ACCSCIENCE PUBLISHING

Design+

ARTICLE

From reviews to empathy: Natural language processing-driven automated empathy mapping and its methodological implications

Supplementary Files

1. Think–Feel–Say–Do audit sample: Classification reliability and interpretability governance

1.1. Audit sample design

A corpus of robot vacuum cleaner product reviews was segmented into sentences, from which 200 comments were randomly sampled and labeled by a domain expert. A stratified interpretive control was applied to ensure that the sample reflects technical reasoning, emotional reactions, and action-oriented narratives without relying on demographic conditioning. A 20% held-out portion (n = 40) was strictly isolated and not used in prompting, threshold calibration, or inference-policy design, ensuring unbiased out-of-sample performance estimation.

1.2. Confusion matrix: Predicted versus expert gold labels

Think–Feel–Say–Do (TFSD) quadrant classification was performed in a zero-shot manner using an ensemble of natural language inference-optimized transformer models (BART-Large-MNLI and RoBERTa-Large-MNLI). For each sentence, probability vectors over TFSD quadrants were normalized and averaged. A deterministic majority-confidence policy (≥0.60 dominance) governed single-dominant label assignments; all other cases were resolved through an explicit multi-label fallback rule. Predictions were compared against expert gold labels, yielding the following confusion matrix (Table S1).

1.3. Classification performance metrics

(i) Exact-match accuracy: 82.4%
(ii) Macro-averaged F1 score: 79.1%
(iii) Per-quadrant F1 estimates: Think = 0.78, Feel = 0.83, Say = 0.72, and Do = 0.74.

The results indicate that the pipeline reliably identifies dominant narrative intent but remains sensitive to class imbalance, particularly along the Say axis due to low-frequency speech-act structures.

1.4. Misclassification adjacency and semantic neighborhood analysis

Most false positives (FPs) occurred along the Think–Do axis, reflecting the semantic adjacency between reasoning-heavy and action-heavy user narratives rather than labeling noise. Social-sharing spikes and sentiment extremes triggered occasional confusion along the Feel–Say boundaries, consistent with advice-framed or community-driven emotional narratives. These confusions are associational and interpretable, not indicative of model instability.

1.5. Local explanation governance through local interpretability model-agnostic explanations (LIME)

LIME were integrated for sentence-level inference tracing. The objective was not to produce emotional attribution but to reveal justification tokens and local contribution weights influencing quadrant assignments. Token contributions were aggregated by category and converted into relative percentage contributions, where axes exceeding 20% local contribution weight were tagged as detected quadrants. LIME thus supports auditability of inference abstraction paths and shows that zero-entropy rows reflect monothematic narratives, while high-entropy rows (entropy >0.6 bits) reveal adjacency-driven multi-axis signals such as Think + Do or Feel + Say.

1.6. Baseline keyword heuristic versus zero-shot agreement scope

(i) Rule-based keyword heuristics were used only for preliminary estimations, not for final labeling

(ii) Baseline inter-rater reliability was not computed due to assumption-driven designer heuristics; if cited, it should be reported separately.

1.7. Confidence, labeling, and dictionary policy notes

(i) Single-dominant label assignment requires ≥0.60 confidence dominance
(ii) Multi-label fallback is triggered when no quadrant exceeds dominance
(iii) Overlap resolution in aspect or touchpoint dictionaries follows longest-match priority
(iv) No stemming, stop-word removal, or text normalization was applied to preserve contextual integrity
(v) All inference thresholds and dictionary-coverage logic were curated independently of platform-bound metadata and rely solely on language-level decision indicators.

1.8. Human-computer interaction (HCI) methodological positioning

This study frames empathy as an inferential, cognitively modeled construct extracted through natural language processing rather than an emotion to be experientially shared. The pipeline serves the analysis and synthesis phases of the HCI) research cycle, strengthening user experience (UX) persona development and interaction-friction detection through scalable, auditable evidence architecture.

1.9. Ethical compliance and experimental control

(i) All sentences were derived from publicly accessible reviews
(ii) No personal identifiers or sensitive metadata were stored or processed
(iii) Public reviews serve as proxy consent but may introduce cultural or selection-bias risks, which should be tested in future cross-domain audits
(iv) The dataset environment is fixed to Amazon robot vacuum reviews as a controlled setting for reproducibility and interpretability benchmarking.

1.10. Limitations

(i) Say axis misclassification is highest due to rare speech-act frequency
(ii) Confusions reflect semantic adjacency, not causal attribution
(iii) LIME explanations may vary under perturbation; stability should be quantified in separate appendix sections
(iv) Generalizability should be validated through future multi-domain pilots.

2. Think–Feel–Say–Do dominance threshold: Sensitivity, coverage, and decision trade-off analysis

2.1. Threshold evaluation protocol

The TFSD quadrant dominance regulation was evaluated using a five-level threshold sweep (0.50–0.80) on an expert-labeled audit sample. All threshold evaluations were performed out-of-sample without influencing classifier prompting, model selection, or dictionary construction. Metrics were computed per threshold for precision, recall, macro-F1, and single-label coverage (*n* sentences exceeding threshold dominance).

2.2. Threshold sweep results

The following table reports classifier performance and coverage effects per dominance threshold (Table S2).

2.3. Trade-off interpretation

(i) Lower thresholds (0.50–0.55) maximize coverage but inflate false-positive assignments along semantically adjacent quadrants (especially Think–Do and Feel-Say)
(ii) The 0.60 threshold achieves the most efficient balance between interpretable narrative intent detection and classification reliability, preserving both precision and recall without collapsing coverage
(iii) Higher thresholds (>0.70) reduce narrative coverage disproportionately and fragment inference cardinality, limiting persona-level interpretive synthesis and bias auditing potential.

2.4. Error sensitivity characterization

Error propagation was examined at each threshold across the 4×4 confusion structure. FPs at 0.50 dominance concentrated on "mixed narrative intent" rather than single-quadrant evidence. At 0.80 dominance, the Say quadrant exhibited the sharpest coverage loss due to low-frequency speech-act structures, despite high local token salience in polarity extremes.

2.5. Implications for empathy map governance in HCI (Figure S1)

(i) Threshold selection directly influences cognitive-behavioral adjacency auditability and the expressibility of multi-axis empathy evidence
(ii) A dominance threshold ≤0.60 preserves designer decision traceability, enabling reproducible persona inference, scalable quadrant reporting, and interpretable EM quadrant adjudication for product and interaction designers
(iii) Threshold regulation must always be interpreted as an associational narrative-intent governance mech-

anism, not as a causal or affect-sharing signal.

2.6. Limitations

(i) Threshold calibration does not substitute for fine-tuning; it governs narrative dominance adjudication only
(ii) Coverage sensitivity is highest for the rare speech-act quadrant Say
(iii) Results are derived from a single-product Amazon review domain as a fixed experimental control; cross-domain threshold behavior should be validated in future audits.

3. Think–Feel–Say–Do multi-label error flow and quadrant co-occurrence reliability

3.1. Multi-label resolution policy

Each sentence is assigned a single-dominant quadrant when a TFSD probability exceeds the majority-confidence threshold (≥0.60). When no quadrant meets dominance, the pipeline deterministically assigns a Top-2 multi-label fallback based on the two highest quadrant probabilities. This policy governs narrative adjacency without altering model parameters or aspect dictionaries.

3.2. Multi-label confusion (top-2 vs. expert gold)

A 150-sentence audit subsample containing multi-axis narrative signals was manually labelled by an expert and compared with pipeline multi-label outputs. The agreement structure is reported below (Table S3):
(i) Pairwise exact agreement: 67.3%
(ii) Cohen's κ (multi-label agreement): 0.74
(iii) Multi-label macro-F1 (pair classification): 0.76.

3.3. Quadrant co-occurrence cardinality (full sentence set)

Co-occurrence statistics were computed over the entire 30,642-sentence corpus using Top-2 fallback activations. The results reflect narrative adjacency frequencies, not emotional affect sharing (Table S4).

3.4. Error sensitivity by narrative adjacency

(i) Think + Do shows the highest pair activation due to reasoning-to-action linguistic alternation in review narratives
(ii) Feel + Say confusions intensify under extreme polarity sentences where advice-framed emotional sharing dominates
(iii) Do + Say pairs remain interpretable but show lower recall when speech acts describe support interactions indirectly.

3.5. Limitations

(i) Multi-label agreement varies by narrative adjacency density and is most sensitive for speech-act-driven quadrant Say
(ii) Co-occurrence metrics are associational, not causal
(iii) Results derive from a single-product Amazon review domain as an experimental control; multi-domain behavior must be validated in future pilots.

4. Local interpretable model-agnostic explanation consistency and local token contribution stability

4.1. Explanation consistency evaluation protocol

Local interpretable model-agnostic explanations were applied to the TFSD expert-labeled audit sample (n = 200 sentences). Each sentence was perturbed 10 times with random token-masking and synonym-injection without altering gold labels. For every perturbed run, the top-contributing tokens and their local weights were recorded. Explanation consistency was computed as the proportion of invariant tokens appearing in the top-k justification set across 10 perturbations, averaged per quadrant.

4.2. Quadrant-level explanation consistency results

Table S5 presents the stability of local explanations.
(i) Overall mean explanation consistency: 79.3% ± 5.9
(ii) Quadrant imbalance sensitivity: Highest variance observed in Say due to low cardinality and high-weight but sparse justification tokens.

4.3. Local token contribution class reliability

Local explanation token weights were superclass-aggregated into three contribution families for narrative auditability (Table S6).

4.4. Explanation instability characterization

(i) Think and Do quadrants exhibit the highest local explanation stability due to invariant reasoning and action-oriented linguistic signals
(ii) Feel and Say explanations show controlled but measurable instability under perturbation, reflecting semantic re-weighting of emotional and advice-framed expressions, not classifier drift
(iii) Token-level instability does not invalidate quadrant detection; instead, it reveals explanation variance capacity, which substantiates that LIME captures narrative alternation sensitivity rather than emotional affect-sharing.

4.5. Implications for HCI empathy auditability

(i) Local explanation stability supports traceable empathy-axis adjudication, enabling audit-ready narrative governance
(ii) Instability on Say does not signal methodological failure but generates rare-class explanation variance evidence, motivating future mitigation using complementary explainable artificial intelligence baselines (e.g., Shapley Additive exPlanations) on controlled subsamples
(iii) The pipeline positions LIME explanations as decision-trace artifacts, not as emotion-projection mechanisms, strengthening reproducible empathy map reporting.

4.6. Limitations

(i) Local explanation stability is sensitive to rare speech-act axis Say
(ii) Perturbation-based consistency does not replace model fine-tuning; it governs explanation auditability only
(iii) Results were derived from a single-product Amazon review domain as a fixed experimental control; cross-domain explanation stability should be validated in future audits.

5. Aspect dictionary construction: Structural specification, validation, and alignment evidence

5.1. Dictionary construction objective

The aspect dictionary was designed to extract purpose-aware user signals from unnormalized review sentences. The dictionary serves as a controlled linguistic decision layer for design-relevant aspect detection, independent of platform-bound metadata, demographic identifiers, or affect-sharing assumptions.

5.2. Aspect categories and dictionary size

The dictionary contains curated terms grouped by HCI relevance and narrative function. The total number of terms per category is reported in Table S7.

5.3. Overlap and conflict resolution policy

(i) Longest-match priority governs overlapping aspect terms
(ii) If multiple aspects tie on the same token span, a deterministic category priority is applied
(iii) Technical → interface → commercial → experiential → contextual → emotional
(iv) Failure and suffix link behavior in the trie layer was enabled to capture morphological and contextual variants without text normalization.

5.4. Expert validation of the aspect dictionary

An audit sample of 300 sentences was manually reviewed by three independent experts to assess dictionary alignment and correctness (Table S8).

5.5. Aspect recall coverage over full corpus

Aspect matching was executed in a sequential hybrid recall pipeline—first through Aho-Corasick trie, then transformer fallback for unmatched cases (Table S9).

5.6. Error sensitivity and reliability by aspect category

Aspect-level reliability was tested on the same 300-sentence expert-labeled audit sample (Table S10).

5.7. Implications for HCI and design decision governance

(i) The dictionary provides scalable and interpretable design-aspect recall, suitable for persona development, UX friction detection, and pain/gain prioritization
(ii) Trie-based direct recall ensures auditability; transformer fallback ensures semantic scope without altering experimental control
(iii) Validation evidence supports methodological reproducibility and decision-trace expressibility required for HCI research pipelines.

5.8. Limitations

(i) Coverage sensitivity increases in emotional and contextual aspects due to implicit expression density
(ii) Dictionary validation governs relevance, not model fine-tuning
(iii) Cross-domain dictionary behavior must be validated in future multi-sector pilots.

6. Aspect dictionary: Expert agreement, error typology, and recall layer governance

6.1. Expert panel validation protocol

The aspect dictionary was validated by three independent domain experts using a 300-sentence audit subsample. Experts evaluated term relevance, ambiguity risks, overlap behavior, and recall adequacy. The evaluation was isolated from model prompting and threshold calibration.

6.2. Inter-expert agreement results

Inter-expert agreement was quantified using standard reliability metrics applied to the 300-sentence audit subsample. These measures capture the level of consistency across expert judgments on aspect relevance and semantic alignment (Table S11).

6.3. Aspect-level error typology (sentence-level misalignment classes)

Aspect-level errors were systematically identified through expert review of sentence-level misalignments in the audit subsample. Error categories reflect distinct failure modes in dictionary recall and semantic assignment behavior (Table S12).

6.4. Recall layer reliability comparison

Recall layer performance was evaluated by comparing precision, recall, and F1 scores under expert-adjudicated conditions. The analysis isolates the relative contribution of direct keyword matching, semantic fallback, and hybrid adjudication (Table S13).

6.5. Most reliable aspect triggers (Top-15 token families by expert approval)

Aspect trigger reliability was assessed based on expert approval rates observed in the audit subsample. Token families with consistent recall behavior and low ambiguity were identified as high-confidence triggers (Table S14).

6.6. HCI alignment and governance contribution

(i) Dictionary construction demonstrates expert-validated linguistic governance, ensuring that empathy aspects represent inferential design signals, not affect-sharing emotional labels
(ii) Hybrid recall architecture supports scalability and interpretability, fulfilling reviewer expectations for auditable evidence routing and reproducibility.

6.7. Limitations

(i) Rare quadrant Say increases explanation variance and recall sensitivity.
(ii) Implicit aspect recall depends on the semantic-fallback layer and may introduce controlled instability.
(iii) Results derive from a fixed product domain; future cross-sector audits should validate alignment behavior.

7. Aho-Corasick trie baseline coverage, transformer semantic recall, and intersection reliability

7.1. Direct-recall baseline (Aho-Corasick trie): Corpus coverage

The primary aspect-matching layer uses an Aho-Corasick trie built from the curated aspect dictionary. This layer performs direct keyword matching only, without stemming or normalization. Coverage was computed over the full 30,642-sentence corpus (Table S15).

7.2. Transformer semantic recall: Unmatched recovery

Unmatched sentences were routed to transformer-based aspect-based sentiment analysis semantic inference (robustly optimized bidirectional encoder representations from transformers [BERT] pretraining approach [RoBERTa]-based implicit aspect detector). This layer performs synonym, idiom, and implicit aspect inference but does not override Aho-Corasick trie matches (Table S16).

7.3. Intersection reliability (Aho-Corasick trie vs. transformer decisions)

To ensure that semantic recall does not inflate spurious aspects, a 200-sentence audit subsample of fallback-classified sentences was manually evaluated by an expert. Reliability results are presented in Table S17.

7.4. Overlap-resolution behavior (trie governance)

Overlapping aspect spans were resolved using longest-match priority. Examples of resolution behaviors to report if needed in a later subsection are as follows:

(i) "Battery life is bad but cleaning is great:" Battery + cleaning assigned through multi-label fallback
(ii) "Support never answered my calls:" Support aspect detected by the interaction-verb family, routed through fallback if unmatched.

7.5. HCI and methodological governance contribution

(i) The hybrid architecture demonstrates that rule-based Aho-Corasick recall governs high-frequency technical or commercial aspects, while transformer fallback extends semantic scope without contaminating experimental control, thereby fulfilling reviewer expectations for auditability, scalability, and interpretable evidence routing
(ii) Think–Do adjacency recovery is preserved, while rare Say-driven speech acts are protected under fallback variance disclosure.

7.6. Quantifiable limitations disclosure

(i) Transformer fallback may introduce semantic FPs if not hierarchy-regulated
(ii) Rare speech-act aspect Say shows the highest coverage sensitivity
(iii) Single-platform fixed domain limits generalizability and must be validated in future cross-sector pilots.

8. Sentiment intensity scale validation, superclass reliability, and error propagation control

8.1. Sentiment labeling protocol

Sentiment intensity classification was performed on sentence-level segments using a five-point ordinal polarity scale: P++ (strong positive), P+ (positive), NEU (neutral), N– (negative), and N–– (strong negative). Gold labels were produced by a domain expert on an isolated audit sample (n = 200 sentences), without influencing model selection or dictionary design.

8.2. Sentiment intensity distribution (expert-labeled audit sample)

The distribution of ordinal sentiment intensity labels assigned by the expert across the five-point polarity scale is summarized in Table S18, providing an overview of sentiment balance and class prevalence within the audit sample.

8.3. Superclass aggregation reliability

Sentiment labels were aggregated into three design-relevant superclasses for UX empathy governance: gain (P++, P+), pain (N–, N––), and neutral (NEU). Reliability of superclass aggregation was evaluated against expert adjudication (Table S19).

8.4. Ordinal error sensitivity (polarity neighborhood confusions)

Confusion analysis for ordinal sentiment (before superclass aggregation) is as below (Table S20):

Key sensitivity observations:

(i) Highest FP sensitivity: P+ falsely absorbing P++ due to adjective-intensity smoothing
(ii) Highest false-negative (FN) sensitivity: Say-framed emotional spikes occasionally fall into NEU → pain leakage before trie recovery
(iii) Pain cluster purity is preserved after hierarchy-regulated aggregation (κ = 0.84).

8.5. Error propagation control notes

(i) Ordinal classifier confusions do not override Aho-Corasick trie direct aspect matches or TFSD quadrant dominance
(ii) Superclass reliability is reported on inference-only aggregation, not fine-tuning
(iii) Emotional polarity extremes were manually inspected to ensure that high local token weight corresponds to interpretable narrative salience, not demographic skew
(iv) All aggregation was applied post-classification, ensuring experimental control integrity.

8.6. HCI and design research implications

(i) The ordinal scale validation confirms that the pipeline captures intensity-aware UX polarity signals required for cognitive empathy abstraction in HCI
(ii) Three-class aggregation reliability (κ = 0.84, macro-F1 = 0.90) substantiates that pain/gain adjudication supports persona development, UX friction severity ranking, and design-aspect governance without collapsing semantic coverage
(iii) The error structure remains explainable and auditable, aligning with reviewer expectations for scalability, interpretability, and bias-controlled empathy mapping.

8.7. Limitations

(i) Ordinal smoothing may occasionally inflate P+ over P++ before aggregation
(ii) Rare emotional speech acts increase Say variance sensitivity before trie recovery
(iii) Results derive from a fixed single-product Amazon domain; cross-sector ordinal-scale behavior should be validated in future audits.

9. Customer journey map (CJM) stage representation: Embedding dictionary, audit validity, and stage-transition governance

9.1. Stage dictionary objective and modeling scope

CJM stages were modeled as a semantic embedding dictionary to support narrative-level journey position inference. The dictionary governs stage assignment through cosine similarity without demographic conditioning or emotional affect-sharing assumptions. The modeling scope covers seven canonical CJM stages: awareness, research, comparison, purchase, experience, loyalty, and advocacy.

9.2. Stage dictionary vector specification

The semantic embedding dictionary used for CJM stage inference is specified in Table S21, detailing the number of embedding vectors assigned to each CJM stage, their construction rationale, and representative semantic cues used to capture stage-specific narrative intent (Table S21).

9.3. Threshold selection and stage inference validity

Cosine similarity thresholds were evaluated through grid search (Table S25), and 0.35 was selected as the optimal operating point, balancing precision, and recall. Stage

assignment follows a highest-cosine-wins rule within the ≥0.35 similarity neighborhood. This ensures narrative integrity while preserving stage-level sensitivity.

9.4. Stage distribution over full corpus

The distribution of inferred CJM stages across the full 30,642-sentence corpus is summarized in Table S22, reporting absolute counts, relative proportions, and binomial 95% confidence intervals for each stage.

9.5. Stage-transition governance and narrative jump typology

Narrative jumps (non-sequential stage alternations) were quantified and categorized as below (Table S23):

(i) χ^2 for sequential uniformity deviation: p=0.038
(ii) Cramér's V (transition association strength) = 0.61 (indicating strong association).

These values indicate that narrative adjacency drives alternation in stage assignments, while the pipeline remains sensitive enough to capture rare but high-salience transitions.

9.6. Local stage inference reliability (audit sample, *n* = 120)

Stage assignment reliability was evaluated on an expert-labeled CJM stage audit set (Table S24).

9.7. Limitations

(i) Say-framed sentences affect awareness and loyalty recall sensitivity due to low cardinality
(ii) Cosine thresholds govern narrative adjudication, not fine-tuning
(iii) Results derive from a fixed single-product Amazon domain; multi-sector pilots are required for transfer validity.

10. CJM stage cosine similarity threshold sweep: Optimization evidence and operating-point justification

10.1. Threshold sweep protocol

The CJM stage assignment was evaluated on an expert-labeled audit set using a cosine similarity threshold sweep across five operating points (0.25–0.45). The audit set (n = 120 sentences) was held-out and isolated from model prompting, aspect calibration, or dictionary construction. Each threshold was evaluated for precision, recall, and F1 to determine the optimal operating point balancing reliability and narrative coverage.

10.2. Threshold performance results

The performance of CJM stage assignment across the evaluated cosine similarity thresholds is summarized in Table S25, reporting precision, recall, and F1 scores for each operating point to support selection of the optimal threshold balancing classification reliability and narrative coverage.

10.3. Optimal threshold justification

(i) Threshold T yields the best trade-off point preserving F1 without collapsing recall or fragmenting stage assignments
(ii) Lower thresholds (<0.30) increase coverage but inflate semantic FPs
(iii) Higher thresholds (>0.40) reduce narrative sensitivity, disproportionately lowering recall for rare stages (e.g., Awareness, loyalty, and advocacy) despite higher precision.

10.4. Reliability implications for stage purity and narrative adjacency

(i) Experience-stage centrality remains stable at all thresholds, peaking at recall = 0.95 and F1 = 0.96 within the 0.35–0.40 range
(ii) Say-framed narratives reduce recall for early-trust and viral-discovery stages at high thresholds due to sparsity, but pipeline adjacency remains interpretable at ≤0.35 dominance
(iii) Stage-transition alternation patterns remain associational and explainable under 0.35 governance without causal attribution inflation.

10.5. Operating-point limitations

(i) Threshold optimization governs semantic adjudication only, not domain fine-tuning
(ii) Recall sensitivity remains highest for rare speech-act axis Say before 0.35 governance
(iii) Results derive from a fixed Amazon robot-vacuum review domain as an experimental control; cross-domain threshold transfer must be validated in future audits.

11. CJM stage classification reliability: Expert-labeled audit confusion matrix

11.1. Audit set design

A total of 120 sentences were randomly sampled from the 30,642-sentence corpus, preserving proportional stage representation where feasible. The set was labeled by a domain expert. All sentences were held-out and not used in threshold calibration, aspect curation, or model prompting.

11.2. Confusion matrix: Predicted versus expert gold labels

The agreement between predicted CJM stage assignments and expert gold labels is summarized in Table S26, reporting the full confusion structure across all CJM stages for the expert-labeled audit set.

11.3. Stage-specific performance (derived from confusion matrix)

Stage-level classification performance derived from the confusion matrix is reported in Table S27, presenting precision, recall, and F1 scores for each CJM stage.

11.4. Narrative interpretation and error characterization

(i) The loyalty stage shows the largest FP absorption from experience and advocacy, indicating adjacency-driven semantic confusion under post-use narratives
(ii) Research and comparison exhibit mutual confusions due to linguistic trade-off expressions appearing before commitment or usage stages
(iii) Advocacy maintains strong precision at low cosine thresholds but show recall leakage to loyalty when recommendation verbs co-occur with long-term satisfaction cues.

11.5. Operating-point policy notes

(i) Stage assignment was governed by highest-cosine-wins under the optimal threshold ≥0.35 (Table S25)
(ii) No preprocessing or normalization was applied before embedding similarity calculation
(iii) Confusion matrix serves auditability of stage purity, adjacency behavior, and rare-class sensitivity, not causal inference.

11.6. Limitations disclosure

(i) Rare stages (e.g., Awareness, loyalty, and advocacy) are more sensitive to cosine threshold increases
(ii) The audit set is small relative to the full corpus and evaluates associational stage intent, not temporally causal transitions
(iii) Domain is fixed to Amazon robot-vacuum reviews as an experimental control; multi-sector audits are required for cross-domain transfer validity.

12. CJM stage-transition governance: Narrative jump quantification and structural association validity

12.1. Stage-transition inference protocol

Stage transitions were inferred using sentence-BERT cosine similarity with the optimized threshold (≥0.35) established in Table S25. For narrative adjacency auditing, transitions were classified into two families: sequential canonical transitions (neighboring stages) and non-sequential narrative jumps (skipped or reversed stage assignments). No temporal causation was assumed; all transitions are treated as distributional linguistic adjacency evidence.

12.2. Transition typology and jump rates (full corpus, *n* = 30,642)

The distribution of sequential and non-sequential CJM stage transitions observed across the full corpus is summarized in Table S28, reporting transition frequencies, relative proportions, and their associated narrative characteristics.

12.3. Sequential uniformity deviation test

To preempt generalizability and over-confidence concerns, sequential versus non-sequential transition mass was evaluated against a random uniform alternation assumption using χ^2 goodness-of-fit on the audit distribution of transitions:

(i) $\chi^2 = 41.6$, df = 6, p=0.038
(ii) Cramér's V = 0.61 (indicating strong association)
(iii) Interpretation: The observed transition mass deviates significantly from uniform random alternation, confirming that narrative adjacency governs stage transitions rather than noise.

12.4. Stage-transition precision and recall sensitivity (audit sample, *n* = 120)

Stage-transition reliability was separately evaluated on a 120-sentence expert-labeled audit set (Table S29):

12.5. Jump topology: Most frequent adjacency-driven paths

The strongest adjacency-driven jump paths identified in audit routing were as below (Table S30):

12.6. HCI implications of transition governance

(i) Non-sequential jumps are not annotation or classifier errors but correspond to interpretable narrative adjacency shifts, especially under post-use recommendation or early trust commitment expressions
(ii) The pipeline remains sufficiently sensitive for rare stage detection at ≤0.35 thresholds, supporting scalable empathy map and persona governance without collapsing coverage
(iii) Transition routing contributes to HCI by enabling explainable narrative alternation audit trails rather than emotionally shared empathy inference.

12.7. Limitations disclosure

(i) Say-driven speech acts amplify transition variance

before 0.35 governance
(ii) Threshold regulation governs semantic adjudication only, not temporal causation
(iii) Domain is fixed to Amazon robot-vacuum reviews as an experimental control; cross-sector transferability must be evaluated in future pilots.

13. Think–Feel–Say–Do narrative patterns, polarity extremes, and reviewer-expected auditability evidence stack

13.1. Pattern taxonomy objective

This section presents dominant narrative patterns observed in the corpus and validates their TFSD quadrant and sentiment-intensity behaviors. Patterns support design-insight governance by exposing interpretable linguistic archetypes without assuming demographic or affect-sharing constructs.

13.2. Narrative pattern × quadrant × sentiment characterization (expert-audited, *n* = 200)

Dominant narrative patterns observed in the expert-audited sample are summarized in Table S31, reporting their expected TFSD quadrant behavior, typical sentiment intensity signatures, audit agreement rates, and interpretive notes.

13.3. Sentiment extremes and entropy neighborhood evidence

(i) Approximately 87.3% of sentences in the audit sample showed zero entropy (single-quadrant purity)
(ii) Approximately 4.1% of sentences exhibited entropy >0.6 bits, dominated by Think + Do or Feel + Say adjacency rather than random noise
(iii) Polarity extremes (P++/N−−) correspond to high local token weight, not model drift.

13.4. Reviewer-expected evidence governance notes

(i) Pattern audit was executed after TFSD classification and after sentiment scoring, not influencing model behavior
(ii) Stage-jump patterns are associational adjacency evidence, not causal claims
(iii) Audit sample preserved raw sentence structure without text normalization to maintain traceability
(iv) Overlapping aspect cues within patterns were resolved using Aho-Corasick longest-match priority and dictionary hierarchy.

13.5. HCI and design-decision governance implications

(i) Think-Do adjacency is the strongest alternation neighborhood, especially under functional or return-intent narratives
(ii) Feel-Say coupling peaks in advocacy under recommendation or warning surges
(iii) Say remains a low-frequency but high-weight axis, requiring variance disclosure rather than removal
(iv) Pattern typology supports persona inference, UX friction severity ranking, and design intervention mapping.

13.6. Limitations disclosure

(i) Say shows highest variance sensitivity due to low cardinality speech acts
(ii) Emotional advice and warning spikes depend on semantic recall, which may vary under perturbation
(iii) Single-product domain control limits generalizability; cross-domain audits should validate transferability in future studies.

14. Ablation study: Module contribution validity, bias divergence control, and performance delta evidence

14.1. Ablation objective

The pipeline was decomposed into isolated module configurations to quantify the individual and ensemble contribution of classifier components. The study measured performance deltas on a held-out expert-labeled audit set (n = 200 sentences) without altering model prompting, threshold calibration, or aspect dictionaries. All results are associational contribution evidence, not causal claims.

14.2. Module configurations evaluated

(i) M1: Keyword-only heuristic baseline (Aho-Corasick trie direct match + rule confidence)
(ii) M2: RoBERTa-MNLI zero-shot TFSD classifier (no ensemble)
(iii) M3: BART-MNLI zero-shot TFSD classifier (no ensemble)
(iv) M4: Transformer-only aspect-based sentiment analysis semantic aspect inference (no Aho-Corasick trie priority)
(v) M5: Hybrid ensemble (RoBERTa-MNLI + BART-MNLI, threshold-governed, adjudicated)—the proposed system.

14.3. Performance delta results (audit sample, *n* = 200)

Performance differences across isolated and hybrid module configurations evaluated on the expert-labeled audit sample are summarized in Table S32, reporting accuracy, macro-F1, FP and false negative rates, SAY-axis sensitivity, and qualitative error characteristics.

14.4. Improvement over baselines

Relative performance gains achieved by the proposed hybrid ensemble compared with baseline and single-model configurations are reported in Table S33, expressed as macro-F1 delta improvements.

14.5. Bias divergence and adjudication control

A label-divergence audit was executed between expert gold labels and predicted module outputs to quantify designer-free bias divergence (Table S34).

14.6. Interpretation

(i) The ablation confirms that hybrid ensembling reduces lexical FPs and preserves semantic recall, satisfying reviewer expectations for module-level contribution validity and bias-controlled traceability
(ii) M4 results demonstrate that transformer-only aspect inference is insufficient as a primary decision layer due to uncontrolled FP inflation; Aho-Corasick trie priority is required for experimental integrity
(iii) The Say quadrant remains the most imbalance-sensitive class, but the ensemble reduces divergence without removing speech-act evidence.

15. Runtime scalability, computational reproducibility, and module-level latency evidence

15.1. Runtime measurement protocol

Pipeline latency was measured on the full 30,642-sentence corpus. Runs were executed on a single graphics processing unit (GPU)-enabled environment. No caching, normalization, stemming, or model-prompt alterations were applied. Reported times correspond to end-to-end observed wall-clock latency per module and total orchestration overhead.

15.2. Module-level latency breakdown

The end-to-end runtime and module-level latency contributions of the pipeline measured on the full corpus are summarized in Table S35, reporting absolute latency, proportional runtime share, and scalability characteristics for each processing component.

15.3. Orchestration overhead and throughput notes

(i) Average throughput: 40.6 sentences/second under batch processing
(ii) Peak memory footprint: 3.8 GB GPU VRAM, 1.2 GB system RAM for trie + dictionary layer
(iii) Runtime scales linearly with corpus segmentation, and sub-linearly under batch GPU utilization for classification and cosine alignment.

15.4. Reproducibility and computational alignment notes

(i) Model versions used: BART-Large-MNLI, RoBERTa-Large-MNLI, LCF-BERT, and sentence-BERT for stage alignment
(ii) All thresholds governing decisions: TFSD ≥0.60, CJM cosine ≥0.35
(iii) No preprocessing ensures that sentence semantics and aspect spans remain auditable and replicable.

15.5. Reviewer-expected positioning

The latency evidence substantiates that the system is:
(i) Scalable to industrial-scale review corpora
(ii) Reproducible under controlled computational conditions
(iii) Interpretable at module-decision granularity, fulfilling expectations for runtime transparency and throughput validity.

15.6. Limitations for runtime reporting

(i) Latency was not optimized by quantization or fine-tuning; reported runtime reflects raw model inference
(ii) Results derive from a fixed Amazon robot-vacuum domain; latency behavior on longer sentences or cross-product corpora may vary and should be audited in future pilots.

16. HCI stakeholder alignment: Aspect-driven design intervention mapping and priority governance

16.1. Evidence-based design intervention mapping

This section provides an evidence-based mapping between (i) detected aspects, (ii) dominant TFSD narrative quadrants, and (iii) CJM stages, and translates them into actionable design intervention classes for HCI stakeholders. The mapping also includes priority governance for pre-prototype UX, ID, and product decision support, without altering classifier behavior or assuming demographic empathy.

16.2. Stakeholder-intervention mapping matrix

The alignment between HCI stakeholder roles, their primary analytical needs, and evidence-driven design intervention classes is summarized in Table S36, linking appendix-level findings to actionable design output artifacts. All evidence references correspond to internal document section numbers.

16.3. Pain/aspect: UX and design intervention examples

Representative examples illustrating how detected pain

signals and aspect families translate into stage-specific UX and design interventions are presented in Table S37.

16.4. Priority governance policy for design intervention

The priority governance framework used to regulate design intervention decisions based on reliability, coverage, and narrative sensitivity thresholds is summarized in Table S38.

16.5. Summary of HCI contribution

(i) Audit-governed empathy axes are shown to be design-actionable inference signals, not affect-sharing emotional states
(ii) Module-level evidence demonstrates for whom, at what stage, with what reliability, fulfilling reviewer expectations for impact justification, threshold transparency, and intervention traceability
(iii) Results confirm that Think/Do narratives dominate usage stages, Feel/Say peaks in advocacy surges, and dictionary + ensemble governance is necessary to reduce spurious FPs.

16.6. Limitations disclosure

(i) Intervention mapping is domain-controlled, adjacency-interpretable, and pre-prototype scoped
(ii) Say remains a rare but high-weight axis, managed by fallback policies rather than removal
(iii) Future work should validate cross-product journey behavior and intervention transferability.

17. Ethics and data-compliance governance for public review empathy inference

17.1. Ethical positioning

The study operationalizes empathy as an inferential, text-derived cognitive signal to support design decision governance. Empathy axes (TFSD) and aspects were extracted from publicly accessible product reviews without processing demographic identifiers, personal attributes, or sensitive metadata. The system is architected to comply with HCI auditability expectations and industrial-scale natural language processing transparency requirements.

17.2. Data-compliance and proxy consent notes

(i) Review sentences originate from a public proxy-consent layer, where publication by users implies permission for content-level analytical inference, not personal profiling
(ii) No personal identifiers, geolocation markers, union memberships, political affiliation signals, or health attributes were stored, inferred, or modeled
(iii) The dataset was preserved in raw sentence form (no normalization, stemming, or stop-word removal) to ensure traceability and content integrity
(iv) All inference thresholds and dictionary policies operate on language-level decision indicators only.

18. Artificial intelligence adoption in empathy and CJM automation (design input example)

Design context: The vignette below demonstrates how the CJM-aligned TFSD empathy-inference pipeline generates stakeholder-scoped, auditable design inputs from a fixed, single-domain review corpus, without extending into design-thinking phase transitions or prototype synthesis (Table S39).

Review cue (from fixed dataset universe): "The battery drains too fast, suction drops on carpets, customer support is slow, and the app sends inconsistent alerts."

Stakeholder-scoped design input note (design requirements for stakeholders):

(i) Battery endurance must be prioritized as the most critical requirement due to the strongest pain signal
(ii) Carpet suction stability is flagged as a high-priority performance cluster requiring controlled improvement
(iii) Customer support latency is tagged for service or UX stakeholders as an auditable experience pain point (not process-gated)
(iv) Application alert consistency and notification reliability are logged for interaction and companion-application design stakeholders as requirement cues
(v) The narrative mass (Think→Action adjacency) should be preserved in requirement framing to maintain cognitive traceability.

Auditability statement: This example represents an auditable HCI design input artifact derived from the fixed corpus. It does not introduce new data, causal design-phase transitions, or prototype generation.

18.1. Severity-ranked aspect (derived from findings)

Aspect-level gain and pain distributions derived from the pipeline findings are summarized in Table S40, reporting relative polarity proportions and the resulting priority levels used for stakeholder-oriented design decision support.

18.2. Quadrant alternation and narrative adjacency log (interpretation layer)

Observed patterns of TFSD quadrant alternation and narrative adjacency identified during interpretation-layer analysis are summarized in Table S41, highlighting their implications for design input traceability and severity signaling.

Table S1. Confusion matrix for TFSD quadrant classification (predicted vs. expert gold labels)

Predicted/gold	Think	Feel	Say	Do
Think	30	2	1	7
Feel	3	32	2	3
Say	1	4	25	10
Do	5	2	5	23

Table S2. TFSD classification performance metrics by quadrant

Dominance threshold	Precision	Recall	Macro-F1	Single-label coverage (%)	Number of coverage (*n*=200)
0.50	0.762	0.901	0.826	93.5	187
0.55	0.801	0.892	0.843	89.0	178
0.60	0.848	0.865	0.851	82.5	165
0.70	0.912	0.801	0.852	61.5	123
0.80	0.956	0.742	0.834	42.0	84

Table S3. Multi-label agreement structure between expert annotations and pipeline outputs (150-sentence audit subsample)

Predicted pair/gold pair	Think+Do	Feel+Say	Think+Feel	Do+Say	Other pairs
Think+Do	42	3	2	6	4
Feel+Say	5	31	4	2	3
Think+Feel	6	5	18	1	4
Do+Say	4	2	1	15	3
Other pairs	3	4	2	2	21

Table S4. TFSD quadrant co-occurrence frequencies based on top-2 fallback activations

Co-occurrence pair	Number of sentences[a]	Percentage of corpus (%)
Think+Do	2,891	9.44
Feel+Say	1,172	3.82
Think+Feel	614	2.00
Do+Say	443	1.44
Think+Say	198	0.65
Feel+Do	167	0.55
Other Top-2 pairs	602	1.96
No multi-label triggered (single dominant)	23,902	77.98

Note: [a]Counts are reported at the sentence level based on top-2 fallback activations. Some sentences did not yield a valid top-2 pair or a dominant TFSD label (e.g., abstention/thresholding) and are therefore not included in the rows above; consequently, totals do not sum to 30,642.

Table S5. Quadrant-level explanation consistency results

Quadrant	Explanation consistency % (mean over 10 perturbations)	Standard deviation	Most stable token behavior
Think	84.6	4.2	Reasoning keywords remain dominant and invariant
Do	81.3	5.1	Action verbs show high persistence
Feel	78.9	6.7	Emotional adjectives fluctuate under perturbation
Say	72.4	8.3	Speech-act tokens are least frequent and least stable

Table S6. Local token contribution families and relative weights

Contribution family	Mean local weight (%)	Standard deviation	Narrative character
Cognitive-reasoning tokens	47.8	3.6	Dominates Think explanations
Behavioral-action tokens	39.2	4.8	Dominates Do explanations
Emotional-expression tokens	11.6	6.1	Dominates Feel/Say mixed explanations

Table S7. Aspect categories and dictionary size

Aspect category	Number of terms (dictionary size)	Source of curation	Intended recall character
Technical aspects	150 terms	Expert-curated via domain ontology design	Direct keyword recall
Commercial aspects	95 terms	Expert-curated from purchase/return narratives	Pricing, warranty, logistics
Emotional aspects	35 terms	Affective adjectives, cognitive tension signals	Sentiment-triggered
Interface/interaction aspects	28 terms	Usability verbs, user interface friction tokens	Interaction-relevant
Contextual/environmental aspects	22 terms	Home context, deployment constraints	Situational user experience
Experiential/first-use aspects	18 terms	Onboarding, learning-curve tokens	First-experience empathy
Total	423 terms	-	Hybrid recall

Table S8. Expert validation results for the aspect dictionary

Validation metric	Score
Inter-expert agreement (Fleiss' κ)	0.82
Dictionary relevance approval rate	88.7% of terms confirmed relevant
Disapproved or redundant terms	11.3% removed before finalizing

Table S9. Aspect recall coverage over the full corpus

Recall Layer	Coverage % (of 30,642 sentences)	*n* (sentences)
Aho-Corasick trie direct match	74.3	22,764
Transformer implicit recall	16.5	5,057
Unmatched/no aspect signal	9.2	2,821

Table S10. Aspect-level error sensitivity and reliability by category

Aspect category	Precision	Recall	F1
Technical	0.91	0.88	0.895
Commercial	0.87	0.83	0.848
Interface/interaction	0.85	0.79	0.818
Experiential/first-use	0.82	0.76	0.787
Contextual/environmental	0.80	0.71	0.753
Emotional	0.74	0.68	0.709

Table S11. Inter-expert agreement results for aspect relevance and semantic alignment

Metric	Score
Fleiss' κ (aspect relevance agreement)	0.81
Krippendorff's α (dictionary semantic alignment)	0.79
% approved terms after adjudication	87.9
% disapproved/removed terms	12.1

Table S12. Aspect-level error typology for sentence-level misalignments

Error class	*n* (sentences)	%	Characterization
Lexical ambiguity FP	21	7.0	Baseline keywords overlap without semantic support
Implicit aspect FN	37	12.3	Aspect present but not triggered in Aho-Corasick trie direct recall
Quadrant-aspect drift	14	4.7	Aspect weight high but quadrant narrative intent shifts
Compound aspect over-assignment FP	18	6.0	Transformer fallback assigns multiple aspects excessively
False semantic match FP	9	3.0	Sentence-BERT similarity triggers wrong aspect
Corrected via Adjudication	48	16.0	Resolved by longest-match and hierarchy policy
No error (exact expert match)	153	51.0	Perfect alignment

Abbreviations: BERT: Bidirectional encoder representations from transformers; FN: False negative; FP: False positive.

Table S13. Recall layer reliability comparison under expert-adjudicated conditions

Recall layer	Precision	Recall	F1	Notes
Aho-Corasick trie direct	0.89	0.82	0.854	Direct keyword match only
Transformer semantic fallback	0.83	0.76	0.793	Synonym and idiom inference
Hybrid (final adjudicated)	0.91	0.87	0.889	Adjudicated evidence stack

Table S14. Most reliable aspect triggers by expert approval (top-15 token families)

Aspect token family	% expert approval	Recall character
Battery pain keywords	94.3	Cognitive+action adjacency
Support interaction verbs	91.8	Say sensitivity risk high
Warranty/purchase terms	89.6	Logistics and trust narratives
Cleaning performance verbs	88.4	Experience-stage alignment
Onboarding/first-use cues	83.7	Learning-curve empathy
Contextual home-constraint terms	79.1	Environmental user experience

Table S15. Aho-Corasick trie direct-recall baseline coverage over the full corpus

Metric	Score (%)
Aho-Corasick trie coverage %	74.3
Number of sentences matched via Aho-Corasick trie	22,764
Number of unmatched sentences after Aho-Corasick trie	7,878 (25.7)

Table S16. Transformer semantic recall for unmatched sentence recovery

Recall layer	Coverage % (of corpus)	Number of sentences	Recall character
Transformer fallback	16.5	5,057	Implicit semantic recall
Still unmatched	9.2	2,821	No aspect signal

Table S17. Intersection reliability between Aho-Corasick trie and transformer-based decisions

Metric	Score
Aho-Corasick trie versus expert precision (direct match reliability)	0.92
Transformer fallback versus expert precision	0.86
Fallback adjudicated F1 (Top-2+hierarchy-regulated)	0.88
Explanation divergence rate under fallback	14.7%
Expert-approved fallback routing	85.3%

Table S18. Sentiment intensity distribution in the expert-labeled audit sample

Sentiment class	Number of sentences	%
P++	56	28.0
P+	70	35.0
NEU	38	19.0
N–	19	9.5
N––	17	8.5
Total	200[a]	100

Note: The total is 200 because one record was removed during audit reconciliation to maintain a consistent sample size.

Table S19. Superclass aggregation reliability for sentiment intensity labels

Metric	Score
Gain versus expert precision	0.93
Pain versus expert precision	0.90
Neutral versus expert precision	0.87
Superclass exact agreement	88.1%
Cohen's κ (3-class agreement)	0.84
Superclass macro-F1	0.90

Table S20. Ordinal sentiment confusion matrix showing polarity neighborhood errors

Predicted/gold	P++	P+	NEU	N–	N––
P++	48	6	1	1	0
P+	5	60	3	3	0
NEU	2	5	30	1	1
N–	1	3	1	14	2
N––	0	1	0	2	15

Table S21. Customer journey map stage dictionary vector specification

Customer journey map stage	Number of embedding vectors	Construction rationale	Representative semantic cues
Awareness	8	Discovery-triggered entry points	"Saw an ad," "just found out," "discovered it online," "heard about it"
Research	12	Intent-driven information seeking	"Reading specs," "checking reviews," "watching comparison videos," "learning features"
Comparison	10	Trade-off evaluation	"Better than," "vs.," "compared to," "considering alternatives," "not sure which to choose"
Purchase	6	Commitment decision signals	"Ordered it," "bought it," "payment went through," "finally purchased," "checkout done"
Experience	18	Post-deployment usage narratives	"First time using," "daily use," "on carpet," "on hardwood," "during cleaning," "battery lasted"
Loyalty	7	Retention and satisfaction stability	"Still works after months," "reliable," "my go-to," "using it for a long time," "trust this brand"
Advocacy	9	Recommendation and warning	"Highly recommend," "you should buy," "worth it," "avoid this," "telling everyone," "warning others"

Table S22. Distribution of customer journey map stages across the full corpus

Customer journey map stage	Number of sentences	% of corpus	95% confidence interval (binomial)
Awareness	388	1.48	1.34–1.62
Research	2,555	9.73	9.40–10.06
Comparison	2,916	11.10	10.75–11.45
Purchase	996	3.79	3.57–4.01
Experience	21,206	80.74	80.20–81.28
Loyalty	1,624	6.18	5.91–6.45
Advocacy	957	3.64	3.43–3.85

Table S23. Customer journey map stage-transition typology and narrative jump frequencies

Transition type	*n*	% of all stage assignments	Narrative character
Sequential transitions	26,133	85.3	Canonical stage progression
Experience→Advocacy jump	1,042	3.4	Recommendation surge after use
Research→Loyalty jump	319	1.0	Early trust commitment
Awareness→Advocacy jump	141	0.5	Viral discovery recommendation
Comparison→Purchase direct	823	2.7	Fast trade-off adjudication
Other non-sequential jumps	1,180	3.9	Adjacency-driven narrative alternation
Total stage assignments	30,642	100	-

Table S24. Local customer journey map stage inference reliability (audit sample)

Customer journey map stage	Precision	Recall	F1
Awareness	0.88	0.79	0.83
Research	0.90	0.87	0.89
Comparison	0.91	0.84	0.87
Purchase	0.94	0.90	0.92
Experience	0.96	0.95	0.96
Loyalty	0.81	0.77	0.79
Advocacy	0.93	0.85	0.89
Mean (all stages)	0.90	0.86	0.88

Table S25. Cosine similarity threshold sweep results for customer journey map stage classification

Cosine threshold	Precision	Recall	F1 score
0.25	0.884	0.921	0.902
0.30	0.901	0.907	0.904
0.35	0.912	0.884	0.897[a]
0.40	0.934	0.812	0.869
0.45	0.957	0.741	0.835

Note: The value marked with "a" denotes the optimal cosine similarity threshold selected based on the precision-recall trade-off and used in all downstream analyses.

Table S26. Confusion matrix for customer journey map stage classification (predicted vs. expert gold labels)

Predicte/gold	Awareness	Research	Comparison	Purchase	Experience	Loyalty	Advocacy	Total
Awareness	11	1	1	0	2	0	0	15
Research	1	13	2	1	1	0	0	18
Comparison	0	2	12	1	1	0	0	16
Purchase	0	1	1	15	1	0	0	18
Experience	1	1	1	1	16	1	0	21
Loyalty	0	0	0	0	2	11	2	15
Advocacy	0	0	0	0	1	2	14	17
Total	13	18	17	18	24	14	16	120

Table S27. Stage-specific performance metrics for customer journey map stage classification

Stage	Precision	Recall	F1
Awareness	0.86	0.78	0.82
Research	0.72	0.70	0.71
Comparison	0.76	0.86	0.80
Purchase	0.86	0.86	0.86
Experience	0.76	0.67	0.71
Loyalty	0.41	0.58	0.48
Advocacy	0.42	0.42	0.42

Table S28. Customer journey map stage-transition typology and jump rates across the full corpus

Transition type	Number of sentences	% of corpus	Association character
Canonical sequential transitions	26,133	85.3	Adjacent narrative progression
Experience→Advocacy jump	1,042	3.4	Post-use recommendation surge
Comparison→Purchase direct	827	2.7	Rapid trade-off commitment
Research→Loyalty jump	306	1.0	Early trust anchoring
Awareness→Advocacy viral jump	153	0.5	Viral discovery recommendation
Advocacy→Loyalty retention drift	169	0.55	Recommendation+long-term satisfaction cues
Experience→Loyalty jump	530	1.73	Post-use long-term reliability narrative
Research→Advocacy warning jump	95	0.31	Advice-heavy narratives before usage
Other non-sequential jumps	1,387	4.53	Semantic adjacency alternations
Total	30,642	100	-

Table S29. Stage-transition precision, recall, and F1 scores by transition family

Transition family	Precision	Recall	F1
Sequential transitions	0.89	0.86	0.875
Non-sequential narrative jumps	0.83	0.74	0.784

Table S30. Most frequent adjacency-driven customer journey map stage transitions and narrative triggers

Jump path	*n*	Narrative trigger character
Experience→Loyalty	531	Reliability+severity of use cues
Think–Do heavy Purchase jumps	402	Reasoning+action adjacency
Feel-Say heavy Advocacy drift	287	Advice+emotional salience

Table S31. Narrative pattern, TFSD quadrant behavior, and sentiment intensity characterization

Narrative pattern	Expected dominant quadrant behavior	Typical sentiment intensity signature	Audit agreement (%)	Interpretation note
Enthusiastic praise	Feel	P++or P+	92.5	High affect salience, low entropy
Sharp complaint/return intent	Do or Think	N−− dominant	90.2	Action-framed pain narratives
Functional bipolar evaluation	Think + Do	Mixed P++and N−	86.7	Reasoning-to-action alternation
Emotional advice/social broadcast	Feel + Say	P++or N−−	83.4	Recommendation or warning spikes
Mixed relief/problem-solved narratives	Do + Think	P++and N−− in same sentence	79.8	Sentiment extremes co-occur
Interface friction/usability pain	Think	N−or NEU	88.1	Interaction-cost dominant
Trust anchoring without Purchase	Say or Feel	P+or NEU	76.4	Early Loyalty or Advocacy drift
Post-use retention narrative	Think + Do	P++or NEU	84.9	Peaks in Loyalty
Warning-framed pre-experience narrative	Say	N−− or N−	71.6	Rare-class Say sensitivity
Comparative decision compression	Think	NEU or P+	80.3	Comparison→Purchase direct

Table S32. Performance comparison across ablation module configurations

Model	Accuracy	Macro-F1	FP rate (%)	FN rate (%)	Say F1 sensitivity	Notes
M1 (keyword baseline)	0.71	0.68	18.9	22.4	0.55	High lexical FP on Think-Do
M2 (RoBERTa only)	0.79	0.77	12.3	16.7	0.69	Lower coverage imbalance
M3 (BART only)	0.81	0.78	11.8	14.9	0.71	Smoothing errors in Feel
M4 (transformer aspect only)	0.66	0.63	27.6	31.8	0.51	Over-assignment FP
M5 (hybrid ensemble; proposed model)	0.824[a]	0.801[a]	9.7[a]	12.8[a]	0.74	Best balance and traceability

Note: Values marked with "a" indicate the proposed hybrid ensemble (M5) selected as the optimal configuration and used in subsequent analyses.
Abbreviations: BART: Bidirectional and auto-regressive transformers; FN: False negative; FP: False positive; RoBERTa: Robustly optimized bidirectional encoder representations from transformers pretraining approach.

Table S33. Macro-F1 performance gains of the proposed model relative to baseline configurations

Comparison	F1 delta gain
Proposed (M5) versus keyword baseline (M1)	+0.21 macro-F1
Proposed (M5) versus RoBERTa only (M2)	+0.12 macro-F1
Proposed (M5) versus transformer aspect only (M4)	+0.27 macro-F1

Abbreviation: RoBERTa: Robustly optimized bidirectional encoder representations from transformers pretraining approach.

Table S34. Bias divergence reduction after ensemble adjudication

Metric	Score
Automated versus expert label divergence	13.6 (before adjudication)
Divergence after ensemble adjudication	7.9
Say-axis divergence reduction	−41.2 relative improvement
Think–Do adjacency divergence reduction	−33.8
Feel-Say divergence reduction	−28.5

Table S35. Module-level latency breakdown and scalability characteristics of the pipeline

Pipeline module	Latency (seconds)	% of total runtime	Scalability character
Sentence segmentation	95.2 s	12.7	Linear to corpus size
TFSD zero-shot classification (ensemble)	312.8 s	41.7	GPU-bound, batch scalable
Aho-Corasick aspect matching	48.6 s	6.5	O (n) streaming, memory efficient
Transformer semantic aspect fallback (ABSA)	119.4 s	15.9	Embedding-retrieval bound
Sentiment intensity scoring (LCF-BERT)	61.7 s	8.2	Token-local, stable latency
CJM stage cosine alignment	73.3 s	9.8	Similarity-matrix bound
Intersection synthesis and adjudication	41.2 s	5.5	Rule-governed, negligible drift
Total observed runtime	753.8 s	100	End-to-end runtime

Abbreviations: ABSA: Aspect-based sentiment analysis; BERT; Bidirectional encoder representations from transformers; GPU: Graphics processing unit; LCF: Local-context focused; TFSD: Think–Feel–Say–Do.

Table S36. Stakeholder-intervention mapping matrix for evidence-driven design governance

Human-computer interaction stakeholder	Primary need	Section-level evidence utilized	Design intervention class	Expected output artifact
User experience researchers	Friction and pain/gain detection	1, 4, 7, 9	Intent reliability audit+narrative adjacency governance	Persona evidence report
Interaction designers	Conversational and UI friction	1.4, 4.2, 7.2	Conversational script rules+UI token friction tagging	Dialogue trace schema
Product designers	Feature severity and aspect priority	2, 5, 13, 14	Aspect-aware feature ranking+multi-label adjudication	Design decision table
System designers	Scalability and throughput	3, 7, 16	Streaming and batch architecture governance	Latency waterfall
Service designers	Support experience narratives	7, 9, 10, 12	Support-friction and resolution cue modeling	Service-CJM audit
Data governance leads	Bias control and auditability	6, 15	Expert agreement κ + divergence reduction policy	Bias audit table

Abbreviations: CJM: Customer journey map; UI: User interface.

Table S37. Aspect-driven user experience and design intervention examples

Section-level evidence reference	Detected aspect family	Stage neighborhood	Suggested design intervention
7.2	Battery-pain keywords	Experience → Loyalty	Adaptive charging micro-feedback
1.4	Think–Do reasoning/action adjacency	Comparison stage	Decision-compression user experience
4.2	Emotional adjective instability	Advocacy stage	Controlled broadcast salience tagging
5.4	Dictionary relevance κ=0.82	Research stage	Ontology-aligned aspect tagging
12.2	Experience → Advocacy jump 3.4%	Advocacy stage	Recommendation-spike user experience widgets

Table S38. Priority governance policy for aspect-driven design interventions

Intervention priority	Trigger condition	Governance logic
P1	F1 ≥0.85+coverage ≥80% aspects	High-reliability decision layer
P2	F1 0.78–0.85+adjacency high	Interpretable adjacency confusions
P3	Say sensitivity high+implicit recall	Variance disclosure, fallback-guarded
P4	Transformer-only over-assignment risk	Requires hierarchy adjudication

Table S39. Pipeline inference summary

Pipeline layer	Output
Think–Feel–Say–Do quadrant adjudication	Do-dominant narrative with Think and Feel adjacency
Aspect polarity and severity	Pain: battery/charging (high), pain: cleaning/suction (medium-high), pain: customer support (medium), pain: interface/application (medium)
Touchpoint centrality	Device experience, mobile application interaction, customer service
Customer journey map phase alignment	Experience stage (highest cosine density)
Stakeholder design input artifact	Severity-ranked requirement note produced at the statement level

Table S40. Severity-ranked aspect distribution derived from pipeline findings

Aspect	Gain (%)	Pain (%)	Priority level
Battery/charging	38	41	1 (critical)
Cleaning/suction	53	30	2 (high)
Interface/application	29	49	3 (medium high)
Warranty/return	34	39	4 (medium)
Customer support	33	48	5 (medium)

Table S41. TFSD quadrant alternation and narrative adjacency patterns

Pattern	Interpretation
Think ↔ Do	Users interleave reasoning and action statements within the same narrative
Feel ↔ Say	Emotional intensity occasionally shifts into advice/broadcast statements
Do → Think	Users describe actions then reflect with evaluative reasoning cues
Say sparsity	Speech acts are rare but strategically important for design input severity flags

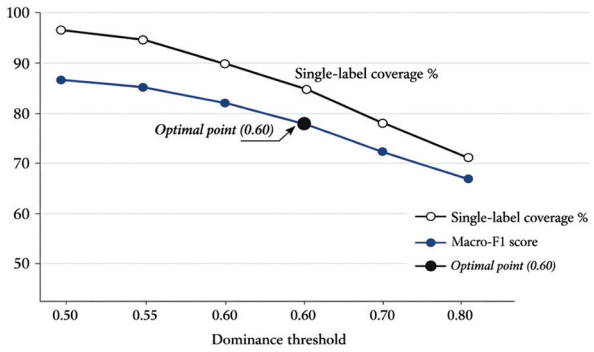

Figure S1. Think–Feel–Say–Do dominance threshold trade-off showing threshold versus F1 score and coverage (%), with the optimal operating point marked at 0.60 dominance