멀티모달 AI 에서는 문제가 더 복잡해진다

한소연

우리가 뉴스 기사를 읽을 때는 글에 적힌 내용만 살펴보지 않는다. 함께 실린 사진이나 그래프를 확인하고, 때로는 사진 속 인물의 표정이나 주변 상황을 통해 기사에서 미처 설명하지 않은 내용까지 짐작한다. 누군가와 대화할 때도 마찬가지다. 우리는 상대방이 하는 말뿐 아니라 목소리의 높낮이와 표정, 몸짓 등을 종합해 그 말의 의미를 이해한다.

최근 인공지능도 이처럼 여러 종류의 정보를 함께 처리하는 방향으로 발전하고 있다. 과거의 언어 모델이 주로 텍스트를 중심으로 작동했다면, 오늘날의 멀티모달 AI 는 이미지와 글, 음성, 영상, 문서의 구성과 배치 등 서로 다른 형태의 정보를 동시에 받아들이고 해석할 수 있다. 덕분에 AI 가 이해하고 수행할 수 있는 일은 훨씬 다양해졌지만, 그만큼 새로운 문제도 등장했다. 여러 정보가 서로 다른 이야기를 할 때, AI 는 과연 무엇을 근거로 판단해야 할까?

예를 들어 식당 메뉴판에는 '순한 맛'이라고 적혀 있지만, 함께 실린 음식 사진에는 매운 고추가 잔뜩 올라가 있을 수 있다. 누군가 "나 정말 괜찮아"라고 말하더라도 표정이나 목소리에서는 전혀 다른 감정이 느껴질 수도 있다. 사람은 이런 상황에서 여러 단서를 종합하고 맥락을 고려해 판단하려 한다. 그렇다면 AI 는 어떨까? 글과 이미지가 서로 다른 정보를 전달할 때 어느 쪽을 더 신뢰해야 하며, 말의 내용과 표정이 일치하지 않을 때는 무엇을 더 중요하게 받아들여야 할까?

최근 나와 동료들이 관심을 가지고 연구해 온 것도 바로 이러한 문제다. 우리는 AI 가 어떤 정보를 받아들이는지를 넘어, 최종 판단을 내릴 때 각 정보에 얼마나 의존하는지에 주목했다. 이를 정보에 대한 의존, 즉 *reliance* 의 문제라고 부른다.

멀티모달 AI 가 내놓는 하나의 답 뒤에는 실제로 여러 종류의 근거가 존재할 수 있다. 사진을 보고 질문에 답하는 경우라면 사진 속 사물뿐 아니라 사물 사이의 관계, 사람들의 행동, 색이나 옷차림 같은 외형적 특징, 주변 배경 등이 모두 판단의 단서가 될 수 있다. 이미지에 포함된 글자를 읽어야 한다면 문자 정보와 주변 장면을 함께 살펴야 할 수도 있고, 음성과 영상을 통해 사람의 감정을 판단할 때는 얼굴 표정이나 목소리의 억양, 움직임은 물론 화자의 외형이나 주변 배경까지 영향을 미칠 수 있다.

우리가 수행한 연구에서는 이러한 정보들을 의미, 관계, 행동, 외형, 배경, 문자, 음향 등의 기능적 범주로 구분해 살펴보았다. 물론 이는 AI 내부에 각 정보를 담당하는 완전히 독립된 영역이 존재한다는 뜻은 아니다. 여러 종류의 정보가 최종 판단에 얼마나 기여하는지 분석하고, 그 영향력을 조절하기 위해 정보 처리 과정을 구조화한 것이다.

이러한 관점에서 보면 AI 의 신뢰성을 이해하기 위해서는 단순히 어떤 정보를 인식할 수 있는지만 살펴보는 것으로는 충분하지 않다. **실제로 답을 만들어내는 과정에서 어떤 정보를 얼마나 중요하게 사용했는지**까지 확인할 필요가 있다.

같은 답을 하는 두 AI 가 사실은 다를 수 있다



이를 이해하기 위해 간단한 상황을 하나 생각해 보자. 두 개의 AI 에게 같은 사진을 보여주었다고 가정해 보자. 사진 속에는 케이크 앞에서 촛불을 끄는 사람이 있고, 주변 사람들은 환하게 웃으며 박수를 치고 있다. 두 AI 에게 지금 무슨 일이 벌어지고 있는지 물었더니, 둘 다 "생일을 축하하고 있습니다"라는 동일한 답을 내놓았다.

겉으로 보기에는 두 AI 모두 사진을 정확하게 이해한 것처럼 보인다. 하지만 같은 정답을 내놓았다고 해서 반드시 같은 근거를 사용했다고 볼 수는 없다.

첫 번째 AI 는 케이크 위의 촛불과 이를 끄는 행동, 주변 사람들이 축하하는 모습처럼 사람과 사물 사이의 관계 및 행동을 종합해 생일이라고 판단했을 수 있다. 반면 두 번째 AI 는 알록달록한 풍선이나 고깔모자, 벽에 걸린 장식처럼 생일 파티에서 흔히 등장하는 배경과 외형적 특징에 주로 의존했을 수도 있다. 평범한 생일 파티 사진에서는 이러한 차이가 드러나지 않는다. 두 AI 가 서로 다른 단서를 사용하더라도 결과적으로는 같은 답을 내놓기 때문이다.



그렇다면 사진 속 상황은 그대로 유지하면서 주변 환경만 바꾸면 어떻게 될까? 케이크와 촛불, 사람들의 행동은 그대로 두고 풍선과 장식을 없앤 뒤, 장소를 화려한 파티 공간에서 평범한 회의실로 바꾸어 보는 것이다.

이 경우 첫 번째 AI 는 여전히 촛불을 끄는 행동과 사람들의 반응을 근거로 생일을 축하하는 상황이라고 판단할 가능성이 높다. 반면 배경에 크게 의존했던 두 번째 AI 는 이전보다 확신을 잃거나, 아예 다른 상황이라고 판단할 수도 있다. 사진에서 실제로 벌어지는 일은 달라지지 않았지만, AI 가 중요하게 받아들였던 단서가 사라졌기 때문이다.

이 예시는 AI 의 판단을 이해하는 데 중요한 점을 보여준다. 처음 사진에서는 두 AI 모두 정답을 맞혔으므로 차이를 발견하기 어렵지만, 판단에 사용되었을 것으로 예상되는 정보를 바꾸어 보면 서로 다른 반응이 나타날 수 있다. 결국 AI 가 무엇을 근거로 답했는지는 정답만 확인해서는 알기 어려우며, 특정한 정보가 달라졌을 때 판단이 얼마나 흔들리는지를 함께 살펴볼 필요가 있다.

우리가 연구에서 주목한 것도 바로 이 지점이다. **같은 답을 유지하면서도 AI 가 그 답에 도달하기 위해 사용하는 근거를 바꿀 수 있을까?** 이 질문은 AI 가 얼마나 정확한 답을 내놓는지를 넘어, 그 판단이 어떤 정보에 의해 뒷받침되는지를 조절할 수 있는가에 관한 문제로 이어진다.

그렇다면 그것을 보지 말라고 말하면 되지 않을까?

오늘날 생성형 AI 는 사용자가 자연어로 전달하는 지시를 상당히 능숙하게 처리한다. 원하는 답변 형식을 지정하거나 특정한 관점에서 설명하도록 요청할 수도 있고, 어떤 정보에 집중해야 하는지 알려줄 수도 있다. 이러한 기능을 생각하면 AI 가 부적절한 정보에 의존하고 있을 때도 간단한 지시로 문제를 해결할 수 있을 것처럼 보인다.

가령 생일 파티 사진을 이해하는 AI 에게는 “배경은 무시하고 사람들이 무엇을 하고 있는지에 집중해”라고 말할 수 있다. 사람의 감정을 분석하는 AI 에게는 “외모보다는 표정과 목소리를 근거로 판단해”라고 요청할 수도 있다. 실제로 우리는 생성형 AI 를

사용할 때 질문의 표현을 바꾸거나 역할을 지정하고, 무엇을 중요하게 고려해야 하는지 설명하는 방식을 자주 활용한다. 이른바 프롬프트 엔지니어링이다.

프롬프트는 AI 의 행동을 조절하는 강력하고 편리한 도구다. 그러나 여기에는 한 가지 주의해야 할 점이 있다. **AI 가 지시에 맞는 답변을 생성했다고 해서 반드시 그 지시에서 강조한 정보를 중심으로 판단했다고 볼 수는 없다는 것이다.**

예를 들어 "행동에 집중하라"고 요청했을 때 AI 가 적절한 답을 내놓더라도, 그 판단은 여전히 배경이나 외형적 특징에 상당 부분 의존하고 있을 수 있다. 마찬가지로 "배경은 무시하라"고 지시한 뒤 답변의 표현이 달라졌다고 해서 실제 판단에서 배경의 영향이 기대한 만큼 줄어들었다고 보장할 수는 없다.

우리가 수행한 실험에서도 이러한 차이가 관찰되었다. 자연어 프롬프트만 제공한 경우에는 AI 내부의 정보 처리 경로와 각 정보에 대한 의존 정도가 상대적으로 적게 변화했다. 또한 특정 정보를 의도적으로 바꾸거나 교란했을 때, 내부의 정보 사용 방식까지 조절한 방법에 비해 판단의 안정성이 낮게 나타났다.

그렇다고 프롬프트가 효과적이지 않다는 뜻은 아니다. 프롬프트는 AI 가 어떤 내용을 말하고 어떤 형식으로 답해야 하는지 전달하는 데 매우 유용하다. 다만 답변의 내용이나 형식을 바꾸는 것과 답변을 만들어내는 데 사용되는 정보의 영향력을 바꾸는 것은 구분해서 생각할 필요가 있다. 사용자의 지시가 결과에 반영되는 것만으로는 그 결과에 이르는 판단 과정까지 원하는 방향으로 달라졌다고 확신하기 어렵기 때문이다.

지시와 습관

나는 이러한 차이를 사람의 '지시'와 '습관'에 비유하고 싶다.

누군가에게 "앞으로 사람을 첫인상으로 판단하지 마"라고 조언했다고 생각해 보자. 상대방은 그 말에 충분히 공감하고 앞으로 그렇게 하겠다고 진심으로 대답할 수 있다. 그러나 오랫동안 형성된 판단 습관이 그 한마디로 사라지는 것은 아니다. 겉으로 드러나는 표현은 달라질 수 있지만, 실제로 무엇에 먼저 눈이 가는지, 어떤 특징을 더 중요하게 받아들이는지, 여러 정보가 충돌할 때 무엇을 더 신뢰하는지는 쉽게 바뀌지 않을 수도 있다.

AI 에서도 이와 유사한 현상을 생각해 볼 수 있다. 프롬프트가 AI 에게 전달하는 지시라면, 학습 과정에서 형성된 정보 처리의 경향은 일종의 계산상의 습관에 가깝다. 물론 AI 가 사람처럼 의식적으로 습관을 형성하거나 특정한 믿음을 갖는다는 의미는 아니다. 학습 과정에서 반복적으로 유용했던 정보의 조합이나 패턴이 이후의 판단에도 영향을 미칠 수 있다는 뜻이다.

앞서 살펴본 생일 파티 사진을 다시 떠올려 보자. 수많은 사진을 학습하면서 풍선이나 화려한 장식이 생일 파티를 판단하는 데 유용한 단서였던 AI 에게 어느 날 "배경은 무시하고 사람들이 무엇을 하고 있는지만 봐"라고 지시한다고 가정해보자. AI 는 그 지시를 처리하고 그에 맞는 답변을 생성할 수 있지만, 그렇다고 해서 풍선이나 장식이 판단에 미치는 영향이 곧바로 사라지는 것은 아닐 수 있다.

이러한 비유는 프롬프트의 한계를 이해하는 데 도움이 된다. 사람에게 새로운 판단 기준을 설명하는 것과 오랫동안 형성된 판단 습관을 바꾸는 것이 서로 다른 일이듯, AI 에게 특정한 정보에 집중하라고 지시하는 것과 실제 계산 과정에서 그 정보의 영향력을 조절하는 것도 서로 다른 문제일 수 있다.

그렇다면 자연어로 무엇을 보라고 알려주는 데서 한 걸음 더 나아가, AI 가 실제로 사용하는 정보의 비중까지 바꿀 수는 없을까? 이 질문이 우리가 수행한 연구의 출발점이었다.

AI 의 '생각'을 통제한다는 것

여기에서 'AI 의 생각'이라는 표현은 조심해서 사용할 필요가 있다. 현재의 인공지능이 인간과 동일한 방식으로 생각하거나 의식적인 판단을 내린다고 볼 수는 없기 때문이다. 다만 AI 에도 입력된 정보를 처리하고 최종 결과를 생성하는 계산 과정은 존재한다. 이미지와 글, 음성 등의 정보는 모델 내부의 여러 층을 거치면서 표현되고 결합되며, 이 과정에서 어떤 정보는 더 크게 반영되고 다른 정보는 상대적으로 덜 반영된다. 이러한 계산의 결과가 최종 답으로 이어진다.

이 관점에서 보면 AI 를 조절한다는 것은 크게 두 가지로 나누어 생각할 수 있다. 첫 번째는 AI 가 내놓는 결과를 조절하는 것이다. 특정한 형식으로 답하도록 하거나 어떤 내용을 포함하거나 제외하도록 요청하는 방식이 여기에 해당한다. 우리가 일상적으로 사용하는 프롬프트의 상당 부분이 이러한 역할을 한다.

두 번째는 그 결과를 만들어내는 과정에서 어떤 정보를 사용할지 조절하는 것이다. 예를 들어 사진 속 풍선이나 장식보다 사람들이 실제로 무엇을 하고 있는지를 중요하게 고려하도록 하거나, 문서의 표를 분석할 때 표의 모양이나 위치보다 그 안에 적힌 숫자에 더 집중하도록 하는 방식이다. 사람의 감정을 판단할 때도 외형적 특징보다는 표정이나 목소리의 억양을 더 중요한 근거로 삼도록 할 수 있다.

두 번째 문제는 첫 번째보다 훨씬 어렵다. AI 가 어떤 답을 내놓아야 하는지 알려주는 것을 넘어, 그 답을 만들어내는 계산 과정에 영향을 주어야 하기 때문이다.

우리는 이러한 문제를 해결하기 위해 자연어 지시를 단순한 출력 명령이 아니라 AI 내부의 정보 흐름을 조절하기 위한 개입(*intervention*)으로 활용할 수 있는지 연구했다. 핵심은 모델 내부의 모든 계산을 하나의 덩어리로 취급하는 대신, 의미와 관계, 행동, 외형, 배경, 문자, 음향처럼 서로 다른 종류의 정보에 대응하는 작은 기능적 경로들을 추가하고, 자연어 지시에 따라 각 경로의 기여도를 조절하는 것이다.

예를 들어 사용자가 "행동에 집중하라"고 요청하면 행동과 관계에 관련된 경로의 기여를 높이고, "배경은 무시하라"고 요청하면 배경과 관련된 경로의 영향력을

낮추도록 설계할 수 있다. 이는 AI 에게 “정답은 생일 파티라고 말해”라고 알려주는 방식과는 다르다. 어떤 답을 내놓아야 하는지를 직접 지정하는 대신, 그 답을 판단하는 데 어떤 정보를 더 중요하게 활용해야 하는지를 조절하는 것이다.

실제 연구에서는 모델 전체를 처음부터 다시 학습시키지 않고 비교적 작은 추가 경로를 이용해 이러한 조절을 구현했다. 같은 과제와 같은 정답의 기준을 유지하면서도, 지시에 따라 서로 다른 종류의 정보가 최종 판단에 기여하는 정도를 바꾸는 것이 목표였다.

그러나 내부의 정보 처리 경로를 조절했다고 해서 곧바로 AI 의 실제 판단 근거가 달라졌다고 결론 내릴 수는 없다. 특정 경로가 더 강하게 작동하더라도 최종 답변에서는 여전히 다른 정보가 중요한 역할을 하고 있을 가능성이 있기 때문이다. 따라서 우리는 내부에서 나타나는 변화뿐 아니라, 그 변화가 실제 판단에 어떤 영향을 미치는지도 확인할 필요가 있었다.

이를 위해 사용자가 중요하게 보거나 무시하도록 지시한 정보를 의도적으로 바꾸거나 제거하는 실험을 수행했다. 배경에 덜 의존하도록 조절했다면 배경을 바꾸어보고, 외형적 특징의 영향력을 낮추도록 했다면 색이나 시각적 속성을 변화시키는 식이다. 이때 AI 의 답이 얼마나 흔들리는지를 살펴보면, 단순히 내부 경로의 작동 정도만 달라진 것인지, 아니면 실제 판단에서 해당 정보가 차지하는 비중도 달라졌는지를 분석할 수 있다.

실험에서는 특정 정보를 강조하도록 조절했을 때 그 정보가 교란되면 모델의 판단이 더 민감하게 변하는 경향이 나타났다. 반대로 사용하지 않도록 한 정보를 바꾸었을 때는 상대적으로 더 안정적인 판단을 보였다. 이는 자연어 지시에 따라 내부 경로의 기여도뿐 아니라, 최종 답변을 뒷받침하는 정보의 영향력도 함께 달라질 수 있음을 보여준다.

이러한 변화는 사진을 보고 질문에 답하는 시각적 질의응답뿐 아니라, 이미지 속 글자를 읽는 문제, 음성과 영상을 함께 이용해 감정을 판단하는 문제, 이미지를 문장으로 설명하는 문제 등 다양한 과제에서도 관찰되었다. 서로 다른 정보를 함께 처리하는 AI 에서 어떤 정보에 의존할 것인지를 조절하는 접근이 여러 상황에 적용될 가능성을 확인한 것이다.

물론 이러한 결과가 AI 가 언제나 올바른 근거로 판단하게 되었다는 뜻은 아니다. 우리가 구분한 의미, 행동, 외형, 배경 등의 범주도 어디까지나 인간이 정의한 기능적 구분이다. 실제 AI 내부에서 정보들이 어떻게 결합하고 상호작용하는지를 완벽하게 설명하거나, 각각의 경로가 완전히 독립된 인과적 메커니즘이라고 주장할 수는 없다. 연구에서 사용한 경로들은 정보에 대한 의존을 실험적으로 분석하고 조절하기 위한 구조화된 장치로 이해하는 것이 적절하다.

그럼에도 이러한 연구는 AI 의 신뢰성을 바라보는 새로운 가능성을 제시한다. 지금까지 우리는 AI 가 내놓은 답을 보고 그 이유를 사후적으로 설명하려는 데 많은 관심을 기울여왔다. 앞으로는 여기서 더 나아가, **AI 가 판단에 사용하는 정보 자체를 조절하고 그 효과를 검증하는 것**도 중요한 연구 과제가 될 수 있다.

정답의 시대에서 근거의 시대로

인공지능 연구에서 정확도는 앞으로도 중요한 평가 기준으로 남을 것이다. 정답을 제대로 맞히지 못하는 AI 라면 아무리 그 판단 근거가 적절하더라도 실질적인 도움을 주기 어렵기 때문이다. 그러나 그 반대 역시 마찬가지다. 높은 정확도를 달성했다는 사실만으로 AI 가 언제나 신뢰할 만한 방식으로 판단한다고 보장할 수는 없다.

이러한 차이는 AI 가 실험실의 벤치마크를 넘어 실제 사회의 의사결정 과정에 참여할수록 더욱 중요해진다. 예를 들어 의료 영상을 분석하는 AI 가 높은 진단 정확도를 보이더라도, 실제 병변의 특징이 아니라 병원마다 다른 촬영 방식이나 장비의 특성에 의존하고 있다면 새로운 환경에서 예상치 못한 오류가 발생할 수 있다. 채용을 돕는 AI 역시 지원자의 업무 능력보다 직무와 관련이 없는 특징을 판단 근거로 삼는다면, 겉으로 드러나는 성능만으로는 그 판단을 충분히 신뢰하기 어렵다.

이미지와 텍스트, 음성과 영상을 함께 처리하는 멀티모달 AI 에서는 이러한 문제가 더욱 복잡해진다. 서로 다른 정보가 일치할 때는 모델이 무엇에 의존하는지 알아차리기 어렵지만, 정보가 충돌하거나 특정한 단서가 사라졌을 때는 그동안 드러나지 않았던 판단의 취약성이 나타날 수 있기 때문이다. 따라서 AI 가 어떤 정보를 활용하는지 살펴보는 것과 함께, 우리가 특정한 정보를 사용하지 말라고 요구했을 때 실제로 그 영향이 줄어드는지도 확인해야 한다.

지금까지 인공지능의 발전을 평가하는 대표적인 방법은 일종의 시험을 치르게 하고 몇 문제를 맞혔는지 확인하는 것이었다. 이러한 평가는 기술의 발전을 측정하는 데 중요한 역할을 해왔으며 앞으로도 필요할 것이다. 다만 AI 가 더욱 복잡한 상황에서 판단을 내리고 그 결과가 사람들의 삶에 영향을 미치게 된다면, 평가의 범위 역시 넓어져야 한다. AI 가 왜 그런 답을 내놓았는지, 실제로 어떤 정보를 사용했는지, 환경이 달라져도 적절한 근거에 의존할 수 있는지, 그리고 필요하다면 그 정보 사용 방식을 바꿀 수 있는지까지 살펴볼 필요가 있다. 단순히 정답을 맞혔는지를 확인하는 데 그치지 않고, 그 답이 얼마나 신뢰할 만한 판단 과정에서 나왔는지를 함께 평가해야 한다는 뜻이다.

사실 이러한 기대는 우리가 인간의 전문적인 판단에 요구하는 기준과 크게 다르지 않다. 우리는 좋은 의사에게 단순히 진단을 잘 맞히는 능력만을 기대하지 않는다. 어떤 증상과 검사 결과를 근거로 진단했는지, 그리고 그 근거가 충분하고 적절한지도 중요하게 생각한다. 과학자의 연구 결과를 평가할 때도 결론 자체뿐 아니라 그 결론을 뒷받침하는 증거와 연구 방법을 살펴본다. 좋은 판단에는 그 판단을 뒷받침할 좋은 근거가 필요하기 때문이다.

AI 가 우리 사회의 다양한 판단 과정에 참여하게 된다면, 이러한 기준은 점점 더 중요해질 것이다. 물론 AI 의 계산 과정을 인간의 사고와 동일하게 바라볼 수는 없으며, AI 가 사용하는 모든 정보를 완벽하게 설명하거나 통제하는 것 또한 아직은 어려운 과제다. 그럼에도 정확도를 높이는 것과 함께 판단에 사용되는 정보의 적절성을 확인하고, 필요할 때 그 영향력을 조절할 수 있도록 만드는 연구는 신뢰할 수 있는 인공지능을 향한 중요한 방향이라고 생각한다.

앞으로 우리가 필요로 하는 AI 는 단순히 더 많은 문제에서 정답을 맞히는 시스템만은 아닐 것이다. 중요한 정보와 그렇지 않은 정보를 구분하고, 적절한 근거에 의존해 판단하며, 상황에 따라 그 정보의 사용 방식을 조절할 수 있는 AI 가 더욱 중요해질 것이다. 이러한 능력은 AI 가 인간을 대신해 판단하는 것이 아니라, 인간이 그 판단을 이해하고 검증하며 필요한 경우 바로잡을 수 있도록 하는 데서 의미를 갖는다.

결국 AI 의 신뢰성을 평가하기 위해서는 지금까지 익숙했던 질문에 하나를 더해야 한다. AI 가 맞는 답을 내놓았는지를 확인하는 데서 나아가, 그 답에 도달하는 과정에서 적절한 근거를 사용했는지까지 살펴야 한다는 것이다.

AI 는 맞는 답을 했는가? 그리고 그 답을 알맞은 이유로 내놓았는가?