Pebblous

AADS LLM 파인튜닝용 QA 데이터셋 구축: 제조 분야

- **작성일:** 2025년 11월 29일
- **기획:** (주)페블러스 데이터 커뮤니케이션팀
- **인터랙티브 콘텐츠:** https://blog.pebblous.ai/

I. 서론 및 구축 목표

이 보고서는 Agentic AI Data Scientist (AADS) 과제에서 대규모 언어 모델(LLM)의 제조 도메인 전문성 강화를 목표로, 제공된 데이터셋 문서(데이터 설명서 및 활용 가이드라인)를 기반으로 구축한 질의응답(QA) 쌍 샘플을 요약합니다. QA 쌍은 도메인 지식, 데이터 구조, AI 모델링 및 품질 관리 절차 등 4가지 유형으로 분류되었습니다.

II. 논리적 데이터 그룹별 QA 샘플 (총 14개 그룹, 28쌍)

1. 제조 현장 이송 장치(OHT/AGV) 탄화 예지보전 데이터

유형	질의 (Query)	응답 (Answer)	출처
A. 도메인 정의/목적	**OHT 및 AGV 열화 데이터셋의 궁극적인 구축 목적**은 무엇이며, 이 서비스를 활용할 주요 정책 대상 기업은 무엇인가?	이 데이터셋의 구축 목적은 반도체, 디스플레이, 자동차, 의료 등 **제조 현장의 OHT, AGV 장치**를 대상으로 실시간 데이터를 기반으로 **탄화 발생을 사전 예측하고 방지하는 서비스 개발을 지원**하기 위함입니다. 주요 정책 대상은 복잡한 공정과 고가의 장비로 인해 화재 위험에 노출된 반도체, 디스플레이, 자동차, 의료기기 제조 기업입니다.	
B. 데이터 구조/	이 데이터셋의 **멀티모달 구성 요소**와 라벨링 유형인 **태깅**의 탄화 상태 분류 기준	데이터셋은 **센서 값** (PM10, NTC, CT1~CT4, 외부 온도, 습도, 조도 등)과 **이미지** (열화상 이미지)의 멀티모달 데이터로 구성됩니다,. 탄화 상태는 **0: 정상, 1: 관심, 2: 경고, 3: 위험**의 네 가지 상태로 분류됩니다,.	,

구성	4가지는 무엇인가?		

2. 김치 생산공정별 비파괴 품질검사 데이터

유형	질의 (Query)	응답 (Answer)	출처
A. 도메인 정의/목적	김치 생산공정 데이터셋의 최종 **AI 활용 목표**는 무엇이며, 원물 배추 데이터는 초분광 이미지를 통해 어떤 품질 지표를 추정하는가?	AI 모델의 주요 임무는 **김치 제조/생산공정의 비파괴적 품질등급 예측 AI 모델 개발 및 서비스**를 지원하는 것입니다. 원물 배추의 초분광 이미지는 **무게, 크기, 당도, 수분 함량**의 품질 지표를 초분광 이미지를 기반으로 추정 AI 학습을 위해 구축되었습니다.	
C. AI 모델/임무	김치 품질 지표를 평가하기 위해 제시된 **MultiTask EfficientNetV2 모델**은 등급을 어떻게 분류하며, 절임배추 품질 예측 모델의 성능 목표는 무엇인가?	MultiTask EfficientNetV2 모델은 품질 지표를 평가하기 위해 사용되며, 모든 지표는 기본적으로 **상(0), 중(1), 하(2)로 나뉩니다**. 절임배추 품질등급 예측모델 성능의 **정량 목표는 F1-score 70% 이상**이며, 염도 등급 예측과 당도 등급 예측 모두 해당됩니다.	’

3. 3D 프린팅 출력물 형상 보정용 데이터

유형	질의 (Query)	응답 (Answer)	출처
B. 데이터 구조/구성	3D 프린팅 출력물의 품질을 검증하기 위한 **외형품질이미지 데이터**와 **수축분석 데이터**의 총 구축 수량은 각각 얼마이며, 사용된 프린터 유형은 무엇인가?	외형품질이미지 데이터는 **165,780장** 구축되었고, 수축분석 이미지 데이터는 **55,260장** 구축되었습니다. 사용된 프린터 유형으로는 G_FDM, I_FDM, SLA, DLP, MJP, SLS 등이 있습니다,.	,,
D. 품질/공	3D 프린팅 데이터셋의 **가공(라벨링) 검수 절차**에서 **재현률(recall) 검사**는 어떤 오류를 중점적으로 확	재현률 검사는 파일 내 **바운딩 박스를 그릴 객체가 누락되었는지 검사**하는 단계입니다. 재현률 검수에서 불통과 처리된 이미지는, **정밀도 검사**	

정관리	인하며, 이 검사에서 불통과 시 어떻게 처리되는가?	**를 통과한 바운딩 박스가 유지된 채 작업자들에게 돌아가 재작업**됩니다.	

4. 금속 3D 프린팅 스파크 이미지 및 기계적 물성 예측 데이터

유형	질의 (Query)	응답 (Answer)	출처
A. 도메인 정의/목적	금속 3D 프린팅 스파크 이미지 데이터 구축의 **최종 목적**은 무엇이며, 이 데이터셋을 통해 얻을 수 있는 **기대 효과 2가지**는 무엇인가?	구축 목적은 **금속 3D프린팅 스파크 이미지 분석을 통한 금속 적층 가공물의 기계적 물성 예측이 가능**한 데이터를 구축하는 것입니다. 기대 효과로는 **금속 3D프린팅 출력물의 품질 신뢰도와 생산성 향상** 및 **최적 공정 조건 추천 서비스** 개발을 통한 출력 비용 절감이 있습니다.	
B. 데이터 구조/구성	금속 3D 프린팅 스파크 이미지 데이터의 **총 수량**은 얼마이며, **원천 데이터**를 구성하는 파일 포맷 3가지를 제시하라.	금속 3D 프린팅 스파크 이미지 데이터의 총 수량은 **51,267,476**개입니다. 원천 데이터를 구성하는 파일 포맷은 **.bmp** (이미지) 및 **.json** (기계적 특성치), **.txt** (melt_temperature) 입니다.	

5. 건설기계 무인 운행 자율 작업 데이터

유형	질의 (Query)	응답 (Answer)	출처
C. AI 모델/임무	건설기계 데이터셋을 활용한 **굴착기의 내부 작업 순서 데이터**에 대한 AI 모델 Task와 라벨링 방법은 무엇이며, 이 모델의 성능 지표는 무엇인가?	굴착기의 내부 작업 순서 데이터의 AI 모델 Task는 **Activity Recognition (활동 인식)**이며, 라벨링 방법은 **Frame-by-Frame Video Annotation**입니다. 시계열 분류 모델(LSTM)을 사용한 이 모델의 유효성 검증 성능 지표는 **F1-score 89.98%**를 달성했습니다,.	,,
D. 품	건설기계 무인 운행 데이터의 **팬옵틱 분할 모델**의	팬옵틱 분할 모델은 **Mask2Former**를 사용하며, 이는 트랜스포머(Transformer) 아키텍처를 기반으로	

질/공정 관리	유효성 검증 모델로 사용된 알고리즘의 작동 방식과 **성능 지표(PQ)**를 설명하라.	**Attention 메커니즘**을 도입하여 이미지의 각 픽셀에 대한 분할(segmentation) 마스크를 예측할 수 있습니다,. 이 모델은 유효성 지표 **PQ (Panoptic Quality)**를 사용하며, **78.41%**를 달성했습니다,.	,,

6. 고품질 연구개발용 리튬이온 이차전지 데이터

유형	질의 (Query)	응답 (Answer)	출처
A. 도메인 정의/목적	이 데이터셋의 **핵심 구축 방법**은 LLM을 어떻게 활용하는 것이며, 이 과정을 통해 추출된 텍스트의 **원본 소스**는 무엇인가?	이 프로젝트는 **대규모 언어 모델(LLM), 특히 GPT 모델**을 사용하여 과학 출판물에서 의도된 정보를 추출합니다. 추출된 텍스트의 원본 소스는 **과학 출판물의 PDF 파일**이며, 이 파일에서 **실험/방법과 결과 및 토론 섹션**의 텍스트만 추출하고 나머지는 폐기합니다.	,
C. AI 모델/임무	과학 출판물에서 **LLM의 성능과 효율성**을 향상시키고 정확한 표 형식 출력을 얻기 위해 사용된 **파인 튜닝 기법**은 무엇인가?	LLM을 미세 조정(파인튜닝)하기 위해 **퓨샷 학습(Few-Shot Learning)과 프롬프트 엔지니어링 기법**이 사용되었으며, 이 과정을 통해 모델이 필요한 데이터를 정확하게 추출하여 **표 형식으로 출력**하도록 훈련합니다.	

7. 전기 인프라 지능화를 위한 가전기기 전력 사용량 데이터

유형	질의 (Query)	응답 (Answer)	출처
A. 도메인 정의/목적	이 데이터셋이 목표로 하는 **AI 기반 기술**인 **NILM 기술**은 무엇을 의미하며, 데이터 구축 목표량 중 **가전기기 전력 사용량 데이터**의 규모는 얼마인가?	이 데이터셋은 **AI 기반 NILM (Non-Intrusive Load Monitoring) 기술 개발**을 위한 것입니다. NILM은 분전반의 총 전력 사용량 데이터에서 개별 가전기기의 전력 사용 패턴을 추정하는 기술입니다. 구축 목표량은 총 40,641건 중 **가전기기 전력 사용량 데이터 37,231건**입니다.	
C.	이 데이터셋을 활용하여 구	두 가지 주요 AI 임무는 **기기별 유효전력 분해**와 **기기**	

AI 모델/임무	현할 수 있는 **두 가지 주요 AI 임무(Task)**와 각각에 적용할 수 있는 **알고리즘**의 예시 및 유효성 검증 결과는 무엇인가?	**의 활성-비활성 상태 분류**입니다. 유효전력 분해 모델에는 **seq2points**가, 활성-비활성 탐지 모델에는 **unet**이 적용되며, 상태 분류 성능의 유효성 검증 결과는 **F1-점수 95.5%**입니다.	,

8. CMF(Color, Material, Finish) 이미지 식별 데이터

유형	질의 (Query)	응답 (Answer)	출처
A. 도메인 정의/목적	CMF 식별 데이터셋 구축의 **목표 임무 유형**은 무엇이며, 학습 모델 후보 중 **1순위 알고리즘**과 선정 사유는 무엇인가?	이 데이터셋의 임무 유형은 **이미지 분류(Image Classification)**입니다. 학습 모델 후보 중 1순위 알고리즘은 **Swin Transformer**이며, 이는 Shifted window 방식으로 이미지를 분할하여 attention을 계산하는 계층적 트랜스포머 구조를 가지며, 작은 물체부터 큰 물체까지 효율적으로 검출하여 높은 성능을 달성할 수 있기 때문입니다.	
B. 데이터 구조/구성	CMF 라벨링 세부 정보 (`annotations.label`) 에 포함되는 **필수 정보** 3가지와 `material_finishing` 속성의 **설명 범위**는 무엇인가?	라벨링 세부 정보에는 **바운딩 박스 정보** (`bndbox`), **색깔 정보** (`color`), 그리고 **material_finishing 정보** 등이 필수적으로 포함됩니다. `material_finishing` 속성의 범위는 **0부터 32까지**의 코드를 가지며, 이는 **거울광택이 있는 금속 느낌**부터 **엷은 목재느낌**까지의 33가지 유형을 포함합니다,.	,

9. 실험 기반 재료 물성 데이터

유형	질의 (Query)	응답 (Answer)	출처
A. 도메인 정의/목적	이 데이터셋의 **주요 활용 목적**은 무엇이며, **Hardness 예측 모델**의 학습 알고리즘은 어떤 방식인가?	이 데이터셋은 **인공지능 기반의 금속 물성 예측 모델에 활용**될 수 있도록 물성 데이터를 데이터베이스화하는 것을 목표로 합니다. Hardness 예측 모델의 학습 알고리즘은 **Random Forest Regressor**이며, 이는 여러 개의 결과를 합쳐 최종 결과를 도출하는 **앙상블(Ensemble) 모델** 방식을 사용합니다.	

C. AI 모델/임무	Hardness 예측 모델의 **AI 모델 사용 데이터 비율**과 이 모델의 **성능 지표 및 목표**는 무엇인가?	Hardness 예측 모델에는 전체 구축 데이터 대비 **100%** (1,000 행)의 데이터가 사용되며, 이 중 **Training Set 비율은 80% (800 행)**, **Test Set 비율은 20% (200 행)**입니다. 성능 지표는 **Predicted R^2**이며, 목표값은 명시되어 있지 않지만, 우수한 예측을 목표로 합니다.	

10. 배터리 불량 이미지 진단 데이터

유형	질의 (Query)	응답 (Answer)	출처
A. 도메인 정의/목적	**배터리 불량 이미지 데이터**가 구축됨으로써 기존의 **전기적 시험을 보완**하여 연구할 수 있는 새로운 표준 마련의 방향은 무엇인가?	이 데이터셋을 통해 **기존 전기적 시험을 넘어 내부 상태 등을 종합적으로 평가하는 새로운 배터리 등급 표준 마련이 가능**하며, **전기적 시험 등급과 배터리 내부 결함 간의 상관관계를 밝히는 연구**에 활용될 수 있습니다.	
D. 품질/공정 관리	배터리 불량 이미지 데이터 구축 과정에서 **라벨링 검수**의 **최소 크기 기준**은 무엇이며, **CT 데이터셋**의 최종 **결함 검출 성능 (mIoU)**은 얼마인가?	라벨링 검수 시 **폴리곤의 최소 크기는 가로 세로 4픽셀** 이상으로 설정되었습니다. CT 데이터셋의 최종 결함 검출 성능(AI 모델 학습 결과)은 **mIoU 92.79%**를 달성했습니다.	

11. LNG 탱크 부품 품질 검사 영상 데이터

유형	질의 (Query)	응답 (Answer)	출처
B. 데이터 구조/구	LNG 탱크 품질 검사 데이터의 **라벨링 유형 3가지**와 JSON 라벨링 데이터에 포함된 **LNG 탱크의 고유 속성 (attributes)** 3가지를 설명하라.	라벨링 유형은 **폴리곤, BB(바운딩 박스), 분류**입니다,. JSON 라벨링 데이터의 속성에는 **탱크 유형** (`tank_type`), **용량** (`volume`), **소재** (`material`), **위치** (`location`), **부품** (`part`), **품질** (`quality`) 등이 포함되며, 예를 들어 **tank_type, volume, material**을 들 수 있습니다.	’

성			
C. AI 모델/임무	LNG 탱크 품질 검사 이미지 데이터셋의 유효성 검증 결과, **세그멘테이션 객체 탐지 모델**의 mAP 결과값과 정량 목표는 각각 얼마이며, 사용된 알고리즘은 무엇인가?	세그멘테이션 객체 탐지 모델의 유효성 검증 결과는 **mAP 95.72%** (목표 79.43% 이상)를 달성했습니다,. 이 모델은 **Mask DINO**를 활용하며, 트랜스포머 아키텍처를 기반으로 합니다.	''

12. 조선·해양플랜트 P&ID 심볼 식별 데이터

유형	질의 (Query)	응답 (Answer)	출처
A. 도메인 정의/목적	조선·해양플랜트 P&ID 심볼 식별 데이터셋을 활용하여 AI 모델이 **설계 및 품질 관리 측면**에서 수행할 수 있는 주요 기능 2가지를 설명하라.	AI 모델은 학습된 모델을 활용하여 **설계 도면의 P&ID 심볼을 자동으로 분류**하고, **P&ID별 수량, 도면상의 위치**를 출력하며, **불필요하거나 잘못된 P&ID를 판단**할 수 있습니다.	
B. 데이터 구조/구성	P&ID 심볼 객체의 JSON 라벨링 데이터에 포함되는 **필수 속성 (attributes)** 세 가지를 제시하고, 이 속성 중 **대상 유형** (`shipType`) 의 예시 3가지를 제시하라.	필수 속성은 **설계사 부호** (`vendor`), **대상 유형** (`shipType`), **P&ID 심볼 문자** (`pidLabel`)입니다,. 대상 유형 (`shipType`)의 예시로는 **"FPSO"**, **"Drillship"**, **"Semi-Flg"** 등이 있습니다.	

13. 선박 도장 품질 측정 데이터

유형	질의 (Query)	응답 (Answer)	출처
A. 도메인 정	**선박 도장 품질 측정 데이터셋**의 **구축 목적**은 무엇이며, 이 데이터셋에 포함된 도장	구축 목적은 **선박 도장의 손상 정도를 파악하고 도장 품질의 검사 정확도를 개선**하는 것입니다. 불량 유형 중 **용접 손상**은 **20,352건**으로 **19.8%**를 차지합	'

의/목적	불량 유형 중 **용접 손상**이 차지하는 비율은 얼마인가?	니다.	
D. 품질/공정 관리	선박 도장 품질 데이터셋의 **2차 검차**는 어떻게 수행되며, **유효성 검사**에서 사용되는 성능 지표 2가지를 제시하라.	2차 검차는 **선별 검사**로 진행되며, **라벨링 값이 참 값(Ground Truth)과 일치하는지**를 크라우드소싱 플랫폼을 이용하여 Eye Checking으로 수행합니다. 유효성 검사에서 사용되는 주요 성능 지표는 **Top-1 Accuracy, mAP@50, MIOU** 등입니다.	

14. 용접 AI 학습 데이터 (육안 및 방사선 검사)

유형	질의 (Query)	응답 (Answer)	출처
C. AI 모델/임무	용접 AI 학습 데이터를 활용하여 용접 결함을 탐지하는 **AI 학습 모델**로 어떤 알고리즘이 제시되었으며, 이 모델이 수행하는 주요 임무와 라벨링 유형은 무엇인가?	적용 모델(알고리즘)은 **YOLOv5x-seg**이며, 이 모델은 **용접 자동화를 위한 검사 유형별 용접 이미지 데이터 구축**에 사용됩니다. 데이터의 라벨링 유형은 **폴리곤**입니다,.	'
B. 데이터 구조/구성	**일반 강재(Steel) 모재**에 대한 **육안검사(VTST)** 데이터의 **총 원천 데이터 수량**은 얼마이며, 가장 많이 구축된 결함 유형 2가지는 무엇인가?	일반 강재 모재(VTST)에 대한 육안검사 데이터의 총 원천 데이터 수량은 **74,019장**입니다. 가장 많이 구축된 결함 유형 2가지는 **용입부족(Incomplete penetration) 16,180개**와 **언더컷(Undercut) 12,195개**입니다.	

III. 질의-응답 유형 최종 통계

LLM 학습 데이터 생성을 위해 총 **14개**의 논리적 데이터 그룹에 대해 **28개**의 질의응답 쌍을 구성했습니다. 데이터 과학의 제조 도메인 적용 측면을 반영한 질의 유형 통계는 다음과 같습니다.

질의 유형	정의	사용 횟수	비율

A. 도메인 정의/목적	데이터의 최종 목적, 비즈니스 목표, 도메인 정의	**7회**	25.0%
B. 데이터 구조/구성	데이터 규모, 포맷, 라벨링 구성 요소 및 분포 등	**7회**	25.0%
C. AI 모델/임무	적용 알고리즘, AI Task 정의, 예측 목표 및 성능 지표	**7회**	25.0%
D. 품질/공정 관리	데이터 획득/가공/검수 절차 및 품질 관리 기준	**7회**	25.0%
총합		**28회**	**100.0%**

- **특징:** 제조 도메인 데이터 과학의 핵심 요소인 **목적 (A)**, **데이터 특징 (B)**, **기술 적용 (C)**, **품질 관리 (D)** 영역에 대해 균등하게 질문을 배분하여 LLM이 전 영역에 걸친 종합적인 지식을 학습하도록 설계되었습니다.

IV. 도메인 LLM 보고서 생성을 위한 프롬프트 템플릿

이 프롬프트는 다른 도메인(예: 도메인 로봇, 헬스케어, 자율주행 등)의 학습 데이터 문서가 주어졌을 때, 해당 도메인의 전문 지식을 LLM이 학습할 수 있도록 구조화된 QA 데이터셋 보고서를 생성하는 데 사용될 수 있습니다.

Report Generation Prompt Template (Korean/English Hybrid)

```
[지시사항]
당신은 Agentic AI Data Scientist (AADS) 과제에서 대규모 언어 모델(LLM) 파인튜닝을 위한 전문 QA 데이터셋을 구축하는 전문가입니다. 아래에 제시된 [INPUT: 분석 대상 문서]의 내용을 분석하여, **'논리적 데이터 그룹'** 단위로 묶어 QA 보고서를 생성해야 합니다.

**[보고서 구성 요소]**
1.  **보고서 제목:** 도메인 및 목적에 맞게 작성하십시오.
2.  **논리적 데이터 그룹 식별:** 문서 내에서 동일한 프로젝트나 목표를 공유하는 문서들을 하나의 '논리적 그룹'으로 묶습니다.
3.  **QA 쌍 생성:** 각 논리적 그룹별로 **2개**의 질의응답(QA) 쌍을 생성해야 합니다.
4.  **질의 유형 분류:** 생성된 QA 쌍은 다음 4가지 핵심 유형 중 하나로 분류되어야 합니다.
    *   **A. 도메인 정의/목적 (Domain Definition/Goal):** 해당 데이터가 해결하려는 산업 문제와 비즈니스 목적에 관한 질문.
    *   **B. 데이터 구조/구성 (Data Structure/Composition):** 데이터의 규모, 포맷, 라벨링 구성 요소(Attributes) 및 분포에 관한 사실적 질문.
    *   **C. AI 모델/임무 (AI Model/Task):** 적용 알고리즘, AI 임무 정의, 예측 목표, 학습 조건 및 성능 지표에 관한 질문.
    *   **D. 품질/공정 관리 (Quality/Process Control):** 데이터 획득/가공/검수 절차, 라벨링 기준, 품질 관리 기준(예: mAP, F1-score 목표치)에 관한 질문.
5.  **출처 표기:** 응답의 모든 문장은 [i] 형식으로 원본 문서의 출처(Source Index)를 명확하
```

```
게 표기해야 합니다.
6.  **최종 통계:** 생성된 모든 QA 쌍을 대상으로, 사용된 **A, B, C, D 유형의 최종 횟수와 비율**을 정리해야 합니다.

**[출력 형식]**
다음 구조를 따라 보고서를 생성하십시오.

### **보고서 제목: [도메인 명] LLM 파인튜닝용 QA 데이터셋 구축 보고서**

#### **I. 논리적 데이터 그룹별 QA 샘플**

| 데이터셋 명 (논리적 그룹) | 유형 | 질의 (Query) | 응답 (Answer) | 출처 |
| :--- | :--- | :--- | :--- | :--- |
| [그룹 1] | [유형] | [질문 1] | [응답 1] | [출처] |
| [그룹 1] | [유형] | [질문 2] | [응답 2] | [출처] |
| [그룹 2] | [유형] | [질문 1] | [응답 1] | [출처] |
| ... | ... | ... | ... | ... |

#### **II. 질의-응답 유형 최종 통계**

| 질의 유형 | 정의 | 사용 횟수 | 비율 |
| :--- | :--- | :--- | :--- |
| A. 도메인 정의/목적 | ... | [횟수] | [비율]% |
| B. 데이터 구조/구성 | ... | [횟수] | [비율]% |
| C. AI 모델/임무 | ... | [횟수] | [비율]% |
| D. 품질/공정 관리 | ... | [횟수] | [비율]% |
| **총합** | | [총 횟수] | 100.0% |
```

Pebblous

> Pebblous Makes Data Tangible

contact@pebblous.ai