When Ground Truth Disagrees: A Human-in-the-Loop Audit of Annotation Errors in High-Stakes Crash Narratives

Md Sajjad Hossain[1], **Lin Li**[1], **Judy A. Perkins**[2], **John Clary**[3], **Joel Meyer**[3]
[1]Department of Computer Science [2]Department of Civil & Environmental Engineering
Prairie View A&M University, Prairie View, Texas
[3]Austin Transportation & Public Works, Austin, Texas
mhossain2@pvamu.edu, lilin@pvamu.edu, juperkins@pvamu.edu
{John.Clary, joel.meyer}@austintexas.gov

Abstract

Linguistic annotation of high-stakes narrative data is often constrained by data confidentiality, domain expertise, and the lack of large-scale multi-annotator pipelines. We present a human-in-the-loop framework for auditing annotation discrepancies in crash narratives, combining structured labels, narrative-based annotation, and expert adjudication. Using 9,387 crash reports, we conduct a multi-layer analysis of disagreement across annotation sources. Nearly half of the records (49.4%) exhibit discrepancies between structured and narrative labels, driven mainly by unsupported structured assignments. In contrast, narrative-based annotation achieves near-perfect agreement with adjudication ($\kappa = 0.990$), indicating strong consistency when grounded in textual evidence. We introduce a taxonomy of discrepancies, showing refinement opportunities and missing details are the most common, while linguistic factors such as hedging and underspecification contribute to ambiguity. We further show that annotator-reported uncertainty strongly predicts annotation difficulty, with uncertain records nearly nine times more likely to disagree with structured labels. These findings highlight limitations of administrative coding and support a scalable, uncertainty-guided annotation paradigm for restricted-access domains.

1 Introduction

Linguistic annotation of natural language corpora is a foundational component of modern NLP systems, enabling both supervised learning and evaluation of language understanding models. In high-stakes domains such as transportation safety, annotation quality is particularly critical (Klie et al., 2024; Kumar and Sangwan, 2025; Khasanah et al., 2025), as labeled data directly informs policy decisions, risk analysis, and system design. However, annotation in such settings is often constrained by data confidentiality, limited access to domain expertise, and reliance on administrative coding systems, which may not fully capture the underlying textual evidence (Di Martino et al., 2020; Salami, 2023; Campbell and Giadresco, 2020).

Crash reporting provides a representative example of this challenge. Police officers are given the option to assign structured contributing-factor codes to each crash, while accompanying narratives describe the event in natural language. Discrepancies frequently arise due to ambiguity, reporting practices, and limitations of predefined coding schemes. As a result, structured labels may introduce systematic biases, omit relevant information, or encode interpretations not replicated in the narrative. Recent work has established that human disagreement in annotation is pervasive across NLP tasks (Pavlick and Kwiatkowski, 2019; Uma et al., 2021; Weerasooriya, 2024), and that quality management practices remain inconsistent even in major dataset creation efforts (Klie et al., 2024). At the same time, traditional multi-annotator approaches are often infeasible in restricted-access domains, where sensitive data cannot be widely shared and annotation requires specialized knowledge.

In this work, we propose a human-in-the-loop annotation framework designed to audit and analyze discrepancies between structured administrative labels and narrative-derived interpretations. Our approach combines (i) structured codes assigned during reporting, (ii) narrative-based annotation performed independently from structured codes, and (iii) expert adjudication guided by explicit uncertainty marking. This design enables controlled comparison across annotation layers while preserving data security. We evaluate this framework on a dataset of 9,387 crash narratives and conduct a comprehensive analysis of annotation discrepancies. Our study is guided by the following research questions:

Proceedings of the 20th Linguistic Annotation Workshop (LAW XX), pages 241–256

- RQ1: How do discrepancies between structured administrative labels, narrative-based annotations, and adjudicated corrections reveal systematic annotation error and linguistic ambiguity in crash reporting?
- RQ2: What role does explicit uncertainty marking and human adjudication play in resolving annotation ambiguity and improving label consistency in high-stakes narrative corpora?

To address these questions, we perform five complementary analyses examining agreement patterns, discrepancy types, linguistic sources of ambiguity, and the role of uncertainty in annotation.

Our analysis reveals that 49.4% of records contain discrepancies between structured and narrative labels, with Distracted Driving exhibiting an 86.8% unsupported rate in structured codes. Narrative-based annotation achieves near-perfect agreement with adjudication ($\kappa = 0.990$), and annotator-reported uncertainty predicts structured-code disagreement with an odds ratio of 8.91 ($p < 10^{-124}$).

This work makes the following contributions:

- A human-in-the-loop annotation framework for restricted-access domains, combining single-annotator labeling, explicit uncertainty marking, and expert adjudication to ensure annotation quality under confidentiality constraints.
- A large-scale empirical analysis of annotation discrepancies across structured and narrative representations.
- A taxonomy of annotation discrepancies, distinguishing errors, omissions, ambiguity, and compression, and revealing class-specific patterns of annotation failure.
- An uncertainty-driven analysis of annotation difficulty, demonstrating that annotator-reported uncertainty strongly predicts both adjudication corrections and disagreement with structured codes.

2 Related Work

Annotation Error and Quality. Annotation quality directly shapes the reliability of NLP systems, yet systematic errors persist even in widely-used benchmarks. Northcutt et al. (2021) found an average of 3.4% label errors across 10 popular test sets, demonstrating that label noise is pervasive rather than exceptional. Klie et al. (2023) reimplemented 18 annotation error detection methods and evaluated them across 9 datasets, establishing that no single method reliably detects errors across tasks. Weber-Genzel et al. (2024) introduced VariErr NLI, a two-round protocol that formally separates annotation errors from legitimate human label variation. Swayamdipta et al. (2020) proposed Dataset Cartography, showing that training dynamics can identify ambiguous and mislabeled instances. These works focus on annotator-produced labels in controlled settings. Our work extends this line of research to administratively produced labels created under operational constraints.

Annotator Disagreement and Perspectivism. A parallel line of research argues that disagreement between annotators carries meaningful signal rather than noise. Plank (2022) articulated this position comprehensively, arguing that human label variation impacts all stages of the ML pipeline and that the assumption of a single ground truth is often inappropriate. This view has been formalized as data perspectivism (Basile et al., 2021). Leonardelli et al. (2023) operationalized this view through the LeWiDi shared task at SemEval-2023, promoting models that learn from disagreement rather than resolving it through majority vote. Davani et al. (2022) showed that multi-task models predicting individual annotator ratings outperform majority-vote approaches on subjective tasks. Frenda et al. (2024) surveyed perspectivist approaches to NLP, documenting growing adoption of non-aggregated annotation across subjective tasks. However, existing disagreement research focuses almost exclusively on annotator-vs-annotator comparisons. Our work introduces a different type of comparison—annotation system vs. annotation system.

Crash Narrative NLP. NLP has been increasingly applied to crash narratives for automated safety analysis. Jaradat et al. (2024) demonstrated that text mining can uncover contributing factors from crash reports, while Bhagat et al. (2025) found that even expert-aligned LLM evaluation reveals persistent divergence between model predictions and human judgment on crash narratives. Oliaee et al. (2023) applied BERT to injury classification from police reports. Despite this progress, all prior crash NLP work treats structured police codes as ground truth labels. No study has audited the annotation quality of these codes by comparing them

against independent narrative-derived labels—the gap our study addresses.

3 Methodology

We propose a multi-layer, human-in-the-loop annotation framework designed to systematically audit discrepancies between structured administrative labels and narrative-derived interpretations in high-stakes crash reporting. The overall workflow is illustrated in Figure 1.

3.1 Problem Formulation

Let the dataset be defined as

$$\mathcal{D} = \{(x_i, s_i)\}_{i=1}^{N} \tag{1}$$

where x_i denotes a crash narrative and $s_i \in \{0,1\}^K$ represents the structured contributing-factor labels assigned by reporting officers, with K contributing-factor categories.

Our objective is not to predict labels, but to analyze annotation consistency across three layers:

- s_i: structured labels (Layer 1)
- n_i: narrative-based annotations (Layer 2)
- a_i: adjudicated labels (Layer 3)

Each layer is represented as a multi-label vector:

$$s_i, n_i, a_i \in \{0,1\}^K \tag{2}$$

To quantify differences between layers, we define a set-level discrepancy function:

$$\Delta(s_i, a_i) = \{(s_i \setminus a_i),\ (a_i \setminus s_i)\} \tag{3}$$

where $s_i \setminus a_i$ denotes labels present in the structured set but absent from the adjudicated set (unsupported labels), and $a_i \setminus s_i$ denotes labels present in the adjudicated set but absent from the structured set (omissions).

3.2 Three-Layer Annotation Framework

To systematically study annotation discrepancies, we design a three-layer annotation process:

- **Layer 1 (Structured Labels):** Administrative contributing-factor codes selected by reporting officers as part of the crash report. These codes are not extracted from the narrative; they are recorded as separate structured fields in the reporting system.
- **Layer 2 (Narrative-Based Annotation):** Labels assigned independently using only officer written narrative text, with structured labels hidden to prevent bias.
- **Layer 3 (Adjudicated Labels):** Final labels obtained through expert-guided adjudication, representing an independent interpretation of narrative evidence.

This layered design enables controlled comparison between administrative reporting, text-derived interpretation, and expert validation, allowing us to isolate discrepancies arising from omission, ambiguity, and reinterpretation.

3.3 Annotation Protocol

Annotation was conducted over approximately 14 months by a trained annotator, with quality maintained through regular review sessions with domain specialists from the City of Austin Transportation and Public Works department. In this context, domain specialists are the co-authors of this paper, who are transportation safety professionals from Austin Transportation and Public Works with practical experience reviewing crash reports, interpreting contributing-factor codes, and using crash data for traffic safety analysis. During these sessions, annotation decisions were audited and guidelines were iteratively refined to ensure consistency and domain alignment. Unlike conventional multi-annotator setups, we adopt a single-annotator-with-expert-oversight paradigm, motivated by:

(a) data confidentiality constraints, (b) restricted access to sensitive records, and (c) the need for domain-specific expertise.

The annotation process follows three principles:

- **Narrative-Only Labeling:** Labels are assigned based solely on textual evidence.
- **Multi-Label Assignment:** Multiple contributing factors may be assigned to each narrative.
- **Explicit Uncertainty Marking:** Annotator flag cases where the narrative is ambiguous or insufficient for confident labeling.

Layer 2 annotation used only the officer-written narrative text; structured contributing-factor codes and other report fields were hidden to prevent label leakage. To support this workflow, we developed a custom annotation interface that presents narrative text to the annotator with structured codes hidden,

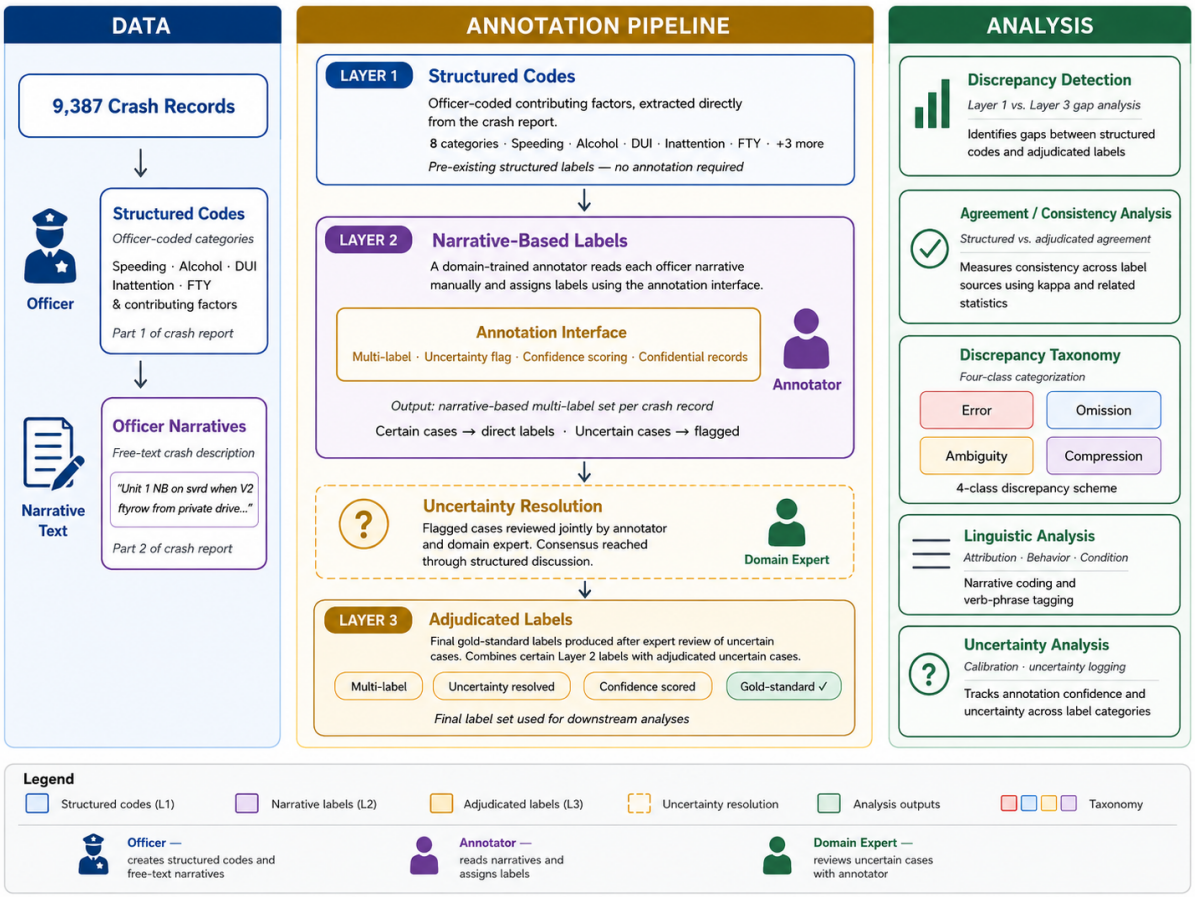


Figure 1: Overview of the proposed three-layer human-in-the-loop annotation framework. Structured officer labels (Layer 1) are contrasted with narrative-only annotations (Layer 2), while uncertain cases undergo expert adjudication to produce Layer 3. The resulting layers support discrepancy, taxonomy, linguistic, and uncertainty analyses.

supports multi-label selection, and includes an explicit uncertainty flag to capture ambiguous cases. A description of the interface and its design is provided in Appendix A.7.

3.4 Uncertainty-Guided Adjudication

Narratives marked as uncertain are routed to an adjudication stage, where the annotator and a domain expert jointly review the case. During adjudication, annotator has access to both the narrative and structured labels, allowing discrepancies to be examined and resolved in context.

Let $u_i \in \{0, 1\}$ denote the uncertainty flag for sample i, where $u_i = 1$ indicates ambiguity.

Adjudication is applied only to cases where the annotator explicitly marks uncertainty ($u_i = 1$). During adjudication, labels may be:

(i) confirmed, (ii) revised, or (iii) left unresolved if insufficient evidence exists.

This process produces the adjudicated label vector a_i, which serves as a reference for evaluating annotation consistency.

The overall annotation workflow can be summarized as:

$$x_i \to n_i, u_i = \phi(x_i, n_i), \quad a_i = \begin{cases} \psi(x_i, n_i, s_i), & \text{if } u_i = 1 \\ n_i, & \text{otherwise} \end{cases} \tag{4}$$

3.5 Analytical Framework

To address our research questions, we design five complementary studies:

- Study 1: Comparison between structured labels s_i and narrative-based annotations n_i
- Study 2: Comparison between narrative annotations n_i and adjudicated labels a_i
- Study 3: Taxonomic classification of discrepancies $\Delta(s_i, a_i)$
- Study 4: Linguistic analysis of ambiguity in narratives
- Study 5: Analysis of uncertainty signals u_i and their relationship to annotation correction

Each study provides a distinct perspective on annotation quality, discrepancy patterns, and ambiguity in high-stakes narrative data.

3.6 Label Space

We define a multi-label space with $K = 8$ contributing-factor categories derived from an original set of 72 fine-grained police contributing-factor codes. The mapping was developed with domain specialists. This consolidation reduces sparsity while preserving interpretability. The categories include Speeding, Impaired Driving, Distracted Driving, Failure to Yield, Red Light Running, Access-Related Crashes, Visual Obstruction, and Other. Because both structured labels and narrative-based annotations were converted to the same 8-class label space before comparison, the consolidation itself did not create cross-layer disagreements. Rather, discrepancies reflect differences between the officer-selected structured codes and the narrative-derived labels within the same shared category scheme.

Detailed definitions, annotation guidelines, and edge cases for each category are provided in the Appendix A.6.

3.7 Data

We evaluate our approach on a real world dataset of police crash narratives obtained from the City of Austin Transportation and Public Works Department (TPW). The dataset contains 9,387 records. Each record contains two annotation sources: (i) a structured contributing-factor field with up to five codes selected from 72 predefined options, and (ii) a free-text narrative describing the crash circumstances in natural language.

4 Results and Analysis

4.1 Study 1: Structured vs. Narrative Agreement

To address RQ1, we compare structured contributing-factor labels with narrative-based annotations at the set level.

Across all records, 50.6% (4,747) show full agreement, while 49.4% (4,640) exhibit at least one discrepancy, indicating that structured labels fail to capture narrative evidence in nearly half of all cases. As shown in Figure 2, omission (present in narrative but absent in structured labels) and unsupported labeling (present in structured labels without narrative evidence) patterns occur in 33.3% and 45.4% of records, respectively.

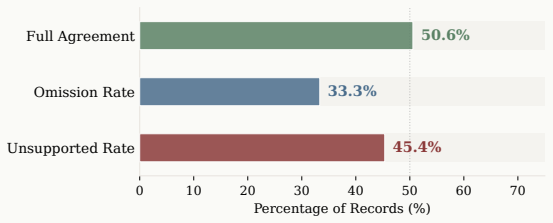


Figure 2: Overall structured vs. narrative agreement rates.

Agreement varies substantially across categories. As illustrated in Figure 3, the macro-average Cohen's κ is 0.446, indicating moderate overall agreement. Clearly defined behaviors such as Red Light Running ($\kappa = 0.696$), Failure to Yield ($\kappa = 0.677$), and Speeding ($\kappa = 0.624$) exhibit substantial agreement, while Distracted Driving ($\kappa = 0.167$) and Access-Related Crashes ($\kappa = 0.000$) show minimal to no agreement.

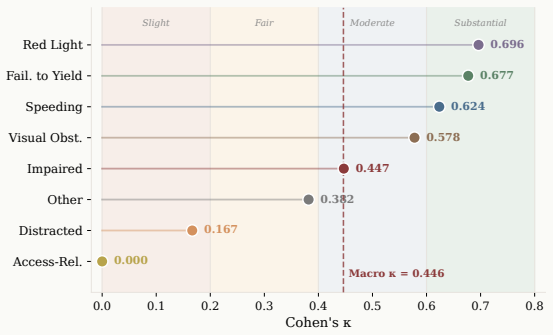


Figure 3: Per-class Cohen's κ with interpretation bands.

To further characterize disagreement, we distinguish between omission and unsupported labeling. Figure 4 shows that these patterns are highly asymmetric across categories.

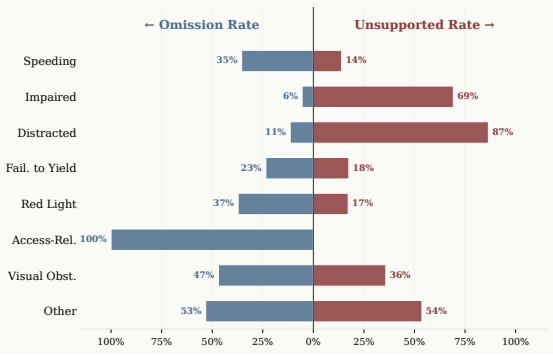


Figure 4: Per-class omission and unsupported rates.

In particular, Distracted Driving (87% unsup-

ported) and Impaired Driving (69% unsupported) are frequently assigned without supporting narrative evidence, while Access-Related Crashes (100% omission) and Red Light Running (37% omission) are often present in narratives but absent in structured codes.

A detailed confusion analysis in Appendix A.1 further reveals that certain structured labels consistently replace alternative contributing factors, indicating systematic substitution rather than random noise.

4.2 Study 2: Narrative vs. Adjudicated Agreement

To address RQ2, we compare narrative-based annotations with adjudicated labels.

Of all records, 852 (9.1%) were marked uncertain, while 8,535 (90.9%) were annotated with confidence.

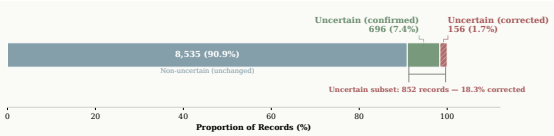


Figure 5: Distribution of annotation outcomes (uncertain vs. non-uncertain).

As shown in Figure 5, overall agreement between narrative and adjudicated labels is 98.3%, corresponding to a 1.7% correction rate across the full dataset.

This distribution indicates that uncertainty marking is conservative: The annotator flags potential ambiguity early, and most decisions remain valid under expert review. The observed corrections therefore reflect genuine annotation difficulty rather than systematic error.

Agreement remains consistently high across all categories. As reported in Table 1, the macro-average Cohen's κ is 0.990, indicating near-perfect agreement. All classes achieve $\kappa \geq 0.966$, with Visual Obstruction (1.000) showing perfect agreement and Other (0.966) exhibiting the lowest consistency.

The contrast with Study 1 ($\kappa = 0.446$) is substantial, indicating that disagreement primarily arises from structured administrative labels rather than inconsistencies in narrative-based interpretation. Adjudication produces targeted and asymmetric corrections. As illustrated in Figure 6, the largest net increase occurs for Failure to Yield

Category	κ
Visual Obstruction	1.000
Distracted Driving	0.998
Impaired Driving	0.997
Red Light Running	0.997
Access-Related Crashes	0.991
Speeding	0.989
Failure to Yield	0.983
Other	0.966
Macro Average	0.990

Table 1: Per-class agreement between narrative and adjudicated labels.

(+42), followed by Other (+24) and Speeding (+14), while Red Light Running shows a net decrease (-4).

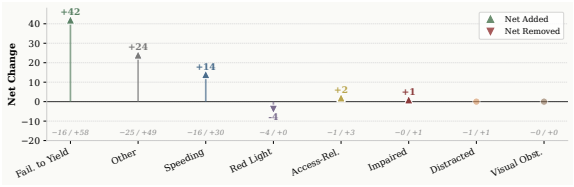


Figure 6: Net label changes during adjudication (additions vs. removals).

A detailed analysis of correction pathways (Appendix A.2) shows that adjudication primarily resolves ambiguous labels into more specific interpretations, with Other → Failure to Yield being the most frequent transition.

4.3 Study 3: Discrepancy Taxonomy

To move beyond aggregate agreement metrics, we classify discrepancies between structured codes and adjudicated labels into a four-category taxonomy aligned with the LAW special theme of errors in annotation: Error (unsupported structured label), Omission (missing structured label), Ambiguity (insufficient narrative evidence), and Compression (structurally unrepresentable cases). Categories are not mutually exclusive.

In this study, an unsupported label refers to a structured contributing-factor code that is present in Layer 1 but not supported by the narrative evidence after Layer 3 adjudication. This does not necessarily imply that the officer made a factual mistake; rather, it indicates that the structured code could not be verified from the narrative text under our annotation guidelines. Our analysis therefore treats the narrative as the evidence source for narrative-based annotation, not as a complete reconstruction of the officer's knowledge or intent during reporting.

Of all records, 50.6% (4,754) show no discrep-

Category	Records	% of All	% of Discrepancies
None	4,754	50.6%	—
Error	4,225	45.0%	91.2%
Omission	2,911	31.0%	62.8%
Ambiguity	799	8.5%	17.2%
Compression	238	2.5%	5.1%

Table 2: Discrepancy taxonomy distribution. Categories can overlap; percentages of discrepancies sum to more than 100%.

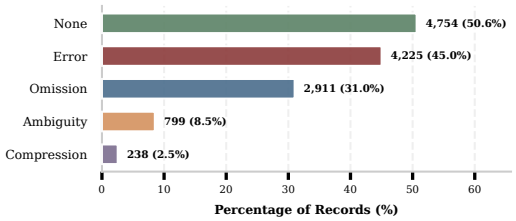


Figure 7: Distribution of discrepancy taxonomy categories.

ancy, while 49.4% (4,633) exhibit at least one discrepancy. As shown in Figure 7 and Table 2, Error is the dominant discrepancy type, appearing in 45.0% of all records and 91.2% of discrepant cases, followed by Omission (31.0%). Ambiguity (8.5%) and Compression (2.5%) occur less frequently but reflect qualitatively different failure modes.

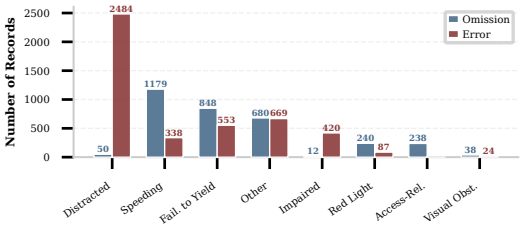


Figure 8: Per-class breakdown of error vs. omission discrepancies.

As shown in Figure 8, Distracted Driving is dominated by errors (2,484) with minimal omissions (50), confirming systematic over-coding in structured labels. Speeding and Failure to Yield show more balanced profiles, while Access-Related Crashes are entirely attributable to omission, reflecting a structural limitation of the coding system.

A consolidated view of annotation fragility is shown in Figure 9, which ranks classes by discrepancy rate. These results show that discrepancies are not uniformly distributed but concentrated in specific classes and driven by distinct mechanisms. In particular, structured over-coding is the dominant source of disagreement, indicating that annotation

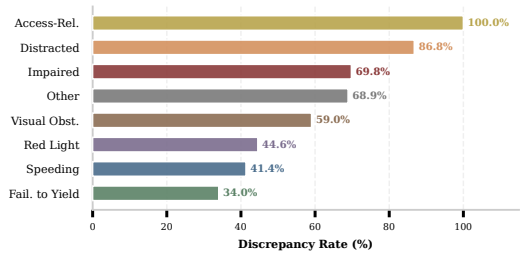


Figure 9: Per-class discrepancy rates across contributing factors.

inconsistencies arise primarily from unsupported structured labels rather than missing information.

A detailed per-class breakdown of discrepancy behavior across classes is provided in Appendix A.3.

4.4 Study 4: Linguistic Sources of Ambiguity

To understand why discrepancies arise, we identify five recurring linguistic phenomena that systematically introduce annotation ambiguity: hedging, implicit causality, underspecification, multi-event compression, and perspective markers. Here are some narrative examples.[1]

"Unit 1 stopped at a stop sign before entering the intersection. Unit 2 proceeded through on the ABC street. Both drivers indicated a stop sign lying on the ground that would have controlled Unit 2's lane."

Here, the downed sign's causal role is never confirmed, the narrative does not establish whether it fell before or after the collision (implicit causality), and the key evidence is attributed to the drivers rather than verified by the officer (perspective markers), leaving the annotator to choose between Failure to Yield and Other based on ambiguous evidence.

"Unit 1 was blocked in a parking stall. The driver possibly backed up slightly and may have collided with Unit 2, then went forward in a panic and struck Unit 3."

Here, hedging expressions ("possibly," "may have") signal that even the reporting officer could not confirm whether the first collision occurred, forcing the annotator to assign labels based on unverified events. Additionally, the narrative describes two separate collisions in sequence, but the

[1]Due to the data privacy agreement, all police narrative examples referenced in this paper are paraphrased and de-identified.

Metric	Uncertain	Non-uncertain
Correction Rate	18.3%	0.0%
Discrepancy Rate	88.1%	45.5%

Table 3: Correction rate and structured-adjudicated discrepancy rate by uncertainty status.

	Discrepancy	Agreement	Total
Uncertain	751	101	852
Non-uncertain	3,882	4,653	8,535
Total	4,633	4,754	9,387

Table 4: Cross-tabulation of uncertainty status and structured-adjudicated discrepancy.

structured coding system captures only a single contributing factor (multi-event compression).

"Unit 1 was traveling southbound in the far left lane when it struck the left side cable barrier."

Here, the narrative provides no explanation for why the vehicle departed its lane, the annotator cannot determine whether the cause was Speeding, Distracted Driving, Impaired Driving, or a mechanical failure, as the description lacks any behavioral or contextual detail (underspecification).

These phenomena are not random sources of noise but reflect systematic properties of narrative reporting based on incomplete information. Detailed examples and analysis for all five phenomena are provided in Appendix A.4.

4.5 Study 5: Uncertainty Analysis

We evaluate whether annotator-reported uncertainty reflects meaningful annotation difficulty. Among all records, 852 (9.1%) were marked uncertain, while 8,535 (90.9%) were annotated with confidence. As shown in Table 3, all observed corrections occur within the uncertain subset because adjudication was applied only to records explicitly marked as uncertain. Non-uncertain records were carried forward unchanged by design. Uncertain records also exhibit substantially higher disagreement with structured codes (88.1% vs. 45.5%).

This relationship is statistically significant. As summarized in Table 4, a chi-square test yields $\chi^2 = 562.37$ ($p < 10^{-124}$), with an odds ratio of 8.91, indicating that uncertain records are nearly nine times more likely to exhibit disagreement with structured labels.

Uncertainty is unevenly distributed across classes. As shown in Figure 10, the Other category exhibits the highest uncertainty rate (30.6%), followed by Failure to Yield (10.0%) and Visual Obstruction (9.9%). In contrast, Distracted Driving (1.8%) and Impaired Driving (4.5%) show low uncertainty despite high structured-code error rates (Study 3), indicating that these factors are linguistically explicit in narratives but systematically underrepresented in structured coding.

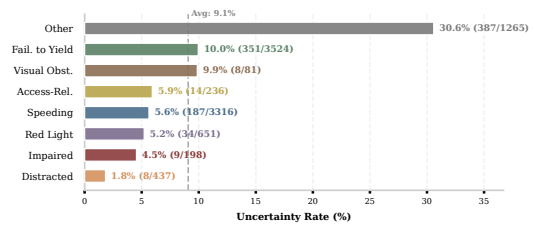


Figure 10: Per-class concentration of uncertainty in annotation.

5 Discussion

Our results show that annotation discrepancies in crash narratives are systematic rather than random. Importantly, these discrepancies should be interpreted as differences between administrative coding and narrative-grounded annotation, rather than as direct evidence of officer error. Structured crash codes may reflect reporting conventions, agency requirements, or contextual information available to the officer but not explicitly stated in the narrative. Label categories also differ in how directly they can be inferred from narrative text. For example, impaired driving may be supported by explicit evidence such as alcohol or drug involvement, whereas distracted driving often depends on more indirect behavioral descriptions such as inattention, failure to perceive hazards, or phone use. These differences help explain why some categories exhibit higher ambiguity or unsupported-label rates than others. Most disagreements are driven by unsupported structured labels, suggesting that administrative coding practices can introduce bias rather than simply omit information. At the same time, narrative-based annotation combined with uncertainty-guided adjudication achieves near-perfect agreement ($\kappa = 0.990$), demonstrating that reliable annotation is possible even with a single trained annotator when decisions are grounded in textual evidence and expert review. We further show that many disagreements arise from linguistic ambiguity, such as hedging and underspecification, highlighting limitations of purely rule-based annotation. Importantly, annotator-reported uncertainty emerges as a strong signal of annotation dif-

ficulty, effectively identifying cases that require expert adjudication. Together, these findings support a human-in-the-loop annotation paradigm for high-stakes, restricted-access domains. This framework generalizes to other high-stakes settings, such as clinical records, legal documents, and safety-critical reporting systems, where data access and domain expertise impose similar constraints.

6 Conclusion and Future Work

We present a human-in-the-loop framework for analyzing annotation discrepancies in crash narratives using structured labels, narrative annotation, and expert adjudication. Our results show that discrepancies are systematic and largely driven by unsupported structured labels and linguistic ambiguity. Across five studies, we demonstrate that narrative-based annotation achieves high consistency under expert review, that discrepancies can be organized through a clear taxonomy, and that annotator-reported uncertainty is a reliable signal of annotation difficulty. These findings highlight limitations of administrative coding and the importance of text-grounded, human-centered annotation. More broadly, this work offers a practical annotation paradigm for restricted-access domains, where traditional multi-annotator approaches are not feasible.

For future work, we plan to extend this framework in several directions. First, we aim to explore semi-automated annotation support, including models that predict uncertainty and assist annotators during labeling. Second, we will investigate error-aware training strategies that leverage discrepancy patterns to improve downstream predictive models. Third, subject to data-sharing constraints, we intend to release a curated subset of annotated data and the annotation tool to support reproducibility and further research.

7 Limitations

This study also has several limitations. First, annotation was carried out by a single trained annotator due to restricted data access and the need for domain expertise. Although this design was motivated by confidentiality constraints and the need for domain expertise, it may still introduce annotator-specific bias. To partially assess consistency, the annotator re-labeled a random subset of records after a time gap, and the results were reviewed for stability. Future work should extend this check to a larger systematically sampled subset and report formal intra-annotator agreement statistics. Instead of conventional multi-annotator redundancy, our framework relies on uncertainty-guided adjudication and expert verification. Second, the dataset is confidential and cannot be fully shared, which limits reproducibility, although we mitigate this by providing detailed guidelines and taxonomy definitions. Third, the taxonomy compresses 72 original contributing-factor codes into 8 classes, which may reduce some fine-grained distinctions. Finally, adjudication was applied only to uncertain cases, reflecting a practical workflow but potentially underestimating total disagreement.

Ethics Statement

This study uses 9,387 crash records provided by a municipal transportation agency under a formal data-sharing agreement for traffic safety research. As the data is administrative and not directly collected from human subjects, IRB approval was not required. All work complies with the ACL Code of Ethics. To protect privacy, no personally identifiable information was included in the dataset, all processing was conducted on secure local infrastructure, and no raw data was shared with external services or commercial models. All examples are paraphrased to prevent identification of real incidents. The system is designed for safety analysis, and while dual-use risks exist, its primary goal is to improve data quality for infrastructure-level decision-making.

Acknowledgments

The authors acknowledge the support provided by the U.S. Department of Transportation through the National Center for Infrastructure Transformation under grant numbers 69A3552344813 and 69A3552348318. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government.

References

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21.

Sudesh Ramesh Bhagat, Ibne Farabi Shihab, and Anuj Sharma. 2025. Accuracy is not agreement: Expert-aligned evaluation of crash narrative classification models. In *arXiv preprint arXiv:2504.13068*.

Sharon Campbell and Katrina Giadresco. 2020. Computer-assisted clinical coding: A narrative review of the literature on its benefits, limitations, implementation and impact on clinical coding professionals. *Health Information Management Journal*, 49(1):5–18.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Beniamino Di Martino, Fiammetta Marulli, Pietro Lupi, and Alessandra Cataldi. 2020. A machine learning based methodology for automatic annotation and anonymisation of privacy-related items in textual documents for justice domain. In *Conference on Complex, Intelligent, and Software Intensive Systems*, pages 530–539. Springer.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: A survey. *Language Resources and Evaluation*.

Shadi Jaradat, Taqwa I. Alhadidi, Huthaifa I. Ashqar, Ahmed Hossain, and Mohammed Elhenawy. 2024. Exploring traffic crash narratives in jordan using text mining analytics. *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*, pages 1–6.

Wirdatul Khasanah, Hale Yılmaz, and Benjamin White. 2025. The validity of automated essay scoring using NLP compared to human raters in the context of language certification exams. *JILTECH: Journal International of Lingua & Technology*, 4(3).

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–866.

Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023. Annotation error detection: Analyzing the past and present for a more coherent future. *Computational Linguistics*, 49(1):157–198.

Akshi Kumar and Saurabh Raj Sangwan. 2025. Introduction to natural language processing in high-stakes domains. In *Transformative Natural Language Processing: Bridging Ambiguity in Healthcare, Legal, and Financial Applications*, pages 1–22. Springer.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. Semeval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Amir Hossein Oliaee, Subasish Das, Jinli Liu, and M. Ashifur Rahman. 2023. Using bidirectional encoder representations from transformers (BERT) to classify traffic crash severity types. *Natural Language Processing Journal*, 3(S1).

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10671–10682, Abu Dhabi, United Arab Emirates.

Aishat O. Salami. 2023. Leveraging natural language processing to detect non-compliance in clinical documentation: Current advances, challenges, and future directions. *International Journal of Scientific Research in Science, Engineering and Technology*, pages 459–473.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2256–2269, Bangkok, Thailand.

Tharindu Cyril Weerasooriya. 2024. *Learning from Disagreement in Human-Annotated Datasets*. Ph.D. thesis, Rochester Institute of Technology.

A Appendix

A.1 Study 1: Extended Analysis of Structured vs. Narrative Discrepancies

This section provides additional analysis supporting the results presented in Section 4.1.

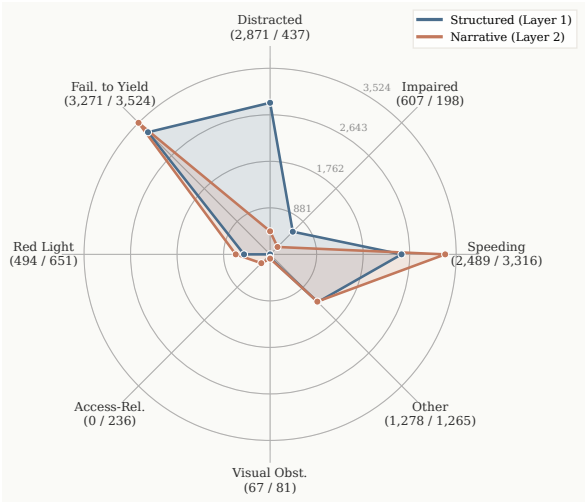


Figure 11: Label distribution comparison between structured and narrative annotations.

Label Distribution Analysis. Figure 11 compares the distribution of contributing-factor labels across structured and narrative annotations. The chart highlights substantial imbalances across several categories. In particular, Distracted Driving is heavily overrepresented in structured labels (2,871) compared to narrative annotations (437), while Access-Related Crashes are present in narratives (236) but entirely absent from structured coding. These patterns indicate systematic differences in how contributing factors are represented, with certain categories being over-assigned in structured data and others omitted entirely.

Confusion Matrix Analysis. Figure 12 provides a detailed view of label transitions between structured and narrative annotations. The matrix reveals systematic substitution patterns beyond simple disagreement. High agreement is observed for categories such as Speeding (91%) and Failure to Yield (88%), while Distracted Driving shows only 18% agreement. In cases of disagreement, Distracted Driving labels are frequently reassigned to other categories, most commonly:

- Speeding (32%)
- Failure to Yield (23%)

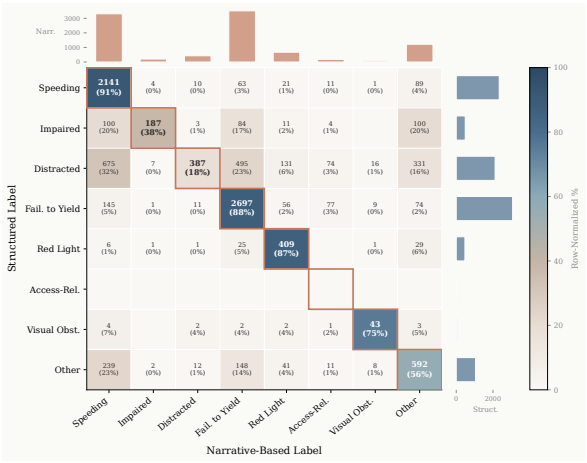


Figure 12: Row-normalized confusion matrix with marginal distributions: structured → narrative labels.

- Other (16%)

These patterns indicate that certain structured labels do not merely introduce noise but systematically replace alternative contributing factors.

Additional Observations.

- Categories with low agreement (e.g., Distracted Driving) correspond to high unsupported labeling rates observed in the main analysis.
- Categories with high omission rates (e.g., Access-Related Crashes) are consistently underrepresented in structured data.

A.2 Study 2: Extended Analysis of Adjudication Effects

This section provides additional analyses of label corrections and class-level transitions during adjudication, complementing the summary results in Section 4.2.

Figure 13 presents the correction flow matrix between narrative and adjudicated labels, revealing structured transition patterns during adjudication. The most frequent transition is from Other → Failure to Yield (17 cases), indicating that ambiguous cases are often resolved into more specific contributing factors. The reverse transition (Failure to Yield → Other, 7 cases) occurs less frequently, confirming that adjudication tends to refine rather than generalize labels.

A.3 Study 3: Extended Taxonomy Analysis

This section provides additional detail on the discrepancy taxonomy introduced in Section 4.3, in-

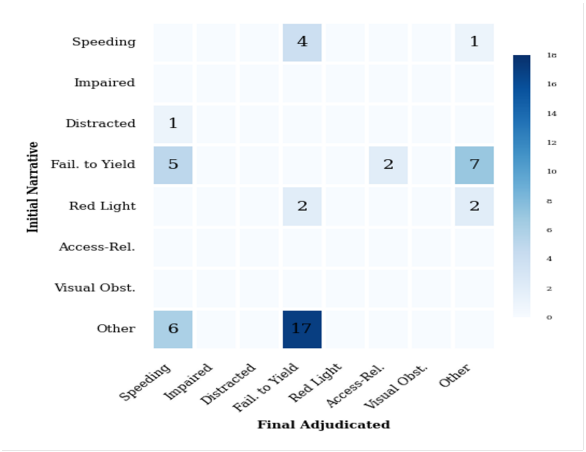


Figure 13: Correction flow between narrative and adjudicated labels. Each cell represents the number of transitions from an initial narrative label (rows) to a final adjudicated label (columns) within the uncertain subset.

cluding formal definitions and class-level breakdowns.

Taxonomy Definitions. Each discrepancy between structured and adjudicated labels is assigned one or more of the following categories:

- **Error:** A structured label is present but unsupported by narrative evidence.
- **Omission:** A contributing factor is present in the narrative but absent from structured codes.
- **Ambiguity:** The adjudicated label includes Other and a discrepancy exists, indicating insufficient narrative clarity.
- **Compression:** The narrative contains an information that cannot be represented in the structured coding system.

These categories are not mutually exclusive, and a single record may exhibit multiple discrepancy types.

Detailed Per-Class Discrepancy Breakdown. Table 5 provides a detailed view of discrepancy behavior across classes. The results confirm strong asymmetry in several categories:

- Distracted Driving is dominated by errors (2,484) with minimal omissions (50), indicating systematic over-coding in structured labels.
- Impaired Driving shows a similar pattern (420 errors vs. 12 omissions).
- Speeding and Failure to Yield exhibit more balanced discrepancies, with both omission and error contributing substantially.
- Other shows near-symmetric behavior (680 omissions vs. 669 errors), consistent with its role as a residual category.

Additional Observations.

- Error is the dominant discrepancy type across nearly all classes, reinforcing that structured labels frequently introduce unsupported information.
- Omission remains substantial in several categories, particularly Speeding and Failure to Yield, indicating that structured codes also fail to capture important narrative evidence.
- Compression is confined to Access-Related Crashes, reflecting a limitation of the reporting schema rather than annotator behavior.

A.4 Study 4: Extended Linguistic Ambiguity Analysis

This section provides detailed examples of linguistic phenomena that introduce annotation ambiguity, complementing the summary presented in Section 4.4.

Hedging Language. Example: *"Unit 1 appeared to be traveling at an excessive speed before losing control."*

Hedging expressions (e.g., "appeared to be") signal that the reported behavior is inferred rather than directly observed. This creates uncertainty in determining whether the evidence is sufficient to assign a label such as Speeding. In practice, structured labels often treat such inferences as definitive, while narrative-based annotation may apply stricter evidence thresholds.

Implicit Causality. Example: *"The roadway was wet from recent rainfall. Unit 2 crossed the center line and struck Unit 1."*

Environmental conditions are described but not explicitly linked to causation. The annotator must decide whether such conditions constitute contributing factors or merely contextual information, leading to inconsistent interpretations.

Contributing Factor	Omitted by Police	Erroneously Coded	Total	Disc. Rate
Distracted Driving	50	2,484	2,534	86.8%
Other	680	669	1,349	68.9%
Impaired Driving	12	420	432	69.8%
Speeding	1,179	338	1,517	41.4%
Failure to Yield	848	553	1,401	34.0%
Red Light Running	240	87	327	44.6%
Access-Related Crashes	238	0	238	100.0%
Visual Obstruction	38	24	62	59.0%

Table 5: Per-class discrepancy breakdown showing omissions and errors for each contributing factor.

Underspecification. **Example:** *"Unit 1 was traveling northbound on the highway when it drifted across the center line and struck Unit 2 head-on. Unit 1's driver was transported to the hospital. No further information was available at the scene."* The narrative describes the crash outcome but provides no behavioral, environmental, or mechanical context for why the vehicle crossed the center line. The annotator cannot distinguish between Speeding, Distracted Driving, Impaired Driving, or a medical event, as the description omits the causal detail needed for confident classification. The phrase "no further information was available" confirms that the underspecification is inherent to the report itself, not an oversight by the annotator.

Multi-Event Compression. **Example:** *"Unit 1 was using a cell phone, failed to notice stopped traffic, and struck Unit 2 before veering into another lane."*

Multiple contributing factors are described, but structured coding often captures only one. This results in systematic loss of information and contributes to discrepancies between structured and narrative annotations.

Perspective Markers. **Example:** *"The witness stated that Unit 1 ran the red light, while the driver claimed the light was yellow."*

Conflicting accounts attributed to different sources introduce ambiguity. The annotator must determine which account to prioritize, reducing consistency and increasing uncertainty.

Summary. These phenomena, hedging, implicit causality, underspecification, multi-event compression, and perspective markers, represent recurring linguistic patterns that systematically produce annotation ambiguity. They reflect inherent properties of narrative reporting rather than annotation error alone.

A.5 Study 5: Extended Uncertainty Analysis

This section provides detailed statistical and class-level analyses supporting the results in Section 4.5.

Uncertainty is unevenly distributed across contributing factor classes. The Other category exhibits the highest uncertainty rate (30.6%), followed by Failure to Yield (10.0%) and Visual Obstruction (9.9%), reflecting classes that require complex situational interpretation.

In contrast, Distracted Driving (1.8%) and Impaired Driving (4.5%) show low uncertainty despite high structured-code error rates (see Study 3). This indicates that these factors are linguistically explicit in narratives but systematically misrepresented in structured coding.

Additional Observations.

- Uncertainty is concentrated in classes requiring contextual reasoning.
- Linguistically explicit behaviors exhibit low uncertainty even when structured labels are unreliable.
- This supports the use of uncertainty as a targeted signal for annotation difficulty.

A.6 Annotation Guidelines and Label Definitions

We follow a three-layer annotation framework in which crash narratives are labeled using an 8-class contributing-factor taxonomy derived from an original set of 72 police codes. The mapping was developed in collaboration with domain specialists to reduce sparsity while preserving interpretability.

General Annotation Principles. Annotations are assigned based solely on textual evidence in the narrative, without reference to structured administrative labels. Multiple contributing factors

Contributing Factor	Total Records	Uncertain	Uncertainty Rate
Other	1,265	387	30.6%
Failure to Yield	3,524	351	10.0%
Visual Obstruction	81	8	9.9%
Access-Related Crashes	236	14	5.9%
Speeding	3,316	187	5.6%
Red Light Running	651	34	5.2%
Impaired Driving	198	9	4.5%
Distracted Driving	437	8	1.8%

Table 6: Per-class uncertainty concentration.

may be assigned when supported by the text. When evidence is insufficient or ambiguous, records are explicitly flagged using an uncertainty indicator. All examples used in this study are paraphrased to preserve data confidentiality.

Label Definitions. We summarize the eight contributing-factor classes below:

- **Speeding:** The driver operated the vehicle at an unsafe speed or followed too closely, including failure to control speed or rear-end collisions where no alternative cause is specified.
- **Impaired Driving:** The driver was under the influence of alcohol, drugs, or medication, with observable evidence linking impairment to the crash.
- **Distracted Driving:** The driver's attention was diverted (e.g., phone use, inattention, fatigue), resulting in failure to perceive or react to hazards.
- **Failure to Yield:** The driver violated right-of-way rules or executed unsafe maneuvers such as improper turns, lane changes, or passing.
- **Red Light Running:** The driver entered an intersection in violation of a red signal or stop control.
- **Visual Obstruction:** The driver's visibility was physically blocked by environmental or structural factors (e.g., glare, weather, obstacles).
- **Access-Related Crash:** The crash occurred while entering or exiting a private complex, parking lot, or private access point.
- **Other:** Residual category for contributing factors not captured by the above classes (e.g., animals, medical events, vehicle issues, road rage).

These definitions reflect a balance between semantic clarity and compatibility with existing reporting standards.

The original reporting system includes 72 fine-grained contributing-factor codes, which were consolidated into the 8-class taxonomy described above. The mapping from 72 original contributing-factor codes to the 8-class taxonomy was developed in collaboration with domain experts, ensuring alignment with real-world reporting practices. The complete code-to-class mapping is provided in Table 7.

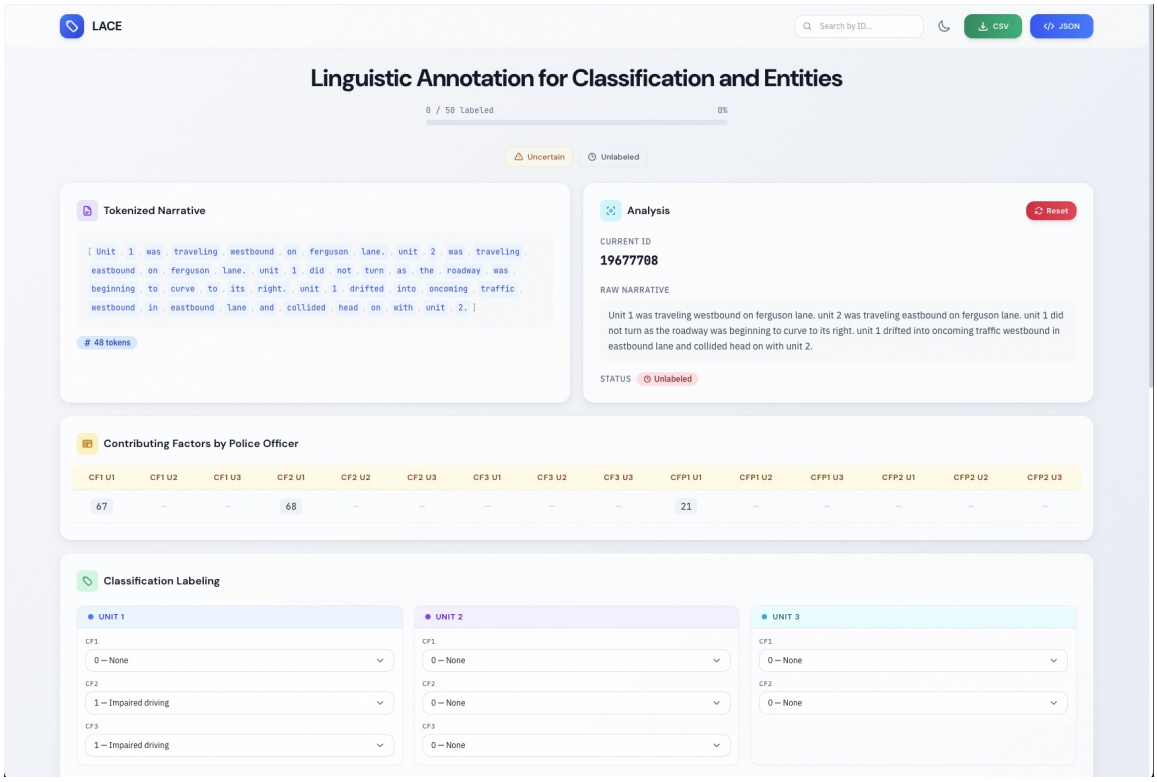
A.7 Annotation Interface

We developed a custom annotation interface to support multi-label assignment and explicit uncertainty marking in crash narratives. The tool allows annotators to select contributing factors, flag ambiguous cases, and review previously labeled records within a unified workflow.

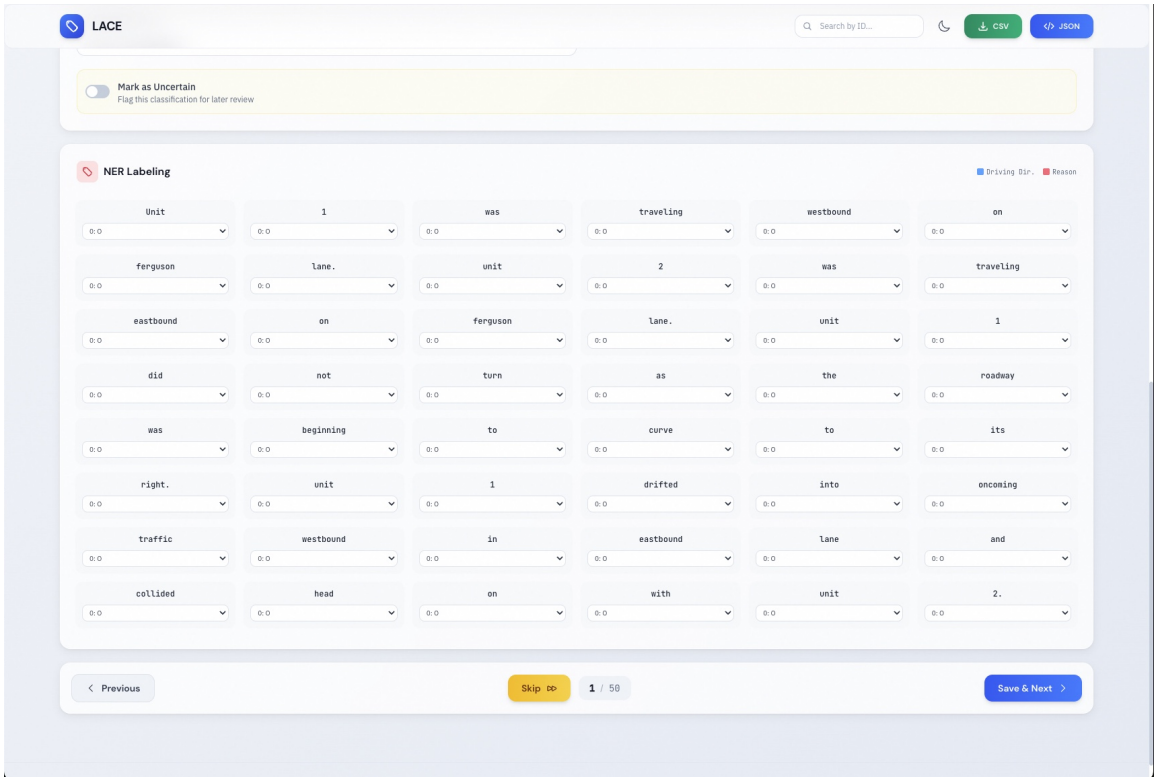
The interface was designed to operate in a restricted-access environment to ensure data confidentiality while maintaining annotation consistency.

ID	Original Contributing Factor Code	8-Class Category
0	NONE	—
1	ANIMAL ON ROAD - DOMESTIC	Other
2	ANIMAL ON ROAD - WILD	Other
3	BACKED WITHOUT SAFETY	Failure to Yield
4	CHANGED LANE WHEN UNSAFE	Failure to Yield
14	DISABLED IN TRAFFIC LANE	Other
15	DISREGARD STOP AND GO SIGNAL	Red Light Running
16	DISREGARD STOP SIGN OR LIGHT	Red Light Running
17	DISREGARD TURN MARKS AT INTERSECTION	Failure to Yield
18	DISREGARD WARNING SIGN AT CONSTRUCTION	Failure to Yield
19	DISTRACTION IN VEHICLE	Distracted Driving
20	DRIVER INATTENTION	Distracted Driving
21	DROVE WITHOUT HEADLIGHTS	Other
22	FAILED TO CONTROL SPEED	Speeding
23	FAILED TO DRIVE IN SINGLE LANE	Failure to Yield
24	FAILED TO GIVE HALF OF ROADWAY	Failure to Yield
25	FAILED TO HEED WARNING SIGN	Failure to Yield
26	FAILED TO PASS TO LEFT SAFELY	Failure to Yield
27	FAILED TO PASS TO RIGHT SAFELY	Failure to Yield
28	FAILED TO SIGNAL OR GAVE WRONG SIGNAL	Failure to Yield
29	FAILED TO STOP AT PROPER PLACE	Failure to Yield
30	FAILED TO STOP FOR SCHOOL BUS	Failure to Yield
31	FAILED TO STOP FOR TRAIN	Failure to Yield
32	FAILED TO YIELD ROW - EMERGENCY VEHICLE	Failure to Yield
33	FAILED TO YIELD ROW - OPEN INTERSECTION	Failure to Yield
34	FAILED TO YIELD ROW - PRIVATE DRIVE	Failure to Yield
35	FAILED TO YIELD ROW - STOP SIGN	Failure to Yield
36	FAILED TO YIELD ROW - TO PEDESTRIAN	Failure to Yield
37	FAILED TO YIELD ROW - TURNING LEFT	Failure to Yield
38	FAILED TO YIELD ROW - TURN ON RED	Failure to Yield
39	FAILED TO YIELD ROW - YIELD SIGN	Failure to Yield
40	FATIGUED OR ASLEEP	Distracted Driving
41	FAULTY EVASIVE ACTION	Failure to Yield
42	FIRE IN VEHICLE	Other
43	FLEEING OR EVADING POLICE	Other
44	FOLLOWED TOO CLOSELY	Speeding
45	HAD BEEN DRINKING	Impaired Driving
46	HANDICAPPED DRIVER	Other
47	ILL	Other
48	IMPAIRED VISIBILITY	Visual Obstruction
49	IMPROPER START FROM PARKED POSITION	Other
50	LOAD NOT SECURED	Other
51	OPENED DOOR INTO TRAFFIC LANE	Failure to Yield
52	OVERSIZED VEHICLE OR LOAD	Other
53	OVERTAKE AND PASS INSUFFICIENT CLEARANCE	Failure to Yield
54	PARKED AND FAILED TO SET BRAKES	Other
55	PARKED IN TRAFFIC LANE	Other
56	PARKED WITHOUT LIGHTS	Other
57	PASSED IN NO PASSING LANE	Failure to Yield
58	PASSED ON RIGHT SHOULDER	Failure to Yield
59	PEDESTRIAN FAILED TO YIELD ROW TO VEHICLE	Failure to Yield
60	UNSAFE SPEED	Speeding
61	SPEEDING - (OVERLIMIT)	Speeding
62	TAKING MEDICATION	Impaired Driving
63	TURNED IMPROPERLY - CUT CORNER ON LEFT	Failure to Yield
64	TURNED IMPROPERLY - WIDE RIGHT	Failure to Yield
65	TURNED IMPROPERLY - WRONG LANE	Failure to Yield
66	TURNED WHEN UNSAFE	Failure to Yield
67	UNDER INFLUENCE - ALCOHOL	Impaired Driving
68	UNDER INFLUENCE - DRUG	Impaired Driving
69	WRONG SIDE - APPROACH OR INTERSECTION	Failure to Yield
70	WRONG SIDE - NOT PASSING	Failure to Yield
71	WRONG WAY - ONE WAY ROAD	Failure to Yield
72	CELL/MOBILE PHONE USE	Distracted Driving
73	ROAD RAGE	Other
74	OTHER (EXPLAIN IN NARRATIVE)	Other
75	CELL/MOBILE DEVICE USE - TALKING	Distracted Driving
76	CELL/MOBILE DEVICE USE - TEXTING	Distracted Driving
77	CELL/MOBILE DEVICE USE - OTHER	Distracted Driving
78	CELL/MOBILE DEVICE USE - UNKNOWN	Distracted Driving
79	FAILED TO SLOW FOR EMERGENCY LIGHTS	Failure to Yield
80	DROVE ON IMPROVED SHOULDER	Failure to Yield

Table 7: Mapping from original police contributing-factor codes to the 8-class taxonomy.



(a) Main labeling interface.



(b) Uncertainty marking view.

Figure 14: Annotation interface used for multi-label labeling and uncertainty marking.