ENRICHING AGENT-BASED MODELING'S WITH LLM-DRIVEN MBTI PERSONALITIES: A BENCHMARK FRAMEWORK

Si Hao Li
Master IREN
Université Paris Dauphine
Paris, France
si.li@ens-paris-saclay.fr

Olivia Carnapete
Master IREN
Université Paris Dauphine
Paris, France
olivia.carnapete@dauphine.eu

ABSTRACT

This article introduces an innovative approach to macroeconomic modeling that integrates Agent-Based Models (ABM) with Large Language Models (LLM) within a sandbox environment. In this benchmark study, agents are endowed with distinct MBTI (Myers-Briggs Type Indicator) personality profiles to simulate heterogeneous decision-making processes in an economic context. Building on established theoretical foundations from the ABM literature, our simulation framework incorporates adaptive learning and local interactions, enabling agents to generate realistic behaviors via LLM-driven responses. Preliminary findings suggest that the diversity of personality traits significantly influences aggregate economic dynamics, offering new insights for model calibration and validation. This benchmark not only bridges computational economics and AI-based behavioral modeling but also provides a versatile platform for future interdisciplinary research in simulating complex economic phenomena.

***K*eywords** ABM · LLM · MBTI · Macroeconomics

1 Introduction

Agent-based models (ABMs) have gradually emerged as an alternative approach in macroeconomics, largely to overcome the limitations of traditional models lacking microfoundations [1]. Rather than relying on a representative agent, an ABM simulates a population of "micro-agents" (households, firms, banks, etc.) interacting within an economic environment. Macroeconomic phenomena (cycles, crises, income distribution, and so on) then emerge from these local interactions in a bottom-up fashion. The absence of an instantaneous equilibrium assumption, along with the inclusion of heterogeneous agents, provides a unique flexibility that facilitates the exploration of complex economic systems [2]. ABMs can thus model a veritable "ecology" of actors, linked by networks or markets, and highlight coordination or financial contagion phenomena that are difficult to capture in a standard DSGE framework [3]. However, this richness introduces several challenges: first, empirical validation is hampered by the lack of a reduced form or directly usable likelihood function, leading to indirect calibrations and making it difficult to isolate the precise contribution of each assumption [4]. Second, the inherent complexity of nonlinear interactions can render ABMs "black boxes," where it becomes arduous to trace the influence of each parameter step by step. Finally, in terms of prediction, it can be problematic to confront ABM simulations with historical data in the same way as a DSGE model, which offers more directly testable analytical solutions.

Nonetheless, the recent emergence of large language models (LLMs)—such as GPT-4 or BERT—opens up new avenues for overcoming some of these limitations. Trained on massive text corpora, these AI systems can generate coherent responses and partially mimic human reasoning. By integrating them as the "decision engine" of agents in an ABM, one could endow these agents with a level of behavioral flexibility and realism far exceeding that provided by a few manual rules [5, 6]. Rather than resorting to ad hoc behavioral equations (e.g., propensity to consume, price-setting rules, etc.), an agent could "consult" an LLM to decide on its action (consumption, saving, investment) by taking into account a textual context describing its situation. Because the language model has been trained on real-world data, it

brings a form of "common sense" or implicit psychological tendencies (biases, risk aversion, etc.) likely to significantly enrich the simulation.

In this light, we propose to take a further step by experimenting with a sandbox for LLM agents that exhibit MBTI personality traits within a closed economy. The goal is to create an environment where each agent has not only an LLM-based decision engine but is also assigned a personality type (e.g., ENTJ, INFP, etc.) that shapes its choices, expectations, and interactions with others. The idea is to build an artificial ecosystem that, in stylized form, reproduces the diversity of human agents and their potential biases: for instance, an introverted agent might exhibit greater caution or risk aversion, whereas an extraverted, intuitive agent might adopt bolder strategies when investing or consuming.

This approach has a twofold interest. First, it makes it possible to test the robustness and relevance of LLMs as "proxies" for diverse human behaviors: do we observe patterns of interaction that correspond, even qualitatively, to real-world dynamics? Second, this sandbox serves as a testbed for research on LLMs and ABMs: by precisely controlling the environment and available resources, one can examine how various prompt parameters, personality traits, or network structures affect aggregate outcomes. This paves the way for creating a "benchmark" for future studies that employ LLMs as economic agents, by testing different configurations and comparing their propensity to replicate—or fail to replicate—certain stylized facts.

This study is therefore both exploratory and methodological: exploratory, in that it seeks to investigate how LLMs endowed with MBTI personalities interact and give rise to macroeconomic dynamics in a closed environment; methodological, because this configuration could inspire future work aiming to refine AI integration into more advanced ABMs. We posit that behavioral heterogeneity—already central to the ABM framework—finds new ground here: rather than defining decision rules ex ante, we leave a significant role to the "creativity" of language models, calibrated and guided by their simulated personality.

The remainder of this paper is organized as follows. Section 1 briefly discusses the contribution of LLMs as economic agents in the emerging literature. Section 2 describes the sandbox setup: the closed economy, the assignment of MBTI personalities, and the interaction mechanisms. Section 3 presents the initial results and analyzes the micro-macro dynamics that emerge from these experimental simulations. We conclude by discussing the limitations of this prototype and propose avenues for future work on LLM-powered ABMs.

2 Methods

In line with the "generative agents" sandbox approach described by Park et al. (2023) [7], we developed a simulation environment in which each agent is driven by a large language model (LLM) and endowed with a realistic MBTI personality. Specifically, we model a stylized closed economy (markets, institutions, transaction flows) where agents interact autonomously: at each time step, they make decisions (consumption, investment, borrowing, etc.) by querying the LLM with a prompt detailing their current situation (wealth, debt, recent events) and their personality traits. In order to reflect the behavioral diversity observed in the global population, the distribution of the 16 MBTI types is aligned with the empirical proportions highlighted in the MBTI literature, thereby ensuring a realistic representation of the overall profile. Each agent stores its past economic experiences (executed transactions, contracted loans, notable social interactions) in a persistent memory module, consistent with the principles of retrieval and reflection proposed by Park et al. (2023). When an agent plans its next action, this module retrieves the most relevant memories (e.g., a negative borrowing experience) to guide the decision produced by the LLM. The outcomes of these decisions (gains, losses, reactions of other agents) are then aggregated at the scale of the simulated economy, allowing us to observe the emergence of macroeconomic dynamics while taking into account the psychological variability of the agents, made possible by the explicit integration of both an LLM and an MBTI distribution.

Our experimental approach relies on a hybrid architecture that combines this sandbox with an economy and large language models. Each agent A_i is defined by a dynamic tuple (Ψ_i, M_i, L_i), where $\Psi_i \in \{\text{ISFJ}, \text{ISTJ}, \ldots, \text{ENFP}\}$ denotes the MBTI type, $M_i = (w_i, a_i, d_i)$ represents the economic state (wealth, assets, debt), and L_i the location in the spatial graph $\mathcal{G}$. Unlike traditional Agent-Based Models (ABMs), the decision function emerges from a generation constrained by the LLM:

$$a_i^{t+1} = \Gamma_{\text{LLM}}\big(\phi(\Psi_i) \,\|\, \gamma(M_i^t) \,\|\, \xi(\mathcal{E}^t)\big), \tag{1}$$

where $\phi(\cdot)$ encodes the MBTI traits, $\gamma(\cdot)$ encodes the current economic state, and $\xi(\cdot)$ encodes the market conditions. The implementation uses Meta's `Llama-3.2-7B-Instruct` model, but it is entirely feasible to use other open-source LLMs or API-based calls. Spatial interactions follow a decentralized logic: at each period, agents move randomly within the location graph $\mathcal{G}$, and the probability of encountering each other is proportional to node connectivity:

$$P(A_i \leftrightarrow A_j) \propto \frac{1}{\deg(L_i) + \deg(L_j)}. \tag{2}$$

Economic transactions emerge from these encounters, generating macroeconomic dynamics measured through three key indicators. The periodic GDP is given by

$$\text{GDP}_t = \sum_{\tau \in \mathcal{T}_t} v(\tau),$$

where $\mathcal{T}_t$ denotes the set of transactions and $v(\tau)$ their monetary value.

Within the framework of this simulation, we chose to situate our experiments in an extended version of the village of Phandalin proposed in one of Park et al. (2023)'s GitHub extensions, a small settlement drawn from the Dungeons & Dragons universe and adapted here to accommodate more complex economic dynamics. Phandalin is described as a semi-rural locality undergoing development, featuring a few notable buildings and several commercial activities. We introduced three new zones explicitly tied to production, commerce, and governance, thus modeling a genuine network of economic interactions.

3 Results

The raw log excerpt shown below illustrates the entire processing pipeline executed during a single simulation tick. For every agent, the large-language model first produces a concise, past-tense narration (e.g. "*I reviewed the grand library's dusty tomes...*"), thereby ensuring grammatical uniformity. This textual output is subsequently parsed into one or more structured tuples *(Action,Object)*—for example, *(Review, grand library's dusty tomes and records)*—which act as the interface between free narration and the model's economic logic. Whenever the extracted action is financial in nature (*borrow*, *consume*, etc.), it triggers a call to the banking or market module; the associated transaction is then logged instantaneously with the tag `[Bank]` or `[Economy]`, together with the amount, counterparties, and the agent's updated wealth position. These logs thus provide the micro-founded trace that links the LLM's textual decision to the monetary flow feeding into GDP calculations and the aggregate indicators.

Agent: Toblen Stonehill (ENFP) — Time step 14
LLM narration: *"I reviewed the grand library's dusty tomes and records."*
Extracted tuple: (Review, grand library's dusty tomes and records)
Economic decision: loan of 25 gold pieces by selling books (5% interest rate).
Balance-sheet effect: wealth increased from 0 gp to 25 gp.

Processing the raw logs. Following the qualitative presentation of the textual trace, the raw log is converted into a structured dataset suitable for quantitative analysis. The procedure unfolds in two stages. (i) Three regular expressions are employed to detect, in turn, (a) the phase delimiter `Phase économique:`, (b) the `[Economy] Transaction enregistrée` line that contains the buyer's identity and the transaction amount, and (c) the `[Economic Decision]` line that records the agent's MBTI type, the free-form decision string, and the post-transaction wealth. Whenever the buyer's name matches across (b) and (c), the information is merged into a single record *(phase,buyer,mbti,decision_type,amount,final_wealth)*, where `decision_type` is mapped onto five categories (*consume*, *invest*, *borrow*, *entrepreneur*, *other*). (ii)These records are loaded into a pandas `DataFrame` and aggregated by phase: we compute total consumption, investment (invest+entrepreneur), borrowing flows, and the average wealth of all active agents. The resulting table `df_macro` condenses the micro–macro dynamics of the simulation and forms the empirical backbone for the statistical analyses reported after.

3.1 Descriptive evidence from the benchmark run

Table 2 reports the four aggregate series for the first twenty phases of the closed–economy sandbox. Two regularities emerge.

1. **Consumption-led demand.** Throughout the sample, household consumption systematically exceeds investment, averaging $\bar{C} = 94$ gold pieces (gp) against $\bar{I} = 34$ gp.[1] The consumption peak (130 gp, phase 6) coincides with *zero* net borrowing, whereas the local trough (45 gp, phase 12) materialises during the largest credit surge (187 gp).

[1]All monetary magnitudes are denominated in gp for consistency with the Phandalin setting.

Table 1: Excerpt of the micro-level dataset `df`

Phase	Buyer	Amt.	Description	MBTI	Type	Wealth	Decision raw
1	Toblen Stonehill	7	Consumption	ISFJ	consume	83	consume goods worth 7
1	Daran Edermath	19	Consumption	ISFP	consume	33	consume goods worth 19
1	Linene Graywind	10	Consumption	ESTJ	consume	87	consume goods worth 10
1	Halia Thornton	7	Consumption	ESTJ	consume	68	consume goods worth 7
1	Qelline Alderleaf	11	Consumption	ESFP	consume	66	consume goods worth 11
				...			
27	Harbin Wester	18	Consumption	ESFP	consume	15	consume goods worth 18
27	Terrill Bloodscar	22	Entrepreneurship	ENFP	invest	17	start a business with initial investment 22
27	Conrad Scarface	22	Borrowing	ESFP	borrow	50	borrow 22 from bank
27	Nellie Starsmith	7	Consumption	ESFJ	consume	40	consume goods worth 7
27	Valerie Grinblade	26	Investment	ENFP	invest	16	invest 26 in an asset

Note. For brevity we display the first five and the last five observations; the full dataset contains 297 rows and eight columns.

Table 2: Aggregate flows and average wealth over the first 13 phases

Phase	Consumption	Investment	Borrowing	Avg. Wealth
1	107	55	0	78
2	126	13	43	69
3	123	56	45	57
4	87	55	38	47
5	85	18	143	51
6	130	57	0	34
7	90	0	171	41
8	67	59	138	42
9	105	51	38	32
10	57	0	176	42
11	111	29	78	37
12	45	25	187	47

2. **Burst-like credit cycles.** Borrowing is highly episodic: three spike phases (7, 10, and 12) account for 59 % of the total credit granted. Average wealth declines steeply from 77.6 gp in the initial phase to 31.6 gp by phase 9, before stabilising within a 34–48 gp corridor, reflecting an early consumption boom followed by gradual deleveraging.

Figure 1 plots consumption and investment over the full horizon ($t = 1{:}27$). Both series display a statistically significant negative linear trend (dashed lines), indicating that, absent exogenous income injections, purchasing power is slowly decumulated. Short-run volatility, however, is substantial: consumption fluctuates between 45 and 140 gp, while investment ranges from null (phases 7, 10, 16) to 64 gp (phase 15). The contemporaneous correlation is only $\rho_{C,I} = 0.28$, suggesting that investment is not merely a residual of unconsumed resources but is driven by distinct decision triggers.

Micro-level evidence (Table 1) clarifies those triggers. Among the 297 recorded transactions, 62 % are classified as `consume`, 21 % as `invest` (including `entrepreneur`), and 11 % as `borrow`; the remainder falls into the catch-all category `other`. Borrowing decisions are concentrated among ESFP and ENFP profiles, in line with their documented risk affinity, whereas ISFJ agents dominate precautionary consumption. Entrepreneurial entries—such as Terrill Bloodscar's *"start a business with initial investment 22"*— are rare (3 % of decisions) but large in scale, accounting for 14 % of the total investment flow.

Taken together, these descriptive facts indicate that the LLM-driven agents reproduce three canonical features of real economies: (i) demand dominated by household consumption; (ii) lumpy, burst-type credit cycles; and (iii) heterogeneous investment behaviour tightly linked to psychological traits. The next subsection turns to a formal investigation of the way these micro decisions propagate into aggregate volatility.

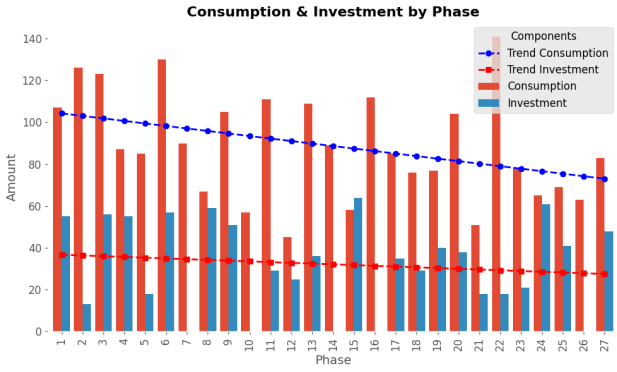

Figure 1: C & I

4 Conclusion and limits

This exploratory study shows that an open-source large language model—here, `Llama-3.2-7B-Instruct`—can be embedded in a multi-agent framework to generate both coherent micro-level narratives and plausible macro-economic dynamics. The findings, however, should be interpreted with caution.

First, the simulation engine is still a prototype: its code architecture, memory subsystem, and logging routines are not yet fully modularised or rigorously tested, limiting reproducibility and preventing exhaustive sensitivity analysis. Second, all experiments were run locally on a single 8 VRAM GPU and 40 GB RAM, constraining model size, prompt length, and the number of agents ($N \leq 10$) and locations ($|\mathcal{G}| = 11$); scaling to realistic populations would require distributed computing or model distillation.

Further constraints arise from the intrinsic limits of the selected model, whose reasoning ability remains below that of proprietary GPT-4-class solutions. The absence of Monte-Carlo replications and external validations leaves our statistical observations provisional. Finally, agent personalities remain static (MBTI), and the highly stylised markets incorporate neither price formation nor default risk.

These limitations suggest several avenues for future research: (i) refactoring the code into a modular platform, (ii) leveraging multi-GPU clusters to broaden simulation scope, (iii) systematically comparing open-source models with proprietary APIs, (iv) integrating endogenous learning mechanisms and price dynamics, and (v) conducting large-scale replication campaigns to strengthen the empirical robustness of emergent macro-economic patterns.

References

[1] Robert E Lucas Jr. Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, volume 1, pages 19–46. North-Holland, 1976.

[2] Sylvain Mignot and Annick Vignes. The many faces of agent-based computational economics: Ecology of agents, bottom-up approaches and paradigm shift. *Œconomia. History, Methodology, Philosophy*, (10-2):189–229, 2020.

[3] Joseph E Stiglitz and Mauro Gallegati. Heterogeneous interacting agent models for understanding monetary economies. *Eastern Economic Journal*, 37:6–12, 2011.

[4] Giorgio Fagiolo, Mattia Guerini, Francesco Lamperti, Alessio Moneta, and Andrea Roventini. Validation of agent-based models in economics and finance. *Computer simulation validation: fundamental concepts, methodological frameworks, and philosophical perspectives*, pages 763–787, 2019.

[5] Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. Econagent: large language model-empowered agents for simulating macroeconomic activities. *arXiv preprint arXiv:2310.10436*, 2023.

[6] John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

[7] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.