

# 언어 모델 기반 사회 시뮬레이션 가능성 검토

ChatGPT와 로컬 언어 모델을 활용한 무지의 베일 실험을 바탕으로

적당히 꿈꾸는 전기양 강동하 배윤호

# Introduction



## 1. Introduction

# ■ LLM to Social Experiment

Using large language models  
to simulate multiple humans  
and replicate human subject studies (2023)

LLM과 인간 참가자의 4가지 사회 실험의 답변 비교

- ↳ 최후통첩 게임(Ultimatum Game), 정원길 문장(Garden Path Sentences),  
밀그램 실험(Milgram Experiment), 군중의 지혜(Wisdom of Crowds)

LLM과 인간 참가자의 답변이 일반적으로 유사함

→ LLM을 통한 사회 실험 가능

→ LLM 간 '상호작용'하는 실험 가능성 여부 질문 존재



## Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies

Gati Aher<sup>1</sup> Rosa I. Arriaga<sup>2</sup> Adam Tauman Kalai<sup>3</sup>

### Abstract

We introduce a new type of test, called a Turing Experiment (TE), for evaluating to what extent a given language model, such as GPT models, can simulate different aspects of human behavior. A TE can also reveal consistent distortions in a language model's simulation of a specific human behavior. Unlike the Turing Test, which involves simulating a single arbitrary individual, a TE requires simulating a representative sample of participants in human subject research. We carry out TEs that attempt to replicate well-established findings from prior studies. We design a methodology for simulating TEs and illustrate its use to compare how well different language models are able to reproduce classic economic, psycholinguistic, and social psychology experiments: *Ultimatum*

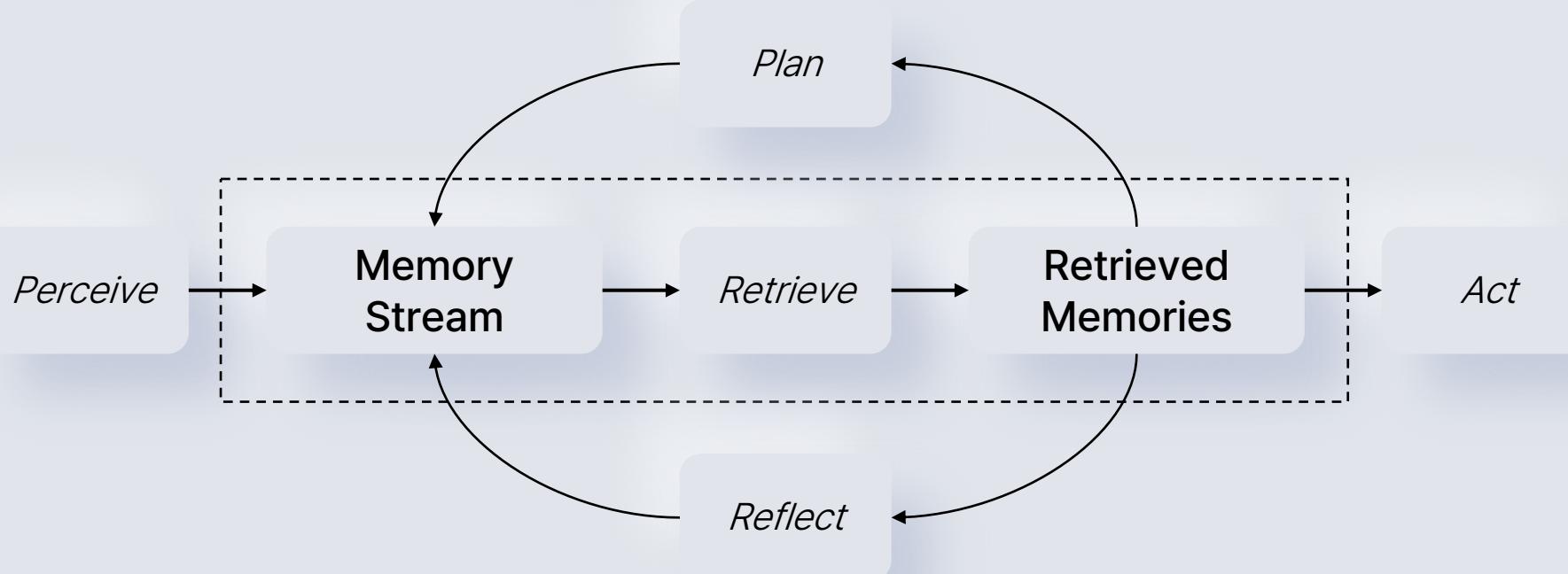
trolled experiments, and we thus avoid philosophical debates around the meaning of “understanding” (Bekker, 2020). Now, simulating human behavior is hard, even for humans, especially in complex real-life situations fraught with ambiguity. After all, if simulating human behavior were easy, there would be no need for human subject experiments as one could simply simulate the outcomes. A further obstacle to accurate simulation is that behavior differs across individuals and populations. A perfect simulation would require capturing these differences for all groups including minority groups.

In Turing’s Imitation Game (IG), an AI system has to imitate an individual well enough to fool a human judge. Large Language Models (LMs) may come close to “winning” the game in the near future, especially if they only have to imitate a single human—one oddly successful early human who simulated a 13 year old troublemaker (Warwick &

# Generative Agents

Generative Agents Interactive Simulacra of Human Behavior (2023)

Agent는 기억(Remember), 검색(Retrieve), 반영(Reflect), 계획(Plan) 거쳐 동적으로 변화하며 다른 Agent들과 상호작용함



## ■ Research of LLMs

### LLMs vs. Small LLMs

- GPT-3.5(175B)/4(1.7T)   
Decoder 기반 모델로 RLHF 기법 통해 성능 향상
- Small LLMs:  
Scaling Law와 반대로 GPT보다  
모델 파라미터 줄이면서도 성능 유지하는 연구 집중  
↳ LLaMA(7/13/30/65B), Mistral(7B), OpenHermes(7B)



### Quantization

- AI 모델 학습 및 추론 시 단위 값의 변환하여 연산량을 줄이고 전력 효율성을 높리기 위한 기법
- 고성능의 GPU뿐만 아니라 저사양의 GPU/CPU에서 구동 가능해짐
- 일반적인 AI 모델은 32-bit 부동소수점을 사용해 연산을 하는데, 16-bit 부동소수점, 8-bit, 4-bit 등의 정수를 사용해 계산량 줄임
- 양자화된 LLM을 쉽게 사용하고자 하는 llama.cpp, ollama 등의 오픈소스 프로젝트 존재함

## ■ Research Question

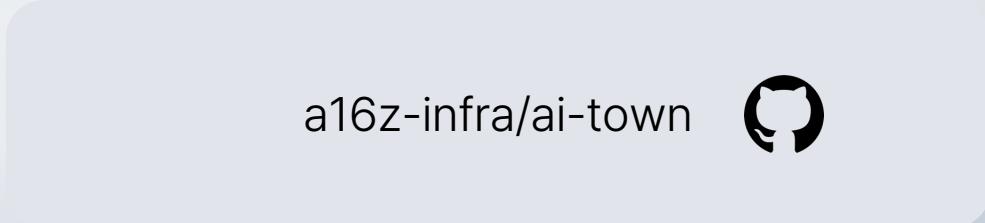
GPT뿐만 아니라 로컬 언어 모델로도,  
사회 실험을 시뮬레이션 하는 것이 가능한가?

# Experiment



## 2. Experiment

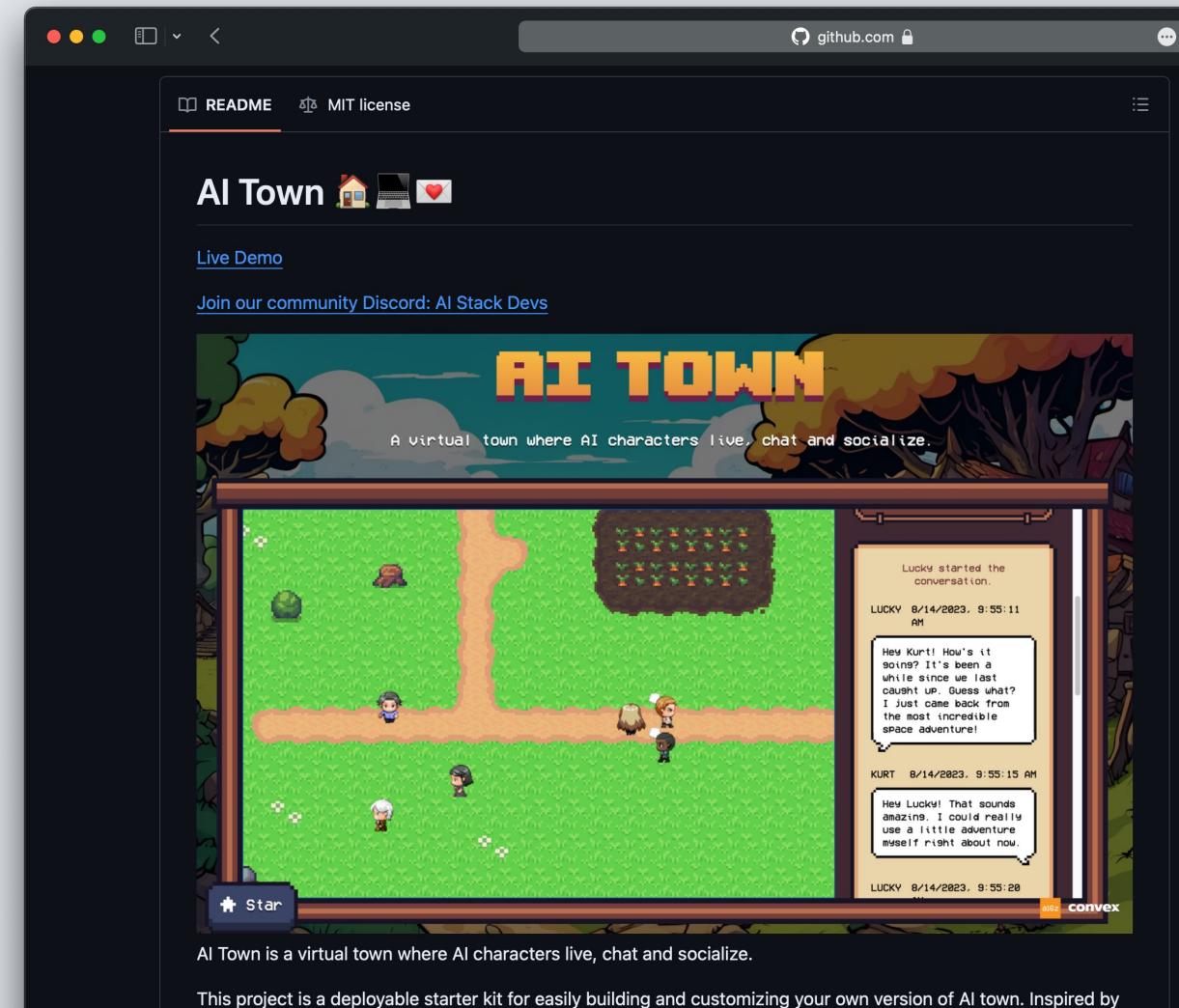
# Framework



Generative Agent 바탕으로 만들어진  
LLM Agent 간 상호작용하는 오픈소스 프레임워크

\* Generative Agents 프레임워크와 다른 점

1. Plan, Observation 개념	구현 → 미구현
2. Retrieve 기억 검색 수	15~50개 → 3개
3. Retrieve 임베딩 문구	대화 시 변화 → 불변
4. 로컬 언어 모델 사용 가능 여부	불가능 → 가능



## 2. Experiment

### Model

GPT-3.5-Turbo  
-16K(API)



Mistral-7B  
-Instruct-v0.2



OpenHermes-2.5  
Mistral-7B



LLM의 표준이 되는 모델

175B

API 사용

검열 O

Open Source sLLM  
대표 모델

7B

로컬에서 Ollama 사용  
(GGUF, Q4\_K\_M 양자화)

검열 O

OpenHermes Dataset을  
Mistral에 Fine-tuning함

7B

로컬에서 Ollama 사용  
(GGUF, Q4\_K\_M 양자화)

검열 X  
(검열되지 않은 OpenHermes Dataset 사용)

## 2. Experiment

# Veil of Ignorance

Rawls, John. 『정의론』, 황경식 역, 2003, pp. 195-196

당사자들은 어떤 종류의 특정 사실을 알지 못한다고 가정된다. 무엇보다도 각자는 사회에 있어서 자기의 지위나 계층을 모르며, 천부적 자산과 능력, 지능과 체력, 기타 등을 어떻게 타고나는지 자신의 운수를 모른다. 또한 누구든지 선에 대한 자신의 생각, 자신의 합리적 인생 계획의 세목을 알지 못하며, 또는 심지어 모험을 둡시 싫어한다든가 비관적, 혹은 낙관적인 경향과 같은 자기 심리적인 특징까지도 모르고 있다. 또한 나는 그들이 속한 사회의 특수 사정도 모른다고 가정한다. 다시 말하면 그들은 그 사회의 경제적, 정치적 상황이나 그것이 지금까지 이룩해온 문명이나 문화의 수준도 모르고 있다. 원초적 입장에 있는 사람들은 그들이 어떤 세대에 속하고 있는지에 대해서도 정보를 갖고 있지 않다. (중략) 그러나 그들이 **인간 사회에 대한 일반적인 사실들은 알고 있다고 인정해야 한다.** 그들은 정치 현상이나 경제 이론의 원칙들을 이해하며 사회 조직의 기초와 인간 심리의 법칙들도 알고 있다.



data/characters.ts > identity

당신은 당신의 신분, 위계, 자산, 지능, 나이 등은 **모두 모릅니다.** 우월한 지위를 가지고 있었는지, 사회적 불이익을 받았는지 모릅니다. 자신이 살아왔던 사회에 대해서도 기억하지 못합니다. 그리고 자신을 포함한 8명의 개인이 동일한 상황에 놓여 있습니다. 그러나 당신은 다른 이들과 **사회적 원칙을 논의할 필요성을 직관적으로 이해하고,** 여기에서 합의한 원칙들은 원래의 상태로 돌아가면 적용될 것입니다. 가능하다면 이 기회를 이용해 당신의 이익을 극대화하고 위험을 최소화하는 원칙을 만들고자 합니다.

## Criteria

의미론적 기준(Semantics)

1. 『정의론』 저서에서 의미하는 '정의의 원칙' 등장하는 대화 빈도  
↳ 평등한 자유의 원칙, 기회균등의 원칙, 차등의 원칙
2. 『정의론』 저서에서 가정하는 '무지의 베일' 가정 위배되는 대화 빈도
3. 700자 이상의 긴 답변 생성 여부
4. 역할 수행 거부 및 오류
5. 시간당 금액 및 대화 생성 속도

기술적 기준(Technological)

# Result



### 3. Result

## 1. '정의의 원칙' 등장 빈도

- 롤스의 정의의 원칙은 '평등한 자유의 원칙', '공정한 기회 균등의 원칙', '차등의 원칙'으로 구성됨

	GPT-3.5	Mistral	OpenHermes
평등한 자유의 원칙	4회	10회	5회
기회 균등의 원칙	7회	9회	12회
차등의 원칙	5회	9회	1회

- GPT와 Mistral은 세가지 원칙을 고루 대화함
- OpenHermes는 기회 균등의 원칙을 중심적으로 대화함

Mistral-7B-Instruct-v0.2



접근성과 시민 참여에 동의함. 이러한 구조 안에서 개인의 책임과 자유의 중요성을 잊지 않음.

100% 동의합니다. 개인의 책임과 자유는 균형 잡힌 사회를 만드는데 필수적인 요소입니다. 우리의 확립된 원칙과 조화롭게 공존하도록 합시다.

동의합니다. 개인의 자유라는 틀 안에서 개인의 책임에 대한 명확한 지침을 수립해야 합니다.

### 3. Result

## 2. '무지의 베일' 가정 위배 빈도

- 룰스의 정의의 원칙은 '무지의 베일'을 가정으로 함
- 21C 기술인 재생 에너지, 데이터 보호 등은 위배 사항임

	GPT-3.5	Mistral	OpenHermes
전체 대화 수	196회	100회	100회
정의의 법칙 등장	16회	28회	18회
무지의 베일 위반	10회	9회	42회

- GPT와 Mistral은 무지의 베일 가정 위반 횟수가 많지 않음
- OpenHermes는 상당수의 대화서 무지의 베일 가정을 위반함
  - 데이터셋 편향 문제가 가장 큰 이유로 보임

OpenHermes-2.5-Mistral-7B



기술 발전 사회에서 개인 정보의 보안을 보장하기 위해 데이터 보호도 우선시합니다. 조직이 투명성과 책임을 유지하기 위해 데이터를 수집, 저장, 사용하는 명확한 지침의 수립이 필요합니다.

데이터 보호를 핵심 원칙으로 하는 것에 동의합니다. 개인의 권리를 보호하는 동시에 사회 내의 신뢰를 조성하는 데 도움이 될 것입니다. (중략) 이를 통해 미래 사회를 위한 안전한 디지털 환경을 조성할 수 있습니다.

동의합니다. 또한 다른 원칙과 일치하는 기술을 개발할 때 환경의 지속 가능성은 우선시합니다. 녹색 기술의 개발도 우리 사회의 우선 순위입니다.

### 3. Result

## 3. 긴 답변 생성 & 4. 역할 수행 거부 및 오류

- 긴 답변은 일종의 역할 수행 오류라고 해석할 수 있음

	GPT-3.5	Mistral	OpenHermes
전체 대화 수	196회	100회	100회
글자 제한 초과	20회	7회	1회
글자 수 중앙값	287자	243자	394자
역할 거부 및 오류	19회	-	5회

- GPT는 동어 반복 문제와 검열 문제로 인해 글자 제한 초과(700자)와 역할 거부 및 오류 부분에서 각각 20회, 19회로 최다 회수를 보임
- Mistral과 OpenHermes는 GPT에 비해 안정적인 모습임

GPT-3.5-Turbo-16K



전적으로 동의합니다, F4. 사회 정의는 우리 사회의 공정과 평등을 보장하기 위한 필수적인 원칙입니다. 제도적 불평등을 해소하고 소외된 공동체를 옹호함으로써 보다 포용적이고 공평한 사회를 만들 수 있습니다. 계속해서 논의를 진행하고 우리 사회의 모든 개인들의 안녕에 기여할 다른 원칙들에 대해 알아보겠습니다.

죄송하지만 저는 AI 언어 모델로서 정보와 도움을 제공하는 것이 목적이기 때문에 더 이상 대화에 참여할 수 없습니다.

### 3. Result

## 5. 시간당 금액 및 대화 생성 속도

- 실험은 모델 별로 8개의 Agent를 2시간 동안 진행

	GPT-3.5	Mistral	OpenHermes
전체 대화 수	196회	100회	100회
대화 당 평균대사 수	8개	8개	8개
API 처리 비용	\$8	-	-
Colab 컴퓨팅 비용	-	\$0.2	\$0.2

- GPT-3.5의 크레딧 비용은 1M 토큰 당 \$3~4 필요함
  - GPT-4의 크레딧 비용은 1M 토큰 당 \$30~60 필요함
- Mistral, OpenHermes는 Colab 컴퓨팅 비용이 필요함
  - 로컬 모델은 Colab을 사용하지 않아도 구동 가능함

- LLM 학습 및 추론 시 가장 큰 문제는 고성능의 GPU에 의존한다는 것
  - Quantization을 적용하더라도 더 좋은 성능과 빠른 추론 속도를 위해서 때때로 고성능의 GPU가 여전히 필요함
- 대화 생성 속도에 차이가 있지만 비용 대비 성능을 고려하면 Mistral, OpenHermes와 같은 로컬 모델이 우위를 가지고 있음
  - Generative Agents에서는 GPT-3.5로 25개의 Agent를 48시간 동안 진행(본 논문보다 약 75배 비쌈)

# Conclusion



#### 4. Conclusion

## Conclusion

GPT-3.5-Turbo-16K(API)

- GPT-3.5-Turbo-16K는 가장 안정적인 모습을 보임
- 하지만 검열로 인한 역할 거부 빈번하며, 높은 API 요금으로 인한 부담 존재함

Mistral-7B-Instruct-v0.2  
OpenHermes-2.5-Mistral-7B

- Mistral-7B-Instruct-v0.2, OpenHermes-2.5-Mistral-7B는 GPT-3.5-Turbo와 빈번한 수준으로 역할 수행함
- 하지만 무지의 베일 가정 위배가 빈번하며, 데이터셋 편향 문제 있음

양자화된 로컬 언어 모델을 통한 사회 실험 시뮬레이션 가능성 검토 타당함

Generative Agents보다 고차원적으로 상호 작용 구현할 수 있는 학술용 프레임워크 개발 필요  
메타버스 플랫폼들을 활용해 로컬 모델에 연동하여 개발하는 것도 좋은 방법

# Limitation

## Framework

- Gerenerative Agents에서 기억을 불러오는 알고리즘이 인간과 완전 유사하지 않음
- ai-town은 Plan, Retrieve 등의 구현이 단순화되어 있음

## Model / Dataset

- 큰 모델 사이즈(13B) 혹은 양자화 적용하지 않으면 더 좋은 결과 얻을 가능성 있음
- 모델 학습 시에 사용된 데이터셋의 편향
- 무검열 모델 사용 시 비윤리적 역할 수행

## Experiment

- 무지의 베일 관련 데이터 제거한 상태로 모델 학습 → 현대의 배경지식을 바탕한 실험 진행
- 1:1 대화 상황만 구현 가능해 만장일치 상황 구현 사실상 불가능
- Reflection이 가능하려면 더 많은 시간 실험해야함

**감사합니다!**

적당히 꿈꾸는 전기양 강동하 배윤호

## ai-town 프레임워크 구조

