



BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Agentic AI: can we streamline economic policy briefing?

by Giuseppe Bruno

March 2026

Number

1001



BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Agentic AI: can we streamline economic policy briefing?

by Giuseppe Bruno

Number 1001 – March 2026

The series Occasional Papers presents studies and documents on issues pertaining to the institutional tasks of Banca d'Italia and the Eurosystem. The Occasional Papers appear alongside the Working Papers series which are specifically aimed at providing original contributions to economic research.

The Occasional Papers include studies conducted within Banca d'Italia, sometimes in cooperation with the Eurosystem or other institutions. The views expressed in the studies are those of the authors and do not involve the responsibility of the institutions to which they belong.

The series is available online at www.bancaditalia.it.

AGENTIC AI: CAN WE STREAMLINE ECONOMIC POLICY BRIEFING?

by Giuseppe Bruno*

Abstract

Central banks increasingly rely on large volumes of policy documents—ranging from monetary policy statements to regulatory guidelines and internal reports—to communicate decisions and maintain institutional accountability. These documents help ground decisions in the most up-to-date economic signals, support internal coordination and strengthen transparency toward markets and the public. Yet their complexity and frequency impose significant administrative and analytical burdens.

The emergence of agentic AI, which combines large language models with autonomous goal setting, structured workflows and tool-calling capabilities, offers the potential to automate key tasks such as drafting, summarization, cross-document consistency checks, integration with data sources and the generation of tables and figures.

In this paper, we address the question posed in the title by examining the role that recent advances in AI—enhanced by agent-based architectures and coordinated task execution—can play in central bank workflows. We present two simple, ground-up examples showing how open-source tools can be leveraged effectively while keeping development and operating costs under control. The paper concludes by discussing how agentic AI could reshape document processing practices within central banks, highlighting both the efficiency gains and the governance safeguards required for its trustworthy adoption.

JEL Classification: C82, C87, O36.

Keywords: Language Models, LangChain, LangGraph, agent.

DOI: 10.32057/0.QEF.2026.1001

* Bank of Italy, Economics and statistic department; giuseppe.bruno@bancaditalia.it

The views and positions expressed in this paper are those of the author only and do not necessarily reflect those of the Bank of Italy.

1. Introduction

going in a circle, and taking that for granted, which is the very point in question. David Hume (1748)

Despite the remarkable progress of large language models (LLMs) and their multimodal extensions—which now generate text, audio, and video, and can address a wide variety of legal, coding, and forecasting tasks (see Tom B. Brown, 2020 and Aakanksha Chowdhery, 2022) - it remains unclear whether such models truly reason, and if so, to what extent and in what sense their reasoning is comparable to human logical inference. Understanding the reasoning capabilities of AI systems is crucial for evaluating their reliability in domains where coherent decision-making and robust inferential processes are essential. The role of logical reasoning in artificial intelligence is fundamental to the development of systems that can simulate human-like decision-making and problem-solving abilities. Logical reasoning has long been a foundational component of artificial intelligence. It enables systems to interpret information, derive inferences, and make decisions based on structured relations among data. As a core element of intelligent behavior, logical reasoning underpins the ability of AI systems to manipulate information in meaningful and internally consistent ways.

Classically, logical reasoning in AI can be categorised into three main forms: deductive, inductive, and abductive (see for example Abdaljalil et al., 2025).

Deductive reasoning applies general rules to specific cases to derive conclusions, much like a mathematical proof (see Morishita, 2023). It plays a central role in rule-based and expert systems, where correctness and internal consistency are essential.

Inductive reasoning, by contrast, generalises from specific instances to broader patterns. This form of inference is fundamental to machine learning, where systems infer statistical regularities from data; for example, analysing large sets of consumer transactions to predict future purchasing behaviour (cfr. Chen & et al, 2025).

Abductive reasoning seeks the most plausible explanation for a set of observations (for example Wang-Zhou et al., 2019). It is especially relevant in diagnostic applications, such as medical decision-support systems that infer the likely cause of a patient’s symptoms from incomplete information.

Logical reasoning therefore plays a central role in a wide range of AI applications. In finance, reasoning-enabled AI systems can optimise trading strategies, assess credit risk, or detect anomalies in transaction flows. In healthcare, logical inference supports diagnostic processes, personalised treatment recommendations, and resource management. Autonomous vehicles rely on real-time forms of reasoning to interpret sensor data and navigate complex environments safely.

The rest of this paper is organized as follows.

In paragraph 2 we show how the limitations present in the classical AI system spurred the development of Autonomous Agent. Then in the paragraph 3 we provide a definition of an Agentic AI system and the way it can be used for Economic Policy applications. In paragraph 4 we sketch the present status of the software components available for developing Agentic AI system. In the following paragraph 5 we provide a synthetic comparison between Agentic and classical AI applications. In paragraph 6 we describe the two empirical applications of Agentic AI in streamlining the production of economic policy report. Finally in the section 7 we draw some final considerations.

2. From LLMs to Autonomous Agents: Limitations as a Catalyst

Large Language Models (LLMs) such as OpenAI’s GPT-5 (OpenAI, 2025), Qwen2.5-Omni (Qwen-Team, 2025), DeepSeek-R1 (DeepSeek-AI, 2025, and Meta’s LLaMA (see Abdullah & et al, 2025), along with Large Visual (LVM) and Multimodal Models (MM) have revolutionized AI by enabling human-like text generation and advanced natural language processing, spurring innovation in conversational agents, automated content creation, and real-time translation.² However, their dependence on static pre-training data can lead to outdated outputs and, sometimes to hallucinated responses (see Rawte et al., 2023 and Ji & et al, 2023), a limitation that Retrieval-Augmented Generation (RAG) addresses by incorporating real-time data from knowledge bases, Application Programming Interfaces (APIs), or the web (cfr. Lewis & et al, 2021).

Building on this foundation, the emergence of intelligent agents capable of reflection, planning, and multi-agent collaboration has led to the development of agentic AI systems. These dynamically orchestrate information retrieval and iterative refinement, enabling them to manage increasingly complex workflows with greater efficiency.

The trajectory leading to contemporary agentic AI can be traced back to the steady evolution of natural language processing (NLP) over the past two decades. Before the deep learning revolution, NLP was dominated by statistical approaches such as n-gram models (see Mikolov & et al., 2013), topic models (see Blei et al., 2003), and latent semantic analysis (cfr. Deerwester et al., 1990). While effective for tasks such as text classification or basic machine translation, these methods suffered from inherently limited context windows³ and exhibited brittle generalization capabilities.

The introduction of neural architectures — first recurrent neural networks (RNNs) (cfr. Elman, 1990) and then long short-term memory networks (LSTMs) (cfr. Sutskever et al., 2014 — allowed models to capture sequential dependencies in language more effectively. However, they remained constrained by training instabilities and difficulties in handling

²for LVM consider Dosovitskiy & et al., 2021 and Bao & et al., 2021, while for MM you can see Radford & et al., 2021 and Jia & et al., 2021

³Here we refer to the admissible size of the input prompt.

very long sequences.⁴

A pivotal breakthrough arrived with the transformer architecture (see Vaswani & et al., 2017). By relying on self-attention mechanisms, transformers removed the bottleneck of sequential processing and enabled models to handle long-range dependencies in parallel. This innovation opened the door to unprecedented scaling, both in terms of data and parameters.

In the picture 1 we show the basic differences between recurrent neural network and the transformer architecture.

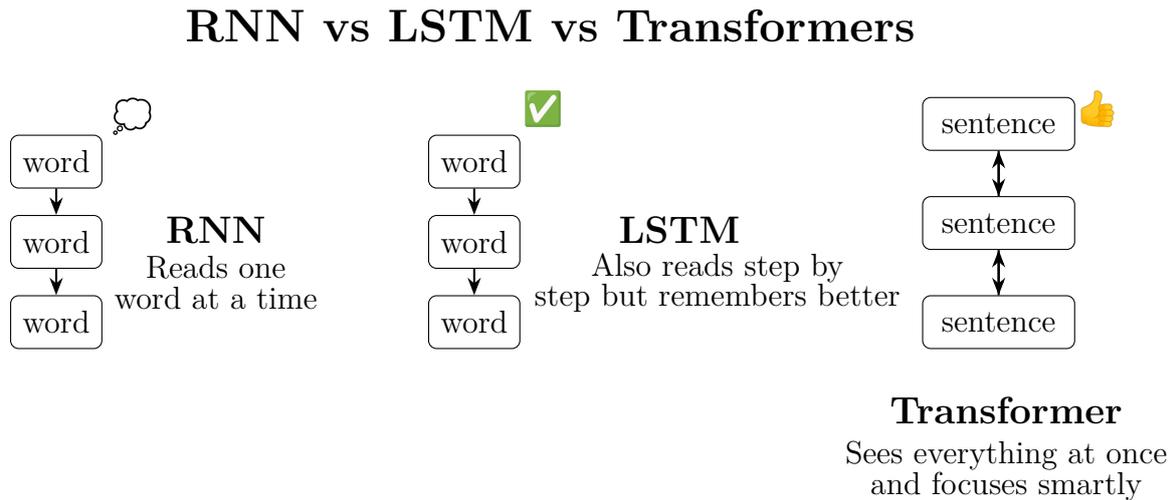


Figure 1: Schematic diagram of the differences on AI Neural Network Architecture.

The most relevant difference in these architectures is that in the RNN and the LSTM architectures each word is sequentially processed by keeping memory of a given word length. The transformer architecture takes advantage of the memory growth on CPUs and GPUs by processing in parallel long set of sentences with multiple attention kernel⁵.

2.1 The Scaling Era and the Emergence of Foundation Models

The concept of the foundation models emerged at the beginning of the 2020s⁶: large pretrained networks that could be adapted across domains. Their generality made them attractive not just to the AI research community but to industry and government agencies seeking versatile tools for knowledge work.

Despite these advances, early LLMs were essentially passive predictors. They generated coherent text given a prompt but lacked structured mechanisms to interact with external tools, data sources, or longer-term workflows. This passivity limited their usefulness in complex decision-making settings, where reasoning, memory, and autonomous action were as important as fluent text generation.

⁴The vanishing and exploding gradients problem.

⁵The attention is a learned, differentiable mechanism that selects relevant input elements at each processing step by assigning them weights, inspired by human selective attention

⁶see Bommasani & et al., 2022

2.2 Adding Instruction Following and Alignment

The next major leap was the refinement of instruction following. Through techniques like supervised fine-tuning and reinforcement learning from human feedback (RLHF), models were adapted to align more closely with user intent (cfr Stiennon & et al., 2020 and Ouyang & et al., 2022). This alignment stage was crucial: it transformed LLMs from generic text generators into systems capable of engaging in structured dialogue, answering questions reliably, and respecting basic safety norms.

Instruction-tuned LLMs such as InstructGPT (2022) and ChatGPT (see OpenAI, 2024) showed that conversational agents could provide practical assistance across a wide variety of domains. For central banks, this period coincided with experiments in using LLMs to summarize policy reports, extract key indicators from regulatory filings, and assist with public communications. Yet even these conversational systems remained limited: they could only respond within the immediate conversational context and had no inherent capacity to plan, use external data, or execute actions.

2.3 The Rise of Tool Use and Plugins

To overcome these limits, developers began integrating LLMs with external tools.⁷ Early demonstrations included connecting models to calculators, databases, and search engines. In 2022–2023, commercial systems introduced plugin architectures, allowing LLMs to query APIs, fetch documents, or trigger business workflows (cfr. Schick & et al., 2023). These integrations marked a conceptual shift: the LLM was no longer just a text generator but the reasoning core within a broader software ecosystem.

This capability also aligned with advances in RAG, where models could access external knowledge bases to reduce hallucinations and ground outputs in verifiable data. For policy institutions and regulators, this meant that a model could draft reports not just from training data but by pulling the latest figures from statistical databases, thereby improving both accuracy and trust.

2.4 From ReAct to Agents

The logical next step was to embed reasoning and acting in a single loop. Research like the ReAct framework (see Yao & Jeffrey Zhao, 2023) showed that LLMs could interleave reasoning (planning steps) with actions (tool use), guided by natural language. Above all this approach simplified substantially the code writing task, furthermore enabled models to decompose tasks, call external functions, and refine answers iteratively.

The community began referring to such systems as agents: AI entities capable of perceiving instructions, reasoning about possible actions, executing them via tools or APIs, and then reflecting on the results. Frameworks such as Autogen (see Qingyun,

⁷These are applications or plugins that an AI can connect to in order to perform actions or access data beyond its knowledge cutoff.

2023), LangChain (see Vasilios, 2024) and LangGraph (see LangGraph, 2024) provided developers with the building blocks for constructing these agents, while commercial offerings embedded them in productivity software and domain-specific applications.

AI agents are autonomous software systems optimized for the pursuit of goal-directed task execution under the constraints of a delimited digital environment (for example Acharya et al., 2025 and Sado & et al., 2023). These agents are characterized by their ability to acquire, interpret, and encode information from both structured and unstructured data sources (cfr. Heer & et al., 2019).

3. Agentic AI for Policy Brief Generation

3.1 The global perspective

Recent progress in *agentic artificial intelligence*—LLM and LVM systems endowed with goal-setting, planning, tool-use, and self-correction abilities—is redefining analytic workflows in economics and public policy. Distinct from single-prompt chat systems, agentic AI operates through iterative reasoning, execution, and feedback cycles (see for example Shinn et al., 2023; Yao & Jeffrey Zhao, 2023 and Qingyun, 2023). These architectures can support macroeconomic analysis by autonomously retrieving data, generating tables and visualizations, and drafting coherent policy narratives.

Figure 2 sketches the canonical pipeline for producing a macroeconomic brief on issues such as inflation, credit, or unemployment.

The pipeline is composed by a planning phase followed by the Data Discovery, Data Representation, a narrative generation step potentially inserted into a loop with an editor agent and finally we have the brief production with all the table and graphs.

3.2 Functional Specializations

In moving from Generative AI towards agentic AI it is highly relevant to build up an ecosystem with functional specializations for improving efficiency, accuracy, and adaptability. This is achieved by having different agents handle specific tasks, much like specialists in a company or team. A typical AI agent simply communicates with a chosen LLM/SLM API endpoint by making requests to centralized cloud infrastructure that hosts these models (cfr. Masterman et al., 2024). LLM API endpoints are specifically designed to serve a large volume of diverse requests using one generalist LLM. This approach moves AI from a monolithic system to a multi-agent system (MAS) where specialized agents can collaborate more effectively, leading to optimized workflows, higher task precision, and more scalable and resilient AI systems (see Li & et al., 2023).

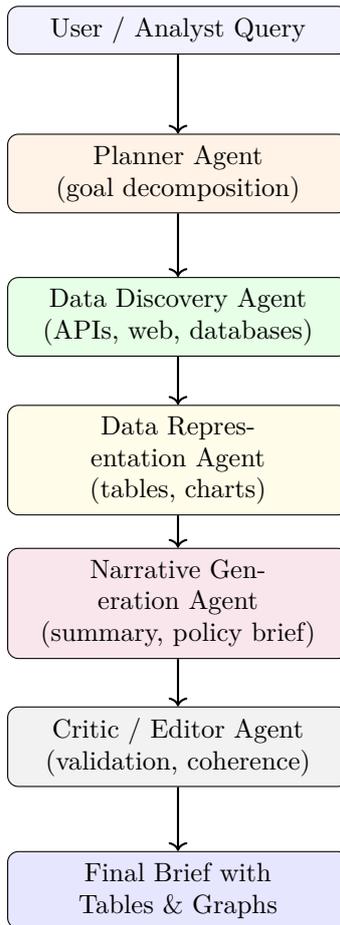


Figure 2: Schematic workflow of agentic AI in policy brief generation.

3.2.1 Data Discovery Agents

These agents locate and validate macroeconomic evidence from statistical portals and research repositories. For increasing the reliability of these reasearches often spacial reasoning environment are employed. The *ReAct* framework (see Yao & Jeffrey Zhao, 2023) demonstrated interleaved reasoning and acting for retrieval, later extended by *WebGPT* (Nakano, 2022) to incorporate browsing and citation collection. Recent systems such as *Gorilla* (Patil & et al., 2024) and the *DSPy* compiler (see Omar Khattab, 2023) improve API selection and query optimization, enabling structured extraction from IMF or Eurostat databases.

3.2.2 Data Representation Agents

A second specialization converts raw series into publishable tables and visualizations. Vision-language models like *ChartX/ChartVLM* and *ChartGPT* (Renqiu Xia, 2025, Yin & et al., 2024, Yuan Tian, 2024) illustrate progress in chart understanding and generation.

Surveys of generative visualization (see Yilin Ye, 2024) map the full pipeline from data transformation to graphical encoding, underscoring the importance of validation to prevent mis-aggregation or incorrect labeling.

3.2.3 Narrative-Generation Agents

Drafting agents synthesize the discovered evidence into structured prose—executive summaries, key-message boxes, and policy options. Approaches such as *Reflexion* (cfr. Shinn et al., 2023) introduce self-critique and revision loops that improve factual consistency. Multi-agent frameworks like *AutoGen* (see Qingyun, 2023) allow role-specific collaboration among “Analyst,” “Fact-Checker,” and “Editor” agents until a final brief meets the adopted acceptance criteria.

3.3 Text-to-Data Interfaces

Parallel advances in text-to-SQL and semantic querying (see for example Xuemei & et al., 2023) make it possible for non-technical economists to query microdata or time-series databases directly through natural language. Embedding these components within agent loops ensures schema validation, provenance tracking, and reproducible code for statistical transparency.

4. Current Status of Software Libraries, APIs and Features of Agentic AI

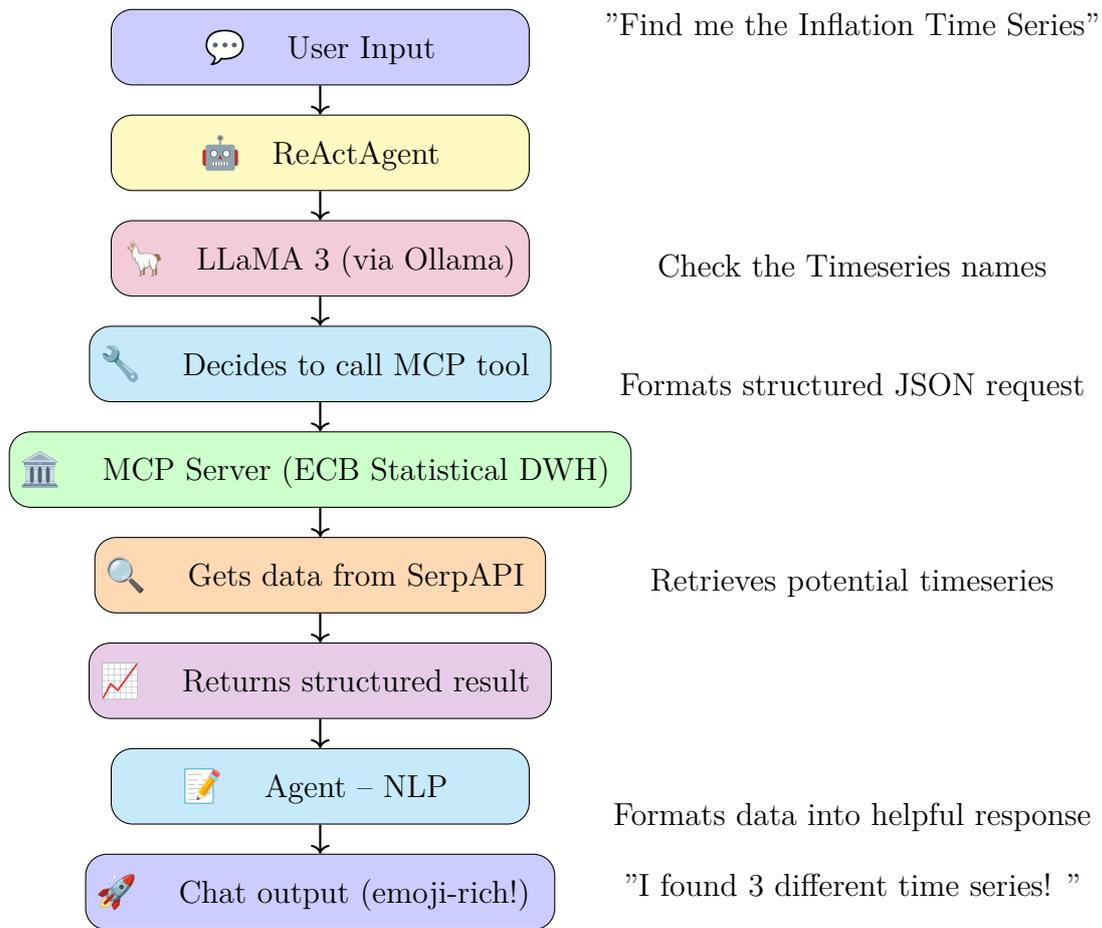
4.1 Major software libraries and APIs

A vibrant ecosystem of frameworks and libraries has emerged to support agentic AI. For example, frameworks like LangChain, AutoGen, CrewAI, and LangGraph are discussed in recent literature comparing their architectures, communication mechanisms and memory management (see Derouiche et al., 2025)

Emerging protocols such as Model Context Protocol (MCP)⁸ facilitate standardized integration between large language models (LLMs), tools and environments, (see for example (Hou et al., 2025)). MCP plays a role for AI systems that is conceptually analogous to the role SQL played for databases: it provides a unified, standardized interface through which heterogeneous tools, data sources, and services can be accessed. While SQL is a declarative query language formalizing operations over relational databases, MCP functions as an interoperability protocol enabling large language models to interact with diverse external resources in a consistent manner.

⁸This protocol, originally introduced by Anthropic in November 2024, is an open-source standard for connecting AI applications to external systems.

Model Context Protocol (MCP)



Open-source libraries like GitHub catalogue dozens of agentic projects and tools for building custom agents.

4.2 Key features of agentic AI

Below are core features typical of agentic AI systems:

1. **Autonomy & goal complexity** Agentic systems can accept a high-level goal and break it into sub-goals, plan their actions, and act without constant human supervision. For instance, one review defines agentic AI as "... capable of initiating goal-directed actions, making context-sensitive decisions, and adapting over time with minimal human oversight." (see Awashti, 2025)
2. **Environment & operational complexity** These agents operate in dynamic environments, possibly multi-modal, tool-enabled, with external API interactions, and changing constraints. They may interact with web APIs, databases, User Interface actions, sensors or even physical actuators. The ability to adapt to external conditions distinguishes them from static models.

- Independent decision making & adaptability Agentic AI emphasises reasoning, reflection, memory, tool use, planning, learning, and goal pursuit (see for example Bandi & et al., 2025).

They can select which tool to call, monitor outcomes, adjust strategy, and collaborate with other agents.

5. Agentic AI vs classical/traditional

In this section we draft a comparison table among the classical AI and traditional agents against the agentic AI framework.

Dimension	Classical AI / Traditional Agents	Agentic AI
Goal setting	Human-specified, narrow tasks	High-level goals, sub-goal decomposition
Autonomy	Low autonomy, human supervision	High autonomy, minimal human oversight
Environment interaction	Static or constrained environment	Dynamic, multi-modal, tool-enabled environment
Decision Making	Rule-based or static learned model	Adaptive decision making, tool invocation, memory, reasoning
Adaptability	Low: limited learning in deployment	High: continuous adaptation, reflection, memory
Libraries / API	ML toolkits, expert-system shells	Agentic frameworks (LangChain, AutoGen, etc.), orchestration APIs

Agentic AI extends the capabilities of traditional agents and classical AI by combining autonomy, goal-driven behaviour, dynamic environment interaction and adaptability

Traditional Multi Agent Systems could interact and negotiate, but their decision logic was typically pre-specified and their environment relatively constrained.

6. The empirical applications

In this section, we present two stylized examples of agentic AI systems designed to support economists in the preparation of policy briefs: one focused on an inflation report and the other on a banking-stability online index. These examples are intentionally simplified

and are not intended to provide a comprehensive economic assessment. Rather, their purpose is to illustrate how agentic AI systems can autonomously retrieve data from relevant datasets, perform the required econometric computations, and generate a narrative grounded in the statistical properties of the data.

6.1 An Agentic AI “Inflation Reporter” for the Euro-Area Big-4

Producing timely, consistent, and well-documented analyses of inflation developments across major economies is a recurring task for central banks. While conceptually straightforward, this activity requires the repeated execution of several operational steps—such as data retrieval, validation, visualization, and synthesis—which are time-consuming and may lead to fragmentation or inconsistencies when carried out manually.

The motivation for developing an agentic AI system is therefore not to replace economic judgement, but to support it. By delegating routine and well-defined components of the reporting process to an automated framework, economists can concentrate on interpretation, cross-country comparison, and policy-relevant assessment, while benefiting from improved consistency, reproducibility, and timeliness of standard analytical outputs.

The realization of our agentic AI reporter is based on a multi-phase cooperation with two different LLMs. In the first phase, we interact with ChatGPT (LLM1) to design and develop the software components required to construct the Inflation Reporter. This interaction is guided by a set of high-level functional requirements and focuses on translating analytical needs into an operational reporting framework. In the second phase, the previously developed software⁹ is autonomously used to carry out the reporting tasks originally envisaged, including data collection and processing, visualization, and finally, by means of an open source LLM (LLM2)¹⁰ for the generation of a concise narrative summary. In this sense, the system moves from a development-oriented interaction with LLM1 to an execution phase in which the AI application operates as an integrated reporting tool.

We develop an agentic AI system that is divided into the following components:

- programmatically reads the Harmonised Index of Consumer Prices (HICP) (all-items) year-over-year (YoY) series for Germany (DE), France (FR), Italy (IT), Spain (ES) from the ECB Data Portal / SDW
- builds a monthly picture (chart/table) of YoY inflation for each country,
- drafts an analytical briefing.

Data source and series keys (annual rate of change, not seasonally adjusted) are¹¹:

- Germany: ICP.M.DE.N.000000.4.ANR (HICP YoY, monthly)

⁹The initial prompt submitted to LLM1 is shown in Appendix.

¹⁰Querying open source model doesn't require licenses. We have employed a light model from the Llama family with 8 billion parameters.

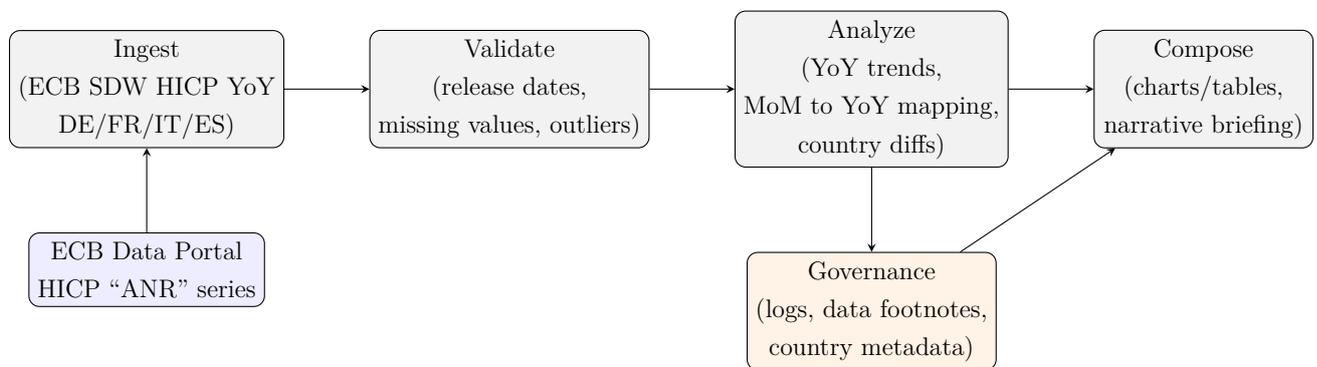
¹¹The names of the time series have been autonomously determined by ChatGPT

- France: ICP.M.FR.N.000000.4.ANR.
- Italy: ICP.M.IT.N.000000.4.ANR.
- Spain: ICP.M.ES.N.000000.4.ANR.

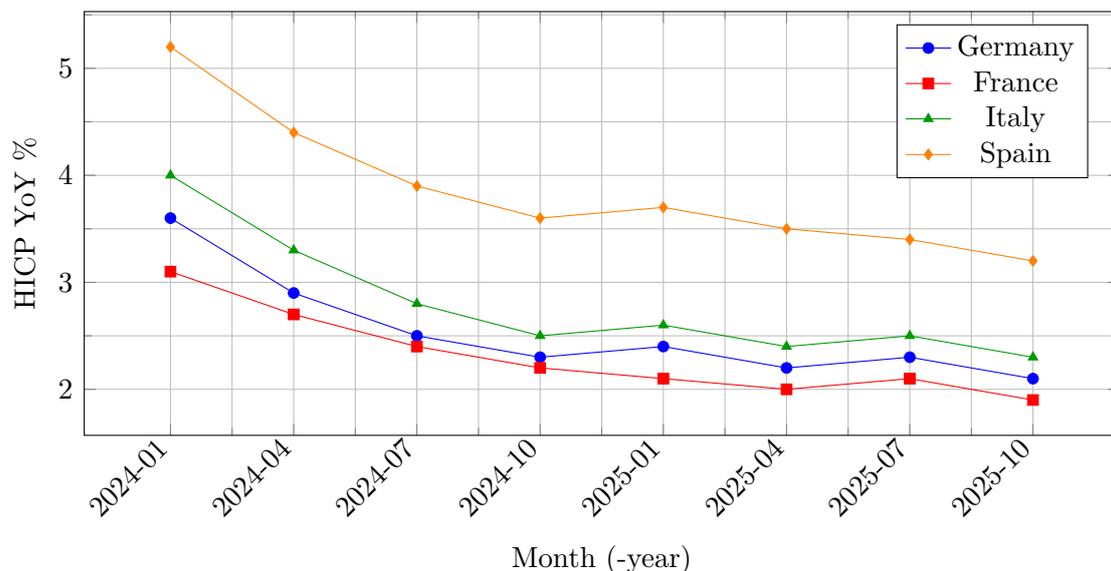
The ECB Data Portal documents HICP concepts, dimensions and latest refresh cycles.¹²

6.2 Pipeline design (agentic workflow)

The agentic workflow is decomposed into four agents with distinct roles: Ingest, Validate, Analyze, Compose. It operates on a monthly cadence aligned to the ECB/Eurostat HICP release (latest “ANR” refresh dates are listed on the ECB Data Portal pages).



After the data discovery there is the call to the drawing agent which takes care of plotting the evolution of the monthly inflation in the four main european economies.



The final step is devoted to writing down a short narrative on the current inflation perspectives building on the official downloaded data.

¹²Programmatic access can be done via the ECB SDW/EDP APIs or community libraries (e.g., the R `ecb` package demonstrates retrieving the exact HICP keys above, including wildcarding for multiple countries)

6.3 Governance & reproducibility

Ensuring accountability and reproducibility is essential for the governance of routine analytical outputs in a central banking context. Policy briefs and monitoring reports often inform internal discussions and decision-making processes, making it necessary that underlying data sources, analytical assumptions, and output vintages can be clearly identified and, if required, reconstructed. For this reason, the design of the agentic reporting system explicitly incorporates mechanisms to document data provenance and preserve successive versions of both data and narrative outputs.

Provenance: Every figure/table footnotes the exact series keys and the “last updated” timestamps shown on the ECB portal (data.ecb.europa.eu).

Versioning: Retain monthly snapshots of CSV & narrative to track any vintage changes (see example of euro-area revision following German updates).

6.4 Sample open source LLM generated briefing

This section summarises recent inflation developments in the four largest euro-area economies, drawing on harmonised consumer price indicators. The narrative provides a concise cross-country comparison of inflation levels and dynamics, followed by short country-specific snapshots. The purpose of this summary is to offer a structured and timely overview of inflation trends, rather than a detailed decomposition of underlying drivers. **Inflation direction** (YoY, HICP all-items): Germany, France, Italy, and Spain remain in the 2–3% corridor in late-2025, with Germany around the mid-2s most recently reported by the Bundesbank; Spain typically at the higher end of the group; France at the lower end; Italy subdued through late-2024 and rebounding modestly in 2025.

Volatility drivers: energy base effects fading; food & services now dominate contributions (country-specific weight updates and category dynamics per HICP methodology).

Data confidence: ECB/Eurostat monthly ANR series; note revisions risk (euro-area aggregate recently re-stated after DE revision).

Country snapshots (YoY HICP, all-items)

Germany — Latest communication places HICP YoY at 2.3% (Oct-2025 estimate), consistent with the ECB “ANR” series trajectory.

France — The ECB France page for ICP.M.FR.N.000000.4.ANR shows low-single-digit YoY through 2025, with a trough in mid-2025 and gentle re-acceleration into autumn.

Italy — The Italy YoY series ICP.M.IT.N.000000.4.ANR documents pronounced disinflation in 2024 (e.g., press noted HICP -1.4% in Dec-2024), stabilising in 2025.

Spain — The Spain YoY series ICP.M.ES.N.000000.4.ANR exhibits a comparatively higher profile among the four, with peaks in mid-2025 and softening thereafter.

6.5 A Financial Stability index for the Banking system

Monitoring financial stability conditions is a core and recurrent activity for central banks, requiring the periodic assessment of a broad set of indicators spanning macro-financial conditions, financial markets and banking-sector soundness. While individual indicators are typically well understood, the need to aggregate, update, and interpret them on a frequent basis makes the process operationally intensive and highly standardised.

Against this background, automation offers clear benefits in supporting routine monitoring tasks. An agentic AI system can be designed to systematically retrieve relevant indicators, apply a consistent aggregation methodology, and generate a concise synthetic measure that facilitates comparison over time. The objective is to provide a transparent and reproducible baseline assessment that can support internal monitoring and communication.

Our second empirical application consists in the construction of a very stylized Financial Stability Index (FSI) with an agentic AI subsystem.

In the appendix we report the different prompts used to query ChatGPT 5.1 for preparing our FSI. ChatGPT prepared the scheme shown in fig. 3. The first step consists

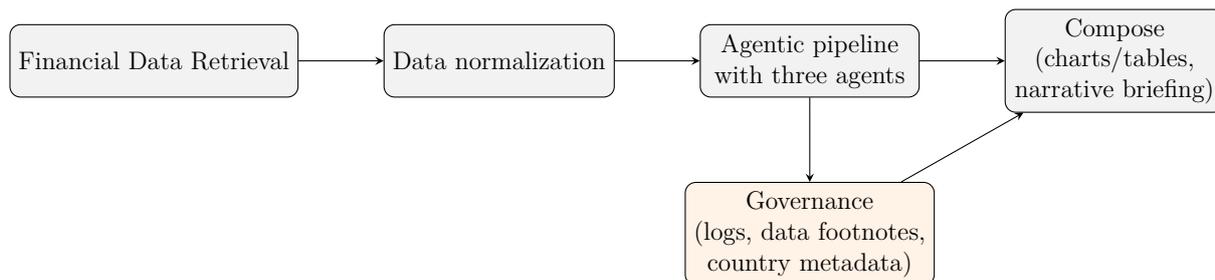


Figure 3: Scheme of agentic AI system for evaluating the Financial Stability Index.

in collecting the required data and successively to normalize these data before starting the operational pipeline. In this example the orchestration of the agents consists in a consecutive scheduling of the different steps involved in the policy report preparation.

6.6 The adopted variables

As in the inflation-reporting application, the development of the Financial Stability Index (FSI) follows a structured, multi-step process in which the initial design choices are supported by an interaction with a first LLM (ChatGPT). In this first stage, the role of the LLM is to assist in identifying a parsimonious set of indicators that capture key dimensions of financial stability, drawing on established central banking practice and standard monitoring frameworks. This interaction provides a transparent and replicable starting point for the construction of the index, while leaving the final specification and interpretation under the responsibility of the analyst.

The choice of the relevant variables for our FSI come from an interaction with our LLM. The following list shows the suggested variables:

- C_t : **total Banking credit** (total loans outstanding of Italian banks to all residents);
- NPL_t **Total non-performing loans**;
- $CET1_t$: **Common Equity Tier 1** capital ratio (Normalized with Risk Weighted Assets);
- LCR_t : **Liquidity Coverage Ratio**;
- MR_t : Market based risk indicator;

These variables are successively **risk-oriented** so that the higher is worse.

- $x_{1,t} = NPLR_t$ Higher NPL ratio = more credit risk \rightarrow worse;
- $x_{2,t} = -CET1_t$ Higher NPL ratio = more credit risk \rightarrow worse
- $x_{3,t} = -LCR_t$ Higher LCR = safer; negative sign makes lower LCR \rightarrow higher risk;
- $x_{4,t} = MR_t$ Higher market spread / volatility = worse by construction.

6.7 Z-score standardization and final variables

Before aggregating the selected variables into a single FSI, it is necessary to place them on a common scale. The underlying variables differ in units of measurement, volatility, and typical ranges, making direct aggregation inappropriate and potentially misleading. Standardisation ensures that each variable contributes to the composite index in a comparable manner, allowing the resulting measure to reflect co-movements in financial conditions rather than differences in scale. In this way, the composite measure captures changes in overall financial stability while preserving the informational content of the underlying variables. We choose the whole available time range as **baseline sample** and compute means and standard deviations for all our variables¹³ :

$$\mu_i = \frac{1}{|T_0|} \cdot \sum_{t \in T_0} x_{i,t} \quad (1)$$

The standard deviation will be

$$\sigma_i = \sqrt{\frac{1}{(T_0 - 1)} \sum_{t \in T_0} (x_{i,t} - \mu_i)^2} \quad (2)$$

Our normalized variables will be:

$$z_t^{NPLR} = -\frac{NPLR_t - \mu_{NPLR}}{\sigma_{NPLR}} \quad z_t^{CET1} = \frac{CET1_t - \mu_{CET1}}{\sigma_{CET1}} \quad (3)$$

¹³In our example we have chosen the period 2015-2025.

$$z_t^{LCR} = \frac{LCR_t - \mu_{LCR}}{\sigma_{LCR}} \quad z_t^{MR} = \frac{MR_t - \mu_{MR}}{\sigma_{MR}} \quad (4)$$

The values are directly interpreted as:

- $z_{i,t} = 0$ baseline risk level;
- $z_{i,t} = +1$ one standard deviation *worse* than baseline;
- $z_{i,t} = -1$ one standard deviation *better* than baseline;

6.8 Weighting scheme and the FSI index

In our example the AI assistant suggests us to combine the four variable in a weighted average. A typical macro-financial approach assigns weights reflecting systemic importance.

Using the weights shown in table 1 we can build our Financial Stability Index as shown

Indicator	Weight	sign
NPL ratio	0.30	negative
CET1	0.30	positive
LCR	0.20	positive
MR	0.20	negative

Table 1: The weights chosen by the AI assistant

in equation 5 whose plot is reported in the graph 4:

$$FSI_t = 0.3 \cdot z_t^{NPLR} + 0.3 \cdot z_t^{CET1} + 0.2 \cdot z_t^{LCR} + 0.2 \cdot z_t^{MR} \quad (5)$$

where:

- **FSI > 0** above normal stability;
- **FSI < 0** below normal stability;
- **-1 < FSI < 1** within normal stability;
- **FSI < -2** strong systemic stress;
- **FSI > 2** exceptional resilience;

At this point we have all the elements for producing our narrative. We put together a multimodal prompt containing some explanatory text and the graph generated by our plotting agent. Here it is quite straightforward to generate a short narrative for this graph by asking a small open source LVM to comment the FSI profile. For this example

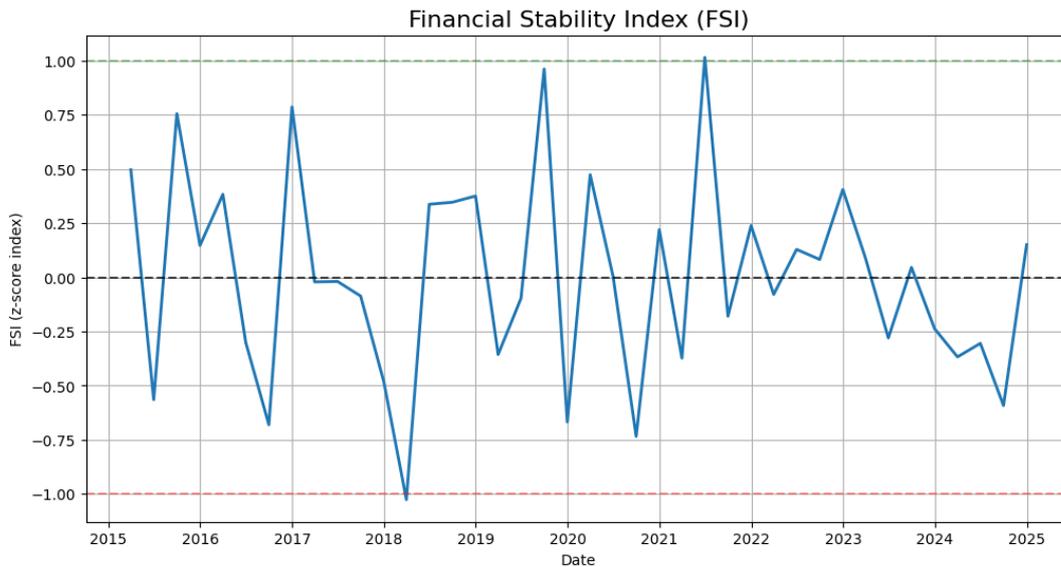


Figure 4: Time profile of the financial stability banking index

we have chosen Llava-v1.6-Vicuna-7b¹⁴ which allow us to prompt the model with pictures and text. The prompt employed to comment the graph and the generated narrative is detailed in appendix. Here it is enough to emphasize that the model correctly pinned down the time range, peaks and troughs, and the trends.

7. Concluding Remarks

The promise of agentic AI in economics and finance is enormous, but so are the challenges. Used wisely, these systems can streamline the development of economic reports and reduce inefficiencies. However, excessive reliance on AI-driven decision-making without robust human oversight could undermine trust, introduce new risks, amplify biases, exacerbate discrimination, and contribute to volatility and systemic instability in financial markets.

In this paper, we have shown how the new paradigm of agentic AI can be harnessed to develop short-term outlooks efficiently and at scale. We described the various phases involved in designing short-term policy reports that draw heavily on multimodal AI systems and autonomous task coordination. We have also demonstrated how the availability of stable, shared metadata systems can foster swift and reliable data discovery.

AI-driven autonomy in economic and financial analysis is no longer a futuristic concept—it is already becoming an integral part of decision-making workflows. The sustainable path forward requires investments not only in adopting agentic AI, but also in cultivating the right skill mix, building high-performance explainability frameworks, strengthening algorithmic accountability, and developing rigorous governance models. It is equally important to reflect on—and actively debate—the broader socio-economic and public policy implications of the widespread deployment of AI agents in our economies.

¹⁴This is the 7 billion version of the multimodal language model.

Ultimately, it is of paramount importance to review and adapt the organisational and governance frameworks of our public institutions so they can effectively meet the challenges posed by this new technological landscape.

Data Availability

The data used for this work are essentially the prompt employed for querying the LLMs. A sample of these prompts are available in the Appendix.

Acknowledgments

The author thanks Juri Marcucci and Sivia Vori for helping in the preparation of the manuscript. I am responsible for all remaining errors

References

- Aakanksha Chowdhery, e. a. (2022). PaLM: Scaling Language Modeling with Pathway. *ArXiv*, 1–87.
- Abdaljalil, S., Kurban, H., Quaraqe, K., & Serpedin, E. (2025). Theorem-of-Thought: A Multi-Agent Framework for Abductive, Deductive, and Inductive Reasoning in Language Models. *ArXiv*, 1–9.
- Abdullah, A. A., & et al. (2025). Evolution of Meta’s Llama models and Parameter-Efficient Fine-Tuning of Large Language Models: a Survey. *ArXiv*, 1–20.
- Acharya, B., K., K., & B., D. (2025). Agentic AI: Autonomous Intelligence for Complex Goals- a comprehensive Survey. *IEEE Access*, 18912–18936.
- Awashti, Y. (2025). gentic AI Redefined: A New Paradigm in Artificial Intelligence. *Journal of Software Engineering and Simulation*, 79–86.
- Bandi, A., & et al. (2025). The Rise of Agentic AI: A Review of Definitions, Frameworks, Architectures, Applications, Evaluation Metrics, and Challenges. *Future Internet*, 1–50.
- Bao, h., & et al. (2021). BEIT: BERT Pre-Training of Image Transformers. *ArXiv*, 1–18.
- Blei, D., A. Y., N., & M. I., J. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Bommasani, R., & et al. (2022). On the Opportunities and Risks of Foundation Models. *ArXiv*, 1–214.
- Chen, K., & et al. (2025). A Survey of Inductive Reasoning for Large Language Models. *ArXiv*, 1–20.
- DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *ArXiv*, 1–22.

- Deerwester, D., A. Y., N., & M. I., J. (1990). Using Latent Semantic Analysis to Improve Access to Textual Information. *Journal of Americal Society for Infomation Science*, 41, 391–407.
- Derouiche, H., Zachi, B., & Haithem, M. (2025). Agentic AI Frameworks: Architectures, Protocols, and Design Challenges. *ArXiv*, (4), 387–414.
- Dosovitskiy, A., & et al. (2021). An Image is worth 16X16 words: Transformers for Image Recognition at Scale. *ArXiv*, 1–22.
- Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science*, 14, 179–211.
- Heer, J., & et al. (2019). Agency plus automation: Designing Artificial Intelligence into Interactive Systems. *Proceeding National Academy Science*, 1844–1850.
- Hou, X., Y., Z., S., W., & Wang, H. (2025). Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions. *ArXiv*, 1–37.
- Ji, Z., & et al. (2023). Survey of Hallucination in Natural Language Generation. *ACM computing surveys*, (55), 1–38.
- Jia, C., & et al. (2021). Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. *ArXiv*, 1–14.
- LangGraph. (2024). LangGraph: Build resilient language agents as graphs. <https://docs.langchain.com/langgraph>, 1–xx.
- Lewis, P., & et al. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *ArXiv*, 1–19.
- Li, G., & et al. (2023). CAMEL: Communicative Agents for “Mind” Exploration of Large Language Model Society. *ArXiv*, 1–77.
- Masterman, T., S., B., M., S., & A., C. (2024). The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey. *ArXiv*, 1–13.
- Mikolov, T., & et al. (2013). Efficient Estimation of Word Representations in Vector Space.. *ArXiv*, 1–12.
- Morishita, T. e. a. (2023). Learning Deductive Reasoning from Synthetic Corpus based on Formal Logic. *ArXiv*, 1–22.
- Nakano, R. e. a. (2022). WebGPT: Browser-assisted question-answering with human feedback. *ArXiv*, 1–32.
- Omar Khattab, e. a. (2023). DSPy: Compiling Declarative Language Model Calls into self-improving Pipelines. *ArXiv*, 1–87.
- OpenAI. (2024). GPT-4 Technical Report. *ArXiv*, 1–100.
- OpenAI. (2025). GPT-5 System Card. <https://cdn.openai.com/gpt-5-system-card.pdf>, 1–60.
- Ouyang, L., & et al. (2022). Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems*, 27730–27744.
- Patil, S., & et al. (2024). Gorilla: Large Language Model Connected with Massive APIs. *Associatin for Computing Machinery*, (4020), 126544–126565.

- Qingyun, W. e. a. (2023). AutoGen: Enabling next-gen LLM Applications via Multi-Agent Conversation. *ArXiv*, 1–43.
- Qwen-Team. (2025). Qwen2.5-Omni Technical Report. *ArXiv*, 1–19.
- Radford, A., & et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. *ArXiv*, 1–48.
- Rawte, V., Amit, S., & Amitava, D. (2023). A Survey of Hallucination in Large Foundation Models. *ArXiv*, 1–11.
- Renqiu Xia, e. a. (2025). ChartX ChartVLM: A Versatile Benchmark and Foundation Model for Complicated Chart Reasoning. *ArXiv*, 1–27.
- Sado, F., & et al. (2023). Explainable Goal-driven Agents and Robots - A Comprehensive Review. *ACM Computing Surveys*, 1–41.
- Schick, T., & et al. (2023). Toolformer: Language Models Can Teach Themselves to Use Tools. *ArXiv*, 1–17.
- Shinn, N., Gopinath, F., Gonzalez, J., & Narasimhan, K. (2023). Reflection: An Autonomous Agent with Verbal Reinforcement Learning. *ArXiv*, 1–19.
- Stiennon, N., & et al. (2020). Learning to Summarize with Human Feedback. *Advances in Neural Information Processing Systems*, 3008–3021.
- Sutskever, I., O., V., & Q.V., L. (2014). Sequence to Sequence Learning with Neural Networks. *ArXiv*, 179–211.
- Tom B. Brown, e. a. (2020). Language Models are Few-Shot Learners. *ArXiv*, 1–75.
- Vasilios, M. (2024). LangChain. *Research Gate*, 1–15.
- Vaswani, A., & et al. (2017). Attention is All you Need. *ArXiv*, 1–15.
- Wang-Zhou, D., Qiuling, X., Yang, Y., & Zhi-Hua, Z. (2019). Bridging Machine Learning and Logical Reasoning by Abductive Learning. *NeuroIPS proceedings*, 1–12.
- Xuemei, D., & et al. (2023). C3: Zero-shot Text-to-sql with ChatGPT. *ArXiv*, 1–15.
- Yao, S., & Jeffrey Zhao, e. a. (2023). REACT: Synergizing Reasoning and Acting in Language Models. *ArXiv*, 1–33.
- Yilin Ye, e. a. (2024). Generative AI for visualization: State of the art and future directions. *Visual Informatics*, (8), 43–66.
- Yin, S., & et al. (2024). A Survey on Multimodal Large Language Models. *Arxiv*, 1–18.
- Yuan Tian, e. a. (2024). ChartGPT: Leveraging LLMs to Generate Charts from Abstract Natural Language. *Visual Informatics*, 43–66.

Appendix

A. Input prompts

In the following you have the textual prompt used to build our agentic AI systems for the computation of an Inflation reporter and a Financial Stability Index.

A.1 The Inflation Reporter

You are a very skilled Central Bank economist. I would like to create an agentic AI system to build an inflation reporter for the four largest European Economies (Germany, France, Italy and Spain). Would you please devise such an agentic system by considering a data retrieval agent, a plotting agent and finally a generative narrative agent providing some comments for the different countries. For the time series you can use the Statistical Data Warehouse of the ECB. Please provide me a clean code that I can immediately test on the field.

A.2 The Banking Financial Stability Index

<input prompt>

Hi friend. Good afternoon. I know you are a very knowledgeable economic policy advisor. I would like you to build a financial stability index for the Italian Banking system. You should provide me an agentic AI system which picks the right variables from the Statistical Data warehouse and then you build the financial stability index considering the whole banking credit and the total Non Performing Loan. I am confident you will provide me a simple set of code snippets to build these agents and test them on a machine equipped with a robust GPU A100.

B. Employing a Multimodal Model

In the following code snippet we show how to put together an image and some textual prompt to generate the short narrative for the evolution of the banking Financial Stability Index.

```
1 model_name = "llava-hf/llava-v1.6-vicuna-7b-hf"
2 processor = LlavaNextProcessor.from_pretrained(model_name)
3 # Load model
4 model = LlavaNextForConditionalGeneration.from_pretrained(
5     model_name,
6     device_map="auto", # CPU or GPU
7     torch_dtype=torch.float16, # or float32 if CPU only
8 )
9 image = Image.open("FSI_index1.png")
10 prompt = "USER: <image>\nPlease comment the evolution of this Financial
11     Stability Index.\nASSISTANT:"
12 inputs = processor(images=image, text=prompt, return_tensors="pt").to(model
13     .device)
14 # Generate
15 output = model.generate(**inputs, max_new_tokens=300)
16 response = processor.batch_decode(output, skip_special_tokens=True)[0]
17 print(response)
```

Listing 1: Python code for feeding a Multimodal Model

Here we have employed Llava: the visual version of Llama for the simple availability of examples. Many other multimodal models might be chosen, but unfortunately prompt composition is not yet a standard task.

B.1 Generated FSI graph comment

In the following we list the explanatory paragraph generated by our Multimodal model. *From the graph, we can observe the following:*

1. ***Peaks and Troughs**: The index shows periods of both stability and volatility. There are peaks that indicate periods of financial stability, and troughs that suggest periods of financial stress or instability.*
2. ***Recent Trends**: The most recent data points show a decline in the index, which could indicate a deterioration in the stability of the financial system.*
3. ***Time Frame**: The graph covers a period from 2016 to 2024, which suggests that the data is current as of the time the graph was created.*
4. ***Date Labels**: The graph has date labels along the horizontal axis, which help to contextualize the data points and understand the timing of the index's fluctuations.*
5. ***Interpretation**: The exact meaning of the index values would depend on the specific methodology used to calculate it. Generally, higher values indicate greater stability*