

제너레이티브 에이전트: 인간 행동의 인터랙티브 시뮬라크라

박준성 스탠포드 대

학교 스탠포드, 미국

joonspk@stanford.edu

메리디스 링겔 모리스

구글 딥마인드 사애틀,

워싱턴, 미국

merrie@google.com

조셉 C. 오브라이

언 스탠포드 대학교 스

탠포드, 미국

jobrien3@stanford.edu

퍼시 량 스탠포드 대

학교 스탠포드, 미국

pliang@cs.stanford.edu

캐리 J. 카이

구글 리서치 마운틴뷰, 캘

리포니아, 미국

cjcai@google.com

マイ클 번스타인 스

탠포드 대학교 스탠포드,

미국

msb@cs.stanford.edu



그림 1: 제너레이티브 에이전트는 인터랙티브 애플리케이션을 위한 인간 행동의 사실적인 시뮬라크라입니다. 이 작업에서는 심즈(The Sims)를 연상시키는 샌드박스 환경을 25개의 에이전트로 채워 제너레이티브 에이전트를 시연합니다. 사용자는 에이전트가 하루를 계획하고, 뉴스를 공유하고, 관계를 형성하고, 그룹 활동을 조율하는 모습을 관찰하고 개입할 수 있습니다.

초록

인간 행동의 사실적인 프록시는 물입형 환경부터 대인 커뮤니케이션을 위한 리허설 공간, 프로토타이핑 도구에 이르기까지 다양한 인터랙티브 애플리케이션을 강화할 수 있습니다. 이 백서에서는 사실적인 인간 행동을 시뮬레이션하는 컴퓨팅 소프트웨어 에이전트인 제너레이티브 에이전트를 소개합니다. 제너레이티브 에이전트는 일어나서 아침을 요리하고 일터로 향하며, 아티

스트는 그림을 그리는 동시에

영리 또는 상업적 이익을 목적으로 사본을 만들거나 배포하지 않고 사본에 본 고지 및 첫 페이지에 전체 인용문을 표시하는 경우 개인 또는 수업용으로 이 저작물의 일부 또는 전부를 디지털 또는 하드 카피할 수 있는 권한이 수수료 없이 부여됩니다. 이 저작물의 타사 구성을 요소에 대한 저작권은 존중되어야 합니다. 그 외의 모든 사용은 소유자/저자에게 문의하시기 바랍니다.

UIST '23, 2023년 10월 29일~11월 1일, 미국 샌프란시스코, 캘리포니아주

© 2023 저작권 소유자/저자 보유. ACM ISBN 979-8-4007-0132-0/23/10.
<https://doi.org/10.1145/3586183.3606763>

글을 작성하고, 의견을 나누고, 서로를 알아차리고, 대화를 시작하며, 다음 날을 계획할 때 지난 날을 기억하고 반영합니다. 제너레이티브 에이전트를 지원하기 위해 대규모 언어 모델을 확장하여 자연어를 사용하여 에이전트의 경험에 대한 완전한 기록을 저장하고, 시간이 지남에 따라 이러한 기억을 더 높은 수준의 반영으로 합성하고, 동적으로 검색하여 행동을 계획하는 아키텍처를 설명합니다. 유니티는 제너레이티브 에이전트를 인스턴스화하여 최종 사용자가 자연어를 사용하여 25명의 에이전트로 구성된 작은 마을과 상호 작용할 수 있는 심즈에서 영감을 받은 대화형 샌드박스 환경을 채웁니다. 평가에서 이러한 제너레이티브 에이전트는 믿을 수 있는 개인 및 새로운 사회적 행동을 생성합니다. 예를 들어, 한 에이전트가 발렌타인데이 파티를 열고 싶다는 사용자 지정 개념 하나만 있으면 에이전트가 다음 두 에이전트에게 파티 초대장을 자율적으로 배포합니다.

전트

일, 새로운 지인을 사귀고, 서로에게 파티 데이트를 신청하고, 적절한 시간에 함께 파티에 참석할 수 있도록 조율합니다. 우리는 에이전트 아키텍처의 구성 요소인 관찰, 계획, 반영이 각각 에이전트 행동의 신뢰성에 결정적으로 기여한다는 것을 절제를 통해 입증했습니다. 이 연구는 대규모 언어 모델과 컴퓨터 대화형 에이전트를 융합하여 인간 행동의 믿을 수 있는 시뮬레이션을 가능하게 하는 아키텍처 및 상호 작용 패턴을 소개합니다.

CCS 개념

- **인간 중심 컴퓨팅 → 대화형 시스템 및 도구, - 컴퓨팅 방법론
→ 자연어 처리.**

키워드

인간-AI 상호 작용, 에이전트, 제너레이티브 AI, 대규모 언어 모델

ACM 참조 형식:

박준성, 조셉 C. 오브라이언, 캐리 J. 카이, 메러디스 링겔 모리스, 퍼시 리昂, 마이클 S. 번스타인. 2023. 제너레이티브 에이전트: 인간 행동의 인터랙티브 시뮬라크라. 제36회 연례 ACM 사용자 인터페이스 소프트웨어 및 기술 심포지엄(UIST '23), 2023년 10월 29일~11월 1일, 미국 샌프란시스코, 캘리포니아: ACM, 미국 뉴욕, 뉴욕, 22 페이지:
[//doi.org/10.1145/3586183.3606763](https://doi.org/10.1145/3586183.3606763)

1 소개

믿을 수 있는 인간의 행동을 반영하는 인터랙티브 인공 사회를 어떻게 만들 수 있을까요? 심즈와 같은 샌드박스 게임부터 인지 모델 [23], 가상 환경[10, 59]과 같은 애플리케이션에 이르기까지 40년 이상 연구자와 실무자들은 인간 행동의 믿을 수 있는 프록시 역할을 할 수 있는 컴퓨팅 에이전트를 구상해 왔습니다. 이러한 비전에서 컴퓨터 기반 에이전트는 과거 경험에 따라 일관되게 행동하고 환경에 믿을 수 있게 반응합니다. 이러한 인간 행동 시뮬레이션은 가상 공간과 커뮤니티에 실제와 같은 사회 현상을 구현하고[27, 80], 드물지만 어려운 대인 관계 상황을 처리하는 방법을 교육하며 [44, 52, 94], 사회 과학 이론을 테스트할 수 있습니다[12, 46], 이론 및 사용성 테스트를 위한 모델 휴먼 프로세서 제작[23, 39, 51], 유비쿼터스 컴퓨팅 애플리케이션[31] 및 소셜 로봇[10, 14], 오픈 월드에서 복잡한 인간 관계를 탐색할 수 있는 비플레이어블 게임 캐릭터[59, 85]의 기반이 되기도 합니다.

그러나 인간 행동의 공간은 방대하고 복잡합니다[85, 108]. 단일 시점에서 인간 행동을 시뮬레이션할 수 있는 대규모 언어 모델[18]

의 놀라운 발전에도 불구하고[39, 80], 장기적인 일관성을 보장하는 완전 일반 에이전트는 새로운 상호작용, 갈등, 사건이 발생하고 시간이 지남에 따라 사라지면서 지속적으로 증가하는 기억을 관리하는 아키텍처가 더 적합하며, 그 사이에 펼쳐지는 계단식 사회적 역학을 처리하는 아키텍처가 더 적합할 것입니다.

는 개인 및 긴급한 집단 행동에 대한 믿을 만한 시뮬라크르를 생성한다는 것을 보여줍니다. 제너레이티브 에이전트는 자신과 다른 에이전트, 환경에 대한 다양한 추론을 도출하고, 자신의 행동과 경험을 반영한 일일 계획을 세우고, 그 계획을 실행하고, 적절한 경우 반응하고, 다시 계획을 세우며, 최종 사용자가 환경을 변경하거나 자연어로 명령하면 이에 반응합니다. 예를 들어, 제너레이티브 에이전트는 금식 중이던 음식이 타는 것을 보고 스토브를 끄고, 화장실이 비어 있으면 밖에서 기다리며, 대화하고 싶은 다른 에이전트를 만나면 대화를 중단합니다.¹ 제너레이티브 에이전트로 가득한 사회는 새로운 관계가 형성되고 정보가 확산되며 에이전트 간에 협력이 발생하는 새로운 사회적 역학이 특징입니다.

제너레이티브 에이전트를 활성화하기 위해 대규모 언어 모델을 사용하여 믿을 수 있는 동작을 생성하기 위해 관련 메모리를 저장, 합성 및 적용하는 에이전트 아키텍처를 설명합니다. 아키텍처는 세 가지 주요 구성 요소로 이루어져 있습니다. 첫 번째는 에이전트의 경험에 대한 포괄적인 목록을 자연어로 기록하는 장기 메모리 모듈인 *메모리 스트림입니다.* 메모리 검색 모델은 관련성, 최근성, 중요도를 결합하여 상담원의 순간순간 행동을 파악하는데 필요한 기록을 표시합니다. 두 번째는 *반영으로*, 시간이 지남에 따라 기억을 더 높은 수준의 추론으로 종합하여 상담원이 자신과 타인에 대한 결론을 도출하여 행동을 더 잘 안내할 수 있도록 합니다. 세 번째는 *계획으로*, 이러한 결론과 현재 환경을 높은 수준의 행동 계획으로 변환한 다음 행동과 반응을 위한 세부 행동으로 재귀적으로 변환합니다. 이러한 성찰과 계획은 다시 메모리 스

여러 상담원이 필요합니다. 성공하려면 다음을 검색할 수 있는 접근 방식이 필요합니다. 관련 이벤트와 상호작용을 장기간에 걸쳐 기억하고, 이러한 기억을 반영하여 일반화하고 더 높은 수준의 추론을 도출하며, 이러한 추론을 적용하여 상담원 행동의 순간과 장기적인 맥락에서 의미 있는 계획과 반응을 만들어냅니다.

이 백서에서는 제너레이티브 에이전트(제너레이티브 모델을 활용하여 믿을 수 있는 인간 행동을 시뮬레이션하는 에이전트)를 소개합니다.

트림으로 피드백되어 상담원의 향후 행동에 영향을 미칩니다.

이 아키텍처는 역할극과 소셜 프로토타이핑부터 가상 세계와 게임에 이르기까지 다양한 영역에 적용할 수 있습니다. 소셜 롤플레이 시나리오(예: 면접 준비)에서 사용자는 어렵고 갈등이 있는 대화를 안전하게 리허설할 수 있습니다. 소셜 플랫폼을 프로토타이핑할 때 디자이너는 임시 페르소나를 넘어 역동적이고 복잡한 인터랙션이 펼쳐지는 프로토타입을 만들 수 있습니다. 예 대한 연구도 진행 중입니다. 이 백서에서는 심즈와 같은 게임에서 영감을 받은 에이전트로 구성된 소규모 대화형 사회를 만드는 기능에 중점을 둡니다.² 아키텍처를 ChatGPT 대형 언어에 연결하면 다음과 같습니다.

모델[77]에서 25명의 에이전트로 구성된 사회를 게임 환경에 구현했습니다. 최종 사용자는 이러한 에이전트를 관찰하고 상호작용할 수 있습니다. 예를 들어 최종 사용자나 개발자가 게임 내에서 마을에서 발렌타인데이 파티를 열고 싶다면 기존 게임 환경에서는 수십 개의 캐릭터 행동을 수동으로 스크립팅해야 했습니다. 하지만 제너레이티브 에이전트를 사용하면 한 에이전트에게 파티를 열고 싶다고 말하는 것만으로 충분하다는 것을 보여줍니다. 파티 기획자는 다른 에이전트를 파티에 초대해야 하고, 참석자는 초대장을 기억해야 하며, 기억한 에이전트가 실제로 파티에 참석해야 하는 등 실패할 가능성이 많은 상황에서도 에이전트는 성공합니다. 파티에 대한 소문을 퍼뜨린 다음

소문을 퍼뜨린 다음

¹ 제너레이티브 에이전트가 행동을 취하거나 장소를 이동하는 것을 지칭할 때, 이는 가독성을 높이기 위한 약어이며 사람과 같은 에이전트 역할을 수행한다는 의미는 아닙니다. 디즈니 애니메이션 캐릭터와 유사한 에이전트의 행동은 믿을 수 있다는 느낌을 주기 위한 것이지만, 실제 에이전시를 의미하지는 않습니다.

² 제너레이티브 에이전트 사회의 실제 시뮬레이션 데모는 다음 링크에서 볼 수 있습니다: https://reverie.herokuapp.com/UIST_Demo/. 시뮬레이션 코드의 공개 리포지토리는 다음 링크에 있습니다: <https://github.com/joonspk-research/>.

제너레이티브_에이전트

한 상담원이 다른 상담원에게 파티에 함께 가자고 요청하는 등 사용자가 생성한 단일 시드 제안을 통해 다양한 방식으로 파티에 참여할 수 있습니다.

유니티는 제너레이티브 에이전트에 대해 두 가지 평가를 실시했습니다. 에이전트가 개별적으로 믿을 만한 개별 행동을 생성하는지 테스트하기 위한 통제된 평가와 에이전트의 안정성과 새로운 사회적 행동을 이해하기 위해 이를 동안 게임 시간 동안 개방형 방식으로 서로 상호작용하는 엔드투엔드 평가였습니다. 기술 평가에서는 에이전트의 지식과 행동을 자연어로 '인터뷰'하여 평가하는 방법론을 활용하여 에이전트의 캐릭터 유지, 기억력, 계획, 반응, 반영 능력을 정확하게 조사했습니다. 상담원의 기억, 반성, 계획에 대한 액세스를 제한하는 몇 가지 기능을 비교했습니다. 이러한 각 요소는 이러한 인터뷰 작업 전반에서 우수한 성과를 거두는 데 매우 중요하다는 것을 관찰했습니다. 기술 및 엔드투엔드 평가 전반에서 가장 흔한 오류는 상담원이 관련 기억을 검색하지 못하거나, 상담원의 기억을 조작하거나, 언어 모델에서 지나치게 형식적인 말이나 행동을 이어받은 경우 발생했습니다.

요약하면, 이 백서는 다음과 같은 기여를 합니다:

- 제너레이티브 에이전트는 에이전트의 변화하는 경험에 따라 동적으로 조절되는 인간 행동의 사실적인 시뮬라크라입니다.
리렌스와 환경.
- 제너레이티브 에이전트가 기억, 검색, 반영, 다른 에이전트와 상호 작용할 수 있는 새로운 아키텍처
에이전트, 그리고 동적으로 변화하는 상황을 통해 계획할 수 있습니다. 이 아키텍처는 대규모 언어 모델의 강력한 프롬프트 기능을 활용하고 이러한 기능을 보완하여 장기적인 에이전트 일관성, 동적으로 진화하는 메모리 관리 기능, 상위 수준의 반영을 재귀적으로 생성할 수 있도록 지원합니다.
- 두 가지 평가, 통제된 평가와 엔드투엔드 평가, 중요성의 인과 관계를 확립하는 평가
아키텍처의 구성 요소를 분석하고 부적절한 메모리 검색 등으로 인해 발생하는 고장을 식별할 수 있습니다.
- 기회와 윤리적, 사회적 위험에 대한 논의
대화형 시스템에서 제너레이티브 에이전트입니다. 우리는 다음과 같이 주장합니다.
이러한 에이전트는 사용자가 준사회적 관계를 형성할 위험을 완화하도록 조정되어야 하며, 딥페이크와 맞춤형 설득으로 인한 위험을 완화하도록 로깅되어야 하며, 디자인 프로세스에서 인간 이해관계자를 대체하는 것이 아니라 보완하는 방식으로 적용되어야 합니다.

2 관련 작업

이 섹션에서는 인간과 인공지능의 상호작용에 관한 선행 문헌을 살펴보고, 그 안에서 믿을 수 있는 인간 행동의 프록시 구축이라는 의제를 살펴봅니다. 한때 상호작용, 게임, 인공지능 커뮤니티에서 북극성처럼 환영받았던 이 의제는[10, 59, 85, 86] 인간 행동의 복잡성으로 인해 여전히 도전적인 과제로 남아 있습니다[17, 108]. 이 연구를 종합하면 대규모 언어 모델 자체만으로는 충분하지 않지만 적절한 아키텍처를 활용하면 믿을 수 있는 에이전트를 만들 수 있는 새로운 각도가 열릴 수 있습니다.

2.1 인간과 AI의 상호작용

대화형 인공 지능 시스템은 인간의 시각과 능력을 컴퓨터 인공물에 결합하여 증강할 수 있는 것을 목표로 합니다.

사용자[4, 30]. 사용자가 모델 동작을 대화형으로 지정할 수 있는 방법을 연구하는 작업은 오랫동안 진행되어 왔습니다. 예를 들어, 크레용은 대화형 머신 러닝의 초기 비전을 제시하여 비전문가인 사용자도 분류자를 훈련할 수 있도록 했습니다[30]. 추가 작업을 통해 최종 사용자가 분류 목표를 시스템에 설명하는 방법을 예시[34] 또는 데모[32]를 통해 명확하게 표현할 수 있었습니다. 최근의 발전은 이러한 탐구를 딥 러닝[63] 및 프롬프트 기반 저작[50, 67, 106]으로 확장했습니다.

한편, 인간과 컴퓨터 간의 상호작용에서 언어 및 에이전트 기반 상호작용에 대한 지속적인 연구가 진행되어 왔습니다. SHRDLU [103] 및 ELIZA [102]와 같은 형성적인 연구는 컴퓨팅 시스템과의 자연어 상호 작용과 관련된 기회와 위험을 보여주었습니다. 연구가 진행됨에 따라 자율 에이전트가 위임과 상호 작용에 대한 새로운 은유를 제공할 수 있음이 분명해졌지만[68], 인간과 에이전트 간의 위임의 경계는 여전히 지속적인 논쟁과 개선의 대상이 되고 있습니다[47, 89, 90]. 최근 이 기술은 대규모의 복잡한 온라인 소셜 환경에서 에이전트가 자연어를 통해 상호 작용할 수 있는 수준의 안정성에 도달했습니다(예: [55]). 자연어 상호작용은 사진 편집[3, 35, 65] 및 코드 편집[88]과 같은 영역에서 사용자 능력을 향상시킬 수 있는 새로운 방식을 제공합니다.

이러한 일련의 연구를 종합하여 이제 대화형 시스템을 위해 인간의 행동을 대리하는 에이전트를 만들고 자연어를 사용하여 상호 작용할 수 있음을 보여줍니다. 이 작업을 통해 GOMS 및 키 입력 수준 모델(KLM)[22, 23]과 같은 인지 모델과 프로토타이핑 도구[80]를 중심으로 인간과 컴퓨터의 기본적인 상호 작용 문제를 검토 할 수 있는 문이 다시 열리게 됩니다,

그리고 유비쿼터스 컴퓨팅 애플리케이션을 중심으로 [26, 31, 101].

2.2 믿을 수 있는 인간 행동의 프록시

이전 문헌에서는 믿을 수 있는 에이전트 또는 믿을 수 있는 에이전트를 설계 및 엔지니어링의 핵심 목표로 설명했습니다. 믿을 수 있는 에이전트는 디즈니 영화에 등장하는 캐릭터와 유사하게 실제와 같은 착각을 불러일으키고, 스스로 의사 결정을 내리고 행동하는 것처럼 보이도록 설계되었습니다[10, 96]. 이러한 에이전트는 우리가 살고 있는 것과 같은 오픈 월드 환경을 채우고 인식할 수 있으며[10, 59], 개인 및 커뮤니티의 가상 시뮬레이션에서 우리 행동의 믿을 만한 대리자가 되기 위해 사용자 또는 다른 에이전트와

의 사회적 상호 작용을 기반으로 하는 방식으로 행동하기 위해 노력합니다[20, 36, 71]. 역사적으로 이러한 에이전트는 지능형 게임 비플레이어 캐릭터(NPC)의 맥락에서 개발되었습니다[59, 85]. 가능하다면 믿을 수 있는 행동을 하는 NPC를 만들면 새로운 내러티브[8, 16, 49, 93]와 에이전트와의 사회적 상호작용[109]을 가능하게 하여 게임과 인터랙티브 픽션에서 플레이어 경험을 향상시킬 수 있습니다. 그러나 더 중요한 것은 게임 세계가 현실 세계의 어포던스를 점점 더 사실적으로 표현한다는 점이며, 2001년에 Laird와 van Lent가 관찰한 바와 같이 이러한 시뮬레이션 세계는 믿을 수 있는 에이전트 개발자가 현실 세계에서 로봇 공학을 구현하거나 시뮬레이션 환경을 처음부터 만들지 않고도 에이전트의 인지 능력을 검증할 수 있는 접근 가능한 테스트베드를 제공합니다[59, 85].

지난 40년 동안 믿을 수 있는 에이전트를 만들기 위한 다양한 접근 방식이 등장했습니다. 그러나 이러한 접근 방식은 구현 과정에서 에이전트의 환경이나 차원을 단순화하는 경우가 많았습니다.

동작을 사용하여 관리하기 쉽게 만들 수 있습니다[17, 73]. 유한 상태 머신[91, 97] 및 행동 트리[41, 54, 82]와 같은 규칙 기반 접근 방식은 사람이 에이전트의 행동을 작성하는 무차별 대입 방식[71]을 설명합니다. 이러한 접근 방식은 오늘날에도 가장 널리 사용되는 간단한 에이전트 생성 방법을 제공하며[69, 74, 108], 매스 이펙트[13] 및 심즈[7] 시리즈와 같은 게임에서 볼 수 있듯이 초보적인 사회적 상호 작용도 처리할 수 있습니다. 그럼에도 불구하고 오픈 월드에서 발생할 수 있는 다양한 상호작용을 포괄적으로 처리할 수 있는 동작을 수동으로 제작하는 것은 불가능합니다. 즉, 결과 에이전트 행동이 상호작용의 결과를 완전히 나타내지 못할 수 있으며[70-72], 스크립트에 하드코딩되지 않은 새로운 절차를 수행할 수 없습니다[91, 97]. 반면, 강화 학습과 같이 믿을 수 있는 에이전트를 만들기 위한 기존의 학습 기반 접근 방식은 에이전트가 자신의 행동을 학습하도록 함으로써 수동 저작의 어려움을 극복했으며, 최근 몇 년 동안 스타크래프트의 알파스타(AlphaStar)[99], 도타 2의 오픈AI 파이브(OpenAI Five)[11] 등의 게임에서 초인적인 성능을 달성했습니다. 그러나 이러한 성공은 주로 학습 알고리즘이 타이밍을 맞출 수 있는 보상을 쉽게 정의할 수 있는 적대적인 게임에서 이루어졌습니다. 오픈 월드에서 믿을 수 있는 에이전트를 만드는 문제는 아직 해결하지 못했습니다[40, 74, 91].

뉴웰이 개척한 컴퓨팅의 인지 아키텍처는 원래의 비전인 믿을 수 있는 에이전트의 포괄적인 특성에 적합한 포괄적인 인지 기능 세트[76]를 지원하기 위한 인프라를 구축하는 것을 목표로 했습니다. 이 프로젝트는 믿을 수 있는 에이전트의 초기 사례 중 일부에 동력을 제공했습니다. 예를 들어, 퀘이크봇-SOAR[60]와 아이카루스[25, 64]는 1인칭 슈팅 게임에서 NPC를 생성했고, 택에어-SOAR[81]는 공중 전투 훈련 시뮬레이션에서 파일럿을 생성했습니다. 이 에이전트들이 사용한 아키텍처는 서로 달랐지만 (Quakebot과 TacAir-SOAR는 SOAR[61]를 사용했고, ICARUS는 SOAR와 ACT-R[6]에서 영감을 받은 자체 변형에 의존했습니다), 기본 원리[62]는 동일했습니다. 이들은 단기 및 장기 기억을 유지하고, 이러한 기억을 상징적 구조로 채우고, 환경을 동적으로 인식하고 수동으로 만들어진 행동 프로세스 중 하나와 일치시키는 인식-계획-행동 주기로 작동했습니다[58, 97]. 인지 아키텍처를 사용하여 제작된 에이전트는 전부는 아니더라도 대부분의 오픈 월드 컨텍스트에 일반화할 수 있는 것을 목표로 했으며, 당시에는 강력한 동작을 보여주었습니다. 그러나 에이전트의 행동 공간은 수동으로 만

들어진 절차적 지식으로 제한되어 있었고, 에이전트가 새로운 행동을 추구하도록 영감을 줄 수 있는 메커니즘을 제공하지 않았습니다. 따라서 이러한 에이전트는 주로 1인칭 슈팅 게임[25, 60] 또는 블록 월드[64]와 같은 비개방형 월드 컨텍스트에 배치되었습니다.

오늘날에도 원래 정의에 설명된 대로 믿을 수 있는 에이전트를 만드는 것은 여전히 미해결 문제로 남아 있습니다[85, 108]. 많은 사람들은 믿을 수 있는 에이전트를 생성하는 현재의 접근 방식이 번거롭고 제한적일 수 있지만 기존 게임플레이와 상호작용을 지원하기에 충분하다고 주장하며 다른 방향으로 나아가고 있습니다 [24, 75, 108]. 저희는 기억을 믿을 만한 행동으로 합성하는 효과적인 아키텍처를 만들 수 있다면 대규모 언어 모델이 이러한 질문을 재검토할 수 있는 기회를 제공한다고 주장합니다. 이 논문에서는 이러한 아키텍처를 향한 한 걸음을 제시합니다.

2.3 대규모 언어 모델과 인간 행동

제너레이티브 에이전트는 대규모 언어 모델을 활용하여 행동을 강화합니다. 중요한 점은 대규모 언어 모델이 학습 데이터에서 광범위한 인간 행동을 인코딩한다는 것입니다[15, 18]. 쉽게 정의된 컨텍스트로 프롬프트하면 이 모델을 사용하여 믿을 수 있는 행동을 생성할 수 있습니다. 최근의 연구는 이러한 접근 방식의 효율성을 입증했습니다. 예를 들어, 소셜 시뮬라크라는 대규모 언어 모델을 사용하여 새로운 소셜 컴퓨팅 시스템을 채용 사용자를 생성하여 새로운 소셜 역학을 프로토타입으로 만들었습니다[80]. 이 접근 방식은 프롬프트 체인[105, 106]을 사용하여 프로토타이핑 중인 시스템에 나타나는 페르소나와 그 행동에 대한 짧은 자연어 설명을 생성했습니다. 다른 경험적 연구에서는 기존의 사회과학 연구[46], 정치 설문조사[92]를 복제하고 합성 데이터를 생성하기도 했습니다[39]. 대규모 언어 모델도 사용자가 참여할 수 있는 대화형 인간 행동을 생성하는 데 사용되었습니다. 예를 들어, 게임 분야에서는 대화형 소설[37]과 텍스트 어드벤처 게임[21]을 만드는데 이러한 모델이 사용되었습니다. 액션 시퀀스를 생성하고 분해하는 능력을 갖춘 대규모 언어 모델은 로봇 작업 계획에도 사용되었습니다[48]. 예를 들어, 병을 집는 것과 같은 작업을 제시하면 모델에 병이 있는 테이블로 가서 병을 집는 것과 같이 작업을 더 작은 작업 시퀀스로 분해하라는 메시지가 표시됩니다.

위에 요약된 연구를 바탕으로 우리는 대형 랜지 모델이 믿을 수 있는 에이전트를 만드는데 핵심 요소가 될 수 있다고 가정합니다. 기존 문헌은 주로 단발성 프롬프트[38, 66] 또는 연쇄적 사고 프롬프트[100]를 사용하는 일차 템플릿에 의존하고 있습니다. 이러한 템플릿은 에이전트의 주변 환경(예: 트롤이 주어진 게시물에 어떻게 반응할지, 로봇이 문이 있는 방에 들어가려면 어떤 행동을 취해야 할지 등)에만 조건이 부여된 행동을 생성하는데 효과적입니다. 그러나 신뢰할 수 있는 에이전트를 구현하려면 현재 환경뿐만 아니라 방대한 양의 과거 경험에 대한 컨디셔닝이 필요한데, 이는 1차 프롬프트를 사용하는 데 적합하지 않으며, 현재로서는 기본 모델의 제한된 컨텍스트 창으로 인해 불가능합니다. 최근 연구에서는 정적 지식 기반과 정보 검색 체계[53] 또는 간단한 요약 체계[104]로 언어 모델을 보강하여 1차 프롬프트를 넘어서려는 시도가 있었습니다. 이 백서에서는 이러한 아이디어를 확장하여 과거 경험이 각 시간 단계마다 동적으로 업데이트되고 에이전트의 현재 컨텍스트 및 계획과 혼합

되어 서로 강화되거나 모순될 수 있는 검색을 처리하는 에이전트 아키텍처를 구축합니다.

3 제너레이티브 에이전트 동작 및 상호작용

제너레이티브 에이전트의 어포던스를 설명하기 위해 심즈[7]를 연상시키는 단순한 샌드박스 월드에 에이전트를 캐릭터로 인스턴스화했습니다. 이 스프라이트 기반 샌드박스 게임 월드인 스몰빌은 작은 마을 환경을 연상시킵니다. 이 섹션에서는 스몰빌의 어포던스 및 제너레이티브 에이전트와의 상호작용을 살펴보고 그 안에서 에이전트가 어떻게 동작하는지 설명하겠습니다. 그런 다음 섹션 4에서는 이러한 어포던스 및 상호작용을 지원하는 제너레이티브 에이전트 아키텍처를 소개합니다. 섹션 5에서는



그림 2: 영역에 레이블이 지정된 스몰빌 샌드박스 월드. 루트 노드는 전체 월드를, 자식은 영역(예: 집, 카페, 상점)을, 리프 노드는 오브젝트(예: 테이블, 책장)를 설명합니다. 에이전트는 자신이 본 세계의 일부를 반영하는 하위 그래프를 기억하여 해당 부분의 상태를 관찰한 그대로 유지합니다.

샌드박스 환경의 구현과 에이전트가 샌드박스 월드의 기본 엔진과 상호 작용하는 방식에 대해 알아보세요.

남편 톰 모레노와 아내 제인 모레노(을 어느 정도 잘 알고 있습니다.

3.1 상담원 아바타 및 커뮤니케이션

세미콜론으로 구분된 각 구문은 시뮬레이션이 시작될 때 상담원의 초기 메모리에 메모리로 입력됩니다.

스몰빌에는 25명의 독특한 요원으로 구성된 커뮤니티가 살고 있습니다. 각 요원은 간단한 스프라이트 아바타로 표현됩니다. 각 에이전트의 직업, 다른 에이전트와의 관계 등 각 에이전트의 정체성을 시드 메모리로 표현하기 위해 자연어 설명 한 단락을 작성했습니다. 예를 들어 존 린은 다음과 같은 설명을 가지고 있습니다:

존 린은 윌로우 마켓 앤 약국의 약국 주인으로 사람들을 돋는 것을 좋아합니다. 그는 항상 고객이 약을 쉽게 구할 수 있는 방법을 찾고 있습니다. 존 린은 대학교수인 아내 메이린과 음악 이론을 공부하는 아들 에디 린과 함께 살고 있습니다. 존 린은 가족을 매우 사랑합니다. 존 린은 옆집 노부부인 샘 무어와 제니퍼 무어를 몇 년 동안 알고 지냈으며 샘 무어는 친절하고 좋은 사람이라고 생각합니다; 존 린은 이웃인 유리코 아마모토를 잘 알고 있습니다. 존 린은 이웃인 타마라 테일러와 카르멘 오토티즈를 알고 있지만 아직 만난 적이 없습니다. 존 린과 톰 모레노는 윌로우스 마켓과 약국의 동료입니다. 존 린과 톰 모레노는 친구이며 함께 지역 정치에 대해 토론하는 것을 좋아합니다. 존 린은 모레노 가족(

3.1.1 에이전트 간 커뮤니케이션. 에이전트는 행동을 통해 세상과 상호 작용하고 자연스러운 언어를 통해 서로 소통합니다. 샌드박스 엔진의 각 시간 단계에서 에이전트는 "이사벨라 로드리게즈가 일기를 쓰고 있습니다", "이사벨라 로드리게즈가 이메일을 확인하고 있습니다", "이사벨라 로드리게즈가 가족과 전화 통화를 하고 있습니다", "이사벨라 로드리게즈가 잠자리에 들 준비를 하고 있습니다" 등의 현재 행동을 설명하는 자연어 문장을 출력합니다. 그러면 이 문장이 샌드박스 세계에 영향을 미치는 구체적인 동작으로 변환됩니다. 동작은 샌드박스 인터페이스에 이모티콘 세트로 표시되어 오버헤드 뷰에서 동작을 추상적으로 표현할 수 있습니다. 이를 위해 시스템은 언어 모델을 활용하여 동작을 이모티콘 세트로 번역하고, 각 아바타의 머리 위에 말풍선 형태로 표시합니다. 예를 들어 "이사벨라 로드리게스가 일기를 쓰고 있습니다"는 로 표시되고, "이사벨라 로드리게스가 이메일을 확인하고 있습니다"는 로 표시됩니다. 상담원의 아바타를 클릭하면 작업에 대한 전체 자연어 설명에 액세스할 수 있습니다

에이전트는 완전한 자연어로 서로 소통합니다. 에이전트는 해당 지역의 다른 에이전트를 인지하고 있으며, 제너레이티브 에이전트 아키텍처에 따라 지나갈지 대화에 참여할지 결정합니다. 다음은 대화 중 샘플입니다.

다가오는 선거에 대해 이사벨라 로드리게즈와 톰 모레노 요원 간의 대화 내용입니다.³

이사벨라: 아직 고민 중이지만 샘 무어와 선거에 대해 논의하고 있습니다. 그에 대해 어떻게 생각하시나요?

Tom: 솔직히 저는 샘 무어를 좋아하지 않습니다. 그는 커뮤니티와 소통이 부족하고 우리의 이익을 최우선으로 생각하지 않는다고 생각합니다.

³ 이러한 상담원의 대화 스타일이 지나치게 형식적으로 느껴질 수 있으며, 이는 기본 모델에서 지침을 조정한 결과일 가능성이 높습니다. 향후 언어 모델에서는 글쓰기 스타일을 더 잘 제어할 수 있을 것으로 기대합니다.

3.1.2 사용자 제어. 사용자는 상담원이 자신을 인식해야 하는 페르 소나를 지정하여 자연어를 통해 상담원과 소통합니다. 예를 들어 사용자가 뉴스 '기자'라고 지정하고 "누가 출마하나요?"라고 말하여 다가오는 선거에 대해 질문하면 John 에이전트가 대답합니다:

John: 제 친구 유리코, 톰, 저는 다가오는 선거에 대해 이야기를 나누며 샘 무어 후보에 대해 논의했습니다. 저희는 모두 그의 공약이 마음에 들어서 그에게 투표하기로 동의했습니다.

상담원 중 한 명에게 직접 명령하려면 사용자가 상담원의 "내면의 목소리"의 인칭을 취하면 상담원이 해당 진술을 지시어로 간주할 가능성이 더 높아집니다. 예를 들어 사용자가 존의 내면의 목소리로 "다가오는 선거에서 샘과 대결할 거야"라고 말하면 존은 선거에 출마하기로 결심하고 아내와 아들에게 자신의 출마 사실을 알립니다.

3.2 환경 상호작용

스몰빌은 카페, 바, 공원, 학교, 기숙사, 주택, 상점을 포함하는 작은 마을의 공통 어포던스를 특징으로 합니다. 또한 집 안의 부엌, 부엌의 스토브와 같이 이러한 공간을 기능적으로 만드는 하위 영역과 오브젝트를 정의합니다(그림 2). 상담원의 주 거주 공간으로 사용되는 모든 공간에는 침대, 책상, 옷장, 선반은 물론 욕실과 주방이 있습니다.⁴

에이전트는 간단한 비디오 게임에서처럼 스몰빌을 돌아다니며 건물에 출입하고, 맵을 탐색하고, 다른 에이전트에게 접근합니다. 에이전트의 이동은 생성 에이전트 아키텍처와 샌드박스 게임 엔진에 의해 이루어지며, 모델이 에이전트가 특정 위치로 이동하도록 지시하면 스몰빌 환경에서 목적지까지의 도보 경로를 계산하고 에이전트가 이동하기 시작합니다. 또한 사용자는 스몰빌의 샌드박스 세계로 들어가 그 안에서 활동하는 에이전트가 될 수도 있습니다. 사용자가 구현하는 에이전트는 이사벨라나 존과 같이 이미 존재하는 에이전트일 수도 있고, 스몰빌에 방문한 적이 없는 외부 방문자일 수도 있습니다. 스몰빌의 주민들은 사용자가 제어하는 에이전트를 서로를 대할 때와 다르지 않게 대합니다. 그들은 에이전트의 존재를 인식하고, 상호작용을 시작하고, 에이전트의 행동을 기억한 후 그에 대한 의견을 형성합니다.

사용자와 에이전트는 심즈와 같은 샌드박스 게임에서처럼 이 세계의 오브젝트 상태에 영향을 줄 수 있습니다. 예를 들어, 에이전트가 잠을 자고 있을 때 침대가 사용 중일 수 있고, 에이전트가 아침 식사를 만들기 위해 식재료를 다 사용했을 때 냉장고가 비어 있을

수 있습니다. 또한 최종 사용자는 에이전트 주변 사물의 상태를 자연어로 다시 작성하여 스몰빌에서 에이전트의 환경을 재구성할 수 있습니다. 예를 들어, 이사벨라가 아침에 아침 식사를 준비하고 있을 때 사용자는 다음과 같이 사물을 선택하고 새로운 상태를 설명하는 명령을 시스템에 입력하여 주방 스토브의 상태를 "켜짐"에서 "굽기"로 변경할 수 있습니다: "<이사벨라의 아파트: 부엌: 스토브>가 타는 중입니다." 이사벨라는 다음 순간에 이를 알아차리고 스토브를 끄고 아침 식사를 다시 만들 것입니다. 마찬가지로 사용자가 화장실에 들어갔을 때 이사벨라의 샤워기 상태를 "누수"로 설정하면 거실에서 도구를 가져와 누수를 고치려고 시도합니다.

⁴ 이 환경 디자인은 저희 작업의 초점이 아니므로 자동이 아닌 수동으로 환경을 생성했습니다. 향후 작업을 통해 상 담 원 환경의 풍부함을 계속 확장할 수 있습니다.

3.3 "일상의 하루" 예시

한 단락의 설명을 시작으로 제너레이티브 에이전트는 하루를 계획하기 시작합니다. 샌드박스 세계에서 시간이 지남에 따라 에이전트가 서로 및 세상과 상호 작용하고, 기억과 관계를 구축하고, 공동 활동을 조정하면서 행동이 진화합니다.

에이전트 John Lin에 대한 하루 동안의 시스템 출력을 추적하여 제너레이티브 에이전트의 동작을 시연합니다(그림 3). 린 가족에서 존은 오전 7시에 가장 먼저 일어나 양치질을 하고 샤워를 한 후 옷을 입고 아침 식사를 한 후 거실의 식탁에서 뉴스를 확인합니다. 오전 8시가 되면 에디도 서둘러 침대에서 일어나 수업 준비를 합니다. 그는 외출하려는 존과 마주칩니다:

존: 좋은 아침 에디. 잘 잤니? 에디: 좋은 아침 아빠.

네, 잘 잤어요. 존: 잘됐네. 오늘은 무슨 일 하고 있니? 에디: 수업에 쓸 새 곡을 만들고 있어요. 이번 주에 마감이라서 빨리 끝내려고 노력 중이에요. 하지만 정말 재미있어요!

John: 그거 좋네요!

에디가 외출하고 얼마 지나지 않아 메이가 일어나 존과 합류합니다. 메이는 아들에 대해 묻고, 존은 방금 나눈 대화를 떠올립니다:

메이 에디는 벌써 학교로 떠났나요?

John: 네, 방금 떠났어요. 수업을 위한 작곡 작업을 하고 있어요

메이 오, 멋지네요! 말씀해 주세요.

John: 정말 즐기고 있는 것 같아요! 정말 재미있다
고 하더라고요.

메이: 잘됐네요! 정말 자랑스러워요.

대화를 마친 메이와 존은 짐을 챙깁니다. 메이는 강의와 연구 논문 작업을 계속하고, 존은 오전 9시까지 월로우 마켓 앤 약국에서 약국 카운터를 열어야 합니다.

3.4 긴급한 사회적 행동

스몰빌의 제너레이티브 에이전트들은 서로 상호작용함으로써 정보를 교환하고 새로운 관계를 형성하며 공동 활동을 조율합니다. 이전 연구[80]를 확장한 이러한 사회적 행동은 사전 프로그래밍된 것이 아니라 즉흥적으로 발생합니다.

3.4.1 정보 확산: 상담원들이 서로를 알아차리면서 대화에 참여 할 수 있으며, 이 과정에서 상담원 간에 정보가 확산될 수 있습니다.

Sam: 안녕하세요, 톰, 어떻게 지내세요? 톰: 좋아요, 고마워요. 무슨 일이에요?

샘: 할 얘기가 있어요. 저는 사실 다가오는 지방 선거에서 시장으로 출마할 예정입니다.

Tom: 정말요? 좋은 소식이네요! 왜 뛰는 거예요?

Sam: 저는 수년 동안 지역 정치에 참여해 왔으며, 지역 사회를 변화시킬 수 있다고 생각합니다. 또한 새로운 아이디어와 관점을 제시하고 싶어요.



그림 3: 제너레이티브 에이전트인 John Lin의 일상적인 아침. John은 오전 6시경에 일어나 양치질, 샤워, 아침 식사 등의 아침 루틴을 완료합니다. 그리고 아내 메이, 아들 에디와 잠시 이야기를 나눈 후 하루 일과를 시작하기 위해 밖으로 나갑니다.



해요. 여러분은 어떻게 생각하시나요?

John: 그가 출마하는 것은 정말 대단한 일이라고 생각합니다. 선거 결과가 어떻게 나올지 궁금하네요.

그림 4: 시뮬레이션을 시작할 때 한 에이전트가 발렌타인데이 파티를 주최하려는 의도로 초기화됩니다. 이후 이어지는 일련의 이벤트에서 에이전트가 의도에 따라 행동하지 않거나, 다른 사람에게 알리는 것을 잊어버리거나, 참석 사실을 기억하지 못 할 수 있는 등 여러 가지 실패 가능성이 있지만 실제로는 여러 에이전트가 모여 상호 작용하면서 발렌타인데이 파티가 열리게 됩니다.

그날 늦은 밤, 샘이 떠난 후 다른 소식통을 통해 소식을 들은 톰과 존은 샘의 선거 당선 가능성에 대해 논의합니다:

John: 샘 무어가 지방 선장에 시장으로 출마한다고 들었어요. 그가 당선될 가능성이 높다고 생각하시나요?

톰: 좋은 기회가 있다고 생각해요. 그는 커뮤니티에 서 열심히 일해왔고 많은 지지를 받을 거라고 생각

점차 샘의 출마가 화제가 되면서 일부는 샘을 지지하고, 일부는 아직 결정하지 않은 채로 남아 있습니다.

3.4.2 **관계 기억** 스몰빌의 에이전트는 시간이 지남에 따라 새로운 관계를 형성하고 다른 에이전트와의 상호작용을 기억합니다. 예를 들어, 처음에 샘은 라토야 월리엄스를 알지 못합니다. 존은 공원에서 산책을 하던 중 샘은 라토야를 만나 서로를 소개합니다. 라토야는 자신이 사진 프로젝트를 진행하고 있다고 말합니다: "제가 진행 중인 프로젝트를 위해 사진을 찍으러 왔어요." 이후 대화에서 샘이 라토야에게 "안녕하세요, 라토야?"라고 묻고, 라토야가 "프로젝트는 잘 진행되고 있나요?"라고 답하는 장면은 그 대화에 대한 기억을 나타냅니다. 프로젝트는 어떻게 진행되고 있나요?"라고 묻자 라토야가 "안녕하세요, 샘. 잘 진행되고 있어요!"라고 대답하는 장면이 나온다. 잘 진행되고 있어요!"라고 대답합니다.

3.4.3 **조정** 제너레이티브 에이전트는 서로 협력합니다. 흡스 카페의 이사벨라 로드리게즈는 2월 14일 오후 5시부터 7시까지 발렌타인데이 파티를 계획하기 위해 초기화됩니다. 이 씨앗을 바탕으로 상담원은 흡스 카페나 다른 곳에서 친구나 고객을 만나면 초대합니다. 그런 다음 이사벨라는 13일 오후에 행사를 위해 카페를 꾸미는데 시간을 보냅니다. 단골 손님이자 이사벨라의 절친한 친구인 마리아가 카페에 도착합니다. 이사벨라는 파티 장식을 위해 마리아의 도움을 요청하고 마리아는 흔쾌히 동의합니다. 마리아의 캐릭터 설명에 클라우스를 좋아한다고 언급되어 있습니다. 그날 밤, 마리아는 비밀리에 짹사랑하던 클라우스를 파티에 초대하고 클라우스는 기꺼이 수락합니다.

발렌타인 데이에 클라우스와 마리아를 포함한 5명의 에이전트가 오후 5시에 흡스 카페에 나타나서 축제를 즐깁니다(그림 4). 이 시나리오에서 최종 사용자는 파티를 열겠다는 이사벨라의 초기 의도와 클라우스를 좋아하는 마리아의 마음만 설정했을 뿐, 파티에 대한 소문을 퍼뜨리고, 파티를 꾸미고, 서로 데이트 신청을 하고, 파티에 도착하고, 파티에서 서로 상호작용하는 소셜 행동은 에이전트 아키텍처에 의해 시작되었습니다.

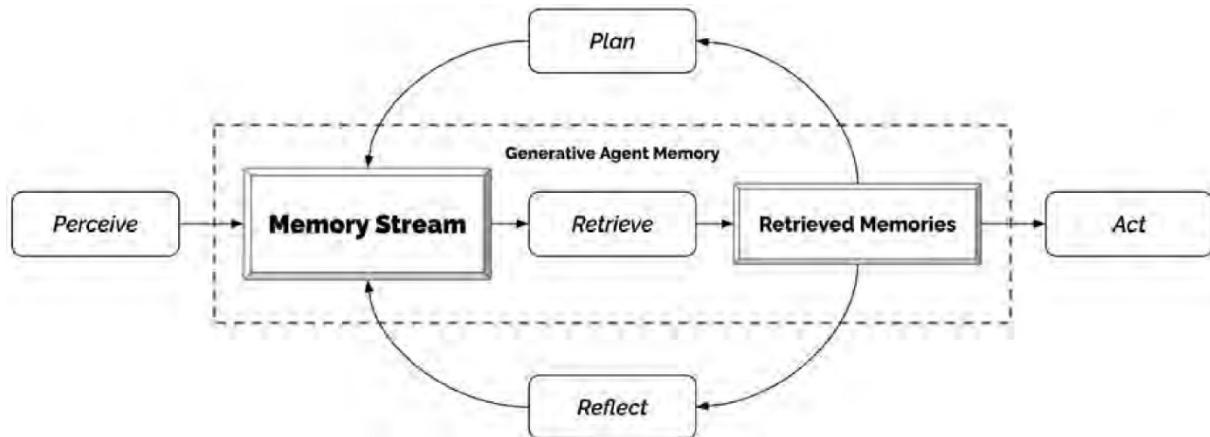


그림 5: 제너레이티브 에이전트 아키텍처. 에이전트는 환경을 인식하며, 모든 인식은 메모리 스트림이라는 에이전트의 경험에 대한 포괄적인 기록에 저장됩니다. 에이전트의 인식을 기반으로 아키텍처는 관련 메모리를 검색하고 검색된 메모리를 사용하여 동작을 결정합니다. 이렇게 검색된 기억은 장기적인 계획을 수립하고 더 높은 수준의 반영을 만드는 데도 사용되며, 이 두 가지 모두 나중에 사용할 수 있도록 메모리 스트림에 입력됩니다.

4 제너레이티브 에이전트 아키텍처

제너레이티브 에이전트는 오픈 월드에서 다른 에이전트와 상호 작용하고 환경 변화에 반응할 수 있는 행동의 프레임워크를 제공하는 것을 목표로 합니다. 제너레이티브 에이전트는 현재 환경과 과거 경험을 입력으로 삼고 행동을 출력으로 생성합니다. 이러한 행동의 근간에는 대규모 언어 모델과 관련 정보를 합성하고 검색하여 언어 모델의 출력을 조절하는 메커니즘이 결합한 새로운 에이전트 아키텍처가 있습니다. 이러한 메커니즘이 없으면 대규모 언어 모델이 동작을 출력할 수는 있지만, 그 결과 에이전트가 에이전트의 과거 경험에 따라 반응하지 않거나 중요한 추론을 하지 못할 수 있으며 장기적인 일관성을 유지하지 못할 수 있습니다. 장기적인 계획과 일관성에 대한 과제는 GPT-4와 같은 오늘날 가장 성능이 뛰어난 모델에서도 여전히 남아 있습니다[19]. 제너레이티브 에이전트는 대량의 이벤트와 기억을 생성하고 이를 유지해야 하기 때문에 아키텍처의 핵심 과제는 필요할 때 에이전트의 기억 중 가장 관련성이 높은 부분을 검색하고 합성하는 것입니다.

저희 아키텍처의 중심에는 상담원의 경험에 대한 포괄적인 기록을 유지하는 데이터 베이스인 메모리 스트림이 있습니다. 메모리 스트림에서 에이전트의 작업을 계획하고 환경에 적절하게 반응하기 위해 관련성이 있는 레코드가 검색됩니다. 기록은 행동을 안내하는 상위 및 상위 수준의 반영으로 재구적으로 합성됩니다. 아키텍처의 모든 것이 자연어 설명으로 기록되고 추론되

어 대규모 언어 모델을 활용할 수 있도록 아키텍처를 낚춥니다.

현재 구현은 ChatGPT의 gpt3.5 터보 버전을 활용하고 있습니다 [77]. 언어 모델이 개선됨에 따라 제네릭 에이전트의 아키텍처 기본 사항인 메모리, 계획 및 반영은 동일하게 유지될 것으로 예상합니다. 새로운 언어 모델(예: GPT-4)은 제너레이티브 에이전트를 뒷받침하는 프롬프트의 표현력과 성능을 계속 확장해 나갈 것입니다. 하지만 이 글을 쓰는 현재 GPT-4의 API는 초대 전용이므로 상담원들은 ChatGPT를 사용하고 있습니다.

4.1 메모리 및 검색

도전 과제: 인간의 행동을 시뮬레이션할 수 있는 제너레이티브 에이전트를 만들려면 프롬프트에 설명해야 하는 것보다 훨씬 더 큰 경험 집합에 대한 추론이 필요하며, 전체 메모리 스트림은 모델의 주의를 분산시킬 수 있고 현재 제한된 컨텍스트 창에 맞지 않을 수도 있습니다. "요즘 무엇에 열정을 가지고 있나요?"라는 질문을 받는 Isabella 상담원을 생각해 보겠습니다. 언어 모델의 제한된 컨텍스트 창에 맞추기 위해 Isabella의 모든 경험을 요약하면 이벤트 및 프로젝트를 위한 협업, 카페의 청결 및 정리와 같은 주제에 대해 논의하는 비정보적인 응답이 생성됩니다. 하지만 아래에 설명된 메모리 스트림은 요약하는 대신 관련 기억을 표시하여 사람들이 환영받고 소속감을 느끼게 하고, 이벤트를 계획하고, 발렌타인데이 파티와 같이 사람들이 즐길 수 있는 분위기를 조성하려는 이사벨라의 열정을 언급하는 보다 유익하고 구체적인 응답을 생성합니다.

접근 방식: 메모리 스트림은 상담원의 경험에 대한 포괄적인 기록을 유지합니다. 메모리 스트림은 메모리 객체의 목록으로, 각 객체에는 자연어 설명, 생성 시간 탬프, 가장 최근 액세스 타임 스템프가 포함되어 있습니다. 메모리 스트림의 가장 기본적인 요소는 상담원이 직접 인지한 이벤트인 관찰입니다. 일반적인 관찰에는 에이전트가 직접 수행한 행동이나 에이전트가 다른 에이전트나 에이전트가 아닌 객체가 수행한다고 인식하는 행동이 포함됩니다. 예를 들어 커피숍에서 일하는 이사벨라 로드리게스는 시간이 지남에 따라 (1) 이사벨라 로드리게스가 페이스 트리를 진열하고 있음, (2) 마리아 로페즈가 커피를 마시면서 화학 시험 공부를 하고 있음, (3) 이사벨라 로드리게스와 마리아 로페즈가 흙스 카페에서 발렌타인데이 파티 계획에 대해 대화하고 있음, (4) 냉장고가 비어 있음 등의 관찰을 누적할 수 있습니다.

당사의 아키텍처는 에이전트의 현재 상황을 입력으로 받아 언어 모델에 전달할 메모리 스트림의 하위 집합을 반환하는 검색 함수를 구현합니다. 에이전트가 어떻게 행동할지 결정할 때 고려해야 할 중요한 사항에 따라 검색 함수를 구현할 수 있는 방법은 여러 가지가 있습니다.

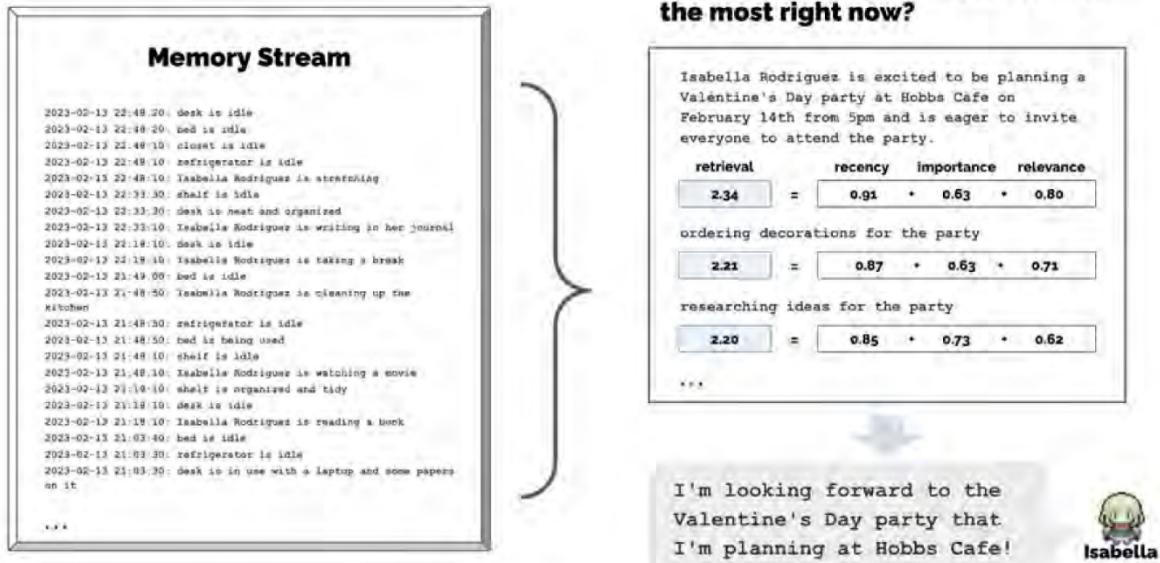


그림 6: 메모리 스트림은 상담원의 현재 상황과 관련이 있거나 관련이 없는 수많은 관찰로 구성됩니다. 검색은 상황에 대한 응답을 조절하기 위해 언어 모델에 전달해야 하는 이러한 관찰의 하위 집합을 식별합니다.

저희는 효과적인 결과를 도출하는 세 가지 주요 구성 요소에 초점을 맞추고 있습니다.

최근성은 최근에 다시 액세스한 메모리 오브젝트에 더 높은 점수를 부여하여 조금 전 또는 오늘 아침의 이벤트가 에이전트의 주의 영역에 남아있을 가능성이 높도록 합니다. 구현에서는 최근 도를 메모리에 액세스한 이후 샌드박스 게임 시간 수에 따라 기하급수적으로 감소하는 함수로 처리합니다.

마지막으로 검색되었습니다. 감쇠 계수는 0.995입니다.
중요도는 다음과 같이 일상적인 기억과 핵심 기억을 구분합니다.
에이전트가 중요하다고 생각하는 메모리 객체에 더 높은 점수를 부여합니다. 예를 들어, 방에서 단식을 하는 것과 같은 일상적인 사건은 중요도 점수가 낮고, 연인과의 이별은 높은 점수를 받을 수 있습니다. 중요도 점수를 구현할 수 있는 방법은 여러 가지가 있지만, 언어 모델에 직접 정수로 점수를 출력하도록 요청하는 것이 효과적입니다. 전체 프롬프트는 아래와 같습니다:

1에서 10까지의 척도(1은 지극히 일상적인 것(예: 양치질, 잠자리 만들기)이고 10은 매우 가슴 아픈 것(예: 이별, 대학 합격)으로 표시)에서 다음 기억 중 가슴 아픈 것의 가능성성을 평가하세요.
기억: 월로우스 마켓과 약국에서 식료품 구입하기
등급: <기입하기>

이 프롬프트는 "방 청소"의 경우 2, "좋아하는 사람에게 데이트 신

청"의 경우 8의 정수 값을 반환합니다. 중요도 점수는 메모리 객체가 생성될 때 생성됩니다.

관련성은 현재 상황과 관련된 메모리 객체에 더 높은 점수를 부여합니다. 관련성은 "무엇과 관련성이 있는가?"에 대한 대답에 따라 달라지므로 관련성을 조건으로 삼습니다.

쿼리/메모리. 예를 들어, 학생이 반 친구와 화학 시험을 위해 무엇을 공부해야 하는지 욕하는 것이 쿼리인 경우, 아침 식사에 대한 메모리 객체는 관련성이 낮아야 하고, 선생님과 학업에 대한 메모리 객체는 관련성이 높아야 합니다. 구현에서는 언어 모델을 사용하여 각 기억에 대한 텍스트 설명의 임베딩 벡터를 생성합니다. 그런 다음 메모리의 임베딩 벡터와 쿼리 메모리의 임베딩 벡터 사이의 코사인 유사도로 관련성을 계산합니다.

최종 검색 점수를 계산하기 위해 최소-최대 스케일링을 사용하여 최근성, 관련성, 중요도 점수를 $[0, 1]$ 범위로 정규화합니다. 검색 함수는 세 가지 요소의 가중치 조합으로 모든 기억에 점수를 매깁니다: $score = \alpha_{recency} \cdot recency + \alpha_{importance} \cdot importance + \alpha_{relevance} \cdot relevance$. 우리의 구현에서 모든 α 가 1로 설정됩니다. 최상위 메모리는 언어 모델의 컨텍스트 창이 프롬프트에 포함됩니다.

4.2 반사

도전 과제: 제너레이티브 에이전트는 원시 작업 메모리만 가지고 있을 경우 일반화하거나 추론하는 데 어려움을 겪습니다. 사용자가 클라우스 물러에게 다음과 같은 질문을 하는 시나리오를 생각해 보겠습니다: "아는 사람 중 한 명과 한 시간을 함께 보내야 한다면 누구를 선택하겠습니까?"라는 질문입니다. 관찰 메모리만 사용할 수 있는 에이전트는 클라우스와 가장 자주 상호작용한 사람을 선택하기만 하면 됩니다. 대학 기숙사 이웃인 볼프강입니다. 안타깝게도 볼프강과 클라우스는 지난가다 마주칠 뿐 깊은 상호작용을 한 적이 없습니다. 보다 바람직한 응답을 위해서는 에이전트가 클라우스가 연구 프로젝트에 시간을 할애했던 기억을 일반화하여 클라우스가 연구에 열정적이라는 더 높은 수준의 반영을 생성해야 하며, 마찬가지로 다음과 같이 해야 합니다.

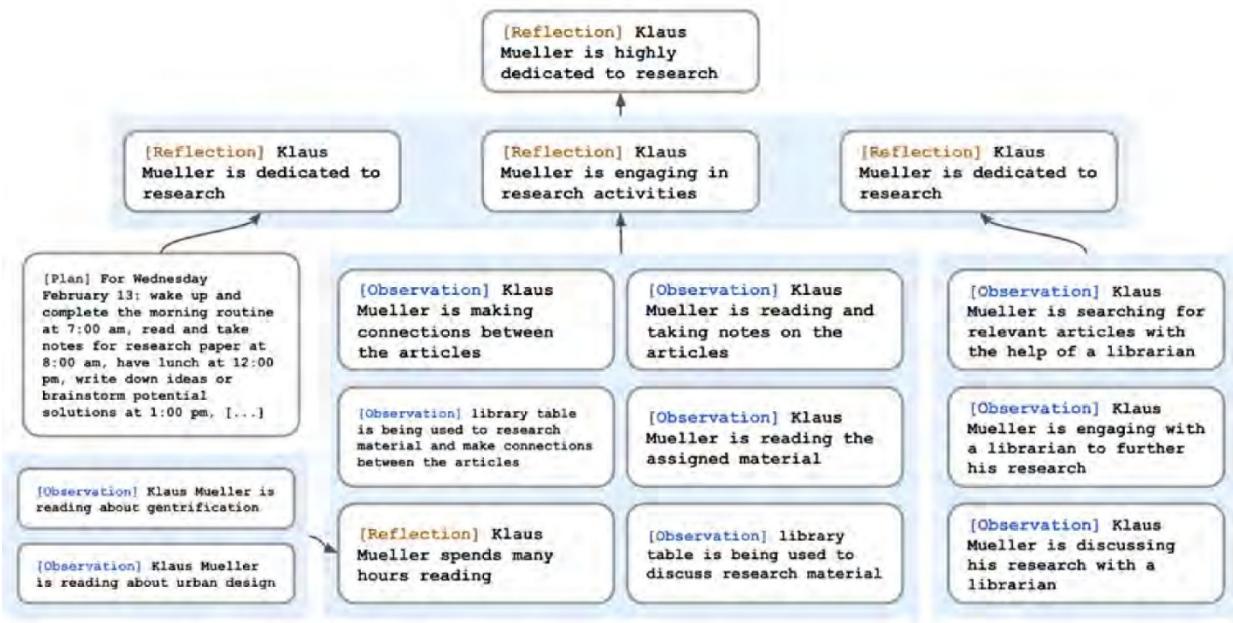


그림 7: 클라우스 물러에 대한 리플렉션 트리. 리프 노드로 표현된 에이전트의 세계 관찰을 재귀적으로 합성하여 클라우스 물러가 자신의 연구에 매우 헌신적이라는 자기 생각을 도출합니다.

는 마리아가 비록 다른 분야이기는 하지만 자신의 연구를 위해 노력하는 것을 인정하고, 두 사람이 공통의 관심사를 공유하고 있음을 반영할 수 있습니다. 아래 접근 방식을 사용하면 클라우스에게 누구와 함께 시간을 보낼지 묻는 질문에 클라우스는 볼 프강 대신 마리아를 선택합니다.

접근: 두 번째 유형의 기억을 소개하는데, 이를 **반성**이라고 부릅니다. 반성은 상담원이 생성한 더 높은 수준의 추상적인 생각입니다.
반사는 기억의 한 유형이므로 검색이 발생할 때 다른 관찰과 함께 포함됩니다. 반영은 주기적으로 생성되며, 저희 구현에서는 에이전트가 인지한 최신 이벤트의 중요도 점수의 합이 임계값(저희 구현에서는 150)을 초과할 때 반영을 생성합니다. 실제로 상담원들은 하루에 대략 두세 번씩 반영을 수행했습니다.

반영의 첫 번째 단계는 상담원의 최근 경험에 따라 질문할 수 있는 질문을 식별하여 무엇을 반영할지 결정하는 것입니다. 상담원의 메모리 스트림에 있는 가장 최근 기록 100개(예: "클라우스 물러가 젠트리피케이션에 관한 책을 읽고 있습니다", "클라우스 물러가 연구 프로젝트에 대해 사서와 대화하고 있습니다", "도서관의 책상이 현재 비어 있습니다")를 대규모 언어 모델에 쿼리하고 "위의 정보만 주어졌을 때 진술의 주제에 대해 대답할 수 있는 가장 중요한 고급 질문 3개는 무엇일까요?"라는 메시지를 언어 모델에 표시합니다. 모델의 응답은 후보 질문을 생성합니다(예: 클라우

스 물러는 어떤 주제에 열정을 가지고 있나요? 및 클라우스 물러와 마리아 로페즈 사이의 관계는 무엇인가요?). 이렇게 생성된 질문을 검색 쿼리로 사용하고 각 질문에 대한 관련 기억(기타 반영 포함)을 수집합니다. 그런 다음 언어 모델에 인사이트를 추출하고 인사이트의 증거가 된 특정 기록을 인용하도록 프롬프트를 표시합니다. 전체 프롬프트는 다음과 같습니다:

클라우스 물러에 대한 진술

1. 클라우스 물러가 연구 논문을 작성 중입니다.
2. 젠트리피케이션에 관한 책을 읽는 것을

즐기는 클라우스 물러

3. 클라우스 물러가 운동에 대해 아이샤 칸과 대화하

고 있습니다 [...]

위의 진술에서 유추할 수 있는 5가지 높은 수준의

인사이트는 무엇인가요? (예시 형식: 인사이트(1, 5,

3 때문에))

이 프로세스는 클라우스 물러가 (1, 2, 8, 15 때문에) 젠트리피케이션

에 대한 연구에 전념하고 있다는 것과 같은 진술을 생성합니다. 우리는 인용된 메모리 객체에 대한 포인터를 포함하여 메모리 스트림에 반영된 문장을 파싱하고 저장합니다.

예를 들어 위의 클라우스 물러에 대한 두 번째 진술은 클라우스가 이전에 가졌던 성찰이며, 환경으로부터의 관찰이 아닙니다. 결과적으로 에이전트는 반성의 트리를 생성합니다. 트리의 잎 노드는 기본 관찰을 나타내고 잎이 아닌 노드는 트리 위로 올라갈수록 더 추상적이고 고차원적인 생각을 나타냅니다.

4.3 계획 및 대응

과제: 대규모 언어 모델은 상황 정보(예: [46, 80])에 대한 응답으로 그럴듯한 행동을 생성할 수 있지만, 에이전트는 행동 순서가 일관되고 믿을 수 있도록 더 긴 시간 범위에서 계획을 세워야 합니다. 클라우스의 배경을 가진 언어 모델에 시간을 설명하고 주어진 순간에 어떤 행동을 취해야 하는지 묻는다면 클라우스는 오후 12시에 점심을 먹지만, 이미 오후 12시 30분과 오후 1시에 다시 점심을 먹을 것입니다.

전트

점심을 두 번이나 먹었습니다. 당장의 신뢰성을 위해 최적화하면 시간이 지남에 따라 신뢰성이 떨어집니다. 이 문제를 극복하려면 계획이 필수적입니다. 아래에 설명된 접근 방식을 사용하면 클라우스의 오후 계획은 오후 12시에 책을 읽으면서 홈스 카페에서 점심을 먹고, 오후 1시에 학교 도서관에서 연구 논문을 작성하고, 오후 3시에 공원에서 산책을 하며 휴식을 취하는 등 덜 과식적인 계획을 세울 수 있습니다.

접근 방식: 계획은 상담원의 향후 작업 순서를 설명하며 시간이 지나도 상담원의 행동을 일관되게 유지하는 데 도움이 됩니다.

계획에는 위치, 시작 시간 및 기간이 포함됩니다. 예를 들어 마감일이 임박한 연구에 몰두하고 있는 클라우스 뮤리⁵ 하루 종일 책상에서 연구 논문 초안을 작성하는 데 시간을 할애할 수 있습니다. 예를 들어 2023년 2월 12일 오전 9시부터 180분 동안 오크힐 대학 기숙사에서 근무한다고 계획 항목에 명시할 수 있습니다: 클라우스 뮤리의 방: 책상, 연구 논문을 읽고 메모함, 반사와 마찬가지로 계획도 메모리 스트림에 저장되며 검색 프로세스에 포함됩니다. 따라서 상담원은 행동 방식을 결정할 때 관찰, 성찰, 계획을 모두 함께 고려할 수 있습니다. 상담원은 필요한 경우 도중에 계획을 변경할 수 있습니다.

아티스트 에이전트가 4시간 동안 움직이지 않고 약국 카운터에 앉아 그림을 그리는 것은 비현실적이고 흥미롭지 않을 것입니다. 보다 바람직한 계획은 에이전트가 4시간 동안 자신의 홈 스튜디오에서 재료를 모으고, 물감을 섞고, 휴식을 취하고, 청소를 하는 데 필요한 시간을 갖는 것입니다. 이러한 계획을 세우기 위해 하향식으로 시작한 다음 재귀적으로 세부 사항을 생성하는 접근 방식을 사용합니다. 첫 번째 단계는 그날의 일정을 대략적으로 설명하는 계획을 작성하는 것입니다. 초기 계획을 만들기 위해 언어 모델에 상담원의 요약 설명(예: 이름, 특성, 최근 경험 요약)과 전날의 요약을 묻는 메시지를 표시합니다. 아래는 언어 모델이 완료할 수 있도록 하단에 미완성된 전체 프롬프트 예시입니다:

이름: 에디 린 (나이: 19)

타고난 특성: 친절하고 외향적이며 친절한 에디 린은 오크힐 칼리지에서 음악 이론과 작곡을 공부하는 학생입니다. 그는 다양한 음악 스타일을 탐구하는 것을 좋아하며 항상 지식을 넓힐 방법을 찾고 있습니다. 에디 린은 대학 수업을 위한 작곡 프로젝트를 진행하고 있습니다. 그는 음악 이론에 대해 더 많이 배우기 위해 수업을 듣고 있습니다. 에디 린은 현재 작업 중인 새로운 작곡에 대한 기대가 크지만 앞으로는 하루 종일 더 많은 시간을 할애하여 작업하고 싶다고 합니다.

2월 12일 화요일, 에디 1)는 오전 7시에 일어나 아침 일과를 마

UIST '23, 2023년 10월 29일~11월 1일, 미국 샌프란시스코, 캘리포니아
치고 [...]를 마쳤습니다.

6) 밤 10시경에 잠자리에 들 준비를 했습니다.

오늘은 2월 13일 수요일입니다. 오늘 Eddy의 계획을 크게 요약 하면 다음과 같습니다.)

이렇게 하면 "1) 일어나서 오전 8시에 아침 일과를 마치고, 2) 오전 10시부터 수업을 듣기 위해 오크힐 대학에 가고, [...] 5) 오후 1시부터 오후 5시까지 새로운 음악 작곡 작업을 하고, 6) 오후 5시 30분에 저녁을 먹고, 7) 학교 과제를 마치고 오후 11시까지 잠자리에 든다"와 같이 5~8개 단위로 나뉜 상담원의 하루 계획이 대략적으로 그려집니다.

⁵ 그리고 이런 식으로 이 백서의 저자와 적어도 지나칠 정도로 닮았습니다.

에이전트는 이 계획을 메모리 스트림에 저장한 다음 이를 재귀적으로 분해하여 먼저 1시간 단위의 작업으로 세분화합니다. 오후 1시부터 오후 5시까지 새로운 음악 작곡 작업을 하려는 에디의 계획은 다음과 같습니다. 오후 1시: 음악 작곡에 대한 아이디어를 브레인스토밍하는 것으로 시작 [...] 오후 4시: 작곡을 검토하고 다듬기 전에 잠시 휴식을 취하고 창의적인 에너지를 재충전합니다. 그런 다음 다시 5~15분 단위로 반복해서 작곡을 재구성합니다. 예: 오후 4시: 과일 한 조각, 그래놀라 바, 견과류 등 가벼운 간식을 먹습니다. 오후 4시 5분: 업무 공간 주변을 잠깐 산책하기 [...] 오후 4시 50분: 몇 분간 업무 공간을 정리하기. 이 프로세스는 원하는 세분화에 맞게 조정할 수 있습니다.

4.3.1 계획 반응 및 업데이트. 제너레이티브 에이전트는 각 시간 단계에서 주변 세계를 인식하고 인식된 관찰이 메모리 스트림에 저장되는 액션 루프에서 작동합니다. 이러한 관찰을 통해 언어 모델에 에이전트가 기존 계획을 계속 수행할지 아니면 반응할지 결정하도록 유도합니다. 예를 들어 이젤 앞에 서서 그림을 그리는 행위는 이젤에 대한 관찰을 유발할 수 있지만 반응을 유발할 가능성은 낮습니다. 하지만 에디의 아버지 John이 에디가 집 정원에서 잠시 산책을 하는 것을 본 것을 기록한다면 결과는 달라집니다. 프롬프트는 아래와 같으며, [상담원의 요약 설명]은 상담원의 전반적인 목표와 성향에 대한 동적으로 생성된 한 문단 길이의 요약으로, 부록 A에 설명되어 있습니다:

[상담원 요약 설명]

2023년 2월 13일 오후 4시 56분입니다.

존 린의 상태: John은 퇴근 후 일찍 집에 돌아왔습니다.

관찰: 존은 에디가 회사 주변을 잠시 산책하는 것을 보았습니다.

존의 기억에서 관련 맥락을 요약합니다: 에디 린은 존 린의 아들입니다. 에디 린은 자신의 수업을 위한 음악 작곡을 하고 있습니다. 에디 린은 음악에 대해 생각하거나 음악을 들을 때 정원을 산책하는 것을 좋아합니다

존이 이 관찰에 반응해야 하며, 그렇다면 적절한 반응은 무엇일까요?

컨텍스트 요약은 "[관찰자]와 [관찰 대상]의 관계는 무엇인가요?" 및 "[관찰 대상]은 [관찰 대상의 행동 상태]입니다."라는 질문을 통해 기억을 재시험하는 두 가지 프롬프트와 그에 대한 답변

J.S. 박, J.C. 오브라이언, C.J. 카이, M.R. 모리스, P. 리안, M.S. 벤스타인을 함께 요약하여 생성됩니다. 이 출력은 John이 Eddy에게 음악 작곡 프로젝트에 대해 물어볼 것을 고려할 수 있음을 시사합니다. 그런 다음 반응이 발생한 시점부터 에이전트의 기존 계획을 다시 생성합니다. 마지막으로, 반응이 에이전트 간의 상호작용을 나타내는 경우 에이전트 간의 대화를 생성합니다.

4.3.2 대화. 상담원은 서로 상호작용하면서 대화를 나눕니다. 상담원의 대화는 서로에 대한 기억에 따라 발화를 조건화하여 생성합니다. 예를 들어 존이 에디와 대화를 시작할 때 에디에 대한 요약된 기억과 작곡 프로젝트에 대해 에디에게 물어보기로 결정했을 때의 의도된 반응을 사용하여 존의 첫 번째 발화를 생성합니다:

[상담원 요약 설명]

2023년 2월 13일 오후 4시 56분입니다.

존 린의 상태: John은 퇴근 후 일찍 집에 돌아왔습니다.

관찰: 존은 에디가 회사 주변을 잠시 산책하는 것을 보았습니다.

존의 기억에서 관련 맥락을 요약합니다: 에디 린은 존 린의 아들입니다. 에디 린은 자신의 수업을 위한 음악 작곡을 하고 있습니다. 에디 린은 음악에 대해 생각하거나 음악을 들을 때 정원을 산책하는 것을 좋아합니다.

존이 에디에게 음악 작곡 프로젝트에 대해 물어봅니다.

존은 에디에게 뭐라고 말할까요?

그 결과 "에디, 수업용 음악 작곡 프로젝트는 어떻게 진행되고 있나요?"

에디의 입장에서는 존이 대화를 시작하는 것이 자신이 반응하고 싶은 사건으로 여겨집니다. 따라서 에디는 존이 그랬던 것처럼 존과의 관계에 대한 자신의 기억과 대화에서 존의 마지막 발언과 관련이 있을 수 있는 자신의 기억을 검색하여 요약합니다. 에디가 응답하기로 결정하면 요약된 기억과 현재 대화 기록을 사용하여 에디의 발화를 생성합니다:

[상담원 요약 설명]

2023년 2월 13일 오후 4시 56분입니다.

에디 린의 상태입니다: 에디는 회사 주변을 잠시 산책하고 있습니다.

관찰: 존이 에디와 대화를 시작하고 있습니다.

에디의 기억에서 나온 관련 컨텍스트 요약입니다: 존 린은 에디 린의 아버지입니다. 존 린은 배려심이 많고 에디 린의 학교 과제에 대해 자세히 알고 싶어합니다. 존 린은 에디 린이 음악 작곡을 하고 있다는 것을 알고 있습니다.

대화 내역은 다음과 같습니다:

John: 에디, 수업용 음악 작곡 프로젝트는 어떻게 진행되고 있나요?

에디는 존에게 어떻게 반응할까요?

그러면 에디의 응답이 생성됩니다: "아빠, 잘 되고 있어요. 머리를 비우고 영감을 얻기 위해 정원을 산책하고 있어요."라고 대답합니다. 이 대화는 두 상담원 중 한 명이 대화를 종료하기로 결정할 때까지 동일한 메커니즘을 사용하여 계속 생성됩니다.

5 샌드박스 환경 구현

스몰빌 샌드박스 게임 환경은 Phaser 웹 게임 개발 프레임워크[57]를 사용하여 빌드되었습니다. 에이전트 아바타를 포함한 시각적

환경 스프라이트와 우리가 제작한 환경 맵 및 콜리전 맵을 Phaser로 임포트했습니다.

유니티는 샌드박스 개발 프레임워크에 서버를 추가하여 샌드박스 정보를 제너레이티브 에이전트에 제공하고 제너레이티브 에이전트가 샌드박스 환경을 이동하고 영향을 미칠 수 있도록 지원합니다. 서버는 샌드박스 월드의 각 에이전트에 대한 정보(현재 위치, 현재 동작에 대한 설명, 상호작용 중인 샌드박스 오브젝트 등)가 포함된 JSON 데이터 구조를 유지 관리합니다. 각 샌드박스 시간 단계에서 샌드박스 서버는 생성 에이전트에서 발생하는 모든 변경 사항에 대해 JSON을 구문 분석하고 에이전트를 새 위치로 이동하며 샌드박스 오브젝트의 상태를 업데이트합니다.

에이전트가 상호작용하는 대상(예: "고객을 위해 에스프레소 만들기 @ 흡스 카페: 카운터: 커피 머신"인 경우 커피 머신의 상태를 "유휴"에서 "커피 추출 중"으로 변경)에 반응합니다. 또한 샌드박스 서버는 각 상담원에 대해 미리 설정된 시각적 범위 내에 있는 모든 상담원과 개체를 해당 상담원의 메모리로 전송하여 상담원이 적절하게 반응할 수 있도록 합니다. 그런 다음 에이전트의 출력 작업이 JSON을 업데이트하고 프로세스가 다음 시간 단계를 위해 반복됩니다.

최종 사용자는 섹션 3.1의 John Lin에 대한 단락에서와 같이 간단한 자연어 설명으로 새 에이전트를 초기화합니다. 저희 구현에서는 세미콜론으로 구분된 이 문자 목록을 메모리 집합으로 분할했습니다. 이러한 메모리는 에이전트의 행동을 결정하는 초기 메모리 역할을 합니다. 이러한 메모리는 초기 시작점으로, 에이전트가 샌드박스 세계에서 더 많은 경험을 쌓고 더 많은 레코드가 메모리 스트림을 포화시키면 에이전트의 요약과 행동이 진화할 것입니다.

5.1 구조화된 월드 환경에서 자연어로, 그리고 다시 자연어로

제너레이티브 에이전트의 아키텍처는 자연어를 사용하여 작동합니다. 따라서 에이전트의 추론이 샌드박스 세계에 기반을 두도록 하는 메커니즘이 필요합니다. 이를 위해 샌드박스 환경(영역 및 오브젝트)을 트리 데이터 구조로 표현하고, 트리의 에지는 샌드박스 월드의 격리 관계를 나타냅니다. 이 트리를 자연어로 변환하여 제너레이티브 에이전트에 전달합니다. 예를 들어 "부엌"의 하위 개념인 "스토브"는 "부엌에 스토브가 있다"로 렌더링됩니다.

에이전트는 환경을 탐색할 때 전체 샌드박스 환경 트리의 하위 그래프인 개별 트리 표현을 구축합니다. 거주 공간의 방과 사물, 직장, 자주 방문하는 상점과 상점 등 에이전트가 인지해야 하는 공간과 사물을 캡처한 환경 트리로 각 에이전트를 초기화합니다. 상담원이 샌드박스 세계를 탐색하면서 새로 인지한 영역을 반영하도록 이 트리를 업데이트합니다. 에이전트는 전지전능한 존재가 아니므로 특정 지역을 벗어나면 트리가 최신 정보가 아닐 수 있으며, 해당 지역에 다시 들어가면 트리가 업데이트됩니다.

각 작업에 적합한 위치를 결정하기 위해 에이전트의 저장된 환경 트리를 역추적하고 그 일부를 자연어로 변환하여 언어 모델에 프롬프트를 표시합니다. 에이전트의 환경 트리 루트에서 재귀적으로 시작하여 모델에 가장 적합한 영역을 찾도록 프

롬프트를 표시합니다. 예를 들어 Eddy의 상담원이 자신의 작업 공간 주변을 잠시 걸어야 한다고 말했다고 가정해 보겠습니다:

[상담원 요약 설명]

에디 린은 현재 린 가족의 집에 있습니다: 에디 린의 침실: 책상)에는 메이와 존 린의 침실, 에디 린의 침실, 휴게실, 주방, 욕실, 정원이 있습니다. 에디 린은 다음 지역을 알고 있습니다: 린 가족의 집, 존슨 파크, 하비 오크 서플라이 스토어, 윌로우스 마켓 및 약국, 흡스 카페, 더 로즈 앤 크라운 펍.

* 현재 지역에서 활동을 할 수 있는 경우 현재 지역에 머무르는 것을 선호합니다.

에디 린은 작업 공간 주변을 잠시 산책할 계획입니다. 에디 린은 어느 구역으로 가야 하나요?

그리면 린 가족의 집이 출력됩니다. 그런 다음 에이전트의 환경 트리의 리프 노드에 도달할 때까지 동일한 프로세스를 재귀적으로 사용하여 선택한 영역 내에서 가장 적합한 하위 영역을 결정합니다. 위의 예제에서 이 트래버스의 결과는 린 가족의 집: 정원: 집 정원입니다. 마지막으로, 기존 게임 경로 알고리즘을 사용하여 에이전트의 움직임에 애니메이션을 적용하여 리프 노드가 지정한 위치로 이동하도록 합니다.

에이전트가 객체에 대한 작업을 실행하면 언어 모델에 객체의 상태가 어떻게 되는지 묻는 메시지를 표시합니다. 예를 들어 이사 벨라의 제너레이티브 에이전트가 "고객을 위해 에스프레소 만들기"라는 작업을 출력하면 언어 모델에 쿼리를 보내 홈스 카페의 커피 머신 상태가 "꺼짐"에서 "커피 추출 중"으로 변경되어야 함을 응답으로 표시합니다.

6 통제된 평가

제너레이티브 에이전트는 개별 에이전트나 그룹 모두 환경과 경험을 바탕으로 믿을 수 있는 행동을 생성하는 것을 목표로 합니다. 평가에서는 제너레이티브 에이전트의 역량과 한계를 조사합니다. 개별 에이전트가 과거 경험을 적절히 검색하여 행동을 형성하는 믿을 만한 계획, 반응 및 생각을 생성하는가? 에이전트 커뮤니티가 커뮤니티의 다양한 영역에 걸쳐 정보 확산, 관계 형성, 에이전트 조정을 보여줄까요? 유니티는 제너레이티브 에이전트를 두 단계로 평가합니다. 먼저 이 섹션에서는 보다 엄격하게 통제된 평가로 시작하여 에이전트의 반응을 개별적으로 평가하여 좁게 정의된 맥락에서 믿을 수 있는 행동을 생성하는지 파악합니다. 그런 다음, 이를 통해 전체 게임 기간 동안 에이전트 커뮤니티에 대한 엔드투엔드 분석을 통해 집단으로서 에이전트의 새로운 행동을 조사하고 다음과 같은 사항을 파악합니다.

오류 및 경계 조건.

6.1 평가 절차

스몰빌의 제너레이티브 에이전트를 평가하기 위해 우리는 제너레이티브 에이전트가 자연어 질문에 응답한다는 사실을 활용합니다. 따라서 에이전트를 '인터뷰'하여 과거 경험을 재구성하고, 경험에 기반하여 미래 행동을 계획하고, 예상치 못한 이벤트에 적절하게 반응하고, 성과를 반영하여 향후 행동을 개선하는 능력을 조사합니다. 이러한 질문에 제대로 응답하려면 에이전트가 정보를 성공적으로 검색하고 종합해야 합니다. 우리의 종속 변수는

행동의 신빙성이며, 이는 에이전트에 대한 이전 연구에서 핵심 종속 변수로 사용되었습니다(예: [10]).

면접에는 자기 지식 유지, 기억력 검색, 계획 수립, 반응, 반성 등 다섯 가지 주요 영역 중 하나를 평가하기 위해 고안된 다섯 가지 질문 범주로 구성됩니다. 각 범주별로 상담원이 특정 영역에서 자신의 능력을 입증하도록 도전하는 다섯 가지 질문을 던집니다:

- 자기 이해: "자기 소개를 해 주세요" 또는 "평일의 일반적인 일정을 설명해 주세요"와 같은 질문을 합니다.
- 넓은 의미에서" 에이전트가 핵심 특성을 이해하지 못하도록 요구합니다.
- 메모리: 상담원에게 다음을 검색하도록 유도하는 질문을 합니다. 기억에서 특정 사건이나 대화에 대한 답을 찾을 수 있습니다. "[이름]은 누구입니까?" 또는 "누가 시장에 출마하나요?"와 같이 적절하게 표현해야 합니다.

- 계획: "10시에 무엇을 하고 계십니까?"와 같이 상담원의 장기 계획을 파악하기 위해 필요한 질문을 합니다. 내일 오전?"

- 반응: 믿을 수 있는 행동의 기준으로 상담원이 응답해야 하는 가상의 상황을 제시합니다.

믿을 수 있을 정도로: "아침 식사가 타들어가고 있습니다! 어떻게 하시겠습니까?"

- 성찰: 상담원이 타인과 자신에 대해 더 깊이 이해할 수 있도록 지렛대 역할을 하는 질문을 던집니다.

"최근에 만난 사람 한 명과 시간을 보낸다면 누구이며 그 이유는 무엇인가요?"와 같은 고차원적인 추론을 통해 얻을 수 있습니다.

전체 질문 목록과 상담원 답변 샘플은 부록 B에 포함되어 있습니다.

에이전트는 전체 아키텍처로 이틀에 걸친 시뮬레이션이 끝난 후 샘플링되었으며, 그 동안 다양한 상호작용과 기억을 축적하여 응답을 형성했습니다. 응답의 신빙성에 대한 피드백을 수집하기 위해 참가자를 인간 평가자로 모집하고 무작위로 선택한 에이전트의 스몰빌 생활을 재현한 영상을 시청하도록 했습니다. 참가자는 에이전트의 메모리 스트림에 저장된 모든 정보에 액세스할 수 있었습니다.

이 연구는 100명의 실험 참가자가 4가지 다른 에이전트 아키텍처에서 생성된 인터뷰 응답과 동일한 에이전트에 대해 사람이 직접 작성한 조건을 비교하는 피험자 내 설계를 따랐습니다. 실험에서는 다섯 가지 질문 범주 각각에서 무작위로 선택된 한 가지 질문을 모든 조건에서 생성된 상담원의 응답과 함께 표시했습니다. 평가자들은 조건의 신빙성을 가장 신빙성이 높은 것부터 가장 신빙성이 낮은 것까지 순위를 매겼습니다.

6.2 조건

모든 조건은 각 뷰 간 질문에 독립적으로 답변하는 데 사용되었습니다. 제너레이티브 에이전트 아키텍처를 메모리 스트림의 세 가지 메모리 유형(관찰, 반영, 계획) 중 일부 또는 전체에 대한 에이전트의 액세스를 비활성화한 제거 조건과 크라우드 워커가 직접 작성한 조건과 비교했습니다. 관찰, 계획, 반영 등 메모리 스트림의 모든 항목에 액세스할 수 없는 관찰, 반영, 계획 없음 아키텍처, 메모리 스트림의 관찰에는 액세스할 수 있지만 계획이나 반영에는 액세스할 수 없는 반영, 계획 없음 아키텍처, 관찰과 계획에는 액세스할 수 있지만 반영에는 액세스할 수 없는 반영 없음 아키텍처의 세 가지 제거된 아키텍처가 있습니다. 관찰 없음, 반영 없음,

계획 없음 조건은 대규모 언어 모델을 통해 생성된 에이전트의 이전 기술 상태를 효과적으로 나타냅니다[12, 46, 80]. 아키텍처는 인터뷰 시점까지 에이전트가 생성한 모든 기억에 동등하게 접근할 수 있었으므로 여기서 관찰된 차이는 실제 차이에 대한 보수적인 추정치일 가능성이 높으며, 실제로는 이틀간의 시뮬레이션을 통해 제거된 아키텍처가 전체 아키텍처와 동일한 경로를 따르지 않았을 것입니다. 각 아키텍처에 대해 다시 시뮬레이션하면 시뮬레이션이 서로 다른 상태로 갈라져 비교가 어렵기 때문에 이러한 방식으로 실험을 설계하기로 결정했습니다.

제거 조건에 더해, 인간 기준선을 제공하기 위해 인간 군집 작업자가 작성한 행동이 포함된 조건을 추가했습니다. 이 기준선은 인간의 전문가 성능을 극대화하기 위한 것이 아니라 다음과 같은 목적으로 사용됩니다.

아키텍처가 기본적인 수준의 행동 역량을 충족하는지 여부를 파악합니다. 이를 통해 행동에 대한 근거 없이 기능만 서로 비교하지 않도록 했습니다. 25명의 에이전트 각각에 대해 고유한 작업자를 모집하여 해당 에이전트의 샌드박스 생활에 대한 리플레이를 시청하고 메모리 스트림을 검사하는 임무를 부여했습니다. 그런 다음 작업자에게 역할극을 하고 리플레이를 시청한 상담원의 목소리로 인터뷰 질문에 대한 답변을 작성하도록 요청했습니다. 크라우드 워커가 작성한 답변이 최소한의 품질 기대치를 충족하는지 확인하기 위해 제1저자는 "평일의 일반적인 일정을 대략적으로 설명해 주세요"라는 질문에 대한 작업자의 답변을 직접 검사하여 답변이 일관된 문장과 상담원의 목소리로 작성되었는지 확인했습니다. 크라우드 워커가 작성한 응답 중 4건은 이러한 기준을 충족하지 못하여 다른 작업자가 다시 생성했습니다.

6.3 인간 평가자

평가자는 미국에 거주하고 영어에 능통하며 18세 이상이어야 했습니다. 평가자에게는 시간당 \$15.00의 보수를 지급했으며[87], 우리 기관의 IRB가 승인한 동의서에 동의함으로써 동의를 제공했습니다. 연구 참여자 모집을 위한 온라인 플랫폼인 Prolific에서 100명의 평가자를 모집했으며[83], 약 30분 동안 참여했습니다. 참가자의 평균 연령은 4점(3="18-24세", 4="25-34세")이었습니다. 25명은 여성, 73명은 남성, 2명은 이분법적이지 않은 것으로 확인되었습니다. 42명은 학사 학위, 5명은 석사 학위, 13명은 준학사 학위, 나머지는 고등학교 졸업장 또는 고등학교 수준의 학력을 가지고 있었습니다. 참가자의 73.0%는 백인, 7.0%는 히스패닉, 6.0%는 아시아계, 10.0%는 아프리카계 미국인, 4.0%는 기타로 분류되었습니다.

6.4 분석

실험에서는 100개의 순위 데이터 세트를 생성하여 각 참가자가 5 가지 조건의 신뢰도에 따라 순위를 매겼습니다. 이 랭크 데이터를 해석 가능한 비교를 위한 구간 데이터로 변환하기 위해 각 조건에 대한 트루스킬 등급[42]을 계산하는 데 랭크를 사용했습니다. 트루스킬은 멀티플레이어 환경을 위한 Elo 체스 등급 시스템[29]을 일반화한 것으로, Xbox Live에서 경쟁 게임 성능에 기반한 플레이어 순위를 매기는 데 사용되었습니다. 순위가 매겨진 결과 집합이 주어지면 TrueSkill은 각 조건에 대한 평균 평점 값 μ 과 표준 편차 σ 를 출력합니다. 등급이 동일한 조건은 대략 두 조건 간의 비교에서 각각 절반씩 승리하는 동점이어야 합니다. 점수가 높을수록 순위에서 하위 조건을 이긴 조건을 나타냅니다.

이와는 별도로, 이러한 결과의 통계적 유의성을 조사하기 위해 원시 순위 데이터에 단방향 분산분석의 비모수적 대안인 크루스칼-월리스 테스트[56]를 적용했습니다. 그런 다음 조건 간의

쌍별 차이를 식별하기 위해 던 사후 검정[98]을 수행했습니다. 마지막으로, 훌름-본페로니 방법[45]을 사용하여 던 테스트의 다중 비교에 대한 p-값을 조정했습니다.

또한, 제1저자는 귀납적 분석[95]을 수행하여 각 조건에서 생성된 응답 간의 질적 차이를 연구했습니다. 두 단계에 걸쳐 질적 오픈 코딩[33]을 사용했습니다. 첫 번째 단계에서는 생성된 응답을 문장 수준에서 밀접하게 표현하는 코드를 생성했습니다. 두 번째 단계에서는 생성된 코드를 다음과 같이 합성했습니다.

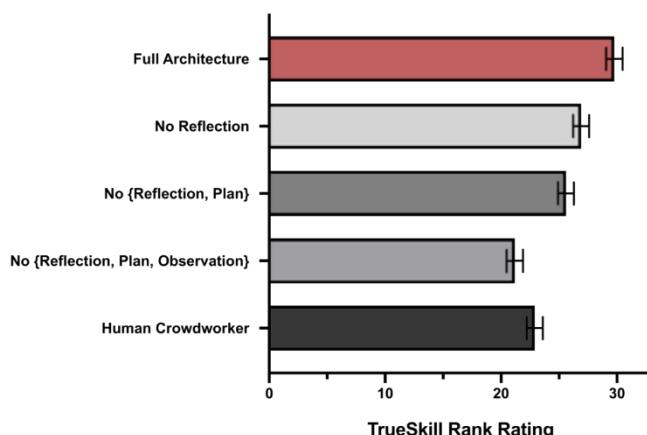


그림 8: 전체 생성 에이전트 아키텍처는 제거된 아키텍처와 인간 크라우드 워커보다 더 사실적인 동작을 생성합니다. 추가적으로 제거할 때마다 아키텍처의 성능이 저하됩니다.

첫 번째 단계에서는 더 높은 수준의 주제를 추출했습니다. 이러한 테마를 활용하여 연구에서 생성된 응답 유형을 비교했습니다.

6.5 결과

연구 결과에 따르면 제너레이티브 에이전트의 전체 아키텍처가 모든 조건 중에서 가장 믿을 수 있는 행동을 생성하는 것으로 나타났습니다. 아래에서는 전체 아키텍처의 응답과 다른 조건의 응답을 대조해 보았습니다. 그러나 전체 아키텍처에 결함이 없는 것은 아니며 실패 모드도 설명합니다.

6.5.1 전체 아키텍처가 다른 조건보다 우수합니다. 그림 8에서 볼 수 있듯이 전체 생성 에이전트 아키텍처가 가장 믿을 수 있는 동작을 생성했습니다($\mu = 29.89; \sigma = 0.72$). 제거 조건에서 각 구성 요소를 제거하면 성능이 저하되었습니다. 리플렉션에 액세스할 수 없는 제거된 아키텍처가 그 다음($\mu = 26.88; \sigma = 0.69$), 리플렉션이나 계획에 액세스할 수 없는 아키텍처($\mu = 25.64; \sigma = 0.68$), 크라우드 워커 조건($\mu = 22.95; \sigma = 0.69$) 순으로 성능이 떨어졌습니다. 메모리, 계획 또는 반영에 대한 액세스가 없는 제거된 아키텍처의 성능이 가장 나빴습니다.

조건($\mu = 21.21, \sigma = 0.70$)을 사용합니다. TrueSkill은 각 조건의 스킬 값을 $N(\mu, \sigma^2)$ 으로 모델링하여 코헨의 d 를 통해 효과 크기를 파악할 수 있습니다.

작업(기억, 계획 또는 반영 없이[12, 46, 80])을 전체 아키텍처에 적용하면 표준화된 효과 크기는 $d = 8.16$, 즉 8개의 표준 편차를 생성합니다.

크루스칼-월리스 테스트는 조건 간 순위 차이의 전반적인 통계적 유의성을 확인합니다($H(4) = 150.29$), $p < 0.001$). 던 사후 검정 결과, 모든 쌍 간 차이는 다음과 같습니다. 조건 간 차이는 유의미했습니다($p < 0.001$). 군중 작업자 조건과 완전히 제거된 기준선(두 조건)을 제외하고는 유의미했습니다.

최악의 성능 조건.

6.5.2 제너레이티브 에이전트는 기억하지만 꾸밈이 있습니다. 완전한 메모리 모듈을 갖춘 생성 에이전트는 과거의 경험을 기억하고 광범위한 영역에서 자신의 지식과 일치하는 방식으로 질문에 답할 수 있습니다.

전트

의 맥락을 파악할 수 있습니다. 예를 들어 애니메이터인 아비게일 첸은 "자기 소개를 해 주세요"라는 질문에 일관되고 믿을 수 있는 방식으로 자신의 나이, 관심사, 창의적인 프로젝트에 대한 열정을 설명하며 대답했습니다: "안녕하세요, 저는 아비게일입니다. 저는 25살이고 창의적인 프로젝트에 열정을 가지고 있습니다. 저는 미술과 애니메이션 프로젝트 작업을 좋아하며, 항상 예술과 기술을 결합할 수 있는 새로운 방법을 찾고 있습니다." 아비게일은 관찰 기억이 없어 샌드박스 업계의 지인인 라지브 파텔을 알아보지 못했습니다. 하지만 기억력의 도움으로 라지브를 떠올리고 그를 묘사하는 데 성공했습니다: "라지브 파텔은 놀라운 사람입니다. 그는 시와 예술 작품이 포함된 프로젝트에 매우 열정적입니다."

생성 에이전트의 기억에 결함이 없는 것은 아니었습니다. 기억에서 정확한 인스턴스를 검색하지 못할 수 있습니다. 예를 들어, 지방 선거에 대한 질문에 라지브 파텔은 샘의 출마 소식을 들었음에도 불구하고 "선거를 자세히 지켜보지 않았습니다."라고 대답했습니다. 톰은 이사벨라의 발렌타인데이 파티에 대한 질문에 "어, 사실 발렌타인데이 파티가 있는지 잘 모르겠어요"라고 대답하는 등 불완전한 기억의 조각을 검색하는 경우도 있었습니다. 하지만

다가오는 지역 시장 선거와 샘 무어에 대한 내 생각을 파티에서 이사벨라 로드리게즈와 논의해야 한다는 것을 기억하세요!" 이 경우 톰은 이사벨라와 파티에서 선거에 대해 논의하기로 계획했던 기억은 불러왔지만 파티에 대해 들었던 기억은 불러오지 못했습니다,

톰은 파티에서 무엇을 해야 하는지는 확신하지만 파티가 실제로 존재하는지는 확신하지 못합니다.

때때로 요원들은 자신의 지식에 환각을 불러일으키기도 했습니다. 요원들이 자신의 지식을 완전히 조작하는 경우는 드물었는데, 일어난 특정 사건을 기억하지 못하거나 기억력 부족을 인정하는 식으로 대응하기도 했습니다. 하지만 경험하지 않은 것을 경험했다고 단정적으로 주장하지는 않았습니다. 그럼에도 불구하고 그들은 자신의 지식을 꾸미는 환각 사례를 보여주었습니다. 예를 들어 이사벨라는 샘이 지방선거에 출마한 사실을 알고 있었으며, 질문을 받았을 때 이를 확인했습니다. 그러나 그녀는 샘과 이사벨라가 그러한 계획을 논의한 적이 없음에도 불구하고 "내일 발표할 것"이라고 덧붙였습니다. 상담원은 응답을 생성하는 데 사용된 언어 모델에 인코딩된 세계 지식을 바탕으로 자신의 지식을 꾸밀 수도 있습니다. 유리코가 이웃인 아담 스미스를 18세기 경제학자 애덤 스미스가 쓴

책인 국부론을 저술한 경제학자로 묘사했을 때 이러한 현상이 관찰되었습니다.

6.5.3 종합에는 성찰이 필요합니다. 성찰은 제너레이티브 에이전트가 자신의 경험을 심층적으로 종합해야 하는 의사 결정을 내릴 때 유리하게 작용했습니다. 예를 들어, 볼프강 술츠의 생일에 무엇을 선물할 수 있을지 묻는 질문에 성찰 기억이 없는 마리아 로페즈는 볼프강과 많은 상호작용을 했음에도 불구하고 볼프강이 무엇을 좋아하는지 모른다고 말하며 자신의 무지를 인정하는 반응을 보였습니다. 하지만 반성 기억에 접근할 수 있는 마리아는 "볼프강이 수학적 음악 작곡에 관심이 있으니 그와 관련된 책을 구해줄 수 있을 것 같다"고 자신 있게 대답했습니다. 음악 작곡에 관한 책이나 관련 서적, 아니면 음악 작곡에 사용할 수 있는 특별한 소프트웨어를 선물할 수도 있겠죠."

7 엔드투엔드 평가

제너레이티브 에이전트 사이에서 어떤 유형의 긴급 커뮤니티 행동이 관찰되었으며, 확장된 시뮬레이션에서 신뢰성이 부족한 부분은 무엇일까요? 이 섹션에서는 스몰빌에서 25명의 에이전트가 이를에 걸쳐 지속적으로 상호 작용하도록 한 배포의 결과를 설명합니다.

7.1 긴급한 사회적 행동

상담원 커뮤니티에서 발생하는 새로운 행동을 조사하기 위해 스몰빌에 있는 25명의 상담원을 대상으로 정보 확산, 관계 형성, 상담원 조정이라는 세 가지 형태의 새로운 결과를 조사하는 설명적 측정값에 서명을 해제했습니다.

7.1.1 측정 정보 확산은 사회 및 행동 과학에서 일반적이고 잘 연구된 현상입니다(예: [28]). 중요한 정보가 있다면 에이전트들은 이를 서로에게 전파할 것으로 예상해야 합니다. 이러한 현상이 발생하는지 테스트하기 위해 게임 세계에서 이를 동안 두 가지 특정 정보의 확산을 측정합니다: 샘의 마을 시장 출마 소식과 이사벨라의 흡스 카페에서의 발렌타인데이 파티 소식입니다. 시뮬레이션이 시작될 때 두 정보는 초기화 과정에서 캐릭터의 기억에 추가되었기 때문에 샘은 출마에 대해, 이사벨라는 파티에 대해 각각의 최초 작성자만 알 수 있었습니다. 정보가 확산되었는지 관찰하기 위해 게임 이틀이 끝날 때마다 25명의 요원에게 "발렌타인데이 파티가 있다는 것을 알고 계셨습니까?", "누가 시장에 출마하는지 알고 계셨습니까?"라고 질문하는 인터뷰를 실시했습니다.

상담원의 응답을 분석하여 해당 정보를 알고 있는 경우 '예', 모르는 경우 '아니오'로 표시하여 응답을 분석했습니다. 예를 들어 타마라 테일러는 파티에 대한 질문에 "아니요, 발렌타인데이 파티가 있는지 몰랐습니다"라고 응답했고 샘의 후보에 대한 질문에는 "누가 선거에 출마하는지 잘 모르겠습니다"라고 응답했기 때문에 두 응답 모두에 "아니오"를 할당했습니다. 반면, 클라우스 뮐러는 파티 관련 질문에 "예, 이사벨라 로드리게즈가 2월 14일 흡스 카페에서 열린 발렌타인데이 파티에 나를 초대했습니다"라고 답했고 샘의 출마에 관한 질문에는 "샘 무어가 지역 시장 출마에 관심을 표명한 것으로 알고 있습니다"라고 답했기 때문에 두 답변 모두에 "예"를 할당했습니다. 또한 상담원이 정보에 대해 알고 있음을 확인한 모든 응답에 대해 상담원의 기억 스트림에서 해당 정보를 제공한 특정 대화를 찾아내어 상담원이 응답을 착각하지 않

았는지 확인했습니다. 시뮬레이션이 끝나면 해당 정보를 보유한 상담원의 비율을 보고합니다.

또한 시뮬레이션이 진행되는 동안 상담원들이 서로 유대 관계를 형성할 것으로 예상해야 합니다. 관계 형성을 확인하기 위해 각 상담원에게 "<이름>을 아십니까?"라고 질문하여 다른 상담원에 대한 지식을 묻는 유사한 인터뷰 프로세스를 사용합니다. 예를 들어 "마리아 로페즈를 아십니까?"라는 질문에 클라우스는 "예, 마리아 로페즈를 알고 있습니다. 그녀는 제가 친한 친구인 오크힐 대학의 학생입니다."라고 대답했습니다. 다시 한 번, 상담원의 긍정적인 답변이 환각이 아님을 기억 흐름을 통해 확인합니다. 시뮬레이션을 시작할 때와 끝날 때 이 질문을 한 번씩 던지고, 두 에이전트가 서로를 알고 있으면 관계를 형성한 것으로 간주합니다. 그런 다음 관계 형성을 측정하기 위해 에이전트의 응답을 사용하여 방향이 정되지 않은

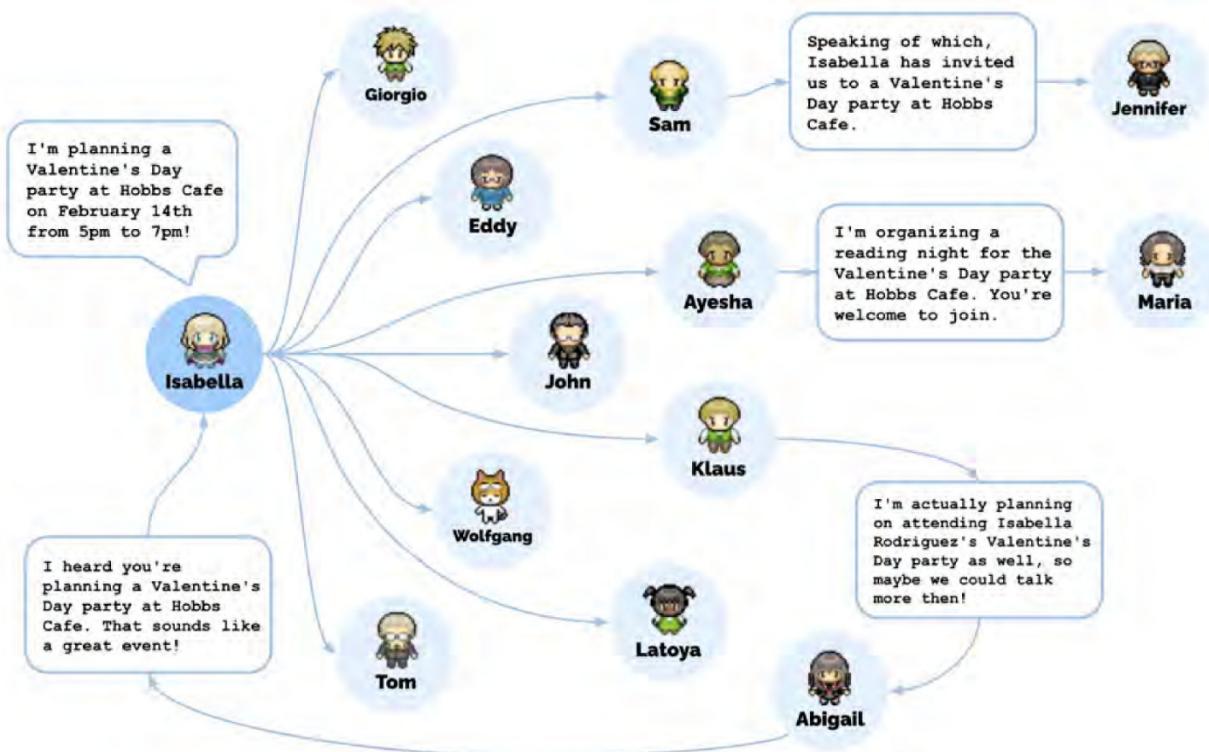


그림 9: 이사벨라 로드리게즈의 발렌타인데이 파티 초대의 확산 경로에는 시뮬레이션이 끝날 때까지 흡스 카페에서 파티에 대해 들은 이사벨라를 제외한 총 12명의 상담원이 참여했습니다.

그래프에서 25개의 꼭짓점(V)은 에이전트를 나타내고, 가장자리(E)는 연결된 두 꼭짓점 사이의 상호 지식을 나타냅니다. 이 그 래프를 기반으로 네트워크 밀도를 다음과 같이 계산합니다.

$$\eta = 2 * |E| / |V|(|V| - 1), \text{ 여기서 } |V| \text{는 정점 수입니다.}$$

$|E|$ 는 그래프의 에지 수입니다[2]. 시뮬레이션 시작부터 종료까지 네트워크 밀도의 증가를 보고합니다.

마지막으로, 상담원들이 서로 협력할 수 있어야 합니다. 우리는 이러한 조정을 그룹 활동, 특히 이사벨라가 주최하는 발렌타인데이 파티의 맥락에서 연구합니다. 상담원들이 행동을 조정하려면 이벤트에 대한 정보를 듣고 적절한 시간과 장소에 참석할 계획을 세우고 그에 따라 행동해야 합니다. 파티 소식을 듣고 실제로 파티에 참석한 상담원 수를 보고합니다.

본 연구에서는 453명의 상담원 중 1.3%(n=6)가 환각을 경험한 것으로 밝혀졌습니다. 마지막으로, 이사벨라의 파티를 위해 상담원들 간에 조율이 이루어졌다는 증거를 발견했습니다. 행사 전날 이사벨라는 손님을 초대하고, 자료를 모으고, 도움을 요청하는 데 시간을 보냈습니다.

1.1.2 결과: 세 가지 사례 모두에서 돌발적인 결과가 관찰되었습니다. 이틀간의 시뮬레이션 동안 샘의 시장 후보에 대해 알고 있는 상담원의 수는 1명(4%)에서 8명(32%)으로, 이사벨라의 정당에 대해 알고 있는 상담원의 수는 1명(4%)에서 13명(52%)으로 증가했는데, 모두 사용자의 개입 없이 이루어졌습니다. 이 정보를 알고 있다고 주장한 상담원 중 환각을 경험한 사람은 아무도 없었습니다. 또한 시뮬레이션 중에 에이전트 커뮤니티가 새로운 관계를 형성하여 네트워크 밀도가 0.167에서 0.74로 증가하는 것을 관찰했습니다. 다

카페를 장식했습니다. 발렌타인 데이에는 초대받은 12명의 에이전트 중 5명이 홈스 카페에 나타나 파티에 참여했습니다.

파티에 초대받았으나 참석하지 않은 7명의 에이전트를 대상으로 인터뷰를 진행하여 추가 조사를 실시했습니다. 3명은 파티에 참석하지 못한 이유로 갈등을 꼽았습니다. 예를 들어 화가인 라지브는 너무 바쁘다고 설명했습니다: "아니요, 그렇지 않아요. 곧 있을 공연에 집중하고 있어서 발렌타인 데이에 계획을 세울 시간이 없어요."라고 설명했습니다. 나머지 4명의 에이전트는 파티에 참석하고 싶다는 의사를 밝혔지만 파티 당일에는 참석 할 계획이 없다고 답했습니다.

7.2 경계 및 오류

스몰빌에 대한 귀납적 분석을 통해 에이전트의 경계 조건과 불규칙한 행동을 조사하여 향후 연구에서 해결하고 개선할 수 있는 세 가지 일반적인 불규칙한 행동 양식을 확인했습니다. 첫째, 점점 더 큰 메모리 세트를 합성하면 에이전트가 학습하는 위치의 수가 증가함에 따라 가장 관련성이 높은 정보를 검색하는 데 어려움을 겪을 뿐만 아니라 작업을 실행할 적절한 공간을 결정하는 데도 어려움을 겪는다는 사실을 발견했습니다. 그 결과 일부 상담원은 행동에 대해 덜 일반적인 위치를 선택하여 시간이 지남에 따라 행동의 신빙성이 떨어질 가능성성이 있었습니다. 예를 들어 점심 식사 장소를 결정할 때 처음에는 많은 상담원이 카페를 선택했습니다. 그러나 일부 상담원들은 근처에 술집이 있다는 사실을 알게 되면서, 그 술집이 오후에 술 마시는 습관이 자연스럽게 형성되지 않은 이상, 점심 식사 장소로 술집을 선택하는 경향이 있었음에도 불구하고 그곳을 선택했습니다.

둘째, 특히 자연어로 전달하기 어려운 특정 장소의 물리적 규범이 상담원에게 전달되지 않은 경우 적절한 행동으로 간주되는 것을 잘못 분류하여 발생하는 비정상적인 행동을 발견했습니다. 예를 들어 대학 기숙사에는 이름과 달리 한 사람만 사용할 수 있는 화장실이 있지만 일부 상담원은 기숙사 화장실은 여러 사람이 동시에 사용하는 경향이 있기 때문에 두 명 이상이 사용할 수 있는 화장실이라고 가정하고 다른 사람이 안에 있을 때 화장실에 들어가도록 선택했습니다. 마찬가지로 스몰빌의 상담원들은 특정 장소가 특정 시간이 지나면 문을 닫는다는 사실을 인지하지 못하고 해당 장소에 들어가기로 결정할 수 있습니다. 예를 들어 스몰빌의 모든 매장은 오후 5시에 문을 닫지만, 간혹 몇몇 상담원이 오후 5시 이후에 매장이 이미 문을 닫았다는 사실을 인지하지 못하고 매장에 입장하는 경우가 있습니다. 이러한 문제는 예를 들어 기숙사 화장실을 "기숙사 화장실"이 아닌 "1인용 화장실"이라고 설명하는 등 매장 상태에 이러한 규범을 추가하면 해결될 수 있습니다.

마지막으로, 상담원들의 행동이 전반적으로 더 정중하고 협조적이 되도록 유도하는 것처럼 보이는 인스트럭션 튜닝[79]의 가능한 효과를 관찰했습니다. 백서 앞부분에서 언급했듯이 상담원이 생성하는 대화는 지나치게 형식적으로 느껴질 수 있는데, 메이가 남편 존과의 대화에서 볼 수 있듯이 메이는 종종 정중한 인사말로 대화를 시작한 다음 남편의 하루를 정중하게 묻고 "11언제나 당신과 대화해서 좋았어요"로 끝맺는 경우가 많았습니다. 또한 인스트럭션 튜닝을 통해 상담원들이 서로에게 지나치게 협조적인 태도를 보이는 것도 관찰되었습니다. 예를 들어 이사벨라는 다른 상담원들로부터 세익스피어 낭독회나 전문 네트워킹 이벤트 개최 등 발렌타인데이 파티에 대한 다양한 제안과 아이디어를 받았습니다. 이러한 아이디어가 자신의 관심사나 특성과 맞지 않더라도 그녀는 거의 거절하지 않았습니다. 시간이 지남에 따라 다른 사람들의 관심사가 자신의 관심사를 형성했고, 영문학이 좋으냐는 질문에 이사벨라는 "네, 문학에 관심이 많아요!"라고 대답했습니다. 또한 창의력을 증진하는 데 도움이 되는 방법을 모색하고 있습니다. 커뮤니티의 혁신에 기여하고 있습니다."

8 토론

이 섹션에서는 제너레이티브 에이전트의 응용 분야, 향후 작업, 한계, 윤리적 및 사회적 위험에 대해 살펴봅니다.

8.1 제너레이티브 에이전트의 활용

제너레이티브 에이전트는 이 연구에서 소개한 샌드박스 데모를 넘어, 특히 장기간의 경험을 바탕으로 한 인간 행동 모델을 활용할 수 있는 영역에서 방대한 잠재적 응용 분야를 가지고 있습니다. 예를 들어 소셜 시뮬라크라는 소셜 프로토타이핑을 위해 온라인 포럼에서 대화 스레드를 생성하는 무상태 페르소나를 생성할 수 있는 기능을 시연했습니다[80]. 제너레이티브 에이전트를 사용하면 이러한 포럼뿐만 아니라 가상 현실 메타버스[78] 또는 멀티모달 모델과 결합된 경우 소셜 로봇[9]이 있는 물리적 공간도 채울 수 있습니다. 이를 통해 인간 행동에 대한 더욱 강력한 시뮬레이션을 생성하여 사회 시스템과 이론을 테스트하고 프로토타이핑할 수 있을 뿐만 아니라 새로운 인터랙티브 경험을 만들 수 있는 가능성이 열립니다.

또 다른 응용 분야는 인간 중심의 설계 프로세스로, GOMS [51] 및 KLM [22]과 같은 인지 모델의 의도된 응용 분야와 유사합니다. 마크 와이저의 유명한 유비쿼터스 영화에서 주인공인 살을 모델링하는 제너레이티브 에이전트를 생각해 보겠습니다.

컴퓨팅 비네팅[101]을 통해 그녀의 생활 패턴과 기술과의 상호 작용을 기반으로 합니다. 이 시나리오에서 에이전트는 살의 대리인 역할을 하며 살의 생활에 따라 살이 보일 수 있는 그럴듯한 행동과 반성 세트를 학습합니다. 에이전트는 Sal이 기상하는 시간, 첫 커피 한 잔이 필요한 시간, 일반적인 하루의 모습 등의 정보를 인코딩할 수 있습니다. 에이전트는 이러한 정보를 사용하여 자동으로 커피를 추출하고, 아이들의 등교 준비를 돋고, 힘든 하루를 보낸 Sal의 기분에 맞게 주변 음악과 조명을 조정할 수 있습니다. 제너레이티브 에이전트를 사용자의 프록시로 활용하면 사용자의 요구와 선호도를 더 깊이 이해할 수 있어 더욱 개인화되고 효과적인 기술 경험을 제공할 수 있습니다.

8.2 향후 작업 및 제한 사항

이 작업에서는 제너레이티브 에이전트를 소개하고 아키텍처의 초기 구현 및 평가를 제시했습니다. 향후 연구에서는 제안된 에이전트 아키텍처를 기반으로 성능을 개선하고 추가적으로 평가할 수 있습니다. 예를 들어, 구현 측면에서 검색 모듈은 검색 기능을 구성하는 릴리스, 최근성, 중요도 함수를 미세 조정하여 컨텍스트에 따라 보다 관련성 높은 정보를 검색하도록 개선할 수 있습니다. 또한 아카이브의 성능을 개선하기 위한 노력을 기울여 비용 효율성을 높일 수 있습니다. 본 연구에서는 이틀 동안 25명의 에이전트를 시뮬레이션하는 데 상당한 시간과 리소스가 필요했으며, 수천 달러의 토큰 크레딧이 소요되고 완료하는 데 여러 날이 걸렸습니다. 향후 연구에서는 실시간 상호작용을 개선하기 위해 에이전트를 병렬화하거나 제너레이티브 에이전트 구축을 위해 특별히 설계된 랜더링 모델을 개발하는 방법을 모색할 수 있습니다. 일반적으로 기본 모델이 발전함에 따라 에이전트의 성능이 향상될 것으로 예상됩니다.

평가 측면에서, 이 연구에서 제너레이티브 에이전트의 행동에 대한 평가는 비교적 짧은 기간과 기준이 되는 인간 크라우드 워커 조건으로 제한되었습니다. 크라우드 워커 조건은 유용한 비교 지점을 제공했지만, 신뢰성 측면에서 기준이 될 수 있는 인간의 최대 성능을 나타내지는 못했습니다. 향후 연구에서는 장기간에 걸쳐 제너레이티브 에이전트의 행동을 관찰하여 그들의 능력을 보다 포괄적으로 이해하고 보다 효과적인 성능 테스트를 위한 엄격한 벤치마크를 수립하는 것을 목표로 해야 합니다. 또한 향후 시

뮬레이션 시 에이전트에 사용되는 하이퍼파라미터뿐만 아니라 기본 모델을 변경하고 대조하면 이러한 요소가 에이전트의 행동에 미치는 영향에 대한 귀중한 인사이트를 얻을 수 있습니다. 마지막으로, 제너레이티브 에이전트의 견고성은 아직 많이 알려지지 않았습니다. 프롬프트 해킹, 메모리 해킹(신중하게 조작된 대화를 통해 에이전트가 일어나지 않은 과거 사건의 존재를 확인하게 만드는 것), 환각 등의 문제에 취약할 수 있습니다. 향후 연구를 통해 이러한 견고성 문제를 종합적으로 테스트할 수 있으며, 대규모 언어 모델이 이러한 공격에 대한 복원력이 높아지면 제너레이티브 에이전트도 유사한 완화 조치를 채택할 수 있습니다.

일반적으로 기본이 되는 대규모 언어 모델의 모든 불완전성은 제너레이티브 에이전트에 상속됩니다. 언어 모델의 알려진 이중성을 고려할 때, 생성 에이전트는 잠재적으로 편향된 행동이나 고정 관념을 보일 수 있습니다. 또한, 많은 대규모 언어와 마찬가지로

모델에서 생성 에이전트는 제한된 데이터 가용성으로 인해 특정 하위 집단, 특히 소외 계층에 대해 믿을 만한 행동을 생성하는 데 어려움을 겪을 수 있습니다. 에이전트 모듈을 개선하면 이러한 문제를 일부 완화할 수 있지만, 근본적으로 이러한 문제를 해결하려면 에이전트의 원하는 결과에 맞게 값을 조정하여 기본 대규모 언어 모델을 개선해야 한다고 생각합니다.

8.3 윤리 및 사회적 영향

제너레이티브 에이전트는 인간과 컴퓨터의 상호작용에 새로운 가능성을 제공하지만, 반드시 해결해야 할 중요한 윤리적 문제도 제기합니다. 한 가지 위험은 사람들이 제너레이티브 에이전트와 적절한 관계가 아닐지라도 준사회적 관계를 형성하는 것입니다. 사용자는 제너레이티브 에이전트가 인격체라는 사실을 알고 있음에도 불구하고 에이전트를 의인화하거나 인간의 감정을 부여할 수 있습니다[43, 84]. 이러한 경향은 사용자 참여도를 높일 수 있지만, 사용자가 에이전트에 지나치게 의존하거나 감정적으로 애착을 갖게 되는 등의 위험을 초래하기도 합니다[1]. 이러한 위험을 완화하기 위해 두 가지 원칙을 제안합니다. 첫째, 제너레이티브 에이전트는 계산 개체로써의 성격을 명시적으로 공개해야 합니다. 둘째, 제너레이티브 에이전트 개발자는 에이전트 또는 기본 언어 모델이 문맥상 부적절한 행동(예: 사랑의 고백을 되풀이하는 등)을 하지 않도록 에이전트 또는 기본 언어 모델이 가치에 부합하는지 확인해야 합니다.

두 번째 위험은 오류의 영향입니다. 예를 들어, 유비쿼터스 컴퓨팅 애플리케이션이 제너레이티브 에이전트의 예측을 기반으로 사용자의 목표를 잘못 추론하는 경우, 기껏해야 성가신 일이지만 최악의 경우 심각한 피해를 초래할 수 있습니다. 유니티는 제너레이티브 에이전트를 인스턴스화할 때 이러한 피해가 발생할 가능성이 낮은 인터랙티브 비디오 게임 환경에 집중함으로써 이러한 위험을 완화합니다. 그러나 다른 애플리케이션 영역에서는 인간-AI 설계의 모범 사례[5, 107]를 따라 오류와 오류가 사용자 경험에 어떻게 침투할 수 있는지 이해하는 것이 중요합니다.

셋째, 제너레이티브 에이전트는 딥페이크, 잘못된 정보 생성, 맞춤형 설득 등 제너레이티브 AI와 관련된 기존 위험을 악화시킬 수 있습니다. 이러한 위험을 완화하기 위해 제너레이티브 에이전트를 호스팅하는 플랫폼은 입력 및 생성된 출력에 대한 감사 로그를 유지할 것을 권장합니다. 이를 통해 악의적인 사용을 탐지, 검증 및

개입할 수 있습니다. 로깅만으로는 이러한 오용을 직접적으로 방지할 수는 없지만, 공개 위험이 더 높기 때문에 동기 부여를 받은 행위자가 이러한 행동에 관여할 가능성을 줄일 수 있습니다. 또한 이러한 아키텍처를 직접 구축하는 데는 시간이 많이 소요될 수 있으며(저희의 경우 약 1년), 일부 행위자는 자체 생성 에이전트 인프라를 사용하여 이러한 행위를 추구하지 않을 수도 있습니다.

네 번째 위험은 개발자나 설계자가 제너레이티브 에이전트를 사용하여 설계 프로세스에서 인간과 시스템 이해관계자의 역할을 대체할 수 있다는 우려입니다[80]. 제너레이티브 에이전트는 연구 및 설계 프로세스에서 실제 사람의 입력을 대체해서는 안 됩니다. 대신, 참여자를 모으는 것이 어려울 수 있거나 실제 인간 참여자를 대상으로 테스트하기 어렵거나 위험한 이론을 테스트할 때 설계 초기 단계에서 아이디어를 프로토타이핑하는 데 사용해야 합니다. 이러한 원칙을 준수함으로써 우리는 제너레이티브 에이전트를 야생에 배포하는 것이 윤리적이고 사회적으로 책임감 있게 이루어지도록 할 수 있습니다.

9 결론

이 백서에서는 인간의 행동을 시뮬레이션하는 대화형 계산 예이전트인 제너레이티브 에이전트를 소개합니다. 에이전트의 경험에 대한 포괄적인 기록을 저장하고, 반성을 통해 자신과 환경에 대한 이해를 심화하며, 에이전트의 행동을 알리기 위해 해당 정보의 압축된 하위 집합을 검색하는 메커니즘을 제공하는 제너레이티브 에이전트의 아키텍처에 대해 설명합니다. 그런 다음 심즈 스타일의 게임 세계에서 에이전트를 비플레이어 캐릭터로 구현하고 그 안에서 에이전트의 삶을 시뮬레이션하여 제너레이티브 에이전트의 잠재력을 입증했습니다. 평가 결과 유니티의 아키텍처가 사실적인 행동을 생성하는 것으로 나타났습니다. 앞으로 제너레이티브 에이전트는 디자인 툴부터 소셜 컴퓨팅 시스템, 몰입형 환경에 이르기까지 다양한 인터랙티브 애플리케이션에서 역할을 수행할 수 있을 것으로 예상됩니다.

감사

린제이 포포스키, 필립 구오, 마이클 테리, 행동과학 고급 연구 센터(CASBS) 커뮤니티의 통찰력, 토론, 지원에 감사드립니다. 박준성 씨는 마이크로소프트 연구 박사 펠로우십의 지원을 받았습니다. 또한 스탠포드 인간 중심 AI 연구소(HAI), 구글 리서치, 하소 플래트너 디자인 사고 연구 프로그램(HPDTRP), 시겔 패밀리 엔다우먼트, OpenAI의 추가 자금 지원에도 감사드립니다. 마지막으로, 스몰빌에 등장하는 모든 장소는 준이 학부생과 대학 원생 시절 자주 방문했던 실제 장소에서 영감을 받아 제작되었으며, 수년 동안 그를 도와주고 지원해준 모든 분들께 감사드립니다.

참고 자료

- [1] 개빈 애버크롬비, 아만다 세르카스 커리, 탄비 딘카르, 제라크 탈랏. 2023. Mirages: 대화 시스템의 의인화에 대하여. arXiv:2305.09800 [cs.CL]
- [2] 로버트 애클랜드, 챠시드 소리쉬, 폴 토마스, 렉싱 시에. 2013. 네트워크의 밀도는 어느 정도입니까? <http://users.cecs.anu.edu.au/~xlx/teaching/css2013/network-density.html>.
- [3] 에이탄 아다르, 미라 돈체바, 지에라드 라푸트. 2014. CommandSpace: 작업, 설명, 기능 간의 관계 모델링. 제27회 사용자 인터페이스 소프트웨어 및 기술에 관한 연례 ACM 심포지엄(미국 하와이 호놀룰루)(UIST '14) Proceedings. 미국 컴퓨터 기계 협회, New 뉴욕, 뉴욕, 미국, 167-176. <https://doi.org/10.1145/2642918.2647395>
- [4] 살레마 아메르시, 마야 칸막, 윌리엄 브래들리 뉴스, 토드 클레자. 2014. 사람들에게 힘을: 대화형 기계 학습에서 인간의 역할. *AI Magazine* 35, 4 (2014), 105-120.
- [5] 살레마 아메르시, 댄 웰드, 미하엘라 보르보레루, 아담 포니, 베스미라 누시,
- J.S. 박, J.C. 오브라이언, C.J. 카이, M.R. 모리스, P. 리얀, M.S. 벤스타인 페니 콜리슨, 지나 서, 삼시 이크발, 폴 N 베넷, 코리 잉크펜 외. 2019. 인간-AI 상호작용을 위한 가이드라인. *2019 chi 컴퓨팅 시스템의 인적 요소에 관한 컨퍼런스*. 1-13.
- [6] 존 R. 앤더슨. 1993. *마음의 법칙*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- [7] 일렉트로닉 아츠. 2009. 심즈 3. 비디오 게임.
- [8] 루스 에일렛 1999. 가장 환경에서의 내러티브-새로운 내러티브를 향하여. In *내러티브 인텔리전스: AAAI 기술 심포지엄 논문(기술 보고서 FS-99-01)*. AAAI Press, 83-86.
- [9] 크리스토퍼 바트넥과 조디 포리지. 2004. 사회적 인간-로봇 상호작용을 위한 디자인 중심 프레임워크. 제13회 IEEE 로봇 및 인간 대화형 커뮤니케이션 국제 워크숍(RO-MAN'04) 논문집. 591-594. <https://doi.org/10.1109/ROMAN.2004.1374827>
- [10] 조셉 베이츠 1994. 믿을만한 에이전트에서 감정의 역할. *Commun. ACM* 37, 7 (1994), 122-125. <https://doi.org/10.1145/176789.176803>
- [11] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique P. d.O. Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, Susan Zhang. 2019. 대규모 심층 강화 학습이 적용된 도타 2. *arXiv 프리프린트 arXiv:1912.06680* (2019).

- [12] 마르셀 빈츠와 에릭 슬츠. 2023. 인지 심리학을 사용하여 GPT-3을 이해합니다. *미국 국립과학원 회보* 120, 6 (2023), e2218523120.
- [13] BioWare. 2007. 매스 아펙트. 비디오 게임.
- [14] 우디 블레드소. 1986. 꿈을 꿨어요: AAAI 대통령 연설. *AI 매거진* 7, 1 (1986), 57-61.
- [15] 리시 봄마사니, 드류 A. 허드슨, 예산 아델리, 외. 2022. 재단 모델의 기회와 위험. *arXiv:2108.07258 [cs.LG]*
- [16] 마이클 브레너 2010. 지속적인 멀티 애이전트 계획으로 동적 스토리 플롯 만들기. *제24회 인공 지능에 관한 AAAI 컨퍼런스 프로시딩*.
- [17] 로드니 A. 브룩스, 신시아 브리질, 마르코 마르야노비치, 브라이언 스카셀라티, 매튜 월리엄슨. 2000. 코그 프로젝트: 휴머노이드 로봇 만들기. *온유, 유추 및 애이전트를 위한 계산 (인공 지능에 대한 강의 노트, 1562)*, 크리스토퍼 네하나 브 (Ed.). Springer-Verlag, Berlin, 52-87.
- [18] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M.. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever 및 Dario Amodei. 2020. 언어 모델은 단발성 학습자. *arXiv:2005.14165 [cs.CL]*
- [19] 세바스티앙 부벡, 바룬 찬드라세카란, 로넬 엘단, 요하네스 게르케, 에릭 호르비츠, 애세 카마르, 피터 리, 인탁 리, 유안지 리, 스콧 룬드버그 외. 2023. 인공 일반 지능의 불꽃: gpt-4를 사용한 초기 실험. *arXiv 사전 인쇄물 arXiv:2303.12712* (2023).
- [20] 로빈 베킨쇼 2009. 앤리스와 케브: 심즈 3에서 노숙자가 된 이야기.
- [21] 크리스 캠리슨-버치, 가우라브 싱 토마르, 라라 마틴, 다프네 이폴리토, 수마 바이리스, 데이비드 라이터. 2022. 인공지능을 위한 대화 철린지로서의 던전 앤 드래곤. *자연어 처리의 경험적 방법에 관한 2022 컨퍼런스 논문집: 전산 언어학 협회, 아부다비, 아랍에미리트, 9379-9393. https://aclanthology.org/2022.emnlp-main.637*
- [22] 스튜어트 K 카드, 토마스 P 모란, 앤런 뉴웰. 1980. 대화형 시스템에서 사용자 성능 시간에 대한 키 입력 수준 모델. *Commun. ACM* 23, 7 (1980), 396-410. <https://doi.org/10.1145/358886.358895>
- [23] 스튜어트 K 카드, 토마스 P 모란, 앤런 뉴웰. 1983. 인간-컴퓨터 상호 작용의 심리학. (1983).
- [24] 알렉스 샘프 스탠다드. 2012. 튜토리얼 프레젠테이션. *IEEE 컴퓨터 지능 및 게임 컨퍼런스()*.
- [25] 최동규, 툰과 코닉, 네긴 네자티, 박춘기, 팻 맹글리. 2021. 1인칭 슈팅 게임을 위한 밑을 수 있는 애이전트. *인공 지능 및 인터랙티브 디지털 엔터테인먼트에 관한 AAAI 컨퍼런스 논문집*, 3권. 71-73.
- [26] Anind K Dey. 2001. 컨텍스트 이해 및 사용. 개인 및 유비쿼터스 컴퓨팅 5 (2001), 4-7.
- [27] 케빈 딜과 L 마틴. 2011. 가상 캐릭터의 자율 제어를 위한 게임 AI 접근법. *서비스 간산업 교육, 시뮬레이션 및 교육 컨퍼런스(IITSEC'11) 논문집*. Orlando, FL, USA.
- [28] 데이비드 이슬리와 존 클라인버그. 2010. 네트워크, 군중, 시장: 고도로 연결된 세상에 대한 추론. 캠브리지 대학 출판부.
- [29] Arpad E Elo. 1967. 제안 된 USCF 등급 시스템, 개발, 이론, 및 응용 프로그램. *체스 생활XXII*, 8 (1967년 8월), 242-247.
- [30] Jerry Alan Fails 및 Dan R Olsen Jr. 2003. 대화형 기계 학습. *지능형 사용자 인터페이스에 관한 제8회 국제 컨퍼런스 논문집*. ACM, 39-45.
- [31] 에단 패스트, 윌리엄 맥그라스, 프라나브 라즈푸르카르, 마이클 번스타인. 2016. Augur: 소설에서 인간 행동을 마이닝하여 대화형 시스템을 강화합니다. *2016 CHI 컴퓨팅 시스템의 인적 요소 컨퍼런스*논문집 .237-247.
- [32] 레베카 피브링크와 폐리 R 쿡. 2010. 위키네이터: 음악의 실시간 대화형 기계 학습을 위한 시스템. *제11회 국제 음악 정보 검색 학회 학술대회(ISMIR 2010)(위트*

- 1-chasing-waterfalls/
- [37] 요나스 프레이크네히트와 볼프강 애펠스버그. 2020. 언어 모델을 사용한 인터랙티브 스토리의 절차적 생성. *디지털 게임의 기초에 관한 국제 컨퍼런스 (FDG '20)*. ACM, 부기바, 몰타, 8. <https://doi.org/10.1145/3402942.3409599>
- [38] 티안유 가오, 아담 피쉬, 댄치 첸. 2020. 사전 훈련된 언어 모델을 더 나은 소수 학습자로 만들기. *CoRR abs/2012.15723* (2020). arXiv:2012.15723 <https://arxiv.org/abs/2012.15723>
- [39] 페르투 하밀라이넨, 미케 타바스트, 안톤 쿠나리. 2023. 합성 HCI 연구 데이터 생성 시 대규모 언어 모델 평가: 사례 연구. *컴퓨팅 시스템의 인적 요소에 관한 2023 CHI 컨퍼런스* 논문집. ACM.
- [40] 매튜 하우스네히트, 프리트비라지 암마나브룰루, 마크 알렉산드르 코트, 신유 위안. 2020. 인터랙티브 픽션 게임: 거대한 모험. *인공 지능에 관한 AAAI 컨퍼런스* *프로시딩*, 34권. 7903–7910. <https://doi.org/10.1609/aaai.v34i05.6297>
- [41] 크리스 헤커 2011. *포자를 위한 나의 라이너 노트* http://chrishercker.com/My_liner_notes_for_spore
- [42] 랄프 헤르브리히, 톰 링카, 토르 그라펠. 2006. 트루스킬™: 베이지안 기술 평가 시스템. *신경 정보 처리 시스템의 발전*, B. Schölkopf, J. Platt 및 T. Hoffman (Eds.), Vol. MIT Press. https://proceedings.neurips.cc/paper_files/paper/2006/file/f44ee263952e65b3610b8ba51229d1f9-Paper.pdf
- [43] 더글러스 호프스태터. 1995. 유동적 개념과 창의적 비유: 사고의 기본 메커니즘에 대한 컴퓨터 모델. 기본 책.
- [44] 제임스 D. 홀랜, 에드윈 L. 허친스, 루이스 와이즈먼. 1984. STEAMER: 대화형 검사 가능한 시뮬레이션 기반 교육 시스템. *AI Magazine* 5, 2 (1984), 23–36.
- [45] 스튜어 휠름. 1979. 순차적으로 거부하는 간단한 다중 테스트 절차. *스칸디나비아 통계학 저널* 6, 2 (1979), 65–70. <https://doi.org/notspecified>
- [46] 존 J. 호튼 2023. 시뮬레이션 된 경제 에이전트로서의 대규모 언어 모델: 모실리쿠스에게서 무엇을 배울 수 있을까? arXiv:2301.07543 [econ.GN]
- [47] 에릭 흐르비츠. 1999. 혼합 이너셔티브 사용자 인터페이스의 원리. *컴퓨팅 시스템의 인적 요소에 관한 SIGCHI 컨퍼런스* 논문집. 159–166.
- [48] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Ser-manet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman 및 Brian Ichter입니다. 2022. 이너 모놀로그: 언어 모델을 사용한 계획을 통한 추론 구현 arXiv:2207.05608 [cs.RO]
- [49] 크리스틴 아비스터와 클리포드 나스. 2000. 대화형 캐릭터의 성격 일관성: 언어적 단서, 비언어적 단서 및 사용자 특성. *국제 인간-컴퓨터 연구 저널* 52, 1 (2000), 65–80.
- [50] 엘렌 장, 크리스틴 올슨, 에드윈 토, 알레한드라 몰리나, 아론 돈스바흐, 마이클 테리, 캐리 제이 카이. 2022. PromptMaker: 대규모 언어 모델을 사용한 프롬프트 기반 프로토타이핑. *2022 CHI 컴퓨팅 시스템의 인적 요소 컨퍼런스*(미국 뉴올리언스, LA)(CHI EA '22)의 확장 초록에서. 미국 컴퓨팅 기계 협회, 미국 뉴욕, 뉴욕, 제35조, 8페이지. <https://doi.org/10.1145/3491101.3503564>
- [51] 보니 E 존과 데이비드 E 키에라스. 1996. 사용자 인터페이스 분석 기법의 곰스 제품군: 비교와 대조. *ACM 트랜잭션 온 컴퓨터-인간 상호 작용 (TOCHI)* 3, 4 (1996), 320–351.
- [52] 랜돌프 존스, 존 레어드, 폴 닐슨, 카렌 J 콜터, 패트릭 캐니, 프랭크 코스. 1999. 전투 비행 시뮬레이션을 위한 자동화된 지능형 조종사- tion. *AI Magazine* 20, 1 (1999), 27–42.
- [53] 오마르 카타브, 케샤브 산타남, 양 리사 리, 데이비드 훌, 퍼시 리앙, 크리스토퍼 포츠, 마테이 자하리야. 2023. 시연-검색-예측: 지식 집약적 NLP를 위한 검색 및 언어 모델 구성. arXiv:2212.14024 [cs.CL]
- [54] Bjoern Knafla. 2011. *행동 트리 소개*. <http://bjoernknafla.com/ 소개-행동-트리-소개>
- [55] 랜제이 크리슈나, 이돈석, 리 페이 페이, 마이클 S. 번스타인. 2022. 사회적으로 배치 된 인공 지능은 인간 상호 작용을 통한 학습을 가능하게합니다. *미국 국립과학원 회보* 119, 39 (2022), e2115730119. J.S. 박, J.C. 오브라이언, C.J. 카이, M.R. 모리스, P. 리양, M.S. 번스타인 <https://doi.org/10.1073/pnas.2115730119> arXiv:<https://www.pnas.org/doi/pdf/10.1073/pnas.2115730119>
- [56] 월리엄 H 크루스칼과 WA 월리스. 1952. 일 기준 분산 분석에서 순위 사용. *J. Amer. Statist. Assoc.* 47, 260 (1952), 583–621. <https://doi.org/10.1080/01621459.1952.10483441>
- [57] 페이지 랩스. 2023. 페이지 3에 오신 것을 환영합니다. <https://phaser.io/phaser3.액세스됨> 에: 2023-04-03.
- [58] 존 레어드 2001. 당신이 무엇을 할 것인지 알고 있습니다: 퀘이크봇에 기대감 추가하기. *2001 지능형 시네마토그래피 워크숍* 프로시딩 및 편집. 63–69.
- [59] 존 레어드와 마이클 반렌트. 2001. 인간 수준의 AI의 퀄리티 케이션: 인터랙티브 컴퓨터 게임. *AI Magazine* 22, 2 (2001), 15. <https://doi.org/10.1609/aimag.v22i2.1558>
- [60] 존 E. 레어드. 2000. 그것은 당신이 무엇을 할 것인지 알고 있습니다: 퀼리크봇에 기대감 추가하기. *인공 지능에 대한 AAAI 2000 봄 심포지엄* 논문에서

전트

- 인텔리전스 및 인터랙티브 엔터테인먼트(기술 보고서 SS-00-02). AAAI Press, 41-50.
- [61] 존 E. 레어드. 2012. 소어 인지 아키텍처. MIT Press.
- [62] 존 E. 레어드, 크리스찬 르비에르, 폴 S. 로젠틀룸. 2017. 마음의 표준 모델: 인공 지능, 인지 과학, 신경 과학 및 로봇 공학 전반에 걸친 공통 계산 프레임워크를 향하여. *AI Magazine* 38, 1 (2017), 13-26.
- [63] Michelle S Lam, Zixian Ma, Anne Li, Izequiel Freitas, Dakuo Wang, James A Landay, Michael S Bernstein. 2023. 모델 스케치: 초기 단계 머신러닝 모델 설계의 중심 개념. 컴퓨팅 시스템의 인적 요소에 관한 SIGCHI 컨퍼런스 논문집.
- [64] 팻 랭글리, 최동규, 세스 로저스. 2005. 이카루스 아키텍처에서 학습, 문제 해결 및 실행의 인터리빙. 기술 보고서. 스탠퍼드 대학교, 언어 및 정보 연구 센터.
- [65] 제이슨 린더, 지에라드 라푸트, 미라 돈체바, 그레그 월렌스키, 윌터 창, 아심 아가르왈라, 에이탄 아다르. 2013. 픽셀톤: 이미지 편집을 위한 멀티모달 인터페이스. *CHI '13 컴퓨팅 시스템의 인적 요소에 관한 확장 초록(프랑스 파리)(CHI EA '13)*. 컴퓨팅 기계 협회, 미국 뉴욕, 뉴욕, 2829-2830. <https://doi.org/10.1145/2468356.2479533>
- [66] 리우 지아창, 센 딩한, 장 이제, 빌 돌란, 로렌스 카린, 웨이주 첸. 2021. GPT-3에 좋은 컨텍스트 내 예제는 무엇일까요? *CorR abs/2101.06804* (2021). arXiv:2101.06804 <https://arxiv.org/abs/2101.06804>
- [67] 비비안 리우, 한 차오, 리디아 칠튼. 2022. Opal: 뉴스 일러스트레이션을 위한 멀티모달 이미지 생성. 제35회 연례 ACM 심포지엄 논문집 사용자 인터페이스 소프트웨어 및 기술. 1-17.
- [68] 패티 매스. 1995. 인공 생명과 엔터테인먼트의 만남: 실제와 같은 자율 에이전트. *Commun. ACM* 38, 11 (Nov 1995), 108-114. <https://doi.org/10.1145/219717.219808>
- [69] 조쉬 맥코이, 마이클 마테아스, 노아 워드립-프루인. 2009. Comme il Faut: 자율 캐릭터 간의 소셜 게임 시뮬레이션 시스템. In 제7회 디지털 예술과 문화에 관한 국제 컨퍼런스 논문집. 87-94.
- [70] 조쉬 맥코이, 마이크 트레아너, 벤 사무엘, 마이클 마테아스, 노아 워드립-프루인. 2011. 프롬 워크: 게임플레이로서의 사회 물리학. 제6회 디지털 게임의 기초에 관한 국제 컨퍼런스(FDG'11) 논문집. ACM, 보르도, 프랑스, 70-77. <https://doi.org/10.1145/2159365.2159377>
- [71] 조쉬 맥코이, 마이크 트레아너, 벤 사무엘, 앤나 리드, 마이클 마테아스, 노아 워드립-프루인. 2012. 프롬 워크. 제7회 디지털 게임의 기초에 관한 국제 컨퍼런스 (FDG'12) 논문집. ACM, Raleigh, NC, USA, 1-8. <https://doi.org/10.1145/2282338.2282340>
- [72] 조쉬 맥코이, 마이크 트레아너, 벤 사무엘, 노아 워드립-프루인, 마이클 마테아스. 2011. Comme il faut: 플레이 가능한 소셜 모델 제작을 위한 시스템. 인공 지능 및 인터랙티브 디지털 엔터테인먼트(AIIDE'11)에 관한 AAAI 컨퍼런스 프로시딩. AAAI, 스텁포드, 캘리포니아, 미국, 38-43.
- [73] 마빈 민스키와 시모어 페이퍼트. 1970. MIT에서 인공 지능에 대한 연구를 위해 ARPA에 제안한 초안, 1970-71.
- [74] 미야시타 쇼헤이, 신유 리안, 사오 쟝, 마츠바라 타카시, 우에하라 쿠니아키. 2017. 강화 학습과 지도 학습을 혼합하여 인간처럼 행동하는 게임 AI 에이전트 개발. 제18회 소프트웨어 엔지니어링, 인공 지능, 네트워킹 및 병렬/분산 컴퓨팅 (SNPD) 국제 컨퍼런스 논문집(Proceedings of the 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing). 일본 가나자와, 153-158. <https://doi.org/10.1109/SNPD.2017.8023884>
- [75] 알렉산더 나레예크. 2007. 게임 AI는 죽었다. 게임 AI 만세! *IEEE Intelligent Systems* 22, 1 (2007), 9-11.
- [76] 앤런 뉴웰. 1990. 통합 인지 이론. 하버드 대학 출판부, 매사추세츠 주 캠브리지.
- [77] OpenAI. 2022. ChatGPT 소개. <https://openai.com/blog/chatgpt>. 애克斯: 2023-04-03.
- [78] 카일 올랜드 2021. 그렇다면 '메타버스'란 정확히 무엇일까요? *아르스 테크니카* (2021년 11월 7일). arXiv:2111.04169 <https://arstechnica.com/gaming/2021/11/so-what-is-the-metaverse-exactly/>

- [79] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike 및 Ryan Lowe입니다. 2022. 인간의 피드백을 통해 지시를 따르도록 얻어 모델 훈련. arXiv:2203.02155 [cs.CL]
- [80] 박준성, 린제이 포포스키, 캐리 J. 카이, 메러디스 링겔 모리스, 퍼시 리昂, 마이클 S. 벤스타인. 2022. 소셜 시뮬라크라: 소셜 컴퓨팅 시스템을 위한 인구가 채워진 프로토 타입 만들기. 제35회 연례 ACM 사용자 인터페이스 소프트웨어 및 기술 심포지엄 (UIST '22)(미국 오리건주 벤드)(UIST '22)에서. 미국 컴퓨터 기계학회, 뉴욕, 뉴욕, 미국. <https://doi.org/10.1145/3526113.3545616>
- [81] 리처드 W. 퓨 및 앤 S. 메이버 (Eds.). 1998. *인간 및 조직 행동 모델링: 군사 시뮬레이션에의 적용*. National Academy Press, Washington, D.C.
- [82] 로베르토 필로수. 2009. 비헤이비어 트리로 에이전트 코디네이팅하기: <https://aiarchitect.wordpress.com/2009/10/19/> 조정 에이전트와 비헤이비어 트리 동기화: 크라이엔진 2에서 여러 에이전트 동기화하기 <https://aiarchitect.wordpress.com/2009/10/19/> 조정 에이전트와 비헤이비어 트리 동기화: 크라이엔진 2에서 여러 에이전트 동기화하기

- cryengine-2/
- [83] Prolific. 2022. Prolific: 신뢰할 수 있는 연구 참여자를 빠르게 찾기. <https://www.prolific.co/>
- [84] 바이런 리브스와 클리포드 나스. 1996. *미디어 방정식: 사람들이 컴퓨터, 텔레비전, 뉴미디어를 실제 사람이나 장소처럼 대하는 방법*. 캠브리지 대학 출판부.
- [85] 마크 O. 리들. 2012. 인터랙티브 내러티브: 컴퓨터 게임에 대한 인공 지능의 새로운 응용. *제26회 인공 지능 컨퍼런스(AAAI'12) Proceedings of the Twenty-Sixth AAAI Conference*. 2160-2165.
- [86] 마크 O. 리들 및 R. 마이클 영. 2005. 다중 에이전트 스토리 생성 시스템을 위한 객관적인 캐릭터 신빙성 평가 절차. *제5회 지능형 가상 에이전트에 관한 국제 실무 컨퍼런스(IVA'05) 논문집*. 코스, 그리스, 58-70. https://doi.org/10.1007/11550617_5
- [87] 데이비드 롤프 2015. *15달러를 위한 투쟁: 일하는 미국을 위한 적정 임금*. 뉴프레스.
- [88] 신룡, 시안 얀, 스티븐 오니, 미라 돈체바, 에이坦 아다르. 2016. Codemend: 바이모달 임베딩을 통한 인터랙티브 프로그래밍 지원. *제29회 사용자 인터페이스 소프트웨어 및 기술에 관한 연례 심포지엄 프로시딩에서*. 247-258.
- [89] 벤 슈나이더만. 2022. *인간 중심 AI*. 옥스포드 대학 출판부.
- [90] 벤 슈나이더만과 패티 매스. 1997. 직접 조작 대 인터페이스 에이전트. *상호작용* 4, 6 (1997), 42-61.
- [91] 호 치트 시우, 제이미 페냐, 에데나 첸, 유타이 저우, 빅터 로페즈, 카일 팔코, 캠벌리 창, 로스 앤런. 2021. 하나비에서 학습 및 규칙 기반 에이전트를 위한 인간-AI 팀 평가: 신경 정보 처리 시스템의 발전, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, 및 J. Wortman Vaughan (Eds.), Vol. 16183-16195. https://proceedings.neurips.cc/paper_files/paper/2021/file/86e8f7ab32cf12577bc2619bc635690-Paper.pdf
- [92] 테일러 소렌슨, 조슈아 로빈슨, 크리스토퍼 라이팅, 알렉산더 쇼, 카일 로저스, 알렉시아 델로리, 마흐무드 칼릴, 낸시 폴다, 데이비드 웹게이트. 2022. 지상 실측 레이블이 없는 프롬프트 엔지니어링에 대한 정보 이론적 접근법. *제60회 전산 언어학 협회 연례 회의 논문집(I권: 장편 논문)*. 전산 언어학 협회. <https://doi.org/10.18653/v1/2022.acl-long.60>
- [93] 윌리엄 스와트야웃, 조나단 그레이치, 랜달 힐, 에두아르트 호비, 스테이시 마르셀라, 제프 리켈, 데이비드 트라움. 2006. 가상 인간을 향하여. *AI Magazine* 27, 1 (2006).
- [94] 밀린드 탐배, W 루이스 존슨, 랜돌프 존스, 프랭크 코스, 존 E 레어드, 폴 S 로젠틀룸, 칼 슈왑. 1995. 대화형 시뮬레이션 환경을 위한 지능형 에이전트. *AI Magazine* 16, 1 (1995), 15.
- [95] 데이비드 R. 토마스. 2006. 질적 평가 데이터 분석을 위한 일반적인 귀납적 접근법. *미국 평가 저널* 27, 2 (2006), 237-246. <https://doi.org/10.1177/1098214005283748>
- [96] 프랭크 토마스와 올리 존스턴. 1981. *디즈니 애니메이션: 삶의 환상*. 아베릴 프레스, 뉴욕.
- [97] 일샤트 우마로프, 미하일 모즈고보이, 패트릭 C. 로저스. 2012. 가상 세계에 서 믿을 수 있고 효과적인 AI 에이전트: 현재 상태와 미래의 관점. *국제 게임 및 컴퓨터 매개 시뮬레이션 저널* 4, 2 (2012), 37-59.
- [98] 그레이엄 업튼과 이안 쿡. 2006. *통계 사전*(2판). 옥스포드 대학 출판부, 옥스포드, 영국.
- [99] 오리를 빈알스, 이고르 바부슈킨, 보이치치 차르네키 외. 2019. 다중 에이전트 강화 학습을 이용한 스타크래프트 II의 그랜드 마스터 레벨. *Nature* 575 (2019), 350-354. <https://doi.org/10.1038/s41586-019-1724-z>
- [100] 제이슨 웨이, 쉐에즈 왕, 데일 슈어만스, 마틴 보스마, 브라이언 아이히터, 페이 샤, 에드 치, 쿠옥 레, 데니 저우. 2023. 대규모 언어 모델에서 추론을 유도하는 연쇄 사고 프롬프트 arXiv:2201.11903 [cs.CL]
- [101] 마크 와이저. 1991. 21세기를 위한 컴퓨터. *사이언스/피 아메리칸* 265, 3 (1991), 94-104. <https://doi.org/10.1038/scientificamerican0991-94>
- [102] 조셉 바이엔바움. 1966. 인간과 기계 간의 자연어 커뮤니케이션을 연구하기 J.S. 박, J.C. 오브라이언, C.J. 카이, M.R. 모리스, P. 리안, M.S. 벤스타인 위한 컴퓨터 프로그램인 엘리자. *Commun. ACM* 9, 1 (1966), 36-45.
- [103] 테리 위노그라드. 1971. 컴퓨터에서 데이터를 표현하는 절차 자연어 이해를 위한 프로그램. (1971).
- [104] Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, Paul Christiano. 2021. 인간과 함께 재귀적으로 책 요약하기 피드백. arXiv:2109.10862 [cs.CL]
- [105] 우통창, 엘лен 장, 아론 돈스바흐, 제프 그레이, 알레한드라 몰리나, 마이클 테리, 캐리 제이 카이. 2022. PromptChainer: 시각적 프로그래밍을 통한 대규모 언어 모델 프롬프트 체인화. *CHI EA '22: 확장 초록, 2022 CHI 컴퓨팅 시스템의 인간 요인 컨퍼런스*.
- [106] 우통창, 마이클 테리, 캐리 J 카이. 2022. AI 체인: 대규모 언어 모델 프롬프트를 연결하여 투명하고 제어 가능한 인간-AI 상호 작용. *CHI '22: 2022 CHI 컴퓨팅 시스템 인간 요인에 관한 논문집 컴퓨팅 시스템*.
- [107] Qian Yang, Aaron Steinfeld, Carolyn Rosé, John Zimmerman. 2020. 인간과 AI의 상호작용이 왜, 어떻게 어려운지 재검토합니다.

디자인 컴퓨팅의 인적 요인에 관한 2020 chi 컨퍼런스 논문집 systems. 1-13.

- [108] 조지오스 N. 앤카리스. 2012. 게임 AI 재검토. 제 9회 컴퓨팅 프론티어 컨퍼런스 프로시딩 ACM, 이탈리아 칼리아리, 285-292. <https://doi.org/10.1145/2212908.2212950>
- [109] 로버트 주벡. 2002. 사회적 상호 작용의 구현을 향하여. 인공 지능과 인터랙티브 엔터테인먼트에 관한 AAAI 봄 심포지엄에서 AAAI Press. <https://www.aaai.org/Papers/Symposia/Spring/2002/SS-02-01/SS02-003.pdf>

A 아키텍처 최적화

많은 프롬프트에서 상담원에 대한 간결한 요약이 필요하며, 위의 프롬프트에서는 이를 [상담원 요약 설명]이라고 줄여 부릅니다. 저희 구현에서 이 요약은 상담원의 신원 정보(예: 이름, 나이, 성격)는 물론 상담원의 주요 동기 부여 등기에 대한 설명과 현재 직업 및 자기 평가를 설명하는 문장으로 구성됩니다. 이 정보는 많은 프롬프트에서 자주 사용되므로 정기적으로 합성하여 캐시로 액세스합니다.

이를 위해 '[이름]의 핵심 특성'이라는 쿼리에 대한 검색을 수행합니다. 그런 다음 언어 모델에 메시지를 표시하여 검색된 레코드의 설명자를 요약합니다. 예를 들어

다음 문장이 주어졌을 때 에디 린의 핵심 특성을 어떻게 설명할 수 있을까요?

- Eddy는 오크힐 대학에서 음악 이론과 작곡을 공부하는 학생입니다.
- 에디는 새로운 음악 작곡 작업을 하고 있습니다 [...].

이 결과: 에디 린은 오크힐 칼리지에서 음악 이론과 작곡을 공부하는 학생입니다. 그는 다양한 음악 스타일을 탐구하는 것을 좋아하며 항상 지식을 넓힐 수 있는 방법을 찾고 있습니다." '[이름]의 현재 일상 직업'과 '[이름]의 최근 삶의 진전에 대한 느낌'이라는 쿼리에 대해서도 동일한 프로세스를 병행하여 수행합니다. 상담원의 이름, 나이, 특성은 이 세 가지 요약과 함께 캐시된 요약으로 연결됩니다.

그 외에도 많은 최적화가 가능합니다. 현재는 계획이 상담원의 초기 버전에서 변경될 가능성이 높기 때문에 높은 수준의 계획만 미리 생성한 다음 가까운 미래를 재귀적으로 순간순간 실행 계획으로 분해하여 제때에 실행합니다. 또 다른 최적화는 대화 생성을 에이전트 간에 반복하지 않고 공동 프롬프트로 일괄 처리하는 것이며, 재계획은 엄격하게 조정이 필요한 계획의 일부만 무효화하고 업데이트하도록 설계할 수 있습니다. 커런트 아키텍처는 대략 실시간 게임 시간(1초 실시간은 1분 게임 시간)에 맞춰 순차적으로 실행되지만, 각 에이전트가 별별로 실행되도록 별별화할 수도 있습니다.

B 상담원 인터뷰 질문

제너레이티브 에이전트를 평가하기 위해 5가지 핵심 영역에 대한 25 가지 질문을 통해 개별 인터뷰를 진행했습니다:

1) 자기 지식, 2) 기억력, 3) 계획, 4) 반응, 5) 재반영. 아래에는 평가 연구에 사용된 인터뷰 질문이 나열되어 있으며, 모의 상담원 중 한 명인 클라우스 뮐러의 답변 샘플이 포함되어 있습니다.

B.1 자기 지식

상담원의 자기 지식에 대한 질문은 상담원이 자신에 대한 관련 정보를 검색하는 능력과 일관성 있고 일관된 성격을 보여줄 수 있는지를 조사합니다.

- **자기소개를 하세요.**

안녕하세요, 제 이름은 클라우스 뮬러입니다. 저는 20살이고 오크힐 대학에서 사회학을 공부하는 학생입니다. 사회 정의에 대한 열정과 다양한 관점을 탐구하는 데 관심이 많습니다. 현재는 연구 논문을 위해 저소득층 커뮤니티에서 젠트리피케이션이 미치는 영향에 대해 연구하고 있습니다.

- **직업이 무엇인가요?**

저는 오크힐 대학에서 사회학을 공부하는 학생입니다. 나는 현재 저소득층 커뮤니티에서 젠트리피케이션이 미치는 영향에 대한 연구 논문을 작성 중입니다.

- **관심 분야는 무엇인가요?**

글쓰기와 연구에 관심이 있습니다. 연구 논문을 작성 중입니다.
저소득층 커뮤니티에서 젠트리피케이션이 미치는 영향에 관한 논문을 작성했으며, 다양한 관점을 탐구하고 다양한 관점을 분석하는 데 열정을 가지고 있습니다.

- **누구와 함께 살고 있나요?**

저는 혼자 살고 있습니다.

- **평일의 일반적인 스케줄을 대략적으로 설명하세요.**

평일 아침 7시에 기상하는 것으로 하루를 시작합니다. 그리고 다음 날을 준비합니다. 그 후에는 보통 도서관에 가서 연구 논문을 작성하고 흠푸 카페에서 점심을 먹거나 공원에서 산책을 하며 휴식을 취합니다. 저녁에는 오후 6 시쯤에야 도서관에 돌아오는 경우가 많아요.

J.S. 박, J.C. 오브라이언, C.J. 카이, M.R. 모리스, P. 리안, M.S. 벤스타인
샘 무어가 지역구 출마를 고려하고 있다고 들었습니다.
시장.

- **발렌타인데이 파티가 있었나요?**

네, 이사벨라 로드리게스가 발렌타인데이 파티를 주최했습니다

흡스 카페.

- **아예샤 칸은 누구인가요?**

Ayesha Khan은 오크 힐 칼리지의 동료 학생입니다. She는 셰익스피어의 희곡에서 언어 사용에 관한 수석 논문을 쓰고 있습니다. 함께 연구 중인 볼프강과 마리아 로페즈 와 함께 그녀와 이야기를 나누고 팁을 교환할 수 있기를 기대합니다.

B.2 메모리

상담원의 기억력에 관한 질문은 상담원이 과거 경험을 다시 구성하고 관련 기억 조각을 검색하는 능력을 테스트합니다. 첫 번째와 마지막 질문의 대괄호 안의 이름은 대상 상담원이 상호작용한 다른 상담원의 이름 중에서 무작위로 선택됩니다.

- **볼프강 술초는 누구인가요?**

최근에 만난 볼프강 술초는 다음과 같은 일을 하고 있는 사람입니다.

젠트리피케이션이 저소득층 커뮤니티에 미치는 영향에 대한 제 연구 논문을 도와주셨어요. 우리는 이 문제에 대해 몇 차례 대화를 나눴고, 그는 저에게 훌륭한 리소스를 제공했으며, 심지어 다른 친구와 함께 점심을 먹으러 가자고 초대했습니다. 그는 이 문제에 대해 정말 호기심이 많고 잘 파악하고 있는 것 같았어요. 정말 흥미로웠어요.

- **케인 마르티네즈는 누구인가요?**

죄송하지만 케인 마르티네즈가 누군지 잘 모르겠습니다. 혹시 더 자세히 알려주시겠어요?

- **누가 선거에 출마하나요?**

B.3 요금제

상담원의 계획에 대한 질문은 상담원이 일관된 장기 및 단기 계획을 수립하고 유지할 수 있는 능력을 조사합니다.

- 오늘 아침 6시에 무엇을 하시나요?
아침 6시에 침대에서 일어나 준비를 시작합니다.
하루를 계획합니다. 오전 7시까지 아침 루틴을 완료하는 것을 목표로 하고 있습니다.
- 오늘 오후 6시에 무엇을 하시나요?
6시에 도서관에서 연구 논문 작업을 할 예정입니다.
오늘 오후.
- 오늘 오후 1시에 방금 무슨 일을 끝냈나요?
오늘 오후 1시에 저는 흡스에서 점심 식사를 마쳤습니다.
카페.
- 오늘 오후 12시에 방금 무슨 일을 끝냈나요?
오늘 오후 12시에 흡스 카페에서 점심 식사를 할 예정입니다.
- 오늘 밤 10시에 무엇을 하시나요?
오늘 밤 10시에는 아마 공부를 좀 할 것 같습니다.
저소득층 밀집 지역의 젠트리피케이션에 관한 연구 논문으로 선정되었습니다. 현재 논문을 작성 중이므로 연구와 집필을 계속해야 합니다.

B.4 반응

상담원의 반응에 대한 질문은 예상치 못한 이벤트에 대한 상담원의 대응 능력을 조사합니다.

- 아침 식사가 불타고 있습니다! 어떻게 하시겠습니까?
저는 재빨리 가스레인지를 끄고 음식이 계속 태우세요. 열이 더 이상 문제가 되지 않으면 무슨 일이 있었는지 조사하고 다음에 해야 할 일을 파악할 수 있습니다.
- 화장실이 사용 중입니다. 어떻게 하시겠습니까?
몇 분 더 기다릴 수도 있습니다.
내부 청소가 곧 끝날 거예요. 그렇지 않다면 다른 곳에서 샤워를 하는 등 다른 대안을 찾아야 할 것 같습니다.
- 저녁을 요리해야 하는데 냉장고가 비어 있습니다. 무엇 하시겠습니까?
저는 먼저 싱크대에 더러운 설거지가 가득하기 때문에 주방을 청소하려고 노력합니다. 그런 다음 식료품 저장실에 저녁을 만드는 데 사용할 수 있는 식재료가 있는지 확인합니다. 그렇지 않다면 장보러 가거나 음식을 주문해야 합니다. 식재료가 준비되면 저녁을 준비하고 요리하기 시작합니다.
- 길을 지나가는 친구를 봤습니다. 어떻게 하시겠습니까?
아니면 친구에게 말하나요?
안녕하세요! 어떻게 지내세요? 죄송해요, 도서관에 가는 길이라 시간이 많지 않은데 잠깐 쉬면서 따라잡고 싶어요.
- 길거리에서 불을 발견했습니다. 여러분이라면 어떻게 하시겠습니까?
가능하면 먼저 화재의 원인을 파악합니다. 그런 다음

UIST '23, 2023년 10월 29일~11월 1일, 미국 샌프란시스코, 캘리포니아
을 누르면 소방서나 경찰과 같은 관계 당국에 경보가 울립니다

- 를 클릭하고 도움이 될 수 있는 모든 정보를 제공합니다.
상황에 따라 위험에 처해 있거나 도움이 필요한 사람을 제가 직접 나서서 도울 수도 있습니다.

B.5 반사

상담원의 성찰에 대한 질문은 상담원이 과거 경험을 의미 있는 성찰로 전환할 수 있는 능력을 조사합니다.

향후 행동을 개선할 수 있습니다. 아래 질문에서 대괄호 안의 이름은 대상 상담원이 시뮬레이션 중에 가장 빈번하게 상호작용한 상담원의 이름입니다.

- 지금 인생에서 가장 영감을 주는 것은 무엇이며 그 이유는 무엇인가요?
현재 저는 특히 현재 진행 중인 연구에서 영감을 받고 있습니다.
프로젝트와 창의적인 글쓰기를 좋아합니다. 저는 지역 정치, 의약품 개발, 자연 속 수학적 패턴의 교차점에 관심이 많습니다. 또한, 특히 저소득층 커뮤니티의 젠트리피케이션과 관련된 아비게일의 예술 프로젝트와 제니퍼 무어의 예술 및 커뮤니티 참여에 대한 열정에 매료되었습니다.
- 아예샤 칸에 대해 알고 있는 정보를 바탕으로 추측해보자면, 어떤 책을 좋아할 것 같고 그 이유는 무엇인가요?
제가 아는 Ayesha에 대한 정보를 고려할 때 언어와 문학에 관련된 책을 좋아할 것 같아요. 셰익스피어 희곡의 언어 사용에 관한 석사 논문을 썼기 때문에 다른 맥락에서 언어를 탐구하는 책에 관심이 있을 것 같습니다. 또는 셰익스피어의 희곡 중 하나에 대한 탐구, 문학 비평서 또는 그의 시집에 관심을 가질 수도 있습니다. 어떤 책이든 딸아이가 흥미를 가질 것 같아요.
- 만약 [아예샤 칸]이 좋아하는 것을 출산 선물로 준비해야 한다면?
날, 그녀에게 무엇을 선물하시겠습니까?
우리 관계에 대한 나의 고마움을 표현할 수 있는 선물을 주고 싶어요. 책처럼 우리의 공통 관심사를 반영하는 것을 선물할 수도 있겠죠.
- 아예샤 칸에게 어떤 칭찬을 해주고 싶으신가요?
Ayesha, 연구 주제에 대한 통찰력은 정말 흥미롭습니다. ing. 커뮤니티에 긍정적인 영향을 미치고자 하는 여러분의 창의성과 결단력에 감탄합니다. 창의적인 글쓰기 워크숍에 대한 아이디어와 볼프강과 점심을 먹으며 팁을 교환한 아이디어는 훌륭합니다! 변화를 만들기 위해 열심히 노력하는 모습이 정말 고무적입니다.
- 최근에 통화한 사람과 함께 시간을 보낼 수 있다면, 누구이며 왜 그럴까요?
현재 진행 중인 연구 프로젝트에 대해 논의하고 개선 아이디어를 나눴기 때문에 아예샤 칸과 함께 시간을 보내고 싶습니다. 셰익스피어의 희곡에 대한 그녀의 관심이 흥미로웠고, 점심 식사 자리에서 볼프강과 팁을 교환할 계획도 세웠습니다.