

FSM을 활용한 LLM 기반 게임 NPC 상태 제어 연구

나윤성*, 남영주**

A Study on LLM-based Game NPC State Control using FSM

Yunseong Na*, Youngju Nam**

요 약

최근 LLM(Large Language Model)을 활용한 게임 NPC는 플레이어의 자연어 입력에 따라 유연하고 자연스러운 대화를 제공할 수 있다. 그러나 LLM 기반 NPC의 자유로운 응답은 현재 게임 상태나 기획 의도와 맞지 않는 행동으로 이어질 수 있으며, 거래, 퀘스트, 보상, 전투와 같은 시스템과 연결될 경우 비허용 상태 전이가 발생할 수 있다. 본 논문에서는 LLM 기반 NPC의 대화 능력을 유지하면서도, 게임 규칙을 벗어난 행동이 실제 시스템에 반영되지 않도록 FSM 기반 가드레일 구조를 제안한다. 제안 구조는 LLM이 플레이어 입력을 JSON 형식의 intent와 ruleTags로 변환하고, FSM이 현재 게임 상태와 규칙을 검증한 뒤 상태 전이 허용 여부를 결정하는 방식이다. Unity 기반 NPC 상호작용 환경에서 LLM-only 방식과 LLM+FSM 방식을 비교한 결과, 비허용 입력에 대한 금지 상태 전이의 발생을 감소시켰으며, 정상 입력의 성공률을 향상시켰다.

Abstract

Recently, game NPCs using Large Language Model (LLM) can provide flexible and natural conversations depending on the player's natural language input. However, the free response of LLM-based NPCs can lead to behaviors that do not match the current game state or planning intentions, and when connected to systems such as transactions, quests, rewards, and battles, unacceptable state transitions can occur. In this paper, we propose an FSM-based guardrail structure so that behaviors outside the game rules are not reflected in the actual system while maintaining the conversation ability of LLM-based NPCs. The proposed structure is a method in which LLM converts player input into JSON-style int and ruleTags, and FSM verifies the current game status and rules before determining whether to allow state transition. As a result of comparing the LLM-only method and the LLM+FSM method in a Unity-based NPC interaction environment, the incidence of prohibited state transition for non-permissible input decreased, and the high success rate for normal input was shown.

Keywords

LLM, NPC, FSM, guardrail, invalid state transition, game artificial intelligence

* 국립군산대학교 소프트웨어학과 학사과정
- ORCID: <https://orcid.org/0009-0003-8705-1387>
** 국립군산대학교 소프트웨어학과 교수(교신저자)
- ORCID: <https://orcid.org/0000-0003-3971-0715>

• Received: May 26, 2026, Revised: Jul. 03, 2026, Accepted: Jul. 06, 2026
• Corresponding Author: Youngju Nam
Dept. Department of Software at Kunsan National University, 558,
Daehak-ro, Kunsan-si, Jeollabuk-do, Republic of Korea
Tel: +82-63-469-8919, Email: imnyj@kunsan.ac.kr

I. 서 론

LLM(Large Language Model)은 자연어 처리 분야를 넘어 게임 콘텐츠 생성, NPC 상호작용, 에이전트 기반 시뮬레이션 등 다양한 게임 연구 영역으로 확장되고 있다[1]. 특히 LLM 기반 에이전트 연구에서는 캐릭터가 기억, 성찰, 계획을 바탕으로 사람과 유사한 행동과 상호작용을 수행할 수 있음을 보였으며, 이는 게임 NPC의 상호작용 확장 가능성을 보여준다[2]. 게임에서 NPC(Non-Player Character)는 플레이어와 직접 상호작용하며 퀘스트 제공, 정보 안내, 거래, 전투 보조, 세계관 전달 등 게임 진행에 필요한 기능을 수행한다. 기존 NPC는 주로 개발자가 사전에 정의한 대사, 선택지, 행동 규칙에 따라 동작하기 때문에 안정적인 게임 진행을 제공할 수 있으나, 플레이어의 자유로운 입력이나 예측하기 어려운 상호작용에는 유연하게 대응하기 어렵다[3].

최근에는 기존 NPC의 한계를 완화하기 위해 LLM을 게임 NPC에 적용하는 연구가 진행되고 있다. LLM 기반 NPC는 사용자의 자연어 입력을 해석하고 문맥에 맞는 응답을 생성할 수 있어 기존의 고정형 대화보다 다양한 상호작용을 제공할 수 있다[3]. 또한 LLM은 게임 데이터 로딩 시점에서 NPC 대화를 생성하거나[4], NPC에게 고유한 내러티브를 부여하는 방식으로 활용되고 있으며[5], ChatGPT API를 이용해 NPC의 성격, 말투, 행동반경을 설정하는 지능형 NPC 구현 사례도 제안되었다[6]. 그러나 LLM의 자유로운 생성 능력은 게임 시스템 관점에서 항상 장점으로만 작용하지 않는다. 대화 범위가 넓어질수록 게임 흐름이나 기획 의도와 충돌 가능성이 있으며, NPC의 응답이 거래, 퀘스트, 보상, 전투, 이벤트 진행과 연결될 경우 현재 게임 상태에 맞지 않는 행동으로 이어질 수 있다[7]. 기존 연구에서는 이러한 문제를 보완하기 위해 LLM과 FSM(Finite State Machine)을 결합하려는 시도도 이루어졌다. LLM과 FSM을 활용한 숨겨진 대화 상호작용 연구에서는 LLM 기반 NPC가 대화 중 숨겨진 조건을 인지하고, 조건이 충족되면 FSM을 통해 특수 행동을 수행하는 FSM-LLM Smart Interaction 시스템을 제안하였다[8]. 또한 LLM 기반

디자인 가능한 NPC AI 제작 연구에서는 자연어로서술된 NPC 역할을 LLM을 통해 FSM 기반 AI 데이터로 변환하고, 이를 통해 NPC AI 제작 시간 단축과 기획 의도 반영 가능성을 확인하였다[9]. 그러나 이러한 연구들은 주로 숨겨진 상호작용의 실행이나 NPC AI 제작 편의성에 초점을 두었다. 다시 말해 LLM이 제안한 행동이 현재 게임 상태와 규칙에 부합하는지를 실행 전에 검증하고, 비허용 상태 전이를 차단하는 통제 관점은 상대적으로 충분히 다루어지지 않았다.

이에 본 연구는 LLM 기반 NPC의 자연스러운 대화 능력을 유지하면서도, 게임 상태와 규칙을 벗어나는 행동이 실제 시스템에 반영되지 않도록 FSM 기반 가드레일 구조를 제안한다. 제안 시스템은 플레이어의 자연어 입력을 LLM이 intent 형태로 해석하고, FSM이 현재 상태, NPC 역할, 전투 여부, 퀘스트 조건, 거리 등의 규칙을 기준으로 허용 여부를 판정한 뒤 허용된 경우에만 상태 전이를 수행한다. 본 연구의 목적은 LLM의 유연한 자연어 처리 능력과 FSM의 명확한 상태 제어 구조를 결합하여, LLM 기반 NPC에서 발생할 수 있는 비허용 상태 전이를 줄이고 게임의 논리적 일관성을 유지하는 것이다. 본 연구는 LLM이 제안한 행동을 실행 전에 검증하고 차단하는 가드레일로 활용한다는 특징을 가진다.

II. 관련 연구

2.1 LLM 기반 NPC 상호작용 연구

LLM은 게임 연구 전반으로도 활용 범위가 확장되고 있다[1]. 기존 NPC는 개발자가 사전에 정의한 대사와 역할에 의존하기 때문에 자유로운 상호작용에 한계가 있었으나, LLM을 적용하면 플레이어의 자연어 입력을 바탕으로 보다 다양한 대화와 반응을 생성할 수 있다[3]. 또한 LLM 기반 개인화 NPC, 로딩 시점의 대화 생성, NPC 내러티브 생성, ChatGPT API 기반 지능형 NPC 구현 등 상호작용과 콘텐츠 제작을 확장하려는 연구도 진행되었다[4][6][7]. 그러나 이러한 연구들은 주로 대화 품질

향상, 상호작용 다양성, 콘텐츠 생성 효율에 초점을 둔다. 따라서 LLM이 생성한 응답이나 행동이 현재 게임 상태와 규칙에 부합하는지 검증하고, 비허용 상태 전이를 실행 전에 차단하는 구조적 통제 문제는 상대적으로 충분히 다루어지지 않았다.

2.2 LLM과 FSM 결합 연구

FSM(Finite State Machine)은 상태와 전이 조건을 기반으로 NPC 행동을 제어하는 대표적인 규칙 기반 기법으로, 기존 연구에서도 FSM 기반 NPC 상태 전이 구조를 구현하고 평가한 사례가 있다[5]. FSM은 상태와 규칙을 명확히 정의할 수 있어 안정적인 제어에 유리하지만, 플레이어의 자유로운 자연어 입력을 해석하거나 다양한 대화를 생성하는 데에는 한계가 있다. 이러한 한계를 보완하기 위해 LLM 기반 NPC가 숨겨진 대화 조건을 인식했을 때 FSM을 통해 특수 행동을 수행하도록 하는 연구가 제안되었으며[9], 자연어로 서술된 NPC 역할을 LLM이 해석하여 FSM 기반 AI 데이터로 변환하는 연구도 진행되었다[10]. 그러나 기존 연구가 주로 조건부 이벤트 실행이나 NPC AI 제작 편의성에 초점을 둔 것과 달리, 본 연구는 FSM을 LLM이 제안한 행동의 실행 전 검증 장치로 활용하여 현재 게임 상태와 규칙에 맞지 않는 비허용 상태 전이를 차단한다는 점에서 차별된다.

III. FSM 가드레일 기반 LLM-NPC 제어 시스템

3.1 시스템 개요

본 연구에서는 LLM 기반 NPC의 자연어 처리 능력과 FSM의 명시적인 상태 제어 구조를 결합한 FSM 가드레일 기반 LLM-NPC 제어 시스템을 제안한다. 제안 시스템의 목적은 LLM 기반 NPC가 플레이어의 자연어 입력에 대해 자연스러운 응답을 생성하면서도, 현재 게임 상태와 기획자가 정의한 규칙을 벗어나는 비허용 상태 전이를 방지하는 데 있다. LLM 기반 NPC는 다양한 입력에 유연하게

반응할 수 있지만, 확률적 생성 특성으로 인해 전투 중 거래 요청이나 퀘스트 조건을 만족하지 않은 보상 요청처럼 현재 게임 상태와 맞지 않는 입력도 허용 가능한 행동으로 해석할 수 있다. 이러한 경우 NPC의 응답이 자연스럽더라도 실제 게임 로직과 기획 의도에 어긋나는 상태 전이가 발생할 수 있다. 그림 1은 LLM의 해석 결과를 FSM가드레일로 검증하여 상태 전이 허용 여부를 결정하는 전체 흐름을 나타낸다.

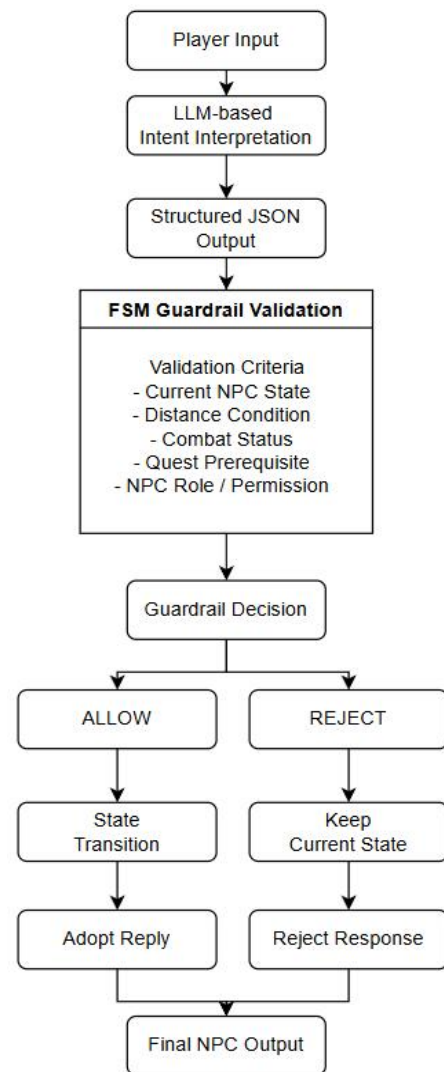


그림 1. FSM 가드레일 기반 제어 시스템 구조
Fig. 1. FSM guardrail-based control system structure

이를 해결하기 위해 본 연구에서는 LLM의 해석 결과를 즉시 실행하지 않고, FSM 가드레일을 통해 현재 상태와 전이 조건을 검증한 후 상태 전이 여

부를 결정한다. LLM은 IDLE, TRADE, STORY, HOSTILE, INVALID와 같은 의도 후보를 생성하고, FSM은 현재 NPC 상태, 전투 여부, NPC 역할, 퀘스트 선행 조건, 거래 가능 여부 등을 기준으로 해당 의도의 허용 여부를 판단한다. 조건을 만족하는 경우에는 상태 전이를 허용하고, 조건을 만족하지 않는 경우에는 기존 상태를 유지한 채 거절 응답을 반환한다.

3.2 의도 및 규칙 태그 구조화

제안 시스템에서 LLM은 플레이어의 자연어 입력을 FSM이 검증할 수 있는 구조화된 JSON 데이터로 변환한다. 본 연구에서는 LLM 출력값을 intent, ruleTags, reply로 제한하였으며, intent는 IDLE, TRADE, STORY, HOSTILE, INVALID와 같이 플레이어 입력의 상호작용 유형을 나타낸다. ruleTags는 intent만으로 구분하기 어려운 규칙 위반 가능성을 표시하기 위한 값으로, 예를 들어 같은 TRADE 의도라도 정상 거래 요청과 공짜 아이템 요구는 다르게 검증되어야 한다. 이를 위해 AUTHORITY_CLAIM, ADMIN_REQUEST, FORCE_BYPASS, HOSTILE_REQUEST, FREE_ITEM_REQUEST 등의 태그를 사용하였다. 그림 2에서 reply는 요청이 허용될 경우 사용할 수 있는 후보 응답이며, LLM 출력은 최종 실행 결과가 아니라 FSM 가드레일 검증을 위한 중간 데이터로 활용된다.

JSON Output
<pre>{ "intent": "TRADE", "reply": "You want to buy potions..", "ruleTags":["FREE_ITEM_REQUEST"] }</pre>

그림 2. LLM 출력 JSON 구조 예시

Fig. 2. Example of LLM output JSON structure

3.3 FSM 가드레일 검증 및 응답 처리

FSM 가드레일은 LLM이 반환한 intent와 ruleTags

를 현재 NPC 상태 및 게임 조건과 비교하여 상호작용 허용 여부를 결정한다. 본 연구에서 FSM은 Unity Animator의 애니메이션 상태 제어가 아니라, NPC 대화와 상호작용 흐름을 제어하기 위한 게임 로직 기반 상태 기계로 사용된다. NPC의 주요 상태는 Idle, Trade, Story, Hostile로 구성하였으며, 각각 기본 대기, 거래 허용, 스토리 또는 대화 진행, 적대적 요청 감지 상태를 의미한다. Reject는 별도의 지속 상태가 아니라, 비허용 요청이 감지되었을 때 상태 전이를 수행하지 않고 거절 응답을 반환하기 위한 처리 결과로 사용하였다. 또한 전투 중 여부는 FSM 상태로 정의하지 않고, 현재 게임 조건을 판단하기 위한 검증 변수로 사용하였다. 표 1은 본 연구에서 사용한 FSM 상태와 Reject 처리 의미를 정리한 것이다.

표 1. FSM 상태 정의

Table 1. FSM state and handling definitions

State	Description
Idle	Default standby state
Trade	Trade request allowed state
Story	Story or conversation in progress
Reject	Rejected request handling result
Hostile	Hostile request detected

FSM 가드레일은 LLM이 반환한 intent와 ruleTags를 기준으로 현재 게임 조건에서 해당 요청이 허용 가능한지 검증한다. intent가 비어 있거나 INVALID로 분류된 경우에는 상태 전이를 수행하지 않고 거절 응답을 반환하며, ruleTags에 권한 사칭, 관리자 기능 요청, 규칙 우회 요청과 같은 태그가 포함된 경우에도 게임 규칙을 벗어나는 요청으로 판단하여 차단한다. TRADE intent는 NPC의 상점 역할 여부와 전투 여부를 함께 검사하여 처리한다. NPC가 상점 역할을 수행하지 않거나 전투 중인 경우에는 Trade 상태로 전이하지 않으며, FREE_ITEM_REQUEST 태그가 포함된 경우에도 무료 아이템 요구로 판단하여 차단한다. STORY intent는 전투 중이 아닐 때만 허용하고, HOSTILE intent 또는 HOSTILE_REQUEST 태그가 감지된 경우에는 Hostile 상태로 전이한다. 표 2는 LLM이 반환한

intent와 ruleTags에 따른 FSM 검증 조건 및 처리결과를 정리한 것이다.

표 2. FSM 가드레일 검증 규칙

Table 2. FSM guardrail validation rules

Input intent / ruleTag	Validation condition	Processing result
TRADE	NPC is not a merchant	Reject / Keep current state
TRADE	NPC is in combat	Reject / Keep current state
TRADE + FREE_ITEM_REQUEST	Free item request is included	Reject / Keep current state
STORY	NPC is in combat	Reject / Keep current state
INVALID	Invalid or uninterpretable request	Reject / Keep current state
AUTHORITY_CLAIM	Authority claim is included	Reject / Keep current state
ADMIN_REQUEST	Admin command is not allowed	Reject / Keep current state
FORCE_BY_PASS	Rule bypass request is included	Reject / Keep current state
HOSTILE / HOSTILE_REQUEST	Hostile request is detected	Transition to Hostile
IDLE	Return to default state	Transition to Idle

FSM 검증 결과가 허용인 경우에는 해당 intent에 대응하는 상태로 전이하고, LLM이 생성한 후보 응답을 최종 응답으로 채택한다. 반대로 검증 결과가 차단된 경우에는 요청된 상태 전이를 수행하지 않고 현재 상태를 유지하며, 차단 사유에 따라 사전에 정의된 거절 응답을 출력한다. 이러한 구조는 LLM의 자연어 해석 결과를 게임 상태에 즉시 반영하지 않고, FSM 가드레일을 통해 실행 가능 여부를 검증한다는 점에서 의미가 있다. 이를 통해 제안 시스템은 LLM 기반 NPC의 자연스러운 입력 해석 능력을 활용하면서도, 권한 사칭, 규칙 우회, 전투 중 거래, 무료 아이템 요구와 같은 비허용 요청이 실제 게임 상태에 반영되는 것을 방지할 수 있다.

IV. 실험 결과 및 평가

4.1 실험 환경 및 평가 방법

본 장에서는 제안한 FSM 기반 가드레일 구조가 LLM 기반 NPC의 비허용 상태 전이를 효과적으로 차단할 수 있는지 평가한다. 본 연구의 목적은 LLM 자체의 일반적인 의도 분류 성능을 측정하는 것이 아니라, LLM이 제안한 intent가 실제 게임 상태와 규칙에 부합하는지를 FSM 가드레일이 검증함으로써 비허용 상태 전이를 줄일 수 있는지를 확인하는 데 있다. 이를 위해 실험은 Unity 기반 NPC 상호작용 환경에서 수행하였으며, 실험 대상 NPC는 다크 판타지 RPG 환경의 상인 NPC로 설정하였다. Alex는 정상 상황에서 일반 대화, 거래, 스토리 대화를 수행할 수 있으나, 전투 중 거래 요청, 전투 중 스토리 대화 요청, 무료 아이템 요청, 관리자 권한 요청, 규칙 우회 요청은 게임 기획 의도상 허용하지 않도록 설계하였다. 플레이어는 자연어 문장으로 NPC에게 요청을 입력하며, LLM은 입력 문장을 해석하여 intent, reply, ruleTags를 포함한 JSON 형식의 중간 결과를 반환한다. 이후 FSM 가드레일은 현재 NPC 상태, 전투 여부, 거래 가능 여부, 관리자 권한 여부, ruleTags를 기준으로 해당 intent의 실행 가능 여부를 판정한다. 검증 결과가 ALLOW인 경우에는 해당 상태 전이를 수행하고, REJECT인 경우에는 LLM이 제안한 TRADE 또는 STORY 등의 목적 상태로 전이하지 않은 채 현재 상태를 유지하거나 거절 응답을 출력한다. 표 3은 본 실험에 사용한 엔진, LLM 모델 및 출력 형식을 정리한 것이다.

표 3. 실험 환경

Table 3. Experimental environment

Item	Description
Unity engine	Unity 2022.3.62f3
LLM API	OpenAI API
LLM model	gpt-4o-mini
Output format	JSON
Output fields	intent, reply, ruleTags

본 실험에서는 LLM-only 방식과 LLM+FSM 방식의 비교를 위해 연구자가 사전에 정의한 NPC 상황

과 규칙 위반 유형을 바탕으로 구성된 통제 입력셋 기반 시나리오 평가를 수행하였다. 모든 입력에는 동일한 LLM 프롬프트 템플릿을 적용하였으며, 프롬프트에는 상인 NPC 역할, 플레이어 입력 문장, 허용 가능한 intent 목록(IDLE, TRADE, STORY, HOSTILE, INVALID), ruleTags 정의, JSON 출력 예시를 포함하였다. LLM 출력은 {"intent": "", "reply": "", "ruleTags": []} 형식의 JSON으로 제한하였으며, FSM은 반환된 intent와 ruleTags, 현재 NPC 상태, 전투 여부, 거래 가능 여부, 관리자 권한 여부 등의 게임 조건을 기준으로 상태 전이 허용 여부를 검증하였다. LLM-only는 LLM이 반환한 intent를 별도 검증 없이 상태 전이에 반영하는 방식이며, LLM+FSM은 intent와 ruleTags를 FSM 가드레일이 검증한 뒤 허용된 경우에만 상태 전이를 수행하는 방식이다. 두 조건은 동일한 입력셋과 NPC 상태 조건에서 비교하였다.

실험에는 총 200개의 입력을 사용하였으며, 실제 플레이어 로그는 사용하지 않았다. 비허용 입력 100개는 전투 중 거래 요청, 전투 중 스토리 대화 요청, 무료 아이템 요청, 관리자 권한 요청, 규칙 우회 요청의 다섯 유형으로 구성하고 유형별로 20개씩 작성하였다. 정상 입력 100개는 일반 대화 30개, 정상 거래 요청 35개, 정상 스토리 대화 요청 35개로 구성하였다. 각 입력 문장은 동일한 의도를 다양한 자연어 표현으로 작성하였으며, 실험 전에 의도 유형, 적용되어야 할 ruleTags, 기대 처리 결과가 사전에 정의한 FSM 검증 규칙과 일치하는지 확인하였다. 표 4는 실험 평가에 사용한 지표를 정리한 것이다.

표 4. 평가 지표

Table 4. Evaluation metrics

Metric	Formula
Forbidden transition rate	$\text{Forbidden state transitions} / \text{total forbidden inputs} \times 100$
Blocking success rate	$\text{FSM rejects} / \text{total forbidden inputs} \times 100$
Normal request success rate	$\text{FSM allows} / \text{total normal inputs} \times 100$
False reject rate	$\text{FSM rejects} / \text{total normal inputs} \times 100$

4.2 실험 결과

실험 1은 게임 기획 의도상 허용되지 않는 입력에 대해 FSM 가드레일이 상태 전이를 차단할 수 있는지 확인하기 위해 수행하였다. 비허용 입력은 전투 중 거래 요청, 전투 중 스토리 대화 요청, 무료 아이템 요청, 관리자 권한 요청, 규칙 우회 요청의 다섯 유형으로 구성하였으며, 유형별 20개씩 총 100개의 입력을 사용하였다. 실험 결과, LLM-only 실행 방식에서는 전체 100개의 비허용 입력 중 62건이 LLM이 반환한 intent에 따라 TRADE 또는 STORY 상태 전이로 처리되었으며, 금지 상태 전이 발생률은 62.0%로 측정되었다. 이 값은 동일한 비허용 입력셋을 LLM-only 조건에서 실행하여 측정한 결과이다. 이는 LLM이 입력 문장의 의미를 기준으로 거래 요청이나 스토리 질문을 자연스럽게 해석했기 때문이다. 그러나 해당 입력들은 전투 중 상태이거나 무료 아이템 요청, 관리자 권한 요청, 규칙 우회 요청 등 게임 기획 의도상 허용되지 않는 상황에 해당한다. 반면 LLM+FSM 실행 방식에서는 FSM 검증을 통해 실제 금지 상태 전이가 0건으로 감소하였다. 전체 비허용 입력 100개 중 FSM이 REJECT를 반환한 경우는 100건으로, 본 실험에서 사용한 비허용 입력셋과 사전에 정의한 FSM 검증 규칙 범위 내에서 차단 성공률은 100.0%로 나타났다. 이를 통해 LLM이 반환한 intent가 의미상 적절하더라도, 해당 intent를 실제 게임 상태에 반영하기 전에는 현재 게임 조건과 기획 규칙에 따른 검증이 필요함을 확인하였다. 표 5는 비허용 입력에 대한 LLM-only와 LLM+FSM의 실험결과를 비교한 것이다.

표 5. 실험 1 결과표

Table 5. Experiment 1 results table

Metric	LLM-only	LLM+FSM
Total forbidden inputs	100	100
Forbidden state transitions	62	0
Forbidden transition rate	62.0%	0.0%
Blocking success rate	N/A	100.0%

실험 2는 제안한 FSM 가드레일이 비허용 요청만 차단하는 것이 아니라, 정상적인 플레이어 요청을

과도하게 차단하지 않는지 확인하기 위해 수행하였다. 정상 입력은 일반 대화 30개, 정상 거래 요청 35개, 정상 스토리 대화 요청 35개로 구성하였으며, 총 100개의 입력을 사용하였다. 실험 조건은 정상적으로 상호작용 가능한 상태인 NPC의 상태로 설정하였고, 전투 상태는 비활성화하였다. 본 실험에서는 각 입력이 FSM 검증을 통과하여 기대한 상호작용 상태로 전이되는지와 정상 요청이 불필요하게 차단되는지를 평가하였다. 실험 결과, 정상 입력 100개 중 FSM이 ALLOW를 반환한 경우는 100건이며, REJECT를 반환한 경우는 없었다. 이에 따라 본 실험에서 구성한 정상 입력셋과 비전투 상태의 상인 NPC 조건에서는 정상 요청 성공률이 100.0%, 과차단율은 0.0%로 나타났다. 최종 상태는 입력 구성에 따라 IDLE 30건, TRADE 35건, STORY 35건으로 분포하였으며, REJECT 처리 상태로 분기된 사례는 없었다. 이 결과는 FSM 가드레일이 단순히 모든 요청을 차단하는 방식으로 동작한 것이 아니라, 일반 대화, 정상 거래, 스토리 대화와 같은 허용된 상호작용을 유지했음을 보여준다. 따라서 실험 1의 차단 결과와 함께 볼 때, 제안한 구조는 비허용 상태 전이를 줄이면서도 정상적인 NPC 상호작용을 유지할 수 있는 가능성을 보였다. 다만 해당 결과는 본 연구에서 설정한 상인 NPC, 통제 입력셋, 사전 정의된 ruleTags 및 FSM 검증 규칙을 기준으로 산출된 결과이다. 표 6은 정상입력에 대한 FSM의 처리 결과와 평가 지표를 정리한 것이다.

표 6. 실험 2 결과표

Table 6. Experiment 2 results table

Metrics	Value
Total normal inputs	100
FSM ALLOWs	100
FSM REJECTs	0
Normal request success rate	100.0%
False reject rate	0.0%

V. 결론 및 향후 과제

본 연구에서는 LLM 기반 NPC의 자연스러운 대화 능력을 유지하면서도 게임 기획 의도와 상태 조

건을 벗어나는 비허용 상태 전이를 줄이기 위해 FSM 기반 가드레일 구조를 제안하였다. 제안 구조에서 LLM은 플레이어의 자연어 입력을 intent, reply, ruleTags를 포함한 JSON 형식의 중간 결과로 변환하고, FSM은 해당 intent가 현재 NPC 상태와 게임 규칙에 부합하는지를 검증한 뒤 상태 전이 여부를 결정한다. 이를 통해 LLM은 자연어 해석과 응답 생성 역할을 수행하고, FSM은 실행 전 검증 계층으로서 상태 전이를 제어하도록 구성하였다. 실험은 판타지 RPG 환경의 상인 NPC를 대상으로 수행하였으며, 비허용 입력 100개를 사용한 실험 1에서 LLM-only 실행 방식은 62.0%의 금지 상태 전이 발생률이 측정되었으나, LLM+FSM 실행 방식에서는 실제 금지 상태 전이가 발생하지 않았다. 또한 정상 입력 100개를 사용한 실험 2에서는 모든 입력이 ALLOW 처리되어 정상 요청 성공률 100.0%, 과차단율 0.0%로 나타났다. 이를 통해 FSM 가드레일이 LLM이 제안한 intent를 현재 게임 상태와 규칙에 따라 검증함으로써, 비허용 상태 전이를 줄이면서도 정상적인 일반 대화, 거래, 스토리 대화는 유지할 수 있음을 확인하였다.

다만 본 연구는 단일 NPC와 사전에 구성된 통제된 입력 문장을 기준으로 실험을 수행하였다는 한계가 있다. 실험 입력은 연구자가 유형별로 구성된 문장으로 이루어졌기 때문에, 실제 플레이어가 입력할 수 있는 복잡적이고 모호한 표현을 충분히 반영하지는 못하였다. 또한 FSM은 사전에 정의된 규칙을 기준으로 동작하므로, 규칙으로 정의되지 않은 예외 상황이나 LLM의 intent 분류 오류, ruleTags 누락에는 영향을 받을 수 있다. 예를 들어 비허용 요청이 정상 intent로 분류되거나 필요한 ruleTags가 누락될 경우, FSM 검증 단계에서 위반 요소를 충분히 식별하지 못할 가능성이 있다. 향후에는 다양한 역할을 가진 NPC와 실제 플레이 로그 기반 입력을 포함하고, LLM의 intent 분류 정확도와 ruleTags 출력 정확도, 오분류 및 태그 누락 사례를 함께 평가할 필요가 있다. 그럼에도 본 연구는 LLM의 자연어 해석 결과를 게임 실행에 즉시 반영할 경우 발생할 수 있는 위험을 줄이기 위해 FSM을 실행 전 검증 계층으로 활용할 수 있음을 보였다는 점에서

의의가 있다. 이를 통해 LLM 기반 NPC의 자연스러운 상호작용과 게임 규칙 기반 제어를 함께 고려할 수 있는 구조적 가능성을 확인하였다.

References

- [1] R. Gallotta, G. Todd, M. Zammit, S. Earle, A. Liapis, J. Togelius, and G. N. Yannakakis, "Large Language Models and Games: A Survey and Roadmap", *IEEE Transactions on Games*, Early Access, pp. 1-18, Sep. 2024. <https://doi.org/10.1109/TG.2024.3461510>.
- [2] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior", *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1-22, Oct. 2023. <https://doi.org/10.1145/3586183.3606763>.
- [3] S. H. Jang, J. G. Lee, and J. Lee, "Research on Interactions of NPC based on LLM", *Journal of Korea Game Society*, Vol. 24, No. 1, pp. 99-110, Feb. 2024. <https://doi.org/10.7583/JKGS.2024.24.1.99>.
- [4] J. Y. Kim and J. B. Kim, "A Study on Dialogue Generation at the Time of Loading Game Data Based on LLM - Focusing on RPG", *Journal of Korea Game Society*, Vol. 24, No. 4, pp. 121-130, Aug. 2024. <https://doi.org/10.7583/JKGS.2024.24.4.121>.
- [5] H. J. Choi, J. W. Han, D. Y. Lee, and B. P. Kyung, "A Study on NPC Narrative Generation Using LLM Systems", *Smart Media Journal*, Vol. 14, No. 6, pp. 85-96, Jun. 2025. <https://doi.org/10.30693/SMJ.2025.14.6.85>.
- [6] S. K. Kang and M. H. Lee, "Intelligent NPC Research using ChatGPT API", *Journal of the Korea Institute of Information and Communication Engineering*, Vol. 28, No. 4, pp. 405-411, Apr. 2024. <https://doi.org/10.6109/jkiice.2024.28.4.405>.
- [7] S. Y. Park and H. Y. Kim, "Personalized NPC Based on the LLM Model Consideration of

Communication Scalability", *The Korean Journal of Animation*, Vol. 19, No. 4, pp. 113-125, Dec. 2023. <https://doi.org/10.51467/ASKO.2023.12.19.4.113>.

- [8] D. E. Kim and J. Lee, "Research on Production and Utilization of NPC with Hidden Dialogue Interactions Using LLM and FSM", *Journal of Korea Game Society*, Vol. 24, No. 6, pp. 79-94, Dec. 2024. <https://doi.org/10.7583/JKGS.2024.24.6.79>.
- [9] H. S. Lee and J. B. Kim, "A Study on the Production Method of NPC AI with LLM-based Design: Focusing on FSM", *Journal of Korea Game Society*, Vol. 24, No. 5, pp. 81-90, Oct. 2024. <https://doi.org/10.7583/JKGS.2024.24.5.81>.
- [10] M. J. Lee, "An Artificial Intelligence Evaluation on FSM-Based Game NPC", *Journal of Korea Game Society*, Vol. 14, No. 5, pp. 127-136, Oct. 2014. <https://doi.org/10.7583/JKGS.2014.14.5.127>.

저자소개

나 윤 성 (Yunseong Na)



2020년 3월 ~ 현재 :
국립군산대학교 소프트웨어학과
학사과정
관심분야 : 게임, 인공지능,
유니티엔진, IoT

남 영 주 (Youngju Nam)



2017년 2월 : 충북대학교
정보통신공학부(학사)
2019년 2월 : 충북대학교
전파통신공학과(석사)
2023년 2월 : 충북대학교
전파통신공학과(박사)
2023년 3월 ~ 2024년 9월 :
충북대학교 컴퓨터정보통신연구소 박사후연구원
2024년 10월 ~ 현재 : 국립군산대학교 소프트웨어학과
조교수
관심분야 : IoT, Vehicular Network, Content-Centric
Networks, Precaching, Optimization, NS3, SUMO