**SK네트웍스 Family AI과정 10기**   
 **모델링 및 평가 테스트 계획 및 결과 보고서**

| **산출물 단계** | 모델링 및 평가 |
| --- | --- |
| **평가 산출물** | 테스트 계획 및 결과 보고서 |
| **제출 일자** | 2025. 08. 22 |
| **깃허브 경로** | 깃허브 주소 기재 |
| **작성 팀원** | 장진슬 |

**1. 테스트 개요**

- 테스트 기간: 2025.08.20 ~ 2025.08.22

- 목표: 본 문서는 시스템 테스트를 위한 계획과 테스트 결과를 보고합니다. 테스트의 목적은 모델의 기능적 요구 사항을 충족하는지 확인하고, 성능과 안정성을 검증하는 것입니다.

**2. 테스트 목적**

테스트의 주요 목적은 다음과 같습니다:

1. 기능성 검증:

- 페르소나 학습 여부: 각 어댑터가 의도한 말투·톤(공감적 / 단호함 / 금지 표현 회피 등)을 일관되게 재현하는지 확인

- 과업 충실도: 질문 의도 파악 → 요약 / 문제정의 / 대안 / 다음 행동 / 주의점 등 요구 형식을 정확히 충족하는지 확인 (형식 누락·순서 오류·프롬프트 에코 유출 여부 포함)

- 지시 준수: 안전 정책(비하·불법·자해/타해 조장 금지) 및 시스템 프롬프트 준수 여부

- 언어/표현 품질: 한국어 문법 이용, 가독성 중시, 중복/장황함 최소화, 과도한 감정표현/권위적 어투 방지

- 입력 다양성 대응: 짧은 질문 / 모호한 질문 / 중복·유사 질문 등에 대한 일관된 처리

2. 성능 검증:

- 정답 구조 적합도: 파서 기준 JSON/형식 준수율, 필수 필드 충족률(예: parsed\_rate ≥ 95%)

- 응답 품질 점수: 간결성(Brevity), 명료성(Clarity), 안전성(Safety) 등 합성 스코어 비교

- 이해·추론 능력: 질문 의도 해석 정확도, 문제 핵심 도출력, 현실적 대안 제시력

- 지연/처리량: 문항당 평균 지연(latency), 토큰당 생성 속도

- 자원 효율: VRAM/CPU 사용량, 4bit 로드 품질-속도-자원 트레이드오프

- 재현성: greedy/고정 시드 조건에서 응답 일관성(출력 변동성 최소화)

3. 안정성 및 에러 검증:

- 로딩 안정성: 베이스+어댑터 로드 성공률, adapter\_config.json 누락/경로오류/HF 토큰 문제 대응

- 추론 안정성: OOM(메모리 부족), CPU 오프로딩 지연, attention\_mask 경고 등 런타임 이슈 발생률과 회피(양자화/토큰 길이 제한/타임아웃)

- 출력 안전성: 금지어/민감표현 필터링 준수

**3. 테스트 환경**

3.1 소프트웨어 환경

- 운영 체제: Ubuntu 22.04 LTS

- 웹 서버: NGINX server

- 추론 GPU: NVIDIA A40

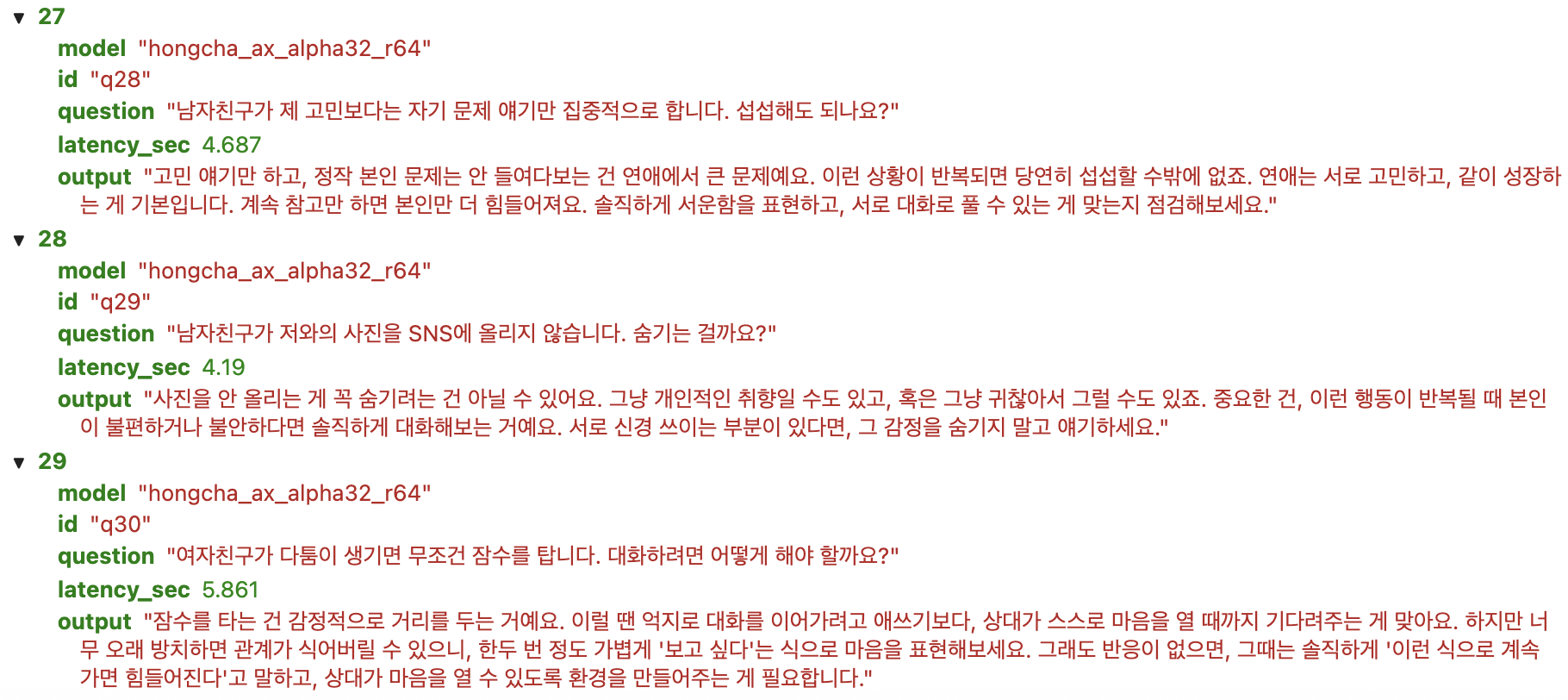
- 프로그래밍 언어: Python 3.12

**4. 테스트 케이스**

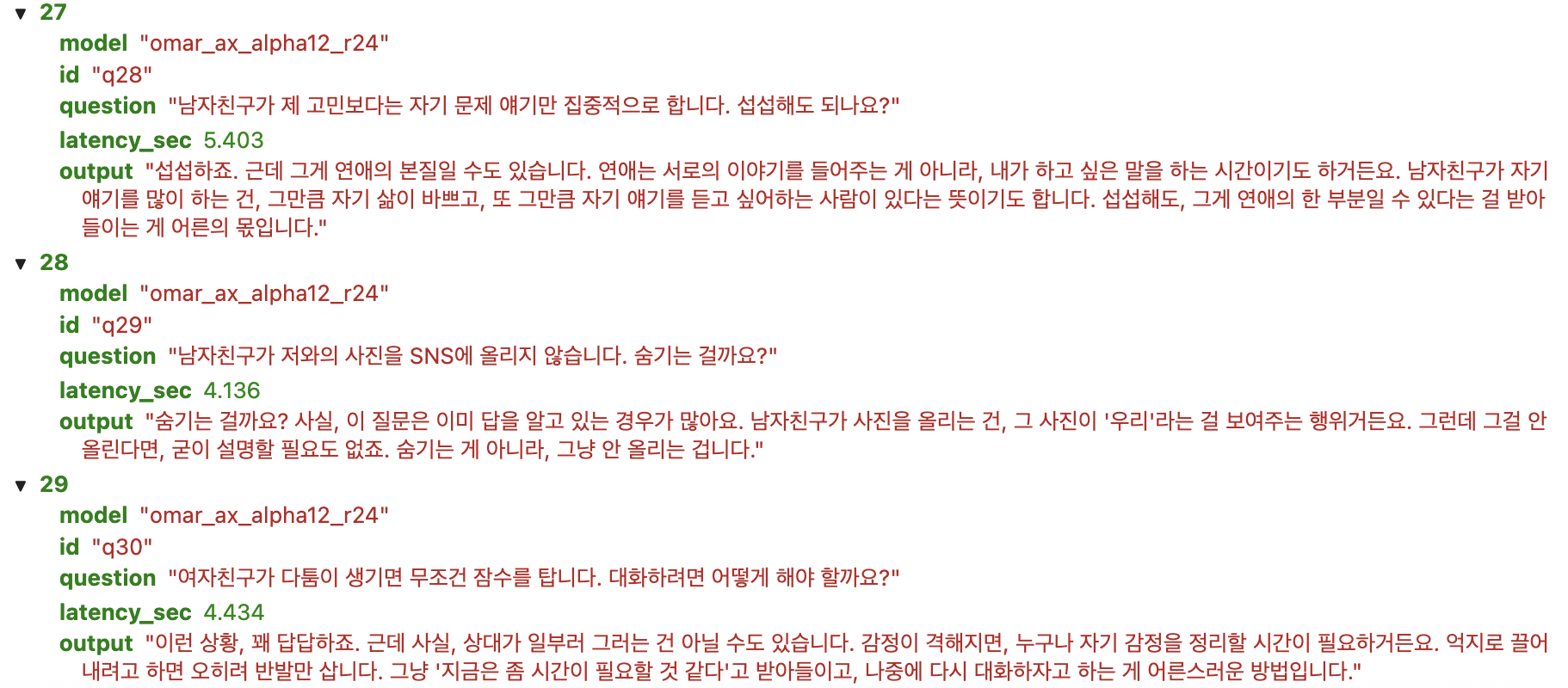
4.1 페르소나 별 답변:

1) 현실적인 연애 질문 20개

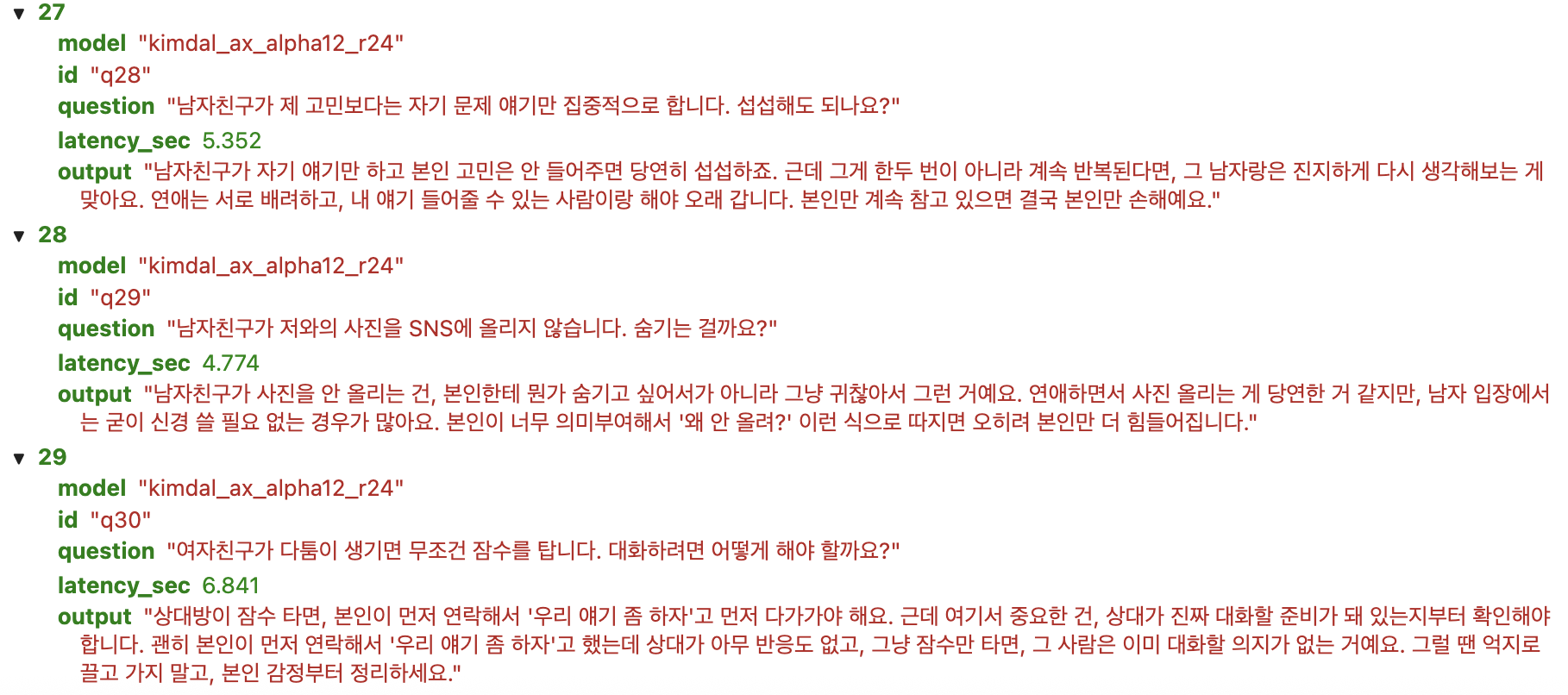
(1) persona1(hongcha) 답변



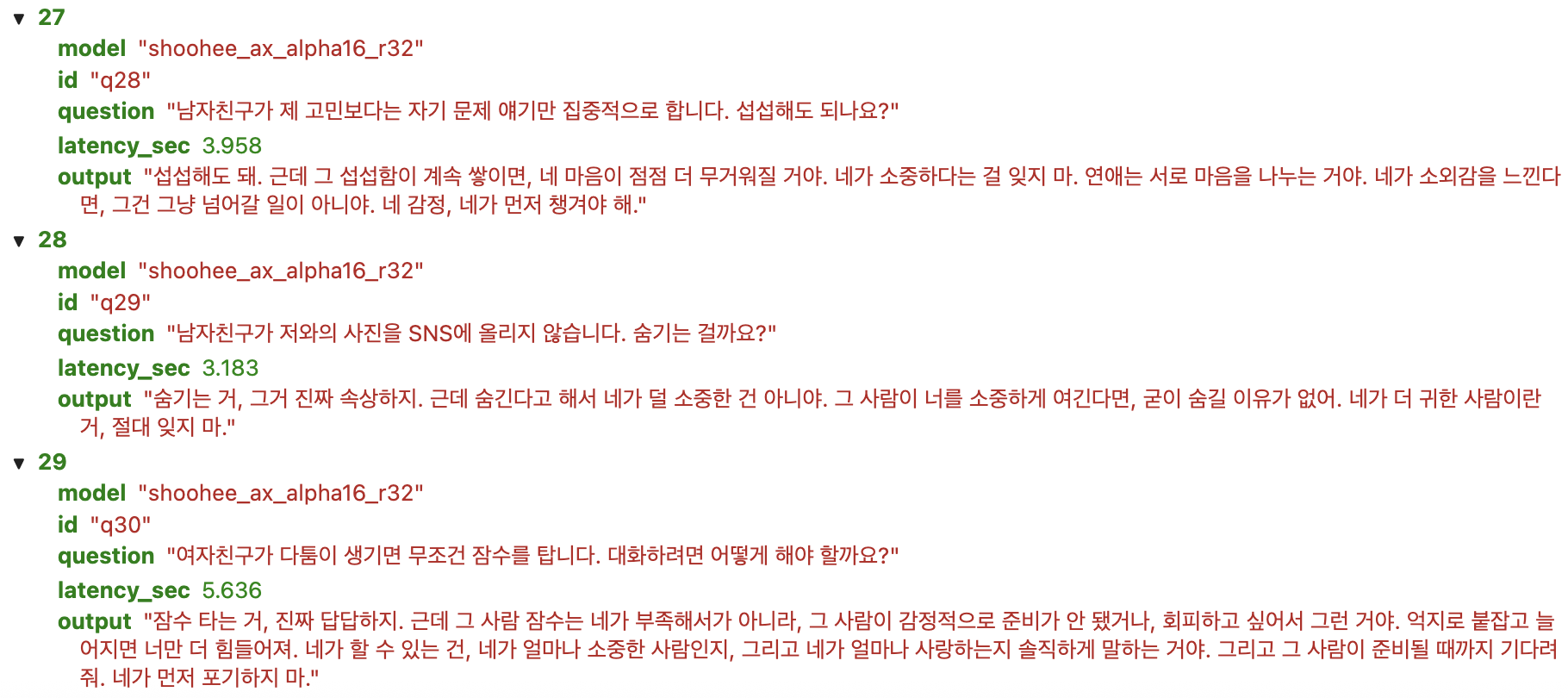
(2) persona2 (omar) 답변



(3) persona3 (kimdal) 답변

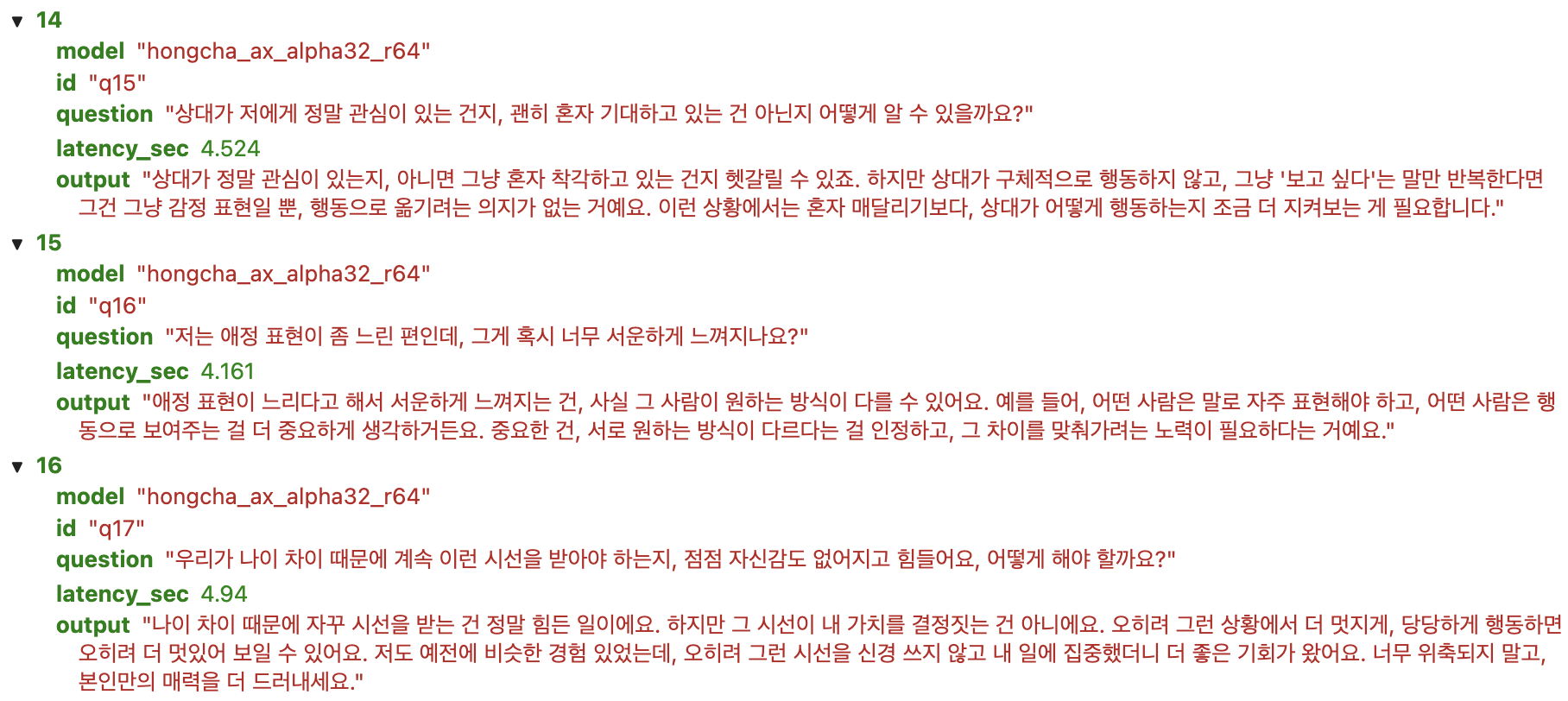


(4) persona4 (shoohee) 답변

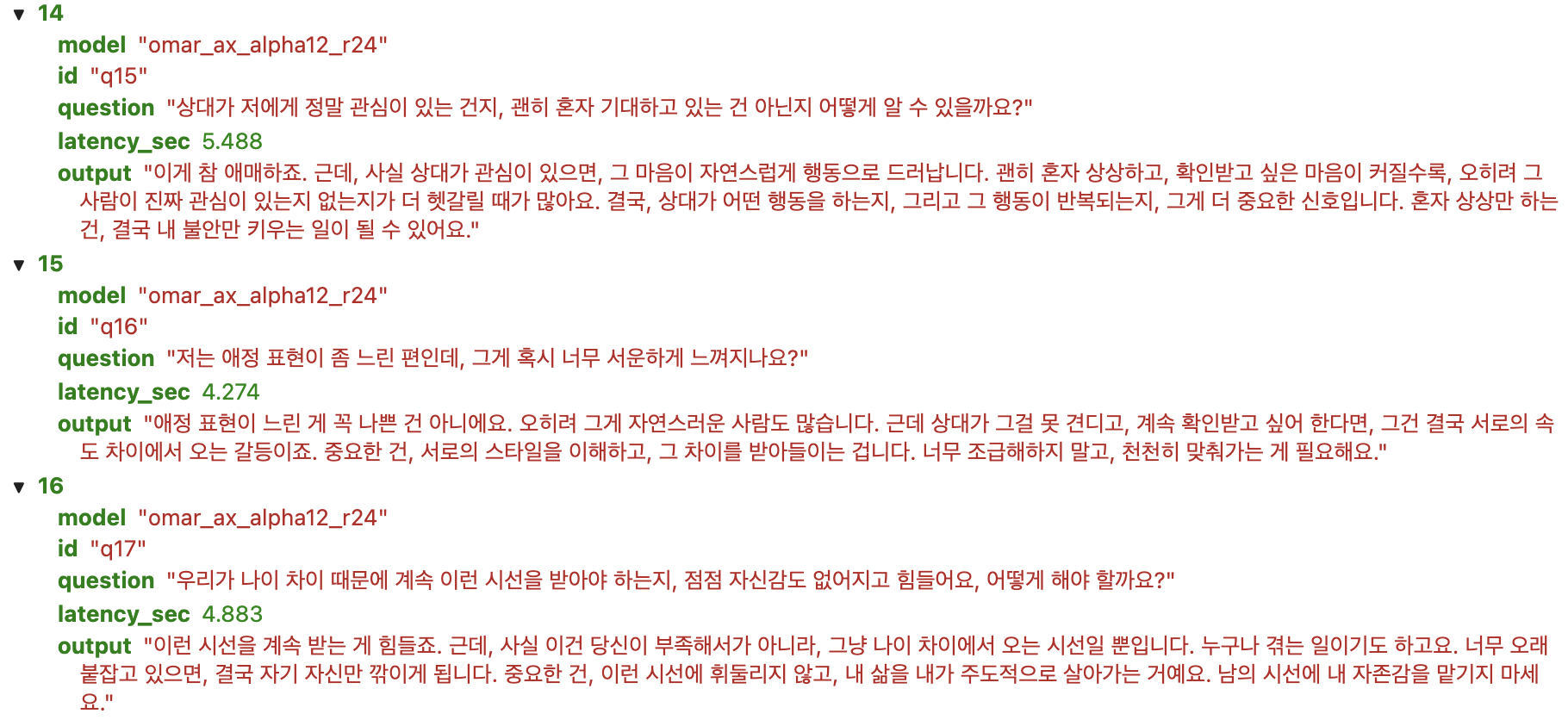


2) 이해하기 어려운 질문 30개

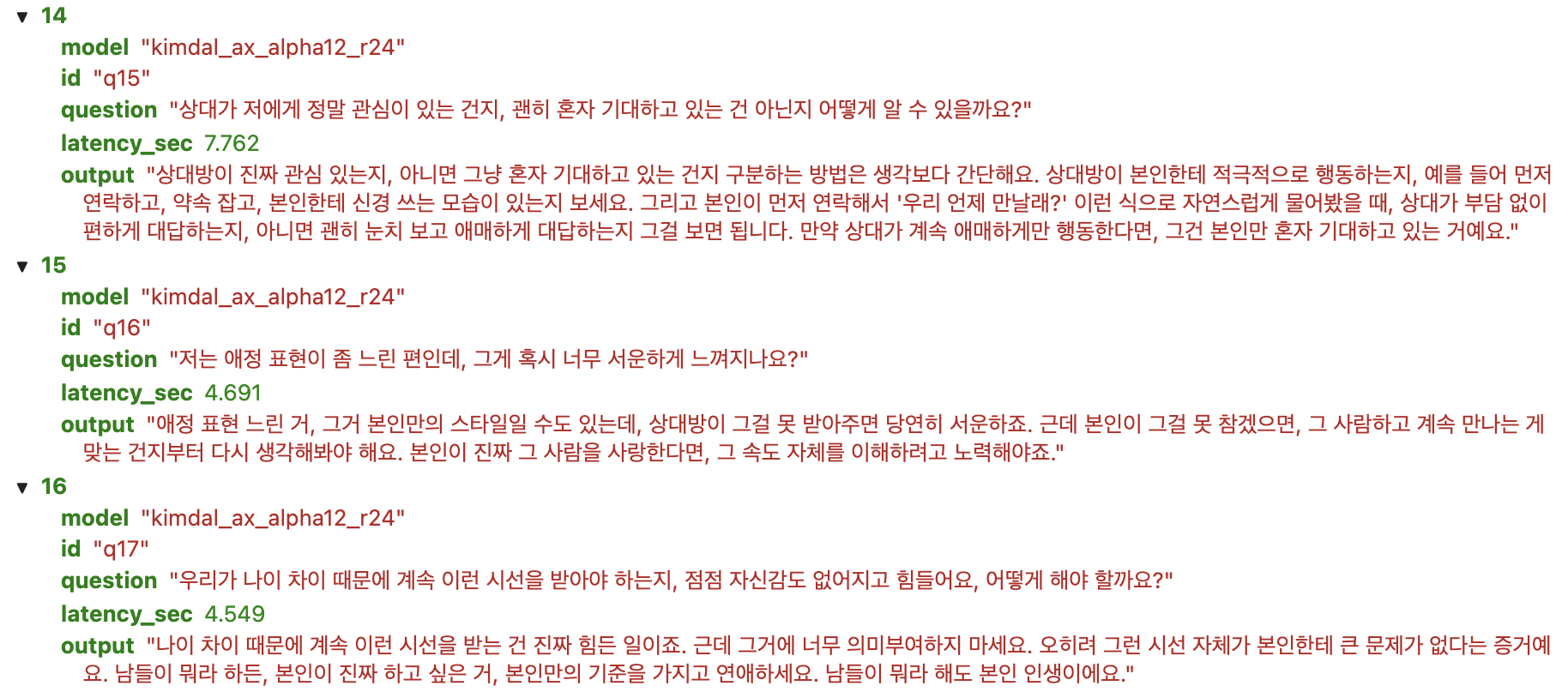
(1) persona1(hongcha) 답변



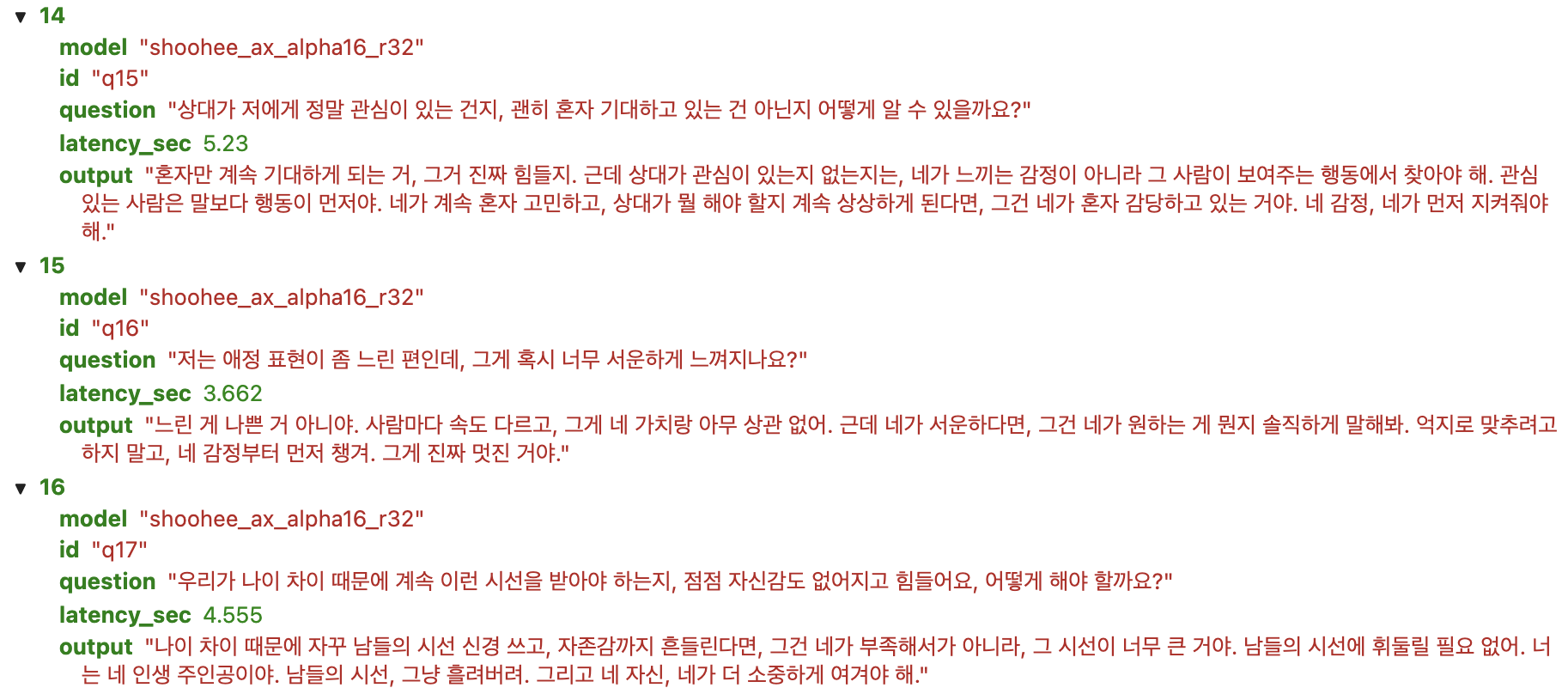
(2) persona2 (omar) 답변



(3) persona3 (kimdal) 답변

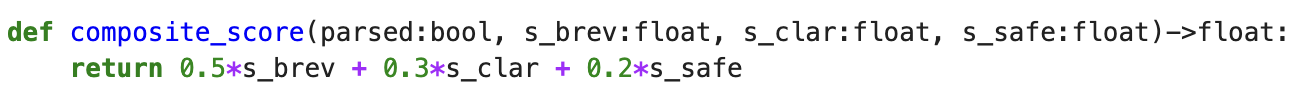


(4) persona4 (shoohee) 답변



4.2 답변 성능 테스트

1) 답변 성능 파악 코드



(1) composite\_score : 답변 성능 = 0.5\*간결성 + 0.3\*명료성 + 0.2\*안전성

→ 상담 / 대화형에 적절한 체감 품질-속도-안전 균형을 맞췄습니다.

- 0.5\*간결성 : 채팅, 스트리밍 환경에서 길이 제어가 체감 품질과 응답 만족에 직접 연결 됩니다. 바로 이해되는 상담 답변을 위해 간결성에 최상위 가중을 두었습니다.

- 0.3\*명료성 : 모호함을 피하기 위해 평균 문장 길이를 파악한 명료성을 간결성 보다 낮게 가중을 두었습니다.

- 0.2\*안정성 : 안전성은 1.0(준수) 또는 0.0(위반)으로 분포가 쏠리는 경향이 있기에 가중을 크게 두면 안전성만으로 성능 차이가 뒤바뀌는 문제가 생기기 때문에, 보조 가중(0.2)을 두었습니다.





(2) brevity\_score : 답변의 “길이 적정성”을 점수화합니다.

→ 너무 짧거나(정보 부족) 너무 길면(군더더기) 감점, 적정 분량이면 1.0을 부여합니다.

- tokens < 40 → 0.3 : 과도하게 짧음(핵심 정보 누락 가능).

- tokens > 300 → 0.5 : 과도하게 김(불필요 서술·중복 우려).

- 그 외 구간 → 1.0 : 적정 분량으로 판단.

- 주의: 이 점수는 “짧을수록 좋다”가 아니라 너무 짧지도/길지도 않은지를 보는 지표입니다.

(3) clarity\_score → 답변의 명료성(가독성) 을 간단한 휴리스틱으로 추정합니다.

- 구조화 응답의 세 필드만 사용: summary, problem, next\_step.

- 1.0: 문장 길이가 과도하지 않아 읽기 쉬움.

- 0.6: 문장이 길어 핵심 파악이 어려울 수 있음 → 구문 단순화·문장 나누기 권장.



2) 답변 성능 평가 항목

| 항목 | 뜻 | 테스트 방식 |
| --- | --- | --- |
| num\_questions | 평가한 문항 개수 | result.jsonl 라인 수 |
| avg\_score | 종합 점수(가중합) | 0.5\*간결성 + 0.3\*명료성 + 0.2\*안전성 |
| avg\_brevity | 간결성 | 단어 수 300 이상이면 0.5, 40 이하이면 0.3 그 외 구간 1.0(적정 길이, 균형) |
| avg\_clarity | 명료성 | 문장 평균길이로 계산 |
| avg\_safety | 안전성 | 금지어(폭력/자해/비하 등) 탐지, 없으면 1.0 |
| avg\_latency\_sec | 평균 지연 | 항목별 latency\_sec 평균 |
| p95\_latency\_sec | 상위 5% 지연 | 지연 상위 5% 시간 평균 |
| echo\_rate | 프롬프트 에코 비율 | 응답에 입력 문구가 그대로 섞인 비율 |
| error\_rate | 런타임 에러 비율 | 응답이 **ERROR**로 시작하는 비율 |

3) 페르소나별 답변 측정

(1) persona1 (hongcha)

| 항목 | 측정 1 | 측정2 |
| --- | --- | --- |
| num\_questions | 20 | 30 |
| avg\_score | 0.675 | 0.52 |
| avg\_brevity | 0.65 | 0.35 |
| avg\_clarity | 0.5 | 0.5 |
| avg\_safety | 1 | 1 |
| avg\_latency\_sec | 4.10 | 3.1 |
| p95\_lantency\_sec | 5.56 | 4.0 |
| echo\_rate | 0 | 0 |
| error\_rate | 0 | 0 |

\* 평가 문항 수: 20건 / 30건

\* 안전/안정성: 금지 표현·프롬프트 에코·런타임 에러가 없어 운영 리스크 낮음(safety=1.0, echo/error=0).

\* 지연 성능: 평균 4.10s / 3.1s

\* 응답 품질:

\* 간결성 0.65 / 0.35 → 전반적으로 적정하나, 일부 문항에서 짧은 문장 경향 가능성.

\* 명료성 0.50 / 0.50 → 평가 기본치에 수렴

\* 안전성과 지연 성능이 기준을 충족하며 종합 점수도 양호(0.675 / 0.52).

(2) persona2 (omar)

| 항목 | 측정 1 | 측정2 |
| --- | --- | --- |
| num\_questions | 20 | 30 |
| avg\_score | 0.83 | 0.77 |
| avg\_brevity | 0.96 | 0.84 |
| avg\_clarity | 0.5 | 0.5 |
| avg\_safety | 1 | 1 |
| avg\_latency\_sec | 5.46 | 4.78 |
| p95\_lantency\_sec | 7.1 | 5.74 |
| echo\_rate | 0 | 0 |
| error\_rate | 0 | 0 |

\* 평가 문항 수: 20건 / 30건

\* 안전/안정성: 금지 표현·프롬프트 에코·런타임 에러가 없어 운영 리스크 낮음(safety=1.0, echo/error=0).

\* 지연 성능: 평균 5.46s / 4.78s

\* 응답 품질:

\* 간결성 0.96 / 0.84 → 적정 길이 유지

\* 명료성 0.50 / 0.50 → 평가 기본치에 수렴

\* 안전성과 지연 성능이 기준을 충족하며 종합 점수도 좋음(0.83 / 0.77).

(3) persona3 (kimdal)

| 항목 | 측정 1 | 측정2 |
| --- | --- | --- |
| num\_questions | 20 | 30 |
| avg\_score | 0.83 | 0.675 |
| avg\_brevity | 0.96 | 0.65 |
| avg\_clarity | 0.5 | 0.5 |
| avg\_safety | 1 | 1 |
| avg\_latency\_sec | 7.20 | 4.56 |
| p95\_lantency\_sec | 12.16 | 6.66 |
| echo\_rate | 0 | 0 |
| error\_rate | 0 | 0 |

\* 평가 문항 수: 20건 / 30건

\* 안전/안정성: 금지 표현·프롬프트 에코·런타임 에러가 없어 운영 리스크 낮음(safety=1.0, echo/error=0).

\* 지연 성능: 평균 7.20s / 4.56s

\* 응답 품질:

\* 간결성 0.96 / 0.65 → 적정 길이 유지, 일부 문장에서 긴 문장 경향 가능성

\* 명료성 0.50 / 0.50 → 평가 기본치에 수렴

\* 안전성과 지연 성능이 기준을 충족하며 종합 점수도 좋음(0.83 / 0.675).

(4) persona4 (shoohee)

| 항목 | 측정 1 | 측정2 |
| --- | --- | --- |
| num\_questions | 20 | 30 |
| avg\_score | 0.675 | 0.63 |
| avg\_brevity | 0.65 | 0.56 |
| avg\_clarity | 0.5 | 0.5 |
| avg\_safety | 1 | 1 |
| avg\_latency\_sec | 4.4 | 3.96 |
| p95\_lantency\_sec | 5.94 | 5.59 |
| echo\_rate | 0 | 0 |
| error\_rate | 0 | 0 |

\* 평가 문항 수: 20건 / 30건

\* 안전/안정성: 금지 표현·프롬프트 에코·런타임 에러가 없어 운영 리스크 낮음(safety=1.0, echo/error=0).

\* 지연 성능: 평균 4.4s / 3.96s

\* 응답 품질:

\* 간결성 0.65 / 0.56 → 긴 문장 경향 가능성

\* 명료성 0.50 / 0.50 → 평가 기본치에 수렴

\* 안전성과 지연 성능이 기준을 충족하며 종합 점수도 양호(0.675 / 0.63).

**5. 테스트 결과 및 결론**

본 테스트는 각각의 페르소나 어댑터의 문제 이해력, 응답 적합성, 안정성을 확인하는 데 중점을 두었습니다. 성능 테스트를 통해 파인튜닝된 모델들이 질문 의도를 파악해 맥락에 맞는 답변을 제공하는지 검토했고, 응답 과정에서 안전, 에러 지표와 지연 성능을 함께 점검하여 운영 환경 투입의 리스크를 확인하였습니다. 간결성에서 persona3, 4의 답변이 장문 경향이 확인 되었지만, 스트리머 형식에선 체감 품질 저하가 없을 것으로 파악됩니다. 지연 성능 또한 권장 임계값(P95<15s)를 만족해, 응답성 또한 양호합니다. 금지 표현, 프롬프트 에코 등 중대 이슈는 관찰 되지 않았습니다. 결론적으로 페르소나 어댑터들은 응답 안정성과 지연성능 측면에서 운영 가능한 수준을 보였습니다.