사람이 수집한 AI 데이터
이를 활용해야 하는 5가지 이유

모델을 학습시키기 위해서는 합성 데이터든 사람이 수집한 데이터든 또는 이 둘의 조합이든 데이터가 필요합니다. 왜 사람이 수집한 데이터가(전부는 아니더라도 대부분일 때) 더 강력하고 안정적인 모델 성능을 얻을 수 있는지 그 이유를 알아보세요.

1 품질

고품질 데이터는 일반적으로 더 정확하고 최신 상태이며, 깨끗하고 일관성 있으며, 체계적이고 풍부한 맥락 정보를 담고 있어 모델 성능을 향상시킵니다. 이러한 데이터로 학습된 모델은 언어와 관련 주제를 보다 섬세하게 이해할 수 있으므로 더 높은 수준의 성능을 발휘할 수 있습니다.

예시

라이온브리지는 한 이러닝 솔루션 제공업체를 지원하여 기계로 트랜스크립션을 삽입한 300건 이상의 동영상에 품질 문제가 있는지 검토했습니다. 검토자들은 AI가 전사한 자막을 수정하여 동영상 트랜스크립션의 정확도를 더욱 개선했습니다.

2 다양성

전 세계 다양한 기여자 기반을 보유한 공급업체는 보다 광범위한 실제 상황을 망라하며 편향 없는 다양한 관점을 반영한 데이터를 제공합니다.

예시

라이온브리지는 한 대형 기술기업을 지원하기 위해 Aurora AI Studio™와 당사의 채용팀을 활용해 특정 감정을 표현한 문장을 백만 개 이상 수집했습니다. 여러 언어와 방언을 사용하는 광범위한 인구 집단을 아우르는 참가자들이 녹음에 참여했습니다.

3 속도

우수한 공급업체를 만나면 속도를 크게 높일 수 있습니다. Lionbridge Aurora AI Studio에서는 전 세계 50만 명에 달하는 기여자에게 다가갈 수 있으며, 필요한 경우 당사의 강력한 채용팀을 통해 크라우드 외부에서 기여자를 찾을 수도 있습니다. 며칠 또는 심지어 몇 시간 만에 작업을 수주하고 완료할 수도 있습니다.

예시

라이온브리지는 한 스마트폰 제조업체를 지원하기 위해 Aurora AI Studio를 활용해 8개 언어로 '실제' 대화 샘플을 수집할 수 있도록 지원했으며, 그 결과 대화 1건당 최대 5명이 참여해 20만여 건의 대규모 데이터를 확보할 수 있었습니다. 이 플랫폼 덕분에 모든 대화 데이터를 전달하기까지 4주도 채 걸리지 않았습니다.

4 가격

합성 데이터만으로는 모델 성능이 저하되므로 추가 검증 절차를 거치거나 사람이 수집한 데이터가 필요한 경우가 많습니다. 처음부터 사람이 수집한 데이터를 활용해 비용을 절감하세요.

예시

라이온브리지는 한 온라인 동영상 서비스 제공업체를 지원해 여러 언어로 제작된 다량의 동영상을 영어로 번역했습니다. 번역사는 욕설이 들어 있거나, 모욕적이거나, 혐오 표현이 있는 콘텐츠에 플래그로 표시를 남겼습니다. 이러한 꼼꼼한 검토를 통해 LLM은 혐오 표현 없이 더 빨리 최적화된 방식으로 작동할 수 있었습니다. 추가 데이터 수집이나 검증 작업도 필요하지 않았습니다.

5 개발 프로세스

개발에는 비용이 많이 듭니다. 개발자의 시간과 노동력을 절감하세요. 처음부터 신뢰할 수 있는 인간이 수집한 데이터를 선택하세요.

예시

라이온브리지는 한 AI 개발사를 지원하여 검토자인 사람이 여러 언어에 걸친 광범위한 필수 학습 데이터를 LLM에 입력하고 프롬프트에 대한 가장 적절한 응답을 선택한 후 여러 요소를 기준으로 해당 응답을 평가했습니다.

AI 데이터 수집 서비스에 대해 궁금하시면 지금 바로 Lionbridge AI™에 대해 알아보세요. 자세한 내용은 lionbridge.com에서 확인할 수 있습니다.