LLM 기반 에이전트 사회에서 외부 자극에 따른 협력의 붕괴와 회복 과정 정량 분석

소훈*, 황수호*, 윤영준*, 정현준**

Quantitative Analysis of Cooperation Collapse and Recovery under Shock in LLM-Based Agent Societies

Hoon So*, Suho Hwang*, Yoon Youngjoon* and Hyunjun Jung**

요 약

대규모 언어 모델(Large Language Model, LLM)은 사회적 시뮬레이션에서 자율 에이전트로 활용되며, 협력과 규범 형성 같은 집단적 상호작용을 재현하는 데 활용되고 있다. 이 연구는 외부 충격을 사건이 아닌 설계 변수로 다루는 실험 프레임인 VOW-ICD를 제안하고, 충격의 강도(intensity)와 지속시간(duration)이 LLM 기반 에이전트 사회에서 협력의 붕괴와 회복에 미치는 영향을 정량 분석하였다. 실험 결과, 강한 충격은 협력률(cooperation_rate) 하락과 회복 지연(average_recovery_time)을 초래했으며, 지속시간이 길수록 shock 이후 신뢰(post_shock_trust)의 안정성이 높게 유지되었다. 또한 메시지와 행동 간의 괴리(message_action_mismatch)는 강도 증가에 따라 확대되었고, 리더십 기여(avg_contribution) 분석에서는 충격이 클수록 리더십이 분산되는 경향이 나타났다. 제안 시스템은 단일 로컬 LLM 기반의 반복 실험이 가능하며, 재현성과 비용 효율 측면에서 실용적 실험 틀로 작용할 수 있다.

Abstract

Large Language Models (LLMs) are increasingly used to simulate autonomous agents in virtual societies, enabling the study of collective phenomena such as cooperation, norm formation, and trust. This study proposes the VOW-ICD framework, which treats external shocks not as singular events but as tunable design variables—specifically their intensity and duration—to quantify the dynamics of cooperation collapse and recovery. Experimental results show that stronger shocks significantly reduce the cooperation rate and delay recovery, while longer-duration shocks contribute to more stable post-shock trust. Moreover, message-action mismatch increases with shock intensity, revealing the risk of performative cooperation. Contribution-based analysis also indicates that strong shocks lead to decentralized leadership and role turnover. The proposed architecture, implemented using a local LLM, supports repeatable and cost-efficient experiments, offering a practical and reproducible platform for analyzing social responses under varying shock conditions.

Key words

LLM, Multi-Agent Simulation, Social Cooperation, Shock Dynamics, Experimental Framework

* 군산대학교 소프트웨어학과 학사과정, hoonso20@kunsan.ac.kr, 2201347@kunsan.ac.kr, 2201067@kunsan.ac.kr,

** 군산대학교 소프트웨어학과 교수(교신저자), junghj85@kunsan.ac.kr

※본 연구는 2025년도 과학기술정보통신부 및 정보통신기획평가원의 "SW중심대학사업" 지원을 받아 수행되었음(2023-0-00065)

Ⅰ. 서 론

대규모 언어 모델(Large Language Model, LLM)은 언어 이해와 생성 능력을 바탕으로 다양한 자율 시뮬레이션 환경에서 활용되고 있다. 특히 다수의 LLM 기반 에이전트를 하나의 사회 시스템처럼 구성해 규범 형성, 정보 확산, 협력 구조 등을 관찰하는 연구 사례가 증가하고 있다[1][2][3]. 최근에는 장기 기억과 반영(reflection) 기능을 갖춘 에이전트를 활용해 일상적 상호작용과 집단 이벤트를 시뮬레이션하거나[1], 공공재 기반 환경에서 협력의 붕괴 현상을 탐색하거나[2], 실물 경제 충격을 재현해 회복 전략을 분석하는 시도도 보고되고 있다[3].

이 연구에서의 'Shock'은 감정적 사건이 아닌, 실험자가 설계 가능한 외부 개입 변수(intensity × duration)로 정의된다.

이러한 연구들은 다중 에이전트의 자율적 상호작용을 통해 사회적 현상을 관찰하는 데 중점을 두었으며, 협력과 규범의 자발적 형성, 다수 의견에 대한 언어적 동조, 장기 기억의 유지 등의 주제를 다루어 왔다[4][5][6]. 그러나 대부분의 연구는 외부 충격 이후의 에이전트 반응을 현상적으로 기술하는 데 머무르며, 충격의 강도와 지속시간을 설계 변수로 체계적으로 조합하여 협력 붕괴 및 회복을 정량 비교한 사례는 확인되지 않았다.

또한 언어적 약속이 실제 행동으로 이어지는지를 평가할 수 있는 지표가 부족하여, 에이전트가 "협력하자"고 발화한 메시지가 실제 자원 기여로 연결되었는지를 판단하기 어려운 문제가 있다. 협력은 단순한 선언이 아닌 실천을 필요로 하며, 이러한 언어-행동 간의 괴리를 정량적으로 측정하는 틀은 아직 충분히 구축되지 않았다.

이 연구는 위와 같은 공백을 해결하기 위해 다음의 목적을 갖는다. 첫째, 자율 에이전트 시뮬레이션 환경인 Village of Words(VOW)에서 외부 충격의 강도와 지속시간을 교차 조합한 시나리오를 설계하고, 이를 기반으로 협력률, 회복 시간, 신뢰 변화 등의 사회적 반응을 정량적으로 비교한다. 둘째, 언어적 약속과 실제 행동의 일치 여부를 측정하기 위해 메시지-행동 불일치율이라는 지표를 정의하고, 충격 속성 변화에 따른 괴리 양상의 변화를 분석한다. 셋째, 각 에이전트의 페르소나에 따른 리더십 전환을 기여량 및 신뢰도 지표를 통해 추적함으로써 사회 내 역할 변화의 양상을 탐색한다. 마지막으로, 로컬 환경에서 실행 가능한 LLM을 기반으로 반복 실험을 수행함으로써 외부 API 의존도를 줄이고, 비용 효율성과 재현성을 동시에 확보한다[7].

이 연구의 기여는 충격의 강도와 지속시간을 독립 변수로 설정하고, 협력 동역학을 수치 기반 지표로 정량적으로 분석함으로써 기존의 기술 중심 시뮬레이션 연구에서 실험 기반 계량 분석으로의 전환 가능성을 제시한다. 이를 통해 LLM 기반 자율 사회 내에서의 협력 형성, 붕괴, 회복 과정을 실험적으로 비교하고, 언어와 행동 사이의 괴리까지 아우르는 해석틀을 구축하는 것을 목표로 한다.

Ⅱ. 관련 연구

2.1 가상 사회 시뮬레이션과 충격 모델링의 공백

Generative Agents는 장기 기억과 일과 계획 구조를 갖춘 에이전트들이 자율적으로 일상을 구성하는 방식을 제시하였으나, 이벤트 충격은 대부분 사전 스크립트에 의해 고정된 형태로 처리되었다[1]. GovSim은 공공 자원 관리 실패를 모사하며 집단 행동 변화를 관찰했지만, 외부 충격의 강도나 지속시간을 설계 변수로 조정하지는 않았다[2]. Shachi 또한 단일 세율 충격을 실험 요소로 포함했지만, 강도·지속시간을 반복 조정하거나 장기 영향력을 계량하는 구조는 없었다[3]. 이 연구는 동일한 페르소나 구성을 유지한 채 강도와 지속시간을 교차 조합함으로써 이 공백을 보완하고자 한다.

2.2 사회적 동조와 메시지 편향 연구

언어적 협력 선언과 실제 행동의 괴리는 협력 양상을 이해하는 핵심 요소이나, 기존 연구들은 이를 정량적으로 측정하는 지표를 활용하지 않았다. LLM 기반 에이전트 집단이 다수 의견에 순응하거나

자율적으로 규범을 형성하는 경향이 보고되었지만 [4][5][6], 발화된 메시지와 실제 행동(예: 자원 기여)의 일치 여부는 체계적으로 분석되지 않았다. 이 연구는 “협력하겠다”는 발화와 실질적 기여 행동의 일치 여부를 계량하기 위해 메시지-행동 불일치율을 정의하고, 충격 강도에 따른 불일치율 증가 경향을 실험적으로 확인한다.

2.3 페르소나 기반 협상 및 협업 프레임워크

기존의 다중 에이전트 협상 연구는 페르소나 별 자원 배분 전략을 분석하거나, 역할 간 상호작용 양상을 정성적으로 기술하는 데 중점을 두었다[2][6]. 그러나 충격 이후 협력 구간에서 나타나는 기여 기반 리더십 전환이나 신뢰도 회복 양상을 정량적으로 추적한 사례는 드물다. 이 연구는 동일한 언어 모델로 구성된 에이전트 집단을 기반으로, 기여량과 신뢰도 지표를 활용해 충격 전후 리더십의 변화를 계측할 수 있도록 실험 프레임을 구성하였다.

III. 제안하는 VOW-ICD 시스템

3.1 시스템 구조

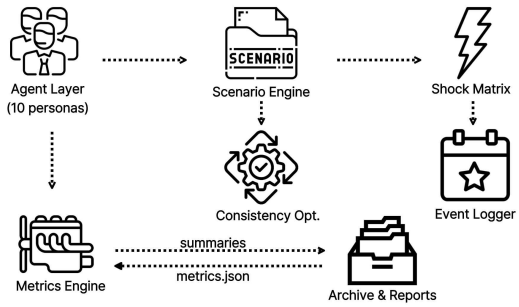

그림 1. VOW-ICD 시스템 아키텍처 개요
Fig. 1. Overview of the VOW-ICD System Architecture

이 논문이 제안하는 VOW-ICD(Intensity × Duration) 시스템은 [그림 1]이 보여주듯이 외부 충격의 강도와 지속시간을 설계 변수로 다루어 협력 동역학을 정량 분석하기 위한 최소 구성의 아키텍처이다. 에이전트 계층은 메시지 채널(발화)과 행동 채널(자원 기여)을 분리해 수집하고, 시나리오 엔진은 [그림 2]와 같이 이어지는 5단계의 흐름을 관리한다. Shock Matrix는 Shock window에 따라 강도×지속시간 조합(soft, baseline, double, extended)을 외부 이벤트로 삽입하며, 발생한 이벤트는 Event Logger로 기록된다. Consistency Optimizer는 메시지와 행동 간 불일치를 평가하고, Metrics Engine은 협력률(cooperation_rate), 평균 회복시간(average_recovery_time), shock 이후 신뢰(post_shock_trust), 메시지 행동 불일치율(message_action_mismatch), 평균 기여량(avg_contribution) 등 핵심 지표를 계산한다. 계산 결과는 metrics.json과 run_summary.csv로 저장되며, Archive & Reports 모듈은 반복 실행 로그를 보관해 후속 분석과 보고서 작성을 지원한다.

3.2 시나리오와 충격 설계

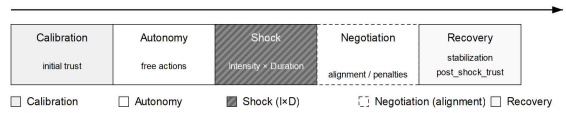

그림 2. 시나리오 충격 타임라인
Fig. 2. Scenario and Shock Timeline

표 1. Shock 설계 메트릭스(강도x지속시간, 최소 구성)
Table 1. Shock Design Matrix(Intensity x Duration, Minimal)

Variant	△stone	Duration(turns)	Injection (turn index)
Soft	-2	1	9
Baseline (single)	-3	1	9
Double (two waves)	-4	1x2	9,12
Extended	-3	3	9-11

실험은 20턴을 기본으로 하며 필요에 따라 30/40/50턴으로 확장된다. 초기에는 Calibration 단계에서 신뢰와 협력 규범을 형성하고, Autonomy 단계에서 제약을 완화하여 자율 행동을 관찰한다. Shock 단계에서 외생 충격을 투입하고, Negotiation 단계에서 메시지 행동 정합성을 높이는 규칙을 적용해 진정성 있는 합의를 유도한다. Recovery 단계에서는 충격 이후의 회복과 안정화 과정을 추적한다. Shock의 강도와 지속시간은 soft/baseline/double/extended 네

가지 변형으로 조합되며, 각 변형의 시간 창과 적용 방식은 [그림 2]와 [표 1]에 정리하였다.

3.3 지표와 분석 절차

협력률은 턴별 협력 행동 비율의 평균으로 정의하고, 평균 회복시간은 Shock 창 종료 이후 정상화 임계치(전충격 평균의 90%)에 재도달하기까지의 평균 턴 수로 측정한다. post_shock_trust는 Shock 이후 10턴의 평균 신뢰로 요약하며, 메시지 행동 불일치율은 “협력” 취지의 발화가 있었음에도 행동이 협력이 아닌 비율로 계산한다. 평균 기여량은 에이전트별/조건별 자원 기여의 평균이다. 조건별 비교는 평균과 표준편차를 기본으로 하고 반복 실행(≥3회) 간 변동을 함께 제시한다. 장기 러닝에서는 UNKNOWN 비율 임계치를 두어 결과의 신뢰성을 확보한다.

3.4 입증 결과

강도 효과를 확인하기 위해 변형 간 협력률과 평균 회복 시간을 비교하였다. Shock window의 위치와 강도×지속 조합은 [그림 2]에 정리되어 있으며, 결과 요약은 [그림 3]에 제시한다. double 변형은 협력률이 가장 낮고(예:0.627±0.030), 평균 회복 시간이 가장 길어(예: 2.95턴) 충격 강도가 협력 붕괴 폭과 회복 지연을 동시에 키운다는 점을 보여준다. 반대로 soft 변형에서는 협력률이 높고(예: 0.703±0.047), 평균 회복 시간이 가장 짧아(예: 1.46턴) 완만한 충격이 빠른 복원을 돕는 경향이 관찰되었다. 지속시간 효과에서는 extended 변형이 평균 회복시간이 다소 길지만(예: 1.93턴) post_shock_trust 평균이 높게 유지되는 양상을 보여 안정적 신뢰 복원이 가능함을 시사한다(예: 0.551). 언어와 행동의 괴리는 강한 충격에서 더 두드러졌으며, 협상 단계의 정합성 규칙 적용 시 완만히 줄어드는 경향이 확인되었다. 리더십 측면에서는 평시·완만한 충격에서 소수 역할 중심의 기여가 유지되지만, 강한 충격(double)에서는 최고 기여자가 교체되는 사례가 나타나 리더십 분산이 발생하였다. 요컨대 [그림 2]의 Shock window 설계와 [그림 3]의 지표 요약을 함께 보면, 충격 설계가 협력의 붕괴 - 회복 곡선뿐 아니라 역할 분배에도 실질적 영향을 준다는 점이 드러난다.

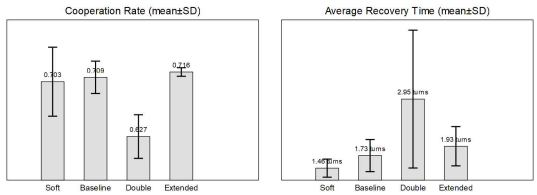

그림 3. 협력률·평균 회복시간(조건별, 평균±SD)
Fig, 3. Cooperation Rate and Average Recovery Time by Condition (mean±SD)

Ⅳ. 실험 결과 및 평가

4.1 실험 환경

표 2. 실험 환경
Table 2. Experiment environment

Software	Version
GPU	NVIDIA RTX A6000
CPU	Intel Core i9-13900K
RAM	48GB
Local LLM	meta-llama-3-8b-instruct(gguf)
Simulation Platform	VOW-ICD(Python 3.10)
LLM Runtime	LM Studio 0.3.15
Runs per Condition	3 repetition per condition(total 12)

이 연구에서 제안한 VOW-ICD 시스템은 표 2에 요약된 동일한 하드웨어 및 소프트웨어 환경에서 실행되었으며, 모든 조건을 공정하게 비교할 수 있도록 동일 설정으로 실험을 반복 수행하였다.

4.2 실험 결과

실험 결과는 VOW ICD 설계의 핵심 가설을 지지한다. 먼저 강도(intensity) 효과에서, shock 직후 `cooperation_rate` 하락과 `average_recovery_time` 지연이 강도에 비례해 커졌다. double 조건은 협력률이 가장 낮고(0.627±0.030), 회복시간이 가장 길었으며(2.95턴), 반대로 soft 조건은 협력률이 가장 높고(0.703±0.047), 회복시간이 가장 짧았다(1.46턴). 이는 강

도가 붕괴 폭과 복원 지연을 동시에 증폭한다는 점을 보여준다.지속시간(duration) 효과에서는 extended 조건에서 `post_shock_trust`가 높게 유지되어(0.551), 장기 신뢰 안정성을 확보하는 데 기여함을 시사한다. 반면 회복 속도는 상대적으로 느렸다(1.93턴).

`message_action_mismatch`는 soft < baseline < double 순으로 소폭 증가해, 강한 충격일수록 협력 발화가 실제 행동으로 이어지지 않는 경향이 확인되었다. negotiation 단계에서 괴리는 일부 줄었으나, double 조건에서는 잔존 효과가 남았다.

`avg_contribution` 기준 역할 분석에서는 강한 충격일수록 리더십이 분산되고, 최고 기여자가 교체되는 사례가 나타났다. 이는 충격 설계가 협력 회복뿐 아니라 역할 구조에도 영향을 미친다는 점을 의미한다. 모든 조건은 3회 반복 실행되었으며, 주요 지표의 분산은 제한적이었다. UNKNOWN 비율, shock 시점(t=9 vs. 10), 정합성 규칙 유무 등에서도 핵심 경향은 유지되었고, 동일 LLM(meta-llama-3-8b-instruct) 기반의 일관된 설정으로 재현성도 확보되었다.

Ⅴ. 결론 및 향후 연구

이 연구는 외부 충격을 사건이 아닌 설계 변수로 다루는 VOW-ICD 틀을 통해, 협력 동역학을 계량적으로 비교 가능한 실험 프레임을 실증하였다. 실험 결과, 강도(intensity)는 `cooperation_rate`의 단기 붕괴 폭과 `average_recovery_time`의 지연을 결정짓는 주요 요인으로 작동했고, 지속시간(duration)은 shock 이후 신뢰의 안정 수준(`post_shock_trust`)을 좌우했다. 충격 강도 증가에 따라 `message_action_mismatch`가 소폭 증가하여 언어적 합의가 행동으로 즉시 환류되지 않는 위험이 드러났고, `avg_contribution` 분석에서는 강한 충격 조건에서 최고 기여 주체가 교체되며 리더십 분산 경향이 관찰되었다.

이러한 결과는 설계적 함의를 제공한다. 단기 복원 속도와 운영 안정성을 중시할 경우에는 약한·단발 shock(soft)과 정합성 규칙을 병행하는 구성이 유리하며, 장기 신뢰의 안정화를 목표로 할 경우에는 완만하지만 지속적인 shock(extended)이 효과적일 수 있다. 즉, 목표 지표(`cooperation_rate`, `average_recovery_time`, `post_shock_trust`)에 맞춰 강도×지속시간(I×D)을 시뮬레이션 기반 정책 조정 전략 수립에 이론적 시사점을 제공한다. 또한 단일 로컬 LLM, 5단계 시나리오, shock matrix, 표준 지표 및 아카이브로 구성된 최소 아키텍처는 반복 실험과 재현성 측면에서 실용적 기반이 될 수 있음을 보여주었다.

한계도 존재한다. 단일 모델 설정에 따른 외적 타당도, 규칙 기반 `message_action_mismatch` 정의의 언어적 제약, 100턴 이상 장기 실행에서의 실패·지연 관리가 그 예이다. 후속 연구에서는 (i) reasoning 특화/일반 instruction 모델의 교차 비교를 통한 문화적 차이 정량화, (ii) 불일치율을 직접 제어하는 정합성 개입 정책 설계 및 평가, (iii) 다단계·다중 파동 shock 시나리오를 통한 임계 강도·지속값 추정으로 확장할 예정이다. 아울러 표준화된 로그·지표·아카이브 규약을 정교화하여 타 연구와의 비교 가능성도 높일 계획이다.

참 고 문 헌

[1] Park, J. S., et al. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442, 2023.

[2] Piatti, G., et al. Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents. arXiv:2404.16698, 2024.

[3] Kuroki, T., et al. Shachi: An Agent-Based Simulation Evaluating Economic Policy on Real-World Data. arXiv:2501.03920, 2025.

[4] Zhong, H., et al. Disentangling the Drivers of LLM Social Conformity: An Uncertainty-Moderated Dual-Process Mechanism. arXiv:2508.14918, 2025.

[5] Zhu, X., et al. Conformity in Large Language Models. ACL 2025.

[6] Flint Ashery, A., et al. Emergent Social Conventions and Collective Bias in LLM Populations. Science Advances, 11(20), 2025.