Journal of Computing and Information Science in Engineering. Received October 01, 2025;
Accepted manuscript posted July 29, 2026. doi:10.1115/1.4072427

Generative AI for Classification, Prediction, and Discovery of Consumer Product Recalls

Diana Bolanos[1]
Department of Mechanical Engineering
University of California
Berkeley, California 94720
Email: dbolanos@berkeley.edu

Mohammadmehdi Ataei[2]
Autodesk Research
Toronto, ON M5G 1M1, Canada
Email: mehdi.ataei@autodesk.com

Daniele Grandi
Autodesk Research
San Francisco, CA 94105
Email: daniele.grandi@autodesk.com

Kosa Goucher-Lambert
Department of Mechanical Engineering
University of California
Berkeley, California 94720
Email: kosa@berkeley.edu

Historical information offers valuable insights throughout product development, and recalled products in particular reveal potential design risks, yet they remain underutilized. We curate RECALL-MM, a multimodal dataset from the US Consumer Product Safety Commission (CPSC) recall database, augmented using generative methods. First, we employ generative artificial intelligence (GenAI) to classify hazards from textual descriptors; using GPT-4o, we achieve a relaxed accuracy of 0.73 against reference hazard labels. Second, we benchmark five state-of-the-art vision–language models (VLMs) on hazard prediction from product images, where Top-1 accuracy ranges from 23–40% and Top-3 reaches 67% with GPT-5. Although classical retrieval baselines (Sentence-BERT and ResNet50 nearest-neighbors) exceed VLM accuracy, the foundation models still identify the correct hazard among three plausible guesses well above chance, zero-shot, from a single image. The near-equivalence of image-feature and LLM-distilled text retrieval suggests that visual interpretation carries most of the predictive signal, motivating continued development of flexible multimodal models. To complement these metrics, we apply an LLM-as-Judge framework that scores long-form predictions on a two-axis rubric, highlighting GenAI strengths and challenges in capturing hazard mechanisms and modalities. Finally, two application studies present interactive clustering maps that embed all recalls into a shared latent space to support hazard discovery. Together, these studies show how computational and GenAI-driven methods can leverage recall archives to classify, predict, and discover hazards, ultimately supporting risk-informed tools in early-stage engineering design.

Keywords: Generative AI, Failure Modes, Datasets, Product Development

1 Introduction

The product development process faces four distinct types of risk: technical, market, collaborative, and financial [1]. The likelihood of these risks manifesting is highest in the early stages of the product life cycle and decreases steadily over time. Conversely, the cost of addressing risks begins low but rises dramatically as the product advances toward manufacturing and release [2]. As such, engineers and designers are challenged to mitigate uncertainty as early as possible, particularly technical risks that could lead to costly rework or product failure.

Several techniques are utilized by industry professionals to guide a systematic approach for predicting and preventing risks. These include, but are not limited to, Failure Modes and Effects Analysis (FMEA) [3], Fault Tree Analysis [4], Event Tree Analysis [5], and Cause-Consequence Analysis [6]. While these methods are well established, they rely heavily on expert judgment and intuition of potential failure modes. This ultimately limits the ability to capture unanticipated hazards that emerge across diverse products and contexts. In practice, such analyses are often performed late in the design process, constrained by limited historical data, or scoped narrowly to known risks, reducing their effectiveness in preventing downstream failures [7]. Moreover, because these approaches are typically applied on a per-product basis, they rarely exploit cross-product evidence from prior failures, making it difficult to identify recurring hazard patterns or risk factors that manifest only when failures are examined at scale. As such, we see an opportunity to learn from publicly available recalled product information to form an empirical basis for risk-informed decision-making. However, raw recall data, such as that maintained by the United States Consumer Product Safety Commission (CPSC), are unstructured, inconsistently labeled, and difficult to analyze at scale. As a result, they remain underutilized in design practice and research.

Recent advances in Generative Artificial Intelligence (GenAI) provide new opportunities to transform unstructured recall records into machine-readable datasets. Large language models (LLMs) are particularly well-suited for tasks such as entity extraction, data normalization, and classification from mixed-modality sources [8]. In this work, we employ OpenAI's GPT-4o [9] to curate RECALL-MM, a multimodal dataset of 6,874 recalls spanning 2000–2024. The dataset includes direct retrieval fields (recall date, description, images), LLM-modified textual fields (de-branded product descriptions and product names), and LLM-categorized fields (hazard, product, and remedy classes). It also includes entirely LLM-generated fields, such as visual product descriptions derived from associated images. Collectively, this dataset represents over 546 million affected product Stock Keeping Units (SKUs) [10], underscoring the real-world impact of the failures documented.

To demonstrate the usefulness of the RECALL-MM dataset for computational design methods, we present a series of experiments and application studies around hazard identification. First, we investigate the ability of LLMs to predict hazard classifications from textual product descriptions, establishing a benchmark for text-only hazard prediction. We then extend this analysis to the visual modality, evaluating five state-of-the-art VLMs for their ability to predict hazards from product images alone. To assess the quality of these predictions, we introduce an LLM-as-Judge approach that scores model outputs along two axes: which component is likely to fail and how it is likely to fail. Finally, we move from prediction to discovery, constructing an embedding-based representation of the recall dataset that demonstrates an example of situating a new product concept within a latent space of historical recalls, enabling designers to explore neighboring products and anticipate potential risks in early-stage design.

Our contributions are: (1) the development of **RECALL-MM**,

[1]Corresponding Author.
Version 1.18, July 27, 2026
[2]Current affiliation: NVIDIA

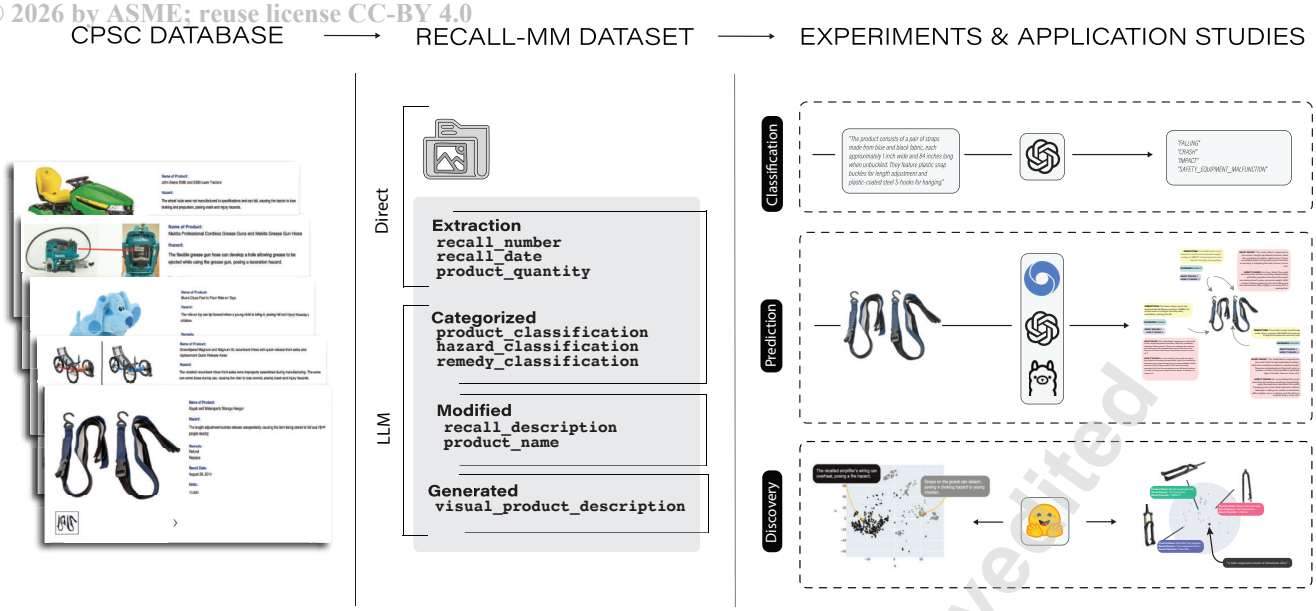

Fig. 1 Overview of the research workflow. Left: US CPSC public recall records (2000–2024) were curated into the RECALL-MM dataset (6,874 entries). Middle: dataset construction involved direct extraction and LLM-based categorizations, modifications and generations. Right: applied studies included (1) classification of hazards, (2) prediction of failure mechanisms and modes, and (3) discovery of risk patterns via embeddings and interactive exploration.

a curated multimodal dataset of 6,874 consumer product recalls combining CPSC raw fields, LLM-modified textual descriptors, LLM-categorized labels, and LLM-generated visual descriptions, validated against human annotations on a 100-item subset; (2) a **three-stage workflow for design-risk discovery from recall archives** – hazard classification, multimodal hazard prediction with long-form mechanism-and-mode reasoning, and embedding-based exploration – along with an evaluation framework comprising a relaxed-accuracy metric for single-vs-multi-label mismatches and an LLM-as-Judge rubric for long-form failure narratives; and (3) a **characterization of where generative AI adds value** for recall-based hazard analysis.

To support further development and evaluation of our dataset, we make the RECALL-MM dataset and accompanying experimental code publicly available on GitHub[3].

2 Related Work

2.1 Design Datasets. Over the last decade, several large design datasets have been curated and released to support data-driven design efforts for product design and related tasks. The classes of objects collected, sample size, and modality of the data are the main differentiators between datasets in this field.

Shapenet [11], the ABC Dataset [12], DeepCAD [13], and the Fusion 360 Gallery dataset [14] are among the largest datasets that contain geometry data and class labels for individual parts and whole assemblies. The datasets have been widely used for design automation and geometry generation tasks, and have also supported other work around ancillary design tasks such as materials selection [15]. Other smaller datasets have also been curated around more specific classes of objects, such as car bodies, mechanical components, and bicycles [16–18]. While these datasets provide valuable structured information on product features, making them a useful starting point for analysis, they also come with limitations, including incorrect or missing semantic information, a limited number of object classes, and a lack of design context, intent, and criteria, which limits their utility for comprehensive risk assessment.

Other design datasets have focused on different modalities, such as hand-drawn sketches [19,20], or textual descriptions of designs [21–24]. Although these datasets emphasize design rationale by describing the features and aesthetics of the design solutions, they do not explicitly consider design feasibility, and some are limited by the number of object classes and data quality.

A few multimodal design datasets have been published, built around graphic design [25], design requirement documents [26], and descriptions of design changes [27]. These datasets combine textual descriptions, geometry, images, and design requirements in various ways to support tasks such as editing 3D geometry, interpreting design requirements, and style classification.

The multimodal dataset collected in this work differs from prior datasets as it provides textual descriptions of designs, images, and recall information related to a failure mode of the product over a wide range of product classes. Table 1 summarizes the datasets discussed in this section.

2.2 Data-Driven Risk Analysis. Product recalls have been used to investigate trends in consumer product safety, with a prevailing focus on children's toys [29–32]. These studies similarly leverage the CPSC database, along with the global recalls dataset from the Organisation for Economic Co-operation and Development (OECD) [28]. Wai and Uttama [32] present a series of machine learning approaches for predicting a binary classification of children's toy safety, showcasing the potential for data-driven hazard prediction. Our study expands on this work by investigating trends across multiple product categories, moving beyond a single domain focus.

The automotive industry has also seen a shift towards data-driven design for predictive maintenance and hazard prevention. Yorulmus et al. [33] present machine learning approaches for predicting brake defects from vehicles passing quality-assurance checkpoints, yet exhibiting high rates of customer complaints. Similarly, [34] demonstrates a multi-label transfer learning approach by implementing a series of pretrained Convolutional Neural Networks (CNNs) to predict the binary failure status of multiple engine components. Both studies rely on failure data to improve the future design of automotive components, demonstrating the effectiveness of leveraging data for failure mitigation and design improvement.

[3] https://github.com/dianabolanos/RECALL-MM

Table 1 Comparison of design and recall datasets discussed in Section 2.1 and related recall resources, organized by domain and stage of the product design process. ML-Ready indicates datasets that are structured and labeled in a form directly usable for machine learning without substantial manual preprocessing.

Dataset	Domain	Design Stage	Multimodal	ML-Ready	Primary Use	Recall Info
ShapeNet [11]	CAD objects	Conceptual	×	✓	Geometric deep learning	×
ABC Dataset [12]	CAD assemblies	Detail	×	✓	Geometric deep learning	×
DeepCAD [13]	CAD objects	Detail	×	✓	Geometric deep learning	×
Fusion 360 Gallery [14]	CAD assemblies	Detail	×	✓	Geometric deep learning	×
Mechanical Components Benchmark [16]	CAD objects	Detail	×	✓	Component classification and retrieval	×
BIKED [17]	Bicycles	Detail	✓	✓	Geometric deep learning	×
DrivAerNet++ [18]	Automotive	Validation	✓	✓	Simulation prediction	×
Design Concept Generation [21]	Product design	Conceptual	×	✓	Concept generation	×
What's in a name? [22]	CAD assemblies	Conceptual	×	✓	Concept generation	×
Patent Data for Engineering Design [23]	Patents	Conceptual	×	✓	Concept generation	×
Crowdsourcing Inspiration [24]	Product design	Conceptual	×	✓	Concept generation	×
DesignProbe [25]	Graphic design	Conceptual	✓	✓	Multi-fidelity generation	×
DesignQA [26]	Engineering documentation	Conceptual	✓	✓	Design requirement understanding	×
ShapeTalk [27]	CAD objects	Detail	✓	✓	Geometry editing	×
OECD Global Recalls [28]	Product design	Post-market	✓	×	Regulatory recall reporting	✓
RECALL-MM (ours)	**Product design**	**Conceptual**	✓	✓	**Hazard classification, prediction, and discovery**	✓

2.3 Hazard Identification in Design. The field of risk analysis has been extensively studied and continues to evolve within various organizational and engineering domains. A fundamental objective of risk analysis in engineering design is to proactively identify and mitigate hazards before they manifest as failures or safety incidents. Failure Mode and Effects Analysis (FMEA) is among the most widely adopted methodologies employed by organizations to structure systematic risk assessments, prioritize potential hazards, and implement preventive actions [35–37]. First developed in the 1960s by the aerospace industry [38], FMEAs now serve as an industry standard tool across various applications, including automotive design, aerospace engineering, and product development, demonstrating its versatility in supporting product reliability and safety. Despite its maturity and widespread use, FMEA remains fundamentally dependent on expert elicitation and product-specific knowledge available at the time of analysis. As a result, its effectiveness is constrained by the completeness of anticipated use scenarios and the prior experience of the analysis team. In many practical settings, FMEAs are conducted within narrow organizational boundaries, limiting the ability to leverage empirical evidence from failures observed in other products or domains. Consequently, recurring hazard patterns that manifest across large populations of consumer products are rarely incorporated into early-stage design risk assessments.

More recent advancements in FMEA methodologies incorporate computational approaches, such as fuzzy logic, machine learning, and integrated decision-making frameworks, enhancing traditional FMEA practices by addressing uncertainties and subjectivity inherent in risk scoring [39]. While these approaches enhance the internal consistency of FMEA evaluations, they typically operate on hazards already identified by experts and do not address the upstream challenge of discovering previously unanticipated failure modes at scale. Moreover, the lack of structured, large-scale failure datasets accessible to designers limits the practical integration of data-driven insights into routine hazard identification workflows. In this work, we argue that systematically structured recall data can complement existing FMEA practices by providing empirical evidence of real-world failures across thousands of products. The dataset introduced in this paper transforms unstructured recall records into a machine-interpretable resource that enables automated hazard categorization, cross-product analysis, and exploratory risk discovery. Rather than replacing established FMEA methodologies, the proposed dataset and experiments are intended to augment hazard identification by exposing designers to patterns of failure that are difficult to anticipate through expert judgment alone.

3 Methods

We begin this section by detailing the steps used to clean and augment the CPSC database into a multimodal dataset. Then we introduce two experiments to explore hazard classification and prediction based on textual and visual evidence using GenAI. Lastly, we position this work to support future studies of user interaction with multimodal data through embedding-space visualizations.

3.1 Dataset. The dataset used in this analysis is a curated subset of the US CPSC recalls database [10]. To ensure consistency and feasibility for analysis, recalls were filtered based on API accessibility, the presence of product images, and the time frame of 2000 to 2024. This filtering process yields a dataset comprising 6,874 recall entries, which accounts for 92.4% of total recalls within the specified time frame.

Each entry in the dataset includes essential recall attributes such as hazard classifications, product categories, remedy types, historical recall dates, and an associated product image (Table 6). The raw CPSC database entries lack labeled classifications for each entry. To address this, we leveraged a generative pre-trained model, GPT-4o [9], to enrich and structure the data. We chose to use GPT-4o as it was the latest model released by OpenAI at the time of preparing this dataset. Figure 1 illustrates which fields were directly extracted from the CPSC database and which were augmented using GPT-4o.

Fields such as the *recall_number*, *recall_date*, and *product_quantity* were directly extracted from the database, along with the first image in each entry. We denote this as **Direct** in Fig. 1. Next, we leverage GPT-4o to assign each recall entry a product

Journal of Computing and Information Science in Engineering. Received October 01, 2025;
Accepted manuscript posted July 29, 2026. doi:10.1115/1.4072427

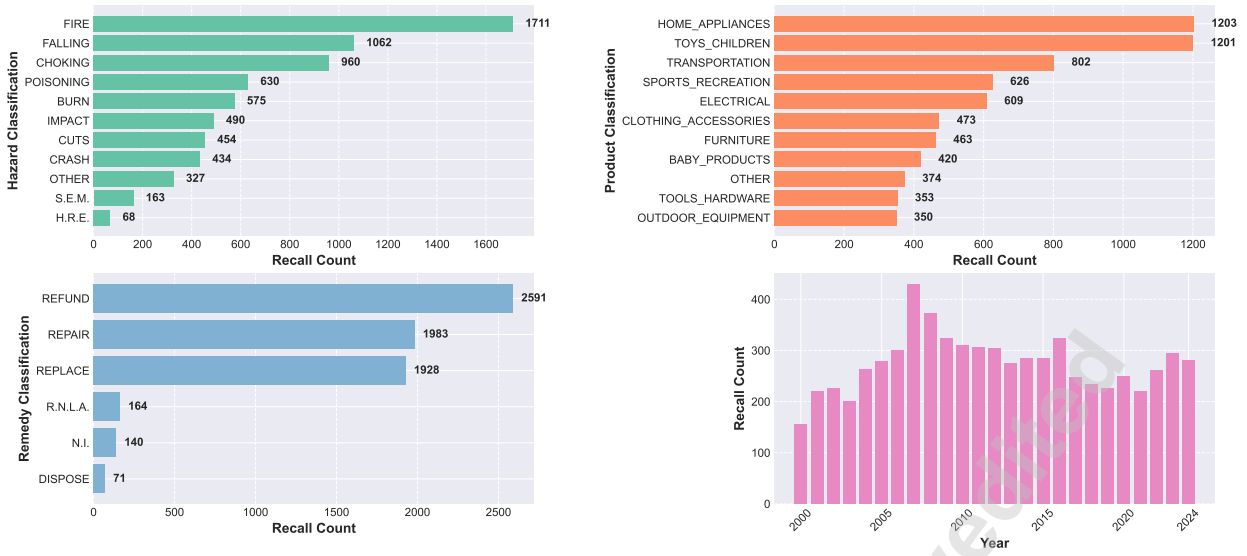

Fig. 2 Summary statistics from the curated RECALL-MM dataset of 6,874 US CPSC product recalls (2000–2024). Top left: distribution by hazard classification. Top right: distribution by product classification. Bottom left: distribution by remedy classification. Bottom right: annual recall counts.

category, hazard category, and remedy type, selecting from predefined lists. We label this as **LLM Categorized**. Hazard and remedy categories were aligned with CPSC's own labels, with definitions provided in Appendix B (Tables 7 and 8). Since the CPSC does not offer standardized product categories, we developed an 11-category scheme based on domain understanding, ensuring it broadly covers the landscape of recalled products while maintaining consistency with existing definitions for global recalls published by the OECD [28]. The **LLM Modified** fields, *recall_description* and *product_name*, were refined to remove brand references and retain generic descriptions. For example, original descriptors often included specific brand names associated with the product. To remove bias in future analysis, we prompted the LLM to perform this modification task by simply removing any branding from the textual descriptors of each item. Lastly, we label the *visual_product_description* field as **LLM Generated**, as it was generated entirely by prompting GPT-4o to describe the product based on its associated image.

During data cleaning, all records were standardized to a predefined schema, ensuring consistent representation of attributes such as product description, hazard type, and remedy details. GPT-4o outputs were validated against this schema, with type and value constraints applied to ensure reliability. Each record retains its original recall ID, preserving traceability to the source data and supporting future reference.

We use consistent terminology for the three categories of labels appearing throughout this work. **CPSC raw fields** (recall_number, recall_date, product_quantity, and the original product images) are factual records taken directly from the source database. **LLM-derived reference labels** are the hazard, product, and remedy classifications assigned by GPT-4o across all 6,874 recalls (the "LLM Categorized" fields in Fig. 1). These serve as the reference targets for the full-dataset evaluations in Sections 4.2 and 4.3.1. **Human-annotated validation labels** are majority-vote categorizations produced by three independent annotators on a 100-item subset and are used to validate the LLM-derived reference labels. The recall_description field, used as the textual comparison target for long-form scoring in Section 4.3.2, is derived directly from the CPSC database which GPT-4o modified only to remove brand references; we refer to it directly as the recall description. Throughout the paper, we reserve "ground truth" for the CPSC raw fields and avoid it elsewhere to prevent conflation with derived labels.

3.2 LLM Hazard Classification.

We evaluate the utility of LLMs to predict potential hazards directly from product image descriptions. Specifically, we focus on the *visual_product_description* field from the dataset, which contains a textual description of each product image generated by GPT-4o.

To perform hazard prediction, we then prompt GPT-4o to analyze the textual description of the image and output all applicable hazard classifications. The model selects hazard labels from a predefined set of 11 hazard categories (Table 7), ensuring consistency with existing CPSC classifications. The full prompt used for LLM prediction is provided:

```
You are a product safety expert. Identify all
potential hazards for the given product. Provide
output in valid JSON format only, structured as:

{
   'product': {product_description},
   'predicted_hazards': [List of all applicable
hazards]
}

The predicted_hazards field must only contain haz-
ards from this set: {all_hazards}. Do not leave
any predicted_hazards fields empty. If multiple
hazards apply, include all relevant ones. No ex-
planations - return JSON only.
```

Outputs are returned in strict JSON format, listing the product description and the predicted hazards. The prompt enforces that the model must not leave any hazard fields empty, however we do include an "OTHER" category in the provided list of hazard choices to encourage comprehensive identification of potential risks.

Evaluation Metric. To quantify the model's performance, we introduce a Relaxed Accuracy (RA) metric:

$$\delta_i = \begin{cases} 1, & \text{if } g_i \in P_i \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$$\text{Relaxed Accuracy (RA)} = \frac{1}{N}\sum_{i=1}^{N} \delta_i \tag{2}$$

where

- N = Total number of products per hazard class.
- g_i = Single reference hazard classification for product i.
- P_i = Set of predicted hazard classifications for product i.
- δ_i = Indicator function that equals 1 if g_i is present in the predicted set P_i, and 0 otherwise.

The RA metric accounts for cases where the LLM predicts multiple hazards, but the recall dataset provides only one reference hazard label per entry. The recall dataset only labels a single hazard per product (single-class), even though products may exhibit multiple concurrent hazards (multi-class). Thus, RA prioritizes capturing all realistic hazards that could apply to a product, rather than penalizing extra predictions, acknowledging that the recall labels do not list every possible hazard. Overall RA is reported as the macro-average (unweighted mean) of the per-class RA values. Because RA checks whether g_i appears in the predicted set P_i, it tends to increase as $|P_i|$ grows.

It is important to note that we intentionally do not impose restrictions on the number of predictions required of the LLM. The reasoning for this lays in the nature of hazard predictions, where a single product may truly be capable of exhibiting multitudes of potential hazards. As such, we choose not to restrict the predictions as this could lead to underprediction. Nonetheless, further analysis of the distribution of predictions confirms that the LLM did not simply inflate the results by predicting all possible hazards for each item. This is presented in detail in Section 4.3.1.

Classical Text Baselines. To contextualize the LLM classification result and isolate what generative methods contribute beyond traditional supervised learning, we additionally evaluate two non-generative text baselines on the same hazard classification task. The first is a Sentence-BERT k-nearest-neighbors classifier. Each product's *visual_product_description* is encoded with `all-MiniLM-L6-v2` (the same encoder used in Section 3.4), and predictions are formed by majority vote among the $k = 10$ nearest neighbors in the training set by cosine similarity. The second is TF-IDF feature extraction followed by multinomial logistic regression, trained on the same split. Both baselines operate on the LLM-generated *visual_product_description*, matching the LLM classification setup for direct comparability. We note that this means neither baseline is fully independent of generative components in data preparation, but both are non-generative at inference time and require no LLM API calls. Results are reported in Section 4.3.1.

3.3 VLM Hazard Prediction. This study explores the use of language models for long-form hazard prediction using the images directly extracted from the CPSC database. We benchmark five vision–language models: OpenAI GPT-4o, OpenAI GPT-4o-mini, OpenAI GPT-5 (2025 API release), Google DeepMind Gemini 1.5 Flash, and LLaVA 1.6 (7B parameters) on a random sample of n=100 entries as a representative portion of the full RECALL-MM dataset. These models were selected to represent both state-of-the-art commercial models and an open-source baseline. This multi-step approach begins by prompting the model to consider each of the 11 hazard categories and determine whether there is potential for the hazard to surface based on the product's visual image. To analyze the results of this experiment, rather than prompting the model to select *any* hazard classification from the list of 11, we now prompt the model to only consider the top 3 most probable hazards. This approach was chosen as a compromise to allow for richer exploration of recall predictions through long-form responses, while balancing the single-class nature of the reference labels versus actual multi-class nature of hazards attributed to a product. Using a Top-k metric, we first compare the model's interpretation of potential hazard classes against the reference classifications.

The next stage prompts the model to provide long-form text descriptions of the failure mechanisms (*what* part or component might fail) and the failure modes (*how* it might fail). The full prompt is found in Appendix C. This step provides additional interpretability by exposing the reasoning behind the model's hazard predictions, rather than limiting the output to a categorical label. It also enables qualitative comparison between the model's predicted failure narratives and the original recall descriptions, allowing us to assess whether the model identifies plausible mechanisms and modalities of failure. This structure mirrors how engineers typically conduct failure analysis, where both the failing component and the mode of failure are explicitly articulated.

As seen in Fig. 10, each of the three predictions are provided in long-form text, ultimately making it a tedious human task to judge the model's accuracy against the *recall_description*. To facilitate evaluation of these generated predictions, we introduce an LLM-as-Judge [40] to evaluate its own model predictions. This approach enables a large-scale evaluation otherwise expensive to conduct across human validations, and is consistent with emerging practices in generative model assessment. Each prediction is rated from score 1-3 on a two-axis system, focused on failure mechanism accuracy and failure modality accuracy, which aligns with the structure of the recall_description and allowing for direct comparison. The prompt used to retrieve these scores is provided in Appendix C. A score of 3 denotes an exact match, a score of 2 refers to a related part or failure mode, and a score of 1 represents an incorrect or unrelated description of the failure mechanism or modality. In addition to raw scores, we also prompt the models to provide a rationale for each score in natural language text. While we acknowledge the limitations of relying on GenAI evaluations, such as model bias and hallucinations, this approach serves as a time-efficient technique otherwise difficult to conduct with human experts.

Classical Image Baselines. To enable compare against the VLM Top-k results, we additionally evaluate two non-generative image baselines on the same $n = 100$ test set. Image features are extracted from the penultimate layer (2048-dim avgpool output) of a ResNet50 [41] pretrained on ImageNet (`IMAGENET1K_V2` weights). ResNet50 is a supervised deep CNN that predates the foundation-model era and is distinct from the CLIP-family encoders used in modern VLMs, providing a non-GenAI vision feature extractor. We then run two classifiers on these features: (i) k-nearest-neighbors with $k = 10$ and majority voting over neighbors, and (ii) multinomial logistic regression. Both are trained on the remaining recalls in RECALL-MM that have associated product images.

3.4 Embedding Space Discovery and Exploration. We now shift our methods to exploring a different way of interacting with the RECALL-MM dataset. Rather than relying exclusively on commercial large language and vision models, we introduce an embedding-based approach that leverages representation learning to enable discovery and exploration. Embedding techniques fall under the broader umbrella of generative AI, as they transform raw text into structured latent representations that can be combined with generative models to support downstream tasks such as synthesis, retrieval, and hazard prediction. In the context of engineering design, embeddings provide a structured way to represent unstructured recall narratives, enabling products to be grouped and compared by

similarity. Visualizing these embeddings offers a practical means of exploring relationships between products and hazards, allowing researchers to identify clusters and trends that would be difficult to see through traditional keyword search.

3.4.1 Text Embeddings. Product descriptors, specifically *recall_description* and *product_name*, were embedded into a numerical latent space using the `all-MiniLM-L6-v2` model from Sentence-BERT [42]. Sentence-BERT is a pre-trained model that generates fixed-length dense vector representations optimized for capturing semantic similarity between text inputs. We selected the `all-MiniLM-L6-v2` model because it offers an effective balance between model size, computational efficiency, and embedding quality, making it well-suited for large-scale analyses without sacrificing performance.

3.4.2 Dimensionality Reduction for Visualization. To visualize relationships among products and recall reasons, we applied dimensionality reduction techniques. Specifically, we employed the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm, chosen for its strength in preserving local structure and effectively capturing complex, non-linear relationships in high-dimensional data [43]. Compared to linear methods such as Principal Component Analysis (PCA) [44], which primarily maintain global variance, t-SNE excels at revealing dense clusters and neighborhood groupings, which are paramount features for identifying semantically similar products and localized recall patterns. Product embeddings obtained via Sentence-BERT were projected from their original vector space into two and three-dimensional coordinates. In the latter example shown in Section 4.4.2, the choice of three dimensions is motivated by visualization clarity rather than analytical necessity, as two-dimensional projections exhibited greater overlap between nearby clusters in this example. Importantly, the underlying embedding remains high-dimensional, and the insights derived from this demonstration are not dependent on the dimensionality of the visualization.

The resulting visual maps (Fig. 8 and Fig. 9) enable exploration of recall clusters, offering insight into product similarities and hazard trends. While other methods like UMAP [45] could also be considered, we prioritized t-SNE for its well-established use in exploratory visualizations where fine-grained local structure is of primary interest.

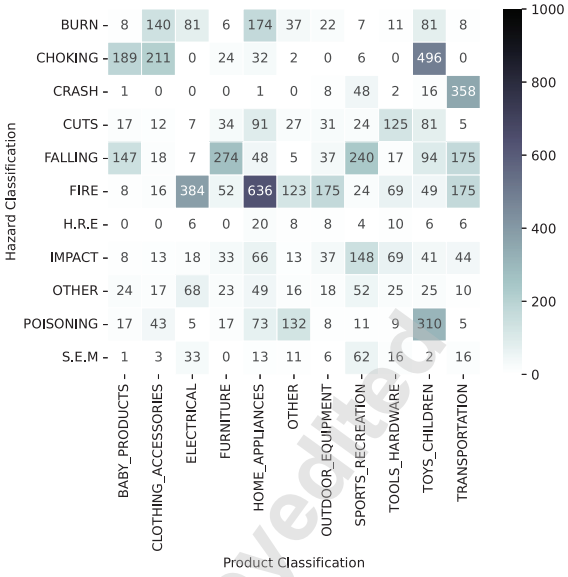

Fig. 3 Correlation matrix of the LLM-derived reference product and hazard classifications across RECALL-MM (assigned by GPT-4o from textual recall data; validated against human-annotated labels in Section 4.1).

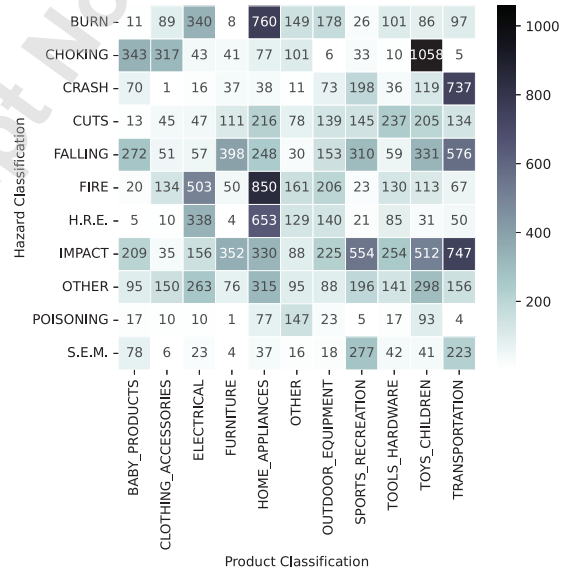

Fig. 4 Correlation matrix of GPT-4o-predicted hazard classifications across all recalls, generated from visual_product_description (multi-label predictions aggregated per recall). Compare to Fig. 3, which shows the reference labels.

4 Results and Discussion

To evaluate the reliability of the dataset, we first validate the LLM-generated categorizations against human-annotated annotation labels across product, hazard, and remedy classifications. We then analyze aggregate trends, examining common hazards, product categories, and remedies over time. Building on these results, we assess LLM hazard classification using a relaxed accuracy metric, benchmark five VLMs with Top-k and LLM-as-Judge evaluations, and present qualitative examples of model reasoning. Finally, we construct embedding-based visualizations that reveal cross-domain hazard clusters and situate new product concepts within the recall landscape.

4.1 Human Evaluation of Dataset Categorizations. To validate the reliability of the LLM-generated categorizations, we conducted a human annotation study to establish human-annotated validation labels for a subset of the dataset. Three independent annotators participated in the study: two industry research scientists with experience in product design and engineering analysis, and one doctoral student in engineering design. All annotators were familiar with consumer product descriptions and hazard-related terminology. Each annotator performed 100 classifications for each task, amounting to 900 annotations. For each record, annotators were provided with the original long-form textual descriptions associated with the recall (including product description, hazard description, and remedy description, respectively) and were asked to assign a single class label from a predefined taxonomy corresponding to each task. No time limits were imposed, and annotators worked independently without communication or adjudication during labeling.

To establish validation labels, we employed majority voting across the three annotators' labels for each classification task (product, hazard, and remedies). We assessed inter-rater reliability by evaluating Fleiss's Kappa, which yielded coefficients of 0.71 for product classification, 0.80 for hazard classification, and 0.85 for

Table 2 Per-class classification metrics for all hazard classifications. Results show strong predictive performance overall (RA = 0.73), with highest accuracy for *crash* and *choking* hazards and lowest for *poisoning*. Note that in our case of a single reference label

, Recall and Relaxed Accuracy are identical.

Hazard Class	Precision	Recall	F1	Relaxed Accuracy	*n*
Burn	0.18	0.59	0.28	0.59	565
Choking	0.47	0.91	0.62	0.91	1055
Crash	0.29	**0.93**	0.44	**0.93**	413
Cuts	0.22	0.65	0.33	0.65	456
Falling	0.36	0.85	0.50	0.85	1041
Fire	**0.56**	0.74	**0.64**	0.74	1706
Heat Related Explosion	0.04	0.74	0.07	0.74	72
Impact	0.13	0.87	0.22	0.87	499
Other	0.08	0.46	0.14	0.46	325
Poisoning	0.52	0.32	0.39	0.32	657
Safety Equipment Malfunction	0.10	0.49	0.17	0.49	157
Macro Avg.	0.27	0.69	0.34	0.69	–
Weighted Avg.	0.37	0.73	0.46	0.73	–
Overall	–	–	–	**0.73**	6874

remedies classification. These scores indicate substantial to almost perfect agreement, based on standard interpretation thresholds. Given this high level of consistency, majority voting was deemed appropriate to consolidate the annotations. Items where no majority agreement was reached (5 for product, 4 for hazard, and 0 for remedy) were excluded from further analysis to maintain the integrity of the validation labels.

With the validation labels established, we compared against the LLM categorizations. Using Cohen's Kappa, we observed almost perfect agreement across all three classification tasks, with coefficients of 0.82 for product classification, 0.91 for hazard classification, and 0.90 for remedies classification. These strong agreement levels indicate that the LLM's predictions align closely with human judgment, achieving a level of consistency comparable to expert annotators. These results indicate that the LLM's categorizations align closely with expert human annotations, supporting the reliability of the automated labeling approach for downstream analysis.

4.2 Data Analysis of Recall Classifications. To analyze patterns in product recalls, we first aggregated the dataset across four dimensions: hazard classifications, product categories, remedy classifications, and recall year. These aggregations are visualized in Fig. 2, allowing us to identify prevalent hazards, frequently recalled product types, common industry remedies, and temporal trends in recall activity.

Additionally, to explore relationships between product categories and associated hazards, we generated a hazard-product correlation heatmap (Fig. 3). This visualization highlights where certain hazards are disproportionately concentrated within specific product types, offering insight into recurring failure modes and industry-specific safety concerns.

Examining **hazard classification**, *fire*, *falling*, and *choking* emerge as the most prevalent hazards, suggesting a need for improved foresight in anticipating these failure modes. Focusing on *fire*, we also see from Fig. 3 that the strongest correlations come from *electrical* and *home_appliances*. This aligns with existing literature indicating heightened risks of household fires due to electrical failures in sockets, plugs, and wiring, as opposed to householder carelessness [46].

Within **product classification**, the high frequency of recalls in *home_appliances* and *toys_children* indicates particular vulnerability to hazards faced within the average US household, underscoring the need for safer design of products intended for vulnerable or high-use demographics. A study conducted by Anwar [47] also relied on the CPSC database to examine the harms resulting from high recalls in the toy industry. This study found that while most toys were manufactured in China, a vast quantity of toys were designed in the US, with choking and lead poisoning as primary concerns. This analysis aligns with our findings, emphasizing the importance of stronger safety precautions when designing consumer products. Interestingly, the heatmap also shows lower recall frequencies for categories such as *tools_hardware* and *outdoor_equipment*, suggesting heightened risk awareness and more conservative design practices within these industries.

The prevalence of *refund* as a remedial action indicates a preference for immediate consumer safety, likely chosen when repairs or replacements are insufficient for risk mitigation. The substantial use of *repair* and *replace* remedies further suggests a widespread practice of addressing safety concerns through corrective product interventions rather than solely financial compensation. Kübler et al. [48] observed the effect of product recalls on brand loyalty, and found that consumers valued transparency and convenient handling of the recalled product.

The temporal analysis reveals an apparent peak in recall incidents around 2005, followed by a general downward trend with fluctuations thereafter. This trend may reflect enhanced regulatory interventions, evolving industry standards, or changes in manufacturing practices and quality-control measures.

Taken together, these results emphasize key areas for improved product safety. Recalls in categories such as *home_appliances*, *electrical*, and *toys_children* underscore the need for more proactive hazard anticipation during the product development process. Furthermore, the correlation patterns revealed by the heatmap suggest potential value in cross-domain learning: designers may benefit from examining hazard trends in adjacent product sectors to better anticipate potential risks.

4.3 Experiments.

4.3.1 LLM Classification. To assess predictive performance, we evaluate the model using precision, recall, F1 score, and RA across all hazard classes, summarized in Table 2. The model achieves an overall RA of 0.73, with weighted-average precision of 0.37, recall of 0.73, and F1 of 0.46. We note that RA is equivalent to recall in this evaluation setting because each reference label contains exactly one hazard label, the condition "at least one predicted hazard

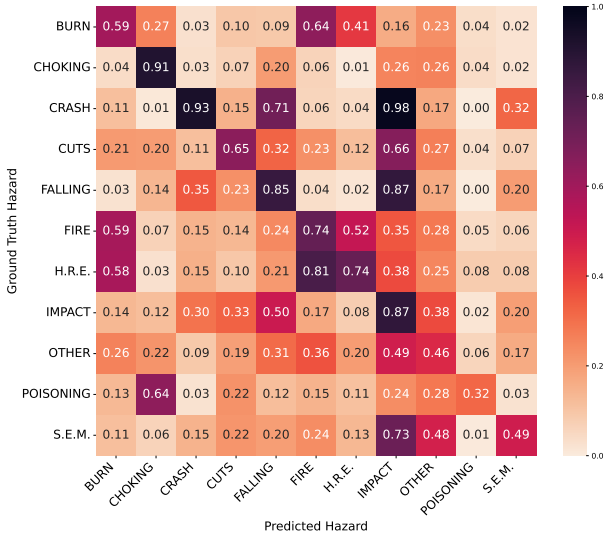

Fig. 5 Prediction frequency matrix showing the proportion of samples with a given reference hazard (row) where the LLM predicted a specific hazard (column). Diagonal values indicate per-class relaxed accuracy, whereas off-diagonal values indicate co-prediction frequency.

matches the reference label" reduces to "the single reference hazard was correctly predicted", which is the definition of recall. We report the distribution of $|P_i|$ (mean = 2.855 / median = 3.000 / IQR = 1.000) to contextualize these metrics. These values indicate the model is not inflating performance by over-predicting all 11 hazard classes for every product. Nonetheless, the observed asymmetry between recall/RA (0.73) and precision (0.37) suggests a prediction strategy in which the model emphasizes hazard detection at the expense of increased false positives.

Per-class results reveal substantial variation in predictive performance. Hazards with clear physical manifestations achieve high recall and strong F1 scores: *choking* (recall: 0.91, F1: 0.62), *crash* (recall: 0.93, F1: 0.44), and *fire* (recall: 0.74, F1: 0.64). In contrast, hazards that are less visually apparent, such as *poisoning* (recall: 0.32) and *other* (recall: 0.46), exhibit lower recall, suggesting the model struggles to infer risks not directly observable from product descriptions.

To further examine the model's multi-label prediction behavior, we construct a conditional prediction frequency matrix (Fig. 5), where each cell represents *P(Predicted=Y | Reference=X)* – the proportion of samples with reference hazard X where the model predicted hazard Y. Diagonal entries correspond to per-class recall, while off-diagonal entries reveal co-prediction patterns. The matrix exposes semantically meaningful clusters. For example, when the reference label is *crash*, the model frequently co-predicts *impact* and *falling*, reflecting physical similarities between mechanical hazards. Similarly, *fire* is often co-predicted with *burn* and *heat-related explosion*. These patterns suggest the model has learned domain-relevant associations between hazard types rather than predicting independently.

Figures 3 and 4 display the co-occurrence frequencies between product categories and hazard classifications for the reference labels and LLM predictions, respectively. Each cell value represents the count of products within a given category (column) that are associated with a specific hazard (row). Figure 3 reflects the single reference label per product, and Fig. 4 aggregates all hazards predicted by the model. Further examination of Figure 4 reveals that the model rarely predicts *poisoning* hazards, particularly for children's toys. Cross-referencing with the reference labels (Fig. 3) shows a substantial number of poisoning-related recalls in this category – a discrepancy that highlights the challenge of predicting non-visible hazards from visual descriptions. The prediction frequency matrix corroborates this: the *poisoning* row shows minimal off-diagonal co-predictions, indicating the model does not conflate *poisoning* with other hazard types, yet the low diagonal value (0.32) confirms these hazards are frequently missed. Notably, when the model does predict *poisoning*, it achieves relatively high precision (0.52), suggesting conservative but accurate behavior. This pattern mirrors how consumers often rely on visual inspection to assess product safety, underscoring a critical need for transparent hazard communication and proactive design strategies that address latent risks not apparent through appearance alone. Prediction of non-visible hazards might be improved through additional metadata such as material composition, functional descriptions, or 3D models that reveal internal mechanisms.

Textual Comparison to Non-GenAI Classical Baselines. For comparison against non-GenAI baselines on the same dataset, we evaluate Sentence-BERT k-NN and TF-IDF + logistic regression operating on the *visual_product_description*. On the $n = 100$ test set used for VLM evaluation in Section 4.3.2, Sentence-BERT k-NN achieves Top-1 accuracy of 69.0% and TF-IDF + logistic regression achieves 64.0%. Although the RA of 0.73 reported here is not directly comparable – RA permits credit for any label in a multi-label prediction whereas Top-1 requires the single highest-ranked prediction to match, and the test sets differ – the LLM's score, computed under the more relaxed metric and on the full dataset, is in a similar range to the baseline Top-1 accuracies, suggesting that generative models do not provide a substantial classification advantage over classical retrieval on this task.

4.3.2 VLM Prediction. We choose a Top-3 cap as a result of the distribution of recalls seen in the LLM classification experiment. The results of the VLM prediction experiment are shown in Table 3. The strongest Top-1 score reaches 40% using GPT-4o-mini, while GPT-5 aligns with the reference hazard classification 67% of the time within three responses (Top-3). All evaluated VLMs outperform a random baseline (11/22/33%) and the stronger models exceed a most-frequent-class baseline (27/43/54%), confirming that the VLMs use product information rather than dataset priors.

Visual Comparison to Non-GenAI Classical Baselines. We additionally compare against classical non-GenAI baselines (Table 3, ResNet50 rows). A ResNet50 k-NN classifier on raw images

Table 3 Hazard classification Top-k accuracy (%) with 95% Wilson confidence intervals on a test set of $n = 100$ recalls. VLMs operate on product images; classical image baselines operate on raw images via ResNet50 features; classical text baselines operate on *visual_product_description*.

Method	Top-1	Top-2	Top-3
Random baseline	11.0	22.0	33.0
Most-frequent baseline	27.0	43.0	54.0
Product Image input			
GPT-4o	$38.0^{\pm 9.4}$	$49.0^{\pm 9.7}$	$56.0^{\pm 9.6}$
GPT-4o-mini	$\mathbf{40.0^{\pm 9.5}}$	$58.0^{\pm 9.5}$	$65.0^{\pm 9.2}$
GPT-5	$39.0^{\pm 9.4}$	$\mathbf{59.0^{\pm 9.5}}$	$\mathbf{67.0^{\pm 9.1}}$
Gemini-1.5-Flash	$39.0^{\pm 9.4}$	$52.0^{\pm 9.6}$	$55.0^{\pm 9.6}$
LLaVA 1.6	$23.0^{\pm 8.2}$	$36.0^{\pm 9.3}$	$49.0^{\pm 9.7}$
ResNet50 + LogReg	$54.8^{\pm 9.9}$	$72.0^{\pm 9.0}$	$82.8^{\pm 7.6}$
ResNet50 k-NN	$\mathbf{67.7^{\pm 9.3}}$	$\mathbf{76.3^{\pm 8.5}}$	$\mathbf{83.9^{\pm 7.4}}$
visual_product_description input			
TF-IDF + LogReg	$64.0^{\pm 9.2}$	$78.0^{\pm 8.0}$	$83.0^{\pm 7.3}$
Sentence-BERT k-NN	$\mathbf{69.0^{\pm 8.9}}$	$\mathbf{83.0^{\pm 7.3}}$	$\mathbf{88.0^{\pm 6.4}}$

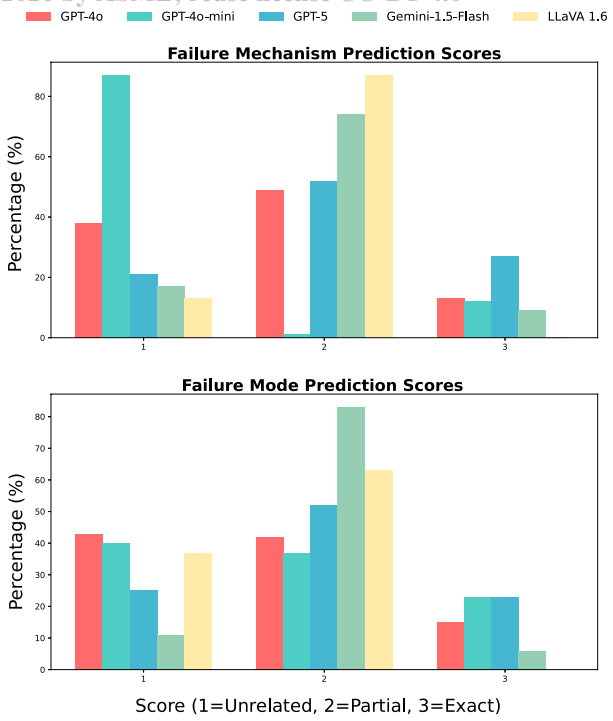

Fig. 6 Best-of-three hazard-prediction scores across five VLMs. For each recall, the highest score among three predictions is taken, and the percentage distribution over scores 1 (unrelated), 2 (partial), and 3 (exact) is shown by model. Top: Failure mechanism distribution of "what failed" scores; Bottom: Failure mode distribution of "how it failed" scores. Higher proportions at 2–3 indicate better alignment with the recall descriptions.

achieves Top-1 of 67.7%, exceeding all five VLMs with non-overlapping confidence intervals. A parallel Sentence-BERT k-NN baseline on the textual *visual_product_description* (Section 4.3.1) achieves similar Top-1 of 69.0%. The near-equivalence between image-feature retrieval and text-description retrieval suggests that visual features carry most of the predictive signal for hazard category, since both raw-image features and the visual description observe comparable performance. While the VLMs do not match retrieval baselines on classification accuracy, they identify the correct hazard among three plausible guesses two-thirds of the time from a single image, zero-shot and without access to a labeled corpus and performs well above the random (33%) and most-frequent (54%) lower bounds. The VLMs additionally produce long-form failure-mechanism and failure-mode narratives that classical baselines cannot generate, which we examine next.

Long-form Predictions and LLM-as-Judge Scoring. Figure 6 shows best-of-three score distributions, where we take the highest of three predictions per recall to reflect each model's best shot. Overall averages are still modest against the reference hazard classifications, but the best-of-three view shows that models like GPT-5 and Gemini-1.5-Flash often produce at least one reasonably correct prediction (score ≥ 2) and a noticeable number of exact matches. In contrast, LLaVA 1.6 never reaches a score of 3, which points to a stricter notion of what counts as an exact match. To assess the optimism introduced by best-of-three selection, Table 4 compares average best-of-three scores against the average across all predictions. The gap of 0.2–0.5 points reflects that not all three predictions match the reference label, which is an expected outcome. Each recall entry documents a single hazard, yet products may pose multiple plausible risks. A model generating three diverse predictions - one matching the documented hazard (score 3) and two identifying other potential hazards - would appear "inconsistent" despite functioning correctly. The mean penalizes this diversity, while best-of-three asks the more practical question: did the model identify the documented hazard among its predictions? We report both metrics for transparency, but argue that best-of-three better reflects real-world utility where identifying at least one correct hazard is the goal.

We highlight the scoring dynamics of the LLM-as-Judge through this qualitative example:

Recall Description: The recalled area rugs violate the mandatory federal flammability regulations for carpets and rugs, posing a fire hazard.

VLM Prediction 2: The synthetic fibers of the rug could ignite if exposed to an open flame or heat source, causing a FIRE hazard because many synthetic textiles are flammable and may rapidly catch fire.

- What Failed: 3
- How it Failed: 2

Rationale: The recall states that the area rugs themselves violate flammability regulations, and the prediction also identifies the rug (specifically its synthetic fibers) as the failing component, so this is an exact match (score 3). For the failure mode, the recall cites violation of mandatory flammability standards (a regulatory noncompliance leading to fire hazard), while the prediction describes the fibers igniting when exposed to flame or heat (a physical ignition mechanism). Both are related to flammability and fire hazard, but the recall emphasizes regulatory failure while the prediction emphasizes material ignition. Thus, they are related but not identical, earning a score of 2.

This recall describes a rug fire hazard from failure to comply with flammability regulations. One of the predictions by GPT-5 is presented, which shows strong accuracy in the mechanism of failure, receiving a score of 3. However, the modality of failure scoring appears strict in expecting closer alignment to the recall description. In this case, the recall description is more general, whereas the prediction specifically denotes the fibers as a source of flammability. We see this as a weakness in validating the accuracy of LLM predictions given this dataset. It is a difficult task to quantify, and we place lower emphasis on the LLM-as-Judge scores relative to the qualitative insights and capabilities of GenAI hazard prediction.

4.4 Application Studies. In addition to evaluating classification and prediction performance, we present application studies that show how the RECALL-MM dataset can be explored through embedding spaces. These studies are intended as a first step toward

Table 4 Impact of aggregation strategy on evaluation scores. For each recall, VLMs generate three hazard predictions, each scored from 1 (unrelated) to 3 (exact match). *Avg. Best-of-3* takes the maximum score among the three predictions per recall, then averages across all recalls. *Avg. All* averages all prediction scores directly.

Model	Mechanism (Avg.)		Mode (Avg.)	
	Best-of-3	All	Best-of-3	All
GPT-4o	1.75	1.52	1.72	1.45
GPT-4o-mini	1.25	1.09	1.83	1.35
GPT-5	2.06	1.54	1.98	1.47
Gemini-1.5-Flash	1.92	1.57	1.95	1.58
LLaVA 1.6	2.00	1.80	1.50	1.30

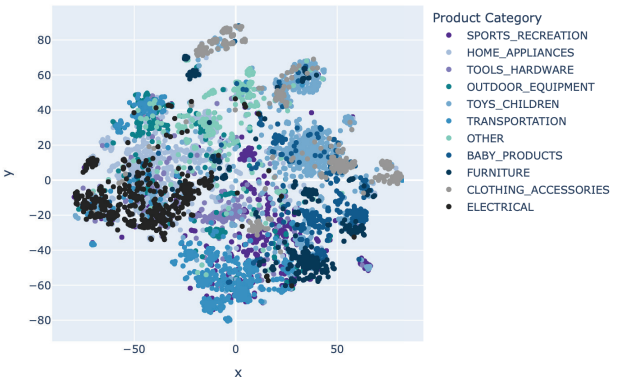

Fig. 7 Two-dimensional embedding of recall descriptions using dimensionality reduction. Markers are colored by product category, showing clustering patterns that reflect category similarities, differences, and overlaps.

discovery-oriented tools, where designers can investigate patterns in recalled products and situate new concepts within a broader recall landscape. While our current work focuses on building and demonstrating these embeddings, we envision future user studies where designers interact with these visualizations directly to support hazard awareness and design decision-making.

4.4.1 Two-Dimensional Embedding Space. In the first application study, we examine the embedding space of the *recall_description* field to explore how textual recall data can reveal underlying patterns beyond predefined hazard classifications. We specifically chose to embed *recall_description* to augment existing predefined groupings, offering an opportunity to observe unseen relationships within the dataset. Figure 7 presents a latent space representation of all 6,874 *recall_description* texts, revealing clusters that broadly correspond to different product categories.

To illustrate how product domains influence the structure of this space, we highlight two specific examples in Fig. 8. In Fig. 8a, we focus on the product categories *electrical* (black) and *clothing_accessories* (gray). Here, we observe minimal overlap in the embedding space, reflecting distinct recall descriptions and associated hazards. For example, recalls within the *electrical* category predominantly reference fire hazards, resulting in a more cohesive clustering pattern. In contrast, the *clothing_accessories* category exhibits multiple distinct clusters, reflecting a wider variety of recall reasons, such as choking hazards due to detachable components. This divergence underscores how certain domains exhibit unique, domain-specific risks, while others exhibit larger diversity in risk possibilities.

Figure 8b highlights a case where recall descriptions from different domains show considerable overlap. Specifically, the categories *baby_products* and *sports_recreation* share similarities in recall descriptions, often referencing issues related to unsecured assembly constraints that pose risks of falling or entrapment. Despite the differing nature of these product types, the commonality in risk profiles suggests valuable cross-domain learning opportunities; studying hazards across domains allows designers to anticipate failure modes that might otherwise remain overlooked within their own field.

To contextualize these findings, we calculate the spread of recall descriptions for each product category using Convex Hull analysis. Table 5 reports the normalized areas for each category, offering a metric for the diversity of recall descriptions. This approach, informed by prior work on diversity metrics in embedding spaces [49], enables a comparative assessment of risk variability across categories. Notably, the *toys_children* category exhibits the largest normalized area, indicating a wide range of distinct hazards associated with children's products. This reinforces the need for cautious safety considerations in the design of children's toys.

4.4.2 Three-Dimensional Embedding Space. We demonstrate the potential to move beyond passive data exploration by demonstrating an interactive visualization approach that situates new product ideas within the context of historical recall data (see Fig. 9). In this example, we embed all product descriptors, specifically *product_name*, into a 384-dimensional latent space. For visualization purposes, we reduce the dimensionality to three dimensions using t-SNE, although a similar analysis could be conducted in two dimensions as well.

When a designer inputs a new product concept, the system first embeds the input text into the same high-dimensional semantic space as the recalled products. For visualization purposes, the new embedding is then combined with the existing recall embeddings, and a low-dimensional layout is recomputed to display the new concept alongside past recalled products. Specifically, both recalled products and new product concepts are encoded using the same Sentence-BERT model (`all-MiniLM-L6-v2`) to obtain fixed-length vector representations, which are subsequently scaled using the same feature normalization fitted on the recall dataset. To visualize the augmented set, the new scaled embedding is concatenated with the existing scaled recall embeddings, and t-SNE is re-fit on the combined set using fixed hyperparameters and a fixed random seed. Because t-SNE does not provide a direct out-of-sample mapping, previously plotted points may shift slightly when a new concept is introduced. Accordingly, the visualization is intended for qualitative, exploratory analysis rather than as a fixed coordinate system.

We chose to embed *product_name* as it typically conveys precise, yet high-level semantic information, making it particularly suitable for early-stage ideation. As an example, an engineering team might propose a product idea described simply as *"a bike suspension made of aluminum alloy"* without fully developed specifications. This approach allows the user to position that idea within the landscape of similar historical recalls, providing insight into neighboring products and their associated hazard classifications. By visualizing these relationships interactively, designers can proactively identify potential risks, explore relevant precedents, and make more informed design decisions. Altogether, this embedding approach transforms a static archive of recall records into an interactive design resource. By visualizing relationships among products and hazards, it complements traditional analytical methods and demonstrates how embedding-based representations can support exploration, comparison, and hazard awareness in safety-critical design contexts.

Table 5 Recall description diversity of each product category as measured by Convex Hull area in 2D scaled embedding space.

Category	Normalized Area
TOYS_CHILDREN	11.848
SPORTS_RECREATION	11.009
TRANSPORTATION	10.486
HOME_APPLIANCES	10.169
TOOLS_HARDWARE	9.776
OUTDOOR_EQUIPMENT	9.485
BABY_PRODUCTS	9.219
FURNITURE	9.111
OTHER	8.872
CLOTHING_ACCESSORIES	8.712
ELECTRICAL	6.688

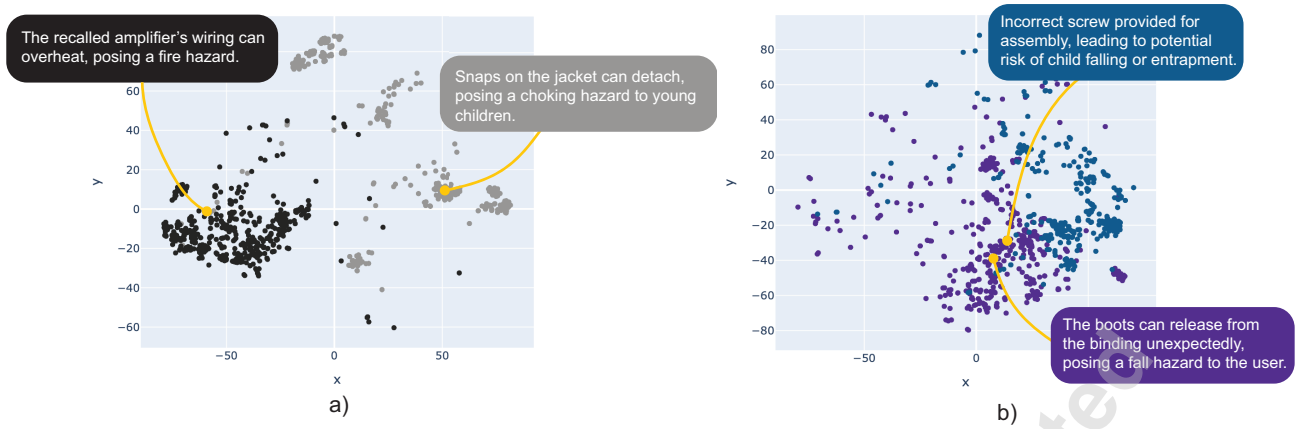

Fig. 8 a) Embedded recall descriptions of *electrical* (black) and *clothing_accessories* (gray). b) Embedded recall descriptions of *baby_products* (blue) and *sports_recreation* (purple). Descriptions are denoted for each example, showcasing distant (a) and near (b) recall descriptions across different product categories.

5 Limitations

While this work demonstrates the feasibility of using recall data for hazard classification, prediction, and discovery, several limitations are worth noting. The dataset is currently limited to recalls reported by the CPSC, which constrains the breadth of hazards represented. Extending to international sources such as the OECD Global Recalls Portal [28] would provide a more comprehensive picture of product risks across regulatory environments.

A core limitation comes from the role of LLMs in augmenting dataset fields. Generative models enabled us to enrich product descriptions and classifications at scale, but natural language remains variable and sometimes ambiguous. Although our human annotation study showed strong agreement with these generated fields, additional validation will be important to ensure their reliability when used in practice. Moreover, the outputs of LLMs are sensitive to prompt design; refining prompts or systematically comparing alternative prompting strategies could help quantify how much these choices affect downstream scores and overall dataset quality.

Our pipeline does not include formal uncertainty quantification at the per-prediction level. We report aggregate accuracy with 95% confidence intervals over a fixed evaluation set, but do not provide calibrated confidence scores or risk-coverage analyses for individual predictions. As discussed in Section 4.3.2, this design reflects the intended use case of the workflow as a candidate-hazard discovery tool for designer review rather than as an autonomous classifier with abstention. Per-prediction uncertainty estimates – derived from multi-run consistency, model-reported logprobs, or cross-model agreement – and accompanying risk-coverage analyses are a valuable direction for future work, particularly in studies that evaluate the workflow with practicing designers.

Our evaluation metrics capture only part of the challenge of hazard prediction. Relaxed accuracy, Top-k scoring, and LLM-as-Judge provide useful signals, but they cannot fully account for the subtleties of hazard reasoning or design context. Similarly, while we illustrate potential applications of embeddings and predictive models, this study does not yet evaluate their effectiveness with practicing designers. Validating their effectiveness in real-world design and safety processes remains an open opportunity.

6 Future Work

In addition to analyzing past recalls, this dataset and classification pipeline could also be used proactively to improve future product design. For instance, integrating recall-informed hazard classifications into early-stage product requirement generation may improve safety considerations from the outset. This could involve coupling the dataset with LLM generated user requirements to ensure risk factors identified in past recalls are embedded into design requirements.

The dataset can further support developing detailed user personas based on specific recall incidents and hazard types (integrating into works such as [50]). By examining product recalls through the lens of potential user interactions, designers could construct user personas representative of individuals most likely to encounter or exacerbate certain hazards. For example, examining products recalled due to choking hazards might inform the creation of personas representing families with young children or elderly individuals with limited mobility, allowing designers to conduct roleplay analysis and proactively consider how different user behaviors and demographics might interact with products to induce hazardous situations.

While the dataset is multimodal and provides both text and visual descriptions of the products, 3D models of the products were not collected by the CPSC, or any other global entity responsible for handling product recalls. Collecting this data modality could further improve the models' ability to predict hazards by providing more context about the design, although passing this data to the model efficiently would pose technical challenges.

Future work also includes per-prediction uncertainty quantification and risk-coverage analyses for deployment contexts, as discussed in Section 5.

7 Conclusion

This work contributes the development of the RECALL-MM dataset, a curated collection of 6,874 historical product recalls that integrates direct, modified, categorized, and generated fields for computational analysis. The dataset supports a three-stage workflow for hazard classification, prediction, and discovery, together with an evaluation framework designed to characterize where computational methods - both generative and non-generative - can support risk-informed early-stage design. We advocate for integrating publicly available recall datasets into early product development, where risk identification is often most critical yet underinformed by historical evidence.

Comparison against non-GenAI retrieval and classification baselines, in both text and image modalities, clarifies where each approach contributes. Classical methods such as Sentence-BERT k-NN on textual product descriptions and ResNet50 k-NN on raw images achieve higher single-label Top-k accuracy than the five evaluated VLMs by exploiting redundancy in the recall archive. At the same time, the strongest foundation models still identify

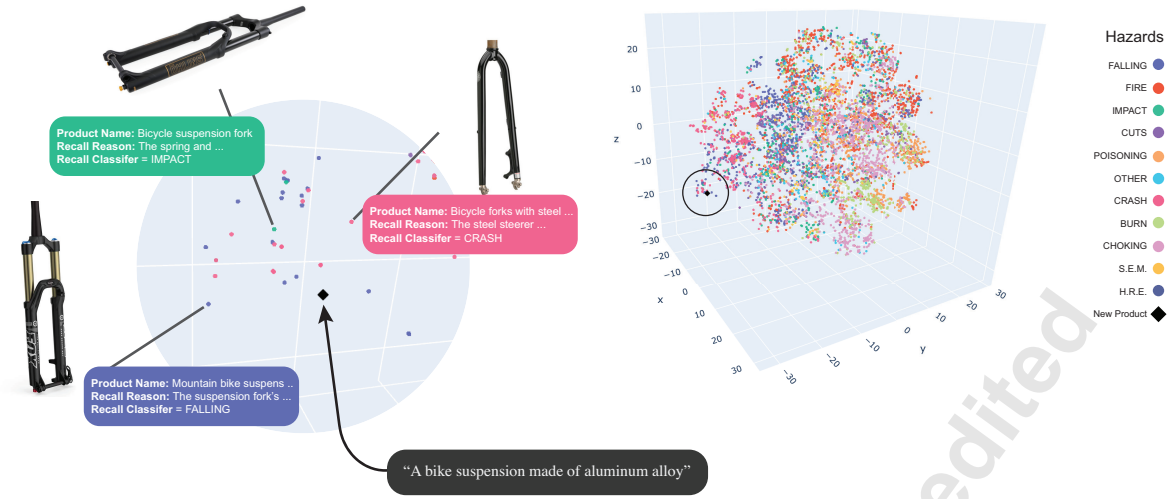

Fig. 9 Three-dimensional embedding of product names, with markers colored by hazard class. A new product (black diamond) is projected into the space and shown to cluster with recalled products of similar hazard types, illustrating embedding-based hazard discovery.

the correct hazard among three plausible guesses approximately two-thirds of the time, from a single image with no labeled training corpus. The near-equivalence between image-feature retrieval and text-description retrieval further suggests that visual interpretation carries most of the predictive signal for hazard category – an observation that motivates continued development of multimodal models capable of reasoning over product images directly.

Beyond classification, generative methods enable capabilities that classical baselines cannot produce: long-form failure-mechanism and failure-mode narratives, natural-language workflows for interacting with recall archives at scale, and integration with embedding-based discovery interfaces that situate new product concepts within historical recall landscapes. This research lays groundwork for risk-informed computational tools in design workflows, where historical recall data and generative methods can each play complementary roles in proactive hazard mitigation.

Acknowledgment

Authors DB and KGL would like to acknowledge financial support from the Society of Hellman Fellows and Autodesk Research. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the aforementioned sponsors. A preliminary version of this study was presented at the 2025 ASME IDETC – Design Automation Conference [51].

References

[1] Schuh, G., Gartzen, T., Soucy-Bouchard, S., and Basse, F., 2017, "Enabling agility in product development through an adaptive engineering change management," Procedia cirp, **63**, pp. 342–347.
[2] Lough, K. G., Stone, R., and Tumer, I. Y., 2009, "The risk in early design method," Journal of engineering design, **20**(2), pp. 155–173.
[3] Department of Defense, 1980, "Procedures for Performing a Failure Mode, Effects, and Criticality Analysis," Department of Defense, Washington, DC, Military Standard MIL-STD-1629A.
[4] Vesely, W. E., Goldberg, F. F., Roberts, N. H., and Haasl, D. F., 1981, "Fault Tree Handbook," U.S. Nuclear Regulatory Commission, Rockville, Maryland, NUREG NUREG-0492.
[5] Schüeller, G. I., ed., 1999, *Safety and Reliability: Proceedings of ESREL '99, the Tenth European Conference on Safety and Reliability, Munich-Garching, Germany, 13–17 September 1999*, Vol. 1, Balkema, Rotterdam.
[6] Bohnenblust, H. and Schneider, T., 1987, "Risk Appraisal—Can It Be Improved by Formal Decision Models?" *Uncertainty in risk assessment, risk management, and decision making*, Springer, pp. 71–87.
[7] Hora, M., Bapuji, H., and Roth, A. V., 2011, "Safety hazard and time to recall: The role of recall strategy, product defect type, and supply chain player in the US toy industry," Journal of Operations Management, **29**(7-8), pp. 766–777.
[8] Jaimovitch-López, G., Ferri, C., Hernández-Orallo, J., Martínez-Plumed, F., and Ramírez-Quintana, M. J., 2023, "Can language models automate data wrangling?" Machine Learning, **112**(6), pp. 2053–2082.
[9] OpenAI, 2024, "GPT-4o," https://openai.com/gpt-4o, Accessed: 2025-03-21.
[10] U.S. Consumer Product Safety Commission, 2025, "CPSC Recalls Database," Accessed: 2025-03-04, https://www.cpsc.gov/Recalls
[11] Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., and Yu, F., 2015, "ShapeNet: An Information-Rich 3D Model Repository," Stanford University — Princeton University — Toyota Technological Institute at Chicago, Tech. Rep. arXiv:1512.03012 [cs.GR].
[12] Koch, S., Matveev, A., Jiang, Z., Williams, F., Artemov, A., Burnaev, E., Alexa, M., Zorin, D., and Panozzo, D., 2019, "ABC: A Big CAD Model Dataset For Geometric Deep Learning," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[13] Wu, R., Xiao, C., and Zheng, C., 2021, "Deepcad: A deep generative network for computer-aided design models," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6772–6782.
[14] Willis, K. D., Jayaraman, P. K., Chu, H., Tian, Y., Li, Y., Grandi, D., Sanghi, A., Tran, L., Lambourne, J. G., Solar-Lezama, A., et al., 2022, "Joinable: Learning bottom-up assembly of parametric cad joints," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15849–15860.
[15] Bian, S., Grandi, D., Liu, T., Jayaraman, P. K., Willis, K., Sadler, E., Borijin, B., Lu, T., Otis, R., Ho, N., et al., 2024, "HG-CAD: hierarchical graph learning for material prediction and recommendation in computer-aided design," Journal of Computing and Information Science in Engineering, **24**(1), p. 011007.
[16] Kim, S., Chi, H.-g., Hu, X., Huang, Q., and Ramani, K., 2020, "A large-scale annotated mechanical components benchmark for classification and retrieval tasks with deep neural networks," *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, Springer, pp. 175–191.
[17] Regenwetter, L., Curry, B., and Ahmed, F., 2022, "BIKED: A dataset for computational bicycle design with machine learning benchmarks," Journal of Mechanical Design, **144**(3), p. 031706.
[18] Elrefaie, M., Morar, F., Dai, A., and Ahmed, F., 2024, "DrivAerNet++: A Large-Scale Multimodal Car Dataset with Computational Fluid Dynamics Simulations and Deep Learning Benchmarks," *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, eds., Vol. 37, Curran Associates, Inc., pp. 499–536, https://proceedings.neurips.cc/paper_files/paper/2024/file/013cf29a9e68e4411d0593040a8a1eb3-Paper-Datasets_and_Benchmarks_Track.pdf
[19] Toh, C. A. and Miller, S. R., 2016, "Creativity in design teams: the influence of personality traits and risk attitudes on creative concept selection," Research in Engineering Design, **27**(1), p. 73–89.
[20] Gryaditskaya, Y., Sypesteyn, M., Hoftijzer, J. W., Pont, S., Durand, F., and Bousseau, A., 2019, "OpenSketch: a richly-annotated dataset of product design sketches," ACM Trans. Graph., **38**(6), pp. 232:1–232:16.
[21] Zhu, Q. and Luo, J., 2022, "Generative Pre-Trained Transformer for Design

Concept Generation: An Exploration," Proceedings of the Design Society, **2**, p. 1825–1834.
[22] Meltzer, P., Lambourne, J. G., and Grandi, D., 2023, "What's in a Name? Evaluating Assembly-Part Semantic Knowledge in Language Models Through User-Provided Names in Computer Aided Design Files," Journal of Computing and Information Science in Engineering, **24**(011002).
[23] Jiang, S., Sarica, S., Song, B., Hu, J., and Luo, J., 2022, "Patent Data for Engineering Design: A Critical Review and Future Directions," Journal of Computing and Information Science in Engineering, **22**(060902).
[24] Goucher-Lambert, K. and Cagan, J., 2019, "Crowdsourcing inspiration: Using crowd generated inspirational stimuli to support designer ideation," Design Studies, **61**, p. 1–29.
[25] Lin, J., Huang, D., Zhao, T., Zhan, D., and Lin, C.-Y., 2024, "Design-Probe: A Graphic Design Benchmark for Multimodal Large Language Models," (arXiv:2404.14801), arXiv:2404.14801 [cs].
[26] Doris, A. C., Grandi, D., Tomich, R., Alam, M. F., Ataei, M., Cheong, H., and Ahmed, F., 2025, "Designqa: A multimodal benchmark for evaluating large language models' understanding of engineering documentation," Journal of Computing and Information Science in Engineering, **25**(2), p. 021009.
[27] Achlioptas, P., Huang, I., Sung, M., Tulyakov, S., and Guibas, L., 2023, "ShapeTalk: A Language Dataset and Framework for 3D Shape Edits and Deformations," *Conference on Computer Vision and Pattern Recognition (CVPR)*.
[28] Organisation for Economic Co-operation and Development (OECD), n.d., "Global Recalls Portal," https://globalrecalls.oecd.org/#/, Accessed: 2025-03-08.
[29] Durrett, J., 2015, "A decade of data: an in-depth look at 2014 and a ten-year retrospective on children's product recalls," Kids In Danger.
[30] Niven, C. M., Mathews, B., Harrison, J. E., and Vallmuur, K., 2020, "Hazardous children's products on the Australian and US market 2011–2017: an empirical analysis of child-related product safety recalls," Injury prevention, **26**(4), pp. 344–350.
[31] Kirschman, K. B. and Smith, G. A., 2007, "Resale of recalled children's products online: an examination of the world's largest yard sale," Injury Prevention, **13**(4), pp. 228–231.
[32] Wai, H. T. and Uttama, S., 2024, "Effectiveness of Classifying Unsafe Children's Toys Using NLP, Deep Learning and Ensemble Learning," *2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE)*, IEEE, pp. 261–267.
[33] Yorulmuş, M. H., Bolat, H. B., and Bahadır, Ç., 2022, "Predictive quality defect detection using machine learning algorithms: A case study from automobile industry," *Intelligent and Fuzzy Techniques for Emerging Conditions and Digital Transformation: Proceedings of the INFUS 2021 Conference, held August 24-26, 2021. Volume 2*, Springer, pp. 263–270.
[34] O'Donnell, J. and Yoon, H.-S., 2024, "Determination of Multi-Component Failure in Automotive System Using Deep Learning," Journal of Computing and Information Science in Engineering, **24**(2).
[35] Stamatis, D. H., 2003, *Failure mode and effect analysis: FMEA from theory to execution*, Quality Press.
[36] Carlson, C. S., 2012, *Effective FMEAs: achieving safe, reliable, and economical products and processes using failure mode and effects analysis*, John Wiley & Sons.
[37] Modarres, M., 2006, *Risk analysis in engineering: techniques, tools, and trends*, CRC press.
[38] Bowles, J. B. and Peláez, C. E., 1995, "Fuzzy logic prioritization of failures in a system failure mode, effects and criticality analysis," Reliability engineering & system safety, **50**(2), pp. 203–213.
[39] Liu, H.-C., Liu, L., and Liu, N., 2013, "Risk evaluation in failure mode and effects analysis with extended VIKOR method under fuzzy environment," Expert Systems with Applications, **40**(5), pp. 1795–1803.
[40] Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al., 2024, "A survey on llm-as-a-judge," arXiv preprint arXiv:2411.15594.
[41] He, K., Zhang, X., Ren, S., and Sun, J., 2016, "Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.
[42] Reimers, N. and Gurevych, I., 2019, "Sentence-bert: Sentence embeddings using siamese bert-networks," arXiv preprint arXiv:1908.10084.
[43] Van der Maaten, L. and Hinton, G., 2008, "Visualizing data using t-SNE." Journal of machine learning research, **9**(11).
[44] Abdi, H. and Williams, L. J., 2010, "Principal component analysis," Wiley interdisciplinary reviews: computational statistics, **2**(4), pp. 433–459.
[45] McInnes, L., Healy, J., and Melville, J., 2018, "Umap: Uniform manifold approximation and projection for dimension reduction," arXiv preprint arXiv:1802.03426.
[46] Taylor, M. J., Fielding, J., and O'Boyle, J., 2024, "Electrical Home Fire Injuries Analysis," Fire, **7**(12), p. 471.
[47] Anwar, S. T., 2014, "Product recalls and product-harm crises: A case of the changing toy industry," Competitiveness Review: An International Business Journal incorporating Journal of Global Competitiveness, **24**(3), pp. 190–210.
[48] Kübler, R. V. and Albers, S., 2012, "The impact of product recall communication on brand image, brand attitude and perceived product quality," Brand Attitude and Perceived Product Quality (May 16, 2012).
[49] Ma, K., Grandi, D., McComb, C., and Goucher-Lambert, K., 2025, "Do Large Language Models Produce Diverse Design Concepts? A Comparative Study with Human-Crowdsourced Solutions," Journal of Computing and Information Science in Engineering, **25**(2).
[50] Ataei, M., Cheong, H., Grandi, D., Wang, Y., Morris, N., and Tessier, A., 2025, "Elicitron: A Large Language Model Agent-Based Simulation Framework for Design Requirements Elicitation," Journal of Computing and Information Science in Engineering, **25**(2).
[51] Bolanos, D., Ataei, M., Grandi, D., and Goucher-Lambert, K., 2025, "RECALL-MM: A Multimodal Dataset of Consumer Product Recalls for Risk Analysis Using Computational Methods and Large Language Models," *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 89237, American Society of Mechanical Engineers, p. V03BT03A050.

Appendix A

Table 6 Example of a recall entry from the curated dataset.

Field	Value
recall_number	1751
recall_date	2014-08-20
recall_description	The length adjustment buckles release unexpectedly, causing the item being stored to fall and injure people nearby.
product_name	Kayak and watersports storage hanger
product_quantity	10,000
remedies	Consumers should stop using the recalled storage hangers and return them to the place of purchase for a full refund or replacement.
visual_product_description	The product consists of a pair of straps made from blue and black fabric, each approximately 1 inch wide and 84 inches long when unbuckled. They feature plastic snap buckles for length adjustment and plastic-coated steel S-hooks for hanging.
product_classification	SPORTS_RECREATION
hazard_classification	FALLING
remedies_classification	REPLACE
product_image	

Appendix B

Table 7 Hazard classifications and definitions.

Hazard Classification	Definition
Fire	Use of the product may lead to a fire or the product violates federal fabric flammability regulations.
Burn	Use of the product may lead to experiencing burns.
Heat-Related Explosion (H.R.E.)	The product may explode unintentionally.
Falling	Use of the product may cause an unintentional fall.
Poisoning	Use of the product may lead to poisoning.
Crash	Use of the product may lead to an unintentional crash.
Choking	Use of the product may lead to choking, or the product violates federal toy safety standards, or the product violates federal children clothing standards (drawstrings).
Cuts	Use of the product may lead to unintentional cuts and/or lacerations.
Safety Equipment Malfunction (S.E.M.)	The safety product does not operate as intended and use of the product may lead to injury or death.
Impact	Use of the product may lead to an unintentional impact that may cause injury or death.
Other	Other

Table 8 Remedy classifications and definitions.

Remedy Classification	Definition
Refund	A customer may receive a full or partial refund, or gift card for the recalled product.
Repair	The company is offering a repair to the recalled product.
Replace	The company is offering a replacement for the recalled product in the form of a new product or other products of similar value.
Dispose	The product should be thrown out or recycled.
New Instructions (N.I.)	The company will issue new instructions on how the customer can make the recalled product safe.
Remedy No Longer Available (R.N.L.A.)	The recalled product should be thrown out or recycled.

Appendix C

VLM Hazard Prediction from Image Prompt

```
Analyze this product image and identify potential safety hazards. Consider each of these hazard categories:

FIRE,
BURN,
FALLING,
HEAT_RELATED_EXPLOSION,
POISONING,
CRASH,
CHOKING,
CUTS,
IMPACT,
SAFETY_EQUIPMENT_MALFUNCTION,
OTHER


For each hazard category, determine if it's a realistic concern for this product. Then select the TOP 3
most probable hazards and provide specific predictions.

  • A specific component or part that could fail

  • The exact failure mode or hazard mechanism

  • Why this poses a safety risk

Format each prediction as: [Component/part] could [specific failure mode], posing [EXACTLY ONE hazard type
from the list above] risk because [mechanism].

IMPORTANT: Use exactly one hazard type from the predefined list. Do not combine categories (e.g., don't
use "FALLING/IMPACT" - pick either "FALLING" or "IMPACT").

Return exactly 3 predictions, numbered as PREDICTION 1, PREDICTION 2, PREDICTION 3. Do not include any
additional commentary or questions.
```

LLM-as-Judge Prompt

```
Compare this recall description with this prediction:

RECALL: {recall_description}
PREDICTION: {prediction_text}

Score the prediction accuracy:

WHAT FAILED (component/part):
3 = Exact match (same part name)
2 = Related part (subassembly, nearby component, same system)
1 = Wrong/unrelated part

HOW DID IT FAIL (failure mode):
3 = Exact match (same failure mechanism)
2 = Related failure (same family, similar process, different trigger)
1 = Wrong/unrelated failure

Return JSON only:
{{"accuracy": {"what_failed": [1-3], "how_failed": [1-3]}, "rationale": "Compare the actual component
names and failure modes from both texts above, explaining why each score was given."}}
```

Journal of Computing and Information Science in Engineering. Received October 01, 2025;
Accepted manuscript posted July 29, 2026. doi:10.1115/1.4072427

Appendix D

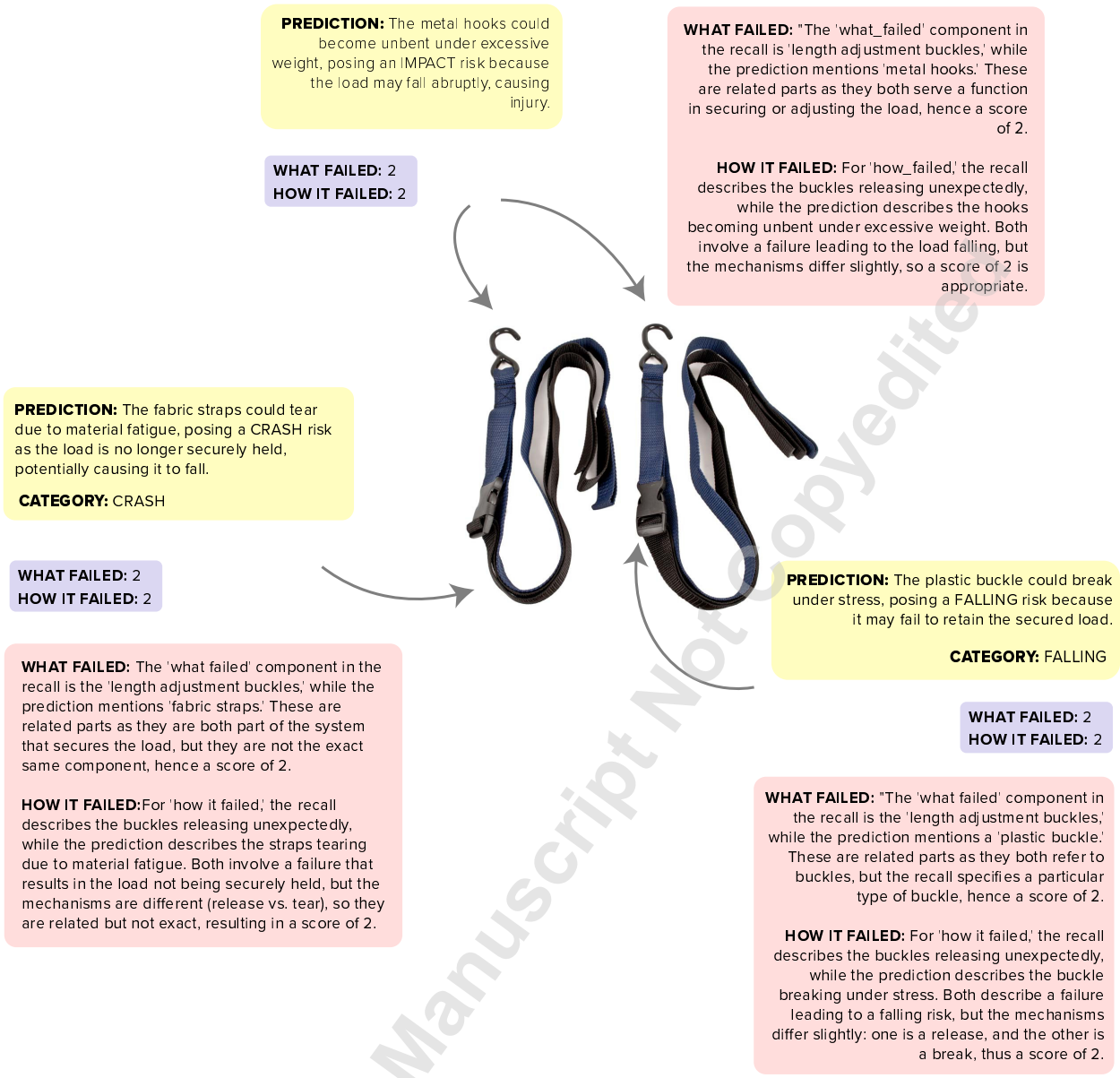

Fig. 10 Example of LLM-as-Judge evaluation of VLM hazard predictions for a recalled storage hanger. Three distinct predictions (metal hooks, fabric straps, plastic buckle) are compared against the recall description of length adjustment buckles releasing unexpectedly. Each prediction is scored on a two-axis rubric for what failed and how it failed, with scores of 2 assigned for partial alignment.

Appendix E

Reproducibility and Implementation Details. To support reproducibility, we report model versions, sampling parameters, and implementation details below.

Call windows. Experiments using GPT-4o were conducted in two distinct windows. (1) Dataset curation (Section 3.1) and the LLM hazard classification experiment (Section 3.2 / Section 4.3.1) were run in approximately March 2025. (2) The VLM hazard prediction experiments across all five models (Section 3.3 / Section 4.3.2) were run in approximately late September 2025.

OpenAI models.. GPT-4o, GPT-4o-mini, and GPT-5 were accessed via the OpenAI Python SDK. The model identifiers used were `gpt-4o`, `gpt-4o-mini`, and `gpt-5-chat-latest`. The `gpt-4o` alias was used in both the March 2025 dataset curation and LLM classification runs and the September 2025 VLM runs; during these windows, the alias served snapshot `gpt-4o-2024-08-06` (the snapshot active from August 2024 until OpenAI's November 2025 update). The `gpt-4o-mini` alias during our September 2025 VLM window served snapshot `gpt-4o-mini-2024-07-18`. The `gpt-5-chat-latest` identifier is a rolling alias that OpenAI updates without versioning to point to the current GPT-5 chat variant; the underlying weights at our late September 2025 call window are not independently retrievable. All calls used default sampling parameters (temperature = 1.0, top-p = 1.0, no explicit `max_tokens` limit, no seed set). JSON-formatted output was requested via the prompt only; we did not enable structured-output enforcement via `response_format`.

Gemini. Gemini 1.5 Flash was accessed via the Google Generative AI Python SDK using the `gemini-1.5-flash` alias. The specific snapshot served by Google for this alias may have changed since our call window. Calls used default sampling parameters (temperature = 1.0, top-p = 0.95, top-k = 40, max output tokens = 8192). The Gemini SDK does not expose a deterministic seed parameter. We did not attempt to enforce reproducibility across repeated calls.

LLaVA 1.6. LLaVA 1.6 was served locally via Ollama (model tag `llava:7b`), running the Mistral-7B-based variant. Sampling used temperature = 0.0 (greedy decoding), `num_predict` = 400, and Ollama defaults for `top_p` (0.9) and `top_k` (40). No random seed was set, but greedy decoding renders outputs effectively deterministic.

Image preprocessing. For all VLM experiments, product images were submitted in their original CPSC JPEG format without client-side resizing or compression. Images were base64-encoded for the OpenAI and Ollama endpoints and submitted as in-memory PIL objects for Gemini. Server-side preprocessing (such as OpenAI's automatic detail-level selection) was left at default settings.

Baselines. The Sentence-BERT k-NN baseline used the `all-MiniLM-L6-v2` model from the `sentence-transformers` library, with $k = 10$ and cosine similarity over L2-normalized embeddings. The TF-IDF baseline used `scikit-learn`'s `TfidfVectorizer` with n-gram range $(1, 2)$ and `min_df` = 2, followed by multinomial `LogisticRegression` with `max_iter` = 2000 and `random_state` = 42. The ResNet50 baseline used `torchvision`'s `ResNet50_Weights.IMAGENET1K_V2` weights, with images resized to 256 pixels on the shorter edge, center-cropped to 224×224, and normalized with ImageNet statistics; 2048-dimensional features from the penultimate (avgpool) layer were used as input to both k-NN (same configuration as Sentence-BERT) and logistic regression (same configuration as TF-IDF).

Confidence intervals. All reported confidence intervals are 95% Wilson score intervals for the binomial proportion, computed as

$$\frac{\hat{p} + z^2/(2n) \pm z\sqrt{\hat{p}(1-\hat{p})/n + z^2/(4n^2)}}{1 + z^2/n}$$

with $z = 1.96$. We chose the Wilson interval over the Wald (normal-approximation) interval because it remains valid near the boundaries of the proportion range at our evaluation size of $n = 100$.

Handling of invalid model outputs. We performed a post-hoc audit of all VLM prediction outputs and LLM-as-Judge score files to characterize invalid outputs. Across all individual VLM prediction slots, 4.22% returned null or empty predictions and 0.07% returned an explicit error from image processing. For the LLM-as-Judge evaluations, 0.73% of expected fields were null and were excluded from the mean and best-of-three aggregations reported in Table 4; no item-level count mismatches occurred between prediction and judge-output records.

List of Figures

List of Tables