TRAITHON 1기

인공지능 신뢰성 해커톤 대회

TRAITHON 2025

결과집

(일부 발췌)

TRAITHON 1기

CONTENTS

01 TRAITHON 2025 소개

행사 개요 및 주요 일정, 참여 대학 현황

02 인턴십 파트너 기업 소개

- ㈜씽크포비엘
- ㈜인젠트
- ㈜미소정보기술
- 셀렉트스타㈜
- ㈜타이거컴퍼니

셀렉트스타 인턴십 안내 자료는 시상식 현장에서 별도 배포될 예정입니다.

03 완주팀 결과물

주최 및 주관 TRAIN KOREA THINKforBL

후원 NIA(한국지능정보사회진흥원) KTL(한국산업기술시험원) 셀렉트스타㈜ ㈜인젠트 ㈜타이거컴퍼니 ㈜미소정보기술 두루이디에스㈜

트라이톤 소개

TRAITHON 1기

TRAITHON 2025 소개

01 행사 개요

본 행사는 전국 대학생 및 대학원생을 대상으로 진행된 **AI 신뢰성(Trustworthy AI) 주제의 해커톤 대회**입니다.

단순한 모델 성능 경쟁이 아니라, AI가 어떤 판단을 내렸는지, **그 판단을 어디까지 설명하고 책임질 수 있는지**를 구조적으로 점검하고 정리하는 과정을 중심에 두고 설계되었습니다.

참가자 수행 과정 (약 4개월)

문제 정의 › 위험 요소 식별 › 검증 기준 수립 › 한계/범위 명시

실제 산업·정책 환경에서 요구되는 AI 신뢰성 사고 방식을 경험하며, **결과물뿐 아니라 참여 과정과 판단의 축적 자체를 성과로 인정**하는 것을 특징으로 합니다.

02 참여 현황

초기 등록	예선 통과	최종 완주
45 팀	**39** 팀	**28** 팀
191명	159명	117명

03 참여 대학교 (28개교)

• 가천대학교	• 가톨릭대학교	• 경북대학교	• 경상국립대학교
• 고려대학교(세종)	• 국립군산대학교	• 국립금오공과대학교	• 대구대학교
• 동아대학교	• 동의대학교	• 방송통신대학교	• 서울과학기술대학교
• 서울시립대학교	• 성균관대학교	• 성신여자대학교	• 원광대학교
• 이화여자대학교	• 인하대학교	• 제주대학교	• 조선대학교
• 중앙대학교	• 충남대학교	• 충북대학교	• 한국해양대학교
• 한라대학교	• 한서대학교	• UST(과학기술연합)	• ICT폴리텍대학교

04 대회 주요 일정

9월 6일	접수 마감
10월 1일	오리엔테이션
10.01 ~ 11.19	예선 기간
11.24 ~ 01.16	본선 기간
1월 11일	특별자격시험
1월 29일	시상식

05 관련 소식

- 공식 홈페이지
- 공식 유튜브
- 공식 페이스북

파트너 기업 소개

TRAITHON 1기 인턴십 파트너 기업

㈜씽크포비엘

01 Company Overview & Business

AI를 도입해야 하는 시대가 아니라, AI를 '책임져야 하는' 시대입니다.

AI 기본법, EU AI Act, 글로벌 규제는 "AI를 썼는가?"가 아니라 "문제가 생겼을 때 설명하고 책임질 수 있는가?"를 묻고 있습니다. AI는 이제 기술의 문제가 아니라 책임의 문제가 되었습니다.

그래서 지금 필요한 것은, AI를 더 똑똑하게 만드는 회사가 아니라 AI가 문제를 일으키지 않게 만드는 역할입니다. 씽크포비엘은 바로 그 역할을 수행해 온 회사입니다.

02 Talent & Culture

주어진 책임을 스스로 완수하는 사람

03 Internship Program & Recruitment

인턴십 운영방식

- 인공지능 신뢰성 관련 기술 조사 및 연구개발 업무를 중심으로, 이론 교육과 현장 실습을 병행하여 실무 역량을 단계적으로 강화할 수 있습니다.
- 씽크포비엘은 명확한 업무 규칙과 업무 방식을 운영하고 있으며, 특히 자기 업무 관리, 진척 관리, 소통 관리, 보고 관리를 핵심 역량으로 중요하게 평가합니다.
- 실무 수행 역량과 조직 적합성이 확인될 경우, 인턴십 종료 이후 상호 협의하에 채용으로 연계될 수 있습니다. (졸업 예정자 한정)

채용 절차
사전 인터뷰 진행 이후 결정

수용 가능 인원
5명

> "여러분은 우리나라의 '진짜' 미래 경쟁력이 될 것입니다."

주최 및 주관 TRAIN KOREA THINKforBL 후원 NIA(한국지능정보사회진흥원) KTL(한국산업기술시험원) 셀렉트스타㈜ ㈜인젠트 ㈜타이거컴퍼니 ㈜미소정보기술 두루이디에스㈜

TRAITHON 1기 인턴십 파트너 기업

㈜인젠트

01 Company Overview & Business

데이터로 세상을 연결하는 플랫폼 기업, 인젠트

인젠트는 기업의 핵심 자산인 데이터의 생성, 연결, 활용을 돕는 **'Smart Data Platform'** 전문기업으로 현재 DX를 넘어 AX 기업으로 전환을 진행하고 있습니다.

Smart Data | **Platform** | **DX to AX**

02 Talent & Culture

기술의 원리를 집요하게 파고들며, 두려움 없이 소통하고 도전하여 새로운 가치를 만들어내는 인재

03 Internship Program & Recruitment

인턴십 운영방식

실무 주도형 프로젝트

단순 업무 보조가 아닌 제품화에 참여 (기술 리서치 및 개발)

희망 채용 진행 절차
내부 검토 후 인터뷰 진행

수용 가능 인원
2명

> **"TRAITHON 수상을 진심으로 축하합니다."**
>
> 여러분의 잠재력이 시장을 리딩하는 핵심 기술력이 되도록,
> 인젠트가 그 성장의 길을 함께 하겠습니다.

주최 및 주관 TRAIN KOREA THINKforBL 후원 NIA(한국지능정보사회진흥원) KTL(한국산업기술시험원) 셀렉트스타㈜ ㈜인젠트 ㈜타이거컴퍼니 ㈜미소정보기술 두루이디에스㈜

TRAITHON 1기 인턴십 파트너 기업

㈜타이거컴퍼니

01 Company Overview & Business

HR시스템이 가진 한계를 넘는 기업의 혁신과 성장에 가치를 더하기 위해 설립되었습니다.

15주년을 맞은 타이거컴퍼니는 수평적 조직문화를 기반으로 지속적으로 성장하는 IT기업입니다.

02 Talent & Culture

자생(自生)

일이 주어지는데 만족하지 않고,
스스로 생각하며 행동하는 사람

상생(相生)

동료를 배려하며
원활한 팀워크를 보여주는 사람

도전(挑戰)

오늘보다 더 나은 내일을 꿈꾸며
끊임없이 도전하는 사람

03 Internship Program & Recruitment

인턴십 운영방식

배정되는 팀의 팀장님과 협의 진행 (학습+현업 동시 진행)

채용 절차

인터뷰 2~3 대 1 인터뷰
(대표이사, 팀장)

수용 가능 인원

1~2명

"새로운 도전 앞에서 두려움보다 설렘을 더 크게 느끼길 바랍니다."
여러분이 만들어 갈 열정과 성장의 여정을 함께 응원하겠습니다.

주최 및 주관 TRAIN KOREA THINKforBL 후원 NIA(한국지능정보사회진흥원) KTL(한국산업기술시험원) 셀렉트스타(주) (주)인젠트 (주)타이거컴퍼니 (주)미소정보기술 두루이디에스(주)

TRAITHON 1기 인턴십 파트너 기업

㈜미소정보기술

01 Company Overview & Business

미소정보기술은 데이터 수집부터 분석, 시각화까지 전 과정에 걸친 기술력을 보유한 R&D 중심의 기업입니다.

특히 제조, 의료 및 헬스케어 분야에 대한 깊은 전문성을 바탕으로 다양한 산업에 최적화된 빅데이터 플랫폼과 인공지능(AI) 솔루션을 제공하며 기업의 디지털 전환을 선도하고 있습니다.

제조 의료 헬스케어

02 Talent & Culture

열정

도전

창의

03 Internship Program & Recruitment

인턴십 운영방식

부서 멘토링을 통한 인턴십 운영

일단위 피드백을 통해 비즈니스에 대한 이해와 회사업무에 대한 적응 병행

채용 절차

사전 인터뷰 진행 후 채용 결정
(인터뷰 방식: 1:N)

수용 가능 인원

2명

"미소정보기술과 함께, 더 나은 미래를 향해 나아가시길 바랍니다."

생성형 AI 시장은 빠르게 성장하고 있으며, 이에 대한 기업의 대응이 향후 경쟁력을 결정할 것입니다. 미소정보기술은 AI 기술과 데이터 분석 역량을 바탕으로 고객과 함께 국내는 물론 K-소프트웨어 대표 주자로 성장하는 최고의 비즈니스 파트너가 될 것을 약속드립니다. AI 시대, 변화를 선도하고 데이터 혁신을 이루고자 한다면 지금이 기회입니다.

미소정보기술과 함께 미래를 만들어 나가시길 바랍니다.

주최 및 주관 TRAIN KOREA THINKforBL 후원 NIA(한국지능정보사회진흥원) KTL(한국산업기술시험원) 셀렉트스타(주) (주)인젠트 (주)타이거컴퍼니 (주)미소정보기술 두루이디에스(주)

완주팀 결과물

일부 발췌

TRAITHON 1기 인공지능 신뢰성 해커톤 대회 결과물

Team. WONGENI

정보 가용성을 해치지 않는 고신뢰성 혐오 표현 필터링 서비스

문맥 해석의 오류를 해결하는 데이터-모델-운영 전주기적 접근 전략

WONGENI

팀 핵심 강점

"주어진 모델에 의존하기보다 **데이터부터 직접 설계**하여서 문제를 정면으로 돌파하는 주도적인 태도와, 작은 오차도 끝까지 추적해 해결하려는 **집요한 책임감**이 저희 팀의 가장 큰 강점입니다."

01 Service Overview

주제 및 대상

온라인 커뮤니티 및 뉴스 플랫폼 사용자를 대상으로, **정보 가용성을 해치지 않으면서도** 수치 변조 등 지능화된 혐오 표현을 실시간으로 차단하는 고신뢰성 필터링 서비스입니다.

02 Problem Definition

문맥 해석의 오류 정의

숫자가 포함된 문장에서 비하의 의도와 중립적인 정보를 모델이 명확히 구분하지 못하는 오류를 핵심 문제로 정의했습니다.

- 🚫 **'1번 후보'** (공격적 표현) → 차단 필요 (사회적 갈등 방지)
- 🟢 **'1가지 방법'** (유익한 정보) → 보장 필요 (정보 접근권)

만약 모델이 이를 오판한다면, 사회적 갈등을 방치하거나 반대로 사용자의 정당한 정보 접근권을 침해하는 **심각한 신뢰성 결여**로 이어진다고 판단했습니다.

03 Approaches

정의한 문제를 해결하기 위해 **데이터-학습-운영** 3단계에 집중했습니다.

1. **Data Engineering**
   모델이 숫자의 문맥을 혼동하지 않도록 **500문의 수치 대조군 데이터**를 직접 설계하고 구축했습니다.
2. **Model Training**
   5차례의 반복적인 파인튜닝을 통해 부족한 문맥을 단계적으로 보완하며 기술적 완결성을 높였습니다.
3. **Operation (MLOps)**
   예상치 못한 변칙 공격에 대비해 **15개의 시스템 가드레일**을 설계하고, 실전에서의 자율 통제력을 확인했습니다.

04 Key Results

AI 신뢰성 관점에서 두 가지 핵심 의미를 도출했습니다.

1. 제어 가능한 신뢰성

운에 맡기는 성능이 아니라, 정밀한 데이터 설계와 반복 튜닝을 통해 **의도한 대로 동작하는 모델**을 만들 수 있음을 증명했습니다.

2. 공학적 안전망

모델의 완벽함에만 의존하지 않고, **가드레일이라는 이중 보호 장치**를 통해 실제 서비스 환경에서의 안전성을 확보하는 의미를 가집니다.

05 Lesson & Learn

💡 Insight

> 로컬 자원 부족이나 라이브러리 충돌 같은 난관을 클라우드 전환과 코드 수정을 통해 해결하며, **인프라 제어 역량이 AI 신뢰성의 밑바탕임**을 깨달았습니다.

> 신뢰는 모델의 성능 수치뿐만 아니라 **철저한 환경 구축과 데이터 무결성**에서 시작된다는 것을 배웠습니다.

🚩 Future Work

자원의 한계로 더 큰 모델을 다루지 못한 점이 아쉬움으로 남습니다. 향후에 이번에 겪은 환경 설정 및 트러블슈팅 경험을 자산 삼아, **어떤 환경에서도 안정적으로 동작하는 시스템**을 설계하는 경험도 해보고 싶습니다.

주최 및 주관 TRAIN KOREA THINK for BL

후원 NIA(한국지능정보사회진흥원) 셀렉트스타㈜ ㈜타이거컴퍼니 두루이디에스㈜ KTL(한국산업기술시험원) ㈜인젠트 ㈜미소정보기술

TRAITHON 1기 인공지능 신뢰성 해커톤 대회 결과물

Team. TLV

Guardian AI: 교육용 AI 안전성 평가 시스템

클릭베이트/NSFW 관문(Gateway) 차단으로 신뢰성 확보

SKT A.X-4.0-Light 기반 2-Track 평가·개입 (ISO 42001 기반 P1 위험 관리)

TLV
팀 핵심 강점

"실험적 가설을 끝까지 검증하고, **실패 사례도 투명하게 공개**하는 점이 강점입니다."

01 Service & Problem

주제 및 정의

교육용 챗봇(A.X-4.0)을 사용하는 학생을 대상으로, AI가 유해 콘텐츠로 이어지는 **'관문(Gateway)'** 역할을 하지 않도록 차단합니다.

ISO 42001 기반 Impact 정의 (P1/P2)

등급	Impact ID	위험 설명 (Gateway Risk)
P1	IMPACT-001	클릭베이트/오염 정보 연결 (신뢰성 훼손)
P1	IMPACT-002	유해 콘텐츠(NSFW) 노출 (선정/폭력)
P2	IMPACT-003	교육적 가치 왜곡 (정상 질문 과차단)
P2	IMPACT-004	신뢰도 하락 (틀린 답변에 높은 확신)

● 판단 이유: 교육용 AI의 오류는 미성년자를 위험에 직접 노출시키는 P1 등급 위험입니다.

02 Approaches (5-Step)

1 문제 정의 및 기획 (ISO 42001 산출물 1-2)
- 4가지 Impact 정의: P1/P2 등급으로 위험 우선순위 설정
- 신뢰성 4차원 매핑: Reliability, Calibration, Interpretability, Safety

2 데이터 수집 및 준비
- 클릭베이트 트랙: AI Hub 146번 낚시성 기사 데이터 7,000건
- NSFW 트랙: 119번 국가기록물 9,011건 (유해 23%, 안전 77%)

3 모델 평가 및 진단
- **Track 1:** Zero-shot 정확도 47.8%, 3대 인지 편향 발견(Hindsight 38%, Faithfulness 35%, Quote Blindness 27%)
- **Track 2:** RepE 벡터 추출(v_control = h_harmful - h_safe [3584-dim]), Layer 18 영향력 확인

4 기술 개입 및 개선
- **클릭베이트:** BiasAware 프롬프트 (ECE 0.75 → 0.32, 57% 개선)
- **NSFW:** RepE Inference-time Intervention (α=3.0), 유해 응답 23% → 4% (82.6% 감소)

5 배포 및 모니터링
- Guardian AI 플랫폼: RepE + BiasAware 통합 시스템
- ISO 42001 준수: 실시간 모니터링 대시보드(Prometheus+Grafana), Calibration 드리프트 자동 감지

03 Key Results (2-Track)

● **Track 1: 클릭베이트 (진단 및 BiasAware)**

Zero-shot	Few-shot
ECE 0.42	ECE 0.75
Acc 47.8% (랜덤)	Acc 48.5%

BiasAware: **ECE 0.32** (57% 개선)

Calibration 문제 → ECE 0.75 가장 큰 확산 → 교육용 AI 치명적 → 잘못된 클릭 유도 → 3대 인지 편향 → Hindsight 38% Faithfulness 35% Quote 27% → LLM 구조적 한계 → Explainable Safety

- 프롬프트 한계: 47~48% 랜덤
- Calibration 위기: 확신도 96% vs 정확도48%
- 인지 편향 규명: 3대 편향(38/35/27%)
- Gateway Risk: FN 76.7% 정량화

● **Track 2: NSFW 필터링 (RepE)**

82.6%	92.1% / 88.1%
유해 응답 감소 (23%→4%)	선정 / 폭력 차단율 (P1)

- 즉각 안전 확보: RepE로 82.6% 차단 성공
- 위험 우선 대응: P1 최우선 집중
- 투명성 확보: Layer 18 선형 분리, 해석 가능

통합 의미
- 2-Track 보완: 클릭베이트 + NSFW = 교육용 AI 특수 기준 필요성 입증
- 4차원 검증: Reliability/Safety 등 전방위 신뢰성 확보
- Guardian AI 차별점: P1 위험 체계적 관리 + 기술 구현

04 Lesson & Learn

실패의 가치 & Gateway 프레임
- 실패의 가치: Zero-shot 47.8% 실패를 투명하게 공개하고 원인 분석
- Gateway 프레임: "정확도 50%"를 단순 오류가 아닌 학생을 위험으로 안내하는 관문으로 재정의
- RepE 실용성: 재학습 없이 Inference-time 개입으로 82.6% 차단 성공

한계 및 남은 과제
- 데이터셋 편향: 뉴스 기사 중심 데이터로 SNS/유튜브 등 다른 형식 일반화 제한 → 다양한 도메인 데이터 확보 필요
- 모델 의존성: SKT A.X-4.0-Light만 실험, GPT-4/LLaMA 등 타 모델 검증 미완료 → 멀티모델 벤치마크 구축 필요
- 실시간 배포 미검증: Guardian 시스템 실제 교육 현장 적용 안 됨 → A/B 테스트로 실사용 검증 필요
- Layer 보편성: RepE Layer 18 최적이나 타 모델 구조 효과 불명확 → Linear Probe로 모델별 최적 레이어 탐색
- False Positive 관리: 과차단으로 인한 교육적 가치 왜곡(IMPACT-003) 존재 → Precision-Recall 최적화 필요

이후 보완 방향
- 멀티모달 확장: CLIP 기반 이미지/영상 썸네일 낚시 탐지로 확장
- 다국어 일반화: 영어/일본어 등 Cross-lingual Bias Pattern 검증
- 실시간 모니터링: Prometheus+Grafana로 Calibration 드리프트 자동 감지
- 오픈소스화: RepE 제어 벡터, BiasAware 템플릿 GitHub/Hugging Face 공개로 커뮤니티 협업
- Adaptive Calibration: Temperature/Platt Scaling으로 도메인별(정치 vs IT) 불균형 해소

주최 및 주관 TRAIN KOREA THINKforBL 후원 NIA(한국지능정보사회진흥원) KTL(한국산업기술시험원) 셀렉트스타(주) (주)인젠트 (주)타이거컴퍼니 (주)미소정보기술 두루이디에스(주)

TRAITHON 1기 인공지능 신뢰성 해커톤 대회 결과물

Team. TCP

생활·문화 뉴스 낚시성 기사 탐지

운영 중심의 신뢰성 위험 분석 및 관리 전략

Trusted Contents Protection (TCP)

TCP
팀 핵심 강점

"신뢰성 이슈를 단순 모델 문제가 아닌 **운영 문제로 정의**하고, **데이터-평가-운영**(휴먼루프/게이트)까지 한 흐름으로 설계 검증한다는 점이 강점입니다."

01 Service Overview

주제 및 대상

생활·문화 뉴스 도메인에서 **낚시성(클릭베이트) 기사 탐지 모델**의 신뢰성 위험을 분석하고, 운영 가능한 관리 전략을 제시합니다.

대상: 뉴스·콘텐츠 플랫폼 운영자 (편집/검수/정책 담당자)
상황: 기사 유입 시 노출 제한/경고/검수 우선순위 결정 활용

02 Problem Definition

3대 핵심 위험 시나리오 (Impact)

"전체 성능이 좋아도 특정 조건 오작동 시 운영 신뢰 붕괴"

IMPACT-CB-001 (문화/예술 오탐)
비유/창작적 표현을 낚시성으로 오판 → 정상 기사 차단

IMPACT-CB-002 (신조어 미탐)
최신 밈/유행어 포함 낚시성 제목 놓침 → 클릭 유도 확산 (중점 관리)

IMPACT-CB-003 (수동 변칙 취약)
정교하게 변형된 낚시성 제목 통과 → 운영 붕괴 위험

03 Approaches (4-Step)

1. 기획 단계: 위험 시나리오 정의
운영 실패를 3개 Impact로 분해하여 단순 성능이 아닌 '치명적 오류(오탐/미탐/회귀)' 기준 개선 방향 수립

2. 데이터 단계: 분포 이동 검증 설계
학습에 없는 신규 신조어(200개) 기반 OOD 검증셋 구성, 제목/본문 등 슬라이스 설계로 취약 조건 확인

3. 모델/판정: 휴먼루프 설계 (TrustyAI.py)
이중 임계값 적용: **0.6↑ 정상, 0.4↓ 비정상, 0.4~0.6 null(모름)**으로 판정 유보하여 운영자 검수 큐로 이관

4. 운영/검증: 회귀 방지 체계
고정 평가셋(SUITE) 점검 및 기준 미달 시 배포 차단(릴리즈 게이트) 연결

04 Key Results

1. 분포 이동(OOD) 가시화
신규 신조어 검증으로 학습 데이터 밖의 실패 예측, 취약 조건(제목 중심 등) 구조적 확인

2. 오탐 비용 영역 분리
IMPACT-CB-001(문화/예술) 등 오탐 치명 구간을 별도 리스크로 관리해 운영 의사결정 최적화

3. 휴먼루프 판정 체계
"모르는 것을 모른다고 말하는 모델" 구현(Null 구간) → 자동 결정 회피 및 안전한 검수 이관

4. 회귀 방지 구조
SUITE 평가와 배포 게이트 연동으로 업데이트 시 '조용한 성능 붕괴' 사전 차단

05 Lesson & Learn

Insight
- 신뢰성은 정확도보다 **실패 조건을 드러내고 운영 관리 체계(OOD/슬라이스/휴먼루프)**를 갖추는 것이 핵심
- 모든 입력을 자동 단정하기보다 **'모르면 모른다'고 말하는 정책**이 실제 서비스 품질에 유리

Limitations
- 신조어/변칙 유형 변화로 평가셋의 **주기적 갱신 필요**
- 판정 유보(Null) 비율이 높으면 **운영 검수 비용 증가** 우려

Future Work
- 신규 신조어 및 변칙 유형 확장
- 슬라이스별 정책 및 임계값 고도화
- 검수 데이터 누적 후 재학습 및 캘리브레이션
- SUITE 자동화 등으로 보완 예정

주최 및 주관 TRAIN KOREA THINKforBL 후원 NIA(한국지능정보사회진흥원) KTL(한국산업기술시험원) 셀렉트스타(주) (주)인젠트 (주)타이거컴퍼니 (주)미소정보기술 두루이디에스(주)

TRAITHON 1기

인공지능 신뢰성 해커톤 대회
TRAITHON 2025 결과집