Critical Analysis: PG-STORY: Taxonomy, Dataset, and Evaluation for Ensuring Child-Safe Content for Story Generation

Kai-Yu Lu

1 Methodological Strengths

1.1 Problem Framing and Taxonomy Design

The paper provides a clear and concrete problem formulation: child safety for narrative text generation, with a specific focus on children under ten years of age. This framing is more specific than generic "toxicity" detection and directly targets the story generation setting, which is methodologically appropriate.

A major strength is the design of a child-centric safety taxonomy with five categories: Profanity & Slurs, Sex & Nudity, Violence & Scariness, Substance Consumption, and Discrimination & Bias. The taxonomy is grounded in guidelines from Common Sense Media and national media rating standards, but is explicitly tailored to narrative text rather than video or audio. The categories are mutually intuitive and minimize overlap, which supports downstream learning and evaluation. The taxonomy goes beyond single-label toxic versus non-toxic schemes and allows models to distinguish different kinds of harmful content, which is crucial for targeted rewriting and analysis.

The paper also distinguishes sentence-level and discourse-level safety. Sentence-level labels capture local lexical and phrasal risks, while discourse-level labels capture cumulative effects such as sustained fear or bias. This dual granularity is an important methodological decision, because harmful impact in stories often arises from multi-sentence build-up rather than isolated terms.

1.2 Construction of UNSAFECORPUS

The UNSAFECORPUS dataset aggregates four offensive language datasets across Reddit, Twitter, Wikipedia talk pages, and YouTube comments. This multi-platform coverage exposes the safety classifier to diverse writing styles and topics and reduces overfitting to a single domain.

A key methodological strength is the enrichment of category labels through harmful lexicons. The authors manually label around 1,690 lexicon entries according to the five safety categories and then match these lexicons in the text. This substantially increases category coverage. For instance, Profanity & Slurs grows from 1,193 to 39,038 unsafe instances, and Violence & Scariness grows from 2,648 to 27,390 unsafe instances. Substance Consumption has no instances without lexicon matching and 993 after lexicon integration. This strategy compensates for the fact that the original datasets do not contain fine-grained category labels aligned with child safety.

The corpus is split into 60 percent training, 20 percent validation, and 20 percent test sets with balanced safe and unsafe counts (97,815 safe and 98,511 unsafe). This supports principled supervised learning and unbiased evaluation.

1.3 Specialized Safety Classifiers and Evaluation Protocol

Two BART-based classifiers are trained: a binary detection model and a multi-class categorization model. Both use a pre-trained BART encoder with additional non-linear and dropout layers before linear classification heads. This reuse of a strong encoder is methodologically sound.

The evaluation protocol on UNSAFECORPUS is carefully designed. For detection, the paper reports precision, recall, and F1 separately for safe and unsafe classes, as well as macro averages. This avoids masking poor unsafe recall by high safe recall. The comparison with Perspective API and Detoxify is systematic: both external tools are evaluated using their toxicity scores with a fixed threshold of 0.5. The results show that the specialized detector attains macro F1 of 96.9 percent, whereas Perspective API and Detoxify reach 67.1 percent. Unsafe recall is particularly improved, from about 41 percent for the external tools to 97.8 percent for the specialized model. This demonstrates the value of a child-centric taxonomy and domain-specific training.

For categorization, the specialized model achieves high F1 scores on four out of five categories. Profanity & Slurs and Discrimination & Bias have F1 around 93 points, Sex & Nudity around 89 points, and Violence & Scariness around 86 points. Substance Consumption has F1 around 49.9 points, reflecting data sparsity. Category-wise recall comparisons with Perspective API and Detoxify show clear superiority for the specialized model, especially for Violence & Scariness and Substance Consumption. These evaluations convincingly support the methodological decision to build specialized classifiers.

1.4 PG-STORY Corpus and Model-in-the-loop Annotation

The PG-STORY corpus is methodologically strong in several aspects. It aggregates four narrative sources with distinct properties: short crowd-sourced commonsense stories (ROCStories), crowd-sourced plot summaries (WikiPlots), and expert-written fairy tales (FAIRYTALEQA narratives and Grimm's tales). Long narratives are segmented into shorter excerpts of about five sentences, which balances annotator workload with the need for discourse context.

The annotation protocol combines model-in-the-loop selection with human judgments. A preliminary detection model assigns sentence-level safety scores, which are aggregated into discourse-level scores. Stratified sampling across the discourse-score spectrum then selects stories for human annotation, increasing the yield of interesting unsafe cases. This is more efficient than uniform random sampling from a large safe-dominated pool.

Each sampled story is annotated by three qualified Amazon Mechanical Turk workers, who are native speakers from selected English-speaking countries. The same worker labels both discourse-level and sentence-level safety for a given story, and is instructed to perform discourse-level labeling first. This reduces inconsistencies between local and global judgments and explicitly accounts for contextual effects such as scary atmosphere or stereotype-laden descriptions.

The model-in-the-loop process is repeated: initial annotations refine the detector, which then guides the selection of additional stories. This iterative design improves the detector and the labeled corpus over time.

1.5 Safe Story Generation Framework

On the modeling side, the paper proposes a coherent framework for safe story generation with three components: plan-to-story generation, safety self-diagnosis, and content rewriting.

The plan-to-story component uses RAKE to extract keywords from reference stories and then conditions BART on a flattened sequence of title, an end-of-title token, keyword list, and an end-of-plan token. This follows well-established plan-and-write techniques and ensures that story generation is guided by structured content.

The conditional generation variant uses special tokens [SAFE] and [UNSAFE] and category tokens [1] through [5] at the start of each sentence during training. This provides an explicit control interface for safety and category-conditioned generation.

The self-diagnosis variant is a methodological innovation. Safety tokens are appended after each sentence's end-of-sentence token and after an [END_STORY] token for discourse-level assessment. The model thus learns to generate a sentence and then label it as safe or unsafe, and to label the entire narrative. This turns safety detection into a generation task, integrating assessment into the same decoder that produces the story.

Finally, the content re-write module uses two controllable generation techniques, Plug-and-Play Language Models and weighted decoding, driven by the specialized safety classifier as an attribute model. Unsafe sentences detected by self-diagnosis are rewritten to reduce predicted unsafety. This closes a loop from detection to generation in a technically well-justified way.

1.6 Evaluation of Generation and Rewriting

The evaluation of story generation and rewriting is multidimensional. Automatic metrics include:

- Perplexity and BERT-F1 for fluency and semantic similarity to reference stories.
- Distinct-n metrics for lexical diversity at unigram, bigram, and trigram levels.
- Keywords Matching Ratio to assess how well generated stories respect planning keywords.
- Toxicity scores from Perspective API as an external proxy for unsafe content.

The paper reports these metrics for self-diagnosis, conditional generation, and both rewriting methods. Self-diagnosis and conditional generation show similar perplexity and BERT-F1, with self-diagnosis having slightly higher diversity. Rewriting significantly increases diversity and perplexity while lowering toxicity and keyword matching, which is an expected pattern under safety-driven modification.

Human evaluation uses 30 unseen stories and four configurations: self-diagnosis, conditional generation, self-diagnosis with PPLM, and self-diagnosis with weighted decoding. Human annotators judge story and sentence safety using the same criteria as for PG-STORY. Results show that self-diagnosis achieves higher discourse-level safety prediction accuracy (63.3 percent) than conditional generation (40.0 percent), and that PPLM roughly doubles rewriting success rates compared to weighted decoding. This combination of automatic and human evaluation strengthens the empirical credibility of the framework.

2 Key Limitations

2.1 Annotation Reliability and Cultural Scope

Inter-annotator agreement on PG-STORY is modest. Cohen's kappa and Fleiss' kappa are around 0.26 to 0.27 for both sentence-level and discourse-level safety. This indicates substantial disagreement about what is unsafe for children under ten. The paper acknowledges annotation difficulty but does not analyze disagreement patterns in depth. Without such analysis, the reliability of fine-grained category labels is uncertain.

In addition, annotators are restricted to a small set of English-speaking countries. Cultural norms regarding violence, sexuality, and acceptable humor differ across societies. The current corpus therefore embeds a specific cultural perspective on safety, which limits its direct applicability to other contexts. This limitation is important for a task explicitly concerned with children's well-being.

2.2 Lexicon-driven Supervision and Category Imbalance

The heavy reliance on lexicon matching introduces systematic weaknesses. Substance Consumption is entirely absent without lexicon matching and only appears when lexicon hits are incorporated. This means that training data for this category is almost entirely determined by the lexicons. Any omissions or overgeneralizations in the lexicons directly affect the classifier.

Lexicons operate at the word level and do not capture context, irony, or educational framing. Words such as wine or beer can appear in neutral or even didactic contexts, yet are treated uniformly as unsafe. The paper does not quantify the extent of false positives introduced by lexicon matching, nor does it explore methods to mitigate this noise.

Category imbalance remains pronounced. Discrimination & Bias and Sex & Nudity have substantially more samples than Substance Consumption. This manifests in the lower F1 score for Substance Consumption, which is about 49.9 points. The impact of imbalance on downstream rewriting and category-wise control is not fully explored.

2.3 Semi-supervised PG-STORY Labels and Noise

Out of the total PG-STORY corpus, only 1,000 stories receive full human annotation. The remaining 100,000 instances are semi-supervised, labeled by the detection model. Although this is a pragmatic choice, the paper does not provide a quantitative analysis of label noise in the semi-supervised portion.

There is no reported study that compares semi-supervised labels with fresh human labels on a held-out subset. Consequently, the degree to which training signals for self-diagnosis and rewriting reflect human judgments rather than model biases is unclear. This is particularly relevant for nuanced categories and borderline stories.

2.4 Limited Safety Dimensions

The taxonomy focuses on content categories that are clearly problematic, but does not account for other dimensions that are important in child development.

Moral framing is not captured. Narratives where violence is clearly condemned and followed by restitution may have different developmental impacts than narratives that glamorize aggression, yet both are labeled under Violence & Scariness. Emotional intensity and potential for distress are not explicitly annotated. A mildly scary scene and a deeply disturbing horror scene receive the same category label.

Subtle and systemic bias that does not use explicit slurs can still shape children's attitudes, but may not be detected by lexicon-based models. The current taxonomy and classifiers are not designed to address such implicit factors.

2.5 Evaluation Coverage and Error Analysis

The evaluation has several blind spots.

The human evaluation of safety prediction and rewriting uses only 30 stories, which limits statistical power. The reported percentages such as 63.3 percent versus 40.0 percent accuracy or 54.5 percent versus 27.2 percent rewriting success lack confidence intervals or significance tests. This makes it difficult to assess the robustness of these differences.

There is little systematic error analysis of self-diagnosis. The paper does not show what types of unsafe content are commonly missed, nor whether misclassifications cluster in certain narrative patterns, such as implied harm or emotional neglect. Similarly, rewriting errors are not categorized beyond a few qualitative examples.

Narrative coherence after rewriting is not thoroughly evaluated. The automatic metrics focus on perplexity, diversity, and keyword coverage, but do not directly capture character consistency or logical flow. Given that PPLM often causes semantic drift, this omission is important.

Computational cost is reported only qualitatively. PPLM is known to be resource-intensive because it performs gradient-based updates at each decoding step. The paper does not present runtime or memory usage for different story lengths or rewriting workloads. This obscures the feasibility of real-time deployment.

2.6 Modeling Choices and Baseline Breadth

The study standardizes on BART as the core architecture for detection, categorization, and generation. While BART is a strong baseline, the absence of comparisons with other architectures such as T5, RoBERTa, or Longformer means that the relative contribution of model choice versus taxonomy and data is unclear.

Baseline coverage for detectors is limited to Perspective API, Detoxify, and ChatGPT 3.5. These are reasonable choices, but they do not represent the full spectrum of current large language models or commercial safety filters. For example, there is no comparison with other instruction-tuned models or specialized safety pipelines.

3 Technical Bottlenecks

3.1 Entanglement of Generation and Self-diagnosis

In the self-diagnosis setting, the decoder performs narrative generation and safety labeling in a single sequence. This creates an entanglement between language modeling and classification tasks.

The decoder must maintain representations that support both creative continuation and accurate risk assessment. When the same hidden states drive token prediction and safety token generation, gradients from the safety objective can influence generation behavior in subtle ways. This may introduce conservative biases that reduce diversity or cause the model to avoid certain topics entirely.

In addition, safety tokens are generated autoregressively after the sentence text. Any misstep in the narrative can influence the self-diagnosis, and vice versa. Miscalibration may arise if the model learns superficial correlations between certain phrases and safety labels instead of deeper situational understanding.

3.2 Attribute Model Dependence in Rewriting

The rewriting mechanisms rely crucially on the attribute model, which is the specialized safety classifier. Several technical bottlenecks follow.

First, the attribute model was initially trained on UNSAFECORPUS, which is dominated by conversational and comment-style text. Although fine-tuning on PG-STORY introduces narrative data, the classifier may still be biased toward patterns of social media and talk pages. This domain mismatch can reduce sensitivity to narrative cues absent from short comments.

Second, the attribute model reduces the rich taxonomy into a scalar probability of unsafety during rewriting. Both PPLM and weighted decoding depend on this scalar, which limits the ability to enforce nuanced, category-specific constraints. For example, lowering violence while preserving suspense is not straightforward when the control signal is a single number.

Third, PPLM modifies hidden states locally with gradients from the attribute model. This can increase safety scores, but it can also cause semantic drift away from the planned content. The high perplexity values and reduced keyword coverage in rewritten stories reflect this drift. Weighted decoding faces similar challenges, although with less flexibility.

3.3 Sentence-level Control and Discourse-level Safety

The rewriting process operates at the sentence level, triggered whenever a sentence is labeled unsafe. This local focus imposes a bottleneck on discourse-level safety.

Certain patterns of harm only emerge across multiple sentences, such as escalating threats or repeated microaggressions. A sentence-by-sentence rewriting strategy may fail to address these patterns adequately. Conversely, individual sentences that are safe in isolation but contribute to a problematic arc remain unchanged.

The discourse-level label emitted at the end of the story does not influence the rewriting process directly. There is no mechanism to search for alternative story trajectories that satisfy global safety constraints while preserving plot coherence.

3.4 Information Loss in the Taxonomy

The five-category taxonomy, while clear and practical, compresses a complex space of child safety concerns. Within Violence & Scariness, narratives range from slapstick accidents to graphic warfare. Within Discrimination & Bias, content ranges from overt slurs to subtle stereotypes. Collapsing these into single labels limits the granularity of both detection and control.

The absence of explicit severity or valence annotations reduces the interpretability of classification scores and the precision of rewriting heuristics. As a result, rewriting strategies treat mild and severe instances of the same category similarly, which can either overcorrect benign content or undercorrect harmful content.

3.5 Trade-offs between Safety and Narrative Quality

The reported metrics reveal a structural trade-off between safety and narrative quality.

Rewriting with PPLM and weighted decoding reduces toxicity scores from around 0.168 to 0.123 and 0.143, respectively, which indicates improved safety according to Perspective API. At the same time, perplexity jumps from roughly 1.59 to above 7.37 and 8.46. Keywords Matching Ratio drops from about 0.71 to under 0.49. These shifts show that rewriting significantly distorts the planned content and makes the text less aligned with reference narratives.

Higher Distinct-n scores suggest greater lexical variety, but without explicit coherence evaluation, it is unclear how much of this diversity corresponds to meaningful variation versus instability in token sequences. The framework thus reveals an inherent tension: more aggressive safety control yields safer but less faithful and potentially less coherent stories.

3.6 Scalability and Latency

PPLM in particular introduces a scalability bottleneck. The method performs gradient-based updates to the hidden state at each decoding step for each rewritten sentence. For long stories or interactive applications, this can create substantial latency and resource demands. The paper does not quantify these costs, but they are implicit in the design.

Weighted decoding is computationally lighter but also less effective in human-judged rewriting success. Balancing safety performance with computational efficiency remains an open technical challenge.

4 Research Implications

4.1 Safety Evaluation beyond Generic Toxicity

The comparative evaluation of Perspective API, Detoxify, ChatGPT 3.5, and the specialized safety model underscores a key implication: generic toxicity detectors are not sufficient for child safety in narrative text. These tools exhibit low recall on unsafe content, particularly for violence, scariness, and substance consumption. Even ChatGPT 3.5, which outperforms the external tools, achieves lower performance than the specialized model on several categories and has difficulty with certain unsafe themes.

This suggests that future safety evaluations for large language models should be grounded in domain-specific taxonomies and datasets, not solely in general-purpose toxic comment benchmarks.

4.2 Narrative Safety as a Distinct Research Area

The study highlights narrative safety as a distinct area within safety research. Unlike short comments, stories involve plot, character development, and emotional arcs. Unsafe elements can arise from sequences of events, framing, and implications rather than explicit taboo words.

The need for both sentence-level and discourse-level annotations demonstrates that safety judgments must consider context and temporal structure. This connects narrative safety to broader research on long-context modeling and discourse coherence and suggests that evaluation protocols for dialogue and instruction following should also incorporate narrative aspects.

4.3 Integration of Generation, Detection, and Control

The self-diagnosis and rewriting framework illustrates the integration of generation, detection, and control within a single system. Models that can both generate and classify their own outputs open new possibilities for safety-aware text generation.

This pattern is relevant beyond child safety. Similar architectures could be designed for misinformation, privacy-sensitive content, or domain-specific harms. The notion of an attribute model steering generation during decoding provides a general blueprint for post-hoc alignment that complements parameter-tuning approaches such as reinforcement learning from human feedback.

4.4 Benchmarking versus Real-world Deployment

The experiments expose a gap between benchmark performance and real-world deployment. On UNSAFECORPUS and PG-STORY, the specialized model achieves high F1 scores and category-wise recall. However, jailbreak-style prompts reveal that 52 percent of ChatGPT-generated stories for children are still flagged as unsafe, even when prompts explicitly mention children and use adversarial adjectives.

This discrepancy indicates that high benchmark scores do not guarantee safety under open-ended or adversarial prompting conditions. Real-world deployment requires robustness to diverse prompts, evolving usage patterns, and adversarial behavior, which benchmarks like PG-STORY only partially capture.

4.5 Parallel Challenges in Other Domains

The PG-STORY framework has implications for other content domains. The combination of taxonomies, specialized detectors, and controlled generation can be adapted to lyrics, scripts, educational content, and social media assistance tools. Each domain will require its own safety dimensions and narrative conventions, but the methodological pattern is reusable.

At a higher level, the work points to a general need for domain-specific safety benchmarks and integrated generation-control pipelines whenever language models are deployed in sensitive settings.

5 Potential Research Directions

5.1 Richer Annotation Schemes

Future work can extend the annotation scheme in several ways.

First, adding severity levels within each category, such as mild, moderate, and severe violence, would allow models to implement graduated responses. Second, annotating moral valence, such as whether harmful actions are punished or rewarded, would help distinguish cautionary tales from glamorizing content. Third, collecting separate safety judgments for different age bands would support age-targeted generation.

Cross-cultural annotation campaigns, where the same stories are assessed by annotators from different regions, would reveal systematic differences and support culturally aware safety settings.

5.2 Improved Semi-supervised Learning and Uncertainty Handling

The semi-supervised portion of PG-STORY can be strengthened by incorporating uncertainty modeling. Safety classifiers could estimate calibrated confidence scores and only assign automatic labels when confidence exceeds a threshold. Low-confidence samples could be routed to human annotators using active learning strategies.

Consistency training and contrastive learning could be used to stabilize representations for borderline cases. Periodic re-annotation of random subsets of semi-supervised data would allow estimation and correction of label noise rates.

5.3 Plan-level and Structural Safety Control

Rather than focusing only on sentence-level rewriting, future systems can integrate safety controls at the planning stage.

Safety classifiers could evaluate story titles and keyword plans before decoding and filter or modify unsafe elements. Structured representations of plots as event graphs or state transitions would allow constraints on event types, outcomes, and causal chains. Constrained decoding over such structures could guarantee that only safe story arcs are realized.

Such approaches would reduce the need for disruptive local rewriting and better preserve narrative coherence.

5.4 Alternative Architectures and Training Objectives

Alternative architectures may address the entanglement issues in self-diagnosis.

One approach is to separate generator and classifier components, with a shared encoder and distinct decoder or classification heads. The classifier could operate on encoder outputs, leaving the decoder focused on language modeling. Long-context encoders such as Longformer could improve discourse-level assessment.

Joint training objectives that reward both story quality and safety could be explored, including reinforcement learning where safety scores act as part of the reward signal. This would allow the generator to internalize safety constraints rather than relying solely on external steering during decoding.

5.5 Multimodal and Interactive Extensions

Many children's stories are multimodal. Future research can extend PG-STORY-style frameworks to text-plus-image or text-plus-audio settings, where illustrations or narration interact with the story.

Interactive storytelling systems where children make choices present additional challenges. Safety mechanisms must monitor the evolving story in real time and adjust future content conditionally. Parental control interfaces that expose high-level safety preferences and explain why certain content is blocked or rewritten could improve transparency and trust.

Explainable safety diagnostics that highlight phrases or narrative elements responsible for unsafe labels would support human oversight and iterative refinement.

6 Conclusion

The PG-STORY paper makes an important contribution by reframing content safety for children as a narrative-specific problem and by proposing a full pipeline that spans taxonomy design, dataset construction, specialized safety modeling, and safe story generation.

Methodologically, the work is strong. The taxonomy is child-centric and multi-dimensional, UNSAFECORPUS and PG-STORY provide complementary training and evaluation resources, and the specialized safety models significantly outperform generic toxicity tools and a widely used conversational model in both detection and category-level recall. The safe story generation framework, combining plan-to-story modeling, self-diagnosis, and controllable rewriting, demonstrates that child safety can be integrated into the generation process in a principled way.

At the same time, the study exhibits important limitations. Annotation agreement is modest and culturally narrow, lexicon-driven supervision introduces category-specific noise, semi-supervised labels are not fully audited, and the rewriting mechanisms reveal a trade-off between safety and narrative fidelity. Evaluation of coherence, severity, and real-world robustness remains limited.

Overall, PG-STORY establishes a strong foundation for research on child-safe narrative generation and provides a valuable benchmark and methodology. The most promising research directions include richer and more nuanced annotation schemes, improved semi-supervised learning with uncertainty modeling, structural and plan-level safety control, alternative architectures for self-diagnosis and control, and multimodal, interactive extensions. Advancing along these directions can move the field closer to story generation systems that are creative, coherent, and reliably aligned with the needs and safety of young readers.