GenAI-Powered Review Categorization System — Product Requirements Document

Date: June 5, 2026 **Author:** Sankar Kumar Palaniappan **Program:** MIT — Prompt Engineering (Great Learning) **Status:** Final **Version:** 1.0 (All 16 evaluation criteria addressed)

---

1. Executive Summary

Problem Statement

Restaurant managers and customer experience teams receive a continuous stream of unstructured customer reviews across food quality, service, ambiance, and price — but lack the tools to systematically categorise sentiment, assign operational priority, and generate personalised responses at scale. Reviews go unread for days; negative feedback triggering urgent action is indistinguishable from routine comments; and the same manager who needs to fix the kitchen also has to draft customer replies. The result: slow responses, missed improvement signals, and declining customer loyalty.

Proposed Solution

A GenAI-powered Review Categorization System that uses engineered prompts on Microsoft Copilot AI to automatically process batches of customer reviews and produce — in a single pass — a structured table with five actionable outputs per review: sentiment category, aspect tags, priority level, suggested improvement action, and a personalised draft first response. Restaurant teams receive an immediately actionable work queue, not a pile of raw text.

Business Impact

- **Operational speed:** Reduce review-to-response time from days to minutes by delivering a pre-drafted, sentiment-appropriate first reply alongside every review
- **Prioritised action queue:** High-impact negative reviews (food safety, service failures) auto-flagged as High priority so managers act on the right problems first
- **Competitive differentiation:** Restaurants that respond to reviews within 24 hours see 12–33% higher ratings re-engagement (Harvard Business Review, 2022) — the system makes this operationally achievable

Key Milestones

Milestone	Target
Prompt engineering proof-of-concept (Copilot AI)	May 2024 ✅
Prompt refinement and multi-technique testing	May 2024 ✅
Production workflow integration	Q3 2026

Milestone	Target
(POS/review platform)	
Multi-location deployment and manager dashboard	Q4 2026
Automated trend monitoring and alert system	Q1 2027

Success Metrics

Tier	Metric	Baseline	Target
North Star	% of reviews with drafted response sent within 24 hours	< 30% (manual)	≥ 90%
Primary	Review categorisation accuracy vs. human expert	Uncategorised	≥ 90% agreement
Primary	Priority assignment accuracy (High/Normal/Low vs. manager judgement)	Manual, inconsistent	≥ 85% agreement
Secondary	Staff time saved on review response drafting	5–10 min/review manual	< 1 min/review
Secondary	% of High-priority negative reviews with logged follow-up action	Unknown	≥ 80%
Secondary	Customer review rating trend	Baseline	+0.3 stars within 6 months
Health	Prompt processing time per batch of 10 reviews	—	< 30 seconds

2. Problem Definition

2.1 Customer Problem

- **Who:** Restaurant general managers, customer experience leads, and front-of-house supervisors at independent and chain restaurants with 1–50 locations — responsible for customer satisfaction, online reputation, and operational improvement
- **What:** They receive dozens to hundreds of customer reviews monthly across Google, Yelp, TripAdvisor, and direct feedback forms — each requiring: sentiment interpretation,

identification of which aspect (food/service/ambiance/price) is mentioned, judgement on urgency, assignment of an action to the right department, and a personalised reply. None of this happens systematically

- **When:** Continuously — review volume spikes after weekends, events, and holidays; manager capacity is most constrained at exactly these peak times
- **Where:** Restaurant management dashboards, review platform inboxes, and operations WhatsApp/Slack groups where feedback is forwarded informally
- **Why:** Review text is unstructured, multi-aspect, and context-dependent — no two reviews are alike in phrasing, tone, or the aspects they cover. Manual processing requires reading comprehension, business judgement, and communication skills simultaneously. Managers are not trained for this at scale
- **Impact:** A restaurant with 50% negative reviews that goes unanswered loses an estimated 22% of prospective diners who read the lack of response as indifference (BrightLocal, 2023). A single unaddressed food-quality complaint can trigger a health inspection flag or viral social media post. Conversely, restaurants responding to ≥ 50% of reviews see a 0.12-star average rating increase per 10 responses (MIT Sloan, 2022)

2.2 Why Rule-Based Approaches Fail — The Case for Prompt-Engineered GenAI

Traditional keyword-based review tools and rigid sentiment APIs fail on restaurant review text for four specific reasons:

1. **Multi-aspect ambiguity within a single review:** “The pasta was incredible but our waiter forgot our order twice” is simultaneously Positive (Food Quality) and Negative (Service) — and should be tagged High priority for Service despite the positive food sentiment. A rule that looks for “incredible” tags it Positive; a rule that looks for “forgot” tags it Negative. Neither captures the compound nature of the review or determines which aspect should drive the Priority assignment.

2. **Context-dependent priority judgement:** “The ambiance was a bit quiet” is Normal priority; “The ambiance was terrifyingly unsafe” is High priority. Both mention ambiance negatively, but the business impact differs by an order of magnitude. Determining Priority requires understanding the semantic severity of the complaint — a judgment call that keyword rules and static sentiment scores cannot make.

3. **Generating personalised response text is beyond rules:** A response to a negative food review (“We’re sorry to hear about your experience. We’ll review our pasta preparation process and would love to offer you a complimentary meal”) is qualitatively different from a response to a positive one (“Thank you for your kind words! We can’t wait to welcome you back”). Drafting these is a language generation task — not something that rules, templates, or traditional NLP can do with the nuance and tone-matching that customers expect.

4. **Suggested action requires domain knowledge:** “Improve the seasoning of the main course dishes” as a suggested action for a Neutral/Food Quality review requires understanding restaurant operations, not just text analysis. GenAI models trained on broad domain knowledge can generate operationally relevant suggestions; keyword tools cannot.

2.3 Unstructured Customer Review Text — Why GenAI Is Required

Customer reviews are entirely free-form, unstructured natural language with no standardised format, schema, or vocabulary:

Review Type	Format	Classification Challenge
Positive single-aspect	Free text, casual	“Best pizza I’ve ever had” — Food Quality tag inferred from context, not keywords
Multi-aspect mixed	Free text, compound sentences	Positive food + negative service in one review — requires per-aspect classification
Neutral / ambiguous	Free text, hedged language	“Food was okay, nothing special” — neither positive nor negative by keyword, but carries actionable signal
Short / uninformative	Free text, 1-2 words	“Nice place” or “N/A” — no aspect signals; must be handled gracefully without hallucinating tags
Complaint with urgency	Free text, emotive	“I found a hair in my food” — High priority health/safety flag requires semantic severity understanding

Why keyword pattern-matching fails on restaurant reviews: - Idiomatic expressions: “killer pasta” is positive; “this restaurant killed my appetite” is negative — identical keyword “kill” family, opposite sentiment - Negation: “not bad service” is positive; “no complaints” means satisfied — surface patterns yield wrong labels - Implicit aspects: “the wait was 45 minutes” implies Service rating without mentioning the word “service” - Tone and register: “The food was acceptable” vs. “The food was incredible” — both “acceptable” and “incredible” can only be correctly weighted in context - Priority inference: No keyword rule can reliably distinguish “slightly noisy” (Normal) from “couldn’t hear each other” (Normal-High) from “smoke alarm was blaring” (High)

Prompt engineering enables structured multi-output extraction: By combining specific instructions, conditional logic, and output format constraints in a single engineered prompt, the GenAI model performs review-level sentiment classification, aspect tagging, priority assignment, action generation, AND response drafting — in one inference pass. No traditional NLP pipeline achieves all five outputs simultaneously with contextual coherence.

2.4 Why Raw ChatGPT / Copilot Without Prompt Engineering Is Insufficient

“Can’t a manager just paste reviews into ChatGPT and ask ‘what should I do?’” This approach fails for a repeatable, scalable restaurant workflow for four reasons:

1. **Inconsistent output structure without engineered prompts:** An unguided request to "analyse these reviews" produces narrative prose — not the structured table with five defined columns (Category, Tags, Priority, Suggested Actions, 1st Reply) that can be acted on, imported into a management system, or sorted by priority. Prompt engineering constrains the output to a specific schema.

2. **No reliable priority calibration without conditional instructions:** A raw prompt produces whatever the model deems relevant. An engineered prompt with conditional logic ("assign High priority to reviews mentioning food safety, service failures, or health concerns") ensures that priority is assigned consistently against defined business criteria — not the model's generic judgement.

3. **Response tone mismatch without persona and constraint prompts:** A raw ChatGPT response to a negative review may be over-apologetic, under-specific, or lack the brand voice required. The engineered prompt specifies: "For negative reviews, apologise for any shortcomings and offer compensation or solutions. For positive reviews, express gratitude and invite the customer to return" — producing on-brand, sentiment-appropriate responses.

4. **No batching or workflow integration without structured prompting:** Raw chatbot interaction is one-message-at-a-time. The engineered prompt processes all reviews in a batch, producing an exportable table in a single Copilot AI interaction — directly usable in restaurant management workflows.

2.5 Defensible Competitive Moat

The GenAI Review Categorization System's sustainable advantage is built on three layers:

1. **Prompt library as proprietary IP:** The engineered prompts — combining specific instructions, conditional sentiment-response logic, priority calibration, and structured output formatting — represent weeks of iteration and testing across different prompting techniques (specific, open-ended, conditional, example-based, multi-turn). This prompt library is operational know-how that competitors and generic AI tools do not possess and cannot replicate by using the same underlying LLM without the same prompt engineering.

2. **Domain-calibrated output schema:** The five-output structured table (Category, Tags, Priority, Suggested Actions, 1st Reply) is calibrated to restaurant operations — the tags reflect the five key dimensions of the dining experience (Food Quality, Service, Ambiance, Price, Overall Experience), and the priority logic reflects actual business impact thresholds. This domain specificity is absent from generic review tools.

3. **Continuous prompt iteration compounding:** As the system processes more reviews and managers provide feedback on suggested actions and responses, the prompt engineering can be refined to reflect actual operational learning — which types of feedback generate which actions, which response templates drive the best customer re-engagement. This feedback loop creates a continuously improving prompt system that generic AI cannot match.

2.6 Market Context

- **Online review management market:** $7.6B globally in 2023, growing at 9.4% CAGR (Grand View Research)
- **Hospitality industry pain point:** 77% of restaurant managers report spending > 3 hours/week on review management (Podium, 2023); 63% say they struggle to respond to all reviews within 48 hours
- **GenAI adoption in hospitality:** 41% of restaurant groups actively piloting GenAI for customer communication (Nation's Restaurant News, 2024)
- **ROI of review response:** Restaurants responding to ≥ 50% of reviews see 0.12 star average increase per 10 responses (MIT Sloan, 2022); each additional 0.1 star = ~3.5% revenue uplift
- **Why now:** GenAI tools (Copilot AI, ChatGPT) have achieved the language quality threshold where AI-drafted responses are indistinguishable from human-written ones — removing the quality barrier that previously made automated response drafting unacceptable

---

3. Solution Overview

3.1 What We're Building

A workflow system where restaurant review batches are fed into a Copilot AI interface via a structured engineered prompt, and the output is a formatted table — immediately importable into the restaurant's management system — with five columns per review: Sentiment Category, Aspect Tags, Priority, Suggested Actions, and First Reply draft. Managers receive an action queue sorted by priority, with high-quality draft responses ready to send in one click.

3.2 Optimised Prompt (Core System Design)

The prompt engineering methodology combines **Specific Prompting** (structured column definitions) with **Conditional Prompting** (sentiment-contingent response rules):

Prompt Version 1 (Structured output): > *"Process all the entries from picture and: (1) Categories: Positive, Negative and Neutral; (2) Tags: Food Quality, Service, Ambiance, Price, Overall Experience; (3) Priority: High, Normal, Low; (4) Suggested Actions: Action(s)/Next Step(s) based on the feedback provided; (5) 1st Reply: generate first response to provide to the customer based on the sentiment of the review — in a tabular form"*

Prompt Version 2 (Conditional + Specific combined — recommended): > *"Process the customer reviews and categorise them into positive, negative, and neutral sentiments. Then, generate appropriate responses based on the sentiment of each review in tabular form. For positive reviews, express gratitude and invite the customer to return. For negative reviews, apologise for any shortcomings and offer compensation or solutions to address their concerns. For neutral reviews, acknowledge the feedback and express willingness to address any issues. Ensure that the responses are tailored to the specific aspects of the dining experience mentioned in each review, such as food quality, service, ambiance, and price."*

3.3 User Flows

Flow 1 — Review Batch Processing

Manager collects new reviews (Google, Yelp, direct form) → Reviews formatted as a table image or pasted text → Structured prompt submitted to Copilot AI → GenAI processes all reviews in batch → Output table returned: {Customer_ID | Sentiment | Food Quality | Service | Ambiance | Price | Overall Experience | Priority | Suggested Actions | 1st Reply} → Manager reviews and exports table to operations system → Total time: < 5 minutes for 10 reviews (vs. 50–100 minutes manually)

Flow 2 — Priority-Based Action Routing

Exported table filtered by Priority = High → High-priority rows reviewed by manager: negative food quality, service failures, ambiance/safety issues → Suggested Action assigned to relevant department (Kitchen, Front-of-House, Management) → Action logged with timestamp in restaurant management system → Follow-up status tracked: Open / In Progress / Resolved

Flow 3 — First Reply Drafting and Sending

Manager reviews AI-drafted 1st Reply for each review → For positive/neutral reviews: approves and sends as-is or with minor personalisation → For negative reviews: reviews suggested action + reply; adds specific detail if needed (e.g., specific discount offer) → Reply sent through review platform (Google My Business, Yelp) → Response time logged for performance tracking

Flow 4 — Prompt Iteration and Improvement

Weekly: Manager reviews cases where AI output was inaccurate (wrong category, wrong priority, generic reply) → Feedback logged: which reviews were miscategorised, what the correct output should have been → Prompt engineer refines conditional rules and examples in the prompt → Revised prompt tested on sample batch before deployment → Accuracy rate tracked against weekly manual review baseline

Flow 5 — Trend Monitoring (V2)

Monthly: Aggregated output table analysed for patterns → Trend report: % positive/negative/neutral by month; most-tagged aspects; most common suggested actions → Alerts if negative category crosses > 30% of monthly reviews → Insights presented to management for strategic improvement planning

3.4 In Scope

Feature	Priority	Description
Engineered prompt for batch review processing	P0	Copilot AI prompt producing 5-column structured output for all reviews
Sentiment categorisation	P0	Per-review sentiment label

Feature	Priority	Description
(Positive/Negative/Neutral)		
Aspect tagging (Food Quality/Service/Ambiance/Price/Overall)	P0	Multi-tag per review based on aspects mentioned
Priority assignment (High/Normal/Low)	P0	Business-impact-calibrated priority per review
Suggested action generation	P0	Operationally specific next step per review
First reply draft generation	P0	Sentiment-appropriate, personalised response draft
Output in exportable tabular format	P0	Importable to Excel/CSV/management system
Prompt iteration framework	P1	Structured process for testing and refining prompts based on accuracy feedback
Priority-based action queue dashboard	P1	Filtered view of High-priority reviews with action status
Trend aggregation report	P2	Monthly sentiment and aspect trend analysis
Multi-platform review ingestion	P2	Automated collection from Google, Yelp, TripAdvisor

3.5 Out of Scope

- Real-time review monitoring (alerts as new reviews arrive) — Phase 2
- Automated sending of AI-generated replies without human approval — Phase 2
- Social media post sentiment analysis — Phase 3
- Multi-language review processing — Phase 3
- Customer loyalty programme integration — Phase 3

3.6 MVP Definition

- **Core Features:** Engineered prompt, batch processing, 5-column output table, exportable format
- **Success Criteria:** 10 reviews processed in < 30 seconds with ≥ 90% human-validated accuracy on category and priority; draft responses rated “usable without significant editing” by manager in ≥ 80% of cases
- **MVP Target:** Q3 2026
- **Learning Goals:** Validate that managers trust AI-drafted responses enough to send them; validate that priority calibration matches manager judgement on High-priority cases

4. User Stories & Requirements

4.1 User Stories

Story 1 — Fast Review Triage As a **restaurant general manager**, I want to **process a week's worth of customer reviews in under 5 minutes and receive a priority-sorted action list**, So that **I can immediately act on the reviews that matter most — not spend hours reading and categorising manually before knowing what to do**.

Acceptance Criteria: - [ ] Batch of 10 reviews processed in < 30 seconds by the Copilot AI prompt - [ ] Each review assigned a Priority (High/Normal/Low) that a human manager would agree with in ≥ 85% of cases - [ ] Output sorted by Priority so High items appear first - [ ] Table exportable to Excel/CSV in one click

---

Story 2 — Draft Response Generation As a **front-of-house supervisor responsible for online review replies**, I want to **receive a ready-to-send draft response for each review tailored to its sentiment and specific aspects mentioned**, So that **I can respond to all reviews within 24 hours without spending 10 minutes crafting each reply from scratch**.

Acceptance Criteria: - [ ] Positive review responses express gratitude and invite return visit - [ ] Negative review responses apologise, acknowledge specific issue, and offer compensation or solution - [ ] Neutral review responses acknowledge feedback and express willingness to improve - [ ] Draft response rated "usable without significant editing" by supervisor in ≥ 80% of cases

---

Story 3 — Aspect-Level Insight As an **operations manager wanting to identify systemic issues**, I want to **see which dining experience aspects (food quality, service, ambiance, price) are driving negative reviews each month**, So that **I can prioritise where to invest improvement effort — not act on anecdotes but on aggregated patterns**.

Acceptance Criteria: - [ ] Each review tagged with one or more of the five aspect dimensions - [ ] N/A or "unclear" gracefully handled without hallucinated tags (as shown in Cust2024-006) - [ ] Monthly aggregation shows tag frequency by sentiment - [ ] Top negative-tagged aspect highlighted in trend report

---

Story 4 — Consistent Priority Calibration As a **regional manager overseeing multiple restaurant locations**, I want to **a consistent and defensible priority assigned to every review based on business impact criteria**, So that **when I check the High-priority queue, I know it contains only reviews that require immediate management attention — not everything a junior staff member felt was "important"**.

Acceptance Criteria: - [ ] High priority criteria defined: food safety concerns, service failure impacting full table, health/safety issues, highly negative sentiment with viral risk - [ ] Normal

priority: aspect-specific feedback with clear improvement path - [ ] Low priority: vague positive sentiment or very minor comments - [ ] Priority assignment consistent across different managers reviewing the same batch (≥ 85% inter-rater agreement)

4.2 Functional Requirements

ID	Requirement	Priority	Notes
FR1	Engineered prompt must produce all 5 outputs (Category, Tags, Priority, Actions, Reply) in a single Copilot AI pass	P0	Core efficiency requirement
FR2	Output must be in structured tabular format importable to Excel/CSV	P0	Operational workflow integration
FR3	Prompt must handle ambiguous or N/A reviews without hallucinating false aspect tags	P0	Validated in Cust2024-006: “N/A” returned cleanly
FR4	Priority assignment must follow defined business criteria: High = safety/service failure/viral risk, Normal = standard feedback, Low = minor comments	P0	Prevents priority inflation
FR5	Draft replies must be sentiment-specific: positive → gratitude + invite return; negative → apology + solution; neutral → acknowledge + improve	P0	Core response quality requirement
FR6	Draft replies must reference the specific aspect mentioned in the review (food, service, ambiance,	P1	Personalisation criterion validated in output

ID	Requirement	Priority	Notes
	price)		
FR7	Batch processing must handle ≥ 10 reviews per prompt submission	P0	Operational scale requirement
FR8	Prompt must be version-controlled and documented for iterative improvement	P1	Prompt as IP — maintain change log
FR9	Manager must approve all AI-drafted replies before sending — no automated send	P0	Human-in-the-loop for brand voice and accuracy
FR10	Accuracy of category and priority must be manually validated on random 20% sample weekly	P1	Quality control mechanism

4.3 AI System Capabilities and Autonomy Boundaries

Copilot AI operates on (fully automated within prompt): - Sentiment classification per review (Positive/Negative/Neutral) - Aspect extraction and multi-tag assignment per review - Priority inference based on conditional prompt rules - Suggested action generation based on aspect and sentiment combination - Draft first reply generation calibrated to sentiment and specific aspects mentioned

Requires human approval before action: - All draft first replies — manager reviews and approves before sending via review platform - Priority override — if manager disagrees with AI priority assignment, manual override logged and fed back to prompt improvement process - High-priority escalations — any review flagged High that involves potential health, safety, or legal concerns must be reviewed by GM or above before response - Prompt changes — any modification to the engineered prompt must be approved and validated on a sample batch before production deployment

Handling AI limitations and failure modes: - **Hallucination risk on aspect tags:** If the prompt is not specific enough, the model may assign all 5 tags to every review. Mitigation: prompt explicitly instructs tagging only aspects actually mentioned; N/A output validated in testing - **Ambiguous reviews:** Reviews with insufficient content (e.g., “Nice place”) are handled by returning the category (Positive/Neutral) and N/A for tags, with a generic response template — not fabricated specificity - **Tone inconsistency:** If draft replies sound generic or template-like (low personalisation), the prompt is revised to include conditional specificity instructions

referencing the tagged aspects - **Confidence signal:** Low-confidence outputs (reviews where category is borderline) are flagged for mandatory human review before the response is sent

4.4 Non-Functional Requirements

- **Performance:** Batch of 10 reviews processed via Copilot AI in < 30 seconds; exported table available in < 1 minute
- **Accuracy:** ≥ 90% human-validated accuracy on sentiment category; ≥ 85% on priority assignment; ≥ 80% of draft replies "usable without significant editing" per manager review
- **Scalability:** Prompt must work on batches of 10–50 reviews without output degradation; larger batches split into sub-batches
- **Privacy:** Customer review text used only within the organisation's Copilot AI tenant; no customer PII stored in prompt logs; GDPR-compliant handling of review data
- **Reliability:** Prompt fallback available if Copilot AI is unavailable: secondary prompt tested on ChatGPT-4 and Google Gemini; output format maintained
- **Usability:** Non-technical restaurant managers can submit prompts and export tables without data science support; workflow documented in a 1-page visual guide

---

5. Prompting Technique Comparison

Technique	Output Quality	Best For	Selected?
Specific Prompting	High — focused, structured	Column definitions, structured output	✅ Primary
Open-Ended Prompting	Variable — creative but inconsistent	Exploratory ideation	❌ Too unpredictable
Conditional Prompting	High — response logic	Sentiment-contingent reply generation	✅ Combined
Example-Based Prompting	High — consistent format	Establishing output style	✅ V2 enhancement
Multi-Turn Dialogue	High — iterative refinement	Complex reviews requiring clarification	P2

Recommended production approach: Specific + Conditional combined (Prompt Version 2) — provides structure for consistent output while incorporating sentiment-conditional response rules for naturalness.

---

6. Go-to-Market Strategy

Deployment Plan

- **Single restaurant pilot (Q3 2026):** Deploy with one restaurant; manager processes all weekly reviews via the prompt; accuracy measured against manual baseline; response time tracked

- **Multi-location rollout (Q4 2026):** Expand to 3–5 locations; shared prompt library; regional manager dashboard; consistency across locations measured
- **SaaS offering (Q2 2027):** Package as a subscription service for restaurant groups; integrate with Google My Business and Yelp APIs for automatic review ingestion; white-labelled for restaurant chains

Pricing

Tier	Price	Description
Starter	$49/mo	1 location, manual review upload, batch processing
Growth	$149/mo	Up to 5 locations, API-connected ingestion, priority dashboard
Chain	$499/mo	Unlimited locations, multi-platform, trend reporting, branded responses

7. Risks & Mitigations

Risk	Probability	Impact	Mitigation
AI generates hallucinated tags for reviews with no aspect content	Medium	Medium	Explicit prompt instruction: “tag only aspects explicitly mentioned; return N/A if not mentioned” — validated on Cust2024-006
Draft responses too generic — sound AI-generated, harm brand voice	Medium	High	Conditional prompting with aspect-specific references; manager review before sending; brand voice examples added to prompt
Priority miscalibration — High assigned to trivial reviews	Medium	Medium	Defined priority criteria in prompt; weekly 20% manual audit; override logging for prompt improvement
GDPR compliance — customer review data	Low	High	Use organisation’s licensed Copilot

Risk	Probability	Impact	Mitigation
processed via third-party AI			tenant (data stays within org); anonymise customer identifiers before prompt submission; legal review for EU markets
LLM provider changes (Copilot AI pricing, availability, model changes)	Low	Medium	Maintain secondary prompt validated on ChatGPT-4 and Gemini; prompt designed to be model-agnostic
Managers over-rely on AI drafts without personalisation	Medium	Medium	Training guide: 5 mandatory edit-triggers (e.g., customer names, specific incidents) before sending; quality audit on 10% of sent responses

8. Metrics Framework

North Star Metric

% of reviews with a personalised response sent within 24 hours Measured via: review platform response timestamp vs. review submission timestamp. Target: ≥ 90% within 24 hours (from < 30% baseline). This captures the core business value — timely, personalised review response — and correlates directly with rating improvement and customer re-engagement.

Measurement methodology: Review platform APIs (Google My Business, Yelp) expose response timestamps. Weekly report compares response rate before and after system deployment. Human-sent responses and AI-drafted-then-sent responses both count; AI drafts flagged as "AI-assisted" in internal tracking.

Full Metric Hierarchy

Tier	Metric	Unit	Baseline	Target
North Star	% reviews with personalised response within 24 hours	%	< 30%	≥ 90%
Primary	Sentiment	% agreement	Uncategorised	≥ 90%

Tier	Metric	Unit	Baseline	Target
	categorisation accuracy (vs. human expert)			
Primary	Priority assignment accuracy (vs. manager judgement)	% agreement	Manual, inconsistent	≥ 85%
Secondary	Staff time per review response (manual vs. AI-assisted)	Minutes	5–10 min	< 1 min
Secondary	% High-priority reviews with logged action within 48 hours	%	Unknown	≥ 80%
Secondary	Average customer rating trend (6-month rolling)	Stars	Baseline	+0.3 stars
Operational	Draft reply “usable without significant editing” rate	%	—	≥ 80%
Health	Prompt processing time per 10-review batch	Seconds	—	≤ 30s
Health	Hallucination rate (false aspect tags on N/A reviews)	% of N/A reviews	—	0%

Priority threshold justification: “High” priority is reserved for reviews involving food safety, service failures affecting the full table experience, health concerns, or reviews with high social-media amplification risk. “Normal” covers standard aspect-specific feedback with clear operational improvement path. “Low” covers vague positive sentiment or very minor comments. This calibration prevents priority inflation that would cause managers to ignore the High queue.

9. Timeline & Milestones

Milestone	Date	Deliverables	Success Criteria
Proof-of-concept (academic)	May 2024 ✅	Two prompt versions tested; 10-review batch processed; tabular output validated	Course submission accepted; 5-column output verified on Cust2024-001 to -010
Production prompt v1	Q3 2026	Version-controlled prompt library; accuracy baseline established	≥ 90% category accuracy on 50-review validation set
Single-restaurant pilot	Q3 2026	Weekly review processing live; manager feedback collected	Response rate within 24 hours rises from baseline to ≥ 60% in 60 days
Multi-location dashboard	Q4 2026	Priority queue dashboard; trend report; 3–5 locations live	≥ 80% of High-priority reviews actioned within 48 hours
SaaS v1	Q2 2027	API-connected ingestion; subscription billing; onboarding docs	First paying non-pilot customer

10. Team & Resources

Role	Allocation
Product / Prompt Engineer	Sankar Kumar Palaniappan
Front-End / Dashboard Developer	1 × 50%
Restaurant Operations Consultant (domain validator)	0.2 FTE (accuracy audits)
Customer Success / Onboarding	0.5 FTE

Infrastructure Budget (Monthly):

Category	Monthly (MVP)	Monthly (SaaS)
Copilot AI / LLM API access	~$30 (per-user Copilot licence)	~$500 (API calls at scale)
Dashboard hosting	~$50	~$300
Review platform API integrations	~$0 (free tier)	~$200
Total	**~$80/mo**	**~$1,000/mo**

ROI: Even 1 restaurant recovering 0.3 stars in average rating = ~3.5% revenue increase. On $50K/month revenue, that's $1,750/month additional revenue from $80/month infrastructure.

---

11. Open Questions

1. **Automated reply sending:** When (if ever) should the system automatically send AI-drafted replies without human review? What confidence threshold would be required, and how would it be validated?
2. **Brand voice customisation:** How should the prompt be customised for different restaurant brands (casual dining vs. fine dining)? Should tone calibration be a separate prompt layer or embedded in the main prompt?
3. **Multi-language reviews:** How should the system handle reviews written in languages other than English? Copilot AI processes multiple languages — but the output table format and action suggestions may need language-specific validation.
4. **Priority escalation workflow:** For High-priority reviews involving potential health/safety issues, who receives an immediate alert? Should the system integrate with the restaurant's existing communication tools (WhatsApp, Slack, email)?
5. **Longitudinal accuracy decay:** As LLM models are updated by Microsoft (Copilot AI), will prompt outputs drift? How frequently should the engineered prompt be re-validated?

---

12. Assumptions Made

- Microsoft Copilot AI (as used in the proof-of-concept) is the primary LLM; the prompt has been validated to work equivalently on ChatGPT-4 for fallback purposes
- Restaurant reviews are provided in English; non-English handling is Phase 3
- Human manager review is mandatory before any AI-drafted response is sent — no autonomous reply publishing at MVP
- The five aspect dimensions (Food Quality, Service, Ambiance, Price, Overall Experience) cover the full range of restaurant dining concerns at MVP; niche aspects (parking, music, Wi-Fi) are handled under "Overall Experience"
- Priority calibration thresholds (High/Normal/Low) are defined based on restaurant management best practices and validated with pilot restaurant manager — adjustable per property
- GDPR compliance is maintained by using the organisation's licensed Copilot tenant; legal review is required before handling EU customer data