



From equations to prompts: integrating LLMs into a hybrid macroeconomic ABM

Georgios Papadopoulos

371

OCTOBER 2026

WORKINGPAPERWORKINGPAPERWORKINGPAPERWORKI

BANK OF GREECE
Economic Analysis and Research Department – Special Studies Division
21, E. Venizelos Avenue
GR-102 50 Athens
Tel: +30210-320 3610
Fax: +30210-320 2432

www.bankofgreece.gr

Published by the Bank of Greece, Athens, Greece
All rights reserved. Reproduction for educational and
non-commercial purposes is permitted provided that the source is acknowledged.

ISSN: 2654-1912 (online)
DOI: <https://doi.org/10.52903/wp2026371>

FROM EQUATIONS TO PROMPTS: INTEGRATING LLMS INTO A HYBRID MACROECONOMIC ABM

Georgios Papadopoulos
Bank of Greece

ABSTRACT

Agent-based models (ABMs) typically rely on equations to incorporate agents' decision-making, ranging from ad-hoc rules or simple heuristics to complex, theory- or data-driven approaches. Recent advances in large language models (LLMs) make it possible to represent economic decision-making through prompts to LLM-driven agents instead of explicitly specified behavioural equations. However, whether such agents can feasibly replace individual components in an ABM and the degree to which they can replicate observed economic dynamics remains insufficiently understood. In particular, it is unclear whether the integration of LLM-driven agents in a macroeconomic ABM can generate historically plausible macroeconomic trajectories and which factors affect their performance. This paper addresses these questions by introducing an LLM-driven bank agent into a data-driven macroeconomic model of the Greek economy, where it determines consumer credit and lending rates. Results show that the LLM-driven agent's decisions can reproduce historical consumer credit dynamics relatively well, although their performance in replicating lending rates is weaker and more heterogeneous; broader economic information generally improves performance, while larger models do not systematically outperform smaller ones. Overall, the study provides encouraging evidence that LLM-driven agents can be feasibly integrated into macroeconomic ABMs, assimilate economically relevant information and produce outcomes that closely match observed data, without requiring explicit functional forms for behavioural relationships - opening the possibility of enabling economic simulations to shift from equations to prompts.

Keywords: LLM-driven agent, agent-based model, data-driven ABM, household credit

JEL classification: C63, E17, G21

Disclaimer: The author would like to thank the participants of the 32nd International Conference Computing in Economics and Finance (Ca' Foscari University of Venice, Italy, July 2026) for useful comments and suggestions. The contents of the paper express the views of the author and do not necessarily reflect the position or opinion of the Bank of Greece or the Eurosystem.

Correspondence:

Georgios Papadopoulos
Economic Analysis and Research Department
Bank of Greece
El. Venizelos 21, 10250, Athens, Greece
Tel.: +30-2103202415
Email: georpapadopoulos@bankofgreece.gr

"Henderson said, "Multivac had nothing to do with victory. It's just a machine."

"A big one," said Swift.

"Then just a big machine. No better than the data fed it."

Isaac Asimov^a

^a*The machine that won the war* in Fantasy & Science Fiction. Vol. 21, no. 4. New York: Mercury Press, 1961 pp. 51–56. ISSN 1095-8258

1. Introduction

Agent-based models (ABMs) offer a natural framework for studying macroeconomic dynamics as the outcome of interactions among heterogeneous decision-makers. Yet they depend critically on how the behaviour of these agents is represented. Conventional macroeconomic ABMs typically rely on behavioural equations, heuristics, or explicitly parameterized decision rules (Dawid and Gatti, 2018; Axtell and Farmer, 2025; Farmer, 2025). Although these approaches are transparent and computationally tractable, they necessarily impose a particular functional form on how agents process information and make decisions. Moreover, the choice of such behavioural rules is often based on common sense and ad hoc assumptions (Farmer and Foley, 2009; Farmer, 2025). Recent advances in large language models (LLMs) provide an alternative approach to representing, at least some, economic decisions. Rather than specifying ex ante how an agent should respond to a particular set of circumstances, the modeller can provide the agent with relevant information and allow the LLM to generate the corresponding decision. This possibility is particularly relevant given the growing experimental evidence that LLMs can reproduce several aspects of human economic behaviour. LLMs perform very well on standardized tests of economics knowledge (Geerling et al., 2023), exhibit substantial economic rationality in budgetary choice experiments (Chen et al., 2023), and can reproduce broad patterns observed in laboratory economic experiments, including deviations from rational expectations (del Rio-Chanona et al., 2025). Horton et al. (2023) consequently describe LLMs as implicit computational models of human behaviour¹, referring to this approach as “Homo silicus” and arguing that their behaviour can be studied through simulation in a manner analogous to the use of Homo economicus. The relevant question for macroeconomic modelling is not simply whether LLMs can make plausible individual decisions, but whether such decisions can be embedded in an economic environment and generate aggregate dynamics that resemble those observed in the real world.

Despite the rapidly expanding literature on LLM-based agents² and agent-based simulation (Gao et al., 2024, and references therein), evidence on these questions remains limited, particularly in the context of macroeconomics³. To the best of the author’s knowledge, the only existing application in that field is the study by Li et al. (2024), which demonstrates that LLM-driven household decisions can generate plausible macroeconomic dynamics and stylized facts within an artificial economic environment. Nevertheless, relatively little is known about whether LLM-driven agents

¹Park et al. (2024) used interviews with 1,052 individuals to create AI agents designed to simulate their attitudes, personality, and decision-making. Remarkably, the generated agents matched participants’ answers on the General Social Survey (GSS) 85% as accurately as participants matched their own answers when retested two weeks later.

²For a recent review on the use of AI agents in economic research as semi-autonomous research assistants, the interested reader is referred to Korinek (2025) and references within.

³One recent example that illustrates the growing interest of policymakers is Project Logos (<https://www.bis.org/project/logos>), launched by the BIS in 2026. The project aims to enable central banks to observe and analyse the behaviour of LLM-based agents acting as portfolio managers within a financial ABM environment.

can replace conventional behavioural rules in a macroeconomic ABM while retaining its ability to reproduce observed economic dynamics. Similarly, it remains unclear which characteristics of the LLM agent and its information environment are most relevant for determining the quality of the resulting decisions.

This paper addresses these questions by integrating an LLM-driven loan officer into a data-driven macroeconomic ABM and assesses whether the hybrid model can reproduce the observed dynamics of consumer credit and lending rates. The model builds on Papadopoulos (2019) and retains its data-driven structure, in which several components of the economy are anchored to historical time series. The main structural extension consists of replacing the conventional representation of bank lending decisions with an LLM-driven agent that determines both the amount of consumer credit offered to each household and the corresponding lending rate. The agent receives borrower-specific information and, depending on the specification considered, additional information concerning bank capitalization and the broader economic environment. Importantly, the prompts provide the agent with relevant information without prescribing how it should respond to changes in these variables. Regulatory constraints and the policy rate define the boundaries within which the decision is made, while the lending decision itself is generated by the LLM. The analysis considers a range of open-weight models spanning several orders of magnitude in parameter size, from models that can operate on consumer-grade hardware to frontier-scale models, and supplies these models with alternative information sets. The simulations are conducted for the Greek economy over the period 2000-2024, providing a long historical setting in which to assess whether an LLM-driven decision rule can reproduce the realized dynamics of consumer credit and the associated lending rates.

The results provide evidence that LLM-driven agents can be integrated into a macroeconomic ABM and generate economically plausible aggregate outcomes. At the same time, the findings indicate that their performance depends on both the economic variable being modelled and the information provided to the agent. This result highlights that the information set provided to the agent is an important determinant of the quality of the resulting behavioural representation.

This paper contributes to the emerging literature on LLM-based economic agents in dynamic simulation environments. Existing evidence indicates that LLMs can exhibit human-like or economically meaningful behaviour in controlled experiments, including rationality, bounded rationality, and responses to economic information (Geerling et al., 2023; Chen et al., 2023; del Rio-Chanona et al., 2025). Other recent studies show that LLMs can reproduce important features of economic forecasting (Hansen et al., 2026) and can be used to represent complex decision-making processes, such as those associated with the Federal Open Market Committee (Kazinnik and Sinclair, 2025).

The current study extends this literature by embedding an LLM-driven agent within an explicit economic environment, allowing its decisions to influence credit-market outcomes throughout a long historical simulation and evaluating the resulting aggregate series against their real-world counterparts. The comparison across model sizes and information sets further indicates that the performance of an LLM agent extends beyond its underlying model capability. Importantly, the information input is itself a significant component of the behavioural representation.

In this respect, the analysis provides an empirical step towards a modelling framework in which conventional behavioural equations can, at least for some economic decisions, be replaced by information-conditioned LLM decisions. The results are encouraging, but they also identify certain challenges. The approach performs substantially better for some aggregate outcomes than for others, shows dependence on the information provided to the agent, and does not yet represent the economy as a system in which all agents are driven by LLM-based decision rules.

The remainder of the paper is organized as follows. Section 2 describes the hybrid macroeconomic ABM and the LLM-driven bank agent. Section 3 presents the data and calibration of the

Greek economic scenario. Section 4 reports the simulation results, first evaluating the ability of alternative LLM-prompt combinations to reproduce consumer credit and lending rates and then examining the role of model size and prompt information. Section 5 discusses the findings, their implications, and the limitations of the approach. Finally, the appendices provide additional details on the feasibility assessment, model selection, and prompts.

2. Model

This section describes the structure of the model, focusing on the agents involved and the rules governing their interactions across the different markets. The model builds on the agent-based model developed in Papadopoulos (2019), retaining most of its original mechanisms and data-driven nature. Consequently, the model remains closely anchored to the observed timeseries. However, this study introduces a fundamental structural extension from a modelling perspective; an LLM-driven agent replaces one of the input series employed in the original model. More specifically, the LLM-driven agent takes on the role of a loan officer at a commercial bank, deciding how much consumer credit to extend to borrowing households and at what interest rate, based on information provided through different prompts. The main components of the model, together with this extension, are described in the following sections.

2.1. Initialisation

The purpose of the initialisation process is to reconstruct the prevailing economic conditions at the selected starting point of the simulation and embed the model within this environment. In this way, the simulation begins from a state that is as representative as possible of the specific country being modelled. The data-driven character of the agent-based model is first reflected in the initialisation of household income. Based on historical data, income is allocated across households according to a $\Gamma(\alpha, 1/\lambda)$ distribution. Empirical evidence indicates that this functional form provides an adequate representation of income distributions across several countries over time (Bandourian et al., 2002). Furthermore, adjusting the shape (α) and scale (λ) parameters enables the simulated distribution to reproduce the historical one in terms of its minimum and average values, as well as its dispersion. Unemployment represents another key variable that is initialised using historical information. The share of unemployed households in the simulation is matched to the real-world counterpart of unemployed persons in the total working population. Consequently, once income has been distributed, the required number of households is selected randomly and their income is set equal to the unemployment benefit.

2.2. Sequence of events

The timeline of events in the simulation is the following:

1. Historical data of household- and bank-related variables are updated.
2. Households collect interest from any deposits they might keep at the bank.
3. Employed households receive their monthly income according to the historical figures of income's growth.
4. The real-world data on unemployment determine the employment status of households.
5. Households form their income expectations.
6. Households try to meet any financial obligations they might have and form their desired consumption plans.
7. Households without the necessary financial resources to achieve the desired level of consumption ask for credit.

8. The bank decides the amount of credit it will provide along with the associated interest rate and makes its offer to the potential borrower.
9. Households decide how much to borrow and the loan's characteristics (size, interest rate) are established.
10. Households fulfill their consumption plans to the maximum extent possible and deposit any excess funds that might remain.

2.3. Agent and market description

The following subsections describe in detail the different types of agents and their behavioural rules that determine their interactions in the various markets.

2.3.1. Households

Expectations formation

Each time step of the simulation starts with households receiving their monthly income exogenously. Their first action is to form expectations about future income. These expectations play a central role in the subsequent determination of households' desired consumption and, consequently, their demand for credit.

Consistent with the literature on expectations formation, households form their income expectations under bounded rationality. More specifically, their expectations are governed by the heuristics switching model (HSM) developed by Anufriev and Hommes (2012a,b). Within the HSM framework, agents rely on a limited set of simple forecasting heuristics and, at each time step, select one of them according to its past performance and agent-specific behavioural characteristics.

The HSM encompasses several established empirical regularities associated with expectations formation within a single framework, including heterogeneity in expectations across individuals (Allen and Taylor, 1990; Mankiw et al., 2003; Branch, 2004; Fehr and Tyran, 2008; Pfajfar and Santoro, 2010; Pfajfar and Žakelj, 2018, among others) and the evolutionary selection of forecasting rules according to their relative performance (Marimon and Sunder, 1995; Arifovic and Sargent, 2003). Moreover, the model has been shown to provide a good fit to both experimental and survey data across a range of settings, spanning financial and macroeconomic applications (Hommes et al., 2005, 2008, 2019; Assenza et al., 2021; Hommes, 2021)⁴. Appendix E elaborates on the mechanics of the HSM.

Desired consumption

Once their income expectations have been formed, households determine their consumption plans. Desired consumption ($C_{h,t}^d$) is determined by expected income and by the household's capacity to maintain itself. The latter depends on whether the household can sustain a level of consumption equal to the required subsistence level, which in turn affects its solvency. This behaviour is captured by the rule presented in Equation 1.

$$C_{h,t}^d = \begin{cases} C_{min,t}, & \text{if } I_{h,t} + D_{h,t-1} < LP_{h,t} + C_{min,t}, \\ \max\{\alpha_y \cdot I_{h,t+1}^e + \alpha_w \cdot D_{h,t-1}, C_{min,t}\}, & \text{if } I_{h,t} + D_{h,t-1} \geq LP_{h,t} + C_{min,t}, \end{cases} \quad (1)$$

with $0 < \alpha_w < \alpha_y < 1$ denoting the marginal propensities to consume out of wealth and income respectively.

⁴The work of Bao et al. (2021) provides a recent review of the literature on the experimental study of expectations formation in macroeconomics and finance.

The amount of liquid wealth held by households determines both their solvency and whether they need to seek additional credit. In particular, current income ($I_{h,t}$) together with accumulated deposits ($D_{h,t-1}$) must be sufficient to maintain the minimum consumption level ($C_{min,t}$). If, however, a household is unable to meet its monthly loan payment ($LP_{h,t}$) while maintaining subsistence consumption, it chooses to reduce consumption as much as possible and forgo any loan payments that may be due. Alternatively, when the household can service its debt, it attempts to consume an amount that depends on its expected income ($I_{h,t+1}^e$) and accumulated deposits. Households first rely on their own financial resources to achieve their desired consumption level. When current income and withdrawals from deposits are insufficient, they subsequently turn to the bank to obtain consumer credit. The Modigliani consumption function (Modigliani and Brumberg, 2005) specified in Equation 1, is commonly adopted in the ABM literature (Assenza et al., 2015; Caiani et al., 2016; Russo et al., 2016; Gurgone et al., 2018; Meijers et al., 2019; Reissl, 2020, 2021; Boitani et al., 2026, among others)⁵. Interestingly, existing research indicates that simple linear, rule-of-thumb consumption functions can provide a good approximation to consumption functions, such as those derived from buffer stock theory, that arise as solutions to utility maximization problems (Allen and Carroll, 2001)⁶.

2.3.2. Bank

A single bank agent represents the economy’s banking system. The bank takes households’ deposits paying an interest of r_t^D and provides them with consumer credit after evaluating their requests. In particular, an LLM-driven agent who assumes the role of the bank’s loan officer decides on the size ($L_{h,t}$) and interest rate ($r_{h,t}^L$) of the provided credit for each potential borrower based on information on borrower-specific, bank-specific and external factors.

Employed LLMs

The selection of LLMs integrated into the ABM focuses on open-weight models that can, subject to available hardware, be executed either locally or through Ollama’s⁷ cloud infrastructure. This provides a common deployment framework while allowing the analysis to consider models with substantially different computational requirements.

The selected LLMs range from relatively small ones that can run on consumer-grade hardware to very large, frontier-scale models that require cloud-class computing resources. Specifically, the analysis considers Gemma 3 in its 4B, 12B and 27B variants (Gemma Team, 2025), Gemma 4’s 31B variant (Gemma Team, 2026), Ministral 3 in its 3B, 8B and 14B variants (Liu et al., 2026), Mistral Large 3 with 675B total parameters⁸, and Qwen3-Coder 480B⁹.

In addition to parameter size, these models differ in their architectures and documented training characteristics. On the one hand, Gemma 3’s models, Gemma 4’s 31B variant and Ministral 3 models have a dense, decoder-only transformers architecture, although they differ in architectural details. On the other hand, Mistral 3 (675B) and Qwen3-Coder (480B) use sparse mixture-of-experts transformers architectures.

Moreover, the available documentation indicates substantial differences in their training data

⁵The work of Dawid and Gatti (2018) provides an overview of the various types of consumption functions implemented in several macroeconomic ABMs.

⁶An important implication of the linear consumption function in Equation 1 is that households maintain a target wealth-to-income ratio of $(1 - \alpha_y)/\alpha_w$. Consequently, when their realized wealth falls below this target, households save to reach it (Godley and Lavoie, 2016).

⁷<https://ollama.com>

⁸Hereafter referred to as *Mistral 3 (675B)* for brevity and described in: <https://docs.mistral.ai/models/model-cards/mistral-large-3-25-12> [Accessed in August 2026].

⁹Described in: <https://qwen.ai/blog?id=qwen3-coder> [Accessed in August 2026].

and objectives. Gemma 3 and 4’s training datasets include web documents, code, images, and -importantly- mathematics, with a cutoff date of August 2024¹⁰ and January 2025¹¹, respectively. As expected, Qwen3-Coder (480B) places particular emphasis on code with approximately 70% of its training dataset consisting of code, with knowledge cutoff date June-July 2025¹². Mistral, in contrast, does not disclose the datasets used to train any of its models. Appendix A provides more information about the LLMs’ computational performance and the factors that guided their selection.

This heterogeneity, and in particular the inclusion of LLMs spanning several orders of magnitude in parameter count, makes it possible to assess whether increasing model scale systematically translates into improved performance, or whether smaller models can achieve comparable results at substantially lower computational cost.

The information sets of the LLM-driven bank agent

To request credit from the LLM-driven loan officer, households use one of four alternative prompts, each of which is used consistently throughout any given simulation exercise. This structure allows the analysis to isolate whether and how different information affects the lending decisions of the former.

The baseline prompt provides borrower-specific information, including the household’s monthly income and any existing loan balance. Arguably, this information is the least controversial since it relates to the potential borrower’s creditworthiness. The bank capitalization prompt extends this baseline by additionally reporting the bank’s current capitalization, proxied by the ratio of capital and reserves to total assets, as well as its annual change in both magnitude and direction. In this way, the LLM-driven agent receives information not only about the characteristics of the borrower but also about the bank’s current financial position and its recent evolution. Indeed, the empirical literature suggests the existence of a link between bank capital adequacy and provided lending rates (Boissay et al., 2019, and references therein) with several macroeconomic ABMs implementing such mechanism (Gatti et al., 2010; Caiani et al., 2016; Cardaci, 2018; Alexandre and Lima, 2020). The economic environment prompt instead supplements the borrower-specific information with the prevailing unemployment rate, together with its annual change in magnitude and direction. It thereby provides the bank agent with an indicator of broader economic conditions, which is likely to be among the types of information that actual loan officers take into consideration. Depending on the LLM’s capabilities, this information could also shape the agent’s internally formed expectations, which may subsequently influence its lending decision. Finally, the full-information prompt combines all information included in the previous prompts, providing the agent simultaneously with borrower-specific characteristics, the bank’s capitalization and its recent evolution, and the current state and evolution of unemployment. All four prompts additionally state the central bank’s current policy rate and the applicable debt-to-income (DTI)¹³ and debt-service-to-income ratio (DSTI)¹⁴ limits. Appendix B provides more details about the prompts supplied to the LLM-driven bank agent, including the various inputs used and their wording.

Importantly, the alternative prompts provide the bank agent with the relevant information but do not prescribe how it should interpret or act upon it. In particular, the prompts report the relevant variables and their recent evolution without specifying whether a given level or change should lead the agent to increase or reduce the loan amount or interest rate offered. The only exceptions concern

¹⁰Described in: https://ai.google.dev/gemma/docs/core/model_card_3 [Accessed September 2026]

¹¹Described in: https://ai.google.dev/gemma/docs/core/model_card_4 [Accessed September 2026]

¹²Not officially disclosed but rather based on the LLM’s release date.

¹³The DTI is the ratio of potential borrower’s outstanding and newly offered debt over its annual income.

¹⁴The DSTI is the ratio of interest payments plus amortisations over the potential borrower’s monthly income.

the constraints that define the boundaries of the lending decision: the agent must comply with the applicable DTI and DSTI limits, while the central bank’s policy rate serves as the lower bound for the interest rate it offers. Subject to these constraints, the lending decision is left to the LLM-driven agent, rather than imposing a predefined response to each piece of information.

2.3.3. Credit market

In the credit market, households without the financial resources to meet their desired consumption level ask for credit from the bank. Equation 2 defines the minimum amount of consumer credit that a household asks:

$$L_{h,t}^{ask} = C_{h,t}^d - (I_{h,t} + D_{h,t-1}) \quad (2)$$

Using the prompts described above, the LLM-driven loan officer decides how much credit to provide and at what interest rate¹⁵. The size of loan offered can be either smaller or, often, larger than the amount asked. In that case, Equation 3 determines the provided amount:

$$L_{h,t} = \begin{cases} L_{h,t}^{offer}, & \text{if } L_{h,t}^{offer} \leq L_{h,t}^{ask}, \\ U(L_{h,t}^{ask}, L_{h,t}^{offer}), & \text{if } L_{h,t}^{offer} > L_{h,t}^{ask}. \end{cases} \quad (3)$$

Consequently, when the loan amount offered is lower than that requested by the household, the latter accepts the amount supplied by the bank. Alternatively, it selects a random amount drawn from a uniform distribution bounded by $L_{h,t}^{ask}$ and $L_{h,t}^{offer}$. Once the household and the bank agree on the loan amount, $L_{h,t}$, the LLM-driven agent is prompted again using the same information set as before, namely the baseline, bank capitalization, economic environment or full-information prompt. This time, however, the agent is instructed to provide the exact loan amount of $L_{h,t}$ and base its calculations on this amount when deciding on the applied lending rate¹⁶.

The stochastic credit assumption mechanism specified in Equation 3 follows the same approach as in Papadopoulos (2019). The intuition underlying the stochastic component of the rule is twofold. On the one hand, the bank has an incentive to offer the maximum permitted loan amount, as this allows it to maximize its interest revenues. On the other hand, households have an incentive to accept a larger amount of credit, since additional borrowing enables them to increase their consumption. From a modelling perspective, this approximation addresses the issue of the implausibly small loan amounts implied by Equation 2, while at the same time allowing the model to reproduce the dynamics of consumer credit satisfactorily.

2.3.4. Labour market

In a full-fledged ABM, firms and households would interact in the labour market, jointly determining, among others, the employment status and income of the latter. Since firms are not explicitly represented in the model, their interactions with households are instead proxied by historical timeseries. The initial monthly income assigned to households is calibrated to reproduce the observed income distribution of the country-specific scenario at the beginning of the simulation. Thereafter, its evolution is governed by the observed growth rate of income in the real economy.

¹⁵If the offered amount and interest rate do not satisfy the DTI, DSTI and minimum interest rate constraints then the prompt is re-submitted for up to ten retries. If the LLM does not yield a valid answer up to that limit, then the simulation proceeds with no change in the asking household’s borrowing.

¹⁶As before, if the offered lending rate does not satisfy the DSTI and minimum interest rate constraints, then the prompt is re-submitted for up to five retries. If the LLM does not yield a valid answer up to that limit, then the simulation assumes the originally offered interest rate.

It is important to note that this imposed income path is applied uniformly across households. Consequently, a change in the aggregate income series results in the same proportional change in the income of each individual household.

The employment status of households is determined in a similar manner using the observed unemployment rate. Changes in aggregate unemployment are translated into changes in the number of unemployed households, with the affected households selected randomly. Unemployed households receive a benefit equal to 80% of the prevailing minimum wage, allowing them to maintain a minimum level of consumption. This modelling choice is primarily driven by the absence of sufficiently granular data describing the monthly evolution of income and unemployment across different household income levels. Nevertheless, the positive skewness of the income distribution has an important implication for this assumption: because unemployment is assigned randomly, households at the lower end of the income distribution will, on average, be affected more frequently than those at the upper end. Hence, despite the absence of an explicit relationship between income and unemployment in the model, the applied mechanism results in an indirect association between the two.

3. Data

A distinguishing characteristic of the ABM developed in this study is its data-driven approach. Rather than representing all components of the economy with the same level of detail, historical data are used to proxy the behaviour of several agents and the outcomes of their interactions in the respective markets. This approach reduces the computational burden of the simulation while allowing the examination of a specific component's performance with the remaining parts of the economy staying anchored to observed data.

The input data are divided into two main categories: scenario data and calibration data. Scenario data are introduced into the simulation and updated at each time step, whereas calibration data are used to initialize the model in accordance with the prevailing economic conditions at a given point in time.

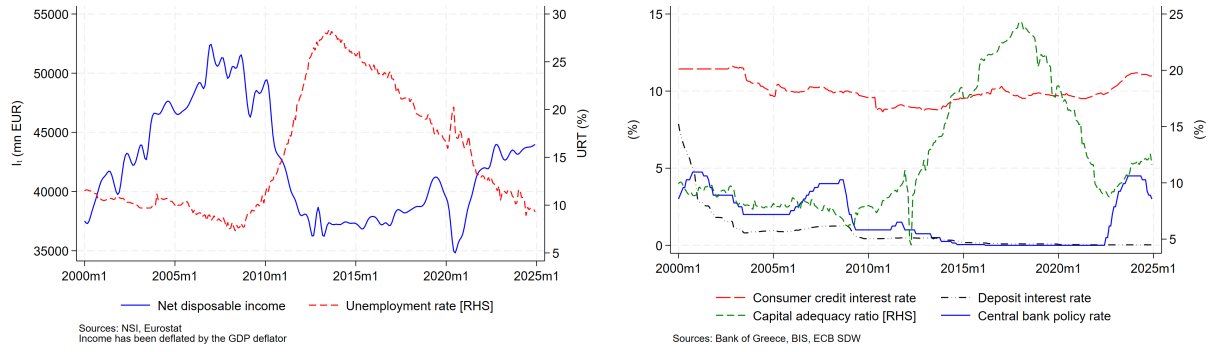
The country-specific scenario considered in this study is based on the Greek economy. This not only demonstrates the model's flexibility in accommodating various economic scenarios, but, more importantly, provides a demanding test of its ability to reproduce the realized evolution of the variables under consideration.

An important observation across both data categories is that their fidelity to the actual economic conditions can significantly affect the results. Naturally, the closer the injected data are to the observed economic conditions, the greater the expected similarity between the simulated and realized series. Consequently, minimizing discrepancies between the model inputs and their real-world counterparts allows for a cleaner assessment of the model's underlying mechanisms and, in particular, of the ability of the LLM-driven agents to reproduce the observed dynamics of consumer credit and lending rates.

3.1. Scenario data

The simulation begins in January 2000 and extends until December 2024, covering twenty five years of data. Each time step in the model represents one month in physical time, therefore any series collected in lower frequencies are converted to monthly using spline interpolation. Figure 1 displays the evolution of the historical series that are injected into the simulation.

Figure 1: Household-related (left panel) and bank-related (right panel) series of the macroeconomic scenario injected into the simulation.



The left panel of Figure 1 presents the series of net disposable income¹⁷ and unemployment, expressed as a percentage of the active population. The latter is originally available at a monthly frequency and is therefore introduced into the simulation using monthly growth rates without any additional processing. In contrast, the raw income data¹⁸ are reported annually and consequently require conversion to a monthly frequency before being used in the model. This transformation is performed in two steps. First, the annual observations are interpolated to a quarterly frequency using the Denton method, with the corresponding seasonally adjusted series of wages & salaries serving as the interpolation pattern. The resulting quarterly series are then deflated using the GDP deflator and subsequently converted to monthly frequency through spline interpolation. The final monthly series are injected into the model in the form of monthly growth rates.

The right panel displays the evolution of bank-related series. These include the banking system's consumer credit and overnight deposit interest rate, the ratio of the sum of bank capital and reserves over total assets, which is used as a proxy for capital adequacy, and the policy rate set by the European Central Bank. The consumer credit interest rate¹⁹, plays a particular role in the analysis. It serves as input only during the calibration of the DSTI series, while it is the benchmark against which the output from the LLM-driven lending rate offers is tested²⁰. All series are originally available at a monthly frequency and are therefore introduced into the model in levels, without requiring interpolation. The only transformation applied concerns the reported annualized deposit interest rates, which are converted to their corresponding monthly values before being used in the simulation.

Overall, both panels in Figure 1 highlight the challenging economic conditions implied by the implemented scenario. Households faced a period of steeply rising unemployment and sharply declining incomes before entering a recovery phase, which has subsequently been briefly interrupted by the COVID-19 pandemic. Similarly, the banking sector underwent a recapitalization followed by a prolonged low interest rate period. This arguably makes for a demanding scenario to simulate and, therefore, a tough test for the LLM-driven agents' ability to replicate the realized credit-related variables.

¹⁷Referred to simply as *income* for brevity for the remainder of the study.

¹⁸The specific series is *IDCM.A.N.GR.W0.S1.S1.B.B6N._Z._Z.XDC.V.N* from the IDCM.

¹⁹More precisely, the series of bank interest rates on outstanding amounts of euro-denominated loans vis-à-vis euro area residents from domestic credit institutions, with over 1 and up to 5 years maturity. Their source is from the Bank of Greece's, publicly available, bank deposit and loan interest rate statistics (Table 2).

²⁰It is worth noting that this is the only series for which historical observations are available starting from September 2002. During the simulation, the missing observations prior to that are replaced with the average value calculated from the first 12 available observations. Naturally, the comparison between the simulated and historical series is conducted only over the period for which actual historical data are available.

3.2. Calibration data

Calibration data are static in nature and are used to initialize the simulation. Their purpose is to embed the model into a specific economic environment and ensure that the simulation begins as closely as possible to the economy it attempts to model.

One part of the calibration data relates to households' income and employment status. The initial distribution of income across households is determined using historical values for minimum and average monthly income, together with the Gini coefficient. Data for the latter are obtained from the World Bank's Poverty and Inequality Platform²¹, while Eurostat's database provides the required data on minimum monthly income²² and the OECD provides data on average annual wages. In addition to the income-related calibration data, information from Eurostat is used to calibrate the initial level of monthly unemployment.

A second subcategory covers the variables related to the banking sector, including deposit interest rates, capital adequacy and the central bank's policy rate. For these variables, the calibration data are drawn from the same sources as the corresponding scenario data, with the calibration based on the observation recorded one month before the beginning of the simulation. Table 1 reports the calibrated values used to define the initial setup for the country-specific scenario.

Table 1: Initial setup of the simulation for the Greek scenario.

Description	Value
Average income	1160
Minimum income	525
Gini coefficient of income distribution (%)	34
Unemployment rate (%)	12.3
Overnight deposit interest rate (%)	5.96
Capital adequacy ratio (%)	10.12
Central bank policy rate (%)	3.0
DSTI (%)	42

3.3. DSTI data

DSTI represents a particular case, as it combines characteristics of both scenario and calibration data. As with the scenario data, the maximum DSTI applied in the model is updated at each simulation time step. Its initial value and subsequent evolution, however, are determined through a calibration procedure described in detail in Appendix D. This calibration is required because precise historical data on the evolution of DSTI are not available. The only publicly accessible source providing a suitable approximation is the BIS database of debt service ratio statistics, which reports quarterly aggregate DSTI series for different countries and sectors using a consistent methodology. As discussed in Drehmann et al. (2015), although the methodology used to construct these series captures the evolution of DSTI reasonably well, the reported levels may differ from the values that would be obtained from micro-level data. Consequently, both the initial level and the trajectory of DSTI need to be calibrated.

The calibration procedure employs a grid search over alternative DSTI paths and associated initial levels, seeking the combination that produces the highest similarity between the consumer credit series generated by the model and its historical counterpart. This approach therefore uses the observed evolution of consumer credit to identify the DSTI trajectory that is most consistent

²¹pip.worldbank.org. [Accessed in December 2025]. UNU-WIDER's World Income Inequality Database, Version of May 2020 provides a similar figure.

²²In particular, Eurostat's series on monthly minimum wages (*earn_mw_cur*).

with the model dynamics. It is worth noting that a key assumption underlying the procedure is that the evolution of the actual DSTI series is sufficiently close to that represented by the aggregate debt service ratio series reported in the BIS database.

3.4. Fixed parameters

Finally, a set of parameters is fixed across every simulation and is not data-related, but their calibration is guided by the literature. These parameters are reported in Table 2.

Table 2: Fixed parameter values.

Description	Parameter	Value
Number of households	N	1000
Propensity to consume out of income	α_y	0.65
Propensity to consume out of deposits	α_w	0.05
WTR expectations extrapolation factor	ω_{wtr}	0.4
STR expectations extrapolation factor	ω_{str}	1.3
ADA expectations parameter	ω_{ada}	0.65
Households' memory strength	η	0.7
Households' expectations rule persistence	δ	0.9
Households' intensity of choice	β	0.4
Loan maturity (months)	m	60

The values assigned to both propensities to consume are consistent with those commonly adopted in the relevant literature (Assenza et al., 2015; Godley and Lavoie, 2016; Meijers et al., 2019; Reissl, 2021, among others). For the HSM-related parameters, the calibration follows the values reported in the original studies of Anufriev and Hommes (2012a,b). Finally, loan maturity is set at 5 years, which represents a reasonable time horizon for this type of credit.

4. Simulation results

The simulation aims to evaluate how well the model-generated series approximate the respective historical ones; i.e., the study uses the method of descriptive output validation to assess the performance of each LLM-driven agent provided with various information sets through different prompts (Bianchi et al., 2008; Cirillo and Gallegati, 2012).

For each LLM-prompt combination, the study conducted 50 simulation runs, collecting the timeseries of total outstanding consumer credit and average lending rate from each run. The comparison between the levels of artificial and observed lending interest rate series is straightforward since it does not require any transformation. However, to facilitate comparison between the simulated and historical consumer credit series²³, the analysis transformed both by their full-sample averages (y_t/y_{mean} , where $y_t = \{L_t^{simulated}, L_t^{historical}\}$).

4.1. Relative performance assessment and validation

This part of the analysis treats each LLM-prompt combination individually and assesses which one replicates more closely the level and dynamics of the real-world series. In particular, it compares the artificial, 50-run median, timeseries with the historical series of consumer credit and

²³The observed consumer credit series comprises consumer and other loans with a maturity of over 1 and up to 5 years, obtained from the Bank of Greece's publicly available aggregate credit statistics. Due to write-offs and reclassifications, the stock series exhibits structural breaks in June 2010, November 2013, and July 2021. To account for these methodological changes, the final aggregate series has been reconstructed using the corresponding flow series, which have been corrected for such adjustments.

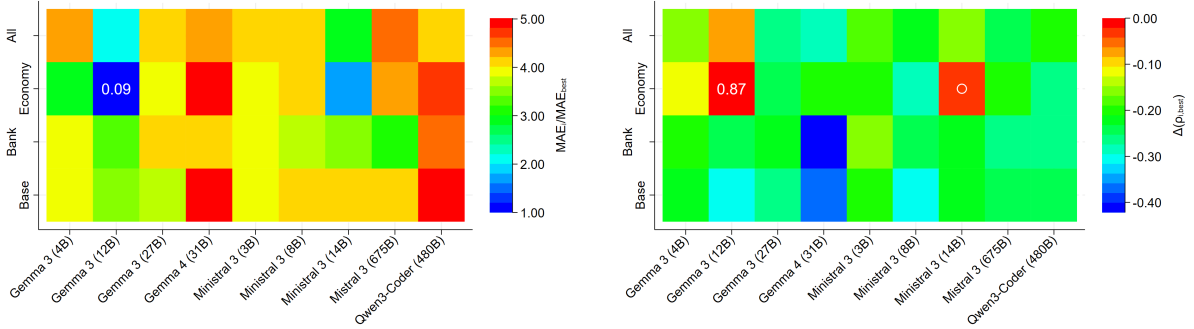
interest rates. It uses the mean absolute error (MAE) and Pearson’s zero-lag correlation coefficient (ρ) between the simulated and observed data to quantify their distance and synchronization, respectively. The study estimates both measures over the full sample.

In addition to identifying the best-performing LLM-prompt combinations, the analysis determines which pairs exhibit statistically equivalent performance to the best-performing ones for each metric. For MAE, the study applies the HAC-adjusted t-test²⁴, while for ρ it uses the Williams (1959) test for equality of correlation coefficients.

4.1.1. Consumer credit

Figure 2 presents the relative performance of each LLM-prompt pair for consumer credit, benchmarked against the best-performing one. In particular, distance is expressed as a ratio of MAEs (MAE_i/MAE_{best}) whereas synchronization as the difference between correlation coefficients ($\Delta(\rho_{i,best})$), where $i = \{LLM, prompt\}$ and $best$ is the top performing combination in each metric.

Figure 2: Distance (left panel) and synchronization (right panel) between simulated and historical consumer credit series for each LLM-prompt combination.



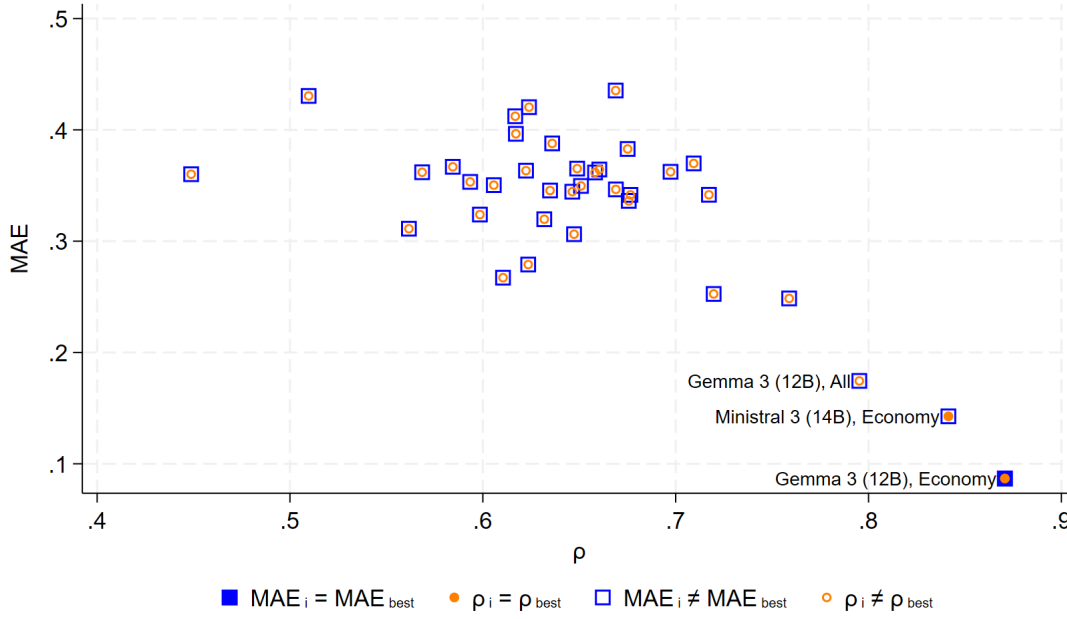
Numbers mark the MAE (left) & ρ (right) of the best-performing configuration and hollow circles the LLM-prompt combinations with statistically equal performance at 95% confidence level.

The results presented in Figure 2 reveal that the small-sized models Gemma 3 (12B) and Ministral 3 (14B), when provided with information on the prevailing economic conditions, produce the closest match to the observed consumer credit series. This result is particularly pronounced in the synchronization metric, for which the two LLM-prompt combinations exhibit statistically indistinguishable performance. In terms of distance, the combination of Ministral 3 (14B) with the economy prompt ranks second, with a MAE 1.6 times larger than that of the best-performing combination, while all other pairs exhibit at least twice the distance. Interestingly, this pattern is not observed for the much larger LLMs Mistral 3 (675B) and Qwen3-Coder (480B): across all prompts considered, their MAEs are more than three and, in some cases, almost five times larger than that of the best-performing combination. Similarly, their correlation coefficients are approximately 20 percentage points lower than the high ρ values of close to 90% and 84% achieved by the two top-performing combinations.

Plotting the two performance metrics together, in Figure 3, provides an informative overview of each pair’s performance.

²⁴Results are identical under the paired t-test and the non-parametric Wilcoxon’s matched-pairs signed-rank test.

Figure 3: Distance versus synchronization across LLM-prompt combinations for consumer credit.

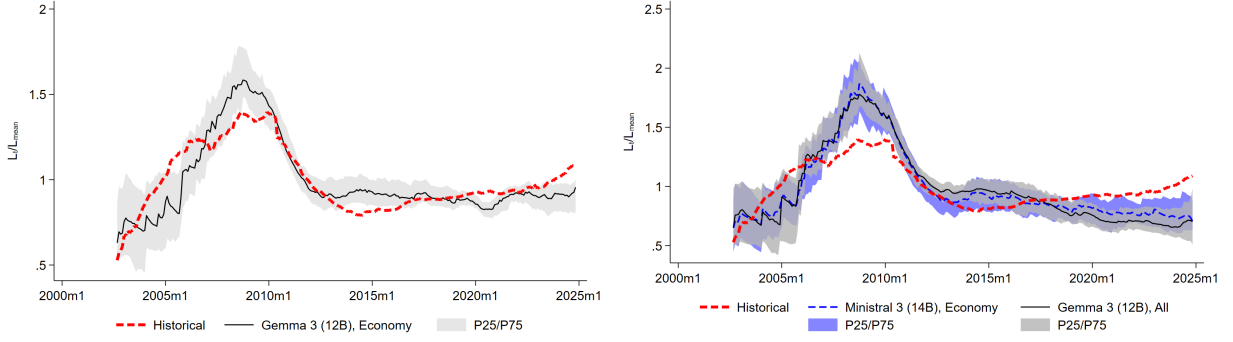


(Hollow)solid markers denote statistically (un)equal performance to the best-performing combination at 95% confidence level, per metric.

The results presented in Figure 3 show two broad performance groups. The top-performing LLM-prompt combination across both metrics is Gemma 3 (12B) provided with the economic environment prompt. Two additional pairs exhibit similar performance, with MAEs below 0.2 and correlations of approximately 80% or higher. In particular, as shown in Figure 2, the combination of Ministral 3 (14B) with the economic environment prompt yields a correlation that is statistically equal to that of the top-performing pair. The second adequately performing combination is again Gemma 3 (12B), this time provided with the full-information prompt. All remaining pairs form a distinct lower-performance cluster, with MAEs ranging from approximately 0.25 to 0.45, which are about three to five times larger than that of the best-performing combination, while their correlations decline to as low as 45%. This separation between the two groups indicates that both the magnitude of the prediction errors and the ability to reproduce the observed dynamics differ substantially across the LLM-prompt combinations.

In addition to the performance metrics discussed above, Figure 4 provides a visual validation of the identified candidates. It compares the evolution of consumer credit generated by the best-performing LLM-prompt combination and the two closest-performing combinations with the observed series.

Figure 4: Timeseries of best-performing (left panel) and adequately performing (right panel) LLM-prompt combinations for consumer credit.



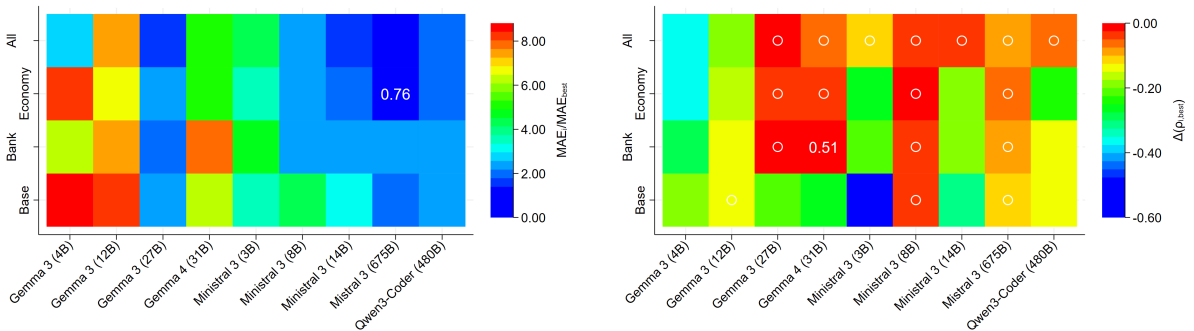
Lines of simulated series denote their 50-run medians while shaded areas their first/third quartiles.

The plots in Figure 4 verify the remarkably good fit between the simulated and historical timeseries. Especially the top-performing pair, Gemma 3 (12B) with the economic environment prompt, closely tracks the historical consumer credit series throughout most of the sample, with only two short periods of divergence. The first occurs around 2015, when simulated credit lies above the observed series, while the second occurs during the final 12 months, when the observed series exhibits a rising trend that is not captured by the simulation. Overall, the simulated median series reproduces both the level and the dynamics of consumer credit well, while its interquartile range encompasses the real-world series over almost the entire sample. The other two candidate pairs also show satisfactory performance and, in particular, capture the decline observed around 2015 more closely until approximately 2018. Thereafter, however, their simulated paths increasingly diverge from the observed series, as the latter follows an upward trend while the former steadily decline.

4.1.2. Lending rate

The study applies the previous metrics to assess the relative performance of LLM-prompt combinations for lending rates and presents the results in Figure 5.

Figure 5: Distance (left panel) and synchronization (right panel) between simulated and historical lending rate series for each LLM-prompt combination.



Numbers mark the MAE (left) & ρ (right) of the best-performing configuration and hollow circles the LLM-prompt combinations with statistically equal performance at 95% confidence level.

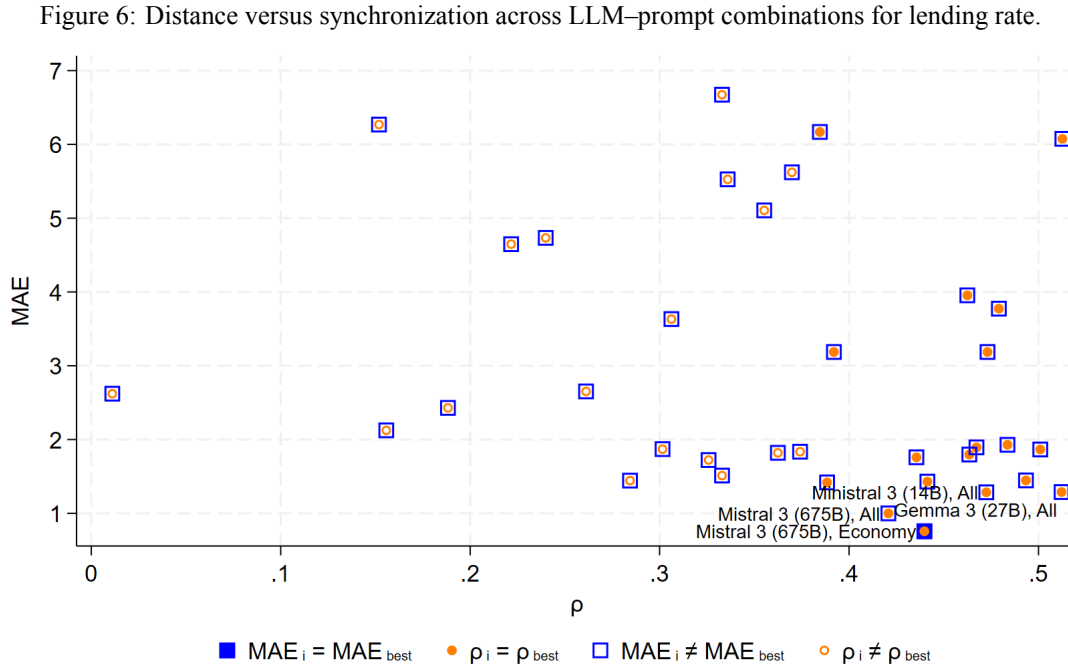
Unlike the results for consumer credit, where smaller models generally outperformed their much larger counterparts, the lending rate results in Figure 5 suggest a different ranking. In terms of distance from the historical series, Mistral 3 (675B) combined with the economic conditions

prompt achieves the best performance. The second largest LLM, Qwen3-Coder (480B) with either the economic environment or the full-information prompt also produces results that are relatively close to the best ones. Importantly, although their MAEs are not statistically equal to the lowest one, the performance of some of the smaller LLMs is sufficiently close to it. In particular, the largest LLMs in the Gemma 3 and Ministral 3 families, Gemma 3 (27B) and Ministral 3 (14B), respectively, when supplied with the full-information prompt, exhibit distances similar to the top-performing combination's.

Performance measured by the temporal synchronization metric in the right panel of Figure 5 presents a similarly differentiated picture. While several LLM–prompt combinations exhibit correlations that are statistically indistinguishable from that of the best-performing one, many have substantially lower ρ s. Interestingly, the baseline prompt consistently produces the weakest performance, whereas the full-information prompt generally ranks among the best-performing specifications across LLMs.

It is worth noting that the lending rate results indicate lower levels of performance than those achieved for consumer credit, both in terms of distance as well as temporal synchronization.

The scatterplot of the two metrics in Figure 6 provides an overview of how various pairs perform.

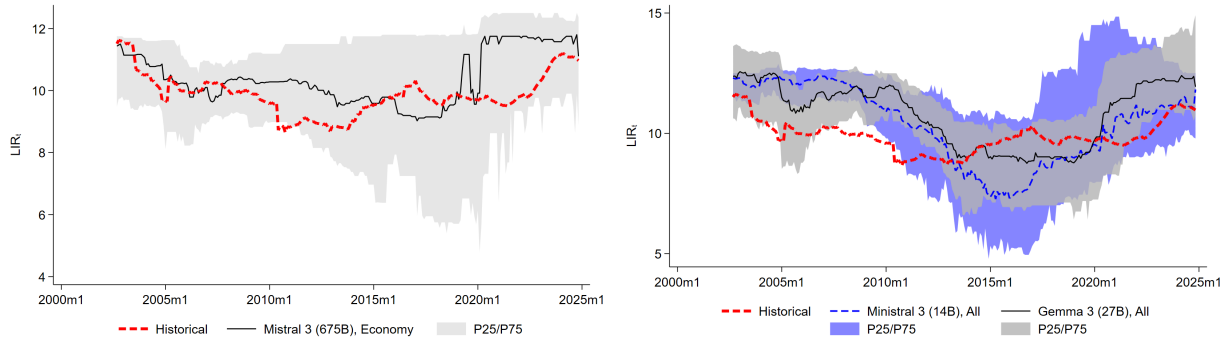


(Hollow)solid markers denote statistically (un)equal performance to the best-performing combination at 95% confidence level, per metric.

Figure 6 reveals a rather dispersed picture, both in terms of distance and synchronization with the observed series. At the lowest distance range, there are four LLM-prompt pairs with MAEs around 1 ± 0.25 , although none is statistically equal to the best combination. Notably, Mistral 3 (675B) appears in two of these four combinations, while the full-information prompt accounts for three. The remaining combinations occupy progressively higher distance tiers, with MAEs ranging from roughly twice to almost nine times that of the best-performing combination. Similarly, correlation results span from almost zero to about 50%. Yet, from correlations of roughly 40% to the highest one, several configurations achieve statistically equal performance.

Figure 7 provides a visual assessment of how closely the lending rate timeseries generated by the best-performing and similarly performing LLM-prompt pairs match the observed series.

Figure 7: Timeseries of best-performing (left panel) and adequately performing (right panel) LLM-prompt combinations for lending rate.



Lines of simulated series denote their 50-run medians while shaded areas their first/third quartiles.

The results in Figure 7 illustrate the performance figures described above. At the left panel, the median simulated series from Mistral 3 (675B), which incorporates information on economic conditions, struggles to closely track real-world lending rates. This is particularly evident after 2020 when deviation increases substantially. Yet, despite the widening interquartile range after the middle of the sample period in 2012, the historical lending rates remain within it throughout the sample.

The right panel in Figure 7 displays a similar behavior by some of the second-best configurations. While not as close as the top pair's, the simulated results from the much smaller Ministral 3 (14B) and Gemma 3 (27B), using the full-information prompt, remain plausibly near the historical lending rates. Moreover, although they exhibit steeper-than-observed declines and increases during the crisis and recovery periods, respectively, their interquartile ranges encompass the real-world series for the whole second half of the sample.

4.2. Performance determinants: model size and prompt information

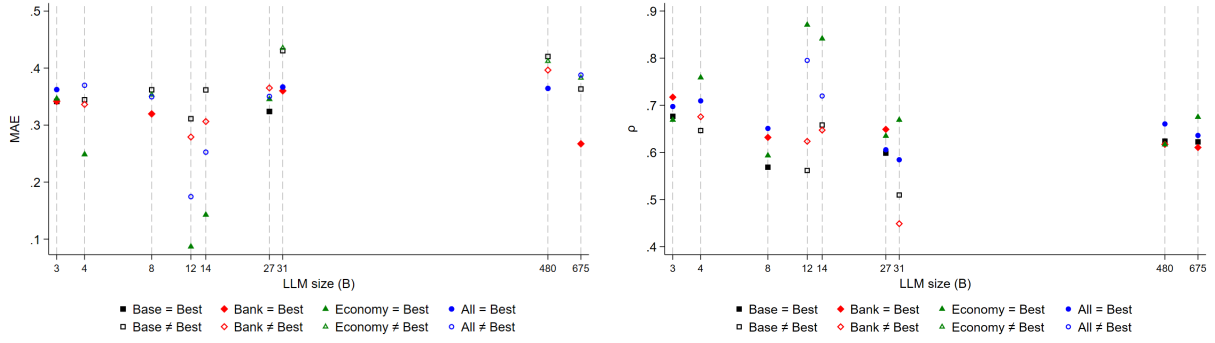
The previous analysis revealed that LLM size only partly plays a role in performance and that, frequently, small- or medium-sized models outperform their much larger counterparts. It has also suggested that, as expected, the information provided to the LLM agent through the prompt affects its ability to replicate the realized series under investigation.

This part of the analysis focuses on how well each prompt performs within each LLM, thus gauging the improvement attributed to the provided information by different models. It uses the previous metrics for measuring distance and synchronization between the simulated and historical timeseries and applies the same statistical equality tests to compare the performance of different prompts within the same LLM.

4.2.1. Consumer credit

Figure 8 presents the results of prompts' performance in matching the magnitude and evolution of consumer credit series within each LLM, by metric.

Figure 8: Semi-log plot of prompt performance within each LLM, measured by its distance from (left panel) and synchronization with (right panel) the historical consumer credit series.



(Hollow)solid markers denote statistically (un)equal performance to the best-performing prompt at 95% confidence level, per metric. Horizontal axis is in logarithmic scale.

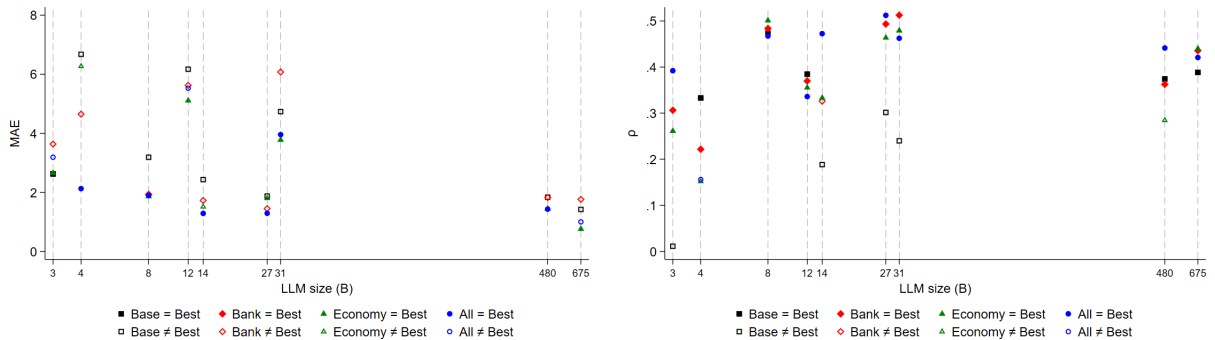
Figure 8 reveals some interesting patterns. It shows that prompt performance is broadly similar across LLMs, with MAE fluctuating around 0.35 and ρ around 65%, despite substantial variation in model size spanning roughly an order of magnitude. Moreover, the variation in performance across prompts remains generally limited within each LLM, in both metrics. There are two notable exceptions, namely for Gemma 3 (12B) and Ministral 3 (14B) in which the full-information and, particularly, the economic environment prompt result in substantial improvements in both performance metrics.

A broad observation from Figure 8 is that prompt-specific gains are more evident in terms of distance, as the best-performing prompt often achieves a clearly lower MAE than the alternatives, whereas synchronization is more frequently the same across prompts. Finally, in terms of the most effective information set, the economic environment prompt delivers the largest performance gains, followed by the full-information one.

4.2.2. Lending rate

The results for lending rates in Figure 9 reveal some differences and some common patterns compared to those for consumer credit.

Figure 9: Semi-log plot of prompt performance within each LLM, measured by its distance from (left panel) and synchronization with (right panel) the historical lending rate series.



(Hollow)solid markers denote statistically (un)equal performance to the best-performing prompt at 95% confidence level, per metric. Horizontal axis is in logarithmic scale.

Figure 9 shows that performance across LLMs is weaker and more dispersed for lending rates than for consumer credit. In terms of distance, some LLMs achieve MAEs around 2, while others range from approximately 3 to 6. Notably, models in the small and medium range of the

scale, such as Ministral 3 (14B) and Gemma 3 (27B), achieve MAEs comparable to those of much larger LLMs. Within-LLM variation in distance is generally limited, except for Gemma 3 (4B), whose MAE ranges from around 3 under the full-information prompt to almost 7 under the worst-performing, baseline, prompt. Synchronization is similarly dispersed, with three LLMs achieving ρ around 50%, while the remaining models range from approximately 40% to below 10%.

However, as with consumer credit, prompts primarily improve performance by reducing distance, whereas most prompts yield statistically similar synchronization, with the baseline prompt being a notable exception. Overall, the full-information prompt performs best across LLMs, followed by the economic conditions prompt.

Considering the results from both Figure 8 and Figure 9, an interesting observation is that prompt-related performance improvements tend to be stronger among small- and medium-sized LLMs, as suggested by their generally higher variation across prompts. However, there is no clear relationship between model size and performance across the metrics and variables examined. In contrast, large LLMs tend to exhibit more consistent performance, with relatively low variation across prompts for both variables and metrics.

5. Discussion

This study investigated whether LLMs can be integrated into macroeconomic agent-based simulations and, more specifically, whether they can reproduce observed patterns in consumer credit and lending rates when asked to act as loan officers using several information sets as input.

The results suggest that such integration is computationally feasible and, importantly, that LLM-driven agents can generate economically plausible outcomes. However, the computational feasibility of LLM integration depends strongly on model architecture. Non-thinking models can be efficiently used, whereas thinking models become prohibitively slow especially when simulations extend over long horizons.

The ability of the models to reproduce historical series varies considerably across the variables examined. Simulated consumer credit series generally exhibit a high degree of similarity with the observed data and, notably, even relatively small LLMs are able to reproduce aggregate consumer credit dynamics well, in some cases outperforming substantially larger models. The results for lending rates are more mixed. While the simulated series remain broadly plausible, LLMs generally struggle to reproduce the historical patterns observed in the data.

Another important finding concerns the information supplied to the LLM through the prompt. The information set appears to be a significant determinant of model performance, with prompts containing information about the broader economic environment, or combining this information with bank-specific characteristics, generally producing the strongest results. By contrast, the baseline borrower-only prompt tends to perform worst. These improvements are particularly pronounced for small- and medium-sized LLMs, whereas larger models tend to exhibit tighter performance across alternative prompts.

Taken together, these findings suggest that model size alone is not a reliable predictor of performance in LLM-driven economic simulations. The fact that relatively small, consumer-grade models can perform particularly well, and occasionally outperform substantially larger models, indicates that the information available to the agent may be at least as important as the underlying capacity of the LLM. Interestingly, the less pronounced gains obtained by larger models when additional information is provided may indicate that their performance is less sensitive to the information explicitly represented in the prompt, potentially allowing them to generate adequate outcomes from a more limited information set. Hence, increasing model scale does not necessarily constitute the most effective way of improving the performance of LLM-driven economic agents.

These findings are particularly noteworthy given the challenging environment in which the models are evaluated. The Greek case represents a demanding setting for economic simulation, as a substantial part of the sample is characterized by severe economic turbulence and considerable changes in economic conditions. This may partly explain the difficulty that LLMs face in reproducing the historical dynamics of lending rates. At the same time, this environment makes the ability of even relatively small LLMs to closely reproduce aggregate consumer credit dynamics particularly remarkable. The result, therefore, provides encouraging evidence that LLM-driven agents can capture meaningful features of economic decision-making even when operating under highly volatile economic conditions.

The weaker performance observed for lending rates merits particular attention. One possible explanation is that the information incorporated into the prompts does not adequately capture the information set and institutional considerations underlying actual loan officers' pricing decisions. Lending rates may depend on a broader range of factors than those represented in the simulation, including financial-market conditions, risk perceptions, and expectations regarding future economic developments. Consequently, the weaker performance should not necessarily be interpreted as evidence that LLMs are unable to reproduce lending rate dynamics. Rather, it may reflect an incomplete specification of the information available to the simulated agents.

The analysis nevertheless has several limitations. While the Greek case provides a challenging testing ground for the models, it also limits the extent to which the findings can be generalized to other countries or economic periods. The computational burden further constrains the range of models and simulation horizons that can be examined, particularly in the case of thinking models. Finally, the analysis focuses on LLM-driven loan officers rather than representing all relevant economic agents through LLMs. Consequently, it does not capture the potentially important interactions that could arise when borrowers and other market participants are themselves heterogeneous, LLM-driven, agents.

These limitations point to several promising directions for future research. A natural extension would be to enrich the information available to loan officers and examine whether additional institutional, financial, and macroeconomic information improves the reproduction of lending rate dynamics. Such an extension would help distinguish between limitations arising from the underlying capabilities of the models and those resulting from the specification of the agents' information sets. Testing the framework across other countries and economic episodes would also be important for determining whether the observed patterns are specific to the Greek experience or instead reflect more general properties of LLM-driven economic agents. Further research could examine whether LLMs can be used to represent other agents in the economy, such as consumers, thereby allowing both sides of the lending interaction to be modeled endogenously. Although such an approach would increase the computational burden, this limitation could potentially be mitigated by focusing on shorter portions of the sample.

Overall, the results provide encouraging evidence that LLMs can act as decision-making agents in economic simulations without requiring the largest or most computationally demanding models. They also highlight an important aspect: successful LLM-driven ABM simulation depends not only on the capabilities of the LLM itself, but also on the information through which its economic environment is represented. The ability of relatively small models to generate plausible, and sometimes notably accurate, aggregate outcomes suggests that carefully designed information sets may be as important as increasing model scale²⁵. LLM-driven agents, therefore, offer a promising avenue for economic simulation, provided that future work continues to address the representation of agents' information sets, the computational costs of long-horizon simulations, and the complexity

²⁵Thus echoing Asimov: the performance of a computational system may depend as much on the information supplied to it as on the sophistication of the machine itself.

of real-world economic decision-making. More broadly, the findings suggest that LLMs' ability to assimilate economically relevant information and yield results remarkably similar to the realized ones, without imposing a specific functional form on the underlying behavioral relationship, may indeed enable economic simulations to shift from equations to prompts.

Appendix A. Feasibility assessment of LLM-driven bank agent

Before integrating LLMs into the full agent-based model, the study conducted a feasibility exercise to assess whether different LLMs could reliably perform the role of a commercial bank’s loan officer by generating valid lending decisions that satisfy both regulatory constraints and prompt instructions, with the results informing the simulation’s modelling choices.

Appendix A.1. Feasibility exercise setup

The analysis first constructs a prototype population of 50 households. It endows each household with a monthly income drawn from a uniform distribution between EUR 300 and EUR 2,000. Their existing debt is also drawn from a uniform distribution, ranging from zero to 2.3 times the household’s annual income. Finally, it assigns a temperature parameter²⁶ to each household, uniformly spanning the interval from 0.05 to 1.45, which is subsequently supplied to the corresponding LLM.

After initialization, each household submits the following prompt to the LLM under investigation:

*Assume you are a loan officer in a commercial bank. A customer with a monthly income of **[income]** EUR and an existing loan balance of **[existing debt]** EUR asks for a consumer loan with 5 years maturity. You will consolidate the existing debt with any amount of new credit you decide to provide into a single fully amortizing loan with 60 equal monthly annuity payments at a fixed annual nominal interest rate. The interest rate you offer will be applied to this total consolidated amount. The central bank’s policy rate is currently 1.5%. Use this as a baseline when determining your offered interest rate. Decide how much new credit to provide and at what annual nominal interest rate (in percent). You must ensure that the total monthly debt service for the consolidated loan (existing loan plus new credit) is strictly less than **[maximum monthly payment]** EUR. You must also ensure that the new credit offered is less than **[maximum new credit]** EUR. Consider also that the current unemployment rate is at the high level of 15%, which represents an increase of 2 percentage points from 12 months ago. Factor this general economic condition into your credit decision, particularly regarding the loan amount and the interest rate risk premium. Respond with exactly two positive numbers: new credit amount first, then interest rate in percent, separated by a single space. Do not include units, punctuation, or explanation.*

To simplify the LLM’s task and eliminate unnecessary calculations, the study pre-computes the maximum permitted monthly payment and the maximum amount of new credit before passing it to the prompt. The maximum monthly payment obeys a debt-service-to-income (DSTI) limit of 50%, while the maximum amount of new credit satisfies a debt-to-income (DTI) limit equal to 2.3 times a household’s annual income after accounting for outstanding debt.

The study considers an LLM response valid only if it returns exactly two numerical values and the proposed lending decision respects both the DSTI and DTI limits. Whenever the LLM returns an invalid response, the same prompt is submitted again, up to a maximum of 15 attempts. If the LLM fails to generate a valid response within this limit, the simulation proceeds to the next household. For each LLM, the experiment is repeated 60 times, resulting in a total of 3,000 lending decisions per model.

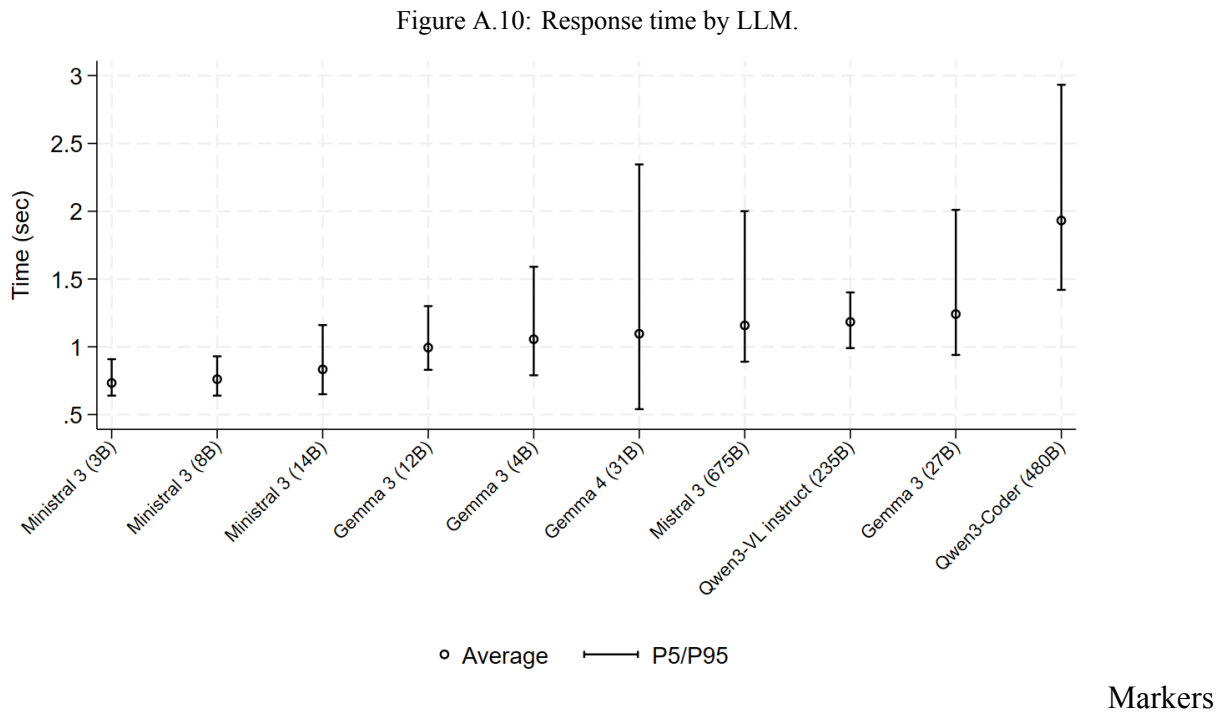
²⁶The temperature parameter controls the degree of randomness in an LLM’s token selection, with lower values producing more deterministic outputs and higher values allowing greater variability. Although there is no universally accepted classification of temperature settings, practitioners commonly distinguish four broad ranges: 0.0-0.3 (deterministic, factual responses), 0.4-0.7 (balanced creativity and accuracy), 0.8-1.5 (creative, more varied outputs), and 1.6-2.0 (highly random or experimental outputs).

The feasibility assessment uses only open-weight LLMs that can be executed — hardware permitting — either locally or through the Ollama platform’s cloud infrastructure²⁷. The selected models span a wide range of computational requirements, from consumer-grade models that can run on commodity hardware to frontier-scale models requiring substantial computing resources. This diversity enables the analysis of whether model size and computational capacity affect the quality and feasibility of LLM-driven lending decisions. In particular, the study examines ten LLMs representing the following model families and parameter sizes: Gemma 3 (4B, 12B and 27B) (Gemma Team, 2025), Gemma 4 (31B) (Gemma Team, 2026), Ministral 3 (3B, 8B and 14B) (Liu et al., 2026), Mistral 3 (675B)²⁸, Qwen3-VL-Instruct (235B) (Bai et al., 2025) and Qwen3-Coder (480B)²⁹.

The analysis below presents the evaluation of the resulting output in terms of computational performance, response validity, and factors that affect them.

Appendix A.2. Prompt execution time

A fundamental aspect of the LLMs’ integration feasibility into agent-based models is the time required to generate lending decisions. Since every credit request requires an LLM inference, response latency directly affects the computational performance of the simulation. Figure A.10 reports the average response time together with the 5th and 95th percentiles for each LLM examined.



denote the average and whiskers the top/bottom 5 percentiles.

The results in Figure A.10 indicate noticeable differences across model families. The Ministral 3 models consistently exhibit the fastest response times, requiring approximately 0.6 seconds on average to generate a lending decision based on the asked prompt. Their response times also display relatively limited variability, with the 5th and 95th percentiles remaining close to one second. The

²⁷<https://ollama.com/>

²⁸Described in: <https://docs.mistral.ai/models/model-cards/mistral-large-3-25-12> [Accessed in August 2026].

²⁹Described in: <https://qwen.ai/blog?id=qwen3-coder> [Accessed in August 2026].

two smaller Gemma 3 models (4B and 12B) follow, with average response times of approximately one second and bottom/top 5th percentile ranges spanning roughly from 0.8 to 1.6 seconds.

The remaining models require slightly longer inference times. Most exhibit average response times of approximately 1.2 seconds, whereas Qwen3-Coder (480B) requires nearly two seconds on average. Moreover, larger models display considerably greater variability in execution time. Their top 5th percentiles frequently exceed two seconds and approach three seconds in the case of Qwen3-Coder (480B), indicating that individual responses can occasionally require substantially longer inference times.

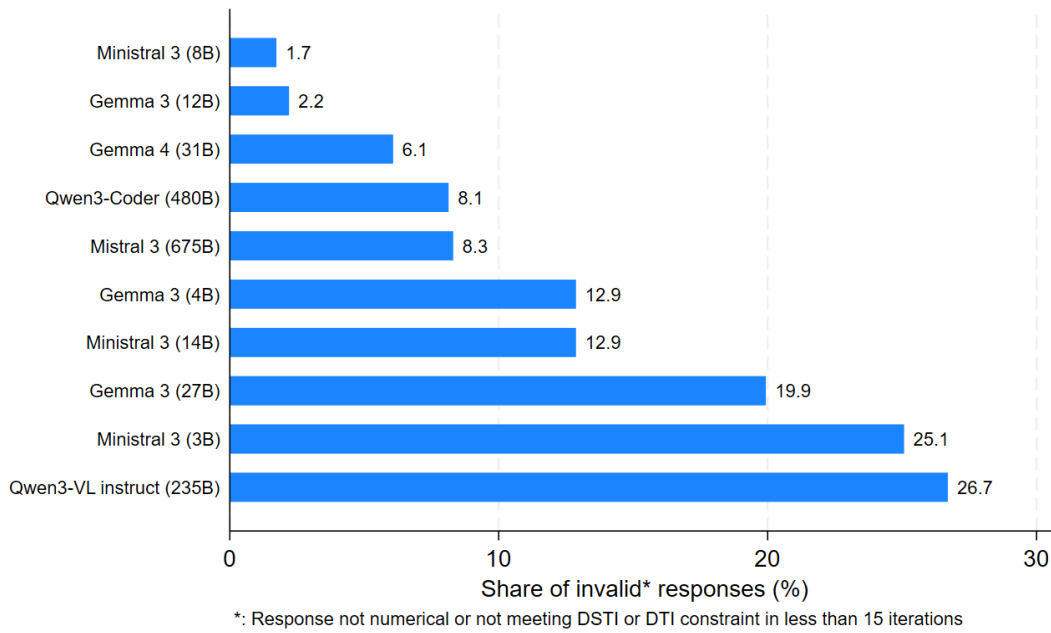
Overall, the results suggest that response latency does not constitute a practical limitation for integrating these³⁰ LLMs into the proposed framework. Even the slowest model examined generates lending decisions within a few seconds, while several smaller open-weight models consistently respond in well under one second. These findings support the computational feasibility of replacing conventional equation-based lending rules with LLM-driven decision mechanisms. Yet, they also demonstrate that the computational cost of LLM inference is non-trivial and should be taken into account when assessing the practical implementation of the integration of LLM-driven agents into ABMs.

Appendix A.3. Valid response rate

Besides response latency, the practical feasibility of integrating LLMs into agent-based models depends on their ability to consistently yield answers that satisfy the imposed prompt requirements. The analysis therefore examines the valid response rate, defined as the share of valid responses relative to the original number of prompts submitted to each LLM, without accounting for any subsequent prompt retries.

³⁰Preliminary experiments also evaluated several "thinking" LLMs. These models were excluded from the analysis because their average inference times were prohibitively high, ranging from approximately 20 seconds (GPT-OSS 20B and DeepSeek-V3.1 671B) to 80 seconds (Gemini 3 Flash) and up to 162 seconds (DeepSeek-V3.2 685B), per prompt. Given that the proposed framework requires thousands of LLM queries, such response times make their integration computationally infeasible.

Figure A.11: Invalid response rate by LLM.



Invalid responses are those which are not numerical or fail to meet DSTI and DTI constraints within 15 retries.

Figure A.11 reveals substantial heterogeneity across the examined models. Ministral 3 (8B) and Gemma 3 (12B) achieve the highest valid response rates, with invalid responses accounting for only approximately 2% of the originally submitted prompts³¹. At the opposite end of the ranking, Gemma 3 (27B) produces invalid responses in approximately one out of every five prompts, while Ministral 3 (3B) and Qwen3-VL-Instruct (235B) exhibit even higher invalid response shares of around 25%.

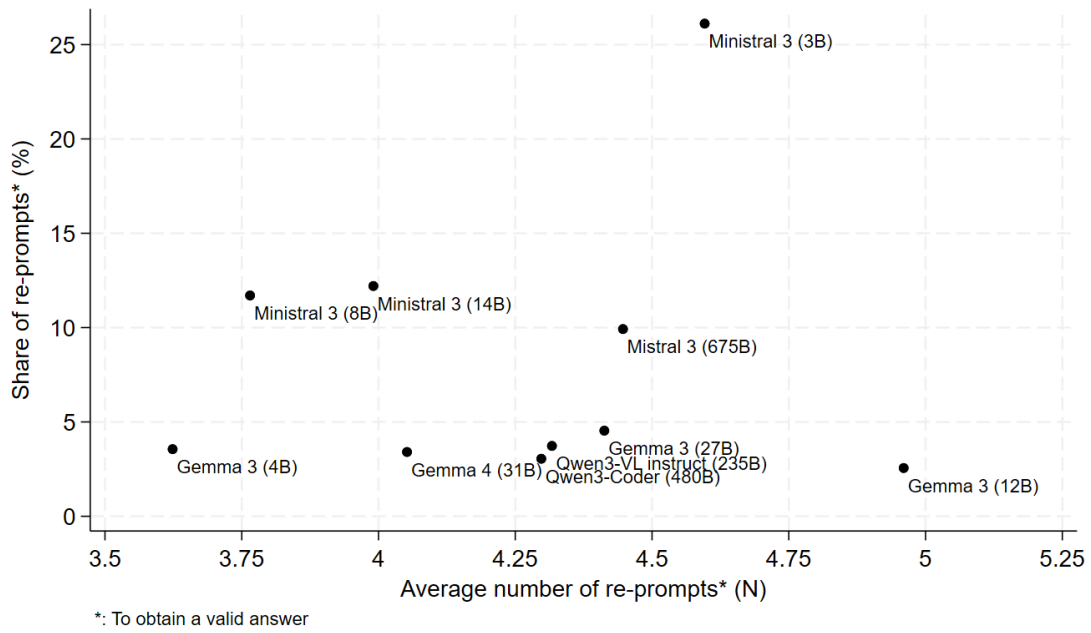
These findings indicate that response validity depends neither exclusively on model family nor on model size. Instead, models from different families and parameter scales appear across all performance tiers, suggesting that the ability to reliably follow structured prompting instructions remains largely model-specific. From a practical perspective, this result highlights the importance of evaluating response validity in addition to speed when selecting LLMs for integration into agent-based models.

Appendix A.4. Prompt retries

Although an invalid response does not necessarily prevent the simulation from proceeding, it increases the computational cost by requiring additional prompt submissions. The analysis therefore examines the frequency of prompt retries required to obtain a valid response together with the number of iterations needed before a successful response is generated.

³¹I.e. not considering the number of times a prompt might have been re-asked to get a valid answer.

Figure A.12: Share versus average number of retries by LLM.



Numbers consider only valid responses.

Figure A.12 reveals three broad performance groups. The first comprises LLMs requiring prompt retries in fewer than 5% of submitted prompts. A second group requires retries in approximately 10% of cases. Finally, Ministral 3 (3B) stands out as the most demanding model, requiring an additional prompt in approximately one out of every four cases before producing a valid response.

Conditional on eventually generating a valid response, the average number of prompt iterations exhibits relatively limited variation across models, ranging from approximately 3.5 to fewer than 5 iterations. Moreover, successful responses are typically obtained after a relatively small number of retries. For all examined models, at least 90% of valid responses are generated within ten prompt iterations. The two best-performing models, Ministral 3 (8B) and Gemma 3 (4B), exhibit even higher performance, with 95% of valid responses obtained with less or equal to nine prompt iterations.

Overall, these findings suggest that, although some LLMs require repeated prompting to produce valid outputs, the additional computational burden remains limited for most models. Combined with the response-time analysis, the results indicate that prompt retries are unlikely to constitute a major practical obstacle for integrating appropriately selected LLMs into agent-based simulations.

Appendix A.5. LLM performance determinants

Appendix A.5.1. Proximity of prompt input to constraints

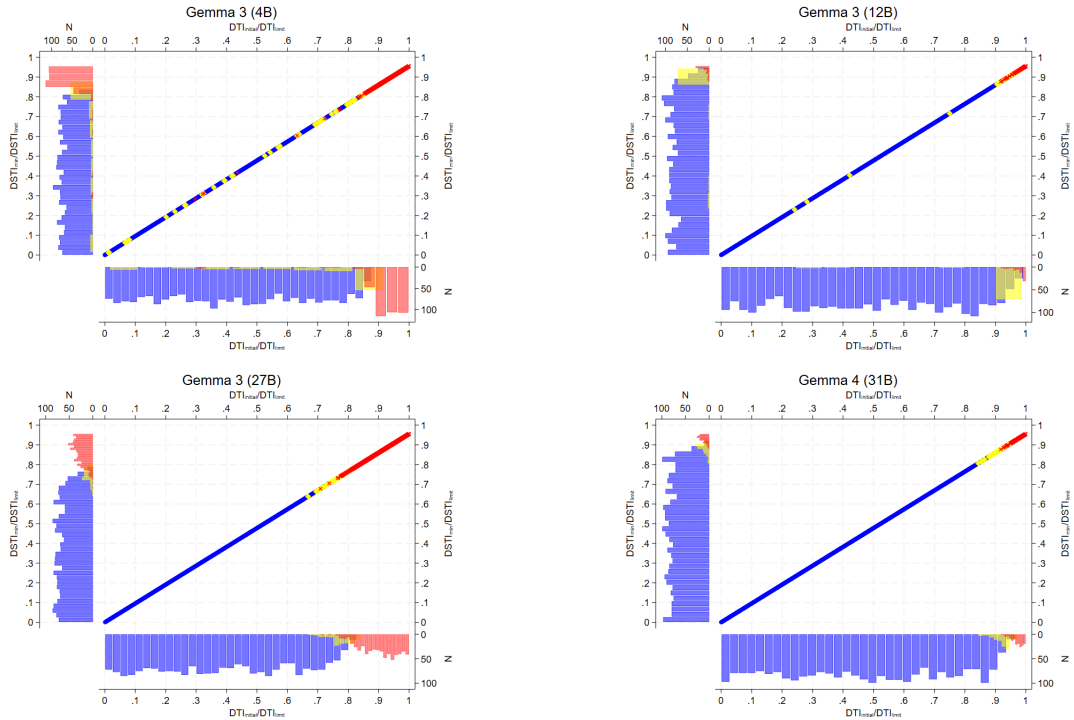
While LLMs demonstrate strong capabilities in language-based tasks, previous studies have documented their limitations in mathematical reasoning, with models achieving only moderate accuracy when solving mathematical problems (Wang et al., 2026). Therefore, an important component of the feasibility analysis examines how effectively LLMs solve the algebraic calculations required to comply with the DSTI and DTI constraints specified in the prompt.

By construction, existing household credit is uniformly distributed between zero and 2.3 times of their annual income, with the upper bound corresponding exactly to the imposed DTI constraint.

Given this distribution and the requirement that the LLM uses the policy rate of 1.5% as the minimum applicable interest rate in its lending decision, the minimum achievable DSTI —assuming that the LLM grants no additional credit and only adjusts the applied interest rate to the minimum level— is also uniformly distributed. This distribution ranges from zero to approximately 48%³², placing the upper bound only two percentage points below the imposed DSTI limit.

Figures A.13 to A.15 illustrate how the validity of LLM responses varies with the proximity of household input values to the prompt-specified DSTI and DTI constraints across the LLMs examined in the study. Proximity is defined as the ratio of the minimum DSTI and the initial³³ DTI to their respective constraint thresholds.

Figure A.13: Relationship between response validity and the proximity of input values to prompt-specified constraints for Gemma-family LLMs.



Blue circles/bars indicate valid responses achieved on the first attempt, yellow diamonds/bars indicate valid responses requiring more than one attempt, and red crosses/bars indicate invalid responses after 15 retries.

The results for the Gemma family in Figure A.13 indicate that invalid responses become increasingly concentrated as the prompt inputs approach the imposed regulatory limits. Nevertheless, performance differs across models. For Gemma 3 (4B) and Gemma 3 (27B), most invalid responses occur when the initial household characteristics exceed 80% of the applicable DSTI or DTI constraint. However, Gemma 3 (4B) exhibits a small number of invalid responses even at substantially lower proximity ratios, reaching values as low as 0.3. Moreover, although prompt retries are most frequently observed for proximity ratios above 0.8, they occur throughout almost

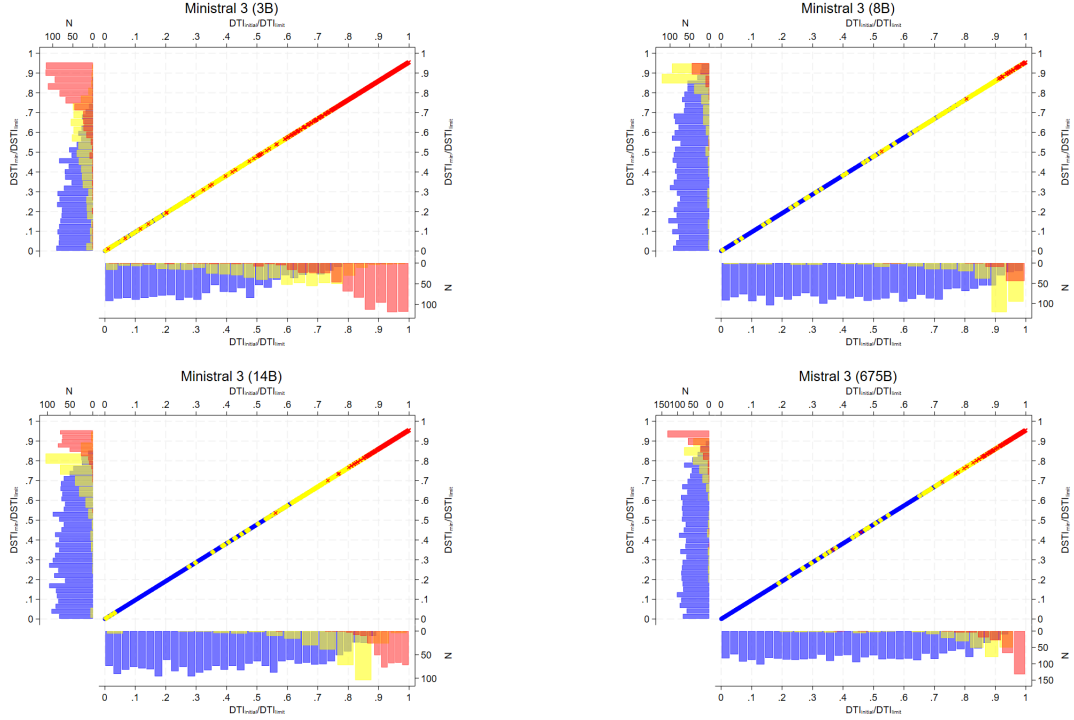
³²The minimum monthly interest rate is $r_{min} = 0.015/12 = 0.00125$. Also, $DTI = \frac{Debt}{12 \cdot I} \rightarrow Debt = 12 \cdot DTI \cdot I$. Substituting the previous in the standard DSTI formula gives $DSTI = \frac{Debt \cdot (0.00125 \cdot (1.00125)^{60}) / (1.00125^{60} - 1)}{I} = \frac{12 \cdot DTI \cdot I \cdot 0.01730989}{I} = 12 \cdot DTI \cdot 0.01730989$. Consequently, in the case of $DTI = DTI_{max} = 2.3$, the later expression yields a maximum of $DSTI_{min} \approx 0.478$.

³³In the exercise, the initial DTI is also the minimum attainable DTI because the existing debt of a household cannot be lower than the originally assigned amount.

the entire distribution of input values.

In contrast, Gemma 3 (12B) and Gemma 4 (31B) exhibit considerably more robust performance. Both models generate fewer invalid responses which occur almost exclusively when the initial prompt inputs exceed 90% of the lending constraints, while prompt retries remain largely concentrated above a proximity ratio of 0.8.

Figure A.14: Relationship between response validity and the proximity of input values to prompt-specified constraints for Mi(ni)stral-family LLMs.

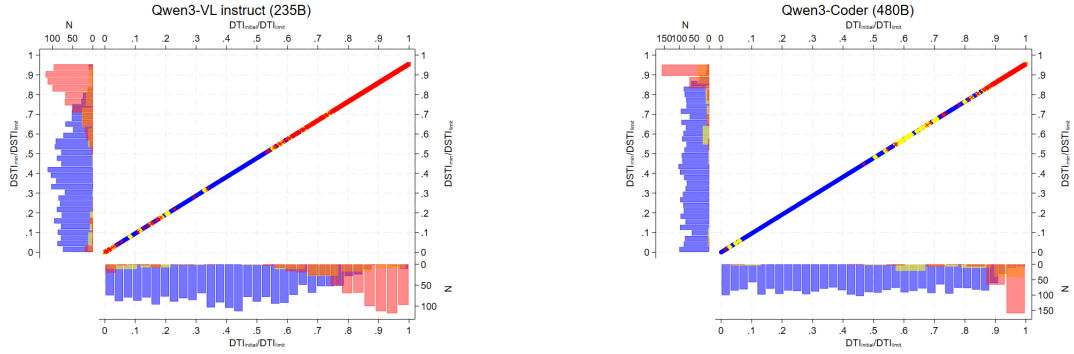


Blue circles/bars indicate valid responses achieved on the first attempt, yellow diamonds/bars indicate valid responses requiring more than one attempt, and red crosses/bars indicate invalid responses after 15 retries.

Results for the Ministral family in Figure A.14 reveal the existence of a relationship between model size and performance near the regulatory constraints. The smallest model, Ministral (3B), exhibits the weakest performance. It produces a non-negligible number of invalid responses beginning at proximity ratios of approximately 0.6, while failed attempts occur even at the lower end of the proximity ratio distribution, below 0.1. Moreover, successful responses frequently require retries across nearly the entire distribution. Beyond a proximity ratio of approximately 0.85, retries cease to recover valid responses, resulting in failed attempts instead.

Larger models improve performance, although the benefits of additional scale seem to plateau. Ministral 3 (8B) and Ministral 3 (14B) concentrate failed attempts near proximity ratios of 0.9, while retries are generally required only for observations with proximity ratios above 0.5, with only sporadic cases occurring below this threshold. Interestingly, the largest model examined, Mistral 3 (675B), delivers only modest additional gains relative to the 8B and 14B models. Although it shows a moderate frequency of invalid responses, it still exhibits occasional failures at relatively low proximity ratios of around 0.4.

Figure A.15: Relationship between response validity and the proximity of input values to prompt-specified constraints for Qwen3-family LLMs.



Blue circles/bars indicate valid responses achieved on the first attempt, yellow diamonds/bars indicate valid responses requiring more than one attempt, and red crosses/bars indicate invalid responses after 15 retries.

Finally, results for the Qwen3 family in Figure A.15 reveal two distinct performance patterns. Qwen3-VL Instruct (235B) exhibits a substantial number of invalid responses beginning at proximity ratios of approximately 0.5, with a smaller but still non-negligible number occurring below 0.3. The distribution of retries closely mirrors this pattern, indicating that additional attempts frequently fail to recover valid responses as the input values approach the regulatory constraints.

In contrast, Qwen3-Coder (480B) performs on par with the largest model examined, Mistral 3 (675B). Both models exhibit a comparable share of invalid responses, with failures distributed similarly across the proximity range.

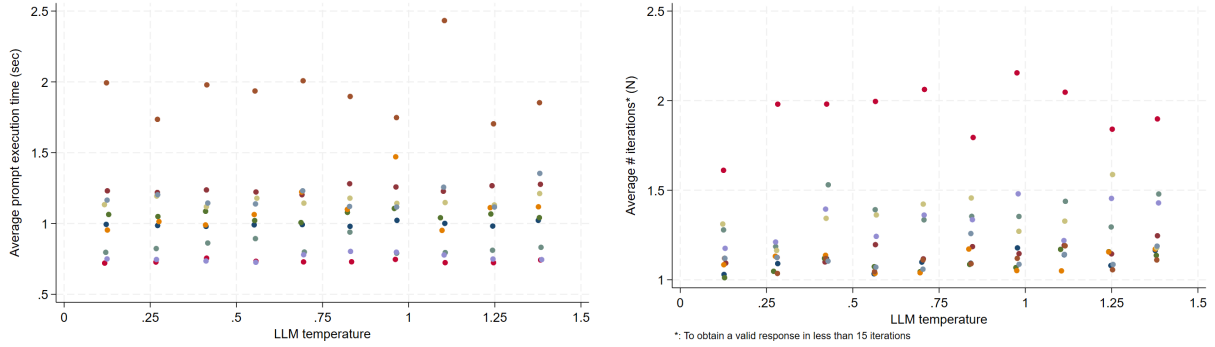
These results suggest that increasing model size alone does not fully explain performance, as models of comparable scale can display markedly different levels of performance, likely depending on their underlying architecture and training.

Appendix A.5.2. Temperature parameter

The feasibility assessment also examines whether the temperature parameter supplied to the LLM affects its computational performance. In particular, the analysis investigates whether temperature influences response latency or the number of prompt iterations required to obtain a valid response.

Given the large number of observations available for each model (3,000 prompts), the analysis summarizes the results using binned scatter plots to facilitate visual interpretation. Specifically, the temperature parameter is partitioned into ten equally sized bins. For each bin, the analysis computes the mean temperature together with the corresponding average prompt execution time and the average number of prompt iterations required to obtain a valid response. Figure A.16 presents the resulting relationships.

Figure A.16: LLM performance as a function of temperature. Average response time (left panel) and average number of prompt iterations (right panel).



Each color denotes a different LLM. Individual labels are omitted for clarity, as the same qualitative pattern emerges across all models.

The results in Figure A.16 provide little evidence that the temperature parameter systematically affects either performance metric. Across all examined models, average response times remain broadly stable over the entire range of temperature values. Similarly, the average number of prompt iterations required to obtain a valid response exhibits no discernible relationship with temperature. Any observed relationship between temperature and the two performance metrics appears very weak, with higher values associated, at most, with marginal increases in average execution time and the average number of prompt retries.

Overall, these findings suggest that, within the examined range, the temperature parameter has a negligible effect on the computational performance of the studied LLMs. This result corroborates the findings of related literature, which documents stable LLM performance in causal reasoning and instruction following (Li et al., 2025) as well as problem-solving tasks (Renze, 2024) across the range of temperatures examined in the current study.

Appendix A.6. Implications for the main simulation

The findings of the feasibility assessment directly inform the modelling choices adopted in the main simulation exercise. In particular, Qwen3-VL Instruct (235B) has been excluded from the subsequent analysis because of its comparatively weak performance near the regulatory constraint boundaries combined with only moderately fast response times. Although Ministral 3 (3B) exhibits similarly limited performance near the lending constraints, its substantially faster response times reduces the associated computational burden and therefore justifies its inclusion in the main analysis.

The feasibility assessment also guides the selection of the temperature parameter. The main simulations draw temperature values from a uniform distribution in the interval $[0.05, 0.85]$. Although the previous analysis found little evidence that temperature systematically affects computational performance, the chosen range maintains responses within the region commonly associated with balanced and moderately creative outputs while avoiding more extreme sampling behaviour.

Finally, the simulation limits the maximum number of prompt retries to ten. This choice balances the computational cost of repeated prompting against the objective of obtaining valid responses. The feasibility assessment showed that, for all examined models, at least 90% of successful prompts produced a valid response within ten iterations, indicating that additional retries would yield only limited gains while increasing computational overhead.

Appendix B. Prompts used in the main simulation

Each simulation exercise uses one of the alternative prompts described below as input to the LLM-driven bank agent. The prompts provide the input required to determine both the amount of credit to extend and the corresponding interest rate. The information included in each prompt is motivated by established empirical findings, common practices in ABM literature, and economic intuition.

In addition to prompt-specific information, each one includes the level of the central bank’s policy rate, DTI and DSTI limits. As discussed in Subsection 3.3, the maximum DSTI’s level and trajectory are calibrated so that simulated consumer credit replicates, to the extent possible, the historical series. Maximum DTI is fixed at 2.2 times the potential borrower’s annual income³⁴. To simplify the LLM’s computational task, each prompt incorporates the pre-calculated figures of the implied maximum monthly loan payment and maximum new loan amount. Finally, every prompt concludes with instructions on how the LLM should report the response.

Appendix B.1. Baseline

The baseline prompt is the simplest one and provides information about the potential borrower’s income and existing loan balance.

Assume you are a loan officer in a commercial bank. A customer with a monthly income of $[\text{income}_{i,t}]$ EUR and an existing loan balance of $[\text{debt}_{i,t}]$ EUR asks for a consumer loan with 5 years maturity. You will consolidate the existing debt with any amount of new credit you decide to provide into a single fully amortizing loan with 60 equal monthly annuity payments at a fixed annual nominal interest rate. The interest rate you offer will be applied to this total consolidated amount. The central bank’s policy rate is currently $[\text{policy rate}_t]\%$. Use this as a baseline when determining your offered interest rate. The upper bound of interest rate for your provided credit should be $[\text{rate}_{i,t}^{\text{max}}]\%$, which is the maximum rate consistent with the debt-service limit at the maximum permissible loan size. Your task is to decide how much new credit to provide and at what annual nominal interest rate (in percent). You must ensure that the total monthly debt service for the consolidated loan (existing loan plus new credit) is strictly less than $[\text{monthly payment}_{i,t}^{\text{max}}]$ EUR. You must also ensure that the new credit offered is less than $[\text{loan size}_{i,t}^{\text{max}}]$ EUR.

The input variables $\text{income}_{i,t}$ and $\text{debt}_{i,t}$ are the borrowing household’s i income and outstanding debt, respectively, at time t . The variable policy rate_t is the central bank’s current policy rate. Based on the applied DTI and DSTI constraints, the variables $\text{loan size}_{i,t}^{\text{max}}$ and $\text{monthly payment}_{i,t}^{\text{max}}$ are calculated as follows:

$$\text{loan size}_{i,t}^{\text{max}} = \text{DTI}^{\text{max}} \cdot 12 \cdot \text{income}_{i,t} - \text{debt}_{i,t} \quad (\text{B.1})$$

$$\text{monthly payment}_{i,t}^{\text{max}} = \text{DSTI}^{\text{max}} \cdot \text{income}_{i,t} \quad (\text{B.2})$$

Finally, $\text{rate}_{i,t}^{\text{max}}$ is the maximum interest rate which is consistent with the DSTI limit³⁵. The

³⁴Data from the OECD (*LESIM_FD4*) indicate that, at its peak, the debt of households and NPISH as a percentage of their net disposable income was about 130% from 2012 to 2014 and close to 100% until 2021. The selected limit provides the LLM-driven agent with sufficient room to capture the borrowing constraints prevailing during the simulation period.

³⁵Technically, the maximum feasible interest rate is estimated by incrementally increasing the interest rate from a predefined lower bound up to a sufficiently high upper bound, and selecting the highest rate that satisfies the DSTI constraint.

prompt provides this figure as the upper bound on the interest rate when the maximum permissible amount of credit is offered. However, it does not restrict the agent from adjusting the interest rate when it chooses to provide a smaller loan amount. Moreover, this information becomes particularly relevant when a specific loan amount is requested, which may occur when the initially offered loan exceeds the amount requested by the household, as described by Equation 3.

It should be noted that if the household requests a specific amount of credit, based on the mechanism described by Equation 3, the second and third sentences of the prompt are modified as follows:

A customer with a monthly income of $[\text{income}_{i,t}]$ EUR and an existing loan balance of $[\text{debt}_{i,t}]$ EUR asks for a consumer loan of exactly $[L_{i,t}]$ EUR with 5 years maturity. You will consolidate the existing debt with this new credit into a single fully amortizing loan with 60 equal monthly annuity payments at a fixed annual nominal interest rate.

where $L_{i,t}$ is the agreed loan amount defined by Equation 3.

Appendix B.2. Bank capitalization

On top of the baseline, bank capitalization prompt provides quantitative and qualitative information about the level and change of the bank's capital adequacy, proxied its equity to assets ratio.

*Consider also that the bank's current capitalization ratio is at the **[capitalization level description]** level of $[\text{equity}_t / \text{assets}_t]\%$, which represents **[capitalization direction description]** of **[capitalization change]** percentage points from 12 months ago. Factor this bank-specific economic condition into your credit decision, particularly regarding the loan amount and the interest rate risk premium.*

The variables *capitalization level description* and *capitalization direction description* provide qualitative information to contextualize the level of capitalization which is proxied by the equity to assets ratio. In particular, the description of the level of capitalization provides a qualitative assessment of its magnitude relative to its historical³⁶ values up to time t and can take one of the following values: {"low", "moderately low", "moderately high", "high", "very high"}, depending on whether $\text{capitalization}_t \in \{< Q1, [Q1, Q2), [Q2, Q3), [Q3, Q4), \geq Q4\}$, where $Q1$ to $Q4$ denote the first through fourth quintiles, respectively. The direction of capitalization changes is categorized as {"an increase", "a decrease", "an absence of change"}, according to the realized annual change³⁷, the latter calculated as $(\text{capitalization}_t - \text{capitalization}_{t-12})$ with its absolute value provided as input in the variable *capitalization change*.

Appendix B.3. Economic environment

Similarly, the economic environment prompt provides quantitative and qualitative information about the level and change of unemployment, as a proxy of the general economic environment, in addition to the borrower-specific input of the baseline.

*Consider also that the current unemployment rate is at the **[unemployment level description]** level of **[unemployment rate]**%, which represents **[unemployment direction description]** of **[un-***

³⁶For the first month only, description takes the generic value of "medium".

³⁷Until 12 months of data accumulate for the proper calculation of annual change, the respective figure is calculated as $12(\text{capitalization}_t - \text{capitalization}_1)/n$, where n the current number of periods in the simulation.

employment change/ percentage points from 12 months ago. Factor this general economic condition into your credit decision, particularly regarding the loan amount and the interest rate risk premium.

The qualitative assessment of the magnitude and direction of unemployment, as well as its changes, follows the same logic as that applied to capitalization.

Appendix B.4. Full information

The full information prompt is the concatenation of all previous information sets, into a single one. It combines borrower-specific, bank-specific and information on general economic environment into one prompt.

Essentially, the full-information set is: *Baseline prompt + Bank capitalization prompt + Economic environment prompt + Reporting instructions prompt.*

Appendix B.5. Reporting instructions

Finally, each prompt concludes with the following reporting instructions that ensure the figures generated by the LLM-driven agent can be consistently processed in the subsequent stages of the simulation.

Respond with exactly two positive numbers: new credit amount first, then interest rate in percent, separated by a single space. Do not include units, punctuation, or explanation.

This instruction is important in order to ensure that the LLM does not produce an answer which contains text explaining e.g. its logic which would hamper its further processing. Rather, two numbers, in specific order, are adequate for the simulation to proceed.

Appendix C. Comparison against a standard rule

To assess the potential improvements associated with the LLM-driven implementations, the study compares them against a conventional, equation-based interest rate setting rule. As a benchmark, the study employs a fixed mark-up rule where banks set the lending rate as a constant premium over the central bank's policy rate.

Appendix C.1. Calibration of the benchmark rule

The fixed mark-up rule is one of the simpler³⁸ and most commonly implemented lending rate mechanisms in the macroeconomic ABM literature, e.g. in the Schumpeter meeting Keynes ABM family (Dosi et al., 2013, 2017, 2018) and others (Palagi et al., 2017; Popoyan et al., 2020). Yet, despite its simplicity, it continues to be adopted (Dosi et al., 2021; Poledna et al., 2023; Di Domenico et al., 2025; Hommes et al., 2025). It should be noted that alternative formulations exist, incorporating borrower-specific characteristics such as creditworthiness or bank-related variables reflecting financial health. However, unlike the broad consensus regarding the implementation of the fixed mark-up rule, there is no generally accepted functional form describing how these additional factors should affect the charged interest rate. Consequently, the fixed mark-up rule provides a transparent and widely used benchmark against which the performance of the LLM-driven mechanisms can be evaluated.

In particular, the analysis examines a fixed mark-up rule of the following form:

³⁸The simplest being the constant lending rate (D'Orazio and Giulioni, 2017; D'Orazio, 2019).

$$r_t = r_{cb,t} + \mu, \quad (C.1)$$

where $r_{cb,t}$ is the central bank's policy rate at time t and μ the applied mark-up by the commercial bank.

The study applies a grid-search procedure to identify the value of the parameter μ that yields the closest match between the historical and the simulated series. The grid-search examines mark-up values ranging from 6.5% to 8.5%, in increments of 0.25 percentage points. Subsequently, it assesses each parameter's performance using the distance (MAE) and synchronization (correlation, ρ) metrics between the 50-run medians of the simulated consumer credit³⁹ and interest rate series and their historical counterparts.

Figure C.17 presents how well each mark-up can reproduce the historically observed series of consumer credit and associated interest rates.

Figure C.17: Distance versus synchronization of simulated consumer credit (blue, circles) and interest rate (red, squares) series, by μ 's value.



Numbers denote the value of the mark-up used, in percent.

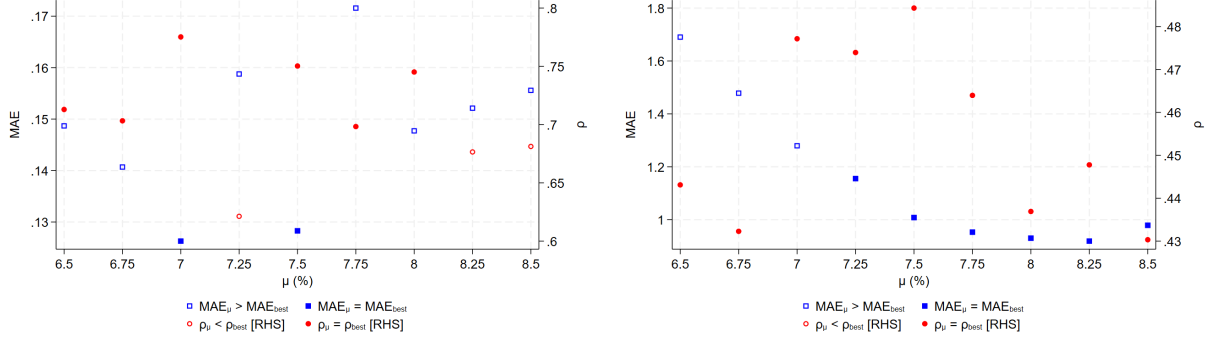
The most noticeable feature in Figure C.17 is the formation of two distinct performance clusters, associated with the two variables examined. At the bottom-right, with low MAEs and high correlations, is the one referring to the consumer credit series. At the top-left, in the high MAE and low correlation region, is the cluster of the interest rate series. This holds for every mark-up value examined, which suggests that while this specific rule can capture reasonably well the evolution of consumer credit, it cannot maintain the same performance for the associated interest rates.

However, the statistical comparison⁴⁰ of the various mark-up rules reveals that many of them exhibit equal performance to the best one. Figure C.18 shows the individual performance of each rule examined and marks the, statistically, top performing ones in each variable and metric.

³⁹As in the main analysis, artificial and historical consumer credit series have been rescaled by their full-sample average.

⁴⁰Using the same statistical tests as in the main analysis.

Figure C.18: Distance and synchronization performance by μ for consumer credit series (left panel) and interest rates (right panel).



Filled/hollow symbols denote mark-up values with statistically equal/worse performance to the best one.

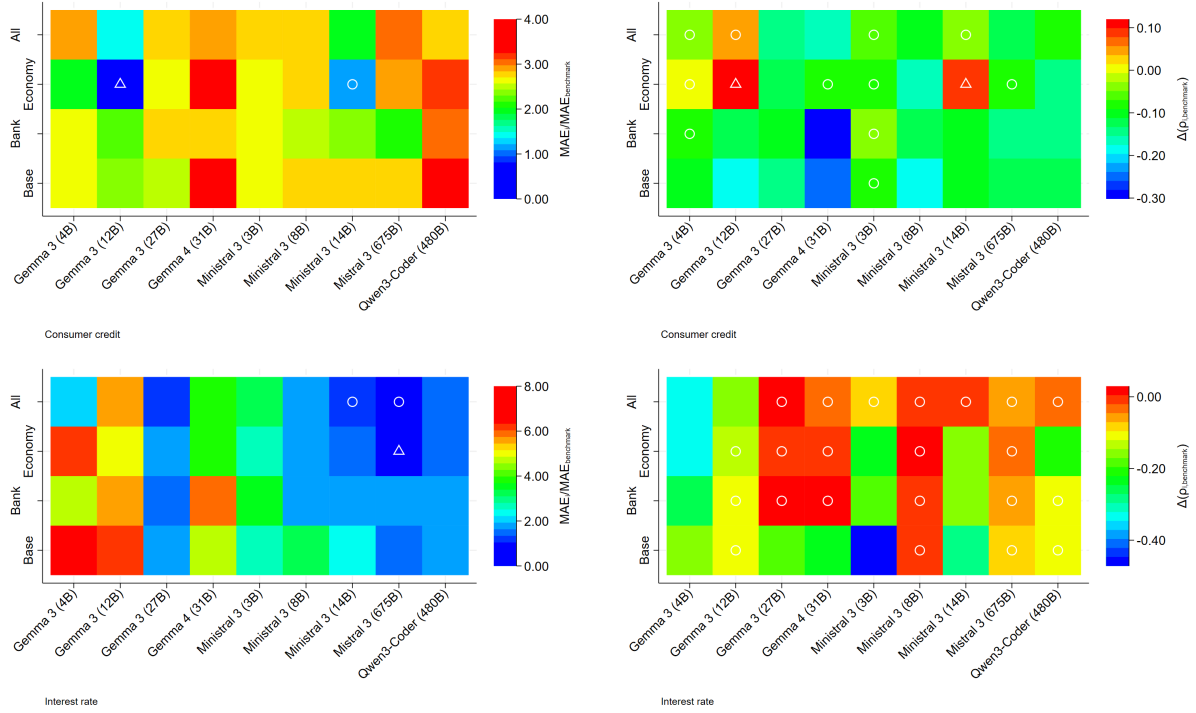
The results in Figure C.18 show that, in terms of correlation, most mark-up values exhibit statistically indistinguishable performance from the best one. Mean absolute errors, on the other hand, exhibit a different pattern. In consumer credit series, only the candidate with $\mu = 7.5\%$ has equal performance with the top one, whereas in interest rates, performance improves with mark-up magnitude up to the value of $\mu = 8.25\%$ before slightly worsening. Nevertheless, all mark-ups beyond $\mu = 7.25\%$ have statistically equal MAEs.

The previous analysis reveals that the mark-up rule exhibiting the strongest joint performance across both consumer credit and lending interest rates is the one with $\mu = 7.5\%$ which makes it the benchmark for subsequent comparison with the LLM-driven results.

Appendix C.2. LLM-driven versus equation-based interest rate setting mechanisms.

Having identified the benchmark, the study compares its performance with that of the LLM-driven interest rate setting mechanisms using the same evaluation methodology as in the main text. The analysis employs the MAE to quantify the distance between the simulated and historical series and the correlation coefficient to assess their synchronization.

Figure C.19: Performance comparison between LLM-driven versus equation-based interest rate setting mechanisms.



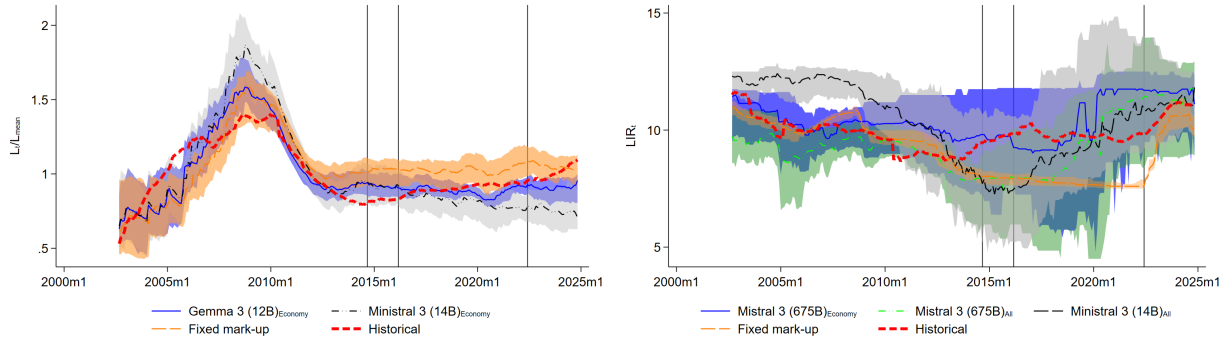
Top(bottom): consumer credit(interest rate). Left(right): $MAE(\rho)$. White triangles(circles) denote better(statistically equivalent) performance relative to the benchmark. MAE is expressed relative to the benchmark, while ρ is reported as the difference from the benchmark.

The patterns emerging from Figure C.19 are similar to those in the main analysis. In particular, Gemma 3 (12B) and Ministral 3 (14B), when provided with information about the economic environment, exhibit lower or statistically equivalent distance to the historical consumer credit series than the benchmark, while achieving statistically higher correlation. The results for interest rates present a slightly different picture. In terms of MAE, Mistral 3 (675B) clearly outperforms the benchmark when using the economic environment prompt. However, when provided with the full-information prompt, both Mistral 3 (675B) and Ministral 3 (14B) achieve performance that is statistically equal to the equation-based rule's. Differences in correlation are less pronounced, with many LLM-prompt combinations exhibiting statistically equivalent performance to the baseline rule's.

Overall, the results indicate that the best-performing LLM-prompt combinations frequently outperform the benchmark or achieve statistically equivalent performance. This finding holds over the full sample, suggesting that LLMs are at least as capable of reproducing the observed dynamics as the conventional fixed mark-up rule.

While the results from the full-sample analysis already indicate that the LLM-driven interest rate setting mechanism can provide a behavioural added value, the comparison becomes even more informative during the zero lower bound period. Figure C.20 presents the simulated trajectories of the benchmark and the best-performing LLM-driven interest rate setting mechanisms alongside the historical series for each variable.

Figure C.20: Evolution of the simulated and the historical timeseries of consumer credit (left panel) and interest rates (right panel).



Shaded areas denote the interquartile range of each simulated series. Vertical lines denote the $r_{cb} = 0.05\%$ (2014m9 - 2016m2) and zero lower bound (2016m3 - 2022m6) policy-rate periods.

The results in Figure C.20 show that in a low policy rate environment, the fixed mark-up rule performs considerably worse than several of the LLM-driven implementations. Specifically, it generates somewhat higher credit volumes while, as expected, produces lending rates substantially below the historical observations. In contrast, the LLM-driven mechanism remains considerably closer to the realized paths of both credit and lending rates. These findings suggest that incorporating richer behavioural decision rules enables bank agents to adjust lending rates in a manner that better reflects observed behaviour when conventional policy rates reach their lower bound. More broadly, they highlight the potential benefits of behavioural realism for improving the empirical performance of agent-based models, particularly during unconventional monetary policy regimes, thereby strengthening the policy relevance of the LLM-integrated framework.

Appendix D. DSTI calibration

An important variable determining the supply of credit is the maximum debt-service-to-income ratio (DSTI) allowed for households. For each credit request received, the LLM-driven bank agent is instructed to propose a loan amount and interest rate such that the resulting DSTI remains below the currently imposed maximum. When this condition is satisfied, the proposed loan is granted. Otherwise, the prompt is submitted again, with up to ten attempts allowed to obtain a response satisfying the DSTI constraint. The maximum DSTI is assumed to be determined by the bank, reflecting its risk tolerance, and is therefore introduced into the model exogenously. Nevertheless, publicly available information on DSTI values remains very limited, making its calibration from observed data very challenging. The only available source for such information is the BIS's debt service ratios statistics⁴¹. The database, as of August 2026, covers 32 countries and provides an estimate of DSTI series at an aggregate level, in quarterly frequency since 1999. While it should correctly capture how the DSTI in a particular country evolves over time, this database has two main limitations; it doesn't cover Greece, which is examined in this study and, "...it does not necessarily accurately measure [DSTI's] level relative to what one could obtain from the correct micro data" (Drehmann et al., 2015, p. 91). The former limitation extends not only to a country being included, but also to the fact that the reported DSTI series at sector level⁴² might not represent

⁴¹<https://data.bis.org/topics/DSR>

⁴²The database distinguishes between the following sectors; households & NPISHs; non-financial corporations; and private non-financial sector, the latter referring collectively to non-financial corporations and households. Naturally, the second one is the least relevant for the purposes of examining household DSTIs.

its sub-components in a sufficient manner. One can reasonably assume that the two components of total household DSTI -mortgage and consumer credit- follow different paths⁴³. Therefore, the calibration procedure aims to overcome these limitations.

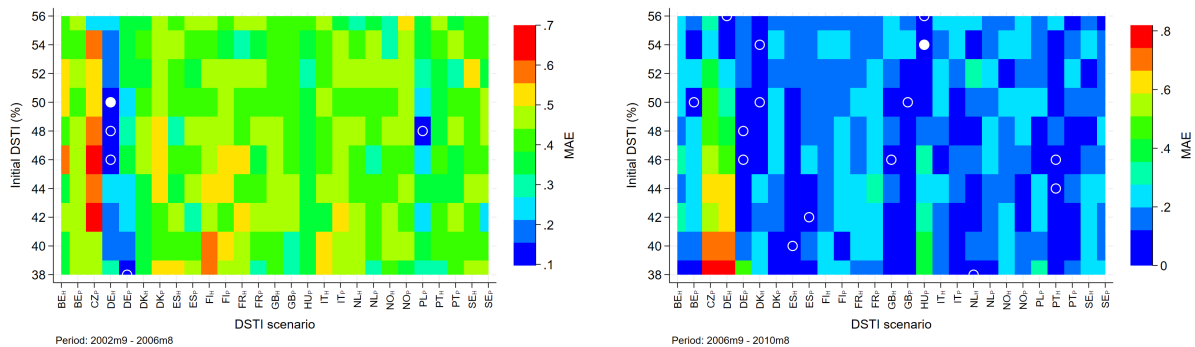
Appendix D.1. The rationale behind DSTI's calibration

The first limitation is the most difficult to address in the absence of more detailed data. Thus, it is circumvented in an axiomatic way; the study assumes that the actual DSTI series for consumer credit in Greece must have roughly similar dynamics with some of those reported in the database. The selection of the individual series to be examined in the calibration exercise is based on their presumed adequacy to resemble the unknown, real-world DSTI in Greece. The second one is approached in a grid-search way; a range of initial DSTI figures is applied on a number of country-specific DSTI growth rates and the resulting path is fed as an input in the model. For every pair of {initial DSTI, DSTI growth rate path}, the study compares the (rescaled) level and trajectory of the simulated consumer credit series with the historical ones. In addition to the country- and sector-specific DSTI scenarios, the analysis constructs several hybrid DSTI scenarios by combining the best-performing DSTI trajectories across six four-year periods⁴⁴. These hybrid scenarios are then subjected to the same grid-search procedure as if they were individual country-specific scenarios. It should be mentioned that during the DSTI calibration procedure, the interest rate was set to follow its historical evolution.

Appendix D.2. Hybrid DSTI scenario construction

To construct the hybrid scenarios, the analysis identifies the best-performing DSTI trajectories within each period and subsequently combines them. These hybrid trajectories are then evaluated alongside the country-specific trajectories in the full-sample grid-search procedure.

Figure D.21: Mean absolute error (MAE) of alternative DSTI trajectories in replicating the historical consumer credit series during the pre-crisis period, split into two subperiods: September 2002–August 2006 (left panel) and September 2006–August 2010 (right panel).



Filled circles mark the best-performing configuration and hollow circles those with statistically equal performance at 95% confidence level. Subscript H refers to the household sector, while subscript P refers to the private non-financial sector.

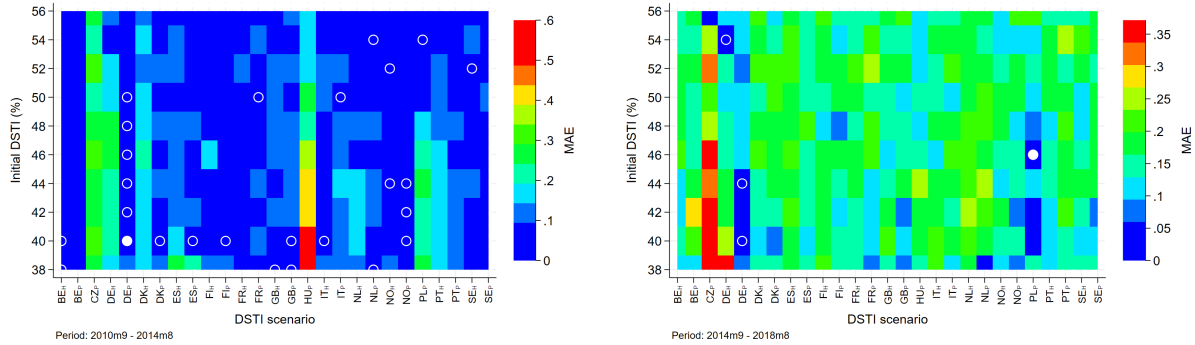
The results for the two pre-crisis subperiods in Figure D.21 reveal a clear difference. In the first subperiod, only a few DSTI trajectories stand out, while in the second subperiod many exhibit

⁴³In (Papadopoulos, 2019, p. 61) the author shows that, for the case of the US, housing and consumer credit DSTIs have indeed very different dynamics.

⁴⁴Dividing the sample into four-year periods balances two objectives: it provides sufficient temporal resolution to distinguish between different economic regimes (two subperiods pre-crisis, two during financial stress, and two during the subsequent recovery) while it maintains adequate sample sizes for robust statistical testing.

performance that is statistically equal to the best one. This contrast is also reflected in the overall distribution of MAEs. In the second subperiod, many trajectories achieve relatively low MAEs of around 0.2, even if they are not statistically equivalent to the best-performing one, whereas in the first subperiod most MAEs are clustered around 0.4.

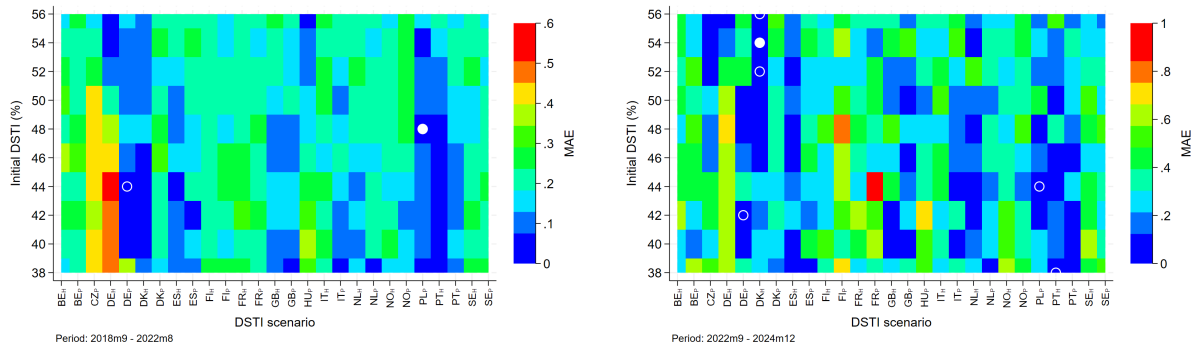
Figure D.22: Mean absolute error (MAE) of alternative DSTI trajectories in replicating the historical consumer credit series during the crisis period, split into two subperiods: September 2010–August 2014 (left panel) and September 2014–August 2018 (right panel).



Filled circles mark the best-performing configuration and hollow circles those with statistically equal performance at 95% confidence level. Subscript H refers to the household sector, while subscript P refers to the private non-financial sector.

Figure D.22 presents a similar contrast between the two subperiods of the crisis period. During the first subperiod (September 2010–August 2014), several DSTI trajectories achieve performance that is statistically indistinguishable from the best-performing one. In contrast, only two trajectories exhibit comparable performance during the second subperiod (September 2014–August 2018). This deterioration is also reflected in the MAEs, which are generally lower in the first subperiod and increase across most DSTI trajectories in the second.

Figure D.23: Mean absolute error (MAE) of alternative DSTI trajectories in replicating the historical consumer credit series during the recovery period, split into two subperiods: September 2018–August 2022 (left panel) and September 2022–December 2024 (right panel).



Filled circles mark the best-performing configuration and hollow circles those with statistically equal performance at 95% confidence level. Subscript H refers to the household sector, while subscript P refers to the private non-financial sector.

Finally, the results in Figure D.23 show that during the recovery period, few DSTI trajectories have the same performance, statistically.

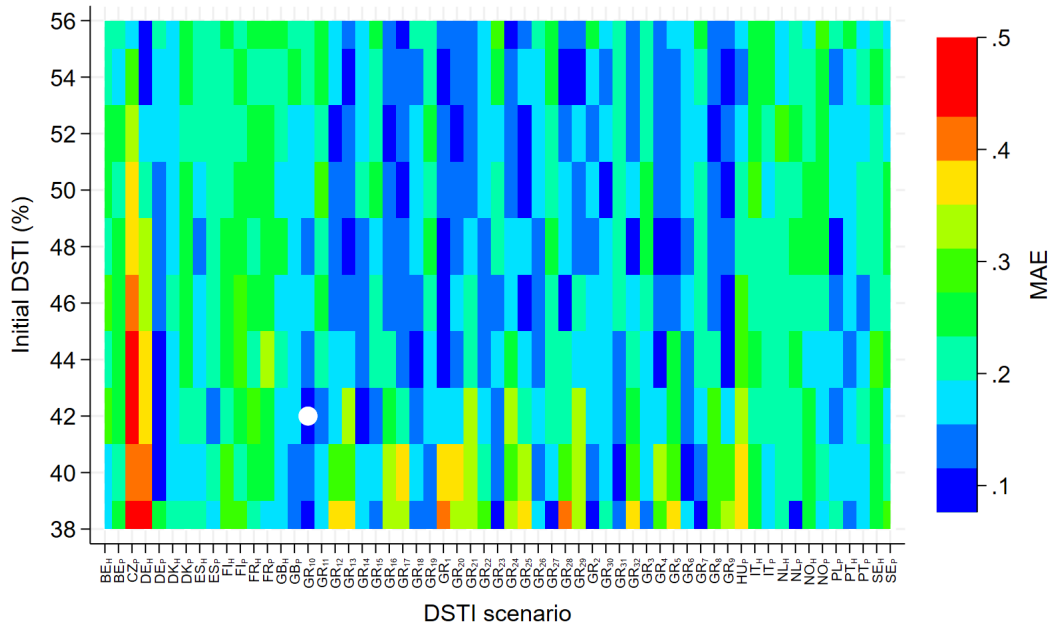
To avoid the combinatorial explosion of the hybrid scenario tree, the analysis considered, at each period, only the best-performing trajectories together with those that had the highest number

of equally well-performing representatives. Specifically, the selected trajectories are DE_H for the first, HU_P , DE_P , DK_H & PT_H for the second, DE_P & NO_P for the third, DE_P & PL_P for the fourth, DE_P & PL_P for the fifth and DK_H for the sixth period. This selection resulted in a total of 32 hybrid scenarios being evaluated.

Appendix D.3. Identification of the best-performing DSTI scenario

Having constructed the hybrid scenarios, the analysis conducts the same grid-search procedure, and evaluates the performance of every scenario over the full sample.

Figure D.24: Mean absolute error (MAE) of alternative DSTI trajectories in replicating the historical consumer credit series over the full sample (left panel) and comparison of the historical and simulated consumer credit series for the best-performing DSTI trajectory (right panel).

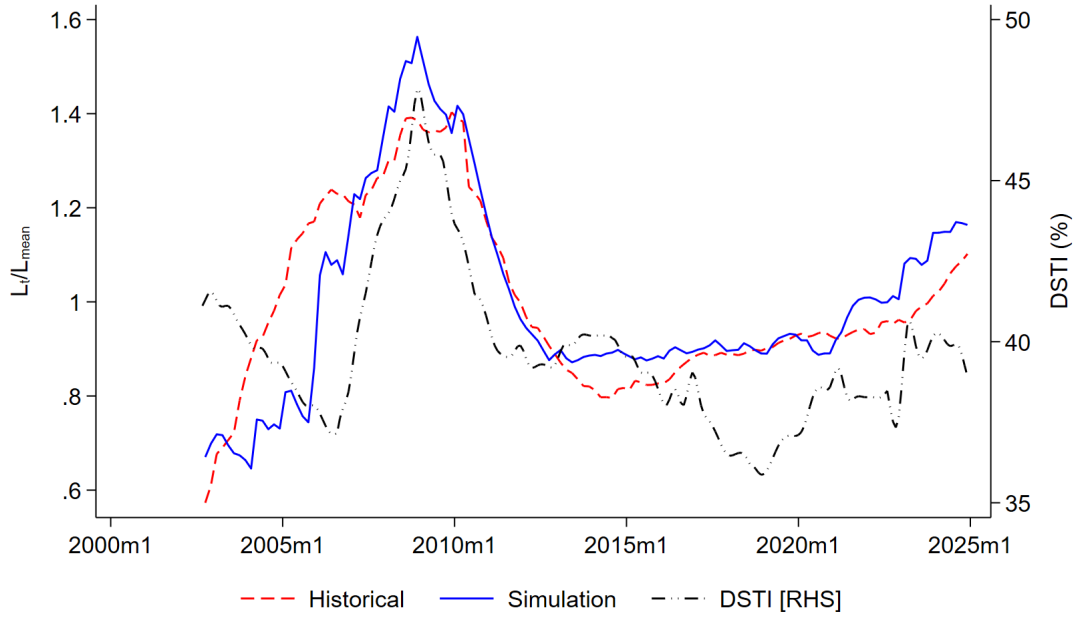


Filled circle marks the best-performing configuration. Subscript H refers to the household sector, while subscript P refers to the private non-financial sector.

The results in Figure D.24 show that the pair of {initial DSTI, DSTI growth rate path} which yields the lowest distance from the historical series of consumer credit is the hybrid scenario marked as GR_{10} , which is the combination of DE_H , DK_H , DE_P , PL_P , DE_P & DK_H individual scenarios, for the respective subperiods.

Figure D.25 presents the evolution of the selected DSTI scenario, along with the simulated and historical consumer credit series.

Figure D.25: Best-performing DSTI trajectory (dash-dotted line) together with the corresponding simulated (solid line) and historical consumer credit series (dashed line).



The DSTI trajectory shown in Figure D.25 initially declines, reflecting the path of the DE_H component used during the first subperiod. It then rises sharply until the onset of the financial crisis in Greece, in 2010, after which it enters a prolonged and pronounced decline. This downward trend is interrupted only briefly between 2013 and 2015 before continuing until around 2019, when the trajectory begins to recover, albeit with several fluctuations. The resulting simulated consumer credit series confirm their high similarity to the historical one, both in terms of distance as well as correlation⁴⁵.

Appendix E. Expectations formation under the Heuristics Switching Model

The Heuristics Switching Model (HSM) is described in detail in Anufriev and Hommes (2012a,b) and the subsequent literature. Nevertheless, a brief overview is provided here. The HSM comprises three main components: a pool of forecasting heuristics, a measure of their performance, and a rule governing the selection of heuristics.

For a generic variable y_t , the forecasting heuristics take the form of the following linear, adaptive, or trend-extrapolating rules:

A weak trend following rule (WTR):

$$y_{wtr,t}^e = y_{t-1} + \omega_{wtr} \cdot (y_{t-1} - y_{t-2}) \quad (\text{E.1})$$

A strong trend following rule (STR):

$$y_{str,t}^e = y_{t-1} + \omega_{str} \cdot (y_{t-1} - y_{t-2}) \quad (\text{E.2})$$

An adaptive expectations rule (ADA):

⁴⁵The correlation coefficient between the two series is 0.85, which is statistically equal to the highest one obtained from the hybrid scenario GR_9 with an initial DSTI of 44%, at 0.87.

$$y_{ada,t}^e = y_{t-1}^e + \omega_{ada} \cdot (y_{t-1} - y_{t-1}^e) \quad (E.3)$$

And a learning anchoring and adjustment rule (LAA):

$$y_{laa,t}^e = \frac{y_{t-1}^{av} + y_{t-1}}{2} + (y_{t-1} - y_{t-2}) \quad (E.4)$$

, where $y_{t-1}^{av} = \frac{1}{t} \sum_{j=0}^{t-1} y_j$.

The first, second and fourth heuristics (WTR, STR and LAA) are trend extrapolating ones. The distinction between the first two lies in the strength of the trend-following behaviour, with STR assigning greater weight to the prevailing trend than WTR ($\omega_{str} > \omega_{wtr}$). The fourth heuristic, LAA, is more complex, as expectations are extrapolated from an anchor that incorporates previously realized values (Anufriev and Hommes, 2012a,b). In contrast, the third heuristic follows a simple adaptive rule, combining past observations with previous expectations to determine the current expectation of y_t .

Each heuristic i 's performance is evaluated according to the following measure:

$$U_{i,t-1} = -(y_{t-1} - y_{i,t-1}^e)^2 + \eta U_{i,t-2} \quad (E.5)$$

The parameter $\eta \in [0, 1]$ determines the extent to which past prediction errors influence households' choices, thus capturing their memory. When $\eta = 0$, households disregard past performance entirely, whereas for $\eta \in (0, 1]$, past prediction errors continue to affect their decisions, although their influence gradually declines over time. This treatment is consistent with the literature on household memory and expectations, which suggests that memories can persist when households make financial decisions (Ampudia and Ehrmann, 2017) and may extend back approximately 10-15 years or even longer before their influence completely fades (Ehrmann and Tzamourani, 2012).

The final element of the HSM is the selection rule which determines the probability $n_{i,t}$ of heuristic i being selected in each period t :

$$n_{i,t} = \delta \cdot n_{i,t-1} + (1 - \delta) \frac{\exp(\beta \cdot U_{i,t-1})}{\sum_{i=1}^{i=4} \exp(\beta \cdot U_{i,t-1})} \quad (E.6)$$

Heuristics' relative performance and two behavioural parameters, δ and β , determine their stochastic selection. The first parameter, $\delta \in [0, 1]$, captures inertia in heuristic selection, reflecting households' tendency to retain their previously chosen heuristic. The second parameter, $\beta \geq 0$, referred to as the "intensity of choice", determines how strongly households respond to differences in heuristic performance and, consequently, how quickly they switch towards the better-performing heuristic. The HSM-related parameters are set to the values identified in the original studies of Anufriev and Hommes (2012a,b), which have subsequently been shown to provide a good fit across a range of different settings (Assenza et al., 2014).

References

- Alexandre, M., Lima, G.T., 2020. Combining monetary policy and prudential regulation: an agent-based modeling approach. *Journal of Economic Interaction and Coordination* 15, 385–411.
- Allen, H., Taylor, M.P., 1990. Charts, noise and fundamentals in the London foreign exchange market. *The Economic Journal* 100, 49–59.
- Allen, T.W., Carroll, C.D., 2001. Individual learning about consumption. *Macroeconomic Dynamics* 5, 255–271.
- Ampudia, M., Ehrmann, M., 2017. Macroeconomic experiences and risk taking of euro area households. *European Economic Review* 91, 146–156.
- Anufriev, M., Hommes, C., 2012a. Evolution of market heuristics. *The Knowledge Engineering Review* 27, 255–271.
- Anufriev, M., Hommes, C., 2012b. Evolutionary selection of individual expectations and aggregate outcomes in asset pricing experiments. *American Economic Journal: Microeconomics* 4, 35–64.
- Arifovic, J., Sargent, T.J., 2003. Laboratory experiments with an expectational Phillips curve. *Evolution and procedures in central banking* , 23–55.
- Assenza, T., Bao, T., Hommes, C., Massaro, D., 2014. Experiments on expectations in macroeconomics and finance, in: *Experiments in Macroeconomics, Research in Experimental Economics*. Emerald Group Publishing Limited, pp. 11–70.
- Assenza, T., Gatti, D.D., Grazzini, J., 2015. Emergent dynamics of a macroeconomic agent based model with capital and credit. *Journal of Economic Dynamics and Control* 50, 5–28.
- Assenza, T., Heemeijer, P., Hommes, C.H., Massaro, D., 2021. Managing self-organization of expectations through monetary policy: A macro experiment. *Journal of Monetary Economics* 117, 170–186.
- Axtell, R.L., Farmer, J.D., 2025. Agent-based modeling in economics and finance: Past, present, and future. *Journal of Economic Literature* 63, 197–287.
- Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al., 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631* .
- Bandourian, R., McDonald, J., Turley, R.S., 2002. A comparison of parametric models of income distribution across countries and over time. *LIS Working Paper Series*, No. 305 .
- Bao, T., Hommes, C., Pei, J., 2021. Expectation formation in finance and macroeconomics: A review of new experimental evidence. *Journal of Behavioral and Experimental Finance* 32, 100591.
- Bianchi, C., Cirillo, P., Gallegati, M., Vagliasindi, P.A., 2008. Validation in agent-based models: An investigation on the CATS model. *Journal of Economic Behavior & Organization* 67, 947–964.
- Boissay, F., Cantú, C., Claessens, S., Villegas, A., 2019. Impact of financial regulations: insights from an online repository of studies. *BIS Quarterly Review*, March .

- Boitani, A., Di Domenico, L., Ricchiuti, G., 2026. Monetary policy and inequality: a heterogeneous agents' approach. *Journal of Economic Dynamics and Control* , 105296.
- Branch, W.A., 2004. The theory of rationally heterogeneous expectations: evidence from survey data on inflation expectations. *The Economic Journal* 114, 592–621.
- Caiani, A., Godin, A., Caverzasi, E., Gallegati, M., Kinsella, S., Stiglitz, J.E., 2016. Agent based-stock flow consistent macroeconomics: Towards a benchmark model. *Journal of Economic Dynamics and Control* 69, 375–408.
- Cardaci, A., 2018. Inequality, household debt and financial instability: An agent-based perspective. *Journal of Economic Behavior & Organization* 149, 434–458.
- Chen, Y., Liu, T.X., Shan, Y., Zhong, S., 2023. The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences* 120, e2316205120.
- Cirillo, P., Gallegati, M., 2012. The empirical validation of an agent-based model. *Eastern Economic Journal* 38, 525–547.
- Dawid, H., Gatti, D.D., 2018. Agent-based macroeconomics. *Handbook of computational economics* 4, 63–156.
- Di Domenico, J., Catalano, M., Riccetti, L., 2025. Scaling and forecasting in a data-driven agent-based model: Applications to the Italian macroeconomy. *Economic Modelling* 147, 107046.
- D’Orazio, P., 2019. Income inequality, consumer debt, and prudential regulation: An agent-based approach to study the emergence of crises and financial instability. *Economic Modelling* 82, 308–331.
- D’Orazio, P., Giulioni, G., 2017. From micro behaviors to macro dynamics: An agent-based economic model with consumer credit. *JASSS* 20, 1–26.
- Dosi, G., Fagiolo, G., Napoletano, M., Roventini, A., 2013. Income distribution, credit and fiscal policies in an agent-based keynesian model. *Journal of Economic Dynamics and Control* 37, 1598–1625.
- Dosi, G., Freeman, R.B., Pereira, M.C., Roventini, A., Virgillito, M.E., 2021. The impact of deunionization on the growth and dispersion of productivity and pay. *Industrial and Corporate Change* 30, 377–408.
- Dosi, G., Napoletano, M., Roventini, A., Treibich, T., 2017. Micro and macro policies in the keynes+ schumpeter evolutionary models. *Journal of Evolutionary Economics* 27, 63–90.
- Dosi, G., Pereira, M.C., Roventini, A., Virgillito, M.E., 2018. Causes and consequences of hysteresis: aggregate demand, productivity, and employment. *Industrial and Corporate Change* 27, 1015–1044.
- Drehmann, M., Illes, A., Juselius, M., Santos, M., 2015. How much income is used for debt payments? A new database for debt service ratios. *BIS Quarterly Review* September .
- Ehrmann, M., Tzamourani, P., 2012. Memories of high inflation. *European Journal of Political Economy* 28, 174–191.
- Farmer, J.D., 2025. Quantitative agent-based models: a promising alternative for macroeconomics. *Oxford Review of Economic Policy* 41, 571–590.

- Farmer, J.D., Foley, D., 2009. The economy needs agent-based modelling. *Nature* 460, 685–686.
- Fehr, E., Tyran, J.R., 2008. Limited rationality and strategic interaction: the impact of the strategic environment on nominal inertia. *Econometrica* 76, 353–394.
- Gao, C., Lan, X., Li, N., Yuan, Y., Ding, J., Zhou, Z., Xu, F., Li, Y., 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications* 11, 1259.
- Gatti, D.D., Gallegati, M., Greenwald, B., Russo, A., Stiglitz, J.E., 2010. The financial accelerator in an evolving credit network. *Journal of Economic Dynamics and Control* 34, 1627–1650.
- Geerling, W., Mateer, G.D., Wooten, J., Damodaran, N., 2023. ChatGPT has aced the test of understanding in college economics: Now what? *The American Economist* 68, 233–245.
- Gemma Team, 2025. Gemma 3 technical report. URL: <https://arxiv.org/abs/2503.19786>, arXiv:2503.19786.
- Gemma Team, 2026. Gemma 4 technical report. URL: <https://arxiv.org/abs/2607.02770>, arXiv:2607.02770.
- Godley, W., Lavoie, M., 2016. *Monetary economics: an integrated approach to credit, money, income, production and wealth*. Springer.
- Gurgone, A., Iori, G., Jafarey, S., 2018. The effects of interbank networks on efficiency and stability in a macroeconomic agent-based model. *Journal of Economic Dynamics and Control* 91, 257–288.
- Hansen, A.L., Horton, J.J., Kazinnik, S., Puzzello, D., Zarifhonarvar, A., 2026. Simulating the survey of professional forecasters. Available at SSRN 5066286 .
- Hommes, C., 2021. Behavioral and experimental macroeconomics and policy analysis: A complex systems approach. *Journal of Economic Literature* 59, 149–219.
- Hommes, C., He, M., Poledna, S., Siqueira, M., Zhang, Y., 2025. CANVAS: A canadian behavioral agent-based model for monetary policy. *Journal of Economic Dynamics and Control* 172, 104986.
- Hommes, C., Massaro, D., Weber, M., 2019. Monetary policy under behavioral expectations: Theory and experiment. *European Economic Review* 118, 193–212.
- Hommes, C., Sonnemans, J., Tuinstra, J., Van de Velden, H., 2005. Coordination of expectations in asset pricing experiments. *The Review of Financial Studies* 18, 955–980.
- Hommes, C., Sonnemans, J., Tuinstra, J., Van de Velden, H., 2008. Expectations and bubbles in asset pricing experiments. *Journal of Economic Behavior & Organization* 67, 116–133.
- Horton, J.J., Filippas, A., Manning, B.S., 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical Report. National Bureau of Economic Research.
- Kazinnik, S., Sinclair, T.M., 2025. FOMC in silico: A multi-agent system for monetary policy decision modeling. Available at SSRN 5424097 .

- Korinek, A., 2025. AI agents for economic research: August 2025 update to ‘generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature* 61.
- Li, L., Sleem, L., Nichil, G., State, R., et al., 2025. Exploring the impact of temperature on large language models: Hot or cold? *Procedia Computer Science* 264, 242–251.
- Li, N., Gao, C., Li, M., Li, Y., Liao, Q., 2024. Econagent: large language model-empowered agents for simulating macroeconomic activities, in: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15523–15536.
- Liu, A.H., Khandelwal, K., Subramanian, S., Jouault, V., Rastogi, A., Sadé, A., Jeffares, A., Jiang, A., Cahill, A., Gavaudan, A., 2026. Ministral 3. *arXiv preprint arXiv:2601.08584*.
- Mankiw, N.G., Reis, R., Wolfers, J., 2003. Disagreement about inflation expectations. *NBER macroeconomics annual* 18, 209–248.
- Marimon, R., Sunder, S., 1995. Does a constant money growth rule help stabilize inflation? Experimental evidence, in: *Carnegie-Rochester Conference Series on Public Policy*, Elsevier. pp. 111–156.
- Meijers, H., Nomaler, Ö., Verspagen, B., 2019. Demand, credit and macroeconomic dynamics. a micro simulation model. *Journal of Evolutionary Economics* 29, 337–364.
- Modigliani, F., Brumberg, R., 2005. Utility analysis and the consumption function: An interpretation of cross-section data, in: *The Collected Papers of Franco Modigliani, Volume 6*. The MIT Press, pp. 3–45.
- Palagi, E., Napoletano, M., Roventini, A., Gaffard, J.L., 2017. Inequality, redistributive policies and multiplier dynamics in an agent-based model with credit rationing. *Italian Economic Journal* 3, 367–387.
- Papadopoulos, G., 2019. Income inequality, consumption, credit and credit risk in a data-driven agent-based model. *Journal of Economic Dynamics and Control* 104, 39–73.
- Park, J.S., Zou, C.Q., Kamphorst, J., Egan, N., Shaw, A., Hill, B.M., Cai, C., Morris, M.R., Liang, P., Willer, R., Bernstein, M.S., 2024. Llm agents grounded in self-reports enable general-purpose simulation of individuals. *arXiv preprint arXiv:2411.10109*.
- Pfajfar, D., Santoro, E., 2010. Heterogeneity, learning and information stickiness in inflation expectations. *Journal of Economic Behavior & Organization* 75, 426–444.
- Pfajfar, D., Žakelj, B., 2018. Inflation expectations and monetary policy design: Evidence from the laboratory. *Macroeconomic Dynamics* 22, 1035–1075.
- Poledna, S., Miess, M.G., Hommes, C., Rabitsch, K., 2023. Economic forecasting with an agent-based model. *European Economic Review* 151, 104306.
- Popoyan, L., Napoletano, M., Roventini, A., 2020. Winter is possibly not coming: Mitigating financial instability in an agent-based model with interbank market. *Journal of Economic Dynamics and Control* 117, 103937.
- Reissl, S., 2020. Minsky from the bottom up—formalising the two-price model of investment in a simple agent-based framework. *Journal of Economic Behavior & Organization* 177, 109–142.

- Reissl, S., 2021. Heterogeneous expectations, forecasting behaviour and policy experiments in a hybrid agent-based stock-flow-consistent model. *Journal of Evolutionary Economics* 31, 251–299.
- Renze, M., 2024. The effect of sampling temperature on problem solving in large language models, in: *Findings of the association for computational linguistics: EMNLP 2024*, pp. 7346–7356.
- del Rio-Chanona, R.M., Pangallo, M., Hommes, C., 2025. Can generative AI agents behave like humans? evidence from laboratory market experiments. *arXiv preprint arXiv:2505.07457*.
- Russo, A., Riccetti, L., Gallegati, M., 2016. Increasing inequality, consumer credit and financial fragility in an agent based macroeconomic model. *Journal of Evolutionary Economics* 26, 25–47.
- Wang, R., Wang, R., Shen, Y., Wu, C., Zhou, Q., Chandra, R., 2026. Evaluation of llms for mathematical problem solving. *Next Research*, 101705.
- Williams, E.J., 1959. *Regression analysis*. Wiley, New York.

BANK OF GREECE WORKING PAPERS

- 353. Andreou, P., Anyfantaki, S., Dellis, K., and C. Makridis, “Doing matters more than knowing: evidence from environmental preferences”, December 2025.
- 354. Chrysanthakopoulos, C. and A. Tagkalakis, “Natural disasters and the effects of reconstruction expenditure on output”, December 2025.
- 355. Gazilas, E.T. and Z. Bragoudakis, “Can higher education levels improve tax compliance?”, January 2026.
- 356. Delis, P. and G. Kontogeorgos, “Outlier-robust evaluation of fixed-event macroeconomic survey expectations”, January 2026
- 357. Gautier, E, Conflitti, C., Enderle, D., Fadejeva, L., Grimaud, A., Gutiérrez, G., Jouvanceau, V., Menz, J. O., Paulus, A., Petroulas, P., Roldan-Blanco, P., and E. Wieland, “Consumer price stickiness in the euro area during an inflation surge”, February 2026
- 358. Anyfantaki, S., Giannakidis, H., Malliaropoulos, D., Migiakis, P., and F. Petroulakis, “Bond funds' risk taking and monetary policy”, February 2026
- 359. Karamanis, D., Kotsogiannis, C., and E. Papapetrou, “Threshold-based policies and distortions in housing markets, March 2026
- 360. Louzis, D., “Trend inflation and inflation expectations in high dimensional vector autoregressions”, March 2026
- 361. Degiannakis, S., Filis, G., and G. Siourounis, “Cryptokurtosis: frequent trading fuels higher losses”, May 2026
- 362. Mermelas, G., and A. Tagkalakis, “The transmission of monetary policy through credit standards: evidence on loan application discouragement in the euro area”, May 2026
- 363. Bragoudakis, Z., Karakitsios, A., and E. Kasimati, “Short-term inflation projections: the new BoG’STIP model”, June 2026
- 364. Lampousis, A., and P. Migiakis, “Central bank reserves and banks’ portfolio rebalancing in the euro area”, June 2026
- 365. Karamanis, D., Louzis, D.P., Papapetrou, E., and A. Theofilakou, “Modelling house price dynamics in Greece, June 2026
- 366. Louzis, D. P., “The econometrics of the euro-area natural rate of interest”, July 2026
- 367. Bidder, R., Papadimitriou, D., Rognone, L., and S. Zhang, “Climate risks, short and long”, August 2026
- 368. Dellas, H., Hall, S. and G. Tavlás, “On the transmission of purchasing power parity in a group setting”, September 2026
- 369. Gefang, D., Hall, S. and G. Tavlás, “Dynamics of the currency composition of central bank reserves”, September 2026
- 370. Sideris, D. and G. Pavlou, “Excessive cost-push inflation in Greece: the role of market power”, October 2026