title: "AI Integration & Prompt Engineering Strategy" subtitle: "Phase 3: Syracuse Housing Safety Tracker" author: "Karan C. Salunkhe" date: "January 2026" output: github_document

1. Overview

The Phase 3 AI integration prioritizes statistical accuracy, civic responsibility, and explainability. By combining deterministic metric computation with constrained language generation and automated validation, the system delivers accessible narratives without compromising analytical integrity.

2. Prompt Engineering Strategy

2.1 System-Constrained Grounding

A **System Instruction–first architecture** is used to tightly constrain model behavior. The system role establishes both domain expertise and epistemic limits.

System Instruction (Authoritative Persona)

> You are a Housing Policy Auditor for the City of Syracuse. You are provided with computed, verified metrics related to housing code violations, Unfit properties, and remediation timelines.
>
> **Rules:**
>
> 1. You must only reference statistics explicitly provided in the prompt context.
> 2. You must not estimate, extrapolate, or invent values.
> 3. If the user asks a question that cannot be answered using the provided data, you must explicitly state that the information is unavailable.
> 4. Any explanation of causes must be grounded in observed fields (e.g., `corrective_action`, `permit_status`).
> 5. When data quality is limited, you must disclose uncertainty.

This instruction is injected before all user content and overrides any conversational tendencies of the model.

2.2 Prompt Context Structure

All prompts follow a fixed schema to minimize variance:

```
{
  "city_kpis": {
    "total_open_violations": 2859,
    "remediation_gap_pct": 19.5,
    "median_unfit_days": 895
  },
  "neighborhood_metrics": {
    "name": "Northside",
    "open_violations": 2859,
    "matched_permits_pct": 41.2
  },
  "data_quality_flags": {
    "small_sample_warning": false,
    "date_inconsistencies_present": true
  }
}
```

3. Prompt Iteration Log

Iteration	Prompt Focus	Observed Issue	Mitigation Implemented
V1	Generic summary	Hallucinated 2024 statistics	Introduced strict system instruction
V2	Neighborhood summary	Ignored remediation duration	Injected "Critical Metrics" block
V3	Causal explanation	Political speculation	Added hypothesis restriction rules
V4	Final	Stable, accurate, cautious	Locked schema + unit tests

4. Data-to-Prompt Validation Pipeline

4.1 Metric Computation Layer

All metrics are computed prior to LLM invocation using Python and Pandas. No raw records are passed directly to the LLM to prevent privacy leaks and processing errors.

```
# Example Computation Logic
median_days = (unfit_df["resolution_date"] - unfit_df["designation_date"]).dt.day
backlog_pct = (open_violations / total_violations) * 100
```

4.2 Prompt Injection Mechanism

Computed KPIs are serialized to JSON and injected verbatim into the prompt context to ensure:

- Numerical determinism
- Reproducibility
- Full traceability from source data to narrative output

4.3 Output Verification (Unit Testing)

A post-generation validation step enforces numeric fidelity. If validation fails, the response is discarded and replaced with a system error message.

5. Uncertainty & Data Quality Guardrails

5.1 Small Sample Size Detection

If the match rate between Unfit properties and issued permits falls below a defined threshold, the model is instructed to prepend a warning disclaimer.

5.2 Hypothesis Constraints

The AI may only reference causal explanations that are directly observable in the dataset:

- **Allowed:** Missing permits, repeated inspection failures, administrative lag.
- **Disallowed:** Political decisions, budget allocations, owner financial status.

6. Example AI Narrative Output

User Query: *What is the status of the Northside neighborhood?*

AI Response: > "Based on current city records, the Northside neighborhood contains 2,859 unresolved housing code violations, representing the largest concentration of open cases city-wide. Properties designated as Unfit remain in that status for a median duration of 895 days, contributing significantly to the overall Remediation Gap of 19.5%. While permit activity exists, only 41.2% of Unfit properties show a corresponding repair permit. Data on owner financial capacity or external constraints is not available in this dataset; therefore, observed delays can only be attributed to documented administrative and permitting timelines."

7. Edge Case Handling & Governance

- **Invalid Geographic Queries:** No inference or fallback guessing is permitted for neighborhoods outside city GIS boundaries.

- **Temporal Inconsistencies:** Records where permit dates precede violation dates are flagged as "Inconsistent Entry" and excluded from duration metrics.
- **Auditability:** Every generated narrative can be traced back to a specific source dataset version hash and KPI computation timestamp.

This approach establishes a scalable blueprint for responsible AI use in municipal analytics, ensuring the AI layer functions as an interpretable reporting interface, not a decision-making authority.