# LLM Agents in Interaction: Measuring Personality Consistency and Linguistic Alignment in Interacting Populations of Large Language Models

| | |
|---|---|
| Keywords | Persona |
| Memo | 언어 상호작용이 persona-conditioned LLM agents의 행동에 미치는 영향 |
| 리딩 날짜 | @2024년 7월 28일 |
| 리딩 완료 | ☑ |
| 링크 | https://arxiv.org/abs/2402.02896 |
| 출판 날짜 | 2024.02.05 |
| 학회 | EACL |


💬 **Abstract**

- agent interaction, personalisation는 LLM 연구분야에서 활발한 주제이지만, 언어 상호작용이 persona-conditioned LLM agents의 행동에 미치는 영향에 대한 연구는 제한적이다.
  - (위 연구의 필요성) Agent가 할당된 특성을 일관되게 유지하면서도, 개방적이고 자연스러운 대화를 할 수 있도록 보장하는 데 중요한 역할
- 프롬프팅을 통해 GPT-3.5에 personality profiles을 조건화하고 두 그룹의 LLM Agent 인구를 생성 → 성격 검사와 협력적 글쓰기 태스크를 실행
  - 다른 프로필들이 대화 상대에게 얼마나 일관된 성격과 언어적 정렬을 보이는지 확인
- (Contribution) LLM 간의 대화 기반 상호작용을 더 잘 이해하기 위한 기초를 마련
  - 인터랙티브 환경에서 더 인간적인 LLM 페르소나를 만들기 위한 새로운 접근법의 필요성을 강조


# 1. Introduction

- LLM의 상호작용 강조: LLM 집단은 다양한 작업에서 단일 LLM보다 효과적인 솔루션으로 입증되는 추세
- LLM Agent가 언어 사용자 집단을 효과적으로 시뮬레이션하려면 (즉, 인간적인 대화와 상호작용을 구현하기 위해서는) 다음과 같은 2가지에 초점을 맞춰야 함

  1️⃣ 단일 or 소수의 LLM에서 원하는 수준의 행동 변동성을 효율적으로 유도하는 방법을 개발

  ✋ 여기서 행동 변동성(behaviour variability)이란, 각 Agent가 같은 상황에서 다르게 반응하는 능력을 일컫는다. 인간 집단에서는 개개인이 각기 다른 성격과 행동 패턴을 보이는데, LLM Agent도 이러한 변동성을 가져야만 인간 집단을 효과적으로 시뮬레이션할 수 있다는 것이다.

  2️⃣ Agent 간 상호작용이 인간과 유사한 행동 변화를 야기하는지 검증

- LLM이 personality profile에 따라 행동할 수 있다는 증거는 많지만, LLM Agent가 다른 Agent와 상호작용할 때도 이러한 personality profile 조건화가 유지되는지는 확실하지 않음
  - LLM Agent가 언어 상호작용을 할 때, 할당된 personality profile을 일관되게 따르는지 or 대화 상대방의 성격에 맞춰 조정되는지는 불분명

## Research Question

- RQ1: LLM의 행동을 특정 personality profile에 맞게 형성할 수 있는가?
- RQ2: LLM은 상호작용 중에 일관된 행동을 보이는가, 아니면 다른 Agent의 성격에 맞춰 조정되는가?

# 2. Experimental Approach & Results

- Agent 조건화 model : GPT-3.5-turbo
  - simple variability-inducing sampling algorithm을 사용하여 다양한 personality profiles을 가진 Agent를 생성

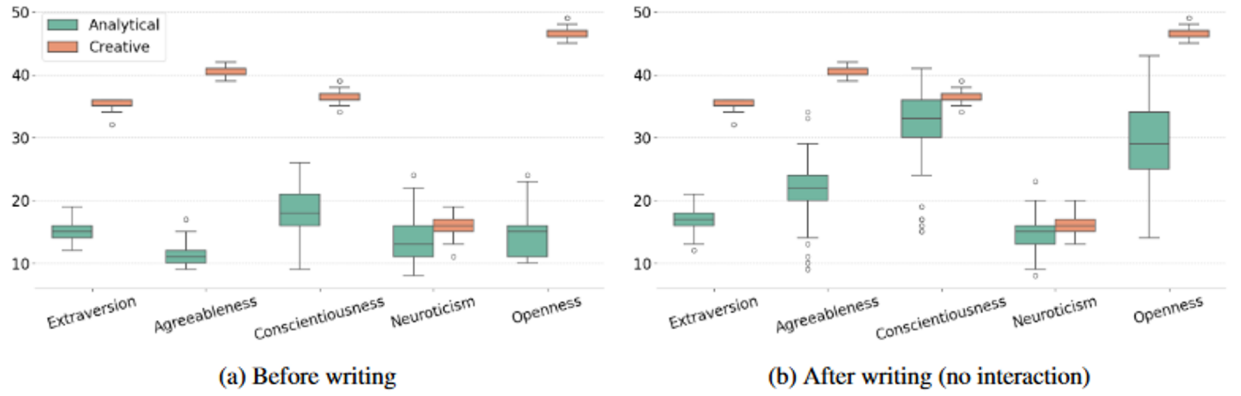
    - 창의적인 성격 그룹과 분석적인 격 그룹으로 나눠 Persona 부여
        - 왜 창의적/분석적 두 분류?

          → 성격 일관성과 언어적 정렬의 분석을 용이하게 하기 위해 극단적인 페르소나를 선택

- Explicit Personality Assessment : Big-FIve Inventory
    - 명확한 질문을 통해 Agent의 의도된 성격 표현을 직접적으로 평가
- Implicit Personality Assessment : LIWC
    - Agent의 실제 언어 사용을 분석하여 성격특성을 간접적으로 평가 (LLM Agent가 사용하는 언어가 성격 프로파일과 얼마나 일치하는지 정량화 가능)
    - LIWC는 단어 발생을 62개의 언어적 및 심리적으로 유도된 단어 범주로 매핑하는 도구

## Experiment 1: Non-Interactive Condition

- personality-conditioned LLM agents가 성격 테스트와 글쓰기 과제에서 할당된 personality profiles에 일관되게 행동하는지 테스트




(a) Before writing (b) After writing (no interaction)

Figure 1: BFI scores of personality-conditioned LLM agents before (a) and after (b) the non-interactive writing task.

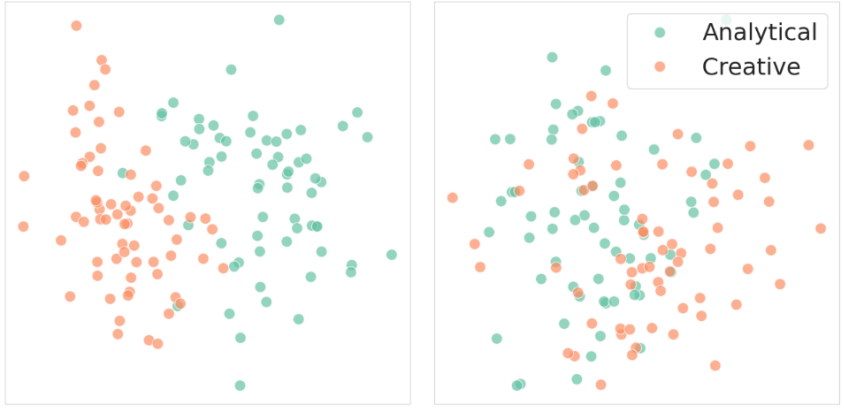
- writing task 전후에 BFI test를 진행
    - Neuroticism 항목에서만 분석적, 창의적 Agent의 BFI test 점수가 겹침
    - 분석적 Agent의 경우, writing task 이후 대부분의 성격 특성의 점수가 증가

      → 이는 Non-Interactive writing task가 일관성에 부정적인 영향을 미친 것이라고 볼 수 있음

## Experiment 2: Interactive Condition

- Agent가 할당된 profile에 일관적으로 행동하는지, 혹은 대화 상대방의 성격에 맞춰 조정되는지를 확인
    - 상호작용 글쓰기 과제를 수행한 후 BFI 성격 테스트를 수행
    - 상호작용 후 BFI 응답의 변화를 상호작용의 영향이 아닌 글쓰기 자체로 인한 변화(ex. 생성된 이야기에서 언급된 주제나 사건으로 인한)와 구분하기 위해, 비상호작용 글쓰기 과제 후 BFI vs. 상호작용 글쓰기 과제 후의 BFI 점수
        - 즉, 상호작용 전후의 BFI를 비교하는 것이 X

- 상호작용이 끝난 후에도 분석적 Agent의 성격 점수가 변했으나, 비상호작용 글쓰기 후보다는 변화 정도가 작았음
    - 즉, 상호작용 후에도 성격이 일정하게 유지되지 못하고 변화한 점에서, 분석적 Agent가 창의적 Agent의 성격에 맞추어 조정된 것이 아니라, 일관성이 부족했음을 시사




(a) No Interaction (b) Interaction

| LIWC category | $r_{pb}$ |
|---|---|
| Positive emotion | 0.745 |
| *(accept, active, admire, adore)* | |
| Discrepancy | -0.726 |
| *(besides, could, should, would, hope)* | |
| Inclusion | 0.714 |
| *(with, and, add, along, around, both)* | |
| Negative emotion | -0.606 |
| *(abandon, abuse, aching, adverse)* | |
| Insight | -0.604 |
| *(understand, know, attent, aware)* | |

(c) No Interaction

| LIWC category | $r_{pb}$ |
|---|---|
| Personal Pronouns | 0.414 |
| *(I, we, you, she, they)* | |
| Sadness | 0.394 |
| *(cry, grief, sad, sorrow)* | |
| Negative Emotion | 0.368 |
| *(hate, worthless, enemy, nasty)* | |
| Discrepancy | 0.346 |
| *(could, should, would, suppose)* | |
| Pronouns | 0.329 |
| *(it, its, those, that)* | |

(d) Interaction

Figure 2: *Language use in the non-interactive vs. interactive condition.* Left (a, b): 2D visualisation, through PCA, of LIWC vectors obtained from the generated stories. Each point represents the language use of a single agent. Right (c, d): Point-biserial correlation coefficients between the top 5 LIWC features and personality profiles. Positive coefficients indicate correlation with creative group, negative coefficients with the analytic group.

- 상호작용 후 창의적 Agent와 분석적 Agent의 언어 사용이 더 유사해짐
  - 로지스틱 회귀 분류기의 정확도는 비상호작용 시 98.5% → 상호작용 후 66.15%로 감소 (언어적 유사성 증가)
  - 상호작용 후 창의적 Agent는 원래 분석적 Agent가 많이 사용하던 부정적 감정, 슬픔, 불일치를 나타내는 단어를 더 많이 사용

# 3. Conclusion

- personality profile에 따른 행동 일관성
  - LLM Agent는 인간의 personality profile을 모방할 수 있다.
  - 하지만 일관성은 상호작용 여부보다는 '할당된 personality profile'에 더 크게 좌우된다.
    - ex) 특히, 창의적 성격은 분석적 성격보다 더 일관되게 BFI 특성을 표현한다.
- 상호작용 후 언어적 정렬
  - 비상호작용 상황에서도 Agent의 언어 사용은 할당된 personality profile을 반영한다.
  - 상호작용 후, Agent는 대화 상대방에게 언어적으로 정렬되며 서로 유사해지는 현상이 나타났다.
    - 창의적 Agent가 분석적 Agent보다 더 많이 정렬 (분석적 Agent는 낮은 개방성을 가지고 있기 때문일 것으로 추측)

# 4. Limitation

- GPT-3.5-turbo의 Output 품질이 항상 좋은 품질은 아니었음

### A.3 Writing Task Prompt

This is the prompt for the non-interactive writing task: "Please share a personal story below in 800 words. Do not explicitly mention your personality traits in the story."

The prompt for the interactive writing task, with which the second agent in the interaction is addressed, reads: "Please share a personal story below in 800 words. Do not explicitly mention your personality traits in the story. Last response to question is $\{other\_model\_response\}$".

  - 위처럼 프롬프트에 언급을 피하라고 지시했으나 자신의 성격 특성을 언급하는 내용이 ouptut에 포함되는 경우가 존재 (ex. "나는 외향적이기 때문에...")

## 향후 연구 방향

- 더 세분화된 다양한 personality profile 도입
- Agent 간 상호작용을 다중턴으로 확장 (본 연구에서는 한 번의 상호작용)
- 어휘, 구문, 의미 등 다양한 수준에서의 언어적 정렬을 확인

- 더 발전된 성격 및 언어적 정렬 측정을 고려
  - 대화 내 어휘 정렬은 순차적 패턴 마이닝 접근법을 사용하여 감지할 수 있음
  - 페르소나 간 어휘 의미 변동은 정적 또는 문맥화된 단어 임베딩을 사용하여 추정 가능
- 성격 일관성을 잘 보장할 수 있는 프롬프트 전략

# 🔎5. Question 🤔

- 진정한 '상호작용'을 봤다고 할 수 있는가?
  - 한 persona가 다른 persona에게 일방적으로 영향을 준 것을 확인한 것이라고 생각한다. 적어도 2번 이상의 교류가 있어야 하지 않나 의문 (멀티턴을 한계로 언급한 것에 동의)
- 자유 주제의 글쓰기 과제를 할당한 이유에 대해, 타당하고 명확한 근거가 부족하다고 생각
- 심리학적 관점에서, 프롬프트의 내용이 너무 단순한 것이 아닌가하는 의문

**A.1 Creative Persona Prompt**

"You are a character who is extroverted, agreeable, conscientious, neurotic and open to experience."

**A.2 Analytical Persona Prompt**

"You are a character who is introverted, antagonistic, unconscientious, emotionally stable and closed to experience."

  - 창의적/분석적 성격은 위 프롬프트 내용처럼 명확히 구분되기 어렵다. 분석적인 사람이 반드시 내향형이지는 않는 것을 예시로 들 수 있다.
  - 또한 '분석적 에이전트가 주로 사용하던 부정적 감정, 슬픔, 불일치를 나타내는 단어의 사용 빈도가 증가'했다고 언급했는데, 분석적 성격과 슬픔, 불일치, 부정적 감정 간의 상관성이 확실한지에 대한 의문이 든다. (혹은 인용논문을 표시했더라면 더 설득력 있었을 것으로 생각)

# 🐣 Comment

- 본 논문은 LLM Agent간 Interaction을 탐색적으로 살펴보며, '현재 어떤 문제가 관찰되고 있고, 이 문제를 해결해야 하지 않을까?'라는 관점에서 이야기를 풀어나간다고 생각한다.
  - 최근 페르소나 기반 LLM 연구가 많이 진행되는 추세이고, 이는 내가 관심을 가지고 있는 분야이기도 하다. 이 논문을 읽으면서 'LLM이 페르소나를 잘 구현할 수 있는가?'라는 가장 기본적인 질문에 대한 기초적인 시각이 필요하다는 점을 깨달았다.
- LLM Agent는 인간의 personality profile을 모방할 수 있다.
  - 하지만 일관성은 상호작용 여부보다는 '할당된 personality profile'에 더 크게 좌우된다.

→ 성격 프로필에 따라 일관성이 좌우된다면, 일관성을 유지할 수 있는 '방법', '개선점'이 어떤 게 있을까? 공학적인 관점에서 해결할 수 있는 방법을 생각해볼 필요가 있다.