GuardNER: Evaluating Named Entity Recognition in Harmful and Toxic Domains

Amer K. Mriziq
mriziq@ischool.berkeley.edu
University of California, Berkeley
School of Information
Berkeley, CA, USA

Miguel Arias
miguel.arias@ischool.berkeley.edu
University of California, Berkeley
School of Information
Berkeley, CA, USA

Ankur Garg
ankur_garg@ischool.berkeley.edu
University of California, Berkeley
School of Information
Berkeley, CA, USA

Abstract

As companies have rapidly integrated large language models into products and services, users have become increasingly adept at exploiting these systems' vulnerabilities. AI safety research has introduced numerous datasets as benchmarks for harm and toxicity. We examine lightweight NER models and their out-of-the-box ability to detect harmful utterances, introducing the first benchmark specifically targeting harm/toxic domains for NER models. GuardNER is our NER dataset composed of ten different entity types, with 97 entries in total. This paper evaluates zero-shot NER language models on the GuardNER benchmark, comparing their harm detection performance to that of few-shot general large language models. We found that span-prediction NER models outperform large language models on GuardNER, achieving higher precision (62.05%), recall (16.58%), and F1 scores across most categories.

1 Introduction

Content Warning: When necessary for clarity, this paper directly quotes user content that contains offensive/hateful speech, profanity, and other potentially triggering content.

Large language models (LLMs) are now widely integrated into various applications, yet their potential for misuse, in which individuals or groups intentionally use AIs to cause harm, remains (Hendrycks et al., 2023). LLMs may generate harmful output that presents significant risks (Hendrycks et al., 2023). Adversarial attacks manipulate input prompts to induce machine learning models, particularly large language models, to generate unintended or harmful outputs. Due to alignment design limitations, LLMs remain vulnerable to these attacks, which can exploit optimized adversarial suffixes using greedy and gradient-based techniques to bypass safety filters and produce objectionable content (Zou et al., 2023). The fine-tuning of AI safety alone is not scalable, as increasing the configuration and accessibility of the end user makes these systems more vulnerable (Qi et al., 2023).

1.1 Importance of AI Safety

Change is inherent in complex systems and not all scenarios can be anticipated during training. Models must adapt to an evolving world and learn from new experiences (Taleb, 2012). Hendrycks, Schulman, and Steinhardt (2021) categorize challenges in AI safety into four primary areas: robustness, monitoring, alignment, and systemic safety. These pillars address different aspects of ensuring that AI systems operate reliably and ethically within complex environments. Alignment (Hendrycks et al., 2023), which focuses on guiding AI systems to adhere to human values, can be advanced through methods such as reinforcement learning from human feedback (RLHF). Similarly, monitoring (Hendrycks et al., 2023) includes anomaly detection as a key strategy for identifying unexpected or malicious behavior in AI systems, essential for preempting harmful actions.

Rule-based filtering offers a straightforward solution by restricting outputs based on a preset list of phrases, though it lacks flexibility in handling unexpected or nuanced harmful content. The primary challenge in anomaly detection is that current representation learning methods struggle to identify representations effective for previously unseen anomalies (Hendrycks et al., 2021). Llama Guard, developed by Meta (Inan et al., 2023), is a safeguard for conversational AI applications to identify and handle safety concerns related to both AI-generated responses and queries. Llama Guard

implements a multiclass classification system based on a safety risk taxonomy to classify prompts and responses. However, Llama Guard only performs well in settings where strong predefined safety measures are crucial, such as customer support or educational tools. Achintalwar et al. (2024) (Achintalwar et al., 2024) introduced a framework for using "detectors" as tools to identify harmful outputs in large language models. These detectors function as compact, specialized classifiers capable of identifying specific types of harm, including prejudice, unfaithful content, and bias. They are used in various stages of LLM deployment to serve as guardrails, especially in critical applications where direct model control may be challenging. Detectors are designed to monitor and label outputs, enabling safer AI use by preemptively catching undesirable content and ensuring that models adhere to safety standards (Achintalwar et al., 2024).

1.2 Use of LLMs for harm/toxicity detection

Google Jigsaw's Perspective API (Google Jigsaw, 2018) has become a widely adopted tool for identifying toxic content online. Recent studies have explored alternative approaches to improve toxicity detection performance. He et al. (He et al., 2024) demonstrate that prompting local large language models, such as T5, GPT2-M, and GPT2-L, enhances performance across various toxicity-related tasks. Hanley and Durumeric (Hanley and Durumeric, 2023) show that a DeBERTa classifier trained with contrastive learning slightly outperforms the Perspective API. Similarly, Li et al. (Li et al., 2023) highlight the potential of GPT-3.5 in effectively detecting hate speech in English. However, model size continues to grow. Kumar et al. 2024 (Kumar et al., 2024) shows how LLMs significantly outperform state-of-the-art toxicity classifiers, yet continuous improvements through larger parameters may lead to issues of scalability and deployment (Kumar et al., 2024). In this paper, we focus on small encoder-only models for this purpose. We aim to evaluate <1b parameter language models on their propensity to detect toxicity out-of-the-box.

1.3 Looking to smaller models

GLiNER, a BERT-like architecture, employs a bidirectional encoder architecture optimized through improved training methods, larger datasets, and the removal of the Next Sentence Prediction (NSP) objective, focusing solely on masked language modeling. This design makes GLiNER-based models particularly effective for understanding tasks such as sentiment analysis, text classification, and named entity recognition, where deep contextual comprehension is crucial (Bhukya et al., 2023; Arif et al., 2024). In contrast, GPT models adopt an auto-regressive architecture that generates text by predicting the next word in a sequence based on the prior context, making them well suited for tasks like text generation, summarization, and dialogue systems (Zibin Zheng, 2023; Dhuliawala et al., 2023). In this paper, we compare NuNERZero and NuNERZero-Span (Bogdanov et al., 2024), with popular multibillion parameter models. The NuMind class of models are general, task-specific models that train on more than 200k entity types (Bogdanov et al., 2024).

1.4 Datasets for harm/toxicity

As innovations in Named Entity Recognition and large language models progress, so does the potential for malicious applications across LLM-driven products. To address these risks, companies employ a variety of benchmarks to evaluate their LLMs' susceptibility to misuse. RealToxicityPrompts (Gehman et al., 2020; OpenAI, 2022; Chung et al., 2022) has become an industry standard for assessing toxicity, while HateXplain (Mathew et al., 2021) is widely recognized for hate speech detection. Currently, there are no NER datasets specifically for harm/toxicity detection. In this study, we create a small BIO-annotated dataset of harmful outputs generated by LLMs.

2 Methods

Content Warning: When necessary for clarity, this paper directly quotes user content that contains offensive/hateful speech, profanity, and other potentially triggering content.

To assess the viability of NER in harm detection, we tested each model's performance on NER task on harmful/toxic domains using a newly created dataset, GuardNER.

2.1 Collecting harmful responses

The primary goal of this research was to develop a new evaluation data set, GuardNER, that includes a taxonomy of BIO-annotated harm entity labels and an associated codebook to annotate harmful outputs generated by large language models. To collect the harmful LLM output data, we referenced two recent evaluation benchmarks: SafetyAligner (Huang et al., 2024) and Do-Not-Answer (Wang et al., 2024). These benchmarks focus on risks associated with inappropriate input flagging and the prevention of harmful outputs.

The Do-Not-Answer dataset comprises 938 harmful prompts (inputs) annotated using a three-level hierarchical taxonomy. This taxonomy consists of five major risk areas at the highest level, 12 harm types at the second level, and 61 specific harms at the third level (Wang et al., 2024). The dataset also includes responses (outputs) from six LLMs: GPT-4, ChatGPT, Claude, ChatGLM2, LLaMA2-7b-chat, and Vicuna-7b. Do-Not-Answer captures instances of harmful responses for each model, specifically when the models failed to detect and filter harmful content effectively (Wang et al., 2024).

In contrast, the SafeAligner dataset contains 628 harmful queries (inputs) categorized under a taxonomy of eight labels: Illegal Activity, Hate/Harass/Violence, Malware, Fraud/Deception, Privacy Violation Activity, Physical Harm, Economic Harm, and Adult Content (Huang et al., 2024). For each harmful query, the dataset provides both a "safe" response and a corresponding "harmful" response, facilitating analysis of model behavior in harmful content scenarios.

Our research utilized harmful responses from both the Do-Not-Answer and SafeAligner datasets to perform an iterative clustering exercise aimed at generating a new set of entity labels. These labels were designed to offer greater granularity, enhancing the explainability of harm-detection processes and potentially improving the performance of NER models.

From the Do-Not-Answer dataset, we sampled 104 harmful outputs, ensuring representation across all five major risk areas and a diverse mix of model responses (Wang et al., 2024). Each harmful output was analyzed to extract tokens identified as “harmful,” guided by our research team’s domain-specific intuition (Wang et al., 2024). From the SafeAligner dataset, we sampled 102 harmful outputs, ensuring representation across all eight harm types for comprehensive coverage (Huang et al., 2024). Consistent with our methodology for the Do-Not-Answer dataset, we analyzed each harmful output individually, identifying and extracting tokens considered "harmful" based on our research team’s informed judgment. A sample of the annotated table used for this analysis is provided below.

2.2 Formulating a taxonomy

Building on harm-entity extractions from both Do-Not-Answer and SafeAligner, we conducted a bottom-up clustering exercise to group similar entities and assign appropriate labels to each cluster. Through an iterative process of clustering and refinement, we developed a final taxonomy that effectively organized harmful entities into meaningful categories. This taxonomy served as the foundation for creating an accompanying codebook. The final iteration incorporates peer feedback and insights gained from the process. Key refinements include the removal of the second-tier hierarchy and the decision to avoid the label “illegal” due to the complexities of accounting for varying legal codes across jurisdictions. Labels that overlapped or conflated with one another were consolidated to improve clarity and consistency. Also, the Harm Entity Examples were expanded to include longer spans beyond single-word tokens, emphasizing the importance of context in labeling. This version served as the basis for Codebook v1.

2.3 Developing Annotation Codebook

Codebook v1: Using the finalized taxonomy and drawing on Universal NER’s Annotation Guidelines for best practices, we developed an initial codebook to guide the annotation process (Mayhew, Stephen, 2022). The design prioritized a balance between simplification and granularity—reducing the number of entity labels for usability while avoiding overgeneralization that could hinder NER models’ ability to accurately tag harmful entities.

Codebook v2: Following the first round of annotations, the codebook was refined to address observed challenges and improve clarity. A snapshot of this version is provided below, with the full codebook available in the Appendix.

Table 1: Codebook v1

Concept	Term Description
Fraud	Dishonest, deceptive, and/or deviant behavior
Physical Harm	Actions that inflict corporal harm onto self or others, or terms describing bodily wounds resulting from malicious actions
Privacy Violation	Actions that violate people's privacy, or terms describing PII/private data in the context of securing them w/o permission
Harmful Language	Malicious terms, or terms that refer to malicious terms/actions in a harmful context (including mental/emotional harms)
Cybersecurity	Terms used to describe malicious cybersecurity behavior or tools
Weapons	Traditional weaponry & any specific object described as a "weapon" or being weaponized for physical harm
Sexually Explicit	NSFW sexual terms, or sexually-related terms/names used in NSFW context
Dangerous Materials	Materials or ingredients that are being repurposed or applied in the context of harmful use or creation of dangerous products
Drugs	Drug names, drug products, and drug categories used in the context of conducting harmful or deviant behavior

Table 2: Codebook v2

Concept	Term Description
Fraudulent	Dishonest, deceptive, and/or deviant behavior
Physical Harm	Actions that inflict corporal harm onto self or others, or terms describing bodily wounds resulting from malicious actions
Privacy Violation	Actions that violate people's privacy, actions that deal with unauthorized access, or terms describing PII/private data in the context of securing them w/o permission
Harmful Language	Malicious terms, or terms that refer to malicious terms/actions in a harmful context (including mental/emotional harms)
Cybersecurity	Terms used to describe malicious cybersecurity behavior or specific tools
Weapons	Traditional weaponry & any specific object described as a "weapon" or being weaponized for physical harm
Sexually Explicit	NSFW sexual terms, or sexually-related terms/names used in NSFW context
Dangerous Materials	Materials or ingredients that are being repurposed or applied in the context of harmful use or creation of dangerous products
Drugs	Drug names, drug products, and drug categories used in the context of conducting harmful or deviant behavior
Destruction	Terms describing the act of destruction/damage/vandalism to property/objects/inorganic matter

2.4 Annotation Workflow

With the initial codebook established, we began annotating the dataset. To ensure high-quality NER annotations, we opted to start with a manageable sample size. Instead of working with all 628 rows from the SafeAligner (Huang et al., 2024) dataset, we randomly sampled 13 rows from each of the eight SafeAligner harm categories, resulting in a dataset of 104 rows. This approach provided a diverse and representative sample spanning multiple categories of harm, enabling a more focused and in-depth analysis.

2.4.1 Rounds of Annotations

Before commencing formal annotations, the three researchers involved in the project collaboratively annotated 10 randomly selected rows. This trial phase identified gaps and inconsistencies in the codebook, allowing us to refine the annotation rules, incorporate additional guidelines, and align our annotation strategies. During this phase, we also familiarized ourselves with the annotation software, Labelbox, which was used throughout the project. Labelbox offered valuable insights, such as annotation consensus, time spent per annotation, and the frequency of entity labels applied. The first formal round of annotations followed. Each of the three researchers independently annotated all 104 rows of data.

To address discrepancies, the team convened to review all annotations with a consensus score below 30%. This review led to further refinements of the codebook, alignment on how to address ambiguous cases, and the addition of new labels. Notable updates included introducing the label *Destruction* to capture harms involving non-biological entities and modifying the label *Fraud* to *Fraudulent* for improved descriptiveness and utility in model training.

The second round of annotations was conducted by two of the researchers. Using the updated codebook, we each went through the 104 rows independently. After this iteration, we no longer edited the codebook. Instead, after annotation, we sat down together, investigated discrepancies against the codebook, and determined whose annotations were more accurate for each row. These were the annotations that created the resulting GuardNER Dataset.

2.4.2 Inter-Annotator Agreement(IAA)

Table 3: Annotator Agreement for First and Final Passes (Grouped Metrics)

Metric	First Pass	Final Pass
Cohen's Kappa (1,2) - All Tokens	0.5723	0.8485
Cohen's Kappa (1,3) - All Tokens	0.6052	-
Cohen's Kappa (2,3) - All Tokens	0.7669	-
Average Cohen's Kappa - All Tokens	**0.6482**	**0.8485**
Cohen's Kappa (1,2) - Filtered	0.3186	0.6854
Cohen's Kappa (1,3) - Filtered	0.3688	-
Cohen's Kappa (2,3) - Filtered	0.5818	-
Average Cohen's Kappa - Filtered	**0.4231**	**0.6854**
Inter-annotator Agreement - All Tokens	0.9389	0.9811
Inter-annotator Agreement - Filtered	0.3219	0.7399
Fleiss' Kappa - All Tokens	0.6532	-
Fleiss' Kappa - Filtered Tokens	0.4192	-
Pairwise F1 (1,2) - Filtered	0.6539	0.8803
Pairwise F1 (1,3) - Filtered	0.6830	-
Pairwise F1 (2,3) - Filtered	0.8180	-
Average Pairwise F1 - Filtered	**0.7183**	**0.8803**

To understand the level of agreement between the two rounds of annotations, we decided to pull in a number of metrics to get a better sense of how our annotations performed. The issue with NER annotations, as opposed to other classification problems, is that not every token gets annotated, and therefore there is the unknown of how many total entities to calculate out of. The suggested approach for NER is to use the pairwise F1 score without consideration for tokens where no annotator has labeled (Brandsen et al., 2020). However, there is still much to learn from other, more traditional metrics such as Cohen's Kappa, Fleiss' Kappa, and the simple percent agreement. Cohen's Kappa is useful for the agreement of two annotators. Meanwhile, due to our variable annotator count across the two passes, the Fleiss' Kappa is the standard for three or more annotators (McHugh, Mary L., 2012). To ensure we hit all our bases, we calculated all 3 and compared (Table 3).

The table demonstrates just how flawed some metrics are compared to others. In the table, "all tokens" refers to instances where we kept every single token in the dataset, labeling unclassified tokens with "O." Filtered refers to the condensed version of the data, where only tokens that were labeled at least once were kept for evaluation. Since there are significantly more unlabeled tokens, we can see how the presence of these tokens skews our results.

The most important rows to look at are the *"Average Cohen's Kappa - Filtered"* and the *"Average Pairwise F1 - Filtered"*. In both instances, it can be seen that there is a significant increase in annotator agreement across the two passes, 0.42 to 0.68 and 0.71 to 0.88 respectively. Even taking into account that the average agreement for pass one considers three annotators while the final pass only considers two, the final pass has a significantly higher agreement across every combination of two annotators.

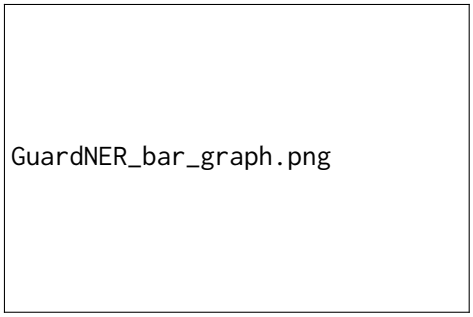
2.5 GuardNER Benchmark

Having completed the annotations, aligning on the best annotations from each, and reflecting all findings in the codebook, we were able to put together the final dataset, which consisted of the original text from SafeAligner, our labeled tokens, and their character location spans.

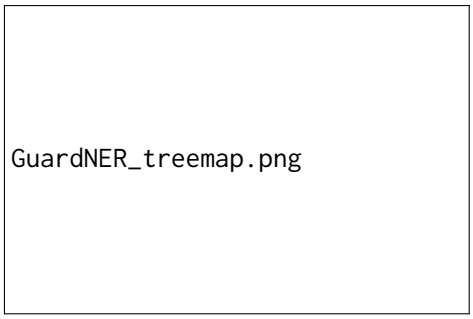
Table 4: GuardNER Dataset: Harm Entity Label Breakdown

Entity Label Type	Count of Labels	Share	Samples/Label
Fraudulent	317	27.95%	54
Physical Harm	74	6.53%	16
Privacy Violation	125	11.02%	24
Harmful Language	88	7.76%	19
Cybersecurity	254	22.40%	32
Weapons	57	5.03%	9
Sexually Explicit	93	8.20%	12
Dangerous Materials	72	6.35%	9
Total	**1080**	**100%**	**175**

Note: There are 97 total samples, and labels may overlap.



(a) Bar Graph Representation of GuardNER Dataset.



(b) Treemap Visualization of GuardNER Dataset.

Figure 1: Visual Representations of the GuardNER Dataset

3 Evaluation Strategy

The evaluation focused exclusively on the GuardNER dataset, which encompasses ten categories of harm and toxicity. We compared NuNERZero and NuNERZeroSpan with Llama 3.1 (8B, 1B) and Mistral 7B models. The NuMind models were evaluated in a zero-shot setting, while the LLMs were pre-prompted with two-shot examples. All evaluations were conducted at the span level, with a threshold of 0.5 applied to filter predicted entities. Results are presented in Table 7 and further dis-

Table 5: Column Descriptions for Dataset

Column Label	Description
Uuid	A unique identifying number for each row.
prompt_SafeAligner	The prompt used as an input query for model response in the SafeAligner dataset (Huang et al., 2024).
harmful_response_SafeAligner	The model's harmful response corresponding to the input query in the SafeAligner dataset (Huang et al., 2024).
harm_labels_and_entities	Harm entity labels assigned to spans within the model's response, represented as [Harm Entity Label: Span].
harm_label_frequency	The total count of each harm entity label associated with a given harmful response, represented as [Harm Entity Label: #].
idx_SafeAligner	The unique ID corresponding to the sample ID in the SafeAligner dataset (Huang et al., 2024).
harm_tags	The character-based start and end positions of a span (including spaces) within a harmful response, paired with its corresponding harm entity label, represented as <[start], [end], [Harm Entity Label]>.

cussed in the Results and Discussion section (Table 7).

3.1 Entity Matching Criteria

Given the inherent variability in entity annotation, especially concerning positional indices in the BIO scheme, we adopt a relaxed matching criterion to account for minor discrepancies. Specifically, a predicted entity is considered a correct match if:

- **Positional Proximity**: The start and end indices of the predicted entity are within ± 2 positions of the ground truth annotations.
- **Entity Type Accuracy**: The predicted entity type matches exactly with the ground truth entity type.

Mathematically, for a predicted entity with start index s_p and end index e_p, and a ground truth entity with start index s_g and end index e_g, the match condition is:

$$|s_p - s_g| \leq 2 \quad \text{and} \quad |e_p - e_g| \leq 2$$

and

$$\text{Entity Type}_p = \text{Entity Type}_g$$

This approach acknowledges minor annotation inconsistencies while maintaining strictness in entity type classification. By allowing a positional tolerance of two indices, we ensure that slight variations in entity boundaries do not disproportionately penalize model performance, thereby providing a more robust evaluation of entity recognition capabilities.

3.2 Experimental Setup

Table 6: Models Used to Conduct Evaluation

Model	Size (Millions)	Architecture Type	Inference Setting	Reference
Llama 3.2 1B	1,000	Transformer w/ GQA	Few-Shot	(Grattafiori et al., 2024)
Llama 3.1 8B	8,000	Transformer w/ GQA	Few-Shot	(Grattafiori et al., 2024)
Mistral 7B	7,000	Transformer w/ GQA & SWA	Few-Shot	(Jiang et al., 2023)
Llama 3.1 405B Instruct Turbo	405,000	Transformer	Few-Shot	(Grattafiori et al., 2024)
NuNER Zero	125	Bidirectional Transformer Encoder	Zero-Shot	(Bogdanov et al., 2024)
NuNER Zero-Span	125	Bidirectional Transformer Encoder	Zero-Shot	(Bogdanov et al., 2024)

3.2.1 Model Configuration and Evaluation Paradigm

Our experimental framework distinguishes between two categories of models based on their evaluation paradigms:

Large Language Models (LLMs)

- **Llama 3.2 1b**: A variant of the Llama series with 1 billion parameters, leveraging few-shot learning capabilities without additional task-specific training.
- **Llama 3.1 8b**: An 8 billion parameter model from the Llama 3.1 series, also utilizing few-shot learning for NER tasks.
- **Mistral 7b**: A 7 billion parameter model, also in a few-shot learning setting.

These three LLMs were evaluated using a **few-shot** learning approach, where each model was provided with a limited number of annotated examples to guide their NER predictions. No further fine-tuning was performed on the GuardNER dataset, allowing the models to leverage their pre-trained knowledge for entity recognition.

Encoder-Only NER-Specific Models

- **NuNerZero**: An encoder-only model specifically designed for NER tasks, operating in a **zero-shot** configuration without any task-specific examples or fine-tuning.
- **NuNERZeroSpan**: A variant of NuNerZero focused on span-level entity extraction, also utilized in a **zero-shot** setting.

These encoder-only models were assessed without providing any annotated examples, relying entirely on their inherent architecture and pre-trained representations to identify and classify named entities within the GuardNER dataset.

3.2.2 Prediction Processing and Annotation Alignment

To ensure consistency across different model outputs and facilitate fair comparison, each model's raw predictions underwent a standardized processing pipeline:

1. **Annotation Formatting**: Predictions from all models were converted into the BIO annotation scheme, aligning with the GuardNER dataset's annotation format. This step ensures uniformity in how entities are represented across different models.
2. **Threshold Application**: A probability threshold of 0.5 was applied to filter out low-confidence predictions. Only entities with a confidence score equal to or exceeding this threshold were retained for subsequent evaluation. This thresholding helps mitigate the impact of uncertain predictions, enhancing the reliability of the evaluation metrics.
3. **Entity Matching**: Processed predictions were then subjected to the entity matching criteria outlined in Section 3.1. This involved checking for positional proximity within ± 2 indices and exact entity type matching against the ground truth annotations.

3.2.3 Evaluation Procedure

All models were evaluated on the same held-out test subset of the GuardNER dataset to maintain consistency in data distribution and characteristics. The evaluation process entailed the following steps:

1. **Prediction Generation**: Each model generated entity predictions for the test set based on its respective configuration (few-shot for LLMs and zero-shot for encoder-only models).
2. **Post-Processing**: As detailed above, predictions were formatted, thresholded, and matched against ground truth annotations to determine correctness.
3. **Metric Computation**: Using the matched entities, precision, recall, and F1-score were calculated to quantify each model's performance on the NER task.

4 Results

Model	Precision (%)	Recall (%)	F1 Score (%)
NuNERZero	41.07%	13.58%	25.33%
NuNERZero-Span	62.05%	16.58%	30.87%
Llama 3.1 8b	28.52%	7.50%	9.74%
Llama 3.2 1b	20.39%	2.73%	3.73%
Mistral 7b	29.25%	2.73%	4.25%

Table 7: *Overall Precision, Recall, and F1 Score for Different Models. n=97*

In evaluating the precision and recall of different models, *NuNERZero-Span* demonstrated superior performance with a precision of 62.05% and a recall of 16.58%, outperforming all other models. *NuNERZero* followed with a precision of 41.07% but showed lower recall at 13.58%. Our evaluation shows that span-prediction NER models, in this case *NuNERZero-Span*, have higher performance on GuardNER than the *Llama* and *Mistral* models. *Mistral 7b* had the highest precision score between the three large language models, with *Llama 3.1 8b* outperforming all LLMs (F1 = $\sim$0.10). In contrast, *Llama 3.1 8b* and *Llama 3.2 1b* achieved lower precision scores of 28.52% and 20.39%, respectively, with recall values of 7.5% and 2.73%. *Mistral 7b* exhibited the weakest performance, with both precision and recall at 2.73%.

Across F1-score evaluations for specific categories, *NuNERZero-Span* consistently achieved the highest scores, particularly in "Sexually Explicit," "Destruction," and "Drugs" categories, while *NuNERZero* showed strong but slightly lower performance. *Llama 3.1 8b* and *Llama 3.2 1b* performed comparatively better in categories such

Figure 2: *Precision and Recall for Different Models*

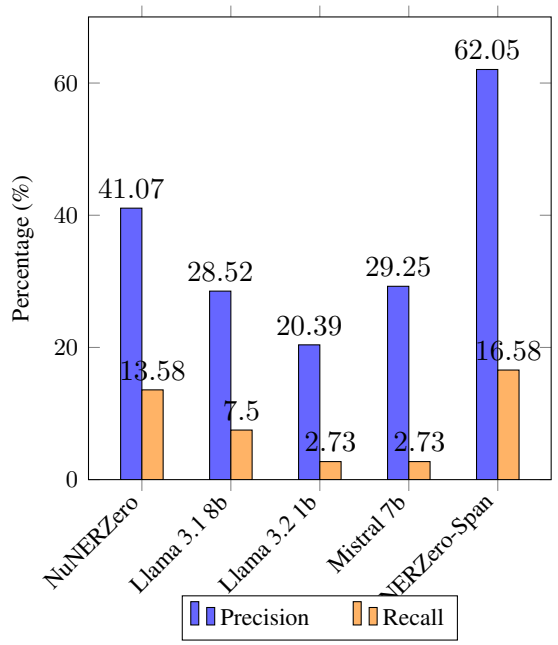


Table 8: F1-Scores of Models on Entity Types (as Percentages)

Entity	NuNERZero	Llama 3.1 8b	Llama 3.2 1b	Mistral 7b	NuNERZero-Span
Privacy Violation	0%	17.8%	4.5%	7.0%	18.5%
Cybersecurity	14.4%	16.7%	5.7%	7.7%	21.3%
Fraudulent	20.9%	8.9%	6.5%	3.0%	23.8%
Harmful Language	25.0%	9.2%	5.3%	4.2%	10.6%
Dangerous Materials	17.5%	0%	2.6%	4.6%	2.7%
Weapons	43.2%	9.2%	0%	9.8%	31.4%
Physical Harm	38.4%	2.1%	0%	0%	47.6%
Drugs	60.0%	0%	0%	0%	66.7%
Destruction	15.4%	17.8%	9.0%	0%	40.0%
Sexually Explicit	18.5%	15.7%	3.8%	6.2%	46.0%

as "Harmful Language" and "Fraudulent" but remained behind *NuNERZero-Span*. *Mistral 7b* underperformed across all categories, indicating its limited ability to capture and recognize harmful entities effectively.

In evaluating the precision and recall of different models, *NuNERZero-Span* demonstrated superior performance with a precision of 62.05% and a recall of 16.58%, outperforming all other models. *NuNERZero* followed with a precision of 41.07% but showed lower recall at 13.58%. In contrast, *Llama 3.1 8b* and *Llama 3.2 1b* achieved lower precision scores of 28.52% and 20.39%, respectively, with recall values of 7.5% and 2.73%. *Mistral 7b* exhibited the weakest performance, with both precision and recall at 2.73%. Across F1-score evaluations for specific categories, *NuNERZero-Span* consistently achieved the highest scores, particularly in "Sexually Explicit," "Destruction," and "Drugs" categories, while *NuNERZero* showed strong but

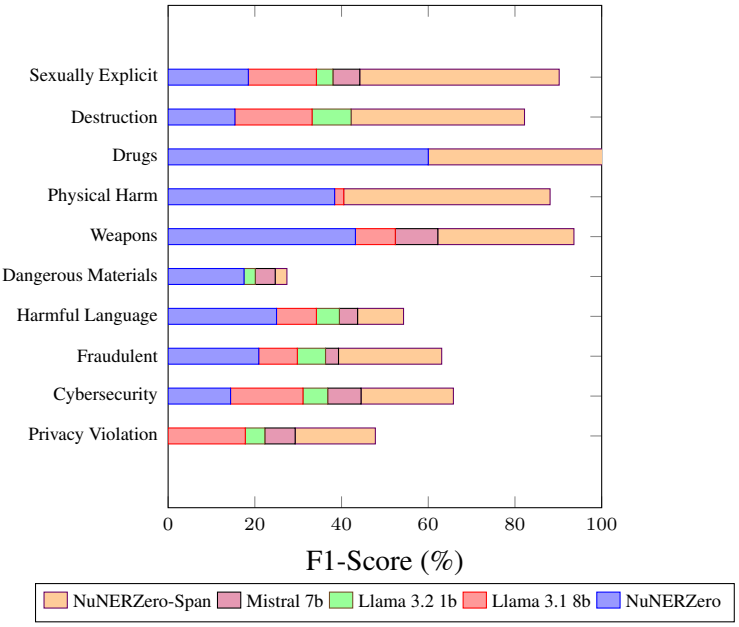


Figure 3: F1-Scores of Models on Entity Types (as Percentages)

slightly lower performance. *Llama 3.1 8b* and *Llama 3.2 1b* performed comparatively better in categories such as "Harmful Language" and "Fraudulent" but remained behind *NuNERZero-Span*. *Mistral 7b* underperformed across all categories, indicating its limited ability to capture and recognize harmful entities effectively.

5 Discussion

The results demonstrate a clear advantage of span-based NER models, specifically NuNERZero-Span, in detecting harmful entities within toxic domains. With a precision of 62.05% and an F1-score of 30.87%, NuNERZero-Span significantly outperformed larger language models such as Llama 3.1 (8B) and Mistral 7B, which achieved considerably lower performance metrics. This disparity highlights the effectiveness of encoder-only architectures in harm-focused NER tasks compared to generative, autoregressive models.

The performance gap can be partially attributed to the behavior of Llama models, which in 15% of the entries, primarily in the "Drugs" and "Weapons" categories, refused to produce outputs. This behavior is likely triggered by the models' built-in safety guardrails and fine-tuning mechanisms, designed to prevent the generation of harmful or sensitive content. While these safety mechanisms are ben-

eficial in real-world applications, they inhibit the models' ability to engage with harm-related entities, resulting in reduced recall and lower overall F1 scores.

Interestingly, the results within the LLM group suggest that larger models, such as Llama 3.1 (8B), yield better performance compared to their smaller counterparts, indicating a performance scaling effect with parameter size. However, as described in Kumar et al. (2024) (Kumar et al., 2024), this reliance on larger models introduces significant challenges for systems where latency, security, and resource constraints are critical. Increased model size exacerbates inference delays and computational overhead, posing limitations for real-time harm detection applications.

In contrast, the NuNER models, despite their relatively small size (125M parameters), achieved superior results without the need for prompt engineering or extensive fine-tuning. NuNERZero-Span's strong performance, particularly in categories like "Drugs" (66.7%) and "Weapons" (31.4%), demonstrates the capability of lightweight, encoder-only models to deliver high precision and recall. These findings emphasize that encoder-only NER models are a plausible alternative to large language models for harm and toxicity detection, achieving competitive results while avoiding the latency and resource trade-offs inherent to larger models.

The category-level results further highlight these advantages. While LLMs displayed strengths in specific domains, such as "Fraudulent" and "Harmful Language," their overall performance was hindered by safety-related refusals and their inability to effectively capture span-level entities. NuNERZero-Span, by comparison, consistently achieved the highest F1 scores across multiple categories, underscoring its robustness and reliability for harm-focused NER tasks.

In conclusion, our findings demonstrate that small, efficient NER models like NuNERZero-Span can outperform larger, resource-intensive LLMs in toxicity and harm detection tasks. Their ability to deliver high performance while addressing latency and security concerns positions them as a scalable, reliable solution for safety-critical applications.

References

Sumeet Achintalwar, Alexandre A. Garcia, Avraham Anaby-Tavor, Ilaria Baldini, Stephan E. Berger, et al. 2024. Detectors for safe and reliable llms: Implementations, uses, and limitations. *arXiv preprint arXiv:2403.06009*.

Samee Arif, Abdul Hameed Azeemi, Agha Ali Raza, and Awais Athar. 2024. Generalists vs. specialists: Evaluating large language models for urdu. *Preprint*, arXiv:2407.04459.

Ramesh Kumar Bhukya, Anjali Chaturvedi, Hardik Bajaj, Udgam Shah, Sumit Singh, and Uma Shanker Tiwary. 2023. Efficiently transferring pre-trained language model roberta base english to hindi using wechsel. *2023 26th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.

Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbe, and Etienne Bernard. 2024. Nuner: Entity recognition encoder pre-training via llm-annotated data. *arXiv preprint arXiv:2402.15343*.

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4573–4577, Marseille, France. European Language Resources Association.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, Danqi Chen, Aakanksha Chowdhery, Ed H. Chi, Zhenzhong Lan, and Jason Wei. 2022. Scaling instruction-finetuned language models. In *NAACL*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *Preprint*, arXiv:2309.11495.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3356–3369.

Google Jigsaw. 2018. Perspective api.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste

Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Pa-

tel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

H. W. Hanley and Z. Durumeric. 2023. Twits, toxic tweets, and tribal tendencies: Trends in politically polarized posts on twitter. *arXiv preprint arXiv:2307.10349*.

X. He, S. Zannettou, Y. Shen, and Y. Zhang. 2024. You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content. In *IEEE Symposium on Security and Privacy*.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*.

Dan Hendrycks, Mantas Mazeika, and Tyler Woodside. 2023. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*.

Caishuang Huang, Wanxu Zhao, Rui Zheng, Huijie Lv, Shihan Dou, Sixian Li, Xiao Wang, Enyu Zhou, Junjie Ye, Yuming Yang, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Safealigner: Safety alignment against jailbreak attacks via response disparity guidance. *arXiv preprint arXiv:2406.18118*.

Huseyin Inan, Keyur Upasani, Jack Chi, Ravi Rungta, Krishna Iyer, Yujia Mao, et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 865–878.

L. Li, L. Fan, S. Atreja, and L. Hemphill. 2023. "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *arXiv preprint arXiv:2304.10619*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pushpak Bhattacharyya, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *arXiv preprint arXiv:2012.10289*.

Mayhew, Stephen. 2022. Universal NER Annotation Guidelines. `https://www.universalner.org/guidelines/`.

McHugh, Mary L. 2012. Interrater reliability: the kappa statistic. *PubMed Central*.

OpenAI. 2022. Aligning language models to follow instructions. `https://openai.com/index/instruction-following`.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Nassim Nicholas Taleb. 2012. *Antifragile: Things That Gain from Disorder*. Random House.

Yingjie Wang, Haoming Li, Xu Han, Preslav Nakov, and Timothy Baldwin. 2024. Do-not-answer: Evaluating safeguards in llms. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 896–911.

Yanlin Wang Jingwen Zhang Dewu Zheng Mingxi Ye Jiachi Chen Zibin Zheng, Kaiwen Ning. 2023. A survey of large language models for code: Evolution, benchmarking, and future trends. *arXiv preprint arXiv:2311.10372*.

Andy Zou, Zhengliang Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Annotation Guidelines and Entity Labels

A.1 Annotation Guidelines

Our guidelines reference Universal NER's Annotation Guidelines (Mayhew, Stephen, 2022) for guidance on how to outline this section.

A.1.1 What Should Be Annotated?

- Any words or terms in the text that correspond to one of the ten Harmful Entity labels in the table below (Table 9).
- Reference the table in "Entity Labels & Descriptions" below for detailed information on each label's use.

A.1.2 Entity Label Tips

Dealing with Edge Cases

- **"And/Or"**: When there is an "and"/"or" separating relevant modifiers to the same harmful term, label them all together as one entity (See Examples 1-2). This is true of lists as well. However, if the listed items are not modifiers of the same harmful term, but rather separate concepts, then assess each item individually (See Example 3).
- **Example 1:** We'd label "inflammatory or harassing messages" as one harmful entity.
- **Example 2:** We'd label "fake names and photos" as one harmful entity.
- **Example 3:** With "... fake news story, a hoax, or a publicity stunt", we'd label "fake news story" and "hoax" individually as [Fraud], but not "publicity stunt".

Adjectives/Nouns

- Only label both the adjective & noun if both provide a level of specificity that help explain the reason for using a particular label on it. If one suffices, then don't group them under one label.
- With "plagiarizing purposes", we wouldn't label the whole thing as [Cybersecurity]. We'd only label "plagiarizing", since "purposes" isn't needed to determine that "plagiarizing" is harmful, and "plagiarizing" is still ultimately an action that happens to be used as an adjective in this sentence (but still describes the act of plagiarizing).
- With "harassing messages", we would label the whole thing as [Harmful Language]. This is because "messages" describes the harassment as linguistic, not physical.
- With "fake emails", we would label "fake emails" entirely as [Fraudulent], since the word to follow "fake" could change the harm entity label (e.g., "fake email" vs "fake porn").
- With "fraudulent activities", we would label just "fraudulent" as [Fraudulent], since that adjective suffices in justifying this particular label (no matter the noun that comes after, it won't change).
- With "malicious links", we would label the whole thing as [Cybersecurity], since "malicious" is an adjective that can be used to describe other harmful behavior as well.
- With "malicious purposes", we would label "malicious" based on local context though.

Verbs

- Only label individual verbs as harmful if, as standalone words in local context, they're still considered harmful.
- With "install the Trojan Horse", "install" would not be labeled as a [Cybersecurity] harm, whereas for "hack the system", "hack" would be labeled as [Cybersecurity].
- With an example like "exploit the system", "exploit" would be labeled as [Cybersecurity].

- **Example of NOT labeled**: “cause damage to” would not be labeled because “damage” does not correspond to a specific GuardNER label.

Dubious Terms

- If a dubious term is used in a non-harmful context (e.g., historical information) and is not NSFW, then it is not considered harmful.
- “Nazi” would not be considered [Harmful Language] if used to respond to a historical question, such as one about WWII German facts. However, it would be labeled [Harmful Language] if used like in a harmful way, such as this: “You’re a Nazi”.

“Attacker” vs “Attack”

- Don’t label “attacker”, but label “attack” if used to describe a particular type of harmful behavior (and label it according to what that type is).
- For “cybersecurity attack”, “attack” would be labeled as [Cybersecurity].

A.1.3 Tips for Managing Ambiguity When Annotating

(Tweaked from Universal NER’s Annotation Guidelines) (Mayhew, Stephen, 2022)

- "Choose the entity type based solely on the local context of the output."
- "When there is ambiguity, either because of lack of context or genuine ambiguity, always choose the literal meaning of the word(s)."
- "If the context doesn’t help, and the surface form is ambiguous, choose the most common usage."

A.2 Entity Labels & Descriptions

The structure of our codebook draws inspiration from Figure 3 in the "NuNER: Entity Recognition Encoder Pre-training via LLM-Annotated Data" research paper (Bogdanov et al., 2024).

B Appendix: Code and Data

Table 9: Codebook

Concept	Term Description	Entity Example	Additional Notes
Fraudulent	Dishonest, deceptive, and/or deviant behavior	Fake (as a verb), scam, impersonate, steal, fake email addresses	For a term like "Plagiarizing Purposes", only label "plagiarizing" as [Fraudulent], since "purposes" doesn't help with determining the label.
Physical Harm	Actions that inflict corporal harm onto self or others, or terms describing bodily wounds resulting from malicious actions	Sexual abuse, suicide, overdose, Lung Damage	"Lung Damage", "Skin Irritation", and "Respiratory Problems" are [Physical Harm] because of the context in which they're mentioned (using chemical weapons for inducing harm in these ways). Terms like "Eating disorder" are not [Physical Harm] if used in the context of discouraging harmful behavior to avoid harms like eating disorders.
Privacy Violation	Actions that violate people's privacy, actions that deal with unauthorized access, or terms describing PII/private data in the context of securing them without permission	Eavesdrop, passwords, sensitive information, credit card information, address, name, unauthorized access, breaking into	With a term like "phishing personal information", "personal information" should not be labeled since it's too broad. But a specific term like "sensitive information" (in the context of stealing) should be labeled as [Privacy Violation].
Harmful Language	Malicious terms, or terms that refer to malicious terms/actions in a harmful context (including mental/emotional harms)	Fat, chubby, harassing messages, derogatory terms, emotional trauma	Terms like "shame" and "damaging" are not harmful terms if they are used in a context that describes why something is wrong (e.g., "sending harassing messages about someone's weight can be damaging to their self-esteem."). Terms like "target area" are too generic to be considered harmful.
Cybersecurity	Terms used to describe malicious cybersecurity behavior or specific tools	Phishing, infect, malicious code	With a term like "phishing personal information", "phishing" should be [Cybersecurity]. Words like "Infect", if used in a harmful cybersecurity context, should be labeled [Cybersecurity].
Weapons	Traditional weaponry & any specific object described as a "weapon" or being weaponized for physical harm	Chemical weapon, knife, gun	"Device" and "homemade device" would not be considered weapons because they're not specific enough.
Sexually Explicit	NSFW sexual terms, or sexually-related terms/names used in NSFW context	Adult industry (case-by-case), Porn, dildo	N/A
Dangerous Materials	Materials or ingredients that are being repurposed or applied in the context of harmful use or creation of dangerous products	Bleach, ammonia, toxic powder, metal/wood (in certain contexts)	N/A
Drugs	Drug names, drug products, and drug categories used in the context of conducting harmful or deviant behavior	Medications, substances, OTC	"Medications" is to remain considered [Drugs] in the context of OTC medication for suicide, since NER models will be able to catch it if we use such straightforward label.
Destruction	Terms describing the act of destruction/damage/vandalism to property/objects/inorganic matter	Destroy, demolish, damage, crash (if about an object)	N/A