Final Report

Development of a Machine Learning and Large Language Model-Based Classification Tool for Automated Analysis of Transportation Public Comments

Milan Knezevic
University of Pittsburgh
Phone: 412-403-2221; Email: mik296@pitt.edu

Trevor Neece
University of Pittsburgh
Phone: 412-403-2221; Email: trn37@pitt.edu

Aleksandar Stevanovic
University of Pittsburgh
Phone: 412-383-3766; Email: stevanovic@pitt.edu

Lev Khazanovich
University of Pittsburgh
Phone: 213-245-9609; Email: lev.k@pitt.edu

Date
March 30, 2026

Prepared for the Sustainable Mobility and Accessibility Regional Transportation Equity Research Center, Morgan State University, CBEIS 327, 1700 E. Coldspring Lane, Baltimore, MD 21251

ACKNOWLEDGMENT

This research was supported by the Safety and Mobility Advancements Regional Transportation and Economics Research Center at Morgan State University and the University Transportation Center(s) Program of the U.S. Department of Transportation.

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, under grant number 69A3552348303 from the U.S. Department of Transportation's University Transportation Centers Program. The U.S. Government assumes no liability for the contents or use thereof.

1. Report No. SM45	**2. Government Accession No.**	**3. Recipient's Catalog No.**
4. Title and Subtitle Development of a Machine Learning and Large Language Model–Based Classification Tool for Automated Analysis of Transportation Public Comments		**5. Report Date** December 2025 **6. Performing Organization Code**
7. Author(s) Milan Knezevic, 0009-0007-9564-4333 Trevor Neece Aleksandar Stevanovic, 0000-0003-1091-3340 Lev Khazanovich		**8. Performing Organization Report No.**
9. Performing Organization Name and Address University of Pittsburgh 4200 Fifth Ave, Pittsburgh, PA 15260		**10. Work Unit No.** **11. Contract or Grant No.** 69A3552348303
12. Sponsoring Agency Name and Address US Department of Transportation Office of the Secretary-Research UTC Program, RDT-30 1200 New Jersey Ave., SE Washington, DC 20590		**13. Type of Report and Period Covered** Final, **14. Sponsoring Agency Code**
15. Supplementary Notes		

16. Abstract

Transportation agencies receive large volumes of free-form public comments describing infrastructure conditions, safety concerns, and service issues. Processing and classifying these comments manually is time-consuming, inconsistent, and difficult to scale, limiting their usefulness for operational decision-making. This study presents a machine learning and Large Language Model (LLM)-based framework for automated classification of transportation public comments across a three-level hierarchical taxonomy: Category, Subcategory, and Final Decision (Dismiss, Refer, Potential Project). Comments are encoded using MPNet sentence embeddings and indexed in a train-only vector database. Three classification strategies are evaluated: (1) an embedding-based k-nearest neighbors (kNN) baseline using cosine similarity, (2) a zero-shot/few-shot LLM with schema-constrained outputs, and (3) a retrieval-augmented LLM that incorporates semantically similar historical examples and concise label guidance during inference. Results indicate that while high-level categories are generally well separated, many public comments contain overlapping or multi-modal semantic content that spans multiple infrastructure types. To address this ambiguity, the retrieval-augmented LLM is evaluated using ordered predictions, allowing multiple plausible labels to be returned. This approach improves overall classification performance and provides richer context for agency review. Overall, the proposed framework demonstrates the potential of retrieval-augmented LLMs to support scalable, transparent, and decision-oriented processing of public transportation comments, enabling faster and higher-quality agency workflows.

17. Key Words: Classification, Machine Learning, Large Language Models, Public Transportation Comments		**18. Distribution Statement**	
19. Security Classif. (of this report) : Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 34	**22. Price**

Table of Contents

Abstract 5
Introduction 5
Problem Statement 5
Research Objectives 6
Literature Review 7
Methodology 8
Problem Formalization 9
Database processing 9
Classification methods 11
kNN Embedding cosine similarity search 11
LLM with Few-shot prompting 12
LLM with Retrieval classification 13
Experimental Setup 16
Results 16
Confusion Analysis 16
Miss-classification Analysis 19
Classification Accuracy 20
Discussion 22
Conclusions 23
References 25
Appendix A: Few-shot prompts per class 27
Appendix B: Retrieval prompts per class 30
Appendix C: Full List of Category-Level Mismatch Comments 33

Abstract

Transportation agencies receive large volumes of free-form public comments describing infrastructure conditions, safety concerns, and service issues. Processing and classifying these comments manually is time-consuming, inconsistent, and difficult to scale, limiting their usefulness for operational decision-making. This study presents a machine learning and Large Language Model (LLM)-based framework for automated classification of transportation public comments across a three-level hierarchical taxonomy: Category, Subcategory, and Final Decision (Dismiss, Refer, Potential Project). Comments are encoded using MPNet sentence embeddings and indexed in a train-only vector database. Three classification strategies are evaluated: (1) an embedding-based k-nearest neighbors (kNN) baseline using cosine similarity, (2) a zero-shot/few-shot LLM with schema-constrained outputs, and (3) a retrieval-augmented LLM that incorporates semantically similar historical examples and concise label guidance during inference. Results indicate that while high-level categories are generally well separated, many public comments contain overlapping or multi-modal semantic content that spans multiple infrastructure types. To address this ambiguity, the retrieval-augmented LLM is evaluated using ordered predictions, allowing multiple plausible labels to be returned. This approach improves overall classification performance and provides richer context for agency review. Overall, the proposed framework demonstrates the potential of retrieval-augmented LLMs to support scalable, transparent, and decision-oriented processing of public transportation comments, enabling faster and higher-quality agency workflows.

Introduction

Problem Statement

Reliable transportation systems are fundamental to the functionality and economic productivity of modern urban areas. Inefficiencies in system operations can lead to substantial time losses, reduced pedestrian and vehicle safety, and increased pollution levels [1]. Consequently, effective planning, monitoring, and operational management are essential to achieving efficient and sustainable transportation systems. Much of the existing academic research in this domain advances through the development of Intelligent Transportation Systems (ITS) and transportation infrastructure, emphasizing the application of modern technologies such as advanced communication systems, artificial intelligence, and machine learning to improve traffic management, mitigate congestion, and enhance overall system performance [2], [3], [4], [5].

Despite substantial progress in these areas, many day-to-day operational challenges faced by transportation agencies, such as routine maintenance issues, traffic signal malfunctions, and service reliability concerns, remain inadequately addressed in practice[6], [7], [8]. These “everyday” operational issues directly shape public experience and significantly influence overall system performance. A large share of these challenges originates from the high volume of free-form public comments submitted by road users (e.g., “the bus is consistently off schedule,” “pothole on 5th Ave,” or “signal is flashing red”). In current practice, transportation agencies

typically rely on manual review and processing of these comments for analysis, classification, and follow-up action. This approach is time-consuming, applied inconsistently across reviewers, and susceptible to fatigue-related errors, which can delay progress from reported issues to field resolution. Moreover, manual screening raises concerns about transparency and efficacy, as prioritization decisions may vary depending on staff workload and individual interpretation.

Accordingly, transportation agencies need an automated, reliable system that can (1) accurately classify public comments into a well-defined domain taxonomy, reducing repetitive manual effort; (2) provide evidence-based rationales to support accountability and oversight; and (3) enable human supervision, auditability, and continuous improvement. Such a system should also help identify recurring issues and emerging trends across neighborhoods, ensure consistent classification to reduce label noise, and generate transparent summaries and performance metrics for use in staff reports, public testimony, and regulatory or policy analysis.

Since the introduction of Large Language Models (LLMs), researchers and practitioners have increasingly explored ways to apply the technology in the transportation sector [9]. Existing applications primarily focus on structured or safety-related text analytics, such as aviation incident mining [10] accident-report automation, and traffic-safety data augmentation [11]. These efforts demonstrate the potential of LLM-based frameworks for extracting insights from textual data; however, their application in day-to-day operational contexts, particularly public-facing reporting systems in road transportation, remains relatively underexplored.

Given that public comments submitted to transportation agencies are predominantly free-form text, the present study investigates the potential of Large Language Models to support the automated triage of transportation-related public comments. The primary objective is to develop a scalable, interpretable, and reliable framework capable of categorizing unstructured text into an operational taxonomy relevant to transportation agencies. Beyond transportation-specific use cases, the proposed system is designed to be domain-agnostic. It incorporates a no-code graphical user interface (GUI) that enables non-technical users to repurpose the framework for other classification domains by supplying alternative taxonomy definitions, policy or routing rules, and a supporting knowledge corpus (e.g., via CSV uploads or database connectors).

Research Objectives

To achieve the overall goal of this study, the main objectives are to define and validate the following components:

- Apply machine learning–based text embedding methods to extract quantitative semantic representations from each free-form public comment, which are then used as input features for classification.
- Develop a baseline machine learning classifier based on a k-nearest neighbors (kNN) approach, and establish a reference LLM-based architecture using few-shot prompting combined with simple classification examples

- Design and evaluate an advanced prediction framework that integrates large language models with machine learning methods to improve classification performance and robustness
- Generation of transparent, evidence-based rationales to support human oversight; and
- Develop a no-code graphical user interface (GUI) that enables users to deploy and adapt the framework without requiring programming expertise.

Literature Review

With the rapid development of Large Language Models (LLMs), substantial potential has emerged for applying these models to complex transportation problems. LLMs have shown promising results in enhancing autonomous driving and driver-assistance technologies, particularly in voice-based assistance and driving-related decision making [12], [13], [14]. Furthermore, Da et al. utilized the interactive capabilities of LLMs to analyze how dynamic features influence traffic flow and proposed a prompt-based grounded action transformer for traffic signal control [15]. Villarreal et al. similarly demonstrated that ChatGPT could address complex mixed-traffic control problems, including ring-road, bottleneck, and intersection scenarios [16]. In addition, recent studies have shown that LLMs can support traffic monitoring by estimating crash risk or hazardous driving conditions [17], [18].

Compared to traditional predictive models such as multinomial logit, random forests, and neural networks LLMs have demonstrated strong reasoning and generalization capabilities in modeling and predicting travel behavior [19], highlighting their potential for transportation planning and behavioral analysis of road users. Building on this direction, Tan et al. introduced an LLM-based framework for urban driver-agent simulation that enables flexible adaptation to traffic rules through natural-language specifications [20]. Wong et al. conceptualized an LLM-powered personal travel assistant designed to support users throughout the entire journey lifecycle, including pre-trip planning (e.g., itinerary generation, route and accommodation recommendations), route assistance (e.g., navigation, translation, and dynamic replanning), and post-trip documentation [21]. In the context of driver behavior and safety analysis, Zhang et al. proposed a vision–language reasoning framework for real-time risk assessment, integrating visual scene understanding with an LLM-based chain-of-thought module to infer driver intentions and potential sources of distraction [22]. Comprehensive overviews of emerging LLM applications in transportation are provided by Wandelt et al. [9] and Hassan et al., [22].

LLMs have demonstrated the capacity to understand complex and even multimodal traffic patterns through their generalization and reasoning capabilities, emulating human-like decision-making in the process [23]. Recent studies have shown that when fine-tuned for traffic management tasks, LLM-based traffic controllers can replicate expert-level decision processes. For instance, LLM-light which is a fine-tuned LLM traffic controller, outperformed both a general-purpose LLM (without domain-specific fine-tuning) and traditional traffic signal control methods across a range of scenarios [23]. Furthermore, Zhang et al. introduced TrafficGPT [24], which combines LLM language understanding with traffic control knowledge to interpret operational contexts and suggest control strategies. Ma et al. proposed the Spatial-Temporal Transformer for Traffic Flow

(STTLM), which disentangles temporal and spatial feature learning through separate transformer blocks integrated with a pre-trained language model, where the temporal module captures location-specific temporal patterns and the spatial module models cross-location correlations at each time step [25].

In terms of text-classification tasks related to transportation, LLMs have been used primarily for analyzing data from various accident reporting systems. A custom aviation-domain BERT model (Aviation-BERT) was developed to perform text-mining tasks and was pre-trained on narrative data from the National Transportation Safety Board (NTSB) and the Aviation Safety Reporting System (ASRS) [10]. Furthermore, Jing et al. investigated multi-label classification problems and demonstrated the superior performance of Aviation-BERT on ASRS data [26]. Andrade and Walsh classified safety reports related to weather and operational procedures by introducing a safety-informed model, SafeAeroBERT [27]. Cheng et al. developed an entity-recognition-based approach to extract key features from traffic-accident text narratives, addressing challenges such as limited sample sizes and reduced accuracy in long-text recognition [10]. Fontes et al. introduced city-wide mobility intelligence using Twitter data and BERT embeddings, achieving high precision in identifying relevant tweets and associated sentiments, and demonstrating the potential of social media as "semantic sensors" for transportation disruptions [28].

Despite these contributions, to the best of the author's knowledge, none of the existing studies have focused on developing and analyzing a framework that uses text-based data to map unstructured public comments to a transportation operations taxonomy, with the goal of improving overall agency workflows.

Methodology

One of the primary objectives of this study is to automate the classification of text-based public comments submitted to transportation agencies. Accordingly, the task is formulated as a supervised text classification problem with label prediction at three hierarchical levels: Category, Subcategory, and Final Decision. In addition, the study seeks to quantify how much model complexity is required for this task by systematically comparing simple non-LLM approaches with Large Language Model–based methods, both with and without retrieval-augmented techniques:

1. **Embedding-based classification.**
   Classification is performed solely based on vector similarity. Each public comment is embedded into a semantic vector space, and predictions are generated using a k-nearest neighbors (kNN) search against vectorized representations of labeled examples.
2. **LLM few-shot classification.**
   A modern Large Language Model (Gemini) is prompted using carefully engineered instructions and few-shot examples to perform three-level hierarchical classification. This approach relies exclusively on the input text and prompt context, without external retrieval.
3. **Retrieval-Augmented Classification (RAC):**
   Following the framework proposed in [29], the LLM retrieves the most semantically similar labeled examples and concise taxonomy/style definitions from a vector database.

These retrieved elements are injected into the prompt as contextual evidence to refine decision boundaries and improve classification consistency and robustness.

Problem Formalization

We let $x \in X$ denote a free-form, text-based public comment submitted to a transportation agency (e.g., "pothole on Forbes Ave near Murray"). Each comment x is mapped to a hierarchical label structure consisting of three components:

- Category $y^C \in C = \{1, \dots, K\}$:
  The top-level classification represents the primary infrastructure or service domain to which the comment pertains. In this study, $K = 5$ corresponding to the categories: Roadway, Transit, Ped/Bike, Bridge, and Freight.
  For example, a comment such as *"At the intersection of Second Avenue and Elizabeth Street there is a pothole that makes driving dangerous"* is classified under the Roadway category, as it relates to roadway infrastructure conditions.
- Subcategory $y^S \in S_{yc} = \{1, \dots, K_{yc}\}$:
  A category-specific label that provides a more granular description of the issue. For instance, within the Roadway category, subcategories may include potholes, pavement markings, lane obstructions, or construction-related issues. This level enables agencies to maintain a more detailed and actionable database of infrastructure concerns.
- Final Decision $y^D \in D = \{1, \dots, L\}$
  An operational label indicating the recommended next step for agency action. In this study $L = 3$ corresponding to: (i) initiation of a potential project, (ii) referral to another responsible agency or municipality (e.g., maintenance authority), and (iii) dismissal of the comment due to insufficient information or irrelevance (e.g., vague complaints or road-rage-related messages).

Importantly, the framework predicts recommended actions rather than final decisions, as the system is intended to support agency staff rather than replace human judgment. Finally, the overall prediction for a comment x is a triplet:

$$y = (y^C, y^S, y^D) \in \bigcup_{c \in C} (\{c\} \times S_c \times D) \tag{1}$$

Database processing

The dataset provided by the transportation agency consists of 900 public comments, each manually labeled at three hierarchical levels: Category, Subcategory, and Final Decision. As required for supervised learning–based evaluation, the dataset is partitioned into separate subsets for model development and performance assessment. While traditional machine learning pipelines typically involve explicit model training, the approaches evaluated in this study including k-nearest neighbors (kNN) [30] and LLMs do not require parameter optimization through gradient-based

training. Instead, labeled data are used as a form of prior knowledge to support similarity-based inference and retrieval-augmented reasoning.

Accordingly, the dataset is divided into two disjoint subsets: a knowledge dataset and an evaluation dataset, as illustrated in Figure 1. Specifically, 700 comments are allocated to the knowledge dataset, while the remaining 200 comments are reserved exclusively for final evaluation. The evaluation dataset is never exposed to the models during inference support or retrieval, ensuring an unbiased assessment of classification performance.

Each comment in the knowledge dataset is converted into a vector representation using a text embedding model. Formally, a vector embedding defines a mapping from a set of objects X to a latent space $\mathbb{R}^d$ of finite dimension d such that semantic or structural relationships in the original space are preserved as geometric relationships in the latent space [31]. Under this formulation, semantically similar objects are mapped to vectors that are close to one another according to a chosen distance metric, such as Euclidean or cosine distance.

In the context of natural language processing, embeddings capture both semantic meaning and contextual usage. For example, words such as "car" and "vehicle" are expected to appear close to each other in the latent space, whereas unrelated concepts such as "banana" or "apple" will be positioned farther away due to differing contextual associations. Embedding techniques are not limited to text and have been successfully applied to images, audio, and video, making them a foundational component of modern artificial intelligence systems [32], [33].

In this study, the MPNet embedding model is employed to generate dense vector representations for all public comments in the knowledge dataset.

Following embedding generation, all vectors are L2-normalized to unit length, enabling stable and efficient cosine-similarity based nearest-neighbor search. The resulting embeddings are stored in a vector database that serves two complementary purposes:

1. **Similarity-based classification:**
   The vector index is used directly by the kNN classifier to identify the most semantically similar labeled comments and infer class labels based on proximity in the embedding space
2. **Retrieval-augmented classification:**
   The same vector database functions as the retrieval source for the retrieval-augmented classification (RAC) framework, supplying the LLM with relevant labeled examples and supporting evidence during inference.

The knowledge database consists of two tightly coupled components: (i) a vector index that stores high-dimensional embeddings of approximately 700 public comments, and (ii) a metadata store that retains the original comment text along with its associated hierarchical labels. This unified database design enables consistent comparison across classical machine learning, LLM-based, and

retrieval-augmented approaches while maintaining strict separation between knowledge support and evaluation data.

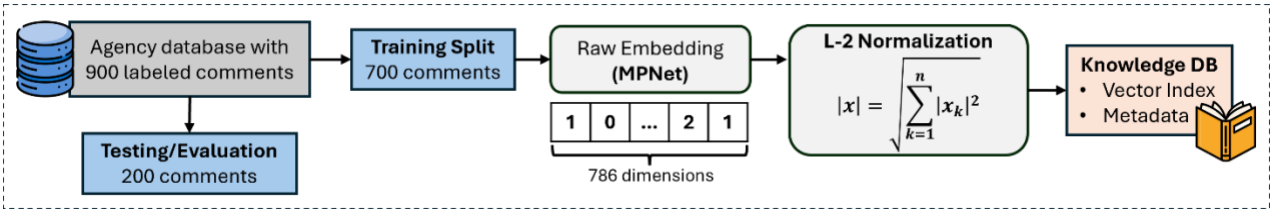

Figure 1. Database processing steps

Classification methods

kNN Embedding cosine similarity search

For each comment in the evaluation set, a vector embedding is generated following the same procedure used to construct the knowledge database. Let $\phi(x)$, denote the L2-normalized sentence embedding of a comment x, such that $\|\phi(x)\|_2 = 1$. Furthermore, cosine similarity between two embeddings reduces to their dot product:

$$\cos(u, v) = u^T v \in [-1,1] \tag{2}$$

where u and v denote unit-length vectors corresponding to normalized embeddings.

Furthermore, given a new comment x we retrieve the k most similar labeled instances are retrieved from the knowledge database K_{db} using cosine similarity:

$$N_k(x) = TopK_{x_j \in K_{db}} \cos\left(\phi(x), \phi\left(x_j\right)\right) \tag{3}$$

Rather than relying on simple majority voting, cosine similarity scores are converted into non-negative normalized weights using a softmax transformation:

$$w_j = \frac{e^{\cos\left(\phi(x), \phi\left(x_j\right)\right)}}{\sum_{j' \in N_k(x)} e^{\cos\left(\phi(x), \phi\left(x_{j'}\right)\right)}}, \forall j \in N_k(x) \tag{4}$$

This formulation assigns higher influence to neighbors that are more semantically similar to the query comment. hierarchical label level $L \in \{C, S, D\}$, (Category, Subcategory, Final Decision), the predicted label, is obtained via weighted aggregation:

$$\hat{y}^L = arg\,\max_l \sum_{j \in N_k(x)} w_j \mathbb{I}\left[y_j^L = l\right] \tag{5}$$

Where:

- y_j^L is the level L label of neighbor j,
- l indexes candidate labels at level L,
- $\mathbb{I}(\cdot)$ is the indicator function

LLM with Few-Shot Prompting

To evaluate the performance of the LLM, this study begins with a pretrained model (Gemini-2.0-Flash-Lite) and applies standard prompting techniques using a small number of few-shot examples. The prompt configuration follows three components:

- **Context Settings:**
  The prompt explicitly places the LLM in the role of a transportation-domain classifier (e.g., *"You are a transportation engineer working in a public agency."*), ensuring that the model interprets the task within the correct domain.
- **Concise task definitions and label descriptions:**
  Each category, subcategory, and decision label is described clearly and succinctly. In addition to definitions, the prompt includes dataset-specific classification rules. For example, although some pothole-related reports could be interpreted as roadway issues, agencies typically route anything involving bridges to the *Bridge* category; therefore, such domain-specific rules are explicitly stated.
- **Few-shot illustrative examples:**
  A small set of representative examples demonstrates how comments should be classified across all three label levels, helping the model infer the structure and expected output format.

This prompting configuration serves as a fair baseline for comparison, requiring neither fine-tuning nor access to the training dataset.

To construct a few-shot baseline prompt all instructions are constant and selected manually. To formulate it let:

- $V^C = C$ denote the category vocabulary.
- For a given category, let $\hat{c} \in C$ and $V^{S|\hat{c}} = S_{\hat{c}}$ denote the Subcategory vocabulary.
- $V^D = D$ denote the Final decision vocabulary.

Therefore, the system prompt defines the relevant label set and includes brief examples. The user message contains only the raw comment x. The model is instructed to output exactly one label from the corresponding vocabulary. The system prompt P is denoted as:

$$P^j(x) = C^j||T^j||E^j_{const}||\ x, \forall j\ \in \{C, S, D\} \tag{6}$$

Where:

- C is a brief context-setting instruction.
- T provides the task definition and allowed label set.
- E_{static} is a fixed set of generic examples,
- x is the raw user comment.

A strict parser accepts only exact matches to the allowed vocabulary at each level. If the output is invalid (e.g., extra text), one automatic retry is issued with a reinforced instruction ("Output only one of: "). If still invalid, a case-insensitive nearest-label mapping is applied, and the instance is

flagged for audit. Figure 2 shows the entire few-shot prompt for category-label classification; the remaining prompts are provided in the Appendix.

```
CONTEXT
You are a transportation engineer agent. Analyze one public comment about
traffic/transportation and classify it into CATEGORY from:
[Bridge, Pedestrian_Or_Bike, Roadway, Transit, Freight, Other].

DEFINITIONS
• Bridge: Bridges/overpasses/viaducts/bridge maintenance or new bridge requests
(including pedestrian bridges).
• Pedestrian_Or_Bike: Sidewalks, crosswalks, bike lanes, trails, micromobility—
without requesting a new bridge structure.
• Roadway: General roadway operations or minor maintenance (potholes, striping,
signage, turn lanes, traffic calming, roundabouts).
• Transit: Buses, bus stops, bus routes, schedules, shelters, rail transit
service.
• Freight: General complaints of operations for freights, trucks, rail.
• Other: Use if the comment is vague, irrelevant, or cannot reasonably be assigned
to the other categories.

DECISION POLICY
1) If the comment explicitly requests/mentions a bridge/overpass/viaduct (e.g.,
"pedestrian bridge"), classify as Bridge (even if it benefits peds/bikes).
2) Otherwise, use clear cues:
  • "bus/stop/route/schedule/transit" → Transit
  • "sidewalk/crosswalk/bike/bikeway/trail/multi-use path" → Pedestrian_Or_Bike
  • "potholes/striping/signage/minor repair/turn lane/roundabout/traffic calming"
→ Roadway
  • "freights/trucks/rails/transport of goods/facilities" → Freight
3) If the comment is vague, pick the Other category.

EXAMPLES
• Bridge: The bridge is in poor condition and will need a full rehabilitation or
replacement.
• Pedestrian_Or_Bike: There is no sidewalk to get from Timberbrook Dr to Jeremiah
Dr.  This missing link would allow many people to walk to Main St.
• Roadway: Heman Road is deteriorating with pot holes and crumbling shoulders. The
road base and subgrade are damaged resulting in weak spots.
• Transit: no public transportation is available and would like to have for
residents to travel to other locations for doctors hospitals food shopping.
• Freight: Trucks traveling through this area have trouble navigating narrow roads
and cause significant traffic congestion.

OUTPUT FORMAT
Now classify the comment: "I want a new pedestrian bridge over the highway."
CATEGORY: <Bridge|Pedestrian_Or_Bike|Roadway|Transit|Freight|Other>
Reason: <1-2 short sentences citing the key phrase(s)>
```

Figure 2. Few-shot prompt for category label prediction

LLM with Retrieval Classification

As previously described, a vector index is constructed over the knowledge database K_{db}, storing tuples of the form $\left(\phi(x_j), y_j^C, y_j^S, y_j^D\right)$. Given a new comment, the top-K nearest neighbors are retrieved under cosine similarity using the procedure defined earlier in equation 4. Furthermore, an evidence pack is then formed by combining (i) the retrieved labeled examples from $N_k(x)$ and (ii) their cosine similarity scores to determine the most similar examples. Thus, for each neighbor $x_i \in N_k(x)$, we assign a rank $r_j \in \{1, \dots, K\}$, where $r_j = 1$ corresponds to the most similar instance.

Because raw cosine similarities may be tightly clustered, we define an effective similarity that penalizes lower-ranked neighbors:

$$\hat{s}_j = \frac{\cos\left(\phi(x), \phi(x_j)\right)}{r_j} \tag{7}$$

Label support for a candidate label l is computed by summing effective similarities over the retrieved set:

$$S(l) = \sum \hat{s}_j \, \mathbb{I}\left[y_j^L = l\right] \tag{8}$$

This is applied at each level (Category, Subcategory conditioned on Category, and Final decision) with the appropriate label sets. The decision margin at a level is reported as the difference between the top two supports (e.g., $\Delta^C = S(c_1) - S(c_2)$). The retrieval stage have two features:

- Support Similarity Table
  For each of the k neighbors, a unique identifier, a short snippet, cosine similarity score s_j, rank r_j, effective similarity $\hat{s}_j$, and historical labels.
- Label Summary
  Aggregated support $S(\cdot)$ for each candidate label at the current level and the resulting decision margin Δ.

Similar to the few-shot baseline, LLM with retrieval pipeline receives task-specific definitions, the allowed label set, minimal policy rules, and only the necessary retrieved snippets to preserve context. The retrieval corpus functions as an external knowledge source for classification, confidence estimation, and active-learning triage. The retrieval-augmented prompt for level j is formulated as:

$$P^j(x) = C^j || T^j || R^j(x) || \, x, \forall j \in \{C, S, D\} \tag{9}$$

Where components C^j, T^j, and x have the same meaning as in equation 6, and $R^j(x)$ denotes the retrieval component containing the label support $S(l)$ and decision margins Δ derived from the nearest neighbors. Figure 3 shows the example of the retrieval LLM prompt for Category label classification. All other prompts are shown in the Appendix.

CONTEXT

You are a transportation engineer ...

DEFINITIONS

| Bridge | Pedestrian_Or_Bike | Roadway | Transit | Freight | Other |

DECISION POLICY

1) If the comment explicitly requests/mentions a bridge/overpass/viaduct (e.g., "pedestrian bridge"), classify as Bridge (even if it benefits peds/bikes).
2) If Top-K retrieved examples and a Support(C) summary are provided:
  • Treat Support(C) = summed weighted similarity per category as prior evidence.
  • If (top Support(C) − second Support(C)) ≥ 10% of total, treat the top category as strong prior evidence.
  • If the margin is smaller, apply rule (3) based on the comment's wording.
3) Otherwise, use clear cues:
  • "bus/stop/route/schedule/transit" → Transit
  • "sidewalk/crosswalk/bike/bikeway/trail/multi-use path" → Pedestrian_Or_Bike
  • "potholes/striping/signage/minor repair/turn lane/roundabout/traffic calming" → Roadway
  • "freights/trucks/rails/transport of goods/facilities" → Freight
4) If the comment is vague, pick the Other category.

RETRIEVAL INFORMATION

For new comment: "I want a new pedestrian bridge over the highway."

Support by category (sum of effective similarities):
Pedestrian_Or_Bike=1.166 (77.9%) | Bridge=0.331 (22.1%)
55.76% >= 10%

Margin difference = True

Top-K evidence table:
[Rank | sim | eff_sim | CATEGORY | Snippet]
1 | 0.663 | 0.663 | Pedestrian_Or_Bike | "There are crosswalks here that…"
2 | 0.662 | 0.331 | Bridge | "Bridge should have safe and accessible…"
3 | 0.661 | 0.220 | Pedestrian_Or_Bike | "Protected bike lanes ple…"
4 | 0.630 | 0.158 | Pedestrian_Or_Bike | "NEED DEDICATED BIKE LANE."
5 | 0.626 | 0.125 | Pedestrian_Or_Bike | "This could be a connection for…"

OUTPUT FORMAT

Now classify the comment: "I want a new pedestrian bridge over the highway."
CATEGORY: <Bridge|Pedestrian_Or_Bike|Roadway|Transit|Freight|Other>
Reason: <1–2 short sentences citing the key phrase(s)>

Figure 3. LLM prompt with retrieval information (the comments in the evidence table are shortened for figure)

Experimental Setup

The dataset used in this study consists of 900 free-form public comments provided by the Southwestern Pennsylvania Commission (SPC). Each comment was manually labeled by agency staff according to the three-level taxonomy described earlier (Category, Subcategory, and Final Decision). These annotations serve as the ground-truth labels for training, evaluation, and prompt conditioning.

For the baseline model, each comment was converted into a dense vector representation using the MPNet sentence encoder, producing a 768-dimensional embedding. A kNN classifier was then applied using $k = 5$, chosen based on a knee-point analysis of validation accuracy. Cosine similarity was used as the distance metric. This baseline provides a useful reference for methods that leverage historical data.

For both few-shot prompting and retrieval-augmented prompting, we used the Gemini-2.0-Flash-Lite model. The model was configured with the following generation parameters across all experiments:

- Temperature: 0.5
- Top-k: 40
- Top-p: 0.9
- Max output tokens: 512

Results

Confusion Analysis

To evaluate the performance of the classification methods, confusion matrices were generated for the Category and Final Decision levels (see Figure 4) as well as for the Subcategory level within each Category (see Figure 5). These matrices reveal several consistent cross-model patterns, highlighting structural ambiguity in the taxonomy and exposing the dominant sources of misclassification.

At the Category level, the most prominent confusion patterns include:

- Bridge frequently misclassified as Roadway
- Pedestrian/Bike misclassified as Roadway
- Roadway misclassified as Pedestrian/Bike
- Transit misclassified as Roadway or Pedestrian/Bike

These errors are largely attributable to linguistic overlap between categories: many comments describe safety, geometry, or maintenance concerns that apply simultaneously to multiple infrastructure types.

The Final Decision confusion matrices show the strongest diagonal dominance, indicating generally robust performance across methods. Most remaining errors are interpretable:

- *Refer* is frequently misclassified as *Potential Project*, reflecting the conceptual proximity of issues requiring referral to another agency versus issues that could be construed as capital projects.
- *Dismiss* is occasionally predicted as *Refer*, consistent with subjective judgment differences and agency policy nuances.
- *Potential Project* is often predicted as *Refer,* particularly when the described issue is legitimate but responsibility lies outside the agency.

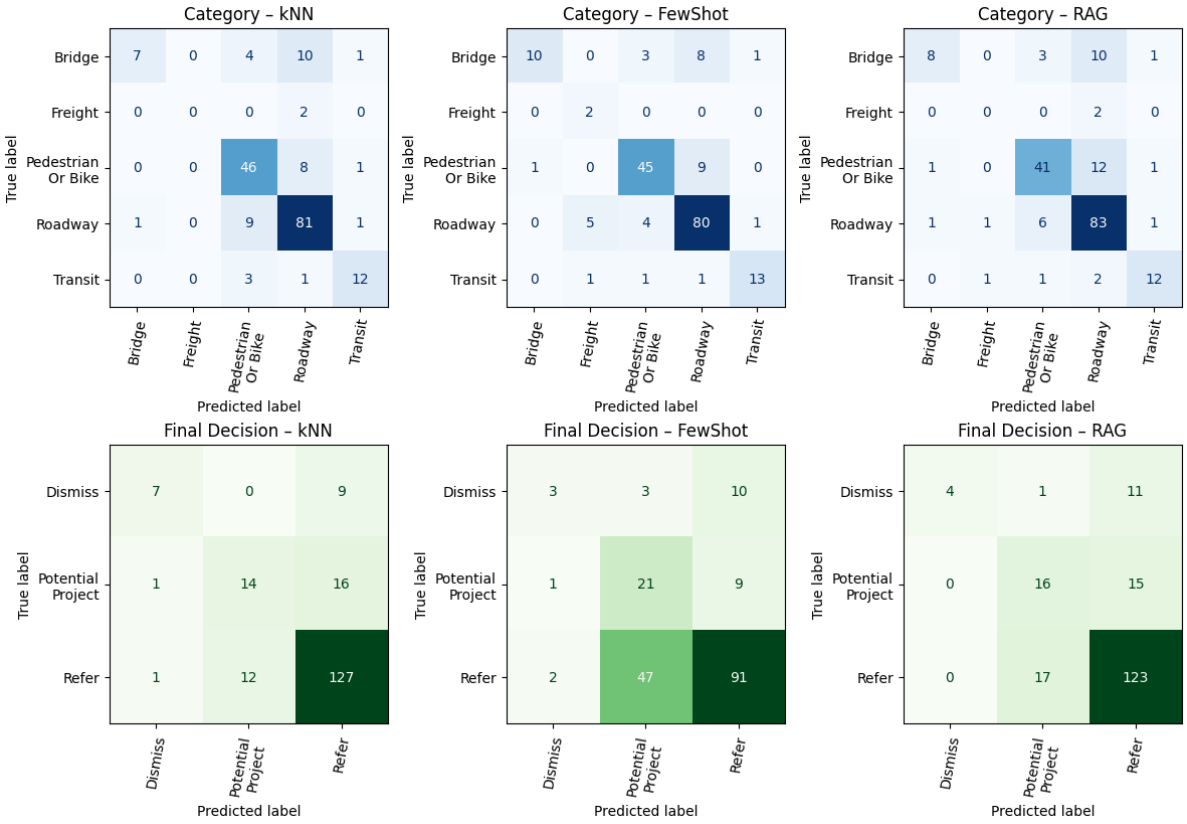

Figure 4. Confusion matrices for Category and Final decision classes.

At the Subcategory level, the confusion patterns are more diffuse, reflecting the fine-grained and lexically overlapping nature of the labels:

- Roadway subcategories exhibit mixing among *Congestion, Intersection & Signals, and Safety & Vulnerable Users*, which often co-occur in real comments (e.g., speeding near an unsafe intersection that also causes congestion).
- Pedestrian/Bike subcategories frequently confuse *New Facilities, Maintenance, and Safety*, as short comments rarely specify whether the issue concerns new infrastructure or deterioration of existing facilities.
- Bridge subcategories show overlap between *Maintenance* and *Congestion*, consistent with comments describing lane closures or work zones on bridges.
- Transit subcategories display mixing between *Service* and *Accessibility*, though small class sizes also contribute to sparse matrices.

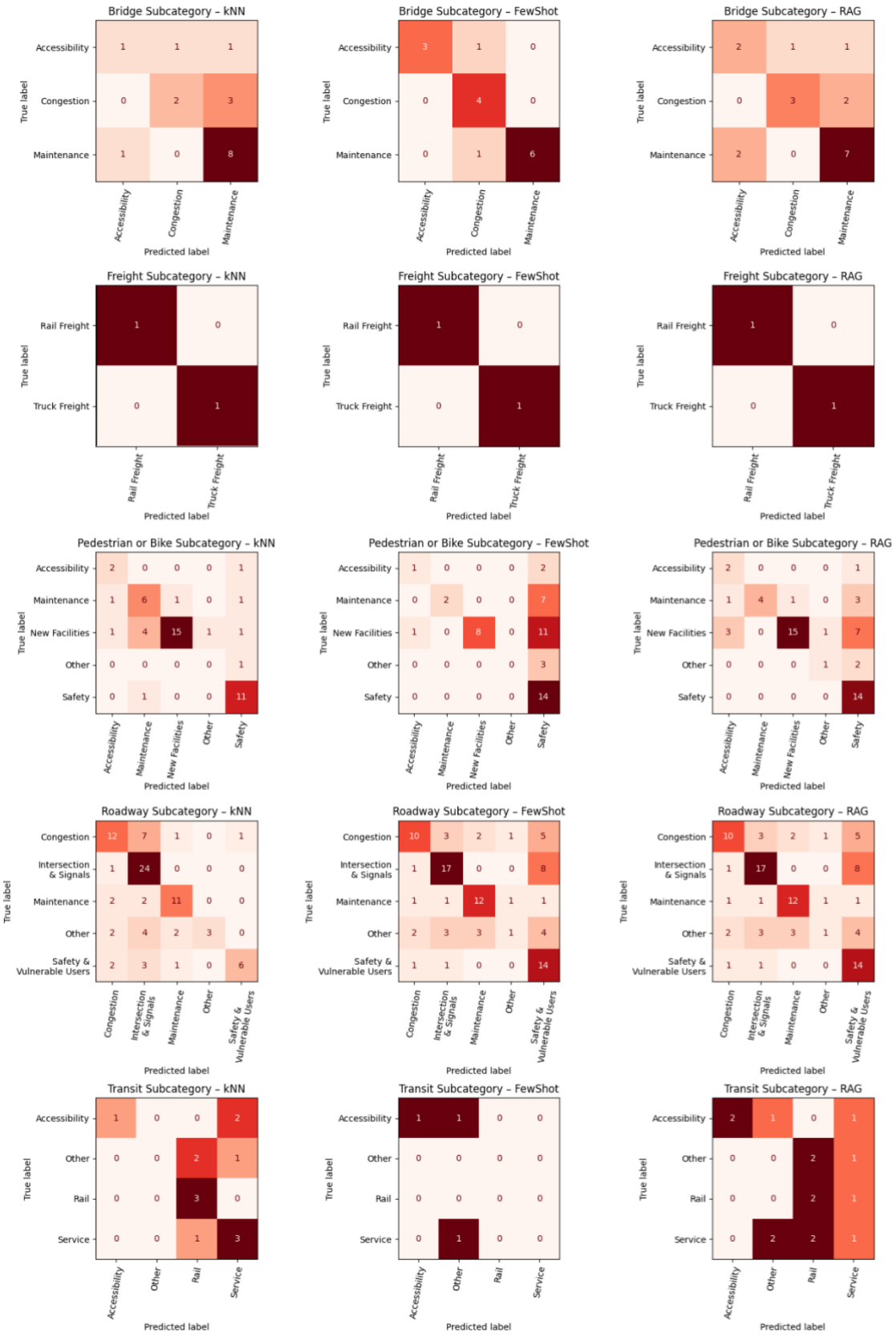

Figure 5. Confusion matrices for Subcategories per Category classes

Misclassification Analysis

While confusion matrices provide a useful high-level view of model performance, they do not fully explain *why* specific errors occur. To address this, we conducted a qualitative review of misclassified comments to assess whether these errors represent true model failures or stem from inherent semantic ambiguity in the comments themselves. Table 1 presents two representative examples for each major category-level mismatch identified in the confusion matrix analysis. The rest of the examples are attached as an appendix.

Table 1. Representative comment examples for major category-level mismatches

Mismatch Pair	Example Comments
1. Bridge category classified as Roadway	"Road resurfacing"
	"Every rainstorm causes flooding in this area. I have hydroplaned multiple times."
2. Pedestrian/Bike category classified as Roadway	"West Carson St has no adequate bike facilities. Traffic frequently exceeds 35 mph and merge lanes are used as passing lanes."
	"Pedestrian killed while crossing McKnight Rd. Please lower speed limit and redesign road to reinforce lower speeds."
3. Roadway category classified as Ped/Bike	"Blind corner on steep hill makes it dangerous for pedestrians, bikes, and drivers alike."
	"People must walk in the vehicle lanes since parts of these roads have no berm. These roads need to be widened so people can safely walk along them."
4. Transit category classified as Rodway	"Traffic moves fast and things get dangerous around here. Consider active features to control traffic better at this location."
	"This corridor is extremely dangerous. High pedestrian traffic conflicts with high car traffic and there is not enough room to handle pedestrians."
5. Transit category classified as Ped/Bike	"Many stations on this transit line are not ADA accessible and are not easily accessible by pedestrians."
	"This corridor is extremely dangerous. High pedestrian volume constantly conflicts with car traffic and pedestrian space is inadequate."

This misclassification analysis focuses primarily on the Category level rather than the Final Decision level. Category labels are directly grounded in the semantic content of the comments (e.g., Bridge, Roadway, Pedestrian/Bike), whereas final decision labels (Dismiss, Refer, Potential Project) partially depend on agency-specific factors such as jurisdiction, policy context, or external decision-making constraints that are not always explicitly expressed in the comment text. For example, an issue may be referred not because of its semantic content, but because responsibility lies with another agency. Consequently, Category-level mismatches provide a clearer and more reliable assessment of the models' semantic reasoning capabilities.

The mismatch pairs in Table 1 correspond directly to the dominant patterns observed in the confusion matrices, such as Bridge to Roadway and Pedestrian/Bike to Roadway. These examples demonstrate that many "errors" arise from overlapping or multi-modal content (e.g., comments about speeding, flooding, or sidewalk gaps that implicate multiple infrastructure types). As a result,

some misclassifications may reflect reasonable alternative interpretations rather than incorrect model behavior.

For this reason, we further analyzed the retrieval-augmented LLM by evaluating both its primary (first) and secondary (second) predicted categories, rather than relying on a single top prediction. Given the frequent presence of overlapping or multi-modal semantic content within public comments, a single label may not fully capture all plausible interpretations. Therefore, accuracy was evaluated based on whether the correct category appeared in either the first or second prediction. Because category and subcategory labels are more directly tied to semantic content, this evaluation was applied only at the category and subcategory level.

Classification Accuracy

Table 1 summarizes the classification performance of the three prediction methods (kNN, Few-Shot LLM prompting, and retrieval-augmented LLM (with just one prediction and with two predictions) across all taxonomy levels. Performance is reported using accuracy, representing the proportion of exact matches with the ground-truth label.

Table 2. Performance Comparison Across Modeling Strategies

Task	Model	Accuracy (%)
Category	kNN	78.07
	Few-shot prompting	80.21
	Retrieval-based prompting	77.01
	Retrieval + second LLM step	95.00
Subcategory	kNN	57.75
	Few-shot prompting	50.80
	Retrieval-based prompting	61.50
	Retrieval + second LLM step	79.00
Final Decision	kNN	79.14
	Few-shot prompting	61.50
	Retrieval-based prompting	76.47

Overall, performance differences across methods are modest, but several clear patterns emerge. At the Category level, few-shot prompting achieves the highest accuracy (approximately 80%), slightly outperforming both kNN and the retrieval-augmented LLM. This indicates that the base LLM, when provided with a small number of domain-specific exemplars, can generalize effectively to broad semantic groupings such as Roadway, Pedestrian/Bike, or Bridge. In contrast, the retrieval-enhanced LLM shows slightly lower performance. This suggests that, for high-level categories, retrieved snippets may introduce noise or conflicting context.

At the finer-grained Subcategory level, the retrieval-augmented LLM yields the strongest results (accuracy 0.615), outperforming both the Few-Shot LLM and kNN. Unlike the high-level taxonomy, Subcategories appear to benefit from external evidence, such as retrieved examples or prior semantic patterns. This is expected: Subcategories are substantially more granular (typically

~5 labels per Category), and their distinctions often rely on subtle semantic cues that a base LLM may not detect without additional context.

However, there is a clear improvement in LLM classification accuracy and in handling overlapping comments when the model is given the opportunity to predict two possible labels for a single comment. Beyond improving overall accuracy, this approach provides additional contextual information to agency staff when reviewing individual comments, supporting faster and higher-quality decision-making. This capability is incorporated into the final application, aligning with the overall goal of improving operational efficiency for transportation agencies.

At the Decision tier (Dismiss, Refer, Potential Project), kNN performs surprisingly well (accuracy 0.791). Many final decisions reflect patterns present in historical human annotations, allowing the kNN classifier to benefit from similarity-based reasoning over the labeled dataset. The few-shot LLM shows the lowest performance (accuracy 0.615) and tends to overpredict actionable categories such as Potential Project. The retrieval-augmented LLM improves substantially over few-shot prompting (accuracy 0.765), suggesting that retrieved examples help anchor decision-making to precedents in the training corpus.

All methods degrade substantially on the Subcategory task. The label space is fine-grained, and many sublabels differ only by a few cue words. Because comments are short and semantically similar within a Category, nearest neighbors frequently include items from competing sublabels, diluting classification signals. This is reflected in the embedding structure: the PCA visualization for the Roadway Category (Figure 6) shows no clear subclusters by Subcategory, suggesting limited separability and substantial lexical overlap between sublabels.

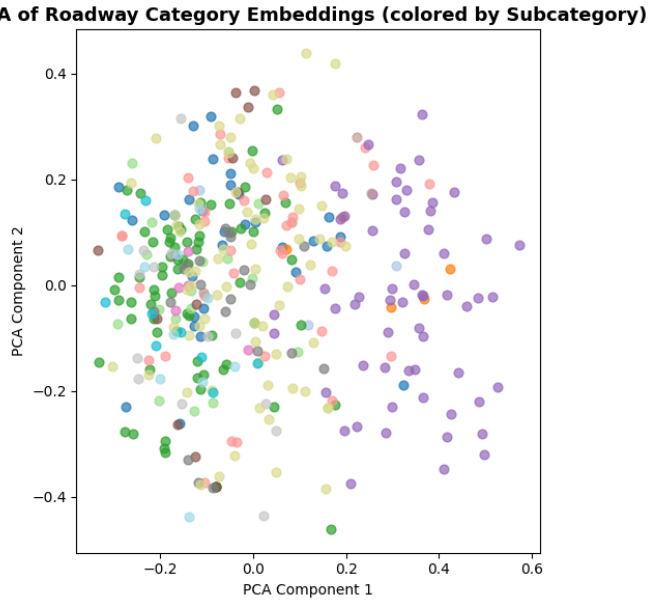

Figure 6. PCA of embeddings within the Roadway Category colored by Subcategory. Lack of distinct subclusters indicates fine grained, overlapping sublabels, which hinders Subcategory classification.

Discussion

Depending on the prediction level (Category, Subcategory, or Final Decision), the overall classification accuracy is broadly comparable across the evaluated methods, while each method exhibits distinct strengths at different levels of granularity. Because vector embeddings encode the semantic meaning of each comment, the high-level Category class is naturally the most separable. At this level, the semantic content of comments belonging to different categories is clearly distinguishable. For example, a comment related to potholes or roadway maintenance is semantically very different from a comment concerning bus scheduling issues in the transit category. These strong semantic differences explain why both the kNN classifier and the few-shot LLM achieve higher performance at the Category level. The high separability benefits kNN due to well-defined clusters in the embedding space, while the pretrained LLM, even with only a few labeled examples and a category definition, is able to generalize effectively at this coarse level without requiring a large historical training database.

On the other hand, at a more granular level (i.e., the Subcategory level within a single Category), the overall semantic meaning of the comments becomes much more homogeneous. For example, most users who report roadway maintenance issues tend to describe a specific road segment or location in similar contextual terms. However, because the subcategories are more specific, such as potholes, damaged traffic signs, or malfunctioning signals, the distinctions between them are often expressed through only a few key words. As a result, the corresponding embedding vectors become highly similar in the latent space (see Figure 6), which significantly reduces class separability. This explains the observed performance degradation of both the kNN classifier and the few-shot LLM at the subcategory level. When kNN-based similarity retrieval is combined with a pretrained LLM through the use of an evidence table, the model benefits from being exposed to highly relevant, correctly labeled examples, leading to measurable improvements in accuracy. Nevertheless, the overall performance remains limited, indicating that fine-grained semantic disambiguation at the subcategory level remains an open challenge and an important direction for future work.

The performance at the Final Decision level is largely driven by the operational policy of the responsible agency (in this case, the Southwestern Pennsylvania Commission (SPC)) and how it processes different types of public comments. Because this classification level consists of only three labels (Potential Project, Refer, and Dismiss), both kNN and retrieval-based methods achieve the strongest overall performance. The comments within each of these three labels exhibit a high degree of internal similarity. For example, comments that are ultimately dismissed are often incomplete, unstructured, or vague, such as general complaints about poor traffic conditions or statements. This leads to highly similar embedding representations and well-defined clusters. Similarly, comments that are referred typically correspond to responsibilities outside the agency's scope (e.g., maintenance projects or traffic signal operations), making them semantically distinct in a manner analogous to the high-level Category classification. This strong semantic separability enables kNN to perform particularly well, both as a standalone classifier and as a retrieval mechanism supporting a pretrained LLM. In contrast, the few-shot LLM struggles at this level because the Final Decision labels are highly agency-specific, and general label definitions with a limited number of examples are insufficient to convey the detailed operational context required

for accurate classification. This further highlights the importance of retrieval-based grounding for decision-level prediction tasks.

The evaluation of misclassified comments provides important insights into the underlying nature of this data. Because the ground-truth labels were manually assigned by human annotators, an inherent degree of human subjectivity and bias is introduced into the classification process. As illustrated by several examples in Table 1, the assigned label may vary depending on the interpretive perspective adopted by the annotator. For instance, Comment 3.1. is labeled as a pedestrian/bicycle issue, although it describes a "blind corner on a steep hill that makes it dangerous for pedestrians, bikes, and drivers alike." From a technically oriented perspective, this comment could be interpreted as a roadway design deficiency requiring geometric redesign. In contrast, from a planning-oriented perspective, the same issue may be viewed as a policy-related deficiency in pedestrian treatment at the intersection. In this sense, some apparent "errors" are not true misclassifications but rather valid alternative interpretations of the same complaint. These ambiguities suggest that strict single-label classification may be insufficient for practical deployment. A promising extension of the proposed framework is to allow LLM to return multiple plausible categories along with supporting rationale, thereby providing broader contextual guidance to agency staff. Ultimately, the objective of the system is not merely to maximize label prediction accuracy, but to improve the efficiency and quality of decision-making within transportation agencies.

Conclusions

This study examined transportation-related public comments using natural language processing techniques and large language models (LLMs) with the goal of developing an effective classification framework for free-form text describing mobility issues in cities.

The findings demonstrate that simple NLP methods, such as sentence embeddings paired with a k-nearest neighbors (kNN) classifier, can achieve strong performance when sufficient historical labeled data are available. At the same time, few-shot LLM prompting using only a small number of labeled examples and concise label definitions also performs well, particularly for higher-level semantic categories. When LLM predictions are further supported by embedding-based retrieval, performance improves for fine-grained labels such as subcategories, where external examples help resolve subtle distinctions.

A key advantage of LLM-based approaches is their ability to handle comments that contain overlapping or multi-modal content, which are often difficult even for human annotators to classify consistently. Our misclassification analysis showed that many "errors'' are semantically reasonable alternatives. For example, comments describing speeding, flooding, sidewalk gaps, or unsafe intersections may legitimately fit multiple infrastructure categories. In such cases, model outputs provide useful insights rather than genuine misclassifications.

Overall, this work demonstrates the potential of LLMs to enhance the automation of public comment processing for transportation agencies. Future research should explore alternative

classification architectures, more advanced retrieval methods, reinforcement learning with human feedback from agency staff, and integration of these tools into operational workflows.

References

[1] C. Steger-Vonmetz, "Improving modal choice and transport efficiency with the virtual ridesharing agency," in *Proceedings. 2005 IEEE Intelligent Transportation Systems, 2005.*, Vienna, Austria: IEEE, 2005, pp. 994–999. doi: 10.1109/ITSC.2005.1520186.

[2] P. T. Weiss, M. Kayhanian, J. S. Gulliver, and L. Khazanovich, "Permeable pavement in northern North American urban areas: research review and knowledge gaps," *International Journal of Pavement Engineering*, vol. 20, no. 2, pp. 143–162, Feb. 2019, doi: 10.1080/10298436.2017.1279482.

[3] N. Mitrovic, I. Dakic, and A. Stevanovic, "Combined Alternate-Direction Lane Assignment and Reservation-Based Intersection Control," *IEEE Trans. Intell. Transport. Syst.*, vol. 21, no. 4, pp. 1779–1789, Apr. 2020, doi: 10.1109/TITS.2019.2943521.

[4] A. Stevanovic, Transportation Research Board, National Cooperative Highway Research Program Synthesis Program, and Transportation Research Board, *Adaptive Traffic Control Systems: Domestic and Foreign State of Practice*. Washington, D.C.: National Academies Press, 2010, p. 14364. doi: 10.17226/14364.

[5] E. Namazi, J. Li, and C. Lu, "Intelligent Intersection Management Systems Considering Autonomous Vehicles: A Systematic Literature Review," *IEEE Access*, vol. 7, pp. 91946–91965, 2019, doi: 10.1109/ACCESS.2019.2927412.

[6] A. S. Krishen, R. L. Raschke, P. Kachroo, M. Mejza, and A. Khan, "Interpretation of public feedback to transportation policy: a qualitative perspective," *Transportation journal*, vol. 53, no. 1, pp. 26–43, 2014.

[7] D. Li, Y. Zhang, and C. Li, "Mining public opinion on transportation systems based on social media data," *Sustainability*, vol. 11, no. 15, p. 4016, 2019.

[8] M. Arlington and C. Santa Ana, "Transportation Agency Self-Assessment of Data to Support Business Needs: Final Research Report," *Transportation Research Board: Washington, DC, USA*, 2015.

[9] S. Wandelt, C. Zheng, S. Wang, Y. Liu, and X. Sun, "Large language models for intelligent transportation: A review of the state of the art and challenges," *Applied Sciences*, vol. 14, no. 17, p. 7455, 2024.

[10] C. Chandra *et al.*, "Aviation-BERT: A preliminary aviation-specific natural language model," in *AIAA Aviation 2023 Forum*, 2023, p. 3436.

[11] O. Zheng, M. Abdel-Aty, D. Wang, Z. Wang, and S. Ding, "ChatGPT is on the horizon: Could a large language model be all we need for Intelligent Transportation," *arXiv preprint arXiv:2303.05382*, vol. 3, no. 4, p. 5, 2023.

[12] L. Lei, H. Zhang, and S. X. Yang, "ChatGPT in connected and autonomous vehicles: benefits and challenges," *Intelligence & Robotics*, vol. 3, no. 2. OAE Publishing Inc., pp. 144–147, 2023.

[13] G. Singh, "Leveraging chatgpt for real-time decision-making in autonomous systems," *Eduzone: International Peer Reviewed/Refereed Multidisciplinary Journal*, vol. 12, no. 2, pp. 101–106, 2023.

[14] Z. Yang, X. Jia, H. Li, and J. Yan, "Llm4drive: A survey of large language models for autonomous driving," *arXiv preprint arXiv:2311.01043*, 2023.

[15] L. Da, M. Gao, H. Mei, and H. Wei, "Llm powered sim-to-real transfer for traffic signal control," *arXiv preprint arXiv:2308.14284*, 2023.

[16] M. Villarreal, B. Poudel, and W. Li, "Can chatgpt enable its? the case of mixed traffic control via reinforcement learning," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 2023, pp. 3749–3755.
[17] T. Driessen, D. Dodou, P. Bazilinskyy, and J. De Winter, "Putting ChatGPT vision (GPT-4V) to the test: risk perception in traffic images," *Royal Society open science*, vol. 11, no. 5, p. 231676, 2024.
[18] E. Qasemi, J. M. Francis, and A. Oltramari, "Traffic-domain video question answering with automatic captioning," *arXiv preprint arXiv:2307.09636*, 2023.
[19] B. Mo, H. Xu, D. Zhuang, R. Ma, X. Guo, and J. Zhao, "Large language models for travel behavior prediction," *arXiv preprint arXiv:2312.00819*, 2023.
[20] S. Tan, B. Ivanovic, X. Weng, M. Pavone, and P. Kraehenbuehl, "Language conditioned traffic generation," *arXiv preprint arXiv:2307.07947*, 2023.
[21] I. A. Wong, Q. L. Lian, and D. Sun, "Autonomous travel decision-making: An early glimpse into ChatGPT and generative AI," *Journal of Hospitality and Tourism Management*, vol. 56, pp. 253–263, Sep. 2023, doi: 10.1016/j.jhtm.2023.06.022.
[22] K. Zhang, S. Wang, N. Jia, L. Zhao, C. Han, and L. Li, "Integrating visual large language model and reasoning chain for driver behavior analysis and risk assessment," *Accident Analysis & Prevention*, vol. 198, p. 107497, Apr. 2024, doi: 10.1016/j.aap.2024.107497.
[23] S. Lai, Z. Xu, W. Zhang, H. Liu, and H. Xiong, "LLMLight: Large Language Models as Traffic Signal Control Agents," 2023, *arXiv*. doi: 10.48550/ARXIV.2312.16044.
[24] S. Zhang *et al.*, "TrafficGPT: Viewing, processing and interacting with traffic foundation models," *Transport Policy*, vol. 150, pp. 95–105, May 2024, doi: 10.1016/j.tranpol.2024.03.006.
[25] J. Ma, J. Zhao, and Y. Hou, "Spatial–Temporal Transformer Networks for Traffic Flow Forecasting Using a Pre-Trained Language Model," *Sensors*, vol. 24, no. 17, p. 5502, Aug. 2024, doi: 10.3390/s24175502.
[26] X. Jing, A. Chennakesavan, C. Chandra, M. V. Bendarkar, M. Kirby, and D. N. Mavris, "BERT for aviation text classification," in *AIAA Aviation 2023 Forum*, 2023, p. 3438.
[27] S. R. Andrade and H. S. Walsh, "SafeAeroBERT: Towards a safety-informed aerospace-specific language model," in *AIAA AVIATION 2023 Forum*, 2023, p. 3437.
[28] T. Fontes, F. Murços, E. Carneiro, J. Ribeiro, and R. J. F. Rossetti, "Leveraging Social Media as a Source of Mobility Intelligence: An NLP-Based Approach," *IEEE Open J. Intell. Transp. Syst.*, vol. 4, pp. 663–681, 2023, doi: 10.1109/OJITS.2023.3308210.
[29] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.
[30] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.
[31] M. Grohe, "word2vec, node2vec, graph2vec, x2vec: Towards a theory of vector embeddings of structured data," in *proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI symposium on principles of database systems*, 2020, pp. 1–16.
[32] J. Cao, J. Fang, Z. Meng, and S. Liang, "Knowledge graph embedding: A survey from the perspective of representation spaces," *ACM Computing Surveys*, vol. 56, no. 6, pp. 1–42, 2024.
[33] Y. Han, C. Liu, and P. Wang, "A comprehensive survey on vector database: Storage and retrieval technique, challenge," *arXiv preprint arXiv:2310.11703*, 2023.

Appendix A: Few-shot prompts per class

```
FEWSHOT CATEGORY
<<SYS>>
You are a transportation engineer agent. Analyze one public comment about
traffic/transportation and classify it into CATEGORY from:
[Bridge, Pedestrian_Or_Bike, Roadway, Transit, Freight, Other].

- Definitions:
• Bridge: Bridges/overpasses/viaducts/bridge maintenance or new bridge requests
(including pedestrian bridges).
• Pedestrian_Or_Bike: Sidewalks, crosswalks, bike lanes, trails, micromobility—
without requesting a new bridge structure.
• Roadway: General roadway operations or minor maintenance (potholes, striping,
signage, turn lanes, traffic calming, roundabouts).
• Transit: Buses, bus stops, bus routes, schedules, shelters, rail transit
service.
• Freight: General complaints of operations for freights, trucks, rail.
• Other: Use if the comment is vague, irrelevant, or cannot reasonably be assigned
to the other categories.

- Decision policy:
1) If the comment explicitly requests/mentions a bridge/overpass/viaduct (e.g.,
"pedestrian bridge"), classify as Bridge (even if it benefits peds/bikes).
2) Otherwise, use clear cues:
  • "bus/stop/route/schedule/transit" → Transit
  • "sidewalk/crosswalk/bike/bikeway/trail/multi-use path" → Pedestrian_Or_Bike
  • "potholes/striping/signage/minor repair/turn lane/roundabout/traffic calming"
→ Roadway
  • "freights/trucks/rails/transport of goods/facilities" → Freight
3) If the comment is vague, pick the Other category.
5) Always output one Category which should be your first best guesse.

- Constraints
• Do not invent categories. Choose exactly one of: Bridge, Pedestrian_Or_Bike,
Roadway, Transit, Freight, Other.
• Be concise and deterministic.

- Output format (exactly three lines)
CATEGORY: <Bridge|Pedestrian_Or_Bike|Roadway|Transit|Freight|Other>
Reason: <1-2 short sentences citing the key phrase(s)>
<<END SYS>>
```

Figure 7. Few-shot prompt for category classification

FEWSHOT SUBCATEGORY

```
<<SYS>>
You are a transportation engineer classifying SUBCATEGORIES for provided CATEGORY.

- Decision policy:
  1) Extract key phrases (entities, facilities, actions, problems).
  2) Map phrases to candidate CATEGORIES.
  3) For each candidate CATEGORY, score each allowed subcategory by lexical cue
overlap and fit to the short definition.
  4) If the comment is vague, pick the Other subcategory.
  5) Always output one Subcategory which should be your first best guesse.

Constraints (must hold in the final lines):
- Do not invent categories. Choose exactly one of provided.
- Be concise and deterministic.
- Use exact label strings as provided (no synonyms, no typos).

<<END SYS>>

---

COMMENT: "The speed of travel through the residential and business district along
Route 30 through Forest HIlls is hazardous. The DOT timing of the lights to
expedite movement at speed from Turtle Creek area to the interstate highway poses
significant disincentives"

CANDIDATE_CATEGORIES AND ALLOWED SUBCATEGORIES:
- Pedestrian_Or_Bike
  * New Facilities: Bike lane needed, trail needed, bike or ped paths do not
connect well, school routes problem
  * Safety: unsafe pedestrian crossing, intersection crossing issues, signane
issues
  * Maintenance: sidewalk missing or in need of repair, existing bike pedestiran
facilities needs repair
  * Accessibility: people with disabilities, bike parking issues
  * Other: unclear or mixed

Output format (exactly these lines; no extra text):
CATEGORY: Pedestrian_Or_Bike
PRED_SUBCATEGORY: <name>
Reason: <1-2 short sentences citing the key phrase(s)>
 <END>
```

Figure 8. Few-shot prompt for subcategory classification

FEWSHOT FINAL DECISION

```
<<SYS>>
You are a transportation engineer agent. Analyze one public comment about traffic/transportation
and you need to give FINAL SUGGESTION from:
[POTENTIAL_PROJECT, REFER, DISMISS, Other].

Context:
- The predicted Category and Subcategory from the Category and Subcategory Agents.
- The agent must use this information—along with the text of the current comment—to provide the
final, most reasonable classification.

- Definitions:
• POTENTIAL_PROJECT: A comment proposing a major new construction or redesign requiring
significant planning or funding.
  - Examples: new lanes, intersections, bridges, bypasses, or major signal installations.
  - Describes large-scale or corridor-wide improvements addressing congestion, safety, or
capacity.
  - Mentions trucks, heavy traffic, or visibility issues requiring infrastructure expansion.

• REFER: A comment describing routine maintenance or minor local improvements handled by city,
county, or local agencies.
  - Examples: potholes, sign repairs, line painting, sidewalks, crosswalks, curb ramps, or
drainage fixes.
  - Small-scale pedestrian, bicycle, or transit safety improvements (e.g., bus stops, benches).
  - If a comment focuses mainly on pedestrian, bike, or transit issues → default to REFER.

• DISMISS: A comment that is vague, irrelevant, or lacks actionable content.
  - General complaints (e.g., "traffic is bad") or opinions without specific solutions.
  - Mentions already resolved or infeasible issues.
  - Spam, unrelated, or off-topic content.

• Other: Use only if the comment cannot reasonably fit any of the above categories or is unclear.

- Decision policy:
1) Assess clarity of the comment:
  - If the comment is vague, irrelevant, or purely opinion-based, classify as DISMISS.
  - If it clearly describes a transportation issue or proposal, continue.

2) Interpret content based on scope:
  - Major new construction, redesign, or large corridor-scale improvement → POTENTIAL_PROJECT.
  - Routine maintenance, local repair, or small-scale fixes (sidewalks, crosswalks, potholes,
signage, transit amenities) → REFER.
  - Unclear or non-actionable content → DISMISS.

3) Special case — Bridges:
  - If the comment explicitly mentions a bridge, overpass, or viaduct, classify as
POTENTIAL_PROJECT, unless it only refers to maintenance or repair, in which case → REFER.

4) Conflict resolution:
  - When unsure between POTENTIAL_PROJECT and REFER, choose REFER.
  - Routine work should never be labeled as POTENTIAL_PROJECT.
  - Always output 1st best decision.

- Constraints
• Do not invent categories. Choose exactly one of: POTENTIAL_PROJECT, REFER, DISMISS, Other.
• Be concise and deterministic.

- Output format (exactly two lines)
FINAL_SUGGESTION: <POTENTIAL_PROJECT|REFER|DISMISS|Other>
Reason: <1–2 short sentences citing the key phrase(s) and, if applicable, Support(C) margin>
<<END SYS>>
```

Figure 9. Few-shot prompt for final decision classification

Appendix B: Retrieval prompts per class

```
RETRIEVAL CATEGORY
<<SYS>>
You are a transportation engineer agent. Analyze one public comment about traffic/transportation and classify it into CATEGORY from:
[Bridge, Pedestrian_Or_Bike, Roadway, Transit, Freight, Other].

- Definitions:
• Bridge: Bridges/overpasses/viaducts/bridge maintenance or new bridge requests (including pedestrian bridges).
• Pedestrian_Or_Bike: Sidewalks, crosswalks, bike lanes, trails, micromobility—without requesting a new bridge structure.
• Roadway: General roadway operations or minor maintenance (potholes, striping, signage, turn lanes, traffic calming, roundabouts).
• Transit: Buses, bus stops, bus routes, schedules, shelters, rail transit service.
• Freight: General complaints of operations for freights, trucks, rail.
• Other: Use if the comment is vague, irrelevant, or cannot reasonably be assigned to the other categories.

- Decision policy:
1) If the comment explicitly requests/mentions a bridge/overpass/viaduct (e.g., "pedestrian bridge"), classify as Bridge (even if it
benefits peds/bikes).
2) If Top-K retrieved examples and a Support(C) summary are provided:
  • Treat Support(C) = summed weighted similarity per category as prior evidence.
  • If (top Support(C) - second Support(C)) ≥ 10% of total, treat the top category as strong prior evidence.
  • If the margin is smaller, apply rule (3) based on the comment's wording.
3) Otherwise, use clear cues:
  • "bus/stop/route/schedule/transit" → Transit
  • "sidewalk/crosswalk/bike/bikeway/trail/multi-use path" → Pedestrian_Or_Bike
  • "potholes/striping/signage/minor repair/turn lane/roundabout/traffic calming" → Roadway
  • "freights/trucks/rails/transport of goods/facilities" → Freight
4) If the comment is vague, pick the Other category.
5) Always output one Category which should be your first best guesse.

- Constraints
• Do not invent categories. Choose exactly one of: Bridge, Pedestrian_Or_Bike, Roadway, Transit, Freight, Other.
• Be concise and deterministic.

- Output format (exactly three lines)
CATEGORY: <Bridge|Pedestrian_Or_Bike|Roadway|Transit|Freight|Other>
Reason: <1-2 short sentences citing the key phrase(s) and, if applicable, Support(C) margin>
<<END SYS>>
---
New comment:
"I want a new pedestrian bridge over the highway."
---
Support by category (sum of effective similarities):
Pedestrian_Or_Bike=1.166 (77.9%) | Bridge=0.331 (22.1%)

55.76% >= 10%

Margin difference = True

Top-K evidence table:
[Rank | sim | eff_sim | CATEGORY | Snippet]
1 | 0.663 | 0.663 | Pedestrian_Or_Bike | "There are crosswalks here that need improvement and sidewalks are needed. People have
carved desired paths under the bridge. Many people wal"
2 | 0.662 | 0.331 | Bridge | "Bridge should have safe and accessible connection to Eliza Furnace Trail which runs directly
underneath north end of bridge but is not legal"
3 | 0.661 | 0.220 | Pedestrian_Or_Bike | "Protected bike lanes please or dont bother. Make them so that when not if WHEN a car makes
a mistake they cannot kill me with their car. Bol"
4 | 0.630 | 0.158 | Pedestrian_Or_Bike | "NEED DEDICATED BIKE LANE."
5 | 0.626 | 0.125 | Pedestrian_Or_Bike | "This could be a connection for pedestrians and bikers to safely travel to Mt Royal
cemetery and Scott elementary school.  Right now it is ru"

---
Return only in this exact format:
CATEGORY: ...
Reason: ...
```

Figure 10. Retrieval prompt for category classification

RETRIEVAL SUBCATEGORY

```
<<SYS>>
You are a transportation engineer classifying SUBCATEGORIES.

Do your reasoning on a hidden scratchpad:
- Extract key phrases (entities, facilities, actions, problems).
- Map phrases to candidate CATEGORIES.
- For each candidate CATEGORY, score each allowed subcategory by:
  (a) lexical cue overlap,
  (b) semantic match with EVIDENCE_TOPK snippets,
  (c) fit to the short definition.
- Apply SUPPORT_BY_CATEGORY only as a weak prior (use it to break ties or confirm).

Constraints (must hold in the final lines):
- You will receive ONE candidate CATEGORIY.
- SUBCATEGORY of provided CATEGORY must be DISTINCT and come from allowed list.
- If best match is outside the provided allowed lists, choose 'Other' within the most appropriate candidate category.
- Use exact label strings as provided (no synonyms, no typos).

Be concise and deterministic. Never add extra commentary outside the required lines.
<<END SYS>>
---
COMMENT: "The speed of travel through the residential and business district along Route 30 through Forest HIlls is hazardous. The DOT
timing of the lights to expedite movement at speed from Turtle Creek area to the interstate highway poses significant disincentives"

SUPPORT_BY_SUBCATEGORY:
Safety & Vulnerable Users=1.441 (79.7%) | Congestion=0.210 (11.6%) | Intersection & Signals=0.157 (8.7%)

68.13% >= 10%

Margin difference = True

EVIDENCE_TOPK:
[Rank | sim | eff_sim | CATEGORY | SUBCATEGORY | Snippet]
1 | 1.000 | 1.000 | Roadway | Safety & Vulnerable Users | "The speed of travel through the residential and business district along
Route 30 through Forest HIlls is hazardous. The DOT timing of the li"
2 | 0.632 | 0.316 | Roadway | Safety & Vulnerable Users | "An increasing amount of traffic accidents are occurring along this
corridor due to excessive speeding by drivers.  This is not safe for othe"
3 | 0.629 | 0.210 | Roadway | Congestion | "Street access offon an interstate highway causes traffic to slow and the flow to be
interrupted."
4 | 0.627 | 0.157 | Roadway | Intersection & Signals | "Intersection is always congested turning lanes are too short traffic signals
are confusing ramps entering and exiting Route 30 are too short"
5 | 0.627 | 0.125 | Roadway | Safety & Vulnerable Users | "Constant speeding headed past neighborhoods in Rebecca Dr  Change speed to
25 and patrol please"

CANDIDATE_CATEGORIES AND ALLOWED SUBCATEGORIES:
- Roadway
  * Intersection & Signals: Issues with Intersections, Traffic signals, Turn-lanes, Signange
  * Congestion: General problems with congestion and traffic, merging, roadways that do not connect well
  * Maintenance: Road in need of repair, drainange issues, shoulder issues, bridge closed, potwholes
  * Safety & Vulnerable Users: traffic accidents, speed limit concerns, bike/pedestrina access issues
  * Other: unclear or mixed

DECISION POLICY:
1) Rank the best and second-best subcategory for EACH listed category using the scoring rubric above.
2) If SUPPORT shows a ≥10% margin for one subcategory, treat it as a weak prior only—override it if text cues clearly favor a
different (Category :: Subcategory).
3) If ambiguous, use 'Other' within the best candidate category.

COMMENT: "The speed of travel through the residential and business district along Route 30 through Forest HIlls is hazardous. The DOT
timing of the lights to expedite movement at speed from Turtle Creek area to the interstate highway poses significant disincentives"
Return only in this exact format:
SUBCATEGORY: ...
Reason: ...

<END>
```

Figure 11. Retrieval prompt for subcategory classification

RETRIEVAL FINAL DECISION

```
<<SYS>>
You are a transportation engineer agent. Analyze one public comment about traffic/transportation and you need to give FINAL SUGGESTION from:
[POTENTIAL_PROJECT, REFER, DISMISS, Other].

Context:
- The predicted Category and Subcategory from the Category and Subcategory Agents.
- A Support Table containing the top 5 most similar historical comments from the database, including their assigned categories, subcategories, and
similarity scores.
- The agent must use this information—along with the text of the current comment—to provide the final, most reasonable classification.

- Definitions:
• POTENTIAL_PROJECT: A comment proposing a major new construction or redesign requiring significant planning or funding.
 - Examples: new lanes, intersections, bridges, bypasses, or major signal installations.
 - Describes large-scale or corridor-wide improvements addressing congestion, safety, or capacity.
 - Mentions trucks, heavy traffic, or visibility issues requiring infrastructure expansion.

• REFER: A comment describing routine maintenance or minor local improvements handled by city, county, or local agencies.
 - Examples: potholes, sign repairs, line painting, sidewalks, crosswalks, curb ramps, or drainage fixes.
 - Small-scale pedestrian, bicycle, or transit safety improvements (e.g., bus stops, benches).
 - If a comment focuses mainly on pedestrian, bike, or transit issues → default to REFER.

• DISMISS: A comment that is vague, irrelevant, or lacks actionable content.
 - General complaints (e.g., "traffic is bad") or opinions without specific solutions.
 - Mentions already resolved or infeasible issues.
 - Spam, unrelated, or off-topic content.

• Other: Use only if the comment cannot reasonably fit any of the above categories or is unclear.

- Decision policy:
1) Assess clarity of the comment:
 - If the comment is vague, irrelevant, or purely opinion-based, classify as DISMISS.
 - If it clearly describes a transportation issue or proposal, continue.

2) Incorporate evidence from Support(C):
 - Treat Support(C) = summed weighted similarity per final decision as prior evidence from knowledge table.
 - If (Top Support - Second Support) ≥ 10% of total Support, accept the top category as strong prior evidence.
 - If the margin is smaller, use step 3 (textual interpretation) to confirm or override.

3) Interpret content based on scope:
 - Major new construction, redesign, or large corridor-scale improvement → POTENTIAL_PROJECT.
 - Routine maintenance, local repair, or small-scale fixes (sidewalks, crosswalks, potholes, signage, transit amenities) → REFER.
 - Unclear or non-actionable content → DISMISS.

4) Special case — Bridges:
 - If the comment explicitly mentions a bridge, overpass, or viaduct, classify as POTENTIAL_PROJECT, unless it only refers to maintenance or repair,
in which case → REFER.

5) Conflict resolution:
 - When unsure between POTENTIAL_PROJECT and REFER, choose REFER.
 - Routine work should never be labeled as POTENTIAL_PROJECT.
 - Always output 1st best decision.

- Constraints
• Do not invent categories. Choose exactly one of: POTENTIAL_PROJECT, REFER, DISMISS, Other.
• Be concise and deterministic.

- Output format (exactly two lines)
FINAL_SUGGESTION: <POTENTIAL_PROJECT|REFER|DISMISS|Other>
Reason: <1-2 short sentences citing the key phrase(s) and, if applicable, Support(C) margin>
<<END SYS>>
---
New comment:
"I want a new pedestrian bridge over the highway."
---
Support by category (sum of effective similarities):
POTENTIAL_PROJECT=0.786 (53.0%) | REFER=0.698 (47.0%)

5.96% >= 10%

Margin difference = False

Top-K evidence table:
[Rank | sim | eff_sim | FINAL_DECISION | CATEGORY | SUBCATEGORY | Snippet]
1 | 0.663 | 0.663 | POTENTIAL_PROJECT | Pedestrian_Or_Bike | Sidewalk_Missing_Or_In_Need_Of_Repair | "There are crosswalks here that need improvement
and sidewalks are needed. People have carved desired paths under the bridge. Many people wal"
2 | 0.662 | 0.331 | REFER | Bridge | Bike_Pedestrian_Access_Issue | "Bridge should have safe and accessible connection to Eliza Furnace Trail which
runs directly underneath north end of bridge but is not legal"
3 | 0.630 | 0.210 | REFER | Pedestrian_Or_Bike | Bike_Lane_Needed | "NEED DEDICATED BIKE LANE."
4 | 0.626 | 0.157 | REFER | Pedestrian_Or_Bike | Unsafe_Pedestrian_Crossing | "This could be a connection for pedestrians and bikers to safely travel
to Mt Royal cemetery and Scott elementary school.  Right now it is ru"
5 | 0.617 | 0.123 | POTENTIAL_PROJECT | Roadway | Bike_Pedestrian_Access_Issue | "Traffic circle going in which is awesome and should help to mitigate
stop and go traffic at the 3 way stop now. Would appreciate a sidewalk "

Return only in this exact format:
FINAL_SUGGESTION: ...
Reason: ...
```

Figure 12. retrieval prompt for final decision classification

Appendix C: Full List of Category-Level Mismatch Comments

Bridge to Roadway
``Road resurfacing."
``clean out roadside drainage ditches"
``It is unclear which lane in the road leads to which ramp. This causes last second lane changes which are dangerous. Lanes should be clearly designated for each exit Bigelow Blvd 579 bridge and Bedford Ave."
``Add an overpass over the PA Turnpike to add a secondary entrance to Graham Park in Cranberry Township. This would alleviate traffic congestion on Rochester Rd."
``Every rainstorm causes flooding in this area. I have hydroplaned multiple times."
``This road is horrible. It is always the last for snow removal. It carries more traffic than ever and the intersections are dangerous."
``Must find a way to reduce traffic only because of the tunnel slow down."
``Clogged catch basins on Route 19 in Peters across from South Hills Subaru."
``Pedestrian/bike path too narrow. South bound cyclists end up on W. Carson Street with no bike lanes."
Pedestrian/Bike to Roadway
``West Carson St is the primary link between downtown Pittsburgh and the West End. There are no adequate bike facilities. Traffic frequently exceeds 35 mph."
``Route 40 from Hopwood to the Summit is a popular cycling route but dangerous due to speeding cars."
``Pedestrian killed while crossing McKnight Rd. Please lower speed limit and redesign road."
``Dangerous and confusing 5-way intersection where vehicles drive too fast."
``Although there are lights, the pedestrian crossing is often missed."
``South Mount Vernon corridor is too wide, poorly marked, and promotes unsafe speeds."
``The sidewalk along Mossfield is poorly maintained, flooding and debris."
``Lack of safe pedestrian infrastructure along east–west roads in the North Hills suburbs."
Roadway to Pedestrian/Bike
``Blind corner on steep hill makes it dangerous for pedestrians, bikes, and drivers alike."
``Very dangerous pedestrian crossing at Centre and Highland; left-turning cars do not see pedestrians."
``Too many cars in this section; pedestrianization of the South Side recommended."
``Four-lane high-speed road segment is unsafe for bikers and pedestrians."
``People must walk in the vehicle lanes since roads have no berm; need a widened shoulder or bike path."
``Increase signage and enforcement due to illegal shoulder passing injuring pedestrians."
``Long crosswalks with no four-way stop make the intersection dangerous for pedestrians."
Transit to Roadway
``Traffic moves fast and things get dangerous; consider active features to control traffic better at this location."
``High pedestrian traffic conflicts with high car traffic; inadequate pedestrian and transit infrastructure."
Transit to Pedestrian/Bike
``Many stations on this transit line are not ADA accessible and not easily accessible by pedestrians."

``Provide Park and Ride access for UPMC Mercy Hospital; no current transit options."
``This corridor is extremely dangerous. High pedestrian volume conflicts with car traffic; pedestrian space is inadequate."
``Stations are poorly accessible; signage is unclear and pedestrian facilities are lacking."