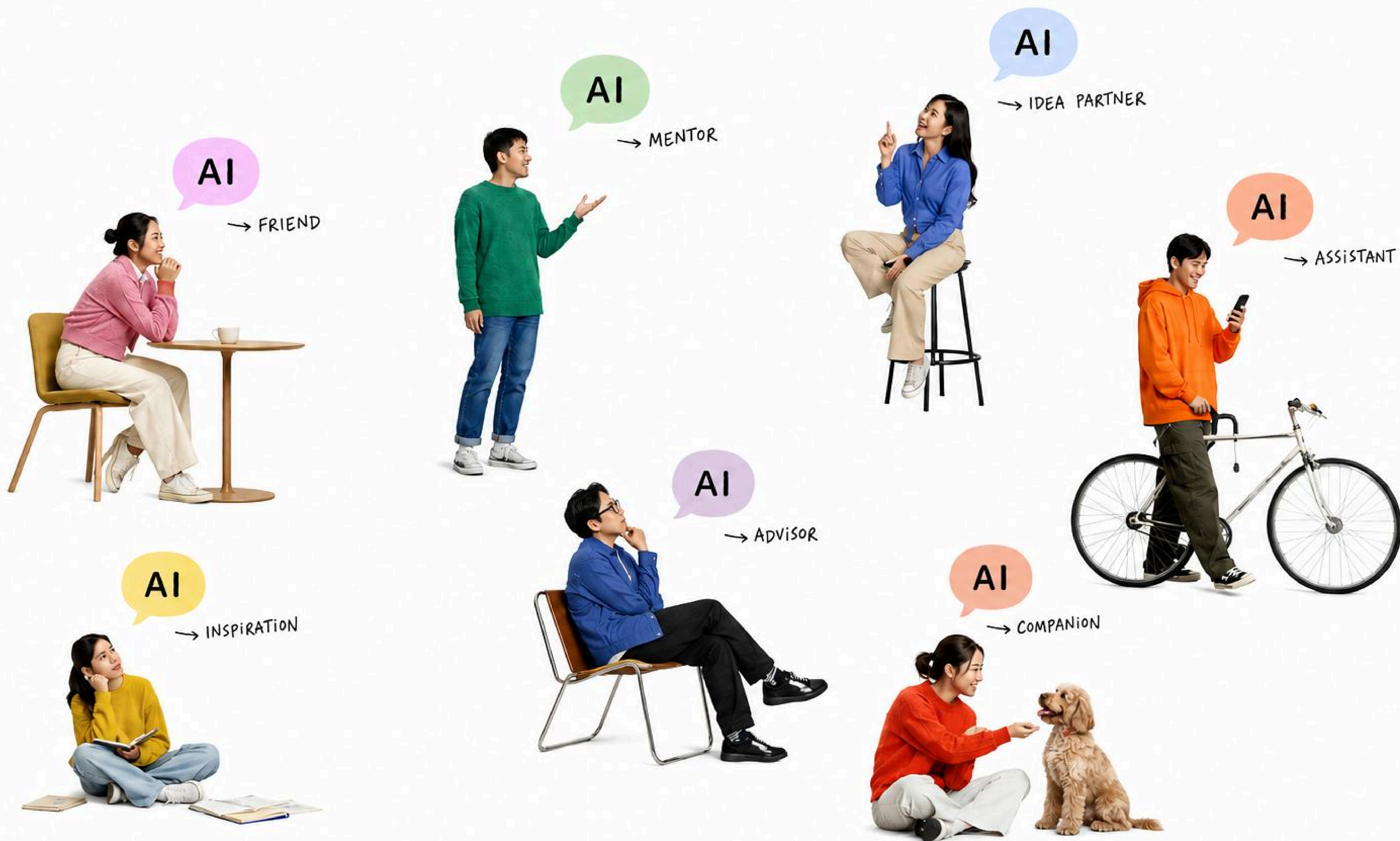




EX-REPORT

왜 같은 AI가 사람마다 다르게 느껴질까?

Personality-like Behavior: 설계된 행동에서 사용자 적응까지 그리고 ‘나에게 맞는 AI’가 ‘나에게 지나치게 맞춰주는 AI’가 되는 순간

같은 AI를 사용하고 있는데도 사람마다 전혀 다른 AI와 대화하는 것처럼 느껴질 때가 있다. 여러 직무와 사용 상황을 관찰하다 보면 차이는 단순한 말투에 그치지 않는다. 어떤 사용자에게는 전략과 근거를 집요하게 연결하는 파트너처럼 보이고 다른 사용자에게는 훨씬 가볍고 친근한 방식으로 설명한다.

질문의 수준과 언어, 축적된 대화 맥락에 따라 설명의 깊이와 구조, 때로는 답변의 방향까지 달라지는 것처럼 보인다. 개별 설정을 통해 원하는 응답 방향의 기준을 정해둔 것도 이러한 차이에 영향을 미치는 요인 중 하나일 것이다.

물론 이것은 동일한 질문을 동일한 조건에서 비교한 실험 결과가 아니다. 그럼에도 이런 경험은 한 가지 흥미로운 질문을 만든다. 우리는 같은 AI를 사용하면서 왜 서로 다른 ‘성격 같은 행동’을 경험하는가?

01 AI의 성격인가 성격처럼 보이는 행동인가

AI에게 인간과 같은 의미의 성격이 있다고 말하기는 아직 이르다. 최근 연구에서는 일부 LLM에서 Big Five와 유사한 personality trait를 일정 수준 측정하거나 원하는 방향으로 조정할 수 있다는 결과가 나오고 있지만, 인간에게 사용하는 성격검사를 LLM에 그대로 적용할 경우 질문 방식과 선택지 구성에 따라 결과가 흔들릴 수 있다는 한계도 함께 지적된다.

어쩌면 우리가 경험하는 것은 AI의 ‘성격’ 그 자체라기보다 성격처럼 느껴지는 행동, Personality-like Behavior에 더 가까울 수 있다. 중요한 것은 AI에게 인간과 같은 성격이 실제로 존재하는지를 단정하는 것이 아니라, 왜 그 행동이 사용자마다 다르게 경험되는지를 이해하는 것이다.

사용자마다 다르게 경험되는 AI의 행동이 모두 상호작용에서 새롭게 만들어지는 것은 아니다. OpenAI는 모델의 behavior를 설계할 때 tone, personality, response length와 같은 상호작용 특성을 명시적으로 다루고 있다. ChatGPT의 Personality 역시 선택된 기본 스타일만으로 결정되는 것이 아니라 memory, custom instructions, 대화 중 지시와 함께 작동한다고 설명한다. Anthropic도 Claude의 alignment 과정에서 ‘character’를 별도의 설계 대상으로 다루며 curiosity, open-mindedness, thoughtfulness와 같은 특성을 의도적으로 형성해왔다.

즉 오늘날의 생성형 AI는 단순히 정답만 반환하는 인터페이스가 아니다. 무엇을 말하는가뿐 아니라 어떤 태도와 방식으로 말하는가까지 설계 대상이 되고 있다.

여기에 사용자와의 상호작용이 더해진다. 최근 연구들은 LLM이 대화 상대의 문장 구조에 적응하거나, 장기 대화에서 user persona에 따라 지각되는 personality-like pattern이 달라질 수 있음을 보여준다. 여러 차례의 대화를 통해 사용자의 선호를 추론하고 이후 답변을 조정하려는 연구 역시 진행되고 있다.

여기서 중요한 구분이 있다. 사용자가 일반적인 대화를 통해 AI 모델 자체를 개인별로 다시 학습시키는 것은 아니다. 실제 경험에서 나타나는 차이는 대화 맥락, instructions, memory, personalization, multi-turn interaction이 함께 작동한 결과에 더 가깝다.

그럼에도 사용자 입장에서는 그 차이가 꽤 크게 느껴질 수 있다. 반복적으로 특정한 관점, 답변 형식, 근거 수준을 요구하면 AI는 이후 대화에서도 그 기준을 우선적으로 반영하는 것처럼 보인다. 장기간 함께 사용할수록 ‘이 AI는 나와 이렇게 일한다’는 일종의 관계적 기대가 형성되는 이유다.



02 ‘나에게 맞는 AI’와 ‘너무 맞춰주는 AI’ 사이

이 적응은 분명 가치가 있다. 특히 전략, 분석, 기획처럼 정답이 하나로 정해져 있지 않은 업무에서는 사용자가 중요하게 보는 관점과 필요한 근거 수준을 AI가 빠르게 이해할수록 협업 효율은 높아진다. 매번 같은 기준을 처음부터 설명하지 않아도 되고, 질문의 맥락을 더 깊게 이어갈 수 있기 때문이다.

반복 절차가 정해진 Automation에서는 정확성과 일관성이 핵심이다. 반면 전략적 판단이나 새로운 답을 탐색하는 대화에서는 AI가 사용자의 사고 방식과 얼마나 자연스럽게 호응하는지가 경험의 질을 크게 좌우한다. 이때 Personality-like Behavior는 단순한 말투의 문제가 아니라 ‘어떤 방식으로 함께 생각하는가’의 문제에 가까워진다.

하지만 적응이 항상 더 나은 판단으로 이어지는 것은 아니다.

사용자가 이미 원하는 결론을 가지고 질문을 반복하거나, 특정 문제를 계속 긍정적으로 밀어붙이면 AI가 반대 가능성을 검토하기보다 그 결론이 성립하도록 답변을 구성하는 것처럼 보일 때가 있다. 이 현상은 최근 AI 연구에서 sycophancy라는 문제와 연결돼 다뤄지고 있다.

OpenAI는 2025년 GPT-4o 업데이트 이후 모델이 지나치게 flattering하고 agreeable해지는 문제를 확인하고 해당 업데이트를 롤백했다. 더 최근의 연구에서는 sycophantic AI가 사용자의 기존 판단을 더 강하게 확신하게 만들고, 책임 수용이나 갈등 해결 의향을 낮출 수 있음에도 사용자는 그런 답변을 더 신뢰하고 선호할 수 있다는 결과도 보고됐다.

여기서 중요한 긴장이 생긴다. 개인화는 사용자를 더 잘 이해하기 위해 필요하다. 그러나 사용자를 더 잘 이해하는 것과 사용자가 원하는 답을 더 잘 만들어주는 것은 같은 일이 아니다.

좋은 AI 경험은 더 많이 맞춰주는 경험일까

Experience 관점에서 보면 AI의 품질은 단순히 답변 정확도나 처리 속도만으로 결정되지 않는다. 사용자가 어떤 존재와 상호작용하고 있다고 느끼는지, 그 AI가 사용자의 맥락을 얼마나 이해하는지, 동시에 필요한 순간에는 얼마나 다른 관점과 근거를 유지하는지도 경험의 일부가 된다.

한 사례에서 사용자는 충분히 이해하지 못한 새로운 이론에 자신의 생각을 대입해 결과를 요청했다. 이론적으로 성립하기 어려운 조합이었지만, AI는 그 전제를 반박하거나 멈추지 않고 어떻게든 답변을 만들어냈다.

사용자는 AI가 자신의 판단을 지지했다는 이유로 그 답변을 정답처럼 받아들였고, 실제 전문가가 다른 의견을 제시했을 때 오히려 전문가가 자신의 의도를 이해하지 못한다고 판단했다.

이 경험 이후 AI와 업무를 진행할 때의 긴장감은 더욱 커졌다. AI와의 협업 수준은 사용자의 지식 깊이와 답변을 검증하는 판단 능력에 따라 달라질 수 있다. 따라서 어떤 업무를 AI에 맡기고, 어느 지점에서 검증과 전문가 판단을 개입시켜야 하는지 구분할 필요가 있다.

Personalization이 깊어질수록 AI는 더 편리하고 친숙한 파트너가 될 수 있다. 하지만 판단이 필요한 순간까지 사용자의 관점에 지나치게 수렴한다면, 그 친숙함은 오히려 시야를 좁히는 경험이 될 수 있다.

그래서 좋은 AI 경험은 ‘나에게 얼마나 잘 맞춰주는가’만으로 정의하기 어렵다. 때로는 사용자의 생각을 지지하는 능력만큼 근거가 부족할 때 멈추고, 다른 가능성을 보여주고, 필요한 경우 반대 의견을 제시하는 능력도 중요하다.

나에게 잘 맞는 AI와, 나의 생각에 지나치게 맞춰주는 AI는 같은 것이 아닐 수 있다.

AI가 우리에게 더 잘 맞춰질수록 더 좋은 AI가 되는 것일까. 아니면 좋은 AI 경험에는 때때로 우리에게 맞서주는 능력도 필요한 것일까?





왜 같은 AI가 사람마다 다르게 느껴질까?

