KAIST AI X 실패 공모전 아이디어 기획서

팀명	팀원	
AI 탐험가	팀장	형유림
아이디어명	팀원	강희주
AI 종 도감 : 실패에서 찾은 공존의 언어	팀원	이승원
	팀원	정수민

기획 배경 및 문제 정의

1. 문제 상황

AI 실패는 이제 일상이 되었습니다. 1 많은 사용자들이 다양한 형태의 AI와 협업하면서 반복적인 좌절을 경험하고 있습니다.

대화형 AI 사용자들은 프롬프트를 여러 번 수정해도 원하는 답을 얻지 못하거나, 자신감 있는 답변이 사실과 다른 환각 현상을 마주합니다. Stanford 연구에 따르면 GPT-3.5의 환각률은 99.6%, GPT-4는 28.6%에 달합니다. 실제로 2025년 노르웨이에서는 ChatGPT가 실존 인물을 아동 살인범으로 날조했고, 미국과 한국에서는 변호사가 AI 생성 가짜 판례를 법정에 제출하여 징계를 받았습니다.

이미지 생성 AI는 손가락이 6개인 인물을 그리거나, 한글 텍스트를 알아볼 수 없게 왜곡합니다. "암호 학회 포스터"를 요청했을 때 글자가 깨지는 현상이 빈번하며, 특히 비영어권 문자는 제대로 학습되지 않아 실무 활용에 큰 장애가 됩니다.

추천 알고리즘은 우연히 본 고양이 영상 하나 때문에 몇 주간 피드를 고양이 콘텐츠로 채우거나, 일 회성 검색을 영구적 선호로 오해하여 수개월간 원치 않는 광고를 노출시킵니다. 이는 사용자를 알고 리즘이 만든 좁은 세계에 가두는 필터 버블을 형성합니다.

이러한 실패들은 단순한 불편함을 넘어 법률·의료 분야의 오판, 전문가의 시간 낭비, 정보 편식과 확증 편향 강화 등 실질적 피해로 이어지고 있습니다. AI 실패는 예외가 아닌 일상이며, AI 유형에 관계없이 공통적인 근본 원인에서 비롯됩니다.

2. 근본 원인: 종 간 언어 차이

AI는 인간과 다른 세계에 살고 있습니다. AI는 "이게 사실인가?"가 아니라 "이 맥락에서 이 단어가 나

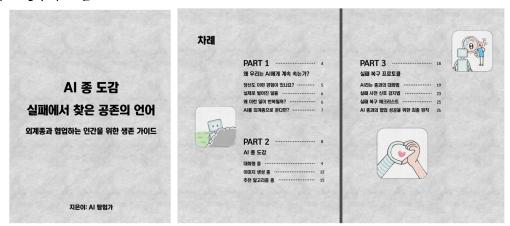
¹ https://tech.co/news/list-ai-failures-mistakes-errors

올 확률이 높은가?"를 계산합니다. 같은 언어를 쓰지만, 정보를 받아들이는 방식이 근본적으로 다릅니다. 기존 관점에서는 AI 실패를 기술적 결함이나 버그로 보았지만, 우리는 이를 종 간 언어 구조의 불일치로 재해석합니다.

아이디어 설명

추가자료 (〈AI 중 도감: 실패에서 찾은 공존의 언어 (외계종과 협업하는 인간을 위한 생존 가이드)〉):

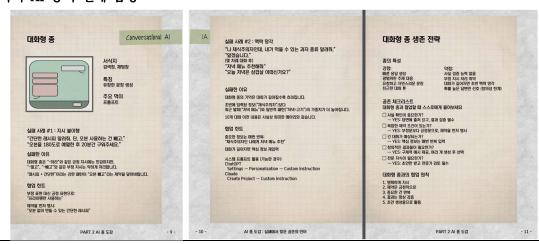
https://drive.google.com/file/d/1csqAOiTIuMAZEiKrVfdFyx9rIW82RivO/view?usp=sharing
전체 구조: 3부작 도감



PART 1 - 왜 우리는 AI에게 계속 속는가?: 문제의 보편성을 입증하고 'AI를 외계종으로 본다면?'이라는 관점 전환을 제시합니다. 실제 통계와 사례를 통해 AI 실패가 예외가 아닌 일상임을 보여주고, 실패를 좋 가 언어 차이의 산물로 재정의합니다.

PART 2 - AI 종 도감: 세 가지 AI 종을 생물 도감처럼 분류합니다. 각 종마다 서식지, 주요 먹이, 대표 실패 사례 2가지, 실패한 이유(종의 특성), 협업 힌트, 공존 체크리스트, 협업 원칙을 담았습니다. PART 3 - 실패 복구 프로토콜: AI 종과의 협업이 무너질 때 필요한 실전 도구를 제공합니다. 각 종의 언어로 말하는 법(프롬프트 설계 지침), 위험 신호를 읽는 법(환각·맥락 오류·과의존 감지), 10가지 복구 체크리스트, 협업의 주도권을 지키는 최종 원칙을 제시합니다.

세 가지 AI 종의 실패 습성



1. 대화형 종 (Conversational AI)

검색창과 채팅창에 서식하며 프롬프트를 먹고 자랍니다. 유창한 문장 생성 능력을 가졌지만 치명적인 약점이 있습니다. 첫 번째 실패 패턴은 '지시 불이행'입니다. "간단한 레시피 알려줘. 오븐 사용하는 건 빼고."라고 요청하면 "오븐을 180도로 예열한 후..."라는 답변을 받게 됩니다. 이 종은 긍정 지시에는 민감하지만 부정 지시("~빼고")는 약하게 처리합니다. 협업 힌트는 "프라이팬만 사용하는"처럼 긍정 표현으로 제약을 먼저 명시하는 것입니다.

두 번째 실패 패턴은 '맥락 망각'입니다. "나 채식주의자"라고 말했는데 몇 턴 후 "삼겹살 어떠세요?"라는 답변을 받습니다. 10개 대화 이전 내용은 희미한 메아리 수준입니다. 중요한 정보는 매번 반복하거나 ChatGPT의 Custom Instruction, Claude의 Project 기능에 저장해야 합니다.

대화형 종과의 협업 원칙: 명확하게 지시, 제약은 긍정적으로, 중요한 건 반복, 결과는 항상 검증, 초 안 생성용으로 활용.

2. 이미지 생성 종 (Image Generation AI)

프롬프트 입력창에 서식하며 텍스트 묘사를 먹고 자랍니다. 픽셀 단위로 예측하는 능력을 가졌지만 인간의 상식은 이해하지 못합니다. 첫 번째 실패 패턴은 '신체 구조 오류'입니다. 손가락이 6개이거나 형태가 왜곡됩니다. 이 종은 픽셀 패턴을 통계적 확률로 이해하고, 인간의 상식은 약하게 처리합니다. 협업 힌트는 "공연장의 기타리스트"처럼 구조를 단순화하거나 "다섯 손가락인 사람"처럼 텍스트로 보 강하는 것입니다.

두 번째 실패 패턴은 '텍스트 왜곡'입니다. 이 종은 텍스트를 의미가 아닌 픽셀 패턴으로 인식합니다. 협업 힌트는 이미지 후처리를 전제로 하거나, 글자 부분은 "빈 공간으로 두기"를 지시하는 것입니다. 이미지 생성 종과의 협업 원칙: 설명은 구체적으로, 시각적 요소 명시, 텍스트 후처리 필수, 반복 교정 지시, 초벌 화가로 활용.

3. 추천 알고리즘 종 (Recommendation AI)

색 주도권 유지, 데이터 흔적 스스로 관리.

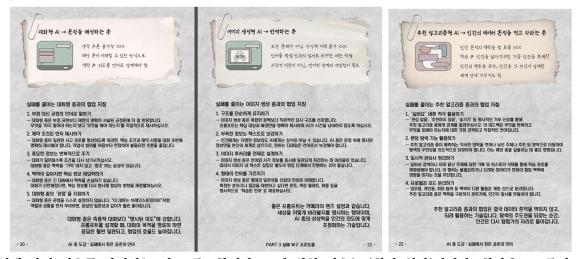
피드와 홈 화면에 서식하며 사용자 행동을 먹고 자랍니다. 클릭 패턴을 추적하는 능력을 가졌지만 맥락을 이해하지 못합니다. 첫 번째 실패 패턴은 '필터 버블 심화'입니다. 고양이 영상 하나를 봤는데 온통 고양이 콘텐츠만 추천합니다. 이 종은 다양성보다 체류 시간을 우선합니다. 협업 힌트는 '싫어요' 버튼을 적극 활용하거나 랜덤 탐색 기능을 사용하는 것입니다.

두 번째 실패 패턴은 '맥락 무시'입니다. "아버지 생신 축하 선물을 사기 위해 잠시 낚시 용품을 검색했을 뿐인데, 이후 수개월간 낚시 관련 광고만 지속적으로 나옵니다." 이 종은 일시적 탐색과 장기적 필요를 구분 못 합니다. 협업 힌트는 일시적 기록을 삭제하거나 프로필/모드를 분리하는 것입니다. 추천 알고리즘 종과의 협업 원칙: '싫어요' 버튼 활용, 알고리즘에 휩쓸리지 않기, 계정/모드 분리, 탐

실패 복구 프로토콜

협업의 실패는 AI 기술 자체의 한계가 아니라 언어 구조의 불일치에서 비롯됩니다. AI는 인간의 말을 이해하는 것처럼 보이지만, 그 정보를 받아들이는 방식이 인간과는 다릅니다. 각 AI는 서로 다른 감각 체계를 지니고 있으며, 인간의 언어를 각자의 알고리즘에 따라 다르게 해석하는 별도의 지적 종이기 때문입니다. 따라서 AI 종과의 성공적인 협업의 핵심은 "무엇을 말하느냐"보다 "어떻게 말하느냐"에 달려 있습니다. 인간은 AI 종의 언어 구조에 맞는 프롬프트 문법을 익혀야 하며, 이 프롬프트는 단순한 명령이 아니라 대화의 문법입니다.

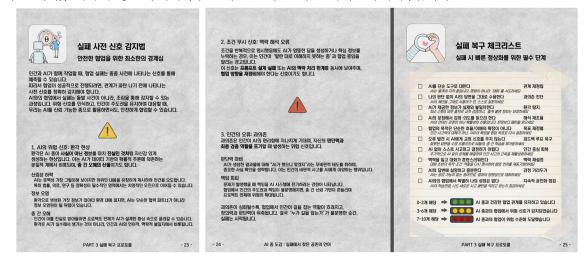
각 종의 언어로 말하는 법은 다음과 같습니다. 대화형 종은 부정 대신 긍정 언어로 말해야 하고, 제약 조건을 먼저 제시해야 하며, 중요한 정보는 반복적으로 주고, 맥락이 길어지면 핵심 정보를 재입력하며, 대화형 종의 '관점'을 지정해야 합니다. 이미지 생성 종은 구조를 단순하게 유지하고, 부족한정보는 텍스트로 보강하며, 이미지 후처리를 전제로 설계하고, 형태의 언어를 가르쳐야 합니다. 추천알고리즘 종은 '싫어요' 버튼을 적극 활용하고, 랜덤 탐색 기능을 사용하며, 일시적 관심사를 정리하고, 프로필과 모드를 분리해야 합니다.



실패 사전 신호를 감지하는 것도 중요합니다. AI의 위험 신호는 '환각 현상'입니다. 환각은 AI 종이 사실이 아닌 정보를 마치 진실인 것처럼 자신감 있게 생성하는 현상으로, 신뢰성 하락, 정보 오염, 종간 오해를 일으킵니다. 조건 무시 신호는 '맥락 해석 오류'입니다. 조건을 반복적으로 명시했음에도 AI가 엉뚱한 답을 생성하거나 핵심 정보를 누락하는 경우, 이는 프롬프트 설계 실패 또는 AI의 맥락처리 한계를 동시에 보여줍니다. 인간의 오류는 '과의존'입니다. AI가 생성한 결과물에 대해 "AI가 했으니 맞겠지"라는 무비판적 태도를 취하거나, 문제 발생 시 책임을 AI 시스템에 전가하려는 경향이나타나면 협업에서 인간이 길을 잡는 역할이 흐려지고 실패가 시작됩니다.

실패 복구 체크리스트 10단계는 다음과 같습니다. 관계 재정립(AI는 별개의 지적 종, 명령이 아니라 대화), 과의존 진단(AI 제안을 그대로 사용 전 스스로 질문), 환각 탐지(최소 2개 외부 출처로 교차 검증), 해석 재조율(AI 언어는 확률성의 산물, 패턴 읽기), 목표 재정렬(협업 목적은 효율이 아닌 이해 확장), 피드백 루프 복구(잘못된 답변을 수정 프롬프트로 되돌리기), 인간 중심 회복(주기적으로 AI 없이 문제 해결), 맥락 재설정(대화 초반 목적·조건 재명시), 감정 거리두기(AI는 공감 기능 없음, 결

과로 대화), 지속적 공진화 점검(나도 새로운 사고 패턴 익히는지 점검)입니다.



협업 성공을 위한 최종 원칙은 세 가지입니다. 첫째, AI를 지능적 존재로 인정하되 협업의 주도권은 인간에게 있으며, 성공은 AI의 알고리즘 언어를 해석하는 데서 시작되지만 최종 검증은 인간의 비판적 언어로 수행해야 합니다. 둘째, AI의 오류는 기술 결함이 아닌 종 간의 언어적 오해로 인식해야하며, 실패를 버그가 아닌 소통 과정으로 이해해야 합니다. 셋째, 궁극적인 협업 목적은 효율이 아닌 공동 사고(Co-thinking)를 통한 지적 창출에 있습니다.

기대 효과 및 사회적 의의

기대 효과는 개인, 관점 전환, 사회적 차원으로 나뉩니다.

1. 개인 차원

실패 패턴 사전 감지 능력이 향상되고, AI 종별 맞춤형 프롬프트 설계 스킬을 습득하며, "내가 겪은 실패는 어떤 종의 습성인가?"라는 메타인지를 제공받습니다.

2. 관점 전환 차원

AI를 완벽해야 할 도구가 아닌 공존해야 할 외계종으로, 실패를 버그가 아닌 종 간 소통 과정으로, 협업을 명령-수행 관계가 아닌 대화-공존 관계로 인식하게 되며, 주도권과 최종 검증 책임은 인간에 게 있다는 인간 중심성을 회복합니다.

3. 사회적 차원

환각 대응 문화 정착(정보 검증 습관화), 책임 소재 명확화, 실패 기록 정상화를 통한 안전한 공존 문화를 조성하고, AI와의 건강한 협업 문법 교육, 종 간 언어 차이 이해 확산, 전 세대 대상 실전 매뉴얼 제공을 통한 디지털 리터러시를 향상시키며, 실패 데이터 체계화 방법론 제시, 사용자 경험 중심 AI 개선 방향, AI 윤리·안전성 논의에 실증적 기여를 합니다.

AI 종 도감은 실패를 통해 인간과 AI가 서로를 더 잘 이해하고, 안전하게 공존할 수 있는 새로운 언어를 찾는 여정입니다. 실패를 '숨겨야 할 버그'가 아닌 '기록해야 할 습성'으로, AI를 '완벽한 도구'가 아닌 '함께 배워가는 외계종'으로, 협업을 '명령-수행'이 아닌 '대화-공존'으로 바라볼 때, 우리는 진정한 공존의 가능성을 발견할 수 있습니다.