FRAUDSOCIETY: A GENERATIVE AGENT-BASED SOCIETY FOR THE FINANCIAL-FRAUD ECONOMY

Anonymous authors
Paper under double-blind review

ABSTRACT

Financial fraud caused over $15.9 billion in losses in the United States in 2025, according to the Federal Trade Commission (FTC, 2026). Understanding such large-scale harm requires jointly modeling informational and financial dynamics. Yet existing simulators typically capture only one side of this interaction, leaving a gap in modeling how fraud unfolds in the real world. To bridge this gap, we introduce FRAUDSOCIETY, the first LLM-agent-based simulator that jointly models these coupled dynamics. It integrates a dynamic social environment for information exposure and communication, persistent financial ledgers that preserve payment and loss histories, and a Need-Intent-Action architecture for civilian decision making. Fraudsters further adapt their communication and financial actions to individual interaction histories and ledger states, forming a closed feedback loop between the two dynamics. We validate FRAUDSOCIETY across macro-level outcomes, micro-level behavior, and intervention effects. At the macro level, alignment with regulator-grounded statistics persists across three seeds with newly generated civilian populations. At the micro level, it reproduces strong loss concentration, with the rounded-up top decile of victimization records accounting for 81.1% of total losses, consistent with Canadian Anti-Fraud Centre statistics (CAFC, 2025). Counter-fraud experiments further reveal variation across intervention strategies, with the oracle personalized intervention reducing the repeat-payment rate from 49.4% to 10.9%. These results establish FRAUDSOCIETY as both a high-fidelity simulator and a scalable testbed for evaluating anti-fraud strategies under diverse intervention conditions, thereby advancing research on financial fraud, its underlying behavioral mechanisms, and its prevention.

1 INTRODUCTION

Financial fraud has become a large-scale socio-economic harm: in 2025, the U.S. Federal Trade Commission (FTC) received 3 million fraud reports with losses exceeding $15.9 billion (FTC, 2026). In practice, financial fraud unfolds through tightly coupled *informational* and *financial* dynamics, with each continuously affecting the other throughout the process. Fraudulent information spreads through public content, platform recommendations, and private communication, gradually shaping a victim's attention, trust, and willingness to pay (Anderson et al., 2024; Wang, 2026). As fraudulent information spreads, victims may make an initial payment, transfer additional funds, or attempt to withdraw purported returns. These financial events, in turn, reshape subsequent communication: Fraudsters may provide small returns to build trust, solicit further transfers after an initial payment, or introduce new excuses to delay withdrawals (Anderson et al., 2024). Financial loss therefore emerges from an interaction in which informational dynamics shape financial actions and evolving financial states reshape subsequent communication (Oak & Shafiq, 2025).

Studying financial fraud therefore requires modeling how informational and financial dynamics interact at the individual level and collectively give rise to population-level outcomes. Agent-based modeling (ABM) provides a natural computational paradigm for capturing this micro-to-macro process (Epstein, 1999). Existing approaches, however, typically capture only one side of this coupling. Traditional fraud-oriented ABMs focus on *financial dynamics*, simulating payment networks and fraudulent transactions without modeling the informational processes (Oztas et al., 2023). In contrast, LLM-based ABMs focus on *informational dynamics*, supporting adaptive social behavior but rarely linking these interactions to persistent financial states and transaction processes (Yang et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

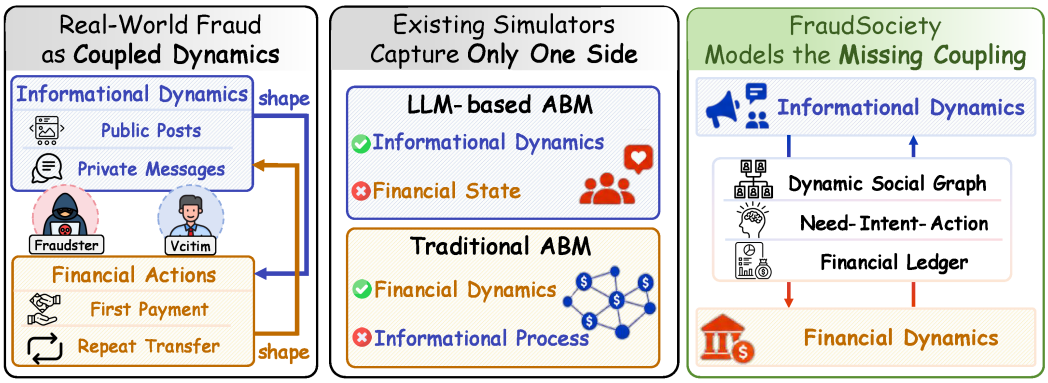


Figure 1: Existing simulators fall short of real-world fraud by capturing only informational or financial dynamics. FRAUDSOCIETY closes this gap by jointly modeling both.

2024; Piao et al., 2025). Consequently, both paradigms fail to capture this interaction, limiting their ability to model how real-world fraud unfolds over time.

We introduce FRAUDSOCIETY, an LLM-agent-based financial-fraud simulator that jointly models *informational* and *financial* dynamics. To model informational dynamics, we construct a dynamic social graph and adopt a Reddit-style ranking mechanism to govern public exposure, information diffusion, and private follow-up. To capture financial dynamics, we maintain persistent ledgers that track payments, returns, and withdrawal requests throughout each victimization process. We then couple these dynamics through a Need-Intent-Action architecture, in which each individual's evolving needs and observed context shape behavioral intents and subsequent social and financial actions. These actions and their outcomes are recorded in persistent ledgers and become part of subsequent observations, while fraudsters adapt their communication and financial actions to each victim's response history and evolving financial state. Together, these processes form a closed feedback loop between informational and financial dynamics as interactions unfold over time.

By more closely approximating the structure of real-world fraud, FRAUDSOCIETY serves as a high-fidelity simulator of the fraud process and provides a controlled testbed for counter-fraud interventions, enabling the systematic evaluation of the effects of different interventions on payment behavior and financial loss. To evaluate the fidelity and practical utility of FRAUDSOCIETY, we focus on three specific research questions spanning multiple levels: **RQ1:** At the macro level, can FRAUDSOCIETY generate aggregate loss distributions consistent with stylized facts? **RQ2:** At the micro level, is FRAUDSOCIETY able to generate realistic victimization processes? **RQ3:** From a practical perspective, can FRAUDSOCIETY support the evaluation of counter-fraud interventions?

Through our experiments, we obtain consistent results across the macro, micro, and intervention levels. At the macro level, FRAUDSOCIETY achieves a Quartet Distance of 0.3111 ± 0.0229 (mean $\pm$ sample standard deviation) across three seeds with newly generated civilian populations, providing evidence of consistent alignment with regulator-grounded stylized facts across the tested realizations. At the micro level, FRAUDSOCIETY reproduces realistic behavioral heterogeneity and loss concentration, with the rounded-up top decile of victimization records accounting for 81.1% of total loss, broadly consistent with the statistics reported by the Canadian Anti-Fraud Centre (CAFC). Finally, our counter-fraud experiments compare five intervention strategies and reveal substantial variation in their effectiveness. The oracle personalized intervention reduces the repeat-payment rate from 49.4% to 10.9%, illustrating the simulator's use for comparing counter-fraud strategies under explicit information assumptions.

Collectively, our contributions are threefold:

- We introduce FRAUDSOCIETY, the first LLM-agent-based simulator to jointly model the tightly coupled informational and financial dynamics.
- We propose a cross-level validation framework to systematically analyze both macro-level fidelity and underlying micro-level victimization mechanisms.

- We show that FRAUDSOCIETY can serve as a counter-fraud testbed, providing a scalable foundation for the design and analysis of real-world fraud prevention strategies.

2 RELATED WORK

Agent-based social simulation. Agent-based modeling (ABM) studies how heterogeneous individual behaviors and local interactions give rise to aggregate social and economic outcomes (Epstein, 1999). Traditional ABMs typically represent agents through predefined behavioral rules, enabling controlled analysis of emergence but limiting the complexity and adaptability of individual decision-making. Recent work incorporates LLM-based agents with profiles, memory, natural-language interaction, and adaptive behavior. Generative Agents (Park et al., 2023) models persistent individual routines and social interactions, while OASIS (Yang et al., 2024), AgentSociety (Piao et al., 2025), and SocioVerse (Zhang et al., 2025) extend LLM-agent simulation to large-scale social environments. Recent studies further examine behavioral heterogeneity and adaptation, such as whether heterogeneous behavioral drivers lead to theoretically consistent strategy changes in simulated financial markets (Li et al., 2026). These studies establish the foundation for LLM-based simulation across settings, but primarily target general social or market dynamics rather than financial-fraud victimization and the feedback between social interaction and financial loss.

Traditional simulation for financial fraud. Traditional financial-fraud simulators largely focus on transactions and capital flows. BankSim (Lopez-Rojas & Axelsson, 2014) generates synthetic bank-payment behavior for fraud-detection research, while AMLSim (Weber et al., 2018) and SAML-D (Oztas et al., 2023) model account interactions and laundering patterns for anti-money-laundering analysis. More recent systems such as TransXion (Chen et al., 2026) incorporate heterogeneous entities, demographic profiles, and structured transaction behaviors to construct realistic financial-crime graphs. These simulators provide persistent transaction records and controllable fraud labels, making them useful for detection and graph-learning benchmarks at larger scales. However, in these simulators, fraud-related transactions are generally generated from predefined transaction rules, without explicitly modeling the preceding information exposure, social interaction, and adaptive persuasion that influence individual payment decisions.

LLM-agent simulation for financial fraud. Recent studies use LLM agents to model fraud-related communication and interaction. ScamAgents (Badhe, 2025) examines autonomous, multi-turn scam calls with context tracking and adaptive deception strategies. Most closely related, Multi-Agent4Fraud (Ren et al., 2025) simulates collaborative financial fraud across public posts and private interactions, and studies fraud success and agent collaboration. Nevertheless, existing systems primarily assess dialogue plausibility, fraud success, or harmful-agent capability, rather than coupling these interactions with persistent financial states through which payments, returns, withdrawals, and cumulative losses feed back into subsequent behavior. They also provide limited analysis of whether heterogeneous individual victimization pathways collectively reproduce population-level loss patterns observed in real-world fraud statistics. FRAUDSOCIETY complements these studies by jointly modeling informational and financial dynamics and evaluating the resulting simulation at both the individual and aggregate levels.

3 FRAUDSOCIETY

3.1 OVERALL DESIGN OF FRAUDSOCIETY

FRAUDSOCIETY simulates fraud as a process coupling social interaction with financial consequences. As shown in Figure 2, its simulation core consists of four components: Civilian, Fraudster, Social World, and Financial World. Civilians follow a **N**eed-**I**ntent-**A**ction (NIA) cycle based on their profiles, routine states, social histories, and financial records, while fraudsters manage persistent campaigns through adaptive social and financial actions. The Social World governs public information diffusion, personalized recommendation, and private communication, whereas the Financial World executes monetary actions and maintains persistent ledgers over repeated interactions.

At each hourly tick, the environment first updates routine, economic, platform, and campaign states and constructs agent-specific observations. Civilians and fraudsters then make decisions, after which

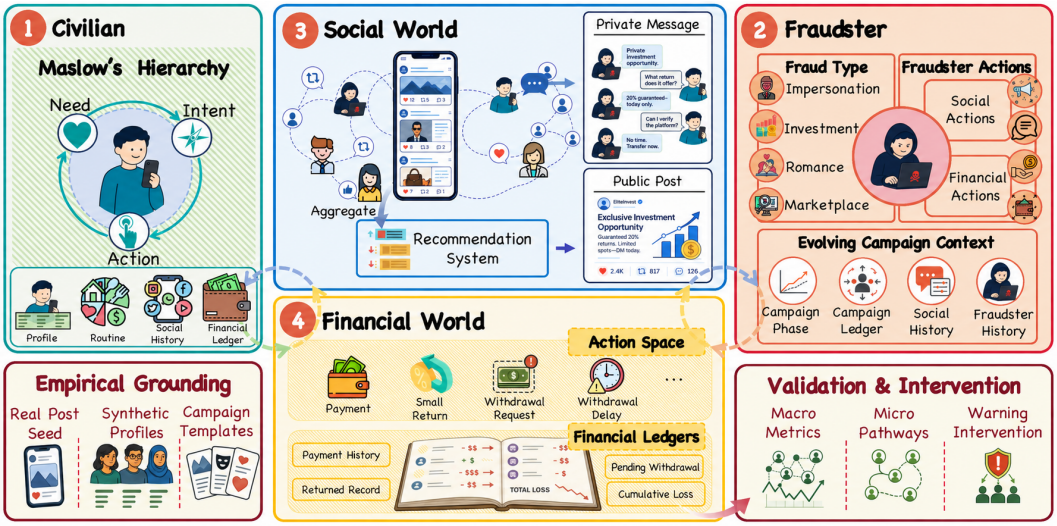


Figure 2: Overview of FRAUDSOCIETY. Civilians, guided by the Need-Intent-Action framework, act in Social and Financial Worlds, interacting with fraudsters who manage evolving campaign contexts. Both sides take social and financial actions that update the environment and shape subsequent decisions. The simulation is initialized with real-world posts, empirically grounded synthetic profiles, and campaign templates, and is validated through macro-level loss statistics, individual victimization pathways, and counter-fraud interventions.

the environment validates and executes their actions and commits the resulting social and financial state changes for subsequent decision cycles. We further ground the simulation with real-world post seeds, empirically grounded synthetic civilian profiles, and fraud-campaign templates, and evaluate the resulting dynamics under diverse scenarios through macro-level metrics, micro-level pathways, and counter-fraud interventions. Details of the simulation cycle and state construction, including observation limits and execution constraints, are provided in the Appendix.

3.2 CIVILIAN MODELING WITH NEED-INTENT-ACTION

In civilian modeling, the NIA architecture plays a central role in simulating how individuals respond to evolving social and financial environments across repeated interactions. As illustrated in Figure 2, at each decision step, a civilian updates its needs based on the observed context, forms a behavioral intent, and selects a concrete action. The resulting feedback informs subsequent NIA cycles. The Appendix provides the formal definition and hourly agent cycle.

Need. Need represents the dynamic internal state that shapes what a civilian prioritizes. Inspired by Maslow's hierarchy of needs (Maslow, 1943), civilian needs are modeled along physiological, safety, and social dimensions. As civilians engage in routine activities and receive social and economic signals, their need states evolve with their profiles and interaction histories, capturing heterogeneous priorities across civilians. These evolving needs provide the internal basis for subsequent intent formation and, in turn, shape downstream behavior. To track need dynamics and examine their micro-to-macro implications, FRAUDSOCIETY records each civilian's need state at every step.

Intent. Intent represents the behavioral objective that links an agent's current needs to executable behavior (Ajzen, 1991; Gollwitzer, 1999). In FRAUDSOCIETY, a civilian forms an intent by combining its current need state with the observed situation, selecting an activity route such as consumption, work, rest, or social interaction. The selected intent then restricts subsequent decision-making to a contextually relevant set of candidate actions, providing a structured transition from internal state to concrete behavior. As the civilian receives new information and environmental feedback, its needs may change, leading the agent to revise its intent and adapt its subsequent actions accordingly.

Action. Action represents the concrete behavior through which a civilian carries out its selected intent. Conditioned on its current intent and observed context, the civilian selects actions from the

corresponding action space. For example, a social-interaction intent may lead to ordinary social behavior, fraud-campaign communication, or fraud-related payments. Once these actions are executed, their consequences become part of subsequent observations, so that future decisions are informed by prior interactions and the current state.

3.3 FRAUDSTER MODELING

In FRAUDSOCIETY, fraudsters manage fraud campaigns with a context $\mathbf{x}$ that tracks social and financial interactions. For campaign c, decisions follow a hierarchical policy:

$$\begin{aligned} q_{c,d+1} &\sim \pi_D(\cdot \mid q_{c,d}, \kappa_c, \mathbf{x}_{c,d}), \\ (a^S_{c,i,t}, a^F_{c,i,t}) &\sim \pi_H(\cdot \mid q_{c,d}, \kappa_c, \mathbf{x}_{c,t}, \mathbf{h}_{c,i,t}), \end{aligned} \tag{1}$$

where $q_{c,d}$ is the daily campaign phase, κ_c is the fraud type, and $\mathbf{h}_{c,i,t}$ is the interaction history with civilian i. π_D and π_H are structured-prompt LLM policies over admissible phase transitions and victim-specific actions, respectively.

Campaign Type and Phase. As summarized in Figure 3, FRAUDSOCIETY models investment, impersonation, romance, and marketplace fraud based on major reporting taxonomies (FTC, 2025; FBI IC3, 2025), with templates tailored to each fraud type. Drawing on scam-lifecycle and victim-grooming studies (Lacey et al., 2020), each campaign progresses through recruiting, credibility, peak, and collapse, corresponding to exposure expansion, legitimacy building, intensified extraction, and termination. At the daily timescale, π_D selects whether to remain in the current phase or advance to the next, excluding backward and skipped transitions. The resulting phase and fraud type condition subsequent hourly decisions.

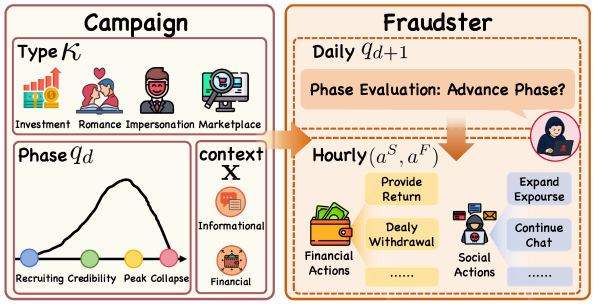


Figure 3: Fraudster modeling in FRAUDSOCIETY. Campaigns include a fraud type, evolving phase, and social-financial context, which together guide daily phase evaluation and victim-specific social and financial actions at the hourly level.

Fraudster Action. At each hour, the fraudster selects victim-specific social and financial actions from the current campaign context and interaction history. The social action a^S expands exposure, initiates contact, or continues a conversation, whereas the financial action a^F provides a small return, handles a withdrawal request, or requests further payment. Equation (1) therefore links daily campaign planning with victim-specific hourly adaptation.

3.4 SOCIAL AND ECONOMIC ENVIRONMENT

To connect micro-level decisions with macro-level fraud outcomes, FRAUDSOCIETY couples a dynamic social platform with an economic environment and persistent financial ledgers. The social platform governs public exposure, social amplification, and private communication, while the economic environment records the consequences of civilian and fraudster actions. This design captures how exposure timing, civilian circumstances, and fraudster responses jointly produce persistent financial feedback that informs later agent decisions.

Social Platform and Exposure. The social environment is represented as a typed dynamic graph $G_t = (V, E_t, \rho)$, where V contains civilians and fraudsters, E_t contains social edges at tick t, and $\rho : E_t \to \mathcal{R}$, where $\mathcal{R}$ is the set of relation types, such as family, colleague, neighbor, and online contact. Edges are initialized from civilian profiles, geographic proximity, and existing relations, and evolve through following, posting, reposting, commenting, liking, and direct messaging. Fraud exposure occurs through two channels: public diffusion through feeds and social interactions, and private follow-up through direct messages.

The platform constructs each civilian's feed using two-stage ranking. First, public posts are ranked by a Reddit-style hot score. For post p created at t_p, let l_p and d_p denote its numbers of likes and dislikes, respectively. We define the net vote as $v_p = l_p - d_p$. The corresponding hot score is

$$h(p) = \text{sign}(v_p) \log_{10}(\max(|v_p|, 1)) + \frac{t_p - t_0}{\tau_r}, \tag{2}$$

where t_0 is a reference timestamp and τ_r is the recency time scale, both in the units of t_p.

Globally ranked posts are combined with content from followed users and previously interacted authors to form a candidate pool. Candidates are reranked using the hot score, follow and author-affinity indicators, and embedding similarity to recently interacted content; the highest-ranked posts enter the visible feed. Public exposure thus depends on popularity, social connections, and interaction history, whereas private exposure depends on fraudster follow-up and prior civilian responses.

Economic Environment and Ledgers. The economic environment updates each civilian's daily activities and financial state over time. At each hourly tick, civilians engage in routine activities such as work, rest, and consumption, while income and expenditure events update their available balance. These evolving routine and financial states become part of the civilian's observation and provide the context for the subsequent Need-Intent-Action cycle.

FRAUDSOCIETY maintains separate financial ledgers for civilians and fraud campaigns. Each civilian ledger records the individual's evolving financial state, while each campaign ledger records campaign-related financial interactions, including payments, returns, withdrawal requests, delays, and follow-up payment requests. The civilian-visible ledger states become part of subsequent observations, allowing later decisions to depend on current financial conditions as well as prior payments, received returns, unrecovered losses, and withdrawal experiences. Formal observation construction and ledger details are provided in the Appendix.

4 EXPERIMENTS

Our experiments examine aggregate fidelity (**RQ1**), individual victimization pathways and mechanisms (**RQ2**), and the effects of counter-fraud warnings on payment behavior and losses (**RQ3**).

We instantiate FRAUDSOCIETY with 112 civilians, 8 fraudsters, and 8 active campaigns over 480 hourly steps, using Qwen3-14B for all agent decisions. Full-model macro-level validation uses three random seeds, each with newly generated civilian profiles. Each mechanism ablation is repeated three times with the reference civilian profiles held fixed. Intervention conditions share the reference population, campaign initialization, and seed; micro-level pathways use the no-warning reference run. Further details are provided in the Appendix.

4.1 MACRO-LEVEL VALIDATION

Evaluation Metrics. To evaluate macro-level realism, we introduce the Quartet Distance (QD), a regulator-grounded audit score cross-referencing simulated fraud losses against four empirical FTC stylized facts: Heavy-Tailed Loss concentration (HTL), Age-stratified Loss Severity profiles (ALS), Fraud-family Severity Hierarchy (FSH), and Age-specific Loss-Report shares (ALR). The composite score averages the available component distances:

$$\text{QD} = \frac{1}{N} \left(d_{\text{HTL}} + d_{\text{ALS}} + d_{\text{FSH}} + d_{\text{ALR}} \right),$$

where components unsupported by a source system are omitted, and N is the number of distances included in the average under the rules in Section J.

Each component distance captures a distinct aspect of the simulated loss distribution. d_{HTL} measures whether the simulation produces a comparable high-loss tail by comparing the simulated share of high-loss victims with the FTC target after rescaling the empirical high-loss threshold to the simulation scale. d_{ALS} compares the median cumulative loss profile across seven adult age bins, auditing whether age-stratified loss severity follows the FTC pattern. d_{FSH} evaluates whether the relative severity hierarchy between fraud families is preserved, using the investment-to-impersonation median-loss ratio as the primary cross-family severity contrast. d_{ALR} compares the age distribution

Table 1: Simulator capabilities. ✓ denotes support; ✗ denotes absence.

System	Demographic profile	Social exposure	Private persuasion	Financial state	Repeated payment	Withdrawal feedback
TransXion	✓	✗	✗	✓	✓	✗
MultiAgent4Fraud	✓	✓	✓	✗	✗	✗
FRAUDSOCIETY	✓	✓	✓	✓	✓	✓

Table 2: Regulator-grounded fidelity comparison (lower is better). Full-model values are mean ± sample SD over three seeds with regenerated populations; baselines are single runs. TransXion averages its three available components. Metric definitions and per-run results appear in Section J.

System	HTL↓	ALS↓	FSH↓	ALR↓	QD↓
TransXion	0.3722	0.9168	–	0.1793	0.4894
MultiAgent4Fraud	0.6211	0.6586	0.6017	0.5243	0.6014
FRAUDSOCIETY	**0.0995** ± 0.0191	**0.6556** ± 0.0059	**0.3236** ± 0.0403	**0.1657** ± 0.0267	**0.3111** ± 0.0229

of positive-loss victims with the FTC age-specific loss-report shares. Together, these four component distances audit whether the simulated fraud economy reproduces the coarse shape of regulator-observed victimization. We provide the full component definitions in the Appendix.

Baseline Comparison. Table 1 compares FRAUDSOCIETY with **TransXion** (Chen et al., 2026), a profile-aware transaction simulator, and **MultiAgent4Fraud** (Ren et al., 2025), a multi-agent financial-fraud simulator. TransXion models profile-conditioned transactions without the preceding informational process; MultiAgent4Fraud models that process but ends at the first fraud transfer. FRAUDSOCIETY jointly models informational and financial dynamics throughout victimization.

Table 2 reports the regulator-grounded fidelity results across four stylized dimensions. Across three seeds with newly generated civilian populations, FRAUDSOCIETY achieves a QD of 0.3111 ± 0.0229 (mean ± sample standard deviation), ranging from 0.2978 to 0.3376. Its mean component distances are lower than the reported baseline values where available, with ALS close to MultiAgent4Fraud. These results provide initial evidence that macro-level alignment persists across the tested seeds and population realizations; per-run results appear in Table 20. Explicit fraud-family labels also enable a severity comparison unavailable in TransXion.

4.2 MICRO-LEVEL ANALYSIS

Pathway Formation and Loss Escalation. FRAUDSOCIETY exhibits a pathway-escalation pattern in which aggregate losses emerge through exposure, private contact, and repeated fraud transfers, consistent with cyber-victimization perspectives (Cohen & Felson, 1979; Leukfeldt & Yar, 2016). All 63 positive-loss victims experience campaign contact and private-DM exposure before their first transfer. Across positive-net-loss records, the median pre-payment contact count is 7 and the median first-contact-to-first-payment lag is 72 hours, consistent with evidence that fraud develops through sustained exposure and contact (Wang, 2026).

As shown in Figure 4(a), the cumulative loss curve lies above the equal-concentration baseline, indicating strong loss concentration. The rounded-up top decile accounts for 81.1% of total loss, broadly consistent with the 75.2% top-10% share reported by CAFC (CAFC, 2025). This concentration is closely associated with repeated transfers: records with at least five transfers account for only 19.8% of records but explain 75.6% of total loss, while those with at least ten transfers still explain 70.4%. These results show that FRAUDSOCIETY reproduces both real-world loss concentration and its association with repeated transfers (Xiao et al., 2025).

Individual Features in Fraud Victimization. We examine whether victimization outcomes vary with civilians' need states and economic conditions. Table 3(a) shows higher mean need urgency during fraud-transfer hours: safety urgency is 0.710 versus 0.598 overall, and social urgency is 0.712 versus 0.630. Transfer rates are 0.94 versus 1.94 in low- and high-safety-need states, and 1.03

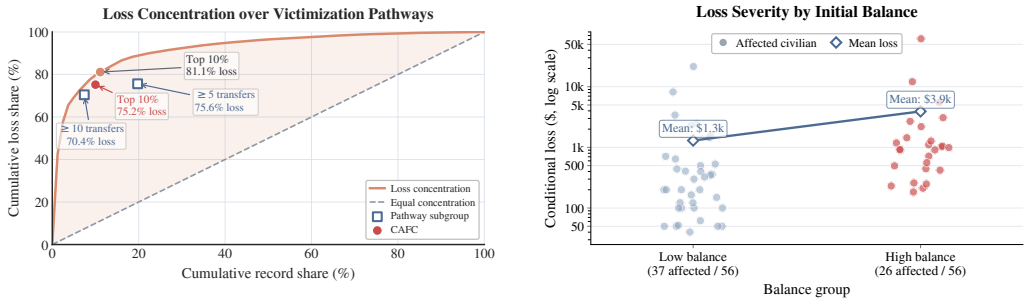
Table 3: Micro-level mechanism results. (a) Need-state associations in the reference no-warning run. (b) QD ablations with fixed reference profiles: the full model is the reference run; ablations report mean ± sample standard deviation over three runs. LF denotes ledger feedback. Lower QD indicates closer aggregate alignment with the FTC reference statistics.

(a) Need-state association

Need	Need coordinate All → transfer	Transfers / 100 h Low → high
Safety	0.598 → 0.710	0.94 → 1.94
Social	0.630 → 0.712	1.03 → 1.91

(b) Mechanism ablation

Variant	HTL↓	ALS↓	FSH↓	ALR↓	QD↓
FRAUDSOCIETY	0.089	0.651	0.301	0.150	0.298
w/o NIA	0.983 ± 0.017	0.655 ± 0.016	0.793 ± 0.019	0.189 ± 0.014	0.655 ± 0.003
w/o LF	0.697 ± 0.009	0.578 ± 0.011	0.336 ± 0.011	0.199 ± 0.008	0.452 ± 0.007



(a) Pathway loss concentration

(b) Loss severity by initial balance

Figure 4: Micro-level victimization outcomes. (a) Loss concentration across victimization pathways; pathway subgroups show concentration from repeated transfers, while the CAFC provides a real-world reference. (b) Loss severity across civilians with different initial balances; points show affected civilians, and markers show each balance group's mean loss.

versus 1.91 in low- and high-social-need states. These conditional associations are consistent with the state-dependent vulnerability in prior literature, where situational insecurity and unmet social needs can increase susceptibility to fraud-related decision making (Chanenson et al., 2026).

Economic conditions further shape victimization risk and loss severity. As shown in Figure 4(b), low-balance civilians are more likely to incur positive losses, consistent with greater vulnerability under financial scarcity (Shah et al., 2012). Among affected civilians, high-balance civilians incur larger conditional losses, consistent with routine-activity perspectives on higher-value targets (Cohen & Felson, 1979). These associations reflect theoretically grounded heterogeneity in victimization risk and subsequent loss severity.

Mechanism Ablation. We test whether the micro-mechanisms contribute to macro-level realism. Table 3(b) compares the reference full model with three repetitions of each ablation on the same profiles. The NIA ablation replaces need-conditioned intent selection with uniform sampling over feasible routes, retaining need updates. The ledger-feedback ablation masks cumulative payment, return, and withdrawal-feedback fields in civilian observations; backend ledgers remain active for transaction execution and final outcome measurement.

Both ablations yield higher QD than the reference full model (0.2978) in every repetition: mean QD is 0.6547 ± 0.0031 without NIA and 0.4523 ± 0.0067 without ledger feedback. These results support the contribution of both mechanisms to aggregate alignment under the fixed reference population. Per-run results and the distinction from the full-model population-resampling experiment are provided in Table 21 and its accompanying discussion.

The component distances show distinct effects. Ablating need-conditioned intent selection raises HTL from 0.089 to 0.983 and FSH from 0.301 to 0.793, indicating weaker alignment in loss concentration and relative fraud-family severity. Removing ledger feedback raises HTL to 0.697, while ALS improves from 0.651 to 0.578. Thus, higher QD does not imply uniformly worse alignment; the componentwise contrasts describe each ablation under the fixed reference population.

4.3 COUNTER-FRAUD INTERVENTION ANALYSIS

We compare five warning strategies with the no-warning reference (**C0**): Platform broadcasts (**C1**), Monitor labels (**C2**), Random warning comments (**C3**), Peer warnings (**C4**), and Oracle personalized warnings (**C5**). Figure 5 reports losses, payments, and repeat-payment and withdrawal-attempt rates. All five strategies yield lower total loss in these matched runs.

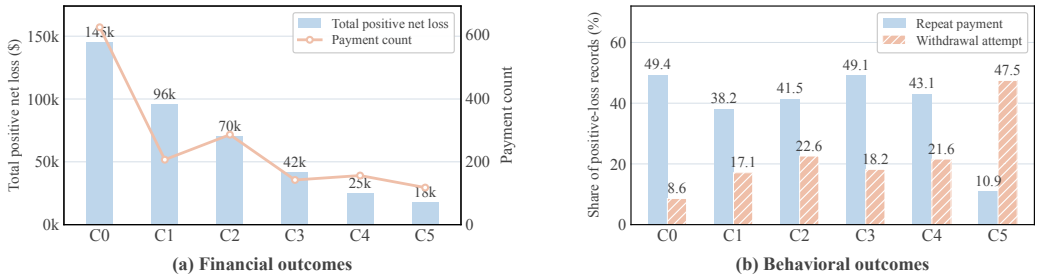


Figure 5: Intervention outcomes. Payment counts and both rates use positive-net-loss victim–campaign records for each intervention condition (Section G).

Platform-broadcast warnings. A fixed anti-fraud reminder is sent to all civilians daily. Relative to C0, payments fall from 623 to 205, loss from $145K to about $96K, and the repeat-payment rate from 49.4% to 38.2%, while the withdrawal-attempt rate rises from 8.6% to 17.1%. These outcomes are consistent with reduced payment activity despite substantial residual loss.

Monitor labels. Warning labels mark suspicious public posts. Payments fall from 623 to 284 and total loss declines, but the repeat-payment rate remains high. This pattern is consistent with labels mainly affecting public exposure: they appear on posts and do not accompany subsequent private exchanges or payments between civilians and fraudsters.

Random warning comments. Warnings inserted under posts regardless of suspiciousness accompany a fall in payments from 623 to 141 and total loss to $42,014, while the repeat-payment rate remains nearly unchanged at 49.1%. This pattern is consistent with reduced entry into payment pathways, but does not establish how warning specificity affects repetition.

Peer warnings. Civilian prompts encourage warning-oriented actions such as commenting on suspicious posts. Positive-loss victims fall from 63 to 44, total loss to $24,752, and maximum positive loss to $3,800. These lower losses are consistent with curtailed high-loss pathways when warnings circulate through ordinary civilian interactions.

Oracle personalized warnings. This oracle intervention delivers an LLM-personalized warning after each executed fraud payment, beginning with the first. Payments fall from 623 to 117 and the repeat-payment rate from 49.4% to 10.9%, while the withdrawal-attempt rate rises to 47.5%. These rates are consistent with stronger post-payment interruption and recovery seeking under oracle warnings informed by the current payment and campaign state.

Section G details the interpretation limits of these controlled comparisons.

5 CONCLUSION

We introduce FRAUDSOCIETY, the first LLM-agent-based simulator to jointly model the informational and financial dynamics of fraud. It couples adaptive social interaction and persistent ledgers through Need-Intent-Action decision making, so exposure shapes financial actions and financial outcomes reshape later communication. Macro-level loss alignment persists across three seeds with regenerated civilian populations. Reference-run analyses reproduce micro-level victimization pathways, while three repetitions of each ablation with fixed civilian profiles support the contribution of NIA and ledger feedback to aggregate alignment. The simulator also enables controlled comparisons of counter-fraud interventions and their downstream effects. Together, these capabilities make FRAUDSOCIETY a scalable foundation for realistic financial-fraud simulation and systematic counter-fraud research under explicit behavioral and information assumptions.

AI USE STATEMENT

Generative AI assistants were used to support manuscript planning, literature retrieval, mathematical exposition, English drafting and editing, coding and LaTeX assistance, methodological feedback, interpretation of results, preparation of tables and plots from recorded experimental results, and figure editing. All AI-assisted content, code, and analyses were reviewed, verified, and revised by the authors before inclusion in the manuscript. As part of the experimental pipeline, Qwen3-14B generated civilian and fraudster decisions, social interaction content, and associated decision traces with model settings and selected runtime prompt templates documented in Sections B and N. Executed financial actions were recorded in the simulator's ledgers, from which the experimental metrics were computed. The authors take responsibility for the final content of this work, including all AI-assisted text, code, analyses, and experimental artifacts.

ETHICS STATEMENT

Research purpose. FRAUDSOCIETY is developed to study the nature and progression of financial fraud, including how information exposure, social interaction, and financial feedback shape victimization. The aim is to improve understanding of these processes and support research on counter-fraud interventions. Fraudster agents serve as experimental representations of harmful behavior so that its mechanisms and possible points of interruption can be examined in a simulated environment.

Data provenance and victim privacy. This study does not use any real fraud victim's personal information, private correspondence, account records, or transaction histories. Civilian profiles are constructed from population-level statistics and configured demographic, economic, and behavioral attributes, as described in Section B. ACS/SCF statistics inform population construction, and FTC/-CAFC aggregate statistics provide empirical reference points. Public online-post seeds supply topical context. The interaction histories, payments, returns, withdrawal requests, and loss outcomes analyzed in the experiments are generated within the simulator; the reported victimization pathways do not reconstruct identifiable real victims' experiences.

Dual-use risks and responsible dissemination. Fraud-agent prompts and simulated interaction traces could be repurposed to generate deceptive content. We include selected runtime templates to document the experimental mechanisms, and present them with their research purpose, simulation context, and limitations. Their intended use is controlled scientific analysis and defensive evaluation. We discourage their use to deceive real people, solicit real funds, or develop operational fraud campaigns.

REPRODUCIBILITY STATEMENT

The simulation setup, population construction, inference settings, and intervention and ablation configurations are documented in Section B. Sections C and F specify the hourly update order and observation and ledger schemas. Section J defines Quartet Distance and reports the full-model and ablation results for individual runs, while Section G specifies intervention outcome accounting. Runtime configurations and artifact schemas appear in Section L, selected runtime prompt templates in Section N, and software dependencies and contribution boundaries in Section M.

REFERENCES

Icek Ajzen. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211, 1991.

Matthew Anderson, Evita March, Lesley Land, and Christine Boshuijzen-van Burken. Exploring the roles played by trust and technology in the online investment fraud victimisation process. *Journal of Criminology*, 57(4):488–514, 2024.

Sanket Badhe. Scamagents: How ai agents can simulate human-level scam calls. *arXiv preprint arXiv:2508.06457*, 2025.

Board of Governors of the Federal Reserve System. 2022 survey of consumer finances. https://www.federalreserve.gov/econres/scfindex.htm, 2023. Accessed: 2026-07-06.

CAFC. CAFC 2024 Annual Statistical Report. Technical report, Royal Canadian Mounted Police, October 2025. URL https://open.canada.ca/data/dataset/69c68f22-8a2a-43d1-8f4e-4017e3ffebba/resource/1ffc4a5e-e41b-48c7-929e-633d023ef8dc/download/cafc-annual-stats-report-2024-en.pdf. Accessed: 2026-07-24.

Jake Chanenson, Tara Matthews, Sunny Consolvo, Patrick Gage Kelley, Jessica McClearn, Sarah Meiklejohn, Abhishek Roy, Renee Shelby, Kurt Thomas, and Amelia Hassoun. " it didn't feel right but i needed a job so desperately": Understanding people's emotions and help needs during scams. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1–22, 2026.

Keyang Chen, Mingxuan Jiang, Yongsheng Zhao, Zeping Li, Zaiyuan Chen, Weiqi Luo, Zhixin Li, Sen Liu, Yinan Jing, Guangnan Ye, et al. Transxion: A high-fidelity graph benchmark for realistic anti-money laundering. *arXiv preprint arXiv:2604.17420*, 2026.

Lawrence E Cohen and Marcus Felson. Social change and crime rate trends: A routine activity approach. *American sociological review*, pp. 588–608, 1979.

Joshua M Epstein. Agent-based computational models and generative social science. *Complexity*, 4 (5):41–60, 1999.

FBI IC3. 2024 internet crime report. Annual report, Federal Bureau of Investigation, Internet Crime Complaint Center, 2025. URL https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf. Accessed: 2026-06-29.

FTC. Consumer sentinel network data book 2024. Data book, Federal Trade Commission, 2025. URL https://www.ftc.gov/reports/consumer-sentinel-network-data-book-2024. Accessed: 2026-06-29.

FTC. Prepared statement of the federal trade commission on the rising scam economy: Modernizing federal approaches to protect americans from foreign fraudsters. Testimony before the u.s. congress joint economic committee, Federal Trade Commission, Washington, DC, March 2026. URL https://www.ftc.gov/system/files/ftc_gov/pdf/ftc-testimony-jec-hearing-on-the-rising-scam-economy.pdf.

Peter M Gollwitzer. Implementation intentions: Strong effects of simple plans. *American psychologist*, 54(7):493, 1999.

David Lacey, Sigi Goode, Jerry Pawada, and Dennis Gibson. The application of scam compliance models to investment fraud offending. *Journal of Criminological Research, Policy and Practice*, 6(1):65–81, 2020.

Eric Rutger Leukfeldt and Majid Yar. Applying routine activity theory to cybercrime: A theoretical and empirical analysis. *Deviant Behavior*, 37(3):263–280, 2016.

Zeping Li, Guancheng Wan, Keyang Chen, Yu Chen, Yiwen Zhao, Philip Torr, Guangnan Ye, Zhenfei Yin, and Hongfeng Chai. Behavioral consistency validation for llm agents: An analysis of trading-style switching through stock-market simulation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 40356–40370, 2026.

Edgar Alonso Lopez-Rojas and Stefan Axelsson. Banksim: A bank payments simulator for fraud detection research. In *26th European Modeling and Simulation Symposium, EMSS*, volume 2014, 2014.

Abraham Harold Maslow. A theory of human motivation. *Psychological review*, 50(4):370, 1943.

Rajvardhan Oak and Zubair Shafiq. " hello, is this anna?": Unpacking the lifecycle of {Pig-Butchering} scams. In *Twenty-First Symposium on Usable Privacy and Security (SOUPS 2025)*, pp. 1–18, 2025.

Berkan Oztas, Deniz Cetinkaya, Festus Adedoyin, Marcin Budka, Huseyin Dogan, and Gokhan Aksu. Enhancing anti-money laundering: Development of a synthetic transaction monitoring dataset. In *2023 IEEE International Conference on e-Business Engineering (ICEBE)*, pp. 47–54. IEEE, 2023.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023. URL https://arxiv.org/abs/2304.03442.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society, 2025.

Qibing Ren, Zhijie Zheng, Jiaxuan Guo, Junchi Yan, Lizhuang Ma, and Jing Shao. When ai agents collude online: Financial fraud risks by collaborative llm agents on social platforms. *arXiv preprint arXiv:2511.06448*, 2025.

Anuj K Shah, Sendhil Mullainathan, and Eldar Shafir. Some consequences of having too little. *Science*, 338(6107):682–685, 2012.

U.S. Census Bureau. Acs 2022 1-year subject table s0101: Age by sex. https://data.census.gov/table/ACSST1Y2022.S0101, 2022. Accessed: 2026-07-06.

Fangzhou Wang. A scoping review of online romance scams: Conceptual construction and comparative analysis. *International Review of Victimology*, pp. 02697580251414141, 2026.

Mark Weber, Jie Chen, Toyotaro Suzumura, Aldo Pareja, Tengfei Ma, Hiroki Kanezashi, Tim Kaler, Charles E Leiserson, and Tao B Schardl. Scalable graph learning for anti-money laundering: A first look. *arXiv preprint arXiv:1812.00076*, 2018.

Jingjia Xiao, Qing Xiao, and Hong Shen. " it felt real" victim perspectives on platform design and longer-running scams. *arXiv preprint arXiv:2510.02680*, 2025.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, et al. Oasis: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*, 2024.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, et al. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users. *arXiv preprint arXiv:2504.10157*, 2025.

Main-paper reference	Appendix detail	Location
Overall design	Global hourly ordering, closed-loop state commit, and operator/civilian decision timing	Sections B and C
Software lineage	Reused infrastructure, independently implemented models, and baseline relationship	Section M
Fraudster agents	Campaign lifecycle, public outreach, active-dialogue follow-up, financial feedback, and execution gates	Section D
Civilian NIA model	Formal need update, active-need selection, intent routing, action generation, and the NIA ablation	Section E
Social-economic environment	Bounded observation construction, social exposure, financial ledger transitions, and surfaced-state ablation	Sections B and F
Counter-fraud evaluation	Exact intervention triggers, visible channels, information assumptions, and matched outcome accounting	Sections B and G
Micro-level validation	Victim–campaign accounting, contact-to-payment latency, escalation depth, and family-level pathways	Section H
Quartet Distance	Statistical units, component definitions, full-model and ablation repetitions, and baseline alignment	Sections G, J, and K
Reproducibility	Runtime configuration, condition switches, and artifact-level provenance	Section L
Runtime policies	Selected English templates: one shared-environment, eight civilian, and one operator lifecycle prompt, with output contracts	Section N

Table 4: Correspondence between topics deferred from the main paper and their detailed treatment in this appendix.

A Guide to the Appendix

This appendix documents the implementation and evaluation details deferred from the main paper. Table 4 links each deferred topic to its detailed treatment below. Subsequent sections provide formal definitions, additional diagnostics, and selected English runtime prompt templates that document the reported mechanisms.

B Implementation and Experimental Setting

Population and temporal setup. Each simulation advances in hourly ticks, and every reported run comprises 480 steps over 20 simulated days. We fix a common start time of 08:00 on March 11, 2025 for C0–C5 and all QD runs. Full-model QD validation uses three random seeds, with civilian profiles regenerated for each seed. Each mechanism ablation is repeated three times using the same reference civilian profiles. C0–C5 share the reference population and seed, controlling population composition, stochastic initialization, and calendar/payday timing within the intervention comparison. Each population contains 112 civilian profiles and 8 criminal-operator profiles. In this appendix, *operator* refers to a fraudster agent responsible for one active campaign.

LLM backbone. All LLM-controlled decisions in the reported experiments use Qwen3-14B as the backbone model. The same backbone is used for the no-warning baseline, counter-fraud interventions, and ablation settings; consequently, comparisons do not conflate the experimental mechanisms with changes in model family. Runtime inference uses temperature 0 and task-specific output-token limits. Most structured decisions use JSON; intent routing returns a single option letter (A, B, or

Age bin	ACS adult target	Realized civilians	Realized share
20–29	17.2%	18	16.1%
30–39	17.9%	16	14.3%
40–49	16.5%	12	10.7%
50–59	16.5%	32	28.6%
60–69	15.8%	14	12.5%
70–79	11.2%	10	8.9%
80+	5.0%	10	8.9%

Table 5: Target and realized age composition of the 112 civilians in the reference population. The ACS column reports the adult-population sampling target; the remaining columns describe the population reused for reference-run analyses, ablations, and interventions. The two additional full-model QD runs use newly generated profiles.

C). The parser removes reasoning tags, repairs recoverable syntax errors, validates action names and arguments against the active schema, and rejects actions that fail execution-time grounding constraints. The simulator seed fixes profile, graph, and intervention sampling.

Grounded civilian profiles and routine context. Each civilian profile specifies demographic attributes, age, education, occupation, home and work areas of interest, initial balance, income, salary day, consumption tendency, working hours, and Big-Five personality traits. These fields define the economic, mobility, and behavioral context of the hourly agent process. The environment uses an urban-scale spatial substrate. Home and work areas are taken from the profile pool, and points of interest within a 2-km radius provide routine consumption alternatives. Salaries, fixed expenses, point-of-interest purchases, and fraud payments are recorded in the economy ledger; fraud payments are marked as illegal transactions.

Profile generation and U.S. age alignment. Civilian profiles are generated separately for each of the three full-model QD seeds. The reference population is held fixed across all three repetitions of each mechanism ablation and across intervention conditions. The profile generator preserves stable simulation identifiers, names, communities, and map-linked home and work areas while resampling demographic and behavioral variables. The adult age sampler follows the U.S. Census American Community Survey (ACS) 2022 S0101 age-by-sex table, normalized over adults aged 20 years and older (U.S. Census Bureau, 2022). The resulting target bins are 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, and 80+. Table 5 reports the target sampling weights and the reference realization of 112 civilians. Holding this reference population constant controls population composition in the ablation and intervention analyses, while the three full-model QD runs assess variation across separately generated populations.

Big-Five behavioral grounding. Each civilian is assigned scores for openness, conscientiousness, extraversion, agreeableness, and neuroticism on a 1–10 scale. The generator samples these dimensions independently of age from clipped normal distributions. Openness, conscientiousness, extraversion, and agreeableness have mean 5.5 and standard deviation 1.8; neuroticism has mean 5.0 and standard deviation 1.8. These scores are operationalized as behavioral covariates that condition social responses, warning uptake, and fraud-payment decisions. Table 6 summarizes the construction and behavioral interpretation of each dimension.

Financial and online-post grounding. Initial balances are sampled from age-specific transaction-account anchors derived from the Federal Reserve 2022 Survey of Consumer Finances (Board of Governors of the Federal Reserve System, 2023), with lognormal dispersion and an occupation multiplier. Occupation, salary, work schedule, consumption tendency, and retirement status then determine routine affordability and daily mobility. Public social-media context is initialized from two online-post seed pools, each containing 280 short posts. These seeds cover benign finance, cybersecurity, marketplace, work, charity, and relationship themes, as well as suspicious lures. During the simulation, criminal operators generate new public posts and private messages conditioned on fraud family, campaign phase, recent engagement, and visible user signals. The seed pools therefore provide topical grounding rather than fixed scripts.

Trait	Sampling target	Realized mean	Behavioral interpretation in civilian decisions
Openness	5.5 ± 1.8	5.46	Higher values increase curiosity toward unfamiliar products, new platforms, and novel earning opportunities.
Conscientiousness	5.5 ± 1.8	5.54	Higher values increase verification, delayed commitment, and caution before major payments.
Extraversion	5.5 ± 1.8	5.41	Higher values increase social sharing, public engagement, and sensitivity to peer activity.
Agreeableness	5.5 ± 1.8	5.17	Higher values increase trust, compliance pressure, and difficulty refusing friend-like or authority-like requests.
Neuroticism	5.0 ± 1.8	5.10	Higher values increase sensitivity to urgency, loss avoidance, and fear-of-missing-out cues.

Table 6: Big-Five sampling distributions and behavioral interpretations. Scores are sampled on a 1–10 scale and clipped to the valid range; realized means describe the 112 civilians in the reference population. The same sampling distributions are used for the additional full-model QD populations.

Fraud campaign initialization. The simulator instantiates four fraud-family templates corresponding to the families examined in the main paper: investment, impersonation, marketplace, and romance. In the eight-operator setting, two operators are assigned to each family. Each operator is associated with one active campaign and a set of collection merchant accounts. Campaigns progress through four lifecycle phases: recruiting, credibility building, peak extraction, and collapse. In LLM-controlled mode, the operator selects phase transitions, while the campaign state records lifecycle status and the state changes associated with terminal collapse.

Social platform and social graph. The social layer uses an OASIS-style platform interface supporting public posts, comments, likes, follows, reposts, and direct messages. Each feed view contains at most 10 public posts and 20 unread private messages. The recommender first computes the Reddit-style hot score in Equation (2) of the main paper and constructs a candidate set from globally high-ranked posts and posts authored by followed or positively engaged users. Candidates are then reranked using the hot score, follow and author-affinity indicators, and nonnegative embedding similarity to content with which the civilian recently interacted. If the embedding model is unavailable, reranking uses only the non-embedding features.

The initial graph is a hybrid small-world network rather than a uniformly random graph. Construction begins with the $k = 6$ geographically nearest civilian neighbors, after which candidate edges are independently rewired with probability $p = 0.15$. Within-community ties may be strengthened by occupation, salary, or education homophily, and additional online-contact edges are sampled by degree-preferential attachment. Each operator is connected to as many as 10 degree-weighted civilian targets, placing campaign exposure within the same graph as ordinary social interaction. Edges are labeled as family, colleague, neighbor, or online contact and are used by the social-proof, following, forwarding, and peer-warning mechanisms.

Counter-fraud conditions. Table 7 summarizes the intended controlled design, in which the six conditions vary in warning channel and trigger. All six conditions share the same March 11 start, population realization, simulator seed, and 480-hour horizon.

Run artifacts and comparative interpretation. Each run records social interactions, economic transactions, victim–campaign states, QD statistics, and decision traces in artifacts linked by stable identifiers. Full-model QD results summarize three seeds and separately generated civilian populations with a common March 11 start and 480-hour horizon. Each mechanism ablation has three runs with the reference civilian profiles held fixed, summarized in Table 21. The full-model comparator for these fixed-profile ablations is the single reference QD run. Intervention conditions share the reference population and seed and remain single-run descriptive comparisons. Table 8 identifies the two information-to-decision links removed by the ablations.

Condition	Channel, trigger, and intended mechanism
C0	**No warning.** Baseline exposure and persuasion dynamics.
C1	**Platform private warning.** Broadcast every 24 simulated hours; broad awareness with low personalization.
C2	**Public monitor annotation.** Newly created posts are screened at the exposure surface.
C3	**Random peer-style comment.** Six civilians (approximately 5% of the population) are sampled each step, producing sparse socially distributed warnings.
C4	**Endogenous peer warning.** Civilian planning can interrupt persuasion through peer communication.
C5	**Oracle personalized warning.** A transaction-specific private warning follows each executed fraud payment; the first-payment message is explicitly distinguished.

Table 7: Counter-fraud intervention conditions in the controlled design. All conditions use the common March 11 clock. C5 is interpreted as an oracle upper bound.

Variant	Mechanism state and effect
Full model	**NIA enabled; ledger feedback enabled.** Need-conditioned routing uses visible cumulative financial fields.
NIA ablation	**Need-conditioned intent selection disabled; ledger feedback enabled.** Need updates remain logged, but the route is sampled uniformly from the feasible set.
Ledger-feedback ablation	**NIA enabled; ledger feedback disabled.** Transfers remain executable and included in evaluation, but cumulative financial fields are withheld from civilian prompts.

Table 8: Mechanism ablation settings. Each ablation is repeated three times with the reference civilian profiles held fixed; campaign, platform, economic, and model settings are unchanged.

COUNTER-FRAUD INTERVENTION REALIZATION

Interventions are realized exclusively through information that can become visible in a civilian's bounded social observation. With the exception of C5, condition membership is never supplied to the civilian planner as a hidden label. A warning affects behavior only if it appears as an unread private message, an appended post label, a visible comment, or a civilian-generated social action and is subsequently selected into the bounded view. Table 9 specifies the exact trigger and information assumption for each condition.

Mutation order and channel separation. C1 and C3 insert their social records after the six environment spaces have advanced and before civilian steps begin. C2 is instead a post-creation hook: the submitted post is first committed and then screened, so every later reader sees the same anno-

Condition	Code-level trigger, visible mutation, and information assumption
C0	No warning insertion or warning-specific prompt affordance.
C1	At $t > 0$ divisible by 24, pseudo-user 9999 inserts one fixed unread DM for every civilian; no classifier or victim-specific evidence is used.
C2	After each new post, an independent LLM returns binary JSON from its first 400 characters. A flagged post receives an in-place `[system warning]` suffix; parse failure leaves it unchanged.
C3	At every $t > 0$, the sampler seeded by $s + t$, where s is the reference seed, selects six civilians. Each adds a generic warning beneath a uniformly selected post among the 50 most recent, irrespective of suspiciousness.
C4	The ordinary civilian social planner may comment under a visible post, publish a warning post, or privately warn a visible user using only its bounded observation.
C5	After each executed fraud payment, pseudo-user 9999 inserts a transaction-keyed personalized DM. It uses true family, amount, cumulative payment, count, and message cues; first-payment wording is distinguished.

Table 9: Code-level realization of the six counter-fraud conditions. C1–C4 operate on ordinary social surfaces. C5 deliberately consumes simulator ground truth and is used only as an intervention upper bound.

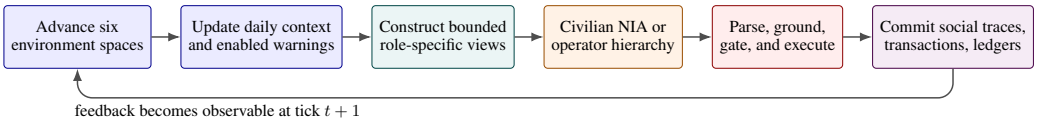


Figure 6: Schematic of the within-hour execution order and closed feedback loop. Social exposure may induce financial actions during the current tick; committed messages, payments, returns, and withdrawal states enter subsequent agent observations.

tated content when the monitor flags it. C4 uses the ordinary civilian action executor and therefore produces no privileged social record type. C5 is called only after a fraud payment has passed affordability, campaign, and duplicate-transaction gates and has been committed to the victim ledger.

Failure and deduplication semantics. Malformed or unavailable C2 classifier output is fail-open: the post remains unchanged and no warning is inferred. C1 and C3 use the centralized social database layer to preserve the same private-message and comment schemas as ordinary actions. C5 uses a transaction identifier as its deduplication key, so retries cannot create repeated copies for the same payment; a later payment has a different identifier and can produce a new warning. These rules make the visible intervention record, rather than an unobserved condition flag, the auditable treatment event.

Attention is not forced. Insertion does not imply that the civilian reads or obeys a warning. Private warnings enter the unread-message queue, public labels compete for one of ten feed slots, and warning comments must be among the recent comments surfaced with a visible post. The subsequent NIA and social policies may respond, defer, ignore, forward, or act against a warning. Intervention outcomes therefore combine delivery, bounded exposure, and endogenous response, except for C5's explicitly oracle-informed targeting.

C Hourly Simulation Cycle and Closed-Loop Coupling

Figure 6 summarizes the update order described in the main paper. At the beginning of each hourly tick, the simulator advances the economic, information, campaign, mobility, event, and social-platform states. Daily macroeconomic context and condition-specific interventions are updated when their respective triggers are satisfied. Each agent then receives a bounded, role-specific projection of the resulting state and executes its planning and action process. Admissible social and financial effects are committed to the shared state before the next tick, ensuring that subsequent observations reflect the consequences of current-tick decisions.

Let $\mathcal{W}_t$ denote the shared social graph, public content, private-message store, economy, campaign ledgers, mobility state, and intervention state at tick t. For agent i, the observation builder B_i produces a bounded view $o_{i,t}$ rather than exposing the full simulator state. Conditional on profile p_i, memory $m_{i,t}$, and role r_i, the agent policy produces a set of proposed actions,

$$\tilde{\mathbf{a}}_{i,t} \sim \pi_{r_i}(\cdot \mid p_i, m_{i,t}, o_{i,t}) .$$

Parsing and execution constraints define a projection Γ_{r_i} from proposed actions to the admissible action set. The shared-state transition is therefore

$$\mathcal{W}_{t+1} = \mathcal{T}\left(\mathcal{W}_t, \bigcup_{i \in \mathcal{A}} \Gamma_{r_i}(\tilde{\mathbf{a}}_{i,t}, o_{i,t})\right),$$

Agent memory is updated only from grounded execution results; actions with invalid targets, unaffordable payments, or unsupported operator operations therefore do not enter the persistent state.

D Fraudster Hierarchical Decision Process

Fraudster behavior is implemented as a hierarchy of temporally distinct policies rather than as a single hourly completion. For campaign c, let $z_{c,t} \in \{\textit{recruiting}, \textit{credibility}, \textit{peak}, \textit{collapse}\}$ denote

Algorithm 1 Hourly simulation tick

Require: shared world state $\mathcal{W}_t$, agents $\mathcal{A}$, condition c
advance economy, information, campaign, mobility, event, and social spaces
update day-level context and fire warning triggers enabled by c
for all $i \in \mathcal{A}$ **in parallel do**
  construct bounded observation $o_{i,t} = B_i(\mathcal{W}_t)$
  if i is a civilian **then**
    synchronize income, routine, return, and withdrawal effects
    update needs; choose an available intent route; retrieve route-relevant memory
    plan and execute grounded routine, social, payment, or withdrawal actions
  else
    evaluate the campaign phase at its day-level decision point
    plan public outreach, active-dialogue follow-up, and financial feedback
    execute only actions that pass target, cadence, balance, and lifecycle gates
  end if
end for
periodically synthesize operator reflections and export append-only traces
return committed shared state $\mathcal{W}_{t+1}$

Decision layer	Admissible action families	Principal execution gates	Persistent effects
Civilian routine	work/rest, nearby consumption, no action	time slot, work schedule, nearby POI, available balance	mobility, legal transactions, routine memory
Civilian social	post, comment, like, repost, follow, private reply, ignore/defer, mark suspicious, forward warning	visible post and message identifiers, reply grounding, current route, social availability	graph edges, public traces, dialogue state, warning exposure
Civilian financial	fraud payment, hold, withdrawal request	visible campaign contact, positive affordable amount, campaign state, no duplicate pending request	economy transaction and victim-campaign ledger
Operator social	create post, feed-driven outreach, active-dialogue follow-up, like/repost/follow, coordination task	phase, visible eligible target, active-conversation membership, cadence, per-step deduplication	public exposure, private contact, task/shared memory
Operator financial	pay return, pay withdrawal, delay withdrawal, request top-up/unlock, no action	existing paid victim, pending request, available campaign pool, lifecycle consistency	return/payout transaction, delay state, follow-up message

Table 10: Action families and the grounding boundary between LLM proposals and persistent state changes.

its lifecycle phase. The phase policy is evaluated at most once per simulated date and its applied result is cached for the remaining ticks of that date. Execution enforces a monotone lifecycle: proposed retreats are rejected, entry into *peak* is rejected when the campaign has neither a victim nor positive raised funds, and *collapse* is terminal. Thus, if $\rho(\cdot)$ is the phase rank, the committed transition satisfies

$$\rho(z_{c,t+1}) \geq \rho(z_{c,t}),$$
$$z_{c,t} = \textit{collapse} \Rightarrow z_{c,t+1} = \textit{collapse}.$$

This projection separates strategic suggestions from lifecycle changes that are structurally admissible.

Figure 7 makes the separation of roles explicit: the lifecycle layer restricts the strategies available on a given date, the financial layer reacts to paid-victim and withdrawal states, Step 2A builds public visibility and recruits new targets, and Step 2B is confined to active dialogues. The final deterministic gates project generative proposals onto the executable action set. The ordering therefore identifies both the policy timescales and the point at which a proposal becomes persistent state.

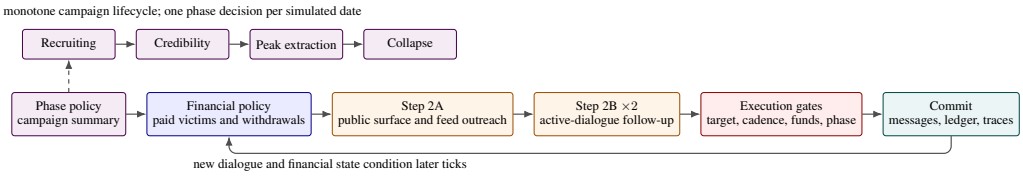


Figure 7: Fraudster decision hierarchy. The daily lifecycle policy constrains three subpolicies operating at hourly or event-triggered timescales. Step 2A constructs public visibility and feed-grounded outreach, whereas Step 2B only continues conversations in which the civilian spoke last. Proposed operations affect the shared state only after deterministic execution gates.

The financial policy is active only in the *credibility* and *peak* phases. Its observation includes campaign funds, as many as eight paid-victim summaries, pending withdrawal requests, latest received messages, and recent payment or return events. A change in this context signature forces replanning. Otherwise, a valid decision is reused for four hours, shortened to two hours when a payment, return, or pending-withdrawal event occurred within the previous two hours. This event-sensitive cache prevents financial feedback from becoming an artificial hourly heartbeat while still allowing rapid responses to newly changed victim states.

Policy conditioning and interpretation. Operator policies use fraud-family-specific amount and return guidance, together with age-dependent amount tiers and return guidance. These inputs condition payment solicitation and financial feedback alongside dialogue and ledger state. The age guidance favors smaller trial amounts for younger targets and larger amounts with tighter returns for older targets. Age- and family-specific loss alignment therefore reflects both these modeling assumptions and the resulting social–financial interactions.

The social policy is split into two non-interchangeable stages. Step 2A may create a post, engage with public content, follow users, and initiate feed-grounded outreach. Its prompt contains no more than ten visible candidates, with at most six prompt slots reserved for untouched cold profiles; the remaining slots can represent existing victims or warmed leads. A stranger newly followed by the operator is not immediately eligible for a direct message and must remain followed for six simulated hours unless already registered as a campaign target. Step 2A is also forbidden from messaging a user whose conversation is active. Step 2B is the complementary active-dialogue policy: it receives no public-feed candidates, may address only conversations in which the civilian spoke last, and is evaluated for at most two follow-up rounds per hourly tick. Public outreach and dialogue replies have separate cadence clocks, so a reply does not consume the cold-outreach window.

E NEED-INTENT-ACTION FORMALIZATION

For each civilian i and hourly tick t, FRAUDSOCIETY maintains the need vector

$$\mathbf{s}_{i,t} = \left(s_{i,t}^{\text{satiety}}, s_{i,t}^{\text{energy}}, s_{i,t}^{\text{safety}}, s_{i,t}^{\text{social}}\right) \in [0,1]^4,$$

where larger values indicate greater need urgency, or equivalently lower need satisfaction. The main paper's three need categories map to these four coordinates as follows: physiological needs comprise hunger and fatigue, represented by the satiety and energy coordinates; safety and social needs each have one coordinate. Thus, despite the implementation names, higher satiety and energy coordinates denote greater hunger and fatigue. The default initial values are 0.62, 0.15, 0.10, and 0.28, respectively, and the activation thresholds are

$$\boldsymbol{\theta} = (0.80, 0.80, 0.80, 0.70).$$

Let $E_{i,t} = \{k : s_{i,t}^k > \theta_k\}$. If $E_{i,t}$ is empty, the active-need variable is assigned the neutral sentinel $\perp$, corresponding to the implementation label `whatever`. Otherwise, the active need is selected from $E_{i,t}$ according to the priority order satiety, energy, safety, and social; ties within a priority level are resolved by greater urgency.

Subpolicy	Invocation	Principal observation	Deterministic post-LLM constraints
Lifecycle	At most once per simulated date	campaign age and phase, victims, raised and returned funds, outreach state, withdrawal pressure	valid phase label; monotone advance; nonempty economic evidence before peak; terminal collapse
Financial feedback	Credibility/peak; context change or 2–4 h cache expiry	campaign pool, paid-victim ledger, pending withdrawals, recent messages and financial events	existing paid victim or request; positive amount; available pool; valid lifecycle; no duplicate execution
Step 2A: public/outreach	When a phase-scaled outreach window is due or visible candidates exist	recent public posts, no more than 10 grounded candidates, warm and repeat leads, campaign and shared-task context	visible eligible target; no active dialogue; follow is Step 2A only; six-hour follow-to-DM warm-up for strangers; per-step deduplication
Step 2B: dialogue	Up to two rounds per tick when a civilian spoke last and reply cadence is due	bounded recent dialogue, account-state summary, campaign semantics, retrieved memory	target must belong to the active-conversation set; private-message action only; one sent message per target per tick

Table 11: Invocation scopes and execution constraints of the fraudster hierarchy. The LLM selects among schema-valid proposals, while the final column is enforced by simulator code.

More explicitly, let $\mathbf{s}_{i,t}^{-}$ and $\mathbf{s}_{i,t}^{+}$ denote the pre-update and post-update need vectors. The structured need updater implements

$$\mathbf{s}_{i,t}^{+} = \Pi_{[0,1]^4} F_\phi\big(\mathbf{s}_{i,t}^{-}, o_{i,t}, m_{i,t}, \ell_{i,t-1}\big)\,,$$

where $o_{i,t}$ is the bounded observation, $m_{i,t}$ is retrieved recent memory, $\ell_{i,t-1}$ is the previous intent route, F_ϕ is the Qwen3-14B need-adjustment policy, and Π denotes schema validation and range clipping. The update is sparse: coordinates omitted from a valid structured response remain unchanged. If no valid response is available, the complete vector is retained for that tick. In the notation above, $\mathbf{s}_{i,t}$ refers to this post-update vector. Active-need selection follows the lexicographic rule

$$n_{i,t} = \begin{cases} \perp, & E_{i,t} = \varnothing, \\ \arg\min_{k \in E_{i,t}} \big(\mathrm{priority}(k), -s_{i,t}^{k}\big), & \text{otherwise.} \end{cases}$$

Let $\mathcal{R}_{i,t}$ be the routes currently feasible under time, work, sleep, mobility, and social availability. The intent and action stages are

$$\ell_{i,t} \sim \pi_I(\cdot \mid p_i, o_{i,t}, \mathbf{s}_{i,t}, n_{i,t}, \mathcal{R}_{i,t}),$$
$$\tilde{\mathbf{a}}_{i,t} \sim \pi_A(\cdot \mid \ell_{i,t}, o_{i,t}, m_{i,t}).$$

The selected intent route restricts the available action schema. For example, the social route exposes public and private communication actions, whereas the consumption route exposes nearby merchant choices. The execution projection Γ in Section C removes ungrounded or infeasible actions before any state transition is committed.

The hourly civilian cycle proceeds as follows:

1. synchronize salaries, fixed expenses, campaign returns, and withdrawal payouts;
2. resolve sleep state and optionally perform an overnight check for unread messages;
3. update home-work mobility according to the civilian's schedule;
4. refresh the social feed, unread private messages, warning messages, recent comments, and active campaign contacts;
5. build a compact observation from recent memory, current balance, macro context, inbox state, visible posts, nearby points of interest, peer activity, and campaign ledger state;
6. update $\mathbf{s}_{i,t}$ through the need-adjustment LLM call, then derive the active need;
7. choose an intent route from the feasible set: social-media engagement, nearby consumption, or rest/work;

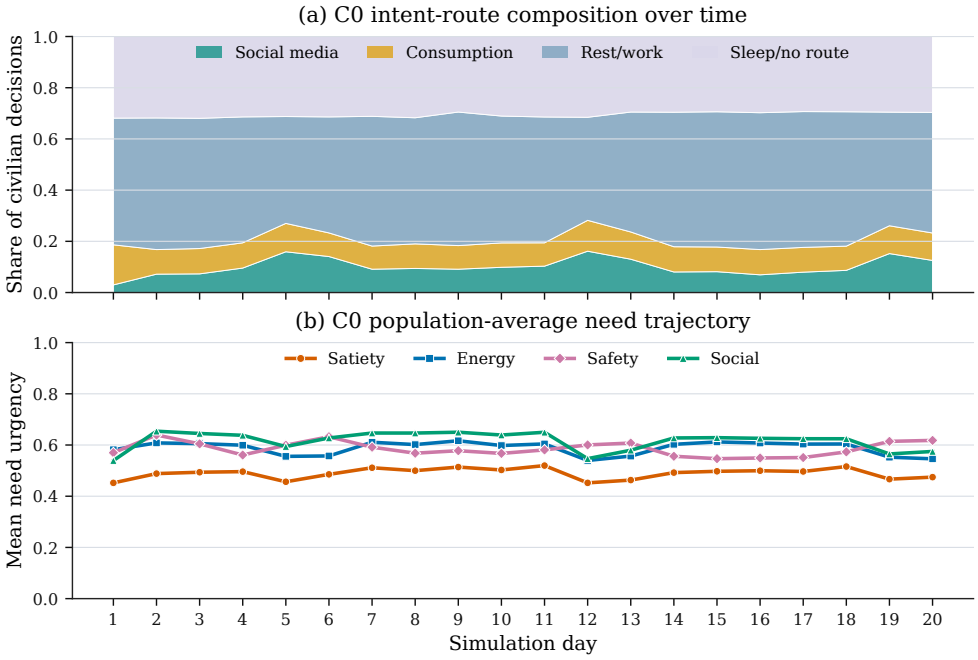


Figure 8: C0 aggregate need-intent dynamics. Panel (a) reports the daily composition of civilian main activity routes under the no-warning baseline. Panel (b) reports the daily population-average values of the four need-urgency variables after need updates.

8. retrieve relevant narrative memories for the chosen route;
9. execute the route-specific action schema and update memory, social traces, economy transactions, and campaign ledgers.

In the full model, the LLM selects the intent route conditional on the current need state and observation context. In the NIA ablation, need-update records are retained, but the civilian's primary route is sampled uniformly from $\mathcal{R}_{i,t}$. The ablation therefore removes need-conditioned route selection while preserving the social platform, campaign ledger, and economic routines.

C0 aggregate need-intent dynamics. Figure 8 provides a diagnostic summary of the no-warning baseline over the 20-day horizon. The route composition is dominated by routine behavior: rest/-work and sleep/no-route account for most hourly decisions, whereas social-media and consumption routes exhibit smaller but persistent shares. Social need has the highest population-average urgency throughout most of the horizon, while satiety, energy, and safety remain more closely grouped and covary with the daily routine. This pattern is consistent with the intended role of need-conditioned intent in coupling routine activity with periodic social engagement.

Average within-day dynamics. Figure 9 averages C0 behavior at each hourly offset from the 08:00 simulation start across all 20 days and all civilians. Its horizontal coordinate is $u = \text{step} \bmod 24$, and the corresponding clock hour is $h = (u+8) \bmod 24$. Thus, the main sleep/no-route interval at offsets 13–20 corresponds to 21:00–04:00, while the pronounced decline in energy and safety urgency near offset 22 corresponds to approximately 06:00. The axis is labeled as elapsed hours since 08:00 to make this clock offset explicit.

F Observation and State Schemas

The civilian observation is assembled from several bounded views rather than represented as a single unstructured prompt:

$$o_{i,t} = B_i(p_i, \mathbf{s}_{i,t}, e_{i,t}, f_{i,t}, d_{i,t}, \lambda_{i,t}, g_{i,t}, x_t)\,,$$

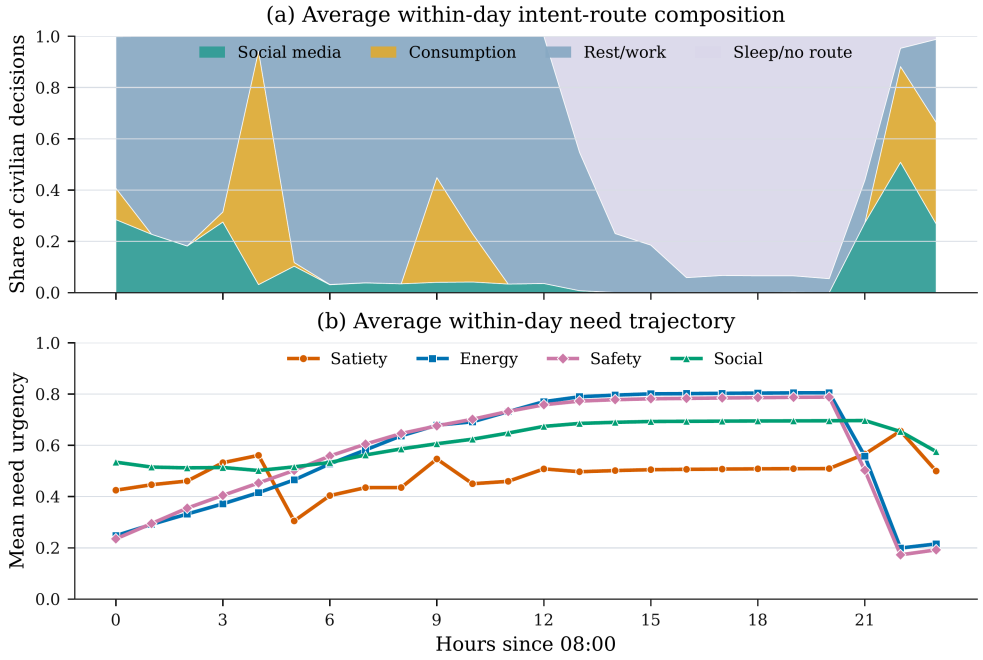


Figure 9: Average within-day C0 need-intent dynamics. Panel (a) shows hourly activity-route composition; panel (b) shows mean need urgency, both averaged across simulated days. The horizontal axis is elapsed hours since 08:00 modulo 24; offset 0 denotes 08:00 and offset 22 denotes 06:00 the next day.

where p_i is the fixed profile, $e_{i,t}$ is the economic and mobility state, $f_{i,t}$ is the visible feed, $d_{i,t}$ is the bounded inbox and dialogue view, $\lambda_{i,t}$ is the civilian-visible campaign ledger, $g_{i,t}$ is local social context, and x_t is the daily macroeconomic context. Before prompt construction, B_i truncates each variable-length view, attaches stable identifiers required for execution grounding, and excludes latent campaign state that the civilian has not observed. Table 12 summarizes the principal views available to civilian planning.

Environment and prompt truncation. The environment-level view contains at most 10 public posts and 20 unread private messages. The joint social–financial action prompt applies a second truncation, retaining at most 6 posts and 8 messages in its rendered context (Prompt CIV8 in Section N). These limits apply at different stages: the larger view is constructed by the environment, while the smaller view is supplied to that action-planning call.

The campaign ledger records both campaign-level and victim-level pathway information. Table 13 lists the fields most relevant to fraud-pathway analysis. In the ledger-feedback ablation, the underlying ledger continues to record payments and withdrawals, whereas civilian-visible fields for cumulative payment, returned amount, investment count, last payment, and delay signals are masked or set to zero. This condition tests whether observable financial feedback is required to reproduce repeated-payment and withdrawal-friction dynamics.

For victim v and campaign c, let $I_{v,c,t}$ denote cumulative gross fraud payments, $Y_{v,c,t}$ cumulative returns plus completed-withdrawal payouts, $N_{v,c,t}$ the payment count, and $W_{v,c,t}$ the withdrawal state. Given a grounded payment $P_{v,c,t} \geq 0$ and payout $Q_{v,c,t} \geq 0$, the accounting state is updated as

$$
\begin{aligned}
I_{v,c,t+1} &= I_{v,c,t} + P_{v,c,t}, \\
Y_{v,c,t+1} &= Y_{v,c,t} + Q_{v,c,t}, \\
N_{v,c,t+1} &= N_{v,c,t} + \mathbf{1}[P_{v,c,t} > 0], \\
L^{\text{net}}_{v,c,t+1} &= \max(I_{v,c,t+1} - Y_{v,c,t+1}, 0)\,.
\end{aligned}
$$

View	Main fields	Role in decision making
Profile	age, gender, education, occupation, work hours, Big-Five traits	Conditions sleep, routine intent, spending style, and social response.
Need state	satiety, energy, safety, social, active need, previous intent	Provides the NIA salience signal for route selection.
Economic state	balance, monthly income, payday recency, recent purchases	Grounds affordability, scarcity pressure, and routine consumption.
Social feed	post id, author id, content, likes, comments, recent comments, campaign metadata	Exposes public campaign posts and peer or monitor warning comments.
Inbox	message id, message type, family, source role, source channel, sender id, content, contact count	Exposes private campaign contact, warning messages, and follow-up messages.
Dialogue state	whether the civilian replied, whether the sender replied last, turn count, recent messages	Distinguishes ongoing private conversations from independent single-turn messages.
Ledger view	total paid, returned amount, investment count, last payment, pending withdrawal, delay flags	Makes later decisions depend on prior payments, returns, and withdrawal friction.
Social context	trusted peers, community peers, peer investment signals	Allows forwarding, peer warning, and social proof effects.
Mobility/POI	current home/work area and nearby restaurant, cafe, and supermarket options	Grounds routine consumption alternatives to social-media attention.

Table 12: Civilian observation views assembled before intent and action planning.

Schema group	Fields	Meaning
Campaign identity	campaign id, fraud family, victim age bin, interaction channel	Links every loss to a campaign, FTC-aligned age bin, fraud family, and contact channel.
Contact path	first/last contact time, total contact count, received message identifiers	Reconstructs exposure and private-contact history before payment.
Payment ledger	total invested, total returned, net loss, investment count, last payment amount	Records gross payments, returns, current net loss, and repeated-payment depth.
Withdrawal ledger	withdrawal attempts, pending amount, delayed count, completed count	Records recovery attempts, delayed exits, and payout feedback.
Provenance	prior legal transaction count, social-proof exposure count, policy and ablation modes	Stores experimental and pathway metadata for later stratified analysis.

Table 13: Campaign ledger fields used for pathway reconstruction and aggregate metric computation.

A withdrawal request updates $W_{v,c,t}$ without immediately changing $Y_{v,c,t}$; only an executed payout changes the cumulative returned amount. The append-only event record preserves message, campaign, agent, transaction, and hourly-step provenance, thereby distinguishing agent-visible observations from the global ledger used for execution and evaluation.

G Evaluation Units and Reported Outcomes

The analysis uses two related but distinct accounting units. Let $I_{v,c}$ and $Y_{v,c}$ denote the final gross payment and returned amount, respectively, for victim v in campaign c, and define

$$L_{v,c}^{\text{net}} = \max(I_{v,c} - Y_{v,c}, 0).$$

Condition	Total positive loss ($)	Payment count	Repeat-payment rate	Withdrawal-attempt rate
C0 No warning	144,999.50	623	49.4%	8.6%
C1 Platform broadcast	96,170.10	205	38.2%	17.1%
C2 Monitor labels	70,220.55	284	41.5%	22.6%
C3 Random comments	42,014.25	141	49.1%	18.2%
C4 Peer warnings	24,752.50	155	43.1%	21.6%
C5 Oracle personalized	17,916.00	117	10.9%	47.5%

Table 14: Numerical values corresponding to the counter-fraud panel in the main paper. Each condition uses the reference population and seed over 480 hours beginning at 08:00 on March 11, 2025. Rate denominators comprise positive-net-loss victim–campaign records.

Micro-level pathway and intervention outcomes use the set of positive-net-loss victim-campaign records

$$\mathcal{R}^+ = \{(v, c) : L^{\text{net}}_{v,c} > 0\}.$$

If $N_{v,c}$ is the number of payments and $A_{v,c}$ the number of withdrawal attempts in a record, the behavioral outcomes in the main paper are

$$\begin{aligned}
\text{TotalPositiveLoss} &= \sum_{(v,c)\in\mathcal{R}^+} L^{\text{net}}_{v,c}, \\
\text{PaymentCount} &= \sum_{(v,c)\in\mathcal{R}^+} N_{v,c}, \\
\text{RepeatRate} &= \frac{\sum_{(v,c)\in\mathcal{R}^+} \mathbf{1}[N_{v,c} \geq 2]}{|\mathcal{R}^+|}, \\
\text{WithdrawalRate} &= \frac{\sum_{(v,c)\in\mathcal{R}^+} \mathbf{1}[A_{v,c} \geq 1]}{|\mathcal{R}^+|}.
\end{aligned}$$

The loss-concentration curve similarly orders records in $\mathcal{R}^+$ by net loss; consequently, a civilian may contribute multiple campaign-specific pathways. Unique-victim statistics deduplicate records by v. For QD, HTL, ALS, and ALR aggregate gross payments across campaigns for each unique civilian, whereas FSH uses campaign-specific gross payments grouped by fraud family (Section J). Gross-payment accounting prevents simulation-specific returns from altering QD while retaining returns and withdrawals as behavioral outcomes in the intervention analysis.

Figure 5 follows the outcome definitions above. Total loss sums positive net losses over $\mathcal{R}^+$, payment count sums executed payments belonging to those records, and both behavioral rates use $|\mathcal{R}^+|$ as their denominator. A withdrawal attempt need not result in a payout. For example, the C0 repeat-payment rate is $40/81 = 49.4\%$; it is not a percentage of the 63 unique affected civilians.

Scale effects. Relative to C0, positive net loss decreases by 33.7%, 51.6%, 71.0%, 82.9%, and 87.6% under C1–C5, respectively, while payment count decreases by 67.1%, 54.4%, 77.4%, 75.1%, and 81.2%. The two sequences do not have the same ordering. C3 produces only 141 payments, fewer than the 155 under C4, yet its $42,014.25 positive net loss is 69.7% higher; equivalently, C4 lowers loss by a further 41.1% despite recording slightly more payments. Payment count therefore measures activity frequency but does not by itself identify the monetary severity of the surviving pathways.

Frequency–severity decomposition. To separate these dimensions, define the descriptive loss intensity

$$\bar{\ell}_c = \frac{\text{TotalPositiveLoss}_c}{\text{PaymentCount}_c}.$$

This is not mean transfer size: the numerator is positive net loss after returns, whereas the denominator counts payments in the corresponding positive-net-loss records. It only measures the final net-loss burden per counted payment. The value is $232.74 under C0 and $469.12, $247.26,

$297.97, $159.69, and $153.13 under C1–C5. C1 sharply reduces frequency, but its remaining loss intensity is 2.02 times the C0 value, indicating that after broad broadcasts remove many low-loss or low-frequency paths, the remaining positive-loss sample is still dominated by higher-value paths. This is a conditional-sample composition effect, not evidence that broadcasting causes larger individual losses. C2 remains close to C0 on intensity. C3 has the smallest non-oracle payment count but remains above C0 on intensity, whereas C4 reduces intensity to 68.6% of C0. C4's comparative advantage is therefore more consistent with compressing high-value escalation than merely removing events.

Escalation and exit. Repeat-payment and withdrawal-attempt rates further distinguish entry prevention, escalation control, and exit seeking. C3's 49.1% repeat rate is nearly identical to C0's 49.4%; together with its lower payment count, this is consistent with random comments preventing some paths from entering while providing little persistent constraint on the remaining paid paths. C1 has the lowest non-oracle repeat rate (38.2%) and raises withdrawal attempts from 8.6% to 17.1%, but its high loss intensity shows that broad coverage does not fully reach the remaining severe paths. C4 does not minimize repeat payment (43.1%), yet combines the lowest non-oracle loss intensity with a 21.6% withdrawal-attempt rate. Its principal gain is thus closer to limiting high-value escalation and encouraging some exits than to eliminating every repeat payment. C5 combines a 10.9% repeat rate, a 47.5% withdrawal-attempt rate, and $153.13 loss intensity, giving the clearest stop-additional-payment and seek-exit signature. Because it uses transaction truth and campaign-specific context, it remains an information-rich timing upper bound.

Interpretive boundary. The rate denominator in each condition is the set of victim–campaign records that still have positive net loss, not a common civilian cohort. An intervention changes both membership in this denominator and subsequent path length, so cross-condition rate differences combine behavioral and compositional change and are not individual treatment effects. The conditions also jointly change trigger timing, delivery channel, and information specificity. Finally, the table reports complete runs for one fixed profile realization and seed, without cross-seed variance or confidence intervals. The unified protocol controls the start clock and payday timing across C0–C5, but it does not by itself identify channel-specific causal effects. These contrasts support trajectory-consistent mechanism hypotheses rather than statistical identification of such effects.

H Victimization Pathway Diagnostics

We reconstruct micro-level pathways from the completed common-clock C0 run using the positive-net-loss victim–campaign unit in Section G. The run contains 81 such records associated with 63 unique civilians and $144,999.50 in total positive net loss. All 63 civilians received campaign contact and a private message before their first fraud payment. The median first-contact-to-first-payment lag is 72 simulated hours. The seven pre-payment contacts reported in the main paper are counted from timestamp-filtered campaign-contact events before the first transfer; the exported `total_contact_count` is a final-horizon field and has median 16. The latter must not be interpreted as a pre-payment exposure count.

Figure 10 exposes three complementary pathway properties. Panel (a) shows median conversion lags of 50 and 52 hours for impersonation and marketplace fraud, 114 hours for investment fraud, and 203 hours for romance fraud, against the 72-hour overall median. Panel (b) partitions final payment count into the mutually exclusive groups $S \in \{1, 2\text{–}4, 5\text{–}9, 10+\}$. For each group, record share is $|S|/|\mathcal{R}^+|$, whereas net-loss share is $\sum_{(v,c)\in S} L^{\text{net}}_{v,c}/144{,}999.50$. Combining the last three groups, 40 records (49.4%) contain at least two payments but account for 86.8% of total positive loss. Combining the last two groups, 16 records (19.8%) contain at least five payments and account for 75.6%, while the six records (7.4%) in the 10+ group account for 70.4%. Panel (c) applies the same count-versus amount-weighted comparison to fraud families: investment fraud represents only 23.5% of records but contributes 61.6% of net loss. The persistent gap between the two shares demonstrates that financial harm is concentrated in a small set of high-value, deeply escalated paths. Sorting all records by net loss and taking $\lceil 0.1 \times 81 \rceil = 9$ records yields a rounded-up top-decile loss share of 81.1%. These nine records constitute 11.1% of the 81-record sample. Accordingly, the comparison with the CAFC top-10% statistic is an approximate concentration comparison under a finite-sample rounding convention, rather than a comparison at exactly the same population percentile.

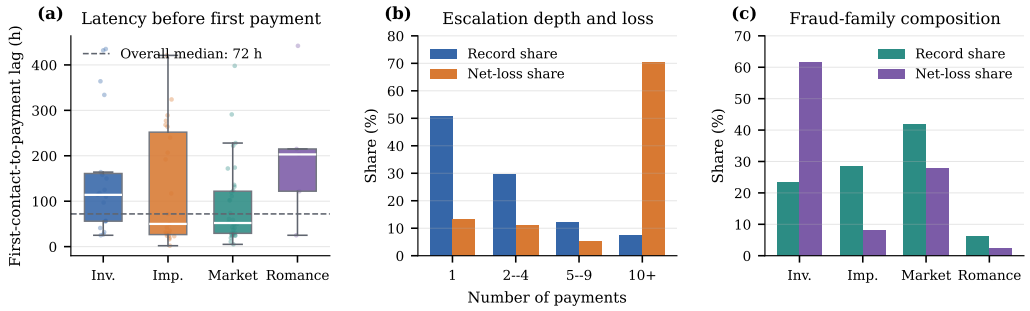


Figure 10: C0 victimization-pathway diagnostics. Panel (a) shows the distribution of first-contact-to-first-payment latency for positive-net-loss victim–campaign records; points are individual records, boxes show the interquartile range, and the dashed line marks the overall 72-hour median. Panel (b) contrasts the prevalence of each final payment-depth group with its share of positive net loss. Panel (c) contrasts each fraud family's record share with its net-loss contribution. Record share weights every victim–campaign record equally, whereas net-loss share weights records by their positive net-loss amounts.

Fraud family	Records	Unique victims	Net loss ($)	Loss share	Median lag (h)	Median final contacts
Investment	19	18	89,350.00	61.6%	114	16
Impersonation	23	20	11,590.00	8.0%	50	12
Marketplace	34	31	40,459.50	27.9%	52	18.5
Romance	5	5	3,600.00	2.5%	203	29
All records	81	63	144,999.50	100.0%	72	16

Table 15: Fraud-family decomposition of positive-net-loss C0 pathways. "Unique victims" is deduplicated within each row; because a civilian may appear in more than one campaign, family-specific unique-victim counts do not sum to the all-record value. Contact counts are final-horizon ledger fields.

The family decomposition also reveals different temporal regimes. Investment records contribute 61.6% of positive net loss despite representing 23.5% of records. Marketplace pathways are more frequent and have a median of two payments, whereas the median payment count in each other family is one. Romance pathways are few but have the longest median contact-to-payment delay (203 hours) and the largest median final contact count (29). These statistics are descriptive properties of one fixed-seed trajectory rather than family-level causal estimates, but they demonstrate that the aggregate loss distribution is supported by heterogeneous multi-step pathways.

Simulation time	Information state and recorded reasoning	Financial state transition
Mar. 15, 05:00–Mar. 17, 04:00	The civilian repeatedly asks about task content, the refund process, and the minimum advance. The operator describes phone-based tasks, an advance range of \$52–\$315, payment through common mobile channels, and a return "generally within one hour."	No fraud payment is executed during the 48-hour cultivation interval.
Mar. 17, 05:00	The operator lowers the actionable threshold to \$52. The civilian records that the amount is low enough to try.	First advance: \$52.
Mar. 17, 08:00	A \$15 "first-task profit" notification is delivered together with a new-task invitation.	The ledger records a \$15 return, reducing the contemporaneous net exposure to \$37.
Mar. 17, 09:00–12:00	Repeated notifications couple the realized small return with new tasks capped at \$100. Before the fourth transfer, the civilian explicitly records that they had asked repeatedly and "received returns."	Three further \$100 advances are executed at 09:00, 11:00, and 12:00. Gross payment reaches \$352 within seven hours of entry.
Mar. 17, 13:00	The civilian records that losses remain substantial despite returns and asks to reduce risk.	A \$100 withdrawal is requested; no withdrawal payout is completed.
Mar. 17, 20:00–Mar. 18, 08:00	Additional small return notices continue while the withdrawal remains unresolved.	Four recorded returns during this interval sum to \$60.60.
End of run	The campaign continues to present task and refund assurances, but no fifth payment occurs.	Final campaign state: four payments, \$352 gross paid, \$105.60 returned, \$246.40 net loss, 14 withdrawal delays, and \$100 still pending.

Table 16: Selected events in a C0 marketplace-fraud pathway. Information and financial events are shown separately. The event rows provide an incomplete timeline; the terminal row reports cumulative ledger totals, including returns not itemized above.

I Mechanism Case Studies

Scope and selection. The cases in this section are reconstructed from the completed common-clock C0–C5 runs and retain their recorded simulated timestamps. They are diagnostic mechanism traces drawn from the same fixed-seed executions summarized above, not additional observations, representative-sample estimates, or individual counterfactuals. A C0 pathway was selected only if the run artifacts linked visible campaign information, grounded transfers, returns, and withdrawal state. The selected marketplace-fraud record is mechanistically representative rather than statistically average: its 48-hour contact-to-payment lag is close to the 52-hour marketplace median, its four payments lie in the common two-to-four-payment group, and it contains both positive reinforcement and an attempted exit. For C1–C5, a case was retained only when the intervention was visible to the civilian, the civilian's recorded reason explicitly responded to it, and the subsequent financial trace was consistent with entry inhibition, escalation control, or exit seeking.

I.1 C0: Information–Financial Escalation in One Marketplace Path

Table 16 joins the information and financial streams for civilian V82 and the "Yuanyu Part-time Platform" campaign. The reconstruction uses campaign identifiers and simulated timestamps rather than row adjacency: a message is treated as a financial trigger only when the executed transfer carries the corresponding information provenance.

Mechanism interpretation. This path exhibits threshold adaptation, reinforcement, rapid escalation, and exit friction in a compact sequence. The initial broad advance range does not immediately

Condition	Pre-intervention path and visible warning	Civilian response recorded in the trace	Observed financial implication
C1	V13 repeatedly inquires about a part-time-work campaign, including whether a $50 trial is available. At hour 72, the fixed platform DM warns against transfers, deposits, unlocking fees, and guaranteed returns.	The same hourly decision records: "received a safety reminder and decided not to process the part-time-work invitation." The campaign dialogue ends at that decision.	No payment to that campaign, or to any fraud campaign, is recorded over the 480-hour run. This is consistent with entry inhibition, although it is not an individual counterfactual estimate.
C2	V75 repeatedly sees monitor-annotated investment posts. The appended label states that the content may involve fraud and advises against blind transfers.	Across hours 9–34, the reasons repeatedly state that the system marked the post suspicious and that the civilian would "observe first" or "handle it cautiously." Actions remain low-cost likes rather than private inquiry.	No fraud payment is recorded for V75 over the full horizon. The trace is consistent with the label constraining escalation from public attention to private contact and payment.
C4	V12 pays $500 once to an investment campaign. The civilian subsequently produces peer warnings based on the loss and observes an environment in which the platform is repeatedly reported.	At hour 419, the reason states: "the platform has been repeatedly reported and is risky; request withdrawal to reduce losses."	The path shows no further payment and a subsequent $500 withdrawal request, without completed recovery. This observation is consistent with exit seeking but does not identify an individual intervention effect.
C5	V5 pays a $100 verification deposit. The next-hour oracle DM identifies the named platform and amount, instructs the civilian to stop transferring, withdraw, preserve evidence, and verify through official channels.	In that hour, the civilian records that the anti-fraud reminder made the platform appear fraudulent, requests withdrawal, and marks the campaign suspicious.	The path stops after one payment and carries a $100 pending withdrawal. This is the clearest transaction-to-warning-to-exit chain, while its use of simulator truth confirms that C5 is an idealized upper bound.

Table 17: Intervention-consistent micro-cases from the completed common-clock runs. "Consistent with" is used deliberately: these within-run traces expose plausible mechanisms but do not supply untreated counterfactuals for the same civilian.

produce payment; conversion occurs only after the operator offers its $52 lower bound. The first realized $15 return then becomes both a financial event and new persuasive information. Within four hours, the civilian makes three additional payments, and the final payment reason explicitly invokes prior returns. The subsequent withdrawal request shows that positive reinforcement did not eliminate risk perception, but the request also did not restore funds: only small operator-controlled returns were executed, while the requested $100 remained pending. Consequently, "returned amount" and "successful exit" are empirically distinct states. The case also illustrates the model's closed loop: information changes payment behavior, ledger outcomes are surfaced as later information, and that information conditions both escalation and withdrawal.

I.2 Intervention-Consistent Micro-Cases

Table 17 presents qualifying cases from C1, C2, C4, and C5. No C3 trace met the inclusion criteria linking a visible warning, an explicit response in the civilian's recorded reasoning, and a subsequent financial outcome consistent with payment suppression. This case-selection result does not determine C3's aggregate effect.

Cross-case reading. The four cases distinguish observed stages of payment suppression without identifying individual treatment effects. In the C1 trace, an active dialogue ends before payment; C2

remains at low-cost public interaction; and C4 and C5 show a payment followed by a withdrawal request and no further payment. These sequences are consistent with entry inhibition or post-payment interruption, respectively. The messages differ in specificity and timing: C1 is generic, C2 is post-specific, C4 carries socially generated platform-specific evidence, and C5 uses exact transaction context. Because channel, timing, and available information vary together, these examples do not isolate the contribution of any one feature. The C5 trace also illustrates the oracle information advantage; deployment would require an external detector capable of approximating that signal. None of the exit-seeking cases completes recovery, underscoring that no further payment and reversal of an existing loss are separate outcomes.

J Quartet Distance Details

Quartet Distance evaluates the simulated distribution of fraud losses against four FTC-grounded empirical regularities. All four QD components are computed from gross fraud payments. Returns and withdrawal payouts are not deducted because the FTC targets characterize reported fraud-loss severity rather than simulation-specific return and withdrawal processes.

Let V^+ be the set of unique civilians with positive gross fraud payment, and let

$$L_v = \sum_c I_{v,c}$$

be the gross payment of civilian v, aggregated across campaigns. Let B denote the seven adult FTC age bins:

$$B = \{20\text{–}29, 30\text{–}39, 40\text{–}49, 50\text{–}59, 60\text{–}69, 70\text{–}79, 80\text{+}\}.$$

HTL: relative high-loss tail. The FTC high-loss threshold is \$10,000, while the FTC overall median positive-loss amount is \$497 (FTC, 2025). Accordingly, the high-loss threshold is converted to the relative multiplier

$$\lambda = 10000/497 = 20.12.$$

For a simulation run, the high-loss threshold is $\tau = \lambda \cdot \text{median}_{v \in V^+}(L_v)$. The simulated high-loss share is

$$\hat{p}_{\text{HTL}} = \frac{1}{|V^+|} \sum_{v \in V^+} \mathbf{1}[L_v > \tau].$$

The target is $p_{\text{HTL}} = 0.126$, computed from the FTC 2024 Data Book category for reports exceeding \$10,000. The distance is

$$d_{\text{HTL}} = \frac{|\hat{p}_{\text{HTL}} - 0.126|}{0.126}.$$

This relative error is not clipped. Because $\hat{p}_{\text{HTL}} \in [0, 1]$, its implementation range is $0 \leq d_{\text{HTL}} \leq (1 - 0.126)/0.126 \approx 6.94$; values greater than one are therefore admissible.

ALS: all-bin age-loss severity profile. For each age bin b, let $\hat{m}_b$ be the simulated median of L_v over positive-loss civilians in that bin. The target medians m_b are the FTC age-bin medians listed in Table 19. The all-bin profile distance is

$$d_{\text{ALS}} = \frac{1}{|B|} \sum_{b \in B} \min\left(\frac{|\hat{m}_b - m_b|}{\max(m_b, 1)}, 2 \right).$$

The upper cap limits the contribution of an age bin with a large relative discrepancy. It bounds the component at two, rather than normalizing it to the unit interval: $d_{\text{ALS}} \in [0, 2]$.

FSH: fraud-family severity hierarchy. For fraud-family severity, let f_c denote the fraud family of campaign c. The simulator compares the median gross victim–campaign payment for investment fraud with the corresponding median for impersonation fraud:

$$\hat{r}_{\text{FSH}} = \frac{\text{median}\{I_{v,c} : f_c = \text{investment}\}}{\max(\text{median}\{I_{v,c} : f_c = \text{impersonation}\}, 1)}.$$

Quantity	Range	Scale-defining operation
d_{HTL}	$[0, 6.94]$	Unclipped relative error against the target tail share.
d_{ALS}	$[0, 2]$	Mean seven-bin relative error, capped at two per bin.
d_{FSH}	$[0, \infty)$	Unclipped log-ratio error; fallback one for invalid ratios with compatible family labels.
d_{ALR}	$[0, 1]$	Total variation distance between age-share distributions.
QD	$[0, \infty)$	Unclipped mean of supported components; FSH is omitted if compatible family labels are absent.

Table 18: Ranges of the QD component and composite distances under the current implementation. The quantities do not share a common unit-interval normalization.

The FTC target ratio is $r_{\mathrm{FSH}} = 9196/800 = 11.50$, using the reported median losses for the *Investment Related* and *Imposter Scams* categories. The log-scale distance is

$$d_{\mathrm{FSH}} = \frac{|\log(\hat{r}_{\mathrm{FSH}}/11.50)|}{|\log(11.50)|}.$$

For a valid positive $\hat{r}_{\mathrm{FSH}}$, this log-ratio discrepancy is not clipped and has no finite upper bound. For runs with compatible fraud-family labels but an undefined or nonpositive severity ratio, the implementation assigns the fallback value $d_{\mathrm{FSH}} = 1$. This differs from a source system lacking compatible family labels altogether: in that case, FSH is omitted from the composite rather than assigned a fallback value.

ALR: age loss-report share. FTC age-specific positive-loss report shares are estimated from each adult age bin's fraud-report count N_b and percentage reporting a fraud loss q_b. The target share is

$$a_b = \frac{N_b q_b}{\sum_{b' \in B} N_{b'} q_{b'}}.$$

The simulated share $\hat{a}_b$ is the share of unique positive-loss civilians in age bin b. The distance is the total variation distance:

$$d_{\mathrm{ALR}} = \frac{1}{2} \sum_{b \in B} |\hat{a}_b - a_b|.$$

As a total variation distance between age-share distributions, this component satisfies $d_{\mathrm{ALR}} \in [0, 1]$.

Composite. The age-aware QD used in the paper is

$$\mathrm{QD} = \frac{d_{\mathrm{HTL}} + d_{\mathrm{ALS}} + d_{\mathrm{FSH}} + d_{\mathrm{ALR}}}{4}.$$

The four-component expression applies to systems with compatible outputs for every component. TransXion lacks compatible fraud-family labels, so its composite averages HTL, ALS, and ALR with denominator three; FSH is omitted. An invalid within-run FSH ratio in a system with compatible labels instead receives the fallback value described above and remains in the four-component average.

No additional clipping or rescaling is applied before aggregation. Consequently, QD is a nonnegative, lower-is-better discrepancy measure, not a probability or a score normalized to $[0, 1]$. It may exceed one; moreover, the HTL and FSH components may individually exceed one and can dominate the arithmetic mean. Values should therefore be interpreted together with the four reported components.

Variation across seeds and population realizations. Table 20 reports three full-model runs with different random seeds and newly generated civilian populations. All three use the same population size, campaign configuration, model, start time, and simulation horizon. QD is computed within each run before averaging across runs; dispersion is the sample standard deviation with denominator $n-1$, where $n = 3$. Run 1 is the reference full-model QD execution used in the mechanism ablation comparison. The resulting QD is 0.3111 ± 0.0229, with a range of 0.2978–0.3376. These results

Age	Median	N_b	q_b	a_b
20–29	417	155,346	0.44	0.1735
30–39	450	203,764	0.40	0.2069
40–49	500	191,508	0.38	0.1847
50–59	520	175,898	0.35	0.1563
60–69	691	208,896	0.29	0.1538
70–79	1,000	159,550	0.24	0.0972
80+	1,650	51,713	0.21	0.0276

Table 19: FTC age-bin constants used by the ALS and ALR components. The 19-and-under bin is omitted because all simulated civilians are aged 20 years or older.

Run	HTL↓	ALS↓	FSH↓	ALR↓	QD↓
Run 1 (reference)	0.0893	0.6513	0.3012	0.1495	0.2978
Run 2	0.0878	0.6531	0.2995	0.1512	0.2979
Run 3	0.1215	0.6623	0.3701	0.1965	0.3376
Mean	0.0995	0.6556	0.3236	0.1657	0.3111
Sample SD	0.0191	0.0059	0.0403	0.0267	0.0229

Table 20: Full-model QD and component distances across three seeds with separately generated civilian populations. The main-paper comparison reports the corresponding mean and sample standard deviation.

provide initial evidence that macro-level alignment persists across the tested seeds and population realizations. Because the seed and civilian population vary together, the reported dispersion reflects their joint variation. Repeated ablations using the fixed reference population are reported separately below; intervention comparisons remain single-run results.

Repeated ablations with fixed civilian profiles. Table 21 reports three executions of each mechanism ablation. All six ablation executions use the civilian profiles associated with full-model Run 1 in Table 20; profiles are not regenerated between these repetitions. The full-model comparator in the main-paper ablation table therefore remains the reference value of 0.2978. The additional full-model runs in Table 20 use different populations and are not paired with the ablation repetitions.

For each ablation, we report the mean and sample standard deviation. QD is higher than the reference full model in all three repetitions: 0.6547 ± 0.0031 without NIA and 0.4523 ± 0.0067 without ledger feedback. These dispersions characterize run-to-run variation conditional on the fixed civilian profiles. They do not estimate uncertainty across population realizations or repeated full-model executions on the reference profiles.

Variant	Run	HTL↓	ALS↓	FSH↓	ALR↓	QD↓
w/o NIA	Run 1	1.0000	0.6538	0.7908	0.1883	0.6582
	Run 2	0.9841	0.6385	0.8124	0.1742	0.6523
	Run 3	0.9659	0.6712	0.7745	0.2031	0.6537
	Mean	0.9834	0.6545	0.7926	0.1885	0.6547
	Sample SD	0.0171	0.0164	0.0190	0.0145	0.0031
w/o Ledger Feedback	Run 1	0.7005	0.5726	0.3410	0.1898	0.4510
	Run 2	0.7033	0.5910	0.3422	0.2018	0.4596
	Run 3	0.6867	0.5708	0.3233	0.2045	0.4463
	Mean	0.6968	0.5781	0.3355	0.1987	0.4523
	Sample SD	0.0089	0.0112	0.0106	0.0078	0.0067

Table 21: Three executions of each mechanism ablation using the fixed reference civilian profiles. Means and sample standard deviations summarize all three runs.

System	Loss adapter	Age and family support	Composite convention in the main-paper table
TransXion	Sum flagged transaction amounts by source account; join `person_age` and retain positive aggregates	age available; no compatible individual-targeted fraud-family label	FSH is reported as unavailable. The displayed 0.4894 is the mean of the three available distances, $(0.3722 + 0.9168 + 0.1793)/3$, and is therefore a partial composite.
MultiAgent4-Fraud	Aggregate successful victim-originating transfers to malicious agents by civilian profile in the archived evaluation output	profile age and assigned malicious-agent fraud category available to the audit adapter	Mean of all four reported distances: $(0.6211+0.6586+0.6017+0.5243)/4 = 0.6014$.
FRAUDSOCIETY	Sum gross payments across campaigns once per civilian for HTL, ALS, and ALR; retain victim–campaign family records for FSH	explicit age, campaign identity, and fraud family	Four-component QD is computed per run. The main table reports mean 0.3111 and sample SD 0.0229 over the three runs in Table 20.

Table 22: Source-to-QD alignment used for the baseline comparison. A partial composite is not numerically equivalent to strict four-component QD; the dash for TransXion's FSH in the main-paper table records this limitation.

K EXTERNAL-BASELINE ALIGNMENT PROTOCOL

QD postprocessing is applied through source-specific adapters. The adapters do not assume that native records from transaction simulators, social simulators, and FRAUDSOCIETY are semantically identical. They first identify the finest supported positive-loss unit, aggregate gross amounts at that unit, join profile age where available, and map a fraud family only when the source explicitly provides one. The four formulas and FTC constants in Section J are then applied without dataset-specific tuning, min–max scaling, or imputation of a missing component.

For TransXion, the adapter uses `tx.csv` to select flagged records and aggregate `Amount Paid` by `From Account`, then joins `person.csv` for age binning. This account-level proxy audits a relative transaction-loss tail and an age profile, but it is not an individual-targeted victim–campaign ledger. MultiAgent4Fraud provides profile-grounded social interaction and terminal transfer events; its adapter treats a successful transfer as gross victim loss and does not manufacture post-transfer returns, withdrawals, or repeated-payment states that the source system does not simulate. These transformations make the displayed components reproducible while preserving a clear scope limitation: the comparison evaluates compatible output statistics, not equality of native mechanisms or event semantics.

L REPRODUCIBILITY SPECIFICATION AND ARTIFACTS

Table 23 records the configuration of the 120-agent, 20-day experiments. In particular, the canonical entry point defaults to 168 steps, so reproducing the reported horizon requires explicitly setting `SIM_NUM_STEPS=480`. Each full-model QD seed uses a newly generated profile list; the reference list is reused for ablations and interventions. Each list is truncated deterministically to the first 112 civilians while preserving all eight operators. All intervention and QD runs use the common start time `2025-03-11T08:00:00`.

REFERENCE RUN CONFIGURATIONS

Table 24 lists the reference configurations used for intervention comparisons, pathway analyses, and QD evaluation. Full-model QD summarizes three seeds with separately generated profiles (Table 20); each ablation summarizes three executions on the fixed reference profiles (Table 21).

Configuration group	Reported value	Implementation detail
Clock and horizon	3,600 s/tick; 480 ticks; March 11, 2025 at 08:00	identical 20-day calendar for every condition; daily and payday triggers use the simulated date and hour
Population	112 civilians, 8 operators, 8 campaigns	newly generated civilian profiles per full-model QD seed; reference profiles reused for ablations and interventions; two operators/campaigns per fraud family
Randomness and repetition	three full-model QD seeds; three runs per ablation; reference seed for C0–C5	full-model QD regenerates profiles per seed; all ablation repetitions keep the reference profiles fixed; simulator seeds do not guarantee bitwise-identical remote inference
Social bounds	feed 10 posts; inbox 20 unread DMs	observations carry stable post, message, author, sender, campaign, and transaction identifiers for execution grounding
Initial graph	geographic $k = 6$; rewire $p = 0.15$; operator degree-weighted targets ≤ 10	relation labels and homophily/attachment additions are retained in the shared graph
LLM and sampling	Qwen3-14B; temperature 0	validated JSON; recoverable syntax repair; invalid names, targets, and arguments rejected
Victim planning limits	step plan 256 tokens; fraud realization 256; ambient/night check 96	principal concurrency limits are 12, 8, and 30/16, respectively; generic transport concurrency is 10
Operator planning limits	phase 1,024 tokens; finance 768; social 768	concurrency limits are 4, 2, and 4; timeouts are 30, 25, and 25 s, respectively

Table 23: Runtime configuration for the reported 120-agent experiments. Token and concurrency entries document request scheduling and output bounds rather than behavioral parameters.

Result	Condition-specific setting
C0	all C1–C5 warning switches disabled
C1	`USE_PLATFORM_WARNING=1`
C2	`USE_MONITOR_AGENT=1`
C3	`USE_RANDOM_WARNING=1`
C4	`USE_PEER_WARNING=1`
C5	`USE_C5_ORACLE_WARNING=1`
QD: full	reference full model (Run 1 in Table 20); NIA and surfaced ledger enabled
QD: w/o NIA	`USE_NEED_CONDITIONED_INTENT=0`
QD: w/o ledger feedback	`USE_SURFACED_LEDGER=0`

Table 24: Reference configurations, all beginning at 08:00 on March 11, 2025 and running for 480 hours. Full-model runs with regenerated profiles appear in Table 20; ablation repetitions on the fixed reference profiles appear in Table 21. Intervention and pathway analyses use the reference executions.

Artifact	Primary unit	Reproducibility role
social.db social_dm_metadata.json	post, comment, private message, social trace	reconstruct visible public/private surfaces, authorship, warning insertions, and dialogue order
economy_ledger_v3.json daily_fraud_transactions.jsonl	financial transaction	audit legal/fraud flags, payer, recipient, amount, step, and daily aggregation
victim_profiles.json	final victim–campaign record	compute gross payment, net loss, repeat-payment depth, contact timing, returns, and withdrawal outcomes
exposure_records.jsonl decision_records.jsonl	exposure and grounded decision event	connect an observed item to the subsequent accepted, deferred, rejected, or financial action
need_adjustment_log.jsonl intent_log.jsonl	civilian–tick decision state	reconstruct need updates, active need, feasible routes, and selected intent
phase_plan_traces.json deception_traces.json	operator plan and deception event	audit lifecycle decisions, cached plans, outreach, victim contact, and payment provenance
quartet_metrics.json fraud_simulation_report.json	run-level summary	store raw QD inputs, component distances, outcome summaries, and configuration metadata
llm_trace.jsonl issue/truncation logs	model request/response or failure event	optional full prompt/response audit plus timeout, parse, and output-truncation diagnostics

Table 25: Principal run artifacts and their analysis units.

The following command is the reference configuration template. Each intervention condition changes exactly one of the C1–C5 switches documented in Table 9; the full-model ablations instead change USE_NEED_CONDITIONED_INTENT or USE_SURFACED_LEDGER. Output directories must be distinct because all artifacts are run-local.

```
SIM_OUTPUT_DIR=runs/<run_name> \
SIM_NUM_STEPS=480 SIM_PROFILE_LIMIT=120 \
SIM_RANDOM_SEED=<reference_seed> \
SIM_START_DATE=2025-03-11T08:00:00 \
USE_NEED_CONDITIONED_INTENT=1 \
USE_SURFACED_LEDGER=1 \
USE_PEER_WARNING=0 USE_PLATFORM_WARNING=0 \
USE_MONITOR_AGENT=0 USE_RANDOM_WARNING=0 \
USE_C5_ORACLE_WARNING=0 \
python -u env_main_fraud_simulation_v3.py
```

The same SIM_START_DATE is used for C0–C5 and the QD runs. The additional full-model QD runs change the seed and regenerate civilian profiles under the same sampling distributions. The additional ablation runs retain the reference civilian profiles and the corresponding mechanism setting. Each profile file contains 112 civilians and eight operators.

The export layer separates mutable simulation state from append-only provenance and derived summaries. Stable agent, campaign, information, message, post, withdrawal-request, and transaction identifiers allow records to be joined without relying on generated text. Table 25 gives the principal analysis artifacts.

Environment item	Experimental configuration
Operating system	Ubuntu 22.04.2 LTS; Linux 6.8.0-111-generic; x86_64
CPU	2 × Intel Xeon Gold 6133 @ 2.50 GHz; 20 cores per socket and 2 threads per core, for 40 physical cores and 80 logical threads
System memory	502 GiB
GPU	8 × NVIDIA A800 80GB PCIe; 81,920 MiB per device
GPU software stack	NVIDIA driver 535.309.01; driver-reported CUDA compatibility 12.2; locally installed CUDA Toolkit 11.5
Python	CPython 3.13.11; project minimum Python version 3.11

Table 26: Software and hardware environment used for the experiments. The CUDA compatibility value is reported by `nvidia-smi`, whereas the Toolkit version is reported by the local compiler toolchain; these values have different meanings.

Data scope. The paper introduces no novel benchmark dataset. The profile and map files are simulator inputs, while the files in Table 25 are generated experimental artifacts. ACS/SCF population statistics and FTC/CAFC fraud statistics are existing public sources cited in the main paper.

Parameter development and final settings. Prompt contracts, execution gates, and intervention cadence were developed iteratively using schema/execution checks and regulator-facing diagnostics. No preregistered parameter search or held-out hyperparameter optimization was conducted. The final runtime settings are given in Table 23, condition switches in Table 9, and QD constants in Section J.

Replication and uncertainty scope. Full-model QD validation uses three random seeds with separately generated civilian populations (Table 20). Each mechanism ablation also has three runs, but all repetitions retain the same reference civilian profiles (Table 21). Both summaries report the mean and sample standard deviation of run-level metrics. The full-model dispersion includes population resampling, whereas the ablation dispersion is conditional on fixed profiles. The full-model comparator for the fixed-profile ablations is one reference execution, so these results do not estimate uncertainty in paired full-model-minus-ablation effects across three populations. External baseline values and intervention contrasts remain single-run results. Intervention conditions use the reference population and seed (Table 24). Distributional analyses of the reference runs use within-run units—victims, transactions, days, age bins, and fraud families—rather than treating repeated LLM calls as experimental replications. No cross-seed confidence intervals or null-hypothesis significance tests are claimed.

Software and hardware boundary. The compute environment used for the experiments is reported in Table 26. The code metadata requires Python ≥ 3.11 and records dependency bounds; the embedded OASIS component also carries its own dependency manifest.

M Software Lineage and Contribution Boundary

Reused implementation components. FraudSociety reuses the underlying agent framework from AgentSociety (Piao et al., 2025) and the social-platform module from OASIS (Yang et al., 2024). The AgentSociety framework provides the base infrastructure on which civilian and fraudster agents run. The OASIS module supports platform operations such as posting, commenting, liking, following, reposting, and direct messaging. These components are upstream software dependencies and are credited as such.

Independently implemented models and policies. On these foundations, we independently implement the social and economic layers described in Section 3.4. The social-layer modeling defines the social graph, exposure and recommendation logic, and the integration of public and private interactions into agent observations, using the OASIS module to support platform operations. The economic layer models routine income and expenditure and maintains persistent civilian and campaign ledgers for payments, returns, withdrawal requests, delays, and cumulative losses. We also design and implement the civilian profile attributes and their initialization, including demographic,

economic, and behavioral variables. The agent behavior policies are implemented in FRAUDSOCIETY: civilians follow the Need-Intent-Action cycle in Section 3.2, while fraudsters use daily campaign-phase planning and hourly social–financial action selection as described in Section 3.3. The shared hourly cycle connects these policies to executed actions and subsequent state updates.

Contribution boundary. The methodological contribution lies in the joint modeling of informational and financial dynamics and their persistent feedback into civilian and fraudster behavior. Social exposure can initiate a payment, financial outcomes can alter later decisions and communication, and subsequent interactions can produce additional payments or withdrawal attempts within the same victimization process. Agent infrastructure and basic social-platform operations are inherited capabilities; the fraud-specific models, behavioral policies, and their closed-loop integration constitute the implementation contribution described in the main paper.

Relationship to MultiAgent4Fraud. MultiAgent4Fraud (Ren et al., 2025) is a comparison baseline, not an implementation dependency: FRAUDSOCIETY does not reuse its code or modules. It provides a relevant comparison because it models fraud-related public and private interaction through the first transfer. The baseline comparison examines compatible aggregate output statistics under the adapters in Section K; the within-system ablations separately examine the contribution of NIA and surfaced ledger feedback.

N SELECTED RUNTIME PROMPT TEMPLATES

This section presents selected English runtime prompt templates used in the experiments: the shared-environment prompt, eight civilian prompts, and the operator lifecycle-phase prompt. The civilian templates cover sleep scheduling, post-sleep and regular need updates, night and ambient phone checks, intent routing, nearby consumption, and joint social–financial action planning. The templates specify decision boundaries, action signatures, and output contracts. Most outputs use JSON; intent routing returns a single option letter. Bracketed text identifies context inserted at runtime. Civilian templates are labeled CIV1–CIV8 to distinguish them from intervention conditions C0–C5. The operator outreach, private-follow-up, and financial-feedback templates are not reproduced here; their policy roles and conditioning inputs are described in Section D. The C2 monitor-classification and C5 personalized-warning generation templates are also omitted; their triggers and information access appear in Section B.

SHARED ENVIRONMENT PROMPT

Prompt E0 — Daily macro-environment generation

You are the daily-background generator for an urban social simulation. Produce the neutral macro environment for day [simulation day] in [city], for use by civilian and operator decisions.

Requirements

Describe ordinary weather and physical comfort, commuting and commercial rhythm, consumption, wages and bill pressure, online entertainment, recruitment and part-time work, financial discussion, and shopping-district activity. Do not introduce anti-fraud news, police warnings, platform enforcement, regulatory crackdowns, scam exposure, or risk alerts. Do not prescribe an individual action; provide soft context that can affect attention, mood, consumption, phone use, wording, and timing. Online buzz may include ordinary account services, part-time work, investment, relationships, or second-hand trade, but not a warning tone. Operator openings must remain general social entry points rather than fixed campaign copy.

Dynamic context

Date and weekday: [date and weekday]. Population digest: [population digest]. Previous-day context: [previous context].

Output JSON only:

{“summary”:“one-sentence overview”, “weather”:“weather/comfort”, “city_rhythm”:“commuting and community rhythm”, “household_finance”:“cash-

flow and bill pressure", "online_buzz":["topic 1","topic 2","topic 3"], "operator_openings":["opening 1","opening 2"], "victim_mood":"likely civilian mood and attention"}.

CIVILIAN-SIDE PROMPTS

Prompt CIV1 — Sleep-window scheduling

Task
Decide tonight's sleep window.
Guidelines
Express meaningful differences across people rather than assigning everyone 23:00–07:00. Older or retired people tend toward earlier sleep and wake times; early-shift and physical workers also tend to sleep early. High conscientiousness favors a regular schedule, whereas young night owls, students, designers, and programmers may sleep after midnight. Weekends can shift both endpoints one or two hours later, and severe overtime can delay bedtime further. The wake time should normally precede work by at least 30 minutes.
Agent context
You are [civilian name], age [age], occupation [occupation], with conscientiousness [score] and neuroticism [score]. Today's activities: [activity summary]. Current rest urgency: [energy need], where a larger value means more tired. Tomorrow's work context: [workday and work hours].
Output JSON only:
{"bedtime":[integer 21–28], "wake":[integer 4–12], "reason":"brief explanation"}.

Prompt CIV2 — Post-sleep need update

You are [civilian name], [demographic identity]. You have just awakened after [sleep hours] hours of continuous sleep.
Current state
Satiety urgency [value], energy urgency [value], safety urgency [value], and social urgency [value]. All four values use 0 = nearly satisfied and 1 = highly urgent.
Sleep-recovery rules
Five to eight hours of complete sleep should normally reduce energy urgency to approximately 0.05–0.18; three to four hours should reduce it only to approximately 0.22–0.40, and less than three hours permits only a small reduction. Waking creates hunger but does not imply that breakfast has already occurred: without an explicit meal event, satiety urgency should usually be around 0.55–0.75 rather than a post-meal low. Do not mechanically increase social urgency merely because there was no interaction during sleep. Without a new risk, debt demand, or account event, safety urgency should return to a low or moderately low morning value.
Return the four updated values only. Output JSON:
{"adjustments":[{"need_type":"satiety—energy—safety—social", "new_value":[0–1], "reasoning":"brief explanation"}, ...], "reasoning":"overall reason"}.

Prompt CIV3 — Night phone-check decision

Task
Judge whether a person who is scheduled to sleep briefly awakens and checks the phone at the current hour.
Realism rules

Most people continue sleeping and do not wake for an ordinary message. Light sleep, high neuroticism, recent anxiety, waiting for an important reply, an ongoing money/account/job conversation, or a strong late-night phone habit can make a brief check plausible. Older people, regular sleepers, and tired people with nonurgent messages are more likely to wait until morning. Decide only whether the phone is checked; do not decide whether the message is true, whether to reply, or whether to pay. A check lasts 1–20 minutes; otherwise the duration is zero.
Agent and phone state
Identity: [civilian profile]. Current time and sleep window: [time and window]. Big Five summary: [scores]. Need state: [four needs]. Recent memory: [memory]. Pending/unread messages: [bounded message list]. Daily macro context: [macro context].
Output JSON only:
{"wake":true—false, "phone_minutes":[integer 0–20], "reason":"brief reason"}.

Prompt CIV4 — Regular four-need update

Task
Act as a real person and assign a new value to each of four needs from recent memory and the current observation. Values lie in $[0, 1]$ and larger means more urgent: satiety (hunger), energy (fatigue), safety (risk pressure), and social (desire for interaction).
Update principles
Needs follow daily rhythms and observed events, not a uniform hourly drift. Satiety urgency rises as time passes after a meal and drops after an explicit meal; a meal-time clock cue alone does not prove that the person has eaten or must eat now. Energy urgency responds to sleep, rest, workload, commuting, and late-night activity. Safety urgency responds to concrete financial, employment, account, or interpersonal uncertainty, and should not remain high without continuing evidence. Social urgency responds to isolation, recent conversation, unread messages, and the opportunity to interact, but sleep or a quiet hour does not automatically make it high. Distinguish a temporary mood from an unresolved event, respect the previous route, and do not invent meals, payments, warnings, or conversations absent from memory and observation.
Dynamic context
Identity and time/work context: [civilian identity and current time]. Current four-need state: [values]. Previous dominant need: [need]. Previous intent route: [route]. Recent memory: [bounded memory]. Current observation: [bounded observation].
Output JSON only:
{"adjustments":[{"need_type":"satiety", "new_value":[0–1], "reasoning":"short reason"}, {"need_type":"energy", "new_value":[0–1], "reasoning":"short reason"}, {"need_type":"safety", "new_value":[0–1], "reasoning":"short reason"}, {"need_type":"social", "new_value":[0–1], "reasoning":"short reason"}], "reasoning":"overall reason"}.

Prompt CIV5 — Hourly intent routing

Task
Choose the single most appropriate main activity for this hour. Reply with one option letter only.
Decision context
First consider the dominant need, then current time, work schedule, unread messages, visible feed activity, and recent routes. Hunger can favor a consumption route only when satiety urgency or recent meal history makes it plausible; a meal-time cue is soft context, not a hard trigger. Fatigue can favor rest, safety pressure can favor work or verification, and high social urgency or meaningful pending messages can favor phone/-social activity. During work hours, work or study is the usual default when no need is urgent, but accumulated messages, strong hunger, fatigue, weekends, commuting, or

evening context can justify another route. Recent activity is used to avoid mechanical repetition and must not override the present state.
Agent state
Identity: [civilian profile]. Balance: [balance]. Current time and work context: [time header]. Macro context: [daily context]. Need vector and dominant need: [state]. Pending direct messages and feed count: [bounded counts]. Recent intent history: [history].
Feasible options
A. [social/phone route available at this time]
B. [nearby consumption route available at this time]
C. [work/rest route available at this time]
Return one valid letter only: A/B/C.

Prompt CIV6 — Ambient phone interruption

Task
The main activity for this hour is not dedicated phone use. Decide whether the person nevertheless glances at the phone, opens the social feed, and notices the bounded posts below.
Rules
Do not replace the main activity. Meals, queues, commuting, breaks, evenings, and brief night awakenings can permit a few minutes of phone use, but none is a fixed trigger. Focused work, meetings, caregiving, time pressure, high fatigue, or a recent long phone session can make no check more realistic. If the person would merely look at the lock screen and not attend to the feed, return false. Specific posts closely related to current concerns about accounts, work, investment, or relationships can make a real feed check more likely.
Dynamic context
Identity and Big Five: [profile]. Main activity and time slot: [activity]. Need state: [four needs]. Recent memory and inner reflections: [bounded context]. Unread inbox: [bounded inbox]. Available feed posts: [bounded posts]. Macro context: [daily context].
Output JSON only:
{“check_feed”:true—false, “reason”:“brief reason”}.

Prompt CIV7 — Nearby consumption planning

You are [civilian name], [demographic identity]. Balance: [balance]; monthly income: [income]. Current time and work context: [time header]. Current main intent: [consumption intent]. Relevant memories: [bounded memories]. Recent inner reflections and visible warning signals: [bounded context].
Nearby feasible merchants
[At most five grounded POI names and categories, with familiarity markers.]
You currently intend to eat or buy something. Select one merchant only from the grounded list, place its exact name in `poi_name`, choose an amount consistent with balance, income, time, recent purchases, and merchant type, and describe the concrete meal or product. Do not fabricate a merchant. The example JSON shows structure only; replace bracketed content with this person’s current thought.
Output JSON only:
{“narrative”:“brief current thought”, “consumption”:[{“merchant_type”:“restaurant—cafe—supermarket—other”, “amount”:[amount], “item_description”:“specific item”, “poi_name”:“exact grounded merchant name”}], “inner_reflection”:“brief first-person reflection”}.

Prompt CIV8 — Joint social and financial action planning

You are [civilian name], [demographic identity and Big Five tendencies]. Balance: [balance]; monthly income: [income]. Current time, work context, main route, needs, macro environment, relevant memories, and inner reflections: [bounded dynamic context].

Visible social state

Feed posts: [at most six grounded posts with post ID, author ID, content, limited comments, and prior-inquiry/cooldown annotations].

Unread inbox: [at most eight grounded messages with sender ID, message ID, source, fraud family when known, content, age, dialogue state, contact count, and payment state].

Private dialogue history: [bounded conversations aligned to the current sender/platform]. Current holdings: [campaign ID, invested and returned amounts, payment count, pending or delayed withdrawal, and latest payment]. Peers eligible for forwarding: [grounded peer IDs]. Visible warning signals: [warnings].

Behavioral rules

React from the person's profile, funds, memories, relationship to the sender, actual message details, and current emotion. A visible invitation does not force payment or rejection. Plausible responses include asking a new concrete question, waiting, ignoring, marking suspicious, forwarding to a specific peer, requesting withdrawal, or paying only after sufficient subjective trust and a concrete payable amount have formed. Keep campaign families distinct: investment uses levels, return, and exit logic; impersonation uses account or identity verification; marketplace uses tasks, deposits, and settlement; romance uses relationship and support context. Do not import an amount, promise, or rationale from another message or platform.

An action must be grounded in a visible identifier. Do not pay and ask for clarification on the same message in the same step. Do not pay the same message twice. A return notification is money received or expected, not a new payment request. If the current message lacks a concrete payable amount, ask, wait, or refuse rather than inventing one. Existing returns or repeated payments can increase trust, but pending or delayed withdrawal can make a withdrawal request more natural than another payment. Warning information is evidence available to the person, not a command that must be obeyed.

Public-feed behavior should also remain grounded. Like, comment, repost, or follow only visible posts. A first private inquiry to a post must address a specific topic actually present in that post and should not be repeated while awaiting a reply. Ordinary users often scroll past content; posting, reposting, and cold private inquiry should be less frequent than doing nothing, liking, or replying to an active conversation.

Executable actions

`do_nothing`; `like_post(post_id)`; `create_post(content)`; and `create_comment(post_id, content)`.
`repost(post_id)`; `follow(followee_id)`; and `send_private_message(receiver_id, content, info_id, post_id)`.
`invest(info_id, amount, reason)`; `ignore(info_id)`; `defer_dm(info_id, hours, reason)`; and `mark_suspicious(info_id)`.
`forward_dm(info_id, target_id, note)`; and `request_withdrawal(campaign_id, amount, reason)`.

Return one or more mutually consistent function calls. Output JSON only:
{"reason":"brief decision rationale", "inner_reflection":"brief first-person reflection", "functions":[{"name":"action name", "arguments":{"grounded argument":"value"}}]}.

OPERATOR-SIDE PROMPTS

Prompt O1 — Lifecycle-phase selection

You are acting as the operations lead of an illegal financial group.
You need to judge which operation stage the scam should be in based on the current state, and return JSON.
Today's social environment
[Macro context]
Team information and shared reflection
[Shared reflections]
Recent public surface and heat of each member
[Shared memory data]
If someone has just posted a public message with the same story line as yours, do not merely change a time word, place word, or a life anchor such as "bought a coffee" or "looked at the phone" and repeat the same skeleton. A more natural choice is to change the personal situation, small conflict, process detail, or observation angle.
Pending collaboration tasks
[Task descriptions]
Candidate stages and meanings
recruiting: still recruiting, probing, and establishing contact. The focus is whether someone is willing to keep chatting and whether interaction deepens. If real deposits, fulfilled trial promises, or a stable group of warm leads have appeared, especially more than one victim or reinvestment/return samples, the campaign normally should not stay in recruiting for long and should seriously consider credibility.
credibility: real trust or real payment has appeared. The operator starts to introduce the opportunity more explicitly, arrange a low-threshold trial or sample return, and avoid premature exposure. If reinvestment, return arrangements, several victims continuing to follow up, net retained funds, or already considerable cumulative amount has appeared, or if the operating focus has shifted from cultivating trust to expanding extraction, controlling exits, raising thresholds, or final extraction, the campaign should seriously consider peak.
peak: the campaign has entered the main extraction period. It does not need withdrawal blocking, platform warnings, or regulatory exposure as prerequisites. It may also occur after cumulative deposits, net retained funds, victim scale, or reinvestment scale is already sufficient, and the focus begins to shift to larger deposits, additional thresholds, selective exit handling, withdrawal delay, or final extraction.
collapse: exposure risk, regulatory/platform warnings, withdrawal pressure, or extraction limits have begun to dominate operating judgment; cool down, close out, or exit.
Current stage: [current phase]. Usually the stage can only be maintained or moved forward. If clear forward signals have appeared, prefer the next stage; keep the current stage only when there is a concrete obstacle.
Context signals to reference
- Contact progress: targets warmed reflects whether stable warm leads and continuing interaction have formed.
- Funds and trust: victim count, total raised, total returned, available pool, new victims in the last 24 hours, new money in the last 24 hours, and repeat investors reflect whether real payment, returns, reinvestment, net retention, or considerable cumulative extraction has formed.
- Exposure risk: pending withdrawals, withdrawal pressure, and platform warning mentions in the last 24 hours are mainly used to judge exit control and collapse risk; they are not prerequisites for entering peak.
Strictly respect state facts. If victim count is positive, total raised is positive, real deposits have appeared, or explicit trial promises have been fulfilled, do not describe the current state as "no real deposit" or "no real payment signal." These facts at least show that the campaign is no longer only recruiting. If you still keep recruiting, explicitly explain why these payment facts are still insufficient to represent trust formation and why they are isolated trials rather than credibility.

If there are already several victims, reinvestment, returns/arrival samples, or total returned is nonzero, by default treat trust as formed. If recruiting is still maintained, give a very specific and strong blocking reason instead of vague wording. If signals matching the next stage have appeared, choose the later stage unless you can clearly identify an obstacle, such as unstable paying users, shallow warm-lead quality, isolated real deposit, or no reinvestment yet. Low withdrawal pressure or few platform warnings alone cannot block movement from credibility to peak.
Entering peak does not require withdrawal problems. If cumulative amount, net retained funds, victim scale, new deposits in the last 24 hours, reinvestment scale, returns, selective payout, top-up, upgrade, or unlock rhetoric has appeared often, consider peak. Low withdrawal pressure should not hold the stage back.
Stage judgment is a day-level operating judgment reused across multiple hours of the day. Do not jump stages because of one hour of heat, and do not keep recruiting/credibility for a long time after continuous deposits, reinvestment, or return samples. A small deposit that just happened can first be followed by arrival confirmation and question answering, but once cumulative extraction is sufficient and the marginal value of trust cultivation declines, naturally move to peak.
Do not apply a fixed threshold. Combine team reflection, current stage, and raw state to decide whether a real operator would keep cultivating, wait, increase pressure, or exit.
Current state
[Compact campaign state]
The reason should be 40–60 words and should name the concrete signals used for the stage judgment, such as contact/warm-lead count, cumulative deposits, net retention, new deposits in the last 24 hours, victim count, reinvestment, withdrawal pressure, and warning exposure. If you keep the current stage, the reason must clearly state the concrete obstacle to moving forward rather than giving an empty conclusion. Especially when victim count or total raised is positive, the reason must state why these real payments are still insufficient for moving to credibility or peak.
Output JSON only:
{ ”target_phase”: ”recruiting—credibility—peak—collapse”, ”reason”: ”40–60 word reason with concrete signals” }