CARO MILKEZ
MIT PROFESSIONAL EDUCATION — APPLIED AI & DATA SCIENCE | MARCH 2026

GENERATIVE AI CAPSTONE | FINAL REPORT

REAL-TIME RETAIL FEEDBACK INTELLIGENCE

EXECUTIVE SUMMARY

ChicStyle processes ~23,000 customer reviews, yet current approaches fail to convert this volume into timely, actionable decisions. The core challenge is not scale—but ambiguity: most reviews contain mixed signals, where customers simultaneously express satisfaction and highlight critical issues. Traditional analysis treats these as neutral. In practice, they represent the highest-risk segment for silent churn.

A Generative AI feedback intelligence system was developed to address this gap. In a single pass the system transforms unstructured reviews into structured, decision-ready outputs—identifying sentiment (including nuanced mixed feedback), product issues, urgency levels, and recommended actions.

The solution enables a shift from descriptive analytics to operational decision-making, allowing the business to prioritize high-risk and high-impact customer segments.

IMPACT

- Faster detection of product and service issues
- Improved customer trust through personalized responses
- Reduction in returns driven by sizing and quality improvements
- Prioritizing issues by urgency improves how teams make and act on decisions.

RECOMMENDATION:

Begin with a pilot in Dresses and Tops, ensure proper governance and monitoring, and gradually roll out across operations.

PROBLEM SUMMARY

ChicStyle is not getting enough value from its customer feedback. Customer reviews contain rich, specific signals about product quality, fit, fabric, and design, but these signals are locked inside unstructured text. Insights do not translate into operational action at the speed required especially in the holiday season. The gap is driven by three key challenges.

1 FRAGMENTED SYSTEMS COMPOUND COMPLEXITY AT SCALE

Traditional NLP approaches require multiple independent models to handle what is effectively a single business task. Category classification, sentiment detection, summarization, response generation, and insight extraction would each require separate models, datasets, and maintenance cycles.

As review volumes surge—often tripling during peak periods—this fragmented setup becomes harder to manage, more expensive to maintain, and too slow to support timely decision-making.

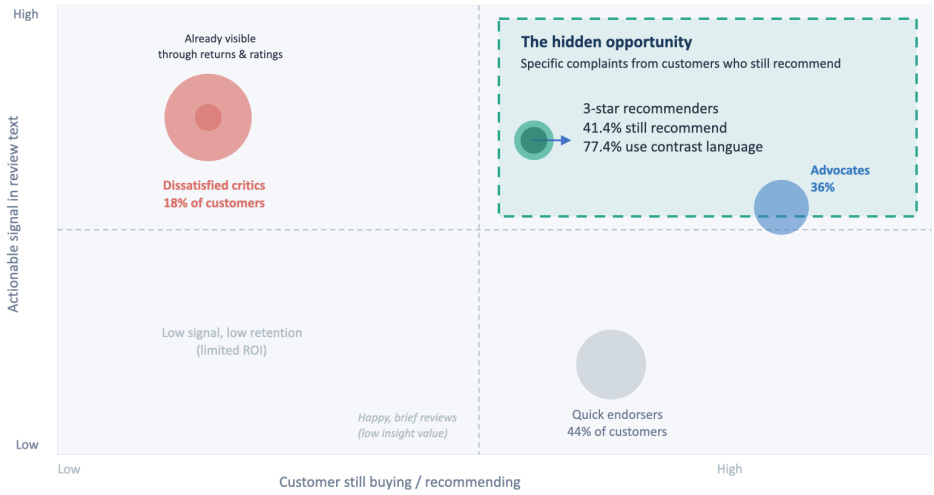
2 CUSTOMER LANGUAGE IS INHERENTLY AMBIGUOUS

Analysis of the full dataset reveals that 77.4% of mid-rated reviews contain contrast language ("love the colour but the sizing is off"). And 41.4% of these customers still recommend the product despite expressing complaints. Traditional three-class sentiment models (positive, negative, neutral) cannot capture this nuance, leading to misrouted or missed issues.

3 INSIGHTS DO NOT TRANSLATE INTO TIMELY ACTION

Customer feedback is effectively trapped within analytics workflows, creating a disconnect between insight and execution. Product, buying, marketing, and customer service teams depend on delayed, periodic reports rather than real-time signals,

resulting in slow response to product issues. This lag allows problems—such as sizing or quality defects—to scale into higher return rates, missed sales opportunities, and avoidable customer churn.



The opportunity is not in obvious dissatisfaction, but in fixing issues from customers who still recommend.

COST OF INACTION

The business impact compounds over time. A sizing issue in a high-volume product line that goes undetected for weeks continues to generate returns, negative reviews, and lost repeat purchases throughout that period. Dissatisfied customers — particularly those who write detailed, visible reviews — influence the purchasing decisions of other shoppers. The longer the delay between customer signal and business response, the more revenue is lost and the harder it becomes to recover customer trust.

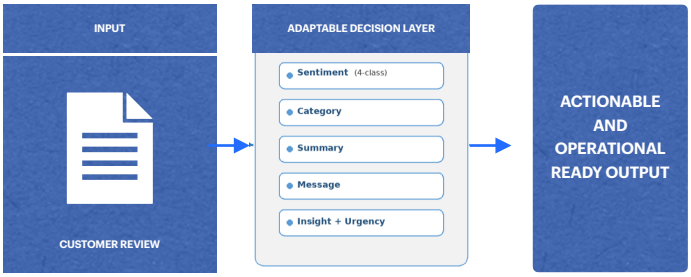
2. SOLUTION

The GenAI system converts each customer review into structured, decision-ready outputs in a single pass, enabling faster and more consistent action across the business.

The system generates five coherently connected outputs in a single step:

OUTPUT FIELD	WHAT IT DELIVERS	WHO ACTS ON IT
CATEGORY	Classifies the product into one of six department categories (Dresses, Tops, Bottoms, Intimate, Jackets, Trend)	Buying, Merchandising
SENTIMENT	Four-label classification (Positive, Negative, Neutral, Mixed-Negative) capturing the nuance that three-class systems miss	Customer Experience, Product
SUMMARY	A concise 25-word capture of the customer's core message	All teams
PERSONALISED MESSAGE	A tone-appropriate, specific response ready for customer communication	Customer Service
RETAIL INSIGHT + URGENCY	Root cause identification, responsible department, concrete action, and a four-tier urgency tag (Critical, Actionable, Monitor, Insight-Only)	Operations, Quality, Leadership

Traditional NLP requires multiple systems, each with its own data and maintenance—making it costly and unreliable in the face of ambiguous, mixed-signal feedback.



The GenAI approach replaces this fragmented architecture with a single, unified decision layer that generates all required outputs in one pass. More importantly, it is inherently adaptable: changes in business requirements—such as introducing new sentiment categories, urgency levels, or product groupings—can be implemented through prompt updates rather than retraining models. This flexibility was validated in practice, where expanding the sentiment framework required only a prompt adjustment, not a system rebuild.

OPERATIONAL READY OUTPUT:
CATEGORY: Blouse | SENTIMENT: Mixed negative
SUMMARY: Customer loved the top's design but found sizing issues with the chest and waist, leading to dissatisfaction. | RETAIL INSIGHT: Actionable | Sizing inconsistencies causing returns — Tops department: review size specifications and adjust fit for better accuracy. | PERSONALIZE MESSAGE: It's great to hear you loved the design! We'll pass your sizing feedback to our team for future improvements.

3. KEY INSIGHTS

Unsupervised learning on the full dataset identified four behaviorally distinct customer types. Each generates different signals and requires a different operational response.

A one-size-fits-all approach misallocates resources.

HIGHEST

~2%

LOYAL POWER REVIEWERS

HIGHEST LEVERAGE

These repeat customers write detailed, thoughtful reviews that other buyers read and trust. Their sentiment is a leading indicator of broader market perception.

Influential | Mixed ratings
Extremely high feedback count (28+ reviews)
Longest reviews (89 words)
Oldest segment (48 avg.)

URGENT

~18%

DISSATISFIED VOCAL CRITICS

REVENUE RISK

These customers are actively signaling dissatisfaction and are at immediate churn risk. Their feedback concentrates on specific, fixable issues.

Unhappy | Rating 2.3
near-zero recommendation rate
moderate review length
Highest conflict rate (4.6%) Decent ratings but still won't recommend

MEDIUM

~36%

SATISFIED DETAILED ADVOCATES

GROWTH OPPORTUNITY

These customers write the kind of specific, persuasive reviews that drive new customer acquisition. Underutilised as a strategic asset.

Moderate detailed | Rating 4.6
100% recommend
Long detailed reviews (84 words)

MONITOR

STABLE BASE

~44%

HAPPY QUICK ENDORSERS

High satisfaction but low engagement depth. No urgent action required, but risk of taking this segment for granted.

Promoter | Rating 4.7
100% recommend
Shortest reviews (38 words)
Lowest feedback count

Sizing and fit are the primary cost drivers

Sizing inconsistencies emerge as the single largest driver of product issues. Across the evaluation, the majority of customer feedback—particularly in high-volume categories such as Tops and Dresses—points to fit-related problems. This is not a perception issue; it is a product specification failure that directly translates into returns, margin erosion, and avoidable customer dissatisfaction.

80% of all reviews —

In the 60-review evaluation sample processed by the recommended technique (Few-Shot V2), fit and sizing appeared in 48 of 60 retail insights, with Tops (28 reviews) and Dresses (14 reviews) accounting for the majority of flagged cases. Fabric/material appeared in 24 of 60 insights (40%), and design in 18 of 60 (30%)

Concrete actions surfaced

The system demonstrates disciplined prioritization of customer issues. Over half of all reviews are classified as **actionable (31 out of 60)**, indicating that the pipeline is consistently surfacing concrete, fixable product and experience issues rather than general sentiment. An overview of the priority level is displayed below.

PRIORITY LEVEL	MEANING	WHO ACTS ON IT	BUSINESS PURPOSE
CRITICAL	Defective, completely wrong item received	Customer Service	Immediate recovery
ACTIONABLE	Recurring product issue (e.g., sizing, quality)	Product & Merchandise	Reduce returns / fix cause issue
MONITOR	Isolated issue	Product Analytics	Track emerging patterns
INSIGHT ONLY	Positive feedback	Marketing	Leverage growth

Ambiguity problem is real and measurable

The expanded four-label sentiment system captures nuance that the standard three-label approach misses entirely. The Mixed-Negative category identifies customers who appreciate aspects of a product but have a concrete complaint that affected their experience. These are operationally the most valuable signals: they point to specific, fixable issues in products that are otherwise successful.

The highest-value product insights are not coming from dissatisfied customers, but from those who still buy and recommend while highlighting specific issues.

77.4%

of mid-rated reviews containing mixed signals

41.4%

still recommending

These customers represent a hidden opportunity: **Targeted product improvements here can reduce returns and strengthen loyalty without the cost of customer recovery.**

4. SYSTEM PERFORMANCE AND VALIDATION

The GenAI decision system was evaluated across three prompting approaches, each tested in baseline and enhanced configurations. The approaches range from minimal

instruction (rules only) to template-guided (standardised examples) to structured reasoning (step-by-step logic before output generation).

CONFIGURATION / METHOD	APPROACH	STRENGTH
MINIMAL INSTRUCTION ZERO-SHOT	Rules and constrains only	Lowest cost, fastest to deploy
	LIMITATION Highest variability on ambiguous reviews; misclassified positive reviews as mixed-negative	
TEMPLATE GUIDE FEW-SHOT	Standardised examples demonstrating expected output quality and format	Most consistent performance across all customer segments; lowest variance (SD = 0.039 under strict evaluation)
	LIMITATION Requires maintenance of example library; higher token cost per review	
STRUCTURED REASONING CHAIN OF THOUGHT	Step-by-step reasoning protocol before output generation	Reason through each review before generating outputs, producing richer analysis on ambiguous cases
	LIMITATION Less predictable results at volume. Adding strict rules constrained the model's reasoning, reducing output quality; highest cost per review	

PRODUCTION RECOMMENDATION

TEMPLATE GUIDE

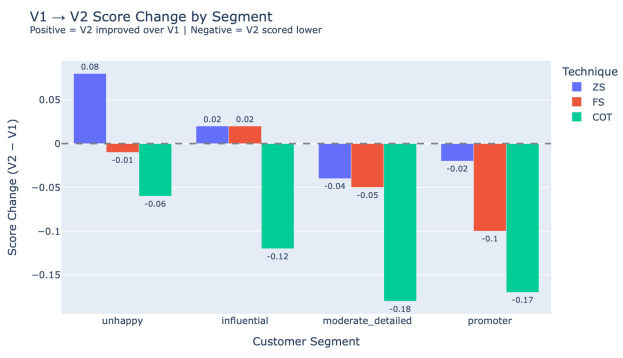
CONSISTENCY ACROSS SEGMENTS

Under realistic business evaluation criteria, consistency becomes more important than peak performance.

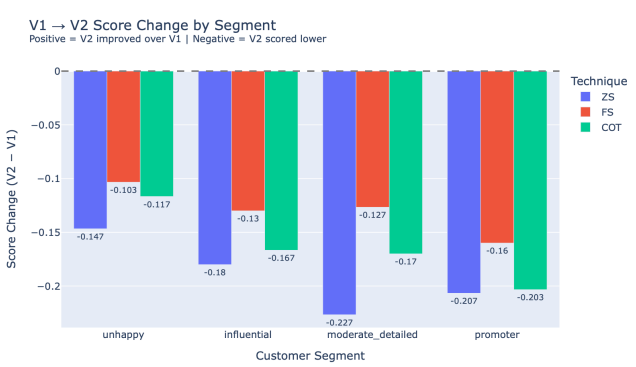
The template-guided (FS_V2) approach delivers the most reliable performance across all customer types, with the lowest variance of any configuration tested. It performs particularly well on the two segments that carry the highest business risk: dissatisfied customers at risk of churn, and high-visibility reviewers who influence purchasing decisions.

TECHNIQUE	OVERALL	UNHAPPY	INFLUENTIAL	MODERATE	PROMOTER
BASELINE EVALUATION (V1)					
Minimal instruction	940	900	940	940	980
Template Guided	898	873	900	873	947
Structure reasoning	905	867	907	880	967
BUSINESS AWARE ENHANCED EVALUATION (V2)					
Minimal instruction	750	753	760	713	773
Template Guided	768	770	770	747	787
Structure reasoning	741	750	740	710	763

Robustness was validated by expanding the evaluation from 20 to 60 stratified reviews. While individual scores shifted, the relative performance of each approach remained stable, confirming that the recommendation is not sample-dependent and can be trusted to scale to the full dataset (~23,000 reviews).



Performance of prompt technique at 20 review sample



Performance of prompt technique at 60 review sample

In contrast, the structured reasoning approach achieved the highest scores under lenient conditions but proved unstable under stricter, business-aligned evaluation. Its customer messages tended toward robotic, repetitive phrasing. This is a critical limitation: a system that performs well on straightforward reviews but fails on

complex, ambiguous cases introduces operational risk. The template-guided approach solves both problems: the examples standardise output structure while also controlling tone and brand voice.

PROMPT DESIGN

PROMPT STRUCTURE	BASELINE V1	ENHANCE PROMPT V2
SYSTEM PROMPT	Generic ("helpful AI assistant")	Domain-specific ("expert customer feedback analyst for ChicStyle")
SENTIMENT LABELS	Uses 3 Positive, Negative, Neutral	uses 4 (adds Mixed-Negative)
RULES AND GUARDRAILS	No explicit rules	Has sentiment decision rules, "do not assume" rules, consistency checks
URGENCY	Does not include urgency classification	Includes 4-tier urgency framework (Critical, Actionable, Monitor, Insight-Only)
CONSISTENCY CHECK	Has none	Has mandatory alignment check between sentiment, urgency, and insight
FEW-SHOT EXAMPLES	Has 3 examples (one per sentiment)	Has 4 examples (including Mixed-Negative)

TEMPLATE DESIGN

Templates span the full output space, with one example per sentiment label to standardize format, tone, and quality. Based on the ~22,600 review dataset, each example illustrates the complete five-field structure.

REVIEW

" I tried this sweater on in the store. I wanted to love it so badly as it is soft and feminine, and would be a great casual layering piece. However, there are pockets in the sweater that are right on seam, and they actually bubble out from the natural drape of the sweater."

EXPECTED OUTPUT

Category: Fine gauge. | **Sentiment:** Mixed-Negative | **Summary:** Soft and feminine sweater, but pocket placement bulges outward and ruins the drape. | **Retail Insight:** Monitor | Seam pocket distorts drape — Tops department: design team to review pocket construction for Fine gauge. | **Personalized Message:** The softness clearly won you over and we love that. We'll share your pocket placement feedback with our design team.

EVALUATION METHODOLOGY

Two different evaluation approaches were used: a general quality check for the baseline models and a stricter, five-criteria rubric for the enhanced versions (covering sentiment, category, summary, personalization, and actionability). Because these use different standards, results are only compared within the same evaluation method. This is a known limitation and has been taken into account in the analysis.

Baseline Quality

JUDGE V1 —
Evaluates the overall quality of the generated outputs against the original review, focusing on fundamental criteria such as accuracy, completeness, sentiment alignment, and clarity. It ensures that all required fields (Category, Sentiment, Summary, Personalized Message, Retail Insight) are present and that the output correctly reflects the customer's intent.

Holistic 0-1 score used for Baseline Prompts

Business-Aware

JUDGE V2 —
Extends the evaluation by incorporating business relevance and decision-making value. In addition to accuracy and completeness, it assesses ambiguity handling (e.g., mixed sentiment), consistency across output fields, and the quality of actionable insights. It emphasizes whether the generated Retail Insight is specific, non-generic, and useful for operational teams, as well as whether urgency and sentiment are interpreted correctly in complex reviews.

Strict 5-criterion rubric (0.20 per criterion). Includes penalty rules and score clamping

INDEPENDENT VALIDATION: RECOMMENDATION PREDICTION

As an independent validation, the system was tested on its ability to detect customer dissatisfaction based solely on review text. It reliably identified all non-recommendation cases, ensuring that no critical negative signals are missed.

88% accuracy / 100 reviews

correctly identified every single non-recommendation case (recall = 1.00), ensuring no critical negative signal is missed.

12% of misclassifications concentrated in reviews with contrast language, the same ambiguity pattern identified in the exploratory analysis, validating that the primary challenge is not detection, but interpretation.

5. RECOMMENDATION FOR IMPLEMENTATION

DEPLOYMENT STRATEGY

The full pipeline spans four stages: data intake, LLM-based structuring (GenAI layer), post-processing enrichment, and business delivery. This report validates the core GenAI layer — the component that transforms unstructured reviews into structured, decision-ready outputs: what the issue is, how urgent it is, who should act on it, and how to respond to the customer.

The recommended rollout begins in the Tops department, where the dataset shows the highest review concentration (10,468 of ~22,600 reviews), the strongest signal density around sizing and fit issues, and the clearest path to measurable impact. Starting narrow ensures faster validation and sharper ROI measurement before expanding to Dresses, Bottoms, and other categories.

PHASE 1: CORE INSIGHT GENERATION

The goal of this phase is to deploy the validated GenAI layer in a controlled pilot, establishing the operational model around it.

Before going live, the evaluation is expanded from the 60-review stratified sample to a larger, more representative set covering all six departments and customer segments. This establishes production-level confidence in the system's outputs before any customer-facing interaction.

OPERATING MODEL

Human-in-the-loop decision system

The operational purpose is a controlled operational model focused on validating the reliability of the GenAI signal layer before scaling. Human teams retain **final decision authority** on customer responses and product interventions

- **Customer Experience (CX) teams** review prioritized outputs—primarily Actionable insights—within the same business day, ensuring accuracy in sentiment, urgency,

and response quality before any customer interaction. The team logs corrections for immediate quality control, and building the labelled dataset that in future stages can enable fine-tuning.

- **Product teams** validate whether surfaced issues (e.g., sizing, fabric) reflect real business problems.
- **Data & Engineering** ensure system stability and monitor for drift.
- **Leadership** evaluates whether outputs are consistently accurate, usable, and actionable enough to proceed, making this phase a validation gate where responsiveness is deliberate and human-controlled rather than automated.

QUALITY MAINTENANCE

The Template Guided approach is based on a Few-Shot prompt that uses a small set of examples to guide how the model behaves. As products evolve and customer language shifts, these examples should be updated regularly.

- **Quaterly prompt review:** aligned with seasonal launches, is sufficient to keep the system accurate and relevant.

If ChicStyle launches a new activewear line in spring, the examples are all about blouses and knits. If customers start complaining about sustainability packaging in Q3, sample should cover that pattern.

- **Output format compliance:** missing fields, out-of-range categories, exceeded word limits, should be tracked continuously as a pipeline health metric.

DRIFT DETECTION

Three signals should be monitored on a rolling basis, they are early warning signals where each requires human investigation.

- **Sentiment distribution drift:** if Mixed-Negative proportions shift more than 10 percentage points from the pilot baseline
- **Urgency inflation or deflation:** the 60-review evaluation showed 52% Actionable, 38% Insight-Only, 10% Monitor, 0% Critical, sustained deviation from these proportions warrants investigation.
- **Output format consistency:** malformed or incomplete structured outputs.

PHASE 2: PROMPT ENGINEERING MATURITY (3–6 MONTHS)

The evaluation showed that different prompt approaches perform better on different tasks. Template-guided (Few-Shot V2) delivers the most consistent output across all segments and is the production standard. Structured reasoning (Chain-of-Thought) achieved the highest peak performance under lenient evaluation (905 on 60 samples) but with higher variance, making it better suited to deep-dive analysis of complex reviews than high-volume processing. Phase 2 builds on this by deploying each approach where it performs best.

The Few-Shot V2 example library is expanded to cover underrepresented departments: Intimate, Trend, and Jackets were effectively untested in the 60-review evaluation and require adjusted examples before reliable processing.

Structured reasoning (Chain-of-Thought) is deployed as an internal analytical tool for product teams, offering deeper investigation of flagged edge cases and ambiguous reviews where the template-guided approach shows its limits .

Other techniques with minimal instruction, such as Zero Shot are implemented for rapid prototyping for quick scans of new product categories or departments before building template examples

Each human correction during this phase; whether in sentiment, urgency, or messaging, becomes a labelled example, building a growing, company-specific dataset that reflects ChicStyle's products, customers, and edge cases.

PHASE 3 — OPERATIONAL INTEGRATION (6–12 MONTHS)

GenAI Core Insights layer reaches its next maturity point. By this phase, two sources of training data are available: the ~22,600 reviews with ground-truth recommendation labels already in the dataset, and the accumulated correction data from months of CX teams overriding sentiment labels and urgency tiers during Phases 1–2.

As human corrections accumulate, they create a growing dataset of company-specific labelled examples. Over time, this data could support fine-tuning a smaller, domain-specific model: one that internalises ChicStyle's brand voice, product language, and edge cases without relying on prompt templates or API dependency. Whether this step is necessary depends on how well the prompted pipeline continues to perform as it scales across all departments and review volumes.

COSTS

API processing costs remain low even with high usage. Processing 1,000 reviews costs approx. $0.24, meaning the full dataset (~22,600 reviews) can be analyzed for around $5, and the entire Tops backlog for under $3. Even with ongoing usage—including evaluation sampling and prompt iteration—annual API costs remain well below $100.

The primary investment lies in **operational adoption** rather than technology: integrating the pipeline into CX workflows, allocating human review capacity during the pilot, and maintaining prompt quality as product categories evolve. These are team-based costs that should be scoped during Phase 1.

Fine-tuning a domain-specific model in Phase 3 would further reduce per-review costs and eliminate the ongoing API dependency.

BUSINESS IMPACT

→ REDUCED DECISION LATENCY DURING PEAK PERIODS

The system enables near real-time identification of recurring product issues (e.g., sizing, fabric) that are currently buried in unstructured reviews. This reduces the time from feedback to insight from days/weeks to same-day visibility, allowing earlier intervention before issues scale.

During peak seasons when review volume triples, speed becomes a competitive differentiator.

→ SCALABLE WITHOUT REBUILDING

Adapting the system to new departments, product categories, or business requirements is a prompt update. This was validated in practice when expanding from three to four sentiment labels required only a prompt adjustment, not a system rebuild.

→ CONSISTENT, BRAND-APPROPRIATE CUSTOMER COMMUNICATION AT SCALE

Each personalised message reflects the customer's specific experience rather than defaulting to generic responses. Through the few-shot examples ChicStyle brand tone is maintained.

→ COST STRUCTURE THAT FAVORS EXPERIMENTATION

At $0.24 per 1,000 reviews, testing the system on a new department or product line is effectively free. This removes the typical barrier to piloting analytical tools, business units can try, learn, and iterate without budget approval cycles.

LIMITATIONS AND FURTHER ANALYSIS

Manageable execution risks, not fundamental barriers to deployment.

RISK & CHALLENGES	DESCRIPTION	MITIGATION
SAMPLE SIZE	The 60-review stratified evaluation is directionally robust but may not capture all edge cases across ~23,000 reviews.	Phased rollout starting with Tops and Dresses. Expand evaluation sample as production data accumulates.
LLM-AS-JUDGE BIAS	Both pipeline and judges use models from the same provider, risking self-preference score inflation.	Periodic human audit of random outputs. The independent recommendation prediction (88% accuracy, 100% negative recall) validates understanding outside the judge framework.
MODEL DRIFT	Customer language, product lines, and seasonal trends change. Accuracy degrades without recalibration.	Quarterly prompt evaluation cycles. Drift detection dashboards monitoring sentiment and urgency distribution shifts.
EVALUATION OVERHEAD	Manual effort to curate evaluation samples and maintain few-shot examples introduces ongoing operational cost.	Assign ownership to analytics team. Build example refresh into quarterly planning. Track template effectiveness.
AMBIGUITY HANDLING	Influential segment (mixed-signal, high-visibility reviews) remains hardest for all configurations. Misclassification carries disproportionate brand risk.	Human-in-the-loop for all flagged influential reviewer outputs. Dedicated handling protocol.

PROVIDER DEPENDENCY	Pipeline relies on GPT-4o-mini via OpenAI API. Behaviour and pricing can change without notice.	Correction data accumulated during Phases 1–2 builds the foundation for fine-tuning a self-hosted model, eliminating this dependency.

FURTHER ANALYSIS REQUIRED

The system demonstrates that a single GenAI layer can reliably transform unstructured customer feedback into structured, decision-ready outputs. The 60-review evaluation validates the approach directionally; the steps below establish the path to production-level confidence.

Validate with human reviewers

Before any customer-facing deployment, LLM-based evaluation should be complemented with human judgment on a representative sample (200–400 reviews) to confirm that outputs meet the quality bar the business expects.

Independent evaluation framework

The current evaluation uses the same model family to both generate and assess outputs, which introduces **self-preference bias**. A production-ready system needs an evaluation layer that works independently of the pipeline. Frameworks like DeepEval offer structured metrics. The current evaluation framework focuses on accuracy, completeness, and actionability. A dedicated faithfulness metric, would add an additional safety layer as the system scales beyond human-in-the-loop oversight.

Monitor hallucination risk in open-ended outputs

The summary, personalised message, and retail insight are are open-ended text generation. Human review catches hallucinations during the pilot, but as volume scales, **automated faithfulness checks** against the original review text would add a valuable safety layer.

Assess bias in sentiment classification

LLMs can carry subtle biases in how they interpret tone, particularly across informal language, non-native English, or culturally specific expressions. The current dataset is relatively homogeneous, but serving a broader customer base would benefit from validating that **sentiment accuracy** holds across different writing styles, ensuring no customer group is consistently misread.

Address data governance for external processing

Each review is sent to an external API, which may require data processing agreements, customer consent or anonymisation depending on applicable regulations and ChicStyle's own governance standards. These should be assessed early in the deployment process.

Assess fine-tuning feasibility

Evaluate whether a smaller, company-specific model trained on accumulated corrections and existing labels can match pipeline quality at lower cost and without API dependency.

REFERENCES

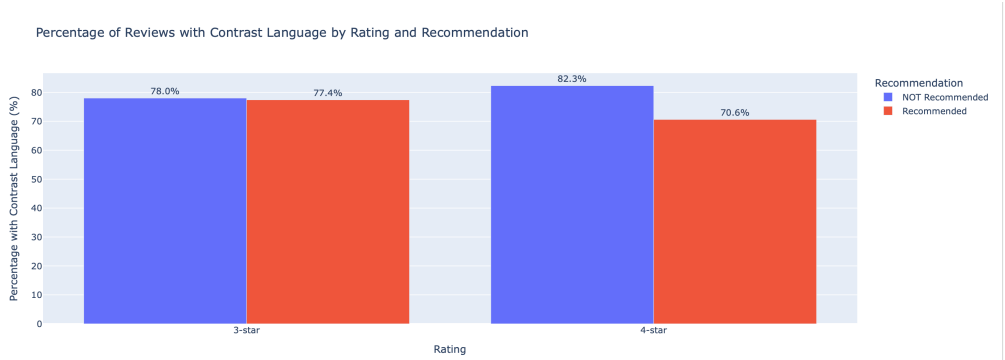
Evaluating fine tuning as a proposed solution: https://www.deeplearning.ai/the-batch/when-to-fine-tune-and-when-not-to/

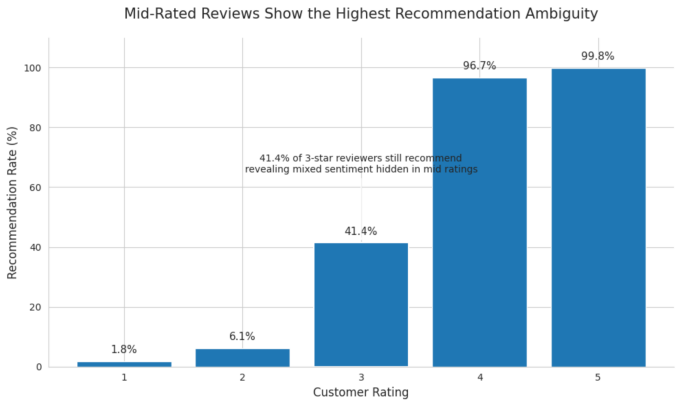
Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge: https://arxiv.org/html/2410.02736v2

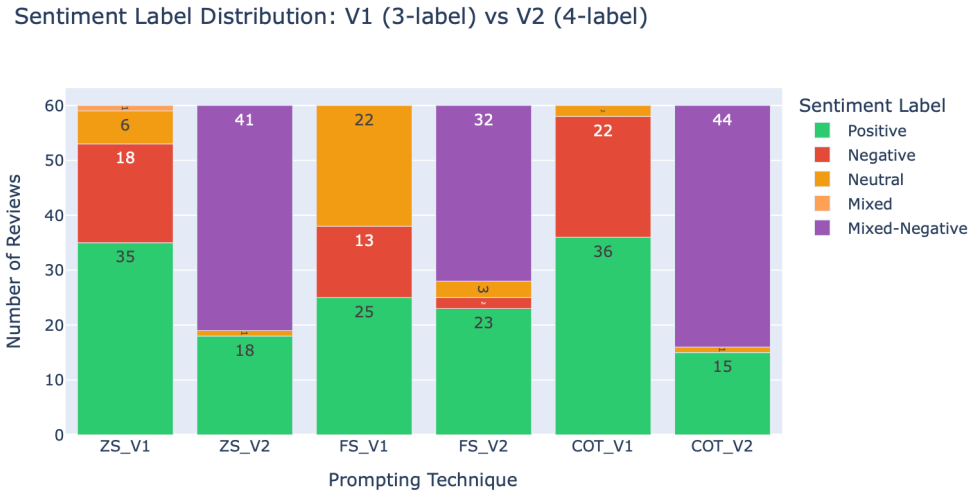
Evaluating LLMs. https://www.kdnuggets.com/top-5-open-source-llm-evaluation-platforms

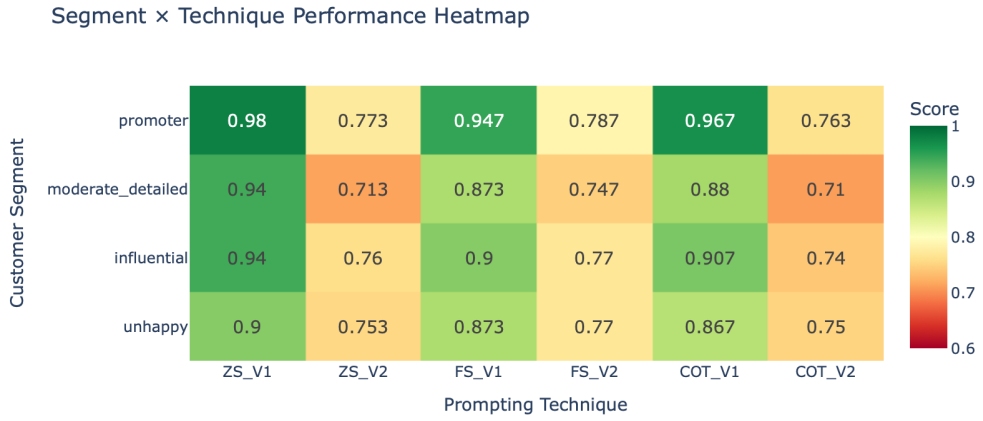
LLM Evaluation: https://github.com/alopatenko/LLMEvaluation/blob/main/LLMEvaluation.pdf

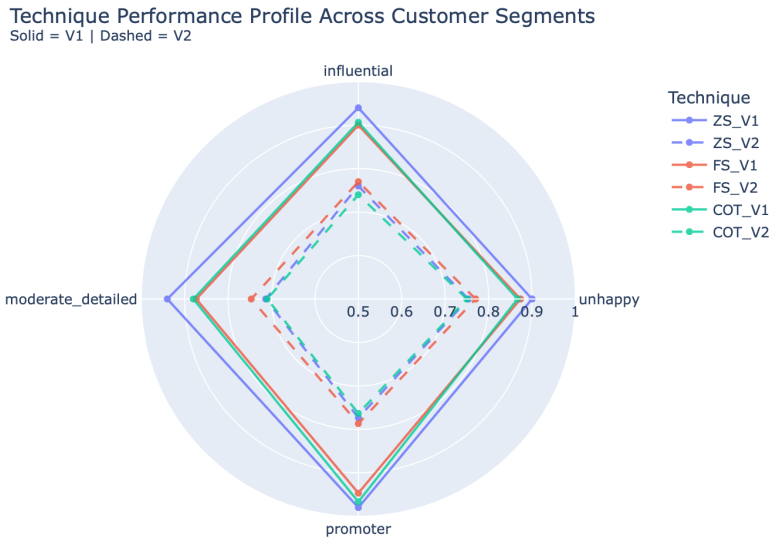
APPENDIX

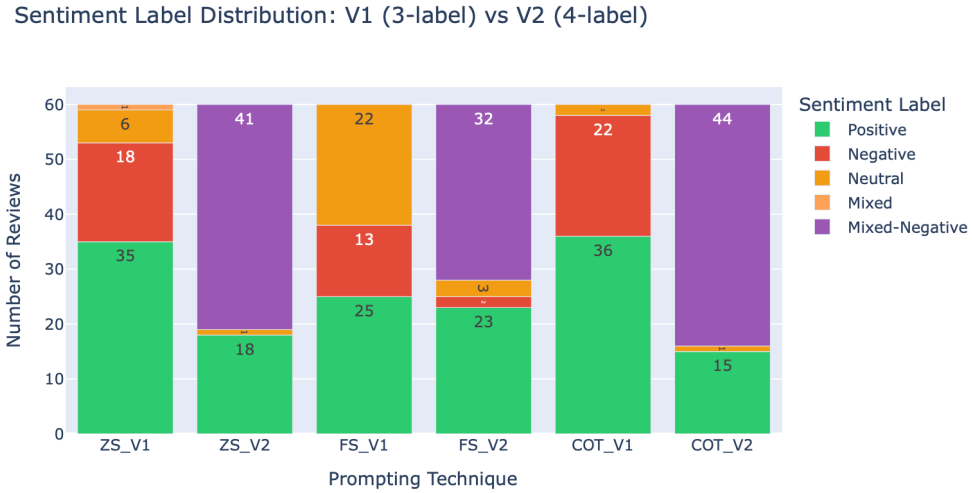

Percentage of Reviews with Contrast Language by Rating and Recommendation
Percentage with Contrast Language (%)
78.0%
77.4%
82.3%
70.6%
3-star
4-star
Rating
Recommendation
NOT Recommended
Recommended


Mid-Rated Reviews Show the Highest Recommendation Ambiguity
Recommendation Rate (%)
1.8%
6.1%
41.4%
96.7%
99.8%
41.4% of 3-star reviewers still recommend
revealing mixed sentiment hidden in mid ratings
Customer Rating


Sentiment Label Distribution: V1 (3-label) vs V2 (4-label)
Number of Reviews
ZS_V1
ZS_V2
FS_V1
FS_V2
COT_V1
COT_V2
Prompting Technique
Sentiment Label
Positive
Negative
Neutral
Mixed
Mixed-Negative


Segment × Technique Performance Heatmap
Customer Segment
promoter
moderate_detailed
influential
unhappy
0.98
0.773
0.947
0.787
0.967
0.763
0.94
0.713
0.873
0.747
0.88
0.71
0.94
0.76
0.9
0.77
0.907
0.74
0.9
0.753
0.873
0.77
0.867
0.75
ZS_V1
ZS_V2
FS_V1
FS_V2
COT_V1
COT_V2
Prompting Technique
Score
1
0.9
0.8
0.7
0.6


Technique Performance Profile Across Customer Segments
Solid = V1 | Dashed = V2
influential
moderate_detailed
unhappy
promoter
0.5
0.6
0.7
0.8
0.9
1
Technique
ZS_V1
ZS_V2
FS_V1
FS_V2
COT_V1
COT_V2


Sentiment Label Distribution: V1 (3-label) vs V2 (4-label)
Number of Reviews
Prompting Technique
ZS_V1
ZS_V2
FS_V1
FS_V2
COT_V1
COT_V2
6
18
35
41
18
22
13
25
32
3
23
22
36
44
15
Sentiment Label
Positive
Negative
Neutral
Mixed
Mixed-Negative

PROMPTS

FEW SHOT V1	FEW SHOT V2
You are a helpful AI assistant	You are an expert customer feedback analyst for ChicStyle, a fashion retail company.
few_shot_prompt_v1_eval = f"""Analyze the following customer review and respond in strict 'Key: Value' format using exactly these five fields: Category: [Select one from {', '.join(list(product_categories_dict.keys()))}] Sentiment: [Select one from {', '.join(sentiment_keys_v1)}. Definitions: {sentiment_guide_v1}] Summary: [Concise summary of the review, maximum 25 words, capturing the main point] Personalized Message: [A short empathetic response to the customer based on their review, maximum 30 words] Retail_Insight: [Identify the root cause, select the responsible department from {', '.join(list(department_categories_dict.keys()))}, and state a concrete action. Structure: root cause → department → action. Maximum 30 words. Do NOT restate the sentiment.] EXAMPLES: {format_examples_v1(few_shot_examples_v1)}"""	few_shot_prompt_v2_eval = f"""Analyze the following customer review and provide the following in a strict 'Key: Value' format: Category: [Select one from {', '.join(list(product_categories_dict.keys()))}] Sentiment: [Select one from {', '.join(sentiment_keys_v2)}. Definitions: {sentiment_guide_v2}] Summary: [Concise summary of the review, maximum 25 words, capturing the main point] Personalized Message: [A short empathetic response to the customer based on their review, maximum 30 words. Avoid generic openers like 'We're thrilled' or 'We're sorry.'] Retail_Insight: [Select urgency tier from {', '.join(list(urgency_framework.keys()))} based on severity, then identify root cause, responsible department from {', '.join(list(department_categories_dict.keys()))}, and a concrete action. Structure: urgency tag | root cause → department → action. Maximum 30 words. Do NOT restate the sentiment.] EXAMPLES: {format_examples_v2(few_shot_examples_v2)}"""
CHAIN OF THOUGHT V1	CHAIN OF THOUGHT V2
You are a helpful AI assistant	You are an expert customer feedback analyst for ChicStyle, a fashion retail company.
cot_prompt_v1_eval = f"""Analyze the following customer review step-by-step before producing your final output. Think through these steps internally: 1. What is the main subject of this review — which product or product type? 2. What is the customer's overall sentiment — are they satisfied, dissatisfied, or neutral? 3. What are the key points that should be captured in a brief summary? 4. What would be an appropriate, empathetic response to this specific customer? 5. What actionable insight can the business extract from this review? After completing your analysis, provide ONLY the following structured output: Category: [Select one from {', '.join(list(product_categories_dict.keys()))}] Sentiment: [Select one from {', '.join(sentiment_keys_v1)}] Summary: [Concise summary of the review, maximum 25 words] Personalized Message: [A short, empathetic message to the customer, maximum 30 words] Retail Insight: [A key insight for the retail business. Structure: root cause → department → action]"""	cot_prompt_v2_eval = f"""Analyze the following customer review step-by-step, then provide ONLY the final structured output. REASONING STEPS (work through internally, do not include in output): 1. SENTIMENT DECISION (apply in this exact order): a. Determine the customer's OVERALL satisfaction — are they happy or unhappy with their purchase? b. Assess whether this customer would buy from ChicStyle again based on this review. c. Apply the label that matches the customer's dominant intent: - Positive: Customer is satisfied. Minor observations (wishes, preferences) do NOT override overall satisfaction. - Mixed-Negative: Customer is conflicted — they appreciate specific aspects but have a concrete complaint that affected their experience (e.g., sizing wrong, fabric quality disappointed). The complaint must be about a real product flaw, not a preference. - Negative: Customer is dissatisfied overall. Even if they acknowledge one positive aspect, their dominant experience is negative. - Neutral: Customer states facts without emotional valence. Rare — most reviews express a clear leaning. 2. Determine the most appropriate product Category from: {', '.join(list(product_categories_dict.keys()))}. 3. Summarize the key points concisely — lead with the product attribute or issue, not "The customer." 4. Formulate an empathetic personalized message. Avoid generic openers like 'We're thrilled' or 'We're sorry.' Reference the specific item or experience. 5. Extract an actionable retail insight. Determine the urgency level: - Critical: defective or wrong item — 4hr response — escalate immediately - Actionable: recurring product quality signal — sizing, material, or photo mismatch - Monitor: isolated design or fit observation, not yet a pattern - Insight-Only: positive feedback or general comment — use for pattern analysis 6. VERIFY: Check that Category, Sentiment, and urgency are internally consistent. A 'Positive' sentiment should not produce a 'Critical' urgency insight. OUTPUT (strict Key: Value format): Category: [Select one from {', '.join(list(product_categories_dict.keys()))}] Sentiment: [Select one from {', '.join(sentiment_keys_v2)}] Summary: [≤ 25 words. Lead with the product attribute or issue.] Personalized Message: [≤ 30 words. Reference the specific item or experience.] Retail Insight: [≤ 30 words. Structure: urgency tag | root cause → department (from {', '.join(list(department_categories_dict.keys()))}) → action. Do not restate sentiment.]"""

ZERO SHOT V1	ZERO SHOT V2
You are a helpful AI assistant	You are an expert customer feedback analyst for ChicStyle, a fashion retail company.

```
zero_shot_prompt_v1 = """Analyze the following customer review and provide the
following:
Category: [Product category]
Sentiment: [Positive, Negative, or Neutral]
Summary: [Concise summary of the review]
Personalized Message: [A short, empathetic message to the customer based on
their review]
Retail Insight: [A key insight for the retail business based on the review]"""
```

JUDGE V2

```
You are a senior retail analyst evaluating whether AI-generated feedback
summaries are accurate and operationally useful. Be strict — most outputs should
score between 0.60 and 0.85.
```

```
judge_prompt_v2 = """You are a strict quality evaluator for a retail customer
feedback analysis system.
You must evaluate how well the Generated Analysis captures the meaning of the
original Customer Review.

Customer Review:
{review_text}

Customer's Actual Rating: {rating}/5
Customer Recommended Product: {recommended}

Generated Analysis:
{formatted_output}

Evaluate on these five criteria (each worth 0.20). Be strict — award full marks
only when the output is genuinely excellent, not merely present.

1. SENTIMENT ACCURACY (0.00–0.20)
   The detected sentiment must align with both the customer's rating and the
review tone.
   Rating 5 → Positive
   Rating 4 = Positive, unless the review text is predominantly negative despite
the high rating
   Rating 3 and Recommended = Neutral
   Rating 3 and Not Recommended = Mixed-Negative
   Rating 2 = Mixed-Negative if any genuine praise exists alongside the
complaint, otherwise Negative
   Rating 1 = Negative

   PENALTY: If the sentiment label is clearly wrong (e.g., "Positive" for a 2-
star review, or "Negative" for a 5-star review), award 0.00 for this criterion.
   PENALTY: If the sentiment label is close but not precise (e.g., "Neutral"
when "Mixed-Negative" was correct), award 0.05–0.10.

2. CATEGORY ACCURACY (0.00–0.20)
   - The Category must match the actual product discussed in the review.
   - It must be a specific product type (e.g., Sweaters, Dresses, Pants), not
vague labels like "Clothing", "Apparel", or "Tops" when a more specific category
exists.
   - Award 0.00 if the category is generic, missing, or clearly wrong.
   - Award 0.10 if the category is in the right area but not the most specific
option.

3. SUMMARY QUALITY (0.00–0.20)
   - Must be concise (under 25 words), accurate, and capture the main point.
   - PENALTY: Deduct 0.05 if it exceeds 25 words.
   - PENALTY: Deduct 0.10 if it omits the key complaint or praise from the
review.
   - PENALTY: Deduct 0.10 if it introduces information not in the original
review.
   - Award 0.20 only if the summary is accurate, concise, and captures the core
customer experience.

4. PERSONALIZED MESSAGE (0.00–0.20)
   - Must be tone-appropriate for the detected sentiment.
   - Must reference something specific from the review (not a generic thank-you
or apology).
   - PENALTY: Award 0.05 if the message is generic (e.g., "Thank you for your
feedback!").
   - PENALTY: Award 0.00 if the message is missing or contradicts the review
sentiment.
   - Award 0.20 only if the message feels genuinely personalized to this
specific review.

5. RETAIL INSIGHT (0.00–0.20)
   - Must identify a specific, actionable finding — not a restatement of
sentiment.
   - Must mention which department or function should act.
   - Must suggest a concrete action (not "improve quality" or "enhance
experience").
   - PENALTY: Award 0.05 if the insight is vague or generic.
   - PENALTY: Award 0.00 if no insight is provided or it merely restates the
sentiment.
   - Award 0.20 only if the insight names a root cause, a responsible team, and
a specific action.

SCORING RULES:
- Sum the five scores. The maximum possible score is 1.00.
- Your score MUST be between 0.00 and 1.00. Never exceed 1.00.
- A score of 0.80+ should be reserved for outputs where ALL five fields are
accurate and well-crafted.
- A score of 1.00 means every field is excellent — this should be rare.

Output ONLY the total score as a decimal (e.g., 0.75). No other text."""
```

```
zero_shot_prompt_v2 = f"""
You are an AI system designed to analyze retail customer reviews for a
fashion e-commerce company.
Your goal is to accurately interpret customer feedback and generate
structured, business-relevant insights.
--------------------------------------
IMPORTANT RULES
- Do NOT rely only on tone — consider the full meaning of the review.
- Reviews may contain mixed signals:
  - Positive beginning but negative ending
  - High rating but negative experience
- If the overall outcome is dissatisfaction, classify as Negative or
Mixed-Negative.
- "Mixed-Negative" = polite wording but underlying disappointment.
--------------------------------------
TASK
--------------------------------------
provide the following in strict 'Key: Value' format:
Generate a structured analysis with the following fields:

1. Sentiment
- One of: Positive, Neutral, Negative, Mixed- Negative
- Reflect the TRUE customer experience, not just wording

2. Category
- Main issue or topic (e.g., sizing, fit, quality, delivery, color,
comfort, design, service)

3. Summary
- 1–2 concise sentences capturing the key points
- Must be accurate and not introduce new information

4. Urgency
- Insight-Only → praise or minor feedback
- Actionable → moderate dissatisfaction or friction
- Critical → strong dissatisfaction, defects, or risk of churn

5. Personalized_Message
- Max 30 words
- Natural, empathetic, and aligned with sentiment
- Avoid generic responses

6. Retail_Insight
- Specific and actionable recommendation for a retail team
- Focus on root cause, operations, product improvement, or customer
recovery

OUTPUT FORMAT
Category: [Select one from {',
'.join(list(product_categories_dict.keys()))}]
Sentiment: [Select one from {', '.join(sentiment_keys_v2)}. Definitions:
{sentiment_guide_v2}]
Summary: [Concise summary of the review, maximum 25 words, capturing the
main point]
Personalized Message: [A short empathetic response to the customer based
on their review, maximum 30 words]
Retail Insight: [One specific, actionable finding. Structure: root cause →
responsible department (from {',
'.join(list(department_categories_dict.keys()))}) → concrete action.
Maximum 25 words. Do NOT restate the sentiment.]"""
```

JUDGE V1

```
You are a helpful AI assistant
```

```
judge_prompt_v1 = """Your task is to act as an impartial judge to evaluate the quality of a
generated review summary and other extracted retail insights.
You will be given:
Customer Review: {review_text}
Generated Analysis: {formatted_output}
Your goal is to rate the output on a scale of 0 to 1, based on:
Evaluation Criteria:
- Accuracy: The summary accurately reflects the sentiment and key points of the original
review.
- Summary faithfully reflects the customer intention.
- A Rating of 4-5 should align with Positive, 1-2 with Negative, 3 with Neutral.
- Completeness: The output includes a Category, Sentiment, Summary, Personalized Message, and
Retail Insight.
- Conciseness: The summary is brief and to the point without unnecessary jargon.
- Actionability: The Retail Insight provides a clear, actionable recommendation for a retail
business. not a generic statement like "improve product quality"
- Personalization: Is the Personalized Message appropriate for the detected sentiment and
appropriate for a customer based on their review.
Scoring:
0.0 - 0.2: Very poor quality. Missing multiple key elements, inaccurate, or completely
irrelevant.
0.3 - 0.5: Below average. Some elements are present but significant flaws in accuracy,
completeness, or actionability.
0.6 - 0.7: Average. Most elements are present and accurate, but there's room for improvement
in conciseness, actionability, or personalization.
0.8 - 0.9: Good quality. All elements are present, accurate, and well-formulated.
1.0: Excellent:  accurate, faithful to the review, actionable, and well-personalized
Based on these criteria, provide a score from 0.0 to 1.0 (decimals are allowed). Only output
the score, no other text."""
```