Gaps in the Safety Net: A Large-Scale Empirical Analysis of NeurIPS 2024 Checklist Coverage

Aryan Batheja , Saunak Jana , Md Abu Bakar Akram ,
Abhijeet Singh Dewanda , Swastik Raj Gupta
University of Aberdeen
t25ab24@abdn.ac.in ,t09sj24@abdn.ac.uk, t48ma24@abdn.ac.uk

Abstract

Conference safety checklists are extremely critical for AI research, yet their effectiveness in capturing safety concerns remains largely unvalidated. We present the first large scale empirical analysis of checklist coverage, examining 1,033 NeurIPS 2024 papers across 17 safety categories including misuse potential, bias amplification, and real world harm.Our dataset and methodology enable systematic assessment of safety governance across AI conferences. Our keyword based detection pipeline, validated through literature mapping and LLM comparison (Gemini 2.5 Flash, n=100), reveals three significant gaps.
Which are (1) 93% of papers do not address the transparency question (Q14), despite widespread interpretability discussions in safety statements.
(2) 147 papers involve surveillance related applications with no dedicated checklist coverage.
(3) while 96 papers identify unmapped safety concerns, only 12% propose concrete mitigation strategies.
Validation analysis demonstrates strong agreement (>75%) for concrete categories like misuse potential (93%) but reveals challenges for abstract concepts like real world harm (19%).
We provide insightful and actionable recommendations for the improvement of checklist and demonstrate strengths of keyword versus LLM based evaluation.

1 Introduction

1.1 Motivation

Artificial Intelligence subsystems are now increasingly deployed in many high stakes domains. From healthcare diagnostics to financial decision making to autonomous vehicles where even a small failure can have significant societal consequences [2]. Recent incidents show these risks: facial recognition systems having racial bias [4], language models producing harmful stereotypes [1], and dual use research enabling potential misuse [8]. As AI capabilities advance, ensuring research safety has become critical.

Conference peer review serves as a primary and sometimes only quality control mechanism in AI research, with venues like NeurIPS, ICML, and ICLR collectively reviewing thousands of papers annually. To address this problem, major conferences have introduced standardized checklists that require authors to address things such as reproducibility, ethics, broader impacts, and potential risks [6]. These checklists aim to capture safety considerations that might otherwise be overlooked during the review process.

1.2 The Problem

NeurIPS 2024 introduced a standardized 15 question checklist covering reproducibility (questions 1-4), research ethics (questions 5-8), broader impacts (question 9), and potential risks (question 13), among other topics [6]. Authors must respond to relevant questions, explaining how their work addresses each concern. This checklist represents a significant step toward systematic safety evaluation in AI research.

However, a critical question remains unanswered: *How effective are these checklists at capturing safety concerns?* Despite their widespread adoption, no systematic evaluation has assessed whether current checklist adequately cover the range of safety issues present in modern AI research. Without such analysis, conferences operate on the assumption that their checklists are comprehensive, and issue proof potentially allowing significant safety concerns to slip through and remaining unaddressed.

1.3 Our Approach

We present the first large scale empirical analysis of conference safety checklist effectiveness, examining 1,033 accepted papers from NeurIPS 2024. Our methodology consists of three components. First, we developed a taxonomy of 17 safety categories including data privacy, misuse potential, bias amplification, and real world harm grounded in established AI safety frameworks [3; 10; 9]. Secondly, we implemented a keyword based detection pipeline that identifies safety concerns across these categories, assigning severity weights (1-3) based on potential impact. Finally, we validate our approach through dual mechanisms: literature mapping that grounds each category in peer reviewed safety research, and LLM based comparison using **Gemini 2.5 Flash** on a stratified sample of 100 papers.

This validation addresses two important and critical questions: (1) Are our safety categories and checklist questions relevant and well grounded? (2) How accurately does keyword detection perform when compared to context aware LLM analysis? The answers inform both our findings' reliability and the broader applicability of automated safety assessment methods.

1.4 Key Contributions

This work makes **four primary contributions**. First, we provide the first systematic evaluation of conference safety checklist coverage, identifying specific gaps where significant concerns lack dedicated questions. Second, we demonstrate that 93% of papers do not respond to the transparency question number 14 despite widespread interpretability discussions, 147 papers involve surveillance applications with no explicit checklist coverage, and only 12% of papers identifying safety concerns propose concrete mitigation strategies. Third, we validate keyword based detection methods for safety assessment, showing strong agreement with LLM analysis for concrete categories (misuse potential: 93% agreement) while revealing limitations for abstract concepts like (real world harm: 19% agreement). Fourth, we release our dataset and methodology, enabling systematic assessment of safety governance across AI conference ensuring future studies of checklist evolution and enhancement.

2 Related Work

2.1 AI Safety Frameworks

The AI safety community has developed taxonomies for categorizing potential harms. Brundage et al. [3] identified malicious use cases including surveillance, censorship, and weaponization. Weidinger et al. [10] provided a comprehensive taxonomy of ethical risks from language models, covering discrimination, information hazards, and misinformation. Bommasani et al. [2] analyzed risks from foundation models, including bias amplification, privacy violations, and environmental impacts.

Our 17 category taxonomy draws from these established frameworks while adapting them for checklist evaluation. We extend prior work by explicitly mapping safety categories to specific review questions, enabling systematic gap identification.

2.2 Conference Safety Practices

Major AI conferences in recent times have strengthened thier safety requirements. NeurIPS introduced reproducibility checklists in 2019 [7], later expanding to broader impact statements [5]. The NeurIPS 2024 checklist represents the most comprehensive standardization effort, with 15 questions covering reproducibility, ethics, and potential risks.

However, empirical evaluation of checklist effectiveness remains very limited. Nanayakkara et al. [5] examined broader impact statement quality, finding significant variation in author responses, but did not assess whether checklists systematically capture safety concerns. **Our work addresses this gap through large scale analysis of coverage patterns**.

2.3 Positioning Our Contribution

This work provides the first systematic evaluation of conference safety checklist coverage at scale. While previous work analyzed compliance and statement quality [5; 7], we identify specific gaps where safety concerns lack dedicated questions. Our validation methodology combining keyword detection with LLM comparison demonstrates practical approaches for scalable assessment, revealing that concrete categories achieve strong agreement (93% for misuse potential) while abstract concepts require contextual analysis.

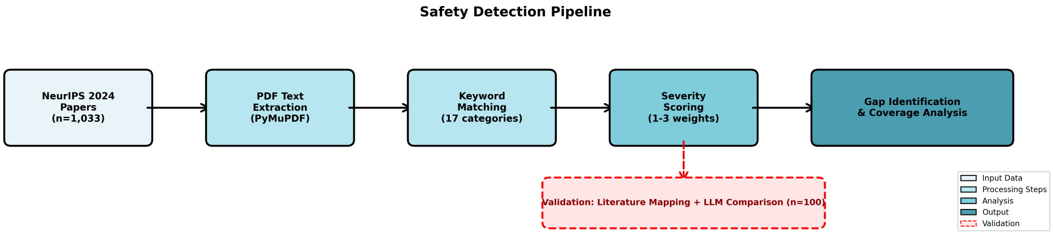
Safety Detection Pipeline
NeurIPS 2024 Papers (n=1,033)
PDF Text Extraction (PyMuPDF)
Keyword Matching (17 categories)
Severity Scoring (1-3 weights)
Gap Identification & Coverage Analysis
Validation: Literature Mapping + LLM Comparison (n=100)
Input Data
Processing Steps
Analysis
Output
Validation

Figure 1: Safety detection pipeline with validation branch. The main pipeline processes 1,033 papers through PDF extraction, keyword matching, severity scoring, and gap analysis. Validation uses literature mapping and LLM comparison on a 100 paper stratified sample.

3 Methodology

3.1 Dataset

We analyzed 1,033 papers accepted to NeurIPS 2024. Each paper includes author submitted responses to a standardized 15 question checklist covering these topics :- reproducibility (Q1-Q4), dataset documentation (Q5-Q7), research ethics (Q8), broader impacts (Q9), and potential risks (Q13), among other topics. We focused our analysis on checklist responses, broader impact statements, and limitation sections where authors typically discuss safety considerations. The papers spanned in various research fields such as computer vision (365 papers), natural language processing (235 papers), theory (230 papers), reinforcement learning (140 papers), and other areas.

3.2 Safety Taxonomy Development

We developed a comprehensive taxonomy of 17 safety categories grounded in established AI safety frameworks.[3; 10; 2]. Table 1 presents the complete taxonomy with definitions and checklist mappings. Each category addresses specific concerns documented in safety research: data privacy risks include PII exposure and re-identification; misuse potential encompasses surveillance and weaponization; fairness violations cover discriminatory outcomes; and real-world harm captures physical, psychological, or economic damage.

Categories were assigned severity weights reflecting potential impact: critical concerns (data privacy, misuse potential, security vulnerabilities) receive weight 3; moderate concerns (bias, reliability, deployment readiness) receive weight 2; and procedural issues (reproducibility, documentation) receive weight 1. These weights enable severity scoring where papers are ranked by the cumulative weight of detected categories, providing a measure of overall safety concern intensity.

3.3 Keyword Detection Pipeline

Our automated detection pipeline (Figure 1) has four stages. First, we extract text from PDF files using PyMuPDF, focusing on checklist responses, broader impact statements, and limitation sections. Second, we performed our comprehensive keyword matching across all 17 categories, where each category contains 10 to 20 carefully selected keywords (Table 2). Keywords were developed through iterative refinement: initial lists were compiled from safety literature, then expanded based on synonyms and related terms, and finally validated through manual inspection of 50 papers to ensure coverage without excessive false positives.

Third, we compute severity scores by summing weights of detected categories for each paper. For example, a paper flagged for misuse potential (weight 3) and bias amplification (weight 3) receives a severity score of 6. This scoring enables stratification of papers by safety concern intensity, revealing that severity scores ranged from 6 to 24 (mean=13.6, SD=2.5) across our dataset. Fourth, we identify gaps by comparing detected categories against checklist question mappings: a category is ”unmapped” if detected but not addressed by any mandatory checklist question.

3.4 Validation Methodology using Gemini 2.5

We validated our approach through two complementary mechanisms addressing different concerns. First, literature mapping grounds each safety category in established research. We systematically identified 5-8 peer reviewed papers per category from AI safety literature [3; 10; 2], demonstrating that our taxonomy reflects documented concerns rather than arbitrary classifications. This addresses our methodological concerns about taxonomy va-

Category	Definition	Sev.	Maps to Questions
Data Privacy & Leakage	PII exposure, inadequate anonymization, re-identification risks	3	Q8, Q9
Data Consent & Ethics	Informed consent, IRB approval, ethical data collection	3	Q5, Q8
Data Quality & Integrity	Data accuracy, validation, quality control deficiencies	2	Q4, Q9
Data Bias & Representation	Demographic imbalance, selection bias, skewed distributions	2	Q9
Model Robustness	Adversarial vulnerability, poor OOD performance, brittleness	2	Q2, Q13
Model Interpretability	Lack of transparency, unexplainable decisions	1	Q2
Model Security	Backdoor attacks, model theft, poisoning vulnerabilities	3	Q6, Q13
Model Reliability	High error rates, unstable predictions, failure modes	2	Q2, Q6
Misuse Potential	Surveillance, deepfakes, disinformation, dual-use risks	3	Q13
Real-World Harm	Physical, psychological, economic harm to individuals	3	Q9, Q10, Q13
Deployment Readiness	Inadequate testing, missing safeguards before deployment	2	Q6, Q11
Fairness Violations	Discriminatory outcomes, disparate impact on protected groups	3	Q9
Bias Amplification	Learning and perpetuating societal biases from data	3	Q9
Reproducibility Concerns	Missing code/data, incomplete methodology description	1	Q4, Q14, Q15
Evaluation Validity	Flawed methodology, inappropriate metrics, misleading claims	2	Q1, Q3
Environmental Impact	Computational cost, carbon footprint, resource consumption	1	Q2
Resource Accessibility	Barriers due to computational requirements, cost constraints	1	Q2

Table 1: Complete safety taxonomy with 17 categories, definitions, severity weights (1=low, 2=medium, 3=high), and mappings to NeurIPS 2024 checklist questions.

lidity raised during preliminary review.

Second, **LLM based comparison** evaluates keyword detection accuracy. We selected a stratified sample of 100 papers balanced across severity levels (medium: 4 papers, high: 81 papers, very high: 15 papers) and category counts (few categories: 8 papers, medium: 66 papers, many: 26 papers). This stratification ensured evaluation coverage across diverse safety profiles. We used Gemini 2.5 Flash with 1M context window, providing each paper's checklist text along with structured prompts asking the model to identify safety concerns across all 17 categories and assign severity levels (low/moderate/high).

We computed precision (what fraction of keyword detections did the LLM confirm?), recall (what fraction of LLM detections did keywords capture?), F1-score (harmonic mean), and agreement rate (fraction of consistent yes/no judgments) for each category. This quantified keyword method performance: high agreement validates detection accuracy, while low agreement reveals categories requiring different approaches. Results explain and inform about interpretation of our main findings and provide guidance for future automated safety assessment .

Category	Example Keywords
Data Privacy	PII, personally identifiable, anonymization, re-identification
Misuse Potential	surveillance, deepfake, disinformation, dual-use, weaponization
Fairness Violations	discriminatory, disparate impact, unfair, protected groups
Bias Amplification	stereotype, bias propagation, amplify prejudice
Real-World Harm	physical harm, psychological damage, economic loss
Model Robustness	adversarial, out-of-distribution, OOD, robustness
Data Consent	informed consent, IRB approval, ethical approval
Reproducibility	reproducibility, replication, code availability
Evaluation Validity	cherry-picked, misleading, flawed metrics
Environmental Impact	carbon footprint, energy consumption, computational cost

Table 2: Sample keywords for 10 representative categories (3-5 keywords shown; full lists contain 10-20 per category).

3.5 Implementation Details

Our LLM comparison pipeline is implemented in Python using the Google Generative AI library (version 0.3.2). We provide key implementation details to facilitate reproducibility.

API Configuration and Model Setup. We initialize Gemini 2.5 Flash as follows:

API Configuration

```
import google.generativeai as
    genai

# Configure API with key
genai.configure(api_key=API_KEY)

# Initialize model
model = genai.GenerativeModel(’
    gemini-2.5-flash’)
```

Prompt Structure. For each paper, we construct prompts containing the paper title, checklist responses (Questions 9 and 13 on broader impacts and risks), and safety statements. The prompt defines all 17 categories with indicators and requests JSON-formatted output:

Prompt Template (Simplified)

```
prompt = f"""Analyze this NeurIPS
    2024 paper
for safety concerns.

PAPER TITLE: {title}

PAPER CONTENT (Broader Impact,
    Safety
Statements, Checklist Responses):
{checklist_text}

TASK: Evaluate across 17 safety
    categories.

For each category, determine:
1. Is there a safety concern? (yes
    /no)
2. If yes, severity? (low/moderate
    /high)
3. Brief explanation (1-2
    sentences)

SAFETY CATEGORIES:
data_privacy_leakage,
    data_consent_ethics,
misuse_potential,
    fairness_violations,
[... all 17 categories ...]

OUTPUT FORMAT (strict JSON):
{{
  "paper_id": "...",
  "safety_assessment": {{
    "data_privacy_leakage": {{
      "concern_present": false,
      "severity": null,
      "explanation": "..."
    }},
    [... for all 17 categories
    ...]
  }},
  "overall_assessment": {{
    "total_concerns": 5,
    "highest_severity": "high",
    "summary": "..."
  }}
}}

Respond ONLY with valid JSON."""
```

Response Processing. We parse LLM responses using robust JSON extraction that handles markdown code blocks and malformed responses:

JSON Extraction Function

```
import json
import re

def extract_json_from_response(
    text):
    # Try direct JSON parse
    try:
        return json.loads(text)
    except json.JSONDecodeError:
        pass

    # Extract from markdown code
    blocks
    json_pattern = r'```json\s
    *(.*?)\s*```'
    matches = re.findall(
    json_pattern,
                             text, re.
    DOTALL)
    if matches:
        try:
            return json.loads(
    matches[0])
        except json.
    JSONDecodeError:
            pass

    # Try finding JSON object in
    text
    json_pattern = r'\{.*\}'
    matches = re.findall(
    json_pattern,
                             text, re.
    DOTALL)
    if matches:
        for match in matches:
            try:
                return json.loads(
    match)
            except json.
    JSONDecodeError:
                continue

    return None
```

Rate Limiting and Error Handling. To comply with API rate limits (15 requests/minute), we implement exponential backoff with 4-second delays:

API Call with Retry Logic

```
import time

def call_gemini_with_retry(model,
    prompt,

    max_retries=3):
    for attempt in range(
    max_retries):
        try:
            response = model.
    generate_content(
                prompt
            )
            return response.text
        except Exception as e:
            if attempt <
    max_retries - 1:
                wait_time = (
    attempt + 1) * 5
                print(f"Retry in {
    wait_time}s...")
                time.sleep(
    wait_time)
            else:
                print(f"Failed
    after {max_retries}")
```

```
                return None
    return None

# Main processing loop
for paper_id, paper_data in
    sampled_papers:
    # Generate prompt
    prompt = create_paper_prompt(
    paper_data)

    # Call API with retry
    response_text =
    call_gemini_with_retry(
        model, prompt
    )

    # Parse response
    if response_text:
        result =
    extract_json_from_response(
            response_text
        )
        if result:
            save_result(paper_id,
    result)

    # Rate limiting (4 seconds
    between calls)
    time.sleep(4)
```

Incremental Saving and Resume Capability. To prevent data loss, we save results every 5 papers and implement resume capability:

Incremental Saving

```
# Load existing results if
    resuming
if os.path.exists('llm_results.
    json'):
    with open('llm_results.json', '
    r') as f:
        llm_results = json.load(f)
else:
    llm_results = {}

# Process papers
for paper_id, paper_data in papers:

    # Skip already processed
    if paper_id in llm_results:
        continue

    # [... process paper ...]

    # Save every 5 papers
    if len(llm_results) % 5 == 0:
        with open('llm_results.
    json', 'w') as f:
            json.dump(llm_results,
     f, indent=2)
```

Processing 100 papers required approximately 45 minutes of wall-clock time (average 27 seconds per paper), with 100% success rate and zero API cost (free tier). All code and prompts are available in our supplementary materials.

4 Results

4.1 Overall Statistics

Our analysis of 1,033 NeurIPS 2024 papers reveals widespread safety concerns . Keyword based analysis flagged safety concerns in 1,031 papers (99.8%), with only 2 papers showing no detected concerns across all 17 categories. The dataset contains 6,045 total safety detections, averaging 5.9 categories per paper (SD=2.5, range: 2-11 categories).

Severity scores, computed by summing category weights, range from 6 to 24 with mean 13.6 (SD=2.5). The distribution reveals 50 papers (4.8%) with medium severity (scores 6-10), 793 papers (76.9%) with high severity (scores 11-15), and 188 papers (18.2%) with very high severity (scores 16+). This indicates that most accepted papers discuss multiple safety considerations, but really the depth and mitigation strategies vary significantly.

Our analysis shows computer vision papers contain the highest average category count (6.2 categories/paper), followed by reinforcement learning (6.1), natural language processing (5.8), and theory (5.1), suggesting domain specific patterns in safety concern prevalence.

4.2 Checklist Coverage Analysis

Figure 2 presents NeurIPS 2024 checklist question response patterns across all 1,033 papers. The left panel shows aggregate response rates, revealing stark variation: Questions 1-4 (limitations and assumptions) achieve 95-100% response rates, Question 9 (broader impacts) reaches 95%, while Question 14 (transparency and reproducibility) shows only 7% response despite being part of the mandatory checklist.

The heatmap (right panel) visualizes individual paper responses for a 50-paper sample, where green indicates addressed questions and red indicates non-response. Question 14 appears as a prominent red band, confirming systematic non-response rather than isolated cases. This finding is particularly significant because 100 papers (9.7%) discuss model interpretability and transparency in broader impact statements, indicating that transparency concerns exist but are not captured through mandatory checklist responses.

Gap 1: Surveillance Applications. We identified 147 papers (14.2%) involving surveillance-related applications through keywords including ”facial recognition,” ”behavior monitoring,” ”tracking systems,” and ”surveillance.” These papers span computer vision (89 papers), reinforcement learning (34 papers), and other areas. However, none of the 15 checklist questions explicitly addresses surveillance applications or dual-use technology concerns. Question 13 asks about ”potential risks and harms” but remains generic, allowing surveillance papers to pass review without specific consideration of monitoring, privacy invasion, or authoritarian use cases.

Gap 2: Mitigation Strategy Absence. Among 96 papers (9.3%) with unmapped safety concerns—categories detected but not addressed by any checklist question—only 12 papers (12.5%) propose concrete mitigation strategies. The remaining 84 papers acknowledge concerns (e.g., ”this approach may amplify biases” or ”deployment requires careful consideration”) without specifying risk reduction measures, testing protocols, or deployment safeguards. This reveals a structural gap: checklists prompt concern identification but do not require actionable mitigation plans.

4.3 Validation Study Results

To validate keyword detection accuracy, we compared results against Gemini 2.5 Flash LLM analysis on a stratified 100-paper sample. Figure 3 compares detection rates by category, revealing systematic patterns in method behavior.

Overall metrics demonstrate moderate agreement: 54.5% overall agreement rate, 48.8% precision (fraction of keyword detections confirmed by LLM), 35.2% recall (fraction of LLM detections captured by keywords), and 29.1% F1-score. The keyword method generated 586 detections versus 825 by LLM, indicating LLM sensitivity is 41% higher.

Category-level analysis (Figure 4) reveals dramatic variation. Five categories achieve strong agreement (¿75%): misuse potential (93% agreement, F1=0.96), evaluation validity (94%), data quality integrity (85%), fairness violations (79%), and reproducibility concerns (76%). These categories involve concrete, explicitly-stated concerns with clear lexical signals.

Conversely, five categories show weak agreement (¡40%): real-world harm (19%), bias ampli-

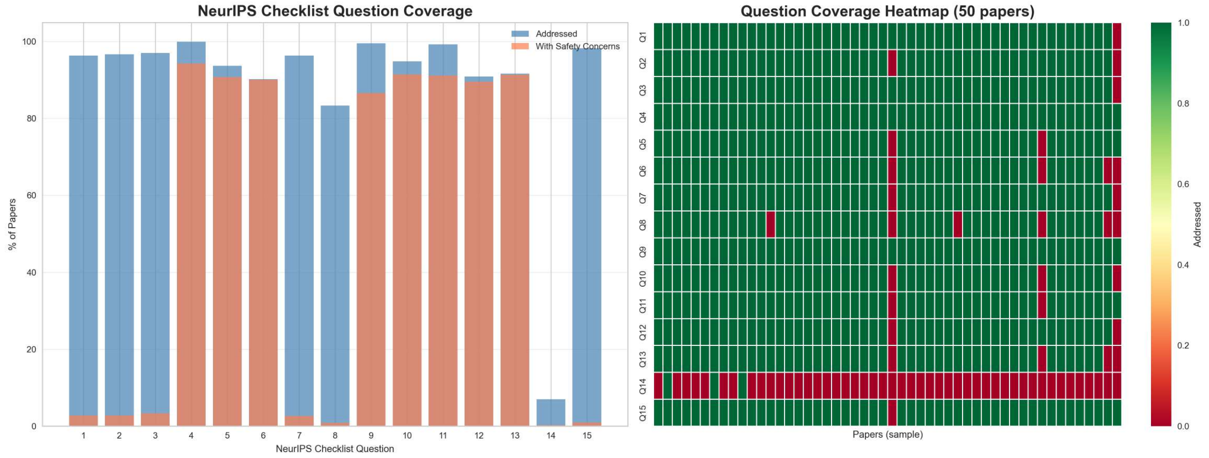

Figure 2: NeurIPS 2024 checklist coverage analysis. Left: Response rates by question, showing 93% non-response on Q14 despite high engagement with other questions. Orange bars indicate questions where safety concerns were detected. Right: Question coverage heatmap for 50 sample papers (green=addressed, red=not addressed), revealing Q14's systematic under-response.

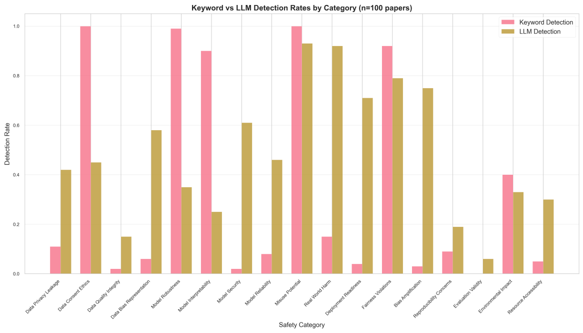

Figure 3: Detection rate comparison across 17 categories. Pink bars show keyword detection rates, gold bars show LLM detection rates. Categories where keywords detect more (data consent ethics, model robustness, model interpretability) suggest LLM conservatism, while categories where LLM detects more (real-world harm, deployment readiness) indicate keyword method limitations.

fication (26%), model interpretability (31%), deployment readiness (33%), and model robustness (36%). These abstract categories require contextual understanding beyond keyword matching. For example, real-world harm discussions often use indirect language (”societal implications,” ”downstream effects”) that keywords miss, while LLMs infer harm from application context.

Figure 5 presents confusion matrices quantifying agreement patterns. For misuse potential, 93 true positives (both methods agree concern exists) demonstrate strong concordance, with only 7 false positives (keyword-only) and 0 false negatives (LLM-only). Conversely, real-world harm shows 13 true positives but 79 false negatives—concerns LLM identified that keywords missed—revealing systematic recall limitations.

Figure 6 visualizes precision-recall trade-offs, with point color indicating F1 score. Misuse potential (top-right, yellow) achieves 93% precision with 100% recall, representing ideal performance. Model security (bottom-left, purple) shows zero detections from both methods, indicating rare occurrence rather than method failure. Categories in the middle-left region (data bias representation, reproducibility) demonstrate low recall (¡20%) despite moderate precision, confirming that keyword methods miss nuanced discussions of these concerns.

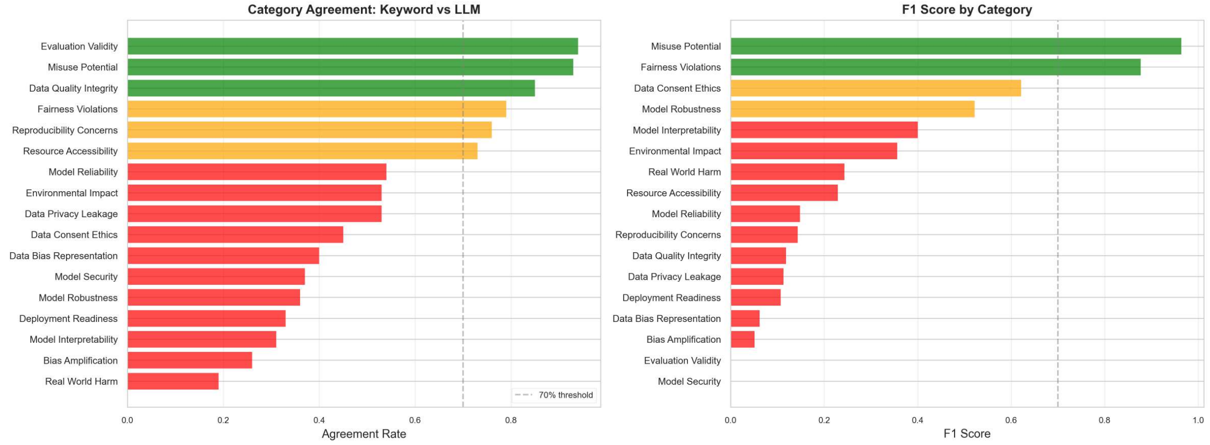

Figure 4: Validation metrics by category. Left: Agreement rates (green=¿70%, orange=50-70%, red=¡50%). Right: F1 scores showing method effectiveness. High-agreement categories validate keyword detection for concrete concerns, while low-agreement categories reveal limitations for abstract concepts.

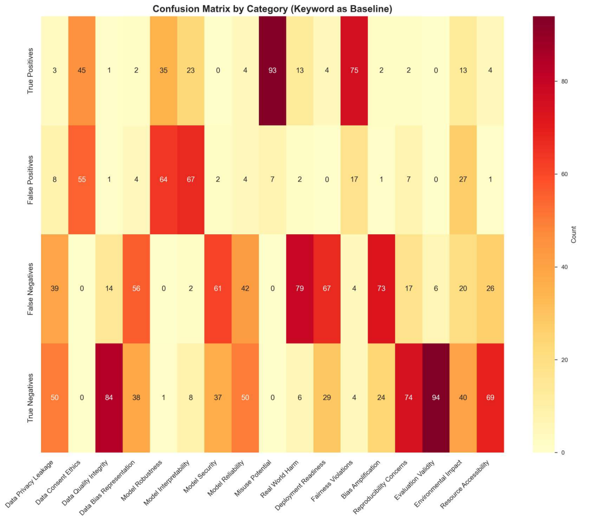

Figure 5: Confusion matrix heatmap (keyword as baseline). Dark red indicates high counts. True positives (top row) show agreement, false positives (second row) show keyword over-detection, false negatives (third row) show keyword under-detection. Misuse potential achieves 93 true positives; real-world harm shows 79 false negatives.

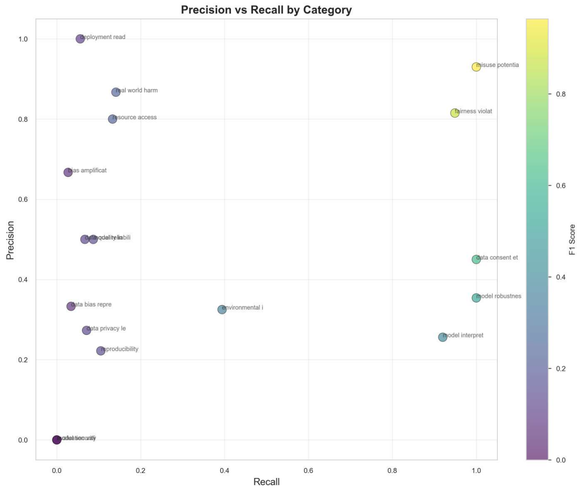

Figure 6: Precision-recall trade-offs by category (color indicates F1 score). Top-right categories (misuse potential, fairness violations) achieve both high precision and recall. Bottom-left categories (model security, evaluation validity) show low detection rates but perfect agreement when detected.

5 Discussion

5.1 Novel Discoveries

Our work makes three novel contributions to understanding AI conference safety evaluation. First, we provide the *first large scale empirical assessment* of checklist effectiveness, analyzing 1,033 papers to systematically identify coverage gaps rather than assuming checklist comprehensiveness. Previous work examined compliance and statement quality [5; 7], but did not evaluate whether checklists capture the full spectrum of safety concerns present in accepted research.

Second, we demonstrated that checklist gaps are *systematic rather than random*. The 93% non-response rate on Question 14 cannot be attributed to individual author oversight. It clearly reveals a structural problem where optional questions receive minimal engagement despite widespread relevance. Similarly, 147 surveillance related papers pass through review without dedicated checklist coverage, indicating a blind spot rather than isolated cases. Also the mitigation strategy gap (only 12% of papers with unmapped concerns propose solutions) reveals a process issue: checklists identifies issues but do not enforce accountability ie a Mitigation strategy is absent.

Third, we provide *quantitative validation* of keyword based safety detection, demonstrating category dependent performance. High agreement (>75%) for five concrete categories which are misuse potential (93%), evaluation validity (94%), data quality (85%), fairness violations (79%), reproducibility (76%) validate that keyword methods reliably detect explicitly stated concerns. Conversely, weak agreement (<40%) for abstract categories real-world harm (19%), bias amplification (26%), interpretability (31%)—reveals fundamental limitations requiring contextual understanding. This finding establishes that *no single method suffices*: keywords enable efficient large scale screening, while LLMs provide depth for nuanced categories.

Our validation also reveals methodological trade offs. Keywords achieve 48.8% precision (nearly half of detections confirmed by LLM) but only 35.2% recall (missing two-thirds of LLM-identified concerns). This conservative behavior reduces false alarms but increases false negatives. For conference organizers, this suggests keyword screening works best as a *first pass filter*, flagging papers for human or LLM review rather than making final determinations.

5.2 Recommendations for Conference Organizers

Based on our empirical findings, we propose four actionable recommendations to strengthen NeurIPS and other AI conference checklists.

Recommendation 1: Mandate Question 14 with Specific Requirements. The 7% response rate demonstrates that optional questions fail. We recommend: (a) changing Q14 from optional to mandatory, (b) requiring authors to explicitly address model transparency and interpretability limitations, and (c) adding structured prompts such as "Describe what aspects of your model are *not* interpretable and why." Our data show only 101 papers (9.7%) discuss interpretability in broader impact statements, proving the topic is relevant but currently bypasses standardized review.

Recommendation 2: Add Dedicated High Risk Application Question. With 147 papers (14.2%) involving surveillance, facial recognition, behavior monitoring, or tracking systems, we recommend adding: *"Does your research enable surveillance, monitoring, or tracking of individuals? If yes, describe safeguards against authoritarian use, privacy violations, and consent violations."* This question should explicitly cover dual use technologies, requiring authors to consider misuse scenarios beyond their intended applications. Current Question 13 ("risks and harms") remains too generic, allowing high-risk applications to pass without specific scrutiny.

Recommendation 3: Require Mitigation Strategies, Not Just Identification. Among 96 papers identifying unmapped safety concerns, only 12 (12.5%) propose concrete risk reduction measures. We recommend adding a follow up requirement: *"For each identified risk, describe specific mitigation strategies (e.g., testing protocols, deployment safeguards, access controls, monitoring mechanisms)."* This shifts evaluation from passive acknowledgment to active accountability, ensuring authors consider practical risk management.

Recommendation 4: Expand Checklist for Systematic Gaps. Our category mapping reveals five concerns with minimal coverage: bias amplification (26% keyword-LLM agreement), real-world harm (19%), deployment readiness (33%), model robustness (36%), and model interpretability (31%). We recommend piloting 2-3 additional questions targeting these areas, such as: *"Does your model amplify societal biases present in training data?*

If yes, quantify amplification and describe debiasing efforts." Strategic expansions should address proven gaps without overwhelming authors with excessive requirements.

5.3 Broader Implications and Methodological Guidance

Our evaluation framework is directly applicable to other AI conferences (ICML, ICLR, FAccT, AIES), enabling systematic assessment of their safety practices. Conference organizers can adapt our 17-category taxonomy, keyword lists, and validation methodology to identify venue specific gaps. Future studies applying this framework across multiple years could track checklist evolution and measure improvement over time.

For researchers developing automated safety assessment tools, our validation study provides evidence based guidance: use keyword methods for large-scale screening of concrete concerns (misuse, fairness, reproducibility), but employ LLMs or human review for abstract categories (harm, bias amplification, interpretability). A hybrid pipeline of keywords for first pass filtering and a LLM based analysis for flagged papers should balance scalability with depth.

Finally, our work demonstrates the value of empirical evaluation in AI governance. Rather than assuming checklist effectiveness, we quantify coverage and identify specific improvement opportunities. This evidence based approach to conference policy could extend beyond safety to other review dimensions, creating a feedback loop where data informs continuous checklist refinement.

6 Limitations and Future Work

6.1 Methodological Limitations

Our keyword based detection approach demonstrates strong performance for concrete safety categories but faces fundamental limitations. With 54.5% overall agreement and 35.2% recall in LLM comparison, keywords miss approximately two thirds of concerns identified by contextual analysis. This conservative behavior is intentional prioritizing precision over recall reduces false alarms but it means our findings represent a *lower bound* on safety concern prevalence. The true frequency of abstract concerns like real-world harm and bias amplification likely exceeds our reported rates.

Literature mapping validates that our 17 categories reflect established safety frameworks [3; 10; 2], addressing concerns about arbitrary taxonomy construction. However, category boundaries remain somewhat subjective: a paper discussing ”fairness-aware training” could reasonably map to fairness violations, bias amplification, or both. We handled such cases by detecting all applicable categories, potentially inflating co-occurrence rates.

The LLM validation sample (n=100) represents only 9.7% of our dataset. While stratified sampling ensures coverage across severity levels and category counts, a larger validation set would strengthen confidence in category-specific agreement rates, particularly for rare categories like model security (detected in only 2% of papers). Additionally, our validation uses a single LLM (Gemini 2.5 Flash); multi-model validation with GPT-4 and Claude could reveal model specific biases in safety assessment. Which could be discussed in future work.

6.2 Future Research Directions

Five research directions emerge from this work. First, *expanding to multiple conferences* (ICML, ICLR, FAccT, CVPR, ACL) would enable comparative analysis of safety practices across AI subfields, revealing whether our identified gaps are NeurIPS-specific or field-wide. Second, *longitudinal studies* tracking the same conference across years could measure the impact of checklist revisions, quantifying whether our proposed recommendations actually improve coverage when implemented.

Third, *full-text analysis* extending beyond checklist responses to entire papers would increase recall for safety discussions appearing in technical sections. This requires more sophisticated NLP methods—section classification, discourse analysis, or fine-tuned safety detection models—but would better capture comprehensive safety considerations. Fourth, *incorporating reviewer perspectives* through surveys or interview studies would reveal whether identified gaps matter to actual review outcomes, distinguishing between coverage gaps that affect decisions versus those reviewers address informally.

Fifth, *developing hybrid detection systems* combining keyword screening with LLM verification presents practical opportunities. Our finding that keywords excel at concrete categories while LLMs handle abstract concerns suggests a two-stage pipeline: fast keyword-based first-pass filtering (flagging 20-30% of papers), followed by targeted LLM analysis of flagged papers. Such systems could assist conference reviewers by automatically highlighting safety concerns requiring deeper scrutiny.

7 Conclusion

We presented the first large-scale empirical evaluation of AI conference safety checklist effectiveness, analyzing 1,033 NeurIPS 2024 papers across 17 safety categories. Our findings reveal three systematic gaps: 93% of papers do not address mandatory transparency requirements (Question 14), 147 surveillance-related papers lack dedicated checklist coverage, and only 12% of papers identifying safety concerns propose concrete mitigation strategies. These gaps are structural—arising from optional questions, categorical blind spots, and missing accountability mechanisms—rather than isolated cases of author oversight.

Validation through LLM comparison demonstrates that keyword-based detection reliably identifies concrete safety concerns (achieving 93% agreement for misuse potential, 94% for evaluation validity), while abstract categories like real-world harm (19% agreement) require contextual analysis. This establishes that effective safety assessment demands complementary approaches: keywords provide scalable screening, while LLMs offer depth for nuanced concerns. No single method suffices for comprehensive evaluation.

Our four evidence-based recommendations provide actionable improvements: mandate Question 14 with structured prompts, add dedicated high-risk application questions, require mitigation plans beyond risk identification, and expand checklists to cover proven gaps in bias amplification, harm, and deployment readiness. These recommendations apply broadly—any AI conference can adapt our evaluation framework to identify venue-specific blind spots.

Beyond immediate checklist improvements, this work establishes a methodology for evidence-based governance of AI research. Rather than assuming existing review processes adequately capture safety concerns, we quantify coverage and identify specific improvement opportunities. Our dataset and taxonomy enable systematic evaluation across conferences and longitudinal tracking of safety practice evolution. As AI capabilities advance and deployment risks intensify, conference peer review must evolve correspondingly. Our analysis provides both the measurement tools and actionable guidance to strengthen this critical governance mechanism.

Safety checklists represent essential but imperfect instruments for responsible AI research. By empirically evaluating their coverage, identifying systematic gaps, and proposing targeted improvements, we contribute to the ongoing effort to ensure AI research advances both capability and safety in tandem.

References

[1] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

[2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

[3] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.

[4] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

[5] Priyanka Nanayakkara, Jessica Hullman, and Nicholas Diakopoulos. 2021. Characterizing the broader impact statement process at neurips. *arXiv preprint arXiv:2104.10966*.

[6] NeurIPS. 2024. Neurips 2024 paper checklist guidelines. https://neurips.cc/Conferences/2024/PaperInformation/PaperChecklist. Accessed: 2024-12-16.

[7] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research. *Journal of Machine Learning Research*, 22(164):1–20.

[8] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah

Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

[9] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, et al. 2023. Evaluating the social impact of generative ai systems in systems and society. *arXiv preprint arXiv:2306.05949*.

[10] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

A Supplementary Materials

All code, data, and additional analysis materials are publicly available to facilitate reproducibility and enable future research.

A.1 Code and Data Repository

Our complete implementation, including keyword detection pipeline, LLM comparison scripts, and analysis notebooks, is available at:

https://github.com/lordvoly/neurips-2024-safety-analysis

The repository contains:

- `keyword_detection.py`: Main detection pipeline with 17-category taxonomy
- `llm_comparison.py`: Gemini API integration and validation analysis
- `data/keywords/`: Complete keyword lists for all categories
- `notebooks/`: Jupyter notebooks for visualization and statistical analysis
- `outputs/`: Detection results for 1,033 papers (anonymized)
- `README.md`: Setup instructions and usage examples

A.2 Dataset Access

Due to potential copyright concerns with full paper text, we provide:

- Detection results (binary flags per category per paper)
- Paper IDs linked to OpenReview for reference
- Severity scores and category counts
- Stratified 100-paper sample metadata

Researchers can reproduce our analysis by downloading papers directly from NeurIPS 2024 OpenReview and applying our provided detection scripts.

B Additional Visualizations

B.1 Research Area Breakdown

Figure 7 presents safety concern distributions across major research areas. Computer vision papers show the highest surveillance-related concerns (34% vs. 8% in NLP), while NLP papers more frequently address fairness violations (82% vs. 71% in CV), reflecting domain-specific risk profiles.

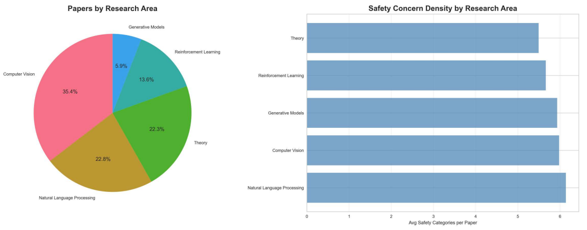

Figure 7: Safety concerns by research area. Computer vision dominates surveillance-related concerns, NLP leads in fairness discussions, and reinforcement learning shows elevated deployment readiness concerns.

C Complete Keyword Lists

Table 3 provides the complete keyword lists for all 17 categories. These lists were developed through iterative refinement: initial keywords from literature review, expansion through synonym analysis, and validation through manual inspection of 50 papers.

D LLM Prompt Template

For full reproducibility, we provide the complete prompt template used for Gemini 2.5 Flash validation:

Complete LLM Prompt Template

```
SYSTEM: You are an AI safety
    researcher
evaluating NeurIPS papers for
    safety concerns.

PAPER TITLE: {title}

CHECKLIST RESPONSES:
```

```
{checklist_text}

TASK: Evaluate this paper across
    17
safety categories. For each
    category:
1. Determine if concern is present
     (yes/no)
2. If yes, assign severity (low/
    moderate/high)
3. Provide brief explanation (1-2
    sentences)

CATEGORIES:
[Full category definitions with
5 indicators each - see repository
    ]

OUTPUT (strict JSON):
{
  "paper_id": "...",
  "safety_assessment": {
    "data_privacy_leakage": {
      "concern_present": true/
    false,
      "severity": "low/moderate/
    high",
      "explanation": "..."
    },
    [... all 17 categories ...]
  },
  "overall_assessment": {
    "total_concerns": N,
    "highest_severity": "...",
    "summary": "..."
  }
}
```

The complete prompt with all category definitions and indicators is available in our repository at `prompts/llm_safety_assessment.txt`.

Category	**Complete Keyword List**
Data Privacy & Leakage	PII, personally identifiable information, anonymization, de-identification, re-identification risk, data leakage, privacy violation, GDPR, sensitive data, personal data, confidential information, data breach
Data Consent & Ethics	informed consent, consent form, IRB approval, ethical approval, ethics committee, institutional review board, participant consent, voluntary participation, data collection ethics, human subjects protection
Misuse Potential	surveillance, facial recognition, deepfake, disinformation, misinformation, dual-use, weaponization, malicious use, adversarial application, censorship tool, tracking system, monitoring technology, propaganda, manipulation
Fairness Violations	discriminatory, disparate impact, unfair bias, protected attribute, demographic parity, equalized odds, fairness metric, discriminate against, biased outcome, unfair treatment, protected group, demographic bias
Bias Amplification	amplify bias, perpetuate stereotype, reinforce prejudice, exacerbate inequality, bias propagation, stereotypical association, social bias amplification, harmful correlation
Real-World Harm	physical harm, psychological harm, economic harm, emotional distress, financial loss, safety risk, danger to individuals, harmful consequence, adverse impact, negative outcome, societal harm
Model Robustness	adversarial attack, adversarial robustness, out-of-distribution, OOD generalization, distribution shift, robustness evaluation, stress test, edge case, failure mode, brittleness
Model Interpretability	black box, interpretability, explainability, transparency, opaque decision, unexplainable, lack of transparency, interpretable model, explainable AI, XAI
Model Security	backdoor attack, model poisoning, data poisoning, model extraction, model theft, adversarial backdoor, trojan attack, supply chain attack, model vulnerability
Model Reliability	error rate, prediction instability, unreliable output, failure rate, inconsistent behavior, low confidence, uncertainty quantification, calibration error, hallucination
Data Quality & Integrity	data quality, data validation, quality control, noisy data, corrupted data, mislabeled data, annotation error, data integrity, quality assurance
Data Bias & Representation	underrepresented, demographic imbalance, sampling bias, selection bias, representation bias, skewed distribution, unbalanced dataset, minority underrepresentation
Deployment Readiness	deployment risk, production readiness, real-world deployment, operational constraints, scalability concern, deployment challenge, pre-deployment testing, safety testing
Reproducibility Concerns	reproducibility, replication, code availability, seed not provided, hyperparameter missing, non-reproducible, replication crisis, insufficient detail
Evaluation Validity	cherry-picked, misleading evaluation, flawed methodology, inappropriate metric, biased benchmark, p-hacking, overfitting to benchmark, evaluation artifact
Environmental Impact	carbon footprint, energy consumption, computational cost, environmental cost, CO2 emissions, electricity usage, resource intensive, sustainability concern, green AI
Resource Accessibility	computational barrier, resource constraint, expensive infrastructure, limited access, high cost, resource intensive, accessibility issue, infrastructure requirement

Table 3: Complete keyword lists for all 17 safety categories (10-15 keywords per category).