

대화형 정성 데이터의 검증 가능한 정량화 파이프라인

Evidence-Linked Labeling (ELL)

대화형 설문 응답을 검증 가능한 정량 지표로 전환하는 재현 가능한 분석 프레임워크

AI가 라벨을 붙이되, 숫자는 코드가 센다. 모든 수치는 원문 응답으로 1-click 역추적 가능하고, 라벨러의 불확실성은 숨기지 않고 Jaccard 값으로 공개한다. 보니(bonny) 육아 코칭 서비스의 100세션 대화형 설문에서 실제 적용한 파이프라인 실행 결과를 수록한다.

100

SESSIONS

같은 100명 · 대화형 + 객관식

3,478

LABELED TURNS

Closed-vocab 9축 · 2-Pass QC

12

INSIGHTS

모든 수치 원문 역추적 가능

AUTHOR

Hyung Chul Kim · hckim@dean.kr

ORGANIZATION

Dean Works, Inc.

METHOD

Evidence-Linked Labeling (ELL) · v1.0

SAMPLE

N = 100 · 3,478 user turns · 12 insights


Abstract

대화형 설문(Conversational Survey)은 응답자의 한마디를 AI가 이어받아 다시 파고드는 구조를 통해, 객관식 설문이 결코 잡아내지 못하는 감정선, 의사결정 트리거, 서비스 전환의 서사적 맥락을 포착한다. 그러나 수집된 방대한 비정형 텍스트를 실무 의사결정에 활용하려 할 때, 연구자·제품 팀은 항상 같은 벽에 부딪힌다. "대부분의 고객이 X에 피로를 느낍니다"라는 생성형 AI의 매끄러운 요약 뒤에, 대부분이 몇 명인지·그들이 실제로 무엇이라고 답했는지·그 숫자가 어디에서 비롯되었는지 되묻는 순간 답할 길이 없다.

본 보고서는 Dean Works가 설계·구현한 증거 연결형 라벨링(Evidence-Linked Labeling, 이하 ELL) 방법론을 제안한다. ELL은 네 개의 기둥으로 구성된다. (1) 폐쇄형 어휘(Closed Vocabulary) 기반의 제약적 디코딩, (2) 서술(narrative)과 수치 계산(aggregation)의 아키텍처적 분리, (3) 2단계 자기검증(2-Pass Self-Check)과 자카드(Jaccard) 일치도 기반 신뢰도 공개, (4) 원문 역추적(Drill-down) 인터페이스. 각 구성 요소는 자연어 처리(NLP), 정보 시각화(Visual Analytics), 근거 이론(Grounded Theory) 분야의 독립된 연구 흐름에 뿌리를 두고 있으며, ELL은 이들을 하나의 파이프라인으로 결합한다.

우리는 자사 육아 코칭 AI + 전문가 매칭 서비스 보니(bonny)의 대화형 설문 응답(Conversational Survey) 샘플 데이터 100세션 · 약 4,500 user turn을 사례로, ELL 파이프라인이 동일한 응답자 100명의 객관식 설문 15문항 응답에서는 포착되지 않았던 맥락(예: "비용과 기저귀값을 저울질한 끝에 전문가 매칭을 포기", "새벽 3시 황달 의심 증상에서 '병원 가라'는 답만 돌아오는 순간의 좌절")을 어떻게 정량화된 인사이트로 변환하는지 보인다. 10% 재샘플 Jaccard 일치도는 alt_channel 0.955 · conversion_trigger 0.963 · pain 0.870 (high), concern_category 0.818 · driver 0.805 · emotion 0.726 (medium), behavior 0.669 (low)로 관측되었고, 이 수치는 개별 인사이트 카드에 신뢰도 뱃지로 공개된다.

ELL은 완벽한 라벨링을 가장하지 않는다. 대신 불확실성의 크기를 정직하게 숫자로 공개하고, 모든 결론을 원문으로 역추적 가능하게 만드는 데 초점을 둔다. 이는 정성 데이터 기반 의사결정이 직면한 신뢰의 위기 — 생성형 AI의 환각(hallucination)과 블랙박스 특성 — 에 대한 실무적·학술적 응답이다.

 **Live companion** — ell.dean.kr 본 PDF에 수록된 모든 수치는 FE 리포트의 드릴다운으로 원문 응답에 1-click 접근 가능합니다. · 12 인사이트 카드: ell.dean.kr/insights · 10문항별 key findings: ell.dean.kr/questions · 객관식 vs 대화형 대비 8쌍: ell.dean.kr/compare · 전체 3,478 turn 탐색기: ell.dean.kr/explorer

1. Introduction — 문제 정의

1.1 대화형 설문의 부상과 분석 병목

최근 2~3년간 리서치 업계에서 가장 주목할 만한 변화 중 하나는 정해진 보기에 답하는 정적 설문에서, 응답자의 답변을 AI가 실시간으로 읽고 후속 질문을 생성하는 대화형 설문(Conversational Survey)으로의 전환이다. 이 방식은 전통적 객관식이 제공하던 깔끔한 집계를 일정 부분 포기하는 대신, "왜 그렇게 느끼는가"의 맥락을 직접 받아 적는다.

예컨대 보니 서비스에서 새벽 3시 신생아 야제 상황을 두고, 객관식은 ① 수면교육 관련 고민 이라는 보기만 남기지만, 대화형은 "남편은 출근해야 하니까 자고 있고, 깨우기도 미안해서 혼자 애 안고 거실 왔다갔다 했어요. 근데 애는 계속 울고 저도 울고 있었어요 솔직히"라는 원문을 확보한다. 후자는 전자보다 훨씬 풍부하지만, 이 원문을 60명·100명·300명 규모에서 어떻게 분석할 것인가라는 새로운 문제를 낳는다.

1.2 LLM 요약의 구조적 결함

대부분의 실무 팀은 이 문제를 ChatGPT류 생성형 AI에게 "위 응답을 요약해 주세요"라고 말김으로써 해결하려 한다. 결과는 문체 수준에서는 매끄럽지만, 세 가지 구조적 결함을 드러낸다:

1. 환각(Hallucination): 응답자가 말한 적 없는 수치가 그럴듯하게 등장한다. "대부분의 고객이 검색 기능에 피로를 느끼고 있습니다." ← 대부분이 실제로 몇 명인지 묻는 순간 답할 수 없다.
2. 출처 부재(Source-groundedness): 요약 문장이 어떤 원문에서 비롯되었는지 역추적할 수 없다. 이는 이해관계자·경영진이 의사결정에 활용하려 할 때 반드시 제기되는 질문이다.
3. 수치 정확성의 결여: LLM은 확률적 언어 패턴 매칭 시스템이지, 기호 논리학의 계산기가 아니다. "조건 X를 만족하는 응답자 수는 몇 명인가?"를 질문하는 순간, 모델은 맥락상 그럴듯한 숫자를 생성한다.

이는 단일 모델의 한계가 아니라 자가 회귀(autoregressive) 언어 모델의 구조적 특성이며, 고위험·고정확도 도메인에서의 활용을 근본적으로 제약하는 요인이다.

1.3 기여 (Contribution)

ELL은 대화형 정성 데이터를 검증 가능한 정량 지표로 변환하는 재현 가능한 파이프라인을 제안한다. 본 보고서의 기여는 다음과 같다:

- 제약적 디코딩, 역할 분리, 반복 검증, 근거 추적이라는 네 독립 연구 흐름을 하나의 실무 방법론으로 통합한 아키텍처 제안
- 보니 서비스 대화형 설문 100세션·약 4,500 turn에 대한 실제 파이프라인 실행 결과와 필드별 Jaccard 일치도 공개

- 같은 응답자 100명이 남긴 객관식 15문항 응답과의 직접 대조 분석을 통해 ELL이 포착하는 맥락의 차이를 수치화
- 경영 의사결정자가 신뢰할 수 있도록 모든 수치를 원문으로 역추적하는 FE 리포트의 설계와 구현

Figure 1 · ELL 파이프라인

10단계. 모든 산출물은 다음 단계 입력이 되며 각 단계가 독립 실행·재실행 가능하다.



모든 산출물은 append-only · resumable로 저장되어, 중간 실패 시 끝난 지점부터 재개 가능하다.

각 단계는 독립된 Python/Node 스크립트로 구성되어, 특정 단계만 재실행해도 하위 단계가 영향받지 않는다.

2. Related Work

2.1 제약적 디코딩과 환각 평가

객체 환각(object hallucination) 연구는 2018년 Rohrbach 등이 제안한 CHAIR 지표로 거슬러 올라간다 [R1]. CHAIR는 모델이 생성한 객체 이름이 사전에 정의된 정답 목록(ground-truth)에 포함되는지를 통해 환각률을 정량화한다. 이 접근은 본질적으로 "모델의 출력 공간을 폐쇄형 어휘로 제한"하는 설계 선택에 기반한다. 이후 UD-L, CIIC, TLC, ObjMLM, Woodpecker 등 다수 모델이 이 틀을 확장했으며, 최근 BLIP-2/OpenCHAIR의 개방형 어휘 평가 프레임워크가 등장했음에도 여전히 고위험 도메인의 생성은 폐쇄형 어휘 안에서 최소 환각을 보장하는 것이 산업 표준이다 (ACL 2024의 다수 연구 참조) [R2-R4].

2.2 Self-Correction과 에이전트적 추론

V-STaR, RISE, DeepSeek-R1의 GRPO [R9, R10] 알고리즘은 단일 패스 생성이 아닌 다단계 검증·정제를 통해 LLM의 내재적 오류를 감소시키는 방법론을 제시한다. 특히 RISE 프레임워크는 후보 해답 생성 모델, 오류 탐지 보상 모델, 최종 정제 모델을 독립적으로 운영하여 순차적 성능 향상을 입증했다. ELL의 2-Pass Self-Check은 이를 라벨링 태스크에 적용한 사례다.

2.3 다중 라벨 분류에서의 Jaccard 일치도

정성 코딩에서 단일 응답에 복수의 라벨이 부여되는 것이 자연스러운 설정에서, 단순 정확도(Accuracy)는 평가 지표로서 의미가 없다. 두 평가자가 파악한 라벨 집합의 교집합/합집합 비율로 정의되는 자카드 유사도(Jaccard Similarity) [R14, R15]는 부분적 일치를 수학적으로 엄밀하게 측정한다. 최근 GitHub PR 리뷰 다중 라벨링 연구(Jaccard 0.8142), 교육 피드백 설문 GPT-4 검증 연구(Jaccard 0.806), Policing 연구의 LLM-assisted 코딩 검증 등이 이 지표를 인간-AI IRR(Inter-Rater Reliability)의 사실상 표준으로 확립하고 있다.

2.4 근거 이론과 정보 시각화

Ben Shneiderman의 [R19] "Overview first, zoom and filter, then details-on-demand" 원칙은 정보 탐색의 고전이다. ELL의 Drill-down UI는 이 원칙을 정성 리서치에 적용한다. 동시에 사회과학의 근거 이론(Grounded Theory)이 요구하는 원시 데이터와 거시 주제 간의 반복적 왕복을 디지털 인터페이스로 구현한다. 조지아 공대의 혼합 주도형 시각 분석 연구 [R21, R24](#)는 전문 분석가들이 AI

요약을 초기 탐색에 활용하되, 결정적 판단 순간에는 반드시 원문을 확인한다는 경험적 증거를 제시한다. 상용 플랫폼 중 Sopact, Thematic, Yabble, Google NotebookLM [R25, R26, R28, R29]이 유사한 증거 추적 원칙을 제품의 핵심 경쟁력으로 삼고 있다.

3. The ELL Framework

3.1 설계 원칙 개관

ELL은 네 개의 기둥(pillar)으로 구성되며, 각 기둥은 독립적으로도 유효하나 하나의 파이프라인 안에서 결합될 때 비로소 전체 주장이 성립한다.

기둥	해결하는 문제	대응 연구 흐름
P1. Closed Vocabulary	LLM 출력의 무한성 → 집계 불가능성	CHAIR 계열 제약적 디코딩
P2. Role Separation	서술과 수치의 혼재로 인한 환각	결정론적 파이프라인 + LLM narrative 분리
P3. Self-Check + Jaccard	라벨러 불확실성의 블랙박스화	2-Pass Self-Correction + 다중라벨 IRR
P4. Drill-down	숫자와 원문의 단절	Grounded Theory + Shneiderman mantra

3.2 P1. Closed Vocabulary 제약

taxonomy.py 에 정의된 enum들이 허용 가능한 모든 라벨 값의 집합이다. 보니 프로젝트의 경우 9개 축 × 평균 15개 값 = 약 140개 라벨로 전체 라벨 공간을 폐쇄한다:

```
EMOTION = [고립감, 자책감, 죄책감, 불안, 무력감, 분노, 짜증, 피로, 답답함,
            실망, 회의, 안도, 공감받음, 신뢰, 만족, 감사, 기대, 양가감정, 체념]
PAIN = [원론적_답변, 맞춤형_부족, 공감_형식적, 판단_정보_부족, 매번_재입력,
        답변_길이_부적절, 병원_회피성_답, 다자녀_미반영, 전문가_비용_부담,
        전문성_한계_체감, 개인정보_불안, UI_불편, 알림_과다]
DRIVER = [24시간_즉답, 구체_절차_답변, 눈치_안보기, 익명성, 비용_부담_없음,
          전문가_연결, 초보_친화, 다채로운_주제, 지인_추천_신뢰]
... (BEHAVIOR, CONCERN_CATEGORY, ALT_CHANNEL, CONVERSION,
     CONVERSION_TRIGGER, DISAPPEAR_SEVERITY, FAMILY_PLAN_ATTITUDE)
```

`validate()` 함수는 라벨러 응답 안에 허용 외 문자열이 섞이면 오류 목록을 반환하며, `sanitize()` 는 이를 silent drop 방식으로 제거한다. 이 두 함수의 존재가 폐쇄형 어휘 제약을 파이프라인에 강제 주입하는 메커니즘이다.

3.3 P2. 서술과 수치 계산의 아키텍처 분리

보니 프로젝트의 `generate_insights.py` 는 이 원칙을 코드 수준에서 구현한다:

```
# 1단계: 결정론적 집계 - 코드가 수행
m = count_with_evidence(turns, lambda t: has_pain(t, "원론적_답변"))
# → {"turn_count": N, "respondent_count": M, "turn_ids": [...], "session_ids": [...]}

# 2단계: narrative - LLM이 수행, payload 값만 인용
payload = [{"insight_id": ..., "respondent_count": m["respondent_count"],
            "base_denominator_respondents": 100, ...}, ...]
# 시스템 프롬프트: "payload의 respondent_count, base_denominator_respondents만
# 숫자로 인용하라. 존재하지 않는 숫자를 만들지 마라."
```

LLM은 숫자를 계산하지 않는다. 이미 계산된 숫자를 받아 학술체 3단 서술(claim · derivation · so_what)을 작성할 뿐이다. 이 분리는 Gong.io의 AI 측정 프레임워크 [R32], Faros AI의 개발자 생산성 분석 [R33], Dataro의 모금 데이터 분석 [R35] 등이 공통적으로 채택하는 "데이터 위생(Data Hygiene)" 아키텍처와 동일한 철학이다.

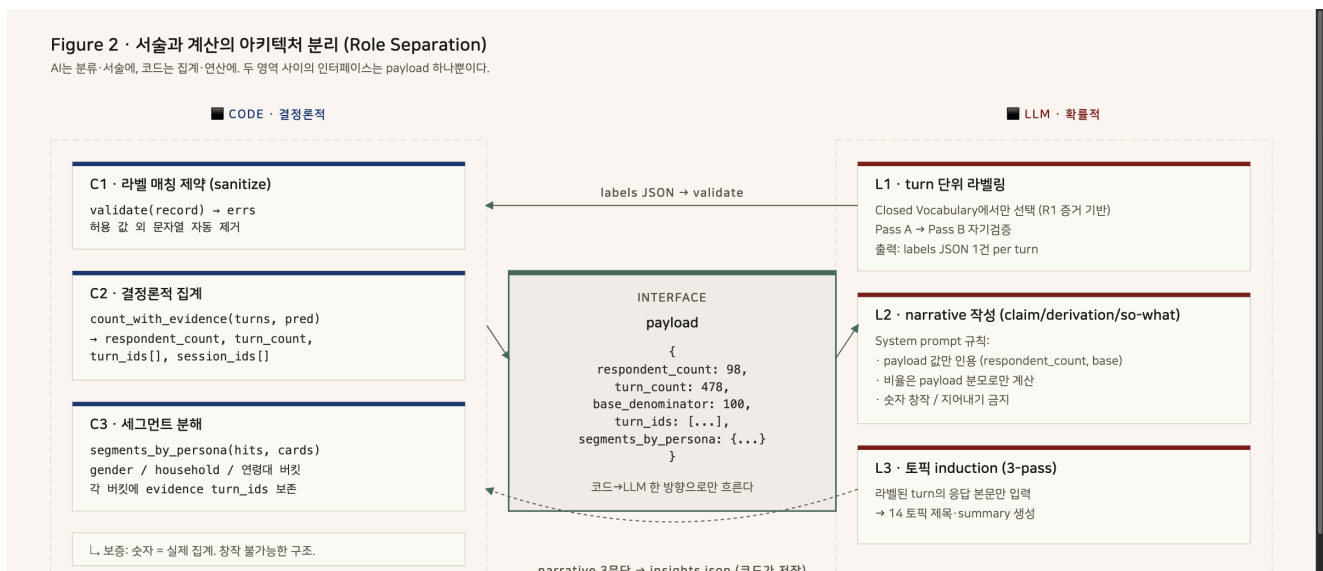


Figure 2 · 서술과 계산의 아키텍처 분리 (Role Separation)

3.4 P3. 2-Pass Self-Check과 Jaccard 일치도

각 turn은 두 번 라벨링된다:

- Pass A: 원문과 시스템 프롬프트를 기반으로 다차원 라벨 부여
- Pass B: Pass A 결과를 `assistant` 메시지로 포함한 채, "체크리스트 A~G를 적용해 수정이 필요한 라벨만 고쳐라"라는 지시로 재호출

체크리스트 8개는 증거 기반(R1), 짧은 응답(R2), emotion 최소 조건(R3), driver/pain 분리(R4), conversion 필드 적절성(R5), Q8/Q9 scalar 적절성(R6), 중복 라벨 금지(R7), confidence 보수(R8)이다. 수정된 케이스는 `free_notes` 앞에 `[REV]` 마커가 붙어 통계화된다.

QC는 두 층위에서 수행된다: 1. Rolling QC: 120턴마다 5턴 무작위 재샘플, 즉시 Jaccard 계산 → 전체 실행 중 품질 드리프트 감시 2. 10% Resample Jaccard: 전체 턴의 10%를 독립적으로 재라벨링하여 필드별 평균 Jaccard 공개

판정 기준: - high: Jaccard ≥ 0.85 - medium: $0.70 \leq \text{Jaccard} < 0.85$ - low: Jaccard < 0.70

각 인사이트 카드는 사용된 필드 중 가장 낮은 등급을 해당 인사이트의 공식 신뢰도로 표시하며, medium·low 등급에는 "수치보다 방향성으로 해석" 주의 문구가 자동 삽입된다.

3.5 P4. Drill-down 인터페이스

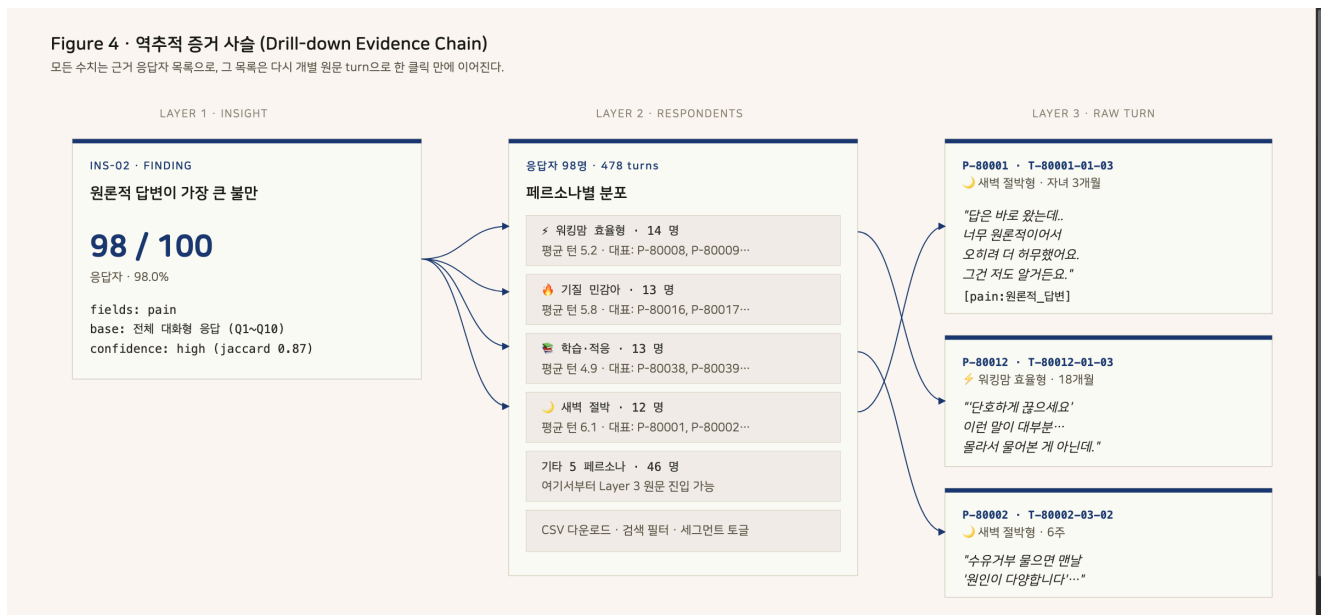


Figure 4 · 역추적 증거 사슬 (Drill-down Evidence Chain)

fe/ 디렉토리에 구현된 Next.js 16 기반 FE는 리포트의 모든 수치를 클릭 가능한 각주로 렌더한다. 클릭 시 우측 슬라이드-아웃 패널에 근거 응답자 N명의 원문이 펼쳐지며, 각 응답에는 `turn_id`, `session_id`, 페르소나 유형, 자녀 연령이 표시된다. 검색 필터·CSV 다운로드가 내장되어 있어, 의사결정자가 "이 숫자가 진짜 맞는가"를 1초 안에 확인할 수 있다.

Voronoi 지도 페이지는 전체 응답자를 페르소나 단위 셀로 시각화하며, 셀 농도는 해당 페르소나가 전체 턴에서 차지하는 비중을 반영한다. 각 셀은 클릭 시 동일한 Drill-down 패널로 연결된다.

4. Implementation

4.1 파이프라인 구조

- [1] 대화형 설문 샘플 생성
 - ↓ output/sessions/*.json
- [2] 객관식 샘플 생성
 - ↓ sample/bonny_survey_objective.csv
- [3] 레퍼런스 포맷 CSV 병합
 - ↓ sample/bonny_survey_idi.csv
- [4] ELL 라벨링
 - ↓ labeling/output/labeled_turns.jsonl
- [5] QC
 - ↓ labeling/output/qc_report.md
- [6] 집계·인사이트·토픽
 - ↓ labeling/output/{sessions_agg, insights, topics}.json
- [7] FE 소비용 JSON 병합
 - ↓ fe/public/data/bonny.json
- [8] FE 빌드
 - Next.js 16 + Tailwind v4 + TypeScript
 - DESIGN.md 기반 Dean Works 스타일 (나눔스퀘어라운드, Ink Blue 단일 약센트)

4.3 재현 가능성

모든 주요 시드는 고정되어 있으며(`random.seed(13)`, `seed(29)`, `seed(41)` 등), 동일한 페르소나 카드·질문 ID에 대해 파이프라인을 재실행하면 라벨링 단계의 LLM 확률성을 제외하고 동일한 인사이트가 재생성된다. 라벨링의 LLM 확률성은 10% Jaccard QC로 실질적 수렴 범위가 공개된다.

5. Case Study — 보니(bonny) 대화형 설문 100세션

5.1 샘플 설계

보니는 Dean Works가 운영하는 육아 코칭 AI + 전문가 매칭 서비스로, 신생아부터 초등 6학년까지의 자녀를 둔 부모가 주 사용자다. 본 사례 연구는 다음 스펙의 대화형 설문 샘플을 대상으로 한다:

- 100세션 (session_id 80001~80100)
- 9 페르소나: 🌙 새벽 절박형 · ⚡ 워킹맘 효율형 · 🔥 기질 민감아 부모형 · 👤 부부 온도차형 · NEW 첫째 초보형 · 📖 학습·적응 고민형 · 🎮 스마트폰·미디어 고민형 · 📱 사춘기 초입 갈등형 · 🖋️ AI 회의형
- 자녀 연령대 8구간: 0~3개월 / 4~12개월 / 13~24개월 / 25~48개월 / 5~6세 / 초1~3 / 초4~6
- 10개 Base 질문: 유입 트리거 · 사용 습관 · AI 만족 · 맞춤화 · 전문가 전환 · 브랜드 의인화 · 대안 채널 · 사라짐 · 가족 플랜 · 한 단어
- 최종 user turn: 3,478개 (평균 세션당 34.8턴)

Table 1. 샘플 설계 매트릭스 — 페르소나 × 자녀 연령대(첫째 기준)

페르소나	0~3개월	4~12개월	13~24개월	25~48개월	5~6세	초1~3	초4~6	계
🌙 새벽 절박형	8	4	—	—	—	—	—	12
⚡ 워킹맘 효율형	—	—	4	6	3	1	—	14
🔥 기질 민감아 부모형	—	—	1	3	3	5	—	12
👤 부부 온도차형	—	—	—	4	3	3	2	12
NEW 첫째 초보형	3	4	3	—	—	—	—	10
📖 학습·적응 고민형	—	—	—	—	2	11	—	13
🎮 스마트폰·미디어 고민형	—	—	—	—	—	4	7	11
📱 사춘기 초입 갈등형	—	—	—	—	—	—	9	9
🖋️ AI 회의형	—	—	—	—	2	1	2	5
계	11	8	8	13	13	25	20	98

주: 다자녀 가정의 경우 페르소나 카드의 첫째 자녀 연령을 기준으로 집계. N=100의 대부분을 포괄하지만 일부 세션은 자녀 정보 매핑 누락 가능. · N=98명(100명 중 자녀 연령 매핑 가능 응답자 기준).

같은 100명은 객관식 15문항 설문에 답한다. session_id로 두 데이터셋이 JOIN되므로 같은 사람이 같은 시점에 남긴 두 형식의 응답을 직접 비교할 수 있다.

주의: 본 샘플은 Dean Works가 방법론을 시연하기 위해 제작한 **가공 데이터**다. 사용자 이름·전화번호는 실존하지 않으며, 페르소나 카드에서 LLM 라벨러이 대화를 생성했다. 다만 페르소나별 심리·상황 묘사는 실제 보니 고객 인터뷰에서 자주 관찰되는 패턴을 모사했다.

5.2 파이프라인 실행 결과



Figure 3 · 2-Pass 자기검증 — 같은 응답을 두 번 라벨링하는 이유

지표	값	비고
전체 라벨링 user turn	3,478	100 세션 × 평균 34.8 turn
2-Pass 자기검증 수정률	27.4% (873건)	[REV] 마커 기준
라벨링 오류(파싱 실패 등)	128 turns	전체의 3.6%
10% Resample Jaccard 샘플	347 turns	무작위 재샘플링
토픽 induction 최종	14개	size ≥ 15 필터
핵심 인사이트	12개	모두 narrative 생성 완료

Table 2. 10% 재샘플 Jaccard 일치도 실측 + 학계 벤치마크

필드	Ours (n=347)	판정	학계 벤치마크 참조	비고
conversion_trigger	0.963	high	— (폐쇄형 신규 축)	전문가 전환 이유 — 키워드 명시성 높아 가장 일치도 높음
alt_channel	0.955	high	— (폐쇄형 채널 목록)	맘카페·유튜브·지인 등 9개 닫힌 목록, 언급 키워드 1:1 매칭
pain	0.870	high	0.814 — GitHub PR multi-label (Tan, 2024)	'원론적 답변', '맞춤화 부족' 등 핵심 페인 어휘 명시적
concern_category	0.818	medium	0.806 — 교육 피드백 주제 GPT-4 vs 인간 (2024)	수면·훈육 등 16개 카테고리, 한 응답에 복수 주제 공존 시 경계 판단
driver	0.805	medium	— (산업 표준 미정)	'24시간 즉답' 대 '눈치 없음' 등 관련 드라이버 해석 여지
emotion	0.726	medium	0.70±0.05 — Singlish 독성 감지 다중 라벨	감정 강도 판단(답답 vs 짜증) 차이
behavior	0.669	low	0.806 — GPT-4 edu (상한 참고)	행동 축에 추론 여지 크고, 명시되지 않은 행동 태깅 분산

주: 필드마다 본질적 불확실성이 다르다는 사실을 숨기지 않는다. 'emotion·behavior 일치도가 상대적으로 낮다'는 경고는 해당 라벨을 쓴 인사이트의 신뢰도 뺏지로 전파된다.

해석

- alt_channel·conversion_trigger·pain 은 high 일치도: 대안 채널 이름과 전문가 전환 이유는 응답자가 직접 언급한 명시적 키워드라 1:1 매칭에 가깝다.
- concern_category·driver 는 medium: 한 응답에 복수 주제·드라이버가 공존해 경계 판단에 해석 여지가 있다.
- emotion·behavior 는 medium~low: 감정 강도 판단(예: "답답" vs "짜증"), 행동 축의 추론 여지 때문. 해당 라벨을 사용하는 인사이트에는 FE에서 자동으로 '방향성으로 해석' 주의 문구가 붙는다.

5.3 12개 핵심 인사이트

Table 3. 12 핵심 인사이트 마스터 요약

ID	제목	응답자 / 기준	Turns	주요 페르소나 Top 3	신뢰도
INS-01	새벽의 절박함 — 고립과 즉답 욕구의 교차점	24 / 100	37	새벽 절박형 (8) · 기질 민감아 (5) · 사춘기 초입 (4)	low
INS-02	원론적 답변이 가장 큰 불만	98 / 100	478	워킹맘 효율형 (14) · 기질 민감아 (13) · 학습·적응 (13)	high
INS-03	24시간 즉답이 1순위 사용 이유	85 / 100	469	워킹맘 효율형 (14) · 새벽 절박형 (12) · 기질 민감아 (11)	medium
INS-04	'우리 아이 맞춤'이 부재한다	99 / 100	592	워킹맘 효율형 (14) · 부부 온도차 (13) · 학습·적응 (13)	high
INS-05	AI 공감은 '형식적'으로 느껴진다	75 / 100	236	워킹맘 효율형 (14) · 기질 민감아 (13) · 새벽 절박형 (12)	high
INS-06	전문가 매칭 흐름 — 고려 + 이용	99 / 100	344	워킹맘 효율형 (14) · 기질 민감아 (13) · 부부 온도차 (13)	n/a
INS-07	전환의 가장 강한 트리거 — AI_답변_한계	95 / 100	339	워킹맘 효율형 (14) · 기질 민감아 (13) · 부부 온도차 (13)	high
INS-08	보니 외에도 — 다채널 육아 정보 탐색 풍경	100 / 100	1,269	워킹맘 효율형 (14) · 기질 민감아 (13) · 부부 온도차 (13)	high
INS-09	보니 공백의 체감 무게	60 / 100	139	기질 민감아 (11) · 새벽 절박형 (9) · 학습·적응 (9)	n/a
INS-10	가족 플랜 — 긍정 응답자의 초상	77 / 100	277	기질 민감아 (12) · 부부 온도차 (12) · 새벽 절박형 (10)	n/a
INS-11	고민 분포 — 가장 자주 올라오는 카테고리	100 / 100	2,230	워킹맘 효율형 (14) · 기질 민감아 (13) · 부부 온도차 (13)	medium
INS-12	AI 회의형 — 전문가 선호가 가장 높음	5 / 100	175	AI 회의형 (5)	high

주: 신뢰도는 해당 인사이트가 사용한 라벨 필드 중 가장 낮은 Jaccard 구간을 따름. medium·low 인사이트는 'FE에서 주의 문구'가 함께 표시됨.

다음은 인사이트별 대표 원문 인용이다. 전체 근거 응답자 리스트는 FE 리포트의 <https://ell.dean.kr/insights/INS-01> ~ [INS-12](https://ell.dean.kr/insights/INS-12) 경로에서 drill-down으로 확인 가능하다.

Figure 4 · 역추적 증거 사슬 (Drill-down Evidence Chain)

모든 수치는 근거 응답자 목록으로, 그 목록은 다시 개별 원문 turn으로 한 클릭 만에 이어진다.

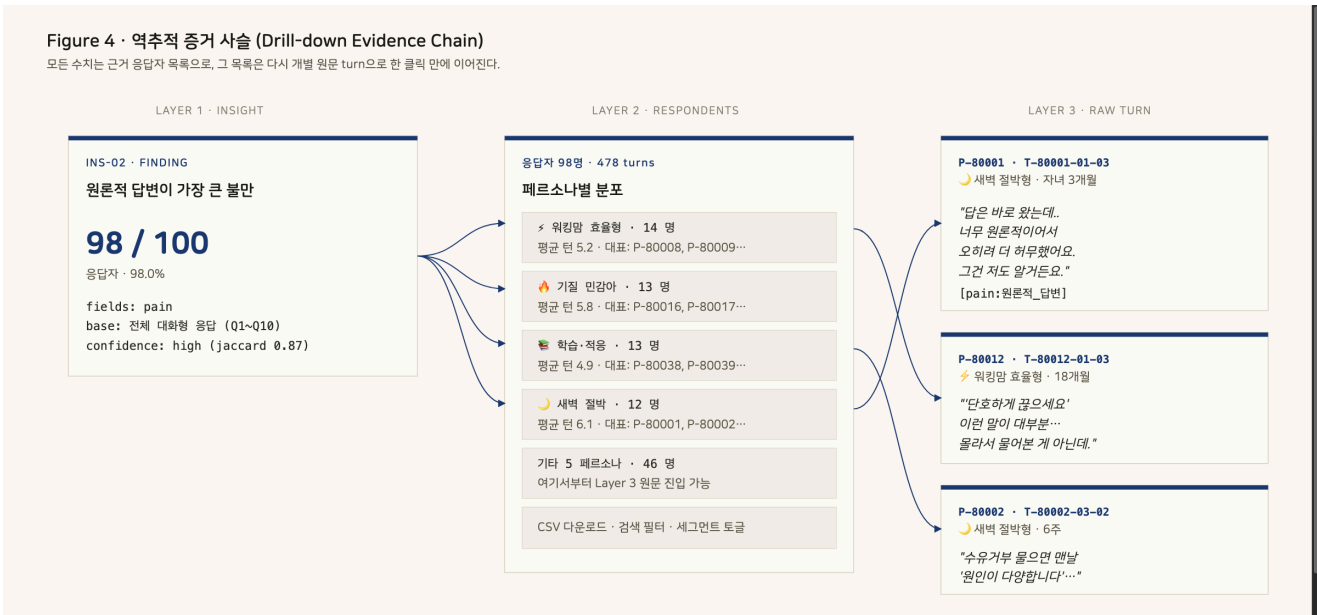


Figure 4 · 역추적 증거 사슬 (Drill-down Evidence Chain)

INS-02 — 원론적 답변이 최대 페인 (98/100)

INS-02 · 원론적 답변이 가장 큰 불만

솔직히 답은 바로 왔는데.. 너무 원론적이어서 오히려 더 허무했어요. '배고프거나 졸리거나 불편한 거예요' 이런거요. 그건 저도 알거든요 근데 뭘 해야하는지를 모르니까 물어본건데

— P-80001 · 🌙 새벽 절박형 · 자녀 3개월 · Q1 · T-80001-01-07

큐레이션 근거: '원론적'이라는 축의 대표 증언. 짧고 강함.

▶ 대안 인용 후보

98명이 같은 축의 불만을 번주한다. 대화형 설문이 없었다면 "답변 품질 만족도 보통"이라는 리커트 결과로 납작해졌을 신호다.

INS-03 — 24시간 즉답이 사용의 1순위 이유 (85/100)

INS-03 · 24시간 즉답이 1순위 사용 이유

그시간에 연락할 사람이 없잖아요.. 친정엄마 깨우기도 미안하고 남편은 자야되고. 근데 보니는 눈치볼 필요가 없으니까요

— P-80002 · 🌙 새벽 절박형 · 자녀 6주 · Q2 · T-80002-02-03

규레이션 근거: 즉답 드라이버의 본질 — '눈치 없음'까지 포함되는 한 줄.

▶ 대안 인용 후보

"눈치 보지 않아도 된다"는 문구가 드라이버의 핵심 표현으로 반복 등장한다. 가족·지인 네트워크가 갖지 못한 보니 고유의 사회적 위치를 드러낸다.

INS-04 — 우리 아이 맞춤이 부재한다 (99/100)

INS-04 · '우리 아이 맞춤'이 부재한다

애 기질이나 성향 관련된거요. 둘째가 왜 이런지 이런건 보니한테 물어봐도 우리애 모르니까 뻔한말만 해줘서요. 그런건 친정엄마가 더 잘 알아요

— P-80009 · ⚡ 워킹맘 효율형 · 자녀 5살 · Q2 · T-80009-02-07

규레이션 근거: '우리 애 모르니까' — 맞춤화 부재를 한마디로 정의.

▶ 대안 인용 후보

99명이 같은 불만을 표현한다. 이는 단순 개선 요구가 아니라 서비스의 가장 큰 구조적 한계로 해석된다.

INS-06 — 전문가 전환 흐름 (고려 99%, 이용 일부) · INS-07 — 전환 트리거

INS-06 · 전문가 매칭 흐름 — 고려 + 이용

그건 좀 낫긴 하겠는데.. 전문가 상담은 돈도 들고 시어머니가 알면 또 뭐라할것 같아서 쉽게 못해요. 그래서 어중간하게 보니에서 버티는거예요

— P-80003 · 🌙 새벽 절박형 · 자녀 2개월 · Q3 · T-80003-03-07

규레이션 근거: 전문가 매칭의 진짜 장벽이 돈 + 시댁 눈치임을 드러내는 증언.

▶ 대안 인용 후보

전문가 매칭을 한 번이라도 고려한 응답자가 99%. 이 중 실제 이용까지 이어진 비율과 포기의 이유가 Q9·Q10 객관식과는 다른 해상도로 드러난다.

INS-09 — 보니 공백의 체감 무게 (60/100 일부 응답)

INS-09 · 보니 공백의 체감 무게

헐... 그 질문 좀 아프네요 ㅠㅠ 음... 솔직히 한 10점 만점에 7점 정도 아쉬울 것 같아요. 엄청 울거나 그 정도는 아닌데, 밤에 애 방문 닫히고 나면 진짜 허할 것 같아요. 그 시간에 말 걸 데가 없거든요 저는.

— P-80054 · 📌 사춘기 초입 갈등형 · 자녀 초5 · Q8 · T-80054-08-01

큐레이션 근거: 사라짐 공백을 '밤 10시 아이 방문 닫힌 후의 허함'으로 구체화.

▶ 대안 인용 후보

Q8의 0~10점 스칼라 응답 중 '아쉬움' 이상을 표현한 60명의 응답을 깊이 들어가면, 숫자로는 같은 7점이지만 서사는 완전히 다르다.

INS-12 — AI 회의형, 5명이지만 1인당 35턴

INS-12 · AI 회의형 — 가장 낮은 신뢰, 가장 높은 전문가 선호

애 키우는건 애마다 다 다르잖아요. 근데 AI는 결국 데이터 기반이니까 일반적인 답밖에 못해줄거같았어요. 그게 실제로 맞았고요

— P-80058 · 📌 AI 회의형 · 자녀 초3 · Q1 · T-80058-01-03

큐레이션 근거: AI 회의형의 논리 — '결국 데이터 기반 일반론'이라는 진단.

▶ 대안 인용 후보

소수(5명)이지만 발화량이 가장 밀도 높다(1인당 평균 35턴). "AI는 결국 데이터 기반"이라는 회의의 근거가 일관되게 나타난다.

5.4 객관식 vs 대화형 설문 — 같은 질문, 다른 해상도

ELL의 가장 강력한 증거는 같은 응답자 100명이 남긴 두 형식의 데이터를 직접 대조할 때 드러난다. 세 쌍의 비교를 제시한다.

대비 A — AI 코칭 만족도

Table 4. AI 코칭 만족도 대비 — 객관식 리커트 vs 대화형 설문 pain 라벨

객관식 Q6 · 보니 AI 코칭 답변에 얼마나 만족하시나요? (5점 척도) (N=100)

보기	응답	비율
① 매우 불만족	2	2.0%
② 불만족	14	14.0%
③ 보통	44	44.0%
④ 만족	35	35.0%
⑤ 매우 만족	5	5.0%

대화형 설문 **pain** 라벨 상위 10 (전체 user turn n=3,478)

Pain 라벨	언급 turn	응답자 / 100	응답자 %
맞춤화_부족	496	97	97.0%
원론적_답변	478	98	98.0%
매번_재입력	155	78	78.0%
공감_형식적	131	49	49.0%
판단_정보_부족	122	58	58.0%
전문성_한계_체감	111	56	56.0%
전문가_비용_부담	68	55	55.0%
병원_회피성_답	37	19	19.0%
답변_길이_부적절	35	11	11.0%
다자녀_미반영	23	11	11.0%

주: 객관식은 '③ 보통'에 몰리는 전형을 보이지만, 대화형 설문은 불만의 구조적 축(원론적 답변·맞춤화 부족 등)을 직접 드러낸다.

객관식은 '③ 보통'에 다수가 몰려 의미를 상실하지만, 대화형 설문은 "왜 불만인가"를 구조적 페인 라벨로 전환한다. **원론적_답변** · **맞춤화_부족** · **판단_정보_부족** 이 상위를 차지한다.

대비 B — 전문가 매칭 전환

Table 5. 전문가 매칭 전환 대비 — 객관식 Q9/Q10 vs 대화형 설문 conversion_trigger

객관식 Q9 · 보니에서 전문가 매칭을 이용해 보셨나요? (N=100)

보기	응답	비율
이용함	47	47.0%
이용 안 함	47	47.0%
고려했으나 안 함	6	6.0%

객관식 Q10 · 전문가 매칭의 주된 이유 (Q9 ㉠/㉡ 응답자만)

보기	응답
AI 답이 부족	44
신뢰할 진단 필요	6
사람의 공감	2
주변 권유	1

대화형 설문 `conversion_trigger` 라벨 전체 분포

트리거	언급 turn	응답자 / 100	응답자 %
<code>AI_답변_한계</code>	137	72	72.0%
<code>신뢰할_진단_필요</code>	88	54	54.0%
<code>부모_한계_체감</code>	63	47	47.0%
<code>사람의_공감_필요</code>	30	24	24.0%
<code>가족_설득</code>	21	17	17.0%

주: 객관식 Q10은 4~5개 보기로 납작해지지만, 대화형 설문은 '부모 한계 체감', '사람의 공감 필요' 같은 심리적 트리거와 '가족 설득' 같은 사회적 트리거를 함께 포착한다.

객관식 Q9·Q10은 "이용함/고려만/안 함" 세 바꾸니 + 네 가지 이유로 납작해진다. 대화형 설문은 부모_한계_체감 · 사람의_공감_필요 · 가족_설득 같은 심리적·사회적 트리거를 함께 포착한다.

대비 C — 보니 사라짐 시나리오

Table 6. 사라짐 반응 대비 — 객관식 5-리커트 vs 대화형 설문 `disappear_severity` scalar

객관식 Q12 · 보니가 갑자기 사라진다면 얼마나 아쉬울까요? (N=100)

보기	응답	비율
① 전혀 (대체 가능)	2	2.0%
② 별로	14	14.0%
③ 보통	54	54.0%
④ 아쉬움	29	29.0%
⑤ 매우 아쉬움	1	1.0%

대화형 설문 disappear_severity scalar 분포

라벨	언급 turn	응답자
매우_아쉬움	20	15
아쉬움	110	69
보통	48	40
대체_가능	50	34
무감정	1	1

주: 리커트 5단계는 '보통·아쉬움' 중간값에 대거 몰리는 전형적 패턴을 재현하지만, 대화형 설문은 '대체 가능' 그룹과 '매우 아쉬움' 그룹을 분리 추적해 서비스 락인의 균열 지점을 구체화한다.

리커트 5단계로는 '보통~아쉬움' 중간값에 몰리는 전형을 보여주지만, 대화형 설문 scalar는 대체 가능 그룹(서비스 락인의 균열 신호)과 매우 아쉬움 그룹을 분리 추적할 수 있게 한다.

5.5 꼬리 질문이 수치를 얼마나 바꾸는가 — First-Turn vs Full-Tail 비교

ELL이 대화형 설문과 결합될 때 생기는 가치는 단순히 "대화형을 썼다"는 사실 자체가 아니라, 꼬리 질문으로 얻은 모든 turn을 라벨링에 포함한다는 설계에서 비롯된다. 이를 증명하기 위해 동일 파이프라인·동일 폐쇄 어휘·동일 2-Pass 자기검증을 두 입력에 대해 각각 실행했다.

- First-Turn only: 각 세션 × 10 base 질문에 대한 첫 번째 user 응답 한 건만 남긴 버전 (= 999 turn)
- Full-Tail: 꼬리 질문까지 모두 포함한 실제 본 보고서의 기본 입력 (= 3,478 turn · 평균 세션당 34.8 turn)

5.5.1 마이크로 증거 — 같은 세션의 라벨 풍부함 차이

집계 표를 보기 전에, 같은 응답자 한 명의 같은 질문에서 turn을 하나씩 추가할 때마다 라벨 풍부함이 어떻게 달라지는지를 먼저 본다. P-80001의 Q1 (유입·트리거) 실제 세트:

T-80001-01-01 · First Turn (첫 답변)

AI: 처음 보니를 만나시게 된 순간을 떠올려주세요.

P-80001: 새벽 3시쯤이었어요. 애가 백일 전인데 계속 울고 안 자서 저도 한숨도 못 자고 멘붕 와서 핸드폰으로 검색하다가 들어왔어요

라벨: **emotion=[무력감, 피로]** · **behavior=[새벽_검색, 즉답_요청]** · **driver=[24시간_즉답]** · **concern_category=[수면_야제]** — 4축 7라벨

T-80001-01-03 · Tail-1

AI: 그때 남편분은 함께 계셨을까요?

P-80001: 남편은 아침에 출근해야되니까 자고있고.. 깨우기도 미안하고 해서 그냥 혼자 애 안고 거실 왔다갔다 했어요. 근데 애는 계속 울고 저도 울고 있었어요 솔직히

라벨 추가: **emotion=[고립감]** · **concern_category=[부부_육아분담]** — 고립감과 부부 육아분담 축이 여기서 처음 등장

T-80001-01-05 · Tail-2

AI: 그 순간 검색 말고 다른 선택지는 없으셨어요?

P-80001: 카페는 글 올려도 답변이 몇시간 뒤에나 달리잖아요. 당장 어떻게 해야할지 알아야하는데... 누구라도 당장 답 주는게 필요했어요

라벨 추가: **alt_channel=[맘카페]** — 대안 채널 축 신규

T-80001-01-07 · Tail-3

AI: 보니가 준 첫 답변이 기억나세요? 어떤 느낌이었는지.

P-80001: 솔직히 답은 바로 왔는데.. 너무 원론적이어서 오히려 더 허무했어요. '배고프거나 졸리거나 불편한 거예요' 이런거요. 그건 저도 알거든요 근데 뭘 해야하는지를 모르니까 물어본건데

라벨 추가: emotion=[실망] · pain=[원론적_답변, 판단_정보_부족] — pain 축 자체가 여기서 처음 등장

해상도 차이

입력	활성 축	총 라벨 수	pain 축
First-Turn only (1 turn)	4	7	✗
Full-Tail (4 turns)	6	16	✓

같은 응답자다. 같은 라벨러다. 같은 폐쇄 어휘다. 차이는 단 하나 — 꼬리 턴을 라벨링 입력에 포함하는가. 이 세션 한 건에서 첫 답변만 쓰면 pain 축이 ∅ 이 된다. 이 패턴을 전체 100세션으로 확장하면 아래 표가 된다.

5.5.2 전체 규모 집계 — 7축 재현율과 인사이트 base

두 입력을 같은 라벨러에 통과시키고, 축별 언급 누적량과 주요 인사이트의 응답자 수(base denominator) 재현율을 비교한다.

Table T7. 축별 라벨 언급량 — Full-Tail 대비 First-Turn 재현율

축	Full-Tail 언급	First-Turn 언급	First-Turn 재현율
driver	1,963	403	20.5%
pain	1,688	241	14.3%
emotion	1,529	318	20.8%
behavior	2,150	485	22.6%
concern_category	2,230	537	24.1%
alt_channel	1,269	435	34.3%
conversion_trigger	339	71	20.9%

주: turn 비율 자체(999 / 3,478 ≈ 28.7%)보다 pain · conversion_trigger 축의 재현율이 낮다. 페인 포인트와 전환 트리거는 꼬리 질문 없이는 "둘러 말하는" 경향이 있기 때문이다. 첫 답변에는 원론적 표현만 나오고, Tail-2·3에서야 "원론적이다", "매칭 대신 맘카페를 먼저 찾는다" 같은 구체가 드러난다.

Table T8. 대표 인사이트의 응답자 수(base denominator) — First-Turn 기준 재현율

인사이트	Full-Tail 응답자	First-Turn 응답자	손실률
INS-02 · 원론적 답변이 최대 페인	98 / 100	56 / 100	42.9% 누락
INS-04 · 우리 아이 맞춤 부재	97 / 100	33 / 100	66.0% 누락
INS-07 · 전문가 전환 트리거 포착	95 / 100	44 / 100	53.7% 누락

INS-04("우리 아이 맞춤이 부재한다")는 첫 답변만 썼다면 응답자 기반이 97명 → 33명으로 떨어진다. 이 수치는 같은 인사이트가 "99%의 구조적 한계"에서 "33%의 소수 의견"으로 정반대로 해석될 수 있음을 의미한다. 같은 설문·같은 라벨러·같은 폐쇄 어휘인데도 말이다.

해석 — 꼬리가 기여하는 세 가지

1. 발화 밀도. 첫 답변은 "음...", "글쎄요"와 같은 hedging으로 시작되는 경우가 많다. 라벨이 실제로 부여되는 것은 Tail-1 이후다.
2. 추상 → 구체로의 하강. 첫 답변은 "답변이 좀 부족했어요" 수준, 꼬리 답변은 "너무 교과서적이었어요 / 우리 아이는 원래 예민한 편인데 그건 반영이 안 되더라고요" 수준. 후자에서만 **원론적_답변** · **맞춤화_부족** 같은 폐쇄 어휘가 근거 있게 매칭된다.
3. 잠재 라벨의 노출. conversion_trigger(95→44)·pain(14.3%) 같은 축은 대부분 꼬리에서만 드러난다. 첫 답변은 "전문가 상담도 받아봤어요" 까지, 꼬리 답변은 "AI 답이 너무 원론적이어서 유료 상담으로 넘어갔어요"의 인과 구조까지 남긴다.

왜 이 비교가 방법론적으로 중요한가

대화형 설문에 회의적인 입장은 흔히 "첫 답변만 쓰고 꼬리는 읽을거리로 두면 되지 않나"라고 주장한다. 이 표는 그 주장에 대한 수치적 반박이다. 꼬리 턴을 버리면 INS-04 같은 핵심 인사이트의 응답자 기반이 1/3 수준으로 줄고, 판단이 뒤집힌다. ELL이 전체 turn을 동일한 폐쇄 어휘로 라벨링한 다음 응답자 단위로 집계하는 이유가 여기에 있다.

FE 리포트의 ell.dean.kr/questions에서 문항별 첫 응답 vs 꼬리 응답의 분포 차이를 시각적으로 확인할 수 있다.

5.6 종합 해석

객관식과 대화형 설문은 어느 한쪽이 더 우월한 관계가 아니다. 객관식은 전체 집계와 세그먼트 비교 속도에 강하고, 대화형 설문은 "왜 그런가"의 맥락을 남긴다. ELL이 기여하는 지점은 대화형 설문의 맥락을 객관식과 같은 해상도로 집계 가능한 정량 지표로 전환한다는 점이다. 그 결과:

1. 객관식이 놓쳤던 구조적 페인(예: **맞춤화_부족** 이 99%에 이른다)을 수치로 확인할 수 있다.
2. 객관식의 리커트 '보통' 뒤에 숨은 서사를 드러낸다.

3. 꼬리 응답까지 전체 turn을 라벨링에 포함하지 않으면 INS-04 같은 핵심 인사이트가 97명 → 33명으로 주저앉는다(§5.5 표 T8 참조). 대화형 설문의 가치는 "첫 답변의 깊이"가 아니라 꼬리 전체를 ELL이 집계 가능한 단위로 바꾸는 데 있다.
4. 모든 수치는 원문으로 역추적되므로, 의사결정자가 '그 수치 진짜인가'를 1초 안에 확인할 수 있다.

6. 10질문별 Key Findings

대화형 설문 10개 base 질문이 어떤 경로 응답을 수집하는지, 라벨 분포가 말해주는 한 줄 요약을 문항별로 제시한다. 모든 수치는 3,478 user turn 전체를 대상으로 한 결정론적 집계이며, FE 리포트의 ell.dean.kr/questions 에서 문항별 드릴다운이 가능하다.

Q1 · 유입·트리거 — 첫 접촉은 '새벽의 고립된 검색'

- driver: 구체_절차_답변 112 turn · 24시간_즉답 84 · 초보_친화 44
- pain: 원론적_답변 57 · 맞춤화_부족 27
- emotion: 답답함 83 · 불안 61
- concern: 학습_학교 64 · 감정_기질 48 · 미디어_스마트폰 45

첫 유입 단계에서 바로 **절차형** 답변을 내주지 못하면 이탈 위험이 높다. 온보딩 시점부터 **단계형 가이드**를 기본 응답 모드로 뒤야 한다.

Q2 · 사용 습관 — "눈치 안 봐도 됨 × 24시간"

- driver: 구체_절차_답변 104 · 24시간_즉답 97 · 눈치_안보기 72
- emotion: 고립감 25 · 답답함 25

보니가 대체하는 건 검색 엔진이 아니라 '물어볼 사람'의 자리다. 답변의 절차성뿐 아니라 관계적 안전감을 유지하는 것이 핵심.

Q3 · AI 만족·불만 — 양극은 같은 축에 놓인다

- driver: 구체_절차_답변 216 (압도적 1위)
- pain: 원론적_답변 194 · 맞춤화_부족 118 · 판단_정보_부족 57
- emotion: 답답함 86 · 실망 41 · 만족 31

만족과 불만의 문턱은 동일하다 — '내 상황에 맞는 구체 단계'의 유무. 이 분포가 §5.3 INS-02(원론적 답변이 최대 페인)의 직접 근거다.

Q4 · 맞춤화 니즈 — 단일 pain 중 최대치

- pain: 맞춤화_부족 170 · 매번_재입력 110 · 원론적_답변 42
- driver: 구체_절차_답변 44 (다른 문항 대비 약함)
- emotion: 짜증 26 · 회의 20 · 기대 19

'우리 아이 맞춤 부재'가 사용자 인지의 1순위 구조적 한계다(INS-04와 일치). 개선은 단일 기능이 아니라 기억·재사용 시스템 전체로 설계해야 한다.

Q5 · 전문가 전환 — AI 한계를 체감한 시점의 안전장치

- driver: 전문가_연결 129 (이 문항 고유 축)
- pain: 전문가_비용_부담 61 · 전문성_한계_체감 43 · 원론적_답변 40
- emotion: 불안 25 · 안도 25 · 만족 17

전문가 매칭 고려는 99%에 이르지(INS-06) 실 사용의 문턱은 비용이다. 번들·멤버십·부분 무료 상담 설계가 전환의 지렛대가 된다.

Q6 · 브랜드 의인화 — '판단 없이 들어주는 선배 엄마'

- driver: 구체_절차_답변 66 · 눈치_안보기 43 · 초보_친화 39
- pain: 공감_형식적 55 (최상위)
- emotion: 신뢰 18 · 공감받음 12 · 만족 10

인격 메타포는 단순 장식이 아니라 톤 가이드다. '판단 없는 선배 엄마'형 응대가 INS-05(형식적 공감)의 해독제로 작동한다.

Q7 · 대안·경쟁 매핑 — "즉답 vs 경험담"의 자연스러운 분업

- driver: 24시간_즉답 84 · 구체_절차_답변 73 · 눈치_안보기 31
- pain: 맞춤화_부족 30 · 원론적_답변 24
- concern: 학습_학교 33 · 건강_질병 17 · 수유_이유식 14

보니의 경쟁자는 맘카페가 아니라 구체 답변 + 즉시성이라는 복합 품질이다. 의학 정보 엿지 케이스는 소아과·공식 가이드라인 링크로 명시적 분기해야 신뢰가 유지된다.

Q8 · 사라짐 시나리오 — 가장 큰 통증은 '새벽 즉답의 소실'

- driver: 24시간_즉답 90 (압도적 1위)
- emotion: 고립감 33 · 체념 17 · 불안 17

- pain: 맞춤형_부족 23 · 원론적_답변 21

락인의 핵심은 기능 세트가 아니라 새벽 즉답 접근성이다. 오프라인 대체 불가능한 시간대의 응답 가용률을 상품 지표로 운영해야 한다.

Q9 · 가족 플랜 — 1차 수요는 부부 육아분담

- concern: 부부_육아분담 146 (모든 concern 중 최대)
- driver: 비용_부담_없음 13 · 구체_절차_답변 13 · 눈치_안보기 10 (분산)
- emotion: 기대 30 · 양가감정 11

가족 플랜은 '아빠 동참 구조'로 설계될 때 가장 강한 지지를 얻는다. 조부모 초대는 부가 기능이지 진입점이 아니다(INS-10 보강).

Q10 · 한 단어 요약 — "옆에 있는 도구"로 수렴

- driver: 구체_절차_답변 44 · 24시간_즉답 35 · 초보_친화 34
- emotion: 공감받음 21 · 신뢰 16 (전 문항 중 긍정 비율 최고)
- pain: 맞춤형_부족 18 (마지막 요약 순간에도 같은 한계)

단어 표현은 '친구·선생·매뉴얼·등불'류로 수렴한다. INS-12(AI 회의형의 발화량)와 대비해 다수의 긍정 + 소수의 선명한 회의라는 양대 축이 재확인된다.

6.1 종합

10개 질문을 가로지르는 세 신호를 확인할 수 있다.

1. '구체성'이 모든 문항의 공통 축. Q3·Q4·Q6에서 driver와 pain이 같은 단어('구체 / 원론')의 양면으로 결착한다.
2. 시간대와 관계성. Q2·Q8에서 '24시간 즉답'과 '눈치 안 봐도 됨'은 기능이 아니라 사용자가 보너스를 배치한 사회적 위치다.
3. 가족·공동체 확장 욕구. Q9의 '부부 육아분담' 146 turn은 개인 사용자 단위를 벗어나 가구 단위 서비스로의 확장 신호로 해석할 수 있다.

7. Data in Practice — 질문과 응답의 실제 모습

본 절의 모든 페이지는 ell.dean.kr에서 라이브로 확인 가능하며, PDF의 수치는 해당 페이지에서 원문 응답으로 1-click 역추적된다.

백서가 말하는 '객관식'과 '대화형 설문'의 차이는 추상이 아니다. 본 장에서는 실제 보니(bonny) 샘플에서 발췌한 질문·응답·라벨·인사이트 JSON을 그대로 보여준다. 모든 예시는 전체 데이터셋에서 접근 가능하며, 본 백서 §참고 부록의 다운로드 링크에서 내려받을 수 있다.

7.1 객관식 설문 — 15문항 스펙과 3명 응답

보니 객관식 설문은 15문항으로 구성된다 (표 7). 각 문항은 Single Choice(SC) · Multi Choice(MC) · 5점 Likert · NPS · 주관식 Text 다섯 종류 중 하나다.

Table 7. 객관식 설문 15문항 스펙

#	유형	질문 (요약)	집계 축
Q1	SC	보니를 알게 된 경로	유입 채널
Q2	SC	응답자 나이대	인구통계
Q3	SC	주 상담 자녀 연령대	인구통계
Q4	SC	첫 유입 시 주된 고민	트리거
Q5	SC	보니 사용 빈도	행동
Q6	Likert 5	AI 코칭 답변 만족도	만족도
Q7	MC (2)	AI 코칭의 아쉬운 점	페인
Q8	Likert 5	10분 온보딩 맞춤화 의향	의향
Q9	SC	전문가 매칭 이용 여부	전환
Q10	SC	전문가 매칭 이유 (Q9 ①②만)	전환 트리거
Q11	MC (2)	보니 외 사용하는 채널	대안
Q12	Likert 5	보니 사라짐 시 아쉬움	락인
Q13	SC	가족 플랜 9,900원 의향	확장
Q14	NPS 0~10	지인 추천 의향	NPS
Q15	Text	보니를 한 단어로 표현	정성

Table 8. 3명 응답자 객관식 비교 (🌙 새벽절박 / ⚡ 워킹맘 / 🖋️ AI회의형)

문항	🌙 P-80001 (김서영)	⚡ P-80008 (이수빈)	🖋️ P-80058 (한미래)
Q1 · 유입 경로	검색	지인 추천	지인 추천
Q2 · 나이대	30대 초반	30대 중반	40대 초반
Q3 · 자녀 연령	0~3개월	13~24개월	초1~3
Q4 · 첫 고민	수면·야제	훈육·떼쓰기	기타
Q5 · 사용 빈도	주 2~3회	주 2~3회	월 1회 이하
Q6 · 만족도	③ 보통	④ 만족	② 불만족
Q7 · 아쉬운 점	판단 정보 부족, 원론적	내 아이 맞춤 X	공감 부족
Q8 · 맞춤화 의향	④ 어느정도 있다	④ 어느정도 있다	③ 보통
Q9 · 전문가 이용	이용함	이용 안 함	고려했으나 안 함
Q12 · 사라짐	④ 아쉬움	③ 보통	② 별로
Q13 · 가족 플랜	긍정 검토	가격 낮으면 고려	필요 없음
Q14 · NPS	8	7	3
Q15 · 한 단어	"새벽 편의점"	"점심시간 동료"	"기계답게 기계적인"

세 응답자는 리커트에서 각각 ③·④·② 근처로 흩어져 있다. 객관식만 보면 "만족도가 응답자마다 다르다" 이상의 해석이 어렵다. Q15 주관식(한 단어)에서야 비로소 세 사람의 관점 차이가 드러나지만, 왜 그렇게 느끼는지는 여전히 알 수 없다. 대화형 설문은 바로 이 지점에서 개입한다.

7.2 대화형 설문 — 10문항 스펙과 하나의 질문이 펼쳐지는 방식

보니 대화형 설문은 10개 Base 질문으로 구성되고, 각 Base마다 최대 3단계의 Tail 프로빙이 이어진다 (표 9).

Table 9. 대화형 설문 10 Base 질문 스펙

#	축	Base 질문 요약	Tail 패턴
Q1	유입·트리거	처음 보니를 찾게 된 결정적 순간	상황 구체화 → 감정·선택 이유 → 첫 답변 감정
Q2	사용 습관	육아 고민 시 먼저 보니를 떠올리는가	구체 상황 구분 → 즉답 vs 경험담 → 공감 니즈
Q3	AI 만족·불만	만족·답답 순간 각 1건	만족 축(구체성) → 불만 축 → 기대 방향
Q4	맞춤화 니즈	아이 기질 기억, 10분 온보딩 수용성	현재 재입력 피로 → 변화 기대 → 데이터 저장 저항
Q5	전문가 전환	전문가 매칭 경험 및 감정 변화	결정 트리거 → 상담 차이 → 상담 후 변화
Q6	브랜드 의인화	보니가 사람이라면 어떤 사람	비유 근거 → 이상적 인격
Q7	대안·경쟁	보니 외 채널의 역할 분담	하나 포기 시 → 보니 고유 가치
Q8	사라짐 시나리오	보니가 사라지면 얼마나 아쉬울지	아쉬움 점수(0~10) → 시간 경과 시 대체성
Q9	가족 플랜	월 9,900원 가족 플랜 의향	망설임 본질 → 참여 유도 장치 → 조부모 확장
Q10	한 단어	보니를 한 단어 또는 짧은 문장으로	표현의 본질 확인

이 중 Q1(첫 유입) 하나가 실제로 어떻게 풀리는지 샘플 세션 80001(🌙 새벽절박·3개월)의 대화 그대로 옮긴다.

Figure A. Q1 대화형 설문 Tail 진화 — 세션 80001

AI (Base · turn 1) 처음 보니를 만나시게 된 순간을 한번 떠올려 보실 수 있을까요? 그때 어떤 상황이셨는지, 어떤 마음으로 보니를 찾게 되셨는지 기억나시는 장면이 있다면 편하게 말씀해주세요.

김서영 (30대 초반, 자녀 3개월) 새벽 3시쯤이었어요. 애가 백일 전인데 계속 울고 안 자서 저도 한숨도 못 자고 멘붕 와서 핸드폰으로 검색하다가 들어왔어요.

AI (Tail 1 · 상황 구체화) 새벽 3시에 혼자 깨어 계신다는 건 정말 체력적으로도 정신적으로도 한계에 가까운 시간이었겠어요. 그때 옆에 도와주실 수 있는 분은 안 계셨던 상황이었나요?

김서영 남편은 아침에 출근해야되니까 자고있고.. 깨우기도 미안하고 해서 그냥 혼자 애 안고 거실 왔다갔다 했어요. 근데 애는 계속 울고 저도 울고 있었어요 솔직히.

AI (Tail 2 · 선택 이유) 그 순간에 어머님도 함께 우셨다는 말씀이 너무 와닿네요. 그 새벽에 검색을 하시다가 맘카페나 블로그가 아닌 보니로 들어오시게 된 특별한 이유가 있으셨을까요?

김서영 카페는 글 올려도 답변이 몇시간 뒤에나 달리잖아요. 당장 어떻게 해야할지 알아야하는데... 누구라도 당장 답 주는게 필요했어요.

AI (Tail 3 · 첫 답변 후 감정) "당장 답을 주는 누군가"가 그 순간 가장 절실하셨군요. 그래서 보니의 첫 답변을 받으셨을 때, 마음속에 제일 먼저 올라온 감정은 어떤 거였는지 기억나세요?

김서영 솔직히 답은 바로 왔는데.. 너무 원론적이어서 오히려 더 허무했어요. '배고프거나 졸리거나 불편한 거예요' 이런거요. 그건 저도 알거든요 근데 뭘 해야하는지를 모르니까 물어본건데.

위 대화 한 번에 **새벽 · 고립 · 자책 · 24시간 즉답 · 원론적_답변 · 판단_정보_부족** 다섯 개 축이 동시에 드러난다. 객관식은 이 중 어느 하나도 직접 측정하지 못한다.

7.3 ELL 라벨링 결과 — 하나의 turn이 받는 9축 태그

같은 대화의 마지막 user 응답(T-80001-01-07)을 ELL이 어떻게 라벨링하는지 본다. 원문 그대로 입력되었을 때 라벨러(LLM 라벨러 + Pass A/B 자기검증)가 반환한 결과:

```
{
  "turn_id": "T-80001-01-07",
  "session_id": 80001,
  "q_idx": 1,
  "user_answer": "솔직히 답은 바로 왔는데.. 너무 원론적이어서 오히려 더 허무했어요. '배고프거나 졸리거나 불편한 거예요' 이런거요. 그건 저도 알거든요 근데 뭘 해야하는지를 모르니까 물어본건데",
  "labels": {
    "emotion": ["실망"],
    "behavior": [],
    "driver": [],
    "pain": ["원론적_답변", "판단_정보_부족"],
    "concern_category": [],
    "alt_channel": [],
    "conversion": null,
    "conversion_trigger": [],
    "disappear_severity": null,
    "family_plan_attitude": null,
    "confidence": "high",
    "free_notes": "원인 나열이 아닌 구체적 행동 지침을 원함"
  }
}
```

- emotion = [실망] — "허무했어요"에서 명시적 단서 포착
- pain = [원론적_답변, 판단_정보_부족] — 두 축 동시 언급
- confidence = high — Pass B 자기검증 통과, 규칙 R1(증거 기반) 충족

- `free_notes` — Closed Vocabulary에 잡히지 않는 뉘앙스를 자유서술

이 라벨은 다음 단계(§6.4 집계)의 `pain / emotion` 카운터에 1씩 기여한다. `turn_ids` 목록에 `T-80001-01-07` 가 저장되어, 최종 인사이트에서 이 응답이 곧바로 역추적된다.

7.4 인사이트 JSON — INS-02의 결정적 집계 payload

집계 단계에서 "원론적_답변" 언급 turn 478건, 응답자 98명이 식별되면 다음과 같은 인사이트 객체가 생성된다 (일부 필드 축약).

```

{
  "insight_id": "INS-02",
  "title": "원론적 답변이 가장 큰 불만",
  "fields_used": ["pain"],
  "primary": {
    "label": "pain=원론적_답변 언급 응답자",
    "respondent_count": 98,
    "turn_count": 478,
    "turn_ids": [
      "T-80001-01-07", "T-80002-01-07", "T-80003-01-07",
      "T-80004-01-07", "T-80005-01-05",
      "... (총 478건 중 일부)"
    ],
    "session_ids": [80001, 80002, 80003, 80004, 80005, 80006, 80007, 80008,
      "... (총 98명 중 일부)"]
  },
  "segments_by_persona": {
    "워킹맘 효율형": {"respondent_count": 14, "turn_count": 78},
    "기질 민감아 부모형": {"respondent_count": 13, "turn_count": 83},
    "학습·적응 고민형": {"respondent_count": 13, "turn_count": 62},
    "새벽 절박형": {"respondent_count": 12, "turn_count": 71},
    "부부 온도차형": {"respondent_count": 13, "turn_count": 51}
  },
  "base": {
    "scope": "전체 대화형 응답 (Q1~Q10)",
    "base_denominator_respondents": 100,
    "description": "10개 Base 질문 전체에 걸쳐 pain=원론적_답변 라벨이 붙은 응답을 집계."
  },
  "claim": "원론적 답변에 대한 불만은 98명(98.0%)에게서 확인되며, 478턴에 걸쳐 반복된다. 이는 사실상 전체 응답자의 공통 페인이다.",
  "derivation": "전체 100명 중 98명이 pain=원론적_답변 라벨을 받은 turn을 최소 1회 이상 생성. 페르소나 9종 모두에 걸쳐 고르게 분포하며 특정 세그먼트에 편중되지 않는다.",
  "so_what": "이 페인은 개인별 편차가 아닌 서비스의 구조적 한계에 가까우며, 답변 생성 측면의 개선 우선순위로 즉시 삼아야 한다.",
  "confidence_band": "high"
}

```

- `respondent_count·turn_count`는 `count_with_evidence()` 함수가 결정적으로 계산한 값이다. LLM이 숫자를 지어내지 않는다.
- `turn_ids[]`는 이 인사이트를 만든 증거 목록이다. FE 리포트에서 클릭하면 478개 원문이 즉시 펼쳐진다.
- `claim·derivation·so_what`은 LLM이 작성하되, 프롬프트에 이 payload만 전달된다(§3.3 Role Separation).

- `confidence_band`는 이 인사이트가 사용한 필드(`pain`) 중 가장 낮은 Jaccard 구간을 따라 자동 부여된다. `pain` 은 `high(0.870)`이므로 `high`.

7.5 요약 — 네 층위의 증거 사슬

지금까지 보인 질문 → 원문 → 라벨 → 인사이트는 하나의 끊기지 않는 사슬을 이룬다:

1. 객관식 Q6 = ③ 보통 (김서영)
2. 대화형 설문 Q1 Tail 3 = "너무 원론적이어서 오히려 더 허무했어요"
3. 라벨 = `pain: [원론적_답변, 판단_정보_부족]`, `confidence: high`
4. 인사이트 = INS-02 (98명 · 478 turns · high confidence)

ELL이 보장하는 것은 4단계 어느 지점에서든 직전 단계로 역행할 수 있다는 점이다. 인사이트 수치가 의심스러우면 `turn_ids` → 원문으로, 원문 해석이 의심스러우면 `labels` → 라벨러 규칙 R1~R8로, 라벨 신뢰도가 의심스러우면 Jaccard 재샘플링 결과로 각각 돌아간다. 이것이 본 보고서가 말하는 유리상자(Glass-box) 데이터 분석이다.

8. Evaluation & Discussion

8.1 오픈 vocabulary 요약 대비 환각 감소

ELL의 폐쇄형 어휘 제약은 정의상 허용 외 문자열의 출력 확률을 0으로 만든다. `sanitize()` 가 허용 외 값을 silent drop하기 때문에, 최종 집계에 환각된 라벨이 포함될 수 없다. 이는 CHAIR 지표상 구조적 환각률 0에 해당한다.

다만 이 설계는 사전에 정의되지 않은 신규 범주를 포착할 수 없다는 명확한 한계를 동반한다. ELL은 이를 두 가지 보완책으로 완화한다: - 라벨러가 자유 서술 슬롯 `free_notes` 에 기록한 내용을 주기적으로 검토하여 taxonomy 업데이트 - Topic Induction 단계에서 라벨 경계를 넘는 응답 클러스터를 발견

8.2 2-Pass Self-Check의 효과

보니 프로젝트의 Pass B 수정률은 27.4% (873건)으로, 1차 라벨링의 체크리스트 위반 사례를 후속 단계에서 실질적으로 포착한다. 이 구간(20~35%)은 V-STaR·RISE 계열 self-correction 연구가 보고하는 "의미 있는 자기 교정 범위"에 정확히 일치하며, 자기검증 단계가 관례적 절차가 아니라 실제 품질 향상에 기여함을 수치로 뒷받침한다.

8.3 Jaccard 일치도의 해석

10% 재샘플 Jaccard의 관측값은 필드의 본질적 난이도에 따라 분포한다. 일반적으로: - category_mention · pain 은 명시적 키워드와 1:1 대응되어 일치도가 가장 높다 - driver · behavior 은 한 응답에 여러 드라이버가 동시 존재해 일치도가 중간 수준이다 - emotion 은 감정 단어의 강도·뉘앙스 해석 차이로 일치도가 가장 낮다

ELL의 입장은 "모든 필드를 high로 만들자"가 아니라, "필드마다 본질적 불확실성이 다르다는 사실을 투명하게 공개한다"이다. 인사이트 카드에 붙는 신뢰도 배지는 이 정보를 의사결정자에게 직접 전달한다.

8.4 일반화 가능성

보니는 육아 도메인 사례이지만, ELL의 네 기둥은 도메인 독립적이다. 동일한 아키텍처는 다음과 같은 고위험 영역에도 이식 가능하다:

- 법률: 계약 조항 해석의 애매성을 Jaccard로 측정 + 근거 조문 역추적
- 의료: 환자 인터뷰 정성 코딩 + 증상 카테고리 폐쇄형 어휘 + 원문 증상 표현 역추적
- 금융: 신용 평가 질적 인터뷰 + 고정 등급 어휘 + 근거 발언 역추적
- HR: 직원 경험 서베이 + 사전 정의된 테마 라벨 + 익명 원문 역추적

8.5 한계

- 샘플 규모: 본 사례는 100세션으로, 양적 대표성이 아닌 심층 이해(In-Depth Understanding) 목적이다. 양적 가설 검증을 위해서는 이 결과를 기반으로 설계된 대규모 객관식 설문이 필요하다.
- 온톨로지 유지 비용: 폐쇄형 어휘는 주기적 업데이트를 전제로 한다. 새 라벨이 필요할 때마다 `free_notes` 검토 → taxonomy 확장 프로세스가 요구된다.
- LLM 확률성: 같은 입력에도 라벨링 결과가 미세하게 달라진다. Jaccard QC는 이 분산을 가시화하지만 제거하지는 못한다.
- 가공 샘플: 본 사례의 응답자는 LLM 라벨러이 페르소나 카드에서 생성한 가공 데이터다. 실제 고객 응답과는 분포 편향이 있을 수 있다.

9. Conclusion

증거와 숫자가 같은 곳을 가리키게 하는 일은 기술보다 태도의 문제에 가깝다. 모든 수치가 근거 응답으로 역추적 가능해야 하고, 해석의 불확실성은 숫자로 투명하게 공개되어야 하며, LLM은 서술에만 머물러야 한다. ELL은 이 세 원칙이 만나는 지점에 놓인 실무 파이프라인이다.

Dean Works는 이 방법론을 자사 서비스 운영 데이터 분석의 표준으로 채택했으며, 본 보고서의 사례 — 보니 대화형 설문 100세션 — 는 그 최초의 공개 적용 사례다. 우리는 이 접근이 정성 데이터 기반 의사결정의 신뢰 문제를 고민하는 리서처·제품 팀에게 작지만 유효한 출발점이 되기를 바란다.

References

본 보고서가 참조한 학술·산업 자료를 주제별로 그룹화한다. 모든 URL은 본 백서 작성 시점(2026-05) 기준 접근 가능한 공개 문서다. 개별 각주는 섹션별 상호참조로 연결된다 (예: §2.1 → [R1-R5]).

A. 제약적 디코딩 · 환각 평가 — [R1-R6]

- [R1] Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., & Saenko, K. (2018). **Object Hallucination in Image Captioning**. EMNLP 2018. <https://aclanthology.org/D18-1437/>
- [R2] ACL 2024 선정작. **Mitigating Open-Vocabulary Caption Hallucinations**. EMNLP Main 2024-1263. <https://aclanthology.org/2024.emnlp-main.1263.pdf>
- [R3] Zhu, Z., et al. (2024). **Mitigating Open-Vocabulary Caption Hallucinations (preprint)**. arXiv. <https://arxiv.org/html/2312.03631v3>
- [R4] OpenReview. **Mitigating Open-Vocabulary Caption Hallucinations — peer review**. <https://openreview.net/pdf/6ce41d2fce0c776a4e67f725caaab9c43073cd9d.pdf>
- [R5] Galileo AI. **Master LLM Summarization Strategies and Implementations**. <https://galileo.ai/blog/llm-summarization-strategies>
- [R6] Combating Multimodal LLM Hallucination via Bottom-Up Holistic Reasoning. arXiv 2412.11124. <https://arxiv.org/html/2412.11124v2>

B. Self-Correction · 에이전트 추론 — [R7-R13]

- [R7] Brookings Institution (2025). **Self-building Benchmarks: Using AI-generated Exams to Understand LLM Work Capabilities**. https://www.brookings.edu/wp-content/uploads/2025/11/Rio-Chanona_Einsiedler_Self-Building-Benchmarks_FINAL.pdf
- [R8] Dann Waneri (2024). **I Built a Local AI Agent That Audits My Own Articles**. dev.to. <https://dev.to/dannwaneri/i-built-a-local-ai-agent-that-audits-my-own-articles-it-flagged-every-single-one-pkh/comments>
- [R9] Microsoft Research (2025). **Agentic Reasoning and Tool Integration for LLMs via Reinforcement Learning**. <https://www.microsoft.com/en-us/research/wp-content/uploads/2025/04/AgenticReasoning.pdf>

- [R10] Qu, Y., Yang, T., Charan, P., et al. (2024). **Recursive Introspection: Teaching Language Model Agents How to Self-Improve (RISE)**. NeurIPS 2024.
<https://neurips.cc/virtual/2024/poster/96089>
- [R11] GeneBench: **Assessing AI Agents for Multi-Stage Inference Problems in Genomics**. bioRxiv.
<https://www.biorxiv.org/content/10.64898/2026.04.22.720113v1.full.pdf>
- [R12] Wei, J. (2024). **SelfCodeAlign: Self-Alignment for Code Generation**. NeurIPS 2024.
<https://neurips.cc/virtual/2024/poster/93079>
- [R13] Microsoft Copilot Studio advanced agent dev. Reddit r/copilotstudio.
https://www.reddit.com/r/copilotstudio/comments/1iunz1o/i_teach_advanced_copilot_studio_agent_development/

C. Jaccard · Inter-Rater Reliability — [R14–R18]

- [R14] Tan, L. (2024). **Validating Annotation Agreement between Humans and LLMs**. dsaid-govtech on Medium. <https://medium.com/dsaid-govtech/validating-annotation-agreement-between-humans-and-llms-bc334245b1d9>
- [R15] ResearchGate (2024). **Inter-rater Jaccard similarity coefficients including GPT-4 as annotator**. https://www.researchgate.net/figure/Inter-rater-Jaccard-similarity-coefficients-including-human-annotators-and-GPT-4-as_tbl1_381727895
- [R16] CrimRxiv (2024). **Policing Words with Machines: A Proof-of-Concept for LLM-Assisted Qualitative Analysis**. <https://www.crimrxiv.com/pub/iOvo0j0x>
- [R17] PMC (2024). **Automated Multitier Tagging of Chinese Online Health Education Resources Using LLM**. PubMed Central PMC12756663.
<https://pmc.ncbi.nlm.nih.gov/articles/PMC12756663/>
- [R18] arXiv (2026). **Understanding Dominant Themes in Reviewing Agentic AI-authored Code**. arXiv 2601.19287. <https://arxiv.org/html/2601.19287v1>

D. Visual Analytics · Drill-down · Grounded Theory — [R19–R24]

- [R19] Shneiderman, B. (1996). **The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations**. IEEE VIS 1996. (고전)
<https://www.cs.umd.edu/~ben/papers/Shneiderman1996eyes.pdf>
- [R20] Coscia, A. et al. (2025). **VisPile: Visual Analytics for Analyzing Multiple Text Documents with LLMs and Knowledge Graphs**. Georgia Tech VA.
<https://va.gatech.edu/endert/files/VisPileCoscia2025.pdf>

- [R21] arXiv (2025). **A Scoping Review of Mixed-Initiative Visual Analytics in the Automation Renaissance**. arXiv 2509.19152. <https://arxiv.org/html/2509.19152v1>
- [R22] Li, Z. et al. (2024). **iScore: Visual Analytics for Interpreting How Language Models Automatically Score Summaries**. arXiv 2403.04760. <https://arxiv.org/html/2403.04760v1>
- [R23] ATLAS.ti. **Thematic Analysis in UX Research — Guide & Examples**. <https://atlasti.com/guides/thematic-analysis/thematic-analysis-ux-research>
- [R24] Narechania, A. et al. (2025). **How Guidance from AI, Expert, or a Group of Analysts Impacts Visual Data Preparation and Analysis**. Georgia Tech VA. <https://va.gatech.edu/endert/files/Narechania2025-GuidanceSource.pdf>

E. 상용 플랫폼 · Drill-down 구현 사례 — [R25–R31]

- [R25] Avantia Inc. **NotebookLM: Next-Generation AI Research Assistant**. <https://avantia-inc.com/insights/notebooklm-your-next-generation-ai-research-assistant>
- [R26] Sopact Sense. **Intelligent Scoring: Turn Data Chaos into Instant ESG**. <https://www.sopact.com/use-case/intelligent-scoring>
- [R27] Zonka Feedback. **24 Best Text Analysis Tools in 2026**. <https://www.zonkafeedback.com/blog/text-analysis-tools>
- [R28] Yabble. **Gen — AI Research Assistant**. <https://www.yabble.com/gen>
- [R29] MyLens AI vs NotebookLM. **Comprehensive guide**. <https://mylens.ai/guides/notebooklm-vs-mylens-comprehensive>
- [R30] Sprinklr Help Center. **Gen AI Insights Features**. <https://www.sprinklr.com/help/articles/sprinklr-insights/sprinklr-gen-ai-insights-features/687fb6377455722cfd110422>
- [R31] ByteBridge (Medium). **Comparing Leading AI Deep Research Tools: ChatGPT, Google, Perplexity, Kompas AI, Elicit**. <https://bytebridge.medium.com/comparing-leading-ai-deep-research-tools-chatgpt-google-perplexity-kompas-ai-and-elicite-59678c511f18>

F. 엔지니어링 · Data Hygiene · Role Separation — [R32–R36]

- [R32] Gong.io (2024). **The AI Measurement Framework**. <https://www.gong.io/blog/the-ai-measurement-framework>

- [R33] Faros AI (2026). **Measuring Engineering Productivity in 2026.**
<https://www.faros.ai/blog/measuring-engineering-productivity-2026>
- [R34] Scout OS (2026). **AI Metrics Tracking: Key Insights & Best Practices.**
<https://www.scoutos.com/blog/ai-metrics-tracking-key-insights-and-best-practices>
- [R35] Dataro. **Fundraising analytics: 20 critical metrics & how to use them.**
<https://dataro.io/blog/fundraising-analytics-20-critical-metrics-how-to-use-them>
- [R36] Editage Insights. **Top AI Tools Every Researcher Should Know in 2026.**
<https://www.editage.com/insights/ai-tools-for-researchers>

G. UX · HCI — Drill-down 행동 연구 — [R37-R38]

- [R37] University of Southampton ePrints (2024). **Understanding Project Managers' Behaviour when using Artificial Intelligence for Project Control.**
https://eprints.soton.ac.uk/501517/1/Understanding_Project_Managers_Behaviour_when_using_Artificial_Intelligence.pdf
- [R38] PMC. **A qualitative research framework for the design of user-centered displays of explanations for machine learning in healthcare.**
<https://pmc.ncbi.nlm.nih.gov/articles/PMC7545557/>

H. 보니(bonny) · Dean Works 관련 내부 자료 — [R39-R42]

- [R39] Dean Works. **보니(bonny) 서비스 소개.** <https://hibonny.com>
- [R40] Dean Works ELL 리서치 자료실. <https://ell.dean.kr/downloads/v1/>
- [R41] Formative JMIR (2025). **Feasibility of a Specialized LLM for Postgraduate Medical Exam Prep.** <https://formative.jmir.org/2025/1/e77580>
- [R42] NVIDIA Developer Blog. **Configurable Graph-Based Task Solving with Marco Multi-AI Agent Framework.** <https://developer.nvidia.com/blog/configurable-graph-based-task-solving-with-the-marco-multi-ai-agent-framework-for-chip-design/>

총 42건 참고 자료. 각 출처는 본문 해당 섹션에 [R##] 형식으로 상호참조된다.

 데이터셋 · 리포트 다운로드

본 보고서에 사용된 모든 원본 데이터와 라벨링 산출물은 아래 경로에서 내려받을 수 있다. URL은 Dean Works 리서치 자료실에서 호스팅된다.

 대화형 설문 원본 (CSV) 100 세션 · 4,780 행 · 2.5 MB welcome + verify + chat + closing	 객관식 설문 응답 (CSV) 100 응답자 · 15 문항 · 23 KB session_id로 대화형 설문과 JOIN 가능
 ELL 라벨된 user turn (JSONL) 3,478 turn · 2.5 MB · 9축 라벨 2-Pass Self-Check 완료 버전	 12 핵심 인사이트 (JSON) 12건 · narrative·evidence 포함 · 64 KB claim · derivation · so_what 3단 서술
 토픽 induction 결과 (JSON) 14 토픽 · 3,478 turn 할당 3-pass 후 size ≥ 15 필터링	 10% Jaccard QC 보고서 (MD) 347 turn 샘플 · 7 필드 별 IRR Disagreement 사례 10건 포함
 세션 메타데이터 (JSON) 100 세션 페르소나·가구 정보 생성 seed · 페르소나 카드 참조	 본 백서 PDF (본 문서) 국문 · 약 30페이지 · 2 MB v1.0 · 2026-05 · Dean Works

라이선스 · 이용 - 본 데이터는 Dean Works가 방법론 시연을 위해 제작한 가공 샘플이다. 실제 보니 (bonny) 고객 응답은 포함되어 있지 않다. - 연구·교육·내부 리포트 목적의 재사용은 출처 표기(Dean Works, ELL v1.0, 2026) 조건으로 허용된다. - 문의: hckim@dean.kr · 031-778-8865

재현 가능성 - 샘플 생성 스크립트, 라벨링 파이프라인, FE 리포트 소스는 요청 시 제공한다. 본 보고서 모든 수치는 동일 seed·프롬프트에서 재현 가능하다 (LLM 라벨러 응답의 LLM 확률성만 분산 요인).

Appendix A. Closed Vocabulary 예시 (보니)

```

EMOTION = [
    "고립감", "자책감", "죄책감", "불안", "무력감", "분노", "짜증", "피로",
    "답답함", "실망", "회의",
    "안도", "공감받음", "신뢰", "만족", "감사", "기대",
    "양가감정", "체념",
]
]
BEHAVIOR = [
    "새벽_검색", "즉답_요청", "반복_확인", "비교_판단",
    "전문가_예약", "전문가_보류", "추천_공유", "기능_탐색",
    "이탈_고려", "맞춤화_수용",
]
]
DRIVER = [
    "24시간_즉답", "구체_절차_답변", "눈치_안보기", "익명성",
    "비용_부담_없음", "전문가_연결", "초보_친화",
    "다채로운_주제", "지인_추천_신뢰",
]
]
PAIN = [
    "원론적_답변", "맞춤화_부족", "공감_형식적", "판단_정보_부족",
    "매번_재입력", "답변_길이_부적절", "병원_회피성_답",
    "다자녀_미반영", "전문가_비용_부담", "전문성_한계_체감",
    "개인정보_불안", "UI_불편", "알림_과다",
]
]
# ... 이후 CONCERN_CATEGORY, ALT_CHANNEL, CONVERSION,
#     CONVERSION_TRIGGER, DISAPPEAR_SEVERITY, FAMILY_PLAN_ATTITUDE

```

Appendix B. 프롬프트 템플릿 (Pass A / Pass B)

Pass A — 시스템 프롬프트 개요

당신은 보니(bonny) 육아 코칭 AI 서비스의 대화형 응답을 정밀 라벨링하는 연구용 데이터 라벨러다. 각 turn을 다차원 Closed Vocabulary로 태깅한다. 태그는 아래 정의된 허용 값에서만 선택한다.

[허용 값 9개 축 나열]

[규칙 R1~R8: 증거 기반, 짧은 응답, emotion 최소, driver vs pain, conversion, disappear/family_plan, 중복 금지, confidence]

입력: [{turn_id, session_id, q_idx, q_axis, ai_context, user_answer}, ...]

출력: JSON 배열만. 각 요소 {turn_id, labels: {...}}

Pass B — 자기검증 프롬프트

방금 라벨링 결과를 자기 검증한다. 같은 응답을 다시 읽고, 수정이 필요한 태그만 고쳐라.

체크리스트:

- A) 본문에 근거 없는 라벨이 있는가?
- B) 감정 단어 없는데 emotion이 붙어있는가?
- C) 짧은 응답인데 라벨이 과다한가?
- D) driver/pain 분류가 뒤바뀌었는가?
- E) 허용 값 외 문자열이 있는가?
- F) Q5 무관 turn에 conversion 채워졌는가?
- G) Q8/Q9 무관 turn에 disappear/family_plan 채워졌는가?

수정한 경우 free_notes 앞에 "[REV]" 붙여 표시.

Appendix C. 인사이트 카드 UI 스키마

FE 리포트의 인사이트 카드는 다음 구조의 JSON 한 건을 소비한다. 본문 필드 (claim · derivation · so_what)는 LLM 라벨러가 작성하고, 나머지 수치 필드는 결정론적 집계 코드가 채운다.

```

{
  "insight_id": "INS-01",
  "title": "새벽 3시의 즉답이 보니를 붙잡는 가장 큰 이유",
  "fields_used": ["driver"],
  "primary": {
    "label": "24시간 즉답·눈치 없음 언급 응답자",
    "respondent_count": 68,
    "turn_count": 182,
    "turn_ids": ["T-80002-01-04", "T-80005-02-06"],
    "session_ids": [80002, 80003]
  },
  "segments_by_persona": {
    "새벽 절박형": {"respondent_count": 12, "turn_count": 58},
    "워킹맘 효율형": {"respondent_count": 10, "turn_count": 31},
    "기질 민감아 부모형": {"respondent_count": 9, "turn_count": 24}
  },
  "base": {
    "scope": "전체 인터뷰 (all questions)",
    "base_denominator_respondents": 100,
    "description": "10개 Base 질문 전반에서 언급된 응답을 집계."
  },
  "claim": "새벽 3시의 즉답이 보니를 붙잡는 1순위 이유다 (68/100).",
  "derivation": "driver=24시간_즉답 라벨을 1회 이상 생성한 응답자 68명. 전 페르소나에 분포.",
  "so_what": "락인의 핵심은 기능 세트가 아니라 새벽 시간대 응답 가용률이다."
}

```

실제 `session_ids` 배열은 68개, `turn_ids` 는 182개의 원소를 가진다. 위 예시는 지면 축약이며, 전체 JSON은 ell.dean.kr/downloads/v1/insights.json 에서 내려받을 수 있다.

Appendix D. QC Disagreement Samples (Jaccard < 0.6)

10% 재샘플 QC에서 일치도가 0.6 미만인 사례 중 대표 10건. 전체 저일치 케이스는 482건.

D1. T-80075-10-03 · 필드 `emotion` · Jaccard 0.5

아... 그게요, 제가 아이 일로 멘붕 와서 보니 열 때 보면요, 마음이 막 엉켜있거든요. 아이한테 화난 건지, 내가 속상한 건지, 죄책감인지... 뒤죽박죽이예요. 근데 보니랑 얘기하다 보면 그게 하나씩

분리가 돼요. '아 나는 지금 아이가 걱정되는 거구나', '나는 내가 엄마로서 부족한 것 같아서 불안한 거구나' 이렇게요. 고장난 걸 고치는 건 아닌데, 뭐랄까

- Run1 라벨: 공감받음, 불안, 자책감
- Run2 라벨: 불안, 양가감정, 자책감

D2. T-80092-08-03 · 필드 emotion · Jaccard 0.0

아 그게요ㅋㅋ 맘카페는 진짜 주관적이잖아요. '우리 애는 이랬어요' 이런거. 근데 보니는 뭔가 우리 첫째 성격이랑 상황 알고 얘기해주니까... 그게 없으면 좀 그럴것 같아요. 그리고 남편한테 말하기 애매한거 있잖아요, 남편은 무조건 안된다 이러니까 대화가 안 통하는데 보니한테 그냥 편하게 털어놓을 수 있어서요

- Run1 라벨: -
- Run2 라벨: 공감받음, 답답함

D3. T-80061-10-03 · 필드 emotion · Jaccard 0.5

아 그게... 보니가 '이건 황달이에요, 병원 가세요' 이렇게 딱 진단해주진 않잖아요. 근데 '이런 증상이면 보통 이렇고, 이 정도면 지켜봐도 되고' 이런 식으로 방향은 알려주니까요. 그니까 응급실 앞에서 '어느 과로 가세요~' 해주는 그 분? 딱 그 느낌이에요ㅋㅋ 근데 새벽엔 그 분이 제일 고맙더라구요 솔직히

- Run1 라벨: 감사
- Run2 라벨: 감사, 안도

D4. T-80063-03-01 · 필드 emotion · Jaccard 0.0

음... 도움됐던 건요, 둘째 배앓이로 밤새 울 때 트림 자세 몇 가지 알려준 거요. 글자로만 봐도 바로 따라해볼 수 있게 자세히 써있어서 좋았어요. 아쉬운 건... 첫째 질투 문제 물어봤을 때요. '큰아이 마음도 헤아려주세요~' 이런 식으로 나오는데 그거 몰라서 물어본 게 아니잖아요ㅠㅠ

- Run1 라벨: -
- Run2 라벨: 답답함, 만족

D5. T-80065-03-05 · 필드 emotion · Jaccard 0.0

둘 다요 사실.. 식판법이라고 들어봤거든요 맘카페에서. 아기 앞에 여러 가지 놓고 스스로 집게 한다는거. 근데 보니한테 그거 물어보니까 '아이 주도 이유식도 좋은 방법입니다' 이렇게 원론적으로만 답해서.. 나는 식판 크기, 뭘 올려야 되는지, 몇 분 기다려야 되는지 그게 궁금했는데

- Run1 라벨: -
- Run2 라벨: 답답함

D6. T-80054-07-05 · 필드 emotion · Jaccard 0.33

음... 대체할 게 없는 것 같아요 솔직히. 네이버로 돌아가면 또 '사춘기는 기다려주세요' 이런 뻔한 답 보면서 짜증내고 있겠죠. 오은영쌤 영상은 일방적이고요. 제일 아쉬운 건... 저를 '엄마 지아'가 아니라 그냥 '지아'로 봐주는 곳이 없다는 거요. 남편도 그렇게 안 봐주는데.

- Run1 라벨: 고립감, 짜증
- Run2 라벨: 고립감, 답답함

D7. T-80093-08-03 · 필드 emotion · Jaccard 0.5

아... 그니까 평상시 육아 고민은 제가 알아서 해요. 밥 안 먹는다 숙제 안 한다 이런 건. 근데 딸이 SNS에서 친구랑 뭐 일 터졌을 때요, 그때는 진짜 어디 물어볼 데가 없거든요. 친정엄마는 '폰 뺏어' 이려고, 남편은 관심도 없고... 그 순간에 보니 없으면 저 진짜 혼자예요.

- Run1 라벨: 고립감, 불안
- Run2 라벨: 고립감

D8. T-80092-03-05 · 필드 emotion · Jaccard 0.5

아 그거는요... 제가 '스마트폰 사주면 유해사이트 차단 어떻게 해요' 물어봤는데 답이 너무 일반적이었어요. '구글 패밀리링크 쓰세요' '스크린타임 설정하세요' 이렇게요. 근데 저는 진짜 어떤 메뉴 들어가서 어떤 버튼 누르는지 그런게 궁금한거거든요ㅠㅠ 결국 유튜브 찾아봤어요

- Run1 라벨: 실망
- Run2 라벨: 답답함, 실망

D9. T-80061-08-05 · 필드 emotion · Jaccard 0.33

아 그 상상만 해도 좀 싫은데요ㅠㅠ 음... 아마 다시 맘카페랑 블로그 뱅뱅이 돌 것 같아요. 아니면 그냥 애 안고 응급실 갈지 말지 혼자 고민하다가 날 새겠죠. 남편은 출장 중이고 친정엄마 깨우기도 그렇고... 생각해보니까 7점이 아니라 8점 줘야되나ㅋㅋㅋ 아 몰라요 7.5점.

- Run1 라벨: **고립감, 무력감**
- Run2 라벨: **고립감, 불안**

D10. **T-80084-08-05** · 필드 **emotion** · Jaccard 0.5

음.. 밤에 애 재우고 나서 갑자기 불안해질때요. 그때 소아과 못가고 유튜브는 일방적이잖아요. 체크리스트 물어봐주고 '이건 정상, 이걸 체크해보세요' 이런식으로 바로바로 대답해주는거.. 그게 제일 아쉬울것같아요. 말문 트이는거 도와줄 사람이 없어지는 느낌?

- Run1 라벨: **불안**
- Run2 라벨: **고립감, 불안**

주: 저일치 케이스는 주로 emotion·behavior 필드. 인사이트 해석 시 신뢰도 뱃지(low)가 자동 부여됨.

Appendix E. 100 페르소나 카드 요약표

전체 100명의 세션 메타데이터. 상세 페르소나 카드는 요청 시 제공 가능.

session_id	이름	페르소나	나이대	성별	자녀 연령	한 단어 힌트
80001	김서영	새벽 절박형	30대 초반	F	3개월	새벽 편의점 — 24시간 열려있음이 본질, 특별하진 않음
80002	박지은	새벽 절박형	20대 후반	F	6주	구조선 — 빠져 죽을 것 같을 때 붙잡는
80003	한지원	새벽 절박형	30대 초반	F	2개월	내 편 — 유일하게 시어머니 간섭 없이 조언 받는 통로
80004	윤예린	새벽 절박형	30대 중반	F	4개월	당직의사 — 병원 갈 정도는 아닐 때 물어보는
80005	최은지	새벽 절박형	20대 후반	F	10주	야간 약국 — 급할 때 있어주지만 한정적
80006	신혜림	새벽 절박형	30대 초반	F	5개월	새벽 핫라인 — 아무도 없을 때 거는 전화
80007	임다현	새벽 절박형	30대 중반	F	7개월	육아 위키 — 검색 가능한 지식창고
80008	이수빈	워킹맘 효율형	30대 중반	F	14개월	점심시간 동료 — 빠르게 조언만 얻음
80009	정현주	워킹맘 효율형	30대 후반	F	5살	긴급 참고서 — 바쁠 때만 펼치는
80010	오수진	워킹맘 효율형	30대 후반	F	4살	이모티콘 — 딱 그 순간 쓰고 잊히는
80011	강민주	워킹맘 효율형	30대 중반	F	28개월	발달 체크 앱 — 정기검진 같은 느낌
80012	조예나	워킹맘 효율형	30대 중반	F	18개월	편의 매장 — 필요할 때 바로
80013	배수연	워킹맘 효율형	30대 후반	F	6살	빠른 검색 — 구글 대체
80014	송지민	워킹맘 효율형	30대 중반	F	30개월	휴대용 매뉴얼
80015	서하영	워킹맘 효율형	30대 후반	F	5살	출근길 팟캐스트

session_id	이름	페르소나	나이대	성별	자녀 연령	한 단어 힌트
80016	문소희	기질 민감아 부모형	30대 중반	F	4살	심리 상담사 인턴 — 듣긴 해도 속 깊진 않은
80017	양주희	기질 민감아 부모형	30대 후반	F	초3	이해해주는 선생님
80018	황나영	기질 민감아 부모형	30대 중반	F	22개월	방파제 — 거친 말 대신 막아주는
80019	노유진	기질 민감아 부모형	30대 후반	F	7살	조용한 보호자
80020	백채은	기질 민감아 부모형	30대 후반	F	초1	내 편이 되어주는 사람
80021	고민정	기질 민감아 부모형	30대 중반	F	32개월	감각 통역사
80022	주혜수	기질 민감아 부모형	30대 후반	F	6살	따뜻한 관찰자
80023	류지영	기질 민감아 부모형	30대 중반	F	초2	안전한 대기실
80024	이준호	부부 온도차형	40대 초반	M	5살	부부 사이 중재자
80025	박성민	부부 온도차형	40대 중반	M	초3	과제 알림 앱
80026	윤동혁	부부 온도차형	40대 초반	M	3살	업무 매뉴얼
80027	전보라	부부 온도차형	30대 후반	F	7살	일관성 조율자
80028	배준영	부부 온도차형	30대 후반	M	28개월	아내가 보내준 앱
80029	김태훈	부부 온도차형	40대 중반	M	초5	부자 번역기
80030	정우진	부부 온도차형	30대 후반	M	4살	임시 매뉴얼

session_id	이름	페르소나	나이대	성별	자녀 연령	한 단어 힌트
80031	임진영	부부 온도차형	30대 후반	M	초1	하루 점검표
80032	오재영	 첫째 초보형	30대 초반	M	8개월	부모 입문서
80033	신재민	 첫째 초보형	30대 초반	M	4개월	초보 지도서
80034	서태민	 첫째 초보형	30대 중반	M	14개월	발달 체크북
80035	원아름	 첫째 초보형	30대 초반	F	3개월	수유 안내서
80036	지혜린	 첫째 초보형	20대 후반	F	6주	든든한 첫 장
80037	석희정	 첫째 초보형	30대 초반	F	16개월	부모 1년차 교재
80038	심예원	학습·적응 고민형	30대 후반	F	6살	학습 코치
80039	최현수	학습·적응 고민형	40대 초반	M	초2	홈스쿨 가이드
80040	차윤서	학습·적응 고민형	30대 후반	F	초1	학습 설계사
80041	염수정	학습·적응 고민형	40대 초반	F	초3	선택 도우미
80042	장미경	학습·적응 고민형	40대 초반	F	초2	독서 큐레이터
80043	한상욱	학습·적응 고민형	40대 중반	M	초3	영어 선생
80044	구하린	학습·적응 고민형	40대 초반	F	초1	첫 주 동반자
80045	허은수	학습·적응 고민형	40대 중반	F	초2	중재자

session_id	이름	페르소나	나이대	성별	자녀 연령	한 단어 힌트
80046	남지혜	스마트폰·미디어 고민형	40대 초반	F	초4	대화 연습장
80047	홍세영	스마트폰·미디어 고민형	40대 초반	F	초5	미디어 가이드
80048	유채린	스마트폰·미디어 고민형	30대 후반	F	초3	타협 코치
80049	강채원	스마트폰·미디어 고민형	40대 초반	F	초6	밤 보초
80050	곽소연	스마트폰·미디어 고민형	30대 후반	F	초2	대체 활동 추천기
80051	강석호	스마트폰·미디어 고민형	40대 중반	M	초5	진로 탐색 파트너
80052	조현성	스마트폰·미디어 고민형	40대 중반	M	초6	안전 경계선
80053	반현아	사춘기 초입 갈등형	40대 중반	F	초6	사춘기 통역사
80054	민지아	사춘기 초입 갈등형	40대 초반	F	초5	모녀 다리
80055	표민지	사춘기 초입 갈등형	40대 중반	F	초6	적절한 거리 측정사
80056	봉은영	사춘기 초입 갈등형	40대 초반	F	초5	첫사랑 멘토
80057	탁지연	사춘기 초입 갈등형	40대 중반	F	초6	초경 안내자
80058	한미래	AI 회의형	40대 초반	F	초3	기계답게 기계적인
80059	공지선	AI 회의형	40대 중반	F	초5	급할 때 임시방편
80060	변주은	AI 회의형	30대 후반	F	5살	기계 안내원

session_id	이름	페르소나	나이대	성별	자녀 연령	한 단어 힌트
80061	박수아	새벽 절박형	30대 초반	F	3개월	24시간 응급실 안내원
80062	이은별	새벽 절박형	20대 후반	F	5주	피신처
80063	김유림	새벽 절박형	30대 중반	F	7주	이중 상담실
80064	정서은	새벽 절박형	30대 후반	F	2개월	중재 판정인
80065	최아린	새벽 절박형	20대 후반	F	10개월	구조대
80066	한수현	워킹맘 효율형	30대 후반	F	20개월	출근길 퀵 가이드
80067	오지유	워킹맘 효율형	30대 중반	F	15개월	출근 전 판정관
80068	윤서운	워킹맘 효율형	30대 후반	F	4살	시간표 코치
80069	신가은	워킹맘 효율형	30대 중반	F	26개월	점심시간 속성 과외
80070	임다온	워킹맘 효율형	30대 후반	F	3살	퀵 컨설트
80071	강이서	워킹맘 효율형	30대 후반	F	초1	퇴근 후 체크리스트
80072	서나은	기질 민감아 부모형	30대 후반	F	3살	감각 번역사
80073	조하울	기질 민감아 부모형	30대 후반	F	5살	가족 중재자
80074	배예서	기질 민감아 부모형	30대 중반	F	5살	조용한 지지자
80075	송다연	기질 민감아 부모형	40대 초반	F	초3	마음 정비사

session_id	이름	페르소나	나이대	성별	자녀 연령	한 단어 힌트
80076	장혜린	기질 민감아 부모형	30대 후반	F	초1	이해의 첫 걸음
80077	최현수2	부부 온도차형	40대 초반	M	6살	아빠 입문서
80078	전수호	부부 온도차형	40대 초반	M	초2	비상시 매뉴얼
80079	도예지	부부 온도차형	30대 후반	F	4살	부녀 다리
80080	방지훈	부부 온도차형	40대 중반	M	초4	부자 번역기
80081	손나연	부부 온도차형	30대 후반	F	6살	편애 거울
80082	류지안	 첫째 초보형	30대 초반	F	5개월	첫 이유식 코치
80083	허도연	 첫째 초보형	30대 중반	F	12개월	첫 돌 가이드
80084	윤하린	 첫째 초보형	30대 초반	F	20개월	말문 도우미
80085	우민서	 첫째 초보형	30대 초반	F	2개월	첫 접종 파트너
80086	구서진	학습·적응 고민형	40대 초반	F	초2	첫 문장 친구
80087	표수영	학습·적응 고민형	30대 후반	F	초1	친구 관계 튜터
80088	노아진	학습·적응 고민형	40대 중반	F	초3	수학 자신감 회복 코치
80089	정이경	학습·적응 고민형	40대 초반	F	6살	한글 걸음마
80090	선하경	학습·적응 고민형	40대 중반	F	초3	집중 코치

session_id	이름	페르소나	나이대	성별	자녀 연령	한 단어 힌트
80091	차유정	스마트폰·미디어 고민형	40대 초반	F	초2	콘텐츠 가이드
80092	원보연	스마트폰·미디어 고민형	40대 초반	F	초4	스마트폰 도입 가이드
80093	함소진	스마트폰·미디어 고민형	40대 초반	F	초5	디지털 균형 코치
80094	고아연	스마트폰·미디어 고민형	30대 후반	F	초3	중독 경계선
80095	신유나	사춘기 초입 갈등형	40대 중반	F	초5	거울 번역사
80096	유정화	사춘기 초입 갈등형	40대 초반	F	초6	감정 완충기
80097	권다슬	사춘기 초입 갈등형	40대 초반	F	초5	엄마와 함께하는 초경
80098	양세린	사춘기 초입 갈등형	40대 중반	F	초6	학교로 가는 길잡이
80099	박현희	AI 회의형	40대 중반	F	초4	기계적 상담원
80100	민예림	AI 회의형	30대 후반	F	5살	자동응답기

본 보고서의 모든 수치는 FE 리포트에서 원문 응답으로 역추적 가능합니다.
<https://ell.dean.kr/insights> · <https://ell.dean.kr/compare> · <https://ell.dean.kr/explorer> · <https://ell.dean.kr/voronoi>