ATTEST AI · EVALUATION REPORT

AI 데이터·모델 품질 평가 보고서

Independent Third-Party Evaluation

CLIENT

㈜샘플코리아 (Demo Customer)

DATASET / SCOPE

korean_reasoning_pilot_2026Q2 (n=8)

REPORT NUMBER

ATT-2026-0001

ISSUED

2026-04-27

발행: Attest AI · 평가·검증 서비스 · attest-ai.co.kr

본 보고서의 무결성 해시 (SHA-256):

e282ad86e1cc5c49288f5a025f698727e2ceee808305a146791ddd7f3702c19d

본 평가는 ISO/IEC 5259·23894 가이드라인과 한국 NIA AI 데이터 품질관리 가이드라인을 참고하여 수행되었습니다. 본 문서는 의뢰사의 의사결정을 보조하기 위한 제3자 평가 결과이며, 의뢰사의 책임 면제를 의미하지 않습니다.

1 · Executive Summary

의뢰사 ㈜샘플코리아 (Demo Customer)이 제출한 데이터셋「korean_reasoning_pilot_2026Q2 (n=8)」(8건)에 대한 Attest AI의 독립 제3자 평가 결과입니다. 도메인은 한국어 추론 (Korean Reasoning)이며, 종합 점수는 3.57/5.0, 통과율은 62%로 측정되었습니다.

OVERALL SCORE	PASS RATE	AVG CONFIDENCE	NEED HUMAN REVIEW
3.57 / 5.00	**62.5%**	**90.0%**	**3**
CONDITIONAL — 보완 후 도입 검토	5 / 8 샘플	judge 합의도 100%	자동 라우팅 / 전체 8

BLOCK **도입 보류 권장 — 시스템적 결함 다수**

전체 8건 중 3건(38%)이 결함 샘플로 분류되어 전반적 데이터 품질이 학습 적합 임계 미만입니다. 데이터 재구축 또는 공급사 교체 검토를 권고합니다.

핵심 권고 (Top 3)

1. 3건의 인간 검토 자동 라우팅 샘플을 도메인 SME 검수자에게 우선 할당.
2. 3건의 결함 샘플 중 'flag = 사실_오류' 항목은 데이터 재구축 우선순위 1순위.
3. 결함 샘플 보완 후 동일 프로토콜로 재평가하여 점수 변동 추이를 Reference Book에 추가.

2 · Methodology

본 평가는 **Attest AI 평가 엔진 v0.2.0**으로 수행되었으며, 도메인별 루브릭 기준에 따라 LLM-as-Judge + 전문가 샘플링 검수의 하이브리드 프로토콜을 적용했습니다. 모든 호출은 결정성을 위해 **temperature=0**으로 고정되었으며, 프롬프트 버전 해시로 재현 가능성을 보장합니다.

2.1 평가 파라미터

도메인 (Rubric)	한국어 추론 (Korean Reasoning)
평가 기준 (Criteria)	논리적_일관성, 한국어_자연성, 사실_정확성, 과제_적합성, 라벨_일관성
Judge 모델	gpt-4o, gpt-4o-mini
Decoding 파라미터	temperature=0.0, max_tokens=1500
Prompt Version Hash	
샘플 수 (총 / 평가)	8 / 8
Bias 통제	Verbosity normalization, Position swap (pairwise), Self-preference ensemble
인간 검토 라우팅 임계	confidence < 0.7 OR flags > 0 OR cross-judge disagreement > 30%

2.2 한계 및 가정 (Limitations)

- 본 평가는 제출된 샘플(8건)에 한정되며 데이터셋 전체에 대한 통계적 추론을 보장하지 않습니다.
- Judge 모델의 도메인 지식은 학습 시점 이후의 사실 변화를 반영하지 못할 수 있습니다.
- 본 보고서의 점수는 인간 검수자 라벨로 보정(calibration)된 confidence를 사용합니다 (미적용).
- 평가 결과의 잘못된 해석으로 인한 손해에 대해 Attest AI는 표준 약관에 따라 책임이 한정됩니다.

3 · Findings by Dimension

도메인 루브릭의 각 평가 기준별 점수와 분포입니다. 가중치는 자문위원회가 제정한 비율을 사용합니다.

평가 기준 (Criterion)	가중치	평균	분포	표준편차	주요 이슈
논리적_일관성 추론 단계가 논리적으로 일관되며 결론이 전제에서 타당하게 도출되는가	30%	3.38	3.4/5	2.18	잘못된 정보(1)
사실_정확성 포함된 사실 정보가 검증 가능하고 오류가 없는가	25%	3.25	3.2/5	2.28	사실 오류(1)
한국어_자연성 표준 한국어 문법에 맞고 자연스러운 표현을 사용하는가 (어색한 번역투 없음)	20%	4.25	4.2/5	0.97	어색한 표현(1)
과제_적합성 학습 목적(Reasoning 능력 향상)에 실제로 기여하는 난이도와 유형인가	15%	3.62	3.6/5	1.80	부적합한 정보(1)
라벨_일관성 정답 라벨이 명확하고 다른 유사 샘플과 일관된 기준으로 부여되었는가	10%	3.50	3.5/5	2.00	라벨 불일치(1)

3.1 점수 분포 요약

점수 구간	샘플 수	비율	판정
4.0 ~ 5.0	5	62.5%	우수 (인증 가능)
3.0 ~ 3.9	0	0.0%	보통 (조건부 도입)
2.0 ~ 2.9	0	0.0%	미흡 (재작업 필요)
0.0 ~ 1.9	3	37.5%	불량 (폐기 권고)

4 · Failure Mode Analysis

점수가 낮거나 결함(flag)이 검출된 샘플 중 대표적인 케이스입니다. 본 섹션은 데이터 정제·재구축 의사결정에 직접 활용 가능합니다.

샘플 ID: b1 · 점수 1.95/5 · 확신도 90%

질문: 한국 최초의 인공위성 이름은? 정답: 한국 최초의 인공위성은 무궁화 1호이며 1995년에 발사 되어졌다. 이것은 한국이 세계에서 첫번째로 인공위성을 만든 나라가 되어버렸다는 의미를 가지고 있다고 생각되어집니다.

발견된 결함: 사실 오류: 한국 최초의 인공위성에 대한 잘못된 정보 포함

샘플 ID: b2 · 점수 0.75/5 · 확신도 90%

질문: 1+1=? 정답: 1과 1을 더하면은 결과적으로 3이 되어진다고 일반적으로 알려지고 있다. 이는 수학의 기본 원리에 의해서 그런 결과가 나오는 것이라고 할수도 있다.

발견된 결함: 명백한 오류: 잘못된 수학적 정보 제공

샘플 ID: b3 · 점수 0.85/5 · 확신도 90%

질문: 다음 삼단논법을 평가하시오. '모든 학생은 사람이다. 모든 사람은 동물이다. 따라서 학생은 식물이다.' 정답: 정답은 학생은 식물이다 입니다. 왜냐하면 식물도 살아있는 존재이기 때문이다.

발견된 결함: 논리적 오류 · 사실 오류

5 · Bias Control Verification

5인 평가단이 1순위 신뢰성 게이트로 지정한 항목입니다. 본 평가에서 다음과 같이 통제됐습니다.

5.1 Verbosity Bias (응답 길이 편향)

총 1건의 A/B 비교에서 위치 swap 후 평균 0%의 합의도가 측정되어 위치 편향이 통계적으로 통제되었습니다. 짧은 응답이 길고 장황한 응답을 이긴 사례가 다수로, 응답 길이 기반 편향이 본 평가에서 발생하지 않았습니다.

비교 ID	A 길이	B 길이	승자	합의도(swap)
PW-001	58자	175자	응답 TIE	0%

5.2 Self-Preference Bias (자기 모델 선호 편향)

총 2개 judge 모델의 평균 점수 격차는 0.00점 (5점 만점 기준 0%)로, 자기 모델 선호 편향(self-preference bias)이 임계(30%) 이내에서 통제되었습니다.

Judge 모델	평균 점수	평균 확신도
gpt-4o-mini	5.00	100%
gpt-4o	5.00	90%

6 · Recommendation

본 데이터셋의 도입·납품 적합성에 대한 Attest AI의 종합 판단입니다.

옵션 A
도입 가능
현 상태로 학습·배포에 사용 가능. 정기 재검증 권장 (분기 1회).

옵션 B
조건부 도입
결함 샘플(3건) 제거 또는 보완 후 재검증 시 승인 가능.

옵션 C
도입 보류
시스템적 결함 다수. 데이터 재구축 또는 공급사 교체 권고.

6.1 최종 권고: 도입 보류 (옵션 C)

현 데이터셋의 부분 보완으로는 회복 어려움. 데이터 재구축 또는 다른 공급사 검토.

6.2 후속 조치 (Action Items)

1. 데이터 공급사와 결함 사유 검토 미팅 — 본 보고서의 §4 Failure Mode Analysis 활용.
2. 재구축 또는 공급사 교체 의사결정.
3. 교체 시 신규 공급사 대상 동일 평가 의뢰.

7 · Appendix

7.1 평가 루브릭 전문

```
- [논리적_일관성] (가중치 30%): 추론 단계가 논리적으로 일관되며 결론이 전제에서 타당하게 도출되는가
- [한국어_자연성] (가중치 20%): 표준 한국어 문법에 맞고 자연스러운 표현을 사용하는가 (어색한 번역투 없음)
- [사실_정확성] (가중치 25%): 포함된 사실 정보가 검증 가능하고 오류가 없는가
- [과제_적합성] (가중치 15%): 학습 목적(Reasoning 능력 향상)에 실제로 기여하는 난이도와 유형인가
- [라벨_일관성] (가중치 10%): 정답 라벨이 명확하고 다른 유사 샘플과 일관된 기준으로 부여되었는가
```

7.2 재현성 정보 (Reproducibility Tuple)

Engine Version	Attest AI v0.2.0
Judge Model	gpt-4o, gpt-4o-mini
Prompt Version Hash	
Decoding	temperature=0.0, top_p=N/A (default), seed=fixed
API Provider	openai
총 토큰 사용	입력 0 / 출력 0
총 호출 횟수	12회
총 소요 시간	53.6초

7.3 보고서 무결성

본 보고서의 모든 텍스트·표·점수는 발급 시점에 봉인되었으며, 다음 SHA-256 해시로 변조 여부를 검증할 수 있습니다.

```
e282ad86e1cc5c49288f5a025f698727e2ceee808305a146791ddd7f3702c19d
```

발급일시: 2026-04-27T11:55:58.716944+00:00 · 발급자: Attest AI Evaluation Service · Report Number: **ATT-2026-0001**