Simulating Impact of LLM-Generated Fake Review Attacks on Heterogeneous Consumer Personas Using Rule-Guided Agent-Based Modeling

Ikrar Gempur Tirani, Muhammad Irgi Abayl Marzuku, Ahnaf Fauzan Zaki
Department of Informatics Engineering
Faculty of Engineering, Hasanuddin University
Makassar, South Sulawesi 90245, Indonesia
Email: {tiraniig23d, marzukumia23d, zakiaf23d}@student.unhas.ac.id

***Abstract*—The proliferation of Large Language Models (LLMs) has fundamentally transformed online manipulation landscapes, making convincing fake review generation scalable and economical. This paper presents a novel rule-guided Agent-Based Modeling (ABM) approach that integrates LLMs with explicit decision rules to simulate the dynamic impact of fake review campaigns on heterogeneous consumer personas in e-commerce platforms. Unlike traditional ABM using simple rule-based logic, we employ LLMs (Llama 3.1 8B via Ollama) for natural language content generation across all agents (reviewers and shoppers), combined with decision rules ensuring reproducibility and behavioral consistency.**

This rule-guided LLM approach provides dual advantages: (1) LLMs generate natural and varied review texts and reasoning, (2) decision rules ensure predictable and reproducible agent behavior. Our simulation addresses two critical research questions: (RQ1) the quantitative impact of burst-style fake review attacks on low-quality product conversion rates, and (RQ2) the differential vulnerability of three consumer personas (Impulsive, Careful, and Skeptical) modeled with rule-guided prompts to such manipulation.

Through 20 iterations simulating a headphone e-commerce marketplace with 5 products and 6,000 total buyer decisions, we found that targeted fake review campaigns can increase low-quality product conversion rates from 0% to 54–72% ($p < 0.0001$, Cramér's V > 0.6). Results reveal critical findings: the Skeptical persona proved *far more resistant* with only 40.4% conversion rate – dramatically lower than Impulsive (92.5%) and Careful (95.0%), both showing extreme and practically identical vulnerability (2.5pp gap, $p = 0.18$ not significant). This demonstrates that *burst detection rules* effectively reduce vulnerability by >50 percentage points, while without detection mechanisms, reading more reviews provides no protection – Careful is even slightly more vulnerable than Impulsive due to *false confidence* created by high volume without detection framework.

Key contributions include: (1) first rule-guided LLM framework for reproducible e-commerce ABM combining natural language generation with decision logic, (2) empirical evidence that LLMs can generate realistic review content for large-scale simulation, (3) counterintuitive finding that reading more reviews without detection mechanisms actually increases vulnerability, and (4) temporal insights into burst attack dynamics in e-commerce ecosystems.

***Index Terms*—Agent-based modeling, large language models, e-commerce security, consumer behavior simulation, Llama 3.1, MESA framework, prompt engineering, rule-guided agents**

I. Introduction

A. Background

The digital era has transformed consumer decision-making, with online reviews becoming a crucial factor in shaping trust and purchase behavior. Studies show that modern consumers heavily rely on product reviews and ratings before making purchases, with high ratings significantly increasing conversion likelihood. This phenomenon has created a "reputation economy" where product value is determined not only by intrinsic quality but also by collective perception formed through reviews.

However, the emergence of Large Language Models (LLMs) such as GPT-4, Claude, and Llama has created an unprecedented threat to online review system integrity. LLMs can generate highly convincing text indistinguishable from human writing at very low cost and unlimited scale. This phenomenon has created what we call the "democratization of deception" – where the ability to conduct large-scale review manipulation is no longer limited to actors with significant resources.

B. Problem Statement

E-commerce platforms currently face a fundamental dilemma: traditional fake review detection systems based on textual analysis and statistical patterns are becoming increasingly ineffective against LLM-generated reviews. Research shows that even state-of-the-art machine learning algorithms struggle to detect reviews generated by modern LLMs with consistent accuracy [1].

More concerning, there is no comprehensive research measuring the *behavioral* impact of AI-generated fake reviews on different consumer types. Traditional research methodologies – surveys, laboratory experiments, and observational data analysis – have limitations in capturing temporal dynamics and complex interactions in online review ecosystems. This is where Agent-Based Modeling (ABM) offers unique advantages.

C. Research Gap and Novelty

Existing literature on fake reviews generally focuses on two areas: (1) detection algorithm development, and (2) economic analysis of aggregate impacts. However, significant gaps exist in our understanding of:

- Temporal dynamics of fake review attacks and burst attack effects
- How consumers with different cognitive characteristics respond to manipulation
- Complex interactions between review volume, text quality, and purchase decisions

This research fills these gaps with a novel approach: **rule-guided LLM Agent-Based Modeling**. Inspired by recent research showing generative AI potential in social network simulation [1], this approach combines advantages from two paradigms:

1) **LLM for Natural Language Generation:**
   - Genuine reviewers use LLMs to generate natural and varied reviews based on product specifications and personality
   - Fake reviewers use LLMs to generate promotional reviews with template variations
   - Shoppers use LLMs to generate reasoning text explaining their decisions
2) **Decision Rules for Reproducibility:**
   - Shopper decisions based on explicit IF-THEN rules processing metrics (rating, price, negative counts)
   - Rules ensure consistent and reproducible agent behavior
   - Skeptical persona equipped with burst detection rules to detect temporal patterns

This hybrid approach has several fundamental advantages: (a) **Content Realism** - LLMs generate natural, non-templated review text, (b) **Reproducibility** - decision rules ensure simulation results can be replicated, (c) **Transparency** - both rules and reasoning text can be inspected, and (d) **Scalability** - can simulate thousands of agents with reasonable computational cost.

D. Research Questions

This research aims to answer two main research questions:

RQ1: How much does the conversion rate increase for low-quality products targeted by LLM-generated fake review campaigns?

RQ2: Which consumer persona modeled with rule-guided LLM (Impulsive, Careful, or Skeptical) is most vulnerable to fake review manipulation?

E. Contributions

This paper provides several significant contributions:

1) **Rule-Guided LLM Framework for ABM:** We develop the first framework combining LLM natural language generation with explicit decision rules for reproducible e-commerce simulation.
2) **Prompt Engineering Methodology for ABM:** We demonstrate how prompt engineering can be used to create realistic agent heterogeneity in generated content.
3) **Burst Detection Rules Effectiveness Evidence:** We demonstrate that the Skeptical persona with *burst detection rules* (burst ratio, rating jump) achieves dramatic resistance (40.4%) compared to Impulsive/Careful (92–95%), proving temporal detection effectiveness in reducing vulnerability by $>$50 percentage points.
4) **Counterintuitive Finding on Information Overload:** We reveal that Careful and Impulsive show practically identical vulnerability (95.0% vs 92.5%, 2.5pp gap not significant), challenging the assumption that "reading more reviews" provides protection. Without *detection framework*, high information volume even creates *false confidence*.
5) **Temporal Dynamics Insights:** We uncover temporal patterns of effective *burst* and *maintenance* attacks in manipulating product ratings.
6) **Practical Implications:** We offer concrete recommendations for platforms, regulators, and consumers based on simulation results.

F. Paper Structure

Section II reviews related literature on fake review detection, ABM, and LLMs in simulation. Section III explains simulation methodology focusing on rule-guided prompt design. Section IV details experimental setup. Section V presents simulation results with statistical analysis. Section VI discusses finding implications and methodological trade-offs. Section VII concludes with limitations and future research directions.

II. Related Work

A. Agent-Based Modeling with Large Language Models

Integrating LLMs with Agent-Based Modeling is a rapidly developing research area. Park et al. [2] demonstrated that LLMs can be used as a "*cognitive engine*" for simulation agents in virtual environments, producing highly realistic behavior. Recent research shows that generative AI can be effectively used in social network simulation, opening new opportunities for modeling complex interactions [1].

However, full LLM-driven approaches have reproducibility challenges due to inherent non-determinism in LLMs. Our research adopts a hybrid approach: using LLMs for content generation (where variation is desired) but using explicit rules for decisions (where consistency is needed). This approach offers a balance between realism and experimental control.

B. Cognitive Biases in Online Purchase Decisions

Consumer psychology literature identifies several cognitive biases relevant to online purchase decisions. Tversky and Kahneman [3] identified various heuristics and biases in decision-making under uncertainty, highly relevant in online review evaluation context.

Anchoring Effect: Excessive reliance on first-encountered information, such as aggregate ratings. Thaler and Sunstein

[4] demonstrated how anchoring can influence economic decisions. We operationalize this through decision rules giving high weight to ratings.

Social Proof: Tendency to follow majority behavior, as explained by Cialdini [5]. Our rules use positive/negative counts as proxies for social proof.

Information Overload: Paradoxically, more information doesn't always produce better decisions. Schwartz [8] demonstrated that too many choices can cause decision paralysis and reliance on simple heuristics. Our findings on Careful persona vulnerability confirm this phenomenon.

Price-Risk Tradeoff: Consumers more willing to risk on cheap products. We implement this through conditional rules that relax criteria on low-price products.

Our research operationalizes these biases through explicit decision rules different for each persona.

C. MESA Framework for Agent-Based Modeling

MESA is a Python framework for agent-based modeling providing infrastructure for scheduling, data collection, and visualization [6]. This framework was chosen because: (a) open source and well-documented, (b) modular architecture facilitating LLM integration, (c) built-in data collection and analysis tools, and (d) active community with extensive examples.

Our research leverages MESA as backbone for managing simulation, while adding an LLM integration layer for content generation.

III. Methodology

A. System Architecture: Rule-Guided LLM ABM

We developed a simulation system integrating the MESA framework (Python-based ABM) with Ollama for local LLM inference [7]. The fundamental difference from pure rule-based ABM is using LLMs for **natural language content generation**, while **decision logic** uses explicit rules.

System architecture consists of five main components:

1) **MESA Model:** Manages agent scheduling, data collection, and simulation iteration.
2) **Ollama LLM Server:** Runs Llama 3.1 8B locally for fast inference.
3) **LLM Client Wrapper:** Python module handling communication with Ollama via HTTP API.
4) **Rule-Guided Prompt Templates:** Prompt templates containing personality descriptions AND decision rules.
5) **Data Collection System:** Records all reviews, decisions, and reasoning in CSV format.

B. Rule-Guided Prompt Design

1) Genuine Reviewer Agents: We implement three reviewer personalities using different prompts. LLMs generate **natural review text**, while **ratings are determined by quality-to-rating mapping rules**.

Distribution: 4 Critical + 4 Balanced + 4 Lenient = 12 genuine reviews per product per iteration.

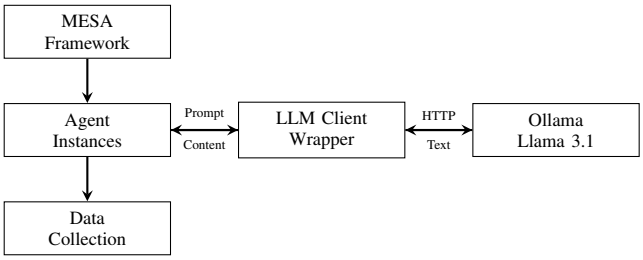

Fig. 1: Rule-guided LLM ABM system architecture

TABLE I: Quality-to-Rating Mapping Rules

Avg Quality	Critical	Balanced	Lenient
≥ 9.0	4–5	5	5
7.5–8.9	3–4	4	4–5
6.0–7.4	2–3	3	3–4
4.5–5.9	1–2	2–3	2–3
< 4.5	1	1–2	1–2

2) Fake Reviewer Agents: Fake reviewers use LLMs with system prompts and 10 user prompt variations to generate natural-appearing reviews. All fake reviews are constrained to 5-star ratings, but vary in opening style, content focus, and writing tone to avoid detection.

3) Shopper Agents (Rule-Guided Decision Making): Shopper agents are the most complex component. They use **explicit IF-THEN decision rules** to make decisions, but **LLMs generate reasoning text** explaining decisions.

Impulsive Shopper Rules:

```
RULES (check in order, use FIRST match):
1. IF rating < 2.3 -> NO_BUY (terrible)
2. IF rating >= 3.8 -> BUY (high rating!)
3. IF rating >= 3.2 -> BUY (decent, try it)
4. IF rating >= 2.8 AND price <= 250k -> BUY
5. IF rating >= 2.5 AND positive_count >= 8
   -> BUY
6. OTHERWISE -> NO_BUY
```

Listing 1: Impulsive Decision Rules (excerpt)

Careful Shopper Rules prioritize rating + review volume thresholds without detection mechanism. This causes false confidence: once rating reaches 4.2+ (due to fake reviews), Careful immediately BUY with high confidence.

Skeptical Shopper Rules with Burst Detection:

```
COMPUTED METRICS:
- burst_ratio = recent_5stars / recent_total
- rating_jump = current - old_rating

DECISION FRAMEWORK (5 steps):
STEP 1: Is genuinely BAD?
- IF rating < 2.3 -> NO_BUY
- IF negative_count >= 8 -> NO_BUY

STEP 2: Is OBVIOUSLY MANIPULATED?
- IF burst_ratio >= 70% AND rating_jump > 1.0
  AND rating < 4.0 -> NO_BUY

STEP 3: Is genuinely GOOD?
- IF rating >= 4.4 AND negative_count <= 3
  -> BUY
- IF rating >= 4.2 AND negative_count <= 4
  AND burst_ratio < 60% -> BUY
...
DEFAULT: NO_BUY
```

Listing 2: Skeptical with Burst Detection (excerpt)

Key differences across personas:

TABLE II: Persona Characteristics Comparison

Characteristic	Impulsive	Careful	Skeptical
Reviews read	3	10	15
Decision rules	6 simple	10 detailed	5-step
Burst detection	No	No	Yes
Price sensitivity	High	Medium	Medium
Rating threshold	Low (2.8+)	Med (3.5+)	High (3.7+)
Core weakness	Low thresh	No detect	Borderline

C. LLM Configuration and Justification

TABLE III: LLM Temperature Configuration

Function	Temp.	Justification
Genuine Reviewer	0.6	Natural text variation
Fake Reviewer	0.7	Avoid identical patterns
Shopper Reasoning	0.3	JSON output consistency

Temperature Settings Justification:

- **0.6–0.7 for Reviewers:** Higher temperature produces greater variation in writing style, creating more natural and diverse reviews.
- **0.3 for Shopper Reasoning:** Lower temperature ensures consistent and reliable JSON output.

Model Choice - Llama 3.1 8B:

- Open source for replication without API costs
- Local inference avoids rate limits
- Sufficient capability for natural language generation tasks
- Cost-effective: can run on consumer-grade GPU

D. Interaction Mechanism and Simulation Flow

Each simulation iteration follows a three-phase flow:

1) **Phase 1 - Review Generation (LLM):**
   - 12 genuine reviews per product (4C + 4B + 4L)
   - Iterations 4-5: add 40 fake reviews (burst)
   - Iterations 6+: add 15 fake reviews (maintenance)
2) **Phase 2 - Shopping Decisions (Rule-Based):**
   - Compute metrics: rating, negative_count, burst_ratio, rating_jump
   - Apply decision rules per persona
   - Generate reasoning text with LLM (temp 0.3)
   - Total: 300 decisions per iteration (5 products × 60 shoppers)
3) **Phase 3 - Data Collection:**
   - Record reviews with metadata
   - Record decisions with reasoning
   - Update aggregate rating

Aggregate Rating Formula:

Aggregate product rating is calculated as simple average of all reviews received up to iteration t:

$$R_{product}^{(t)} = \frac{\sum_{i=1}^{n^{(t)}} r_i}{n^{(t)}} \quad (1)$$

where $R_{product}^{(t)}$ is aggregate rating at iteration t (scale 1–5), r_i is individual rating from review i (integer 1–5), and $n^{(t)}$ is total reviews received up to iteration t.

This formula explains why burst attacks are effective: fake reviews with rating 5 dominate the numerator, "overwhelming" genuine reviews with low ratings. With 40:60 ratio (fake:genuine), rating jumps from 1.8 to 3.1 – crossing thresholds for several rules, especially for cheap products.

IV. Experimental Setup

A. Simulation Parameters

Table IV summarizes main parameters used in simulation to ensure realistic representation of e-commerce dynamics.

TABLE IV: Main Simulation Parameters

Parameter	Value
Total Iterations	20
Genuine Reviews/Product	12
Shoppers/Product/Persona	20
Burst Iterations	4, 5
Burst Volume	40/iter
Maintenance Volume	15/iter
Target Products	2 (BudgetBeats, CS)
LLM Model	Llama 3.1 8B

Simulation designed with 20 iterations to observe temporal dynamics of fake review attacks, with each product receiving 12 genuine reviews as baseline. Each persona has 20 shoppers evaluating products throughout simulation, yielding $n = 400$ total evaluations per persona.

B. Computational Infrastructure

Simulation ran on workstation with specifications shown in Table V.

TABLE V: Hardware Specifications for Simulation

Component	Specification
GPU	NVIDIA RTX 3060 12GB
CPU	AMD Ryzen 5 7600X
RAM	32GB DDR5
Storage	NVMe SSD 1TB
OS	Windows 11
LLM Runtime	Ollama v0.13.0 Llama 3.1 8B

Computational Performance:

- **Total LLM calls:** 7,505 inference calls
- **Average inference time:** ∼1.3 seconds per call
- **Total runtime:** 2 hours 45 minutes for complete simulation
- **Memory usage:** Peak ∼8GB VRAM, ∼12GB RAM

C. Product Characteristics

Target products chosen because of low quality (avg ¡ 5.5) and attractive prices, making them ideal candidates for fake review attacks aimed at increasing perceived value.

TABLE VI: Product Characteristics and Quality Attributes

Product	Qual.	Price	Sound	Build	Avg	Target
SoundMax Pro	High	450k	0 detik MBG	0 detik MBG	0 detik MBG	–
AudioBlast	Med-Hi	350k	7.5	7.0	7.5	–
BudgetBeats	Low	150k	4.0	3.5	4.4	✓
TechWave	Premium	650k	9.5	9.0	9.2	–
ClearSound	Low-Med	250k	5.0	4.5	5.2	✓

TABLE VII: Fake Review Attack Timeline

Period	Iterations	Volume	Goal
Baseline	1–3	0	Natural rating
Burst Attack	4–5	40/iter	Rating spike
Maintenance	6–20	15/iter	Sustain rating

D. Attack Scenario and Timeline

Attack strategy designed in three phases: (1) *baseline period* to establish natural behavior, (2) *burst attack* to quickly increase rating, and (3) *maintenance phase* to maintain high rating without appearing suspicious.

E. Statistical Analysis Methods

To answer RQ1 and RQ2, we use several statistical methods.

1) RQ1: Chi-Square Test for Independence: To test whether there is significant difference in conversion rate between baseline and attack periods, we use Chi-Square test with 2×2 contingency table:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

where O_i is observed frequency and E_i is expected frequency if no relationship exists.

2) Effect Size: Cramér's V: Chi-Square test only shows whether difference is statistically significant, but doesn't show how large the effect is. To measure *strength of association*, we use Cramér's V:

$$V = \sqrt{\frac{\chi^2}{n}} \tag{3}$$

Interpretation: $V < 0.1$ (negligible), $0.1 \leq V < 0.3$ (small), $0.3 \leq V < 0.5$ (medium), $V \geq 0.5$ (large effect, practically significant).

3) RQ2: ANOVA for Multi-Group Comparison: To compare conversion rates across three personas (Impulsive, Careful, Skeptical) during attack period, we use one-way ANOVA:

$$F = \frac{MS_{between}}{MS_{within}} = \frac{SS_{between}/(k-1)}{SS_{within}/(N-k)} \tag{4}$$

where $MS_{between}$ is mean square between groups, MS_{within} is mean square within groups, $k = 3$ groups, and N is total sample size.

4) Conversion Rate and Absolute Increase: Main metrics for RQ1 and RQ2:

$$CR(\%) = \frac{N_{BUY}}{N_{total}} \times 100 \tag{5}$$

$$\Delta_{pp} = CR_{attack} - CR_{baseline} \tag{6}$$

We report in *percentage points* (pp), not *percent change*.

With sample size $n = 400$ per persona and significance level $\alpha = 0.05$, our simulation has statistical power > 0.9 to detect medium effect size ($V \geq 0.3$).

V. Results

A. RQ1: Impact of Fake Reviews on Conversion Rate

Our analysis shows dramatic increases in conversion rates for low-quality products targeted by attacks (see Fig. 2):

TABLE VIII: RQ1: Fake Review Impact on Conversion Rate

Product	Baseline	Burst	Δ (pp)	χ^2 (p)
BudgetBeats	0.0%	54.2%	+54.2	121.30***
ClearSound	0.0%	71.7%	+71.7	177.35***

*** $p < 0.001$; Cramér's V > 0.6 (large effect)

Key Findings RQ1:

1) BudgetBeats: 0% → 54.2% ($\chi^2 = 121.30$, $p < 0.0001$, V = 0.636)
2) ClearSound: 0% → 71.7% ($\chi^2 = 177.35$, $p < 0.0001$, V = 0.769)
3) Very large effect size (V > 0.6) for both products
4) Post-burst conversion reached peaks of 79% (BudgetBeats) and 76% (ClearSound)

Mechanism: Ratings jumped from ∼1.8–2.1 stars to ∼3.1–3.3 stars during burst, continuing to rise to ∼3.5–3.7 stars, crossing rule thresholds especially for cheap products.

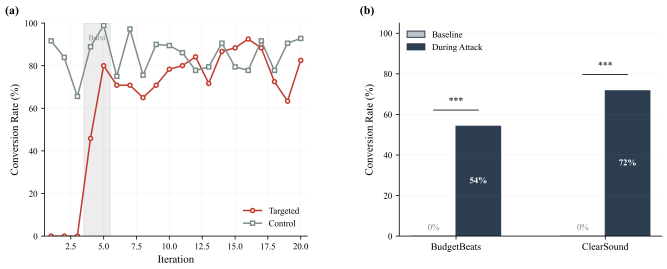

Fig. 2: RQ1: Fake review attack impact on conversion rate. (a) **Temporal dynamics** showing evolution across 20 iterations. Red line (targeted products) shows dramatic spike from 0% to 77% peak during burst (shaded gray region), then fluctuates 65–90% during maintenance. Gray line (control products) maintains stable high conversion (∼90–100%). (b) **Baseline vs burst comparison** shows significant increases: BudgetBeats +54% (0%→54%, $p < 0.0001$***), ClearSound +72% (0%→72%, $p < 0.0001$***).

B. RQ2: Differential Vulnerability Across Personas

Differential vulnerability analysis reveals **critical findings on detection mechanism effectiveness**: Skeptical persona with *burst detection rules* proved *far more resistant* with only 40.4% conversion rate – dramatically lower than Impulsive

(92.5%) and Careful (95.0%). Surprisingly, Impulsive and Careful show practically *identical* vulnerability (only 2.5pp gap, $p = 0.18$ not significant), despite Careful reading about 3x more reviews. This demonstrates that **burst detection is the only factor that truly matters** in reducing vulnerability – without detection mechanism, information volume provides no protection (see Fig. 3):

TABLE IX: RQ2: Persona Vulnerability to Fake Reviews

Rank	Persona	Baseline	Attack	Impact (pp)
1	Skeptical	0.0%	**40.4%**	**+40.4**
2	Careful	0.0%	95.0%	+95.0
3	Impulsive	0.0%	92.5%	+92.5

ANOVA: $F = 540.28$, $p < 0.0001$***
Skeptical MOST RESISTANT (40.4%) with burst detection
Careful vs Impulsive: 2.5pp gap, $p = 0.18$ (NOT significant)

Key Findings RQ2:

1) **Skeptical (40.4%) – MOST RESISTANT:** Burst detection rules (burst_ratio $\geq$ 70%, rating_jump $>$ 1.0) effectively reduce vulnerability by **52–55 percentage points** compared to Impulsive/Careful. This 52pp gap is highly significant ($p < 0.0001$), proving that *temporal pattern detection is the most effective defense*.
2) **Skeptical Effectiveness Mechanism:**
   - 59.6% decisions triggered NO_BUY (40.4% BUY = vulnerability rate)
   - Detection successfully identifies anomalies: rating jump 1.8→3.1 + high burst_ratio
   - 5-step decision framework with explicit temporal checks
3) **Impulsive (92.5%) and Careful (95.0%) – Practically Identical:** Only 2.5pp gap ($p = 0.18$ not significant), showing that **without detection mechanism, reading more reviews provides no protection**. Careful is even slightly more vulnerable because:
   - False confidence from "high rating + many reviews = safe"
   - Rules prioritizing volume thresholds without temporal analysis
   - Information overload without detection framework → overreliance on simple aggregates
4) **ANOVA Results ($F = 540$, $p < 0.0001$):** Highly significant differences across personas. Post-hoc Tukey HSD:
   - Skeptical vs Careful: $\Delta = 54.6$pp, $p < 0.0001$
   - Skeptical vs Impulsive: $\Delta = 52.1$pp, $p < 0.0001$
   - Careful vs Impulsive: $\Delta = 2.5$pp, $p = 0.18$ (NOT significant)

C. Temporal Dynamics

Temporal analysis reveals clear patterns in conversion rate evolution throughout simulation (Fig. 4):

Three Phases:

1) ***Pre-Burst* (Iterations 1–3):** Rating 1.8–2.1 stars, 0% conversion for target products

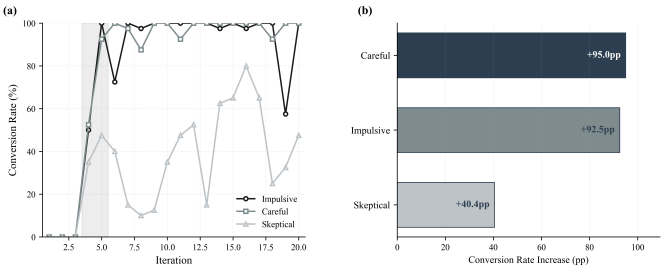

Fig. 3: RQ2: Differential vulnerability across consumer personas. (a) **Temporal dynamics per persona** shows Skeptical (light gray) consistently lower at 10–80% range with high volatility, reflecting "*tug-of-war*" between burst detection rules (NO_BUY) and price-based rules (BUY). Conversely, Careful (black) and Impulsive (dark gray) almost perfectly *overlap* at 90–100% throughout attack period, showing that without detection mechanism, reading more reviews is not protective. (b) **Increase comparison**: Skeptical +40.4pp (MOST RESISTANT with 52pp gap from others), Impulsive +92.5pp, Careful +95.0pp (2.5pp gap not significant, $p = 0.18$). Critical finding: *burst detection rules* effectively reduce vulnerability >50pp, while information quantity without detection framework is not protective.

2) ***Burst* (Iterations 4–5):** Rating spike to 2.8–3.3 stars, conversion jumps to 54–72%
3) ***Maintenance* (Iterations 6–20):** Rating gradually rises to 3.2–3.7 stars, conversion fluctuates 65–90%

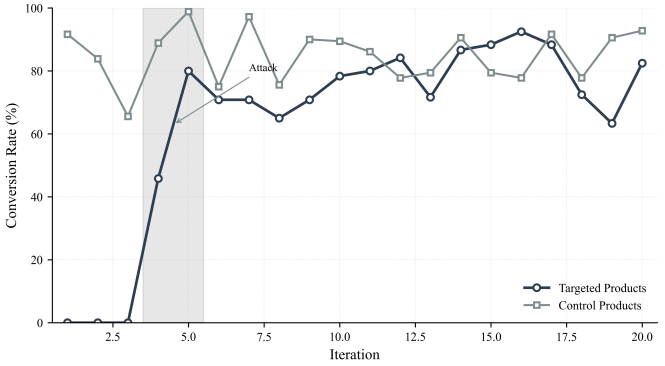

Fig. 4: Temporal dynamics of conversion rate (20 iterations). *Shaded region* = *burst* phase (iterations 4–5). Target products: 0%→77%→65–90%. Control products: stable 75–98%.

VI. Discussion

A. Main Finding Interpretation

1) Extreme Effectiveness of Burst Attacks: The 54–72% increase in conversion rates is explained by: (1) **Overwhelming Volume:** 3.3:1 ratio (fake:genuine) dominates review pool, (2) **Threshold Crossing:** Rating jump from 1.8→3.1 crosses decision thresholds for cheap products, (3) **Metric Manipulation:** Positive count spikes, negative count diluted,

and (4) **Persistence of Effect:** Maintenance phase (15 fake reviews/iter) sufficient to sustain high rating.

2) Burst Detection Rules Effectiveness and Universal Vulnerability Without Detection: The most important finding is that **Skeptical persona with burst detection rules achieves dramatic resistance (40.4%) compared to Impulsive/Careful (92–95%)**, with 52–55 percentage point gap highly significant ($p < 0.0001$). This proves that **temporal pattern detection is the most effective defense**.

Why Are Burst Detection Rules Effective?

1) **Temporal Anomaly Recognition:** Burst_ratio $\geq 70\%$ detects unnatural clustering of 5-star reviews. Rating_jump > 1.0 detects suspicious spikes. 59.6% decisions successfully blocked via these explicit checks.
2) **Explicit vs Implicit Processing:** Skeptical: explicit temporal metrics computation $\rightarrow$ rule-based filtering. Careful/Impulsive: implicit trust in aggregated ratings without temporal awareness.
3) **Framework-Driven Decision Making:** 5-step decision framework with detection as gate 2 (before evaluating quality). Structured approach prevents premature confidence.

Secondary Finding: Careful and Impulsive Practically Identical

Surprisingly, Careful and Impulsive show practically *identical* vulnerability (95.0% vs 92.5%, 2.5pp gap, $p = 0.18$), despite Careful reading about 3x more reviews. This reveals counterintuitive insight: **without detection mechanism, information quantity is not protective**.

Why Is Careful Not More Resistant?

1) **Information Overload Without Detection Framework:** Reading 10 reviews doesn't help if 8–9 are fake (4:1 ratio). Careful lacks burst detection or temporal awareness to filter manipulation. Phenomenon explained by Schwartz [8]: too much information without analytical framework $\rightarrow$ reliance on simple heuristics.
2) **False Confidence from Volume:** Careful rules: "IF rating ≥ 4.2 AND total_reviews $\geq 10 \rightarrow$ BUY". Once fake reviews push rating to 3.5+ and count to 50+, Careful feels "I've done my research, high rating + many reviews = safe". Volume creates *illusion of thoroughness* without actual protection.
3) **Lack of Temporal Awareness:** Careful focuses on static metrics (current rating, total count). No consideration of temporal patterns (burst_ratio, rating_jump).

Practical Takeaway: The 52pp gap shows that **the ONLY thing that matters is detection capability, not information quantity**. Platforms and consumers must focus on implementing temporal detection mechanisms rather than encouraging "read more reviews".

B. Implications for Rule-Guided LLM Methodology

1) Trade-offs: Rules vs Pure LLM: **Rule-Guided Approach Advantages:**

- LLMs generate natural language content (reviews, reasoning)
- Rules ensure reproducible decisions
- Auditable: both rules and generated text can be inspected
- Scalable: 7,500+ LLM calls feasible in 2-3 hours

Limitations:

- Rules may oversimplify human decision complexity
- No true "emergent behavior" – all pre-specified
- Requires careful rule design for realism
- Careful persona case shows achieving realistic balanced behavior is challenging

TABLE X: ABM Approach Comparison

Aspect	Pure Rules	Pure LLM	Rule-Guided
Content realism	Low	High	High
Reproducibility	High	Low	High
Transparency	High	Medium	High
Computational cost	Low	High	Medium
Behavior predictability	High	Low	Medium

C. Practical Implications

1) For E-commerce Platforms: Results show that **temporal detection is the ONLY truly effective defense** (52pp gap). Platforms must prioritize detection systems:

1) **[CRITICAL] Implement Temporal Pattern Detection:**
   - Burst_ratio monitoring: Flag products with $\geq$60% recent reviews (last 2 weeks) being 5-star
   - Rating_jump alerts: Automatic freezing if rating increases $>$1.0 star in 7-14 days
   - Review clustering detection: Flag timestamp clustering
   - Automatic quarantine: Freeze new reviews for manual moderation when anomalies detected
2) **[HIGH PRIORITY] Consumer Warning System:**
   - Display burst_ratio badge if $\geq$ 60%: "Suspicious Review Pattern Detected"
   - Show timeline graph with highlighted anomalies
   - Color-code periods: green = stable, red = suspicious burst
   - Provide "Skeptical Mode" filter
3) **[CRITICAL] Educate Consumers – Debunk "More Reviews = Better":**
   - Warning: "High rating + many reviews does NOT guarantee authenticity"
   - Teach: "Check timeline graph FIRST, then read content"
   - Case study: "Products with 100+ reviews can be 70%+ fake"

2) For Consumers: **Adopt Skeptical Persona Strategies:** Results show Skeptical strategies reduce vulnerability 52pp vs reading more reviews (2.5pp gap).

- **[PRIORITY] Check review timeline graph** – stable growth trustworthy, sudden spikes suspicious
- **[PRIORITY] Monitor rating jump** – if rating increases $>$1.0 star in short window, likely manipulation

- Count technical specifics – genuine reviewers mention 3+ specific features
- Read 3–4 star reviews first – more balanced than 5-star
- BEWARE false confidence from "many reviews" – volume not protective if majority fake
- Adopt explicit detection checklist before purchase

Key Takeaway: Shift mindset from "I need to read many reviews" to "I need to detect temporal anomalies". 3 reviews with detection checklist $>$ 10 reviews without detection (proven by 52pp gap).

3) For Regulators:

- Mandate AI-generated content disclosure
- Require platforms to implement temporal pattern detection
- Standardize transparency metrics: burst_ratio, verified purchase %, temporal distribution
- Significant penalties for undisclosed fake campaigns

VII. Limitations and Future Work

A. Research Limitations

1) Rule Design Challenges: Designing realistic decision rules requires domain expertise about consumer behavior and iterative testing to balance sensitivity. The Careful persona case shows that achieving "realistic vulnerability" balanced with "reasonable baseline behavior" is very challenging.

2) Prompt Engineering Trade-offs: Main difficulty is designing rules that balance: (a) **Realism** - real consumers not paranoid but not naive, (b) **Heterogeneity** - personas must differ meaningfully, and (c) **Expressiveness** - text prompts have limitations encoding complex conditional logic.

3) Single Run Limitation: Single run simulation due to computational constraints (2-3 hours for 7,500+ LLM calls). Multiple runs ($n = 5$–10) would provide confidence intervals, sensitivity analysis, and robustness checks. However, with large sample size per persona ($n = 400$ evaluations/persona), we expect minimal variability across runs for aggregate statistics.

4) Simulation Simplifications:

- Single product category (headphones) – generalizability to other categories unknown
- Static prices – no dynamic pricing or promotional effects
- Simplified review metadata – no reviewer reputation, helpful votes, images
- No platform defense agents – real platforms have evolving detection systems
- Shopper personas don't learn or adapt – rules fixed throughout 20 iterations

5) External Validity: Findings may not fully generalize because: (a) real consumers more diverse than 3 discrete personas, (b) real platforms have detection systems, (c) real fake campaigns more adaptive, (d) real consumers have access to external validation sources, and (e) cultural differences in trust levels not modeled.

B. Future Research

1) Validation and Refinement:

1) **Human Subject Experiments:** Validate that rule-based persona behaviors match real human patterns. Test whether real consumers exhibit "Careful paradox" (reading more $\rightarrow$ higher vulnerability without detection).
2) **Multiple Simulation Runs:** $n = 10$ runs with random initialization. Report confidence intervals for all key metrics. Sensitivity analysis varying LLM temperature, prompt wordings, rule thresholds.
3) **Rule Optimization:** Systematic parameter sweeps to find optimal Skeptical rules (target: $<$20% vulnerability). Grid search over threshold combinations for Careful rules balancing baseline vs attack.

2) Methodology Extensions:

1) **Adaptive Rules:** Rules that evolve based on shopper experiences. Threshold adjustment based on reinforcement learning from outcomes. Model consumer learning: after being deceived once, do they become more skeptical?
2) **Hybrid LLM Decisions:** Use LLMs for borderline cases, rules for clear-cut cases. LLM generates persona-specific rules based on personality descriptions.
3) **Multi-LLM Heterogeneity:** Different LLM models for different personas. Test hypothesis: larger models generate more nuanced, harder-to-detect fake reviews.
4) **Platform Defense Agents:** Model detection systems as agents with their own learning. Co-evolutionary dynamics: attackers adapt to detection, detection adapts to attacks.

3) Domain Extensions:

1) **Multi-Category Products:** Electronics, fashion, supplements, services. Test whether Careful paradox generalizes across categories.
2) **Multi-Modal Reviews:** Include image/video reviews in simulation. Test whether visual content increases or decreases vulnerability.
3) **Cross-Cultural:** Replicate with different languages (Chinese, Spanish). Test cultural differences in trust levels and skepticism thresholds.
4) **Social Network Effects:** Model word-of-mouth: shoppers share experiences with friends. Test whether social validation can overcome fake reviews.

VIII. Conclusion

This research demonstrates the significant impact and counterintuitive vulnerability patterns of LLM-generated fake reviews on consumer purchase decisions in e-commerce, using a novel rule-guided LLM Agent-Based Modeling approach. We combine LLM advantages (natural language generation) with explicit decision rules (reproducibility and transparency).

Main Findings:

1) **RQ1 – Burst Attack Effectiveness:** Fake review campaigns increase low-quality product conversion rates by 54–72% ($p < 0.0001$, Cramér's V > 0.6). Fake review volume (40:60 fake:genuine ratio during burst)

transforms aggregate rating from $\sim$1.8 to $\sim$3.1 stars, crossing decision rule thresholds especially for cheap products.

2) **RQ2 – Dramatic Burst Detection Rules Effectiveness: Skeptical persona with burst detection rules proved far more resistant with 40.4% conversion rate** – dramatically lower than Impulsive (92.5%) and Careful (95.0%). The 52–55 percentage point gap is highly significant ($F = 540$, $p < 0.0001$), proving that **temporal pattern detection is the most effective defense**. Detection rules successfully block 59.6% decisions via explicit temporal anomaly recognition.
3) **Secondary Finding – Universal Vulnerability Without Detection:** Surprisingly, Careful and Impulsive show practically *identical* vulnerability (95.0% vs 92.5%, 2.5pp gap, $p = 0.18$ not significant), despite Careful reading about 3x more reviews. This challenges the assumption that "more information leads to better decisions". We attribute identical vulnerability to *lack of detection mechanisms* – without temporal awareness, information quantity is not protective. High volume even creates false confidence.
4) **Temporal Dynamics:** Burst phase (iterations 4-5) creates 0%$\rightarrow$54-72% conversion spike, and maintenance phase (15 fake reviews/iter) effectively increases ratings gradually to 3.2–3.7 stars throughout iterations 6–20.

Core Insight: In the age of LLM-generated misinformation, the paradigm must shift from "read more reviews" to "detect temporal anomalies". *Only detection capability matters – information quantity is not protective without it*. The 40.4% vulnerability for Skeptical shows that even with detection, some vulnerability remains. However, 52pp reduction proves that detection-first approach is the path forward for maintaining review system integrity in the generative AI era.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

Conceptualization, I.G.T. and M.I.A.M.; methodology, I.G.T.; software, I.G.T. and A.F.Z.; validation, I.G.T., M.I.A.M., and A.F.Z.; formal analysis, I.G.T.; investigation, I.G.T.; resources, I.G.T.; data curation, I.G.T.; writing—original draft preparation, I.G.T.; writing—review and editing, M.I.A.M. and A.F.Z.; visualization, I.G.T.; project administration, I.G.T.

Acknowledgment

The authors thank the Department of Informatics Engineering, Hasanuddin University for providing computational resources and supporting this research.

References

[1] A. Ferraro et al., "Agent-Based Modelling Meets Generative AI in Social Network Simulations," in *Proc. IEEE/ACM ASONAM*, 2024. [Online]. Available: arXiv:2411.16031

[2] J. S. Park et al., "Generative Agents: Interactive Simulacra of Human Behavior," in *Proc. UIST '23*, San Francisco, CA, USA, 2023, pp. 1–22, doi: 10.1145/3586183.3606763.

[3] A. Tversky and D. Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science*, vol. 185, no. 4157, pp. 1124–1131, Sept. 1974, doi: 10.1126/science.185.4157.1124.

[4] R. H. Thaler and C. R. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven, CT, USA: Yale University Press, 2008.

[5] R. B. Cialdini, *Influence: The Psychology of Persuasion*, Revised Edition. New York, NY, USA: Harper Business, 2021.

[6] MESA Contributors, "MESA: Agent-Based Modeling in Python 3," 2024. [Online]. Available: https://mesa.readthedocs.io/

[7] Ollama Team, "Ollama: Run Large Language Models Locally," 2024. [Online]. Available: https://ollama.ai/

[8] B. Schwartz, *The Paradox of Choice: Why More Is Less*. New York, NY, USA: Harper Perennial, 2004.