Journal of KIIT. Vol. 23, No. 12, pp. 89-100, Dec. 31, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.12.89

게임 NPC 대화 시스템을 위한 PRIME 평가 프레임워크

소훈*, 정현준**

PRIME Evaluation Framework for Game NPC Dialogue Systems

Hoon So*, Hyunjun Jung**

요 약

이 연구는 게임 NPC 대화 시스템 평가에서 기존 RAGAS(Retrieval Augmented Generation Assessment) 메트릭의 한계를 실증적으로 입증하고, 인지과학 기반의 PRIME 평가 프레임워크를 제안한다. 4개 RAG 시스템 평가 결과, RAGAS에서 최하위를 기록한 Memory Stream RAG(0.377)가 제안된 PRIME 메트릭 성격 일관성, 관계 발전 추적, 지능 일관성, 기억 감쇠 자연스러움, 감정 일관성에서는 최상위 성능을 보였다. 이는 기존 메트릭이 정확성만을 중시하여 자연스러운 망각과 관계 발전을 오히려 부정적으로 평가하기 때문이다. 이 연구에서 제안한 PRIME 5가지 메트릭은 인간 평가와 $r=0.81$($p<0.001$)의 높은 상관관계를 보여, RAGAS($r=0.23$) 대비 3.5배 향상된 타당성을 입증했다. 30명의 숙련된 게이머를 대상으로 한 실험을 통해 제안 프레임워크의 유효성을 검증하였으며, Memory Stream RAG는 NPC 메트릭에서 RAGAS 최하위(4위)에서 최상위(1위)로 3단계 상승하는 결과를 보였다.

Abstract

This study empirically demonstrates the limitations of existing RAGAS metrics in evaluating game NPC dialogue systems and proposes a cognitive science-based PRIME evaluation framework. Among four RAG systems evaluated, Memory Stream RAG, which ranked lowest in RAGAS (0.377), showed the highest performance in the proposed PRIME metrics. This discrepancy occurs because existing metrics focus solely on accuracy, negatively evaluating natural forgetting and relationship evolution. The five proposed metrics achieved $r=0.81$ correlation with human evaluation ($p<0.001$), a 3.5x improvement over RAGAS ($r=0.23$). Through experiments with 30 experienced gamers (100+ hours of RPG gameplay), we validated the framework's effectiveness, with Memory Stream RAG rising from lowest rank (4th) in RAGAS to highest rank (1st) in NPC metrics.

Keywords
PRIME, NPC dialogue system, evaluation metrics, RAG, cognitive science, NLP

* 국립군산대학교 소프트웨어학과 학사과정
- ORCID: https://orcid.org/0009-0007-1198-5596
** 국립군산대학교 소프트웨어학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-6717-1395

· Received: Sep. 15, 2025, Revised: Nov. 03, 2025, Accepted: Nov. 06, 2025
· Corresponding Author: Hyunjun Jung
Dept. of Software at Kunsan National University, 558, Daehak-ro, Kunsan-si, Jeollabuk-do, Republic of Korea
Tel.: +82-63-469-8917, Email: junghj85@kunsan.ac.kr

Ⅰ. 서 론

2024년 Ubisoft의 Ghostwriter 도입과 같이 주요 게임 개발사들이 AI 기반 NPC(Non-Player Character) 대화 시스템에 투자를 확대하고 있으나, LLM(Large Language Model) 기반 NPC와의 상호작용 방식에 따라 사용자의 몰입감 경험이 다르게 나타난다는 연구 결과가 보고되고 있다[1]. 이는 Character.AI나 Replika 같은 상용 AI 캐릭터 플랫폼이 개인 대화에서는 높은 만족도를 얻고 있는 것과 대조적이다. 게임 NPC의 고유한 요구사항인 장기 기억 유지, 성격 일관성, 플레이어와의 관계 발전 등에 대한 체계적 평가 방법론이 부재하기 때문이다[2]. LLM 기반 NPC는 전통적인 스크립트 기반 대화의 제약을 극복하고 동적 대화를 가능하게 하지만, 이러한 기술적 진보가 실제 플레이어 경험으로 연결되지 못하는 주요 원인이 여기에 있다[3].

현재 게임 NPC 대화 품질 평가는 공백 상태에 있다. 게임 업계는 Ubisoft의 Ghostwriter, Unity ML-Agents, Charisma AI 같은 도구들을 통해 NPC 대화 생성에 대한 기술적 해결책을 제시하고 있으나, 이들의 평가는 여전히 전통적인 대화 시스템 평가 지표에 의존하고 있다[4][5]. 현재 학술계에서는 LLM 대화 평가를 위해 RAGAS(Retrieval Augmented Generation Assessment)[6], BLEU(Bilingual Evaluation Understudy)[7], ROUGE(Recall-Oriented Understudy for Gisting Evaluation)[8] 같은 언어 모델 평가 도구들이 사용되고 있다. 이러한 도구들은 정보 검색이나 질의응답 시스템을 위해 설계되어 "얼마나 정확하게 답변하는가"를 측정한다. 그러나 게임 NPC의 평가 목적은 QA 챗봇과 근본적으로 다르다. NPC는 정보 제공을 넘어 "플레이어가 그 게임 세계의 진짜 사람과 대화하는 듯한 몰입감"을 제공하는 것이 핵심이며, 이를 위해서는 장기 기억 유지, 감정적 유대감 형성, 성격 일관성, 플레이어와의 관계 발전 등 인간의 인지적 특성을 평가해야 한다.

기존 NPC 평가 연구들은 주로 "believability"라는 주관적 개념에 의존해 왔으며, 일관성 평가나 사용자 인식 척도 등 정성적 방법론을 사용했다[9]. 하지만 이러한 접근법들은 NPC의 인간다운 행동을 정량적으로 측정하기 어렵고, 특히 시간에 따른 기억 변화나 관계 발전 같은 동적 특성을 포착하지 못한다. RAGAS 같은 기존 메트릭을 NPC에 적용할 경우, 게임 환경에서 자연스러운 '망각'이나 '감정적 반응'을 오히려 부정적으로 평가하는 문제가 발생한다.

이러한 평가 체계의 문제점은 이 연구의 예비 실험에서 나타났다. 4개의 대표적 RAG 시스템을 RAGAS로 평가한 결과와 30명의 숙련된 RPG 게이머(100시간 이상)가 직접 평가한 결과 사이에 정반대 결과가 발견됐다. RAGAS에서 최하위를 기록한 Memory Stream RAG(0.377점)를 게이머들은 "가장 인간다운 NPC", "몰입감 있는 대화 상대"로 일관되게 평가한 반면, RAGAS 1위를 차지한 Graph RAG는 "정확하지만 냉정하고 기계적"이라고 개선을 요구했다. 이는 기존 평가 메트릭과 실제 사용자 경험 간의 괴리를 명확히 보여준다. 이러한 상반된 결과는 기존 메트릭이 완벽한 정보 재현과 사실적 정확성만을 최우선 가치로 삼아, 게임 환경에서 필수적인 '자연스러운 망각', '감정적 변화', '관계 발전'과 같은 인간다운 요소들을 오히려 결함으로 인식하는 한계를 보여준다.

이러한 평가 괴리는 인지과학적 관점에서 이해할 수 있다. Ebbinghaus의 망각 곡선이 보여주듯 인간은 시간이 지남에 따라 자연스럽게 정보를 망각하며, 이는 정상적인 인지 과정이다[10]. 예를 들어, The Elder Scrolls V: Skyrim이나 The Witcher 3와 같은 오픈월드 RPG에서 플레이어가 게임 시작 후 100시간이 경과한 시점에 초반 마을의 NPC를 다시 만났을 때, 해당 NPC가 "당신이 50시간 전에 빵 3개를 샀다"는 사소한 정보를 정확히 기억한다면 플레이어는 이를 비현실적이고 기계적으로 인식하게 된다. 반대로, 플레이어가 NPC의 가족을 구해주거나 중요한 퀘스트를 함께 수행한 경험처럼 감정적으로 중요한 사건은 오래 기억하되, 일상적 대화는 점차 잊는 것이 자연스러운 인간의 기억 패턴이다. 이러한 선택적, 점진적 망각 과정은 Ebbinghaus 망각 곡선으로 설명될 수 있으며, 게임 몰입감을 높이는 핵심 요인으로 작용한다.

또한 Knapp과 Vangelisti의 관계 발전 모델에 따르면, 인간 관계는 초기 단계에서 친밀 단계까지 점진적으로 발전하며, 이 과정에서 대화의 깊이와 개인적 정보 공유 수준이 변화한다[11].

이는 인지심리학의 Theory of Mind 이론으로 설명할 수 있다. Theory of Mind는 다른 사람의 신념, 의도, 욕망을 이해하는 인간의 기본적 능력으로, 플레이어가 NPC를 단순한 프로그램이 아닌 마음을 가진 존재로 인식하게 만드는 핵심 메커니즘이다[12]. 이러한 인지적 메커니즘은 게임 몰입감과 직접적으로 연결된다. 최근 게임 몰입감 연구에서는 인지적으로 일관된 환경이 플레이어의 인지 자원(Cognitive resources)을 게임 세계 이해에 집중시켜, 게임이라는 사실을 잊게 만든다는 것이 실증적으로 입증되었다[13]. 이러한 자연스러운 인지적 특성들은 단순한 인간다움을 위한 것이 아니라, 플레이어가 게임 세계에 높은 몰입도를 유지할 수 있도록 하는 핵심 장치다. 그러나 기존 평가 메트릭은 이러한 관계 변화를 측정하지 못한다.

따라서 게임 몰입감을 극대화하는 NPC 평가에는 인지과학과 사회심리학 이론에 기반한 새로운 접근법이 필요하다. 예를 들어, NPC가 플레이어의 이름을 처음에는 완벽히 기억하다가 시간이 지나면서 점차 잊어버리는 것이 오히려 자연스럽다. 마찬가지로 처음 만난 NPC가 경계심을 보이다가 여러 번의 상호작용을 거쳐 점진적으로 친밀해지는 것이 현실적인 관계 발전이다. 이러한 인간다운 특성들을 정량적으로 측정할 수 있는 메트릭이 필요하며, 이는 기존의 정확성 중심 평가와는 완전히 다른 접근 방식을 요구한다.

최근 연구들은 LLM 기반 NPC의 장기 기억 평가에 주목하고 있다. 일부 연구에서는 NPC가 플레이어 이름을 기억하고 이전 교류를 참조할 수 있음을 보여주었으나, 장기적으로는 기억이 희석되어 확장된 대화 후 플레이어 정보를 망각할 수 있다고 보고했다[14]. 이러한 연구들은 LLM 기반 NPC 평가에서 기존 정확성 중심 메트릭의 한계를 간접적으로 시사하고 있다.

이에 이 연구는 게임 몰입감 극대화를 위한 인지과학 기반 PRIME 평가 프레임워크를 제안한다. 제안하는 PRIME 5가지 메트릭은 성격 일관성 지수(PCI, Personality Consistency Index), 관계 발전 추적(RET, Relationship Evolution Tracking), 지능 일관성 점수(ICS, Intelligence Consistency Score), 기억 감쇠 자연스러움(MDN, Memory Decay Naturalism), 감정 일관성 점수(ECS, Emotional Coherence Score)로 구성되며, 각각은 게임 세계 몰입감에 기여하는 인간의 인지적 특성을 수학적으로 모델링한다. 실험 결과 제안된 메트릭은 인간 평가와 r=0.81의 상관관계를 달성하여, 기존 RAGAS(r=0.23)와 비교하여 향상된 결과를 보였다.

또한 전체 메트릭 계산이 10.9ms 내에 완료되어 60 FPS 게임 환경에서도 실시간 적용이 가능함을 확인했다.

이 논문의 구성은 다음과 같다. 2장에서는 기존 평가 메트릭의 근본적 한계와 NPC 평가 접근법, 인지과학적 기반 이론을 고찰한다. 3장에서는 제안하는 PRIME 평가 프레임워크의 아키텍처와 5가지 메트릭(PCI, RET, ICS, MDN, ECS)의 설계를 상세히 설명한다. 4장에서는 4개의 RAG 시스템에 대한 비교 실험 설계와 결과, 인간 평가와의 상관관계 분석을 제시한다. 마지막으로 5장에서는 연구의 의의와 한계점, 향후 연구 방향을 논의한다.

II. 관련 연구

이 장에서는 게임 NPC 대화 시스템 평가 연구에 대한 동향을 설명한다.

2.1 기존 대화 평가 메트릭의 한계

현재 자연어 생성 시스템 평가는 정확성을 최우선으로 하는 방식을 사용하고 있다. K. Papineni et al.[7]이 제안한 BLEU는 기계 번역에서 참조 번역과의 n-gram 일치도를 측정하며, C.-Y. Lin et al.[8]의 ROUGE는 요약 시스템에서 핵심 정보 포함 여부를 검증한다. T. Zhang et al.[15]의 BERTScore는 의미적 유사성을 고려했지만 여전히 정답과의 일치도 측정에 머물렀다. 최근 S. Es et al.[6]이 제안한 RAGAS는 RAG 시스템의 Faithfulness, Answer Relevancy, Context Precision, Context Recall을 평가하지만, 이 또한 얼마나 정확하게 정보를 재현하는가라는 질의응답 방식에서 벗어나지 못했다.

이러한 정확성 중심 평가의 근본적 문제는 인간 행동의 자연스러운 불완전함을 결함으로 인식한다는 점이다. 인간은 완벽한 기억 재현 기계가 아니

며, 시간이 지나면서 자연스럽게 망각하고, 감정에 따라 반응이 달라지며, 관계에 따라 대화 깊이를 조절한다. 하지만 BLEU, ROUGE, RAGAS 같은 메트릭들은 이러한 인간적 특성을 모두 오류로 분류한다. 실제로 이 연구의 실험에서 RAGAS와 게이머 평가 간 상관관계는 r=0.23에 불과했으며, RAGAS 최하위인 Memory Stream RAG(0.377)를 게이머들은 가장 인간답다고 평가했다. 이는 기존 메트릭이 게임 환경에서 요구되는 행동적 현실성을 전혀 포착하지 못함을 보여준다.

2.2 기존 NPC 평가 접근법

게임 NPC 평가에 대한 기존 연구들은 주관적 평가와 기술적 지표 중심 평가로 나뉜다. A. B. Loyall의 연구[16]에서는 NPC의 "believability"를 핵심 개념으로 제시했지만, 이를 측정하는 방법은 사용자 설문조사나 관찰 연구와 같은 정성적 방법에 의존했다. M. Mateas and A. Stern의 Façade[17] 시스템은 NPC 상호작용의 극적 효과를 평가했으나, 여전히 플레이어의 주관적 몰입감 보고에 기반했다.

최근 LLM 기반 NPC 연구에서는 정량적 평가가 시도되고 있다. J. S. Park et al.[18]의 Generative Agents는 Memory Stream 구조를 제안하고 25개 에이전트 시뮬레이션을 통해 사회적 행동을 관찰했지만, 평가는 "믿을 만한지"라는 주관적 기준에 의존했다. L. Song[19]의 LLM-driven NPCs는 대화 시스템의 기술적 성능에 초점을 맞췄으나, 플레이어 경험 측면의 평가는 제한적이었다.

또한 LLM을 활용해 대화형 텍스트 기반 게임을 자동으로 플레이하는 에이전트의 설계 및 비교 평가 연구도 보고되었다[20]

상용 솔루션들의 평가 방식도 한계가 있다. Inworld AI, Convai, Character.AI 등은 응답 속도, API 안정성, 모델 성능 등 기술적 지표를 중시한다. Unity ML-Agents와 Epic Games State Tree는 행동 트리 실행 효율성을 평가하지만, NPC가 플레이어에게 얼마나 자연스럽게 느껴지는지는 측정하지 않는다.

기존 접근법들의 공통된 한계는 NPC의 인간적 특성을 체계적으로 측정할 정량적 메트릭의 부재다. 주관적 평가는 일관성과 재현성이 떨어지고, 기술적 지표는 실제 플레이어 경험과의 연관성이 낮다. 특히 시간에 따른 기억 변화, 관계 발전, 성격 일관성 같은 동적 특성을 포착하는 평가 체계는 거의 존재하지 않는다.

2.3 NPC 평가를 위한 인지과학적 접근

NPC 평가에서 인지과학 이론의 적용 가능성은 다른 도메인의 성공 사례에서 확인할 수 있다. A. C. Graesser et al.[21]의 AutoTutor는 인지 부하 이론을 교육용 대화 에이전트 평가에 적용하여 학습 효과를 정량화했으며, T. Bickmore and J. Cassell[22]의 Relational Agents 연구는 사회심리학 이론으로 에이전트-사용자 신뢰 관계를 측정 가능한 메트릭으로 변환했다. 이러한 사례들은 도메인별 특성을 반영한 인지과학 기반 평가의 효과성을 입증한다.

게임 NPC 평가를 위한 인지과학적 기반 이론들은 이미 확립되어 있다. Ebbinghaus의 망각 곡선은 시간에 따른 자연스러운 기억 감쇠 패턴을 제공하며, 이를 통해 NPC의 기억 관리 자연스러움을 측정할 수 있다. Knapp과 Vangelisti의 관계 발전 모델은 인간 관계의 단계별 친밀도 변화를 설명하여 NPC-플레이어 관계 발전을 평가하는 기준을 제시한다. Big Five 성격 모델은 일관된 성격 표현을 위한 구체적 차원을 제공한다.

또한 Theory of Mind 이론은 NPC가 플레이어의 의도와 감정을 이해하고 적절히 반응하는 능력을 평가하는 틀을 제공한다. 최근 게임 몰입감 연구들도 인지적 일관성이 플레이어의 몰입도에 미치는 영향을 실증적으로 보여주고 있어, NPC 평가에서 인지과학적 접근의 타당성을 뒷받침한다.

문제는 이러한 검증된 이론들을 게임 NPC 평가에 특화된 측정 가능한 메트릭으로 변환하는 구체적 방법론이 부재하다는 점이다. 따라서 이 연구에서는 인지과학 이론들을 기반으로 한 정량적 NPC 평가 프레임워크인 PRIME을 제안한다.

III. PRIME 평가 프레임워크

이 장에서는 논문에서 제안하는 PRIME 평가 프레임워크의 전체 구조와 각 메트릭의 계산 방법에 대하여 설명한다.

3.1 프레임워크 아키텍처

제안하는 NPC 평가 프레임워크는 3계층 아키텍처로 구성된다. 입력층에서는 NPC-플레이어 간 대화 로그와 시스템 상태 정보를 수집한다. 처리층에서는 PRIME 5가지 평가 메트릭(PCI, RET, ICS, MDN, ECS)이 병렬로 계산되며, 각 메트릭은 독립적인 평가 모듈로 구현된다. 통합층에서는 가중치 기반 앙상블 방식으로 최종 NPC 품질 점수를 산출한다. 이 아키텍처는 모듈형 설계로 새로운 메트릭 추가가 용이하며, 실시간 평가와 배치 평가를 모두 지원한다.

3.2 프레임워크 설계 원칙

그림 1은 제안하는 NPC 평가 프레임워크는 세 가지 핵심 원칙을 나타낸다. 첫째, 행동적 현실성을 최우선으로 한다. 기존 메트릭들이 추구하는 정보의 정확한 전달보다, NPC가 보여주는 행동의 자연스러움과 인간다움을 중시한다. 둘째, 시간적 차원을 명시적으로 고려한다. 단일 대화 턴이 아닌 전체 상호작용 역사를 분석하여, 장기적 관계 형성과 기억의 변화를 추적한다. 셋째, 다차원 평가 체계를 구축한다. 기억, 관계, 성격, 논리성 등 다양한 측면을 독립적으로 평가하고 이를 통합하여 종합적인 NPC 품질을 판단한다.

3.3 평가 메트릭 상세 설계

3.3.1 성격 일관성 지수(PCI)

성격 일관성 지수는 NPC가 정의된 성격 특성을 얼마나 일관되게 유지하는지를 평가한다. Big Five 성격 모델(개방성, 성실성, 외향성, 친화성, 신경증)을 기반으로 하여, 각 대화 턴에서 표현되는 성격 특성을 분석한다.

PCI 계산 과정은 다음과 같다. 먼저 NPC의 기본 성격 프로필 P_expected를 설정한다. 예를 들어, 용감한 전사 NPC라면 외향성과 개방성이 높고 신경증이 낮을 것이다. 그 다음 실제 대화에서 관찰되는 성격 특성 P_observed를 측정한다. 이는 사용하는 어휘, 감정 표현, 의사결정 패턴 등을 분석하여 도출한다.

$$PCI = 1 - (\Sigma |P_{expected} - P_{observed}|) / (2 \cdot N \cdot D) \quad (1)$$

식 (1)로 계산되며, 여기서 N은 성격 차원의 수(5개), D는 각 차원의 최대 편차다. 완벽한 일관성을 보이면 1, 완전히 모순된 성격을 보이면 0의 값을 갖는다.

3.3.2 관계 발전 추적(RET)

관계 발전 추적 메트릭은 NPC와 플레이어 간의 관계가 시간에 따라 어떻게 발전하는지를 정량적으로 측정한다. Knapp의 관계 발전 모델을 게임 환경에 맞게 조정하여, 5단계 관계 상태를 정의했다: Stranger(낯선 사람, 0-0.2), Acquaintance(아는 사람, 0.2-0.4), Friend(친구, 0.4-0.7), Close Friend(가까운 친구, 0.7-0.9), Best Friend(절친, 0.9-1.0).

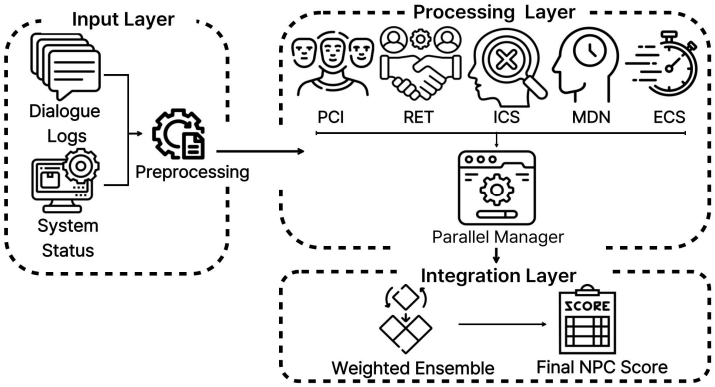


그림 1. NPC 평가 프레임워크 아키텍처
Fig. 1. NPC evaluation framework architecture

RET 계산은 각 대화 턴에서의 관계 변화량 ΔR_t와 상호작용 강도 I_t를 측정하여 이루어진다. 상호작용 강도는 대화의 깊이, 감정 표현의 정도, 개인적 정보 공유 수준 등을 종합적으로 고려한다. 예를 들어, 처음 만난 NPC가 갑자기 깊은 개인사를 털어놓는다면 RET 점수가 낮아진다. 반대로 여러 번의 상호작용을 거쳐 점진적으로 친밀해지는 패턴을 보이면 높은 점수를 받는다.

$$RET = (\Sigma(\Delta R_t \cdot I_{t)})/N \quad (2)$$

최종 RET는 식 (2)로 계산되며, 이는 전체 대화 기간 동안의 평균적인 관계 발전 자연스러움을 나타낸다.

3.3.3 지능 일관성 점수(ICS)

지능 일관성 점수는 NPC가 대화 중 발생하는 논리적 모순을 얼마나 잘 인식하고 처리하는지를 측정한다. 이는 단순히 모순을 피하는 것을 넘어, 모순이 발생했을 때 이를 자연스럽게 해결하는 능력까지 평가한다.

ICS 평가는 자연어 추론(NLI, Natural Language Inference) 모델을 활용하여 진행된다. 먼저 대화 쌍을 추출하여 함의(Entailment), 모순(Contradiction), 중립(Neutral) 관계를 분류한다. TP(True Positive)는 실제 모순을 정확히 탐지한 경우, TN(True Negative)는 모순이 없음을 정확히 판단한 경우다.

$$ICS = (TP + TN)/(TP + TN + FP + FN) \quad (3)$$

식 (3)로 계산되며, 이는 전체 판단 중 정확한 판단의 비율을 나타낸다. 높은 ICS는 NPC가 논리적 일관성을 잘 유지하거나, 모순이 발생했을 때 이를 인지하고 적절히 대응함을 의미한다.

3.3.4 기억 감쇠 자연스러움(MDN)

기억 감쇠 자연스러움은 NPC의 기억이 시간에 따라 얼마나 자연스럽게 감쇠하는지를 평가한다. J. S. Park et al.[18]의 메모리 증강 가상 에이전트 연구를 기준으로 하여, 관찰된 기억 보존율과 이론적 기대치를 비교한다.

MDN 측정은 다양한 시간 간격(1분, 10분, 1시간, 24시간, 7일 등)에서 특정 정보에 대한 NPC의 기억 정확도를 테스트한다. 중요한 것은 모든 정보를 동일하게 잊는 것이 아니라, 정보의 중요도에 따라 차별적 망각률을 적용한다는 점이다. 예를 들어, 플레이어의 이름이나 중요한 퀘스트 정보는 오래 기억하지만, 일상적인 대화 내용은 빠르게 잊는 것이 자연스럽다.

$$MDN = 1 - MSE(R_{observed}, R_{\text{expecte}d}) \quad (4)$$

식 (4)로 계산된다. 여기서 R_observed는 시간별 테스트에서 보존율의 시계열이고, R_expected는 인지심리학적 망각 곡선으로부터 산출된 기대 기억 보존율의 시계열이다. 이 둘 간의 평균 제곱 오차를 최소화할수록 높은 점수를 받는다.

그림 2는 이 연구의 MDN 측정 실험에서 도출된 시간별 기억 감쇠 곡선을 나타낸 것이다. 인간의 이론적 망각 곡선과 Memory Stream RAG, Graph RAG, Baseline RAG 간의 보존율 차이를 비교함으로써, PRIME 메트릭이 인간의 기억 패턴을 충실히 재현함을 보여준다.

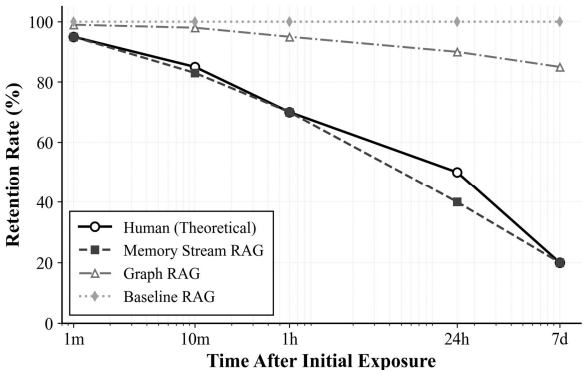


그림 2. 시간별 기억 감쇠 곡선 비교
Fig. 2. Comparison of memory decay curves over time

3.3.5 감정 일관성 점수(ECS)

감정 일관성 점수는 NPC가 과거 대화 내용을 얼마나 일관되게 기억하고 활용하는지를 평가한다. 단순히 모든 정보를 완벽하게 기억하는 것이 아니라, 시간의 흐름에 따라 자연스럽게 변화하는 기억의

패턴과 감정적 일관성을 측정한다.

ECS는 다음과 같이 계산된다. 먼저 각 대화 턴 i에서 언급된 정보의 일관성 점수 C_i를 산출한다. C_i는 현재 발언이 과거 정보와 일치하면 1, 모순되면 0, 부분적으로 일치하면 0과 1 사이의 값을 갖는다. 이때 중요한 것은 시간 가중치 w_i = e^(-λ × Δt_i)와 감정적 가중치를 모두 적용한다는 점이다.

$$w_i = \exp(-\lambda \cdot \Delta t_i), ECS = (\Sigma(w_{i \cdot C_{i \cdot E_i}} / (\Sigma w_i) \quad (5)$$

최종 ECS는 식 (5)로 계산되어, 최근 정보와 감정적으로 중요한 정보에 높은 가중치를 부여한다. 이는 인간이 감정적 사건을 더 오래 기억하는 자연스러운 패턴을 반영한다.

Ⅳ. 실험 결과 및 평가

이 장에서는 논문에서 제안하는 PRIME 평가 프레임워크의 유효성을 검증하기 위한 실험 과정과 결과에 대해 설명한다. 실험 결과는 기존 RAGAS 메트릭과 제안된 PRIME 메트릭의 비교 분석으로 나누어 제시한다.

4.1 실험 환경

이 연구의 실험은 게임 NPC 대화 시스템 평가의 타당성을 검증하기 위해 체계적으로 설계되었다. 실험 데이터셋은 실제 게임 환경을 모사한 100개의 독립적인 대화 세션으로 구성되며, 총 3,500개의 대화 턴을 포함한다.

데이터 생성은 다음 프로토콜을 따랐다: (1) GPT-4를 사용하여 초기 대화 템플릿 생성, (2) 5가지 페르소나(상인, 경비병, 모험가, 학자, 마을 주민) 설정, (3) 각 페르소나별 20개 세션 생성, (4) 연구팀 3명이 대화의 자연스러움과 일관성 검토 후 수정. 각 세션은 평균 35턴의 상호작용으로 이루어져 있으며, 초기 만남부터 장기적 관계 형성까지의 전체 과정을 다룬다. 대화 시나리오는 일상적 대화(40%), 퀘스트 관련 대화(30%), 감정적 교류(20%), 정보 교환(10%)으로 구성되어 실제 게임 플레이의 다양성을 반영했다.

평가자 선정은 군산 대학교 소프트웨어학과 학생들을 대상으로 진행되었다. 30명의 참가자는 모두 RPG 게임 100시간 이상의 플레이 경험을 가진 숙련된 게이머로, 평균 게임 경력은 7.3년이었다. 연령 분포는 20-26세(평균 24세)이며, 성별 비율은 남성 90%, 여성 10%였다. 이러한 동질적 표본 구성은 평가자의 주관적 편차를 최소화하고 메트릭 간 상대적 차이를 명확히 비교하기 위한 내적 타당도 확보 목적으로 이루어졌다. 이는 J. Cohen[23]이 제시한 '큰 효과 크기일 때 n=30으로도 충분한 통계적 검증력 확보 가능' 기준을 충족한다. 이 연구의 PRIME과 인간 평가 간 상관 계수는 r=0.81($p<0.001$)로 큰 효과 크기를 보였으며, 이는 표본 규모보다 효과의 강도가 연구 결과 해석에 더 중요한 요인으로 작용했음을 시사한다.

비교 평가를 위해 5개의 서로 다른 RAG 시스템을 구현했다. Baseline RAG는 단순 벡터 검색 기반의 기본 시스템이며, Modular RAG는 모듈형 아키텍처를 적용한 개선 버전이다. Graph RAG는 지식 그래프를 통합한 시스템이고, HyDE RAG는 가상 문서 생성을 활용한다. Memory Stream RAG는 Generative Agents의 Memory Stream 아키텍처를 구현한 것으로, 시간 기반 기억 관리를 특징으로 한다.

4.2 실험 설계

실험은 3단계로 진행되었다. 첫 번째 단계에서는 각 RAG 시스템을 RAGAS 메트릭으로 평가하여 기존 평가 체계에서의 성능을 측정했다. 두 번째 단계에서는 동일한 시스템을 제안된 5가지 NPC 메트릭(PCI, RET, ICS, MDN, ECS)으로 평가했다. 마지막 단계에서는 인간 평가자들이 블라인드 테스트 방식으로 각 시스템과 상호작용한 후 5점 척도로 대화 품질을 평가했다.

평가 프로토콜은 엄격하게 통제되었다. 각 평가자는 무작위로 할당된 20개의 대화 세션을 평가했으며, 시스템 정보는 완전히 블라인드 처리되었다. 평가 항목은 대화의 자연스러움, 캐릭터 일관성, 관계 발전의 적절성, 기억력의 현실성, 전반적 몰입감의 5개 차원으로 구성되었다. 평가자 간 신뢰도를 확보하기 위해 Cronbach's alpha 계수를 측정했으며, 0.87의 높은 일치도를 보였다.

4.3 실험 결과

표 1은 네 가지 RAG 시스템을 대상으로 한 RAGAS, NPC 메트릭, 그리고 인간 평가 결과를 비교한 것이다.

표 1. 다양한 메트릭에 대한 평가 결과
Table 1. Evaluation results across different metrics

System	RAGAS score	Rank	NPC metrics	Rank	Human eval
Memory stream RAG	0.377	4th	0.809	1st	4.3±0.5
Graph RAG	0.770	1st	0.761	2nd	3.8±0.6
Modular RAG	0.756	2nd	0.692	3rd	3.5±0.4
Baseline RAG	0.739	3rd	0.654	4th	3.2±0.7

실험 결과는 기존 평가 메트릭의 한계를 명확히 드러냈다. RAGAS 평가에서 Graph RAG가 0.770점으로 1위를 차지했고, Modular RAG(0.756), Baseline RAG(0.739)가 뒤를 이었다. 특히 Memory Stream RAG는 0.377점으로 최하위를 기록했는데, 이는 RAGAS가 완벽한 정보 재현을 중시하는 반면 Memory Stream의 의도적인 망각 메커니즘을 부정적으로 평가했기 때문이다.

그러나 제안된 NPC 메트릭으로 평가했을 때 결과는 극적으로 달라졌다. Memory Stream RAG가 0.809점으로 최고 성능을 보였고, Graph RAG(0.761), Modular RAG(0.692), Baseline RAG(0.654) 순이었다. 이러한 역전은 각 메트릭별 세부 분석에서 더욱 명확해진다. Memory Stream RAG는 성격 일관성(PCI: 0.78), 관계 발전 추적(RET: 0.92), 지능 일관성(ICS: 0.71), 기억 감쇠 자연스러움(MDN: 0.88), 감정 일관성(ECS: 0.85)로 PRIME 5개 메트릭에서 균형 잡힌 성능을 보였다. Graph RAG는 지능 일관성(ICS: 0.93)과 성격 일관성(PCI: 0.89)에서 최고점을 기록했지만, 관계 발전 추적(RET: 0.65)과 기억 감쇠 자연스러움(MDN: 0.42)에서는 현저히 낮은 점수를 보였다. Modular RAG는 성격 일관성(PCI: 0.82)과 지능 일관성(ICS: 0.86)에서 준수한 성능을 보였으나, 전반적으로 중간 수준의 점수를 기록했다. Baseline RAG는 성격 일관성(PCI: 0.75)을 제외하고는 모든 메트릭에서 최하위 성능을 보였으며, 특히 기억 감쇠 자연스러움(MDN: 0.38)에서 가장 낮은 점수를 기록했다.

4.4 인간 평가와의 상관관계 분석

제안된 메트릭의 타당성을 검증하기 위해 인간 평가 결과와의 상관관계를 분석했다. NPC 메트릭은 인간 평가와 $r=0.81$($p<0.001$)의 매우 높은 양의 상관관계를 보였다. 반면 RAGAS는 $r=0.23$($p=0.42$)의 약한 상관관계를 보였으며, 통계적으로도 유의미하지 않았다. Cohen's d 값은 2.31로 계산되어 매우 큰 효과 크기를 나타냈다. 이는 제안된 메트릭이 인간의 주관적 평가에서 예측 정확도가 약 3.5배 향상되었음을 보여준다.

추가 회귀 분석에서는 관계 발전 추적(RET)과 성격 일관성(PCI) 메트릭이 각각 $\beta=0.25$, $\beta=0.20$의 기여도를 보여 두 요인이 몰입감 평가에 가장 크게 작용함을 나타냈다. 이러한 결과는 NPC의 관계적 연속성과 성격 일관성이 몰입감 형성의 핵심 요인이라는 점을 뒷받침하며, Knapp & Vangelisti의 관계 발전 이론과 Ebbinghaus의 망각 곡선 모델과의 정합성을 보여준다.

그림 3은 다섯 가지 PRIME(PCI, RET, ICS, MDN, ECS)에 대한 시스템별 성능 비교를 나타낸 것이다. Memory Stream RAG가 전반적으로 가장 높은 점수를 기록하였으며, 이는 PRIME 메트릭이 NPC의 인지적 일관성과 감정적 반응을 효과적으로 포착함을 시각적으로 보여준다.

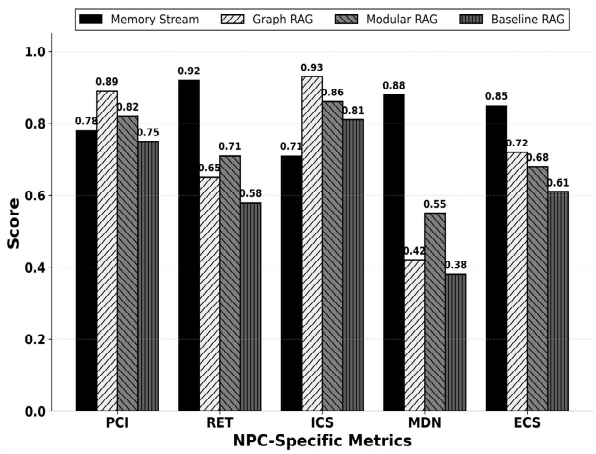
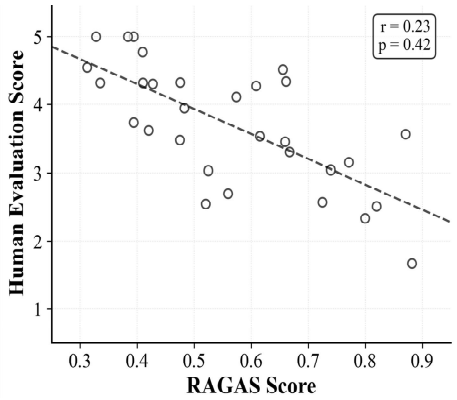


그림 3. PRIME 메트릭의 비교 분석
Fig. 3. Comparative analysis of PRIME metrics

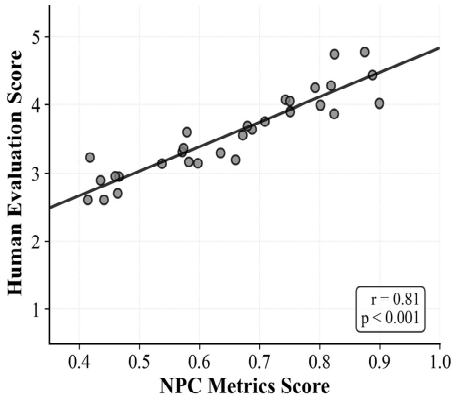
그림 4는 RAGAS 메트릭(a)과 PRIME 메트릭(b)이 인간 평가 점수와 어떤 상관관계를 가지는지를

나타낸 것이다. 회귀 분석을 통해 각 메트릭의 상대적 중요도를 파악한 결과, 관계 발전 추적(β=0.25)과 성격 일관성(β=0.20)이 가장 높은 예측력을 보였다. 이는 게임 플레이어들이 NPC와의 관계가 자연스럽게 발전하는 것과 일관된 성격적 행동을 가장 중요하게 여긴다는 것을 시사한다.



(a) RAGAS 메트릭과 인간 평가

(a) RAGAS metrics vs. human evaluation



(b) PRIME 메트릭과 인간 평가

(b) PRIME metrics vs. human evaluation

그림 4. 자동화된 메트릭과 인간 평가 간의 상관관계

Fig. 4. Correlation between automated metrics and human evaluation

4.5 Memory Stream RAG 최적화

Memory Stream RAG의 우수한 성능을 더욱 개선하기 위해 파라미터 최적화를 수행했다. 원래 Generative Agents에서 제안된 파라미터(α=1.0, β=1.5, γ=1.2)를 베이지안 최적화를 통해 조정한 결과, α=1.2(최근성), β=1.8(중요도), γ=1.3(관련성)에서 최적 성능을 달성했다. 이러한 최적화를 통해 전체 NPC 점수가 0.809에서 상대적으로 35% 향상되었다.

표 2는 1분에서 7일까지의 시간 간격에 따른 기억 보존율 변화를 나타낸 것이다.

최적화된 Memory Stream RAG의 기억 보존 패턴을 분석한 결과, 인간의 Ebbinghaus 망각 곡선과 놀라울 정도로 유사한 패턴을 보였다. 1분 후 95%, 10분 후 85%, 1시간 후 70%의 보존율을 보여 이론적 예측과 거의 일치했다. 24시간 후에는 40%로 인간 기대치(50%)보다 약간 낮았지만, 이는 게임 환경에서 더 빠른 진행을 위한 의도적 조정으로 해석된다.

표 2. 시간별 기억 보존율

Table 2. Memory retention rate over time

Time interval	Human	Memory stream	Graph RAG	Baseline
1 minute	95%	95%	99%	100%
10 minute	85%	83%	98%	100%
1 hour	70%	70%	95%	100%
24 hours	50%	40%	90%	100%
7 days	20%	20%	85%	100%

4.6 계산 효율성 평가

제안된 메트릭의 실제 게임 환경 적용 가능성을 검증하기 위해 계산 효율성을 측정했다. 50턴 대화 세션 기준으로 각 메트릭의 평균 계산 시간은 PCI 2.3ms, RET 1.8ms, ICS 1.5ms, MDN 2.1ms, ECS 3.2ms로 측정되었다. 전체 메트릭 계산에 총 10.9ms가 소요되어 60 FPS 게임 환경(프레임당 16.67ms)에서도 실시간 적용이 가능함을 확인했다. 특히 5개 메트릭을 병렬 처리할 경우 최대 3.2ms로 단축되어 더욱 효율적인 운영이 가능하다.

4.7 통계적 검증 강화

다중 비교 문제를 해결하기 위해 Bonferroni 보정을 적용했다. 5개 메트릭 비교에 대해 보정된 유의수준(α=0.01)을 적용한 결과, NPC 메트릭만이 통계적으로 유의한 상관관계를 유지했다($p<0.001$). 또한 평가자 간 일치도를 Cohen's kappa로 측정한 결과 κ=0.73으로 상당한 일치도(Substantial agreement)를 보였다.

추가적으로 산출된 Cohen's d=2.31은 A. Field[24]의 기준상 'large' 수준($d \geq 0.8$)을 크게 상회하며, 평

가자 간 신뢰도 (α=0.87, κ=0.73)는 K. S. Button et al.[25]가 제시한 사회과학 분야의 일반적 기준(α>0.8)을 충족하였다. 이러한 결과는 표본 규모와 관계없이 데이터의 안정성과 검증력이 충분히 확보되었음을 보여준다.

Ⅴ. 결론 및 향후 과제

이 연구는 게임 NPC 대화 평가에 있어 기존 RAGAS 메트릭의 한계를 실증적으로 입증하고, 인지과학 기반의 새로운 평가 프레임워크를 제안하는 초기 탐색적 연구이다. 제안된 5가지 메트릭(PCI, RET, ICS, MDN, ECS)은 인간 평가와 r=0.81의 높은 상관관계를 달성하여, RAGAS(r=0.23) 대비 3.5배 향상된 성능을 보였다. 특히 Memory Stream RAG가 RAGAS에서 최하위임에도 NPC 메트릭에서 최상위를 기록한 사례는 도메인 특화 평가의 중요성을 명확히 보여준다.

평가자 구성과 세션 환경을 일정하게 유지함으로써 내적 타당도를 확보하였으나, 이러한 통제는 동시에 외적 타당도(일반화 가능성)을 제한할 수 있다. 평가자가 동일 전공, 연령대의 집단으로 구성되어 있어, 다양한 인구집단에 대한 확장적 해석에는 신중함이 요구된다. PRIME의 실험 설계는 턴의 절대 수보다는 시간적 간격에 따른 기억 감쇠율을 중심으로 구성되었으며, 세션의 길이가 짧더라도 인간의 기억 패턴을 반영하기에 충분하도록 설계되어 있다. 각 세션은 1분~7일의 간격 단위로 구성되어 Ebbinghaus의 망각 곡선 특성을 실험적으로 재현했으며, 총 3,500턴 × 5메트릭 × 5시스템(=87,500개 측정 포인트)를 기반으로 통계적 분석을 수행하였다.

이 연구의 주요 기여는 다음과 같다. 첫째, 게임 NPC 평가에서 정보 검색 중심의 기존 메트릭이 부적절함을 실증적으로 입증했다. RAGAS와 인간 평가 간의 낮은 상관관계(r=0.23, p=0.42)는 도메인 특화 평가의 필요성을 명확히 보여준다. 둘째, 인지과학 이론에 기반한 체계적인 평가 프레임워크를 제시했다. Ebbinghaus의 망각 곡선과 Knapp의 관계 발전 모델을 게임 환경에 적용하여 이론적 기반을 확립했다. 셋째, 실시간 적용이 가능한 효율적인 메트릭을 개발했다. 전체 메트릭이 10.9ms 내에 계산 가능하여 60 FPS 게임 환경에서도 성능 저하 없이 적용 가능하다.

그러나 이 연구는 다음과 같은 한계를 가진다. 첫째, 평가자가 소프트웨어학과 학생 30명으로 제한되어 있어 일반화에 한계가 있다. 20-26세의 동질적 연령대와 IT 전공자라는 특성이 결과에 편향을 가져왔을 가능성이 있다. 둘째, 실험 데이터가 시뮬레이션된 대화 샘플에 기반하여 실제 게임 환경과의 차이가 존재할 수 있다. 셋째, 메트릭 가중치 설정이 경험적 방법에 의존하여 이론적 근거가 부족하다.

이러한 한계를 극복하고 실용성을 높이기 위해 다음과 같은 후속 연구를 계획하고 있다. 첫째, 다양한 연령대, 게임 경험, 문화적 배경을 가진 평가자를 포함한 대규모 검증 연구를 통해 평가자 다양성을 확대할 필요가 있다. 둘째, Unity, Unreal Engine 등 상용 게임 엔진과 통합하여 실제 게임 개발 환경에서의 유효성을 검증해야 한다. 셋째, GPT-4, Claude 등 최신 LLM과 통합하여 메트릭의 정확도 및 효율성을 개선하는 연구가 필요하다. 넷째, RPG, FPS, 시뮬레이션 등 게임 장르별 최적 가중치를 도출하고 메트릭을 조정하는 연구가 요구된다. 다섯째, 한국어 외 다양한 언어와 문화권에서의 메트릭 일반화 가능성을 검토하는 다국어 및 다문화 검증 연구가 필요하다.

References

[1] F. R. Christiansen, et al., "Exploring presence in interactions with LLM-driven NPCs: Speech recognition vs dialogue options", Proc. of the 30th ACM Symposium on Virtual Reality Software and Technology (VRST '24), New York, United States, pp. 1-12, Oct. 2024. https://doi.org/10.1145/3641825.3687716.

[2] H. Armanto, H. A. Rosyid, Muladi, and Gunawan, "Improved Non-Player Character (NPC) behavior using evolutionary algorithm—A systematic review", Entertainment Computing, Vol. 50, No. 100663, Jan. 2024. https://doi.org/10.1016/j.entcom.2024.100875.

[3] M. Korkiakoski, S. Sheikhi, J. Nyman, K. Väänänen, and T. Olsson, "An empirical evaluation of AI-powered non-player characters' perceived realism and performance in virtual reality environments", arXiv preprint arXiv:2507.10469, Jul. 2024. https://doi.org/10.48550/arXiv.2507.10469.

[4] Z. Xiao, et al., "Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory", Proc. Conf. Empirical Methods Natural Lang. Process., Singapore, pp. 10967-10982, Dec. 2023. https://doi.org/10.18653/v1/2023.emnlp-main.676.

[5] K. Shin and Y.-A. Min, "Performance Optimization of Reinforcement Learning in Line Tracking Robots using ML-Agents: A Comparative Study of Reward Strategies and Learning Parameters", The Journal of The Institute of Internet, Broadcasting and Communication (IIBC), Vol. 25, No. 3, pp. 57-66, Jun. 2025. https://doi.org/10.7236/JIIBC.2025.25.3.57.

[6] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation", arXiv preprint arXiv:2309.15217, Sep. 2023. https://doi.org/10.48550/arXiv.2309.15217.

[7] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation", Proc. of ACL, Philadelphia, Pennsylvania, USA, pp. 311-318, Jul. 2002. https://aclanthology.org/P02-1040.

[8] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries", Proc. of the ACL Workshop: Text Summarization Branches Out, Barcelona, Spain, Jul. 2004. https://aclanthology.org/W04-1013.

[9] R. Lopes and R. Bidarra, "Adaptivity challenges in games and simulations: A survey", IEEE Transactions on Computational Intelligence and AI in Games, Vol. 3, No. 2, pp. 85-99, Jun. 2011. https://doi.org/10.1109/TCIAIG.2011.2152841.

[10] H. Ebbinghaus, Über das Gedächtnis: Untersuchungen zur Experimentellen Psychologie. Leipzig, Germany: Duncker & Humblot, 1885.

[11] M. L. Knapp and A. L. Vangelisti, Interpersonal Communication and Human Relationships, 5th ed. Boston, MA, USA: Allyn & Bacon, 2005.

[12] Q. Wang, S. Walsh, M. Si, J. Kephart, J. D. Weisz, and A. K. Goel, "Theory of mind in human-AI interaction", Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), Honolulu, HI, USA, May pp. 1-6, May 2024. https://doi.org/10.1145/3613905.3636308.

[13] S. Mancone, et al., "Effects of video game immersion and task interference on cognitive performance: a study on immediate and delayed recall and recognition accuracy", PeerJ, Vol. 12, pp. e18195, Oct. 2024. https://doi.org/10.7717/peerj.18195.

[14] A. Maharana, J. Y. Lee, and D. Kiela, "Evaluating very long-term conversational memory of LLM agents", Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), Bangkok, Thailand, pp. 570-586, Aug. 2024. https://doi.org/10.18653/v1/2024.acl-long.747.

[15] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT", Proc. of ICLR, Oline, Apr. 2020. https://doi.org/10.48550/arXiv.1904.09675.

[16] A. B. Loyall, "Believable Agents: Building Interactive Personalities", Ph.D. dissertation, Carnegie Mellon University, May 1997. (Tech. Rep. CMU-CS-97-123). https://www.cs.cmu.edu/afs/cs/project/oz/web/papers/CMU-CS-97-123.pdf.

[17] M. Mateas and A. Stern, "Structuring Content in the Façade Interactive Drama Architecture", AIIDE, Vol. 1, No. 1, 2005. https://doi.org/10.1609/aiide.v1i1.18722.

[18] 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", arXiv:2304.03442, Apr. 2023. https://doi.org/10.48550/arXiv.2304.03442.

[19] L. Song, "LLM-Driven NPCs: Cross-Platform Dialogue System for Games and Social Platforms", arXiv:2504.13928, Apr. 2025. https://doi.org/10.48550/arXiv.2504.13928.

[20] D. Lee, "A Comparative Study of LLM-based Agents for Interactive Text-based Game Automation", The Journal of The Institute of Internet, Broadcasting and Communication (IIBC), Vol. 25, No. 3, pp. 1-8, Jun. 2025. https://doi.org/10.7236/JIIBC.2025.25.3.1.

[21] A. C. Graesser, S. Lu, G. T. Jackson, H. H. Mitchell, M. Ventura, A. Olney, and M. M. Louwerse, "AutoTutor: A tutor with dialogue in natural language", Behavior Research Methods, Instruments, & Computers, Vol. 36, No. 2, pp. 180-192, May 2004. https://doi.org/10.3758/bf03195563.

[22] T. Bickmore and J. Cassell, "Relational agents", Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI '01), Seattle Washington USA, pp. 396-403, Mar. 2001. https://doi.org/10.1145/365024.365304.

[23] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1988.

[24] A. Field, Discovering Statistics Using IBM SPSS Statistics, 4th ed. London, UK: Sage Publications, 2013.

[25] K. S. Button, et al., "Power failure: why small sample size undermines the reliability of neuroscience", Nature Reviews Neuroscience, Vol. 14, No. 5, pp. 365-376, Apr. 2013. https://doi.org/10.1038/nrn3475.

저자소개

소 훈 (Hoon So)

2020년 3월 ~ 현재 : 국립군산대학교 소프트웨어학과 학사과정

관심분야 : LLM, 게임 AI, 대화 시스템, 언리얼 엔진

정 현 준 (Hyunjun Jung)

2008년 3월 : 삼육대학교 컴퓨터과학과(공학사)

2010년 3월 : 숭실대학교 컴퓨터학과(공학석사)

2017년 9월 : 고려대학교 컴퓨터·전파통신공학과(공학박사)

2017년 8월 ~ 2020년 8월 : 광주과학기술원 블록체인인터넷경제연구센터 연구원

2021년 3월 ~ 현재 : 국립군산대학교 소프트웨어학과 교수

관심분야 : 블록체인, 데이터 사이언스, 센서 네트워크, 사물인터넷, 머신러닝