LIONBRIDGE

모델 평가를 위한 AI 데이터 분석

평가 범주

대규모 언어 모델(LLM)의 성능을 평가하려면 체계적인 접근 방식을 통해 언어 생성 과정의 다차원적 특성을 포착해야 합니다. 다음 평가 범주는 모델 결과물을 품질, 신뢰성, 사용자 관련성 전반에 걸쳐 종합적으로 평가하기 위한 프레임워크를 제공하기 위해 선정되었습니다. 이 범주를 활용하면 겉보기에 정확한 수준을 넘어 유창성, 완전성, 분야별 용어, 문화적 적합성 등 다양한 측면을 평가할 수 있습니다. 따라서 조직은 모델 결과물이 사용자의 구체적인 기대에 부합하는지, 비즈니스 목표와 일치하는지, 윤리적 기준을 준수하는지 확실하게 평가할 수 있습니다.

정확성

답변의 사실 관계가 정확하며 오류가 없는지 측정합니다.

예시: 법률과 관련된 답변을 할 때 올바른 법 조항을 인용합니다.

완전성

질문의 모든 부분을 충분히 다루는지 확인합니다.

예시: 요약을 요청했을 때 모든 요점을 언급함

관련성

답변이 주제에서 벗어나지 않고 질문에 대한 직접적인 답변인지 평가합니다.

예시: 제품 설명에 관련 없는 정보가 포함되지 않음

일관성

답변이 내적으로 논리적이며 모순이 없는지 확인합니다.

예시: 여러 단락으로 이루어진 답변에서 내적 모순이 없음

유창성

문법적으로 정확하며 흐름이 자연스러운지 평가합니다.

예시: 적절한 문장 구조와 구두점을 사용함

환각 오류(반전)

모델이 사실을 꾸며내지 못하도록 하는지 측정합니다.

(반전 시 환각 오류가 적을수록 점수가 높음)

예시: 존재하지 않는 제품 기능을 꾸며내지 않음

용어

분야별 용어와 전문용어가 적절하게 사용되었는지 평가합니다.

예시: 의료 관련 질문에 답변할 때 정확한 의학 용어를 사용

가독성

대상 독자가 내용을 쉽게 읽고 이해할 수 있는지 측정합니다.

예시: 일반 대중을 고려해 명확하고 간결한 표현을 사용함

문화적 관련성

답변이 대상 독자의 문화적 규범에 적절하며 민감한지 평가합니다.

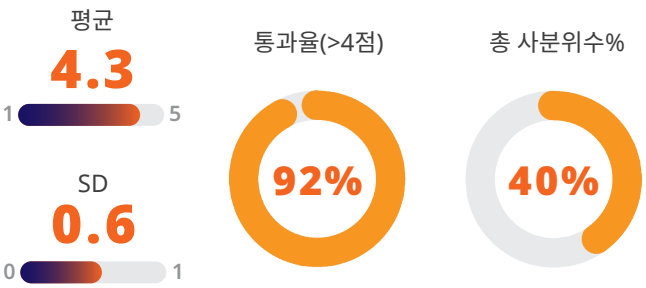
예시: 전 세계에서 통용되기 어려운 관용구나 인용은 사용하지 않음

고객은 5점 리커트 척도를 사용해 이러한 범주에 체계적으로 점수를 부여함으로써 모델이 어떤 부분에서 뛰어난지, 어떤 부분에 미세 조정이나 추가 학습, 사람의 감시가 필요한지 등에 대해 실행 가능한 인사이트를 얻을 수 있습니다. 그에 따라 고객 지원, 콘텐츠 생성, 제품 추천 등 다양한 적용 분야 전반에서 AI가 일관성 있는 양질의 경험을 제공하도록 만들 수 있습니다.

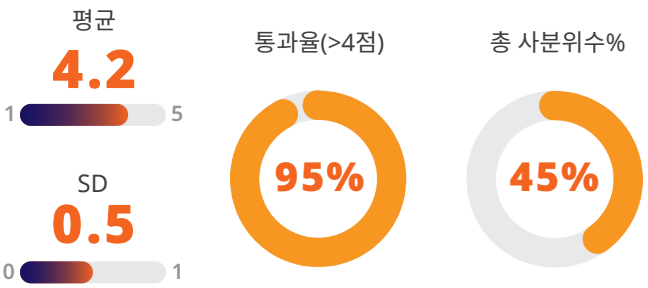
분석 예시 | 보다 심층적인 인사이트 확보

라이온브리지는 AI 학습 프로세스의 품질, 일관성, 효율성에 대한 중요한 인사이트를 얻을 수 있도록 분석을 위한 데이터 서비스를 제공합니다. 이러한 인사이트를 활용하면 데이터의 다양성, 주석 처리자의 동의 수준, 모델의 강점과 약점, 결과물에서의 경향성 등 패턴을 식별하는 데 적용할 수 있으며 AI 시스템이 강력하고 편향되지 않은 높은 성능을 유지할 수 있습니다. 조직은 이 분석 결과를 바탕으로 데이터에 기반한 결정을 내려 모델 학습을 최적화하고, 주석 처리 워크플로를 개선하며, AI 배포를 가속화할 수 있습니다.

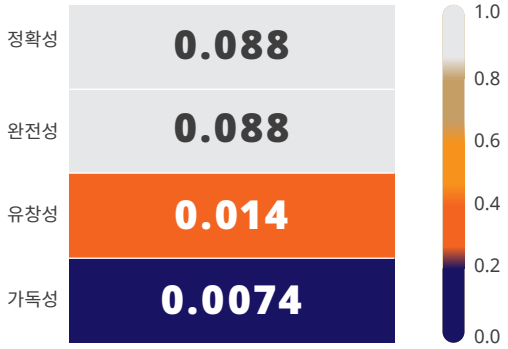
정확성



관련성



평가자 간 동의(플레이스의 카파)



플레이스 카파

평가 범주 간 상관성

자세히 알아보기:
LIONBRIDGE.COM