JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 11, pp. 469-486, November 2026
https://doi.org/10.9708/jksci.2026.31.11.469

Evaluating Utterance-Level Safety Guardrails for Preschool AI Companions: Contextually Benign Utterances and Support-Need Detection

Woomin Nam*, Sudan Prasad Uprety*, Keon Chul Park**

*Research Fellow, Urban AI EH, Tech University of Korea, Siheung, Korea
**Assistant Professor, Business Administration Department, Tech University of Korea, Siheung, Korea

[Abstract]

In this paper, we design an input-utterance safety guardrail for preschool AI companions and compare its classification and detection performance across learning configurations, focusing on contextually benign utterances and the support-need attribute. We annotated 27,605 Korean utterances from children aged 3-5 with semantic categories and safety attributes and evaluated separate learning, joint learning, and joint learning with combined losses using five pretrained encoders. Overall classification performance and false-positive rates for contextually benign utterances were similar across configurations, whereas missed-detection rates for benign utterances requiring support were 2.22%, 41.11%, and 43.89%, respectively. Additional contrastive and margin losses reduced harmful-to-benign misclassification but did not improve support-need detection. Increasing the support-loss weight substantially reduced missed detections under joint learning. These findings highlight the importance of distinguishing harmfulness from support needs and evaluating category-specific errors alongside overall performance. The evaluation is limited to offline prediction from individual utterance texts and does not assess downstream responses or effectiveness in interactions with children.

▸ **Key words**: Preschool AI Companion, Safety Guardrail, Contextually Benign Utterances, Support-Need Detection, Harmful Language Detection

[요 약]

본 논문에서는 유아용 인공지능 동반자의 입력 발화를 평가하는 안전 가드레일을 설계하고, 맥락무해 발화와 지원 필요 속성을 중심으로 학습 구성에 따른 분류 및 탐지 성능을 분석한다. 이를 위해 만 3-5세 유아의 한국어 발화 27,605건에 의미 범주와 안전 속성을 부여하고, 5종의 사전학습 인코더를 활용하여 분리학습, 공동학습 및 복합손실 공동학습을 평가하였다. 실험 결과, 세 학습 구성의 전체 분류 성능과 맥락무해 오탐률은 유사하였으나, 무해·지원 필요 발화의 평균 누락률은 각각 2.22%, 41.11%, 43.89%로 나타났다. 지도대조손실과 마진손실의 추가는 유해 발화를 무해로 예측하는 오류를 줄였으나, 지원 필요 탐지 성능의 개선으로 이어지지 않았다. 반면 공동학습의 지원 필요 손실 가중치를 높이면 해당 발화의 누락률이 크게 감소하였다. 이러한 결과는 유해성 판별과 지원 필요 탐지를 구분하고, 전체 분류 성능과 함께 의미 범주별 탐지 오류를 평가하는 것이 중요함을 보여준다. 본 연구는 유아용 인공지능 동반자의 발화 수준 안전 가드레일 설계 및 평가를 위한 실증적 근거를 제공한다.

▸ **주제어**: 유아용 인공지능 동반자, 안전 가드레일, 맥락무해 발화, 지원 필요 탐지, 유해언어 탐지

• First Author: Woomin Nam, Corresponding Author: Keon Chul Park
*Woomin Nam (applewoods9559@tukorea.ac.kr), Urban AI EH, Tech University of Korea
*Sudan Prasad Uprety (uprety_sud@tukorea.ac.kr), Urban AI EH, Tech University of Korea
**Keon Chul Park (parkkc07@tukorea.ac.kr), Business Administration Department, Tech University of Korea
• Received: 2026. 09. 16, Revised: 2026. 10. 27, Accepted: 2026. 10. 28.

Copyright © 2026 The Korea Society of Computer and Information
http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945
www.kci.go.kr

I. Introduction

대화형 인공지능(Artificial Intelligence, AI)은 정보 제공과 학습 지원을 위한 기능적 도구를 넘어 이용자와 지속적으로 대화를 주고받는 상호작용형 시스템으로 확장되고 있다. 이러한 변화는 유아의 대화형 인공지능 이용에서도 나타나고 있다. 육아정책연구소의 조사에 따르면, 만 3-6세 자녀를 둔 부모의 52.0%는 자녀가 로봇이나 AI 스피커에 인사하거나 질문한 경험이 있다고 응답하였으며, 34.9%는 자녀가 로봇을 친구처럼 대하는 모습을 보았다고 응답하였다[1]. 또한 아동은 음성 기반 대화형 시스템을 정보 탐색뿐 아니라 놀이와 이야기 구성 등 다양한 활동에 활용한다[2]. 이러한 이용 양상은 대화형 인공지능이 유아에게 정보 제공뿐 아니라 반복적인 상호작용의 대상으로 활용되고 있음을 보여준다.

인공지능 동반자(AI Companion)는 지속적인 상호작용을 통해 이용자와 관계적 경험을 형성하는 시스템을 의미한다[3]. 유아 대상 인공지능 동반자는 이용자의 발달적 특성을 반영한 설계가 요구된다. 특히 유아는 언어적, 인지적, 사회정서적 능력이 발달하는 과정에 있으므로, 유아용 인공지능 동반자의 설계에서는 기술적 성능뿐 아니라 상호작용 과정에서 발생할 수 있는 잠재적 위해를 예방하고 최소화하기 위한 안전성 확보가 중요하다[4-5].

유아와 인공지능 간 대화의 안전성을 판단할 때, 유아 발화의 맥락적 특성으로 인해 서로 구별되는 두 가지 쟁점을 고려할 필요가 있다. 첫째, 유아기의 또래 상호작용에서는 관계적 배제와 공격성이 나타나는 한편, 상상놀이에서도 공격과 관련된 주제가 표현될 수 있다[6-7]. 이에 따라 유해 발화와 표면적으로 유사한 표현이라도 실제 갈등 상황에서 사용된 것인지, 놀이와 같은 비유해적 상호작용 맥락에서 사용된 것인지에 따라 안전상 의미가 달라질 수 있다. 이러한 특성으로 인해 표현의 표면적 특징에 의존한 판단은 맥락상 무해한 발화를 유해한 것으로 해석할 가능성을 내포한다. 둘째, 명시적인 유해 표현이 포함되지 않은 발화에서도 부정적 정서나 정서적 곤란을 나타내는 표현이 관찰될 수 있으며, 이러한 발화에 대한 지원적 대응의 필요성이 논의되어 왔다[8-9]. 따라서 유아 발화의 안전성 판단에서는 유해 표현의 식별뿐 아니라, 표면적으로 유해해 보이지만 맥락상 무해한 발화를 구별하는 문제와 유해성 여부와 별개로 지원적 대응이 필요한 발화를 식별하는 문제를 함께 고려해야 한다.

유해언어 탐지는 온라인 환경에서 나타나는 공격·혐오적 표현을 식별하기 위한 기술적 접근으로 발전해 왔다[10]. 한국어에서도 온라인 뉴스 댓글을 중심으로 유해 여부와 공격 대상 및 세부 유형을 판별하기 위한 말뭉치와 분류체계가 제안되었다[11-12]. 그러나 유아 발화에서 맥락상 무해한 표현의 오분류와 지원적 대응이 필요한 발화의 식별을 함께 평가한 연구는 제한적이다. 기존 유해언어 탐지 연구의 적용 범위와 유아 발화의 특성을 고려할 때, 맥락무해 발화의 판별과 지원 필요 탐지를 함께 평가할 필요가 있다.

본 연구는 유아용 인공지능 동반자의 안전 가드레일에 활용할 입력 발화 판단 기능을 대상으로 학습 구성에 따른 맥락무해 발화의 판별과 지원 필요 탐지의 차이를 분석한다. 본 연구에서 평가하는 범위는 개별 입력 발화에 대한 분류 탐지이며, 판단 이후의 생성, 차단, 지원 행동을 실행하는 시스템 전체의 평가와는 구분된다. 맥락무해 발화는 유해 표현과 표면적 특징을 공유하지만 놀이, 감정 표현 또는 안전교육의 맥락에서 무해로 분류되는 발화이며, 지원 필요 속성은 자기지향적 부정표현과 무해 범주의 정서적 곤란 표현(Self-Distress) 유형을 정서지지 응답대상으로 지정한 정책 라벨이다. 기본 의미 범주와 두 안전 속성을 구분한 라벨 체계를 바탕으로 분리학습, 공동학습 및 복합손실 공동학습을 비교한다. 이를 통해 유아 발화의 맥락적 특성과 후속 대응 필요성을 반영한 안전성 평가의 필요성을 검토하고, 유아용 인공지능 동반자의 안전 가드레일의 입력 발화 판단 단계에 대한 평가 근거를 제시하고자 한다.

II. Literature Review

1. Interaction and Safety in Preschool AI Companions

대화형 인공지능은 이용자의 발화에 반응하며 지속적인 대화와 상호작용을 수행하는 시스템이다. 아동과 대화형 인공지능의 상호작용에서는 신뢰와 친밀감을 포함한 사회적·관계적 경험이 형성될 수 있다[13-14]. 유아는 인공지능에 사회적 속성을 부여하고, 놀이와 이야기 구성 및 대화 활동에 이를 활용한다[15-16]. 이러한 연구는 유아용 인공지능 동반자가 반복적인 상호작용을 통해 사회적·관계적 경험을 형성할 수 있음을 시사한다.

취학 전 아동이 로봇에 심리적 속성과 생물학적 속성을 부여하는 양상은 연령에 따라 차이를 보인다[17]. 이는 유아의 인공지능에 대한 사회적 인식과 해석이 발달 수준에 따라 달라질 수 있음을 시사한다. 따라서 유아용 인공지능 동반자의 상호작용과 대응 기준을 설정할 때, 유아의 상호

www.kci.go.kr

작용적 특성과 발달적 특성을 고려할 필요가 있다. 이러한 관점은 아동 대상 인공지능의 설계에서 위해 예방뿐 아니라 아동의 권리와 발달적 특성을 함께 고려해야 한다는 아동 중심 인공지능의 원칙과 부합한다[4]. 또한 UNICEF 역시 아동 대상 인공지능의 설계와 활용에서 아동 최선의 이익과 발달 및 복지를 주요 원칙으로 제시하고 있다[5].

유아용 인공지능 동반자의 발화 수준 안전성에는 부적절한 콘텐츠의 차단뿐 아니라 발화의 의미와 기능에 부합하는 후속 대응이 포함된다. 그러나 유아 발화의 맥락적 특성을 반영하여 유해성을 판단하고, 발화의 의미에 따라 지원 필요성을 구분하는 방법에 관한 실증적 연구는 제한적이다.

2. Contextual Interpretation and Support Needs in Harmfulness Detection

유해언어 탐지 연구는 온라인 환경에서 나타나는 공격적 표현과 혐오적 표현을 식별하는 방향으로 발전해 왔다[10]. 한국어 유해언어 탐지 연구에서도 온라인 뉴스 댓글과 같은 공개 온라인 담화를 대상으로 유해 발화를 식별하고 공격 대상 및 세부 유형을 분류하기 위한 데이터셋과 분류체계가 제안되어 왔다[11-12]. 이러한 연구는 유해 발화를 자동으로 식별하기 위한 기술적 기반을 제공하였으나, 주요 분석 대상이 공개 온라인 담화라는 점에서 유아-인공지능 동반자 간 대화에 동일한 판단 기준을 적용하기에는 한계가 있다. 유아 발화는 놀이와 또래 관계 등 다양한 상호작용 맥락에서 나타나며, 공격적 표현도 상상놀이에서는 가상의 상황을 재현하는 기능을, 실제 또래 관계에서는 관계적 공격의 기능을 수행할 수 있다[6-7]. 이는 표면적으로 유사한 표현이라도 발화의 대상과 기능, 그리고 상호작용 맥락에 따라 유해성의 의미가 달라질 수 있음을 보여준다.

유아 발화의 이러한 특성은 유해언어 탐지에서 논의되어 온 문맥 의존성과 연결되며, 동일하거나 유사한 표현도 주변 문맥에 따라 유해성 판단이 달라질 수 있다[18]. 또한 모델이 특정 어휘나 표현을 유해성의 주요 단서로 과도하게 활용할 경우 문맥상 무해한 표현까지 유해한 것으로 분류할 수 있다[19-20]. 유아 발화에서는 놀이, 감정 표현 등 서로 다른 상호작용 맥락에서 부정적 표현과 공격적 표현이 나타날 수 있으므로, 이러한 표현의 표면적 특징만으로는 해당 발화가 실제로 수행하는 기능을 충분히 구별하기 어렵다. 따라서 기존 유해언어 탐지에서 제기되어 온 문맥 의존성과 특정 언어적 단서에 대한 과도한 의존 문제는 유아 발화의 안전성 판단에서도 고려되어야 하며, 특히 맥락상 무해한 발화를 유해한 것으로 판단하는 오류 가능성을 구분하여 살펴볼 필요가 있다.

유아 발화의 유해성 판단과 후속 대응의 필요성은 서로 구분하여 고려해야 한다. 특히 자기지향적 부정표현과 정서적 곤란 표현은 타인에 대한 공격이나 위협과 구별되는 특성을 가지며, 발화의 유해성 여부와 별개로 지원적 대응의 대상이 될 수 있다. 유아와 성인 간의 정서 대화 및 부정적 정서에 대한 반응이 유아의 정서 이해와 조절과 관련하여 논의되었다[21]. 이러한 논의를 바탕으로 유아 발화의 안전성 판단에서는 맥락상 무해한 표현의 식별과 유해성 여부와 별개로 나타나는 지원 필요성의 식별을 구분하여 다룰 필요가 있다.

3. Joint Learning and Error Evaluation of Safety Attributes

유해언어 탐지 모델의 특정 어휘 의존과 조건별 오분류는 전체 분류 성능만으로 충분히 드러나지 않을 수 있다[19-20]. 하위집합 및 언어적 조건에 따른 평가 [22-23]와 아동의 개인적 발화와 정서적 발화에 대한 대응 문제[24]를 고려하면, 성능을 검토할 때, 전체 분류성능과 특정 발화 조건의 오류를 함께 살펴볼 필요가 있다. 이러한 평가 관점은 의미 범주와 안전 속성을 함께 학습하는 구성의 적절성을 검토하는 근거가 된다.

발화의 기본 의미 범주와 안전 속성은 상호 관련되지만 서로 구별되는 판단 대상이다. 동일한 의미 범주에서도 후속 대응의 필요성이 달라질 수 있으며, 서로 다른 의미 범주가 공통적인 안전 속성을 가질 수 있다. 따라서 의미 범주와 안전 속성의 구분 및 상호 관련성을 고려한 학습 구성이 요구된다.

서로 관련되지만 구별되는 복수의 학습목표를 함께 다루기 위한 대표적인 접근으로 다중과제학습(Multi-Task Learning)이 활용되어 왔다. 다중과제학습은 관련 과제 간 표현을 공유하면서 각 과제의 학습목표를 동시에 최적화하는 접근이다[25-26]. 유해언어 탐지에서도 여러 도메인의 혐오 표현과 감성 및 풍자 분류 정보를 공동으로 활용하기 위해 다중과제학습이 적용되어 왔다[27]. 이는 단일한 분류목표에 여러 유형의 정보를 포괄하기보다, 관련된 정보를 구별된 학습목표로 구성하면서 공유 표현을 통해 상호 관련성을 학습할 수 있음을 보여준다.

지도대조학습(Supervised Contrastive Learning)은 동일 범주의 표상을 가깝게 배치하고, 서로 다른 범주의 표상을 멀리 배치하여 표현 구조를 학습하는 방법이다[28]. 관련 과제의 공동학습과 표현학습은 의미 범주와 안

www.kci.go.kr

전 속성을 함께 다루는 방법론적 근거를 제공한다. 이에 본 연구는 분리학습, 공동학습 및 복합손실 공동학습을 비교하여 전체 분류 성능과 의미 범주 및 안전 속성의 교차 조건에서 나타나는 탐지 오류를 비교하고자 한다.

III. Materials and Methods

1. Dataset and Labeling Scheme

본 연구에서는 만 3-5세 유아로부터 수집한 음성 발화를 전사하여 구축한 한국어 발화 데이터셋 27,605건을 사용하였다. 유아교육, 유아 언어발달, 발달심리, 사회정서발달, 아동안전 및 보호정책 관련 전문가 자문을 통해 유해 언어 분류체계와 라벨링 기준을 설계하였다. 발화는 무해 범주와 5개의 유해 세부 범주로 구성하였다. 무해 범주는 일상적인 대화와 놀이 발화를 포함하며, 유해 범주는 개인 특성에 대한 공격, 또래 관계에서의 배제, 자기지향적 부정표현, 신체·외모에 대한 비하, 위협 및 공격 행동의 선동으로 구분하였다. 각 범주는 유아 발화의 의미와 안전 가드레일의 후속 대응 방향을 함께 고려하여 설정하였다. 데이터셋의 범주별 구성은 Table 1과 같다.

자료 수집은 두 유아교육기관에서 5개월간 이루어졌으며, 연구대상은 만 3세 66명, 만 4세 72명, 만 5세 169명으로 총 307명이다. 유아 발화 데이터는 수업 및 놀이 상황에서의 개별면담, 사회정서적 상황을 활용한 일대일 대화와 상황 발문으로 구축하였다. 자료 수집은 보호자의 동의와 아동의 참여 의사를 확인한 후 진행하였으며, 자발적 참여 중단과 자료 폐기 요청에 관한 절차를 안내하였다.

기본 의미 범주와 함께 맥락무해 여부와 지원 필요 여부를 안전 관련 속성으로 구성하였다. 맥락무해 발화는 표면적으로 유해 발화와 유사한 표현을 포함하지만 놀이, 감정표현 또는 안전교육의 맥락에서 무해로 분류되는 발화이며, 2,379개 모두 무해 범주 11,137개에 포함된다. 지원 필요 속성은 최종 의미 범주와 세부 유형을 기준으로 정서지지 응답 대상을 지정한 정책 라벨이다. 자기지향적 부정표현 3,409건과 무해 범주의 정서적 곤란 표현 427건 등 총 3,836건을 지원 필요 해당 사례(positive)로 분류하였다(Table 2). 무해 범주의 도움 요청 208건과 정당한 감정표현 397건을 포함한 나머지 발화는 지원 필요 비해당 사례(negative)로 분류하였다. 여기서 '비해당'은 본 연구에서 정의한 정서지지 응답 정책의 적용 대상에 포함되지 않음을 뜻하며, 실제 상호작용에서 도움이나 정서적 지원이 불필요함을 의미하지 않는다.

Table 1. Composition of the Dataset

Category		Count
Non-harmful		11,137
Harmful	Identity Attack	3,558
	Social Exclusion	3,550
	Self-Perception Negation	3,409
	Body & Appearance Disparagement	3,293
	Threat & Attack Incitement	2,658
	Subtotal	16,468
Total		27,605

Table 2. Positive instance counts for safety attributes

Safety Attribute	Positive Subset	Count
Support-Need	Self-Perception Negation	3,409
	Non-harmful emotional distress	427
Subtotal		3,836
Contextually Benign	Non-harmful contrast set	2,379

라벨 품질 검증은 라벨과 안전 속성 간 매핑의 일관성, 중복·결측 및 규칙 위반 여부에 대한 검사와 전문가 평가로 구성하였다. 유아교육, 발달심리, 사회정서발달, 아동안전 및 보호정책, 유아 언어발달 분야의 전문가 5명이 전체 27,605개 발화를 독립적으로 평가하였다. 기존 라벨을 제공하지 않은 상태에서 발화를 무작위 순서로 제시하였으며, 전문가들은 발화 텍스트와 함께 상황 설명 및 전후 대화를 참고하여 사전에 정의된 기준에 따라 의미 범주를 판정하였다. 전문가 5인 중 3인 이상이 동일한 범주를 선택한 경우를 과반수 동의로 정의하였으며, 해당 기준에 도달하지 못한 발화는 이견 사례로 구분하였다.

전문가 라벨링 및 검증에는 관련 분야의 교수 및 연구원 5명이 참여하였으며, 성별은 남성 1명과 여성 4명으로 구성되었다. 2025년 라벨링 당시 참여 전문가의 연령대는 30대 3명, 40대 1명, 50대 1명이었다. 관련 전문 분야는 유아교육 및 심리학 교육, 아동발달 지원, 아동·청소년 및 부모 대상 심리상담 등을 포함하였다. 의미 범주에 대한 전문가 간 평가 일치도는 전체 발화를 대상으로 Fleiss' κ와 Krippendorff's α, 전문가 5인으로 구성되는 10개 평가자 쌍의 Cohen's κ를 산출하여 검토하였다. Fleiss' κ는 0.9472, 평균 쌍별 Cohen's κ는 0.9473, Krippendorff's α는 0.9472로 나타났다. 3인 이상의 과반수 판정이 형성되지 않은 사례는 67건으로 전체의 0.24%

www.kci.go.kr

를 차지하였다. 전문가 5인의 최다득표 범주와 최종 분석 라벨 간 일치율은 83.9%로 나타났다.

데이터셋의 라벨링 신뢰성과 정합성을 평가하기 위해 전체 27,605건의 발화를 대상으로 5인의 전문가 라벨링 결과와 최종 분석 라벨을 비교·분석하였다. 발화별 대표 범주는 5인이 부여한 라벨 중 최빈값을 기준으로 결정하였으며, 최빈값이 동일한 경우 저장된 라벨링 결과의 열 순서에서 먼저 나타나는 범주를 대표 범주로 선정하였다. 이를 바탕으로 대표 범주와 최종 분석 라벨 간의 정합성을 분석하고, 과반수의 동의가 형성되지 않은 사례를 대상으로 범주별 분포와 라벨링 특성을 추가적으로 검토하였다.

전체 데이터셋의 발화 식별자, 발화 텍스트 및 분석 라벨은 대조 자료와 일치하는 것으로 확인되었다. 전문가 라벨링 결과와 최종 분석 라벨을 비교한 결과, 전체 27,605건 중 4,456건에서 범주 간 불일치가 관찰되었으며, 처리 상태가 'confirmed' 이외의 값으로 기록된 사례는 4,472건이었다. 두 지표는 각각 라벨 정합성과 데이터 처리상태를 나타내는 변수로 구분하여 분석하였다. 주 평가집합 1,996건에서는 대표 범주와 최종 분석 라벨이 일치한 사례가 1,633건(81.8%), 불일치한 사례가 363건(18.2%)으로 나타났으며, 이 중 과반수의 동의가 형성되지 않은 사례는 5건이었다.

전체 데이터셋에서 전문가 5인의 독립 평정 결과, 과반수 동의가 형성되지 않은 사례는 총 67건이었다. 이들 사례는 평가자별 판단 근거, 발화의 상황 맥락 및 사전에 정의한 범주별 기준을 검토하는 전문가 협의를 거쳐 최종 의미 범주를 확정하였다. 협의 전 독립 평정의 대표 범주와 최종 분석 라벨이 일치한 사례는 16건, 불일치한 사례는 51건이었다. 최종 분석 라벨은 사회적 배제 34건(50.7%), 정체성 공격 17건(25.4%), 위협·공격 유도 8건(11.9%), 신체·외모 비하 6건(9.0%), 무해 2건(3.0%)으로 분포하였다. 협의 전 라벨 선택 분포는 2-2-1 형태가 60건(89.6%), 2-1-1-1 형태가 7건(10.4%)으로, 복수 범주에 걸쳐 판단이 분산되었다.

과반수 동의가 형성되지 않은 67건의 협의 전 독립 평정에서 범주 간 동시 선택 빈도를 분석한 결과, 무해와 정체성 공격이 함께 선택된 사례는 49건, 무해와 사회적 배제가 함께 선택된 사례는 40건, 사회적 배제와 정체성 공격이 함께 선택된 사례는 25건이었다. 동시 선택 빈도는 동일 발화에 대한 5인의 평정에서 두 범주가 모두 나타난 발화 수로 산출하였으며, 하나의 발화가 여러 범주 쌍에 중복 집계될 수 있도록 하였다. 이러한 분포는 일부 발화의 유해성 여부와 세부 범주에 대한 전문가 판단이 서로 달랐음을 보여준다. 전문가 간 일치도, 라벨 선택 분포 및 범주 간 동시 선택 빈도는 모두 협의 전 독립 평정값을 기준으로 산출하였다.

대표적인 라벨링 불일치 사례인 “혼자 둘 거야”는 무해 2명, 사회적 배제 2명, 위협·공격 유도 1명으로 라벨이 분산되었으며, 최종 분석 라벨은 위협·공격 유도로 분류되었다. “나 진짜 별로야”는 전문가 라벨링 결과에서 자기지향적 부정표현이 최빈 범주로 나타났으나, 최종 분석자료에서는 무해 범주의 self_distress로 분류되었다. 또한 과반수 동의가 형성되지 않은 67건 중 8건에서는 최종 분석 라벨이 전문가 5인이 선택한 범주에 포함되지 않는 것으로 확인되었다. 이러한 사례는 전문가 라벨링의 대표 범주와 최종 분석 라벨 간 차이를 보여주며, 라벨 정합성의 특성을 파악하기 위한 주요 분석 대상으로 활용되었다.

발화의 언어적 특성에 따른 전문가 간 판단 차이를 살펴보기 위해 라벨링 불일치 사례에 대한 사후 분석을 수행하였다. “너 때문에 다 망가졌어”는 정체성 공격 2명, 무해 2명, 위협·공격 유도 1명으로 분류되었으며, “왜 친구 없어”는 무해 2명, 정체성 공격 2명, 사회적 배제 1명으로 분류되었다. 전자는 특정 사건에 대한 책임 귀속과 상대방에 대한 비난의 의미가 함께 나타나는 표현이며, 후자는 사회적 관계에 관한 질문과 부정적 평가의 의미가 중첩될 수 있는 표현이다. 두 사례에서 관찰된 라벨링 분포는 발화의 의미적 특성과 문맥적 해석에 따라 전문가의 범주 판단이 달라질 수 있음을 보여준다.

전문가 라벨링의 신뢰성을 정량적으로 검증하기 위해 전체 데이터셋에 저장된 5인의 라벨링 결과를 대상으로 범주별 Fleiss’ κ를 산출하였다. 각 의미 범주에 대해 해당 범주와 나머지 범주를 구분하는 이진 분류체계를 적용하여 전문가 간 일치도를 평가하였다. 분석 결과, 무해는 0.9443, 정체성 공격은 0.9336, 사회적 배제는 0.9140, 자기지향적 부정표현은 0.9825, 신체·외모 비하는 0.9664, 위협·공격 유도는 0.9299로 나타났다. 모든 의미 범주에서 0.90 이상의 높은 일치도가 확인되었으며, 자기지향적 부정표현에서 가장 높은 일치도가 나타난 반면 사회적 배제에서는 상대적으로 낮은 일치도가 관찰되었다.

종합적으로, 의미 범주에 대한 전문가 간 일치도는 높게 나타났으나, 일부 발화에서는 전문가 대표 범주와 최종 분석 라벨 간 차이가 확인되었다. 평가자 간 일치도는 전문가 판단의 일관성을, 최종 분석 라벨과의 정합성은 전문가 대표 범주와 분석에 사용된 라벨의 대응 정도를 나타내므로 두 결과를 구분하여 해석하였다. 주관적 주석 과제에서 평가자 간 불일치는 서로 다른 해석을 반영할 수 있다는

www.kci.go.kr

논의[29]를 고려하여, 이견 사례를 범주 판단의 특성을 검토하기 위한 자료로 분석하였다.

지원 필요 정책 매핑 결과의 일치도를 보조적으로 평가하기 위해, 저장된 평가자별 의미 범주 평정값을 이진값으로 변환하였다. 각 평가자가 발화를 자기지향적 부정표현으로 분류한 경우에는 1을 부여하였다. 무해로 분류한 경우에는 최종 분석자료의 하위유형이 정서적 곤란 표현(Self-Distress)인 발화에 1을 부여하였으며, 나머지는 0으로 처리하였다. 최종 하위유형 정보는 모든 평가자에게 공통으로 적용하였다.

전체 27,605건에 대한 5인의 평가자별 이진값을 분석한 결과, Fleiss' κ는 0.9848, 평가자 쌍별 관측 일치율은 99.63%였다. 5인의 이진값이 모두 일치한 발화는 27,370건(99.15%)이었으며, 나머지 235건에서는 평가자 간 차이가 나타났다. 이 결과는 공통 하위유형 정보를 적용한 조건에서 평가자별 의미 범주 판단을 지원 필요 정책에 따라 변환한 이진값의 일치도를 나타낸다.

Table 3은 최종 데이터셋에 포함된 대표 발화와 해당 라벨의 분류 사례를 제시한다. "나 진짜 별로야"는 무해 범주의 self_distress로 분류되며, 정서적 지원이 필요한 양성 사례에 해당한다. 반면, 도움 요청과 정당한 감정 표현은 지원 필요 여부의 음성 사례로 분류된다. 또한 "넌 짜증쟁이야"와 "짜증날 땐 잠깐 쉬자"는 '짜증'이라는 공통 표현을 포함하지만, 각각 정체성 공격과 무해 범주로 구분된다. 이는 동일한 어휘를 포함한 발화에서도 표현의 의미와 사용 맥락에 따라 서로 다른 라벨이 적용될 수 있음을 보여준다. 본 연구에서는 자기지향적 부정표현 범주 전체와 무해 범주의 정서적 곤란 표현에 지원 필요 양성 라벨을 부여하였으며, 그 외 범주는 음성으로 구분하였다. 이러한 분류체계를 바탕으로 발화의 유해성 유형과 정서적 지원 필요성을 구분하여 분석하였다.

Table 3. Stored examples and support policy mapping

Label	Example
Support-need positive	"I cannot draw."
	"I'm really no good."
Support-need negative	"Can I borrow an eraser for a moment?"
	"My best friend is in a different class now, and I'm so sad."
Contextually benign	"It's not scary at all when the light is on."
	"Let's take a short break when you're annoyed."
Harmful	"You're a grump."

2. Data Splits and Evaluation Tasks

데이터는 학습, 검증 및 평가집합으로 분할하였다. 전체 27,605건 발화 중 맥락무해 발화 2,379건을 제외한 25,226건을 주 데이터셋으로 구성하였다. 주 데이터셋에는 무해 발화 8,758건과 5개 유해 범주의 발화 16,468건이 포함되며, 이를 학습 20,907건, 검증 2,323건, 평가 1,996건으로 나누었다. 주 데이터셋의 범주별 분할 결과는 Table 4와 같다.

맥락무해 발화 2,379건은 학습 1,189건과 진단 평가 1,190건으로 분할하였다. 분할에는 난수시드 42를 적용하였으며, 학습 시드와 관계없이 동일한 분할을 유지하였다. 분리학습, 공동학습 및 복합손실 공동학습에는 주 학습집합과 맥락무해 학습집합을 합한 22,096건 발화를 동일하게 제공하였다. 맥락무해 검증집합은 별도로 구성하지 않았으며, 진단 평가집합은 맥락무해 발화를 유해한 것으로 예측하는 오류를 측정하는 데 사용하였다. 동일한 텍스트는 같은 분할에 배치하여 학습집합과 평가집합 간 직접적인 문자열 중복을 방지하였고, 세 학습 조건에는 동일한 주 평가집합과 진단 평가집합을 적용하였다.

Table 4. Train/Validation/Test split of the Main Dataset

Category		Train	Val	Test
Non-harmful		7,278	809	671
Harmful	Identity Attack	2,946	328	284
	Social Exclusion	2,937	326	287
	Self-Perception Negation	2,824	313	272
	Body & Appearance Disparagement	2,721	301	271
	Threat & Attack Incitement	2,201	246	211
	Subtotal	13,629	1,514	1,325
Total		20,907	2,323	1,996

지원 필요 양성은 학습집합 3,176건, 검증집합 352건, 주 평가집합 308건이었다(Table 5). 각 집합은 자기지향적 부정표현 2,824건, 313건, 272건과 무해 범주의 지원 필요 발화 352건, 39건, 36건으로 구성되었다. 지원 필요 양성은 모두 주 데이터셋에 포함된 발화의 부분집합이며, 맥락무해 학습집합과 진단 평가집합에는 포함되지 않았다.

모델의 입력은 개별 발화의 텍스트로 한정하였으며, 외부 상황 설명이나 앞선 대화 이력은 추가하지 않았다. 따라서 맥락무해 발화의 평가는 입력 텍스트에 포함된 단서에 근거한 판별 성능을 대상으로 하였다. 각 발화에는 기

www.kci.go.kr

본 의미 범주, 유해성 이진 여부, 지원 필요 여부 및 맥락무해 여부의 정답정보를 사용하였다. 유해성 이진 정답은 기본 의미 범주를 기준으로 구성하였으며, 세 안전 속성 관련 예측은 각각 별도의 이진 출력층에서 산출하였다. 맥락무해 학습 발화에는 기본 의미 범주상 무해, 지원 필요 비해당 및 맥락무해 양성의 정답정보를 부여하였다.

안전 속성의 탐지 성능을 의미 범주별로 검토하기 위해 전체 평가와 부분집합 평가를 수행하였다. 주 평가집합에서는 전체 분류 성능, 유해 발화 누락률, 지원 필요 누락률 및 오탐률을 산출하였으며, 진단 평가집합에서는 맥락무해 오탐률을 산출하였다. 지원 필요 탐지는 양성 308건과 음성 1,688건을 대상으로 평가하고, 양성 집합을 자기지향적 부정표현 272건과 무해 발화 36건으로 나누어 누락률을 비교하였다.

Table 5. Distribution of Support-Need Positive Instances

Category	Train	Val	Test
Self-Perception Negation	2,824	313	272
Non-harmful emotional distress	352	39	36
Total	3,176	352	308

3. Learning Configurations and Objectives

학습 구성에 따른 안전 기능의 차이를 분석하기 위해 분리학습, 공동학습 및 복합손실 공동학습을 비교하였다. 세 조건에는 동일한 학습자료와 정답정보를 제공하고, 의미 범주와 안전 속성 간 인코더 공유 여부 및 표현 손실의 적용 여부를 달리하였다.

학습목표는 6범주 분류, 유해성 이진 판별, 지원 필요 탐지 및 맥락무해 발화 구별로 구성하였다. 6범주 분류는 무해와 다섯 유해 세부 범주를 구별하며, 유해성 이진 판별은 다섯 유해 범주를 통합한 이진 정답을 사용한다. 이진 예측은 6범주 예측을 변환하지 않고 별도의 출력층에서 산출하였다. 지원 필요 탐지와 맥락무해 발화 구별은 각각 정서지지 응답 대상으로 지정한 정책 라벨과 맥락무해 속성의 해당 여부를 예측한다.

분리학습에서는 의미 범주를 학습하는 인코더와 안전 속성을 학습하는 인코더를 구분하였다. 의미 인코더는 6범주 분류와 유해성 이진 분류를 수행하고, 속성 인코더는 지원 필요 탐지와 맥락무해 발화 분류를 담당하였다. 두 인코더는 동일한 학습 발화를 입력받되, 각각 담당하는 과제의 손실을 통해 독립적으로 갱신된다. 공동학습에서는 하나의 인코더에 네 과제의 출력층을 연결하여 의미 범주와 안전 속성을 함께 학습하도록 구성하였다. 복합손실 공동학습에서는 발화 간 표현의 구별을 강화하기 위해 기존 공동학습의 목적함수에 지도대조손실과 마진손실을 추가하였다.

분리학습과 공동학습의 비교에서는 과제별 인코더의 분리 및 공유에 따른 성능 차이를 검토하였다. 분리학습은 두 인코더를 사용하므로, 두 구성의 차이에는 모델 용량, 계산량 및 최적화 방식의 차이도 포함된다. 공동학습과 복합손실 공동학습의 비교에서는 공유 인코더 구조를 유지한 상태에서 표현 손실을 추가하였을 때 분류 성능과 안전 관련 오류가 어떻게 달라지는지 분석하였다.

각 인코더의 첫 토큰 표현에 드롭아웃과 과제별 선형 출력층을 적용하였다. 6범주 분류, 유해성 이진 판별, 지원 필요 탐지 및 맥락무해 발화 구별의 교차엔트로피 손실을 각각 L_{σ}, L_{β}, L_{s}, L_{p}로 정의하였다. 분리학습의 의미 인코더와 속성 인코더는 식 (1)과 (2)의 목적함수를 각각 최소화하며, 공동학습은 네 과제의 손실을 결합한 식 (3)을 최소화한다. 지원 필요 손실과 맥락무해 손실의 가중치는 각각 0.30과 0.10으로 설정하고, 세 학습 조건에서 동일하게 유지하였다.

$$L_{sep,\,semantic} = L_{\sigma} + L_{\beta} \quad (1)$$

$$L_{sep,\,attribute} = 0.30L_{s} + 0.10L_{p} \quad (2)$$

$$L_{joint} = L_{\sigma} + L_{\beta} + 0.30L_{s} + 0.10L_{p} \quad (3)$$

복합손실 공동학습에서는 분류 정답에 대한 학습과 함께 발화 간 표현의 구별을 반영하기 위해 지도대조 손실 $L_{contrast}$와 마진 손실 L_{margin}을 추가하였다. 두 손실의 가중치는 각각 0.10으로 설정하였으며, 전체 목적함수는 식 (4)와 같다.

$$L_{combined} = L_{joint} + 0.10L_{contrast} + 0.10L_{margin} \quad (4)$$

지도대조 손실은 같은 유해 세부 범주에 속하는 발화의 표현을 가깝게 배치하고, 서로 다른 유해 세부 범주의 표현을 구별하도록 구성하였다[28]. 마이크로배치 내 정답 유해 발화의 집합을 H, 발화 i의 인코더 출력을 정규화한 표현을 z_i로 정의하였다. P(i)는 H에서 발화 i와 동일한 유해 세부 범주에 속하는 다른 발화의 집합이며, A는 P(i)가 공집합이 아닌 앵커의 집합이다. 지도대조 손실은 식 (5)와 같이 계산하였다. 분모에서는 앵커 자신을 제외하였으며, 온도 τ는 0.10으로 설정하였다.

www.kci.go.kr

$$L_{contrast} = -\frac{1}{|A|}\sum_{i \in A}\frac{1}{|P(i)|} \times \sum_{p \in P(i)} \log \frac{\exp(\frac{z_i \cdot z_p}{\tau})}{\sum_{a \in H, a \neq i} \exp(\frac{z_i \cdot z_a}{\tau})} \quad (5)$$

마진 손실은 유해 발화와 맥락무해 발화가 표면적으로 유사한 표현을 포함하더라도 표현 공간에서 구별되도록 구성하였다. 마이크로배치 내 맥락무해 발화의 집합을 Q로 정의하고, 유해 발화와 맥락무해 발화 사이의 코사인 유사도가 마진 0.20을 초과하는 경우에 손실을 부과하였다. 정규화된 표현의 내적을 이용한 마진 손실은 식 (6)과 같다.

$$L_{margin} = \frac{1}{|H||Q|} \times \sum_{h \in H}\sum_{q \in Q} \max(0, z_h \cdot z_q - 0.20) \quad (6)$$

두 표현 손실은 32개 발화로 구성된 마이크로배치 내에서 계산하였으며, 유효한 발화 쌍이 없으면 해당 손실을 0으로 처리하였다. 지도대조 손실은 자기지향적 부정표현을 포함한 다섯 유해 세부 범주를 대상으로 하며, 마진 손실은 유해 발화와 맥락무해 발화의 구별을 대상으로 한다. 따라서 두 손실은 무해이면서 지원이 필요한 발화의 식별을 직접 최적화하지 않으며, 지원 필요 속성은 해당 과제의 교차엔트로피 손실을 통해 직접 학습된다.

맥락무해 학습 발화는 6범주 분류와 유해성 이진 판별에서 무해 정답으로 학습하고, 맥락무해 발화 구별에서는 양성으로 학습하였다. 주 데이터셋의 발화는 맥락무해 속성의 음성 사례로 사용하였다. 추론 시에는 각 과제의 출력층에서 예측을 산출하였으며, 맥락무해 속성의 예측값을 이용하여 6범주 예측을 변경하는 후처리는 적용하지 않았다. 이에 따라 맥락무해 오탐률은 속성 예측에 따른 보정 결과가 아니라, 학습된 의미 범주 출력에서 발생한 오류를 나타낸다.

교차 엔트로피(Cross Entropy, CE) 기준선의 기존 결과는 맥락무해 학습자료와 안전 속성 감독정보가 없는 조건의 참조값으로 사용하였다. 이 기준선은 주 학습자료에서 6범주 분류와 유해성 이진 판별의 교차엔트로피 합을 최소화하며, 지원 필요 및 맥락무해 속성을 학습하지 않는다. 따라서 학습자료와 감독정보를 통제한 비교는 분리학습, 공동학습 및 복합손실 공동학습 간에 수행하였다.

학습자료와 속성 감독에 따른 차이를 비교하기 위해 동일 인코더, 시드, 고정 분할에서 CE_main, CE_aug 및 공동 학습 조건을 비교하였다. CE_main은 주 학습자료 20,907건으로 6범주 및 유해성 이진 CE를 학습하였다. CE_aug는 맥락무해 학습자료를 포함한 22,096건으로 같은 두 목적함수를 학습한 기존 분리학습의 의미 인코더이다. 여기에 지원 필요 CE 0.3과 맥락무해 CE 0.1을 함께 추가한 조건은 기존 공동학습을 의미한다. 세 조건의 공통 비교 지표는 분류 성능과 맥락무해 오탐률이다. 같은 3에포크에서 CE_main은 654회, CE_aug와 공동학습은 693회 갱신하였다.

4. Training Settings

학습 구성에 따른 차이가 특정 사전학습 인코더에 한정되는지 검토하기 위해 KLUE-RoBERTa와 KLUE-BERT[30], KcBERT[31], KoELECTRA[32], KR-ELECTRA[33]의 다섯 한국어 사전학습 인코더를 사용하였다. 각 인코더에 분리학습, 공동학습 및 복합손실 공동학습을 적용하고, 동일한 데이터 분할에서 세 학습 시드로 반복 실험을 수행하였다. 이를 통해 학습 조건 간 평균 성능의 차이와 함께 인코더 및 학습 시드에 따른 변동을 검토하였다.

세 학습 조건에는 동일한 입력 길이, 배치 크기, 최적화 방법 및 학습 횟수를 적용하였다. 최대 입력 길이는 128토큰으로 설정하였으며, 각 인코더에 대응하는 토크나이저를 사용하였다. 최적화에는 AdamW를 사용하고, 학습률은 2×10^{-5}, 가중치 감쇠는 0.01로 설정하였다. 과제별 출력층에 앞서 적용하는 드롭아웃 비율은 0.10으로 설정하였다. 전체 갱신 단계의 10%를 워밍업 구간으로 사용한 후 학습률을 선형으로 감소시켰으며, 기울기 클리핑의 최대 노름은 1.0으로 설정하였다.

학습 시드는 42, 4337, 2026을 사용하였으며, 데이터 분할은 학습 시드와 관계없이 고정하였다. 동일한 사전학습 인코더와 시드에서는 초기 상태와 학습 발화의 배치 순서를 일치시켰다. 마이크로배치 크기는 최대 32로 설정하고, 일반 배치에서는 세 마이크로배치의 기울기를 누적하여 유효 배치 크기를 96으로 구성하였다. 3.3에서 정의한 지도대조 손실과 마진 손실은 각 마이크로배치 내부의 발화 쌍을 이용하여 계산하였다.

세 학습 조건은 세 에포크 동안 학습하였으며, 인코더별 매개변수 갱신 횟수는 693회로 동일하게 유지하였다. 검증 성능에 따른 조기 종료나 중간 체크포인트 선택은 수행하지 않고, 마지막 에포크의 모델을 평가에 사용하였다. 검증집합은 학습 과정의 성능 확인과 지원 필요 탐지의 사후 탐색적 보조 분석을 위한 임계값 설정에 사용하였다. 손실 가중치, 온도 및 마진은 3.3에서 제시한 값으로 고정

www.kci.go.kr

하였다.

전체 실험은 다섯 인코더, 세 학습 조건 및 세 학습 시드의 조합으로 구성하여 총 45개 시스템을 평가하였다. 분리학습에서는 시스템마다 두 인코더를 학습하므로, 전체 인코더 학습 횟수는 60회이다. 학습은 NVIDIA RTX 6000 Ada GPU에서 수행하였으며, Python 3.12.3, PyTorch 2.8.0+cu128 및 Transformers 4.56.2를 사용하였다.

주 비교에서는 3.3절에 제시한 손실 가중치, 온도 및 마진을 고정된 실험 설정으로 적용하여 학습 구성별 성능과 오류 양상을 평가하였다. 이후 지원 필요 손실 가중치에 따른 성능 변화를 사후 민감도 분석을 통해 검토하였다.

지원 필요 손실 가중치의 민감도를 분석하기 위해 나머지 손실 가중치를 고정하고, 분리학습의 속성 인코더와 공동학습에 0.1, 0.3, 1.0을 적용하였다. 해당 분석은 사전에 적용한 가중치 0.30을 기준으로 설정값에 따른 성능 변화를 탐색적으로 평가하였다. 기존 0.3 결과와 의미 인코더를 재사용하였다. 학습자료 비교의 CE_main 15회와 계수 민감도 비교에서 두 구성의 새 계수 60회를 합쳐 총 75회 학습을 추가하였다. 모든 5인코더 ×3시드 조합을 보고하며, 분리학습의 의미 인코더는 속성 계수별로 다시 학습하지 않았다. 복합손실 공동학습은 이 계수 비교에 포함하지 않았다.

Table 6의 CE 참조값은 별도로 수행된 5개 인코더와 학습 시드 42, 1337, 2026을 조합한 15회 실행의 집계이다. CE의 맥락무해 평가 규모는 시드별 1,189건이며, 세 비교 구성에는 고정 진단 평가집합 1,190건을 사용하였다. 기존 CE 실행의 개별 예측값과 평가 사례 식별정보가 확보되지 않아 평가 사례의 동일성은 확인하지 못하였다. CE 결과는 세 학습 구성의 비교 실험과 구분하여 참조값으로 제시하였다. 추가 CE 실험에서는 주 비교와 같은 시드 42, 4337, 2026과 평가집합 1,190건을 사용하여 기존 참조 결과와 구분되는 대응 비교를 수행하였다.

5. Evaluation Metrics and Decision Thresholds

6범주 분류와 유해성 이진 판별은 각 출력층의 argmax로 결정하였다. 지원 필요 탐지는 양성 확률이 음성 확률보다 큰 경우 양성으로 판정하였다. 추가로 판정 임계값에 따른 오류 변화를 분석하기 위해 검증집합의 지원 필요 음성 1,971건을 기준으로 오탐률 상한 0.1%, 0.5%, 1.0%를 적용하였다. 인코더, 학습 구성, 시드별로 각 상한을 만족하는 최저 임계값을 선택하고, 이를 주 평가집합에 적용하여 지원 필요 누락률과 오탐률을 비교하였다. 해당 분석은 기존 학습 모델을 고정한 사후 탐색적 평가로 수행하였다.

전체 분류 성능은 주 평가집합 1,996건에서 산출하였다. 6범주 Macro-F1은 무해와 다섯 유해 세부 범주의 F1을 평균한 값이며, 유해성 이진 Macro-F1은 무해와 유해의 F1을 평균한 값이다. 유해 세부 범주 Macro-F1은 정답이 유해인 1,325건을 대상으로 다섯 유해 범주의 F1을 평균하였다. 이때 무해로 예측한 사례는 해당 유해 범주의 누락으로 처리하였다.

지원 필요 누락률(False Negative Rate, FNR)은 정답 양성 308건 중 음성으로 예측한 비율이며, 오탐률(False Positive Rate, FPR)은 정답 음성 1,688건 중 양성으로 예측한 비율로 정의하였다. 지원 필요 양성은 자기지향적 부정표현 272건과 무해·지원 필요 발화 36건으로 구분하여 부분 집합별 누락률을 산출하였다. 또한 지원 필요 탐지의 Precision, Recall 및 Macro-F1을 평가하였다. 무해 및 지원 필요 발화 36건과 지원 필요 음성 1,688건을 대상으로 불균형 자료의 평가 특성[34]을 고려하여 AUROC와 Average Precision(AP)을 함께 산출하였다.

맥락무해 오탐률은 진단 평가집합 1,190건에서 정답이 무해인 발화를 6범주 출력에서 유해 범주로 예측한 비율로 산출하였다. 기존 CE 참조 결과에서는 6범주 출력과 유해성 이진 출력의 오탐률을 각각 제시하였다. 유해 발화의 무해 예측률은 주 평가집합의 유해 발화 1,325건 중 6범주

Table 6. Classification performance and contextually benign false-positive rates of the CE baseline

Pretrained Encoder	Macro-F1			False-positive rate on contextually benign utterances	
	Six-class	Fine-grained harmfulness	Binary harmfulness	Six-class (%)	Binary harmfulness (%)
KLUE-RoBERTa	0.9028	0.9207	0.9355	44.21	44.63
KLUE-BERT	0.9011	0.9166	0.9336	57.84	62.43
KcBERT	0.9031	0.9223	0.9298	71.26	72.16
KoELECTRA	0.8919	0.9034	0.9329	54.53	57.67
KR-ELECTRA	0.9008	0.9153	0.9351	60.64	61.03
Overall mean	0.9000	0.9157	0.9334	57.70	59.59

※ Each value is averaged over three random seeds

www.kci.go.kr

출력에서 무해로 예측한 비율로 정의하였다. 인코더별 결과는 세 학습 시드의 평균과 표본 표준편차로 제시하였으며, 학습 구성별 평균은 다섯 인코더와 세 시드로 구성된 15개 실행에 동일한 가중치를 적용하여 산출하였다.

무해 및 지원 필요 발화 36건의 누락률과 학습 구성 간 차이에 대한 불확실성은 동일 발화 ID를 복원추출하는 50,000회 대응 부트스트랩으로 평가하였다. 학습 구성별 15개 실행을 고정하고 발화 단위로 재표집하였으며, 재표집 분포의 2.5 및 97.5 백분위수로 95% 구간을 산출하였다. 신경망 모델 비교에서 실행 간 성능 변동을 고려할 필요성[35]에 따라, 인코더, 시드, 발화의 변동을 반영하는 탐색적 재표집 분석을 추가하였다. 본 분석에서는 인코더와 학습 시드를 계층적으로 재표집하고, 복원추출한 발화 ID를 모든 학습 구성에 공통으로 적용하는 대응 재표집을 20,000회 수행하였다.

결과의 민감도를 검토하기 위해 발화, 인코더 및 학습 시드를 각각 하나씩 제외하는 사후 분석을 수행하였다. 또한 무해·지원 필요 발화 36건 중 최종 의미 범주와 전문가 최다득표 범주가 다른 10건을 제외한 26건에서 누락률을 재산출하였다. 학습자료·속성 감독정보 및 지원 손실 가중치 비교에서도 동일한 기본 판정 규칙을 적용하였으며, 해당 부분집합의 발화 ID를 공통으로 재표집하는 50,000회 대응 부트스트랩으로 오류율 차이의 95% 구간을 산출하였다. 이러한 불확실성 분석은 고정된 학습 실행과 평가 자료를 바탕으로 한 탐색적 분석으로 해석하였다.

마지막으로 의미 범주 예측만으로 지원 필요 여부를 결정하는 범주 기반 규칙을 구성하였다. 6범주 출력에서 자기지향적 부정표현으로 예측한 경우 지원 필요 양성으로 판정하고, 그 외 범주는 음성으로 처리하였다. 이 규칙과 별도의 지원 필요 출력층을 동일한 평가집합에서 비교하여 무해·지원 필요 발화의 탐지 성능을 평가하였다.

IV. Results

1. Cross-Entropy Baseline Performance

CE 기준선의 평균 6범주 Macro-F1은 0.9000, 유해 세부 유형 Macro-F1은 0.9157, 유해성 이진 Macro-F1은 0.9334로 나타났다(Table 6). 맥락무해 발화의 평균 오탐률은 6범주 출력에서 57.70%, 이진 출력에서 59.59%였다. 인코더별 6범주 Macro-F1은 0.8919–0.9031, 맥락무해 오탐률은 44.21–71.26%의 범위로 나타났다.

KcBERT는 6범주 Macro-F1이 0.9031로 가장 높았으나, 맥락무해 오탐률 역시 71.26%로 가장 높았다. 반면 KLUE-RoBERTa는 유사한 분류 성능(0.9028)을 보이면서 가장 낮은 맥락무해 오탐률(44.21%)을 기록하였다. 이진

Table 7. Classification performance by pretrained encoder and learning configuration

Pretrained Encoder	Learning configuration	Macro-F1		
		Six-class	Fine-grained harmfulness	Binary harmfulness
KLUE-RoBERTa	Separate learning	0.9040 ± 0.0019	0.9208 ± 0.0009	0.9341 ± 0.0011
	Joint learning	0.9043 ± 0.0019	0.9214 ± 0.0020	0.9339 ± 0.0035
	Joint learning with combined losses	0.9066 ± 0.0033	0.9241 ± 0.0026	0.9362 ± 0.0020
KLUE-BERT	Separate learning	0.9023 ± 0.0011	0.9168 ± 0.0011	0.9360 ± 0.0017
	Joint learning	0.9016 ± 0.0017	0.9163 ± 0.0028	0.9343 ± 0.0041
	Joint learning with combined losses	0.9027 ± 0.0027	0.9197 ± 0.0011	0.9356 ± 0.0026
KcBERT	Separate learning	**0.9069 ± 0.0038**	**0.9261 ± 0.0019**	0.9339 ± 0.0013
	Joint learning	0.9067 ± 0.0018	0.9254 ± 0.0022	0.9339 ± 0.0020
	Joint learning with combined losses	0.9057 ± 0.0005	0.9253 ± 0.0008	0.9322 ± 0.0013
KoELECTRA	Separate learning	0.8863 ± 0.0029	0.9016 ± 0.0022	0.9326 ± 0.0026
	Joint learning	0.8885 ± 0.0015	0.9036 ± 0.0005	0.9333 ± 0.0029
	Joint learning with combined losses	0.8909 ± 0.0054	0.9069 ± 0.0043	0.9335 ± 0.0019
KR-ELECTRA	Separate learning	0.9055 ± 0.0020	0.9196 ± 0.0028	**0.9387 ± 0.0024**
	Joint learning	0.9030 ± 0.0045	0.9160 ± 0.0053	0.9378 ± 0.0020
	Joint learning with combined losses	0.9036 ± 0.0036	0.9192 ± 0.0042	0.9383 ± 0.0046

www.kci.go.kr

출력에서도 두 인코더의 맥락무해 오탐률은 각각 72.16%와 44.63%로 유사한 경향을 보였다.

2. Classification and Support-Need Detection

분리학습, 공동학습 및 복합손실 공동학습의 평균 6범주 Macro-F1은 각각 0.9010, 0.9008, 0.9019로 나타났다(Table 7). 평균 유해성 이진 Macro-F1은 각각 0.9350, 0.9346, 0.9352였다. 복합손실 공동학습의 6범주 Macro-F1은 공동학습보다 평균 0.0011 높았으며, KcBERT를 제외한 네 인코더에서 성능이 향상되었다.

주 평가집합 1,996건에서 지원 필요 탐지의 평균 Precision, Recall 및 Macro-F1은 분리학습에서 각각 0.9965, 0.9922, 0.9967, 공동학습에서 0.9894, 0.9485, 0.9811, 복합손실 공동학습에서 0.9892, 0.9452, 0.980이었다. 각 학습 구성의 지원 필요 탐지 성능은 다섯 인코더와 세 시드로 구성된 15개 실행의 평균으로 산출하였다.

학습 구성별 안전 관련 오류는 Table 8과 같다. 평균 지원 필요 누락률은 분리학습 0.78%, 공동학습 5.15%, 복합손실 공동학습 5.48%였다. 공동학습의 누락률은 분리학습보다 평균 4.37%p 높았으며, 다섯 인코더 모두에서 같은 방향의 차이가 나타났다. 평균 지원 필요 오탐률은 각각 0.06%, 0.19%, 0.19%로, 분리학습은 평균 누락률과 오탐률이 모두 낮았다.

지원 필요 누락률의 차이는 인코더에 따라 크기가 달랐다. 분리학습의 누락률은 모든 인코더에서 1% 미만이었으나, 공동학습에서는 KLUE-RoBERTa, KLUE-BERT 및 KcBERT가 1.30–2.27%, KoELECTRA와 KR-ELECTRA가 각각 11.80%와 8.77%로 나타났다. 공동학습의 누락률 표준편차는 KoELECTRA에서 0.37%p, KR-ELECTRA에서 4.22%p였다.

공동학습에 두 표현 손실을 추가한 경우 평균 지원 필요 누락률은 약 0.3%p 증가하였다. 인코더별로는 KLUE-RoBERTa, KLUE-BERT 및 KR-ELECTRA에서 증가하였으며, KcBERT와 KoELECTRA에서는 동일하였다. 복합손실 공동학습은 공동학습보다 평균 6범주 Macro-F1이 높았으나, 지원 필요 누락률의 감소는 관측되지 않았다.

3. Harmfulness Detection Errors

평균 맥락무해 오탐률은 분리학습 1.33%, 공동학습 1.40%, 복합손실 공동학습 1.35%로 나타났다. 공동학습에 두 표현 손실을 추가한 경우 평균 오탐률은 0.05%p 감소하였다. 인코더별로는 KLUE-RoBERTa, KcBERT 및 KoELECTRA에서 감소하였으나, KLUE-BERT에서는 증가하고 KR-ELECTRA에서는 동일하였다. 분리학습에서는 KcBERT의 맥락무해 오탐률이 1.71%로 가장 높았으며, 공동학습과 복합손실 공동학습에서는 KoELECTRA가 각

Table 8. Safety-related errors by pretrained encoder and learning configuration

Pretrained Encoder	Learning configuration	Support-need FNR (%)	Support-need FPR (%)	Contextually benign FPR (%)	Harmful-to-benign misclassification rate (%)
KLUE-RoBERTa	Separate learning	0.87 ± 0.19	0.06 ± 0.00	1.09 ± 0.08	5.08 ± 0.04
	Joint learning	2.27 ± 0.65	0.28 ± 0.03	1.06 ± 0.26	5.38 ± 0.27
	Joint learning with combined losses	3.25 ± 0.56	0.24 ± 0.10	**1.04 ± 0.21**	4.75 ± 0.27
KLUE-BERT	Separate learning	0.87 ± 0.37	0.08 ± 0.03	1.20 ± 0.17	5.08 ± 0.34
	Joint learning	1.30 ± 0.00	0.28 ± 0.03	1.32 ± 0.26	5.31 ± 0.09
	Joint learning with combined losses	1.52 ± 0.37	0.30 ± 0.00	1.34 ± 0.30	4.91 ± 0.13
KcBERT	Separate learning	0.87 ± 0.19	0.02 ± 0.03	1.71 ± 0.21	4.96 ± 0.17
	Joint learning	1.62 ± 0.32	0.24 ± 0.06	1.65 ± 0.13	4.91 ± 0.26
	Joint learning with combined losses	1.62 ± 0.32	0.30 ± 0.06	1.54 ± 0.24	4.78 ± 0.21
KoELECTRA	Separate learning	0.87 ± 0.19	0.10 ± 0.12	1.48 ± 0.32	5.53 ± 0.68
	Joint learning	11.80 ± 0.37	0.06 ± 0.06	1.76 ± 0.39	5.23 ± 0.43
	Joint learning with combined losses	11.80 ± 0.37	0.04 ± 0.03	1.62 ± 0.35	4.83 ± 0.49
KR-ELECTRA	Separate learning	**0.43 ± 0.19**	0.06 ± 0.06	1.15 ± 0.10	4.73 ± 0.29
	Joint learning	8.77 ± 4.22	0.10 ± 0.07	1.20 ± 0.13	4.96 ± 0.17
	Joint learning with combined losses	9.20 ± 4.03	**0.10 ± 0.06**	1.20 ± 0.13	**4.60 ± 0.27**

www.kci.go.kr

각 1.76%와 1.62%로 가장 높았다. KLUE-RoBERTa는 세 학습 구성 모두에서 가장 낮은 오탐률을 보였다.

학습자료와 속성 감독의 차이를 살펴본 동일 평가집합 비교에서 CE_main, CE_aug 및 공동학습의 평균 맥락무해 오탐률은 각각 58.08%, 1.33%, 1.40%였다. CE_aug의 오탐률은 CE_main보다 평균 56.75%p 낮았으며, 차이의 95% 대응 구간은 [-58.83, -54.66]%p로 나타났다.

유해 발화의 무해 예측률은 분리학습 5.08%, 공동학습 5.16%, 복합손실 공동학습 4.77%로 나타났다. 복합손실 공동학습은 다섯 인코더 모두에서 공동학습보다 낮은 유해 발화의 무해 예측률을 보였다. KLUE-RoBERTa에서는 5.38%에서 4.75%로, KLUE-BERT에서는 5.31%에서 4.91%로 감소하였다. KcBERT, KoELECTRA 및 KR-ELECTRA에서도 각각 4.91%에서 4.78%, 5.23%에서 4.83%, 4.96%에서 4.60%로 낮아졌다. 복합손실 공동학습의 유해 발화 무해 예측률은 분리학습보다 모든 인코더에서 낮게 나타났다. 인코더와 학습 구성별 오류는 Fig. 1.에 제시하였다.

4. Support-Need Detection by Semantic Category

지원 필요 양성을 의미 범주에 따라 구분한 결과, 학습 구성 간 누락률의 차이는 두 부분집합에서 다르게 나타났다. 자기지향적 부정표현의 평균 지원 필요 누락률은 분리학습 0.59%, 공동학습 0.39%, 복합손실 공동학습 0.39%였다. 반면 무해·지원 필요 발화의 평균 누락률은 각각 2.22%, 41.11%, 43.89%로 나타났다(Fig. 2). 공동학습과 복합손실 공동학습의 누락률은 분리학습과 비교하여 자기지향적 부정표현에서는 낮았으나, 무해·지원 필요 발화에서는 높게 나타났다.

무해 및 지원 필요 발화의 누락률은 다섯 인코더 모두에서 분리학습이 가장 낮았다. KoELECTRA에서는 분리학습 1.85%에 비해 공동학습과 복합손실 공동학습 모두 97.22%

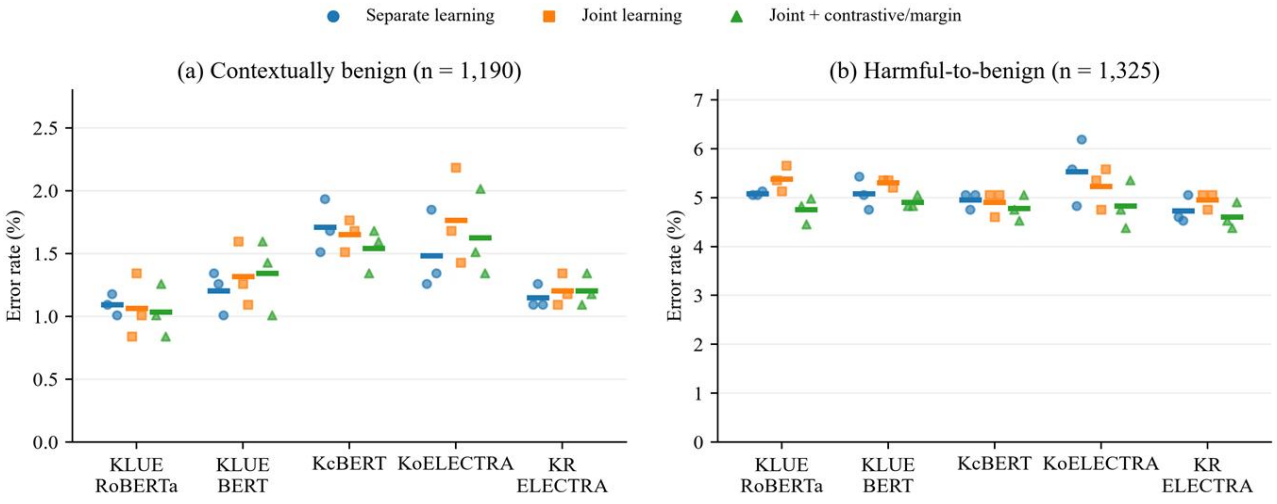

Fig. 1. Encoder-wise errors in (a) contextually benign utterances and (b) harmful-to-benign predictions

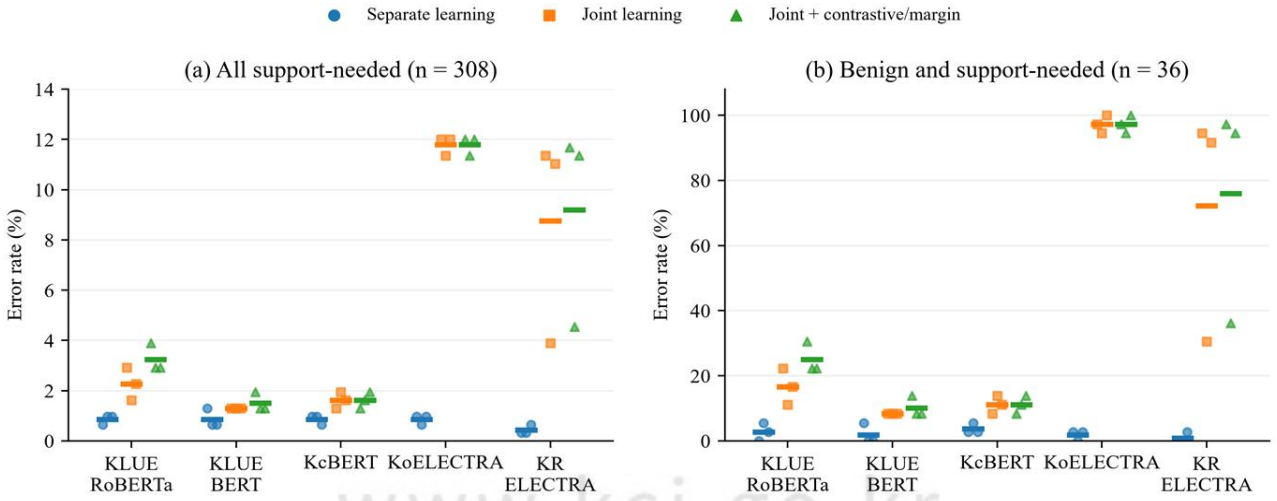

Fig. 2. Missed support needs in (a) all positive instances and (b) benign positive instances

www.kci.go.kr

로 나타났다. KR-ELECTRA에서도 분리학습은 0.93%였으나, 공동학습과 복합손실 공동학습은 각각 72.22%, 75.93%였다. 반면 KLUE-RoBERTa, KLUE-BERT 및 KcBERT의 공동학습 누락률은 각각 16.67%, 8.33%, 11.11%로 상대적으로 낮았다. KR-ELECTRA의 공동학습과 복합손실 공동학습에서는 시드별 누락률의 변동이 크게 나타났다.

인코더, 시드, 발화를 재표집한 보조 분석에서 분리학습 대비 무해·지원 필요 누락률 차이는 공동학습 38.89%p(95% 구간 [6.85, 74.81]), 복합손실 공동학습 41.67%p(95% 구간 [9.63, 76.48])로 나타났다. 반면 복합손실 공동학습과 공동학습 간 차이는 2.78%p(95% 구간 [-0.19, 7.22])로, 구간에 0이 포함되었다. 따라서 두 공동학습 구성 간 차이는 평균값과 실행 간 변동을 함께 고려하여 해석할 필요가 있다.

발화·인코더와 시드를 각각 하나씩 제외한 민감도 분석에서도 공동학습과 복합손실 공동학습의 분리학습 대비 누락률 차이는 모두 양의 값을 유지하였다. 또한 최종 분석 라벨과 전문가 최다득표 범주가 다른 10건을 제외한 26건에서도 누락률은 분리학습 3.08%, 공동학습 44.36%, 복합손실 공동학습 47.69%로 나타났다. 이 사후 민감도 분석에서는 최종 분석 라벨과 전문가 최다득표 범주가 다른 10건을 제외한 후에도 학습 구성 간 누락률 차이의 방향이 유지되었다.

5. Sensitivity to Support-Loss Weight and Decision Threshold

지원 필요 손실 가중치 0.1, 0.3, 1.0에 따른 탐지 성능을 비교하였다(Table 9). 공동학습의 무해·지원 필요 발화 누락률은 각각 83.33%, 41.11%, 3.89%였으며, 분리학습은 1.67%, 2.22%, 1.67%로 나타났다. 공동학습의 가중치를 0.3에서 1.0으로 변경한 경우 누락률은 37.22%p 감소하였으며, 대응 사례 부트스트랩의 95% 구간은 [-41.48, -33.52]%p였다. 가중치 1.0에서 공동학습과 분리학습의 누락률 차이는 2.22%p로 축소되었으며, 인코더·시드·발화를 함께 재표집한 95% 구간은 [-1.48, 7.59]%p였다. 이러한 결과는 학습 구성 간 누락률 차이가 지원 필요 손실 가중치에 따라 달라짐을 보여준다.

학습 가중치와 판정 임계값의 영향을 구분하기 위해 기존 가중치 0.3에서 학습된 모델을 대상으로 임계값 조정 분석을 수행하였다(Table 10). 검증집합의 지원 필요 오탐률 상한을 0.5%로 설정한 결과, KoELECTRA의 공동학습과 복합손실 공동학습에서 무해·지원 필요 누락률은 모두 97.22%에서 92.59%로 감소하였다. KR-ELECTRA에서는 공동학습이 72.22%에서 39.81%로, 복합손실 공동학습이 75.93%에서 50.00%로 감소하였다. 네 조건 모두에서 누락률 감소와 함께 지원 필요 오탐률이 증가하였다.

지원 필요 손실 가중치에 따른 순위 판별력은 AUROC와 AP로 평가하였다(Table 11). 공동학습의 가중치 0.1, 0.3, 1.0에서 AUROC는 각각 0.8429, 0.9838, 0.9993이었으며, AP는 0.4482, 0.7619, 0.9674였다. 이는 지원 필요 손실 가중치에 따라 기본 판정 기준의 누락률뿐 아니라 양성과 음성의 점수 순위도 변화하였음을 보여준다.

Table 9. Sensitivity to the support-loss weight at the fixed decision rule

Config.	Weight	Benign support FNR(%)	All support FNR(%)	Support FPR(%)
I	0.1	1.67	0.74	0.10
	0.3	2.22	0.78	0.06
	1	1.67	0.74	0.07
J	0.1	83.33	10.17	0.08
	0.3	41.11	5.15	0.19
	1	3.89	0.87	0.16

※ I: Separate learning, J: Joint learning

Table 11. Ranking performance for benign support positive utterances

Config.	Weight	AUROC	AP
I	0.1	0.9997	0.9949
	0.3	0.9998	0.9962
	1	0.9997	0.9953
J	0.1	0.8429	0.4482
	0.3	0.9838	0.7619
	1	0.9993	0.9674
M	0.3	0.9764	0.7339

※ I: Separate learning, J: Joint learning, M: Joint learning with combined losses

Table 10. Exploratory comparison of support-need detection errors before and after threshold adjustment for ELECTRA encoders under joint learning configurations at support-loss weight 0.3

Pretrained Encoder	Learning Configuration	Default FNR36(%)	Adjusted FNR36(%)	Default FPR(%)	Adjusted FPR(%)
KoELECTRA	Joint Learning	97.22	92.59	0.06	0.24
	Joint learning with combined losses	97.22	92.59	0.04	0.32
KR-ELECTRA	Joint Learning	72.22	39.81	0.10	0.22
	Joint learning with combined losses	75.93	50.00	0.10	0.20

www.kci.go.kr

6. Comparison of Category-Based Rules and the Support Head

의미 범주 예측을 지원 필요 판정으로 변환하는 범주 기반 규칙과 지원 필요 출력층의 성능을 비교하였다(Table 12). 범주 기반 규칙은 6범주 출력에서 자기지향적 부정표현으로 예측한 경우를 지원 필요 양성으로 판정하였다. 범주 기반 규칙의 무해 및 지원 필요 발화 누락률은 세 학습 구성 모두 평균 99.81%로 나타났다. 반면 지원 필요 손실 가중치 0.3에서 지원 필요 출력층의 누락률은 분리학습 2.22%, 공동학습 41.11%, 복합손실 공동학습 43.89%였다.

범주 기반 규칙은 현재 정책 라벨에서 무해·지원 필요 발화의 대부분을 누락하였다. 다만 두 판정 방식은 지원 필요 속성에 관한 감독정보가 다르므로, 본 비교는 현재 정책 라벨을 기준으로 한 보조 분석으로 해석하였다.

Table 12. Category rule and support head at support-loss weight 0.3

Config.	Decision	FNR36 (%)	All FNR (%)	FPR (%)
I	Category Rule	99.81	12.10	0.063
	Support Head	2.22	0.78	0.063
J	Category Rule	99.81	12.08	0.067
	Support Head	41.11	5.15	0.190
M	Category Rule	99.81	12.06	0.067
	Support Head	43.89	5.48	0.194

※ I: Separate learning, J: Joint learning, M: Joint learning with combined losses

V. Discussion and Conclusions

본 연구는 유아용 인공지능 동반자의 입력 발화 안전 가드레일을 위해 의미 범주와 안전 속성을 구분한 라벨 체계를 구성하고, 학습 구성에 따른 탐지 성능과 오류 특성을 실증적으로 분석하였다. 기존 유해언어 탐지 연구에서는 문맥 의존성과 특정 언어적 단서에 따른 오분류 문제가 논의되어 왔으며[18-20], 하위집합 및 기능별 오류 평가의 중요성이 제시되었다[22-23]. 그러나 유아 발화의 맥락적 특성을 반영한 유해성 판단과 지원 필요 탐지를 함께 평가한 실증적 연구는 제한적이었다. 이에 본 연구에서는 27,605건의 유아 발화에 의미 범주와 안전 속성을 적용하고, 전문가 5인의 독립적인 라벨링 결과를 바탕으로 의미 범주에 대한 전문가 간 일치도와 최종 분석 라벨의 정합성을 구분하여 분석하였다. 또한 분리학습, 공동학습 및 복합손실 공동학습을 비교하여 맥락무해 발화의 오분류와 무해 및 지원 필요 발화의 탐지 오류를 정량적으로 분석하였다.

세 학습 구성의 전체 분류 성능은 유사하였으나, 의미 범주에 따른 지원 필요 탐지 성능은 차이를 보였다. 자기지향적 부정표현의 평균 누락률은 세 학습 구성 모두 1% 미만이었으나, 무해·지원 필요 발화의 평균 누락률은 분리학습 2.22%, 공동학습 41.11%, 복합손실 공동학습 43.89%로 나타났다. 전체 지원 필요 양성 308건 중 272건이 자기지향적 부정표현에 해당하므로, 전체 지원 필요 누락률에는 무해·지원 필요 발화의 식별 실패가 상대적으로 작게 반영된다. 본 연구에서는 기존 연구에서 제시한 하위집합 평가 관점을 유아 발화의 안전 판단에 적용하여, 전체 성능에 가려질 수 있는 의미 범주와 안전 속성의 교차 조건별 탐지 오류를 정량적으로 확인하였다. 이는 안전 기능의 성능을 평가할 때 전체 성능과 함께 지원 필요 발화의 의미 범주별 누락률을 주요 평가 지표로 활용할 수 있음을 시사한다.

본 연구에서는 정서지지 응답 대상으로 지정한 의미 범주와 세부 유형을 기준으로 지원 필요 속성을 정의하고, 유해성 분류와 구분하여 평가하였다. 의미 범주 예측을 지원 필요 판정으로 변환한 범주 기반 규칙에서는 무해 및 지원 필요 발화의 평균 누락률이 세 학습 구성 모두 99.81%로 나타난 반면, 지원 필요 출력층에서는 2.22-43.89%로 나타났다. 두 판정 방식은 지원 필요 속성에 관한 감독정보가 다르지만, 현재 정책 라벨에서 의미 범주 기반 판정과 지원 필요 속성의 직접 예측은 서로 다른 탐지 특성을 보였다. 이러한 결과는 유아용 인공지능 동반자의 안전 판단에서 의미 범주 분류와 지원 필요 탐지를 구분된 과제로 구성하고, 각 판단의 성능을 독립적으로 검토하는 설계 방향을 뒷받침한다.

기존 CE 결과에서 평균 6범주 Macro-F1은 0.9000이었으나, 맥락무해 발화의 평균 오탐률은 57.70%로 나타났다. 동일 평가집합에서는 CE_main의 평균 맥락무해 오탐률이 58.08%였으나, 맥락무해 학습자료를 포함한 CE_aug에서는 1.33%로 감소하였다. 이 비교에서는 맥락무해 학습자료를 포함한 조건에서 해당 발화의 오탐률이 낮게 나타났다. 다만 동일한 3에포크 학습에서 CE_main과 CE_aug의 갱신 횟수가 각각 654회와 693회였으므로, 관찰된 차이에는 학습자료 구성과 갱신 횟수의 영향이 함께 포함될 수 있다. 이러한 결과는 유아 발화의 안전 판단 모델을 학습할 때 유해 발화와 표면적 특징을 공유하는 무해 발화를 학습자료에 포함하는 방안을 뒷받침한다. 또한 맥락무해 발화에 대한 별도의 진단 평가를 통해 전체 분류 성능만으로는 드러나기 어려운 해당 발화의 오탐 양상을

www.kci.go.kr

분석하였다.

다중과제학습은 관련 과제 간 공유 표현을 활용하는 학습 접근으로 제안되어 왔다[25-26]. 본 연구에서는 의미 범주와 안전 속성을 공동으로 학습하는 구성을 평가한 결과, 지원 필요 탐지 성능이 학습 구성과 손실 가중치에 따라 다르게 나타났다. 공동학습의 지원 필요 손실 가중치를 0.3에서 1.0으로 변경한 경우 무해 및 지원 필요 발화의 누락률은 41.11%에서 3.89%로 감소하였으며, AUROC는 0.9838에서 0.9993으로, AP는 0.7619에서 0.9674로 증가하였다. 가중치 1.0에서 공동학습과 분리학습의 누락률 차이도 2.22%p로 축소되었다. 이러한 결과는 안전 속성을 별도의 학습목표로 구성하는 것에 더하여, 해당 과제의 손실 가중치가 특정 발화 조건의 탐지 성능과 관련된 주요 학습 설정임을 보여준다. 따라서 공동학습을 적용하는 경우 안전 기능별 손실 가중치와 판정 기준을 함께 검토하고, 누락률과 오탐률 및 순위 판별력을 종합적으로 평가하는 접근이 적절하다.

공동학습에서 무해 및 지원 필요 발화의 누락률은 인코더에 따라 차이를 보였으며, 특히 KoELECTRA와 KR-ELECTRA에서 높게 나타났다. ELECTRA는 마스킹된 언어모델링(Masked Language Modeling, MLM)과 달리 입력 토큰의 대체 여부를 판별하는 Replaced Token Detection(RTD) 방식으로 사전학습된다[36]. 이러한 사전학습 목표의 차이는 인코더별 표현 학습 특성과 관련될 가능성이 있으며, 공동학습에서 관찰된 성능 차이를 설명하기 위한 기술적 가설로 고려할 수 있다. 다만 인코더별 성능 차이는 사전학습 목표뿐 아니라 모델의 구조 및 학습 특성과도 관련될 수 있으므로, 향후 표현 구조와 과제 간 학습 간섭에 대한 추가 분석을 통해 해당 관계를 검토할 필요가 있다.

분리학습과 공동학습에는 동일한 학습자료와 감독정보를 적용하였으나, 인코더 구성과 최적화 방식에는 차이가 있다. 분리학습은 두 개의 독립적인 인코더를 사용하며, 공동학습은 하나의 공유 인코더를 사용한다. 이에 따라 학습 구성 간 성능 차이에는 표현 공유 여부와 함께 모델 용량 및 최적화 독립성의 차이가 포함된다. 지원 필요 손실 가중치에 따른 민감도 분석에서는 학습 구성 간 무해 및 지원 필요 발화의 누락률 차이가 크게 달라졌다. 이는 지원 필요 탐지 성능이 인코더 구성뿐 아니라 손실 가중치와도 관련되어 있음을 보여준다. 따라서 학습 구성과 손실 가중치를 함께 고려하여 의미 범주별 지원 필요 탐지 성능을 평가할 필요가 있다.

복합손실 공동학습에서는 지도대조손실과 마진손실의 추가로 유해 발화의 무해 예측률이 공동학습 5.16%에서 4.77%로 감소하였다. 반면 무해 및 지원 필요 발화의 누락률은 공동학습 41.11%, 복합손실 공동학습 43.89%로 나타났다. 표현 손실 추가에 따른 성능 변화는 안전 기능별로 다르게 나타났으며, 이는 학습 손실이 직접 대상으로 하는 판단과 실제 개선되는 오류 유형을 대응시켜 검토하는 것이 중요함을 보여준다. 따라서 안전 가드레일의 학습 목표를 구성할 때에는 개별 안전 기능의 목적에 부합하는 손실 설계와 조건별 성능 평가를 함께 고려해야 한다.

본 연구의 학술적 기여는 세 가지로 정리된다. 첫째, 유아 발화의 의미 범주와 안전 속성을 구분한 라벨 체계를 구성하고, 27,605건의 발화에 적용하여 전문가 라벨링 및 검증을 수행하였다. 전문가 간 의미 범주 일치도와 최종 분석 라벨의 정합성을 각각 분석하여 유아 발화의 안전 판단을 위한 평가자료의 특성을 확인하였다. 둘째, 맥락무해 발화와 무해 및 지원 필요 발화를 대상으로 조건별 오류를 정량화하여 전체 분류 성능과 안전 속성별 평균 성능에서 드러나지 않는 탐지 실패를 확인하였다. 셋째, 세 학습 구성의 성능을 비교하고 지원 필요 손실 가중치, 판정 임계값 및 순위 판별력에 대한 추가 분석을 수행하여 학습 구성과 설정 조건에 따른 안전 기능별 성능 변화의 특성을 제시하였다.

본 연구는 다음과 같은 한계가 있다. 첫째, 본 연구는 발화 단위 분할과 개별 발화 텍스트를 기반으로 의미 범주 및 안전 속성의 탐지 성능을 평가하였다. 동일한 텍스트의 분할 간 중복을 방지하고 고정된 평가집합에서 학습 구성별 성능을 비교하였다. 이러한 평가 설계는 발화 수준의 판단 성능에 초점을 두며, 화자 독립 평가 및 대화 맥락을 포함한 판단과는 구분된다. 향후에는 화자 독립 분할과 대화 이력 및 상호작용 상황을 포함한 평가로 확장할 필요가 있다. 이를 통해 후속 응답의 적절성과 유아와의 상호작용 효과를 검증할 수 있다.

둘째, 학습 구성 간 성능 차이를 해석하는 데 한계가 있다. 분리학습은 두 개의 독립적인 인코더를 사용하고, 공동학습은 하나의 공유 인코더를 사용한다. 이에 따라 관찰된 성능 차이에는 표현 공유 여부뿐 아니라 모델 용량과 최적화 방식의 영향이 함께 포함된다. 지원 필요 손실 가중치에 따른 민감도 분석에서도 학습 구성 간 누락률 차이가 달라졌다. 따라서 본 연구의 결과는 적용한 모델 구조와 학습 설정을 고려하여 해석해야 한다. 향후에는 모델 용량을 통제한 비교와 손실 가중치별 분석을 확대하여 각 요인이 탐지 성능에 미치는 영향을 구체적으로 검토할 필요가 있다.

www.kci.go.kr

셋째, 지원 필요 속성의 평가 표본 규모와 라벨 검증 범위에 한계가 있다. 본 연구는 수업 및 놀이 상황에서의 개별면담과 사회정서적 상황을 활용한 일대일 대화 및 상황 발문을 통해 유아 발화를 수집하였다. 이 과정에서 확보된 무해·지원 필요 평가 발화는 36건으로 구성되었다. 지원 필요 속성은 의미 범주와 세부 유형을 기준으로 정서지지 응답 대상을 지정한 정책 라벨이다. 본 연구의 전문가 평가는 의미 범주의 평정 일관성과 최종 분석 라벨의 정합성을 검토한 것으로, 실제 상호작용에서 정서지지 응답이 필요한지를 판단하는 정책 기준의 타당성 검증과 구분된다. 무해 및 지원 필요 발화에 대한 재표집 및 민감도 분석에서는 전문가 최다득표 범주와 차이가 있는 10건을 제외한 후에도 학습 구성 간 누락률 차이의 방향이 유지되었다. 향후에는 다양한 상호작용 상황에서 발화자료를 추가로 확보하고, 데이터 증강으로 생성한 발화의 의미와 라벨 정합성을 검토하여 표현의 다양성을 확대할 계획이다. 일반화 평가에서는 평가 목표와 학습자료 및 평가자료 간 분포 차이를 명확히 설정할 필요가 있다는 논의[37]를 바탕으로, 새로운 화자와 상호작용 상황을 포함하는 독립적인 평가자료에서 지원 필요 탐지 성능을 검토하고자 한다. 아울러 상황 맥락을 고려한 전문가 평가를 통해 정서지지 응답 대상에 관한 정책 기준의 타당성을 검토하고자 한다.

본 연구는 유아용 인공지능 동반자의 안전 가드레일에서 유해성 분류, 맥락무해 발화의 구별 및 지원 필요 탐지를 구분된 판단 과제로 구성하고, 의미 범주와 안전 속성의 교차 조건에서 각 판단의 탐지 성능을 실증적으로 분석하였다. 지원 필요성을 별도의 학습목표로 구성한 조건에서 무해·지원 필요 발화의 누락률은 학습 구성에 따라 차이를 보였으며, 표현 손실 추가에 따른 성능 변화도 유해성 판별과 지원 필요 탐지에서 서로 다르게 나타났다. 이러한 결과를 바탕으로 안전 가드레일의 설계와 평가에서는 의미 범주에 따른 지원 필요 발화의 식별 성능과 학습목표별 안전 관련 오류의 변화를 함께 고려해야 한다. 본 연구는 유아 발화의 의미 범주와 지원 필요 속성을 구분한 판단 구조를 실제 탐지 성능의 검증으로 연결하고, 후속 대응 기능과 연계할 수 있는 입력 발화 판단 단계의 설계 및 평가를 위한 실증적 근거를 제공한다.

INSTITUTIONAL REVIEW BOARD STATEMENT

본 연구에 사용된 음성 발화자료의 수집은 서울대학교 생명윤리위원회의 승인을 받아 수집함(승인번호: 2309/003-020, 최초 승인일: 2023년 9월 20일, 변경 승인일: 2023년 10월 10일).

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) in 2026 (Grant No. RS-2023-00251378, Project Title: Context-centered Artificial Intelligence Companion for Young Children: Development and Demonstration of an Artificial Intelligence Model that Talks Like a Friend to Support Social-emotional Development of Young Children, Contribution: 50%; Grant No. NRF-2026S1A5A8010250, Project Title: Impact of AI Companion's 'Pedagogical Rephrasing' Guardrails on Preschoolers' Moral Habitus and Language Socialization: A Study on the Humanities Utility of Technical Safety Nets, Contribution: 50%).

REFERENCES

[1] Y.K. Choi, J.H. Lee, H.Y. Park, B.R. Lee, and H.S. Kim, Developing a Future Agenda for Child-Centered AI Technology, Research Report 2025-13, Korea Institute of Child Care and Education, pp. 1-358, 2025.

[2] R. Garg, H. Cui, S. Seligson, B. Zhang, M. Porcheron, L. Clark, et al., "The Last Decade of HCI Research on Children and Voice-based Conversational Agents," Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pp. 1-19, 2022, DOI: 10.1145/3491102.3502016.

[3] S.J. Chang, S. Han, and K.C. Park, "Social Features of Artificial Intelligence Companions for Children," Journal of Digital Contents Society, Vol. 24, No. 12, pp. 3199-3211, 2023, DOI: 10.9728/dcs.2023.24.12.3199.

[4] K.S. Jeong, J. Yang, and K.C. Park, "Child-Centered AI: An Ethics Framework and Future Directions," Journal of Digital Contents

www.kci.go.kr

Society, Vol. 25, No. 8, pp. 2225-2237, 2024, DOI: 10.9728/dcs.2024.25.8.2225.

[5] UNICEF, Policy Guidance on AI for Children, ver. 2.0, UNICEF Office of Global Insight and Policy, 2021.

[6] A.K. Marcelo and T.M. Yates, "Prospective Relations of Preschoolers' Prosocial and Aggressive Affect Themes in Pretend Play with Prosocial and Aggressive Behaviors across Contexts," Social Development, Vol. 29, No. 4, pp. 995-1013, 2020, DOI: 10.1111/sode.12470.

[7] N.R. Crick, J.F. Casas, and M. Mosher, "Relational and Overt Aggression in Preschool," Developmental Psychology, Vol. 33, No. 4, pp. 579-588, 1997, DOI: 10.1037/0012-1649.33.4.579.

[8] L.G. Bailes, G. Ennis, S.M. Lempres, D.A. Cole, and K.L. Humphreys, "Parents' Emotion Socialization Behaviors in Response to Preschool-Aged Children's Justified and Unjustified Negative Emotions," PLOS ONE, Vol. 18, No. 4, Article e0283689, 2023, DOI: 10.1371/journal.pone.0283689.

[9] K.M. Zinsser, R.A. Gordon, and X. Jiang, "Parents' Socialization of Preschool-Aged Children's Emotion Skills: A Meta-Analysis Using an Emotion-Focused Parenting Practices Framework," Early Childhood Research Quarterly, Vol. 55, pp. 377-390, 2021, DOI: 10.1016/j.ecresq.2021.02.001.

[10] P. Fortuna and S. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," ACM Computing Surveys, Vol. 51, No. 4, Article 85, 2018, DOI: 10.1145/3232676.

[11] J. Moon, W.I. Cho, and J. Lee, "BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection," Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media, pp. 25-31, 2020, DOI: 10.18653/v1/2020.socialnlp-1.4.

[12] J. Lee, T. Lim, H. Lee, B. Jo, Y. Kim, H. Yoon, et al., "K-MHaS: A Multi-Label Hate Speech Detection Dataset in Korean Online News Comment," Proceedings of the 29th International Conference on Computational Linguistics, pp. 3530-3538, 2022.

[13] Y. Xu, "Talking with Machines: Can Conversational Technologies Serve as Children's Social Partners?," Child Development Perspectives, Vol. 17, No. 1, pp. 53-58, 2023, DOI: 10.1111/cdep.12475.

[14] C.L. van Straten, J. Peter, and R. Kühne, "Child-Robot Relationship Formation: A Narrative Review of Empirical Research," International Journal of Social Robotics, Vol. 12, No. 2, pp. 325-344, 2020, DOI: 10.1007/s12369-019-00569-0.

[15] S. Druga, R. Williams, C. Breazeal, and M. Resnick, "Hey Google Is It OK if I Eat You? Initial Explorations in Child-Agent Interaction," Proceedings of the 2017 Conference on Interaction Design and Children, pp. 595-600, 2017, DOI: 10.1145/3078072.3084330.

[16] J. Choi and K. Jung, "Embodied Conversational Agent Using a Virtual Character to Induce Children's Verbal Communication," Journal of Korea Multimedia Society, Vol. 23, No. 10, pp. 1296-1306, 2020, DOI: 10.9717/kmms.2020.23.10.1296.

[17] E.J. Goldman, A.-E. Baumann, and D. Poulin-Dubois, "Preschoolers' Anthropomorphizing of Robots: Do Human-Like Properties Matter?," Frontiers in Psychology, Vol. 13, Article 1102370, 2023, DOI: 10.3389/fpsyg.2022.1102370

[18] J. Pavlopoulos, J. Sorensen, L. Dixon, N. Thain, and I. Androutsopoulos, "Toxicity Detection: Does Context Really Matter?," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4296-4305, 2020, DOI: 10.18653/v1/2020.acl-main.396.

[19] A. Ramponi and S. Tonelli, "Features or Spurious Artifacts? Data-Centric Baselines for Fair and Robust Hate Speech Detection," Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3027-3040, 2022, DOI: 10.18653/v1/2022.naacl-main.221.

[20] I. Sen, M. Samory, C. Wagner, and I. Augenstein, "Counterfactually Augmented Data and Unintended Bias: The Case of Sexism and Hate Speech Detection," Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4716-4726, 2022, DOI: 10.18653/v1/2022.naacl-main.347.

[21] K. Salmon, I.M. Evans, S. Moskowitz, M. Grouden, F. Parkes, and E. Miller, "The Components of Young Children's Emotion Knowledge: Which Are Enhanced by Adult Emotion Talk?," Social Development, Vol. 22, No. 1, pp. 94-110, 2013, DOI: 10.1111/sode.12004.

[22] D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman, "Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification," Companion Proceedings of the 2019 World Wide Web Conference, pp. 491-500, 2019, DOI: 10.1145/3308560.3317593.

[23] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J.B. Pierrehumbert, "HateCheck: Functional Tests for Hate Speech Detection Models," Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 41-58, 2021, DOI: 10.18653/v1/2021.acl-long.4.

[24] N. Kurian, "'No, Alexa, no!': Designing Child-Safe AI and Protecting Children from the Risks of the 'Empathy Gap' in Large Language Models," Learning, Media and Technology, Vol. 50, No. 4, pp. 621-634, 2025, DOI: 10.1080/17439884.2024.2367052.

[25] R. Caruana, "Multitask Learning," Machine Learning, Vol. 28, No. 1, pp. 41–75, 1997, DOI: 10.1023/A:1007379606734.

[26] E. Kim, "A Multi-task Self-attention Model Using Pre-trained Language Models on Universal Dependency Annotations,"

www.kci.go.kr

Journal of The Korea Society of Computer and Information, Vol. 27, No. 11, pp. 39-46, November 2022. DOI: 10.9708/jksci.2022.27.11.039.

[27] P. Kapil and A. Ekbal, "Leveraging Multi-domain, Heterogeneous Data using Deep Multitask Learning for Hate Speech Detection," Proceedings of the 17th International Conference on Natural Language Processing (ICON), pp. 491-500, Patna, India, December 2020. URL: https://aclanthology.org/2020.icon-main.66/

[28] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, et al., "Supervised Contrastive Learning," Advances in Neural Information Processing Systems, Vol. 33, pp. 18661-18673, 2020.

[29] A. Mostafazadeh Davani, M. Díaz, and V. Prabhakaran, "Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations," Transactions of the Association for Computational Linguistics, Vol. 10, pp. 92- 110, 2022, DOI: 10.1162/tacl_a_00449.

[30] S. Park, J. Moon, S. Kim, W.I. Cho, J. Han, J. Park, et al., "KLUE: Korean Language Understanding Evaluation," Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, Vol. 1, Virtual Conference, December 2021. DOI: 10.48550/arXiv.2105.09680.

[31] J. Lee, "KcBERT: Korean Comments BERT," Proceedings of the 32nd Annual Conference on Human and Cognitive Language Technology, pp. 437-440, Online, Korea, October 2020. URL: https://github.com/Beomi/KcBERT

[32] J. Park, "KoELECTRA: Pretrained ELECTRA Model for Korean," GitHub repository, 2020. URL: https://github.com/monologg/KoELECTRA

[33] S. Lee and H. Shin, "KR-ELECTRA: a KoRean-based ELECTRA model," GitHub repository, 2022. URL: https://github.com/snunlp/KR-ELECTRA

[34] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," PLOS ONE, Vol. 10, No. 3, Article e0118432, 2015, DOI: 10.1371/journal.pone.0118432.

[35] R. Dror, S. Shlomov, and R. Reichart, "Deep Dominance- How to Properly Compare Deep Neural Models," Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2773-2785, 2019, DOI: 10.18653/v1/P19-1266.

[36] K. Clark, M.-T. Luong, Q.V. Le, and C.D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," International Conference on Learning Representations, 2020. DOI: 10.48550/arXiv.2003.10555.

[37] D. Hupkes et al., "A taxonomy and review of generalization research in NLP," Nature Machine Intelligence, Vol. 5, pp. 1161-1174, 2023, DOI: 10.1038/s42256-023-00729-y.

Authors



Woomin Nam received his M.S. degree in Artificial Intelligence Convergence from Sungkyunkwan University, Korea, in 2022. He is currently pursuing his Ph.D. degree in Artificial Intelligence Convergence at Sungkyunkwan University, Korea.

Woomin Nam is currently a Research Fellow at the Urban AI EH, Tech University of Korea, Republic of Korea. From 2022 to 2025, he was a Researcher with the Urban Computing and Innovation Laboratory (UCIL) at the Advanced Institute of Convergence Technology (AICT), Republic of Korea. His research interests include AI cities, urban computing, urban data analytics, machine learning, deep learning, and AI applications in healthcare and urban planning.



Sudan Prasad Uprety received the B.Sc. degree in physics from Tri-Chandra Campus, Kathmandu, Nepal, in 2009, and the M.S. and Ph.D. degrees in business IT from Kookmin University, Seoul, Republic of Korea, in 2013 and 2021, respectively.

He is currently affiliated with Urban AI EH, Tech University of Korea, Republic of Korea.. His research interests include artificial intelligence, deep learning, machine learning, natural language understanding, big data analytics, IoT, data mining, and cloud computing.



Keon Chul Park received his M.S. and Ph.D. degrees in Information Systems from Yonsei University, Korea, in 2011 and 2015, respectively.

Dr. Park is currently an Assistant Professor in the Business Administration Department at Tech University of Korea. From 2021 to 2025, he served as the Head of the Urban Computing & Innovation Laboratory (UCIL) at Advanced Institute of Convergence Technology (AICT). His research interests include urban computing, data innovation policy, AI cities, data-driven urban planning, and artificial intelligence applications in urban environments.

www.kci.go.kr