country by comparing each monthly CPI index with its value twelve months earlier. After that, it computes the cross-sectional median across the countries with available data for each month. This median aggregation is a project-level construction, not an official World Bank aggregate, but it follows a standard robust-summary logic: the middle available country value is less affected by extreme national observations than a simple average. The result is a monthly global inflation rate, stored as `cpi_global_rate`, which can be compared with international contextual signals.

The processed Global CPI file contains 276 monthly observations from 2002-01 to 2024-12. It is especially useful for testing whether broader institutional, commodity, financial, and supply-chain variables add value. Unlike Spain CPI, which is a national target, Global CPI is naturally closer to variables such as global economic policy uncertainty, commodity indicators, the US dollar index, market volatility, Treasury yields, the Federal Funds Rate, supply-chain pressure, and geopolitical risk.

7.1.3. European HICP target

The European target is the Euro Area all-items Harmonised Index of Consumer Prices, obtained from the ECB SDMX data service. The selected series is monthly, not seasonally adjusted, and expressed as a price-level index where 2015 is the reference year. In practice, this means that the index is normalised around the 2015 price level, so later values show how consumer prices have evolved relative to that reference. The processed file stores the variable as `hicp_index`.

The construction process downloads the raw ECB series, parses the monthly dates, filters the 2002-01 to 2024-12 period, checks for missing months, and verifies that the 2015 values are coherent with the index base. A diagnostic plot is also generated as part of the ingestion process to visually inspect the time path and the high-inflation period.

This target is important because it is the most directly aligned with several contextual sources used in the project. European Central Bank rates and communications, European policy uncertainty, Eurozone economic sentiment, energy prices, and GDELT or ECB-related text signals are all economically closer to European HICP than to the Global CPI or Spain CPI targets. For that reason, the European target becomes a key case for testing whether institutional and semantic context can improve the forecast.

The output of this first stage is a set of consistent monthly target datasets. The contextual feature tables are developed in Section 7.3, where their contents can be explained together with the signal pipelines that produce them.

7.2. DATA UNDERSTANDING AND EXPLORATORY ANALYSIS

Once the target series and feature tables had been constructed, the next development task was to understand their behaviour before modelling. This exploratory phase was important because inflation forecasting is not a neutral modelling exercise where data can be passed directly into algorithms. The series contain trend, seasonality, structural changes, and periods of very different volatility. If these properties are not analysed, the comparison between models becomes harder to interpret.

The exploratory analysis was organised around five types of diagnostics: visual inspection of the series, year-on-year inflation behaviour, stationarity tests, seasonality decomposition, and autocorrelation analysis. In addition, the 2021–2024 evaluation period was inspected as a sequence of economic regimes. This helped connect the technical model results with the economic context of the period.

7.2.1. Visual behaviour of the target series

The first step was to plot the three target series and inspect their temporal evolution. Spain CPI and European HICP are price-level indices, so their behaviour is characterised by a long-term upward trend. The Global CPI target is different because it is already constructed as a year-on-year inflation rate. This difference matters because the index-level targets focus on forecasting the evolution of a price index, while the Global CPI target focuses directly on forecasting an inflation-rate series.

The visual exploration confirmed that the 2021–2024 evaluation window is especially demanding. It includes the post-pandemic recovery, the 2022 inflation and energy shock, and the later normalization period. This means that the test window is not a calm continuation of the historical training period. It contains exactly the type of instability that makes inflation forecasting difficult and that gives contextual signals a possibly very im-

portant role.

Table 7.2.: Exploratory summary of the target series

Series	Period	Peak inflation	Main exploratory observation
Spain CPI	2002-01 to 2026-01	10.77% YoY in 2022-07	Strong post-2021 acceleration followed by normalization; national CPI shows clear shock behaviour during 2022.
Global CPI	2002-01 to 2024-12	9.64% in 2008-09	Broad inflation rate with international shock behaviour; less tied to a single domestic seasonal pattern.
European HICP	2002-01 to 2024-12	10.62% YoY in 2022-10	Strong European inflation shock, closely aligned with energy and monetary-policy context.

Figure 7.2 presents the three targets in a year-on-year inflation view so their shock behaviour can be compared on the same visual scale. This view is used only for exploratory comparability: the modelled targets for Spain CPI and European HICP remain the original level indices, while Global CPI is modelled as the constructed year-on-year rate.

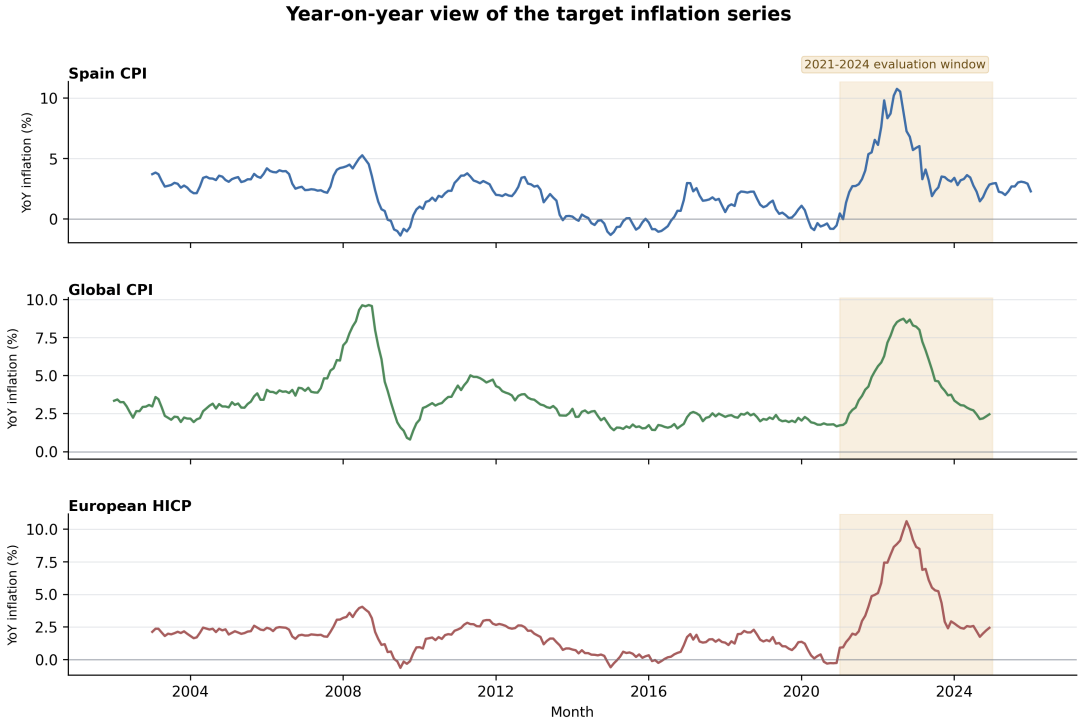

Figure 7.2.: Year-on-year exploratory view of the three target series

7.2.2. Training and evaluation periods

The visual exploration also helped verify that the temporal split was appropriate. The initial training window from 2002-01 to 2020-12 contains several economic cycles, including the financial crisis, low-inflation periods, and the pre-pandemic years. The evaluation period from 2021-01 to 2024-12 is shorter but more difficult, because it concentrates the inflation shock and the later adjustment phase.

This split makes the evaluation stricter. The models are tested on months that resemble the training period and on a period in which inflation moved sharply. This is important for the CDIA contribution because it tests the models under conditions where a simple extrapolation of the past may fail and where contextual variables such as energy prices, policy rates, uncertainty, or institutional communication may become more relevant.

7.2.3. Regime analysis

The exploratory analysis divided the evaluation period into three main regimes: a stability phase in 2021, a shock phase in 2022, and a normalization phase from 2023 to 2024. This segmentation was first inspected on Spain CPI, where the year-on-year inflation rate clearly shows the transition between phases. During 2021, Spanish inflation was already increasing but had not yet reached the 2022 peak. In 2022, the mean year-on-year rate rose sharply, reaching a maximum of 10.77%. In 2023 and 2024, inflation decelerated but remained part of a post-shock adjustment period.

Table 7.3.: Spain CPI regime summary during the evaluation period

Regime	Period	Mean YoY	Max YoY	Months
Stability and reopening	2021-01 to 2021-12	3.09%	6.55%	12
Inflation and energy shock	2022-01 to 2022-12	8.40%	10.77%	12
Normalization	2023-01 to 2024-12	3.17%	6.03%	24

This regime analysis is useful later in the results chapter. Stronger contextual performance during the shock phase would support the idea that external information helps when the historical target series alone is insufficient. Weaker contextual performance during stable phases would indicate that, when inflation is calmer, extra signals can add noise and reduce the quality of the forecast.

Figure 7.3 shows the regime interpretation. Year-on-year inflation is displayed with shaded areas for the 2021 stability period, the 2022 shock period, and the 2023–2024 normalization period.

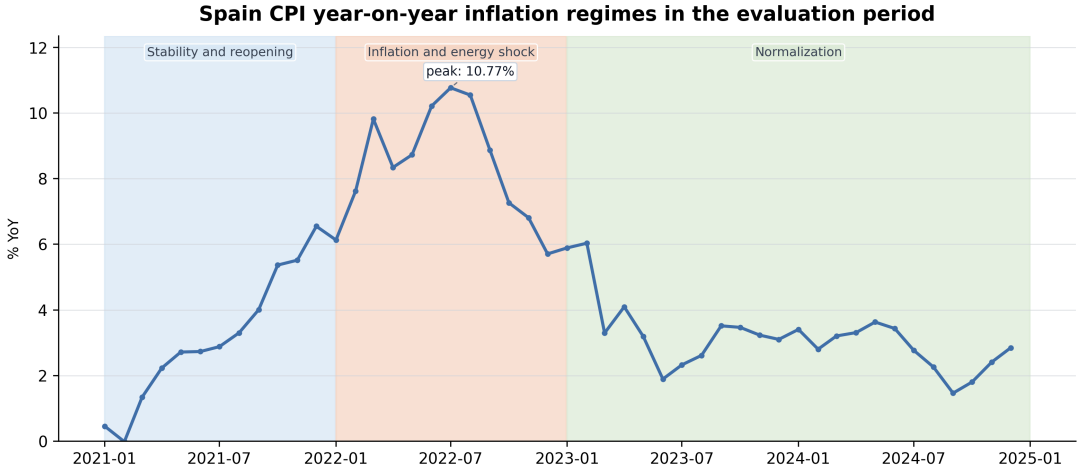

Figure 7.3.: Inflation regimes used to interpret the evaluation period

7.2.4. Stationarity and seasonality

The stationarity and seasonality diagnostics were used to decide how the models should be interpreted and which baselines were necessary. Spain CPI and European HICP are price-level indices, so they show persistent trend behaviour. This makes differencing and seasonal modelling important for the statistical baselines. It also explains why a seasonal naive benchmark is a meaningful reference for MASE: monthly inflation indicators often contain annual patterns that a simple lag-12 rule can capture surprisingly well.

The Global CPI target behaves differently because it is already a year-on-year rate computed from many countries. Its aggregation reduces some country-specific seasonal effects. This difference helps explain why the statistical modelling strategy is not identical across the three series. Spain and Europe naturally motivate seasonal models such as SARIMA, while Global CPI is closer to a broader rate series where dynamic order selection and non-seasonal ARIMA structures may be more competitive.

The seasonality decomposition and autocorrelation analysis also supported the inclusion of multiple model families. Classical models are expected to perform well when persistence and seasonal structure dominate. Deep learning models are included to test whether locally trained neural architectures can capture additional nonlinear behaviour. Foundation models are included because they may transfer useful patterns from pretraining, especially at longer horizons where local structure alone may not be enough.

7.2.5. Context exploration

In the exploratory stage, I also inspected the relationship between each target series and the contextual variables. The purpose was to check whether the external signals had an economic connection with the inflation series before using them in the forecasting experiments. Energy variables, policy rates, uncertainty measures, market indicators, and text-derived signals were examined as possible sources of forecasting context.

One important observation was that high correlation in levels can be misleading. For example, a contextual variable can rise during the same period as inflation because both are affected by the 2022 shock, but that does not mean the variable predicts month-to-month inflation changes. This issue was especially relevant for Spain, where some broad European signals were related to the inflation level but did not necessarily improve the forecast. For that reason, contextual signals are treated as hypotheses to test, not as evidence by themselves.

For European HICP, the link between the target and several contextual variables was more direct. The European Central Bank influences monetary conditions in the Euro Area, energy prices were central during the 2022 inflation shock, and Eurozone sentiment or market expectations describe the same economic environment as the target series. For Global CPI, the relevant context was broader and more international: global uncertainty, commodities, financial conditions, supply-chain pressure, geopolitical risk, and central-bank communication. This reinforced the decision to use target-specific feature tables for each series.

7.2.6. Development outcomes

The exploratory analysis produced three practical outcomes. First, it confirmed that the evaluation period is economically meaningful and difficult, especially because of the 2022 inflation shock. Second, it supported the use of strong classical baselines, including seasonal models and a seasonal naive reference, because the series contain persistence and annual structure. Third, it shaped the framing of contextual signals: external information may help when it is aligned with the target, available at the right time, and connected with the economic regime. When a variable only moves in the same general direction as inflation, without anticipating its monthly changes, it is treated with caution because it can introduce noise rather than useful forecasting information.

The exploratory phase shaped the later modelling decisions and provided the basis for interpreting both positive and negative results. The models are therefore evaluated against the actual behaviour of the data, with particular attention to whether complexity and contextual information are justified by the evidence.

7.3. CONTEXTUAL SIGNAL LAYER

After the exploratory analysis, the next development task was to build the contextual layer used by the `C1` experiments. This layer converts external economic information into monthly variables that can be aligned with each target series. Its implementation separates structured indicators, such as rates or commodity prices, from semantic signals extracted from news and institutional communications. The resulting feature tables remain target-specific because Spain CPI, Global CPI, and European HICP are connected to different economic environments.

7.3.1. MCP acquisition and extraction flow

The Spain and Europe pipeline is organised around a Model Context Protocol (MCP) server implemented with FastMCP, a Python framework for exposing functions as MCP tools [37]. A client communicates with the server through standard input and output. The server exposes tools to acquire Global Database of Events, Language, and Tone (GDELT) events for Spain, retrieve official Really Simple Syndication (RSS) publications from the European Central Bank (ECB), the Spanish National Statistics Institute (INE), and Banco de España, and query the documents stored during the process. The documents are normalised and stored in MongoDB, with their publication date, source, URL, text, and processing status.

The GDELT branch uses the event database directly [38]. Events are filtered by their connection with Spain and aggregated by month into average tone, average Goldstein score, and article volume. The Goldstein score is a GDELT event-impact measure: positive values indicate more cooperative events, while negative values indicate more conflictual events. These variables are already numerical, so they do not require language-model processing.

The official publications follow a different path. Their text is processed locally through Ollama [39], using Qwen3-4B [40] for inference. In this context, inference means that the model is not trained or fine-tuned during the project; it reads each publication and returns a structured interpretation. The output is validated against a predefined Pydantic schema [41], which specifies the expected variables and allowed ranges. Each publication is converted into decision type, magnitude, tone, shock score, uncertainty score, and topic. This schema keeps the extraction consistent across documents and makes the resulting variables usable in the forecasting pipeline.

Figure 7.4 shows the contextual-signal flow from external sources to the target-specific feature tables. It distinguishes the quantitative GDELT branch, the semantic extraction branch, the structured economic branch, the

monthly aggregation step, and the final temporal alignment before forecasting.

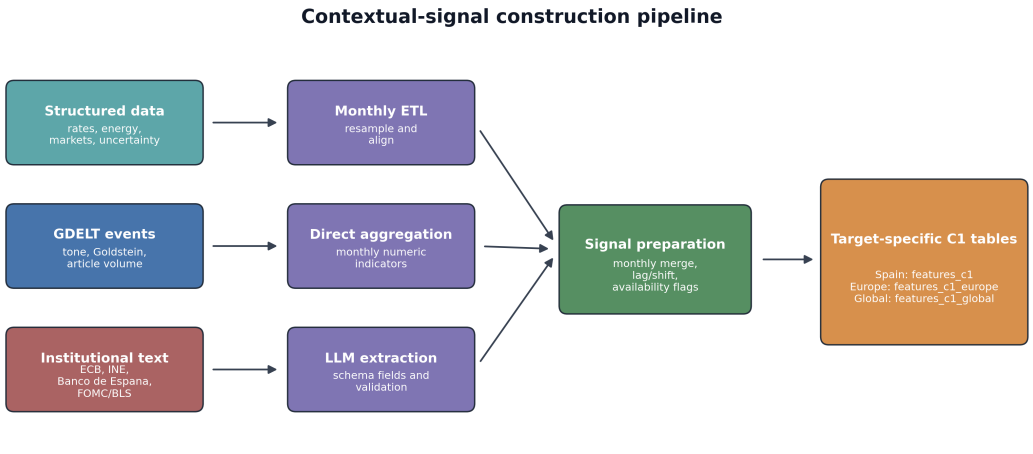

Figure 7.4.: Contextual-signal construction pipeline

Figure 7.5 presents the MCP architecture used for the semantic layer. It shows the host/client interaction with the FastMCP server, the JSON-RPC communication channel, the exposed acquisition tools, MongoDB storage, local Ollama/Qwen extraction, Pydantic validation, and the construction of monthly feature tables.

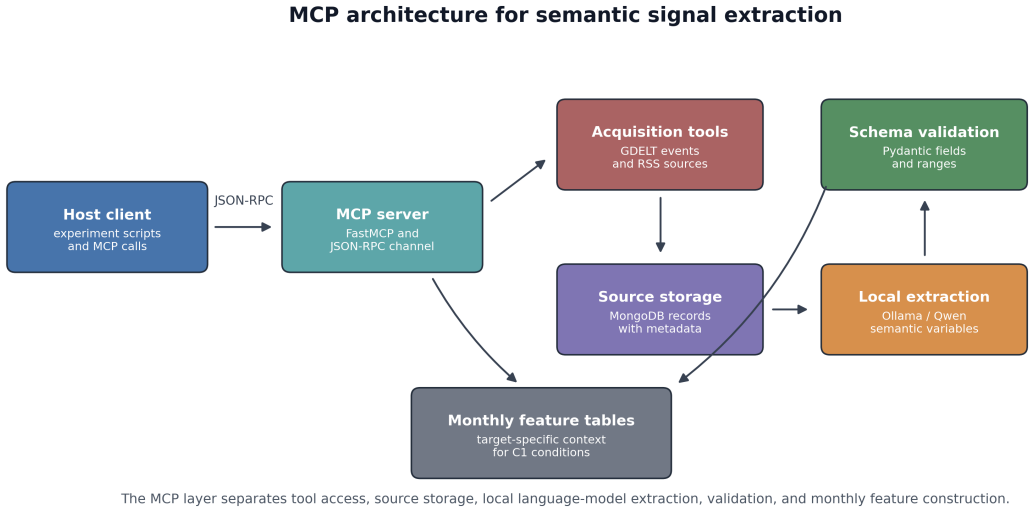

Figure 7.5.: MCP architecture for semantic signal extraction

7.3.2. Spain and European feature tables

For Spain, the script `news_to_features.py` aggregates the semantic layer between 2015 and 2024. ECB publications provide the monthly shock score, uncertainty score, and dominant tone. INE publications provide a

surprise score and the main topic. GDELT contributes monthly tone, Goldstein score, and article count. The script also derives a numerical ECB tone, a cumulative monetary-policy stance, three- and six-month moving averages for GDELT tone, and an availability flag that separates months without semantic history from genuinely neutral observations.

The semantic variables are merged with the structured Spain features to produce `features_c1.parquet`. European HICP uses the same semantic base where it is economically relevant, combined with European variables such as ECB rates, Brent and TTF energy prices, European policy uncertainty, Eurozone economic sentiment, five-year breakeven inflation, and the EUR/USD exchange rate. This process produces `features_c1_europe.parquet`. The European table keeps the contextual selection compact and aligned with the institutional environment of the target.

7.3.3. Global feature table

The Global CPI pipeline uses a broader set of structured indicators and a separate semantic extractor. Its institutional table combines global uncertainty, commodity prices, the dollar index, market volatility, United States Treasury yields, the Federal Funds Rate, supply-chain pressure, geopolitical risk, Brent prices, and the ECB deposit facility rate. The values are aligned to monthly frequency and transformed into lagged, differenced, and smoothed variants when appropriate.

The semantic branch collects official Federal Open Market Committee (FOMC) statements [42], European Central Bank (ECB) monetary-policy statements [43], and United States Consumer Price Index (CPI) releases [44]. A source-specific extraction schema converts these documents into nine numerical variables. For the Federal Reserve and ECB documents, the variables are hawkishness, surprise, and forward guidance. For United States CPI releases, the variables are surprise, inflation direction, and component pressure. The monthly semantic table is stored as `mcp_signals_global.parquet`. The current Global CPI experiments use `features_c1_global_institutional.parquet`; the semantic table is preserved as an additional pipeline output for a later combined-context experiment.

All semantic signals are shifted by one month before they are merged into the forecasting tables. A signal produced during month t is therefore available to the model from month $t+1$. This temporal rule is applied inside the feature-building scripts and preserves the information boundary required by the rolling-origin evaluation.

Table 7.4.: Contextual feature files produced during development

File	Scope	Observed size	Main content
`news_signals.parquet`	Spain / Europe	120 x 16	Monthly GDELT indicators and structured signals extracted from ECB and INE publications.
`features_c1.parquet`	Spain CPI	282 x 34	Spain target and structured variables enriched with the semantic layer.
`features_c1_europe.parquet`	European HICP	276 x 15	European institutional, energy, market, and selected semantic signals.
`features_c1_global_institutional.parquet`	Global CPI	276 x 31	Global institutional, commodity, financial, uncertainty, supply-chain, geopolitical-risk, and energy variables.
`mcp_signals_global.parquet`	Global CPI	276 x 9	Nine monthly semantic variables extracted from FOMC, ECB, and United States CPI releases.

The output of this stage is a set of reproducible contextual tables ready for model execution. Each table preserves the economic scope of its target and the temporal availability of its signals. The next development step is the execution of the classical, deep-learning, and foundation-model experiments under the corresponding `C0` and `C1` conditions.

7.4. MODEL EXECUTION

The model layer was developed as a set of comparable forecasting experiments. Each script loads the corresponding target or feature table, generates forecasts from successive origins, stores the prediction records, and computes the evaluation metrics. The common output format preserves the model name, forecast origin, forecast date, horizon, observed value, predicted value, and error. This makes it possible to compare models after execution without mixing training logic with result analysis.

7.4.1. Classical and deep-learning baselines

The classical layer provides the main reference for the experiment. It includes a seasonal naive benchmark, ARIMA, SARIMA, and SARIMAX variants [5], together with AutoARIMA [6]. The ARIMA, SARIMA, and SARIMAX rolling scripts use an expanding training window with specifications selected during the preliminary analysis and kept fixed across the evaluation period. AutoARIMA is evaluated separately as a dynamic variant: it recalibrates the statistical order at each forecast origin. This comparison shows whether repeated order selection adds value for each target series. Spain CPI and European HICP retain a stronger seasonal component, while

Global CPI is modelled as an aggregated year-on-year rate.

The local neural layer uses LSTM [7], N-BEATS [8], and N-HiTS [9] through the NeuralForecast library. These scripts use a 24-month input window, standard scaling, a fixed random seed, and a maximum of 300 training steps. Training the neural models at every monthly origin would add a large computational cost, so, for this project, their rolling evaluation was set up to use only quarterly origins. The four forecast horizons remain unchanged. This difference in origin frequency is retained when the neural results are interpreted alongside the monthly classical and foundation-model runs.

7.4.2. Foundation-model matrix

The foundation-model layer evaluates Chronos-2 [14], TimesFM [13], and TimeGPT [12]. Each model is first executed under the `C0` condition, using the historical target series as its input. The contextual scripts then evaluate the relevant `C1` variants for Spain CPI, Global CPI, and European HICP. This creates a model matrix that separates the effect of pretrained forecasting knowledge from the contribution of external information.

The implementation of context depends on the interface offered by each model. Chronos-2 accepts covariates natively and generates quantile forecasts, including the median prediction and the 10th and 90th percentiles. TimesFM is combined with a Ridge correction layer [45] fitted at each origin. This layer is a regularised linear adjustment: the pretrained forecast provides the base value, and the contextual variables estimate a controlled additive correction. Before this correction is calculated, the contextual variables are standardised using only the information available at that origin. TimeGPT receives historical and future covariate tables through the Nixtla API [46], which is the external service used to send the target series, the forecast horizon, and the exogenous variables to the TimeGPT model. When the future value of a contextual variable is unknown, the last value available at the forecast origin is propagated through the requested horizon. The TimeGPT scripts also provide a reduced test mode to verify the configuration before running the complete set of API calls.

Table 7.5.: Execution strategy for the model families

Family	**Models**	**Execution strategy**
Classical baselines	Seasonal naive, ARIMA, SARIMA, SARIMAX, AutoARIMA	Expanding-window evaluation with monthly origins and target-specific statistical specifications.
Local neural baselines	LSTM, N-BEATS, N-HiTS	NeuralForecast training with quarterly origins, a 24-month input window, and shared forecast horizons.
Foundation models	Chronos-2, TimesFM, TimeGPT	Monthly rolling evaluation under univariate and compatible contextual conditions, using model-specific covariate integration.

Figure 7.6 summarizes the execution layer. It shows the common input tables, the three model families, the C0 and compatible C1 branches, and the shared prediction and metric outputs.

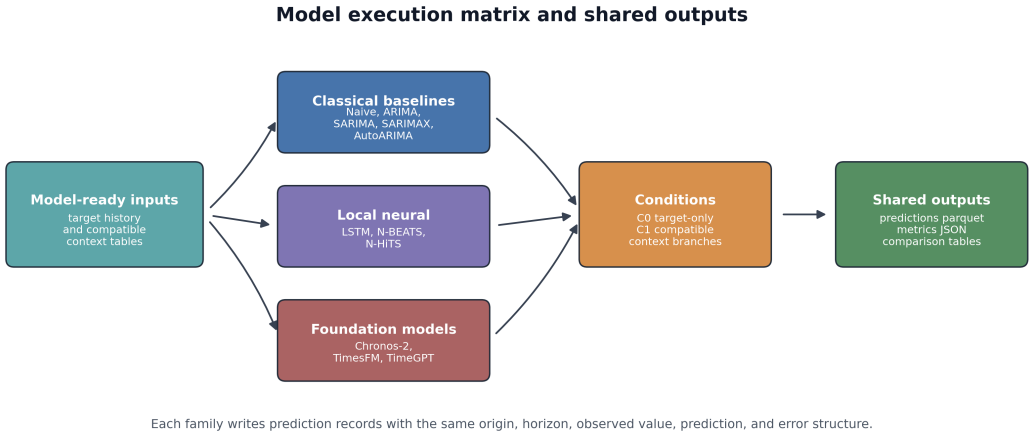

Figure 7.6.: Model execution matrix and shared outputs

7.4.3. Execution controls and outputs

Every final comparison uses executions that satisfy the temporal-availability rule defined in the methodology. When the future path of a contextual variable is unknown, the implementation carries forward the last value available at the forecast origin. The scripts also apply origin-specific scaling before Ridge corrections and use shared integrity checks to detect temporal misalignment, duplicated predictions, and incomplete artifacts.

The scripts store detailed predictions as Parquet files and aggregated metrics as JSON files under `08_results`. This separation supports later comparisons by series, horizon, condition, and economic period. It also allows the result chapter to distinguish between a general model-family pattern and an isolated improvement at a particular horizon.

7.4.4. Reproducibility artifacts

The forecasting implementation is organised into numbered modules that follow the execution flow. The `01_etl` and `02_eda` modules prepare and inspect the datasets. The `03_models_baseline`, `04_models_deep`, and `06_models_foundation` modules execute the three model families. The `05_mcp_pipeline` and `05_mcp_pipeline_global` modules build the semantic signal layer, while `07_evaluation` and `08_results` consolidate the evidence used in the comparison.

The implementation uses a Python environment with pandas, NumPy, statsmodels, scikit-learn, PyTorch, Neu-

ralForecast, and the model-specific dependencies required by the foundation-model scripts. TimeGPT also requires an API key. Together with the stored Parquet predictions, JSON metric summaries, and automated integrity checks, this module structure makes it possible to inspect the origin of each reported result.

7.5. CONSOLIDATED RESULTS

The final stage of the development consolidated the prediction records and metric summaries into a common comparison table. The first reading of the results focuses on the twelve-month horizon, where the differences between the three target series are clearest. Table 7.6 separates the model family, the selected model, the `C0` condition, the `C1` condition, and the relative effect of context when that effect can be computed from a valid same-model pair. This structure avoids mixing model selection with condition selection.

Table 7.6.: Audit-aware C0 and C1 comparison at the twelve-month forecast horizon

Series	**Family**	**Model**	**C0 condition**	**C0 h=12**	**C1 condition**	**C1 h=12**	**△ MAE**
Spain CPI	Classical	ARIMA	`C0`	MAE 1.541 MASE 1.097	–	–	n/a
Spain CPI	Foundation	TimesFM	`C0`	MAE 1.864 MASE 1.326	`C1_inst`	MAE 1.878 MASE 1.337	+0.8%
Global CPI	Classical	ARIMAX	`C0`	MAE 1.523 MASE 1.299	`C1_inst`	MAE 1.636 MASE 1.396	+7.5%
Global CPI	Foundation	Chronos-2	`C0_global`	MAE 1.338 MASE 1.141	`C1_inst`	MAE 1.143 MASE 0.976	-14.5%
Global CPI	Foundation	TimesFM	`C0_global`	MAE 1.233 MASE 1.052	`C1_inst`	MAE 1.191 MASE 1.016	-3.4%
European HICP	Classical	SARIMAX	`C0`	MAE 2.370 MASE 1.628	`C1_inst`	MAE 2.313 MASE 1.589	-2.4%
European HICP	Foundation	TimesFM	`C0`	MAE 2.014 MASE 1.384	`C1_full`	MAE 1.995 MASE 1.370	-1.0%

Note. Positive △ MAE values mean that adding context increased error; negative values mean that context reduced error. The Global Chronos-2 and TimesFM `C0_global` rows are regenerated runs that pass the target-integrity audit against `cpi_global_rate`. TimeGPT `C0_global` is unavailable, so no same-model Global TimeGPT effect is reported.

Figure 7.7 presents the Global CPI comparison at the twelve-month horizon after the foundation-target audit. The left panel answers one narrow question: what happens to the same ARIMAX model when institutional context is added. In that valid classical pair, `C1_inst` increases MAE from 1.523 to 1.636. The right panel has a different purpose: it ranks all verified Global CPI rows available at the same horizon, so that the corrected foundation runs can be compared with the classical references. In that panel, Chronos-2 and TimesFM can be read

as valid same-model foundation `C0_global`/`C1_inst` comparisons, while TimeGPT appears only as `C1_inst` because its clean `C0_global` run remains unavailable.

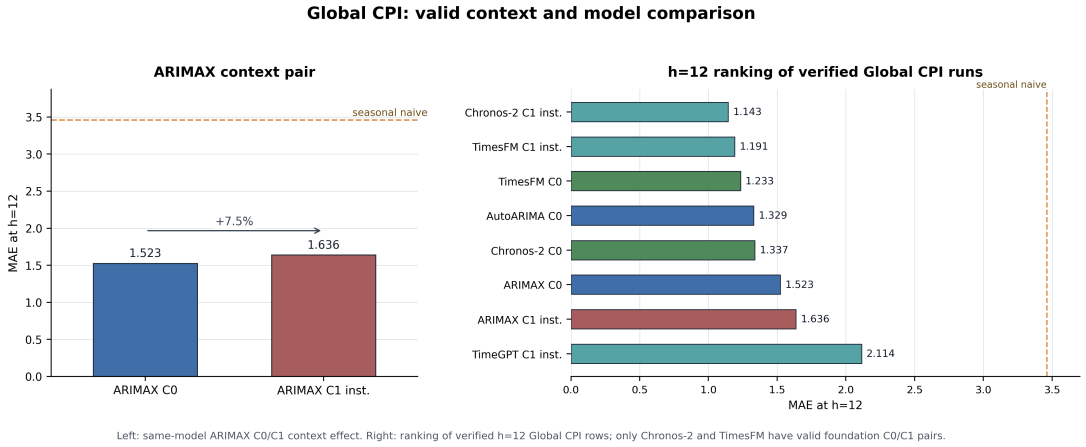

Figure 7.7.: Global CPI context effect and verified model ranking at the twelve-month horizon

Figure 7.8 shows valid same-model foundation context effects at the twelve-month horizon. Spain CPI, Global CPI, and European HICP are compared against their own `C0` references when those references pass the target-integrity audit. Global TimeGPT is omitted from the same-model comparison because its clean `C0_global` run remains pending.

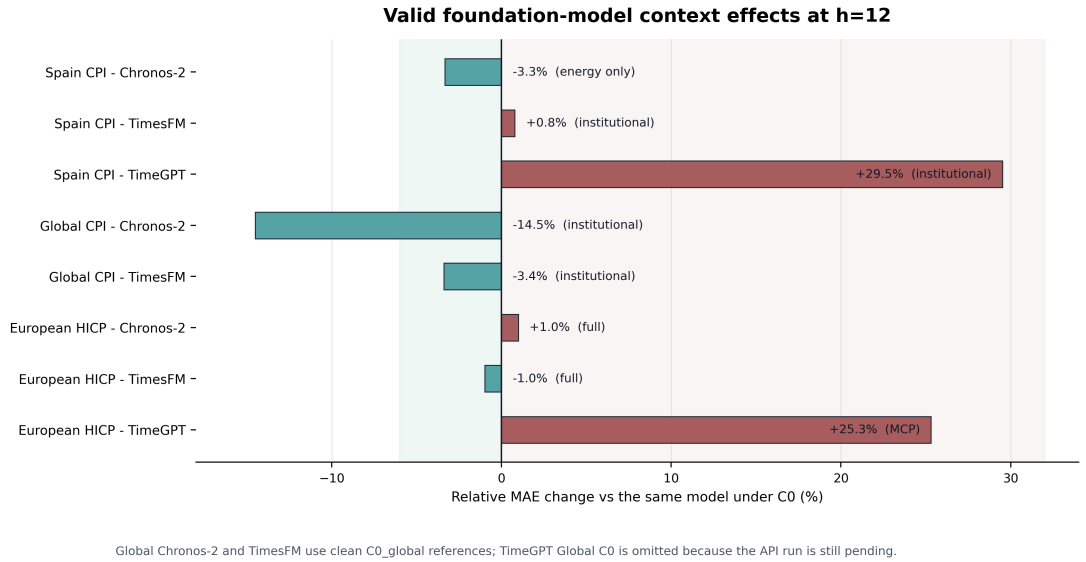

Figure 7.8.: Context effect on foundation time-series models at the twelve-month horizon

The comparison gives a more precise reading than a single “best configuration” summary. The baseline context effect is adverse in the complete classical pair, because Global ARIMAX performs worse with the institutional `C1`

layer than under C0 at twelve months. The foundation pairs are more mixed: Spain CPI does not improve with the selected TimesFM institutional condition, Global CPI improves with institutional context for both Chronos-2 and TimesFM, European HICP obtains a small additional gain from the full TimesFM contextual layer, and some model-context combinations degrade sharply.

Figure 7.9 shows the MAE profile across the four forecast horizons for the selected classical reference, selected foundation configuration, and seasonal naive benchmark in each target series. The curves make the horizon-dependence result visible, because the relative value of each model family changes as the forecast moves from one to twelve months.

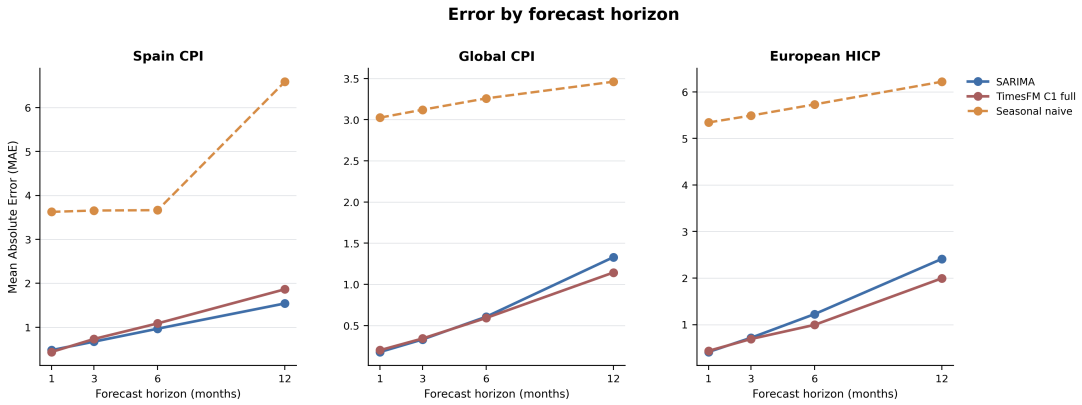

Figure 7.9.: Error-by-horizon comparison for the selected model configurations

Figure 7.10 shows representative one-step-ahead forecast paths from the stored prediction records. The Global CPI panel includes the Chronos-2 prediction interval because that model stores the 10th and 90th percentile forecasts, while the Spain CPI and European HICP panels show point forecasts only.

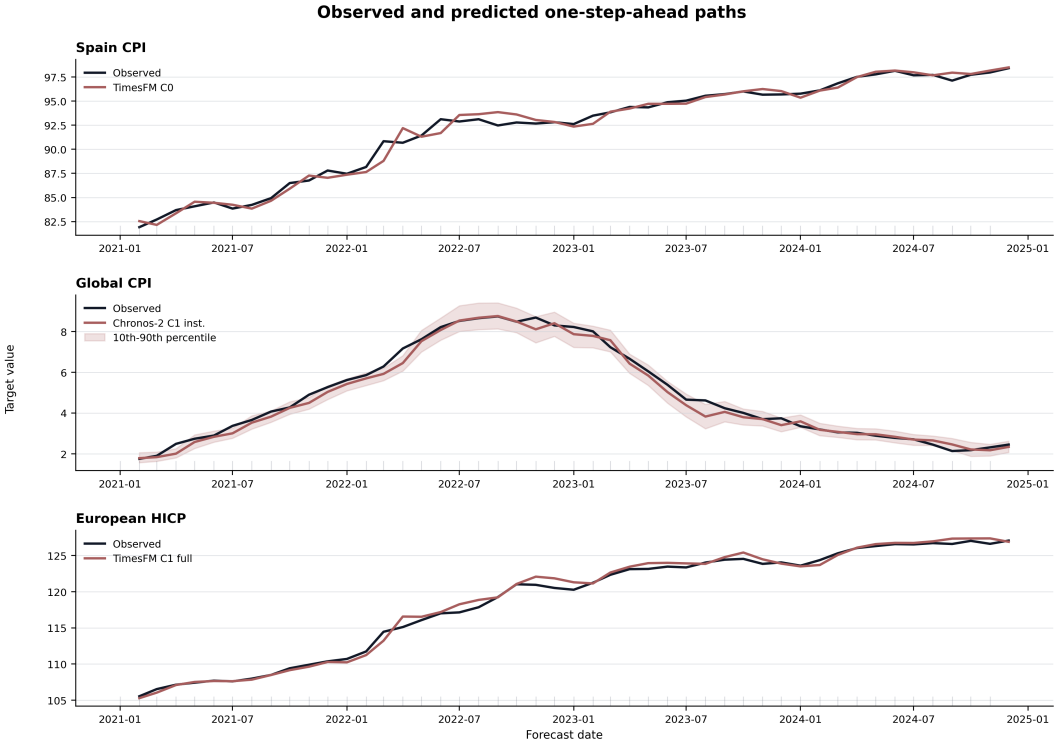

Figure 7.10.: Observed and predicted paths for representative one-step-ahead forecasts

7.5.1. Spain CPI

Spain CPI provides the clearest case in favour of maintaining strong classical baselines. N-BEATS obtains the lowest MAE at one and three months, with values of 0.359 and 0.670 respectively. ARIMA becomes the strongest option at six and twelve months, with MAE values of 0.966 and 1.541. The change in ranking is relevant because it shows that a model that performs well in the immediate forecast does not necessarily preserve that advantage as the forecast horizon increases.

The strongest TimesFM configuration for Spain is the univariate `C0` condition, with a twelve-month MAE of 1.864 and a MASE of 1.326. The institutional variant, `C1_inst`, reaches a twelve-month MAE of 1.878 and a MASE of 1.337. The difference is small, but it does not support an improvement from the available institutional signals. ARIMA remains clearly stronger at the same horizon, with a twelve-month MAE of 1.541. The Spain experiment therefore indicates that the effect of the available external context is neutral to slightly negative, while the historical structure of the national index remains more informative for medium-term and long-term prediction horizons.

7.5.2. Global CPI

Global CPI produces a different ranking. AutoARIMA is the strongest classical model across the four horizons, improving on the fixed ARIMA specification as the forecast extends further into the future. Its twelve-month MAE is 1.329, compared with 1.544 for fixed ARIMA. This is the clearest case in which recalibrating the statistical order at each origin is useful, as the aggregated international rate is exposed to broader changes in inflation dynamics.

Chronos-2 with institutional context already improves the clean `C0_global` run at medium and long horizons. A final Global variant was then evaluated to test whether the future path of the exogenous variables was limiting the original contextual result. Instead of carrying the last observed contextual value flat across all future steps, this variant uses a damped forward path estimated only from information available at the forecast origin. This produces the strongest Global row from the three-month horizon onward. At twelve months, its MAE falls to 1.065 and its MASE to 0.91. Relative to the clean Chronos-2 `C0_global` run, the MAE reductions are 23.6%, 14.1%, 17.5%, and 20.4% at the 1, 3, 6, and 12-month horizons respectively.

Table 7.7.: Global CPI performance across forecast horizons

Model	**MAE h=1**	**MAE h=3**	**MAE h=6**	**MAE h=12**
Fixed ARIMA	0.191	0.357	0.682	1.544
AutoARIMA	**0.179**	0.331	0.606	1.329
Chronos-2 `C0_global`	0.252	0.358	0.642	1.338
Chronos-2 `C1_inst`	0.200	0.342	0.591	1.143
Chronos-2 `C1_fwd`	0.193	**0.307**	**0.530**	**1.065**
TimesFM `C0_global`	0.210	0.376	0.648	1.233
TimesFM `C1_inst`	0.214	0.349	0.607	1.191
TimeGPT `C1_inst`	0.415	0.715	1.180	2.114

Note. Chronos-2 and TimesFM `C0_global` rows are included after passing the target-integrity audit against the Global CPI rate. TimeGPT `C0_global` is not available because the API-limited rerun did not complete.

The corrected Global foundation result now supports a direct same-model reading for Chronos-2 and TimesFM. Chronos-2 `C1_fwd` is the best twelve-month row in the table and improves over Chronos-2 `C0_global` by 20.4%. The original institutional Chronos-2 configuration also improves over `C0_global`, but less strongly. TimesFM provides a smaller contextual gain, reducing twelve-month MAE from 1.233 to 1.191. TimeGPT remains limited

to its `C1_inst` row until a clean `C0_global` API run is available.

Figure 7.11 gives the same Global comparison as a horizon profile and a twelve-month ranking. AutoARIMA remains preferable for the immediate forecast, while the contextual Chronos-2 configurations become more competitive as the horizon increases. The profile also shows that the clean TimesFM `C0_global` run is already strong and that its institutional version provides a smaller but still positive twelve-month improvement. TimeGPT is shown only in the ranking because its clean `C0_global` reference remains unavailable.

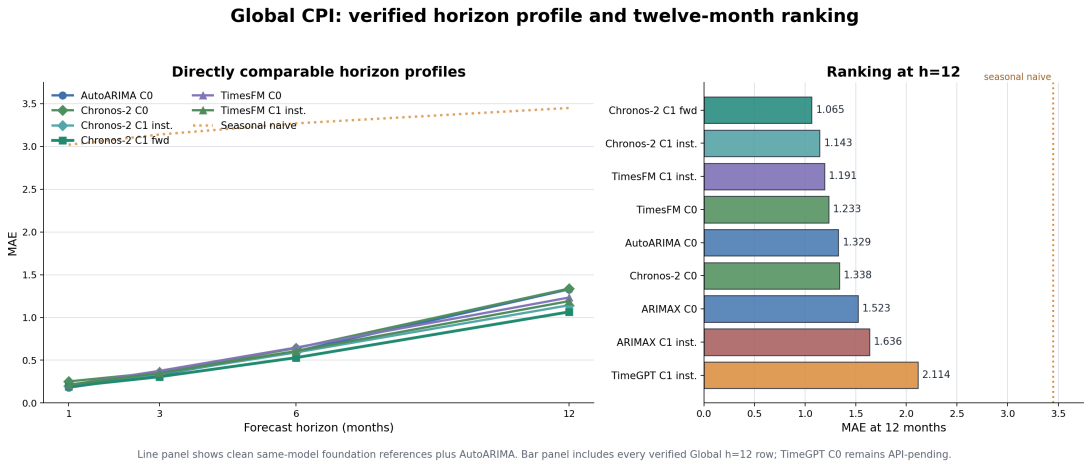

Figure 7.11.: Valid Global CPI horizon profile and twelve-month ranking

The horizon profile is consistent with the economic scope of the target: an aggregated international inflation rate is more naturally connected to global uncertainty, monetary-policy, commodity, and supply-chain variables than a domestic consumer-price index. In this case, the clean Chronos-2 and TimesFM comparisons indicate that the institutional signal layer adds value, especially when the future path of the contextual variables is represented as a forecast path rather than as a flat carry-forward value.

7.5.3. European HICP

European HICP presents a third pattern. TimesFM is the strongest model family for this series. Its univariate configuration already improves on SARIMA from the six-month horizon onward, reducing the twelve-month MAE from 2.411 to 2.014. The stricter Diebold-Mariano comparison confirms that this difference is statistically significant at six months ($p = 0.024$) and twelve months ($p = 0.002$) when the complete forecast paths are compared.

The full TimesFM contextual configuration combines institutional and semantic signals. It obtains the lowest twelve-month MAE in the European experiment, with a value of 1.995 and a MASE of 1.370. This represents a

17.3% reduction in MAE relative to SARIMA. The additional improvement over univariate TimesFM is smaller, at 1.0%. The comparison therefore indicates that the main gain comes from the foundation-model forecast, while the contextual layer provides a more limited refinement.

Table 7.8.: European HICP performance across forecast horizons

Model	**MAE h=1**	**MAE h=3**	**MAE h=6**	**MAE h=12**
SARIMA	0.413	0.717	1.226	2.411
AutoARIMA	0.376	**0.658**	1.147	2.510
TimesFM `C0`	**0.353**	0.697	1.035	2.014
TimesFM `C1_inst`	0.405	0.681	0.998	2.011
TimesFM `C1_full`	0.436	0.691	**0.995**	**1.995**

The ablation results also show that the contextual variables should be interpreted as a combined layer. Institutional signals improve the longer-horizon TimesFM result slightly, while the semantic-only configuration is weaker. Their combination produces the best twelve-month value. However, the additional gain over univariate TimesFM is small. The stricter forecast comparison therefore supports TimesFM as a strong foundation-model baseline for Europe, while treating the incremental contextual gain over `C0` as limited. Against SARIMA, TimesFM `C1_full` remains especially strong at the six-month horizon, where the Diebold-Mariano comparison is significant. These results are consistent with the European scope of the target, where monetary-policy variables, energy prices, uncertainty indicators, market expectations, and European institutional communication describe the same economic environment as the HICP series.

7.6. CONTEXT REPRESENTATION AND ROBUSTNESS ANALYSIS

After the main C0/C1 comparison, an additional robustness stage was carried out to understand why contextual variables helped in some cases but not in others. This stage does not replace the main experiment. It refines the interpretation of the contextual effect by checking whether the results depend on three elements: the way exogenous variables are represented, the future path assumed for those variables, and the statistical comparison used to support the observed differences.

The first diagnostic result is that several contextual variables were more related to the level of the price index than to the next change in the target. This distinction matters because a variable can move together with inflation during a shock period without necessarily improving a forecast. The effect was especially relevant for Spain and Europe, which are modelled as price-index targets. Global CPI is different because the target is already a year-on-year rate, so changes in global macroeconomic, commodity, and institutional variables are closer to

the quantity being forecast.

The second diagnostic result concerns the forward path of the exogenous variables. In several contextual configurations, unknown future values were represented by carrying the last observed value forward. This is a conservative and reproducible assumption, but it can make the contextual path too static during a shock. The Chronos-2 Global forward-path variant was introduced to test this point under the same rolling-origin protocol. The result shows that the future representation is a real modelling lever: the forward-path variant reduces MAE against Chronos-2 `C0_global` by 23.6%, 14.1%, 17.5%, and 20.4% at the 1, 3, 6, and 12-month horizons respectively. The strict Diebold-Mariano test gives marginal support at the one-month horizon ($p = 0.075$); the longer-horizon gains are large in MAE but not statistically significant with the available number of origins.

The third diagnostic result comes from the overlay variants. These variants use stored `C0` forecasts and add a validated contextual correction only when the correction improves a pre-2021 validation baseline. This prevents a contextual layer from being applied merely because it is available. For Spain, the selected correction did not beat the zero-correction validation baseline, so the validated and regime-gated outputs correctly remain identical to `C0`. For Global CPI, the regime-gated Chronos-2 overlay produces smaller effects than the forward-path model, but they are more statistically stable at the path level: the gains are significant at three and six months ($p = 0.019$ and $p = 0.039$) and marginal at twelve months ($p = 0.054$). For Europe, the overlays remain neutral to slightly adverse against `C0`.

Table 7.9.: Robustness variants used to interpret the contextual effect

Variant	**Purpose**	**Models or series**	**Main reading**
Strict paired comparison	Recompute C0/C1 differences by aligned origin, forecast date, and horizon using an HLN-adjusted Student-t Diebold-Mariano test.	Comparable stored prediction pairs.	Supports a more conservative reading of statistical significance.
Validated overlay	Apply a Ridge correction only if it improves a pre-2021 zero-correction baseline.	Chronos-2 and TimesFM where stored C0 forecasts are available.	Spain falls back to no correction; Global and Europe remain selective.
Regime-gated overlay	Apply the correction only in high-volatility regimes defined from the training period.	Chronos-2 and TimesFM for Spain, Global, and Europe.	Global Chronos-2 obtains significant small gains at h=3 and h=6.
Forward-path covariates	Replace flat future covariates with a damped forward path estimated at each origin.	Chronos-2 Global CPI.	Strongest contextual result, with 14–24% MAE reductions against C0.

Figure 7.12 summarizes the same interpretation numerically. It averages the MAE reduction of the selected contextual representation against its comparable C0 forecast across the four horizons. Positive values indicate that the contextual variant improves the univariate forecast, while negative values indicate that it worsens it.

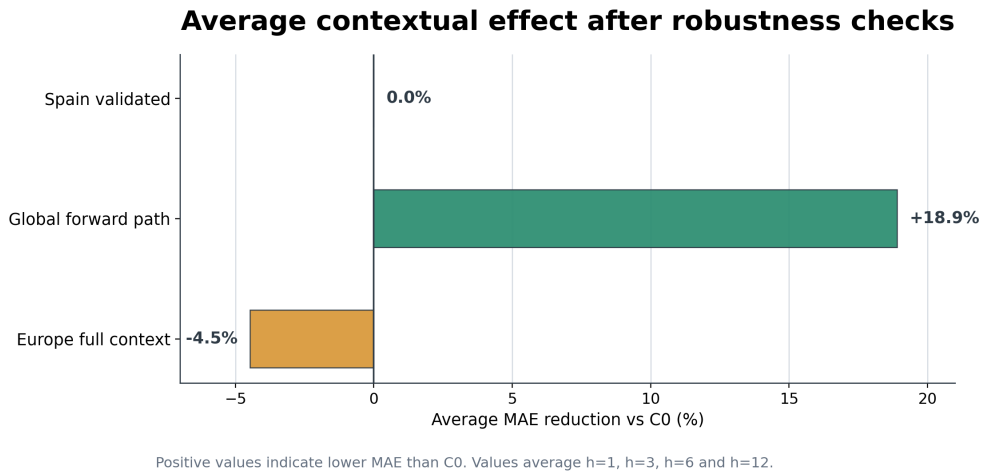

Figure 7.12.: Average MAE reduction versus C0 after robustness checks

This robustness stage also clarifies how the statistical tests should be read. The p-values reported here are forecast-comparison p-values, not coefficient p-values inside a model such as SARIMAX. They are used when two prediction records can be paired on the same origins, forecasted months, and horizons. For models whose role is mainly to provide benchmark error levels, such as the local deep learning baselines, the main evidence remains the MAE, RMSE, and MASE comparison. This keeps the thesis focused on comparable forecast evidence rather than forcing every model family into the same contextual protocol.

7.7. EXPLORATORY PROBES

After the consolidated results, two additional probes were run as a small bridge toward possible extensions of the project. They are not part of the main evaluation and are not used to select the final models of the project. Their purpose is more limited: to test whether two extensions of the current pipeline are technically feasible and whether they deserve further study. The first probe studies mixed-frequency modelling through Mixed Data Sampling (MIDAS) [47], a regression approach designed to combine variables observed at different frequencies without aggregating all information to the lowest-frequency series. The second probe studies whether a Nexus-inspired contextual reviser can adjust a strong TimesFM prior; Nexus [48] is an agentic forecasting framework in which specialised components analyse temporal and contextual information before producing or revising a forecast.

Table 7.10 summarizes the role of these probes. The table separates the objective of each test, the main numerical signal observed, and the interpretation used in the thesis. This helps keep the distinction clear between the controlled experiments reported in Section 7.5 and the exploratory tests used to motivate future work.

Table 7.10.: Exploratory probes beyond the main evaluation

Probe	**Purpose**	**Observed signal**	**Interpretation**
MIDAS monthly proxy	Test whether lagged contextual variables can help Spain CPI nowcasting.	MIDAS-ADL obtains MAE 0.4677 and MASE 0.333 at one month, slightly below ARIMA.	Small nowcasting benefit, but weak longer-horizon behaviour.
Daily MIDAS variants	Test whether daily WTI information adds high-frequency value.	Daily variants do not improve on ARIMA or on monthly-proxy MIDAS.	Daily oil prices are not the right high-frequency signal for this setup.
Nexus heuristic reviser	Test whether a bounded contextual correction can improve TimesFM.	The heuristic improves the three- and six-month TimesFM prior slightly and almost matches it at twelve months.	Feasible and interpretable, but only marginally useful.
Nexus LLM reviser	Test whether a zero-shot agentic reviser can correct the TimesFM prior.	The multi-agent run reaches 34% directional accuracy; Qwen3-8B reaches 51.3% in a single-call probe.	Zero-shot reasoning is not enough; a trained reviser would be needed.

Table 7.11.: MAE by horizon for the exploratory probes on Spain CPI

Model or probe	**h=1**	**h=3**	**h=6**	**h=12**
ARIMA (reference)	0.4781	0.6716	0.9660	1.5410
MIDAS-ADL (monthly proxy)	0.4677	1.0348	1.5707	2.7863
TimesFM prior (C0)	0.4364	0.7320	1.0866	1.8635
Heuristic reviser	0.4400	0.7284	1.0849	1.8634
Multi-agent reviser	0.4398	0.7402	1.0895	1.8639

The MIDAS probe evaluates whether high-frequency or lagged contextual information can improve Spain CPI forecasts. This is relevant for inflation forecasting because CPI is monthly, while market prices, news flow,

and institutional events may be available daily. The first implementation used a MIDAS-ADL specification with monthly proxy variables and Beta-polynomial lag weights. At the one-month horizon, the model obtains a MAE of 0.4677, slightly below the ARIMA reference value of 0.4781. The corresponding MASE values are 0.333 for MIDAS-ADL and 0.340 for ARIMA. This result suggests a small nowcasting benefit; in this context, nowcasting means estimating the very short-term CPI value, close to the next publication date, using the information already available before the official value is observed.

Table 7.11 shows why this result remains limited. The monthly-proxy MIDAS configuration improves the one-month forecast but loses competitiveness as the horizon increases. A second daily-frequency probe used daily West Texas Intermediate (WTI) crude-oil information, a light sweet United States crude-oil benchmark priced at Cushing, Oklahoma [49], in three variants: daily Beta-MIDAS with price levels, daily Beta-MIDAS with returns, and a Ridge-daily version with free lag weights. None of the daily variants improves on ARIMA or on the monthly-proxy MIDAS model. The returns version is slightly better than the daily-level version at twelve months, with MAE 3.4030 instead of 3.5465, but both are clearly weaker than ARIMA. The interpretation is that oil-price information, in this setup, does not provide the kind of high-frequency signal that improves monthly CPI forecasting. A more promising extension would use daily news, GDELT tone, or institutional event flow, closer to the type of mixed-frequency text signal studied in macroeconomic nowcasting work [50].

The second probe explores the Nexus-inspired reviser introduced above. The idea is to keep TimesFM as a frozen time-series prior and add a contextual correction layer. In this project, the first version was a bounded heuristic reviser over the TimesFM `C0` prior for Spain CPI. The numeric comparison in Table 7.11 shows that the heuristic revision only changes the TimesFM prior marginally. It slightly improves the three- and six-month forecasts and almost matches the prior at twelve months, while remaining weaker at one month. It also beats the Ridge-based TimesFM `C1_inst` configuration at all horizons, with relative gains between 0.8% and 2.4%.

The multi-agent version used a locally running Qwen3 model [40] through Ollama [39]. It produced structured reasoning traces for 47 of the 48 forecast origins, which shows that the pipeline is technically feasible. Its MAE profile does not improve consistently on the TimesFM prior or on the heuristic reviser. The post-hoc analysis explains why: the directional signal is weak. In the multi-agent run, directional accuracy is 34% on the decisive origins, with 13 correct and 25 wrong directional calls. A single-call Qwen3-4B reviser improves this to 48.7%, and a Qwen3-8B version reaches 51.3%, but neither model produces a MAE improvement over the TimesFM prior at the one-month horizon. The larger model also shows that model size is not the main bottleneck in this setup. The central problem is that a zero-shot reviser can produce plausible macroeconomic reasoning without learning when the foundation-model prior should be preserved.

These probes are therefore useful mainly as boundary tests. MIDAS shows a small nowcasting signal, but the available high-frequency commodity information is not enough for longer horizons. The Nexus-style reviser shows that contextual revision is technically feasible and interpretable, but zero-shot prompting is not enough to improve a strong time-series prior. Both probes motivate future work, while keeping the main conclusions anchored in the controlled experiments reported above.

7.8. CROSS-SERIES INTERPRETATION

The tables in the previous sections give a compact cross-series reading of the experiment. Spain CPI is led by classical autoregressive models at medium- and long-term forecasting horizons. Global CPI is the clearest positive case for contextual foundation forecasting, especially after the Chronos-2 forward-path variant. European HICP is best read as a foundation-model result: TimesFM is strong against SARIMA at medium and long horizons, while the extra contextual gain over TimesFM `C0` is small.

The results provide strong evidence for horizon dependence. Classical and dynamically selected statistical models remain competitive in immediate forecasts, while foundation configurations become more useful at longer horizons for Global CPI and European HICP. The contextual contribution is more selective. For Spain CPI, the validation guard prevents the contextual overlay from acting because it does not improve the pre-2021 validation baseline. For Global CPI, the forward-path Chronos-2 model shows the largest effect size, and the regime-gated overlay provides smaller but more statistically stable gains at the path level. For European HICP, context does not clearly beat the univariate foundation forecast, although the foundation model itself remains competitive against classical baselines.

The regime analysis adds a more nuanced reading of the contextual layer. External information is most useful when the target is exposed to broad economic shocks and when the model interface can represent the future path of that information. This is why the Global year-on-year rate benefits more clearly than the Spain and Europe price-index targets. Stable periods and weakly aligned signals can expose the model to unnecessary noise, while shock periods make the potential value of contextual information more visible. The effect therefore depends on the target representation, the forecast horizon, the economic period, and the technical way in which the context enters the model.

Figure 7.13 summarizes the valid context-effect pairs across horizons. The heatmap shows that context is not a monotonic improvement mechanism. Spain CPI mostly worsens with the selected institutional TimesFM variant before the validated no-op guard is applied; Global ARIMAX degrades with institutional context; Global Chronos-2 and TimesFM improve with institutional context at longer horizons; and European TimesFM im-

proves descriptively after the first horizon when the full contextual layer is used.

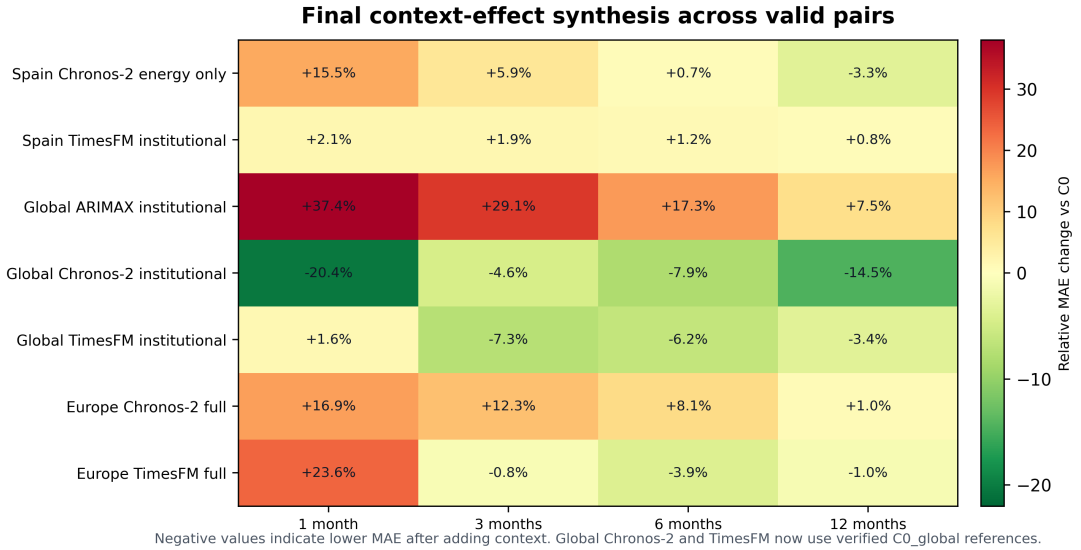

Figure 7.13.: Final synthesis of context effects across valid C0/C1 pairs

Figure 7.14 adds a signal-family ablation view for Chronos-2 in Spain CPI, European HICP, and Global CPI. The bars show the relative MAE change against the same model under C0 at six and twelve months. Negative values indicate an improvement. Global CPI has only the institutional Chronos-2 comparison, so it appears as a compact one-row panel rather than a full signal-family grid.

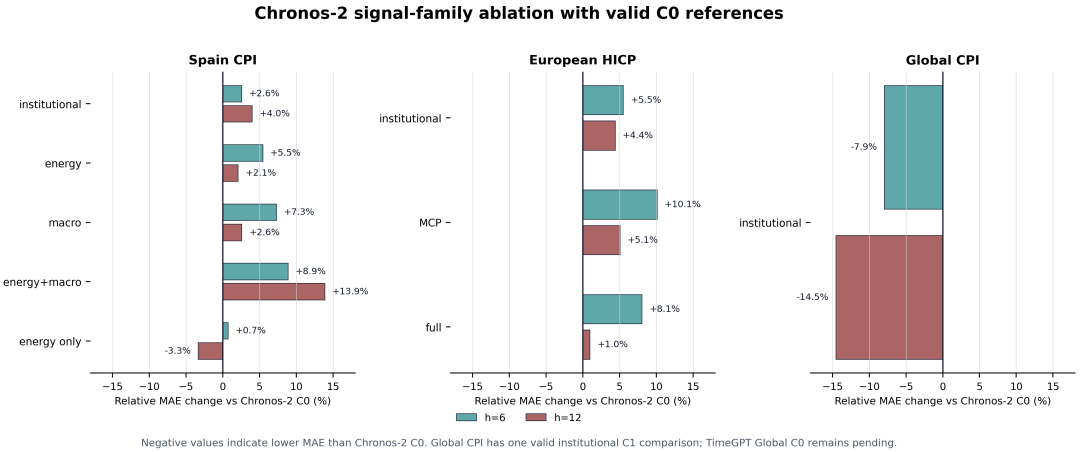

Figure 7.14.: Chronos-2 signal-family ablation for Spain CPI, Global CPI, and European HICP

The ablation clarifies why the contextual result is not transferable across targets. For Spain CPI, most Chronos-2 signal families are small or adverse, with the energy-only variant becoming useful only at the twelve-month horizon. For Global CPI, the institutional signal layer improves the six- and twelve-month Chronos-2 forecasts

against the clean `C0_global` reference, and the forward-path version strengthens that pattern. For European HICP, the complete Chronos-2 contextual layer remains slightly adverse overall, even though some shock-period comparisons point in a more favourable direction. The same modelling family therefore reacts differently depending on the target, horizon, period, and contextual information available for it.

7.9. CHAPTER SUMMARY

This chapter has described the practical development of the CDIA contribution, from target-series construction and exploratory analysis to contextual-signal extraction, model execution, robustness analysis, and consolidated evaluation. The final results show that model selection should consider the target series, the forecast horizon, the economic relevance of the available signals, and the representation of future contextual paths. The following chapters record the ethical implications, incidents, and final conclusions of the project.

8. ETHICAL ASSESSMENT

8.1. PROFESSIONAL RESPONSIBILITY

The ethical assessment of this project is part of the technical work itself. In engineering and data science, design decisions affect how information is produced, interpreted, and used. A model is therefore not only evaluated by its accuracy, but also by the way it handles uncertainty, communicates limitations, and avoids creating unjustified confidence in its outputs.

This responsibility is especially relevant in inflation forecasting. Inflation affects salaries, savings, business costs, public policy, investment decisions, market expectations, and citizens' trust in institutions. A forecast that is wrong, poorly explained, or presented with excessive confidence can influence how the public understand the economic situation. Even when the system is used only for analysis, its results can still shape interpretation.

Professional ethics provides a practical framework for this responsibility. Principles such as beneficence, nonmaleficence, autonomy, and justice are useful because they translate ethical awareness into concrete questions: does the system contribute to better economic understanding, does it reduce avoidable harm, can users interpret the results critically, and are the risks of misleading information handled fairly? The ACM Code of Ethics [51] expresses similar duties through the need to contribute to the public good, avoid harm, be honest and trustworthy, and communicate limitations clearly. In this project, these principles guide the way data sources, models, evaluation metrics, and conclusions are selected and presented.

8.2. MAIN ETHICAL RISKS

The first risk is unreliable information. Some variables used in the project come from official statistical or institutional sources, while others are derived from news, economic communication, uncertainty indicators, or semantic text-processing pipelines. If these inputs are false, biased, exaggerated, badly aligned, or manipulated, the forecast can become misleading. In inflation forecasting this risk is important because expectations matter. News about energy prices, monetary policy, geopolitical risk, or inflation pressure can affect markets and public interpretation. A model that reproduces unreliable information may indirectly amplify confusion.

The second risk is excessive dependence on the algorithm. Inflation forecasting is uncertain by nature. It depends on political decisions, international conflicts, energy shocks, monetary policy, consumption patterns,

expectations, and many other factors that cannot be fully captured by one model. If a forecast is treated as an oracle, users may stop questioning whether the result is plausible, whether the context has changed, or whether simpler explanations are more convincing. This creates a risk of dehumanization in the decision process: economic judgement is reduced to a numerical output, while the people affected by inflation and the uncertainty behind the forecast receive less attention.

The third risk is lack of transparency. Foundation models and artificial intelligence systems can appear more authoritative than they really are, especially when their internal behaviour is difficult to inspect. A reader may see a modern model name and assume that the result is automatically better than a classical statistical forecast. This is ethically problematic because novelty is not the same as validity. Transparency, in this context, means explaining what data are used, which model family produces the forecast, what horizon is being evaluated, and how much error remains. This idea is aligned with the NIST AI Risk Management Framework [52], which emphasizes validity, reliability, transparency, explainability, and accountability as central elements of responsible AI systems.

The fourth risk is methodological error. In forecasting, a technical mistake can become an ethical problem if it creates a false conclusion. Examples include using future information by mistake, aligning variables incorrectly, evaluating models on incomparable forecast dates, failing to scale variables before a correction layer, or reporting a result without checking whether the improvement is stable. These problems can make a model appear better than it really is. If the final thesis presented those results without verification, the reader would be misled.

8.3. MITIGATION IN THE PROJECT

The project mitigates the reliability risk by giving priority to traceable sources and by separating the signal families. Official inflation data, institutional datasets, market variables, uncertainty indicators, and text-derived signals are documented, aligned, and evaluated under different conditions. This separation makes it possible to see whether structured economic variables, semantic signals, or their combination actually improve the forecast. It also avoids assigning the same credibility to all information sources before they are tested.

The risk of algorithmic dependence is addressed by presenting the forecasts as analytical evidence. The thesis compares modern models against strong classical baselines and reports negative or neutral results together with positive ones. This point is important because it avoids technological hype. For example, the Spain CPI results show that adding contextual information to TimesFM does not improve the univariate configuration, and ARIMA remains stronger at the twelve-month horizon. Reporting this result clearly is part of the ethical

responsibility of the project.

Transparency is addressed through the structure of the evaluation. The results are separated by target series, forecast horizon, model family, and contextual condition. The use of MAE, RMSE, and MASE makes the error interpretation explicit, while Diebold-Mariano tests provide additional statistical support where comparable prediction records are available. The thesis also explains the difference between `C0` and `C1` conditions, so that the reader can understand whether a result comes from the base model or from the contextual layer. This communication strategy supports critical interpretation instead of blind trust.

The methodological risks are mitigated through the experimental protocol and the final audit described in Chapter 7 and Chapter 9. Rolling-origin backtesting is used to simulate a realistic forecasting situation, where each prediction only uses information available at the corresponding origin. Contextual variables are shifted or carried forward according to their temporal availability. Variables with different magnitudes are standardised before Ridge correction. The final audit corrected a limited set of implementation issues and regenerated the affected outputs before the results were consolidated. This process strengthens the credibility of the conclusions because methodological problems were identified, corrected, and documented.

The project also preserves human judgement as a necessary part of interpretation. The forecasts are useful only when they are read together with the economic context, the selected horizon, the source of the signals, and the limitations of each model family. This is consistent with the broader European approach to artificial intelligence regulation in the EU AI Act [53], which emphasizes risk management, transparency, and human oversight for AI systems that may affect important social or economic decisions. In the context of this project, human oversight means that the model supports reasoning about inflation, while the final interpretation remains a professional and economic judgement.

8.4. CONCLUSION

The ethical challenge of this project is to evaluate and present an inflation forecasting system responsibly. The main risks are unreliable information, excessive dependence on algorithmic outputs, lack of transparency, and methodological errors that could produce misleading conclusions. These risks are addressed through source traceability, model comparison against strong baselines, explicit evaluation metrics, rolling-origin backtesting, leakage prevention, feature scaling, human interpretation, and transparent reporting of negative as well as positive findings.

In this sense, the ethical value of the project is connected to its technical discipline. A careful forecasting ex-

periment does more than produce numbers: it helps readers understand what can be trusted, what remains uncertain, and where artificial intelligence adds value only under specific conditions. This is the role that professional responsibility plays in the project.

9. INCIDENTS

9.1. EXECUTION INCIDENTS AND SCOPE ADJUSTMENTS

This chapter records concrete events that occurred during the execution of the CDIA work and required a response. General limitations of the study are discussed in Chapter 10; the incidents below are included because they changed the way part of the project was executed, verified, or delimited.

9.1.1. Reduced origin grid for local neural baselines

The first execution incident appeared when the local neural baselines were prepared for the rolling-origin evaluation. The initial methodological idea was to keep all model families on the same monthly origin grid. In practice, retraining LSTM, N-BEATS, and N-HiTS at every monthly origin was too costly for the available computing resources, because each origin required a new training process.

The impact was a deviation from the fully shared origin protocol used by the classical and foundation-model experiments. The response was to keep the same target series and forecast horizons, but evaluate the local neural baselines only at quarterly origins. This preserved their role as useful neural references while avoiding a computational load that would have delayed the rest of the experiment.

The effect on the results is that the neural baselines are interpreted with caution and are not used as the main evidence for the final cross-series conclusions. They remain useful to compare locally trained neural models with statistical and foundation-model approaches, but their smaller origin set is explicitly acknowledged in the methodology and development chapters.

9.1.2. TimeGPT API execution constraints

The TimeGPT experiments depended on an external API key and on the call limits of the service. During execution, this created a practical constraint: some runs could not be treated like local scripts that can be repeated freely, because failed calls, free-tier limits, or large batches of origins would consume time and external API allowance.

The impact was mainly operational. The TimeGPT scripts required a reduced test mode to verify the configuration before launching the full set of calls, especially when exogenous covariates were involved. This response

reduced the risk of wasting API calls on incorrectly configured inputs and made the execution more controlled.

The effect on the project scope was that TimeGPT remained part of the model comparison, but it was handled as an externally constrained model rather than as a fully local baseline. This distinction matters for reproducibility and budget, although it does not change the interpretation of the reported TimeGPT results.

9.1.3. Limited temporal availability of semantic signals

The semantic layer also required an execution adjustment. The target series start in 2002, but the GDELT, RSS, and MCP-derived text signals used in the project were not available with the same complete historical coverage. In particular, the semantic signal layer was built from the period in which those sources could be collected and processed reliably, while earlier months did not contain genuine text-derived observations.

The impact was that missing semantic history could not be treated as ordinary zero values. Doing so would have created a false pre-2015 regime in which the model might interpret the absence of collected text as a meaningful neutral economic signal. The response was to add explicit availability handling: unavailable text signals were represented with neutral or missing values according to the model interface, and a `signal_available` variable was used to mark the period in which the semantic layer actually existed.

The effect on the results is that the contextual experiments remain temporally coherent, but the interpretation of MCP and GDELT signals is narrower than the interpretation of structured macroeconomic variables. The semantic layer is useful where it is available and aligned with the target, especially in the European and global experiments, but its shorter history is part of the reason why semantic context is not presented as a universally reliable source of improvement.

9.1.4. Exploratory extensions kept outside the main evaluation

During the final stage of the work, two additional extensions were tested: MIDAS-style mixed-frequency modelling and a Nexus-inspired contextual reviser. These tests were technically connected to the project, but they were introduced after the main experimental protocol had already been defined and executed.

The impact was a scope decision. Integrating those extensions into the main evaluation would have required a new protocol, more forecast origins, additional data-ingestion work, and, in the case of the contextual reviser, a real training objective linked to realised forecast errors. The response was to keep them as exploratory probes in Chapter 7, separate from the controlled comparison used for the final conclusions.

The effect is that the MIDAS and Nexus results support the future-work direction without changing the main claims of the thesis. They show that richer or learned context integration is technically promising, but the final conclusions remain based on the controlled experiments with the statistical, deep learning, and foundation-model families.

9.2. METHODOLOGICAL AUDIT OF CONTEXTUAL EXPERIMENTS

During the final review of the forecasting pipeline, the contextual experiments were audited with particular attention to temporal availability and feature scaling. This review identified two implementation issues in a limited subset of the foundation-model scripts. The affected executions were corrected and repeated before consolidating the results presented in Chapter 7.

The first issue affected one Spain Chronos-2 contextual script. The future path of two European Central Bank policy-rate variables was read from the complete processed dataset. In a rolling-origin evaluation, later policy-rate decisions are not known at the forecast origin. The corrected implementation carries forward the last observed value through the requested horizon. This keeps the experiment aligned with the information that would have been available in a real forecasting situation.

The second issue affected five TimesFM contextual scripts. Their Ridge correction layers combined variables with substantially different numerical scales without standardising them first. This made the corrections unstable and gave excessive influence to variables with larger raw magnitudes. The corrected implementation fits a scaler on the information available at each forecast origin and applies the transformation before calculating the Ridge adjustment.

9.3. VERIFICATION AND EFFECT ON RESULTS

The affected foundation-model executions were repeated after the corrections, and the downstream metrics, comparison tables, and Diebold-Mariano files were regenerated. An automated integrity script was also added to check temporal causality, duplicated predictions, origin grids, expected artifacts, and the handling of exogenous variables.

The corrections did not change the overall interpretation of the project, but they made the result more precise. For Spain CPI, the institutional TimesFM condition is now neutral to slightly negative compared with the univariate configuration. For Global CPI, the corrected TimesFM institutional condition improves and becomes a com-

petitive second foundation-model result at the twelve-month horizon, although Chronos-2 remains stronger. The European HICP conclusion is unchanged: the combined TimesFM contextual condition remains the best long-horizon configuration for that series.

This incident reinforced the importance of treating temporal alignment and scaling as part of the experimental design rather than as secondary implementation details. Preserving a clear record of the corrections also improves the traceability of the final conclusions.

10. CONCLUSIONS AND FUTURE WORK

10.1. CONCLUSIONS

This project has evaluated whether foundation time-series models and contextual economic signals improve monthly inflation forecasting, and under which conditions that improvement appears. The final answer is conditional: the value of modern models and external context depends on the scope of the target series, the forecast horizon, and the quality of the information available at the forecast origin.

The main synthesis is that inflation forecasting cannot be evaluated only by asking which model is newest or most complex. The useful cases for foundation models and contextual signals appear when the target is broad enough, or institutionally connected enough, for the external information to describe the same economic environment. Classical baselines remain essential because they show whether that added complexity produces a real forecasting gain or only a more elaborate modelling pipeline.

The seasonal-naive benchmark is also central to the interpretation. At the twelve-month horizon, only Global CPI achieves a MASE below one with the selected foundation-model configuration. The best Spain and European configurations still remain above the seasonal naive reference. This is not a weakness of the evaluation; it is one of its strengths. It shows that the experiment is not only reporting improvements over ARIMA, AutoARIMA, or SARIMA, but also checking whether those improvements beat a simple seasonal rule. In Spain and Europe, some gains over classical baselines are gains over baselines that are themselves worse than the seasonal naive benchmark. In Global CPI, the contextual Chronos-2 result is more meaningful because it crosses that threshold.

The research questions from Chapter 3 can therefore be answered as follows:

1. **RQ1.** Foundation time-series models improve on strong classical baselines only in part of the experiment. Chronos-2 gives the clearest improvement for Global CPI, especially when the future path of the contextual variables is represented explicitly, and TimesFM improves European HICP prediction relative to SARIMA. Spain CPI remains better explained by classical autoregressive models.

2. **RQ2.** External economic context can improve forecasts when it is aligned with the target series, but it can also add little value or make the result slightly weaker. The clearest contextual contribution appears in Global CPI. In European HICP, the strongest result is mainly a foundation-model gain over SARIMA, while the additional contextual gain over the univariate TimesFM configuration is limited. Spain CPI shows the

limit of broader signals for a more national target.

3. **RQ3.** The results are different for Spain CPI, Global CPI, and European HICP. This is one of the main findings of the project: the same modelling strategy does not behave equally across a national price index, an aggregated international rate, and a European harmonised index.

4. **RQ4.** The usefulness of foundation models and contextual signals depends on the forecast horizon. Classical and dynamically selected statistical models remain competitive at short horizons, while foundation models become more relevant at longer horizons for Global CPI and European HICP.

5. **RQ5.** A reproducible pipeline that transforms semantic and institutional context into exogenous variables can provide measurable value, but the value is selective. It is clearest when the context describes the same economic environment as the target, when the future path of that context is represented realistically, and when the model interface can use it without introducing leakage or excessive noise.

The working hypotheses can also be adjudicated from the final evidence:

1. **H1, foundation-model transfer: partially supported.** Foundation models transfer useful forecasting knowledge for Global CPI and European HICP at longer horizons, but the Spain CPI results show that local statistical structure can remain stronger.

2. **H2, context relevance: partially supported.** Context is useful when it matches the economic scope and representation of the target, but the evidence is not universal. Global institutional and commodity variables provide the strongest contextual result. European context is relevant to the HICP series, but the additional gain over TimesFM `C0` is small. Spain CPI does not show a reliable contextual improvement.

3. **H3, shock-period benefit: partially supported.** The regime and robustness analysis shows that contextual information is more useful during shock-sensitive periods and targets, especially for Global CPI. However, the effect is not universal across all targets and model interfaces.

4. **H4, stability-period noise: partially supported.** Stable or weakly aligned contexts can reduce the usefulness of additional signals, as shown by the Spain signal set and by the weaker pre-shock contextual comparisons.

5. **H5, horizon dependence: supported.** The results change clearly across 1, 3, 6, and 12-month horizons, with longer horizons giving foundation models and contextual configurations more room to become competitive.

Overall, the CDIA contribution is an empirical evaluation rather than a claim that newer models are always better. Its value lies in building a reproducible comparison, testing context under temporal-availability constraints, and showing where modern forecasting tools genuinely add value. The robustness analysis in Section 7.6 also clarifies how the p-values should be interpreted: they support a selective contextual effect, mainly for Global CPI, rather than a general conclusion that exogenous variables improve every forecast. The most defensible conclusion is that foundation models and contextual signals are promising for broader or institutionally connected inflation targets, but classical baselines remain essential for interpreting whether that promise is real.

10.2. LIMITATIONS

The results should be interpreted with the scope of the experiment in mind. The evaluation period is concentrated on 2021–2024, a highly relevant but relatively short interval marked by the post-pandemic inflation shock, energy-market pressure, and monetary-policy tightening. This period is useful for testing the behaviour of contextual signals under stress, but it does not cover every possible inflation regime.

The contextual layer also has practical limits. Some signals are monthly or lower frequency, while others are derived from text-processing pipelines and must be aggregated before being used in the forecasting models. Not all model families accept exogenous variables directly, so the contextual comparison had to use compatible protocols depending on the model. This makes the evaluation realistic, but it also means that context is not introduced in exactly the same technical form for every model. As discussed in Section 7.6, the statistical p-values are therefore used where forecasts can be paired fairly, while benchmark-only models are interpreted mainly through MAE, RMSE, and MASE.

Finally, the project focuses on point forecasts and standard error metrics. These are appropriate for the objectives of the thesis, but future work could extend the analysis with probabilistic forecasts, prediction intervals, scenario evaluation, and a more detailed economic interpretation of the forecast errors.

10.3. FUTURE WORK

The future work follows from one main limitation: the contextual layer used in the final experiment is monthly, relatively low-frequency, and only weakly informative at that granularity for some targets. The next step is therefore not simply to add more variables. The more coherent direction is to improve both the modelling of contextual information and the robustness of the evidence used to evaluate it.

10.3.1. Model and signal extensions

The first line of future work is a trained contextual reviser. The exploratory probes in Section 7.7 show that a small Nexus-inspired prototype can generate coherent macroeconomic reasoning, but zero-shot prompting is not enough to revise a strong time-series foundation-model prior reliably. For this direction to work, the missing pieces are a stronger data-ingestion layer, more forecast origins, richer higher-frequency contextual records, and a real training objective tied to realised forecast errors. A trained reviser, closer to PostTime-style forecast revision [54], could learn when to preserve the original foundation-model forecast, when to adjust it, and how large the correction should be.

A second line is a richer semantic signal layer. The current MCP and text-processing pipeline converts institutional and news information into monthly variables. Future work could preserve more detail from the original sources: publication timing, event type, source reliability, topic, tone, uncertainty, and market relevance. This would make the context less dependent on monthly averages and more capable of capturing shocks, announcements, and expectation changes before they are absorbed into standard macroeconomic indicators.

A third line is mixed-frequency modelling. MIDAS models [47] are designed to combine variables observed at different frequencies, and machine-learning MIDAS work [50] shows how high-dimensional or text-derived signals can be used in nowcasting tasks. In this project, the exploratory MIDAS probe suggests that this line is most promising when the high-frequency variable contains genuinely new information. Future work should therefore move beyond simple commodity-price variants and test daily news flow, GDELT event intensity, institutional releases, and market-expectation signals.

These extensions point toward a future system that ingests richer daily or event-level context, preserves the timing of that context, and learns how it should affect a foundation-model prior. The current project provides the baseline pipeline and the evidence that context is valuable only under specific conditions. Future work should use that evidence to move from prompted or manually engineered context toward trained, higher-frequency contextual forecasting.

10.3.2. Robustness and contextual representation

A second group of future work concerns the way contextual evidence is represented and validated. The robustness analysis in Section 7.6 suggests that the future path of exogenous variables is itself a modelling decision. Holding unknown future values flat is conservative and reproducible, but it can hide useful signal during periods of rapid change. Future experiments should compare flat carry-forward paths with damped trends, scenario-

based paths, market-implied paths, and small auxiliary forecasts for the exogenous variables.

This robustness line should also extend the evaluation design. A longer evaluation window would provide more forecast origins for Diebold-Mariano tests and would make it easier to distinguish large descriptive MAE improvements from statistically stable effects. It would also allow the same contextual protocol to be tested across more inflation regimes, instead of concentrating mainly on the 2021–2024 shock and normalization period.

Finally, future work could test alternative target representations for Spain and Europe, such as inflation-rate or change-based targets, before adding more contextual variables. The Global results suggest that context is easier to exploit when the target is closer to the economic variation described by the covariates. This would keep future extensions aligned with the main lesson of the thesis: context should be improved through better representation and validation, not only through a longer list of signals.

11. BIBLIOGRAPHY

[1] Y. Liang, H. Wen, Y. Nie, Y. Jiang, M. Jin, D. Song, and S. Pan, "Foundation Models for Time Series Analysis: A Tutorial and Survey," *arXiv preprint arXiv:2403.14735*, 2024, available: https://arxiv.org/abs/2403.14735.

[2] L. J. Tashman, "Out-of-Sample Tests of Forecasting Accuracy: An Analysis and Review," *International Journal of Forecasting*, vol. 16, no. 4, pp. 437–450, 2000, available: https://doi.org/10.1016/S0169-2070(00)00065-0.

[3] R. J. Hyndman and A. B. Koehler, "Another Look at Measures of Forecast Accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006, available: https://doi.org/10.1016/j.ijforecast.2006.03.001.

[4] F. X. Diebold and R. S. Mariano, "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, vol. 13, no. 3, pp. 253–263, 1995, available: https://doi.org/10.1080/07350015.1995.10524599.

[5] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. Wiley, 2015, available: https://www.wiley.com/en-us/Time+Series+Analysis%3A+Forecasting+and+Control%2C+5th+Edition-p-9781118675021.

[6] R. J. Hyndman and Y. Khandakar, "Automatic Time Series Forecasting: The forecast Package for R," *Journal of Statistical Software*, vol. 27, no. 3, pp. 1–22, 2008, available: https://doi.org/10.18637/jss.v027.i03.

[7] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, available: https://doi.org/10.1162/neco.1997.9.8.1735.

[8] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, "N-BEATS: Neural Basis Expansion Analysis for Interpretable Time Series Forecasting," *arXiv preprint arXiv:1905.10437*, 2019, available: https://arxiv.org/abs/1905.10437.

[9] C. Challu, K. G. Olivares, B. N. Oreshkin, F. Garza, M. Mergenthaler-Canseco, and A. Dubrawski, "N-HiTS: Neural Hierarchical Interpolation for Time Series Forecasting," *arXiv preprint arXiv:2201.12886*, 2022, available: https://arxiv.org/abs/2201.12886.

[10] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 11 106–11 115, available: https://doi.org/10.1609/aaai.v35i12.17325.

[11] X. Zhang, R. Roy Chowdhury, R. K. Gupta, and J. Shang, "Large Language Models for Time Series: A Survey," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024, available: https://www.ijcai.org/proceedings/2024/0921.pdf.

[12] A. Garza, C. Challu, and M. Mergenthaler-Canseco, "TimeGPT-1," *arXiv preprint arXiv:2310.03589*, 2024, available: https://arxiv.org/abs/2310.03589.

[13] A. Das, W. Kong, R. Sen, and Y. Zhou, "A Decoder-Only Foundation Model for Time-Series Forecasting," *arXiv preprint arXiv:2310.10688*, 2024, available: https://arxiv.org/abs/2310.10688.

[14] A. F. Ansari, L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S. S. Rangapuram, S. P. Arango, S. Kapoor, J. Zschiegner, D. C. Maddix, H. Wang, M. W. Mahoney, K. Torkkola, A. G. Wilson, M. Bohlke-Schneider, and Y. Wang, "Chronos: Learning the Language of Time Series," *arXiv preprint arXiv:2403.07815*, 2024, available: https://arxiv.org/abs/2403.07815.

[15] M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, and A. Dubrawski, "MOMENT: A Family of Open Time-Series Foundation Models," in *Proceedings of the 41st International Conference on Machine Learning*, 2024, available: https://proceedings.mlr.press/v235/goswami24a.html.

[16] K. Rasul, A. Ashok, A. R. Williams, H. Ghonia, R. Bhagwatkar, A. Khorasani, M. J. D. Bayazi, G. Adamopoulos, R. Riachi, N. Hassen, M. Bilos, S. Garg, A. Schneider, N. Chapados, A. Drouin, V. Zantedeschi, Y. Nevmyvaka, and I. Rish, "Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting," *arXiv preprint arXiv:2310.08278*, 2024, available: https://arxiv.org/abs/2310.08278.

[17] G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo, "Unified Training of Universal Time Series Forecasting Transformers," *arXiv preprint arXiv:2402.02592*, 2024, available: https://arxiv.org/abs/2402.02592.

[18] V. Ekambaram, A. Jati, P. Dayama, S. Mukherjee, N. H. Nguyen, W. M. Gifford, C. Reddy, and J. Kalagnanam, "Tiny Time Mixers (TTMs): Fast Pre-Trained Models for Enhanced Zero/Few-Shot Forecasting of Multivariate Time Series," *arXiv preprint arXiv:2401.03955*, 2024, available: https://arxiv.org/abs/2401.03955.

[19] C. Ghirelli, J. J. Pérez, and A. Urtasun, "A New Economic Policy Uncertainty Index for Spain," Banco de España, Documentos de Trabajo 1906, 2019, available: https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/19/Files/dt1906e.pdf.

[20] P. Aguilar, C. Ghirelli, M. Pacce, and A. Urtasun, "Can News Help Measure Economic Sentiment? An Ap-

plication in COVID-19 Times," Banco de España, Documentos de Trabajo 2027, 2020, available: https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/20/Files/dt2027e.pdf.

[21] J. Ashwin, E. Kalamara, and L. Saiz, "Nowcasting Euro Area GDP with News Sentiment: A Tale of Two Crises," European Central Bank, Working Paper Series 2616, 2021, available: https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp2616~58494f90b7.en.pdf.

[22] G. J. de Bondt and Y. Sun, "Enhancing GDP Nowcasts with ChatGPT: A Novel Application of PMI News Releases," European Central Bank, Working Paper Series 3063, 2025, available: https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp3063~f88c1b73fc.en.pdf.

[23] D. Caldara and M. Iacoviello, "Measuring Geopolitical Risk," *American Economic Review*, vol. 112, no. 4, pp. 1194–1225, 2022, available: https://doi.org/10.1257/aer.20191823.

[24] S. P. Arango, P. Mercado, S. Kapoor, A. F. Ansari, L. Stella, H. Shen, H. Senetaire, C. Turkmen, O. Shchur, D. C. Maddix, M. Bohlke-Schneider, Y. Wang, and S. S. Rangapuram, "ChronosX: Adapting Pretrained Time Series Models with Exogenous Variables," *arXiv preprint arXiv:2503.12107*, 2025, available: https://arxiv.org/abs/2503.12107.

[25] Z. Xu, H. Wang, and Q. Xu, "Intervention-Aware Forecasting: Breaking Historical Limits from a System Perspective," *arXiv preprint arXiv:2405.13522*, 2025, available: https://arxiv.org/abs/2405.13522.

[26] Universidad Internacional de La Rioja, "Pedro Gómez Tejerina," https://www.unir.net/profesores/pedro-gomez-tejerina/, 2026, accessed: 2026-06-12.

[27] R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining," in *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 2000, available: https://www.the-modeling-agency.com/crisp-dm.pdf.

[28] Instituto Nacional de Estadística, "Consumer Price Index (CPI)," https://www.ine.es/dyngs/IPC/en/index.htm?cid=1425, 2026, accessed: 2026-05-30.

[29] World Bank, "A Global Database of Inflation," https://www.worldbank.org/en/research/brief/inflation-database, 2026, accessed: 2026-05-30.

[30] European Central Bank, "Harmonised Index of Consumer Prices (HICP) dataset," https://data.ecb.europa.

eu/data/datasets/HICP, 2026, accessed: 2026-05-30.

[31] ——, “ECB Data Portal,” https://data.ecb.europa.eu/, 2026, accessed: 2026-05-30.

[32] Yahoo Finance, “Historical Market Data,” https://finance.yahoo.com/, 2026, accessed: 2026-05-30.

[33] Federal Reserve Bank of St. Louis, “FRED API: Series Observations,” https://fred.stlouisfed.org/docs/api/fred/series_observations.html, 2026, accessed: 2026-05-30.

[34] Federal Reserve Bank of New York, “Global Supply Chain Pressure Index,” https://www.newyorkfed.org/research/policy/gscpi, 2026, accessed: 2026-05-30.

[35] The GDELT Project, “The GDELT Global Knowledge Graph Codebook,” https://data.gdeltproject.org/documentation/GDELT-Global_Knowledge_Graph_Codebook.pdf, 2026, accessed: 2026-05-30.

[36] D. Harvey, S. Leybourne, and P. Newbold, “Testing the Equality of Prediction Mean Squared Errors,” *International Journal of Forecasting*, vol. 13, no. 2, pp. 281–291, 1997, available: https://doi.org/10.1016/S0169-2070(96)00719-4.

[37] FastMCP, “FastMCP Documentation,” https://gofastmcp.com/, 2026, accessed: 2026-06-08.

[38] GDELT Project, “The GDELT Event Database: Data Format Codebook V2.0,” https://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf, 2015, accessed: 2026-05-31.

[39] Ollama, “Ollama API Documentation,” https://docs.ollama.com/api, 2026, accessed: 2026-06-08.

[40] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, “Qwen3 Technical Report,” *arXiv preprint arXiv:2505.09388*, 2025, available: https://arxiv.org/abs/2505.09388.

[41] Pydantic, “Pydantic Models Documentation,” https://docs.pydantic.dev/latest/concepts/models/, 2026, accessed: 2026-06-08.

[42] Board of Governors of the Federal Reserve System, “Federal Open Market Committee: Meeting Calendars, Statements, and Minutes,” https://www.federalreserve.gov/monetarypolicy/fomccalendars.htm, 2026, accessed: 2026-05-31.

[43] European Central Bank, “Monetary Policy Statements,” https://www.ecb.europa.eu/press/pressconf/, 2026, accessed: 2026-05-31.

[44] U.S. Bureau of Labor Statistics, “Consumer Price Index Archived News Releases,” https://www.bls.gov/bls/news-release/cpi.htm, 2026, accessed: 2026-05-31.

[45] A. E. Hoerl and R. W. Kennard, “Ridge Regression: Biased Estimation for Nonorthogonal Problems,” *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970, available: https://doi.org/10.1080/00401706.1970.10488634.

[46] Nixtla, “TimeGPT Exogenous Variables Documentation,” https://nixtlaverse.nixtla.io/nixtla/docs/capabilities/forecast/exogenous_variables.html, 2026, accessed: 2026-06-08.

[47] E. Ghysels, P. Santa-Clara, and R. Valkanov, “The MIDAS Touch: Mixed Data Sampling Regression Models,” https://econpapers.repec.org/RePEc:cir:cirwor:2004s-20, CIRANO, Working Paper 2004s-20, 2004.

[48] S. S. S. Das, P. Goyal, M. Parmar, N. Peng, V. Tirumalashetty, C.-L. Li, R. Zhang, J. Yoon, and T. Pfister, “Nexus: An Agentic Framework for Time Series Forecasting,” https://arxiv.org/abs/2605.14389, 2026.

[49] U.S. Energy Information Administration, “Crude Oil Prices: West Texas Intermediate (WTI) - Cushing, Oklahoma,” https://fred.stlouisfed.org/series/DCOILWTICO, 2026, retrieved from FRED, Federal Reserve Bank of St. Louis; accessed: 2026-06-12.

[50] A. Babii, E. Ghysels, and J. Striaukas, “Machine Learning Time Series Regressions with an Application to Nowcasting,” *Journal of Business & Economic Statistics*, vol. 40, no. 3, pp. 1094–1106, 2022, available: https://arxiv.org/abs/2005.14057.

[51] Association for Computing Machinery, “ACM Code of Ethics and Professional Conduct,” https://www.acm.org/code-of-ethics, 2018, accessed: 2026-05-12.

[52] National Institute of Standards and Technology, “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” https://www.nist.gov/itl/ai-risk-management-framework, 2023, accessed: 2026-05-12.

[53] European Union, “Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence,” https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, 2024, accessed: 2026-05-12.

[54] H. Liu, Y. Zhou, R. Sen, B. A. Prakash, and A. Das, “Rethinking Post-Training Recipes for Multimodal Time-Series Forecasting,” https://arxiv.org/abs/2605.29401, 2026.

[55] University of Deusto Artificial Intelligence Committee, “Use of Artificial Intelligence,” https://ai-label.org, 2026, guidance sheet provided for students; accessed: 2026-06-12.

[56] Anthropic, “Claude,” https://claude.ai/, 2026, accessed: 2026-06-15.

[57] ——, “Claude Code,” https://code.claude.com/, 2026, aI coding assistant; accessed: 2026-06-15.

DEFINITIONS, ACRONYMS, AND ABBREVIATIONS

ACRONYMS AND ABBREVIATIONS

AI Artificial Intelligence. In this thesis, the term refers mainly to machine-learning and foundation-model approaches applied to time-series forecasting.

API Application Programming Interface. A software interface used to request data or services from another system, such as TimeGPT or external data providers.

ARIMA Autoregressive Integrated Moving Average. A classical statistical time-series model based on autoregressive terms, differencing, and moving-average terms.

AutoARIMA Automatic ARIMA. A procedure that selects the ARIMA specification automatically according to statistical criteria and validation logic.

BCE Banco Central Europeo. Spanish acronym for the European Central Bank, used when referring to Spanish-language institutional sources or variables.

BLS Bureau of Labor Statistics. United States statistical agency used as a source for CPI-related information in the global contextual layer.

C0 Baseline forecasting condition. It uses only the historical target series, without additional contextual or exogenous variables.

C1 Contextual forecasting condition. It introduces external information when the corresponding model or correction protocol allows it.

`C1_energy` Contextual condition based mainly on energy-related variables, such as oil or gas prices.

`C1_full` Contextual condition that combines the main structured and semantic signal layers when they are available and meaningful for the target series.

`C1_inst` Contextual condition based mainly on institutional, macroeconomic, market, uncertainty, and policy-related variables.

`C1_macro` Contextual condition based mainly on macro-financial variables and broader economic indicators.

`C1_mcp` Contextual condition based mainly on semantic or text-derived signals collected and processed through the MCP-related pipeline.

CDIA Ciencia de Datos e Inteligencia Artificial. Degree area of this thesis, focused on the data-science, artificial-intelligence, and forecasting evaluation part of the integrated project.

CPI Consumer Price Index. A price index used to measure the evolution of consumer prices and commonly used as an inflation indicator.

CRISP-DM Cross-Industry Standard Process for Data Mining. A reference methodology for data-mining projects, adapted in this thesis to the forecasting experiment.

CSV Comma-Separated Values. A plain-text data format commonly used for tabular datasets.

DFR Deposit Facility Rate. European Central Bank policy rate used as part of the monetary-policy context.

DM Diebold-Mariano test. A statistical test used to compare predictive accuracy between two competing forecasting methods.

ECB European Central Bank. Institution whose policy rates, data portal, and communications are used in the European and institutional context layers.

EDA Exploratory Data Analysis. Initial analysis of datasets to understand their structure, missing values, trends, and relevant patterns.

EPU Economic Policy Uncertainty. A family of indicators that measure uncertainty related to economic policy.

ETL Extract, Transform, Load. Data-processing workflow used to collect raw data, transform it into a usable form, and store it for later analysis.

FOMC Federal Open Market Committee. United States monetary-policy body whose statements and calendars are relevant for global monetary-policy context.

FRED Federal Reserve Economic Data. Database maintained by the Federal Reserve Bank of St. Louis and used as a source for economic and financial indicators.

GDP Gross Domestic Product. Macroeconomic indicator that measures aggregate economic output.

GDELT Global Database of Events, Language and Tone. Open data project used to obtain event and text-derived

signals such as tone, article volume, and geopolitical or economic event information.

GPR Geopolitical Risk. Indicator or signal family used to capture geopolitical tension and uncertainty.

HICP Harmonised Index of Consumer Prices. European inflation indicator designed to be comparable across European Union countries.

INE Instituto Nacional de Estadistica. Spanish statistical institute used as the source for Spain CPI data.

JSON JavaScript Object Notation. Structured data format used in the repository to store metrics, model traces, and intermediate outputs.

LLM Large Language Model. A language model trained on large text corpora. In this project, LLM-related components are used for contextual signal extraction and exploratory future-work revisers, not as direct inflation forecasters in the main experiment.

LSTM Long Short-Term Memory. A recurrent neural network architecture used as a deep-learning baseline for sequence modelling.

MAE Mean Absolute Error. Error metric that measures the average absolute difference between forecasted and observed values.

MASE Mean Absolute Scaled Error. Error metric that scales the forecast error against a naive benchmark, making comparisons across series easier.

MCP Model Context Protocol. In this project, MCP refers to the software and signal-access layer used to collect, organize, and expose contextual information for the forecasting pipeline.

MIDAS Mixed Data Sampling. A modelling approach designed to combine variables observed at different frequencies, such as daily market data and monthly inflation.

N-BEATS Neural Basis Expansion Analysis for Interpretable Time Series Forecasting. A deep-learning forecasting architecture used as one of the local neural baselines.

N-HiTS Neural Hierarchical Interpolation for Time Series Forecasting. A deep-learning forecasting architecture designed for efficient multi-horizon time-series forecasting.

Parquet Column-oriented data format used to store processed datasets and prediction records efficiently.

PMI Purchasing Managers' Index. Survey-based economic indicator sometimes used in nowcasting and macroeconomic sentiment analysis.

RMSE Root Mean Squared Error. Error metric that gives more weight to large forecast errors than MAE.

SARIMA Seasonal Autoregressive Integrated Moving Average. ARIMA extension that includes seasonal dynamics.

SARIMAX Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors. SARIMA extension that incorporates external variables.

TFG Trabajo Fin de Grado. Final degree project.

TimeGPT Foundation time-series forecasting model accessed through an API.

TimesFM Foundation time-series forecasting model developed for general time-series forecasting.

TSFM Time-Series Foundation Model. A pretrained model designed to generalize across time-series forecasting tasks.

WTI West Texas Intermediate. Oil-price benchmark used in exploratory future-work experiments with daily energy-price signals.

KEY DEFINITIONS

Contextual signal External information used to support the forecast beyond the historical target series. In this project, contextual signals include institutional variables, monetary-policy data, market indicators, energy prices, uncertainty indices, GDELT signals, and text-derived variables.

Exogenous variable Variable introduced from outside the target time series. Exogenous variables can improve forecasting when they are available at the forecast origin and economically related to the target series.

Forecast horizon Number of months between the forecast origin and the value being predicted. The project evaluates 1-, 3-, 6-, and 12-month horizons.

Forecast origin Date from which a forecast is produced. At each origin, the model can only use information that would have been available at that date.

Foundation time-series model Pretrained forecasting model designed to transfer knowledge from large collections of time series to new forecasting tasks.

Inflation forecasting Estimation of future inflation values using historical price data, economic variables, market indicators, and, in this project, contextual information.

Leakage Methodological error that occurs when a model uses information from the future, directly or indirectly, during training or prediction.

Nowcasting Forecasting the current or very near-term value of an economic variable, often before official data are released.

Rolling-origin backtesting Evaluation design in which forecasts are generated repeatedly from successive historical origins. This simulates a realistic forecasting process and avoids evaluating models with information that would not yet have been available.

Semantic signal Numerical variable derived from text sources such as news, institutional releases, or event databases. Examples include tone, uncertainty scores, topic indicators, and shock-related variables.

Shock period Time interval in which the inflation process is affected by strong economic changes, such as energy shocks, post-pandemic disruptions, or monetary-policy tightening.

Univariate forecast Forecast that uses only the historical target series and no external contextual variables.

A. APPENDICES

The appendices collect supporting material that is useful for understanding, reproducing, or reviewing the CDIA forecasting thesis without interrupting the main methodological and results chapters.

A.1. PROJECT REPOSITORY

The complete GitHub repository of this project is available at DiegoRamirezLacalle/tfg-ipc-mcp. It contains the forecasting pipeline, data-processing code, evaluation scripts, generated result artifacts, and supporting implementation used as evidence for this memory.

A.2. DECLARATION OF ARTIFICIAL INTELLIGENCE USE

This appendix declares the use of generative artificial intelligence tools during the development and writing of this thesis. It follows the transparency criterion recommended by the University of Deusto guidance on the use of artificial intelligence in academic work [55]. The objective of this declaration is not to reproduce the prompts used during the project, but to state which tools were used, in which stages, and for what purpose.

I developed the project and wrote the thesis, using the artificial intelligence tools described below only as support for research organization, code assistance, debugging, and writing revision. Their outputs were not accepted automatically. I reviewed them, contrasted them with reliable sources, checked them against the actual project code and metric files when relevant, and rewrote them where necessary to preserve the accuracy and authorship of the work.

Table A.1.: Use of artificial intelligence tools during the project

Tool	**Stage of the project**	**Purpose of use**
Claude [56]	Literature review, state-of-the-art analysis, and technical understanding	I used Claude to support the search, organization, and synthesis of information about recent forecasting models, time-series foundation models, contextual signals, and selected research papers. I checked the resulting suggestions against the original papers, official documentation, and the bibliography before incorporating any content into the thesis.
Claude Code [57]	Software development and experimental support	I used Claude Code as a coding assistant while implementing and reviewing parts of the repository, especially unfamiliar components related to the Model Context Protocol, semantic-signal extraction, data validation, and reproducibility checks. I inspected, adapted, executed, and verified code suggestions before keeping them.
Claude and Claude Code writing assistance [56, 57]	Draft revision and final document preparation	I wrote the thesis text and used these tools to review clarity, grammar, and professional tone in English paragraphs, tables, captions, and explanatory sections. This support was particularly useful because the thesis is written in formal academic English. I made the final wording decisions and edited the text so that it reflected the work I actually carried out.

In all cases, I used the tools to support learning and productivity rather than to replace the core academic work. I decided the research questions, experimental design, data selection, metric interpretation, and final conclusions. Numerical results were taken from the project artifacts and regenerated metric files, not from artificial intelligence outputs. When I used AI assistance for literature or technical explanations, I verified the content against primary sources or official documentation. When I used it for code, I tested the resulting scripts and outputs through the local project environment.

The main safeguards applied during the project were the following:

- AI-generated explanations were checked against papers, official documentation, or project files before being used.
- Forecasting results, tables, and figures were based only on real metric files and prediction artifacts generated by the project pipeline.
- Code suggestions were executed and reviewed before being accepted.
- Textual revisions were used to improve clarity and correctness, not to change the meaning of the work or introduce unsupported claims.
- I retained responsibility for the final content, interpretation, and academic validity of the thesis.

Therefore, the use of artificial intelligence in this thesis was limited, transparent, and supervised. It contributed to research organization, software assistance, and language quality, while the final evidence, reasoning, and conclusions remained grounded in the implemented project and in verified sources.