Generative Agents: Interactive Simulacra of Human Behavior

User Interface Software and Technology, 2023

논문 소개

Generative Agents: Interactive Simulacra of Human Behavior

Joon Sung Park
Stanford University
Stanford, USA
joonspk@stanford.edu

Joseph C. O'Brien
Stanford University
Stanford, USA
jobrien3@stanford.edu

Carrie J. Cai
Google Research
Mountain View, CA, USA
cjcai@google.com

Meredith Ringel Morris
Google DeepMind
Seattle, WA, USA
merrie@google.com

Percy Liang
Stanford University
Stanford, USA
pliang@cs.stanford.edu

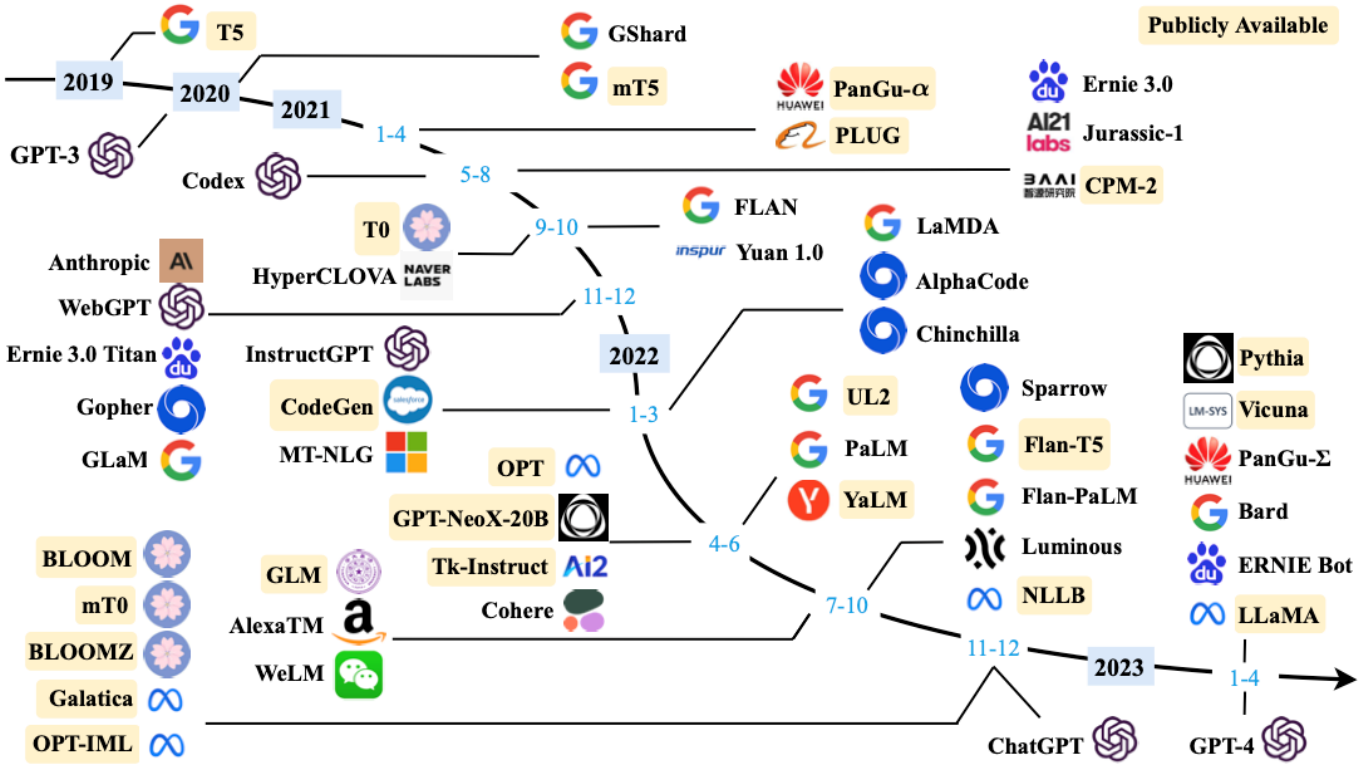
Michael S. Bernstein
Stanford University
Stanford, USA
msb@cs.stanford.edu

Abstract

- GPT-3.5 로 생성한 Generative Agent 25개
- 대화형 샌드박스 환경으로 마을(스몰빌) 형성
- Agent끼리 자연어 사용하여 상호 작용

Related Work

1. LLM (=Large Language Model)



▲ LLM timeline (A Survey of Large Language Models)

- 대량의 텍스트 데이터를 학습하여 인간의 언어 이해 및 생성 능력을 모방하는 인공지능 모델.
- 주로 딥러닝 및 자연어 처리 기술을 기반으로 함.
- 대화 생성, 기계 번역, 질문 응답, 텍스트 요약 등 다양한 자연어 처리 작업에 적용됨.

Related Work

2. Sandbox

: 일반적으로 미리 결정된 목표 없이 또는 플레이어가 스스로 설정한 목표를 통해,
플레이어에게 상호 작용할 수 있는 높은 수준의 창의성을 제공하는 게임 플레이 요소가 포함된 비디오 게임

▲ SIMS

▲ 동물의 숲

▲ 마인크래프트

Behavior and Interaction

1. Agent Avatar

+

2. Environment

Behavior and Interaction

1. Agent Avatar

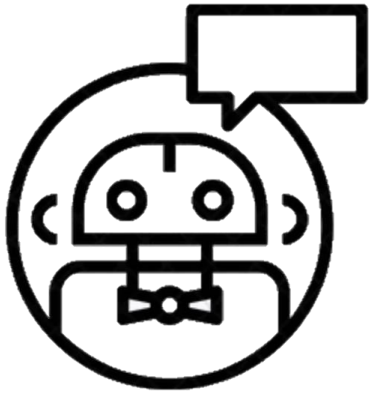
+

2. Environment

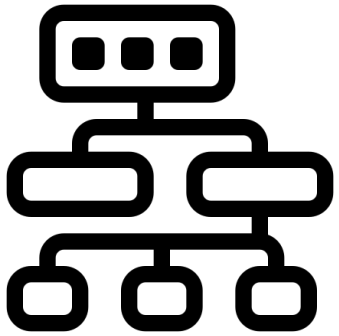
3. Day in the Life

4. Social Behaviors

Contribution



에이전트의 경험과 환경에
따라 동적으로 조절되는
Generative Agent



Generative Agent가
동적으로 변화하는 상황을
조절할 수 있게 해주는
새로운 Architecture

아키텍처 구성 요소 중요성
확인 및 에러를 식별하는
2가지 평가 수행

Behavior and Interaction

1. Agent Avatar

이름 직업

John Lin is a pharmacy shopkeeper at the Willow Market and Pharmacy who loves to help people. He is always looking for ways to make the process of getting medication easier for his customers; John Lin is living with his wife, Mei Lin, who is a college professor, and son, Eddy Lin, who is a student studying music theory; John Lin loves his family very much; John Lin has known the old couple next-door, Sam Moore and Jennifer Moore, for a few years; John Lin thinks Sam Moore is a kind and nice man; John Lin knows his neighbor, Yuriko Yamamoto, well; John Lin knows of his neighbors, Tamara Taylor and Carmen Ortiz, but has not met them before; John Lin and Tom Moreno are colleagues at The Willows Market and Pharmacy; John Lin and Tom Moreno are friends and like to discuss local politics together; John Lin knows the Moreno family somewhat well — the husband Tom Moreno and the wife Jane Moreno.

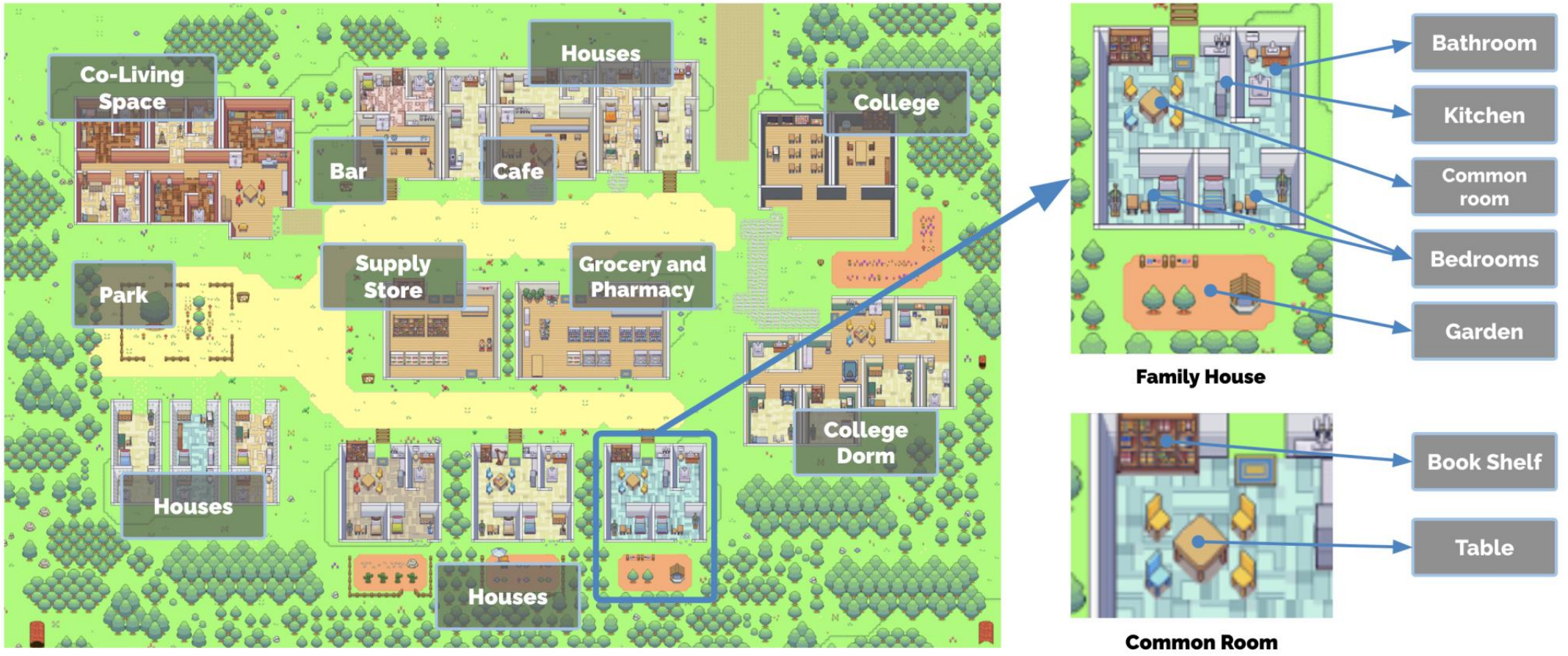
성격

가족 관계

인간 관계

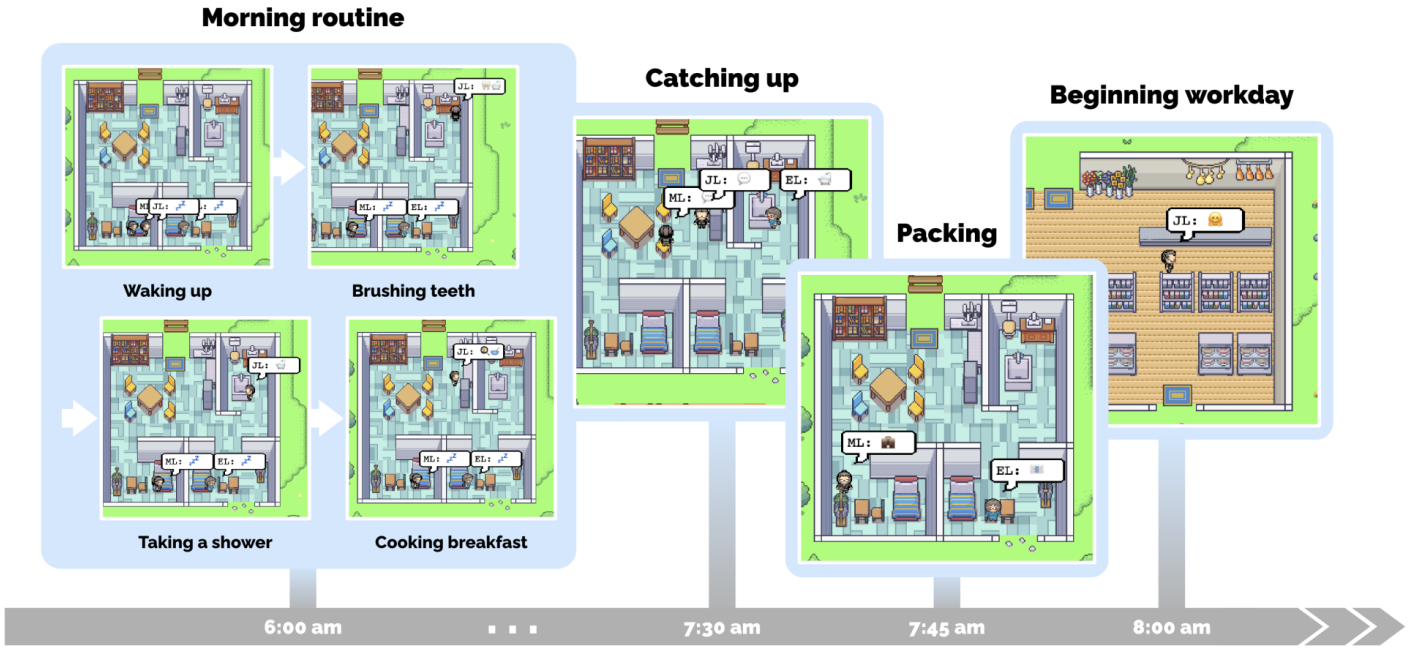
Behavior and Interaction

2. Environment



Behavior and Interaction

3. Day in the Life Example



▲ John Lin 의 하루 일과

▼ John Lin의
기상 직후, 아침 루틴

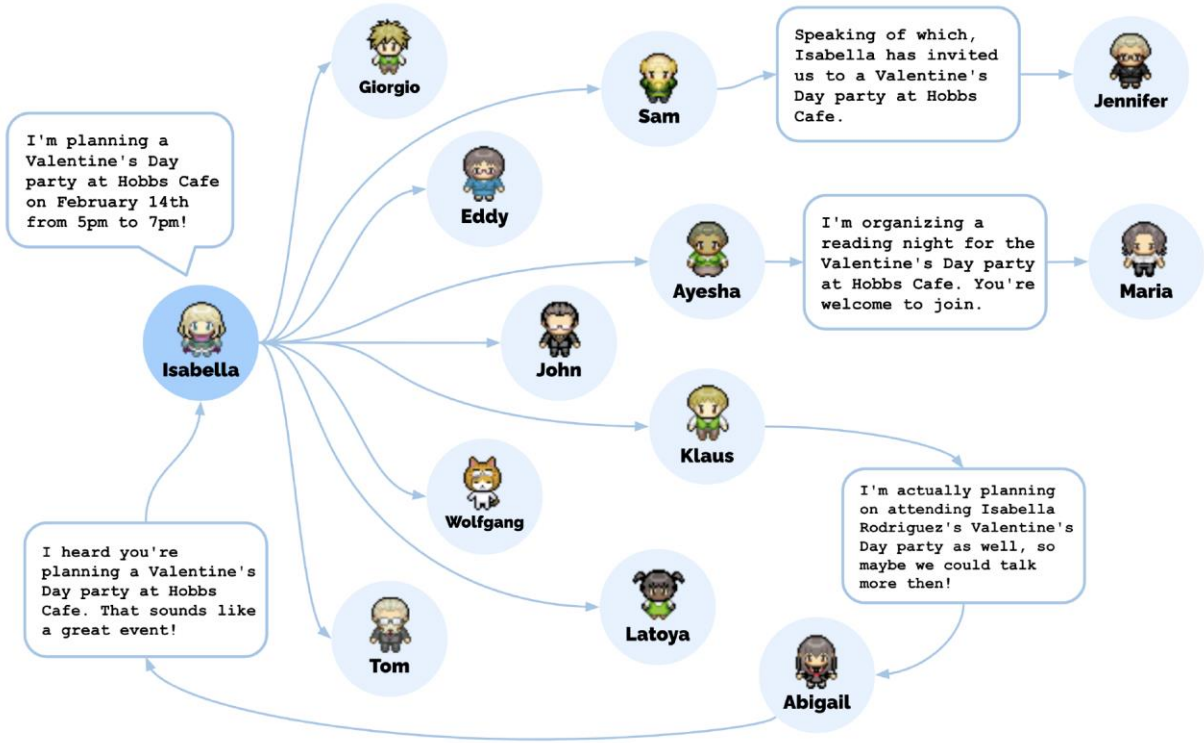
Behavior and Interaction

4. Social Behaviors

1) Information Diffusion

2) Coordination

▲ Isabella의 파티에 참여한 다른 Agent들



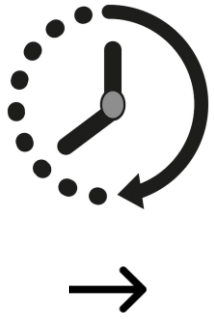
▲ Isabella의 파티 정보가 확산되는 과정

Behavior and Interaction

4. Social Behaviors

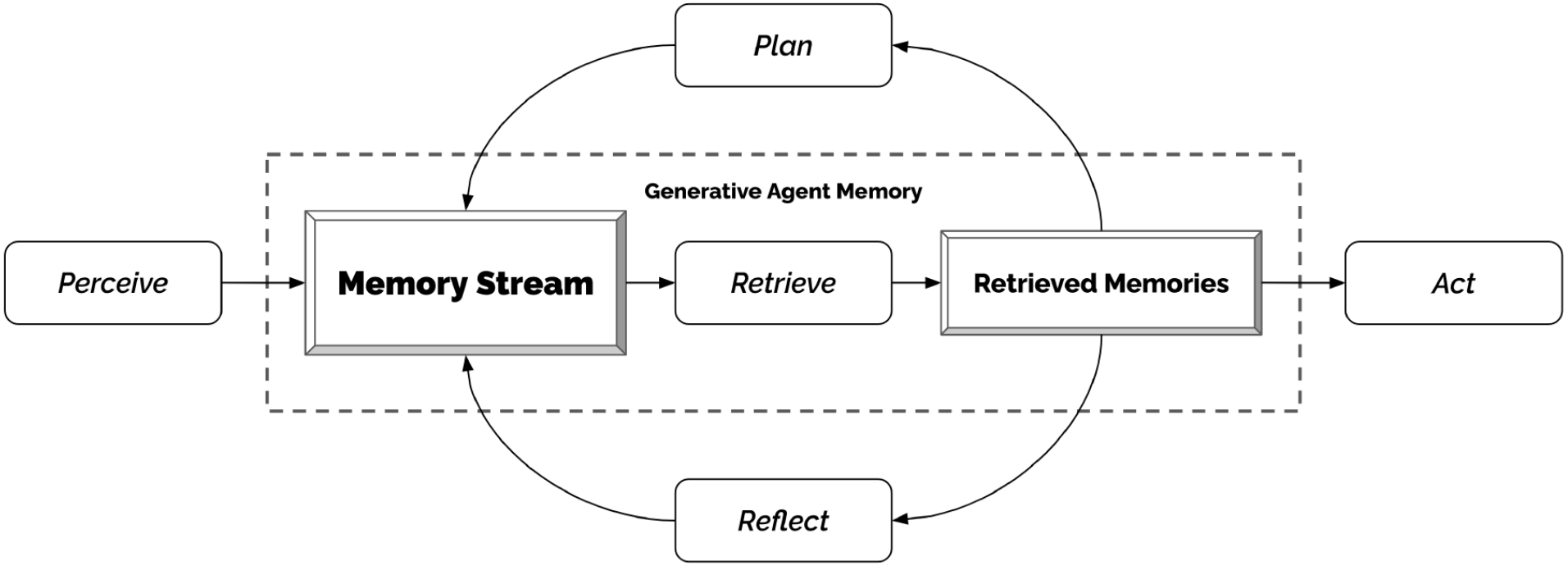
3) Relationship Memory

초면, 인사 나눔

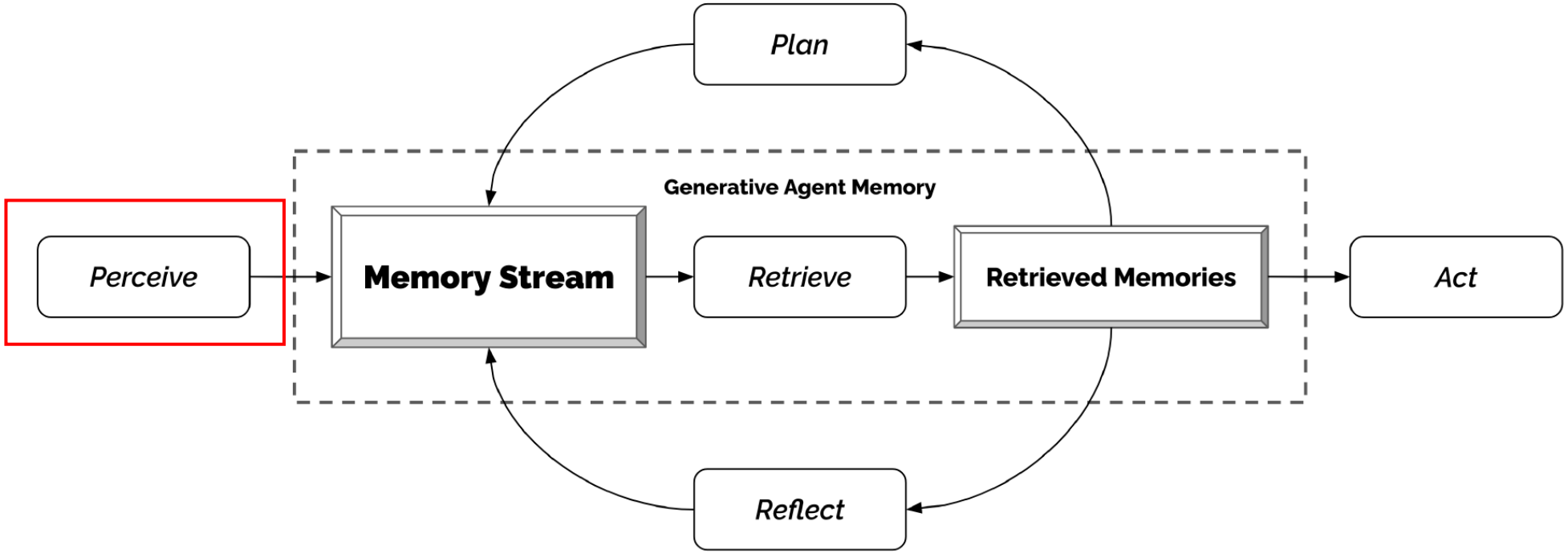


서로 구면임을 인지
서로의 이름 및 정보 기억

Generative Agent Architecture



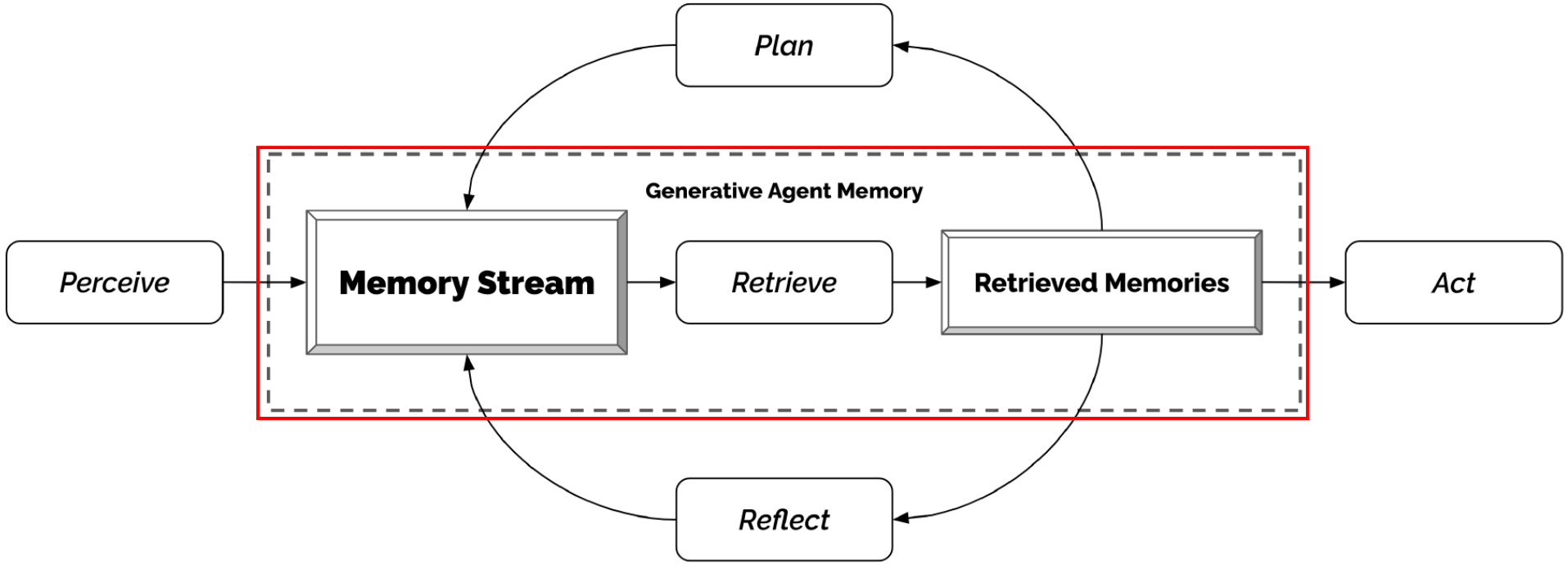
Generative Agent Architecture



1. Agent의 주변 환경 인식

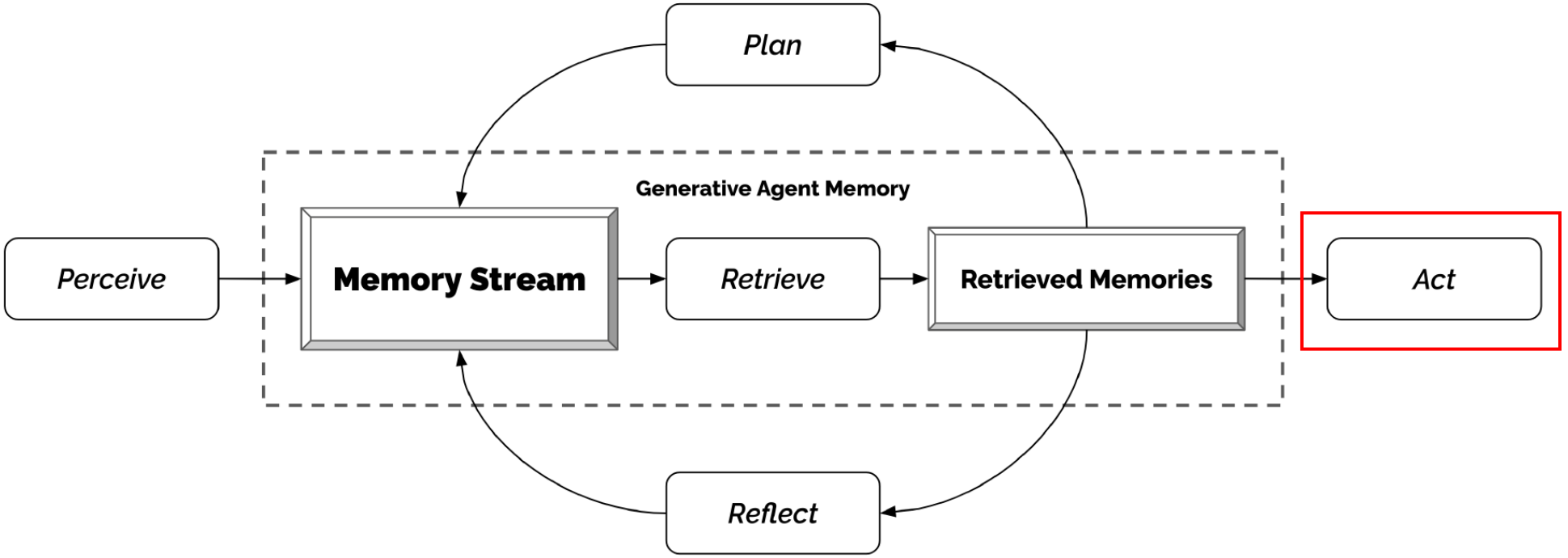
(ex. 다른 Agent와의 대화, 주변 환경 등)

Generative Agent Architecture



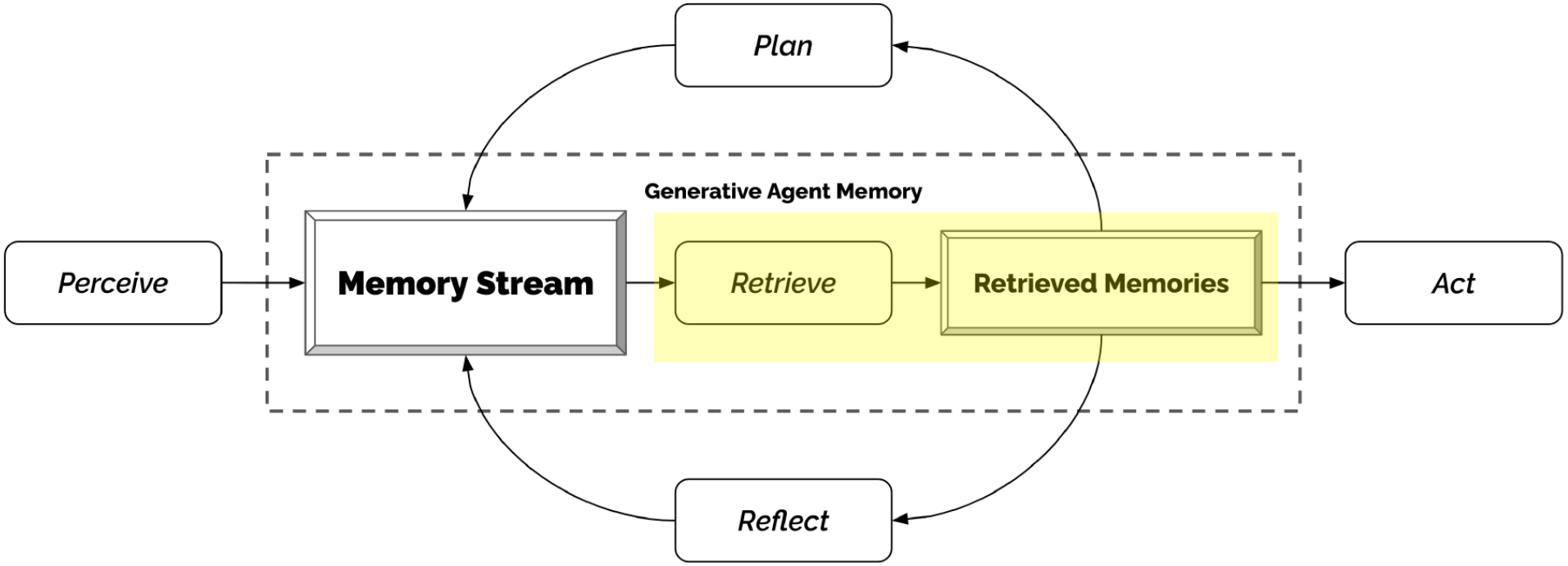
2. Agent의 모든 기억들이 시간순으로 자연어 형태로 저장되어 있는 Memory Stream을 사용하여,
GPT에게 Agent의 다음 행동을 예측해달라는 프롬프트 전송

Generative Agent Architecture



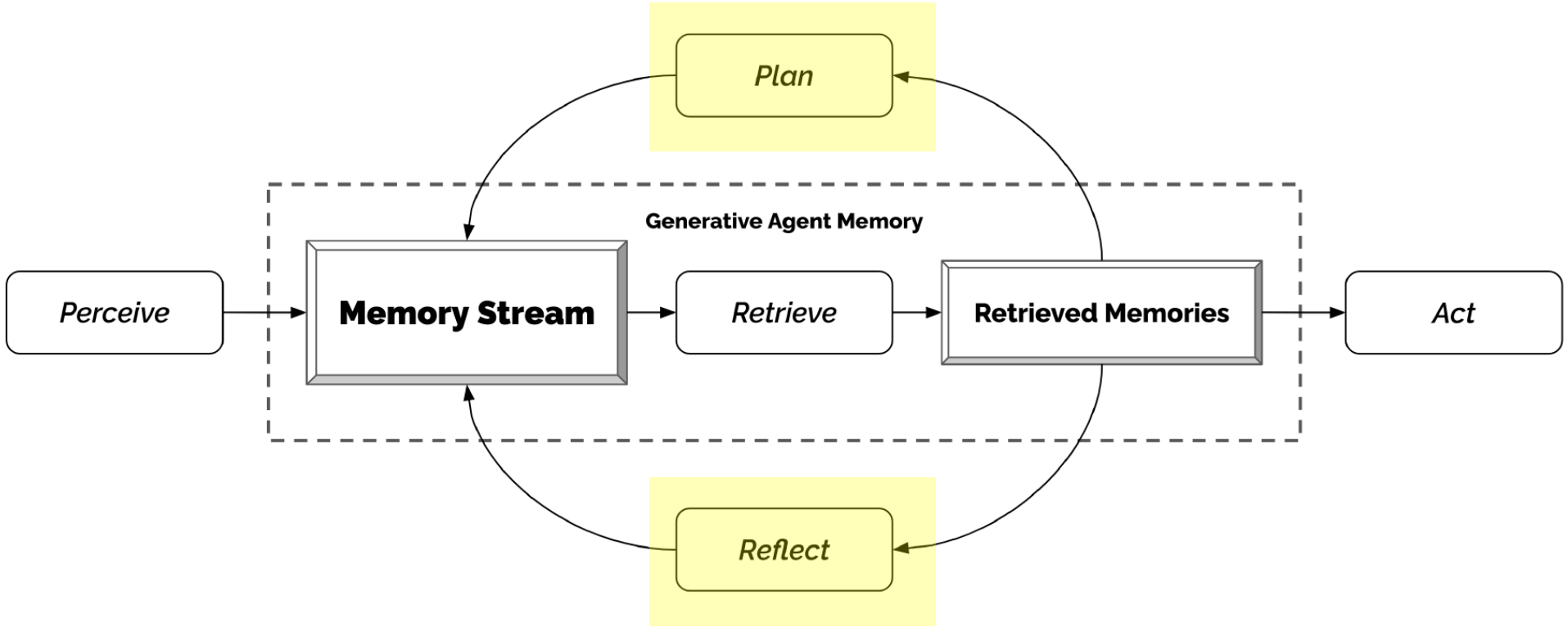
3. GPT가 도출해낸 결과를 행동에 옮김

Generative Agent Architecture



1) Retreive 단계를 통해, 관련성 높은 중요한 기억 추려내기

Generative Agent Architecture

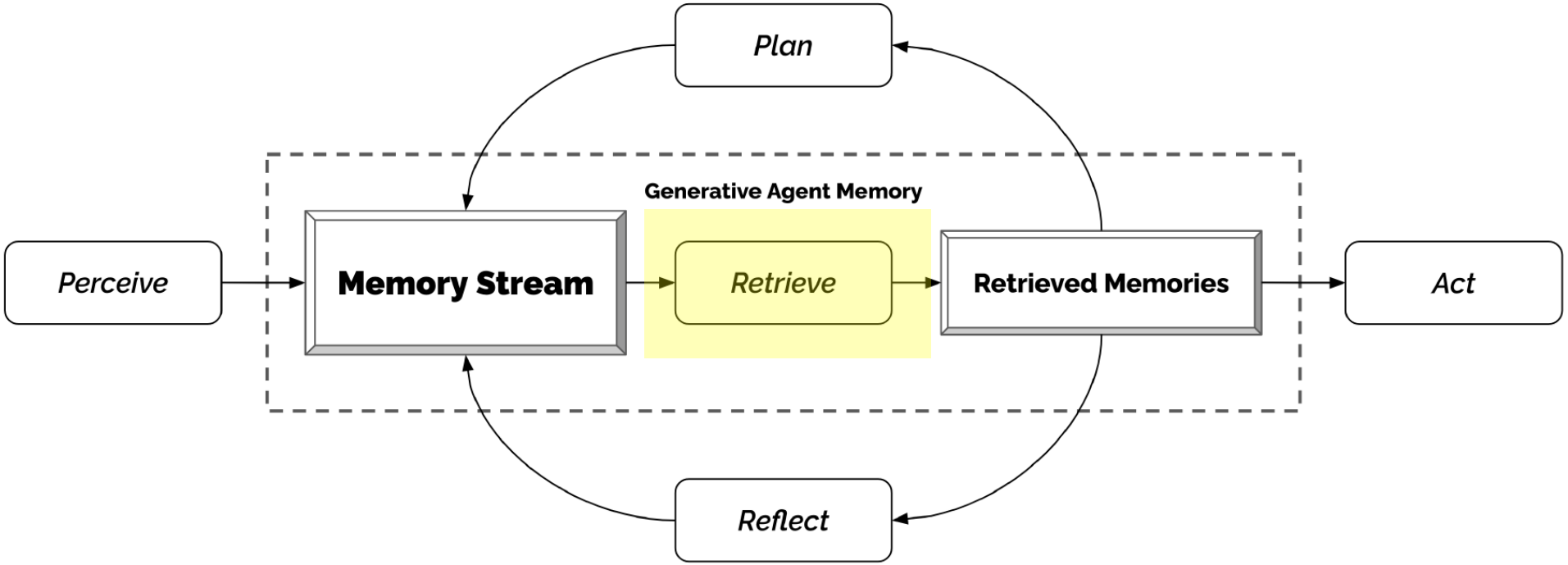


2) 보다 높은 수준의 추상적인 생각을 생성하는 Reflect 단계

3) 미래에 대한 계획을 Memory Stream에 반영하는 Plan 단계

Generative Agent Architecture

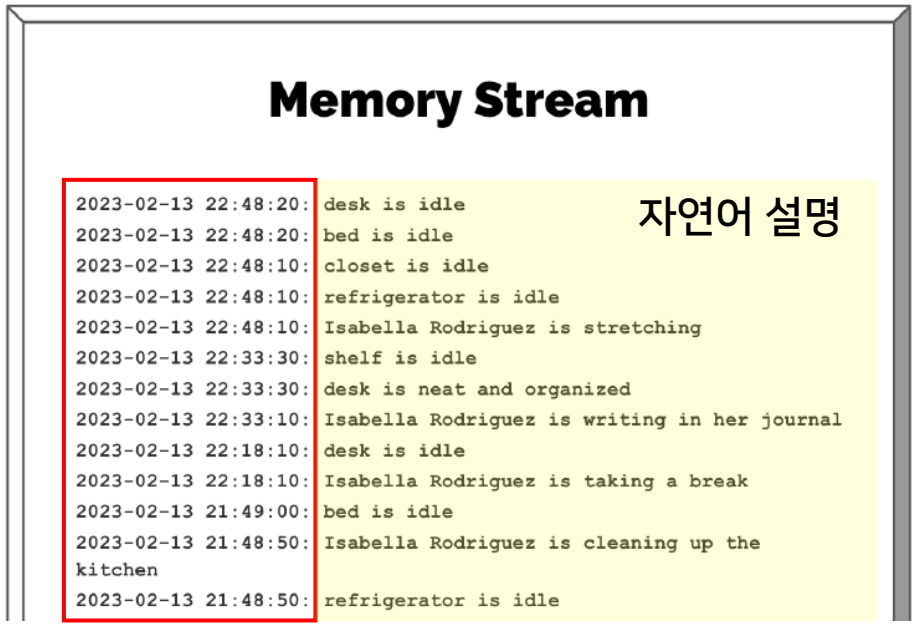
1. Memory and Retrieval



Generative Agent Architecture

1. Memory and Retrieval

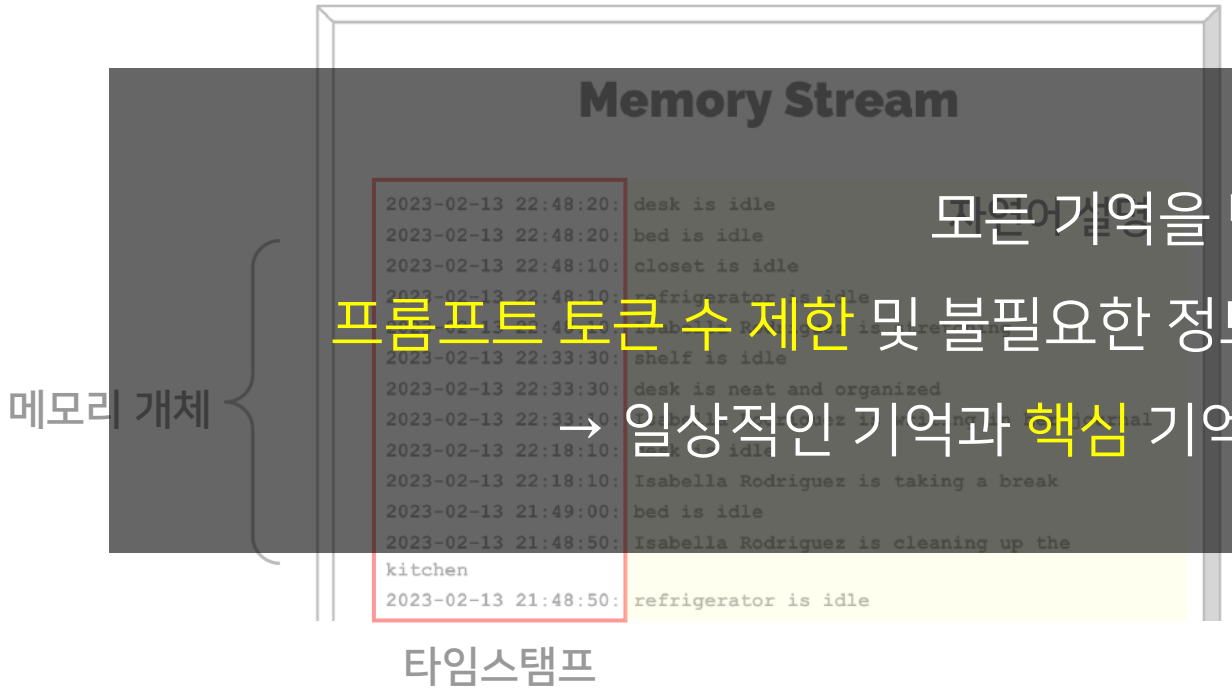
메모리 개체



- Memory Stream = 메모리 개체 리스트
- 메모리 개체에는
  Agent가 인식한 주변 환경, Agent의 행동,
  타 Agent와 나눈 대화 등이 포함되어 있음.
- 생성 타임스탬프, 자연어 설명,
  가장 최근 엑세스 타임스탬프가 포함됨.

Generative Agent Architecture

1. Memory and Retrieval



- Memory Stream = 메모리 개체 리스트
  메모리 개체에는
  Agent가 인식한 주변 환경, Agent의 행동,
  타 Agent와 나눈 대화 등이 포함되어 있음.
- 생성 타임스탬프, 자연어 설명,
  가장 최근 엑세스 타임스탬프가 포함됨.

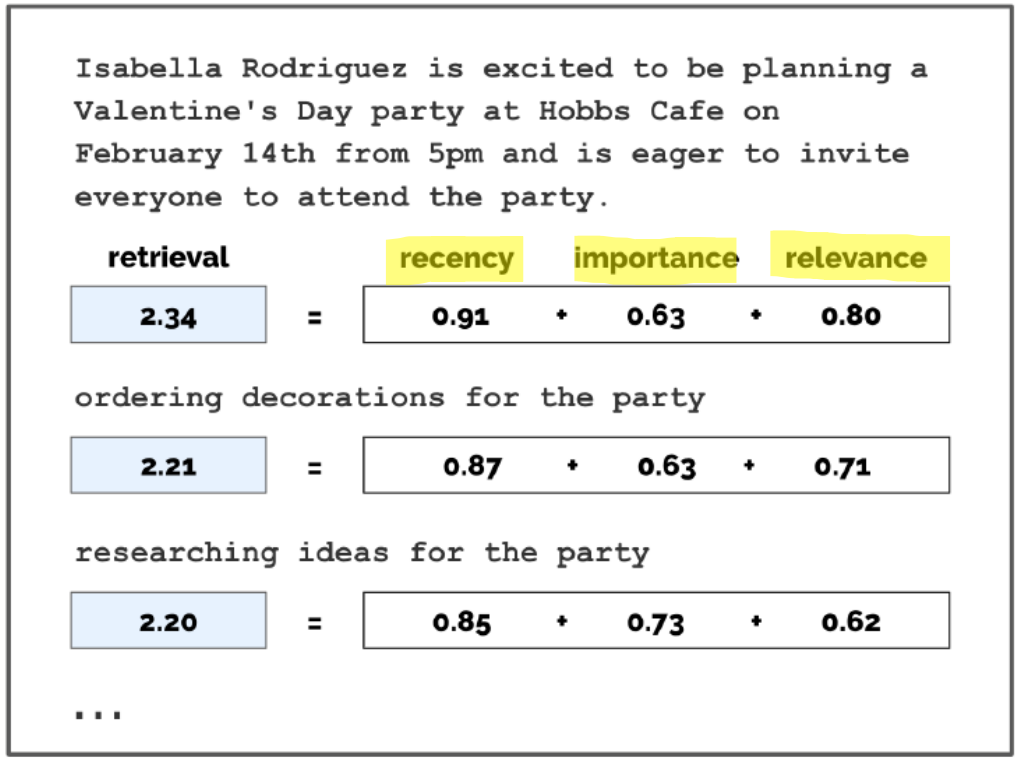
모든 기억을 반영하기엔,

프롬프트 토큰 수 제한 및 불필요한 정보에 의한 답변 정확도 하락 문제 발생

→ 일상적인 기억과 핵심 기억을 구분하여 기억을 추려냄

Generative Agent Architecture

1. Memory and Retrieval



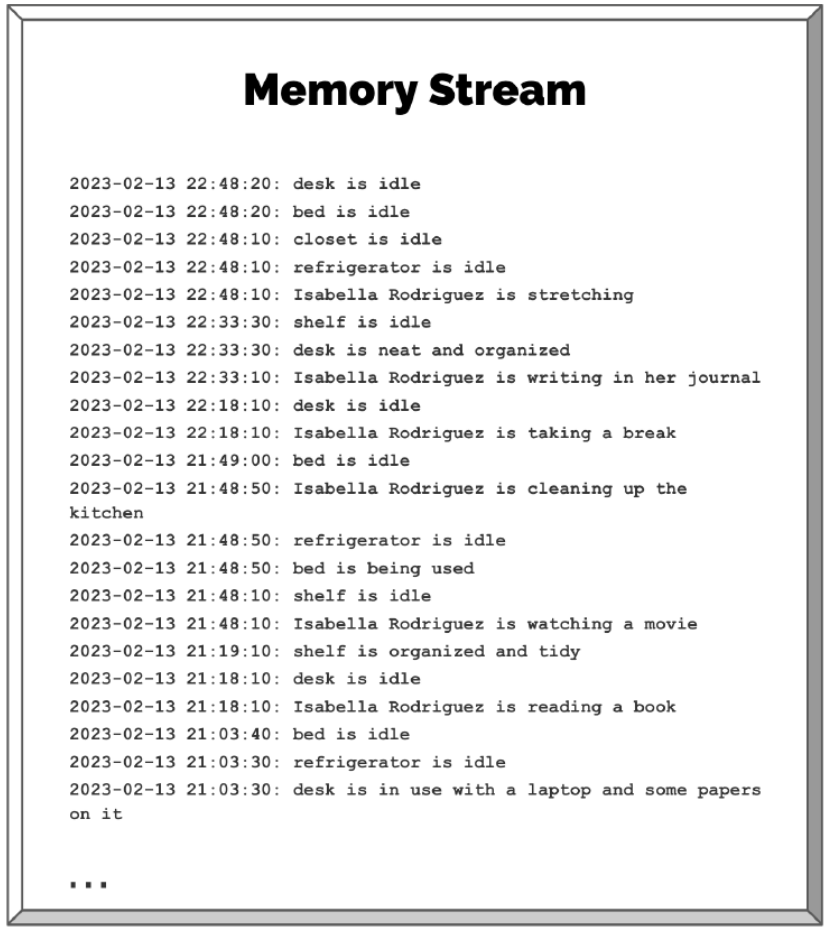
1) 최신 기억(recency) : 마지막으로 접근된 시점부터 계속 0.99를 곱해줌. (weight decay)

2) 중요한 기억(importance) : 기억이 생성될 때마다 GPT에게 이 기억이 얼마나 중요한지 0~10 사이 점수 매겨달라고 요청

3) 질문과 유사한 기억(relevance) : 질문의 embedding 값과 각 기억의 embedding 간 cosine 유사도 계산

Generative Agent Architecture

1. Memory and Retrieval



Q. What are you looking forward to the most right now?

Isabella Rodriguez is excited to be planning a Valentine's Day party at Hobbs Cafe on February 14th from 5pm and is eager to invite everyone to attend the party.

	retrieval		recency		importance		relevance
	2.34	=	0.91	+	0.63	+	0.80
ordering decorations for the party	2.21	=	0.87	+	0.63	+	0.71
researching ideas for the party	2.20	=	0.85	+	0.73	+	0.62

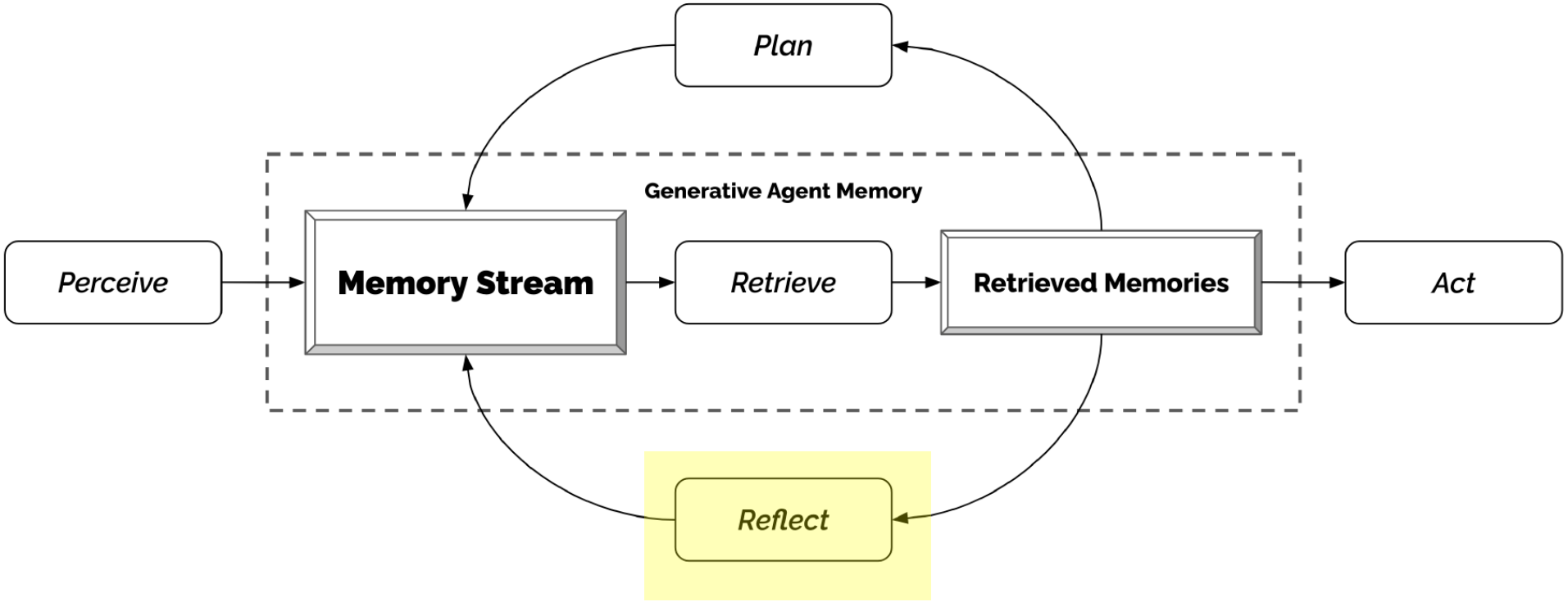
...

I'm looking forward to the Valentine's Day party that I'm planning at Hobbs Cafe!

Isabella

Generative Agent Architecture

2. Reflection



Generative Agent Architecture

2. Reflection

“If you had to choose one person of those you know to spend an hour with, who would it be?"

단순 인지 데이터 기반

WS

가장 잦은 빈도로 마주치는 사람

KM

Reflect를 통한 고차원적 사고 기반

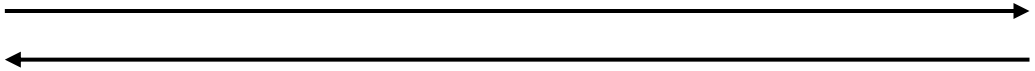
ML

관심사가 가장 비슷한 사람

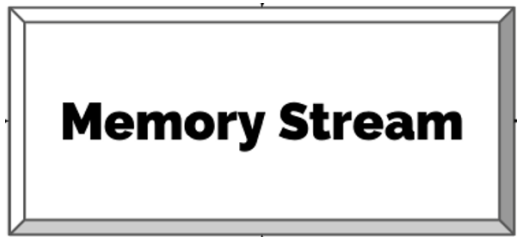
Generative Agent Architecture

2. Reflection

1) Agent의 Memory Stream에서 가장 최근 100개의 기억
+ "위 기억들로 대답할 수 있는 고차원적인 질문 3개 만들어줘."

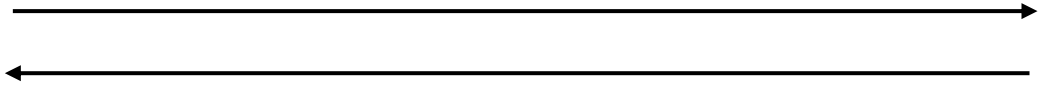


질문 3개 반환



2) GPT로부터 받은 질문 3개에 대한 retrieval 사용하여 관련 기억 추려냄

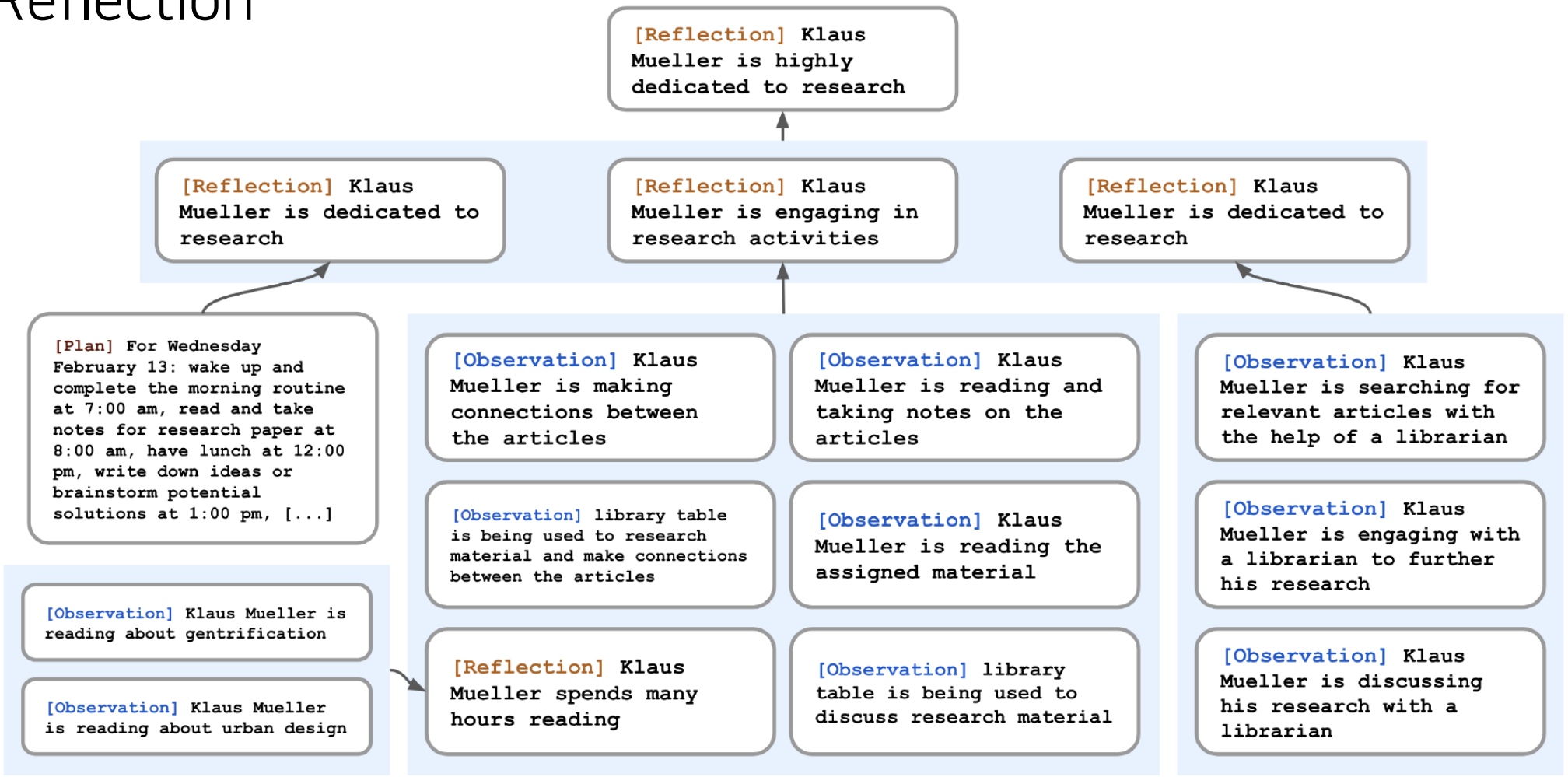
3) 이 기억들로 Agent에 대한 Insight 뽑아줘



Insight 반환
→ Memory Stream에 추가

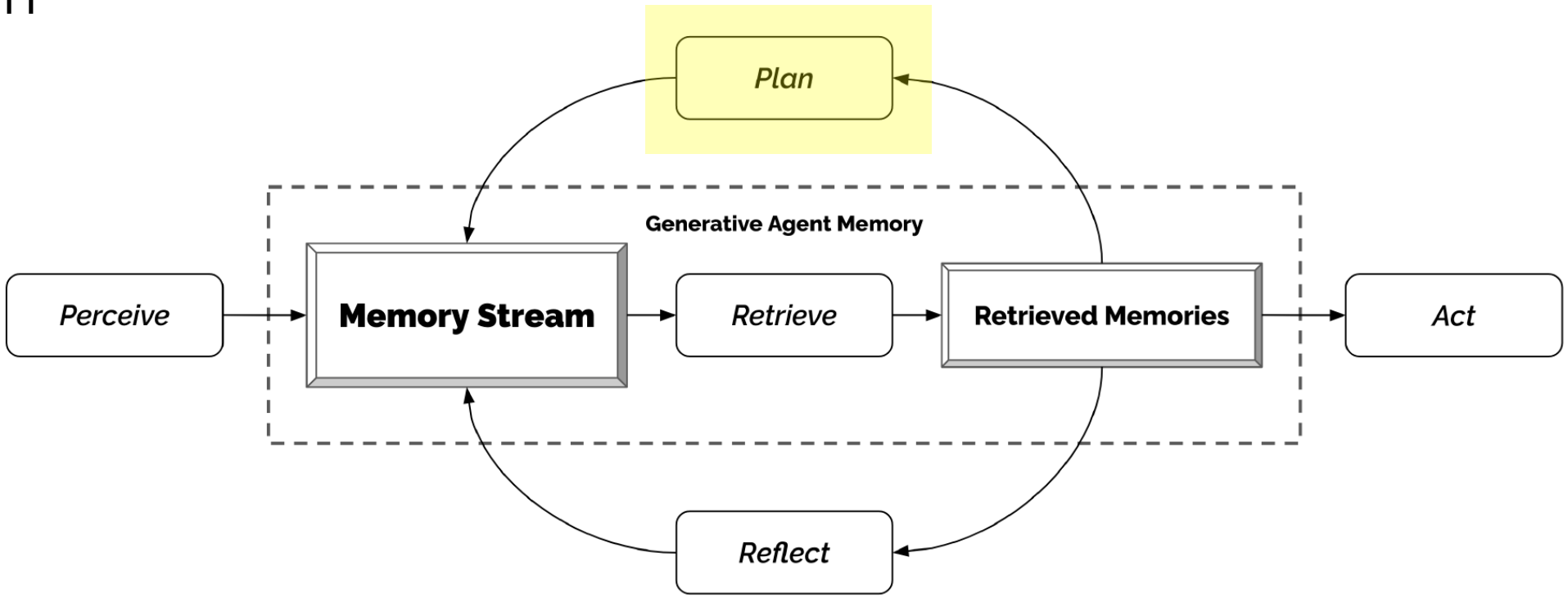
Generative Agent Architecture

2. Reflection



Generative Agent Architecture

3. Plan



Generative Agent Architecture

3. Plan

12:00 PM

12:30 PM

1:00 PM

Agent가 현재 시간대에 중요하다고 생각하는 일 반복하는 현상 발생

Generative Agent Architecture

3. Plan

대략적인 하루 일정을 Recursive하게 GPT에게 짜달라고 하기

Generative Agent Architecture

3. Plan

대략적인 하루 일정을 Recursive하게 GPT에게 짜달라고 하기

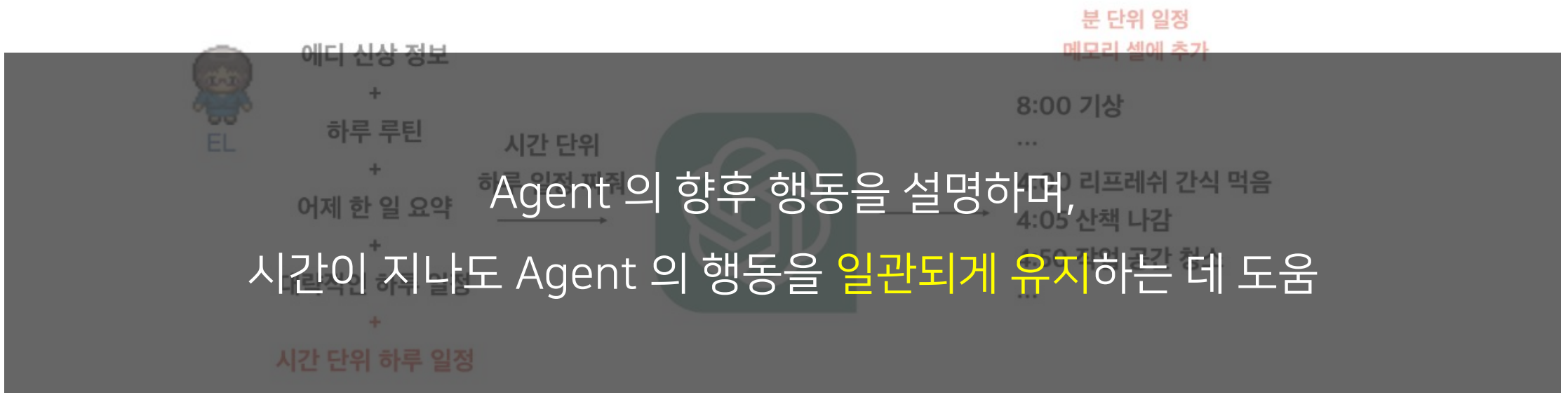
Generative Agent Architecture

3. Plan

대략적인 하루 일정을 Recursive하게 GPT에게 짜달라고 하기

Generative Agent Architecture

3. Plan



Agent 의 향후 행동을 설명하며,

시간이 지나도 Agent 의 행동을 일관되게 유지하는 데 도움

대략적인 하루 일정을 Recursive하게 GPT에게 짜달라고 하기

Experiment

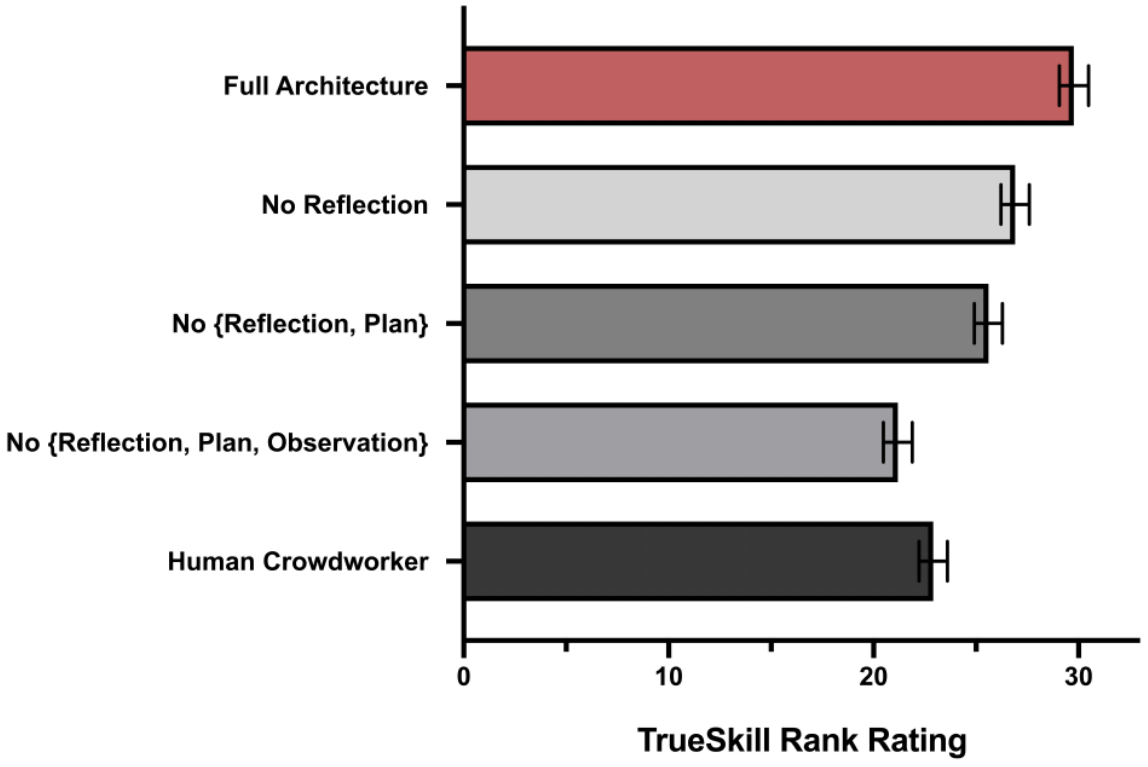
1. Controlled Evaluation

- Generative Agent가 하는 개별 행동들이 믿을만한가

- Agent들에 대해 인터뷰를 진행하여, 그들이 답변을 적절하게 하는지 평가
  1) Self-Knowledge (ex. 자기 소개, 일주일의 평일 일정 설명 등)
  2) Memory (ex. 특정 이벤트나 대화를 기억해야 정확하게 답할 수 있는 질문)
  3) Plan (ex. Agent의 장기 계획을 검색해야 하는 질문)
  4) Reactions (ex. 가상의 상황에 대해 Agent의 반응이 적절한지)
  5) Reflect (ex. 추론을 통해 고차원적 이해를 활용하는가)

Experiment

1. Controlled Evaluation



- Full Architecture가 다른 조건 중 가장 높은 점수를 획득함

- TrueSkill Rank Rating
  : 참가자들이 실험 조건을 순위 매긴 데이터를 기반으로 조건간 상대적 성능을 나타내는 지표

Experiment

2. End-to-End Evaluation

- Agent의 안정성과 새로운 사회적 행동을 파악하기 위해, 게임 시간 기준, 이틀 동안 서로 상호작용 하는 평가

- 이틀 동안 두 가지 특정 정보의 확산 측정
  1) Sam의 마을 시장 출마 소식
  2) Isabella의 홉스 카페에서 발렌타인데이 파티 개최 소식

Experiment

2. End-to-End Evaluation

1) **Sam의 마을 시장 출마 소식**

: 시뮬레이션 동안 Sam 시장 후보 출마 소식을 알고 있는 Agent 수가 1명에서 8명으로 증가. 소식을 아는 Agent들의 Memory Stream 확인 시, Hallucination 없었음.

2) **Isabella의 홉스 카페에서 발렌타인데이 파티 개최 소식**

: 개최 하루 전, Isabella가 손님 초대 및 파티 준비. 준비 과정에서 타 Agent의 도움 받음. 파티에는 초대받은 12명 중 5명 참여. 참여하지 않은 7명 인터뷰 시, 3명은 가고 싶었으나 일정이 있어 가지 못했다고 설명하였고, 나머지 4명은 소식만 들었을 뿐 참석 계획은 따로 없었음.

Applications of Generative Agents

1. **인간 행동 시뮬레이션**

   : 사회 시스템 및 이론 테스트 인간 행동을 이해하기 위한 수많은 사회 심리 실험에 인공지능 투입 가능성

2. **인간 중심 설계**

   : Agent 가 사용자 대신 일상적인 작업을 수행하고, 사용자의 선호도와 필요에 따라 피드백을 제공함으로써 더 나은 기술 경험 제공

Conclusion

- 인간 행동을 시뮬레이션하는 대화형 에이전트인 Generative Agent
- Generative Agent를 위한 Architecture

- Generative Agent를 샌드박스 게임의 NPC로 구현하고 시뮬레이션 함.
- 실험 결과, 제안한 Architecture가 신뢰 가능한 행동 생성함을 확인

- Generative Agent의 디자인 도구, 소셜 컴퓨팅 시스템, 다양한 대화형 응용 프로그램까지 확장 가능성 제안