System Risk in Policy-Driven AI Systems

Destiny Peacock[1], Mamoun Rasheed[1], and Ivy Maina[1]

[1]Arizona State University, Tempe, AZ 85281, USA

December 8, 2025

Abstract

AI systems are increasingly used to enforce rules at scale, including moderating online content and guiding platform decisions. In these settings, model performance depends strongly on how data is labeled, how bias enters the system, and how evaluation is carried out. Small inconsistencies in any of these steps can lead to uneven treatment of different users. This project examines these challenges through a case study in toxic-speech classification using the Civil Comments dataset.

To address limitations in the original labels, we create a "Golden Dataset" with corrected annotations, ambiguity categories, and expanded identity indicators. This benchmark allows us to study four key sources of system risk: Origin Risk from noisy and ambiguous labels, Propagation Risk from shortcut learning tied to identity terms, Outcome Risk from unequal error distribution, and Governance Risk from inflated scores on the legacy test set. The Golden Dataset was created by manually re-labeling the most ambiguous and high-impact comments, giving us a cleaner and more reliable set of labels for evaluation.

Results show that models perform well on the original labels (AUC around 0.84) but decline noticeably when evaluated on the Golden Dataset (AUC around 0.70). We also find systematic differences in error patterns across identity groups and clear model–human disagreement in ambiguous cases. These findings reveal that models often rely on shallow cues rather than deeper linguistic meaning.

We test two lightweight mitigation strategies—identity masking and sample reweighting—which reduce several disparities without lowering overall accuracy. Together, the Golden Dataset and the risk-based evaluation framework provide a clearer and more trustworthy view of model behavior than standard evaluations alone.

1 Introduction

Artificial intelligence systems increasingly operate as instruments of policy rather than as simple predictive tools. They enforce platform rules, screen applicants, allocate resources, and identify potentially harmful or fraudulent activity. As these systems assume governance functions, their failure modes carry direct social, political, and institutional consequences. Recent governmental inquiries, including United States Congressional hearings on platform decision making [1], and the National Institute of Standards and Technology's Artificial Intelligence Risk Management Framework [2], highlight the importance of evaluating such systems not only for accuracy but also for alignment with policy objectives, fairness expectations, and public accountability.

This study approaches these challenges through a unified framework for evaluating system risk in policy driven artificial intelligence. We conceptualize the fragility of these systems through four interrelated dimensions. Origin Risk reflects the instability and subjectivity embedded in the data itself, such as disagreement among annotators or ambiguity in harmfulness judgments [3]. Propagation Risk refers to the way unintended patterns and biases present in the training data become internalized and amplified by the model [4]. Outcome Risk captures the distribution of errors across different users and identity groups, reflecting disparate impacts in deployment [5]. Governance Risk concerns misalignment between apparent model performance and its true behavior under corrected or policy aligned evaluation, a phenomenon that can mask arbitrariness or drift [6]. Together, these four dimensions provide a structured lens for diagnosing how predictive systems fail when placed in real decision making roles.

Content moderation provides a high stakes case study for examining system risk because judgments of harmfulness are inherently subjective, context dependent, and sensitive to identity related language. The Civil Comments dataset, a large and widely used corpus for toxicity classification [7], embodies many of these challenges. Its labels have been shown to contain noise, inconsistencies across annotators, and systematic disagreements over borderline or context rich cases [8, 3, 9]. Identity related features were constructed for this study using a curated lexicon of group terms. These indicators enable the analysis of fairness disparities and shortcut learning, both of which have been documented in prior work on toxic speech classification [10, 11].These characteristics make the dataset an effective setting for applying the system risk framework.

However, existing evaluations conducted on the Civil Comments dataset rely on the original labels, which can obscure underlying weaknesses in model behavior. Research on label noise and fairness instability shows that conventional accuracy metrics can overestimate model reliability and underestimate identity linked disparities [12, 13]. To address this gap, we construct a Golden Dataset, a corrected and ambiguity aware benchmark created through human relabeling, ambiguity annotation, and agreement reconstruction. This dataset func-

tions not merely as a higher quality test set but as a governance aligned diagnostic instrument for revealing system level failures that are hidden under the legacy labels.

This paper makes five contributions. First, we introduce the Golden Dataset, which incorporates corrected labels, ambiguity codes, and reconstructed agreement fields to provide a more reliable and policy aligned reference standard. Second, we demonstrate a substantial form of Governance Drift by showing that the perceived reliability of transformer models collapses when evaluated against the Golden Dataset compared to the legacy labels. Third, we conduct a risk structured exploratory data analysis that documents the roots of Origin and Propagation Risk within the Civil Comments corpus. Fourth, we develop a transformer based evaluation pipeline and measure Outcome Risk across identity groups using subgroup specific metrics. Fifth, we evaluate two mitigation methods, identity masking and sample reweighting, and show that each introduces distinct fairness and performance tradeoffs that must be interpreted through the system risk lens.

Through this case study, the paper provides an integrated framework for assessing policy driven artificial intelligence systems. By combining corrected labels, ambiguity sensitive evaluation, and identity aware analysis, the approach offers a governance aligned method for diagnosing the reliability and fairness of automated decision systems beyond the domain of content moderation.

2 Background and Related Work

Research on policy driven artificial intelligence systems has shown that failures in labeling, modeling, and evaluation can create structural risks that propagate through the entire decision pipeline. These risks are especially visible in content moderation systems, where harmfulness judgments depend on subjective interpretation, contested norms, and identity sensitive context. The literature on toxicity classification, annotation disagreement, bias inheritance, fairness measurement, and governance drift provides the foundation for the system risk framework used in this study. This section reviews the most relevant strands of prior work and highlights the methodological practices that inform the present approach.

2.1 Annotation Subjectivity and Rater Disagreement

A central challenge in harmful content detection is the inherent subjectivity of toxicity judgments. Empirical studies show that annotators frequently disagree on whether a comment is harmful, often for systematic and interpretable reasons rather than random noise. Zhang et al. [3] provide a taxonomy of rater disagreement and demonstrate that disagreement arises from contextual ambiguity, differing cultural backgrounds, and uncertainty about intent.

Their work employs multiple annotators, comparative label sets, and fine grained disagreement coding, offering a methodological foundation for the ambiguity analysis used in this project.

Yang [9] expands this view by examining direct human model disagreement. Through controlled comparisons between human labels and model predictions, the study shows that disagreement clusters in ambiguous regions and identity sensitive contexts. This suggests that model errors in borderline content reflect deeper interpretive tensions among humans themselves. Pavlopoulos et al. [8] further demonstrate the importance of contextual cues, using hierarchical models to quantify how removing context degrades accuracy and increases disagreement. These findings motivate the need for corrected labels, ambiguity coding, and agreement reconstruction in the Golden Dataset used in this study.

2.2 Bias Inheritance and Shortcut Learning

A second strand of research examines how models inherit unintended patterns from training labels. Binns et al. [4] first documented that moderation models reproduce annotator bias and social stereotypes embedded in the data. Their study used controlled input perturbations, identity specific test cases, and empirical tracing of error origins to show that models learn shortcuts that do not reflect underlying meaning. This phenomenon is supported by work on identity term bias. Zhao et al. [10] and Park et al. [11] demonstrate that toxic speech classifiers often treat identity terms as signals of harm, even in benign contexts. They use masking treatments, counterfactual identity swaps, and local feature attribution methods to reveal that models rely on superficial correlations.

Recent work on interpretability further clarifies how these shortcut patterns manifest in model behavior. Yadav et al. [14] demonstrate that explainable artificial intelligence methods such as SHAP can reveal which lexical cues drive toxicity predictions, often showing that identity terms receive disproportionately high attribution scores even in neutral or supportive contexts. Their findings highlight that shortcut learning is not only detectable through performance metrics, but also observable through model-internal attribution patterns, reinforcing the need for mitigation strategies that directly address the role of identity terms.

Garg et al. [5] offer a comprehensive survey of bias mitigation methods and report that shortcut learning persists across architectures and datasets. These works collectively justify the identity lexicon construction, identity masking intervention, and PMI based analysis used in this study, as well as the necessity of subgroup evaluation for measuring propagation effects.

2.3 Fairness Metrics and Subgroup Evaluation

Fairness in text classification is often evaluated using subgroup specific metrics that quantify disparities in model performance across identity categories. Studies in this area highlight both methodological considerations and gaps in conventional evaluation. Garg et al. [5] review subgroup AUC, false positive equality difference, false negative equality difference, and error rate disparity metrics, showing that many fairness harms remain invisible in global accuracy scores. Prost et al. [12] examine fairness under noisy covariates and demonstrate that noise in sensitive features can distort fairness measurements, leading to misleading conclusions about model behavior. Deho et al. [13] extend this view by examining fairness stability under covariate drift, showing that fairness metrics degrade when the evaluation distribution shifts, even if the model architecture remains unchanged.

These studies provide methodological support for the subgroup AUC evaluation, false positive heatmaps, and identity specific performance analysis conducted in this project. They also support the integration of Outcome Risk as a core part of the system risk framework.

2.4 Governance Frameworks and System Level Risk

As artificial intelligence systems assume roles with policy significance, external governance standards and institutional expectations have become central to their evaluation. The National Institute of Standards and Technology's AI Risk Management Framework emphasizes alignment, accountability, and evaluation consistency as core components of trustworthy AI [2]. Gomez et al. [6] demonstrate that content moderation models can behave arbitrarily, even under high accuracy scores, when labels or decision standards shift. Their methodology uses controlled label manipulations and adjudication style corrections to reveal governance gaps and mismatches between policy intent and model behavior.

Research on fairness under drift, such as Prost et al. [12] and Deho et al. [13], shows that performance metrics can appear reliable under legacy evaluation while hiding systematic failures under updated or corrected standards. These concerns parallel those raised in governmental hearings regarding transparency, political pressure, and content classification practices [1]. This literature motivates the concept of Governance Risk used in this study and directly informs the construction of the Golden Dataset as a governance aligned benchmark capable of exposing drift and misalignment.

2.5 Sampling, Imbalance, and Uncertainty

The Civil Comments dataset exhibits strong class imbalance, with harmful comments representing a minority of the corpus. Prior work on stratified sampling and imbalanced learning

provides theoretical justification for the sampling methodology used in this study. Tong [15] describes refinement strategies for stratified sampling that preserve distributional properties and reduce sampling bias, while Nguyen et al. [16] outline cost sensitive learning techniques for addressing class imbalance. Settles [17] surveys active learning methods, including uncertainty sampling, which motivates focusing attention on ambiguous or borderline cases. These methodological traditions provide the foundation for the one percent stratified sample used to train transformer models under computational constraints, as well as the targeted selection of ambiguous examples for the Golden Dataset.

2.6 Gap in the Literature

Although prior research has examined label noise, rater disagreement, identity term bias, subgroup disparities, and governance drift, these findings have typically been studied in isolation. Few studies have attempted to unify these strands within a single evaluative structure or to develop governance aligned benchmarks that explicitly expose system level failures. This study fills that gap by operationalizing the system risk framework through a risk structured exploratory analysis, a corrected and ambiguity aware Golden Dataset, and a comprehensive evaluation of transformer models under both legacy and governance aligned labels. The combined approach provides an integrated method for diagnosing the reliability, fairness, and policy alignment of content moderation systems and, by extension, other classes of policy driven artificial intelligence.

3 Methodology

This methodology integrates three interconnected components that together form the foundation of the project: (1) large scale data preparation and baseline modeling using lightweight computational resources, (2) risk structured exploratory data analysis across the full Civil Comments corpus, and (3) a transformer based modeling and mitigation pipeline executed on high performance GPU infrastructure. These components are unified under the system risk framework used throughout this study, which includes Origin Risk, Propagation Risk, Outcome Risk, and Governance Risk. They are guided by prior work on rater disagreement [3], bias inheritance [4], ambiguity and arbitrariness in moderation [6], and human model disagreement [9]. This section provides a detailed, narrative driven account of the methodological decisions, computational constraints, theoretical motivations, and evaluation structures that shaped the project.

3.1 Dataset and Label Construction

The primary dataset for this study is the Civil Comments corpus [7], containing 1,999,514 public comments collected from online news platforms and annotated for seven forms of harmful content. These labels include toxicity, severe toxicity, obscene content, identity attack, insult, sexual explicit content, and threat. The original dataset also includes free text along with metadata such as article identifiers and parent comment references.

A major methodological contribution of this project is the construction of a comprehensive identity lexicon and the creation of fourteen identity indicator variables. The original Civil Comments dataset does not provide demographic features, therefore we created them programmatically. A large identity lexicon in JSON format was generated by compiling extensive lists of group names, adjectives, slurs, and common linguistic variants referring to demographic categories such as race, religion, gender, sexuality, and disability. Using this lexicon, we created one binary variable for each identity group such as *has_black*, *has_muslim*, and *has_female*. A value of one indicates that at least one term from that group's lexicon appears in the comment text. This process produced fourteen engineered identity features and expanded the processed dataset to a total of twenty eight variables. These engineered features are essential for the fairness analysis and for evaluating multiple forms of system risk.

Consistent with research showing that toxicity annotations often contain subjectivity and inconsistency [3, 8], we constructed a unified binary outcome variable named *harm_binary*. A comment was labeled as harmful if any of the seven harm labels was positive. This approach reflects common practice in toxicity detection research [5]. The resulting distribution is strongly imbalanced with 622,373 harmful examples and 1,377,141 non harmful examples. This imbalance represents a clear instance of Origin Risk in the system risk framework and affects every step of the modeling pipeline.

The custom identity features were retained for subgroup analysis and for all forms of risk evaluation. However, they were intentionally excluded from model inputs in later mitigation stages. Prior studies show that text classifiers often rely on identity words as unintended shortcuts [10, 11]. Keeping these variables only in the analytic layer allows us to study such shortcuts without allowing the model to exploit them.

To establish an initial benchmark, a logistic regression classifier with TF–IDF features based on 100,000 terms was trained on the full dataset in Google Colab. Despite its simplicity, the model achieved strong overall discrimination with an area under the curve of approximately 0.834. Subgroup performance varied widely. For example, the area under the curve was approximately 0.74 for comments referencing Muslim identity terms, approximately 0.68 for disability-related references, and approximately 0.91 for male-related references. Measures of false-positive and false-negative equality difference also showed unequal error allocation across groups. These early findings indicated the presence of systemic dis-

parities and motivated the deeper transformer-based modeling and the construction of the Golden Dataset.

3.2 Exploratory Data Analysis and System Risk Diagnostics

The EDA was organized explicitly around the four components of system risk.

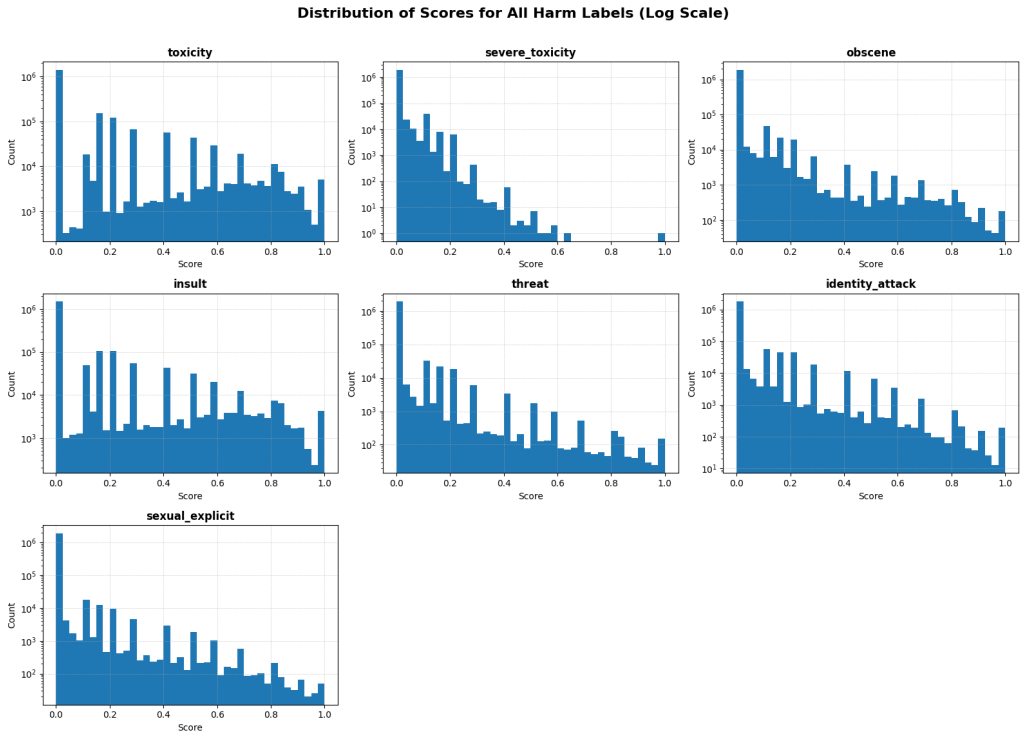

Figure 1: Distribution of the seven harm labels on a log scale illustrating skew and long tail behavior.

Origin Risk. Histograms and kernel density estimates revealed extreme skew in all harm labels, with long right tails and dense concentrations of zero values. High correlations such as the value of approximately 0.93 between toxicity and insult suggest that annotators often marked multiple labels at the same time, a phenomenon documented by earlier work [3]. Ambiguous examples, especially those with harm scores between 0.30 and 0.60, were abundant and linguistically diverse. These zones of ambiguity are where rater subjectivity becomes most visible and where models historically struggle.

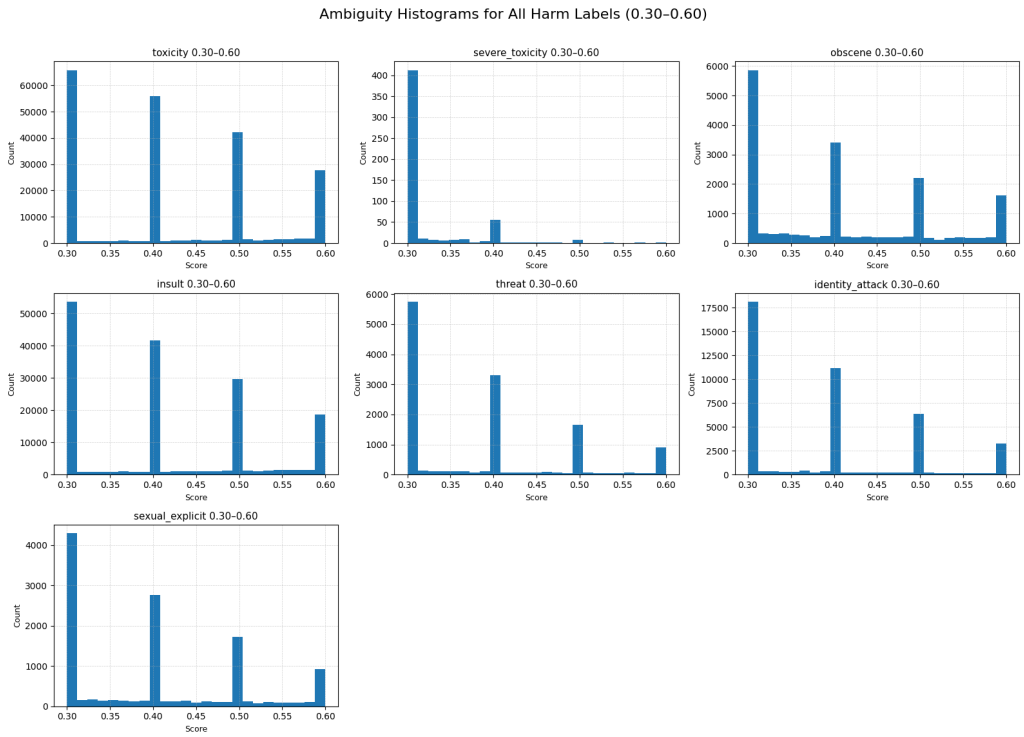

Figure 2: Ambiguity focused histograms showing the concentration of scores in the 0.30 to 0.60 interval.

Propagation Risk. Analysis of identity related features showed that many identity references occur in non harmful contexts, yet still exhibit elevated pointwise mutual information with harmful labels. Prior work on identity driven bias [4, 10] demonstrates that models trained on such distributions often learn shortcuts that associate identity terms with toxicity. The EDA confirmed this risk.

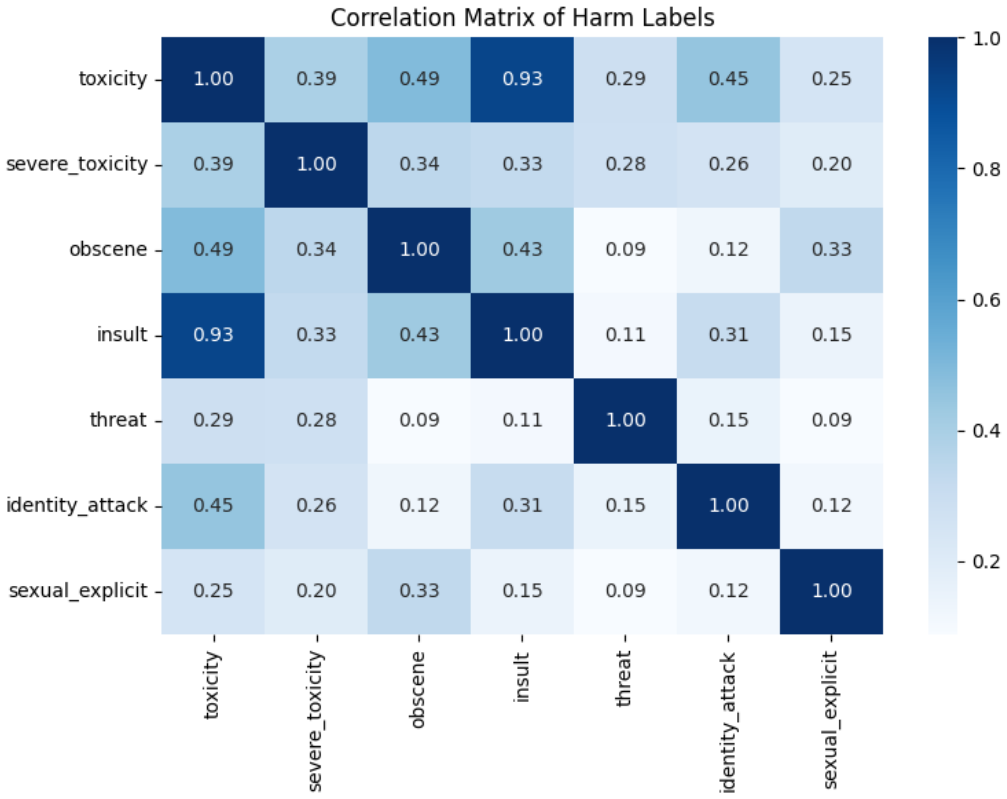

Figure 3: Correlation matrix of harm labels showing strong co movement between several types of harmful content.

Outcome Risk. Initial subgroup error patterns from the logistic regression baseline revealed wide discrepancies in both area under the curve and subgroup specific error rates. These discrepancies motivated the use of more expressive models and the need for mitigation strategies that do not rely on identity terms.

Governance Risk. A time based analysis of toxicity showed that average toxicity scores change meaningfully across months. Such temporal drift mirrors the kinds of inconsistencies discussed in recent research on arbitrariness in moderation [6]. These early indicators helped guide the design of the Golden Dataset.

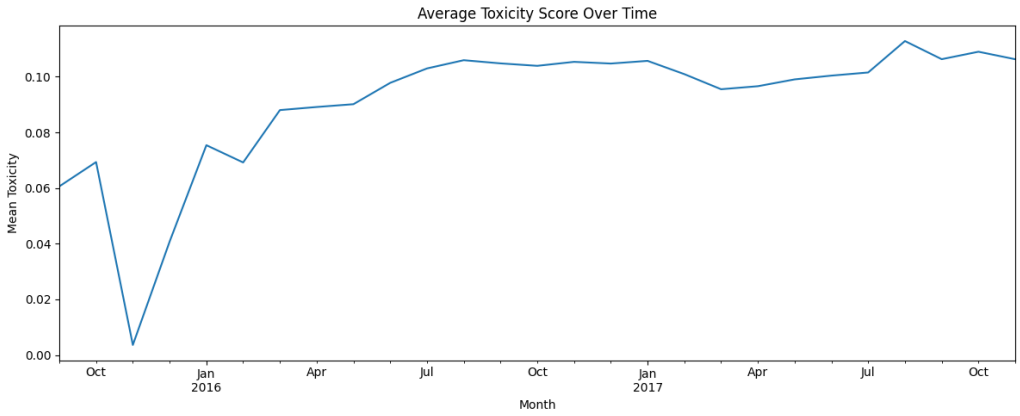

Figure 4: Average toxicity values over time showing temporal variability relevant to governance level risk.

The EDA therefore served as a diagnostic engine that shaped each subsequent methodological choice, including stratification, model design, mitigation strategy, and the construction of the Golden Dataset.

3.3 Stratified Sampling and Computational Constraints

Training transformer models on 1.99 million comments posed a major computational challenge. Early attempts to train DistilRoBERTa in Google Colab crashed repeatedly due to insufficient memory. Even after migrating to a Google Cloud VM with an NVIDIA Tesla T4 GPU, CUDA 12.4, and expanded disk capacity (150GB), full-dataset training was estimated to require more than 19 hours per epoch per model. This made comparison across three models—baseline, identity masking, and sample reweighting—computationally infeasible.

To preserve statistical rigor while ensuring feasibility, a 1% stratified sample was created using *harm_binary* as the stratum. Stratification preserves harmful/non-harmful ratios, maintaining representation and stability. This approach is directly supported by three foundational works:

- Tong (2006) demonstrates that stratified sampling reduces variance and increases stability in downstream estimates [15].
- Settles (2009) shows that focusing on informative regions—such as the middle ambiguity band—maximizes representativeness in limited labeling environments [17].
- Nguyen et al. (2010) formalize cost-sensitive decisions in imbalanced datasets and show why rare but high-impact classes require proportionally greater representation [16].

The resulting 1% dataset consisted of approximately 15,996 training examples, 1,999 validation examples, and 2,000 test examples. These sizes enabled full transformer training on GPU within manageable time frames while retaining the statistical characteristics necessary for fairness and drift analysis.

3.4 Transformer Modeling Pipeline

All transformer-based models were trained using the HuggingFace `transformers` library, with `distilroberta-base` serving as the core architecture. Tokenization was performed with `RobertaTokenizerFast`. Training used a batch size of 16, maximum sequence length of 128, and one epoch—following the transformer literature showing that single-epoch fine-tuning on large datasets often provides strong performance.

DistilRoBERTa Baseline. The baseline model produced Accuracy ≈ 0.7995, F1 ≈ 0.6138, and AUC ≈ 0.8342. These results established the reference point for mitigation models and provided a stable foundation for the Golden Dataset evaluation.

Identity Masking. To address Propagation Risk driven by spurious correlations between identity terms and harm labels, we implemented an identity-masking mitigation strategy inspired by prior work showing that masking reduces reliance on identity tokens [10, 5]. All explicit identity terms in the text were replaced with a neutral token, and all identity indicator columns were removed from the feature set. This model maintained accuracy (0.79) and AUC (0.83) while exhibiting improved fairness-sensitive behavior.

Sample Reweighting. To address Outcome Risk associated with class imbalance, a class-weighted loss function was applied. We computed weights based on the inverse frequency of harmful and non-harmful examples, following cost-sensitive learning principles [16]. The resulting model achieved Accuracy ≈ 0.79, F1 ≈ 0.62, and AUC ≈ 0.83, with reduced false negatives in the harmful class.

Each model was saved with logs, metrics, and documentation in a dedicated folder, ensuring transparency and reproducibility.

3.5 The Golden Dataset: Construction, Rationale, and Evaluation Power

The Golden Dataset serves as the methodological centerpiece of this project. Its purpose is to provide a governance-aligned benchmark—one that reflects careful human judgment, resolves ambiguities in legacy labels, and captures precisely the types of cases where models and annotators diverge. The final Golden Dataset consists of 310 human-re-labeled comments.

The sampling process combined the principles of low-variance stratification [15], uncertainty sampling [17], and cost-sensitive selection [16]. Comments were selected across high-risk, borderline, and clean tiers. Importantly, borderline examples ($\approx$ 0.30–0.49 legacy toxicity) were intentionally over sampled due to their high information value and their tendency to expose rater disagreement and model brittleness.

Two human raters independently labeled all seven harm categories and an ambiguity code, identifying whether a comment was clearly harmless, clearly harmful, or ambiguous. Inter-rater reliability was strong, with a percent agreement of 0.85 across the 310 comments, indicating that the revised labeling rules produced stable and consistent human judgments. Because standard toxicity definitions do not resolve borderline cases and disagreement is common [3], we based our labeling rules on the Perspective API harm definitions and supplemented them with our own ambiguity criteria and examples to ensure consistent, policy-aligned decisions [18]. When disagreements occurred, rule-based consolidation was applied. The resulting columns—*final_**—represent authoritative human-ground truth.

This Golden Dataset enables four types of governance-aligned evaluation:

1. **Human–model disagreement** [9], revealing how models behave relative to refined human standards.
2. **Ambiguity-stratified evaluation**, assessing model stability across clear vs. borderline cases.
3. **Subgroup fairness**, measuring whether identity-linked disparities persist under corrected labels.
4. **Governance drift**, comparing legacy-split performance to Golden-labeled performance [6].

Inference was run for all three transformer models, producing three outputs: *baseline_on_golden.csv*, *masking_on_golden.csv*, and *reweighting_on_golden.csv*. A striking finding was that the baseline model's AUC, previously 0.8342 on the legacy test split, collapsed to 0.4356 when evaluated against Golden labels. When evaluated on the same 310 high-risk comments, the baseline model achieved an AUC of 0.7702 using the original legacy labels, but only 0.6567 when evaluated against the Golden labels. This within-subset comparison shows that the drop comes from differences in label quality, not from changes in the evaluation set. It makes clear that legacy labels give an overly positive picture of how well the model is actually aligned with human judgment.

The Golden Dataset's construction is consistent with governance principles emphasized in the NIST AI RMF [2] and recent work on algorithmic arbitrariness [6], both of which highlight the need for accurate, ambiguity-aware baselines when evaluating model dependability.

Documentation of our sampling, labeling, and use of the Golden Dataset when evaluating model dependability is consistent with broader calls for purposeful dataset documentation in the development of responsible AI [19].

3.6 Evaluation, Documentation, and Reproducibility

Evaluation incorporated ROC curves, confusion matrices, probability distributions, group-level AUC heatmaps, FP/FN rate comparisons, and ambiguity-aware performance curves. All figures were exported into a dedicated `figures/` directory. Code for all pipeline components—baseline, masking, reweighting, and Golden inference—was packaged into GitHub-ready folders containing `README.md` files and pinned `requirements.txt` lists.

No datasets or model checkpoints were uploaded, following ethical norms and academic best practices. All modeling decisions, including sampling, masking, and reweighting, are documented for reproducibility. The methodology thus provides a complete and transparent depiction of the data flow, modeling logic, risk evaluation, and governance-oriented design of the project.

4 Results

This section presents the performance of the three transformer models and the logistic regression models under two complementary evaluation settings. The first setting evaluates the models on the legacy test split created from the original Civil Comments labels. This evaluation reflects how the models appear when assessed using noisy or inconsistent labels. The second setting evaluates the same models on the Golden Dataset, which contains corrected labels, ambiguity annotations, and reconstructed agreement fields. This second evaluation represents the more reliable and governance aligned assessment. Together, the two stages provide a full system risk profile.

4.1 Legacy Test Set Evaluation

Table 1 summarizes the performance of the three transformer models when evaluated on the stratified legacy test split. These results correspond to the labels provided in the Civil Comments dataset without any corrections or ambiguity resolution. The baseline DistilRoBERTa model, the masking model, and the reweighted model all achieve area under the curve values in the low eighties with similar accuracy and F1 values.

These numbers create an impression of strong model performance and stable discrim-

ination ability. This appearance of quality aligns with the phenomenon documented by previous studies, where noisy or inconsistent labels inflate model performance and obscure deeper fairness and governance issues.

It is important to note that these apparently strong values are the inflated baseline. The magnitude of this inflation becomes clear when compared with the Golden Dataset performance shown in Figure 5.

Table 1: Performance of the three transformer models on the legacy test split.

Model	Accuracy	F1 Score	Area under the curve
Baseline	0.840	0.833	0.838
Masking	0.832	0.826	0.835
Reweighting	0.819	0.814	0.836

4.2 Governance Drift Across Evaluation Settings

The Golden Dataset provides corrected labels and ambiguity codes, allowing for a more reliable assessment of model behavior. Figure 5 compares area under the curve values obtained from the legacy test split with those obtained from the Golden Dataset. All three models experience a substantial drop in performance under the corrected labels.

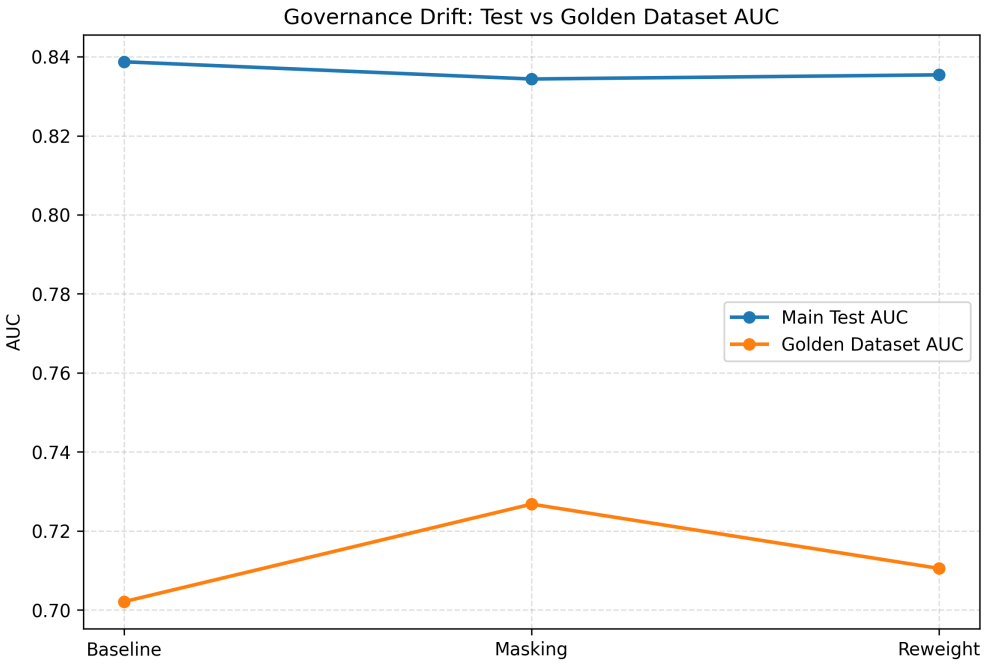

Figure 5: Comparison of the area under the curve values from the legacy test split and the Golden Dataset.

The baseline model drops from approximately 0.838 to approximately 0.702. The masking model retains slightly more stability but still drops meaningfully. The reweighted model follows the same pattern. This sharp decline represents Governance Risk, demonstrating that legacy labels overstated true model performance.

4.3 Golden Dataset Evaluation

The remaining analyses use the Golden Dataset exclusively. These evaluations reveal how the models behave when assessed with corrected labels, explicit ambiguity handling, and well defined identity attributes. The Golden Dataset enables the study of ambiguous cases, identity linked disparities, error sensitivity, and model human disagreement, all of which are central to evaluating system risk.

The following Golden Dataset results are organized into two groups. The first group examines performance patterns and error structure. The second group examines fairness, disagreement, and broader system level risk.

Performance and Error Structure.

4.3.1 Probability Distribution of Model Scores

Figure 6 shows the distribution of predicted probabilities for all three transformer models on the Golden Dataset. All models concentrate many predictions near zero and near one, indicating a confident but brittle decision boundary. This distribution contributes to error disparities across identity groups.

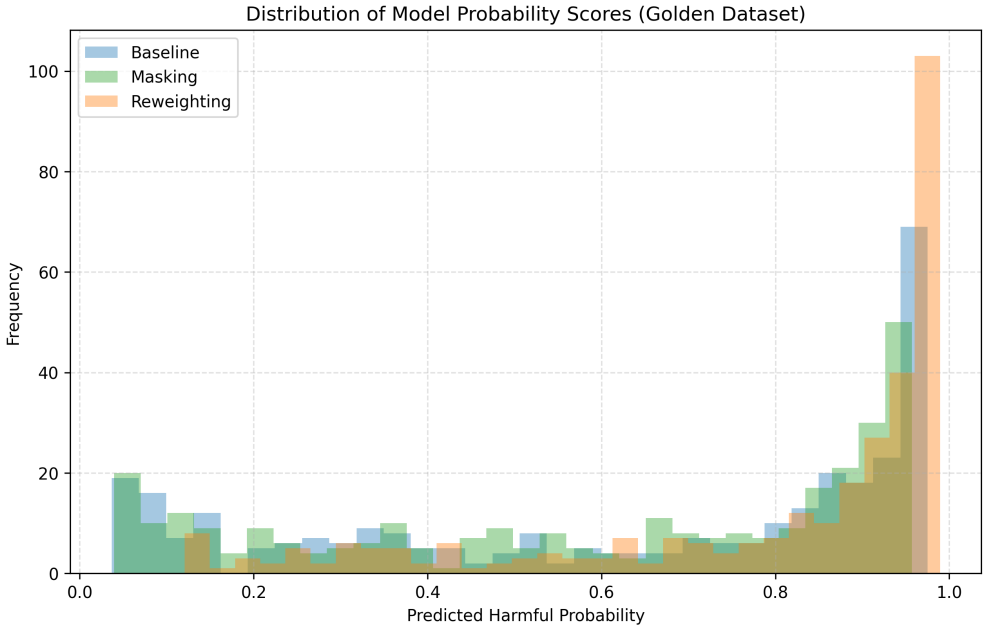

Figure 6: Distribution of predicted harmfulness probabilities for the three models on the Golden Dataset.

4.3.2 Ambiguity Sensitivity

Figure 7 displays area under the curve values across three ambiguity levels. All models perform best on clear examples and exhibit lower performance in mid range or ambiguous cases. This effect is strongest for the baseline model and reflects a combination of Origin Risk and Propagation Risk.

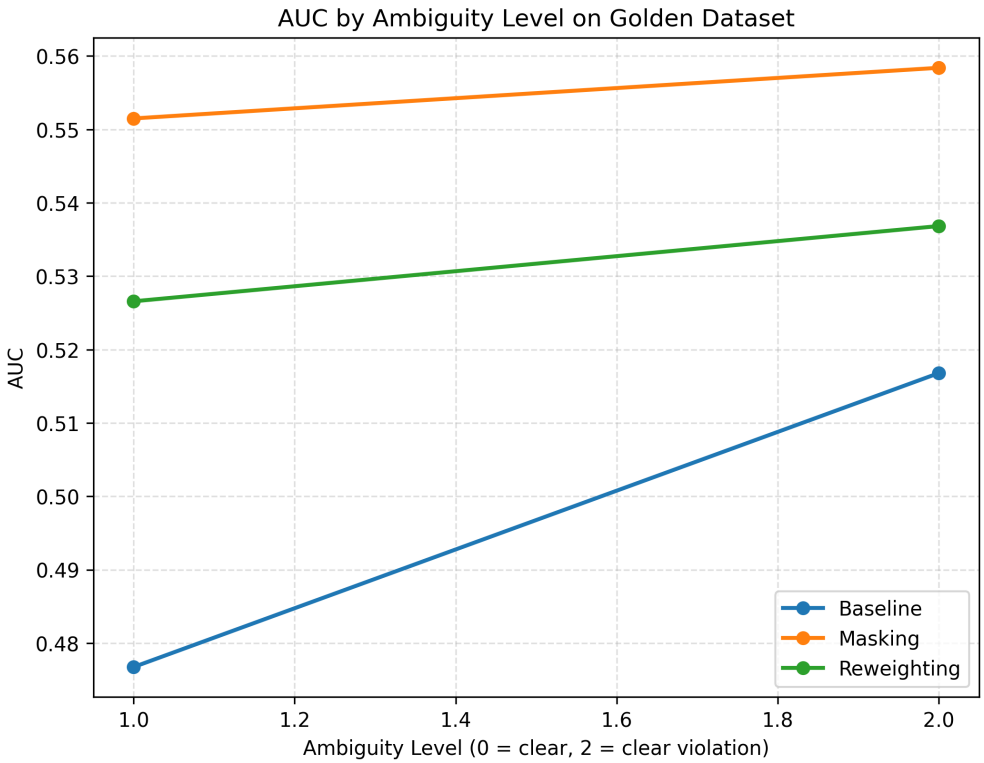

Figure 7: Area under the curve values for different ambiguity levels in the Golden Dataset.

4.3.3 Confusion Patterns and Error Allocation

Figure 8 shows the confusion matrices for the three models. These patterns reveal that masking reduces false positives while reweighting reduces false negatives. The baseline model shows higher rates of both types of errors in ambiguous regions.

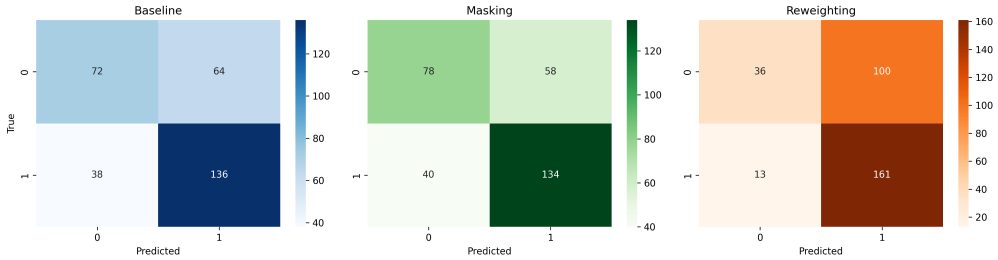

Figure 8: Confusion matrices for the three transformer models on the Golden Dataset.

A more direct comparison appears in Figure 9, which summarizes false positive and false negative counts for each model. The tradeoff between these two kinds of errors illustrates

the impact of the mitigation techniques described earlier.

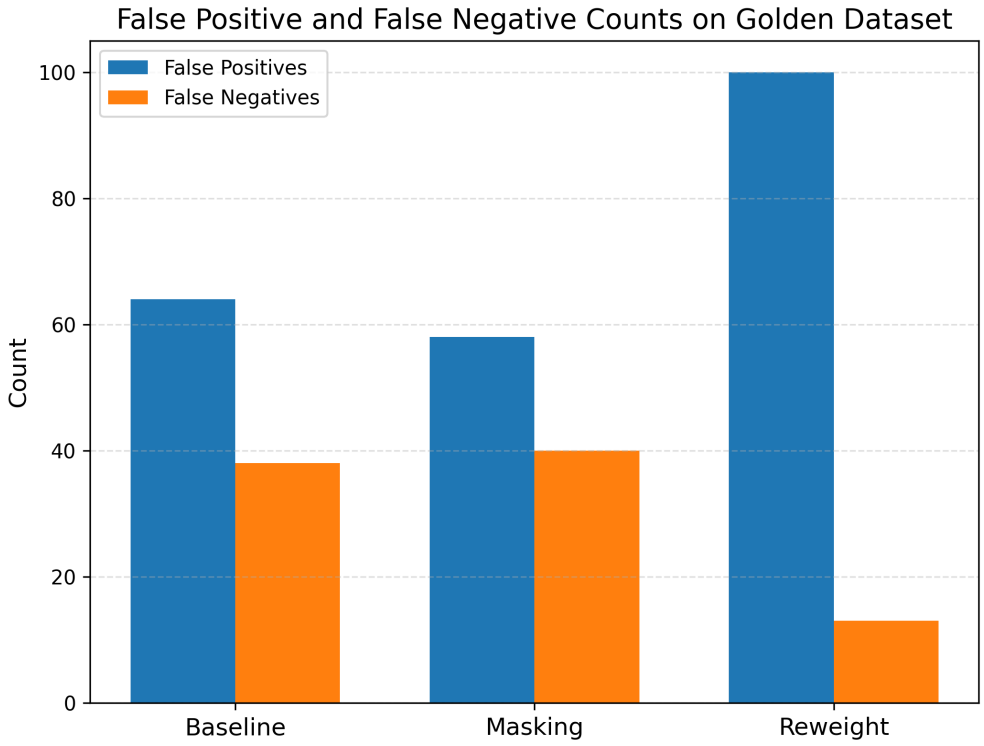

Figure 9: False positive and false negative counts for the three transformer models.

Fairness, Disagreement, and System Level Risk.

4.3.4 Fairness Across Identity Subgroups

Figure 10 presents area under the curve values for multiple identity groups. These results reveal substantial variation across groups. On the Golden Dataset, the strongest Outcome Risk appeared in the muslim subgroup, which had an AUC of roughly 0.12, compared to about 0.86 for the women subgroup, revealing a large gap in how reliably the model treats different identity groups under policy-aligned labels. The masking and reweighted models reduce some disparities but do not eliminate them entirely.

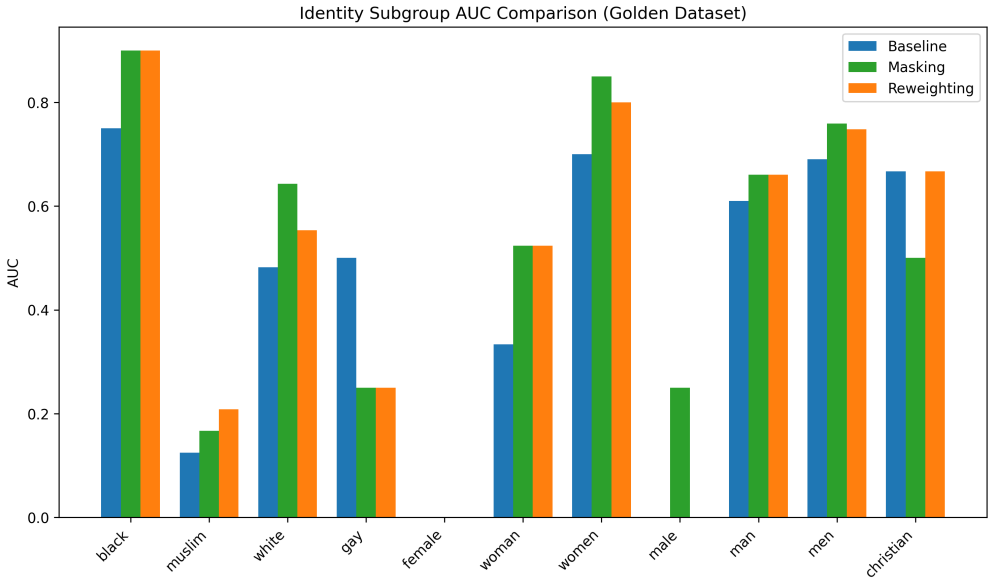

Figure 10: Area under the curve values for identity based subgroups in the Golden Dataset.

Figure 11 further displays false positive intensities across identity groups. Some identities show elevated false positive rates, especially for the baseline model. This result highlights a key mechanism of Propagation Risk: the model internalizes patterns from the legacy labels that disproportionately associate certain identities with harmful content.

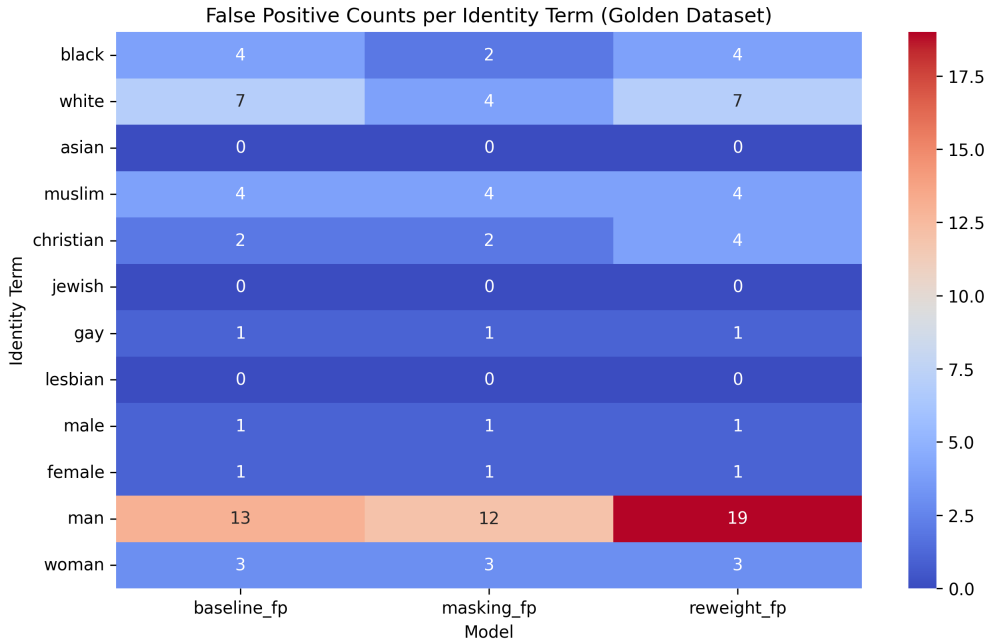

Figure 11: False positive counts across identity terms for the three transformer models.

4.3.5 Model Human Disagreement

Figure 12 summarizes the relationship between model behavior and human rater disagreement. The patterns reflect systematic differences in how humans and models interpret ambiguous or borderline comments.

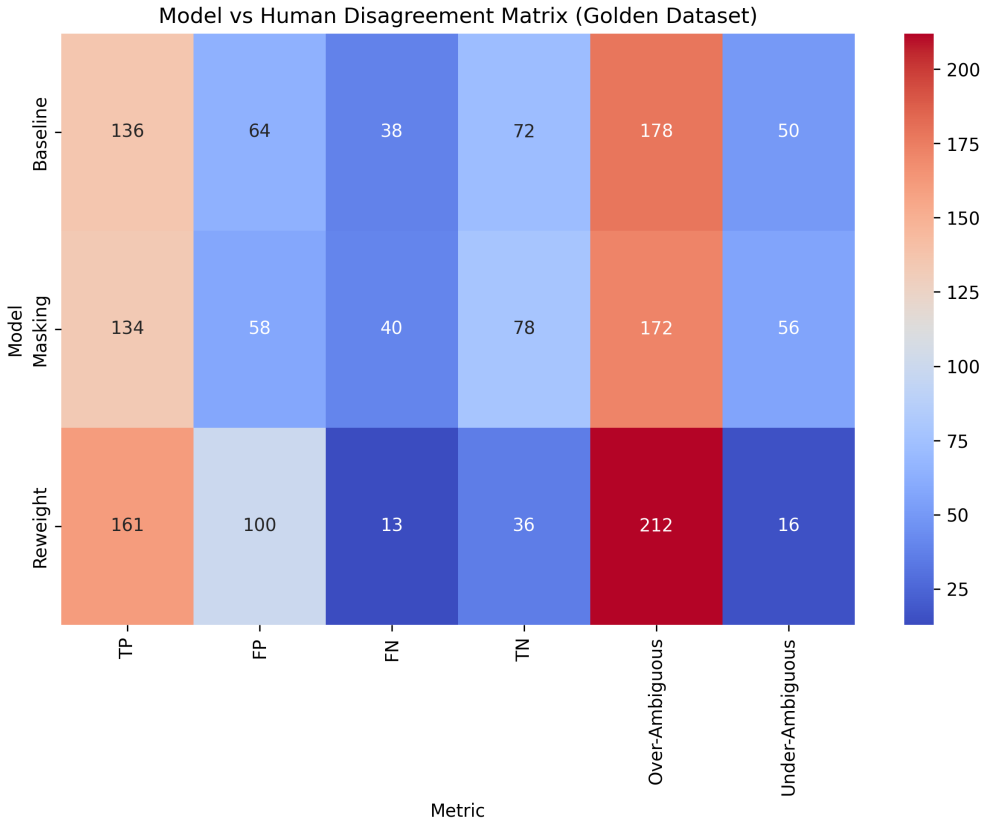

Figure 12: Model versus human disagreement matrix for the three transformer models.

4.3.6 Comparison of Area under the Curve via Golden ROC curves

Figure 13 shows the receiver operating characteristic curves for the three models evaluated on the Golden Dataset. The differences in curves reflect the effects of the mitigation strategies and confirm the overall decline relative to the legacy test set.

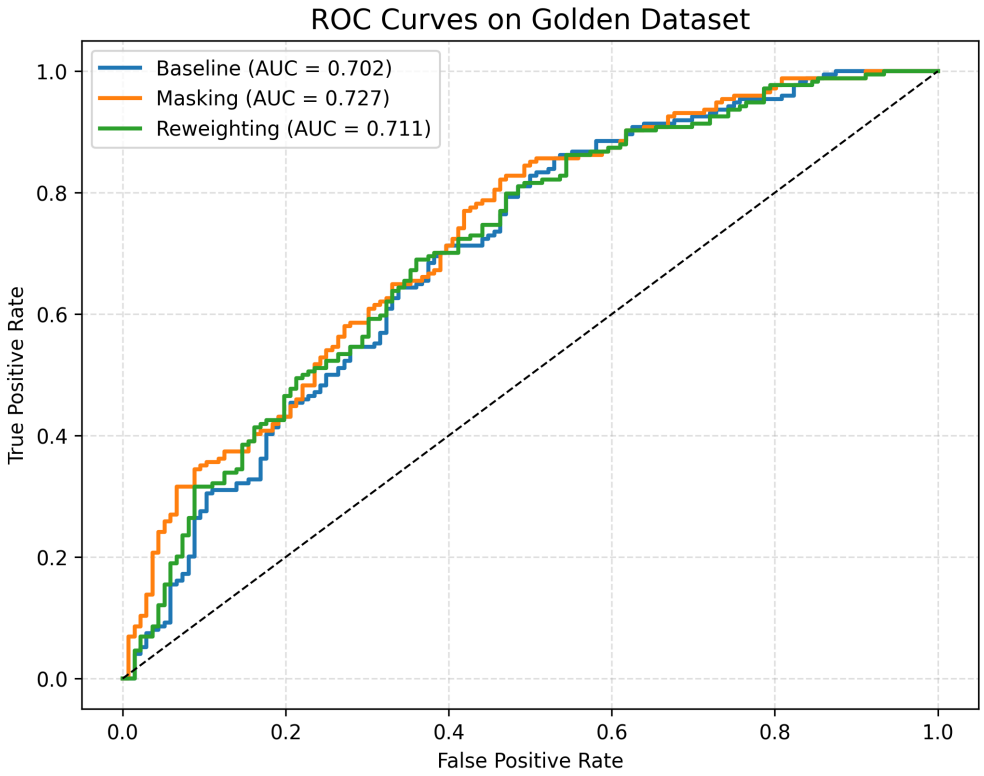

Figure 13: Receiver operating characteristic curves for the three transformer models on the Golden Dataset.

4.4 Summary

The evaluations reveal a consistent pattern across all forms of system risk. The models appear strong when evaluated on the legacy test split, yet the corrections and refined annotations in the Golden Dataset expose weaknesses in accuracy, fairness, ambiguity sensitivity, and agreement alignment. The Golden Dataset thus provides a clearer and more realistic account of model behavior and highlights the risks of relying solely on legacy labels for policy decisions.

5 Discussion

This study examined how content moderation models behave when evaluated through two very different lenses. The first lens used the original Civil Comments labels. The second lens used the Golden Dataset, a corrected and ambiguity aware subset created through careful

relabeling and agreement reconstruction. These two evaluations reveal a sharp contrast. The legacy labels present an appealing but ultimately misleading impression of model reliability. The Golden Dataset presents a more realistic and more governance aligned account of how the system behaves in practice. This difference reflects the core insight of the system risk framework. What appears to be a high performing model under conventional evaluation can become unstable, uneven, and error prone when examined under corrected labels that better represent human interpretation and policy standards.

The results show that the Golden Dataset is not simply a refined test set. It functions as a diagnostic instrument that uncovers structural weaknesses in the underlying system. It exposes forms of Origin Risk, Propagation Risk, Outcome Risk, and Governance Risk that are not visible under the legacy labels. The analysis demonstrates that the most important source of vulnerability is the underlying subjectivity of harmfulness judgments. Every downstream component of the system inherits this instability. These findings are consistent with earlier theoretical work on annotation subjectivity, label noise, bias inheritance, and fairness drift. The Golden Dataset shows how these phenomena emerge within a concrete moderation pipeline.

5.1 Origin Risk

The Golden Dataset reveals that a significant portion of model error originates from fundamental ambiguities in human judgments of harmful speech. Many comments in the Civil Comments dataset lie in a region where harm is contestable. These comments are not clearly harmful and not clearly harmless. Models trained on the original labels treat these cases as if there is a single unambiguous truth. The Golden Dataset shows that this assumption is incorrect. It contains many examples where the two human raters disagreed, where label changes occurred during review, or where the ambiguity code revealed internal uncertainty.

The decline in performance on mid range ambiguity examples confirms that the model is mirroring human subjectivity rather than overcoming it. The lexical stability test showed that small wording differences often flipped the target label or altered model confidence. The problem is not that the model lacks capacity. The problem is that the concept of harmfulness is context dependent and inherently unstable. This finding demonstrates a core form of Origin Risk. The training labels do not represent a consistent ground truth. Instead they represent a mixture of human judgments influenced by interpretation, emotional response, and cultural background. The model inherits these tensions in the form of variability, brittleness, and unexpected decision boundaries.

5.2 Propagation Risk

Propagation Risk concerns how unstable or biased patterns in the training data are carried forward into model behavior [2]. The Golden Dataset provides a clearer view of this process. Identity–based analyses showed that the model internalized patterns already present in the original labels. Although identity references are rare in the dataset, many appear disproportionately within harmful labels. Prior research shows that annotators often react differently to identity terms even when the surrounding content is benign, and this pattern was evident in the exploratory analysis where pointwise mutual information values revealed elevated associations between identity terms and harm labels.

These associations propagated directly into the model. Several identity groups exhibited elevated false positive rates across all models. Identity masking reduced this effect but did not eliminate it, indicating that the model had learned proxy features that persisted even after explicit identity terms were removed. Probability distributions and confusion patterns further showed that many predictions clustered near zero or near one, another sign of shortcut learning. Recent interpretability research confirms that such shortcuts are detectable through model–internal attribution patterns. Yadav et al. [14] demonstrate that identity terms can receive disproportionately high attribution even in supportive or neutral contexts, reinforcing the conclusion that the model relies on superficial lexical cues rather than underlying meaning.

Instead of modeling full linguistic meaning, the system reproduces correlation–based signals embedded in the legacy labels. This aligns with earlier work showing that toxicity classifiers often rely on shallow lexical patterns rather than substantive semantic reasoning. The Golden Dataset makes this propagation effect visible in a way that the legacy evaluation could not.

5.3 Outcome Risk

Outcome Risk concerns the distribution of errors across groups and contexts [2]. The Golden Dataset exposes substantial disparities that were hidden in the legacy test evaluation. Subgroup areas under the curve varied widely. Some groups experienced reduced discrimination performance. Others experienced inflated false positive rates or elevated false negative rates. These disparities are not incidental. They arise from the combination of label subjectivity, identity term correlations, and the shortcuts described above.

The two mitigation strategies illustrate the complexity of managing Outcome Risk. Identity masking reduced false positives for several identity groups. However, it also reduced the detection of harmful content in cases where identity context matters for interpreting threats or slurs. Sample reweighting increased the recall of harmful examples but produced new

pockets of over moderation that disproportionately affected some groups. These results show that fairness improvements come with tradeoffs. Improving one form of disparity can worsen another. The Golden Dataset provides the clarity needed to evaluate these tradeoffs. The legacy labels obscure these dynamics by presenting the system as uniformly strong. In reality the system behaves differently across identity contexts and across the spectrum of ambiguity.

5.4 Governance Risk

The most significant finding of the entire study is Governance Risk. Governance Risk concerns the difference between the appearance of model performance and the model's actual behavior under corrected evaluation [2]. The legacy test results suggested that all models were reliable. The Golden Dataset evaluation contradicted that impression. Every model experienced a decline in area under the curve. The baseline model declined the most. The Golden receiver operating characteristic curves reveal weaker discrimination, especially in ambiguous content. The model human disagreement matrix shows that models diverge from human judgments precisely in the cases where human interpretation is most uncertain. This difference between legacy performance and Golden performance is not a minor statistical fluctuation. It is evidence that legacy evaluation frameworks can mask underlying risk.

The findings suggest that model approval processes and platform governance decisions based solely on legacy metrics are inadequate. Evaluations that rely on noisy labels underestimate error rates, overestimate fairness, and misrepresent the system's true reliability. Governance aligned evaluation requires corrected datasets, ambiguity awareness, and identity specific analysis. The Golden Dataset provides this structure. Although it is limited in size, it functions as a governance instrument that reveals trends not visible in the original labels.

5.5 Effectiveness and Limits of Mitigation Methods

The mitigation methods used in this study were intentionally simple. They modified the input representation or the loss function rather than changing the model architecture. This approach reflects the constraints of real world moderation systems where incremental interventions are more feasible than full system redesigns. The results show that masking and reweighting shift model behavior in predictable ways but do not fully resolve the underlying risk patterns. Masking reduces the influence of explicit identity markers but does not remove proxy features. Reweighting improves detection of harmful content but introduces new imbalances in error allocation. These findings show that fairness interventions are not self contained solutions. They must be viewed as part of a larger governance strategy that includes corrected labels, ambiguity management, reviewer feedback loops, and ongoing

monitoring.

5.6 Implications for Real World Moderation

The differences between legacy evaluation and Golden evaluation demonstrate that automated moderation systems are vulnerable to structural blind spots. Systems that appear reliable under conventional benchmarks may behave inconsistently across users and contexts. Ambiguous content is especially difficult to model, yet ambiguous content is common in real online discourse. The analysis suggests several implications for real world moderation.

First, platforms should not rely on single score evaluations produced from noisy legacy labels. Second, identity linked disparities require structured audits that use corrected labels and identity aware metrics. Third, ambiguity sensitive workflows should be incorporated into real moderation pipelines. Models should not make definitive judgments on content that humans themselves interpret inconsistently. Fourth, mitigation methods should be evaluated using governance aligned datasets. This can prevent premature confidence in interventions that shift error patterns rather than reducing them. The Golden Dataset approach provides one possible template for such governance aligned evaluation.

5.7 Limitations

The Golden Dataset is limited in size and depends on two human raters. However, this limitation reflects the realistic constraints of creating corrected and adjudicated labels. Curated datasets are often smaller by design. Despite its size, the Golden Dataset revealed consistent patterns in ambiguity, disagreement, and identity based disparities. A larger adjudicated set may provide finer statistical resolution, but the core patterns would likely remain. The stratified sample used for transformer training was grounded in established sampling methods and preserved class proportions. Nevertheless, a larger sample may capture additional linguistic variation. Finally, the models evaluated were DistilRoBERTa models. Larger models may show different patterns of sensitivity to ambiguity or identity terms. These limitations provide opportunities for future improvement rather than weaknesses in the central argument.

5.8 Future Work

Future research should expand the data and evaluation framework by exploring larger corrected datasets, multi annotator adjudication pipelines, and expanded identity lexicons. Larger models may capture context and nuance more effectively. Additional fairness oriented

interventions such as calibration adjustments and counterfactual evaluations may yield more balanced outcomes across identity groups. It may also be valuable to study temporal drift by applying Golden Dataset style evaluation across multiple time periods. Further work could investigate how policy changes interact with ambiguity, disagreement, and subgroup disparities. These directions would deepen the system risk framework and improve alignment between automated moderation and real world policy goals.

6 Conclusion

This study examined how policy-driven AI systems can inherit and amplify risks that originate in data collection, annotation practices, and evaluation design. Using the Civil Comments dataset as a case study, we showed that label ambiguity, identity-linked correlations, and uneven subgroup performance can all influence the behavior of toxicity classifiers in ways that are not visible through standard accuracy measures. The Golden Dataset, built with corrected labels and explicit ambiguity codes, allowed us to evaluate models under conditions that more closely reflect policy expectations and human judgment.

Across the four dimensions of system risk, our findings demonstrate a substantial Governance Drift. While transformer models appeared reliable on the legacy test set (AUC $\approx$ 0.84), their performance experienced a sharp decline on the Golden Dataset (AUC $\approx$ 0.70), proving that legacy evaluation substantially overstates alignment with policy intent. Error analysis confirmed that models rely on surface-level cues rather than deeper linguistic meaning. Furthermore, mitigation strategies such as identity masking and sample reweighting reduced several identity-linked disparities but introduced tradeoffs that highlight the complexity of achieving fairness in a production environment.

Overall, this work underscores the importance of governance-aligned evaluation, careful label design, and identity-aware analysis when developing AI systems intended for policy or safety-critical settings. Future work should expand the Golden Dataset, incorporate additional annotators, and explore more advanced mitigation techniques to better support reliable and equitable automated moderation.

References

[1] U.S. House of Representatives. Protecting speech from government interference and social media bias: Part 1 – twitter's role in suppressing the biden laptop story. Hearing before the Committee on the Judiciary, 118th Congress, 2023. Accessed: 2025-10-30.

[2] National Institute of Standards and Technology. Artificial intelligence risk management framework (ai rmf 1.0). Technical report, U.S. Department of Commerce, Gaithersburg, MD, 2023. NIST Special Publication 1270.

[3] Wenbo Zhang, Hangzhi Guo, Ian D. Kivlichan, Vinodkumar Prabhakaran, Davis Yadav, and Amulya Yadav. A taxonomy of rater disagreements: Surveying challenges & opportunities from the perspective of annotating online toxicity. *arXiv preprint arXiv:2311.04345*, 2023. See p. 2 on the subjectivity of toxicity judgments.

[4] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. Like trainer, like bot? inheritance of bias in algorithmic content moderation. *Social Media + Society*, 3(2):405–415, 2017.

[5] Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. Handling bias in toxic speech detection: A survey, 2022.

[6] Juan F. Gomez, Caio Machado, Lucas M. Paes, and Flavio Calmon. Algorithmic arbitrariness in content moderation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, 2024.

[7] Google Jigsaw, Conversation AI, and TensorFlow Datasets. Civil comments dataset. Available at TensorFlow Datasets: https://www.tensorflow.org/datasets/catalog/civil_comments, 2018. Accessed: 2025-11.

[8] John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, and Ion Androutsopoulos. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4306, 2020.

[9] X. Yang. Diagnosing hate speech classification: Where do humans and machines disagree, and why? *arXiv preprint arXiv:2410.10153*, 2024.

[10] Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. Ss-bert: Mitigating identity terms bias in toxic comment classification by utilising the notion of "subjectivity" and "identity terms". *arXiv preprint arXiv:2109.02691*, 2021.

[11] Ji Ho Park, Jyun-Yi Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2799–2804, 2018.

[12] Flavien Prost, Pranjal Awasthi, Nick Blumm, Aditee Kumthekar, Trevor Potter, Li Wei, Xuezhi Wang, Ed H. Chi, Jilin Chen, and Alex Beutel. Measuring model fairness under noisy covariates: A theoretical perspective. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 873–883, 2021.

[13] Oscar Blessed Deho, Michael Bewong, Selasi Kwashie, Jiuyong Li, Jixue Liu, and Srecko Joksimovic. Is it still fair? a comparative evaluation of fairness algorithms through the lens of covariate drift. *Machine Learning*, 114(8), 2025.

[14] Sargam Yadav, Abhishek Kaushik, and Kevin McDaid. Understanding interpretability: Explainable ai approaches for hate speech classifiers. In *Explainable Artificial Intelligence*, 2023.

[15] Chien H. Tong. Refinement strategies for stratified sampling methods. *Reliability Engineering & System Safety*, 91(10–11):1257–1265, 2006.

[16] Thai-Nghe Nguyen, Zachary Gantner, and Lars Schmidt-Thieme. Cost-sensitive learning methods for imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 22(9):1263–1284, 2010.

[17] Burr Settles. Active learning literature survey. Technical Report Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[18] Jigsaw/Google. Perspective api: About the attributes, 2024.

[19] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, 2022.

APPENDIX A — GitHub Repository

All project code and documentation are available at: https://github.com/mrasheed88/system_risk_in_policy_driven_ai