Pebblous

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

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

1. 서론 (Introduction)

본 보고서는 페블러스(Pebblus)의 AADS(Agentic AI Data Scientist) 과제의 일환으로, 로봇 분야의 깊이 있는 전문 지식을 갖춘 커스텀 대규모 언어 모델(LLM)을 개발하기 위해 고품질 질의-응답(QA) 데이터셋을 구축하는 전 과정을 기술하는 것을 목적으로 합니다. 이 과정은 로봇 기술의 복잡성과 전문성을 LLM이 효과적으로 학습하고 활용할 수 있도록 지원하는 핵심 단계입니다.

로봇 기술 관련 기술 문서로부터 체계적인 질의-응답 쌍을 생성하는 것은 전략적으로 매우 중요한 가치를 지닙니다. 이는 단순히 정보를 요약하거나 추출하는 것을 넘어, LLM이 특정 도메인의 복잡한 맥락과 기술적 뉘앙스를 깊이 있게 이해하도록 훈련시키는 과정입니다. 이처럼 정교하게 구축된 QA 데이터셋은 LLM의 도메인 이해도, 답변의 정확성, 그리고 최종적인 신뢰성을 높이는 결정적인 요소로 작용하며, 데이터 과학 분야에 특화된 LLM 파인튜닝의 견고한 기반이 됩니다.

본 보고서는 먼저 데이터의 거시적 청사진인 13개 데이터셋 그룹을 정의하고, 미시적 정보 추출을 위한 4가지 질의 유형 프레임워크를 제시합니다. 이 체계적인 접근법을 바탕으로 생성된 QA 샘플을 통해, 본 데이터셋이 어떻게 로봇 도메인 지식을 구조화하고 LLM의 추론 능력을 극대화하도록 설계되었는지 명확히 증명할 것입니다.

2. 로봇 지능 데이터셋 그룹 및 원천 문서 요약

본 섹션에서는 방대하고 다양한 로봇 기술 문서들을 체계적으로 분석하고 분류하기 위해 13개의 핵심 데이터셋 그룹을 정의한 과정을 설명합니다. 이 분류 작업은 광범위한 로봇 기술 영역을 포괄적으로 다루고, LLM 학습을 위한 구조화된 지식 기반을 마련하는 첫 번째 단계입니다. 각 그룹은 특정 로봇 기술, 임무, 또는 환경 데이터를 대표하며, 이를 통해 LLM이 편향되지 않고 균형 잡힌 지식을 습득할 수 있도록 설계되었습니다.

아래 표는 13개 데이터셋 그룹의 명칭과 핵심 내용, 그리고 해당 정보의 출처가 된 원천 문서를 명시하여 데이터셋의 전체적인 범위와 구성을 요약합니다.

No.	데이터셋 그룹명	주요 내용	원천 문서
1	3D 스캔 객체 데이터	가정, 산업, 물류 등 다양한 환경의 객체를 3D 스캐너로 디지털화하여 객체의 원본 형태와 시각 정보를 포함하는 데이터	가려진 객체 추론 데이터
2	다중 객체 가림 환경 데이터	책상, 선반, 박스 등 복잡한 환경에서 여러 객체가 서로를 가리는 상황을 시뮬레이션한 데이터. RGB, 깊이(Depth), 포인트 클라우드(PCD) 포함	가려진 객체 추론 데이터
3	6D 객체 자세 추정 데이터	로봇이 객체를 정밀하게 조작하기 위해 필요한 3D 위치(x,y,z)와 3D 회전(R) 값을 포함하는 데이터	가려진 객체 추론 데이터
4	로봇-객체 파지 데이터	UR5, Panda 등 다양한 로봇 팔과 그리퍼를 이용해 객체를 파지(grasping)하는 과정을 기록한 데이터	가려진 객체 추론 데이터
5	사람-객체 파지 데이터	사람이 일상적인 물체를 잡는 동작을 기록하여 로봇이 인간의 파지 방식을 학습할 수 있도록 지원하는 데이터	가려진 객체 추론 데이터, 손·팔 협조에 의한 파지-조작 동작 데이터
6	로봇 핸드 객체 특성 데이터	200종의 가정용 물품에 대해 로봇 핸드가 쥐기, 돌리기, 흔들기 등 5가지 임무를 수행하며 얻는 시계열 및 물리량 데이터	로봇 핸드용 객체 특성 식별 데이터
7	비도로 환경 주행 데이터	배송 로봇이 인도, 골목, 공원 등 일반 도로가 아닌 환경에서 자율주행하기 위한 2D 이미지 및 3D LiDAR 센서 데이터	배송로봇 비도로 운행 데이터
8	실내 다중 이용시설 주행 데이터	식당, 전시장, 체육시설 등 복잡한 실내 공공장소에서 4족보행 및 바퀴형 로봇 관점으로 수집된 주행 데이터	로봇 관점 주행 영상 (고도화) 데이터
9	SLAM 및 경로 추정 데이터	로봇이 자신의 위치를 파악하고 동시에 지도를 작성(SLAM)하는 데 사용되는 LiDAR 및 IMU 센서 데이터	로봇 관점 주행 영상 (고도화) 데이터
10	사람 행동 인식 데이터	키오스크 등 자동 서비스 시스템을 사용하는 사람의 행동(탐색, 사용, 종료)을 비디오로 기록하고 사용자의 메타데이터를 결합한 데이터	사람 행동 인식 로봇 자율 행동 데이터

11	손/팔 협조 조작 데이터	사람이 특정 과업(문 열기, 버튼 누르기 등)을 수행할 때 손과 팔의 협응 동작을 기록한 멀티모달 데이터(영상, 손 관절 좌표, 힘 센서)	손·팔 협조에 의한 파지-조작 동작 데이터
12	서비스 로봇 상태 및 운영 데이터	안내, 배송, 청소 등 다양한 서비스 로봇의 실시간 상태(위치, 배터리, 작업 현황)를 기록한 시계열 텍스트(JSON) 데이터	실내공간 유지관리 서비스 로봇 데이터
13	로봇 에러 및 예방정비 데이터	서비스 로봇 운영 중 발생하는 에러(장애물, 충돌, 네트워크 등)의 상태와 원인을 라벨링한 데이터	실내공간 유지관리 서비스 로봇 데이터

위 표는 LLM이 학습할 지식의 전체적인 범위와 각 정보의 출처를 명확히 보여줌으로써, 이어질 QA 생성 과정의 신뢰성과 체계성을 뒷받침합니다.

3. QA 데이터셋 질의 유형 정의

고품질의 원천 문서로부터 일관성 있고 깊이 있는 정보를 효과적으로 추출하기 위해, 본 과제에서는 4가지 핵심 질의 유형을 정의했습니다. 이 표준화된 프레임워크는 각 데이터셋이 가진 기술적, 사업적, 관리적 측면을 다각도로 분석하고, 이를 통해 LLM이 특정 측면에 편향되지 않은 균형 잡힌 전문 지식을 학습하도록 유도하는 중요한 방법론입니다.

4가지 핵심 질의 유형은 다음과 같습니다.

1. **A. 도메인 정의/목적:** 해당 데이터셋이 해결하려는 산업 문제, 비즈니스 목적, 최종 활용 서비스 등 도메인 맥락과 목표에 관한 질문입니다. 이를 통해 LLM은 데이터의 **전략적 가치와 비즈니스 목적**을 학습합니다.
2. **B. 데이터 구조/구성:** 데이터 규모, 파일 포맷, 라벨링 구성 요소(Attributes), 메타데이터 항목, 데이터 분포 등 데이터셋의 물리적/논리적 구조에 관한 질문입니다. LLM은 이 유형을 통해 데이터의 **논리적 스키마와 물리적 레이아웃**을 이해합니다.
3. **C. AI 모델/임무:** 적용된 AI 알고리즘, 학습 임무(Task) 정의, 모델 선정 사유, 예측 목표 및 성능 지표 등 AI 기술 적용 전략에 관한 질문입니다. 이를 통해 LLM은 데이터의 **모델링 임무(Task)와 기술적 활용 방안**을 파악합니다.
4. **D. 품질/공정 관리:** 데이터 획득, 정제, 가공(라벨링) 기준, 검수 절차 및 품질 관리 지표 등 데이터 라이프사이클 관리에 관한 질문입니다. 이 유형은 데이터의 **생성 이력(Provenance)과 품질 보증 기준**에 대한 정보를 제공합니다.

이러한 체계적인 질의 유형 접근 방식은 다음 장에서 제시될 구체적인 QA 샘플 생성의 논리적 기반이 되며, 생성된 데이터셋의 품질과 일관성을 보장하는 핵심적인 역할을 합니다.

4. 그룹별 QA 데이터셋 생성 샘플

이 섹션에서는 앞서 정의한 13개의 데이터셋 그룹과 4개의 질의 유형을 실제로 적용하여 생성한 QA 샘플을 제시합니다. 각 샘플은 원천 기술 문서의 핵심 정보를 바탕으로 구성되었으며, 로봇 분야 LLM 파인튜닝에 직접 사용될 데이터의 구체적인 품질과 형태를 명시하는 실질적인 예시입니다.

1. 3D 스캔 객체 데이터

유형	질문 (Question)	답변 (Answer)
A	'가려진 객체 추론 데이터' 구축의 궁극적인 파급 효과는 무엇인가?	세계 최다, 최고 품질의 가려진 물체 데이터셋 구축을 통해 관련 연구를 활성화하고, 다양한 실제 환경으로 로봇 비전 알고리즘을 확장하는 것입니다. 또한, 로봇팔의 물체 파지, 조작, 이송, 배치를 활용하여 물류 현장 등 관련 산업 분야를 활성화하는 효과를 기대합니다.
B	3D 스캔 데이터 수집 시 Artec 3D Leo 툴을 활용하여 생성되는 원시데이터의 포맷은 무엇인가?	Artec 3D Leo 툴은 RGB-D 물체 3D 스캔 원시데이터와 카메라 파라미터를 메쉬(mesh) 데이터인 'obj' 포맷으로 병합하여 생성합니다.
C	'가려진 객체 추론 데이터'를 활용하여 학습할 수 있는 '물체의 6D 자세 예측' 모델 후보에는 어떤 것들이 있는가?	PoseCNN, PVNet, TemplatePose 모델이 있습니다. PoseCNN은 로봇 환경 내 대표적인 자세 추정 모델이며, PVNet은 가려진 환경에서의 성능을 높이기 위해 제안되었습니다.
D	3D 스캐닝 데이터 처리 과정에서 데이터 보정은 어떤 소프트웨어를 사용하여 이루어지는가?	Artec 3D 스캐너로 물체를 스캐닝한 후, Artec Studio SW를 활용하여 데이터를 보정합니다. 이 소프트웨어는 자체적으로 RGB-D 병합 및 후처리 가공 툴을 통해 메시 및 포인트 클라우드 획득을 지원합니다.

2. 다중 객체 가림 환경 데이터

유형	질문 (Question)	답변 (Answer)
A	'다수 물체 가림 데이터'는 어떤 실제 로봇 환경을 대표하도록 구성되었는가?	로봇이 마주할 수 있는 대표적인 3가지 환경인 책상, 선반, 박스 환경을 구성하여 데이터를 수집했습니다. 특히 IKEA 가구를 활용하여 전 세계 연구자가 동일한 환경을 쉽게 구축할 수 있도록 했습니다.
B	'다수 물체 가림 데이터' 세트의 원시데이터는 어떤 정보와	원시데이터는 scene 정보(scene 장소 종류, scene ID 등)와 물체 종류 정보(semantic class, instance class,

	매칭되는가?	object id 등)와 매칭됩니다.
C	가려진 영역을 포함하여 객체를 분할하는 '아모달 인스턴스 분할'을 위해 고려된 AI 모델은 무엇인가?	ORCNN, ASN, UOAIS-Net 모델이 고려되었습니다. ORCNN은 최초로 제안된 모델이며, ASN과 UOAIS-Net은 각각 2020년과 2022년 기준 최고 성능을 보인 모델입니다.
D	가상 데이터의 라벨링 정제는 어떤 도구를 사용하여 수행되었는가?	'bop_toolkit'을 자체적으로 수정 및 개량한 가상 데이터 정제 프로그램 SW를 사용하여 라벨링 정제를 수행했습니다. 이 프로그램을 통해 GT(Ground Truth) 및 JSON 파일을 자동으로 생성합니다.

3. 6D 객체 자세 추정 데이터

유형	질문 (Question)	답변 (Answer)
A	6D 객체 자세 예측 기술의 주요 응용 서비스는 무엇인가?	펙인홀(peg-in-hole), 물품 조립 등 물체의 정밀 인식 및 정밀 조작이 필요한 자동화 시스템에 응용됩니다. 또한 물류나 가정 환경 내 서비스 로봇의 물체 인식 및 순서 추론에도 활용될 수 있습니다.
B	6D 객체 자세 예측의 결과물인 3D 변위(Translation)와 3D 회전(Rotation) 값은 각각 어떤 형식으로 표현되는가?	3D 변위(T)는 x, y, z 좌표로 표현되고, 3D 회전(R)은 쿼터니언(quaternion) 형식인 x, y, z, w 값으로 표현됩니다.
C	2019년 기준 BOP 데이터셋에서 최고 성능을 보인 6D 물체 자세 추정 모델은 무엇이며, 그 특징은 무엇인가?	Peng 등이 제안한 PVNet 모델입니다. 이 모델은 물체별 Keypoint Vector Field 예측을 통해 가려진 물체의 자세 추정 성능을 향상시킨 특징이 있습니다.
D	6D 객체 자세 추정 데이터의 원시데이터 생성과 보정은 어떤 도구를 통해 이루어지며, 품질을 어떻게 확보하는가?	Artec 3D Leo 툴을 사용해 RGB-D 스캔 원시데이터와 카메라 파라미터를 메쉬(obj) 포맷으로 병합하여 생성하고, Artec Studio SW의 자체 툴을 활용해 데이터를 보정하여 고품질의 메시 및 포인트 클라우드를 획득합니다.

4. 로봇-객체 파지 데이터

유형	질문 (Question)	답변 (Answer)
		먼저 물체의 가시 영역, 가려진 영역, 가려짐 여부를 인식하

A	로봇의 가려진 물체 파지 순서 계획은 어떤 과정을 통해 이루어지는가?	고, 목표 물체가 가려지지 않을 때까지 인접 물체를 순서대로 파지하여 제거한 후, 최종적으로 목표 물체를 파지합니다.
B	로봇-물체 파지 데이터는 어떤 로봇팔(Robot Arm)과 그리퍼(Gripper) 조합으로 구성되어 있는가?	로봇팔은 UR5와 Panda를 사용하며, 그리퍼는 Robotiq 2f, Robotiq 3f, Allegro, Qb_hand, Suction, RG_2, Panda_gripper, Delto_3f 등 다양한 종류를 조합하여 데이터를 구성합니다.
C	로봇 환경에서 6D 자세 추정을 위해 제안된 CNN 기반 모델은 무엇이며, 어떻게 자세를 예측하는가?	Xiang 등이 제안한 PoseCNN 모델입니다. 이 모델은 물체별 Class, Position(위치), Rotation(회전)에 대한 Regression(회귀)을 수행하여 자세를 예측합니다.
D	로봇-물체 파지 데이터 수집 시, 데이터의 유용성을 높이기 위해 그리퍼(gripper)를 어떻게 구성하였는가?	범용 그리퍼와 국내 그리퍼를 모두 포함하고, 1지부터 5지까지 모든 종류를 확보했습니다. 특히 활용도가 높은 2지와 3지 그리퍼는 각각 3종, 2종을 사용하여 데이터의 유용성을 향상시켰습니다.

5. 사람-객체 파지 데이터

유형	질문 (Question)	답변 (Answer)
A	사람-객체 파지 데이터 수집의 목적은 무엇인가?	사람이 다양한 일상생활 용품을 잡는 방식을 기록하여, 로봇이 인간의 파지 및 조작 방식을 학습하고 모방할 수 있도록 하는 데 목적이 있습니다.
B	사람-객체 파지 데이터는 어떤 손의 상태와 파지하는 물체 개수에 따라 분류되는가?	손의 상태는 왼손, 오른손, 양손으로 분류되며, 파지하는 물체 개수는 1개 또는 2개로 분류되어 데이터가 구성됩니다.
C	사람의 손동작을 분류하고 이해하기 위해 어떤 유형의 AI 모델이 사용되는가?	ST-GCN(Spatial-Temporal Graph Convolutional Network) 기반의 행동 분류 모델이 사용됩니다. 시계열적, 공간적 특징을 그래프 형태로 학습하여 동작 클래스를 분류합니다.
D	사람-객체 파지 데이터 수집 시, 실험 참여자는 어떻게 구성되었는가?	손 모양과 크기의 다양성을 고려하여, 성별과 연령에 따라 다양한 실험자 20명을 모집하여 데이터를 수집했습니다.

6. 로봇 핸드 객체 특성 데이터

유형	질문 (Question)	답변 (Answer)
A	'로봇 핸드용 객체 특성 식별 데이터' 구축의 필요성은 무엇인가?	물체를 정밀하게 인식하고 조작하기 위해 시각 정보, 물리량, 조작 시 발생하는 시계열 데이터를 빠르게 획득하여 로봇의 지능 향상에 기여하기 위해 필요합니다.
B	'로봇 핸드용 객체 특성 식별 데이터'에서 하나의 객체에 대해 어떤 종류의 데이터가 구축되는가?	영상 데이터(Hi-RGB, Low-RGB, RGB-D), 3차원 포인트 클라우드 메쉬, 물리량(무게, 크기, 재질), 5가지 임무 수행 시 발생하는 시계열 데이터(촉감, 온도, 역감, 사운드)가 구축됩니다.
C	로봇 핸드가 객체를 안정적으로 파지할 위치(grasping point)를 예측하기 위해 어떤 AI 모델이 사용되는가?	CNN 기반의 파지점 탐색 알고리즘이 사용됩니다. 영상 내 파지 가능한 위치의 중심점, 크기, 기울기를 라벨링한 데이터를 학습하여 최적의 파지점을 예측합니다.
D	로봇 핸드 임무 데이터 수집 시, 힘의 단계를 어떻게 나누어 데이터를 수집하는가?	힘의 단계를 5단계(힘단계 0~4)로 나누고, 각 단계별로 10회씩 임무를 수행하여 다양한 힘 조건에서의 상호작용 데이터를 확보합니다.

7. 비도로 환경 주행 데이터

유형	질문 (Question)	답변 (Answer)
A	'배송로봇 비도로 운행데이터'는 어떤 사회적 배경 하에 구축되었는가?	비대면 거래 증가로 인한 택배 물량 급증과 인력 부족, 배달앱 수수료 부담 문제를 해결하기 위한 대안으로 배송 로봇의 자율주행 기술 개발을 지원하기 위해 구축되었습니다.
B	'배송로봇 비도로 운행데이터'의 2D 이미지 데이터는 어떤 클래스들을 대상으로 세그멘테이션 가공을 하는가?	총 22종의 클래스를 대상으로 가공하며, 주요 클래스로 승용차, 보행자, 도로, 인도, 횡단도로, 건물, 초목 등이 포함됩니다.
C	2D 이미지 데이터를 기반으로 주행가능영역(Drivable Area)을 인식하기 위해 선정된 모델은 무엇이며, 그 선정 사유는 무엇인가?	ERF-PSPNet 모델이 선정되었습니다. 효율적인 Deep Architecture를 사용하여 LinkNet보다 성능이 높다고 판단되었기 때문입니다.
D	데이터 수집 시 개인정보 보호를 위해 어떤 조치를 취하는가?	수집된 데이터에 등장하는 모든 인물의 얼굴(타원형 블러)과 차량 번호판(자동 블러)에 대해 비식별화 조치를 수행합니다.

8. 실내 다중이용시설 주행 데이터

유형	질문 (Question)	답변 (Answer)
A	'로봇 관점 주행 영상 데이터'는 어떤 분야의 기술 개발에 활용될 수 있는가?	자율주행 및 실시간 데이터 처리 연구, 자율주행 로봇의 센서 및 알고리즘 개발, 안내/청소/운반 로봇의 자율주행 기술 고도화에 활용될 수 있습니다.
B	'로봇 관점 주행 영상 데이터'는 어떤 종류의 센서 데이터로 구성되어 있으며, 각각의 포맷은 무엇인가?	RGB-D 이미지 데이터(JPG, PNG), LiDAR 데이터(PCD), 6D IMU 센서 데이터(CSV)로 구성됩니다.
C	이 데이터셋의 3D 객체 검출 유효성 검증을 위해 사용된 학습 알고리즘은 무엇이며, 어떤 특징이 있는가?	FocalsConv (OpenPCDet) 알고리즘이 사용되었습니다. 2D 이미지의 RGB feature와 LiDAR feature를 결합하고 Depth map을 추가로 사용하여 3D 객체를 탐지합니다.
D	라벨링 데이터의 품질을 높이기 위해 어떤 검수 절차를 거치는가?	작업자 간 교차 검수(1단계), 관리자 육안 확인 및 오류 재할당(2단계), 최종 검수(3단계)의 총 3단계 검수 절차를 거칩니다.

9. SLAM 및 경로 추정 데이터

유형	질문 (Question)	답변 (Answer)
A	SLAM 기술의 목적은 무엇이며, 이 데이터셋은 어떻게 기여하는가?	SLAM은 로봇이 미지의 환경에서 위치를 추정하고 지도를 작성하는 기술입니다. 이 데이터셋은 LiDAR와 IMU 데이터를 제공하여 SLAM 알고리즘의 유효성을 검증하는 데 사용됩니다.
B	SLAM 성능 유효성 검증에 사용되는 데이터의 파일 형식은 무엇인가?	ROS 환경에서 사용되는 bag 파일 형식을 사용합니다. LiDAR, IMU 등 여러 센서의 시계열 데이터가 타임스탬프와 함께 저장되어 있습니다.
C	SLAM 성능 유효성 검증에 사용된 알고리즘은 무엇이며, 성능 지표는 무엇으로 측정하는가?	Fast-LIO2 알고리즘이 사용되었으며, 성능 지표는 'End to End RMSE(Root Mean Square Error)'로 측정합니다. (목표 0.2m 이내)
D	데이터 수집 시, 여러 센서(RGB-D, LiDAR, Meta data)	소프트웨어 동기화 방식을 통해 다중 센서 데이터 로깅 시 시간적 동기화를 진행하고, 주기 설정에 따라 슬라이싱하여

	간의 시간적 동기화는 어떻게 이루어지는가?	원천 데이터를 생성합니다.

10. 사람 행동 인식 데이터

유형	질문 (Question)	답변 (Answer)
A	'사람 행동 인식 로봇 자율 행동 데이터'는 정보 취약 계층이 겪는 어떤 문제를 해결하는 데 활용될 수 있는가?	시각 장애인이나 노인, 휠체어 사용자가 키오스크 사용 시 겪는 어려움을 해결하는 데 활용됩니다. 로봇이 사용자의 상태를 인식하여 맞춤형 UI나 자동 높이 조절 기능을 제공할 수 있습니다.
B	이 데이터셋의 JSON 파일에는 사용자의 어떤 메타데이터가 포함되는가?	암호화된 사용자 ID, 나이(유소년/청중장년/노년), 키, 성별, 장애 여부 등의 사용자 정보와 서비스 위치, 카메라 높이 등의 환경 정보가 포함됩니다.
C	이 데이터는 어떤 AI 기술 개발에 직접적으로 활용될 수 있는가?	로봇이 사람의 행동(탐색, 사용, 종료)과 특성을 파악하여 맞춤형 서비스를 제공하는 사용자 상태 인식 및 행동 예측 모델 개발에 활용됩니다.
D	데이터 수집 환경은 어떻게 구성되었는가?	교통 시설, 의료 시설, 교육 시설 등 실제 서비스 환경과 유사한 10종의 환경을 선정하여 데이터를 수집했습니다.

11. 손/팔 협조 조작 데이터

유형	질문 (Question)	답변 (Answer)
A	'손·팔 협조에 의한 파지-조작 동작 데이터'는 로봇 기술의 어떤 분야에 활용될 수 있는가?	가사지원 로봇이나 산업용 협동 로봇과 같이 인간과 유사한 동작이 필요한 지능형 로봇 개발, 메타버스 기반의 가상 객체 상호작용 미디어 제작 등에 활용될 수 있습니다.
B	이 데이터셋의 라벨링 데이터(JSON)에는 손동작과 관련하여 어떤 주요 속성들이 포함되는가?	손 관절의 2D/3D 좌표, 손의 접근 방향, 사용된 손가락 수, 손가락 끝의 객체 접점, 손가락 끝 힘 데이터 등이 포함됩니다.
C	사람의 손동작 분류를 위해 어떤 AI 모델이 사용되며, 이 모델은 데이터를 어떻게 처리하는가?	ST-GCN 모델이 사용됩니다. 손과 팔 관절 데이터를 그래프 형태로 표현하고 관절 간의 연결성과 시간적 변화를 학습하여 동작을 분류합니다.
		스마트 조명을 사용하여 밝기와 색상을 조절하며 총 10가

D	손동작 데이터의 다양성을 확보하기 위해 조명 조건은 어떻게 설정되었는가?	지의 서로 다른 조명 조건을 설정하여 데이터를 수집했습니다.

12. 서비스 로봇 상태 및 운영 데이터

유형	질문 (Question)	답변 (Answer)
A	'실내공간 유지관리 서비스 로봇 데이터'의 구축 목적은 무엇인가?	서비스 로봇의 상태 및 운영 데이터를 기반으로 고장을 사전에 예측하고 선제적 유지보수를 수행할 수 있는 데이터 분석 시스템 및 학습 모델 개발을 목적으로 합니다.
B	이 데이터셋의 JSON 파일에 포함된 'deviceData' 객체는 로봇의 어떤 상태 정보들을 담고 있는가?	로봇의 주 상태, 배터리 잔량, 충전 여부, 장애물 감지 정도, 충돌 정도, 긴급정지 버튼 상태, 위치 정보, 총 운행 정보 등을 포함합니다.
C	로봇의 운영 상태 예측을 위해 어떤 AI 모델을 사용하며, 그 이유는 무엇인가?	트랜스포머 기반의 'LLama3.2-3B-instruct' 모델을 사용합니다. 시퀀스 데이터를 처리하여 다음 상태를 예측하는 데 최적화되어 있고, 성능과 효율성의 균형을 맞춘 모델이기 때문입니다.
D	데이터 정제 과정에서 수행되는 주요 작업은 무엇인가?	원천 데이터 포맷 정의 및 필드 정제, 로봇 ID 기준 데이터 분류, 운영 시간 외 데이터 필터링, 맵핑, JSON 변환 등이 수행됩니다.

13. 로봇 에러 및 예방정비 데이터

유형	질문 (Question)	답변 (Answer)
A	로봇 에러 및 예방정비 데이터는 서비스 로봇 운영 기업에게 어떤 가치를 제공하는가?	로봇 비정상 상태의 원인을 분석하고 고장을 사전에 예측하는 AI 알고리즘 개발에 활용되어, 로봇의 다운타임을 최소화하고 운영 효율성을 극대화합니다.
B	로봇의 비정상 상태는 어떤 에러 코드(errorCode)로 분류되는가?	장애물 에러, 충돌감지 에러, 배터리 에러, 비상정지 에러, 엘리베이터 에러, 자동문 에러, 네트워크 에러, 로봇 SW 에러 등 총 8가지로 분류됩니다.
C	로봇의 에러 발생 여부를 예측하기 위해 어떤 학습 모델이 사용되었으며, 성능 지표는 무엇인가?	Decision Tree 모델이 사용되었으며, 성능 지표는 '정확도(Accuracy)'입니다.

D	데이터 가공(라벨링) 과정에서 로봇의 정상/비정상 상태는 어떤 기준으로 판단하는가?	로봇이 할당된 작업을 정상적으로 수행하고 있는지를 기준으로 판단하며, 비정상 상태일 경우 8가지 에러 항목 중 하나에 할당하여 라벨링합니다.

5. 결론 (Conclusion)

본 보고서는 다양한 로봇 기술 문서로부터 13개의 데이터셋 그룹을 체계적으로 정의하고, 4가지 표준화된 질의 유형을 적용하여 LLM 파인튜닝을 위한 고품질 QA 데이터셋을 성공적으로 구축한 과정을 상세히 기술했습니다. 이 접근법을 통해 방대한 로봇 기술 지식을 구조화하고, LLM이 학습할 수 있는 일관되고 신뢰성 높은 정보 자산을 마련했습니다.

본 데이터셋이 갖는 전략적 의미는 매우 큽니다. 이는 AADS 과제의 핵심 자산으로서, 로봇 분야의 복잡한 맥락과 기술적 뉘앙스를 깊이 있게 이해하는 전문 LLM 개발의 초석이 될 것입니다. 구축된 데이터셋은 단순 정보 검색을 넘어, 로봇의 비정상 상태 원인을 다각적으로 추론하고, 복잡한 가림 환경에서 최적의 파지 순서를 계획하며, 사용자의 행동을 예측하여 선제적 서비스를 제공하는 등, 진정한 의미의 **에이전트(Agentic) AI 데이터 과학자**를 구현하는 데 결정적으로 기여할 것입니다.

향후 계획으로는, 본 보고서에서 제시된 QA 샘플을 바탕으로 전체 데이터셋을 대규모로 확장하는 작업을 진행할 것입니다. 완성된 데이터셋을 활용하여 본격적인 LLM 파인튜닝 실험을 시작하고, 로봇 도메인에서의 성능을 검증할 예정입니다. 또한, 지속적인 데이터 품질 관리와 모델 성능 평가를 통해 로봇 도메인에 특화된 AI의 성능을 지속적으로 고도화해 나갈 것입니다.

Pebblous

> Pebblous Makes Data Tangible

contact@pebblous.ai