Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

Estimating Decision Rules from Large Language Model Agents: An Application to Bank Balance-Sheet Management

Zlatin Sarastov

zlatin.sarastov@unwe.bg

University of National and World Economy

Research Article

Keywords: agent-based modelling, large language model agents, decision-rule estimation, out-of-sample evaluation, bank balance-sheet management, interest rate risk

Posted Date: October 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11429290/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Estimating Decision Rules from Large Language Model Agents: An Application to Bank Balance-Sheet Management

Abstract

This paper examines whether decisions generated by large language model (LLM) agents can be represented by simple equations in an agent-based model of bank balance-sheet management. In a monthly leader–follower model, a stylised, predominantly deposit-funded euro-area bank chooses deposit rates, lending margins and the composition of bond reinvestment, while depositors and borrowers respond to these decisions and to economic conditions. The initial balance sheet is calibrated to selected indicators from regulatory disclosures. We estimate equation-based approximations of LLM-generated decisions and evaluate them against simulations not used in estimation. Across ten out-of-sample bank-year simulations, the final specification reproduces annual return on equity (ROE) within 0.31 percentage points and meets all fit criteria in seven; none of the deposit-run scenarios meets all criteria. In eighteen simulations with a second LLM family, the ROE difference does not exceed 0.71 percentage points, but only six meet all criteria. Against a market-pricing rule, the original optimisation model, and regression and nearest-neighbour approximations estimated on the same data, the final specification meets all criteria in more years and tracks liquidity most closely in deposit-run years. Instructions without behavioural guidance show that withdrawal magnitudes and fixed-rate borrowers' price sensitivity depend materially on the instructions supplied; with customer equations estimated under neutral instructions, a two-month withdrawal episode no longer becomes self-sustaining. The results demonstrate a procedure for approximating and evaluating LLM-generated decisions within a defined simulation environment; they do not establish that these responses represent the behaviour of bank managers or customers.

Keywords: agent-based modelling; large language model agents; decision-rule estimation; out-of-sample evaluation; bank balance-sheet management; interest rate risk

JEL classification: C63, C72, G21, E43, G28

1. Introduction

The monetary tightening of 2022–2023 illustrated the relationship between banks' interest-rate exposures and funding risks. Policy rates rose by more than 400 basis points in the United States and the euro area, reducing the market value of longer-term fixed-rate assets. Banks with deposits that could be withdrawn rapidly faced additional pressure. Silicon Valley Bank provides a prominent example (Jiang et al., 2024; Cookson et al., 2023). Estimates for euro-area banks suggest that deposit franchise value and hedging positions partly offset unrealised losses during this period (Rice and Guerrini, 2025). Assessing these effects requires assumptions about how depositors respond to interest rates, bank conditions and publicly available information.

Three considerations motivate the analysis. First, regulatory assessments depend on assumptions about depositor behaviour that may change with bank pricing and economic conditions. The Basel standard on interest rate risk in the banking book (IRRBB) and the EBA guidelines require banks to assess the repricing and withdrawal behaviour of non-maturity deposits (BCBS, 2016; EBA, 2022), while the liquidity coverage ratio (LCR) specifies runoff assumptions by deposit category (BCBS, 2013). Coulier et al. (2025) find relatively limited changes in euro-area banks' deposit-maturity assumptions during the tightening period. Second, empirical estimates of deposit-rate pass-through, net interest income and economic-value sensitivities provide information about observed outcomes but do not, by themselves, determine the consequences of alternative bank pricing and portfolio choices. Third, much agent-based banking research focuses on contagion and asset sales rather than the decisions of an individual bank and its customers. LLM agents provide an additional source of simulated decisions that may be useful for specifying such models, subject to careful testing of their dependence on prompts and model choice.

Interest rate risk in the banking book is commonly assessed using two related measures. Net interest income (NII) reflects an earnings perspective over a specified horizon, whereas the economic value of equity (EVE) reflects the present value of interest-rate-sensitive cash flows. Two unpublished empirical calibration analyses of euro-area disclosures are used below (denoted Paper 1 and Paper 2). Paper 1 reports a median ΔEVE of −6.1% of Tier 1 capital and a median ΔNII of +0.7% of Tier 1 capital under a +200-basis-point shock, together with deposit-rate pass-through that varies with the level of rates. Paper 2 measures the share of funding potentially subject to withdrawal within 30 days.

Those empirical findings motivate a complementary question: how might a bank's deposit pricing, lending margins and reinvestment duration interact with customers' responses to produce different earnings, valuation and liquidity outcomes? We examine this question using a deliberately parsimonious simulation model.

The study addresses two questions. First, can equation-based decision rules be estimated from LLM-generated choices and reproduce decisions in scenarios not used for estimation? Second, within the simulated banking environment, how do the bank's pricing and duration decisions interact with depositor and borrower responses to affect NII, EVE and liquidity? The second question concerns the properties of the specified model; it is not a direct test of human behaviour.

The principal contribution is methodological. We set out a procedure for estimating decision rules from LLM-generated responses, comparing them with equation-based specifications and evaluating revised specifications on separate simulations. The procedure also specifies when the estimates should be reconsidered after new tests. Its application is a monthly leader–follower model of bank pricing, interest-rate exposure and liquidity management, initialised using selected Pillar 3 indicators. The banking application allows the procedure to be examined in an economically interpretable setting; it does not provide independent behavioural identification or a direct empirical validation of customer decision rules.

The model consists of twelve monthly decisions. The bank determines two deposit rates, two new-loan margins and the allocation of bond reinvestment. Four customer categories respond: relatively rate-sensitive and rate-insensitive depositors, and variable-rate and fixed-rate borrowers. A macroeconomic component specifies interest-rate, yield-curve, unemployment and stress scenarios. Deterministic accounting equations update the balance sheet and calculate simplified measures of LCR, CET1, ΔEVE and ΔNII.

Each decision-making agent has two representations: an explicit equation-based specification and an LLM instructed to act within the same economic setting. We simulate comparable scenarios, estimate equations using the LLM-generated choices and assess the revised equations against additional decisions. Earlier specifications are retained for comparison. This design makes the source of each modification identifiable within the computational experiment, although it cannot establish the underlying decision process of the LLM.

The main results are as follows.

1. Under the original instructions, LLM-generated household deposit flows, rate-sensitive deposit outflows and fixed-rate loan demand respond less to price than the initially calibrated equations, by roughly four-fifths, one-third and nine-tenths (Section 6.2). Without behavioural guidance, the ordering of sensitivities and near-zero flows at market rates remain, but the fixed-rate borrower sensitivity and the withdrawal magnitudes depend substantially on the instructions (Section 6.7).
2. LLM bank choices are approximated more closely by a few conditional pricing and reinvestment rules than by the initial optimisation benchmark (Section 6.3). The final specification (EQ4) meets every fit criterion in seven of ten out-of-sample Claude bank-years, with a maximum ROE difference of 0.31 percentage points; the three exceptions involve deposit withdrawals (Section 6.4). With GPT-5.5 as the bank, the ROE difference stays within 0.71 percentage points, but only six of eighteen years meet all criteria. EQ4 meets the full criteria more often than simpler specifications estimated on the same data; none meets them in deposit-run years (Section 6.8).
3. Extending the maturity of fixed-rate assets from two to six years raises calm-year ROE by about 1.9 percentage points but reduces economic net worth after a +300-basis-point rate increase from EUR 651 million to EUR 135 million (Section 6.6).

These rules describe LLM-generated choices under particular instructions, not observed bank or customer behaviour. Within the model, competitor-rate pass-through matters more for outcomes than the estimated customer price responses, and a published LCR breach can prolong a simulated withdrawal episode, with a timing that depends on withdrawal magnitudes partly specified in the original instructions. They are conditional results that motivate empirical examination rather than conclusions about banking practice.

Section 2 reviews relevant research. Section 3 defines the model, Section 4 explains the empirical calibration and Section 5 describes the estimation and evaluation procedure. Section 6 presents the results. Section 7 examines the conditions under which the

estimated rules provide an adequate approximation, Section 8 discusses their interpretation, Section 9 addresses robustness and limitations, and Section 10 concludes.

2. Related literature

The study brings together three elements discussed in largely separate research areas: bank pricing and interest-rate-risk management, the responses of depositors and borrowers, and the calibration of agent-based models using simulated decisions. We have not identified a published study with precisely this combination and its evaluation procedure, but do not claim priority for the individual components. The following review places each element in the relevant literature.

2.1 The bank as a price-setting firm

The industrial-organisation view of banking starts from the Monti–Klein model, in which a bank with market power sets deposit and loan rates by their own demand elasticities (Klein, 1971; Monti, 1972; Freixas and Rochet, 2008). Empirical work found that deposit rates are rigid, more so in concentrated markets, and adjust upward more slowly than downward (Hannan and Berger, 1991; Neumark and Sharpe, 1992). Drechsler, Savov and Schnabl (2017) turned this into the deposits channel of monetary policy: when policy rates rise, banks with local market power widen deposit spreads and lose deposits. Egan, Hortaçsu and Matvos (2017) estimate a structural model of deposit demand with insured and uninsured depositors and show that deposit competition and fragility are linked. Repullo (2004) shows that franchise value from imperfect deposit competition disciplines risk-taking. For the euro area, Albertazzi et al. (2026) document weak and regime-dependent pass-through to sight deposits, and Koont, Santos and Zingales (2024) show that digital banking raises deposit betas and outflows.

In these models, bank pricing affects customer demand subject to the competitive environment. We extend this framework to a repeated monthly setting with two categories of depositors, two categories of borrowers and regulatory constraints. Within the specific scenarios examined below, competitor-rate pass-through has a larger influence on some simulated outcomes than modest changes in the estimated customer price-response parameters (Section 8.2).

2.2 Interest rate risk: earnings and economic value

A second literature asks why banks hold interest rate risk and how it appears in earnings and value. Drechsler, Savov and Schnabl (2021) argue that low deposit betas make deposits behave like long fixed-rate liabilities, so that maturity transformation hedges income risk; DeMarzo, Krishnamurthy and Nagel (2024) qualify this, since sticky deposits are offset by fixed loan spreads and franchise value falls when rates rise. Di Tella and Kurlat (2021) derive the exposure from optimal hedging, Jermann and Xiang (2023) make deposit maturity endogenous to depositors' beliefs, Greenwald, Schulhofer-Wohl and Younger (2023) show that deposit betas rise with the level of rates, and Abadi, Brunnermeier and Koby (2023) show that rate cuts can compress margins enough to reduce lending.

Empirically, English, Van den Heuvel and Zakrajšek (2018) find that bank equity falls after surprise rate increases, Gomez et al. (2021) show that a bank's income gap predicts its lending response to policy, and Begenau, Piazzesi and Schneider (2015) map bank positions into factor exposures. For the euro area, Hoffmann et al. (2019) find small but heterogeneous exposures shaped by national mortgage conventions; Rice and Guerrini (2025) estimate that deposit franchise value and swaps offset about half of the unrealised losses of 2022–2023; and Coulier et al. (2025) show that behavioural deposit-maturity assumptions barely changed during the tightening. The Basel IRRBB standard (BCBS, 2016) and the EBA guidelines (EBA, 2022) set the shocks, outlier tests and treatment of non-maturity deposits that our accounting model follows in simplified form. This literature supplies the exposures, behavioural assumptions and regulatory measures used here; the present application concentrates on the interaction between one bank's choices and the simulated responses of its customers.

2.3 Runs and run-prone funding

Runs are modelled as coordination failures: in Diamond and Dybvig (1983) demand deposits create a self-fulfilling run equilibrium, and Goldstein and Pauzner (2005) tie the probability of a run to fundamentals through global games. Chen et al. (2024) find that uninsured deposit flows are more performance-sensitive at banks with more liquidity transformation. After the 2023 failures, Jiang et al. (2024) show that mark-to-market losses combined with uninsured leverage can trigger solvency runs, Drechsler et al. (2026) show that uninsured depositors can run on the deposit franchise itself, and Cookson et al. (2023) find that social-media attention amplified outflows. The LCR (BCBS, 2013) fixes 30-day runoff rates by deposit category, and the separate funding analysis (Paper 2) estimates the share of euro-area funding potentially subject to withdrawal.

In our model, depositor withdrawals depend partly on stress news and the bank's previously published CET1 and LCR. This permits a conditional examination of how public disclosures interact with an assumed depositor response function (Section 7.2). The resulting threshold is a property of the model and should not be interpreted as an empirically estimated threshold for actual depositors.

2.4 Agent-based and stress-test models of banks

Agent-based computational economics models markets as interacting rule-following agents (Tesfatsion and Judd, 2006); it was proposed after 2008 as a complement to equilibrium models for financial stability (Farmer and Foley, 2009). Banking applications concentrate on contagion and the leverage cycle: interbank networks that are robust yet fragile (Gai and Kapadia, 2010), multi-layered interbank exposures (Montagna and Kok, 2016), leverage cycles from risk targeting (Aymanns and Farmer, 2015), fire-sale and funding-run dynamics among banks, dealers and funds (Bookstaber, Paddrik and Tivnan, 2018; Hałaj, 2018), and credit in macro agent-based models (Assenza, Delli Gatti and Grazzini, 2015). Central-bank top-down stress tests, such as RAMSI (Burrows et al., 2012) and STAMP€ (Dees, Henry and Martin, 2017), project detailed balance sheets with rule-based bank reactions. Englisch et al. (2023) solve bank treasury management under regulatory constraints with neural networks, taking deposits as exogenous.

Calibration and validation remain central methodological concerns (Fagiolo, Moneta and Windrum, 2007). Gouriéroux, Monfort and Renault (1993), Grazzini and Richiardi (2015) and Platt (2020) develop indirect-inference and simulated-distance estimators, and Lamperti, Roventini and Sani (2018) use machine-learning surrogates to reduce their cost; all fit aggregate moments, which need not determine individual decision rules. Our approach instead estimates equation-based approximations from decisions generated for specified agent categories. These estimates identify patterns in the generated responses, not necessarily the mechanisms that would generate human responses.

2.5 LLM agents in economics and finance

Horton, Filippas and Manning (2023) proposed using LLMs as simulated economic agents ("homo silicus") and showed that they reproduce classic behavioural experiments qualitatively. LLM agents have since been used to replicate human-subject studies (Aher, Arriaga and Kalai, 2023), to simulate survey samples (Argyle et al., 2023) and market research (Brand, Israeli and Ngwe, 2023), to automate hypothesis generation and testing (Manning, Zhu and Horton, 2024), and to populate macro simulations (Li et al., 2024). Generative agents can sustain believable behaviour over long interactions (Park et al., 2023). LLMs can also act as expectation formers: they reproduce the Survey of Professional Forecasters (Hansen et al., 2024) and survey-based return expectations (Bybee, 2023). Korinek (2023) and Gao et al. (2024) survey the uses.

In banking, Fu, Xu and Ren (2026) build FRAIL, a deterministic environment for bank runs and debt rollover with LLM agents, and find collective collapse common across seven models. Ruano and Rajan (2026) place LLM depositors on a social network calibrated to 2023 activity and reproduce a cross-bank tipping point. Kazinnik (2026) shows that bank communication changes LLM depositors' withdrawal intentions. Kazinnik and Sinclair (2025) run an LLM monetary-policy committee beside a rule-based committee and report where they diverge.

A parallel literature documents where LLM agents depart from humans. They can be more consistent with rational choice than people (Chen et al., 2023) and give "hyper-accurate" answers (Aher, Arriaga and Kalai, 2023). They compress the variance of human responses and drift with prompt wording (Bisbee et al., 2024; Park, Schoenegger and Zhu, 2024), and they reflect some groups' opinions more than others (Santurkar et al., 2023). They also fail to reproduce human strategic depth in simple games (Gao et al., 2025). Dillion et al. (2023) conclude that LLMs can complement but not replace human participants. These findings frame our limitations (Section 9.2).

2.6 Method and positioning

Two established concepts inform the design. The strategy method (Selten, 1967) elicits choices for hypothetical states rather than only the realised state; experimental comparisons find similar treatment effects in many, though not all, settings (Brandts and Charness, 2011). We use an analogous procedure to collect LLM-generated responses across economic states that may not arise in an ordinary simulation. The leader–follower formulation follows von Stackelberg (1934). Simon's (1955) account of satisficing and the

Lucas (1976) critique motivate caution when interpreting decision rules outside the conditions under which they were estimated.

Five elements of the design have precedents in this literature: separate deterministic accounting and agent decisions, as in FRAIL; deposit-flow assumptions motivated by the empirical calibration in Paper 1; explicit equation-based and LLM-based agent representations, as in Kazinnik and Sinclair (2025); an optimisation-based benchmark for bank decisions, following Englisch et al. (2023); and an economic-net-worth measure alongside regulatory EVE, following Rice and Guerrini (2025).

Our emphasis differs from work that primarily reports LLM-generated behaviour as a simulation outcome. We use the generated decisions to estimate an equation-based approximation and then evaluate that approximation in additional scenarios. Tieleman (2022) distinguishes agreement in outcomes from evidence about underlying mechanisms. Our comparisons address agreement in model-generated decisions and outcomes; they do not validate the mechanisms of human behaviour. The tests that remove behavioural guidance from the instructions (Section 6.7) further distinguish responses attributable to those instructions from responses generated without them. The contribution is therefore an explicit estimation and evaluation procedure, demonstrated in a stylised banking application.

3. The model

The model is a repeated monthly leader–follower interaction over twelve periods. One bank chooses its actions before four customer categories, while a macroeconomic component specifies external conditions. Deterministic accounting equations implement the resulting transactions and update the balance sheet. Customer responses depend on linear functions of specified price or state differences, subject to bounds. Fixed-rate positions are represented by subportfolios with assigned durations. Interest rates are expressed as annual percentages with monthly accrual; monetary amounts are in EUR millions.

3.1 Agents, balance sheet and timing

The bank's assets comprise cash C, four bond subportfolios distinguished by interest-rate type and accounting classification (fixed-rate and floating-rate securities in either the amortised-cost investment portfolio or the fair-value portfolio: FI, FLI, FM and FLM), variable-rate loans L_s and fixed-rate loans L_i. Funding consists of relatively rate-sensitive deposits D_s, less rate-sensitive deposits D_i, wholesale funding W and CET1 equity E. The initial balance sheet totals EUR 10 billion (Table 1).

Table 1 Opening balance sheet (EUR million)

Assets	Amount	Liabilities and equity	Amount
Cash C	800	Sensitive deposits D_s	3,800
Fixed investment bonds FI	800	Insensitive deposits D_i	5,000
Floating investment bonds FLI	400	Wholesale funding W	400
Fixed market bonds FM	800	CET1 equity E	800

Assets	Amount	Liabilities and equity	Amount
Floating market bonds FLM	400		
Variable-rate loans L_s	3,000		
Fixed-rate loans L_i	3,800		
Total	**10,000**	**Total**	**10,000**

Each period t consists of four sequential stages (Fig. 1):

1. The macroeconomic component specifies the policy-rate change Δr, yield-curve slope change Δκ, unemployment change Δu and stress-news indicator n ∈ {0, 1, 2}.
2. **The bank observes the state and chooses x_t = (d_s, d_i, m_s, m_i, F, M):** two deposit rates, two new-loan margins, the fixed-rate proportion F and fair-value portfolio proportion M of bond reinvestment.
3. Each customer category observes its relevant bank price, the alternative market price and a public stress indicator, and the four categories respond simultaneously.
4. The accounting equations update interest, contractual repayments, new business, deposit flows, profit, securities valuations and funding, after which LCR and CET1 become available for the following period.

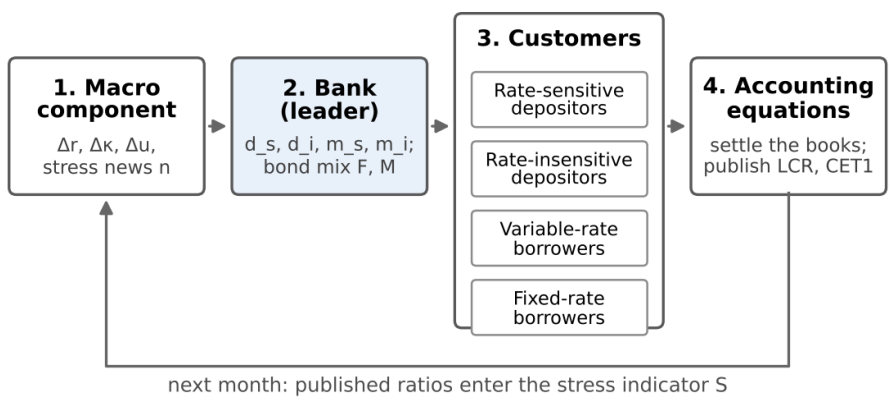

Fig. 1 The monthly stage game: order of moves within each period

The bank acts before the customers. In the optimisation-based specifications (EQ0–EQ2), its decisions are computed using the specified customer-response functions. Reported prudential ratios affect the stress indicator in the following month.

3.2 Market environment

The yield curve is approximated by a straight line between the policy rate and the ten-year yield. Alternative deposit rates follow the state-dependent pass-through coefficients estimated in Paper 1, using household deposits for type i and non-financial corporate deposits for type s; these rates are bounded below by zero:

$$y_t(D) = r_t + \kappa_t \frac{\min(D, 10)}{10}, \qquad a_{j,t} = \max\{0,\ a_{j,t-1} + \left(\alpha_j + \gamma_j r_{t-1}\right) \Delta r_t\}$$

Depositors observe a stress indicator constructed from news and the most recently published prudential ratios:

$$S_t = n_t + w_C\,\mathbf{1}\{CET1_{t-1}^{pub} < 14\%\} + w_L\,\mathbf{1}\{LCR_{t-1}^{pub} < 100\%\}$$

3.3 Depositors

For depositor category j ∈ {s, i}, f_j denotes the fraction of its deposit balance transferred in one month. The specified decision problem compares the incremental yield on the transferred balance with a quadratic switching cost. Its solution gives a linear price response with distinct inflow and outflow coefficients, an inflow bound and a separate response to stress:

$$f_j = \begin{cases} -R_j(S_t) & \text{if } R_j(S_t) > 0 \\ \min\{\varphi_j^{\pm}\left(d_j - a_j\right),\ \bar{f}_j\} & \text{otherwise} \end{cases}, \qquad \varphi_j^{\pm} = \begin{cases} \varphi_j^{+} & d_j > a_j \\ \varphi_j & d_j \le a_j \end{cases}$$

At the alternative market rate, the price-induced net flow is f = 0. This imposed initial condition reflects the calibration assumption motivated by Paper 1; it is not a separately estimated behavioural result. Monthly inflows are limited to f̄, whereas withdrawals respond linearly until the assumed stress-withdrawal limit is reached. R_j(S) specifies withdrawals as a function of the stress indicator; the upper limits g_j use the LCR runoff weights of 40% for type s and 7% for type i.

3.4 Borrowers

For borrower category j, n_j is the monthly flow of new loans as a percentage of the outstanding portfolio, q_j is the prepayment proportion and PD_j is the annual default probability:

$$n_j = \max\{0,\ g_j^0 - \chi_j \Delta u_t + \eta_j\left(\bar{m}_j - m_j\right)\}, \qquad g_j^0 = 100\rho_j + q_j^0 + PD_j^0/12$$

$$q_s = q_s^0 + \zeta_s \max\{0,\ \mu_s - \bar{m}_s\}, \qquad q_i = q_i^0, \qquad PD_j = PD_j^0 + \psi\,(u_t - u_0)$$

Baseline new lending g^0 offsets scheduled repayments, prepayments and defaults, making the outstanding loan portfolio constant in the calm reference state. Variable-rate borrowers pay r + μ_s, where the outstanding portfolio margin μ_s is a weighted average of margins on earlier and new lending. New fixed-rate loans carry the rate y(D_Li) + m_i for their remaining contractual life.

3.5 The bank

In the optimisation-based specification, the bank chooses its actions to maximise projected earnings and a terminal portfolio-value term, subject to penalties for departures from internal risk targets. The forecast uses a horizon of H = 12 months, holds the current choice x and macroeconomic state fixed, and assumes that the initial stress-news effect subsequently subsides:

$$J(x) = \frac{\sum_{h=1}^{H} NI_h + FV_H}{E_t} - \mu \sum_{g} \max_{h} \max\{0, gap_{g,h}\} - \kappa_{dep} \sum_{j} \left(d_j - a_j\right)^2 - \kappa_{loan} \sum_{j} \left(m_j - \bar{m}_j\right)^2 + \lambda_{LCR} \left([\![LCR]\!]_H - 1.30 \right)$$

$$FV = \frac{1}{100}\left[(r - d_s)D_s\ell_{Ds} + (r - d_i)D_iD_{core} + \mu_s L_s \ell_{Ls} + \left(c_{Li} - y(D_{Li})\right) L_i D_{Li}\right]$$

The penalty terms measure deviations from five internal targets: LCR of at least 130%, CET1 of at least 14%, ΔEVE/T1 no lower than −10%, an absolute ΔNII/T1 no greater than 3%, and W/A no greater than 10%. With μ = 10, a deviation of one percentage point produces a penalty equivalent to ten percentage points of ROE. These weights give risk targets considerable influence over the computed choices. Removing the terminal portfolio-value term produces solutions at the boundaries of the admissible choice set. The coefficients κ_dep and κ_loan and the liquidity preference λ_LCR are zero in EQ0 and estimated subsequently. Separate supervisory thresholds are set at LCR 100%, CET1 11% and ΔEVE/T1 −15%. A breach is recorded, and the simulation ends if CET1 falls below 11%.

The optimisation-based bank evaluates each rate within ±50 basis points of its market reference, in increments of five basis points; F ∈ {0, ⅓, ⅔, 1} and M ∈ {0, ½, 1}. A coordinate-search algorithm makes two successive passes through the choices. From EQ2 onwards, estimated conditional decision rules replace parts of the optimisation procedure (Section 6.3).

Equilibrium interpretation. In each period, the optimisation-based bank solves an approximate leader problem given the specified, contemporaneous customer responses. For forecasting it holds its current choices and the macroeconomic state fixed for twelve months, without anticipating its own future revisions. Consequently, the simulation comprises a sequence of approximate one-period Stackelberg solutions rather than a fully dynamic Stackelberg equilibrium. EQ3 and EQ4 generally use estimated decision rules, returning to the optimisation procedure only when the capital target is breached without a liquidity shortfall. They should not therefore be described as solutions of the original Stackelberg optimisation problem.

Numerical approximation. The optimisation-based specifications use two coordinate-search passes on a 0.05-percentage-point pricing grid and examine twelve possible bond allocations. We compare these results with twenty searches from alternative starting points, 3,000 randomly selected points in four states per scenario and a 0.10-percentage-point exhaustive grid in one state per scenario (runs/robust/R7_optimality.csv). For EQ2, whose bond allocation follows a specified rule, the largest observed difference from the best objective value in the comparison is 0.27 percentage points of equity, and the median is zero. For EQ0, which also optimises bond allocation, the median difference is 0.02 percentage points and the maximum is 1.6 percentage points in falling-rate scenarios with relatively flat objective values. Repeating EQ0 with multiple starting points changes ROE by

no more than 0.2 percentage points and minimum LCR by no more than three percentage points. The deposit volumes are more sensitive: in falling-rate and recession scenarios, rate-sensitive deposits decline by 4% and 2%, rather than 11% and 14%, respectively (runs/robust/R7b_optimality_outcomes.csv). EQ0 deposit-volume outcomes in these scenarios should therefore be treated as approximate. The rule-based EQ3 and EQ4 are unaffected by this numerical search. References to Stackelberg solutions apply only to the locally approximated optimisation-based variants.

3.6 Accounting and risk measures

Interest accrues on opening balances. The cost of wholesale funding increases with its outstanding amount and with weaker capitalisation:

$$r_t^W = r_t + w_0 + w_1 \max\{0,\ W/A - 0.04\} + w_2 \max\{0,\ 100\,(0.14 - CET1^{pub})\}$$

Each fixed-rate subportfolio is represented by a uniform maturity distribution with maximum remaining maturity 2D. The resulting monthly principal repayment proportion is 1/(24D); floating-rate securities and variable-rate loans have repayment proportions of 1/(12 × 3 years). All principal repayments from bonds are reinvested, keeping the total securities portfolio constant. The associated decision variables therefore affect interest-rate composition and accounting classification, not the aggregate portfolio amount. Effective portfolio rates are weighted averages of existing and new positions. Fixed-rate securities measured at fair value are recorded at par plus FVA = −D_FM (y(D_FM) − c_FM) B_FM / 100, with the monthly change recognised in other comprehensive income (OCI). Net income is NI = II − IE − CL − opex − tax, and equity changes by NI + OCI. Cash is determined residually subject to a lower bound of 1% of deposits; wholesale funding meets a shortfall, and excess cash reduces wholesale funding to no less than its initial level. The balance-sheet identity A = D_s + D_i + W + E is checked each month with a tolerance of 10^{-6}.

The risk measures approximate the regulatory sensitivities and liquidity quantities used in the two calibration analyses, but are not full regulatory calculations. ΔEVE uses an asset-liability duration approximation based on carrying amounts. The simplified LCR treats all bonds as high-quality liquid assets without valuation haircuts, excludes contractual inflows and assumes a 100% outflow for wholesale funding W. Risk-weighted assets comprise loans only. For a parallel interest-rate shock Δ = +2 percentage points applied to a constant balance sheet:

$$\Delta EVE = -\frac{\Delta}{100}\left(\sum_{k\in A} D_k\, X_k - \sum_{k\in L} D_k\, X_k\right), \qquad \Delta NII = \frac{\Delta}{100}\left(\sum_{k\in A} w_k\, X_k - \beta_s D_s - \beta_i D_i - W\right)$$

$$LCR = \frac{C + \sum_k B_k + FVA}{0.40\, D_s + 0.07\, D_i + W}, \qquad CET1\ ratio = \frac{E}{0.75\,(L_s + L_i)}$$

The repricing weight w equals one for cash, floating-rate securities and variable-rate loans, and 1/(4D) for a fixed-rate subportfolio. Economic net worth adds to accounting equity the estimated unrealised valuation gains or losses on amortised-cost fixed-rate securities and loans.

4. Calibration to selected empirical moments

The initial specification (EQ0) represents a stylised, predominantly deposit-funded euro-area bank that matches three summary statistics from the two empirical calibration analyses: ΔEVE/T1 and ΔNII/T1 from Paper 1 and runoff intensity from Paper 2. At inception, simulated ΔEVE/T1 is −6.10% (reference median −6.1%), ΔNII/T1 is +0.97% (reference median +0.7%, within the reported range up to +1.0%) and runoff intensity is 21.2% (sample mean 20.5%); the bank has a CET1 ratio of 15.7%, leverage of 8.0%, an LCR of 141% and a calm-year ROE of 12.9%. Agreement with these moments is a calibration result, not validation, and it does not make the bank representative of the median euro-area institution (Section 7.1).

The model uses three categories of input parameters, defined as follows (Table 2).

1. **Empirical calibration inputs.** Deposit-rate pass-through coefficients come from the empirical panel analysis labelled Paper 1; their numerical values are reported in Table 2. The assumed LCR runoff weights are 40% and 7%, compared with rates of 49% and 6.9% inferred from Paper 2. The behavioural duration of less rate-sensitive deposits, D_core = 3.35 years, is selected to match ΔEVE and remains within the EBA five-year cap. The twelve-month repricing coefficients (0.75 and 0.27) are selected to match the ΔNII target.
2. **Maturity assumptions.** Fixed-rate investment securities have an assigned duration of four years, fixed-rate fair-value securities of two years and fixed-rate loans of four years. Section 6.6 varies these assumptions.
3. **Initial equilibrium condition.** Customer price-response parameters are selected numerically so that the bank's computed optimal prices equal the alternative market prices under calm initial conditions. The reference state is therefore stationary by construction; changes in subsequent simulated outcomes arise from scenarios and the specified response functions, not from an independent estimate of customer behaviour.

Table 2 EQ0 parameters, values and sources

Group	Parameter	EQ0 value	Rule or source
Market	r_0, κ_0; a_s0, a_i0	2.0%, 1.0%; 1.0%, 0.3%	Scenario start
Market	α_s, γ_s / α_i, γ_i	0.238, 0.069 / 0.092, 0.035	ECB deposit-rate panel estimates (Paper 1)
Market	m̄_s, m̄_i	2.0%, 1.5%	Market new-loan margins
Durations	D_FI, D_FM, D_Li; D_core, D_s	4, 2, 4; 3.35, 0.25 years	Run inputs; fitted to ΔEVE
Depositors	φ_s, φ_i	18, 6	Calibration condition
Depositors	f̄_s, f̄_i; g_s, g_i	2%, 0.5%; 40%, 7%	Inflow caps; LCR weights
Depositors	withdrawal threshold S	2 (s), 3 (i)	All-or-nothing run at threshold
Borrowers	η_s, η_i	1.3, 0.8	Calibration condition
Borrowers	q^0_s, q^0_i; ζ_s; χ_i	0.3%, 0.2%; 1; 0.2	Baseline prepayment; refinancing; income effect
Credit	PD^0_s, PD^0_i; ψ; LGD	1.5%, 0.8%; 0.3; 40%	Retail and SME ranges
Bank	opex, tax	1.3% of assets, 10%	A low euro-area statutory rate; tax scales profit, not the

Group	Parameter	EQ0 value	Rule or source
			comparisons
Bank	w_0, w_1, w_2	0.5%, 5, 0.5	Solvency-linked wholesale cost
Bank	μ, H	10, 12 months	Penalty weight; horizon

Independent comparison with bank-level observations. Matching three aggregate statistics does not uniquely determine deposit behaviour, asset duration, hedging or balance-sheet composition. We therefore compare the resulting risk measures with bank-level outcomes not used as calibration targets, following a specification recorded in advance (runs/paper1/SPEC.md; Online Resource 2 contains the comparison script and processed empirical inputs). The comparison covers 53 banks for June 2025 and 57 for December 2025 with both EBA Transparency balance sheets and Pillar 3 IRRBB disclosures. Their observed shares of household and corporate deposits, other funding, Tier 1 capital, loans and securities replace the initial balance-sheet proportions of the stylised bank. All behavioural and duration parameters remain fixed at their EQ4 values; no bank-level refitting is performed. Table 3 reports the simulated and disclosed sensitivities under a +200-basis-point shock, grouped by customer-deposit share.

Table 3 Simulated and disclosed +200bp sensitivities by customer-deposit share, June and December 2025

Banks by customer-deposit share (June 2025)	Banks	Deposits / assets (median)	ΔEVE/T1 disclosed	ΔEVE/T1 model	ΔNII/T1 disclosed	ΔNII/T1 model
Lowest third	18	27%	−3.2%	−33.1%	0.7%	−6.1%
Middle third	17	52%	−8.6%	−13.5%	0.4%	−1.3%
Highest third	18	70%	−7.5%	0.1%	1.6%	1.6%
All 53: rank correlation, model vs disclosed			−0.28 (Dec 2025: −0.01)		+0.17 (Dec 2025: +0.24)	
Calibrated base bank		88%	−6.1% (median target)	−6.1%	0.7% (median target)	1.0%

The comparison yields three findings.

1. **Comparison of ΔNII for deposit-funded banks.** Within the third of banks with the highest customer-deposit shares, the model's median ΔNII/T1 is close to the disclosed median without refitting: 1.6% versus 1.6% in June 2025, and 1.2% versus 1.3% in December 2025. The correspondence does not extend uniformly to the other balance-sheet groups.
2. **Limited correspondence for ΔEVE.** Disclosed ΔEVE/T1 ranges from about −3% to −9% across the three funding groups, while the simulated values range from about −33% to zero. The rank correlations between model and disclosed sensitivities are −0.28 in June and −0.01 in December, the model's interquartile range is four times the observed range, and the simulated signs of ΔNII and ΔEVE agree with the disclosed combination in 17% of June and 25% of December observations. Asset duration, derivatives and hedging evidently vary among banks in ways that broad balance-sheet shares do not reveal, so the initial moment matching does not identify effective duration; we therefore treat duration as a model input and decision variable rather than infer it from funding composition. In the disclosure sample, 82% of 205 observations have ΔEVE/T1 above −10% and only one is below −15%

(median −6.0%), which is consistent with management of exposures against limits but does not establish clustering around internal limits, which are not observed.

3. **Applicability to predominantly deposit-funded banks.** Customer deposits finance 88% of assets in the stylised bank, against a 52% median in the comparison sample, and the simplified funding equations treat all non-deposit liabilities as short-term and fully subject to withdrawal, which misrepresents covered bonds and senior debt. Applied to the median balance sheet of the least deposit-funded third, the model gives an opening LCR of 56% and a calm-year ROE of −24%; for the middle third 66% and −11%; for the most deposit-funded third 126% and 5.9% (runs/paper1/profiles.csv). The first two reflect the funding assumption rather than predictions. Substantive interpretation is therefore restricted to banks funded predominantly by customer deposits, roughly two-thirds or more of total assets (Section 7.1).

5. Estimation and evaluation of equation-based decision rules

The equation-based specification is used for the main simulations because its output is reproducible for a given input. Decisions generated by LLM agents are used to estimate or revise the specified response functions. Each estimation round produces a revised equation-based model, denoted EQ_k+1, while earlier versions are retained for comparison (Fig. 2). This procedure assesses how well simple equations approximate LLM-generated choices rather than treating those choices as observations of human behaviour.

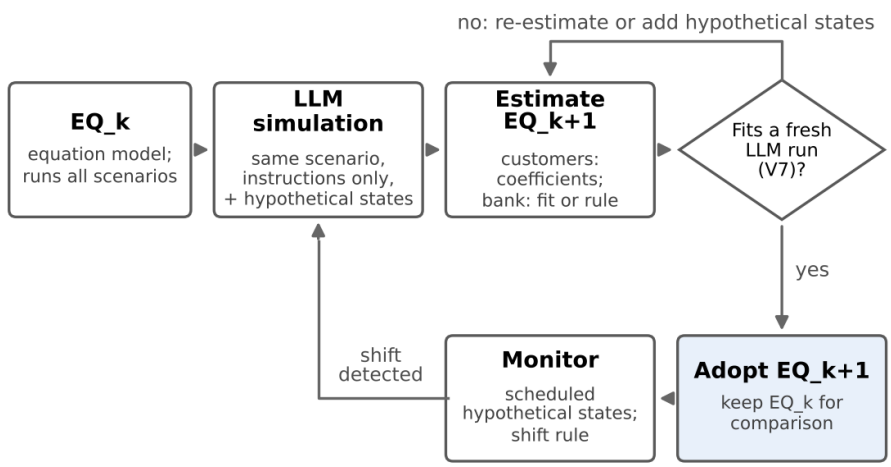

Fig. 2 Estimation and evaluation procedure

A revised specification is retained for further analysis only after comparison with an additional LLM simulation that was not used to estimate it. A failure to meet the stated fit criteria leads to reconsideration of the specification, as occurred for EQ2 following the rising-rate scenario.

5.1 LLM agents

The initial full simulations use five persistent Claude agents, each retaining the information supplied during the preceding twelve months; the cross-family assessment (Section 6.4) substitutes an OpenAI model for the bank, and Table 4 identifies the model behind each data set. Agents receive plain-language instructions describing their circumstances, objectives, available decisions, bounds and response format (one JSON record and up to sixty words of explanation), but not the model equations or coefficients. The original instructions do contain behavioural guidance (deposit-flow limits, an assumed default response and an ordering of bank objectives), whose influence Section 6.7 examines. Each month an LLM agent receives the information available to its equation-based counterpart and the preceding month's outcomes, and its answer is clipped to the same admissible range. Agents ran at default sampling settings without controlled seeds, so each simulation is one realisation; all raw responses are saved.

Table 4 LLM data sets: agent roles, models and dates

Data set	Agents	Model	Date
Full LLM runs pilot1 (calm) and hike1 (hike)	Bank	claude-opus-5-5	15 Oct 2026
Full LLM runs pilot1 and hike1	Four customer categories	claude-sonnet-5-5	15 Oct 2026
Bank questionnaire (bond rule)	pilot1 bank after 12 months	claude-opus-5-5	15 Oct 2026
Bank-only runs calm3, stress4, hikestress5, calm6–calm8	Bank (customers are equations)	claude-opus-5-5	15–16 Oct 2026
Customer questionnaire replications 1–4 (pilot1 and hike1 customer agents after 12 months; two fresh agents)	Customers	claude-sonnet-5-5	15–16 Oct 2026
Model-tier questionnaires (Section 9.1)	Fresh customer agents, one answer each	claude-opus-5-5, claude-sonnet-5-5, claude-haiku-5-5	16 Oct 2026
Evaluation set eval_calm, eval_hike, eval_cut, eval_stress, eval_recession, eval_hikestress (reworded brief)	Bank (customers are equations)	claude-opus-5-5	16 Oct 2026
Cross-family test: 18 bank-only years (3 per scenario, original brief) and the four fresh hypothetical-state questionnaires	Bank (customers are equations); questionnaires: fresh customer agents	gpt-5.5-2026-04-23 (OpenAI)	16–17 Oct 2026
Brief ablation (Section 6.7): four customer questionnaires and an eight-state bank hypothetical-state questionnaire, original and neutral instructions, 6 fresh agents per brief and battery (3 on each of two days)	Fresh customer agents; fresh bank	claude-sonnet-5-5 (customers), claude-opus-5-5 (bank)	17–18 Oct 2026
Brief ablation, second family: same batteries and bank hypothetical-state questionnaire, 5 per brief and battery (2 and 3 on two days)	Fresh customer agents; fresh bank	gpt-5.5-2026-04-23 (OpenAI)	17–18 Oct 2026

Claude agents were executed using Anthropic agent tools and the GPT bank through OpenAI's Responses API, at default sampling settings in each case; the replication package includes the instructions, questions and unedited responses (Appendix A2).

5.2 Elicitation of decisions in hypothetical states

A twelve-month simulation exposes an agent only to states reached along that particular trajectory. We therefore also request decisions for a predetermined collection of hypothetical states. This is analogous to the strategy method of experimental economics (Selten, 1967), although the LLM responses are not incentivised choices and should not be equated with human experimental data. Each collection of customer scenarios was completed four times per category: by agents that had participated in the initial calm and rising-rate simulations and by two newly initialised agents. Depositors receive combinations of bank and market rates and variations in stress news, CET1 and LCR. Borrowers receive alternative loan margins, two outstanding-portfolio margin settings and two unemployment shocks. Bank agents assess seven internally consistent states computed from the accounting model: an initial state, a +100-basis-point rate change, 15% deposit withdrawals, ΔEVE near the specified limit, CET1 near target, LCR near target and a steeper yield curve. Each set is elicited in one agent interaction.

5.3 Estimation

Customer response coefficients are estimated by least squares with the specified zero-flow condition imposed. Separate coefficients describe deposits entering and leaving in response to the interest-rate difference (φ and φ^+), demand responses to loan margins (η), refinancing (ζ) and unemployment (χ). The stress-dependent withdrawal function R_j(S) is estimated using the mean response within each stress category. Because each category has only four observations, these values should be interpreted as descriptive estimates, not precise population parameters. The estimated function for less rate-sensitive deposits is not monotonic. The answers further indicate that a breach of the LCR threshold is treated comparably to severe stress news, motivating w_L = 2.

Bank-decision parameters are selected to minimise differences between the equation-based decisions and those generated by the LLM bank. A parameter grid covers the liquidity preference λ_LCR and the pricing coefficients κ_dep and κ_loan. The objective is evaluated across both the hypothetical decision states and simulated monthly choices:

$$G = \frac{1}{T}\sum_{t}\left(\sum_{k\in\{d_s,d_i,m_s,m_i\}} \left|x_{k,t}^{EQ} - x_{k,t}^{LLM}\right| \; + \; 0.5\left(\left|F_t^{EQ} - F_t^{LLM}\right| + \left|M_t^{EQ} - M_t^{LLM}\right|\right)\right)$$

The decision-discrepancy statistic G sums the mean absolute differences for four pricing decisions and half the corresponding differences in two portfolio-composition decisions. The acceptance threshold therefore applies to the combined statistic, not separately to each decision.

Where the tested parameter values in the optimisation-based formulation do not reproduce the observed LLM choices sufficiently closely, we estimate a conditional decision rule directly from those choices. This occurs for bond reinvestment composition, prices in ordinary conditions and prices during liquidity stress.

5.4 Evaluation criteria and specification revision

The study uses seven prespecified assessment categories (Table 5). Criteria V1–V6 examine internal consistency, calibration and expected directional responses of the equation-based model; V7 measures agreement with an additional LLM simulation under the same scenario. Satisfying these criteria establishes performance under the stated computational tests, not external validity with respect to human decisions.

Table 5 Evaluation criteria V1–V7

Check	Pass condition
V1 Accounting identity	Assets = liabilities + equity in every logged month, recomputed from the log
V2 Static calibration	Opening ΔEVE/T1 ≈ −6.1%, ΔNII/T1 +0.7% to +1.0%, runoff intensity 20–22%
V3 Calm stationarity	Every customer book within ±1% after a calm year
V4 Calm equilibrium	In every calm month the bank prices within 5bp of its normal-times policy
V5 Depositor anchor	Bank within 2.5bp of market implies $\|f\| \leq 0.5\%$ a month
V6 Economic signs	Hike: NII up, value loss, bank betas within 0.1 of market. Cut: ROE down. Stress: run concentrated in sensitive deposits, LCR ≥ 100%
V7 LLM fit	ROE within 1pp, year-end balances within 3%, decision discrepancy G ≤ 0.5pp

Two of the criteria were modified during development. V4 initially required market pricing. Because later rule-based specifications explicitly impose pricing near market, V4 was reformulated to check whether the normal-time decision rule was applied consistently; it consequently has limited independent diagnostic content for those models. V6 originally excluded household withdrawals under stress but was changed after estimated monthly household withdrawals of 0.6% were observed; it now requires them to be less than one-tenth of the withdrawals by rate-sensitive depositors. V7 was extended to include the discrepancy between decisions because similar aggregate balances need not imply similar choices (Section 6.4). These changes should be considered when interpreting overall pass counts.

The specification is reconsidered when additional hypothetical-state responses imply a change of more than 25% in a price-response coefficient or more than five percentage points in a stress-withdrawal entry, or when a separate LLM simulation does not meet criterion V7.

After the customer equations met the separate rising-rate test, subsequent evaluations of bank decisions used those fixed equation-based customer responses. This reduced the computational burden from 60 to 12 LLM interactions per bank-year, but also means that these later evaluations primarily test the bank decision rule rather than the entire five-agent system. Predictions for the separate evaluations were saved before the relevant LLM simulation began, and file hashes were recorded from the evaluation set onwards. We describe these as saved or frozen predictions, not externally preregistered ones: their timing is verifiable from the supplied version history rather than an independent registry.

6. Results

Successive estimation rounds produced specifications EQ0–EQ4. Customer-response coefficients were estimated once and subsequently assessed in a separate rising-rate simulation. The bank preference coefficients were then estimated, after which three conditional decision rules replaced parts of the optimisation procedure. The development data comprise eight LLM bank-year simulations: two with LLMs representing all five decision-makers and six using equation-based customers. The latter include four calm years, one stress year and one rising-rate year combined with a withdrawal episode. Each customer decision questionnaire was repeated four times. These are model-generated observations rather than human or bank transaction data.

6.1 The equation baseline (EQ0)

EQ0 meets the internal consistency and directional criteria V1–V6 in the five tested scenarios. A +300-basis-point rate increase changes simulated annual NII from EUR 276 million to EUR 309 million and ROE from 12.9% to 16.0%, with estimated unrealised fixed-asset losses of EUR 569 million. A −180-basis-point reduction lowers ROE to 10.7%. Severe stress news in month four generates withdrawals equal to 40% of rate-sensitive deposits; the LCR declines to 105% before recovering. Under a rising-rate scenario, however, the initially calibrated price responses induce substantial competition for deposits: rate-sensitive deposit balances increase by 24% over the year. This result is a consequence of the initial calibration condition and motivates estimation of alternative response coefficients.

6.2 Customer responses under the original instructions

Responses elicited under the original instructions indicate lower price sensitivity among depositors and fixed-rate borrowers than in EQ0, together with greater sensitivity to specified stress signals. Variable-rate borrowers display somewhat higher price sensitivity than assumed initially. These comparisons are conditional on the instructions: some explicitly suggest low sensitivities or impose withdrawal bounds. Section 6.7 re-estimates responses without that guidance and finds that both withdrawal magnitudes and the fixed-rate borrower coefficient change substantially. Table 6 reports the EQ1 coefficients subsequently retained through EQ4.

Table 6 Customer-response coefficients: EQ0 calibration, EQ1 estimates and replication means

Parameter	EQ0 (calibrated)	EQ1–EQ4 (pilot answers)	Mean (s.d.) over 4 replications
Sensitive depositors: outflow / inflow slope φ_s, φ_s^+	18 / 18	12 / 6.3	13.2 (1.9) / 6.5 (0.2)
Insensitive depositors: outflow / inflow slope φ_i, φ_i^+	6 / 6	1.3 / 1.0	1.1 (0.2) / 1.0 (0.1)
Variable-rate borrowers: price elasticity η_s	1.3	1.6	2.3 (0.5)
Fixed-rate borrowers: price elasticity η_i	0.8	0.09	0.14 (0.05)
Refinancing response ζ_s	1.0	1.4	1.6 (0.7)
Income effect χ_s / χ_i	0 / 0.2	0.41 / 0.01	0.51 (0.12) / 0.01 (0.01)

Parameter	EQ0 (calibrated)	EQ1–EQ4 (pilot answers)	Mean (s.d.) over 4 replications
Sensitive withdrawal share at stress index 1 / ≥2	0% / 40%	6.5% / 32.5%	8.4% (3.8) / 32.5% (2.0)
Insensitive withdrawal share at stress index 1 / 2 / 3 / 4	0 / 0 / 7 / 7%	0.2 / 0.6 / 0.3 / 7%	0.15 / 1.3 / 0.26 / 6.8%
Weight of an LCR breach in the stress index	1	2	2

Repeated elicitation generally preserves the direction of the differences from EQ0, but the variability differs across parameters. The reported coefficients of variation for deposit-price responses range from 3% to 16%. Loan-demand responses are more variable; the mean replicated estimates of η_s and η_i exceed the pilot estimates by 44% and 67%, respectively. Substituting these means for the pilot coefficients (EQ5) changes reported ROE by no more than 0.2 percentage points and minimum LCR by no more than 1.1 percentage points (Online Resource 1, Section S3). This limited effect reflects the tendency of the simulated bank to price near market, where price-response slopes have relatively little influence. Under the original instructions, mild stress news produces withdrawals of 6.5–14% among rate-sensitive depositors. The estimated fixed-rate loan-demand response is small, but Section 6.7 shows that this feature is not stable when the instructions change.

A separate rising-rate simulation provides an initial test of the revised customer equations. Given the LLM bank's observed choices, the EQ1 customer equations reproduce annual ROE with a difference of 0.03 percentage points and end-of-year balances within 2.3%. The remaining lending difference is associated with an LLM-generated decline in variable-rate loan demand of approximately 0.05 percentage points per month for every 25-basis-point increase in the policy rate, even where the margin remains unchanged. This test concerns the response equations for the stated LLM setting, not customer behaviour in observed data.

6.3 Conditional decision rules estimated for the LLM banks

The eight development simulations display several recurring choices, alongside variation between individual LLM bank realisations. Table 7 distinguishes months without material liquidity pressure from months with stress news or a published LCR at or below the internal 130% target.

Table 7 LLM bank decisions by regime in the eight development simulations

Bank (scenario)	Regime	Months	Sensitive deposit premium	Household deposit premium	Variable-loan margin vs market	Fixed-loan margin vs market	Mean fixed share F
pilot1 (calm)	Normal	12	+10bp	0	0	+28bp	0.67
hike1 (hike)	Normal	12	+1bp	0	−4bp	+22bp	0.58
calm3 (calm)	Normal	12	0	0	+1bp	+39bp	0.83
calm6 (calm)	Normal	12	0	0	+3bp	+41bp	0.72
calm7 (calm)	Normal	12	+1bp	0	+1bp	+43bp	0.86
calm8 (calm)	Normal	12	0	−1bp	+1bp	+38bp	0.89
stress4 (stress)	Normal	5	+8bp	+6bp	+2bp	+27bp	0.67
stress4 (stress)	Liquidity stress	7	+46bp	+26bp	+21bp	+46bp	0.33
hikestress5	Normal	5	+5bp	0	0	+2bp	0.33

Bank (scenario)	Regime	Months	Sensitive deposit premium	Household deposit premium	Variable-loan margin vs market	Fixed-loan margin vs market	Mean fixed share F
(hike + run)							
hikestress5 (hike + run)	Liquidity stress	7	+45bp	+22bp	+23bp	+24bp	0

The normal-period observations in stress4 include two months during which elevated deposit premiums were gradually reduced after the LCR recovered. This explains the mean premiums of eight and six basis points in the corresponding row; before withdrawals began, the bank priced deposits at market. Three recurring patterns are represented by conditional decision rules (Appendix A1).

1. **Bond reinvestment composition (EQ2).** The estimated rule allocates approximately two-thirds of reinvestment to fixed-rate securities in the amortised-cost portfolio, increases the fixed-rate proportion when there is capacity within the EVE target or the yield curve is steep, and reduces it as ΔEVE/T1 approaches −9%. It is based on the initial bank's hypothetical-state responses. Combined with a lower loan-pricing coefficient (κ_loan reduced from 1 to 0.5), it reduces that bank's decision discrepancy from 0.54 to 0.38 percentage points.
2. **Pricing in ordinary conditions (EQ3).** The approximation sets deposit rates and variable-rate lending margins near their corresponding market values and applies a 30-basis-point addition to the fixed-rate loan margin. In seven of eight development simulations the average fixed-rate addition is 22–43 basis points; in five it approaches the admissible 50-basis-point maximum by year end. This is a descriptive rule for the simulated banks, not an estimate of actual bank pricing conventions.
3. **Pricing and reinvestment during liquidity stress (EQ4).** The approximation raises deposit rates by 50 basis points for rate-sensitive deposits and 30 basis points for less rate-sensitive deposits, increases new-loan margins by approximately 30–50 basis points and directs most reinvestment to floating-rate securities in the amortised-cost portfolio. Both stress simulations show broadly similar choices, and the initial bank's hypothetical-state responses imply the 50-basis-point sensitive-deposit premium. These observations differ from the earlier EQ2 bond-composition rule, which allocated more investment to the fair-value portfolio under stress.

The main variation in ordinary conditions concerns a persistent deposit premium: six banks pay approximately zero to one basis point, one pays five basis points and one ten basis points. The fixed-rate reinvestment proportion ranges from about one-third to nine-tenths. No single tested preference coefficient adequately approximates both the premium-paying and market-pricing realisations. EQ4 therefore uses market pricing as its reference rule and evaluates the alternatives in sensitivity analysis.

6.4 Out-of-sample evaluation

These evaluations use the fixed customer equations estimated under the original instructions. They therefore test the bank approximation conditional on those customer equations, rather than independently evaluating all five agents. Section 6.7 shows that the

associated withdrawal magnitudes are partly instruction-dependent. Each model is compared with the available bank simulations using three statistics: the annual ROE difference, the largest year-end balance difference and the combined decision discrepancy G (Table 8). Bold entries meet all elements of V7 (ROE difference no greater than one percentage point, balance difference no greater than 3%, and G no greater than 0.5 percentage points). A dagger identifies observations used in the estimation or specification of the respective model.

Table 8 Fit of EQ1–EQ4 to the LLM bank simulations (ROE gap / largest balance gap / decision discrepancy G)

Model	pilot1 (calm)	hike1 (hike)	calm3 (calm)	stress4 (stress)	hikestress5 (hike + run)
EQ0 calibrated	+0.48 / 9.1% / 0.54	−0.40 / 22.4% / 0.66	−0.06 / 0.4% / 0.63	+1.31 / 12.7% / 0.98	+0.95 / 5.0% / 1.07
EQ1 + estimated customer equations	+0.24 / 2.0% / 0.54 †	−0.41 / 6.5% / 0.64	−0.29 / 7.8% / 0.81	+1.11 / 1.8% / 1.00	+0.61 / 1.0% / 0.85
EQ2 + bond rule	**+0.25 / 2.1% / 0.38** †	−0.33 / 6.5% / 0.59	−0.29 / 7.8% / 0.84	+1.21 / 2.8% / 1.15	+0.90 / 2.3% / 0.87
EQ3 + normal-times pricing	+0.52 / 9.1% / 0.13 †	**+0.03 / 1.4% / 0.31** †	**−0.01 / 0.2% / 0.46** †	+1.22 / 3.4% / 1.20	+0.84 / 3.1% / 0.88
EQ4 + liquidity repair	+0.52 / 9.1% / 0.13 †	**+0.03 / 1.4% / 0.31** †	**−0.01 / 0.2% / 0.46** †	**+0.09 / 2.5% / 0.26** †	−0.31 / 1.5% / 0.57

The initial comparison includes two tests not used to estimate the respective specifications. EQ2, estimated using the calm pilot, does not satisfy V7 in a separate rising-rate simulation (maximum balance difference 6.5%; G = 0.59), which motivated the normal-time pricing rule. For EQ4, the prediction recorded before the rising-rate and deposit-withdrawal simulation differs from the LLM bank's ROE by 0.31 percentage points, but its decision discrepancy of 0.57 exceeds the 0.5 threshold, owing to shorter-duration bond purchases and a two-basis-point rather than 30-basis-point addition to the fixed-rate loan margin; such variation between realisations cannot be represented by a single deterministic rule. Two qualifications apply. The withdrawal simulation uses the same customer equations as EQ4, so agreement on balances is partly built in and bank choices are the independent object of comparison. And improving the fit in one simulation can worsen it in another: the pricing rule improves the rising-rate and calm3 comparisons but increases the balance difference for the pilot bank, which maintained a deposit premium.

Three additional LLM bank agents were evaluated in the calm scenario after EQ4 had been specified (Table 9). The equation model's predictions were saved before those simulations. All three meet the full V7 criteria.

Table 9 Three further calm-year LLM banks against the saved EQ4 prediction

Run (calm year)	ROE, EQ4 / LLM	ROE gap	Largest balance gap	Decision gap G	Full V7
calm6	12.96% / 12.97%	−0.01pp	0.7%	0.25	Pass
calm7	12.96% / 12.91%	+0.04pp	0.9%	0.36	Pass
calm8	12.96% / 13.04%	−0.08pp	0.2%	0.32	Pass

Among these four separate evaluations of EQ4, three satisfy V7 and one satisfies the ROE and balance criteria but not the decision criterion. The remaining discrepancy is associated with variation in bank pricing and portfolio decisions across LLM realisations.

Separate evaluation set. After the EQ4 specification and evaluation thresholds had been recorded, we obtained six further bank-year simulations, one for each scenario, using a reworded instruction document with the same economic objectives, decisions and bounds (runs/bank_brief_v2.txt). Predictions and associated file hashes were saved before those simulations began (runs/eval/preregistration.json). Table 10 reports the results. This assessment examines sensitivity to new realisations and alternative wording, rather than external validity.

Table 10 Evaluation set: six bank-years under the reworded instructions against frozen EQ4 predictions

Year (reworded brief)	ROE, EQ4 / LLM	ROE gap (pp)	Balance gap	Largest monthly LCR gap (pp)	Decision gap G	Full V7
Calm	12.96% / 13.19%	−0.24	0.7%	1.6	0.33	Pass
Hike +300bp	16.43% / 16.58%	−0.16	0.5%	0.9	0.12	Pass
Cut −180bp	10.19% / 10.44%	−0.25	0.7%	1.4	0.26	Pass
Recession	9.64% / 9.94%	−0.30	0.9%	1.8	0.36	Pass
Stress (run in month 4)	10.95% / 10.87%	+0.08	3.1%	1.4	0.69	Fail (G, balances)
Hike with a run in month 6	13.62% / 13.58%	+0.05	0.7%	1.5	0.85	Fail (G)

Across the six additional simulations, the ROE difference does not exceed 0.30 percentage points and the monthly LCR difference remains within 1.8 percentage points. All four simulations without a deposit run satisfy V7. The two withdrawal scenarios exceed the decision-discrepancy threshold, and the stress scenario also exceeds the 3% balance tolerance slightly (3.1%), because the reworded-instruction bank reinvests in floating-rate securities in the fair-value portfolio during liquidity restoration, whereas EQ4 directs one-third of reinvestment to fixed-rate securities in the amortised-cost portfolio. The simplified LCR treats the two portfolios equivalently, so aggregate outcomes are hardly affected. Adopting the reworded-instruction allocation would improve fit to these two cases but worsen it for the original-instruction withdrawal cases, so the stress-time portfolio rule is not stable across the tested instructions; we retain the original rule rather than refit on the evaluation observations.

Across the ten separate Claude bank-year evaluations, the annual ROE difference for EQ4 does not exceed 0.31 percentage points, and seven simulations satisfy all V7 criteria. The median decision discrepancy is 0.35 percentage points (range 0.12–0.85). The three simulations not satisfying all criteria involve deposit runs.

The calm3 result illustrates why a choice-level statistic is needed in addition to aggregate outcomes. EQ0 reproduces year-end balances within 0.4% while selecting different prices;

near-market pricing and minimal customer flows make the balance outcomes relatively insensitive to those differences.

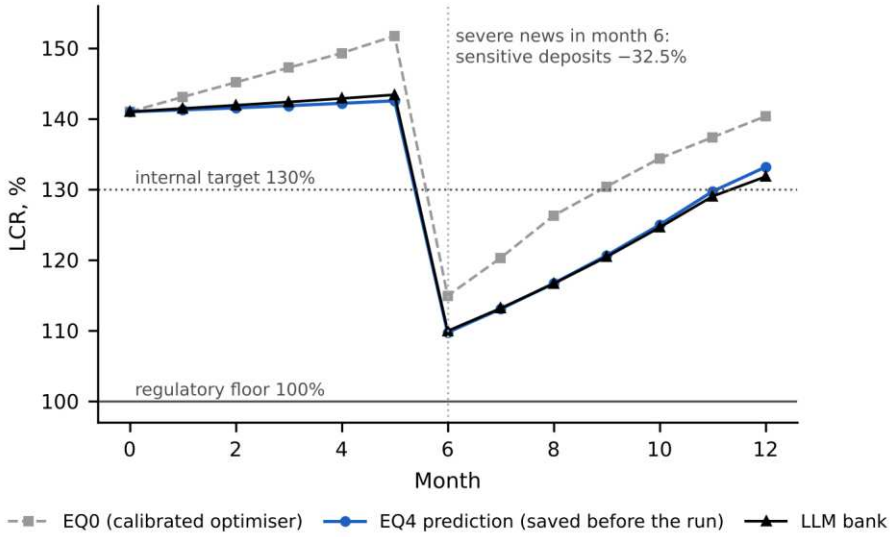

Fig. 3 LCR by month in the out-of-sample rising-rate year with a deposit run in month 6: LLM bank, EQ4 prediction saved before the run, and EQ0. Sources: runs/hikestress5 and runs/eq4

Fig. 3 shows the LCR path in the rising-rate withdrawal year. EQ0 produces a faster recovery in the simulated LCR because its initially calibrated customers respond more strongly to the deposit premium. The revised customer equations and the EQ4 liquidity-pricing rule give a slower, six-month recovery, closer to the corresponding LLM simulation.

Evaluation using a second model family. The initial bank agents all belong to the Claude family. To examine dependence on the language model, we substitute OpenAI's gpt-5.5-2026-04-23 for the bank while retaining the original instruction document, the estimated customer equations and the EQ4 parameters. The model is called at default sampling settings and is not provided with the estimated decision rules or this paper. One bank-year per scenario was followed by two more per scenario under a plan recorded in advance, yielding eighteen bank-years; returned model identifiers and responses are preserved in the replication records (Table 11).

Table 11 EQ4 against eighteen GPT-5.5 bank-years

Scenario (3 GPT years each)	Full V7 passes	ROE gap (pp)	Largest balance gap	Decision gap G
Calm	2 of 3	−0.03 to +0.09	0.9–1.9%	0.36–0.59
Cut −180bp	3 of 3	+0.05 to +0.12	0.6–2.9%	0.31–0.43
Hike +300bp	1 of 3	+0.20 to +0.34	1.6–3.8%	0.35–0.50
Recession	0 of 3	+0.05 to +0.14	3.6–4.6%	0.49–0.70
Stress (run in month 4)	0 of 3	+0.29 to +0.71	2.4–4.0%	0.56–0.95
Hike with a run in month 6	0 of 3	+0.03 to +0.17	2.2–2.6%	0.55–0.80
All 18	6 of 18	median 0.11, max 0.71	within 3% in 12	median 0.50

The EQ4 ROE difference is no greater than 0.71 percentage points across the eighteen GPT bank-years, but only six meet the combined V7 criteria; the corresponding figure for the Claude evaluations is seven of ten. These proportions should not be interpreted as precise between-family differences. Under an independence assumption, the reported 95% Wilson intervals are 16–56% and 40–89%, respectively, and the scenario and instruction compositions differ. The following differences in generated decisions account for much of the reduced fit.

1. **Deposit pricing:** partial correspondence. In the calm, falling-rate and recession scenarios, the GPT banks' average deposit rates are within two basis points of the EQ4 values. During monetary tightening they offer rate-sensitive depositors premiums of two to five basis points above market, leaving balances 1.6–3.8% higher than under EQ4 and contributing to failure of the balance criterion.
2. **Loan margins:** differences across model families. The GPT banks generally begin at the market loan margin rather than applying EQ4's initial 30-basis-point fixed-rate addition. They subsequently increase both margins in five-basis-point increments, reaching average increases of 19 basis points for variable-rate and 29 basis points for fixed-rate loans by month twelve. Higher variable-rate margins reduce the simulated SME loan portfolio by 2–4.6% relative to EQ4. In ordinary conditions, the GPT banks also allocate a larger share to fixed-rate securities (mean F = 0.89, compared with EQ4's 2/3 and the Claude bank mean of 0.69).
3. **Portfolio composition during liquidity restoration:** differences across instructions and model families. From the second month of liquidity restoration, every GPT bank allocates reinvestment to floating-rate securities (F = 0); three of the six relevant simulations use the fair-value portfolio. The same broad choice occurs among Claude banks under the reworded instructions. EQ4's allocation of F = 1/3 and M = 0 is therefore specific to only one of the three tested LLM conditions. The evidence does not support treating that allocation as invariant across model families and instructions.

The cross-family assessment thus shows partial correspondence: EQ4 approximates ROE, deposit pricing and the timing of liquidity responses in the second family, whereas loan margins and reinvestment choices differ. Refitting those choices within the same rule class (EQ4-G; Online Resource 1, Section S6) improves in-sample fit but not the number of held-out years satisfying V7. The chosen LLM, its version and the form of its instructions are therefore components of the simulation specification. Records and results are in runs/crossmodel.

6.5 Scenario results for the final equation specification

EQ4 is used for the subsequent scenario analysis; Table 12 compares it with EQ0 across the five scenarios. Relative to EQ0, its assumed near-market pricing reduces deposit accumulation during monetary tightening, while its liquidity-restoration rule produces larger costs and a slower recovery in the stress scenario.

Table 12 Scenario results, EQ0 and EQ4

Scenario	Model	ROE	12-month NII (EUR m)	ΔEVE/T1 at year end	LCR low	Economic net worth at year end (EUR m)	Sensitive deposits, change
Calm	EQ0 / EQ4	12.9% / 13.0%	276 / 276	−8.2% / −6.7%	141% / 141%	903 / 910	0 / 0
Hike +300bp	EQ0 / EQ4	16.0% / 16.4%	309 / 307	−5.6% / −6.6%	141% / 141%	359 / 379	+23.9% / 0
Cut −180bp	EQ0 / EQ4	10.7% / 10.2%	253 / 251	−8.6% / −6.8%	137% / 141%	1,227 / 1,224	−11.1% / 0
Stress (run in month 4)	EQ0 / EQ4	12.2% / 10.9%	260 / 250	−7.1% / −6.0%	105% / 109%	901 / 897	−32.0% / −24.0%
Recession (−120bp, unemployment +3pp)	EQ0 / EQ4	10.2% / 9.6%	261 / 259	−8.2% / −6.7%	138% / 141%	1,108 / 1,106	−14.3% / 0

The changes in EQ0 rate-sensitive deposits under the falling-rate and recession scenarios depend on the numerical search method; they become −4% and −2% under multiple starting points (Section 3.5). Under rising rates, both EQ0 and EQ4 produce higher NII together with substantial reductions in economic net worth, as the additional earnings are small relative to fixed-asset valuation losses. In the stress scenario, the liquidity-restoration rule reduces annual ROE by approximately 1.1 percentage points (EQ3: 12.1%; EQ4: 10.9%), reflecting the interest expense from deposit premiums maintained over seven months.

6.6 Duration and the NII–EVE trade-off

The effective durations of fixed-rate positions are scenario inputs and can be varied independently of the LLM-generated decisions. We consider values from two to six years for fixed-rate securities held at amortised cost (D_FI) and fixed-rate loans (D_Li), retaining the other EQ4 assumptions (Fig. 4). The exercise illustrates how income and valuation sensitivities vary with these assumptions; it is not a bank-level estimation of actual duration choices.

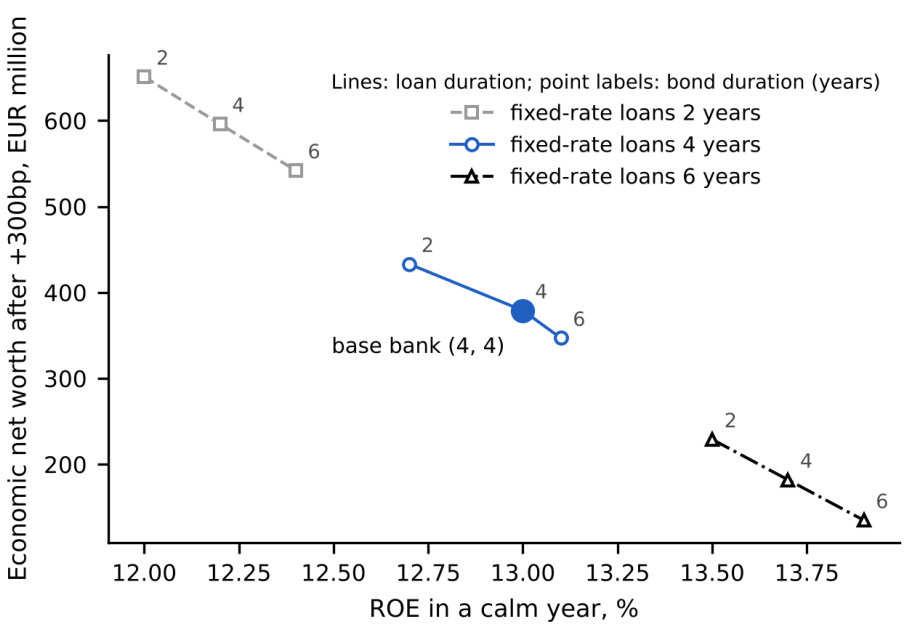

Fig. 4 Calm-year ROE against economic net worth after a +300bp rate increase under EQ4 for nine combinations of fixed-rate loan and bond duration. Source: runs/paper/C_duration_sweep.csv

Longer assumed duration increases simulated earnings in the calm scenario but raises exposure to rate increases. Increasing both durations from two to six years raises calm-year ROE from 12.0% to 13.9% and changes opening ΔEVE/T1 from +16.9% to −29.1%, beyond the model's −15% supervisory threshold. Following a +300-basis-point rate increase, economic net worth falls from EUR 651 million to EUR 135 million. Loan duration contributes more to this effect because outstanding loans are nearly five times larger than the fixed-rate securities portfolio. By comparison, opening ΔNII/T1 changes from +1.7% to +0.7%. The result illustrates why a relatively small NII sensitivity need not imply a small EVE exposure.

Some duration combinations are inconsistent with the assumed internal EVE target. Five of the nine combinations keep opening ΔEVE/T1 above −10%: two- or four-year duration for both categories, or six-year securities with two-year loans. The income–value trade-off persists within this restricted set but is smaller: the maximum calm-year ROE difference is approximately one percentage point and the difference in economic net worth after the rate shock is EUR 272 million.

6.7 Dependence of LLM-generated decisions on agent instructions

The initial instructions define each agent's circumstances and admissible decisions, but also contain guidance about likely behaviour. The depositor instructions label one group rate-sensitive and the other rate-insensitive, state that flows are near zero at market prices, and cap monthly inflows and withdrawals (+2% and 40% for rate-sensitive depositors, +0.5% and 7% for households); the fixed-rate borrower instructions state that a margin above market reduces borrowing "only mildly" and give a default response of 0.3 percentage points per point of unemployment; the bank instructions rank liquidity, capital and interest-rate risk above profit "in strict priority order". Consequently, parameter estimates obtained from those agents may partly reproduce the supplied assumptions rather than reflect decisions generated independently of that guidance. We examine this issue by comparing the original instructions with versions that retain roles, information and decision bounds but omit explicit behavioural suggestions. This is an instruction-sensitivity experiment (an ablation study in machine-learning terminology).

The experiment follows a 2 × 2 design recorded before the additional responses were collected (runs/ablation/SPEC.md), reusing one previously collected GPT-5.5 original-instruction customer replication as specified. The neutral instructions retain agent identity, balances, observed information, available choices, reporting format and regulatory or internal limits, but remove behavioural adjectives, suggested reactions, flow and withdrawal limits, the assumed default coefficient and the ordering of bank objectives; because quantitative constraints are removed as well as wording, the test concerns instructional assumptions, not phrasing alone. Six Claude replications (Sonnet for customers, Opus for the bank) and five GPT-5.5 replications are available per instruction type and questionnaire, all on the same predetermined states. The first three and two were

collected on one day; the enlargement was recorded before the remaining responses were collected the next day, without changing states or instructions, and one truncated GPT bank response was requested again with a larger output allowance (both responses are archived). Customer answers are analysed with the same coefficient estimator; bank answers cover the seven decision states and one additional severe-news state. The design can detect large instruction effects relative to within-instruction variation but cannot characterise the distribution of LLM responses precisely. Table 13 reports the results.

Table 13 Instruction ablation: estimated features under original and neutral instructions in two model families

Behaviour	Claude, original	Claude, neutral	GPT-5.5, original	GPT-5.5, neutral	Equilibrium calibration (EQ0)
Near-zero deposit flow at market price, no news	6 of 6	6 of 6	5 of 5	5 of 5	By construction
Sensitive depositors more price-elastic than households (φ_s / φ_i)	19.7 / 1.1	10.6 / 2.0	7.8 / 1.2	10.5 / 4.8	18 / 6
Sensitive-depositor inflow slope	5.1	9.0	4.9	9.8	18
Variable- more price-elastic than fixed-rate borrowers (η_s / η_i)	2.8 / 0.14	2.4 / 1.11	1.4 / 0.08	1.9 / 0.62	1.3 / 0.8
Sensitive deposits withdrawn under severe news, % a month	28.3	12.3	27.3	13.5	32.5 (EQ4)
Household deposits withdrawn, severe news and weak ratios, %	5.7	3.5	5.0	10.4	7.0 (EQ4)
Default rate per point of unemployment, pp	0.30	0.28	0.30	0.34	0.3 (brief)
Pre-specified check: all three slopes below EQ0 ($\varphi_s < 18$, $\varphi_i < 6$, $\eta_i < 0.8$)	3 of 6	0 of 6	5 of 5	4 of 5	—
Bank: deposits priced at market in a normal month	6 of 6	6 of 6	5 of 5	5 of 5	—
Bank: maximum +50bp deposit premium under severe news	6 of 6	6 of 6	5 of 5	5 of 5	—
Pre-specified check: premium of 25bp or more under severe news, after an outflow and with LCR exactly at its 130% target	2 of 6	1 of 6	5 of 5	4 of 5	—
Bank: no new fixed bonds when ΔEVE/T1 is near its limit	6 of 6	6 of 6	5 of 5	5 of 5	—
Bank: floating bonds while repairing liquidity (mean F)	0.02	0.15	0.00	0.00	1/3 (EQ4)
Bank: fixed-loan markup above market in a normal month	0 of 6	0 of 6	0 of 5	0 of 5	+30bp (EQ4, from runs)

The responses support three distinctions.

1. **Patterns retained without explicit behavioural guidance.** In both model families, the neutral instructions produce approximately zero price-induced deposit flows when the bank offers the market rate, greater price sensitivity among corporate and affluent than household depositors, larger withdrawals under severe news, greater price sensitivity among variable-rate than fixed-rate borrowers, and an increase in defaults of 0.28–0.34 percentage points for each percentage-point rise in unemployment. In the bank questions, near-market deposit pricing, the maximum premium under severe news, shorter-duration purchases near the EVE target and floating-rate reinvestment during liquidity stress also occur without an explicit ordering of objectives. There are qualifications. With LCR exactly at the 130% target

and no stress news, Claude banks offer about 20 basis points under both instructions, below the prespecified 25-basis-point criterion, which is met in two of six original and one of six neutral Claude replications (five of five and four of five for GPT-5.5). The prespecified condition that all three customer price coefficients fall below EQ0 is met in half of the original-instruction Claude replications and none of the neutral ones: fresh Claude sensitive-depositor outflow coefficients range from 12 to 26 (mean 19.7) against EQ0's 18, so a lower elasticity than EQ0 is not reliably reproduced. And in some joint stress states, neutral-instruction GPT households withdraw as much as or more than sensitive depositors do under severe news alone.

2. **Patterns materially influenced by the original instructions.** Without the imposed withdrawal limit, monthly sensitive-depositor withdrawals under severe news are approximately 12–14% in both LLM families, compared with 27–28% under the original instructions and 32.5% in the EQ4 withdrawal table. Removing the phrase "only mildly" increases the fixed-rate borrower price-response coefficient from 0.14 to 1.11 for Claude and from 0.08 to 0.62 for GPT-5.5, approaching or exceeding the original equilibrium-calibration value. The between-instruction differences are appreciably larger than the reported variation within each instruction group: approximately 14–16 percentage points between withdrawal averages, compared with a standard deviation of around three percentage points within groups, and differences of 0.5–1.0 in loan price-response coefficients, compared with standard deviations of 0.02–0.14. Removing the inflow bound makes inflows and outflows more symmetric, and household deposit-price responses rise two- to fourfold, though Claude households remain below EQ0. The low variation across the original four withdrawal estimates therefore reflects repeatability under one instruction, not robustness to instructions.
3. **Features not resolved by one-period hypothetical decisions.** Neither instruction variant produces the approximately 30-basis-point fixed-rate loan markup observed later in several full simulations. That markup develops after the bank agents have received earlier borrower responses. The experiment cannot determine whether it reflects within-simulation adaptation, the behavioural suggestion in the original bank instructions or a combination of the two.

The instruction dependence matters for the liquidity scenarios. Substituting the customer equations estimated from the neutral Claude instructions while retaining EQ4's bank rules changes annual ROE by at most 0.18 percentage points in the calm, rising-rate, falling-rate and recession scenarios (runs/ablation/neutral_model.py). In the stress scenario the minimum LCR rises from 109% to 134% and the fall in rate-sensitive deposits shrinks from 24% to 12%. Two months of severe news no longer trigger withdrawals that outlast the news: LCR reaches 124%, never below 100%, and rate-sensitive deposits fall by 20% rather than 97%. The 100% disclosure threshold is reached only after five consecutive months of severe news, or with monthly sensitive-depositor withdrawals of about 26% or more. The withdrawal persistence reported in Section 7 therefore depends on withdrawal magnitudes that were materially influenced by the original instructions.

6.8 Comparison with alternative specifications

To determine whether the additional estimation procedure improves approximation of the LLM decisions, we compare EQ4 with four simpler specifications using the same simulated environment and fixed customer equations. Where estimation is required, the alternatives use the same four development bank-years (48 monthly decisions); the comparison was specified before scoring (runs/baselines/SPEC.md). The alternatives are a market-pricing rule with the initial portfolio allocation, the original optimisation model EQ0 without LLM-derived parameters, separate regressions of each bank choice on six pre-decision variables, and a three-nearest-neighbours approximation using the same variables. All specifications are evaluated against the ten separate Claude bank-years and eighteen GPT-5.5 bank-years (Table 14).

Table 14 EQ4 against four simpler specifications

Model	Claude passes (of 10)	Claude G (median)	Claude balance gap (median)	Claude ROE gap, max (pp)	Claude LCR gap (median of max, pp)	GPT-5.5 passes (of 18)	GPT-5.5 G (median)
Market-taker heuristic	4	0.54	0.8%	1.46	1.4	5	0.56
EQ0 optimiser (no LLM information)	0	0.69	12.2%	1.28	9.7	0	0.88
Direct regression	0	0.49	5.1%	1.85	5.9	0	0.67
Nearest-neighbour imitation	1	0.38	4.7%	0.75	4.2	1	0.51
EQ4 (calibration loop)	7	0.35	0.7%	0.31	1.4	6	0.50

EQ4 meets all criteria in more years and has the lowest median decision discrepancy and smallest maximum ROE difference in both evaluation sets, but it does not dominate on every measure: the market-pricing rule reproduces the Claude ΔEVE path more closely in some years, and the nearest-neighbours specification has a smaller median LCR-path difference in the GPT years (2.6 versus 5.5 percentage points). In the seven Claude years without substantial withdrawals, the market-pricing rule meets all criteria in four and matches balances and LCR about as well as EQ4, which differs mainly through its bond-allocation and fixed-rate margin rules. In the three withdrawal years, EQ4 has the lowest median maximum monthly LCR difference (1.4 percentage points against 5–21 for the alternatives) and decision discrepancy (0.69 against 0.76–1.62), although none meets all criteria. The regression and nearest-neighbours alternatives are estimated from 48 monthly observations containing one withdrawal episode, which limits their accuracy in stress states. Annual ROE alone discriminates little in ordinary conditions, which supports the decision-level criteria. Results by scenario are in runs/baselines/benchmark_by_run.csv.

Comparison using additional bank-decision information. EQ4 used answers to hypothetical bank-decision states that were not supplied to the initial regression or nearest-neighbours alternatives; all models use the same customer equations. We therefore consider two additional comparisons specified before scoring (runs/baselines/SPEC.md, addendum; Table 15): regression and nearest-neighbours

models re-estimated using the original forty-eight monthly observations plus seven bank hypothetical-state answers, and EQ4 with the bond rule derived exclusively from those hypothetical states omitted.

Table 15 Information-parity comparison

Model	Claude passes (of 10)	Claude G (median)	Claude run years: LCR gap (median of max, pp)	GPT-5.5 passes (of 18)	GPT-5.5 G (median)
Direct regression	0	0.49	6.1	0	0.67
Direct regression with the bank hypothetical-state questionnaire	0	0.45	3.5	2	0.58
Nearest-neighbour imitation	1	0.38	6.9	1	0.51
Nearest-neighbour imitation with the bank hypothetical-state questionnaire	1	0.40	5.0	2	0.51
EQ4	7	0.35	1.4	6	0.50
EQ4 without its questionnaire-only bond rule	6	0.33	1.4	6	0.44

The seven additional answers improve some liquidity comparisons and add two GPT full-criterion cases for the regression and one for nearest neighbours, but those alternatives still meet all criteria in at most one of the ten Claude years, so EQ4's relative performance rests mainly on its conditional structure rather than on extra bank observations. The hypothetical-state bond rule itself has mixed effects: it improves some rising-rate, recession and withdrawal comparisons but reduces fit in calm and falling-rate years, adding one Claude full-criterion case while slightly raising the median decision discrepancy (0.35 versus 0.33 for Claude; 0.50 versus 0.44 for GPT-5.5). In this application, hypothetical-state decisions matter most for the customer equations and the liquidity-restoration rule; their value for the normal-time bond rule is less clear (runs/baselines/benchmark_plus_by_run.csv).

7. Scope and boundary conditions

The conclusions depend on a specific balance-sheet structure, twelve-month horizon, set of possible bank decisions and simulated customer responses. This section states the relevant restrictions and examines how the model behaves outside the scenarios used for estimation.

7.1 Model coverage and exclusions

Table 16 lists what the model assumes and what it excludes.

Table 16 Model coverage and exclusions

Boundary	What the model assumes	What lies outside it
Market structure	One bank sets prices; competitor rates follow the policy rate through Paper 1's pass-through rule	Strategic competitors, deposit-price wars, interbank lending, fire sales and contagion
Horizon and equilibrium	A 12-month year of one-month Stackelberg stage games with an open-loop 12-month forecast	A dynamic (closed-loop) equilibrium, multi-year franchise building, capital raising

Boundary	What the model assumes	What lies outside it
Rate risk	Parallel shifts and a linear curve; durations stand for maturity ladders	Twists, basis risk, embedded options beyond prepayment and behavioural deposit duration
Runs	Severe news for one month, with a stress-withdrawal table estimated from questionnaires	Runs that persist or spread through social media and interbank channels
Credit	Defaults follow unemployment with a fixed loss given default	Concentration risk, collateral values, credit cycles longer than a year
Behaviour	LLM agents with plain-language instructions, estimated back into equations	Human treasurers, depositors and borrowers with real money at stake
Calibration	A stylised, predominantly deposit-funded bank (88% customer deposits) calibrated to selected median IRRBB and liquidity disclosures of Papers 1 and 2	Banks funding less than about two-thirds of assets with customer deposits, whose other funding is not overnight (Section 4: opening LCR 56–66%, negative ROE); bank-level ΔEVE, which depends on hedges the balance sheet does not reveal; very small banks and banks with large trading books

7.2 Behaviour beyond the estimated scenarios

We examine the sensitivity of EQ4 to scenarios more severe than those used to estimate its rules. The first set varies the duration of severe stress news from one to six consecutive months, annual policy-rate increases of zero, 300 or 500 basis points, and competitor-rate pass-through at its estimated value or 50% faster. A second set varies annual unemployment increases up to eighteen percentage points and loss-given-default assumptions of 40% and 60%. Table 17 reports the liquidity scenarios at the estimated pass-through value; complete results appear in runs/robust/R4_boundary.csv and R5_credit_boundary.csv.

Table 17 Liquidity boundary grid: consecutive months of severe news and rate increases

Run (months)	Hike (bp)	ROE	LCR low	Months LCR < 100%	Sensitive deposits, change	ΔEVE/T1 low
1	0	10.95%	109%	0	−24%	−7.2%
1	+500	14.66%	109%	0	−24%	−7.2%
2	0	6.43%	65%	8	−97%	−9.8%
2	+500	7.01%	63%	8	−97%	−10.2%
4	+300	4.60%	56%	8	−97%	−15.2%
6	+500	2.27%	49%	8	−97%	−20.4%

The analysis identifies three boundaries.

1. **Duration of deposit withdrawals and the disclosure assumption.** Within the tested grid, a one-month withdrawal episode leaves LCR above 108% and the deposit-pricing response then restores liquidity. Two consecutive months of severe news can push published LCR below 100%; the stress indicator then stays elevated through its published-LCR term even after the news ends, rate-sensitive deposits fall by 97% over the year, wholesale funding replaces them and LCR stays below 100% for eight months, because the deposit premium cannot offset withdrawals at that stress level. EQ4's liquidity-restoration findings therefore apply to episodes that end before published LCR falls below 100%. The threshold's location depends on withdrawal magnitudes: with neutral-instruction responses, two months of severe news leave LCR at 124% and five months are needed to reach the persistent state

(Section 6.7). Crisis-resolution instruments and central-bank facilities are not represented; Section 7.3 considers alternative disclosure assumptions.

2. **Economic-value exposure following changes in funding composition.** In the tested settings, ΔEVE/T1 falls below the −15% threshold only when four or more consecutive months of withdrawals coincide with a rate increase. Deposits assigned a behavioural duration of 3.35 years are replaced by short-term wholesale funding, shortening the effective liability duration while fixed-rate assets retain their duration. At the assumed initial portfolio composition, the rate shock alone does not produce a breach.
3. **Limited information about capital-related choices.** Even with an eighteen-percentage-point unemployment increase and 60% loss given default, simulated ROE falls to −4.4% but CET1 does not decline below its opening 15.7%. Consequently, the capital-related fallback optimisation rule is not exercised in these scenarios. The model cannot evaluate its accuracy from the available simulations; doing so would require a different capital position, more prolonged losses or a longer horizon.

Online Resource 1, Section S7, tabulates for each main result the conditions under which it holds and those under which it was shown to weaken.

7.3 Liquidity disclosure and securities valuation assumptions

The withdrawal persistence described in Section 7.2 depends partly on the assumption that depositors observe the preceding month's LCR. Actual regulatory disclosures differ. For example, the EU LIQ1 Pillar 3 template reports quarterly averages of the preceding twelve month-end LCR values. The model's HQLA calculation is also simplified, assigning all securities their carrying values without haircuts. We therefore repeat the withdrawal scenarios under four disclosure arrangements and two HQLA valuation assumptions, keeping the bank's internal liquidity measure unchanged (runs/robust/R6_disclosure.csv; Table 18).

Table 18 Two-month run under alternative LCR disclosure and HQLA valuation assumptions

LCR seen by depositors	HQLA	1-month run: LCR low, sensitive deposits	2-month run: LCR low, months < 100%, sensitive deposits	Run self-sustaining?
Last month-end (baseline)	Book value	109%, −24%	65%, 8, −97%	From 2 months
Last quarter-end	Book value	109%, −24%	66%, 8, −96%	From 2 months
12-month average at quarter ends (Pillar 3)	Book value	109%, −24%	89%, 6, −48%	Never within the grid
Not disclosed	Book value	109%, −24%	89%, 6, −48%	Never within the grid
Last month-end	Market value, 15% haircut on bonds	56%, −97%	52%, −97%	From 1 month
Last quarter-end	Market value, 15% haircut on bonds	99%, −18%	58%, −95%	From 2 months
12-month average at quarter ends	Market value, 15% haircut on bonds	99%, −18%	79%, −46%	Never within the grid

Two conditional findings emerge. First, with the estimated withdrawal function, the persistence of withdrawals depends on whether depositors receive a contemporaneous LCR signal. Under twelve-month averaged disclosure, the publicly observed indicator remains above 100% in the tested year, so the modelled stress response subsides when adverse news ends; the outcome is equivalent to the case without public LCR disclosure. Second, applying market valuations and a 15% haircut to securities reduces opening LCR from 141% to 125%, below the 130% internal target. With monthly disclosure, even one month of severe withdrawals can then generate continued withdrawals through the assumed LCR response. Both findings depend on withdrawal magnitudes partly suggested by the original LLM instructions (Section 6.7). They illustrate an interaction between an assumed depositor response and the disclosure rule; they do not establish that actual depositors would behave in this manner.

8. Discussion and implications

The principal contribution of this paper is a procedure for estimating and evaluating equation-based approximations of decisions generated by LLM agents. The banking results illustrate what can be examined using such a model, subject to its simplified funding structure, response assumptions and instruction dependence. They should not be interpreted as direct findings about the behaviour of human bank managers, borrowers or depositors. The following discussion therefore distinguishes computational findings from propositions requiring further empirical investigation.

8.1 Behavioural patterns in the LLM simulations

The original depositor instructions included the weak aggregate quantity response to spreads identified in the empirical calibration (Paper 1), so the low deposit-price sensitivities elicited under them cannot independently confirm it. The neutral-instruction comparison (Section 6.7) retains near-zero flows at market prices and the ordering of depositor sensitivities, while fixed-rate borrowers become substantially more price-sensitive. Under the original instructions, households move about 1.3% of balances a month per percentage point of rate difference and rate-sensitive depositors about 1.2% per ten basis points, far below the sensitivities that the initial equilibrium calibration imposed to make market pricing optimal. The bank agents' near-market pricing persists under neutral instructions, so it is not solely a consequence of those imposed coefficients; the experiment does not, however, identify the objective behind it.

A few conditional rules approximate the LLM banks' decisions more closely than the optimisation benchmark: deposits at market in ordinary conditions, fixed-rate loan margins raised over repeated decisions, bond duration adjusted as the EVE target tightens, and larger deposit premiums under liquidity stress that are withdrawn once liquidity recovers. Comparable responses arise when the ordering of objectives is removed from the instructions, although loan margins and bond allocations vary across simulations. The preference for longer fixed-rate securities, which yield more on the assumed upward-sloping curve and, at amortised cost, keep value changes out of profit, is consistent with the income-versus-value considerations of DeMarzo, Krishnamurthy and Nagel (2024); the

correspondence is conceptual and does not show that banks select portfolios through the estimated rule, or that the agents pursue no alternative objective.

8.2 Theoretical implications

Rule-based approximation of simulated bank decisions. Among the eight development simulations, conditional rules based on liquidity conditions, economic-value headroom and market pricing approximate the decisions more closely than the tested objective function. This is consistent with satisficing in the sense of Simon (1955) without establishing a decision mechanism, and the original instructions ranked the objectives; neutral-instruction answers still weight liquidity restoration heavily, whereas the fixed-rate loan markup appears only after repeated interaction with borrowers (Section 6.7). Rule-based reactions of this kind also appear in top-down stress tests such as RAMSI and STAMP€. Whether human treasuries use comparable rules is untested, and, as Lucas (1976) emphasised, estimated decision relationships may change with the environment, as Section 7.2 illustrates for prolonged withdrawals.

Relative influence of market-rate pass-through and customer price responses. In a Monti–Klein framework, margins and quantities depend on demand elasticities and alternative prices (Klein, 1971; Monti, 1972; Freixas and Rochet, 2008). Because the simulated bank prices near market, outcomes are insensitive to moderate changes in customer price responses: moving any single coefficient by two standard deviations changes ROE by at most 0.3 percentage points (Section 9.1), whereas changing competitor-rate pass-through by ±50% moves rising-rate ROE by about ±2.4 percentage points. The assumed path of alternative deposit rates is therefore the main determinant of the simulated income response, which motivates testing pass-through jointly with depositor sensitivity; it does not show that price responses are unimportant in other settings.

Interaction of liquidity disclosure and depositor responses. Under the estimated withdrawal function, depositors treat a published LCR below 100% like severe news, so withdrawals can continue after the shock has ended (Section 7.2). This resembles the feedback in theoretical run models (Diamond and Dybvig, 1983; Goldstein and Pauzner, 2005), although the deterministic simulation estimates no run equilibrium. With an LCR averaged over twelve month-ends, as in the EU template, the same two-month episode ends with the news and rate-sensitive deposits fall by 48% rather than 97% (Section 7.3); with neutral-instruction depositors, two months of severe news do not cross the threshold at all (Section 6.7). The result illustrates how disclosure frequency may interact with depositor responses; it is not evidence that more frequent disclosure causes runs.

Duration choices under an economic-value constraint. The estimated bond rule lengthens duration while ΔEVE/T1 is comfortably above the internal threshold and shortens it as the threshold approaches, consistent with accounts in which banks balance income stability against economic value (Drechsler, Savov and Schnabl, 2021; DeMarzo, Krishnamurthy and Nagel, 2024). The Pillar 3 disclosures used in the empirical calibration (Paper 1) give only indirect context: ΔEVE/T1 varies little with funding composition, 82% of observations lie above −10% of Tier 1 and one of 205 is beyond −15% (Section 4).

Testing for clustering inside internal limits would require data on those limits and on hedge-adjusted duration.

Use of LLM-generated responses in agent-based modelling. Agent-based financial models usually rely on aggregate calibration targets (Fagiolo, Moneta and Windrum, 2007; Grazzini and Richiardi, 2015; Platt, 2020). LLM choices requested for specified states add decision-level observations from which parsimonious equations can be estimated and then evaluated on separate scenarios. This can support model construction and sensitivity analysis, but complements rather than replaces estimation and validation on observed transactions or incentivised human responses.

8.3 Implications for further empirical and supervisory analysis

The following points are conditional implications for testing with bank, depositor and borrower data, not recommendations established by the LLM experiments.

For bank asset-liability management:

- The income–value trade-off is explicit: lengthening fixed-rate assets from two to six years raises calm-year ROE by 1.9 percentage points but lowers economic net worth after a +300-basis-point shock by EUR 516 million, about two-thirds of opening equity; within the internal EVE target the ranges shrink to 1.0 percentage point and EUR 272 million (Section 6.6).
- Depositor sensitivity to public information merits estimation alongside rate sensitivity: under the estimated withdrawal function, stress news and published ratios move more deposits than a 25-basis-point rate difference, although the magnitude is highly sensitive to the instructions.
- A deposit premium works only early. The 50-basis-point premium restores LCR after a one-month episode at a cost of about 1.1 percentage points of ROE, but once published LCR is below 100% the estimated response sustains withdrawals regardless of price; contingent funding is outside the model, and the threshold moves materially under the neutral-instruction withdrawal equations (Section 6.7).
- Projected NII is more sensitive to assumed competitor pass-through than to modest changes in customer price responses, so both deserve sensitivity analysis; the ranking need not hold for other banks or markets.

For prudential analysis:

- Liquidity and interest-rate exposure interact: in the severe scenarios the EVE threshold is breached through the change in liability composition during withdrawals, not through the rate shock alone. This is an interaction to evaluate in supervisory data, not a demonstrated frequency for actual banks.
- Disclosure arrangements may matter where depositors respond discontinuously to published ratios: a twelve-month averaged indicator removes the withdrawal persistence that monthly point-in-time disclosure generates under the original instructions, and neutral-instruction withdrawals reach the threshold only after a longer episode (Section 6.7). Policy relevance depends on empirical depositor

responses and on the informational benefits of disclosure, neither of which is estimated here.

- Rule-based reactions can serve in top-down stress tests if their parameters and applicability are evaluated empirically; the liquidity-restoration rule changes both the projected deposit path and the earnings cost of holding liquidity, and is offered as a candidate specification, not a validated reaction function.

For model development, a bank-only LLM simulation uses about 0.9 million tokens and yields hypothetical decisions without a human-subject experiment, although cheaper evidence is not equivalent evidence. Outcomes here depend strongly on the stress-withdrawal function and competitor pass-through, which are the priorities for comparison with transaction-level or supervisory data. Instructions should state the setting, the information and the constraints without specifying the reactions that are to be estimated; Section 6.7 shows why.

8.4 Alternative model specifications considered

Six alternative specifications were examined and not retained for the reported model. Four changes to the objective function produced no differences in the tested decisions or outcomes; the alternatives and the reasons for not retaining them are listed in Online Resource 1, Section S5.

Appendix A3 summarises considerations for assigning parameter sources, recording agent instructions and evaluating equation-based approximations in future applications.

9. Robustness and limitations

9.1 Robustness

Twelve sensitivity and robustness analyses examine sampling variation, parameter uncertainty, alternative modelling choices and the accuracy of numerical procedures. Table 19 summarises the twelve tests; the corresponding detailed tables are provided in Online Resource 1, Section S4, with scripts and detailed outputs in runs/robust.

Table 19 Robustness summary

Source of doubt	Test	Effect on outcomes	Verdict
Sampling noise in customer answers	Four replications of each hypothetical-state questionnaire; EQ5 uses their mean	ROE within 0.2pp, LCR low within 1.1pp, net worth within EUR 2m	Robust
Single-parameter error	Each customer coefficient moved by ±2 s.d. of the replications	Price slopes: no change. Borrower slopes: ≤0.3pp ROE, ≤0.5pp LCR low in stress	Robust
Run severity	Run shares ×0.5 and ×1.5	Stress-year LCR low +19 / −8pp, ROE +1.5 / −0.9pp; calm and hike years unchanged	Sensitive; stress results are conditional on the stress-withdrawal table
Market environment	Competitor pass-through ×0.5 and ×1.5	Hike-year ROE ±2.4pp, net worth ±EUR 20m; calm year unchanged	Sensitive; a first-order input
Bank heterogeneity	EQ4 rerun with each of the eight LLM banks' own normal-times	Calm ROE 12.2–13.1%, hike 15.9–16.5%, stress LCR low 109–111%	Robust in sign and ranking

Source of doubt	Test	Effect on outcomes	Verdict
	pricing and bond mix		
Model dependence	Fresh questionnaire answers from three Claude models (Opus, Sonnet, Haiku); baseline customer agents are Sonnet	Calm, hike, cut and recession years within 0.03pp ROE. Stress-year LCR low 103% (fresh Sonnet), 120% (Opus) and 126% (Haiku) against 109%	Normal-times results robust; stress results depend on the model
Prompt wording	Six bank-years with a reworded bank brief, predictions frozen in advance	ROE within 0.30pp and LCR within 1.8pp in every month; deposit premiums identical; repair-time securities portfolio differs	Pricing rules robust; securities portfolio under stress sensitive to wording
Model family	OpenAI GPT-5.5 as the bank in 18 pre-planned bank-years; GPT customer hypothetical-state questionnaires; refit of EQ4's rule class on 12 GPT years, scored on 6 held out	ROE within 0.71pp in all 18; full V7 in 6; loan margins creep up and repair bonds go floating; refit passes 1 of 6 held-out years; GPT customer estimates inside the Claude range on 8 of 11 parameters	Outcomes and deposit pricing transfer; loan-margin and bond-mix rules are family-specific
Identification of the calibration	Balance sheets of 53 (57) banks from EBA balance-sheet and IRRBB disclosures (Online Resource 2) loaded into the engine with no refit; disclosed ΔEVE/T1 and ΔNII/T1 held out	ΔNII/T1 of the deposit-funded third matched (1.6% vs 1.6%); ΔEVE/T1 rank correlation −0.28 (−0.01); wholesale-funded profiles give opening LCR 56–66%	ΔNII identified within scope; ΔEVE not identified by composition; results limited to deposit-funded banks
Behaviour supplied by the instructions	Questionnaires answered under original and neutral instructions by Claude and GPT-5.5 (Section 6.7)	Orderings, near-zero flows at market and the maximum premium under severe news survive (smaller premium at LCR target for Claude); sensitive-deposit runs fall from 27–28% to 12–14% a month; fixed-rate borrower elasticity rises to near or above the calibration's; with neutral-brief customers a two-month run leaves LCR at 124%	Qualitative rules robust; withdrawal magnitudes and fixed-rate borrower elasticity instructed; persistent withdrawals conditional on them
Added value of the loop	EQ4 against a market-taker heuristic, EQ0, direct regression and nearest-neighbour imitation, same training years, and the latter two also given the bank hypothetical-state questionnaire (Section 6.8)	Full V7 passes, Claude out of sample: 7 against 4, 0, 0, 1 (0 and 1 with the bank questionnaire); GPT-5.5: 6 against 5, 0, 0, 1 (2 and 2 with the bank questionnaire); EQ4 without its questionnaire-only bond rule: 6 and 6	EQ4 leads on passes (all in normal years) and on liquidity and decision fit in run years, which no model passes
Optimiser accuracy (benchmark banks)	Multistart and random search against the coordinate search; EQ0 rerun with a multistart bank	ROE ≤0.2pp, LCR low ≤3pp; EQ0 deposit outflow in falling-rate years 2–4% instead of 11–14%	Rule-based EQ4 unaffected; EQ0 deposit volumes approximate

Three points follow. Outcomes in ordinary conditions are insensitive to the uncertain customer price-response coefficients because the bank prices close to market. Liquidity outcomes depend materially on the stress-withdrawal function and competitor pass-through, and the former was partly specified by the original instructions (Section 6.7). And the LLM itself matters: a fifth Sonnet response lies within two standard deviations of the four replications for thirteen of fifteen parameters, but Opus produces a 22.5% and Haiku a 20% severe-news withdrawal against Sonnet's 32.5%, with no monotonic relation to model tier, and GPT-5.5 banks differ in normal-time loan margins and portfolio choices (Section 6.4). Model identity and instructions are therefore part of the specification, particularly in stress scenarios.

The evaluations differ in independence. The customer equations and EQ2 each had one held-out rising-rate test; EQ4 had ten separate Claude bank-years, six of them under reworded instructions after its parameters and the criteria were fixed, with file hashes

recorded in runs/eval/preregistration.json (an internal audit trail, not an external preregistration). Bank-only evaluations share the equation-based customers, so part of the agreement in balances and ROE is built in. All three deposit-run cases fail the decision-discrepancy criterion, two of them because the securities portfolio chosen during liquidity restoration differs between instruction variants. Pass counts also hide distances: median G across the ten Claude evaluations is 0.35 (range 0.12–0.85) against a threshold of 0.5, and at most 0.36 in the seven years without a run. The eighteen GPT-5.5 years test family transfer and are reported separately; the comparisons with simpler specifications (Section 6.8) measure relative computational fit, not validation against actual banking decisions.

9.2 Limitations

The following limitations distinguish the steps undertaken in this study from evidence still required.

1. **Limited numbers of LLM realisations.** The analysis includes fourteen Claude bank-years (eight with original instructions and six with reworded instructions), eighteen GPT-5.5 bank-years used for cross-family assessment, two full simulations with LLM customers and four repetitions of each original customer decision questionnaire. Separate bank-year assessments and sensitivity analyses provide some information about variability. Nevertheless, the sample is insufficient to estimate the frequency of less common decisions, such as a sustained ten-basis-point deposit premium, with precision.
2. **No direct validation of human behaviour.** The procedure evaluates how well equations reproduce LLM-generated decisions. Hypothetical-state answers are not incentivised, and LLMs may reproduce textbook descriptions, regulatory conventions or patterns in their training data; existing research documents reduced dispersion, instruction sensitivity and uneven representation of human responses (Aher, Arriaga and Kalai, 2023; Bisbee et al., 2024; Dillion et al., 2023; Santurkar et al., 2023). The neutral-instruction comparison shows that several original findings depend on supplied assumptions: without the withdrawal bounds, the severe-news share of sensitive depositors falls from roughly 28% to 12–14% a month, below the 25% and 40% regulatory runoff rates for operational and non-operational corporate deposits. In Tieleman's (2022) terms, the comparisons concern agreement with a simulated decision process, not the mechanisms of human decisions; deposit-flow and loan-demand data are needed for that.
3. **Dependence on LLM family and instruction wording.** Most development decisions come from Claude models. Comparisons across Opus, Sonnet and Haiku and the eighteen GPT-5.5 bank-years give a partial assessment (Sections 6.4 and 9.1): ROE and deposit pricing are similar across families, loan margins and stress-time portfolio composition are not, a within-rule-class refit does not satisfy more held-out GPT criteria, and GPT customer estimates fall within the Claude range for eight of eleven coefficients. Separating systematic differences from generation variability would require more families and repeated realisations, and a fixed decision-discrepancy tolerance compares poorly across families with different within-family variation.

4. **Single bank and exogenous alternative prices.** The model contains one decision-making bank; competing deposit prices follow an externally specified pass-through function. It does not include a strategic competing bank, interbank transactions, market-wide asset-price effects or contagion. Varying competitor-rate pass-through by ±50% changes rising-rate ROE by approximately ±2.4 percentage points. A multi-bank extension would be required to study endogenous competitive deposit pricing.
5. **Simplified interest-rate exposure and time horizon.** Effective durations approximate the maturity structure of fixed-rate positions, interest-rate shocks are parallel and the analysis covers twelve one-month decisions rather than a multi-period equilibrium with fully anticipated future strategies. Durations are varied from two to six years, but non-parallel yield-curve changes and longer-term changes in franchise value remain outside the present analysis.
6. **Rule-based representation is not an optimality result.** In the scenarios reached, EQ4 usually applies directly estimated conditional choices rather than maximising the original objective function. Its fit to LLM decisions therefore cannot establish optimal bank policy. The optimisation-based EQ0–EQ2 specifications remain separate comparison cases.
7. **Aggregate-moment calibration does not determine bank-level mechanisms.** Several combinations of deposit behaviour, duration and hedging match the selected ΔNII, ΔEVE and runoff statistics. Against bank-level disclosures, modelled ΔNII/T1 corresponds most closely for predominantly deposit-funded banks, whereas ΔEVE/T1 does not (Section 4); the model therefore restricts its scope and does not infer hedge-adjusted duration for individual banks, which would require derivative positions and a differentiated wholesale-funding block.

9.3 Follow-on research

The following extensions address particular limitations while retaining the main model structure.

- **Duration choice and the NII–EVE trade-off:** examine whether effective duration and ΔEVE/T1 are related to banks' disclosed or observable risk limits, using sufficient information for a formal test of clustering near those limits.
- **Depositor withdrawals and disclosure:** compare the assumed stress-withdrawal function with observed deposit flows and examine the effects of alternative disclosure measures and frequencies.
- **Variation among simulated banks:** obtain enough independent LLM realisations to estimate the distributions of deposit premiums, lending margins and reinvestment choices under fixed instructions.
- **Strategic competition:** introduce a second bank and allow alternative deposit rates to be determined jointly by the banks' decisions and customers' responses.
- **Bank-specific calibration:** use disclosed ΔEVE and other risk information to estimate effective, hedge-adjusted durations, and distinguish covered bonds, senior debt, interbank liabilities and central-bank funding according to their contractual maturities and liquidity characteristics.

- **Human responses and further model-family evaluation:** collect incentivised human choices under comparable scenarios, evaluate the same hypothetical states with additional LLM families and extend cross-family comparisons to full five-agent simulations.

10. Conclusion

This paper develops and evaluates a procedure for representing LLM-generated economic decisions with parsimonious equations, applied to a stylised, predominantly deposit-funded bank whose pricing, lending and investment choices interact with simulated customer responses under interest-rate and liquidity scenarios. Repeated estimation and evaluation yield decision rules that approximate the LLM choices in additional simulations; this establishes fit within the specified computational setting, not a model of observed human behaviour.

The final specification uses a small number of conditional bank rules and customer-response equations. In ten separate Claude bank-year evaluations using fixed equation-based customers, annual ROE differs by no more than 0.31 percentage points, and seven evaluations satisfy all criteria. The three cases not meeting all criteria involve deposit runs. In eighteen GPT-5.5 bank-year evaluations, the corresponding ROE difference remains below 0.71 percentage points but only six satisfy all criteria, mainly because loan-margin and reinvestment decisions vary across model families. Simulations with neutral customer instructions further show that withdrawal magnitudes and fixed-rate borrower price responses are sensitive to information supplied in the original instructions. Accordingly, the simulated persistence of withdrawals following an LCR disclosure breach depends on assumptions that cannot be treated as independently identified depositor behaviour. The duration analysis illustrates a trade-off between earnings and economic value under the chosen balance-sheet structure; it does not estimate the optimal duration decisions of real banks.

The evidence is strongest for the reproducibility of the accounting model and the ability of a structured equation-based specification to approximate selected LLM bank decisions; it is more limited for the joint five-agent system, since most evaluations use fixed equation-based customers, and the correspondence across LLM families is incomplete. There is no direct evidence that the estimated rules describe actual bank managers, borrowers or depositors. The contribution is therefore a transparent method for estimating, comparing and revising simulation-based decision rules, with explicit tests of instruction dependence; assessing its value for behavioural banking research will require transaction data or incentivised human decisions.

Appendix

A1. Conditional decision rules in EQ4

Bond reinvestment rule (introduced in EQ2). In each period the reinvestment proportions (F, M) are determined by the first applicable condition:

1. Published LCR ≤ 1.31: F = 0, M = 1 (floating-rate securities in the fair-value portfolio).
2. ΔEVE/T1 < −9%: F = 0.
3. ΔEVE/T1 between −9% and −8%: F = ⅓.
4. Curve slope ≥ 2pp: F = 1.
5. Otherwise F = ⅔, M = 0 (amortised-cost investment portfolio).

Ordinary-condition pricing rule (introduced in EQ3). In the absence of stress news, with published LCR above 1.30 and published CET1 of at least 14%, the bank sets $d_s = a_s$, $d_i = a_i$, $m_s = \bar{m}_s$ and $m_i = \bar{m}_i + 0.30$.

Liquidity-restoration rule (introduced in EQ4). Under stress news or published LCR no greater than 1.30, the bank sets $d_s = a_s + 0.50$, $d_i = a_i + 0.30$, $m_s = \bar{m}_s + 0.30$, $m_i = \bar{m}_i + 0.50$, F = ⅓ and M = 0. This condition takes precedence over the bond reinvestment rule; the latter's first condition (LCR ≤ 1.31) operates only when published LCR lies between 1.30 and 1.31.

Capital-related fallback. If published CET1 is below 14% without stress news or a liquidity shortfall, the bank maximises J(x) with $\lambda_{LCR} = 0.5$, $\kappa_{dep} = 1.0$ and $\kappa_{loan} = 0.5$. None of the simulated scenarios reaches this condition.

A2. Code, reproducibility and replication procedures

The Python package banksim contains the accounting model, agents, bank objective, calibration, validation and the estimation and evaluation procedure. Every LLM run is saved with its raw answers, monthly log and the prediction made before it. The script runs/paper/paper_results.py reproduces every number in Section 6 from those files.

The replication package contains the banksim Python implementation, the complete LLM instructions and hypothetical-state questions, unedited responses, monthly accounting records, predictions recorded before separate evaluations, and scripts for the reported tables and figures. REPLICATION.md identifies the relevant files and model versions (Table 4). Given the same inputs, the equation-based specification is deterministic and reproduces its reported values. LLM agents ran at default sampling settings without controlled seeds, so new calls may yield different decisions. This motivates reporting repeated responses and evaluations separate from estimation.

A3. Considerations for assigning and estimating agent parameters

The estimation procedure, model-family comparison and bank-level assessment illustrate several principles for assigning parameters to equation-based and LLM agents (Table 20). Each parameter should have an explicit empirical or modelling source, a specified role in the two representations and an appropriate evaluation criterion.

Table 20 Parameter classes, sources and roles in the equation-based and LLM agents

Parameter class	Example in this paper	Identified from	Equation agents	LLM agents	Test before use
Balance-sheet	Deposit, loan	Supervisory bank-	Opening state, bank	Observed state in	Engine outputs

Parameter class	Example in this paper	Identified from	Equation agents	LLM agents	Test before use
composition	and securities shares	level data (EBA Transparency, FINREP)	by bank, only where the engine's funding structure fits	each monthly snapshot	against the bank's own disclosures (Section 4)
Market environment	Competitor pass-through	Panel estimates (Paper 1)	Fixed input	Never as coefficients; only the competitor rates they produce	Sensitivity of ±50%, reported as first-order
Regulatory rules	LCR run-off weights, outlier threshold	Regulation; rates implied in Paper 2's data	Fixed input	As goals and limits in plain language in the brief	None needed; record the version of the rule
Rate-risk position	Asset duration, deposit duration, hedges	Disclosed ΔEVE and ΔNII after hedging	Backed out per bank, or chosen by the bank within its limits	Not loaded; the bank chooses the bond mix	Calibrate on one disclosed sensitivity, predict the other
Counterparty behaviour	Price elasticities, stress-withdrawal table	Strategy-method questionnaires now; human micro data later	Estimated with standard deviations, propagated at ±2 s.d.	Elicited, never told	Fresh replications, three model sizes, a second family
Bank conduct	Pricing, repair and bond rules	LLM bank-year simulations and questionnaires across model families	Rule class shared across families; coefficients per family	Never shown	Holdout frozen in advance; refit on development runs only
Model identity	Family, dated version, brief wording	Run metadata	Not applicable	Pinned dated model IDs, hashed instructions, returned model logged per call	Report within-family spread; set fit thresholds relative to it

Five considerations follow.

1. **Distinguish predetermined inputs from estimated responses.** Equation-based agents receive parameter values; LLM agents receive observable states, objectives and admissible bounds without being told the coefficients or decision rules that the experiment intends to estimate. Otherwise responses may mechanically reproduce the instructions. Record the exact instructions, models and parameter files, including file hashes. Estimate sensitivity to behavioural guidance by repeating the elicitation after removing such guidance (Section 6.7).
2. **Separate calibration and comparison observations.** In this study, three summary statistics were used for initial calibration and the bank-level distribution was reserved for a subsequent comparison. Specifying the mapping and test statistics before examining the comparison results prevented the poor ΔEVE correspondence from being absorbed through additional parameter adjustments.
3. Use balance-sheet quantities appropriate to the institution and separately justify behavioural assumptions. Institution-level asset and liability shares can be incorporated only when the model distinguishes relevant funding instruments. A model used for wholesale-funded banks would need separate maturity and liquidity characteristics for covered bonds, senior debt, interbank liabilities and central-bank borrowing. Effective interest-rate duration should be estimated from risk

information that incorporates hedges rather than inferred solely from broad balance-sheet proportions.

4. **Treat the LLM specification as a source of measurement variation.** Record dated model versions and compare multiple model families and instruction variants where feasible. Estimate variation across repeated responses before choosing decision-discrepancy tolerances: a threshold of 0.5 percentage points is appreciably above the median Claude within-family discrepancy of 0.19 but closer to the GPT-5.5 median of 0.33. LLM responses may inform the functional form or initial values of equations, but consequential coefficients should be compared with human or bank observations when those are available.
5. **Prioritise independent estimation of consequential parameters.** In this application, many outcomes are relatively insensitive to the customer price-response coefficients, but depend substantially on assumed withdrawal responses and competitor-rate pass-through. These are therefore high-priority parameters for testing with depositor-level or supervisory transaction data rather than relying exclusively on LLM-generated answers.

References

Bibliographic details were checked against publisher, NBER, RePEc, PMLR, ACL or arXiv records on 16 October 2026.

Abadi, J., Brunnermeier, M., & Koby, Y. (2023). The reversal interest rate. *American Economic Review*, 113(8), 2084–2120. https://doi.org/10.1257/aer.20190150

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202, 337–371.

Albertazzi, U., Faber, F., Georgescu, O.-M., Gavazza, A., & Lecomte, E. (2026). Bank deposit pricing in the euro area. ECB Working Paper No. 3255.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337–351. https://doi.org/10.1017/pan.2023.2

Assenza, T., Delli Gatti, 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. https://doi.org/10.1016/j.jedc.2014.07.001

Aymanns, C., & Farmer, J. D. (2015). The dynamics of the leverage cycle. *Journal of Economic Dynamics and Control*, 50, 155–179. https://doi.org/10.1016/j.jedc.2014.09.015

Basel Committee on Banking Supervision (2013). *Basel III: The Liquidity Coverage Ratio and liquidity risk monitoring tools*. Bank for International Settlements.

Basel Committee on Banking Supervision (2016). *Standards: Interest rate risk in the banking book*. Bank for International Settlements.

Begenau, J., Piazzesi, M., & Schneider, M. (2015). Banks' risk exposures. NBER Working Paper No. 21334. https://doi.org/10.3386/w21334

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32(4), 401–416. https://doi.org/10.1017/pan.2024.5

Bookstaber, R., Paddrik, M., & Tivnan, B. (2018). An agent-based model for financial vulnerability. *Journal of Economic Interaction and Coordination*, 13, 433–466. https://doi.org/10.1007/s11403-017-0188-1

Brand, J., Israeli, A., & Ngwe, D. (2023). Using LLMs for market research. Harvard Business School Working Paper No. 23-062 (first circulated as "Using GPT for market research").

Brandts, J., & Charness, G. (2011). The strategy versus the direct-response method: A first survey of experimental comparisons. *Experimental Economics*, 14(3), 375–398. https://doi.org/10.1007/s10683-011-9272-x

Burrows, O., Learmonth, D., McKeown, J., & Williams, R. (2012). RAMSI: A top-down stress-testing model developed at the Bank of England. *Bank of England Quarterly Bulletin*, 52(3), 204–212.

Bybee, L. (2023). Surveying generative AI's economic expectations. arXiv:2305.02823.

Chen, Q., Goldstein, I., Huang, Z., & Vashishtha, R. (2024). Liquidity transformation and fragility in the U.S. banking sector. *Journal of Finance*, 79(6), 3985–4036.

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(51), e2316205120. https://doi.org/10.1073/pnas.2316205120

Cookson, J. A., Fox, C., Gil-Bazo, J., Imbet, J. F., & Schiller, C. (2023). Social media as a bank run catalyst. SSRN Working Paper No. 4422754.

Coulier, L., Pancaro, C., Pancotto, L., & Reghezza, A. (2025). Banking on assumptions? How banks model deposit maturities. ECB Working Paper No. 3140.

Dees, S., Henry, J., & Martin, R. (Eds.) (2017). *STAMP€: Stress-Test Analytics for Macroprudential Purposes in the euro area*. European Central Bank.

DeMarzo, P. M., Krishnamurthy, A., & Nagel, S. (2024). Interest rate risk in banking. NBER Working Paper No. 33308.

Di Tella, S., & Kurlat, P. (2021). Why are banks exposed to monetary policy? *American Economic Journal: Macroeconomics*, 13(4), 295–340. https://doi.org/10.1257/mac.20180379

Diamond, D. W., & Dybvig, P. H. (1983). Bank runs, deposit insurance, and liquidity. *Journal of Political Economy*, 91(3), 401–419. https://doi.org/10.1086/261155

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Drechsler, I., Savov, A., & Schnabl, P. (2017). The deposits channel of monetary policy. *Quarterly Journal of Economics*, 132(4), 1819–1876. https://doi.org/10.1093/qje/qjx019

Drechsler, I., Savov, A., & Schnabl, P. (2021). Banking on deposits: Maturity transformation without interest rate risk. *Journal of Finance*, 76(3), 1091–1143. https://doi.org/10.1111/jofi.13013

Drechsler, I., Savov, A., Schnabl, P., & Wang, O. (2026). Deposit franchise runs. *Journal of Finance*, 81(3), 1573–1617.

Egan, M., Hortaçsu, A., & Matvos, G. (2017). Deposit competition and financial fragility: Evidence from the US banking sector. *American Economic Review*, 107(1), 169–216. https://doi.org/10.1257/aer.20150342

English, W. B., Van den Heuvel, S. J., & Zakrajšek, E. (2018). Interest rate risk and bank equity valuations. *Journal of Monetary Economics*, 98, 80–97. https://doi.org/10.1016/j.jmoneco.2018.04.010

Englisch, H., Krabichler, T., Müller, K. J., & Schwarz, M. (2023). Deep treasury management for banks. *Frontiers in Artificial Intelligence*, 6, 1120297. https://doi.org/10.3389/frai.2023.1120297

European Banking Authority (2022). *Guidelines on IRRBB and CSRBB* (EBA/GL/2022/14).

Fagiolo, G., Moneta, A., & Windrum, P. (2007). A critical guide to empirical validation of agent-based models in economics: Methodologies, procedures, and open problems. *Computational Economics*, 30(3), 195–226. https://doi.org/10.1007/s10614-007-9104-4

Farmer, J. D., & Foley, D. (2009). The economy needs agent-based modelling. *Nature*, 460(7256), 685–686. https://doi.org/10.1038/460685a

Freixas, X., & Rochet, J.-C. (2008). *Microeconomics of Banking* (2nd ed.). MIT Press.

Fu, Z., Xu, R., & Ren, Q. (2026). Financial fragility in societies of LLM agents: Coordination failures and stabilizing mechanisms. arXiv:2609.30940.

Gai, P., & Kapadia, S. (2010). Contagion in financial networks. *Proceedings of the Royal Society A*, 466(2120), 2401–2423. https://doi.org/10.1098/rspa.2009.0410

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. https://doi.org/10.1057/s41599-024-03611-3

Gao, Y., Lee, D., Burtch, G., & Fazelpour, S. (2025). Take caution in using LLMs as human surrogates: Scylla ex machina. arXiv:2410.19599.

Goldstein, I., & Pauzner, A. (2005). Demand-deposit contracts and the probability of bank runs. *Journal of Finance*, 60(3), 1293–1327. https://doi.org/10.1111/j.1540-6261.2005.00762.x

Gomez, M., Landier, A., Sraer, D., & Thesmar, D. (2021). Banks' exposure to interest rate risk and the transmission of monetary policy. *Journal of Monetary Economics*, 117, 543–570. https://doi.org/10.1016/j.jmoneco.2020.03.011

Gouriéroux, C., Monfort, A., & Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, 8(S1), S85–S118. https://doi.org/10.1002/jae.3950080507

Grazzini, J., & Richiardi, M. (2015). Estimation of ergodic agent-based models by simulated minimum distance. *Journal of Economic Dynamics and Control*, 51, 148–165. https://doi.org/10.1016/j.jedc.2014.10.006

Greenwald, E., Schulhofer-Wohl, S., & Younger, J. (2023). Deposit convexity, monetary policy and financial stability. Federal Reserve Bank of Dallas Working Paper No. 2315. https://doi.org/10.24149/wp2315

Hałaj, G. (2018). Agent-based model of system-wide implications of funding risk. ECB Working Paper No. 2121.

Hannan, T. H., & Berger, A. N. (1991). The rigidity of prices: Evidence from the banking industry. *American Economic Review*, 81(4), 938–945.

Hansen, A. L., Horton, J. J., Kazinnik, S., Puzzello, D., & Zarifhonarvar, A. (2024). Simulating the Survey of Professional Forecasters. SSRN Working Paper No. 5066286.

Hoffmann, P., Langfield, S., Pierobon, F., & Vuillemey, G. (2019). Who bears interest rate risk? *Review of Financial Studies*, 32(8), 2921–2954. https://doi.org/10.1093/rfs/hhy113

Horton, J. J., Filippas, A., & Manning, B. S. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? NBER Working Paper No. 31122. https://doi.org/10.3386/w31122

Jermann, U., & Xiang, H. (2023). Dynamic banking with non-maturing deposits. *Journal of Economic Theory*, 209, 105644. https://doi.org/10.1016/j.jet.2023.105644

Jiang, E. X., Matvos, G., Piskorski, T., & Seru, A. (2024). Monetary tightening and U.S. bank fragility in 2023: Mark-to-market losses and uninsured depositor runs? *Journal of Financial Economics*, 159, 103899. https://doi.org/10.1016/j.jfineco.2024.103899

Kazinnik, S. (2026). Bank run, interrupted: Modeling deposit withdrawals with generative AI. *Management Science*, articles in advance. https://doi.org/10.1287/mnsc.2024.06807

Kazinnik, S., & Sinclair, T. M. (2025). FOMC in silico: A multi-agent system for monetary policy decision modeling. George Washington University, Center for Economic Research Working Paper 2025-005.

Klein, M. A. (1971). A theory of the banking firm. *Journal of Money, Credit and Banking*, 3(2), 205–218.

Koont, N., Santos, T., & Zingales, L. (2024). Destabilizing digital “bank walks”. NBER Working Paper No. 32601. https://doi.org/10.3386/w32601

Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4), 1281–1317. https://doi.org/10.1257/jel.20231736

Lamperti, F., Roventini, A., & Sani, A. (2018). Agent-based model calibration using machine learning surrogates. *Journal of Economic Dynamics and Control*, 90, 366–389. https://doi.org/10.1016/j.jedc.2018.03.011

Li, N., Gao, C., Li, M., Li, Y., & Liao, Q. (2024). EconAgent: Large language model-empowered agents for simulating macroeconomic activities. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15523–15536. https://doi.org/10.18653/v1/2024.acl-long.829

Lucas, R. E., Jr. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1, 19–46. https://doi.org/10.1016/S0167-2231(76)80003-6

Manning, B. S., Zhu, K., & Horton, J. J. (2024). Automated social science: Language models as scientist and subjects. NBER Working Paper No. 32381. https://doi.org/10.3386/w32381

Montagna, M., & Kok, C. (2016). Multi-layered interbank model for assessing systemic risk. ECB Working Paper No. 1944.

Monti, M. (1972). Deposit, credit and interest rate determination under alternative bank objective functions. In G. P. Szegö & K. Shell (Eds.), *Mathematical Methods in Investment and Finance* (pp. 431–454). North-Holland.

Neumark, D., & Sharpe, S. A. (1992). Market structure and the nature of price rigidity: Evidence from the market for consumer deposits. *Quarterly Journal of Economics*, 107(2), 657–680. https://doi.org/10.2307/2118485

Park, J. S., O’Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. https://doi.org/10.1145/3586183.3606763

Park, P. S., Schoenegger, P., & Zhu, C. (2024). Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, 56(6), 5754–5770. https://doi.org/10.3758/s13428-023-02307-x

Platt, D. (2020). A comparison of economic agent-based model calibration methods. *Journal of Economic Dynamics and Control*, 113, 103859. https://doi.org/10.1016/j.jedc.2020.103859

Repullo, R. (2004). Capital requirements, market power, and risk-taking in banking. *Journal of Financial Intermediation*, 13(2), 156–182.

Rice, J., & Guerrini, G. M. (2025). Why euro area banks stayed resilient amid rising rates: Lessons from the 2022–2023 monetary cycle. SUERF Policy Brief No. 1263.

Ruano, C., & Rajan, S. (2026). Social contagion and bank runs: An agent-based model with LLM depositors. arXiv:2602.15066.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202, 29971–30004.

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In H. Sauermann (Ed.), *Beiträge zur experimentellen Wirtschaftsforschung* (pp. 136–168). J.C.B. Mohr (Paul Siebeck).

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, 69(1), 99–118. https://doi.org/10.2307/1884852

Tesfatsion, L., & Judd, K. L. (Eds.) (2006). *Handbook of Computational Economics, Vol. 2: Agent-Based Computational Economics*. Elsevier/North-Holland.

Tieleman, S. (2022). Towards a validation methodology for macroeconomic agent-based models. *Computational Economics*, 60, 1507–1527. https://doi.org/10.1007/s10614-021-10191-w

von Stackelberg, H. (1934). *Marktform und Gleichgewicht*. Julius Springer.

Data and code availability

The anonymised supplementary materials accompanying this submission comprise Online Resource 1 (supplementary procedures and results) and Online Resource 2 (simulation code, LLM-agent instructions and recorded responses, numerical outputs, evaluation scripts, and processed empirical input datasets). Online Resource 2 contains the processed ECB deposit-rate panel, EBA bank-level and IRRBB panels, normalisation and robustness files, and the inputs required to reproduce the Section 4 bank-level comparisons and representative-bank profiles. The two empirical calibration analyses referred to as Paper 1 and Paper 2 are unpublished. Source-level LIQ1 observations underlying Paper 2's funding-runoff measure are not included; the archive does not independently reproduce every input of those separate empirical analyses from original supervisory disclosures. The model calibration and corresponding assumptions are described in Section 4. A public, attributed repository release is planned after anonymous review.

Use of generative AI

The research uses Anthropic Claude models and OpenAI GPT-5.5 (gpt-5.5-2026-04-23) as simulated decision-making agents, as described in Section 5. Separately, Claude (claude-opus-5-5) assisted with simulation-code development, bibliographic checking and editorial consistency, while ChatGPT (GPT-6) assisted with manuscript editing and condensation. These supporting uses are distinct from the experimental use of LLMs as agents. The

research design, computational analyses, interpretation and final text remain the responsibility of the submitting author.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OnlineResource1AnonymousSupplementaryInformation.docx