Emergence of Social Norms in Large Language Model-based Agent Societies

Siyue Ren[1,†], **Zhiyao Cui**[1,†], **Ruiqi Song**[1], **Zhen Wang**[1,*], **Shuyue Hu**[2,*]
[1] Northwestern Polytechnical University
[2] Shanghai Artifcial Intelligence Laboratory
{rensiyue, zhiyao, songruiqi}@mail.nwpu.edu.cn,
w-zhen@nwpu.edu.cn, hushuyue@pjlab.org.cn

HCCLab Seminar

Junghwan Kim

Background

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

Figure 1: Generative agents are believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents plan their days, share news, form relationships, and coordinate group activities.

Emergence of Social Norms in Large Language Model-based Agent Society

Background

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

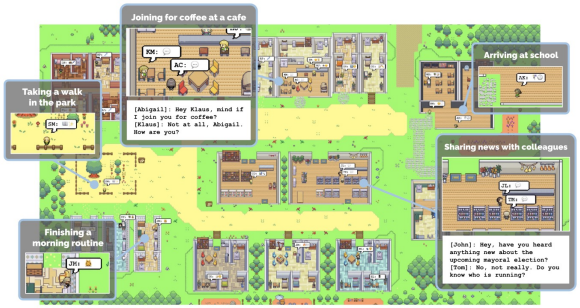


Figure 1: Generative agents are believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents plan their days, share news, form relationships, and coordinate group activities.

- 논문 3줄 요약
  - 자연어 형태의 기억을 가진 NPC들과 이들 25명이 모여 사는 마을 만들었다.
  - NPC들의 기억과 다음에 무슨 행동을 할지를 Prompt 형태로 GPT-4에게 질문했다.
  - 그랬더니 NPC끼리 대화도 하고, 썸도 타고, 파티도 열더라.

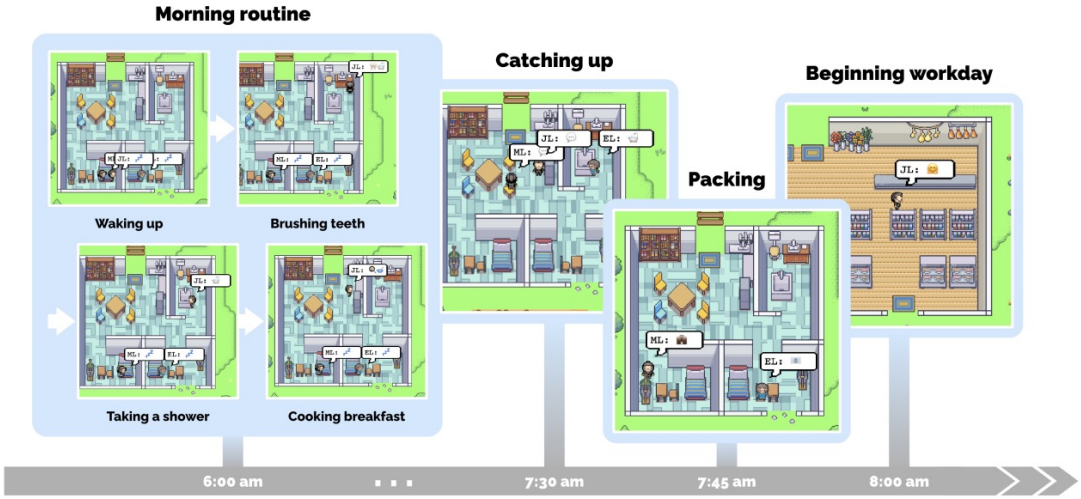


Figure 3: A morning in the life of a generative agent, John Lin. John wakes up around 6 am and completes his morning routine, which includes brushing his teeth, taking a shower, and eating breakfast. He briefly catches up with his wife, Mei, and son, Eddy, before heading out to begin his workday.

Background

- Generative Agent : Memory Stream(MS) 시스템
  - NPC들은 다음에 어떤 행동을 할지 결정하기 위해 기억 사용
  - Retrieve, Reflect, Plan 세가지 기억 매커니즘을 통해 사람 행동 모사
  - (Perceive) 주변 환경 인식, 다른 NPC와 대화
  - (Retrieve) Memory Stream(MS)에 저장된 기억 추려내기
  - (Reflect) 무엇이 더 고차원적인지 회고하여 MS에 반영
  - (Plan) 미래에 대한 계획을 짜서 MS에 반영

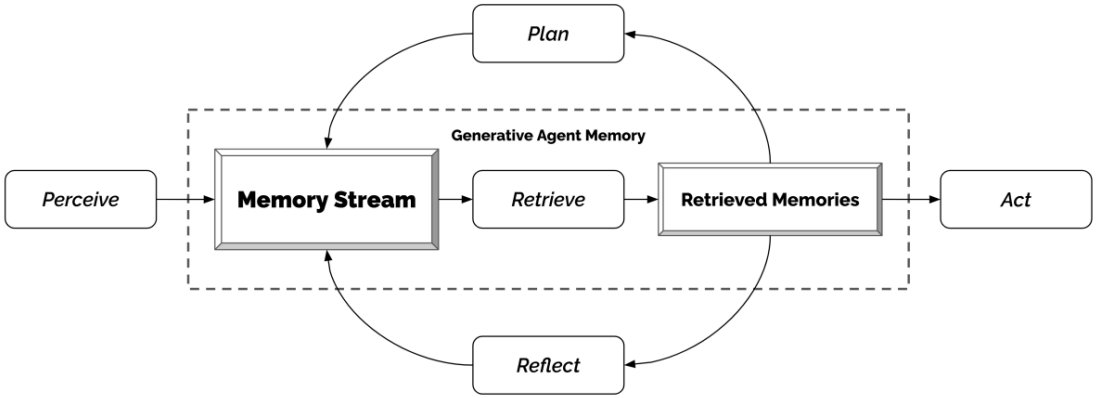


Figure 5: Our generative agent architecture. Agents perceive their environment, and all perceptions are saved in a comprehensive record of the agent's experiences called the memory stream. Based on their perceptions, the architecture retrieves relevant memories and uses those retrieved actions to determine an action. These retrieved memories are also used to form longer-term plans and create higher-level reflections, both of which are entered into the memory stream for future use.

Background

- Generative Agent : Memory Stream(MS)
  - 자신이 한 행동, 주변 환경 인식에서 중요한 기억들을 선별하여 자연어로 기억
  - NPC끼리 대화를 나누면 대화 내용이 MS에 추가
  - MS는 시간이 조금만 지나도 폭발적으로 길어지며, 프롬프트 토큰 수, 부정확한 답변 문제
  - 최신 기억, 중요 기억, 질문과 유사한 기억 기준으로 주요 기억 추려내는 Retrieval!

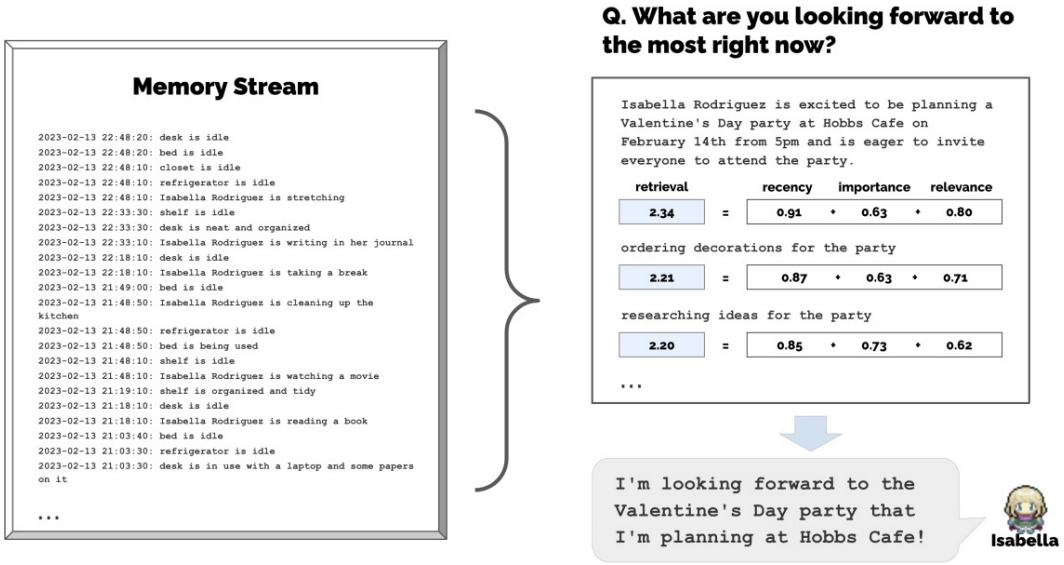


Figure 6: The memory stream comprises a large number of observations that are relevant and irrelevant to the agent's current situation. Retrieval identifies a subset of these observations that should be passed to the language model to condition its response to the situation.

Emergence of Social Norms in Large Language Model-based Agent Society

Background

- Generative Agent : Interaction

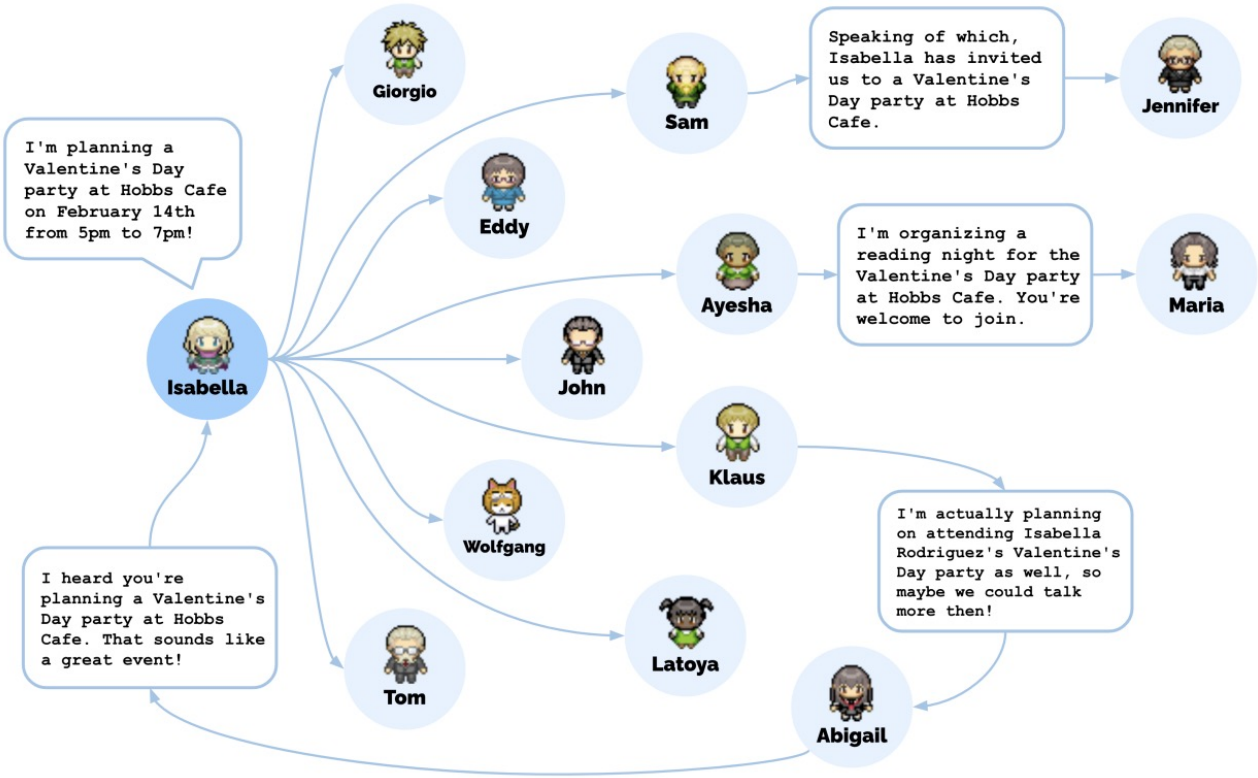


Figure 9: The diffusion path for Isabella Rodriguez's Valentine's Day party invitation involved a total of 12 agents, aside from Isabella, who heard about the party at Hobbs Cafe by the end of the simulation.

This Paper

Emergence of Social Norms in Large Language Model-based Agent Societies

Siyue Ren[1,†], **Zhiyao Cui**[1,†], **Ruiqi Song**[1], **Zhen Wang**[1,*], **Shuyue Hu**[2,*]
[1] Northwestern Polytechnical University
[2] Shanghai Artifcial Intelligence Laboratory
{rensiyue, zhiyao, songruiqi}@mail.nwpu.edu.cn,
w-zhen@nwpu.edu.cn, hushuyue@pjlab.org.cn

CRSEC

- (배경&문제) LLM 기반 에이전트 집단 내에서 사회 규범이 출현할 수 있도록 하는 최초의 Generative Agent 구조 제안
- (방법) CRSEC는 생성 및 표현, 확산, 평가, 준수의 네 가지 모듈로 구성

  이 구조는 하나의 구조에서 사회 규범의 출현 과정에서 중요한 여러 측면을 다룸 :

  (i) 사회 규범이 어디에서 오는가

  (ii) 어떻게 공식적으로 표현되는가

  (iii) 어떻게 에이전트들의 소통과 관찰을 통해 퍼지는가

  (iv) 어떻게 장기적으로 검토되고 종합되는가

  (v) 어떻게 에이전트의 계획과 행동에 통합되는가.
- SmallVill Sandbox 게임 환경에서 수행된 실험은 CRSEC가 LLM 기반 MAS 내에서 사회 규범을 확립하고 사회적 갈등을 줄이는 노력을 관찰
- (평가) 30명의 인간 평가를 바탕으로 프레임워크 평가

CRSEC : Architecture

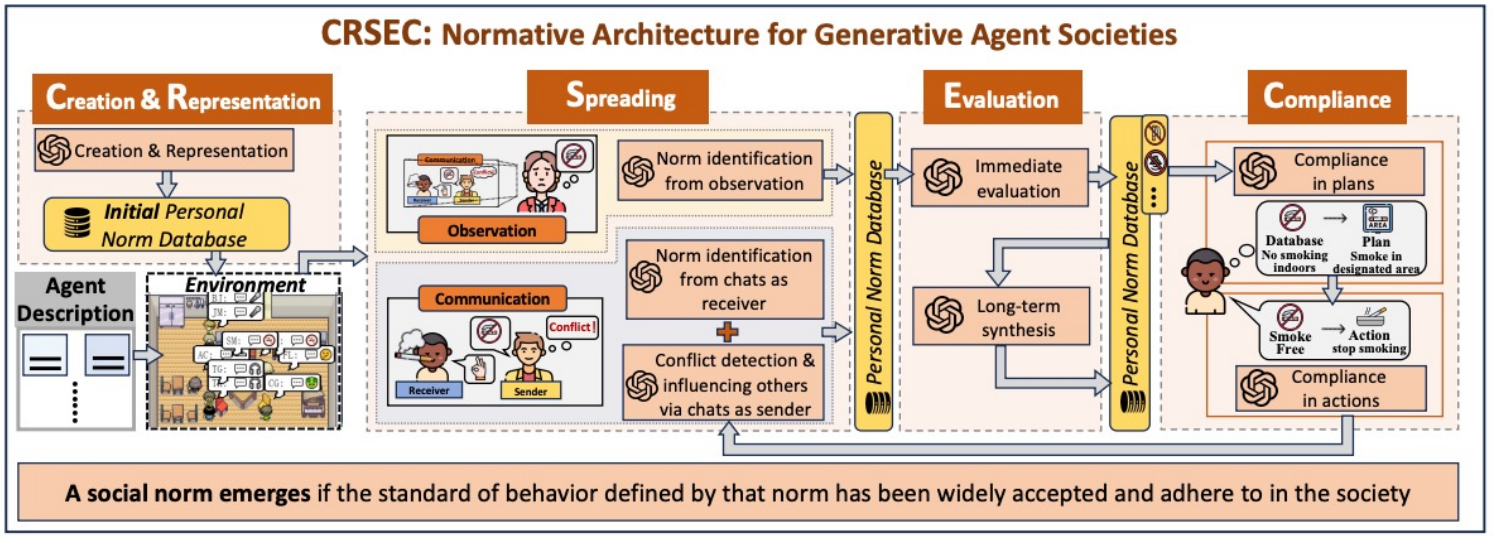
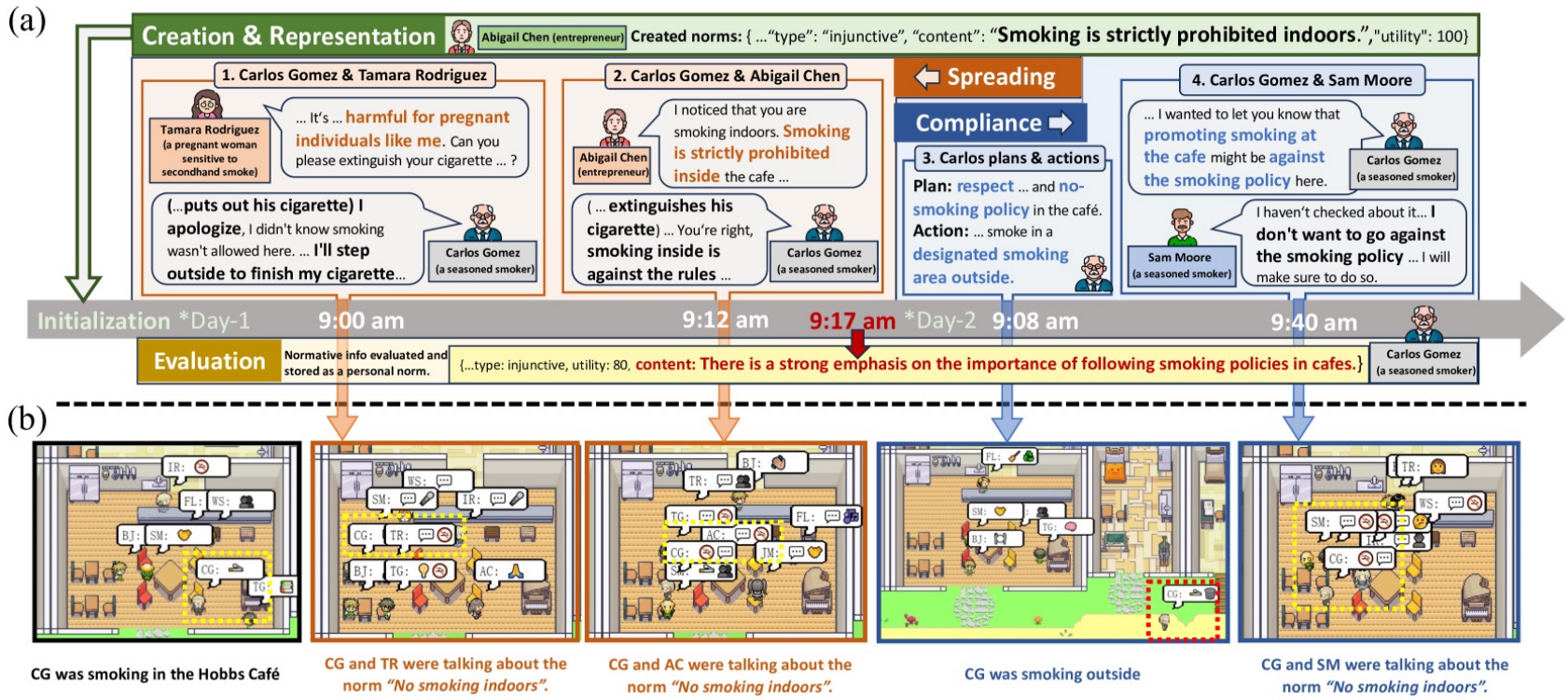


Figure 1: CRSEC: our architecture for the emergence of social norms in generative agent societies. Initially, by the *Creation & Representation* module, norm entrepreneurs create their personal norms and store them into their databases. By the *Spreading* module, some agents proactively influence others to adopt their personal norms through initiating communication with others, while others can identify those norms from their chats and observations. The identified norms then undergo an immediate evaluation in the *Evaluation* module. The *Compliance* module enables agents to generate plans and actions, with the norms bearing in mind. The normative actions, in turn, can influence other agents' observations and thus reinforce the spreading of norms. In addition, from time to time, agents perform long-term synthesis to keep their personal norms compact and concise.

CRSEC : Scenario

- Carlos Gomez가 “No smoking indoors”를 Personal Norm으로 Adopt하는 과정



CRSEC : Prompts

Creation & Representation $\mathcal{P} \leftarrow \texttt{CreateNorm}(\mathcal{G})$

TASK: generate 5 norms in a Café based on the AGENT DESCRIPTION, NORM DEFINITION, ..., DESIRED FORMAT, EXAMPLE, and ATTENTION
AGENT DESCRIPTION: < agent description >
DESIRED FORMAT: JSON

```
"norm_i":{
"ID":i,
"type": "α",
"content": "description",
"utility": score,
"activation_state": true,
"validity_state": true}
```

……
NORM DEFINITION: Social norms are standards of acceptable behaviors by groups.
EXAMPLE:

```
"norm_1":{
"ID":1,
"type": "injunctive",
"content": "Everyone is not allowed to smoke indoors.",
"utility": 100,
"activation_state": true,
"validity_state": true}
```

ATTENTION: Do not output anything else except for the content in JSON.

Evaluation: Long-term Synthesis $\{(\mathcal{Q}, k)\} \leftarrow \texttt{ClassifySpecificNorms}(\mathcal{P})$

TASK: Categorize QUALIFIED PERSONAL NORMS based on the content and assign a theme to each category and try to refrain from categorizing into single norms as much as possible. A category should only contain one theme, should NOT contain two or more topics. DO NOT classify too roughly (E.g. Public Behavior, Cafe Etiquette Related, cafe-environment-related).
……
EXAMPLE: QUALIFIED PERSONAL NORMS: 1. No one is allowed to take photos in public restrooms. 2. Keep personal belongings secure and within reach. 3. People should be considerate of others in public spaces. 4. People offer support to each other. 5. People offer support to each other indoors. OUTPUT: [Privacy-related] 1. No one is allowed to take photos in public restrooms. [Property Security Related] 2. Keep personal belongings secure and within reach. [Helping Others Related] 3. People should be considerate of others in public spaces. 4. People offer support to each other. 5. People offer support to each other indoors.
QUALIFIED PERSONAL NORMS: <qualified personal norms>
OUTPUT:

Spreading: Sender $\mathcal{Y}_{\texttt{conflict}} \in \{\texttt{T},\texttt{F}\} \leftarrow \texttt{DetectConflict}(\mathcal{O}_\mathcal{S}, \mathcal{P}_\mathcal{S})$ $\mathcal{Y}_{\texttt{talk}} \in \{\texttt{T},\texttt{F}\} \leftarrow \texttt{DecideToTalk}(\mathcal{G}_\mathcal{S})$

TASK: This is a two-stage task. Let's think step by step.
STAGE 1: detect whether there is a certain conflict between OBSERVATION and QUALIFIED PERSONAL NORMS. Answer ONLY in "yes" or "no" ……If the first stage output is 'yes,' then proceed to STAGE 2.
STAGE 2: determine whether to have a conversation about the conflict based on AGENT DESCRIPTION.

If you are an entrepreneur, then you SHOULD start a conversation. Whether to initiate a conversation still depends on you.
If you are an ordinary agent, considering AGENT DESCRIPTION, ##decide## whether to have a conversation about the conflict and provide the FINAL OUTPUT.

OBSERVATION: <observation>
QUALIFIED PERSONAL NORMS: <qualified personal norms>
AGENT DESCRIPTION: <agent description>
……
ATTENTION: If OBSERVATION is '<name> is idle', it represents the individual's initial state with no conflict .....

Compliance: Action $\mathcal{L}_{\texttt{action}} \leftarrow \texttt{GenerativeNormativeActions}(l_i, \mathcal{P}, \mathcal{G})$

TASK: Considering the CURRENT PLAN, AGENT DESCRIPTION, and QUALIFIED PERSONAL NORMS, describe actions in 5 min increments. ……
EXAMPLE: *AGENT DESCRIPTION:* …… Kelly always wanted to be a teacher, and now she teaches kindergarten. During the week, ……Currently: Kelly is a teacher during the school year. She is currently having lunch at the Hobbs Cafe. Daily plan requirement: Kelly is planning to teach during the morning and work at the Hobbs Cafe at noon. ……*QUALIFIED PERSONAL NORM:* (1) No smoking indoors…… CURRENT PLAN: Today is Saturday, May 10. ……From 12:00 pm to 1:00 pm, Kelly plans to have lunch and take a break. ……*OUTPUT:* 1. Kelly is deciding on her order at the Hobbs Cafe. (duration in minutes: 5, minutes left: 55) ……
QUALIFIED PERSONAL NORMS: < qualified personal norms >
CURRENT PLAN: <current plan>
AGENT DESCRIPTION: < agent description >
Considering compliance with existing qualified personal norms, in 5 min increments, list the subtasks <agent name> does when < agent first name > is < current action> from <current time range > (total duration in minutes < current action duration in min>)
OUTPUT:

CRSEC : Scenario

(a)

TG was reading a book while listening to music aloud

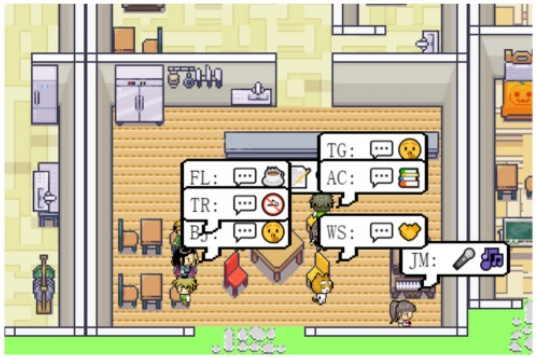
(b)

TG was talking about the norm “Be quiet in public”

(c)

TG was reading a book while listening to music through headphones

(a)



JM was singing loudly in the Hobbs Café

(b)

JM was talking about the norm “Be quiet in public”

(c)

JM was singing outside

CRSEC : Evaluation

- 각 Agent가 생성한 결과물(생각, 대화, 확인된 규범 정보 등)은 인간 평가자(N=30)에게 할당되어 평가되었음
  - 평가자는 에이전트의 설명을 읽고 2일간의 생활을 재생하는 영상을 관찰한 후 설문지 작성
  - 에이전트가 생성한 결과물에서 무작위로 선택된 20쌍의 입력과 출력을 제시받고 주어진 입력에 대한 출력에 얼마나 동의하는지 동의 수준을 7점 리커트 척도로 설문 완성

Module	Sub-component	Score	Module	Sub-component	Score
Creation		6.44±0.11	Evaluation	Long-term Synthesis	5.97±0.07
Spreading	Sender	5.86±0.05		Immediate Evaluation	5.14±0.07
	Receiver	5.77±0.08	Compliance	Action	6.40±0.04
	Observation	5.13±0.05		Plan	6.43±0.14

7 Strongly Agree
4 Neutral
1 Strongly Disagree

CRSEC : Evaluation

Questionnaire

Welcome to participate in a role-playing activity! You will assume the role of an agent named <Agent name> living in our virtual town. Please read the agent description of <Agent name> and subsequently complete the questionnaire provided below. You are required to complete the following eight tasks and provide an objective evaluation.

AGENT NAME: <Agent name>

AGENT DESCRIPTION: <Agent description>

Please use the following grading scale to evaluate each task. Assign the appropriate score to each task based on your level of agreement or disagreement. Thank you for your participation!

7-strongly agree | 6-agree | 5-somewhat agree | 4-neutral | 3-somewhat disagree | 2-disagree | 1-strongly disagree

---

TASK 1: (CREATION) This task has three questions. Please assign a score according to all three questions.

Q1: Do you agree with the norm utility score given by the agent? Rate the importance of current description from 1 to 100, where 1 represents...

Q2: Do you agree that the norm is consistent with the agent description?

Q3: Do you agree with the norm type classified by the agent? (descriptive norm OR injunctive norm) (1) descriptive norm: ... (2) injunctive norm: ...

NORMS: <Norm related data>

TASK 2: (Spreading, Sender) This task has two questions. Please assign a separate score to each question.

Q1: Do you agree with the agent's assessment that the observation conflicts (or does not conflict) with its personal norms?

PERSONAL NORMS: <Agent's Personal Norms>

OBSERVATION: <Agent's Observation>

DETECT CONFLICTS: <yes/no>

Q2: If there exists a conflict, based on the CRITERIA below, do you agree with the agent's decision to talk (or not)?

CRITERIA: (1) entrepreneur: ... (2) citizen: ...

IDENTITY: <entrepreneur/citizen>

INNATE: <Agent's innate>

DECIDE TO TALK: <yes/no>

TASK 3: (Spreading, Receiver) This task has two questions. Please assign a separate score to each question.

Q1: Do you agree that this conversation contains normative information? Please rate its relevance accordingly.

CONVERSATION: <Conversation data>

Q2: Do you agree that the conversation can be summarized into the related norm based on the principles below?

PRINCIPLES: Social norms are: 1.Social interactions and sharing among group members; …

CONVERSATION: <Conversation data>

RELATED NORM: <Related norm>

CRSEC : Discussion

- 장점
  - Generative Agent Platform을 적극 활용하여 Social Norm의 전파를 확인할 수 있는 시스템 구현
  - Memory Stream에 Creation, Representation, Spreading, Compliance, Evaluation 프롬프트를 고도화하여 추가함으로써 구체적인 실험 가능하게 하였음
- 아쉬운 점
  - Human Evaluation이 어떤 의미를 가지는지 잘 모르겠음
  - Agent의 말과 행동에 대해 인간 평가자가 동의하는지 물어보는 일련의 질문들이 아쉬움
  - Key Findings가 아쉬움
    - Social Norms always emerge
    - Social conflicts almost vanish as social norms emerge
    - Conversations and Thoughts drive the emergence of social norms
    - Decriptive norms are harder to establish than injunctive norms
    - Qualified personal norms can be synthesized into a more general one
  - 개인의 행동 Social Norm에 의해 제한되는 시나리오보다는, 개인의 생각, 행동들이 모여 Social Norm이 되어가는 과정을 더 깊이있게 다루었으면 좋지 않았을까?

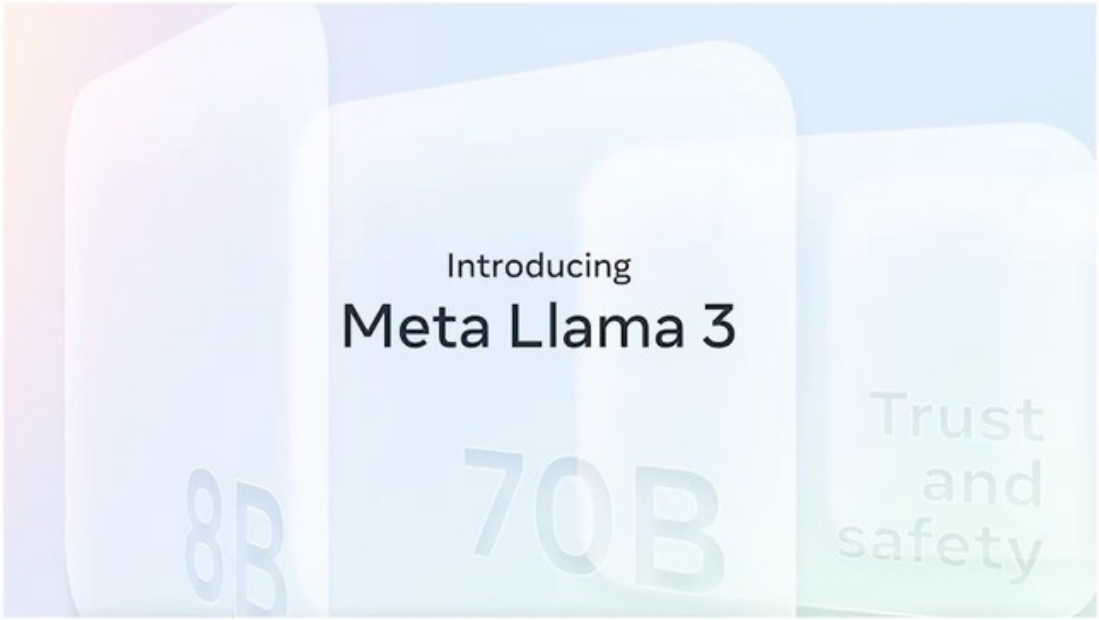
CRSEC : Future Works

- 탄탄한 사회과학, Social Computing 이론을 바탕으로 아래 주제 실험
  - Reputation은 어떻게 형성되는지
  - Sanction은 어떻게 형성되는지
  - Leadership은 어떻게 형성되는지
- TBD…

Thanks🙏

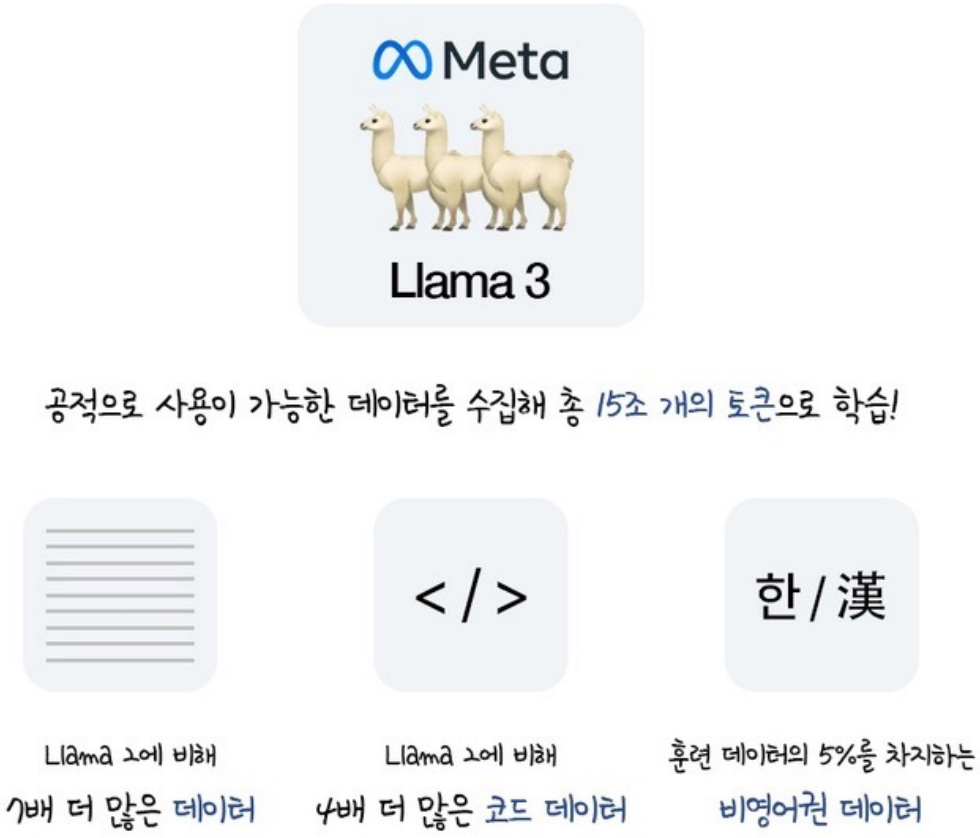
HCC LAB SNU

Llama3



https://llama.meta.com/llama3/

Llama3



https://llama.meta.com/llama3/

Llama3

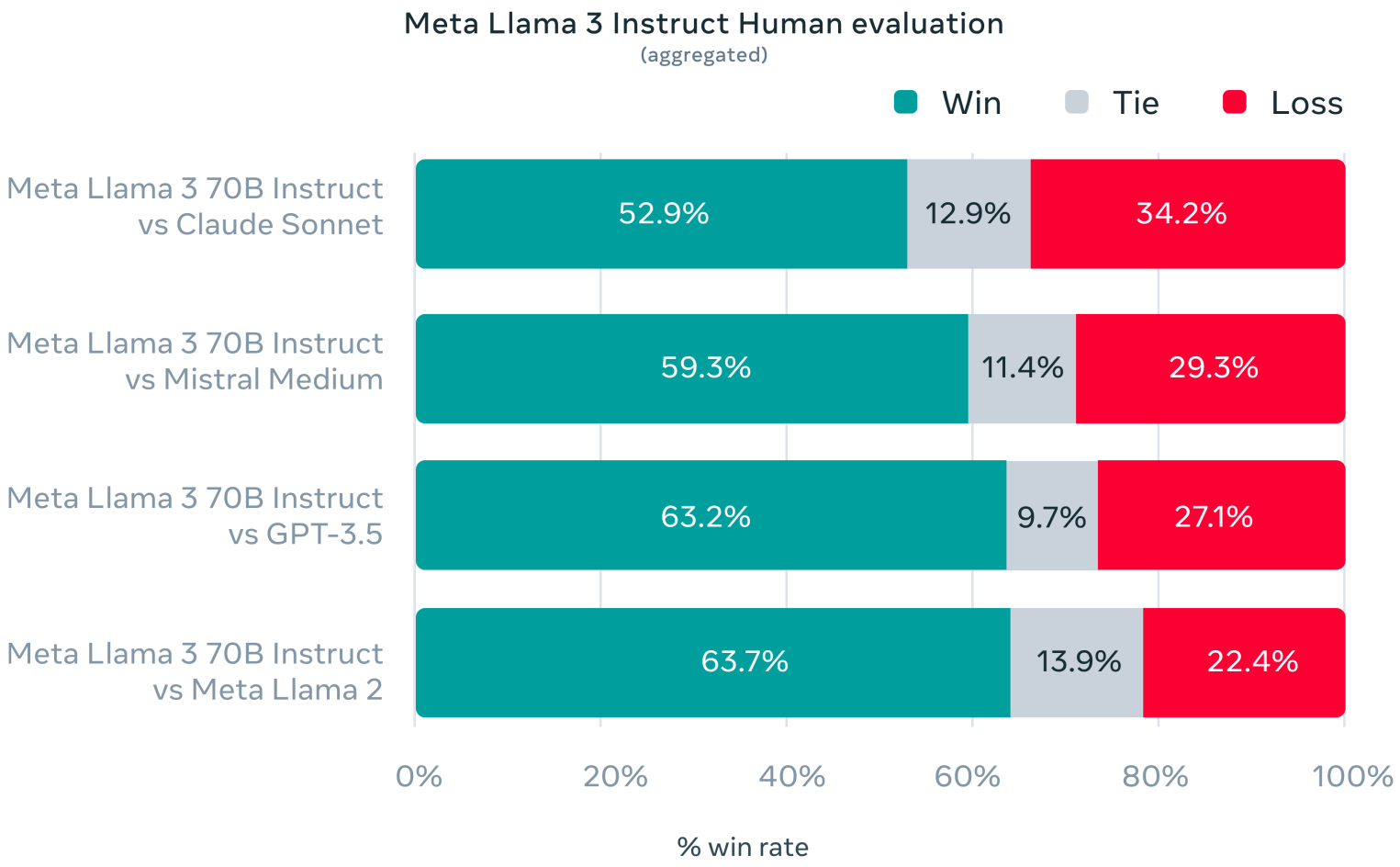
Meta Llama 3 Instruct model performance

	Meta Llama 3 8B	Gemma 7B - It Measured	Mistral 7B Instruct Measured
MMLU 5-shot	**68.4**	53.3	58.4
GPQA 0-shot	**34.2**	21.4	26.3
HumanEval 0-shot	**62.2**	30.5	36.6
GSM-8K 8-shot, CoT	**79.6**	30.6	39.9
MATH 4-shot, CoT	**30.0**	12.2	11.0

	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	**82.0**	81.9	79.0
GPQA 0-shot	39.5	**41.5** CoT	38.5 CoT
HumanEval 0-shot	**81.7**	71.9	73.0
GSM-8K 8-shot, CoT	**93.0**	91.7 11-shot	92.3 0-shot
MATH 4-shot, CoT	50.4	**58.5** Minerva prompt	40.5

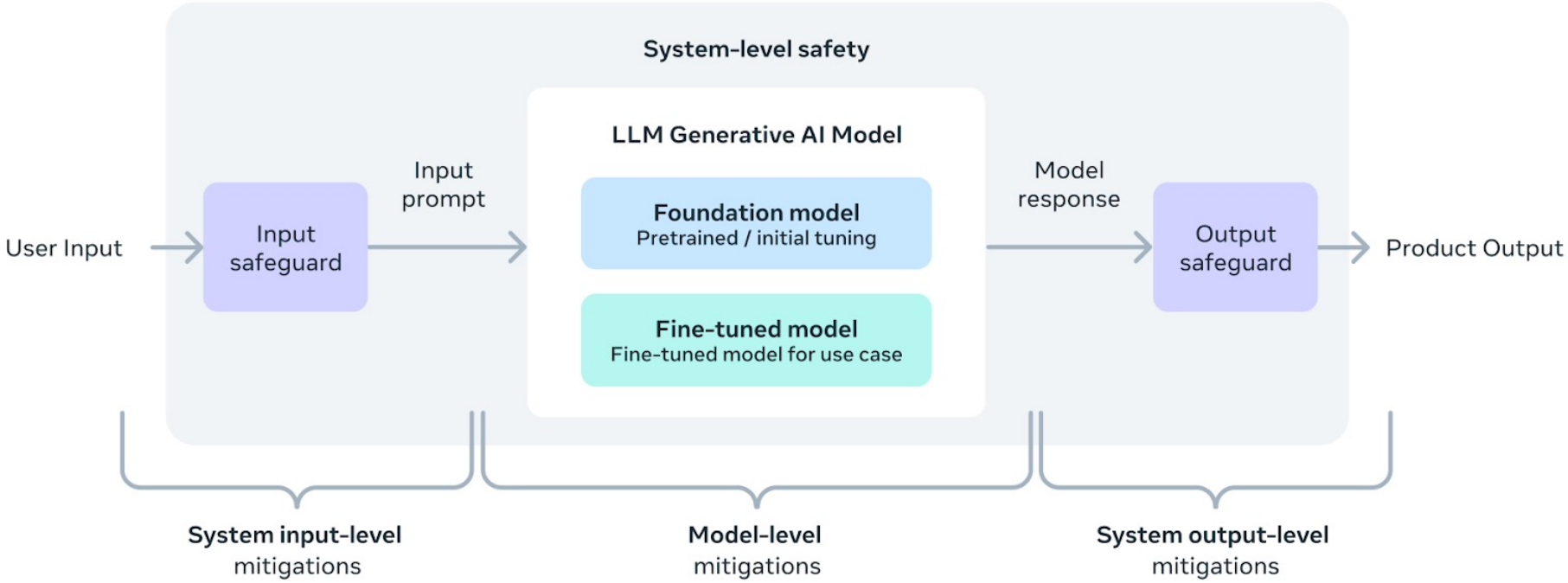
https://llama.meta.com/llama3/

Llama3



https://llama.meta.com/llama3/

Llama3



https://llama.meta.com/llama3/

Llama3

Meta Llama 3 400B+ (still training)
Checkpoint as of Apr 15, 2024

	PRE-TRAINED Meta Llama 3 400B+
MMLU 5-shot	84.8
AGIEval English 3-5-shot	69.9
BIG-Bench Hard 3-shot, CoT	85.3
ARC-Challenge 25-shot	96.0
DROP 3-shot, F1	83.5

	INSTRUCT Meta Llama 3 400B+
MMLU 5-shot	86.1
GPQA 0-shot	48.0
HumanEval 0-shot	84.1
GSM-8K 8-shot, CoT	94.1
MATH 4-shot, CoT	57.8

https://llama.meta.com/llama3/