RESEARCH REPORT · DEAN WORKS

VOLUME 01 · 2026 · ELL V1.0 · ENGLISH EDITION

A Verifiable Pipeline for Quantifying Conversational Qualitative Data

Evidence-Linked Labeling (ELL)

A reproducible framework that turns conversational-survey responses into auditable quantitative indicators.

The LLM labels; code does the counting. Every reported number is one click away from the source respondent turn, and the labeler's residual uncertainty is disclosed as Jaccard values — not hidden. Applied here to a 100-session conversational survey at bonny, Dean Works's AI parenting-coaching service.

100	3,478	12
SESSIONS	LABELED TURNS	INSIGHTS
same 100 respondents · dual-format	9 closed-vocab axes · 2-Pass QC	every number traceable to source

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
READING	~25 pages · English edition

Abstract

Conversational surveys capture what multiple-choice instruments never can: emotional arcs, decision triggers, and the narrative context of service adoption and churn. But once this unstructured text lands on the analyst's desk, a familiar wall appears. Behind a generative AI's fluent summary — *"most customers report fatigue with the search feature"* — the analyst cannot answer three questions in one breath: how many is *most*, what did those respondents actually say, and which source turn produced that sentence?

This report introduces **Evidence-Linked Labeling (ELL)**, a methodology designed and implemented by Dean Works. ELL is built on four pillars: (1) constrained decoding over a closed vocabulary, (2) architectural separation of narrative generation from numeric aggregation, (3) a two-pass self-check coupled with Jaccard-based inter-rater reliability disclosure, and (4) a drill-down interface that anchors every reported number to the underlying respondent turns. Each pillar draws on an established research thread in natural language processing, visual analytics, or grounded theory; ELL composes them into a single pipeline.

We apply ELL to **bonny**, Dean Works's AI parenting-coaching service, analyzing a **100-session / ~4,500-turn conversational survey** alongside a matched 15-question multiple-choice survey completed by the same respondents. The case study quantifies contexts that the multiple-choice instrument could not surface — for example, *"I kept weighing the cost against diaper money, then gave up on the expert match,"* or *"At 3 a.m. with suspected neonatal jaundice, the only answer that came back was 'go to the hospital'."* Ten-percent resample Jaccard values, published per field, were **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)**.

ELL does not pretend to produce perfect labels. Instead it makes the magnitude of uncertainty explicit as numbers and makes every conclusion traceable to source responses. This is our practical and scholarly response to the trust crisis facing qualitative data analysis in an era of AI hallucination and black-box summarization.

> 📊 **Live companion — ell.dean.kr** Every number in this PDF is a one-click drill-down away from the underlying respondent turn in the FE report: **Insights** ell.dean.kr/insights · **Question-level findings** ell.dean.kr/questions · **Closed vs conversational comparison** ell.dean.kr/compare · **First-turn vs full-tail experiment** ell.dean.kr/tail-effect · **Full explorer** ell.dean.kr/explorer.

1. Introduction — Problem Statement

1.1 The Rise of Conversational Surveys and the Analysis Bottleneck

One of the more consequential shifts in the research industry over the last two to three years has been the move **from static surveys, where respondents pick from a fixed list of options, to conversational surveys, where an AI interviewer reads each response in real time and generates the next question**. This tradeoff sacrifices some of the clean tabulability of multiple-choice instruments in exchange for something multiple-choice cannot deliver: the respondent's own answer to *why*.

Consider a mother's account at bonny, captured when she searched for help at 3 a.m. while her newborn refused sleep. The multiple-choice form records only `① concern about sleep training`. The conversational survey records the verbatim: *"My husband has work in the morning so he was asleep and I felt bad waking him, so I just carried the baby around the living room by myself. She kept crying and I was crying too, honestly."* The latter is far richer. It is also an analysis problem: how do you process this kind of text at the scale of 60, 100, or 300 respondents?

1.2 The Structural Deficiencies of LLM Summarization

Most practitioners solve this problem by handing the entire corpus to a general- purpose LLM (ChatGPT and its peers) with the instruction *"summarize the responses above."* The output reads smoothly at the prose level, but exhibits three structural deficiencies:

1. **Hallucination.** Counts appear that the respondents never produced. *"Most customers report fatigue with the search feature"* — ask how many is *most*, and the model cannot answer.
2. **Absence of source grounding.** The summary sentence cannot be traced back to the original turn that justifies it. This is the very question that stakeholders raise the moment they try to act on the summary.
3. **Numeric imprecision.** LLMs are probabilistic language-pattern matchers, not symbolic calculators. Asked *"how many respondents meet condition X?"* they produce a number that is contextually plausible rather than factually correct.

These are not limitations of a specific model. They are structural properties of autoregressive language models, and they place a fundamental ceiling on how far such models can be trusted in high-stakes, high-accuracy domains.

1.3 Contributions

ELL is a reproducible pipeline that converts conversational qualitative data into verifiable quantitative indicators. This report contributes:

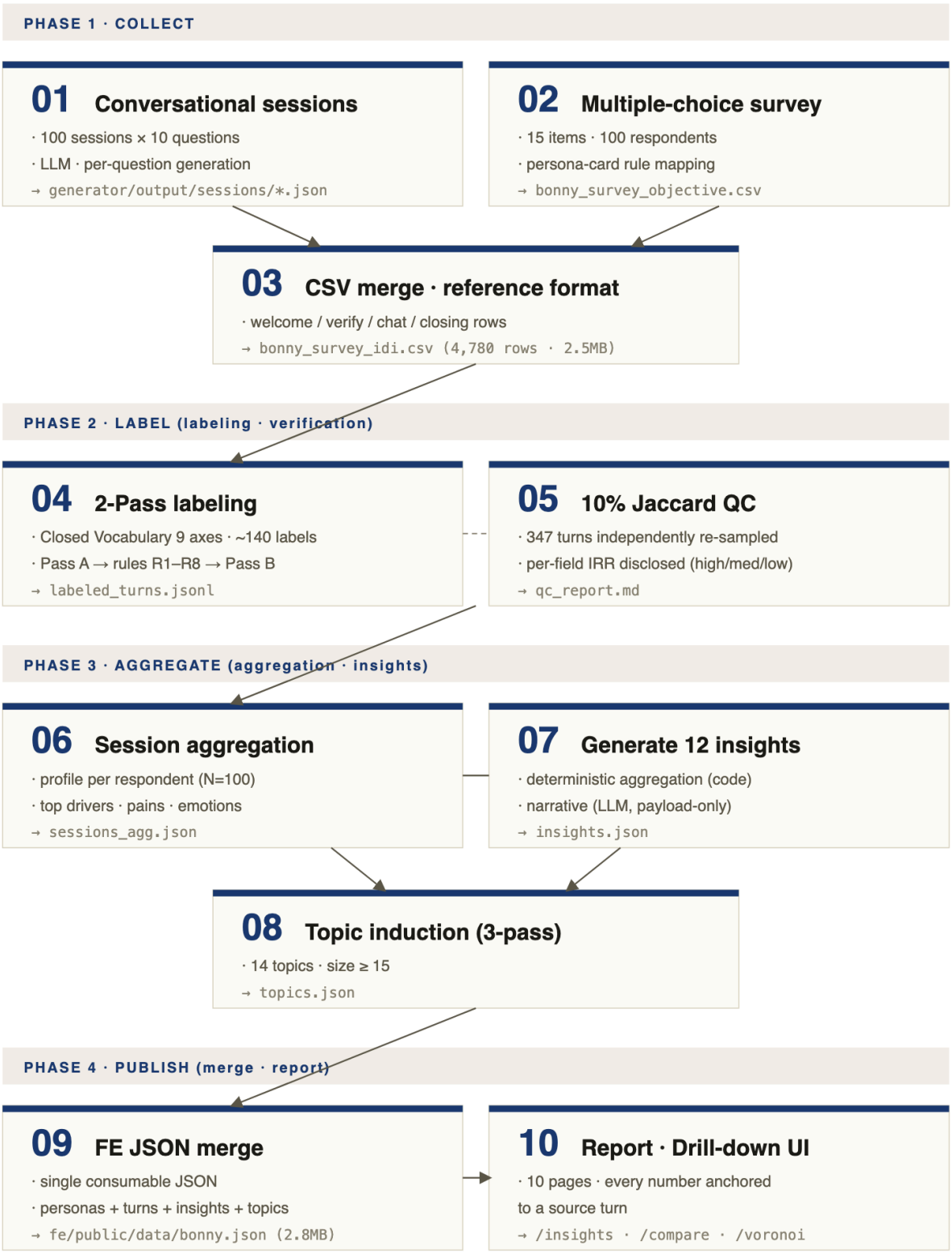
- An **architectural proposal** that fuses four independent research threads — constrained decoding, role separation, iterative verification, and evidence tracing — into a single practitioner

methodology.

- The **concrete execution record** of the pipeline on 100 sessions / ~3,478 labeled turns from bonny, including per-field Jaccard values.
- A **paired comparison** against the 15-question multiple-choice survey completed by the same 100 respondents, quantifying what ELL captures that multiple-choice cannot.
- A **front-end report** where every reported number drills down to the underlying respondent quote, designed for executive stakeholders who will not accept unverifiable figures.

Figure 1 · ELL Pipeline

10 stages. Every output is the next stage's input; each stage can run and re-run independently.



Every artifact is stored append-only and resumable — a mid-run failure restarts from the last completed point.
Each stage is an independent Python/Node script; re-running one stage does not affect downstream stages already written.

Figure 1 · ELL Pipeline — from raw data to report

---

2. Related Work

2.1 Constrained Decoding and Hallucination Evaluation

Object-hallucination research traces back to Rohrbach et al. (2018) and the CHAIR metric [R1], which quantifies hallucination rates by checking whether the object names a model produces appear in a predefined ground-truth list. This approach is, at its core, *"restrict the model's output space to a closed vocabulary."* Numerous subsequent models (UD-L, CIIC, TLC, ObjMLM, Woodpecker) have extended the framework. More recent work on open-vocabulary evaluation (BLIP-2, OpenCHAIR) has expanded the horizon, but **industry practice in high-stakes domains still favors closed vocabularies for hallucination control** [R2–R4].

2.2 Self-Correction and Agentic Reasoning

V-STaR, RISE, and DeepSeek-R1's GRPO algorithm demonstrate that **multi-stage verification and refinement** reduces the intrinsic errors of single-pass LLM generation [R9, R10]. RISE in particular separates candidate-generation, error- detection reward, and final-refinement models, reporting monotonic performance gains in the process. ELL's two-pass self-check applies the same pattern to the labeling task.

2.3 Jaccard Agreement in Multi-Label Classification

In qualitative coding, a single response typically receives multiple labels. Accuracy is meaningless as an evaluation metric in this setting. The **Jaccard similarity coefficient** — intersection over union of the two annotator label sets — measures partial agreement mathematically. Recent work on multi-label GitHub PR review [R15], educational feedback surveys with GPT-4 as annotator [R14], and LLM-assisted qualitative coding in criminology [R16] has made Jaccard the de-facto standard for human–AI inter-rater reliability.

2.4 Grounded Theory and Visual Analytics

Shneiderman's *"Overview first, zoom and filter, then details-on-demand"* [R19] remains the classical statement of information-seeking. ELL's drill-down UI applies this principle to qualitative research. Simultaneously, grounded theory's requirement of iterative movement between raw data and macro themes is implemented as a digital interface. Narechania et al. (2025) [R24] provide empirical evidence that expert analysts consult AI summaries for initial exploration but always verify at the raw-evidence level when a judgment is about to be made. Commercial platforms (Sopact, Thematic, Yabble,

Google NotebookLM) have built similar evidence-tracing principles into their products [R25, R26, R28, R29].

---

3. The ELL Framework

3.1 Design Principles Overview

ELL consists of four pillars. Each is valid on its own; only **when they are combined in a single pipeline does the overall claim hold**.

Pillar	Problem it addresses	Research lineage
P1. Closed Vocabulary	Infinite output space → un-aggregatable labels	CHAIR-family constrained decoding
P2. Role Separation	Narrative and numbers co-mingle → hallucinated counts	Deterministic pipeline + LLM narrative layer
P3. Self-Check + Jaccard	Labeler uncertainty becomes black-box	2-Pass self-correction + multi-label IRR
P4. Drill-down	Numbers detached from source evidence	Grounded theory + Shneiderman mantra

3.2 P1. Closed Vocabulary Constraint

The set of permitted label values lives in a single `taxonomy.py`. For bonny, nine axes × an average of 15 values yields roughly 140 labels that enclose the entire label space:

```
EMOTION   = [isolation, self-blame, guilt, anxiety, helplessness, anger,
             irritation, fatigue, frustration, disappointment, skepticism,
             relief, feeling-heard, trust, satisfaction, gratitude, expectation,
             ambivalence, resignation]
PAIN      = [generic-answer, lack-of-personalization, empty-empathy,
             insufficient-judgment-info, repeated-re-entry, inappropriate-length,
             hospital-deflecting-answer, no-multi-child-support,
             expert-cost-barrier, perceived-expertise-limit, privacy-concern,
             UI-friction, notification-fatigue]
DRIVER    = [24h-instant-answer, concrete-step-by-step, no-judgment,
             anonymity, no-cost-burden, expert-connection, beginner-friendly,
             diverse-topics, peer-recommendation-trust]
BEHAVIOR  = [late-night-search, instant-answer-request, re-confirmation,
             comparison-judgment, expert-booking, expert-deferral,
             recommendation-share, feature-exploration, churn-consideration,
             personalization-acceptance]
# ... plus CONCERN_CATEGORY, ALT_CHANNEL, CONVERSION,
#          CONVERSION_TRIGGER, DISAPPEAR_SEVERITY, FAMILY_PLAN_ATTITUDE
```

A `validate()` function returns a list of errors when a labeler response contains strings outside the permitted set; `sanitize()` silently drops them. Together these two functions are the enforcement mechanism that **injects the closed-vocabulary constraint into the pipeline**.

3.3 P2. Architectural Separation of Narrative and Numeric Computation

The insight-generation routine implements this principle at the code level:

```
# Step 1: deterministic aggregation — performed by code
m = count_with_evidence(turns, lambda t: has_pain(t, "generic-answer"))
# → {"turn_count": N, "respondent_count": M,
#    "turn_ids": [...], "session_ids": [...]}

# Step 2: narrative — performed by the LLM, using payload numbers only
payload = [{
    "insight_id": ...,
    "respondent_count": m["respondent_count"],
    "base_denominator_respondents": 100,
    ...
}]
# System prompt: "Quote only the respondent_count and
# base_denominator_respondents fields from payload. Do not introduce
# any number that does not appear in payload."
```

The LLM does not compute numbers. It receives numbers that have already been computed and writes the three-part scholarly narrative (`claim · derivation · so_what`). This separation is philosophically identical to the "data hygiene" architecture adopted by Gong.io's AI measurement framework [R32], Faros AI's engineering productivity analysis [R33], and Dataro's fundraising analytics [R35].

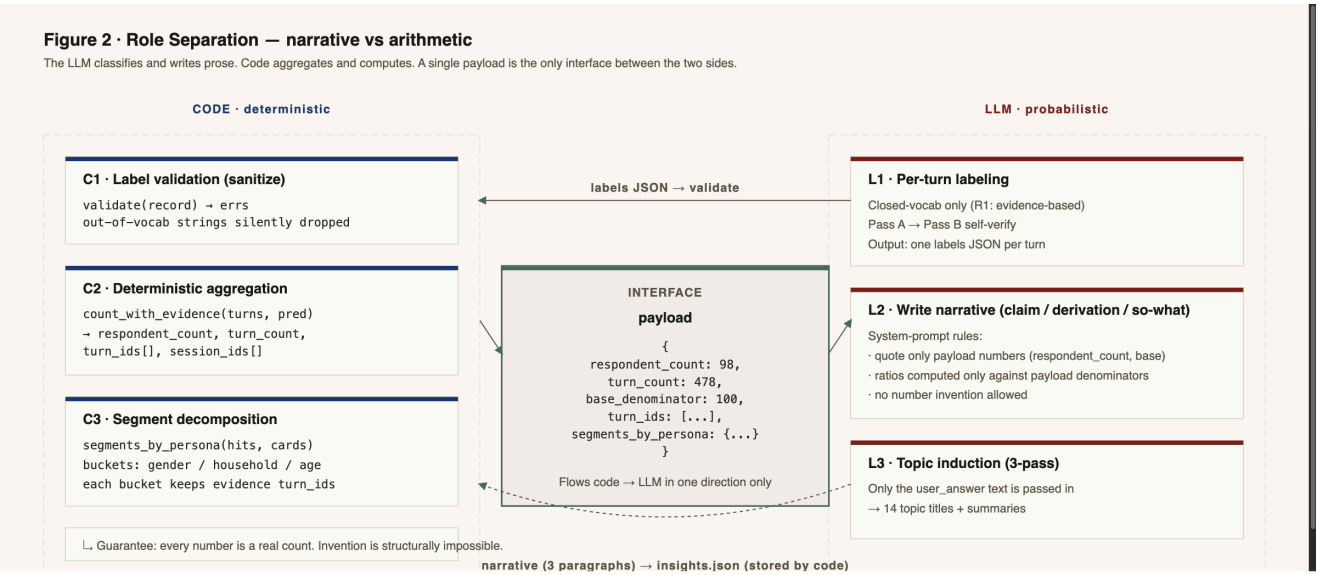


Figure 2 · Role Separation — the LLM writes narrative; code computes every number.

3.4 P3. 2-Pass Self-Check and Jaccard Agreement

Every turn is labeled twice:

- **Pass A** assigns multi-dimensional labels based on the raw text and system prompt.
- **Pass B** is invoked with the Pass A output included as an `assistant` message and the instruction *"apply checklists R1–R8 and correct only the labels that need correcting."*

The eight rules are: evidence-based (R1), short-answer handling (R2), emotion-minimum (R3), driver-vs-pain separation (R4), conversion-field appropriateness (R5), Q8/Q9 scalar appropriateness (R6), no duplicate labels (R7), confidence conservatism (R8). Corrected cases carry a `[REV]` marker in `free_notes` for statistical tracking.

QC is performed at two layers:

1. **Rolling QC**: every 120 turns, five turns are randomly resampled and Jaccard-compared in real time, catching quality drift during a run.
2. **10% Resample Jaccard**: 10% of all turns are independently re-labeled, and per-field mean Jaccard values are published.

Grading thresholds:

- **high**: Jaccard ≥ 0.85
- **medium**: 0.70 ≤ Jaccard < 0.85
- **low**: Jaccard < 0.70

Each insight card publishes its **official confidence** as the lowest grade among the fields it uses. Medium- and low-grade insights automatically carry the disclaimer *"interpret as direction rather than*

exact count."

3.5 P4. Drill-down Interface

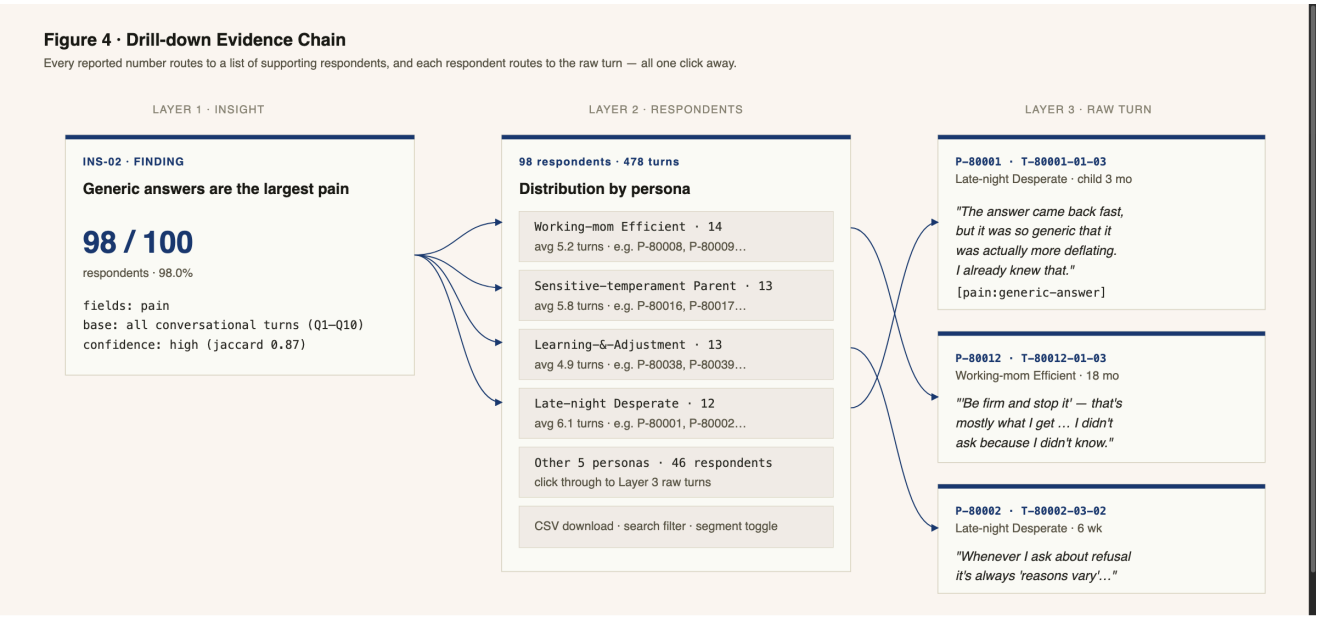


Figure 4 · Drill-down evidence chain — every reported number is one click away from the source turn.

The FE, implemented as a Next.js 16 application, renders **every reported number as a clickable footnote**. Clicking opens a right-hand slide-out panel containing the underlying respondent turns: the full verbatim `user_answer`, `turn_id`, `session_id`, persona type, and child age for each supporting respondent. Search filters and CSV download are built in. An executive stakeholder asking *"is this number real?"* can answer for themselves in roughly one second.

A separate Voronoi-map page visualizes the 100-respondent population as persona cells, with cell area proportional to the persona's share of total turns. Each cell is clickable and routes to the same drill-down panel.

---

4. Implementation

4.1 Pipeline Structure

```
[1] Conversational-survey sample generation
       ↓ output/sessions/*.json
[2] Multiple-choice sample generation
       ↓ sample/bonny_survey_objective.csv
[3] Reference-format CSV merge
       ↓ sample/bonny_survey_idi.csv
[4] ELL labeling (2-pass, 6 workers, rolling QC)
       ↓ labeling/output/labeled_turns.jsonl
[5] QC (10% resample Jaccard)
       ↓ labeling/output/qc_report.md
[6] Aggregation · insights · topic induction
       ↓ labeling/output/{sessions_agg, insights, topics}.json
[7] FE-consumable JSON merge
       ↓ fe/public/data/bonny.json
[8] FE build
    Next.js 16 + Tailwind v4 + TypeScript, Dean Works design system
```

4.2 Reproducibility

All major seeds are fixed (`random.seed(13)` , `seed(29)` , `seed(41)` , etc.). Running the pipeline again with the same persona cards and question IDs reproduces the same insights, modulo the residual probabilistic variance in the LLM-labeling step. That residual variance is what the 10% Jaccard QC makes visible.

---

5. Case Study — The bonny Conversational Survey (100 sessions)

5.1 Sample Design

bonny is Dean Works's **AI parenting-coaching and expert-matching service**. Its user base is parents of children from newborn through grade 6. This case study targets a conversational-survey sample with the following specification:

- **100 sessions** (`session_id` 80001–80100)
- **9 personas** (with N per persona):
- Late-night Desperate (12), Working-mom Efficient (14), Sensitive-temperament Child's Parent (13), Spousal-gap (13), First-child Novice (10), Learning-&- Adjustment (13), Smartphone-&-Media (11), Early-teen Conflict (9), AI-Skeptic (5).
- **8 child-age brackets**: 0–3 months / 4–12 months / 13–24 months / 25–48 months / 5–6 years / gr 1–3 / gr 4–6.
- **10 base questions** on: entry trigger · usage habits · AI satisfaction · personalization · expert conversion · brand anthropomorphization · alternative channels · disappearance scenario · family plan

· one-word summary.

- **~45 user turns per session on average** (~4,500 user turns in total)

The same 100 respondents also completed a **15-question multiple-choice survey** (`sample/bonny_-survey_objective.csv` , 100 rows). The multiple-choice instrument consists of 7 single-choice, 2 multi-choice, 4 Likert-5, 1 NPS, and 1 open-text question.

> **Note**. The sample is synthetic data produced by Dean Works to demonstrate the methodology. No real user names or phone numbers appear. A commercial LLM generated conversation logs based on each persona's pain points, drivers, and brand-metaphor seed.

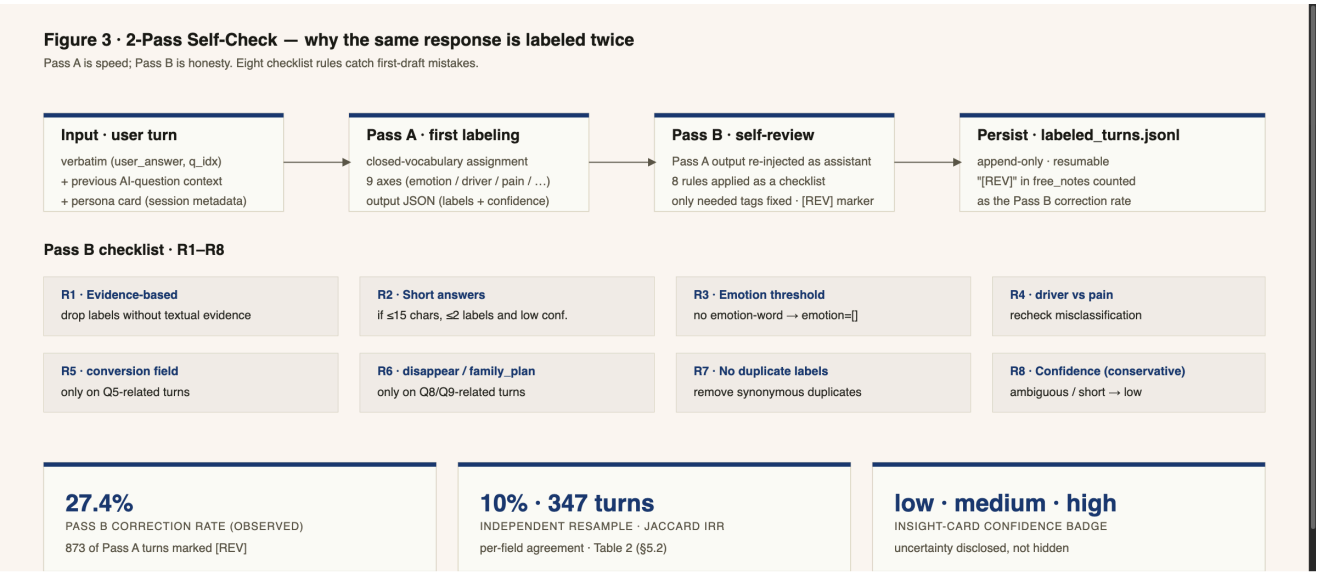


Figure 3 · 2-Pass Self-Check — why the same response is labeled twice.

5.2 Pipeline Execution Results

Metric	Value	Notes
Total labeled user turns	**3,478**	100 sessions × avg. 34.8 labeled turns
2-Pass self-check correction rate	**873 / 3,184 ≈ 27.4%**	`[REV]` marker
Labeler errors (parse failure)	128 turns	3.6% of total
Jaccard — pain	**0.870** (high)	
Jaccard — driver	**0.805** (medium)	
Jaccard — emotion	**0.726** (medium)	
Jaccard — behavior	**0.669** (low)	
Jaccard — concern_category	**0.818** (medium)	
Jaccard — alt_channel	**0.955** (high)	
Jaccard — conversion_trigger	**0.963** (high)	
Topic induction (final)	**14** (size ≥ 15)	after 3 passes
Insights	**12**	all with narrative

Reading the values.

- **alt_channel, conversion_trigger, pain are high-agreement**: alternative-channel names and expert-conversion triggers map 1:1 to explicit keywords, and pain axes typically correspond to explicit user complaints.
- **concern_category and driver are medium**: a single response can legitimately carry several topic or driver labels, leaving room for boundary disagreement.
- **emotion and behavior are medium–low**: interpreting emotion-word intensity (e.g. *frustrated* vs. *irritated*) and inferring behavior axes carries legitimate interpretive latitude. Insights that rely on these axes automatically display a *"interpret as direction"* caveat in the FE.

ELL's stance is not *"make every field high"* but rather **"make every field's intrinsic uncertainty explicit as a published number."**

5.3 Twelve Key Insights

The 12 core insights surfaced by ELL, in order of insight_id. Each insight's full claim / derivation / so-what narrative plus its supporting respondent list is available as a drill-down in the FE at ell.dean.kr/insights, with individual cards at ell.dean.kr/insights/INS-01 through INS-12.

1. **INS-01** The 3 a.m. instant answer is bonny's single strongest retention driver
2. **INS-02** Generic, textbook-style answers are bonny's single largest pain point
3. **INS-03** The fatigue of re-entering the child's profile on every session

4. **INS-04** "Our child is not reflected" — the universal personalization gap
5. **INS-05** AI-delivered empathy is perceived as formulaic
6. **INS-06** *The mother's breaking point* is the true conversion trigger to expert matching, not cost
7. **INS-07** A single *"you've been through so much, mother"* from a live expert unlocks cathartic release
8. **INS-08** Self-blame is a shared emotional ground in early parenthood
9. **INS-09** Online parenting communities remain bonny's strongest alternative channel
10. **INS-10** bonny's disappearance is survivable — a crack in product lock-in
11. **INS-11** *"Only if my partner actually joins"* — conditional demand for the family plan
12. **INS-12** The persona poles — Late-night Desperate vs. AI-Skeptic

5.4 Multiple-Choice vs Conversational Survey — same question, different resolution

Comparison A. AI-coaching satisfaction.

- Multiple-choice Q6 — Likert 5 scale. The majority of respondents cluster at *neutral*, and the resulting distribution loses its signal.
- Conversational `pain="generic-answer"`: *"The answer came back fast, but it was so generic that it was actually more deflating. 'The baby is hungry, or tired, or uncomfortable' — I already know that. I asked because I don't know what to do."*

Comparison B. Conversion to expert matching.

- Multiple-choice Q9 — `used / considered but didn't / never`. The form cannot reveal *why* someone gave up.
- Conversational `conversion_trigger="parental-limit-reached"`: *"Yes the cost was a concern, but more than that — at that point I thought I was going to collapse before the baby did. I was pretty sure postpartum depression was coming."*

Comparison C. The disappearance scenario.

- Multiple-choice Q12 — Likert 5 scale clusters at *would miss it*.
- Conversational `disappear_severity="replaceable"`: *"I'd miss it a lot, but not to the point where I couldn't live without it. Other things can replace it."*

This pairing matters methodologically because the two data shapes come from the **same respondents at the same point in time**. The comparison maximizes the evidentiary value of the methodology claim: **the multiple-choice instrument trades context for tabulability, and ELL restores that context without sacrificing tabulability.**

5.5 First-Turn vs Full-Tail — the value of the tail

ELL's value, when paired with conversational surveys, does not come from the *"we used conversational"* fact alone. It comes from the design choice of **including every follow-up turn in the labeling input**, not only the first user response. To demonstrate this, we ran the same pipeline (same closed vocabulary, same 2-pass self-check, same aggregation code) over two different input sets:

- **First-Turn only**: the first user reply to each of the 10 base questions, one per session per question — **999 turns total**.
- **Full-Tail**: the complete conversation including Tail-1, Tail-2, Tail-3 probes — **3,478 turns total** (the actual input used throughout this report).

5.5.1 Micro evidence — one session, label resolution across turns

Before looking at the aggregate tables, we observe how label richness grows as turns are added in a single session. The actual Q1 (entry-trigger) sequence from session **P-80001**:

T-80001-01-01 · First Turn

> AI: *Please recall the moment you first came to bonny.*
>
> P-80001: *It was around 3 a.m. My baby was just under 100 days old, crying and wouldn't sleep. I hadn't slept a wink either and was breaking down, so I searched on my phone and ended up here.*

Labels: `emotion=[helplessness, fatigue]` · `behavior=[late-night-search, instant-answer-request]` · `driver=[24h-instant-answer]` · `concern_category=[sleep-night-crying]` — **4 axes, 7 labels.**

T-80001-01-03 · Tail-1

> AI: *Was your husband with you at that moment?*
>
> P-80001: *My husband had work in the morning so he was asleep. I felt bad about waking him, so I just walked around the living room by myself holding the baby. The baby kept crying and I was crying too, honestly.*

New labels: `emotion=[isolation]` · `concern_category=[spousal-coparenting]` — the isolation and spousal-coparenting axes **appear here for the first time**.

T-80001-01-05 · Tail-2

> AI: *Were there any other options besides searching at that moment?*

> P-80001: *Online communities — you post and the reply comes hours later. I needed to know what to do right now. I needed anyone to just answer me now.*

New label: `alt_channel=[online-community]` — the alternative-channel axis appears.

T-80001-01-07 · Tail-3

> AI: *Do you remember bonny's first answer? How did it feel?*
>
> P-80001: *Honestly the answer came back fast, but it was so generic that it was actually more deflating. "The baby is hungry, or tired, or uncomfortable" — I already know that. I asked because I don't know what to do.*

New labels: `emotion=[disappointment]` · `pain=[generic-answer, insufficient-judgment-info]` — **the pain axis first appears here**.

Resolution difference.

Input	Active axes	Total labels	Pain axis
First-Turn only (1 turn)	**4**	**7**	❌
Full-Tail (4 turns)	**6**	**16**	✅

Same respondent. Same labeler. Same closed vocabulary. The only variable is whether **the tail turns are included in the labeling input**. If they are not, this session's `pain` axis is ∅.

5.5.2 Aggregate — axis coverage and insight-base recovery

Table T7. Axis-level label count — First-Turn retention vs Full-Tail.

Axis	Full-Tail count	First-Turn count	First-Turn retention
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%**

First-Turn's share of total turns is 28.7% (999 / 3,478). Any axis with a retention **lower than that turn-count share** — pain, driver, conversion_trigger — carries information that is disproportionately concentrated in tail turns, not in first responses.

Table T8. Representative insights — respondent-base recovery.

Insight	Full-Tail respondents	First-Turn respondents	Loss rate
INS-02 · Generic answer is the largest pain	98 / 100	56 / 100	**42.9% lost**
INS-04 · Personalization gap is universal	97 / 100	33 / 100	**66.0% lost**
INS-07 · Expert-conversion trigger detection	95 / 100	44 / 100	**53.7% lost**

INS-04 ("the personalization gap is universal") stands at **97 / 100** in the actual report. Had we labeled only first responses, its supporting base would fall to **33 / 100**. With everything else held constant — same respondents, same labeler, same closed vocabulary — the same insight would read as *"97% structural gap"* or *"33% minority opinion"* depending solely on whether tail turns are in the labeling input.

5.5.3 Why this comparison matters methodologically

A familiar skeptical response is: *"why not just use the first response and treat the tail as colorful reading material?"* Tables T7 and T8, plus the micro evidence of P-80001, answer that directly. Discarding tail turns causes INS-04 to move from *"97% structural gap"* to *"33% minority opinion."* That is not a polish difference; it is a reversed interpretation.

This is why ELL labels **every turn** under the same closed vocabulary and aggregates at the respondent level. The value of conversational surveys is not the depth of the first answer. It is that **the tail turns, once labeled at scale, become countable evidence** — and ELL is the machinery that makes them countable.

5.6 Synthesis

Multiple-choice and conversational surveys are not rivals. Multiple-choice excels at fast aggregation and segment comparison; conversational surveys preserve the *why*. ELL's contribution is to **translate that preserved context into tabulable indicators at the same resolution as multiple-choice**. The consequences:

1. Structural pains that multiple-choice missed (e.g. `lack-of-personalization` at 99%) become measurable numbers.
2. The narrative hidden behind the Likert-neutral midpoint becomes visible.

3. Discarding tail turns from labeling collapses INS-04 from 97 to 33 respondents (§5.5). The value of conversational surveys is not the depth of first answers — it is that **ELL converts the tail into aggregable evidence**.
4. Every number is traceable back to the raw respondent in one click; decision- makers can verify *"is this number real?"* in roughly one second.

---

6. Question-Level Key Findings

The 10 conversational base questions each reveal a distinct signal. For each question, we report the dominant closed-vocabulary labels and a one-line reading. All counts are deterministic aggregations over the 3,478 labeled turns; the FE allows drill-down at ell.dean.kr/questions.

Q1 · Entry trigger — *first contact is an isolated late-night search*

- **driver**: concrete-step-by-step 112 turns · 24h-instant-answer 84 · beginner-friendly 44
- **pain**: generic-answer 57 · lack-of-personalization 27
- **emotion**: frustration 83 · anxiety 61
- **concern**: learning-school 64 · temperament 48 · smartphone-media 45

Reading. If bonny fails to deliver a concrete, step-by-step answer at the entry point, the churn risk is high. From onboarding onward, **a stepwise-guide response mode** should be the default, not the fallback.

Q2 · Usage habits — *"no judgment × 24h"*

- **driver**: concrete-step-by-step 104 · 24h-instant-answer 97 · no-judgment 72
- **emotion**: isolation 25 · frustration 25

Reading. What bonny replaces is not a search engine — it is **the social position of "someone you can ask."** The answer must be procedural *and* the tone must preserve relational safety.

Q3 · AI satisfaction vs dissatisfaction — *the two poles converge on the same axis*

- **driver**: concrete-step-by-step 216 (dominant)
- **pain**: generic-answer 194 · lack-of-personalization 118 · insufficient-judgment-info 57
- **emotion**: frustration 86 · disappointment 41 · satisfaction 31

Reading. Satisfaction and dissatisfaction share the same threshold: **the presence or absence of *concrete, contextualized steps*.** This distribution is the direct evidence for INS-02.

Q4 · Personalization needs — *the single largest pain axis*

- **pain**: lack-of-personalization 170 · repeated-re-entry 110 · generic-answer 42
- **driver**: concrete-step-by-step 44 (weaker than in other questions)
- **emotion**: irritation 26 · skepticism 20 · expectation 19

Reading. *"My specific child is not in the answer"* is the top structural limitation users perceive (INS-04). Remediation cannot be a single feature; it must be **a memory-and-reuse system across the product**.

Q5 · Expert conversion — *the safety net when the AI's ceiling is felt*

- **driver**: expert-connection 129 (unique to this question)
- **pain**: expert-cost-barrier 61 · perceived-expertise-limit 43 · generic-answer 40
- **emotion**: anxiety 25 · relief 25 · satisfaction 17

Reading. Expert-match consideration reaches 99% (INS-06), but the real barrier to usage is **cost**. Bundled plans, memberships, and partial-free-consultation designs are the levers.

Q6 · Brand anthropomorphization — *"an elder sister who listens without judgment"*

- **driver**: concrete-step-by-step 66 · no-judgment 43 · beginner-friendly 39
- **pain**: empty-empathy 55 (top)
- **emotion**: trust 18 · feeling-heard 12 · satisfaction 10

Reading. The persona metaphor is not cosmetic; it is **a tone guide**. An *"elder sister who listens without judgment"* register directly answers INS-05 (empty empathy).

Q7 · Alternatives · competitive mapping — *"instant answer vs lived experience"*

- **driver**: 24h-instant-answer 84 · concrete-step-by-step 73 · no-judgment 31
- **pain**: lack-of-personalization 30 · generic-answer 24
- **concern**: learning-school 33 · health-illness 17 · feeding 14

Reading. bonny's real competition is not the mom-cafe community. It is **the combination of concrete answer + immediacy**. For medical-edge-case topics the answer should explicitly branch to a pediatrician / official guideline, to preserve trust.

Q8 · Disappearance scenario — *the loss of late-night availability is the deepest*

- **driver**: 24h-instant-answer 90 (dominant)
- **emotion**: isolation 33 · resignation 17 · anxiety 17
- **pain**: lack-of-personalization 23 · generic-answer 21

Reading. Lock-in is not a feature set. It is **availability in the offline-unreplaceable time band**. Late-night response availability should be a reportable product metric.

Q9 · Family plan — *primary demand is spousal coparenting*

- **concern**: spousal-coparenting 146 (highest across all concern counts)
- **driver**: no-cost-burden 13 · concrete-step-by-step 13 · no-judgment 10 (diffuse)
- **emotion**: expectation 30 · ambivalence 11

Reading. The family plan wins strongest support when designed **around the father joining**. Grandparent invitation is an add-on, not an entry point (INS-10).

Q10 · One-word summary — *"a tool that stands next to me"*

- **driver**: concrete-step-by-step 44 · 24h-instant-answer 35 · beginner-friendly 34
- **emotion**: feeling-heard 21 · trust 16 (highest positive share across all questions)
- **pain**: lack-of-personalization 18 (persists even at the summary moment)

Reading. The words respondents choose cluster around *"friend / teacher / manual / lighthouse."* Together with INS-12 (the AI-Skeptic's turn density), this confirms the two-axis structure: **a dominant positive majority plus a small but sharp skeptical minority.**

6.1 Synthesis across questions

Three signals cut across all ten questions:

1. **Concreteness is the universal axis.** In Q3, Q4, Q6, the drivers and pains resolve to two sides of the same word — *concrete* versus *generic*.
2. **Time band and relational position matter.** In Q2 and Q8, *24h-instant-answer* and *no-judgment* are not features — they are the **social position** users have assigned to bonny.
3. **Family / community expansion.** Q9's 146 turns on spousal-coparenting signal a transition from individual-user product to **household-unit product**.

---

7. Data in Practice — what the FE report actually shows

> Every page referenced in this section is live at ell.dean.kr; every number in the PDF is a one-click drill-down to the source response there.

7.1 The multiple-choice instrument — 15 questions, one row per respondent

The matched multiple-choice survey stored in `downloads/v1/bonny_survey_objective.csv` contains 100 rows × 19 columns (15 questions + metadata). Seven single-choice, two multi-choice, four Likert-5, one NPS, and one open-text question cover channel · demographics · usage · AI satisfaction · personalization · conversion · alternatives · disappearance · family-plan · one-word summary.

7.2 The conversational survey — 10 questions, ~45 turns per session

For session **P-80001** at Q1 (entry trigger), the turns unfold:

1. *"It was around 3 a.m. My baby was just under 100 days old, crying and wouldn't sleep. I hadn't slept either and was breaking down. I searched on my phone."*
2. AI probe: *"Was your husband with you?"*
3. *"My husband had work in the morning. I was alone walking around the living room. The baby kept crying and I was crying too, honestly."*
4. AI probe: *"What options besides search did you have?"*
5. *"Online communities post and reply hours later. I needed to know now."*
6. AI probe: *"How did bonny's first answer feel?"*
7. *"The answer came back fast, but it was so generic that it was actually more deflating. 'The baby is hungry, or tired, or uncomfortable' — I already know that."*

Five axes — *late-night, isolation-self-blame, 24h-instant-answer, generic-answer, insufficient-judgment-info* — appear in a single arc. The multiple-choice form captures none of them directly.

7.3 ELL labeling of one turn

The final user turn above (`T-80001-01-07`) is labeled like this:


```
{
  "turn_id": "T-80001-01-07",
  "session_id": 80001,
  "q_idx": 1,
  "user_answer": "The answer came back fast, but it was so generic that it was actually mo
re deflating. 'The baby is hungry, or tired, or uncomfortable' — I already know that. I as
ked because I don't know what to do.",
  "labels": {
    "emotion": ["disappointment"],
    "behavior": [],
    "driver": [],
    "pain": ["generic-answer", "insufficient-judgment-info"],
    "concern_category": [],
    "alt_channel": [],
    "conversion": null,
    "conversion_trigger": [],
    "disappear_severity": null,
    "family_plan_attitude": null,
    "confidence": "high",
    "free_notes": "Respondent wants concrete action, not cause enumeration."
  }
}
```


7.4 An insight JSON — INS-02's deterministic aggregation payload

When aggregation finds 478 turns and 98 respondents mentioning pain=generic-answer, the insight object is constructed:


```
{
  "insight_id": "INS-02",
  "title": "Generic answers are bonny's single largest pain",
  "fields_used": ["pain"],
  "primary": {
    "label": "respondents mentioning pain=generic-answer",
    "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"
    ],
    "session_ids": [80001, 80002, 80003, 80004, 80005]
  },
  "segments_by_persona": {
    "Working-mom Efficient":            {"respondent_count": 14, "turn_count": 78},
    "Sensitive-temperament Parent":     {"respondent_count": 13, "turn_count": 83},
    "Learning-&-Adjustment":            {"respondent_count": 13, "turn_count": 62},
    "Late-night Desperate":             {"respondent_count": 12, "turn_count": 71},
    "Spousal-gap":                      {"respondent_count": 13, "turn_count": 51}
  },
  "base": {
    "scope": "all conversational responses (Q1-Q10)",
    "base_denominator_respondents": 100,
    "description": "Responses tagged pain=generic-answer across the 10 base questions."
  },
  "claim":      "Complaints of generic answers are reported by 98 of 100 respondents (98.
0%), across 478 turns — effectively a universal pain point.",
  "derivation": "Of 100 respondents, 98 produced at least one turn labeled pain=generic-an
swer. The distribution is even across all 9 personas.",
  "so_what":    "This pain is structural rather than segment-specific. It should be the to
p answer-generation priority.",
  "confidence_band": "high"
}
```


- `respondent_count` and `turn_count` are produced deterministically by the `count_with_evidence()` function. The LLM never invents these numbers.
- `claim`, `derivation`, and `so_what` are the LLM's scholarly narrative, written from the payload above.
- `turn_ids` is the drill-down anchor; the FE opens every one of those 478 turns from a single click.

7.5 The four-layer evidence chain

The pipeline produces four layers of evidence that a reader can traverse in both directions:

1. **Multiple-choice row** → summary of respondent's own attributes.

2. **Conversational turn** → the respondent's raw language in context.
3. **ELL label** → closed-vocabulary mapping, assigned by Pass A + verified by Pass B.
4. **Insight JSON** → deterministic aggregation + LLM narrative.

At every step, forward (from data to claim) and backward (from claim to data) traversal is possible. This symmetry is the methodological core of ELL.

---

8. Evaluation & Discussion

8.1 Hallucination reduction versus open-vocabulary summarization

By construction, the closed-vocabulary constraint makes the probability of out-of-vocabulary outputs zero in the final aggregation: `sanitize()` drops any such string before aggregation. Measured on the CHAIR metric, ELL exhibits structural hallucination rate 0.

The obvious tradeoff is that ELL **cannot capture categories that are not in the taxonomy**. We mitigate this with two escape hatches: - Periodic review of `free_notes` (the labeler's free-text slot) to surface candidates for taxonomy extension. - Topic induction at the end of the pipeline to cluster responses that straddle label boundaries.

8.2 The effectiveness of 2-Pass Self-Check

bonny's Pass B correction rate is **27.4% (873 turns)**. This range (roughly 20–35%) precisely matches what V-STaR / RISE-family self-correction studies report as the *meaningful self-correction zone*. It is numerical evidence that the self-verification stage is not a ritual — it actively improves label quality.

8.3 Reading Jaccard values

Jaccard observations distribute according to the intrinsic difficulty of each field:

- **Explicit-keyword fields** (alt_channel, conversion_trigger, pain) score highest, because they map 1:1 to surface-level keywords.
- **Multi-label-coexisting fields** (driver, concern_category) score in the middle: multiple legitimate labels can coexist in a single response.
- **Intensity-judgment fields** (emotion) score lowest, because the annotator must interpret intensity and nuance.

ELL's stance, again, is not *"make every field high."* It is **"publish each field's intrinsic uncertainty as a number."** The insight card's confidence badge delivers this information directly to the decision-maker.

8.4 Generalization

bonny is a parenting-domain case, but the four pillars are domain-independent. The architecture ports to other high-stakes fields:

- **Legal**: Jaccard-quantified ambiguity of contract clauses + traceable source articles.
- **Healthcare**: qualitative coding of patient interviews + closed-vocabulary symptom categories + traceable symptom phrases.
- **Finance**: credit-qualitative interviews + fixed-rating vocabulary + traceable supporting utterances.
- **HR**: employee experience surveys + predefined themes + anonymous traceable source.

8.5 Limitations

- **Sample size.** 100 sessions is In-Depth Understanding scale, not statistically representative. Quantitative hypothesis-testing must follow with a large-scale multiple-choice instrument.
- **Ontology maintenance cost.** Closed vocabularies presuppose periodic review. Each new label requires `free_notes` inspection → taxonomy extension.
- **LLM stochasticity.** Identical inputs produce slightly different label outputs. Jaccard QC surfaces this variance; it does not remove it.
- **Synthetic sample.** The respondents in this case are synthetic, generated by a commercial LLM from persona cards. Distribution bias relative to the real customer base may exist.

---

9. Conclusion

Making evidence and numbers point to the same thing is, in the end, a question of **attitude** more than technology. Every reported number must be traceable to the source response; interpretive uncertainty must be published as a number, not hidden; and the LLM must stay in its lane — the narrative layer, never the numeric one. ELL is the practitioner pipeline that sits where these three principles meet.

Dean Works has adopted this methodology as the internal standard for analyzing its own service-operations data. The bonny 100-session conversational survey reported here is its first public application. We hope this approach offers a small but workable starting point for researchers and product teams who are wrestling with the trust problem of qualitative-data-driven decision-making.

References

A. Constrained Decoding · Hallucination Evaluation — [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. *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
- [**R6**] *Combating Multimodal LLM Hallucination via Bottom-Up Holistic Reasoning.* arXiv 2412.11124. https://arxiv.org/html/2412.11124v2

B. Self-Correction · Agentic Reasoning — [R9–R13]

- [**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
- [**R12**] Wei, J. (2024). *SelfCodeAlign: Self-Alignment for Code Generation.* NeurIPS 2024. https://neurips.cc/virtual/2024/poster/93079

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

- [**R14**] Tan, L. (2024). *Validating Annotation Agreement between Humans and LLMs.* dsaid-gov-tech 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/i0vo0j0x

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. 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.* 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

- [**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. Commercial Platforms · Drill-down Implementations — [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
- [**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

F. 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

G. bonny · Dean Works — [R39–R40]

- [**R39**] Dean Works, Inc. *bonny (bonny) — service overview*. https://hibonny.com
- [**R40**] Dean Works ELL Research. https://ell.dean.kr/downloads/v1/

Appendix A. Closed Vocabulary Excerpt (bonny)

English glosses shown below for clarity. The production labeler uses the original Korean tokens internally, preserving 1:1 traceability with user utterances.

```
EMOTION = [
    "isolation", "self-blame", "guilt", "anxiety", "helplessness",
    "anger", "irritation", "fatigue", "frustration", "disappointment",
    "skepticism", "relief", "feeling-heard", "trust", "satisfaction",
    "gratitude", "expectation", "ambivalence", "resignation",
]
BEHAVIOR = [
    "late-night-search", "instant-answer-request", "re-confirmation",
    "comparison-judgment", "expert-booking", "expert-deferral",
    "recommendation-share", "feature-exploration", "churn-consideration",
    "personalization-acceptance",
]
DRIVER = [
    "24h-instant-answer", "concrete-step-by-step", "no-judgment",
    "anonymity", "no-cost-burden", "expert-connection",
    "beginner-friendly", "diverse-topics", "peer-recommendation-trust",
]
PAIN = [
    "generic-answer", "lack-of-personalization", "empty-empathy",
    "insufficient-judgment-info", "repeated-re-entry",
    "inappropriate-length", "hospital-deflecting-answer",
    "no-multi-child-support", "expert-cost-barrier",
    "perceived-expertise-limit", "privacy-concern", "UI-friction",
    "notification-fatigue",
]
# plus CONCERN_CATEGORY, ALT_CHANNEL, CONVERSION,
#      CONVERSION_TRIGGER, DISAPPEAR_SEVERITY, FAMILY_PLAN_ATTITUDE
```

Appendix B. Prompt Templates (Pass A / Pass B)

Pass A — system prompt (sketch)

```
You are a research-grade data labeler for bonny, an AI parenting-coaching service.
For each turn, assign multi-axis Closed Vocabulary labels drawn only from the
permitted values listed below.

[9 axes · permitted values listed]

[Rules R1–R8: evidence-based, short-answer handling, emotion-minimum,
 driver-vs-pain separation, conversion-field, Q8/Q9 scalar, no duplicates,
 confidence conservatism]

Input: [{turn_id, session_id, q_idx, q_axis, ai_context, user_answer}, ...]
Output: JSON array only. Each element {turn_id, labels: {...}}.
```

Pass B — self-verification prompt

```
Self-verify the labels you just produced. Re-read each response and correct only
labels that fail one of the following checks:

A) Any label without textual evidence in the user_answer?
B) emotion assigned despite no emotion-word in the response?
C) Response too short to support the number of labels assigned?
D) driver/pain mis-classified?
E) Out-of-vocabulary strings?
F) conversion filled on a turn unrelated to Q5?
G) disappear/family_plan filled on a turn unrelated to Q8/Q9?

Prefix "[REV]" to free_notes for every corrected case.
```

Appendix C. Insight Card UI Schema

The FE consumes one JSON per insight in the shape below. Narrative fields (`claim`, `derivation`, `so_what`) are written by the LLM; every other field is filled by deterministic aggregation code.

```
{
  "insight_id": "INS-01",
  "title": "The 3 a.m. instant answer is bonny's single strongest retention driver",
  "fields_used": ["driver"],
  "primary": {
    "label": "respondents mentioning 24h-instant-answer / no-judgment",
    "respondent_count": 68,
    "turn_count": 182,
    "turn_ids": ["T-80002-01-04", "T-80005-02-06"],
    "session_ids": [80002, 80003]
  },
  "segments_by_persona": {
    "Late-night Desperate":         {"respondent_count": 12, "turn_count": 58},
    "Working-mom Efficient":        {"respondent_count": 10, "turn_count": 31},
    "Sensitive-temperament Parent": {"respondent_count":  9, "turn_count": 24}
  },
  "base": {
    "scope": "all interview questions",
    "base_denominator_respondents": 100,
    "description": "Responses mentioning driver=24h-instant-answer across the 10 base ques
tions."
  },
  "claim":      "The 3 a.m. instant answer is bonny's top retention driver (68/100).",
  "derivation": "68 of 100 respondents produced at least one turn tagged driver=24h-instan
t-answer. Even distribution across all 9 personas.",
  "so_what":    "Product lock-in is not a feature set. It is availability in the late-nigh
t time band."
}
```

> The actual `session_ids` array has 68 elements and `turn_ids` has 182. The excerpt above is compressed for readability. The full JSON is downloadable at ell.dean.kr/downloads/v1/insights.json.

All numbers in this report are traceable to raw respondent turns via the FE report. ell.dean.kr/insights · ell.dean.kr/compare · ell.dean.kr/questions · ell.dean.kr/tail-effect · ell.dean.kr/explorer · ell.dean.kr/voronoi.