Pebblous

AADS LLM 파인튜닝용 QA 데이터셋 구축: 사회안전 분야 데이터 품질 관점

- **작성일:** 2025년 11월 30일
- **기획:** (주)페블러스 데이터 커뮤니케이션팀
- **인터랙티브 콘텐츠:** https://blog.pebblous.ai/

서론

본 보고서는 Pebblus의 AADS(Agentic AI Data Scientist) 프로젝트의 성공적인 수행을 위한 핵심 전략을 기술합니다. AADS 프로젝트의 궁극적인 목표는 데이터 과학자의 작업을 효과적으로 보조할 수 있는 자율 AI 에이전트를 개발하는 것입니다. 이 목표를 달성하기 위한 핵심 과제는 데이터 과학 분야의 깊이 있는 지식과 추론 능력을 갖춘 맞춤형 대규모 언어 모델(LLM)을 확보하는 것입니다.

이를 위해 본 프로젝트에서는 LLM 파인튜닝(Fine-tuning) 전략을 채택하였으며, 그 기반이 되는 고품질 질의응답(QA) 데이터셋 구축에 집중하고 있습니다. 본 보고서는 그 첫 단계로 '사회안전' 분야의 주요 기술 문서들로부터 체계적으로 생성된 QA 데이터셋의 구체적인 구성과 내용을 상세히 기술합니다. 이 데이터셋은 AADS LLM이 전문 분야의 지식을 정확하게 학습하고 활용하는 능력의 초석이 될 것입니다.

---

1. 데이터셋 및 출처 문서 요약

LLM 학습용 데이터셋을 구축할 때, 그 범위와 출처를 명확히 하는 것은 매우 중요합니다. 이는 데이터셋의 근거를 투명하게 공개하고, AI가 학습할 정보의 신뢰성을 확보하는 첫걸음이기 때문입니다. 아래 표는 본 보고서에서 '사회안전' 분야의 QA 데이터셋 구축을 위해 기반으로 삼은 8개의 핵심 데이터셋과 그 출처 문서를 요약한 것입니다.

데이터셋 명칭	출처 문서
기반암 시추 시료를 이용한 암반 등급 분류 데이터	'23년 인공지능 학습용 데이터 활용 가이드라인(기반암 시추 시료를 이용한 암반 등급 분류 데이터) v1.2
지능형 관제 서비스 CCTV 영상 데이터	'24년 초거대AI 확산 생태계 조성 활용 가이드라인(04. 지능형 관제 서비스 CCTV 영상 데이터)_V1.0
건설용 자갈 품질관리 데이터	'23년 인공지능 학습용 데이터 활용 가이드라인(건설용 자갈 품질관리 데이터) v1.3

자연발생 석면 탐지 데이터	'24년 초거대AI 확산 생태계 조성 활용 가이드라인(자연발생 석면 탐지 데이터) v1.0
SOC 시설물 균열패턴 이미지 데이터	'23년 인공지능 학습용 데이터 활용 가이드라인(SOC 시설물 균열패턴 이미지 데이터) v1.0
놀이기구 및 시설 이용자 위험 상황 인식 데이터	'23년 인공지능 학습용 데이터 활용 가이드라인(놀이기구 및 시설 이용자 위험 상황 인식 데이터) v3.5
내륙습지 탄소흡수원 데이터	내륙습지 탄소흡수원 데이터 설명서 (2024)
화학물질 위험성 예측 데이터	화학물질 위험성 예측 데이터 설명서 (2024)

이처럼 다양한 성격의 데이터셋에서 핵심 정보를 추출하여 QA 쌍을 구성함으로써, LLM은 특정 주제에 치우치지 않고 다각적인 데이터 과학 지식을 균형 있게 학습할 수 있습니다.

2. QA 유형 정의 요약

체계적인 QA 프레임워크는 포괄적이고 균형 잡힌 학습 데이터셋을 구축하는 데 필수적입니다. 단순히 정보를 나열하는 것을 넘어, 질의를 명확한 유형으로 분류함으로써 문서의 여러 측면을 깊이 있게 탐색할 수 있습니다. 이는 LLM이 특정 도메인 데이터를 단편적으로 암기하는 것이 아니라, 종합적으로 이해하고 맥락에 맞게 추론하는 능력을 기르도록 유도합니다. 본 보고서에서는 질의를 아래 네 가지 핵심 유형으로 분류하여 QA 데이터셋을 생성했습니다.

질의 유형	설명
A: 도메인 정의/목적	데이터셋의 구축 목적, 활용 분야, 해결하고자 하는 문제 등 프로젝트의 근본적인 **'왜(Why)'**에 해당하는 정보를 질의합니다.
B: 데이터 구조/구성	데이터의 종류, 형식, 수량, 분포, 메타데이터 구조 등 데이터셋의 구체적인 **'무엇(What)'**에 해당하는 정보를 질의합니다.
C: AI 모델/임무	데이터를 활용하여 수행할 AI 임무(Task), 적용 모델, 알고리즘, 성능 목표 등 **'어떻게(How)'**에 해당하는 정보를 질의합니다.
D: 품질/공정 관리	데이터 수집, 정제, 가공, 검수 절차 및 품질 관리 기준 등 데이터 구축 과정의 **신뢰성**에 관한 정보를 질의합니다.

이 프레임워크는 이어질 각 데이터셋별 상세 QA 생성의 일관된 기준이 되며, LLM이 각 문서의 핵심 정보를 구조적으로 학습하는 데 기여합니다.

---

3. 데이터셋별 상세 QA 생성

본 섹션에서는 앞서 정의된 프레임워크를 기반으로, 분석된 8개의 사회안전 분야 데이터셋 각각에 대해

생성된 구체적인 QA 쌍을 제시합니다. 이 자료들은 AADS LLM이 실제 데이터 과학 문서의 핵심 정보를 정확히 이해하고, 개념을 연결하며, 논리적으로 추론하는 능력을 학습하는 데 직접적으로 사용될 핵심 자산입니다.

3.1. 기반암 시추 시료를 이용한 암반 등급 분류 데이터

유형	질의	응답
A	'기반암 시추 시료를 이용한 암반 등급 분류 데이터' 구축의 궁극적인 활용 목표는 무엇인가?	이 데이터의 핵심 목표는 암반 등급 판정의 정확도와 객관성을 높여, 터널 및 비탈면 공사 등에서 발생할 수 있는 붕괴 사고를 예방하는 것입니다. 이는 AI를 통해 기존의 주관적 판단에 의존하던 암반 분석을 표준화하고 자동화하여 건설 안전성을 확보하는 데 기여합니다.
B	이 데이터셋은 어떤 두 가지 주요 데이터로 구성되며, 각각의 구축 규모는 어떻게 되는가?	데이터셋은 '암종 분류 데이터셋'과 '절리 탐지 데이터셋'으로 구성됩니다. 암종 분류 데이터는 550,080쌍의 이미지(JPG)와 라벨링 데이터(JSON)로, 절리 탐지 데이터는 111,345쌍의 이미지와 라벨링 데이터로 구축되었습니다.
C	암종 분류와 절리 탐지라는 두 가지 다른 임무에 각각 ResNet과 Deeplab V3 모델이 사용된 기술적 이유는 무엇인가?	두 모델은 각기 다른 임무에 최적화된 아키텍처를 가지고 있습니다. ResNet은 깊은 신경망에서도 안정적인 학습이 가능한 잔차 연결(residual connection) 구조 덕분에 이미지 전체의 특징을 보고 특정 클래스로 분류하는 '이미지 분류(Classification)' 임무에 널리 사용됩니다. 반면, Deeplab V3는 Atrous Convolution을 사용하여 다양한 크기의 컨텍스트 정보를 효율적으로 포착하므로, 이미지 내 특정 영역(절리/단층)을 픽셀 단위로 정밀하게 분할하는 '세그멘테이션(Segmentation)' 임무에 더 적합합니다.
D	데이터 정제 과정에서 중복 이미지 파일을 제거하기 위해 해시(hash) 값을 사용한 이유는 무엇이며, 이 방법의 장점은 무엇인가?	Python의 Hashlib 라이브러리를 사용한 해시 값 비교는 이미지의 내용이 단 1비트라도 다르면 전혀 다른 해시 값이 생성되는 특성을 이용합니다. 이 방법은 픽셀 값 기반의 단순 비교보다 계산적으로 훨씬 효율적이며, 파일명이나 메타데이터가 다르더라도 내용이 완전히 동일한 중복 파일을 정확하고 신속하게 식별하여 제거할 수 있어 데이터의 순수성을 보장하는 데 매우 효과적입니다.

3.2. 지능형 관제 서비스 CCTV 영상 데이터

유형	질의	응답
A	'지능형 관제 서비스 CCTV 영상 데이터'를 구축하는 근본적인 목적은 무엇인가?	이 데이터셋의 목적은 지능형 CCTV가 침입, 싸움, 쓰러짐, 군집, 인파밀집, 침수 등 6종의 주요 안전사고 및 재난 상황을 자동으로 탐지하고 분석할 수 있는 AI 모델을 학습시키기 위함입니다. 이를 통해 관제 효율성을 높이고, 사건 발생 시 조기 감지 및 신속한 대응 체계를 마련하여 시민의 안전을 강화하는 것이 최종 목표입니다.
B	이 데이터셋은 총 몇 건의 영상으로 구성되며, 이벤트 유형별 분포는 어떻게 되는가?	총 300건의 영상으로 구성됩니다. 세부적으로는 일반 안전사고 4종(침입 55건, 쓰러짐 45건, 싸움 30건, 군집 70건) 200건과, 특화 이벤트 2종(인파밀집 40건, 침수 60건) 100건으로 분포되어 있습니다.
C	이 데이터셋의 AI 모델 임무가 '감시 영상 내 이상 행위 탐지'일 때, 성능 지표로 AUC를 사용하는 이유는 무엇이며, 목표치 80%는 어떤 의미를 가지는가?	'이상 행위 탐지'는 정상 상황 대비 비정상 상황의 발생 빈도가 현저히 낮은 불균형 데이터셋인 경우가 많습니다. AUC(Area Under the Curve)는 모델이 얼마나 이상 상황을 잘 판별하는지를 모든 가능한 임계값(threshold)에 대해 평가하므로, 특정 임계값에만 의존하는 정확도(Accuracy)보다 모델의 전반적인 판별 성능을 더 강건하게 측정할 수 있습니다. 목표치 80%는 모델이 무작위 추측(AUC 50%)보다 월등히 높은 수준으로 이상 행위를 탐지할 수 있음을 의미하는 실용적인 성능 목표입니다.
D	CCTV 영상에서 개인정보 비식별화 조치는 데이터 품질 관리 측면에서 어떤 중요한 역할을 하는가?	개인정보 비식별화(블러링, 마스킹)는 '개인정보 보호법' 준수라는 법적 요구사항을 충족하는 동시에, 모델의 성능과 일반화 능력을 높이는 핵심적인 품질 관리 조치입니다. 만약 안면, 차량 번호판 등의 정보가 그대로 노출되면, AI 모델이 문제의 본질(예: 싸움 행위)이 아닌 특정 인물이나 차량과 같은 부수적 정보에 과적합(overfitting)될 수 있습니다. 비식별화는 이러한 편향을 제거하여 모델이 순수하게 상황과 행동 패턴에만 집중하여 학습하도록 유도합니다.

3.3. 건설용 자갈 품질관리 데이터

유형	질의	응답
A	'건설용 자갈 품질관리 데이터' 구축은 어떤 산업적 문제를 해결하고, 어떤 시스템	이 데이터는 SOC 건설 시 콘크리트 품질에 직접적인 영향을 미치는 자갈의 품질을 AI로 자동 분석하기 위해 구축되었습니다. 이는 육안 검사에 의존하던 기존 방식의 비효율성과 주관성을 해결하며, 불량 골재 유통을 차단하기 위한 '골재 이력관리 시스템'과 연동되어 건설

	에 활용될 수 있는가?	자재의 신뢰도를 높이는 데 활용될 수 있습니다.
B	이 데이터셋은 어떤 두 가지 주요 데이터로 나뉘며, 각각의 라벨링 방식은 무엇인가?	데이터셋은 '자갈 암석 종류 분석 데이터'와 '편장석 비율 비 분석 데이터'로 나뉩니다. '자갈 암석 종류 분석'은 암석의 복잡한 경계를 따라 정밀하게 영역을 지정하는 폴리곤 세그먼테이션(Polygon Segmentation) 방식으로, '편장석 비율 비 분석'은 회전된 형태의 객체를 감싸는 로테이티드 바운딩 박스(Rotated Bounding Box) 방식으로 라벨링되었습니다.
C	편장석(flat and elongated particle) 비율 분석에 YOLO-OBB(Oriented Bounding Box) 모델이 사용된 기술적 이유는 무엇이며, 이 모델의 성능 목표인 'IoU@50, mAP 75%'는 무엇을 의미하는가?	자갈과 같은 편장석 입자는 길고 납작하여 일반적인 축 정렬 바운딩 박스(axis-aligned bounding box)로는 객체를 정확하게 감싸기 어렵습니다. YOLO-OBB는 회전된 경계 상자를 사용하므로, 이러한 비정형 객체의 방향과 형태를 더 정밀하게 탐지할 수 있어 기술적으로 적합합니다. 성능 목표에서 'IoU@50'은 예측된 박스와 실제 정답 박스 간의 중첩 영역(Intersection over Union)이 50% 이상일 때 '정답'으로 간주하겠다는 기준을 의미하며, 'mAP 75%'는 모든 클래스에 대한 평균 정밀도(Average Precision)를 종합한 값이 75%에 도달하는 것을 목표로 함을 뜻합니다.
D	데이터 정제 과정에서 중복 제거를 위해 'Dup Detector'를, 오류 제거를 위해 'Sony image edge view'를 사용한 이유는 무엇인가?	이는 정제 작업의 효율성과 정확성을 모두 고려한 전략입니다. 'Dup Detector'와 같은 자동 검사 도구는 대규모 이미지 데이터셋에서 내용이 동일한 중복 파일을 신속하게 찾아내는 데 효과적입니다. 반면, 초점이 맞지 않거나 수분으로 인한 빛 반사 오류 등은 기계가 판단하기 어려운 질적 문제이므로, 'Sony image edge view'를 통해 이미지를 확대하여 전문가가 육안으로 전수 검사하는 방식을 병행함으로써 데이터의 정량적, 정성적 품질을 모두 확보하였습니다.

3.4. 자연발생 석면 탐지 데이터

유형	질의	응답
A	'자연발생 석면 탐지 데이터' 구축의 궁극적인 목표와 사회적 기여는 무엇인가?	이 데이터 구축의 궁극적인 목표는 자연 환경에 존재하는 석면 및 석면 함유 가능 암석을 신속하고 정확하게 탐지하고 분류하는 AI 모델을 개발하는 것입니다. 이를 통해 석면으로 인한 잠재적 건강 위협을 사전에 관리하고, 관련 환경 및 보건 안전 정책 수립에 과학적 근거를 제공하여 사회 안전에 기여하고자 합니다.
	이 데이터셋은 어떤 세 가지 유	

B	형의 이미지 데이터로 구성되어 있는가? 이는 왜 멀티모달(multi-modal) 데이터로 간주되는가?	데이터셋은 광학(jpg), 가시근적외선 초분광영상(VNIR, tif), 단파적외선 초분광영상(SWIR, tif)의 세 가지 데이터로 구성됩니다. 각 데이터는 서로 다른 센서와 파장대역(가시광선, 근적외선 등)에서 정보를 수집하므로, 동일한 대상을 다른 관점(modality)에서 포착한 멀티모달 데이터로 간주됩니다.
C	이 멀티모달 데이터를 학습시키기 위해 '초기 결합(Early Fusion)'과 '후기 결합(Late Fusion)' 기법을 비교 평가하는 전략을 세운 이유는 무엇인가?	초기 결합(Early Fusion)은 광학, VNIR, SWIR 같은 여러 이종(multi-modal) 이미지 데이터의 특성을 입력 단계에서부터 통합하여 풍부한 피처를 함께 학습시키려는 전략입니다. 이는 각 데이터 간의 저수준 상호작용을 포착하는 데 유리할 수 있습니다. 반면, 후기 결합(Late Fusion)은 각 모달리티별로 독립적인 모델을 학습시킨 후 예측 결과를 결합하는 방식으로, 각 데이터의 고유한 특성을 깊이 학습하는 데 강점이 있습니다. 두 방식을 비교 평가함으로써, 석면 탐지라는 특정 임무에 어느 방식이 더 효과적인지 검증하고 최적의 모델 아키텍처를 결정하기 위함입니다.
D	초분광 영상 데이터에 대해 반사도, 대기, 기하 보정과 같은 정제 과정을 수행하는 이유는 무엇인가?	초분광 영상은 센서 자체의 노이즈, 대기 중의 수증기나 에어로졸에 의한 산란, 그리고 위성이나 드론의 촬영 각도로 인한 왜곡 등 다양한 외부 요인에 의해 데이터가 오염될 수 있습니다. 반사도, 대기, 기하 보정은 이러한 왜곡을 제거하여 물질 고유의 순수한 분광 특성(spectral signature)을 복원하는 필수적인 전처리 과정입니다. 이 과정을 통해 데이터의 정확성과 일관성을 확보해야만 AI 모델이 신뢰할 수 있는 패턴을 학습할 수 있습니다.

3.5. SOC 시설물 균열패턴 이미지 데이터

유형	질의	응답
A	'SOC 시설물 균열패턴 이미지 데이터'는 어떤 구체적인 서비스 및 의사결정 지원 도구 개발에 활용될 수 있는가?	이 데이터는 인력 접근이 어려운 교량, 터널 등 위험 지역의 시설물 균열을 자동으로 진단하는 AI 서비스 및 탐지 로봇 개발에 직접적으로 활용될 수 있습니다. 또한, 균열의 종류, 심각도 등을 정량적으로 분석하여 시설물의 유지·보수 우선순위를 결정하고 장기적인 관리 계획을 수립하는 의사결정 지원 도구를 개발하는 데 핵심적인 기반 데이터로 사용됩니다.
	데이터셋의 라벨링 클래스에는	

B	어떤 10가지 종류의 균열 및 손상 패턴이 포함되어 있는가?	데이터셋에는 균열, 망상균열, 박리, 박락, 백태, 누수, 철근노출, 재료분리, 들뜸, 파손 등 총 10가지 종류의 대표적인 시설물 손상 패턴이 클래스로 정의되어 있습니다.
C	이 데이터셋을 활용하여 '이미지 분류'와 '이미지 분할'이라는 두 가지 임무를 수행하기 위해 각각 CvT와 SegFormer 모델을 사용한 이유는 무엇인가?	'이미지 분류(Classification)'는 이미지 전체에 어떤 종류의 균열이 있는지를 판단하는 임무로, 이미지의 전역적인 특징을 효과적으로 학습하는 CvT(Convolutional Vision Transformer) 모델이 적합합니다. 반면, '이미지 분할(Segmentation)'은 균열이 이미지 내 어느 위치에 어떤 형태로 존재하는지를 픽셀 단위로 정확히 구분해내는 임무로, 경량화된 구조이면서도 높은 분할 성능을 보이는 SegFormer 모델이 더 효율적입니다. 이처럼 두 가지 연관되면서도 다른 임무를 위해 각각에 최적화된 모델을 적용하여 전체 시스템의 성능을 극대화한 것입니다.
D	동영상에서 프레임을 추출할 때 SSIM(구조적 유사성 지수)을 사용한 이유는 무엇이며, 이것이 AI 모델 학습에 어떤 긍정적 영향을 미치는가?	영상 데이터에서 프레임 간 유사도가 높은 중복 이미지를 제거하기 위해 SSIM(구조적 유사성 지수)을 사용했습니다. SSIM 지표가 0.9 이하인 프레임만 추출함으로써, 모델이 거의 동일한 이미지를 반복적으로 학습하여 발생하는 과적합(overfitting)을 방지하고, 한정된 데이터셋 내에서 최대한 다양한 균열 패턴을 학습하도록 데이터의 다양성을 확보했습니다. 이는 모델의 일반화 성능을 높이는 데 중요한 품질 관리 기법입니다.

3.6. 놀이기구 및 시설 이용자 위험 상황 인식 데이터

유형	질의	응답
A	'놀이기구 및 시설 이용자 위험 상황 인식 데이터' 구축의 전략적 목적은 무엇인가?	이 데이터의 전략적 목적은 놀이기구 및 관련 시설에서 발생할 수 있는 다양한 위험 상황과 시설물의 파손 상태를 AI가 자동으로 인식하고 분석하도록 학습시키는 것입니다. 이를 통해 사고 발생 가능성을 사전에 예측하여 경고하거나, 사고 발생 시 즉각적인 대응을 지원함으로써 이용자의 안전을 선제적으로 확보하는 지능형 안전 관리 시스템을 구축하는 것이 목표입니다.
	이 데이터셋의 라벨링이 바운딩박스, 키포인트, 세그멘테이	세 가지 라벨링 유형은 서로 다른 수준의 정보를 제공하여 복합적인 상황 인식을 가능하게 합니다. '바운딩박스'는 객체(사람, 놀이기구)의 위치와 존재를 신속하게 탐지하고, '키포인트'는 사람의 관절 위치를 추적하여 자세나 행동을 정밀하게 분석하며, '세그멘테이션'은 시설물의 파손 영역을

B	션 세 가지 유형으로 구성된 이유는 무엇인가?	픽셀 단위로 정확하게 지정합니다. 이 세 가지를 조합함으로써 AI는 "누가(바운딩박스) 어떤 위험한 자세로(키포인트) 파손된 영역(세그멘테이션) 근처에 있는지"와 같은 복합적인 시나리오를 종합적으로 이해할 수 있습니다.
C	놀이시설 객체 탐지 모델로 오픈소스 Foundation 모델인 'InternImage-L'을 기반으로 한 모델을 사용한 전략적 이점은 무엇인가?	'InternImage-L'과 같은 대규모 사전학습된 Foundation 모델을 기반으로 사용하는 것은 전이 학습(transfer learning)의 이점을 극대화하는 전략입니다. 이 모델은 이미 방대한 양의 일반 이미지 데이터로부터 시각적 특징을 추출하는 능력을 학습했기 때문에, 놀이시설이라는 특정 도메인에 대해 상대적으로 적은 데이터로도 빠르고 효과적으로 미세조정(fine-tuning)하여 높은 성능을 달성할 수 있습니다. 이는 모델 개발 시간과 비용을 단축시키는 효율적인 접근 방식입니다.
D	데이터 구축 공정에서 '데이터 가공 → 검수 → 재가공 → 검증'의 4단계 품질 관리 절차를 따르는 이유는 무엇인가?	이 4단계 절차는 체계적인 피드백 루프를 형성하여 데이터 품질을 지속적으로 향상시키기 위함입니다. 1차 '가공' 후 전문가의 '검수'를 통해 오류를 식별하고, 불합격된 데이터는 피드백과 함께 '재가공' 단계로 돌아가 수정됩니다. 이 과정을 반복하여 데이터의 정확성을 높인 후, 최종적으로 외부 전문기관의 '검증'을 통해 객관적인 품질을 확보합니다. 이러한 반복적 개선 및 검증 프로세스는 대규모 데이터 구축 프로젝트에서 일관되고 높은 품질을 유지하는 데 필수적입니다.

3.7. 내륙습지 탄소흡수원 데이터

유형	질의	응답
A	'내륙습지 탄소흡수원 데이터'를 구축하는 목적은 기후 변화 대응과 어떤 연관이 있는가?	이 데이터는 내륙습지의 식생, 수역 등을 분석하여 탄소 흡수 및 저장량을 추정하고 그 변화를 모니터링하는 AI 모델을 개발하기 위해 구축됩니다. 습지는 중요한 탄소흡수원이므로, AI를 통해 그 기능을 정량적으로 평가하고 관리하는 것은 국가 온실가스 감축 목표 달성에 기여하고, 기후 변화 대응을 위한 과학 기반 환경 정책을 수립하는 데 핵심적인 역할을 합니다.
	이 데이터셋은 위성영상, 드론영상, 수치자료 등 다양	각 데이터는 서로 다른 공간적, 시간적 해상도와 정보를 제공하여 상호 보완적인 역할을 합니다. Sentinel 위성영상은 넓은 지역을 주기적으로 관측하여 거시적인 변화를 파악하는 데 유리하고, 고해상도 드론 영상은 특정 지역

B	한 종류의 원천 데이터로 구성되어 있는데, 그 이유는 무엇인가?	의 세밀한 식생 및 지형 분석에 적합합니다. 여기에 강수량, 기온, 수위와 같은 수치 자료를 결합함으로써, 영상 데이터만으로는 알 수 없는 환경적 요인을 함께 분석하여 탄소흡수량 추정 모델의 정확도를 높일 수 있습니다.
C	드론과 위성 이미지 분석에 각각 'Trans UNet'과 'Modified Trans UNet'을 사용한 이유는 무엇인가?	Trans UNet은 이미지의 전역적 맥락을 파악하는 데 강한 Transformer와 지역적 특징 추출에 뛰어난 U-Net(CNN 기반)을 결합한 모델로, 고해상도 드론 이미지 내 객체 영역을 정밀하게 분할하는 데 효과적입니다. 위성 이미지의 경우, 드론 이미지보다 해상도가 낮고 분광 특성이 다르며 대기 효과 등 노이즈가 더 많기 때문에, 이러한 데이터 특성에 맞게 모델 구조나 전처리 부분을 조정한 'Modified Trans UNet'을 사용한 것으로 추정됩니다. 이는 각 데이터 소스의 고유한 특성에 맞게 모델을 최적화하는 전략입니다.
D	Sentinel-1 위성영상 데이터에 대해 SNAP 소프트웨어를 활용한 표준 전처리 절차를 수행하는 이유는 무엇인가?	Sentinel-1은 레이더(SAR) 위성으로, 원시 데이터는 방사 왜곡 및 지형에 의한 기하 왜곡을 포함하고 있어 분석에 바로 사용하기 어렵습니다. 유럽우주국(ESA)에서 제공하는 공식 소프트웨어인 SNAP을 활용하여 방사 보정, 지형 보정 등 표준 전처리 절차를 수행하면, 이러한 왜곡을 제거하고 분석에 적합한(Analysis-Ready Data) 후방산란계수 값을 얻을 수 있습니다. 이는 데이터의 과학적 정확성과 여러 영상 간의 일관성을 확보하기 위한 필수적인 품질 관리 단계입니다.

3.8. 화학물질 위험성 예측 데이터

유형	질의	응답
A	'화학물질 위험성 예측 데이터' 구축은 어떤 실질적인 안전 관리 문제 해결을 목표로 하는가?	이 데이터는 화학물질의 물리화학적 특성(증기압, 연소열, 인화점)과 위험성 간의 관계를 AI가 학습하여, 알려지지 않은 물질의 위험성을 예측하는 모델을 구축하는 것을 목표로 합니다. 이는 실험 데이터가 부족한 신규 화학물질의 잠재적 위험을 사전에 평가하고, 화학물질을 취급하는 현장에서 안전 관리 기준을 수립하며, REACH와 같은 국제 규제에 효과적으로 대응하는 데 기여합니다.
B	이 데이터셋이 포함하는 세 가지 핵심 물리/화학적 특성은 무엇이며, 각각의 라벨링 데이터 구축량은	데이터셋은 화학물질의 위험성을 판단하는 데 중요한 세 가지 핵심 특성인 '증기압'(5,110건), '연소열'(5,005건), '인화점'(10,010건)에 대한 라벨링 데이터로 구성됩니다.

	얼마인가?	
C	이 데이터를 기반으로 최종적으로 '웹 기반 예측 서비스'를 개발하는 이유는 무엇이며, 이는 사용자에게 어떤 가치를 제공하는가?	'웹 기반 예측 서비스'를 개발하는 것은 전문적인 소프트웨어나 화학 지식이 없는 사용자(예: 현장 안전 관리자, 소방관)도 화학물질의 위험성을 손쉽게 예측하고 접근할 수 있도록 하기 위함입니다. 사용자가 웹 브라우저에서 분자 구조를 그리거나 입력하기만 하면 AI 모델이 즉시 위험성을 예측해주므로, 정보 접근성이 극대화됩니다. 이는 화학 안전 정보를 민주화하고, 신속한 현장 의사결정을 지원하여 사고 예방에 직접적으로 기여하는 높은 가치를 제공합니다.
D	GHS(Globally Harmonized System) 정보를 기반으로 화학물질의 위험성을 '위험', '자료없음', '해당없음' 등으로 표준화하여 관리하는 이유는 무엇인가?	GHS는 국제적으로 통용되는 화학물질 분류 및 표기 시스템입니다. 수집된 GHS 유해성 정보를 '위험', '자료없음', '해당없음'과 같은 명확하고 일관된 라벨로 표준화하는 것은 데이터의 정합성과 품질을 확보하는 핵심적인 데이터 거버넌스 활동입니다. 이 과정을 통해 AI 모델이 모호하거나 일관성 없는 텍스트 정보가 아닌, 명확하게 정의된 범주(category)를 학습하게 되어 예측 성능과 신뢰도를 높일 수 있습니다.

결론

본 보고서는 AADS 프로젝트의 LLM 파인튜닝을 위해 '사회안전' 분야의 8개 핵심 데이터셋으로부터 체계적으로 구축된 QA 데이터셋을 제시했다는 점에서 중요한 기여를 합니다. 특히, 본 데이터셋은 단순한 정보 추출을 넘어, 정의, 구조, AI 모델의 기술적 선택 이유, 품질 관리의 전략적 의미를 묻는 추론 기반 질의를 포함함으로써 LLM이 복잡한 데이터 과학 문서의 핵심 정보를 구조적으로 이해하고 논리적으로 사고하는 능력을 기르도록 설계되었습니다.

이 데이터셋은 AADS 프로젝트의 성공적인 LLM 파인튜닝을 위한 전략적 자산으로서, 데이터 과학자의 작업을 지능적으로 보조할 AI 에이전트 개발에 실질적인 진전을 가져올 것입니다. 나아가, 본 보고서에서 제시된 심층적이고 체계적인 QA 데이터셋 구축 방법론은 향후 다른 전문 분야로 확장될 수 있는 검증된 모델로서, AADS 프로젝트의 지속적인 발전에 기여할 잠재력을 지니고 있습니다.

Pebblous

> Pebblous Makes Data Tangible

contact@pebblous.ai