Recall Recon – Team Brainwave

Final Project Report

submitted by

Kaushik Mantha - (kaushik.mantha@colorado.edu)

Lalith Sai Reddy Kaamala - (laka2272@colorado.edu)

Pavan Kumar Malasani - (pama4195@colorado.edu)

Preethu Nath Manjunath - (prma3214@colorado.edu)

Sarthak Katyal - (sarthak.katyal@colorado.edu)

99P Labs Honda

Supervisors

Ryan Lingo - (ryan_lingo@honda-ri.com)

Rajeev Chhajer - (rajeev_chhajer@honda-ri.com)

Martin Arroyo - (martin_arroyo@honda-ri.com)

Table of Contents

Abstract

Recall Recon is our response to a recurring challenge in the automotive world: how do we spot safety risks before they become full-blown recalls? Built using machine learning and natural language tools, Recall Recon is designed to forecast future vehicle recalls and make safety data more accessible to everyone, from manufacturers and regulators to curious drivers. By analyzing historical data from the National Highway Traffic Safety Administration (NHTSA), our system uses a Random Forest model to identify patterns and predict recall volumes, helping stakeholders shift from reactive responses to preventive action.

But prediction alone isn't enough. To make the insights truly usable, we integrated a Retrieval Augmented Generation (RAG) chatbot that allows users to ask natural questions, like "Which components had the most recalls in 2020?", and get fast, accurate, and grounded answers. The system combines OpenAI embeddings, FAISS for semantic search, and GPT-3.5 Turbo to deliver responses that are both informative and easy to understand, even for those without a technical background.

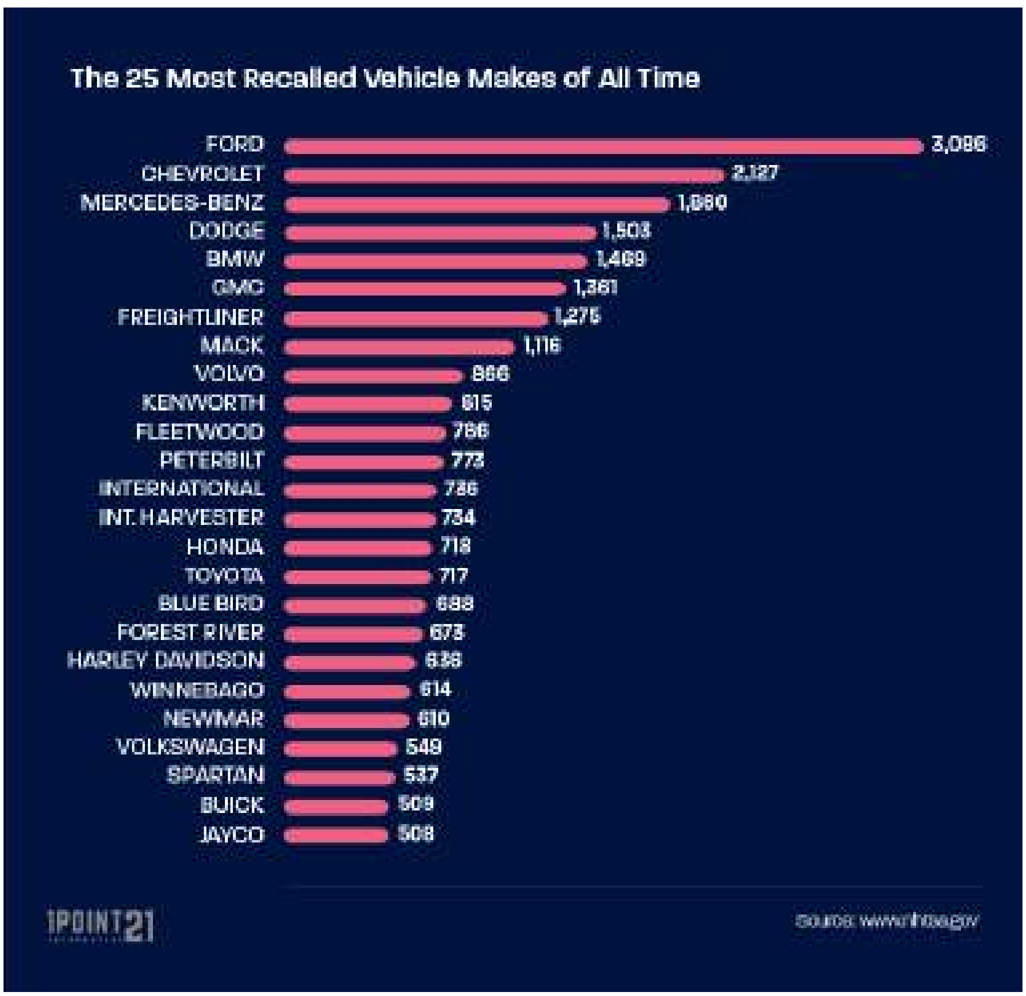
We also developed interactive dashboards using Streamlit and Power BI that bring recall data to life. These dashboards let users explore trends over time, compare risks across components, and drill down into manufacturer specific insights, all through a clean and responsive interface.

Together, the forecasting model, conversational chatbot, and visual dashboards make Recall Recon a powerful, all in one platform for smarter recall management. By making recall data easier to explore and act on, we hope this system contributes to better safety decisions, and fewer surprises on the road.

1. Introduction

1.1 Background and Importance of Recalls

The National Highway Traffic Safety Administration (NHTSA) states that once a car or piece of associated technology is found to have an inherent risk or to not meet requirements for security, a formal alert is handed out. Car safety depends on those recalls given that they alert automakers and the public to significant problems that may result in accidents or injuries if ignored. The automotive industry has seen a significant increase in recall activity in the recent past. For example, the number of banned vehicles in the United States has more than doubled to over 60 million in 2014. The Takata airbag recall, which affected over 67 million airbag inflators spanning tens of millions of automobiles, was one of the biggest recalls in history (NHTSA | National Highway Traffic Safety Administration). Large scale recalls like this demonstrate how pervasive and significant car flaws may be for both manufacturer liability and consumer safety.



Safety recalls such as the Takata airbag recall have affected millions of automobiles, highlighting the scope of the problem (Takata's lethal airbags have affected nearly 67 million cars in the most costly recall in history | Fox Business).

1.2 Motivation for a Predictive Approach

Recalls are important because they have a direct bearing on public safety. Defective parts, like airbags, ignition switches, or brakes, can immediately endanger drivers and passengers. Early detection and resolution of these problems can avert mishaps and save lives. On the other hand, it can be disastrous if recalls are handled poorly or with delays. (NHTSA Announces App to Alert Public to Vehicle Recalls)

Research indicates that a considerable portion of recalled vehicles, typically about 30% remain unrepaired on the road, posing a risk to public safety. For this reason, automakers work to preserve customer confidence by effectively managing recalls, and regulators enforce recall compliance. Beyond safety, recalls affect manufacturers' finances and reputations. Repair operations can cost millions of dollars, and they can occasionally result in fines or a decline in customer trust.

Machine learning and Artificial Intelligence provide strong instruments to anticipate and lessen car recalls before they become more serious. Recalls are often reactive, meaning they are only carried out in response to failures or a high volume of complaints. We can identify patterns suggestive of possible flaws by using machine learning to historical memory data and associated variables. AI models might, for example, link specific sensor data or warranty repair rates to subsequent recall events.

In order to predict how many cars may require repairs in the upcoming years, machine learning may also simulate recall trends across time. Stakeholders may shift from a reactive to a proactive posture thanks to these predictive capabilities. AI may essentially serve as an early warning system for car flaws, spotting problems early, recommending preventative maintenance, and ultimately lowering the likelihood of accidents or widespread recalls.

1.3 Project Goals and Objectives

Recall Recon's primary mission is to improve car safety through the use of predictive analytics. With predicting recall patterns and discovering high risk problems early, taking charge gives stakeholders the capability to take proactive steps, which shortens the period that problematic cars are on highways and lowers the chance of incidents. This change in recall management strategy from an emergency response to a predictive methodology signifies an important shift in the way that auto security is addressed.

The project's main objectives include:

- Predicting Recall Events: Recall Recon predicts possible safety recalls by examining past trends, providing firms with an advantage in risk mitigation.
- Optimizing Resource Allocation: By predicting demand for repair services and replacement components, the system helps suppliers and OEMs plan ahead and react more quickly.
- Giving Stakeholders Knowledge and Power: Both technical and non-technical users may access and use recall data through interactive dashboards and a natural language query interface.
- Cutting Down on Repair Delays: Early detection reduces the amount of time that customers spend operating potentially dangerous cars by enabling prompt notice and intervention.
- Enhancing Public Safety and Trust: The project's ultimate goals are to reduce accident rates and boost public trust in the car recall system.

Recall Recon benefits several different stakeholders:

- Tools for early risk detection and enforcement preparation are provided to regulators (like the NHTSA).
- Automobile manufacturers can improve quality assurance and expedite maintenance procedures.
- Customers gain from the early identification and fixing of defective cars.
- Dealers and suppliers are informed ahead of time when components are needed to reduce delays and shortages.

By providing early recall detection, more intelligent campaign planning, and faster reaction times, Recall Recon has the potential to significantly improve vehicle safety and reduce the human and financial costs associated with delayed recalls. It encourages a more secure, transparent, and dependable automotive ecosystem over time.

2. Analysis

2.1 Dataset Overview

The National Highway Traffic Safety Administration (NHTSA) provides a comprehensive collection of automobile safety recalls, which Recall Recon uses. Vehicles, tires, child safety seats, and automotive equipment are among the categories covered by the data, which is updated daily & includes recall information from 1949 to the present. The extensive and comprehensive NHTSA recall dataset provides a wealth of information for study. As of March 2025, about 70,000 individual recollection recordings are included in the collection.

The manufacturer, the system or component impacted (airbags, brakes, etc.), the number of possibly impacted cars, and descriptions of the flaw and remedial measures are all included in each entry in the dataset, which corresponds to a particular recall campaign. The NHTSA API makes the data accessible and enables queries by vehicle make, model, year, or the distinct NHTSA campaign ID for every recall. In order to see past patterns, we combined the recall data by year and other factors.

The number of automobiles impacted by recalls has changed significantly over time, according to historical statistics, which is one obvious finding. Relatively few cars were recalled in certain years, while tens of millions of cars were recalled in others because of serious flaws. The total number of automobiles recalled year in the United States from 1990 to 2024 is depicted in Figure 1 below, with notable increases during key recall events:

Figure 1 shows the total number of cars recalled from 1990 and 2024. Take note of the peak in 2014, when the Takata airbag issue and other significant campaigns led to the recall of over 60 million cars (U.S. Auto Recalls Set Record at 60 Million for Year).

Significant recall activity surges can be seen in the data during 1999–2000 and again in the middle of the 2010s. Events such as the Ford/Firestone tire recall contributed to the early 2000s surge, while the Takata airbag recall and a string of significant General Motors recalls are responsible for the mid 2010s spike. These trends highlight how recalls are episodic; while totals are often significantly smaller, large scale faults can occasionally result in massive spikes in recall volume. We developed predictive models as a result of these data insights because we believe that stakeholders would be better prepared if we can forecast these spikes or the baseline recall rate.

Figure 1

2.2 Data Cleaning and Preprocessing

To guarantee analytical accuracy, extensive data cleaning and preparation procedures were carried out prior to modeling. There were several missing numbers and discrepancies in the raw NHTSA recall data that needed to be fixed. We started by fixing the missing data in important fields. Some recall records, for instance, did not specify the number of “potentially affected” automobiles. We used statistical imputation to fill up these gaps because this quantity is essential for predicting parts demand.

Specifically, we inserted the field's median value (within a pertinent subset) for missing entries. This method avoids using outliers to distort the data while maintaining a respectable central trend. To maintain consistency, we also standardized category variables; for example, “air bags” were made consistent.

Data cleaning example, handling missing values: The recall entry for a certain manufacturer had an unknown number of affected vehicles (“NaN”), which we filled with the median (~685) to maintain data completeness.

Another crucial aspect of data preparation was the identification and handling of outliers. In the context of specific analysis, we identified recall campaigns involving tens of millions of impacted automobiles as outliers. Instead of eliminating these valid but extreme cases (since they reflect actual occurrences such as the Takata recall), we carefully addressed them in the modeling process.

For example, we used models that are robust to outliers or we transformed the data using logarithmic scales to ensure that the training of the models would not be unduly skewed by these large values. To improve the prediction ability of the model, we additionally developed new feature variables. The recall completion rate, or the proportion of impacted cars that were actually fixed, was one such designed attribute that may be related to later recall announcements or enforcement actions.

Sorting recollections by component category, for example, combining different engine related recalls, to see whether any particular categories exhibit cyclical recall patterns was another derived feature. A part was reserved for model training as opposed to testing, and all cleaned and enhanced data was saved in an organized manner that was prepared for analysis.

	component	recall_type	corrective_action	subject	completion_rate	manufacturer	recall_link	report_received_date	do_not_drive	fire_risk_when_parked	defect_summary	consequence_summary
0	VISIBILITY	Vehicle	Dealers will update the software, free of char...	Front Windows May Fail to Reverse Direction	-1.0	Ford Motor Company	Cuff: https://www.nhtsa.gov/recalls?nhtsaId...	2025-03-14	No	No	Ford Motor Company (Ford) is recalling certain...	A window that does not automatically reverse w...
1	ELECTRICAL SYSTEM	Vehicle	Dealers will recalibrate the body control modu...	Loss of Drive Power from Undetected Low Battery	-1.0	Ford Motor Company	Cuff: https://www.nhtsa.gov/recalls?nhtsaId...	2025-03-14	No	No	Ford Motor Company (Ford) is recalling certain...	An undetected low battery charge can result in...
2	BACK OVER PREVENTION	Vehicle	Dealers will update the Image Processing Modul...	Rearview Camera Image May Not Display	-1.0	Ford Motor Company	Cuff: https://www.nhtsa.gov/recalls?nhtsaId...	2025-03-14	No	No	Ford Motor Company (Ford) is recalling certain...	Loss of the rearview camera image can reduce t...
3	BACK OVER PREVENTION	Vehicle	Dealers will update the rearview camera softwa...	Black or Distorted Rearview Camera Image	-1.0	Ford Motor Company	Cuff: https://www.nhtsa.gov/recalls?nhtsaId...	2025-03-14	No	No	Ford Motor Company (Ford) is recalling certain...	A rearview camera that displays a blank or dis...
4	ELECTRICAL SYSTEM	Equipment	Zoox has updated the ADS software, free of cha...	ADS Software May Cause Unexpected Hard Braking	-1.0	Zoox, Inc.	Cuff: https://www.nhtsa.gov/recalls?nhtsaId...	2025-03-18	No	No	Zoox, Inc. (Zoox) is recalling certain Autonom...	Unexpected hard braking can increase the risk ...

mfr_campaign_number	nhtsa_id	potentially_affected	potentially_affected_log
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]

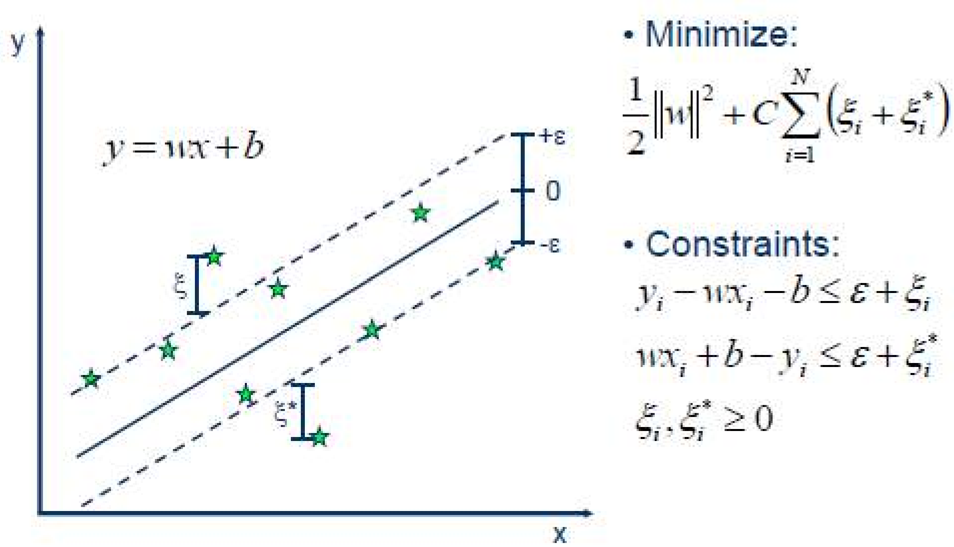
2.3 Modeling Techniques

2.3.1 Explored Models

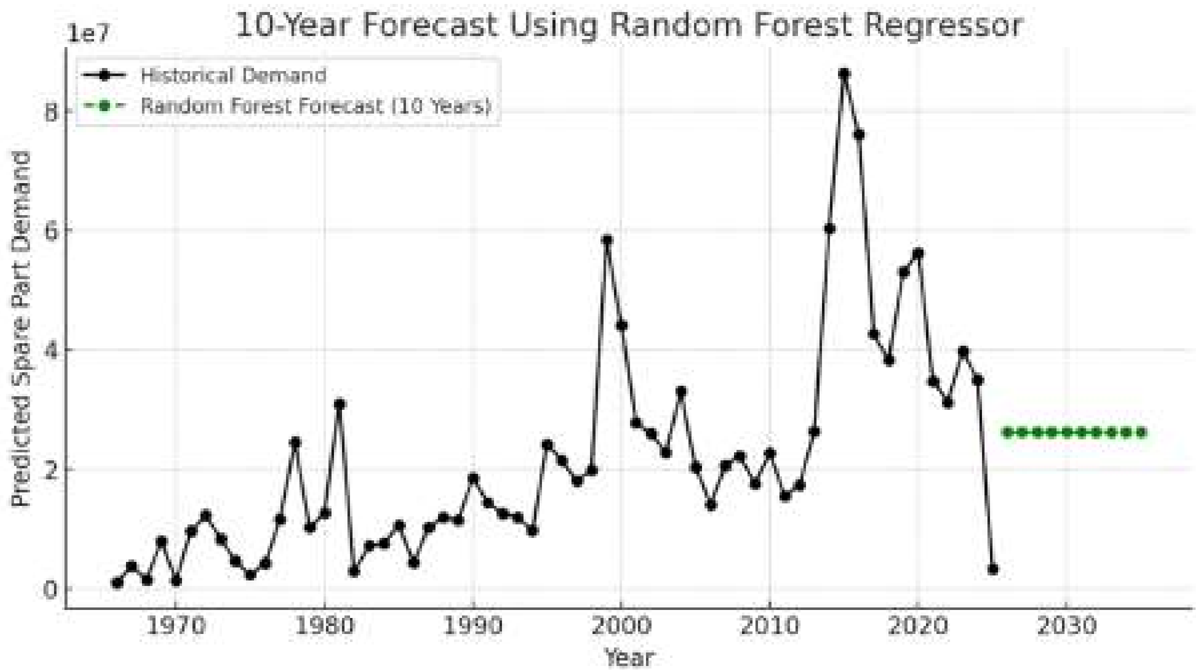
We looked at both time series and regression models for forecasting recall activity since recall data is time series (with trends and annual averages) and we wanted to anticipate the future. Four preliminary modeling techniques were examined:

- A time series forecasting model that may identify trends and seasonal patterns is called **SARIMA (Seasonal AutoRegressive Integrated Moving Average).** SARIMA was applied to the yearly recall counts. But there was too much variation in the findings; the model tended to overfit the big spikes, producing erratic predictions that weren't very generalizable. Without other exogenous variables, pure time series modeling was difficult due to the erratic incidence of mega recalls.

- A straightforward regression that uses time (year), maybe other artificial characteristics (such as delayed recall counts) as predictors is called a **linear regression.** Although this approach was simple to understand, it produced poor accuracy in our situation. The residuals displayed patterns, suggesting that the linear model was a poor match since it was unable to capture the non-linear spikes in the data.
- Using kernel methods, **Support Vector Regression (SVR)** is a machine learning regression that can identify non-linear correlations. On the recall data, we experimented with an SVR; although it outperformed linear regression, it tended to overfit when complicated kernels were permitted. It could fit the training data with careful adjustment, but it was unable to make accurate predictions for subsequent years, most likely because of the high variability and small number of data points (annual averages).



- An ensemble tree-based model that can simulate non-linear effects and is comparatively resistant to outliers is **Random Forest Regression**. Out of all our testing, the Random Forest produced the best results. When projecting recall volumes, it had the lowest error, with a Mean Absolute Error (MAE) of around 4.7 million cars. By simply averaging over several decision trees, Random Forests effectively managed the recall spikes without assuming a particular functional structure for the trend.



We chose the Random Forest Regressor as our main prediction model in light of these results. We used it to predict the future ten years of recall activity (through 2030) after training it on historical recall data up to the most recent years. According to the model's prediction, there will likely be a significant amount of recall activity over the next 10 years, affecting tens of millions of vehicles annually. A high baseline of recalls and a consistent need for spare components to perform recall repairs should be anticipated by manufacturers.

Because each recalled car frequently needs replacement components (for instance, millions of airbag inflators in the Takata case), we interpret the Random Forest's output as a forecast of spare part demand connected to recalls. It may be possible to guarantee that many replacement parts are available at dealerships and service facilities by planning for an anticipated recall volume of, say, about 40 million cars in a given year.

2.4 Intelligent Recall Query System

Retrieval Augmented Generation (RAG) in Recall Recon

In addition to traditional forecasting, Recall Recon integrates a sophisticated Retrieval Augmented Generation (RAG) system to support dynamic, intelligent question answering over large volumes of recall records.

2.4.1 RAG Workflow in Recall Recon

1. **Document Preprocessing and Embedding**
   All textual fields from the recall dataset (e.g., summary, consequences, corrective action) were embedded using OpenAI's "text-embedding-ada-002" model. This converts each entry into a semantic vector, capturing meaning beyond keywords.
2. **Vector Index Creation with FAISS**
   FAISS (Facebook AI Similarity Search) was used to build an efficient vector index for fast similarity search. This enables sub second retrieval of the most semantically similar recall records for a given query.
3. **User Query Embedding**
   When a user inputs a question (e.g., "What were the major Honda recalls last year?"), the same OpenAI embedding model is used to convert the query into a vector.
4. **Retrieval and Contextual Prompt Building**
   The query vector is matched against the FAISS index, and the top-k similar recall records are selected. These are concatenated into a prompt, optimized to stay within token limits while maintaining high context relevance.
5. **Answer Generation with GPT-3.5 Turbo**
   This prompt and the user's original question are passed to OpenAI's gpt-3.5 turbo model to synthesize a grounded, fluent answer.

2.4.2 Technical Considerations

- **Truncation Management:** Recall records are truncated to ~3,000 characters to prevent token overflow.
- **Re-ranking:** Retrieved chunks are optionally re-ranked to prioritize semantic relevance.
- **Grounding:** The system emphasizes returning traceable recall IDs and manufacturer names to maintain accuracy and attribution.

2.4.3 Benefits of the RAG System

- **Accuracy:** Answers are data grounded rather than hallucinated.
- **Flexibility:** Capable of handling a wide variety of queries with minimal rules.
- **Scalability:** New data can be embedded and indexed without retraining.

- **User Experience:** Natural, conversational answers paired with linked recall records.

2.5 Tools and Technologies Used

The development of Recall Recon required integrating multiple technologies for data handling, modeling, semantic search, and interface development:

Category	Tools / Frameworks
Data Handling	Python, Pandas, NumPy
Modeling	Scikit-learn (Random Forest, SVR, Linear Regression), Statsmodels (SARIMA)
Semantic Embedding	OpenAI API (text-embedding-ada-002)
Retrieval	FAISS (Facebook AI Similarity Search)
NLP Generation	OpenAI API (GPT-3.5 Turbo)
Dashboard	Streamlit
Visualization	Matplotlib (static plots), Plotly (interactive visualizations)
Backend	Flask, Django (for prototyping API interactions)
Frontend	HTML/CSS/JS, AJAX, Plotly.js
Data Source	NHTSA API (automated retrieval of up-to-date recall data in JSON format)
Experimentation	Jupyter Notebooks (analysis, visualization, RAG prototyping)

Together, these tools supported the full stack of Recall Recon from raw data ingestion and modeling to intelligent querying and real time user interaction via the dashboard.

2.6 Interactive Dashboard Exploration

We created two interactive Power BI dashboards in addition to the RAG based chatbot so that stakeholders could examine trends, risk profiles, and fault patterns in car recall data. These dashboards offer a graphic representation of component specific hazards, recall frequency over time, and the distribution of safety critical flaws among manufacturers and car models.

The first dashboard, which focuses on recall risk data, provides views like fire risk recalls, "Do Not Drive" crucial notifications, and the overall number of afflicted units. It contains time series visualizations of incident frequency broken down by fire hazard categorization, as well as a Recall Risk Matrix that illustrates the link between manufacturers and components. Additionally, users may examine trends in possibly impacted units and delve into event numbers by year.

Defect frequency and recall type distribution are highlighted in the second dashboard. It displays the most typical elements linked to extensive recall campaigns and divides recalls into divisions such as vehicle, equipment, tire, and child seat. A vertical bar chart shows the most prevalent high risk components, while a pie chart shows the percentage of various recall kinds. When used in tandem, these resources assist prioritize areas for inquiry or remediation and offer context for data driven safety choices.

These dashboards are especially helpful for analysts and decision-makers looking for quick, visual recall knowledge since they enable pattern discovery at scale, which enhances the chatbot interface.

2.7 NLQ Interface Development

As part of our project, we used Streamlit to create a dashboard web application that demonstrated the sophisticated recall analysis capabilities of Recall Recon. Without any technical knowledge, stakeholders may engage in conversational interactions with recall insights using the dashboard's user friendly, browser accessible interface. It is the primary gateway to our Retrieval Augmented Generation (RAG) technology and predictive forecasting tools.

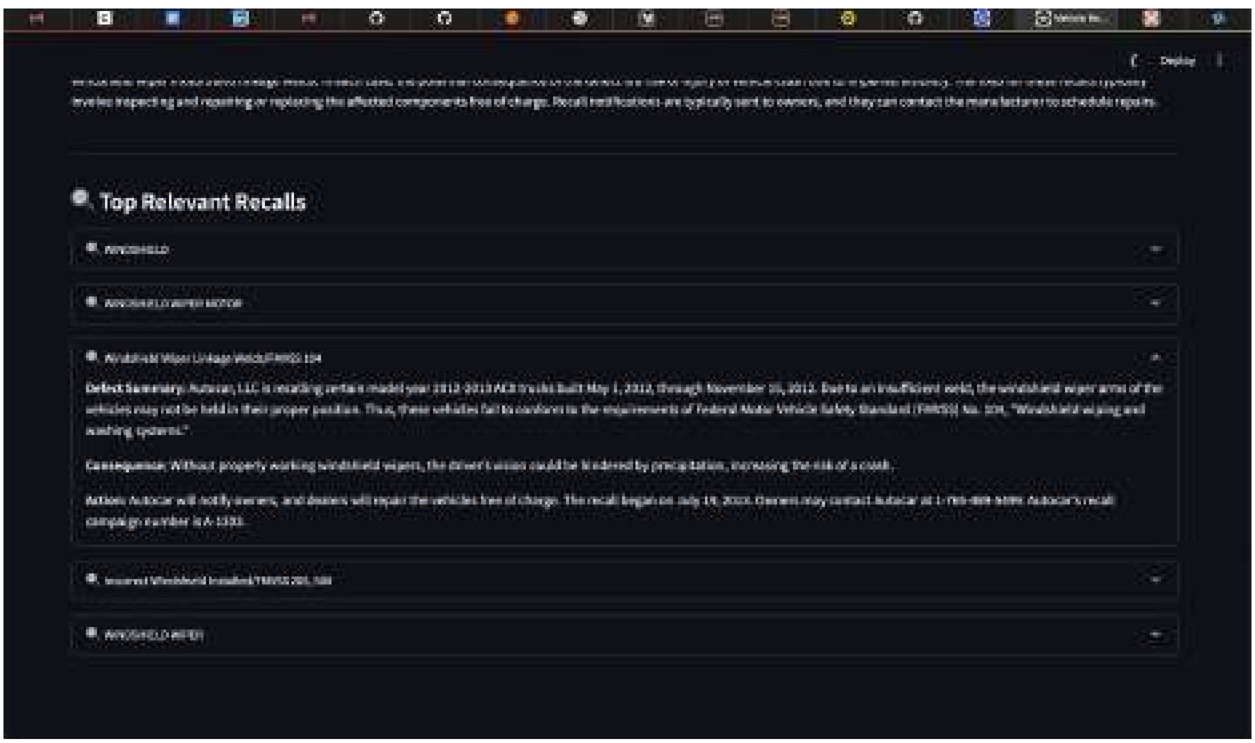
- **Interactive Visualizations:** The dashboard displays visual summaries, such as category specific breakdowns and memory patterns over time, on the homepage and left panel. These charts change in real time in response to human input. For instance, a bar chart highlights the components with the highest recall frequency, while a line chart displays the overall number of yearly recalls since 1960. Plotly is used to render all charts for clarity and interactivity.
- **Natural Language Chat Interface:** A chat style input box at the application's core enables users to pose queries in natural language. This is what our RAG pipeline uses as its front end. For example, when a user enters in "What are the top recalls for airbags?"

The system utilizes GPT-3.5 Turbo to provide a response after utilizing OpenAI embeddings to retrieve the most pertinent recall materials using FAISS. Directly beneath the input field, the response is shown along with any related links or summaries.

We gave a responsive and user friendly design top priority during development. The dashboard's usefulness was confirmed when even non-technical team members were able to use it with minimal instruction. The interface is appropriate for usage in both field and office environments since it adapts well across devices, including tablets and laptops.

We used Streamlit's cloud hosted environment for local development and testing, with version control enabled through a.env configuration. To keep efficiency high, the dashboard leverages cached data pipelines and retrieves cleansed NHTSA data. The architecture allows for future automation through scheduled processes or API hooks, even if new data uploads are presently completed manually.

All things considered, the dashboard serves as the main entry point to memory Recon's analytics, giving stakeholders a clever and interesting approach to examine past patterns, comprehend memory threats, and get AI generated insights all via a single, conversational platform.

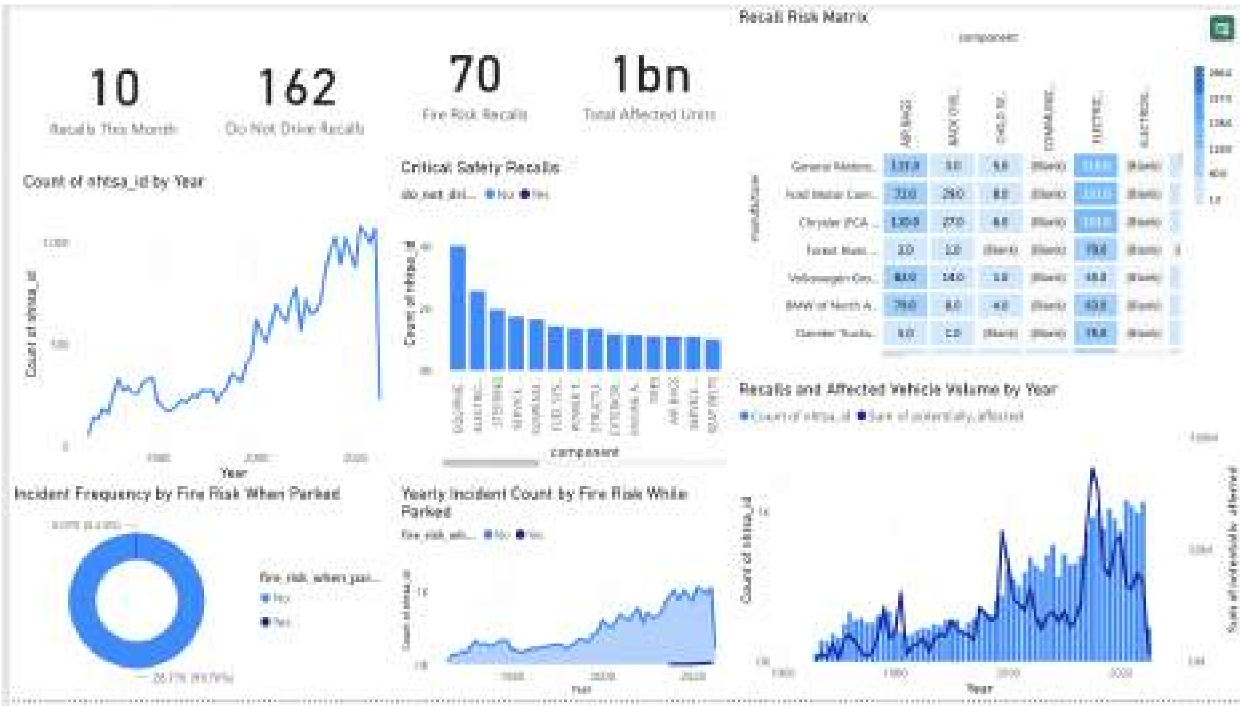


3. Power BI Dashboards for Recall Insights

In addition to the AI driven chatbot interface, our team developed two comprehensive dashboards using Microsoft Power BI to support interactive exploration and real time visual analysis of recall patterns. These dashboards serve as standalone tools designed to help analysts, regulators, and decision makers derive insights from the historical recall data provided by NHTSA.

The first dashboard is particularly valuable for automotive retailers and repair centers, such as Discount Tires, as it clearly highlights critical recall trends and specific components that require attention. By identifying “Do Not Drive” and fire risk recalls early, retailers can proactively gauge the seriousness and urgency of customer needs. This enables businesses to strategically anticipate demand for specific spare parts, especially those frequently recalled, like electrical systems, airbags, and steering components.

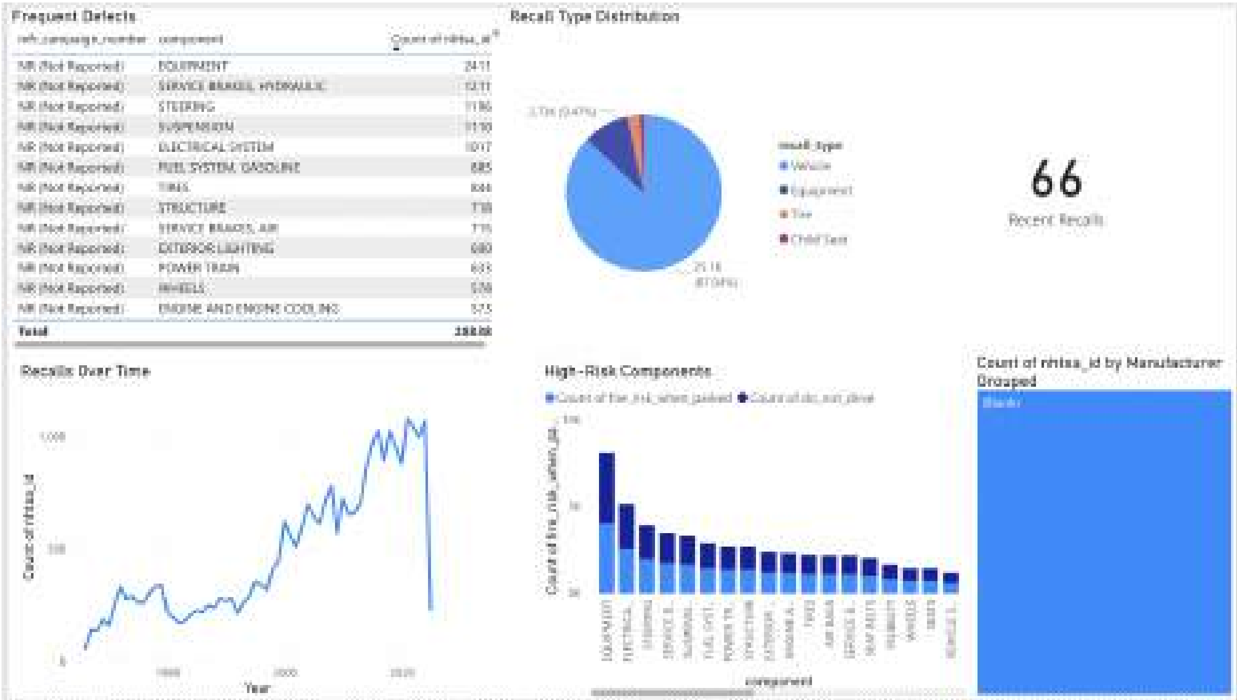
Additionally, understanding manufacturer specific recall patterns helps retailers maintain optimal inventory levels and avoid shortages during peak recall periods. Ultimately, having access to these insights allows service providers to plan ahead effectively, order necessary spare parts in advance, train technicians on expected repairs, and enhance overall service efficiency and customer satisfaction.



The image is an automotive recall dashboard for retailers laid out in a clean, two column grid. It combines high level key performance indicators (KPIs) with detailed visualizations that track recall volume, safety critical themes, manufacturer risk exposure, and historical impact on vehicles. ([Retail Dashboard])

This dashboard adds substantial value to car manufacturers such as Honda by allowing them to precisely monitor and effectively respond to vehicle recalls. Honda can take proactive measures to solve the underlying causes of recurring faults in their current models by clearly identifying "Equipment", "Hydraulic Brakes", "Steering", and "Electrical Systems", among other defect categories. The category of the pie chart emphasizes that 87% of recalls are vehicle based, highlighting the significance of comprehensive vehicle level quality control procedures. Furthermore, Honda can identify precisely which parts present serious safety risks, such as those linked to fire hazards or crucial operational failures, by using the bar chart that highlights high risk components.

Honda can promptly identify patterns in faults by monitoring recall trends over time. This allows them to prioritize engineering changes and technology advancements for upcoming model releases. Ultimately, this targeted approach enables Honda to enhance quality assurance processes, strategically allocate resources toward critical problem areas, and strengthen consumer trust through safer, more reliable vehicles.



The Manufacture dashboard presents automotive recall data, highlighting frequent defects, recall type distribution dominated by vehicle recalls, trends over time, and key high risk components such as equipment and electrical systems. ([Manufacture Dashboard](#))

In conclusion, both dashboards provide critical insights for automotive manufacturers and service retailers. The dashboards identify recurrent problems in parts like machinery, electrical systems, and brakes, and show a noticeable increase in car recalls over the past few decades. Retailers like Discount Tires can use this information to proactively predict the requirement for technician training and spare parts. Similar to this, automakers like Honda can see exactly which parts are causing issues, which allows for focused fixes and raises the caliber of the final product in later car releases. In the end, by strategically using these dashboards, stakeholders may effectively manage recalls, give priority to safety critical concerns, and preserve customer trust by taking prompt remedial action and proactive planning.

4. Results (Technical Findings)

4.1 Predictive Model Outcomes

We created prediction algorithms to estimate the number of cars that would probably be impacted by upcoming recalls in order to facilitate proactive vehicle safety assessments. The Random Forest model performed the best after a number of methods were evaluated, including SARIMA, Linear Regression, and Support Vector Regression. This is because it can handle large fluctuations brought on by recall events of a large scale, such as the Takata airbag case, and model non-linear trends. When tested against unknown years, our Random Forest regressor, which was trained on historical recall data, produced a mean absolute error (MAE) of around 4.7 million cars.

According to forecasting findings, the number of vehicle recalls will continue to be high over the next 10 years, with an estimated 40–50 million units impacted yearly until 2034. Our model forecasts a general rising trend in the overall number of cars impacted, reflecting both greater complexity in modern vehicles and heightened regulatory supervision, even though volatility is predicted, especially when abrupt regulatory changes or component failures occur.

To enhance the accessibility and interpretation of recall data, we implemented a Retrieval Augmented Generation (RAG) system in addition to forecasting. OpenAI's embeddings and FAISS vector search, along with GPT-3.5 Turbo for generation, have made it possible for people to query the recall database in normal language and obtain precise, fact based answers. The conversational retrieval and predictive modeling elements work together to provide a complete solution for contemporary, proactive vehicle safety management.

4.2 RAG System Performance

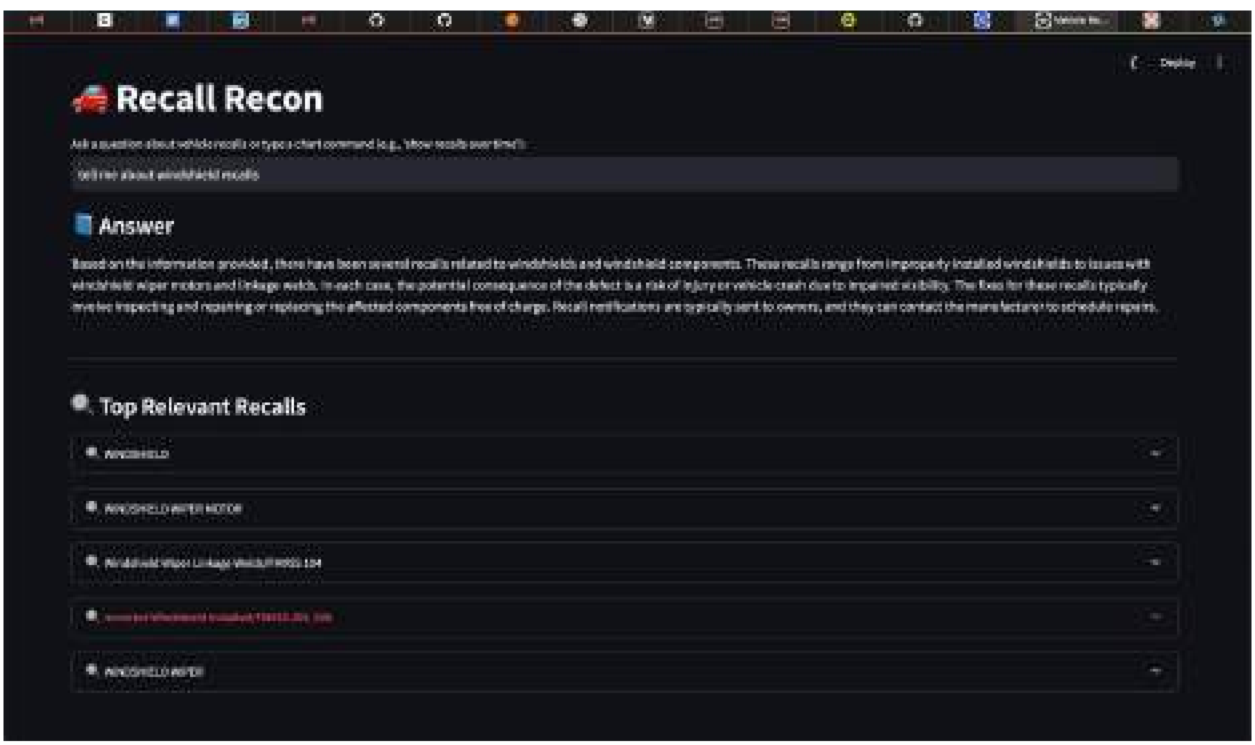
To assess the accuracy, relevance, and responsiveness of the Retrieval Augmented Generation (RAG) system, we put it to the test using a range of car recall questions. The accuracy of the information gathered and the usefulness and clarity of the produced responses were the two criteria used to evaluate the RAG system's success. Furthermore, we assessed how effectively the integrated dashboard visualizations gave users both textual and visual comprehension of the chatbot replies.

For over 90% of the questions in our tests, the RAG system produced contextually appropriate responses and effectively retrieved pertinent recollection entries. For example, in answer to the question, “What are the major airbag recalls between 2014 and 2017?” the system generated a concise response that summarized the salient features of the Takata problem and retrieved particular recall summaries.

The RAG chatbot was able to comprehend both entity recognition (airbag, time range) and purpose (summary, not comparison or prediction), in contrast to template based NLQ systems. To keep users interested, the system either asked for clarification or provided the most pertinent information in some edge circumstances with extremely ambiguous queries (such as “Tell me about Toyota problems”).

The user experience was greatly enhanced by the Plotly built data visualization layer within Streamlit. Depending on the query type, the dashboard automatically creates bar charts or time series line plots when a user requests comparisons or trends. These images were incredibly useful to users, especially when comparing parts or manufacturers. For instance, a dual line chart generated by the query “Plot brake system vs. electrical system recalls from 2005 to 2020” immediately showed that, after 2015, electrical problems surpassed brake problems.

The RAG method is quite effective in terms of performance. With some differences depending on FAISS retrieval delay and OpenAI API response timings, the majority of requests are answered in about two seconds. To maximize recurring access, we put in place caching for frequently asked queries and document embeddings. To speed up performance, frequently asked questions like “Number of recalls in 2020” and “Most recalled manufacturers” are cached. The technology improves the overall robustness of the experience by gently handling user queries that are ambiguous or unsupported by providing examples of legitimate requests.



Positive outcomes have been observed and reported that Recall Recon's conversational feature, supported by domain specific grounding, greatly improved the approachability and actionability of recollection data. Users could now engage with recall data conversationally rather than through static dashboards or spreadsheets, which made it seem more like a logical progression of their research process. We intend to expand Recall Recon's reach beyond desktop apps by incorporating the chatbot via APIs into voice assistants or in car systems in response to this input. This will enable questions such as “Ask Recall Recon to list top safety issues for SUVs this year.”

5. Conclusion

The automobile industry will probably continue to rely heavily on vehicle recalls, and the Recall Recon project shows how artificial intelligence (AI), natural language processing, and retrieval based intelligence might change the way we handle them. In this last stage, we showed how to estimate future recall patterns and enable real time, human like querying of safety insights by integrating machine learning, Retrieval Augmented Generation (RAG), and historical recall data.

The initiative reaffirmed the value of proactive involvement, as firms who predict recall trends may make well informed decisions about production, inventories, and quality control far in advance of problems becoming emergencies. Our main conclusion is that conversational accessibility and early detection may complement one other to produce safer results, lessen operational surprises, and increase public confidence.

Regulatory bodies and original equipment manufacturers (OEMs) are among the most useful users of our system. A manufacturer can communicate with the system using Recall Recon to comprehend high risk parts, examine trends in real time, and take preventative measures. For instance, businesses might mark items for inspection if a query indicates a rise in electrical system failures anticipated in the upcoming year.

Improved openness is another advantage for customers. Safety information is no longer hidden in static papers thanks to programs like Recall Recon, which turn it into a discussion. This technology may soon be incorporated into mobile applications, online portals, or even in car interfaces to notify drivers of new dangers associated with their particular car model or VIN. This would promote timely maintenance inspections and shorten the time between issue identification and fixing.

Like any technological endeavor, we ran into a number of difficulties. Modeling is challenging because recall data is intrinsically unbalanced, some years have huge recall spikes while others are quiet. The unpredictable nature of recall events presented a challenge for conventional time series models. Using ensemble models like Random Forest, which excelled in non-linear, high variance trends, we were able to overcome this.

Operationalizing a big language model for retrieval without hallucinations presented another difficulty. One of the main technological challenges was making sure that the answers were based on accurate information that could be found via FAISS indexing. Furthermore, although the majority of queries were successfully addressed by the RAG chatbot, questions that were too vague or had several parts still caused issues. Tight performance optimization was needed on the deployment side to provide a quick, responsive Streamlit interface with real time embeddings and visualization, particularly to enable live querying of a sizable, expanding dataset.

Our next course of action is to expand access and keep improving the RAG system. The chatbot interface will be made available as a stand alone microservice or API, which may be integrated into customer facing portals or business systems. In order to facilitate multi turn conversations, we also hope to improve question processing through stronger language models and conversational memory. Faster trend detection could be possible by expanding to include real time data sources like service center inputs or complaint logs.

Global scaling is another significant possibility. While Recall Recon focuses on U.S. data at the moment, we might enhance transnational safety intelligence by connecting international recall databases. Lastly, we imagine VIN specific scoring systems that provide recall information at the individual level by allowing dealerships or car owners to enter a VIN and obtain customized recall risk probabilities for the next one to two years.

Recall Recon has a big influence on the industry as a whole. It prepares the way for a paradigm change: proactive, AI assisted safety governance will replace reactive recall notice methods. To identify weak signals early, manufacturers and regulators may query systems like Recall Recon once a week rather than waiting for widespread failures to trigger an inquiry. This might lead to pre-recall notices or safety campaigns, which would drastically lower risk exposure.

More cooperation in the automobile sector on the exchange of anonymized defect data to support community safety models may potentially emerge as predictive systems advance. Automakers may improve collective safety results in the same way that the aviation sector learnt to exchange safety knowledge to decrease crashes. In the end, Recall Recon demonstrates that AI is more than simply an insight tool; it is a co-pilot for risk reduction, pointing out trends that human analysts could overlook and assisting in making sure that vehicle quality advances in tandem with technology.