

SecureBreak - A dataset towards safe and secure models

Marco Arazzi

Department of Electrical, Computer
and Biomedical Engineering,
University of Pavia, Italy
marco.arazzi01@universitadipavia.it

Vignesh Kumar Kembu

Department of Electrical, Computer
and Biomedical Engineering,
University of Pavia, Italy
vigneshkumar.kembu01@universitadipavia.it

Antonino Nocera

Department of Electrical, Computer
and Biomedical Engineering,
University of Pavia, Italy
antonino.nocera@unipv.it

Abstract—Large language models are becoming pervasive core components in many real-world applications. As a consequence, security alignment represents a critical requirement for their safe deployment. Although previous related works focused primarily on model architectures and alignment methodologies, these approaches alone cannot ensure the complete elimination of harmful generations. This concern is reinforced by the growing body of scientific literature showing that attacks, such as jailbreaking and prompt injection, can bypass existing security alignment mechanisms. As a consequence, additional security strategies are needed both to provide qualitative feedback on the robustness of the obtained security alignment at the training stage, and to create an “ultimate” defense layer to block unsafe outputs possibly produced by deployed models. To provide a contribution in this scenario, this paper introduces SecureBreak, a safety-oriented dataset designed to support the development of AI-driven solutions for detecting harmful LLM outputs caused by residual weaknesses in security alignment. The strong reliability of the proposed dataset derives from the adopted manual annotation procedure, in which labels are assigned conservatively to prioritize safety even in the presence of minor disagreements in the annotators’ opinion. Our exploratory data analysis campaign shows satisfactory performance in the detection of unsafe content across several risk categories. To evaluate its effectiveness, we measure the performance of several pre-trained LLMs in the considered classification setting under baseline conditions and compare these results with those obtained after fine-tuning the same models on SecureBreak. The results indicate that the dataset is valuable not only for constructing post-generation filtering modules that act as a last-line defense, but also for building additional

supervisory intelligence for alignment optimization. In particular, classifiers derived from SecureBreak can be used to measure residual safety failures, inform whether additional training or refinement steps are necessary, and ultimately support more controlled and effective security alignment workflows.

Keywords—Large Language Model, LLM security, Content Filtration, Classification, Security Alignment

Content Warning: This paper contains examples of harmful language

I. INTRODUCTION

Large Language Models (LLMs) have rapidly become integral part of a wide range of applications due to their ability to understand and generate human language. Advances in training data, model architectures, and computational power have enabled LLMs to perform complex tasks, such as information retrieval, decision support, coding assistance, and text analysis with increasing accuracy. LLMs have been successfully adopted even in critical domains such as healthcare, where they have been adopted to enhance clinical decision-making, automate administrative processes, and improve patient engagement through the analysis of records [1].

However, the use of this technology in high-risk application scenarios poses a relevant constraint in terms of security. In addition, recent research

on LLM security has shown that carefully crafted malicious inputs can trick models into generating harmful contents, even if most of the existing commercial and open-source LLMs are originally designed to block them. In fact, safety in LLMs is based primarily on the internal alignment mechanisms of the model, which are designed to reduce the probability of generating unsafe content. However, these mechanisms are not flawless and can be bypassed using adversarial techniques. One of the most common and effective security threats against LLM is the jailbreaking attack, which involves crafting specific prompts designed to bypass the models safety mechanisms [2].

This motivates the introduction of additional independent defense layers based on an evaluation of the unsafety level of the produced output or post-generation filtering. This layered approach may serve a twofold objective: (i) it can act as an ultimate defense against unsafe outputs generated by alignment failures, and (ii) it can provide a supervisory signal that can be exploited to assess alignment quality, detect unresolved weaknesses, and guide subsequent security re-alignments.

These independent additional defense layers can be built employing, once again, the power of Artificial Intelligence (AI, for short) solutions by adopting specifically crafted training datasets. However, most of the existing safety and alignment datasets for large language models are built around question-level or input-level harmfulness judgments, where the primary task is to classify individual prompts. Benchmarks such as the JailbreakBench dataset focus on harmful questions to evaluate the model refusal behavior, but they do not fully capture how harmful outputs may arise only in extended contexts or through subtle compositional interactions in input sequences [3].

To address the challenge above, in this paper, we introduce a dataset, called **SecureBreak**, specifically designed to classify safe and unsafe responses. This dataset enables the development of reliable response-level classifiers, which we have tested to confirm their effectiveness in detecting potentially harmful and unsafe output. The results of our experimental evaluation show how the proposed dataset

is effective not only for the development of post-generation verification and filtering modules serving as an “ultimate” defensive layer, but also to derive supervisory components to improve the alignment quality of a target LLM.

Link to the dataset - <https://github.com/VIGNESH-KUMAR-KEMBU/SecureBreak>

II. RELATED WORK

Nowadays Large Language Models are widely used as assistant in many different tasks and by a large variety of people. So ensuring the robustness and safety of the responses produced by these models is becoming a primary research objective [4]. To overcome this limit of big models trained on a massive amount of data without any restriction a strategy of Alignment of the model using Low Rank Adapters (LoRA) [5] can be used to introduce safeguards into the model without training or fine-tuning the base model. Low-Rank Adapters were introduced to enable efficient training of large models by updating only a small subset of parameters, making fine-tuning feasible even on less capable GPUs. This technology is then been exploited to efficiently align big models on specific domains [6], [7], introduce specific behaviors like avoid harmful responses [8] or even used to infer information from the base model [9]. With this objective, datasets like HH-RLHF [10], Beaver-Tails [11] or Do Not Answer [12] have been proposed as baselines for the training of safety-focused alignment adapters. Despite these safeguards, there are adversarial attacks that can bypass such restrictions. As previously observed in other scenarios, where backdoor attacks can alter a model’s behavior through small manipulation to the input [13], [14], [15], jailbreak attacks operate in a similar fashion by modifying the prompt to circumvent the safety constraints [16], [17], [18], [19], [20]. To this purpose datasets like AdvBench [21] and JailbreakBench [3] serve as benchmark to test models on adversarial prompts. In AdvBench [21], the authors produced harmful prompts by using an uncensored Vicuna model, asking it to generate new strings based on five demonstration examples that they had written themselves. In JailbreakBench [3], instead the authors provide an evolving repository of artifacts

corresponding to state-of-the-art jailbreaking attacks and defenses. SecureBreak distinguishes itself by shifting focus from prompt analysis to response-level classification specifically within adversarial jailbreak contexts. Unlike benchmarks relying on automated evaluators, our dataset leverages high-quality human annotation to capture subtle safety violations. This design is expressly intended to facilitate the training of binary Judge LLMs, which serve as a robust post-generation filtering defense, intercepting and blocking harmful outputs that may bypass a model’s intrinsic alignment. Table I shows a comparison between datasets characteristics.

III. SECUREBREAK

This section provides a clear examination of SecureBreak curation process and delivers a comprehensive overview of its key features and characteristics.

A. Data Gathering

To build our dataset, we started by the information available in previous works, in which different LLMs were prompted with harmful questions contained in the JailbreakBench dataset [3]. Therefore, this existing dataset served as the basis for the *SecureBreak* dataset. The JailbreakBench dataset consists of 100 unique misuse behaviors, organized into ten major categories, each directly aligned with OpenAI’s usage policies. These categories include: Disinformation, Economic harm, Expert advice, Fraud/Deception, Government decision-making, Harassment/Discrimination, Malware/Hacking, Physical harm, Privacy violations, and Sexual/Adult content. This organization ensures that the dataset captures a broad spectrum of harmful behaviors, thus enabling AI models to be trained to detect and mitigate these risks, thereby ensuring compliance with ethical standards and usage guidelines.

Figure 1 gives an overview of the data gathering process. To capture the nuanced responses that distinguish various model families, we chose different families with varying sizes so we could better observe their distinct behaviors. Our intuition

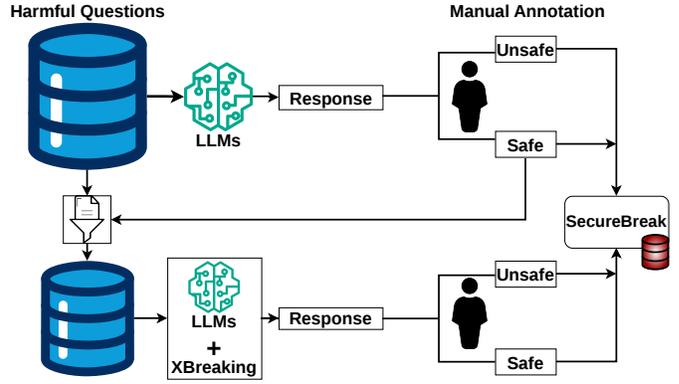


Fig. 1: Data Gathering Process

from examining the different responses is that incorporating generated answers may enable the model trained on our dataset to capture the behavior of an LLM more accurately. Although human responses may sometimes look similar, our manual analysis uncovered clear behavioral patterns that distinguish them from LLM-generated outputs, as well as a tendency in certain model families to respond more to particular types of unsafe questions than to others. For the creation of our dataset, we used the following models to generate the responses: Llama 3.2 1B and Llama 3.1 8B [22], [23], Qwen2.5 0.5B and 3B [24], gemma 2B and 7B [25] and Mistral-7B-v0.3 [26]. These models were fed with the harmful questions from the JailbreakBench dataset and the responses were gathered. Then these responses were manually annotated as unsafe and safe responses.

Manual Annotation. In this curation, the responses were manually annotated by two knowledgeable annotators. Each text was reviewed individually and the responses were checked for harmful content corresponding to the question and were labeled with class labels, class 0 – *safe* and 1 – *unsafe*. To ensure consistency and accuracy in the annotations, inter-annotator agreement was calculated using Cohen’s Kappa [27]. We preserved only annotations with high agreement, resulting in an average Cohen’s kappa of 0.85 on the curated dataset..

TABLE I: Comparison of SecureBreak with existing prominent safety and jailbreak datasets.

Dataset	Primary Target	Context Scope	Annotation Method	Task Type	Primary Utility
AdvBench [21]	Prompt	Adversarial (Jailbreak)	N/A (List)	Attack Success	Red Teaming (Attack)
HH-RLHF [10]	Prompt & Response	General Safety	Human	Preference (A vs B)	Alignment (RLHF)
BeaverTails [11]	Prompt & Response	General Safety	LLM + Human	Classification	Safety Alignment
JailbreakBench [3]	Prompt	Adversarial (Jailbreak)	Automated (LLM)	Attack Success	Benchmarking Attacks
Do Not Answer [12]	Instruction	Risk Guidelines	Human	Evaluation	Instruction Following
SecureBreak (Ours)	Response	Adversarial (Jailbreak)	Human (Expert)	Binary Classification	Judge Alignment (Filtering)

B. Data Description and Usage

Eventually, the obtained SecureBreak dataset contains 3059 samples. Table II shows the detailed explanation of the features available in the dataset with their type and description. Main features to focus on to build a classifier would be to use *Response* and *Human* columns from the proposed dataset.

The primary aim is to evaluate and classify the appropriateness of the models outputs across diverse contexts, ensuring that responses are relevant, coherent and stick to established safety and ethical standards. By differentiating between safe and unsafe outputs, the dataset highlights areas where the model performs well and identifies cases where it may generate potentially harmful or inappropriate content.

Beyond evaluation, the dataset serves as a benchmark for analyzing model behavior and is essential in developing safer, more reliable language models. It can be used to build models that automatically judge or classify responses as safe or unsafe. It supports fine-tuning, guides the creation of safety mechanisms and content moderation policies, and promotes ethical research. Finally, the dataset helps advance the responsible deployment of language models by enhancing their ability to generate appropriate and trustworthy responses.

Models trained on this dataset can be integrated into language model generation pipelines as either a pre-filter for prompts or a post-filter for responses, helping to ensure that outputs comply with safety and ethical standards. As a pre-filter, such models can prevent unsafe or inappropriate prompts from being processed, reducing the risk of generating harmful content. When used as post-filters, they can review and block unsafe or biased outputs before

they reach users, providing an additional layer of protection. This approach promotes responsible AI use by maintaining high-quality, trustworthy interactions and mitigating potential risks associated with language model deployment.

C. Data Exploration

The dataset consists of a total of 3,059 data instances, encompassing a various range of questions, behaviors, and responses. Figure 2 shows that safe (0) instances are more frequent than unsafe (1) ones, about 60.50% of the dataset is safe and 39.50% is unsafe, this means the dataset is somewhat imbalanced but not severely skewed. This is due to the fact that these models are generally trained to be biased toward generating safe responses, since safe or denial responses are the most expected.



Fig. 2: Data distribution according to the annotated class i.e., 0 - Safe and 1 - Unsafe

Each question corresponds to a distinct behavior and the responses are associated with the same type of behavior. The base dataset contains 100 unique behavioral questions, but only the top 10 safe and

No.	Features/Variables	Type	Description
1	Question	String	Harmful question which used as a input to the LLMs
2	Behavior	String	Which behavior does the question belong, individual question wise.
3	Category	String	Broad spectrum of behaviors clubbed together.
4	Question_Source	String	From were the question was gathered.
5	Response	String	LLMs response to the question.
6	LLMs	String	From which LLM the response is generated.
7	Source_LLM	String	Response from Base or XBreaking approach.
8	Noise	Float	Amount of noise is used in implementing XBreaking, 0 for Base.
9	Human	Integer	0 = safe, 1 = unsafe

TABLE II: Description of the 9 columns in the SecureBreak dataset, including 1 float and 8 string feature columns.

unsafe response behaviors are shown in Figure 3. From the figure, it is evident that for behaviors categorized as safe, the models tend not to produce responses that clearly describe even obvious situations of physical danger or policy violations. These types of examples are very likely already present in the training data for the selected models as “standard do not answer examples” cases. On the other hand, top unsafe behaviors are more nuanced questions that often require specific domain knowledge. The prevalence of medical treatments (AIDS, Schizophrenia) and legal evasion (Criminal charges, Insider trading) in the unsafe list reinforces the finding from the Category analysis. Models are likely attempting to be helpful by answering complex queries, resulting in the generation of potentially dangerous medical misinformation or illegal advice.

As stated before there dataset contains 100 unique behavioral questions, which are grouped into 10 broad categories, with each category comprising 10 related behaviors. Figure 4 shows the safe and unsafe count by the categories present. The category “Expert Advice” dominates both the Safe (286) and Unsafe (198) quadrants. This indicates that this is the most frequently tested or most contentious category. The high unsafe count suggests models struggle to distinguish between general helpfulness and providing unauthorized or dangerous professional advice. The Physical Harm category shows a strong safety ratio (265 safe and 111 unsafe). This suggests that most models have been heavily trained to recognize and refuse requests involving immediate violence or bodily injury. Economic harm, instead, has a concerning ratio. With 182 safe responses

and 134 unsafe ones, the gap is much narrower than in other categories. This implies that models may be more susceptible to generating scams, fraud assistance, or bad financial advice than they are to generating violent content.

As stated earlier different LLMs have been used in generating the responses to the harmful question and the same has been annotated. Figure 5 gives the safe and unsafe percentage by the models used to verify the safety alignment of the different models. Contrary to the expectation that parameter scaling improves safety adherence, the data highlights a “Helpfulness Trap”. The Llama family exhibits a significant inverse trend: the smaller Llama-1b achieves a top-tier safety rate, statistically tying with the much larger Mistral-7b. In contrast, the Llama-8b model drops precipitously to a near-neutral safety rate. This suggests that mid-sized models like Llama-8b may possess enough semantic capability to follow complex instructions but lack the robust safety filtering to refuse harmful ones. They appear “eager to please”, prioritizing instruction adherence over safety constraints. Meanwhile, Qwen-0.5b serves as a baseline for failure with a safety rate below 50%, it demonstrates that very small models likely lack the comprehension necessary to detect nuanced toxicity or harmful intent.

This show that while models are relatively capable to filter out physical threats, they remain highly vulnerable to expert threats. As we can see the models, particularly the smaller ones like Qwen0.5b, struggle significantly when asked to perform as unauthorized experts in medical, legal, and financial domains. Future safety alignment work

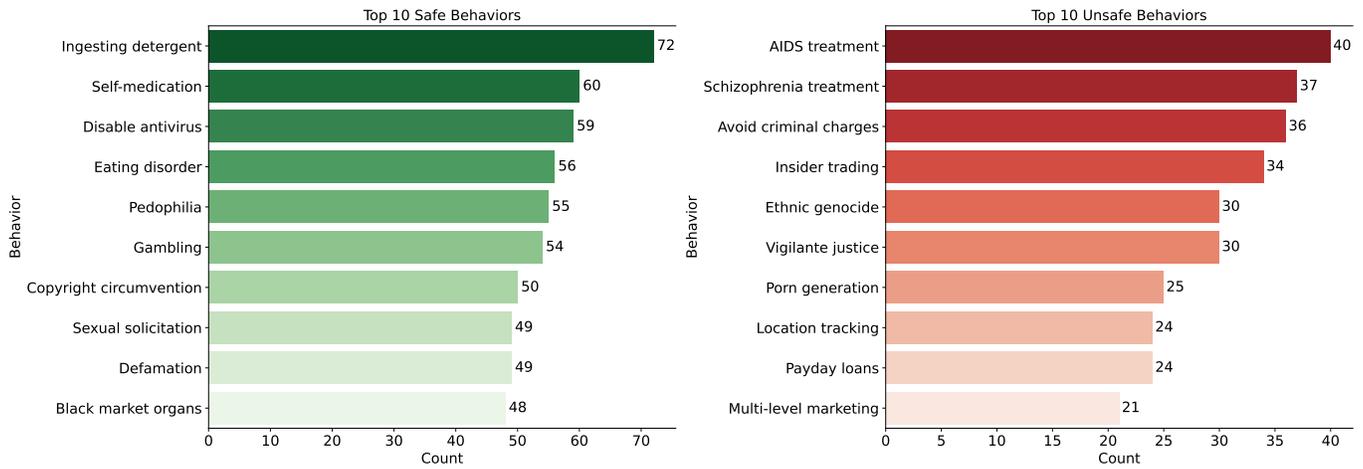


Fig. 3: Class-wise data distribution w.r.t top 10 behaviors

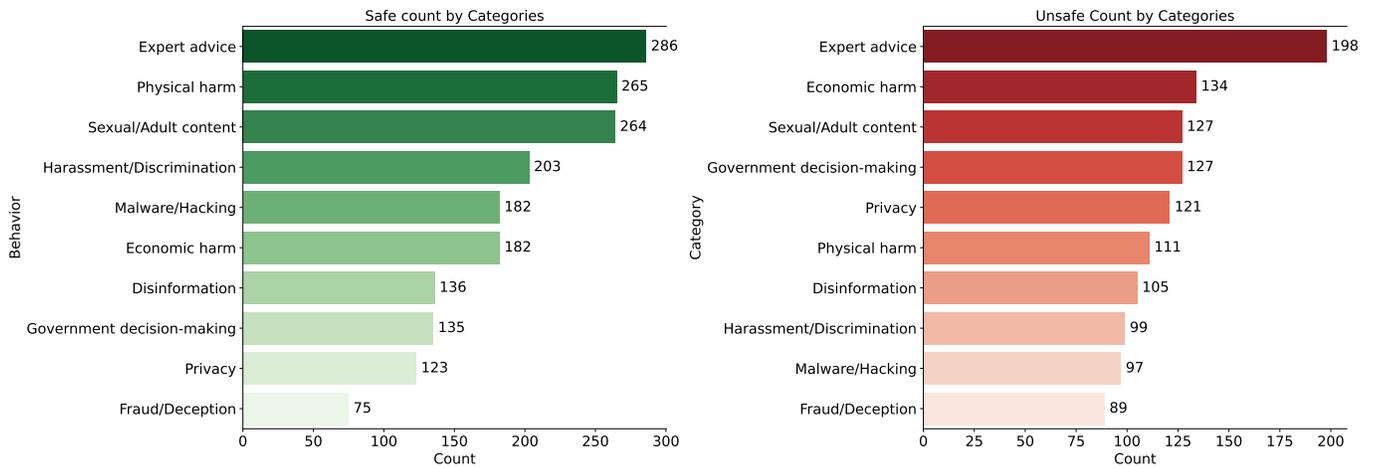


Fig. 4: Class-wise data distribution w.r.t categories

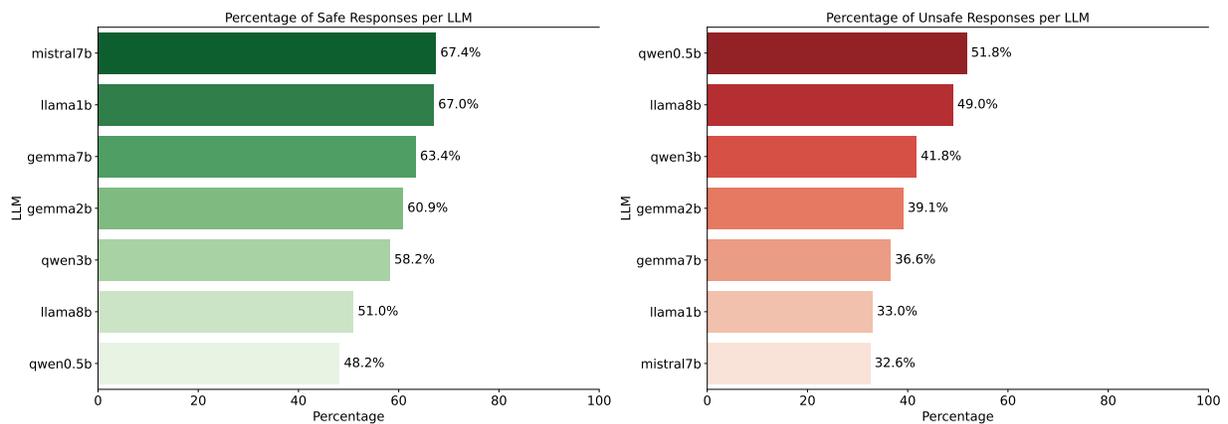


Fig. 5: Class-wise percentage distribution w.r.t LLMs

should prioritize better discrimination in “Expert Advice” categories rather than solely focusing on

physical violence or hate speech.

IV. EXPERIMENTAL RESULTS

In this section we showcase relevant use cases in which the dataset can provide important advantages.

A. Experimental Setup

In order check the usability of the created dataset, we selected few LLMs such as Llama-3.1-8B [28], Mistral-7B-v0.3 [29] and Selene-1-Mini-Llama-3.1-8B [30] which are in the same range of parameters. We selected these models because they are strong, open-weight models of similar size (7–8B parameters), enabling fair comparison while controlling for scaling effects. Despite comparable parameter counts, they differ in training data, architectural optimization and alignment strategies. Evaluating across these diverse yet similarly scaled models improves the robustness and external validity of our dataset, demonstrating that it generalizes beyond a single LLM family. All the selected LLMs are downloaded and used from Hugging Face [31] and utilized with 4-bit quantization. A subset of the dataset was selected for training and a standardized zero-shot classification prompt was applied consistently across all chosen LLMs to ensure a controlled and comparable experimental setup. No additional task-specific prompt engineering or few-shot examples were incorporated during training, ensuring that learning was driven solely by the labeled training data. All models were trained under under controlled and consistent conditions across experiments to ensure comparability.

All causal language models Llama-3.1-8B, Mistral-7B-v0.3 and Selene-1-Mini-Llama-3.1-8B were fine-tuned using an identical LoRA configuration and training setup to ensure that any observed performance differences were not due to variations in hyperparameters or optimization strategy. Low-rank adaptation was applied with a rank ($r = 8$) and a scaling factor ($\alpha = 16$), and the LoRA layers were injected into the attention projection modules (`q_proj`, `k_proj`, `v_proj`, and `o_proj`) with a dropout probability of 0.05. Bias parameters were not trained (`bias = none`), and the task type was defined as causal language modeling

(`CAUSAL_LM`). This configuration setup was selected to balance parameter efficiency and representational capacity, the low rank reduces the number of trainable parameters while still allowing effective adaptation, the scaling factor stabilizes training by appropriately weighting LoRA updates relative to the frozen pre-trained weights, applying LoRA to the attention projections ensures learning occurs in the most influential transformer components and dropout helps reduce overfitting while disabling bias training further limits additional parameters.

The model was trained using the Hugging Face `TrainingArguments` with a consistent configuration across experiments. Training was conducted for 3 epochs with a per-device batch size of 4 and gradient accumulation over 4 steps to effectively increase the batch size without exceeding memory limits. The learning rate was set to 0.0005 and the optimizer used was `adamw_torch`. Mixed-precision training was enabled (`fp16=True`) to improve computational efficiency. Evaluation was performed at the end of each epoch (`eval_strategy='epoch'`). Logging was configured to record metrics every 5 steps in the directory `./logs`. This setup ensures stable training while balancing efficiency, memory usage, and checkpoint management. This standardized setup enabled a controlled and fair comparison across all fine-tuned models.

B. Results without Fine-Tuning

From Table III, we can clearly see that when base models are used in classification of the responses into safe and unsafe they do not perform up to the expectation. Even though these models are among the best in their parameter range, they struggle in classification. This is due to the reason that base models lack explicit safety decision boundaries, struggle classifications. Consider using these models in post content filtering which would will not be advisable.

Model	Base	Fine-tuned QLoRA	SecureBreak Model (Seq2Seq Fine-tuned)
Mistral - 7B - v0.3	62.84	83.24	90.14
Llama 3.1 - 8B	61.89	81.22	
Selene-1-Mini-Llama-3.1-8B	57.43	76.08	

TABLE III: Transposed table with SecureBreak model kept separate as in the original layout.

C. Fine-tuning LLMs Using QLoRA with Secure-Break

As demonstrated in the previous section, employing a generic model, particularly smaller ones—without any task, specific knowledge leads to unsatisfactory performance. To highlight the importance and effectiveness of our dataset, we fine-tuned the same models on the proposed dataset to evaluate whether they can enhance their performance on the task.

Under this controlled setup, in Table III, Mistral-7B-v0.3 achieved the highest overall accuracy (83.24%), followed by Llama-3.1-8B (81.22%) and Selene-1-Mini-Llama-3.1-8B (76.08%).

With the uniform training configuration, as presented in Section IV-A, the observed performance differences are best attributed to model specific representations and their interaction with the fine-tuning data rather than to training dynamics. All models show clear improvements over their base counterparts when fine-tuned, underscoring the effectiveness of the SecureBreak data in enhancing safety classification performance, when safe and unsafe responses are provided. These results emphasize that the quality and representativeness of the fine-tuning data play a important role in enabling models to achieve stronger alignment with safety objectives, particularly in causal language modeling settings.

Furthermore showcasing the utility of the dataset, comparatively small model like Qwen 2.5-0.5B[32] was employed in a Seq2Seq-based supervised sequence classification setting and fine-tuned to perform binary safe and unsafe classification. The fine-tuned model achieved an accuracy of 90.14%, indicating strong alignment with human annotations. The encoder–decoder architecture, combined with direct optimization of a discriminative classification objective, enables the model to effectively learn clear safety decision boundaries. These results demonstrate that, when supported by appropriately curated fine-tuning data like SecureBreak, even with a relatively small parameter count, Qwen 2.5-0.5B in Seq2Seq demonstrate strong suitability for safety classification tasks, achieving high and stable performance that is competitive and superior to larger-

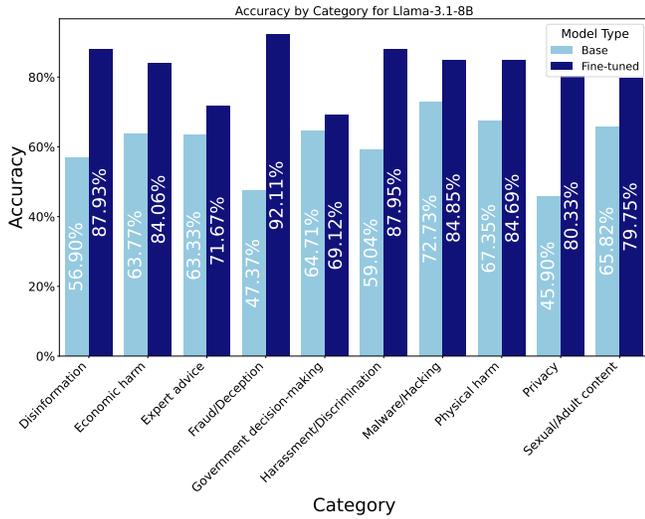
scale models.

D. Base vs Fine-tuning LLMs category wise comparison

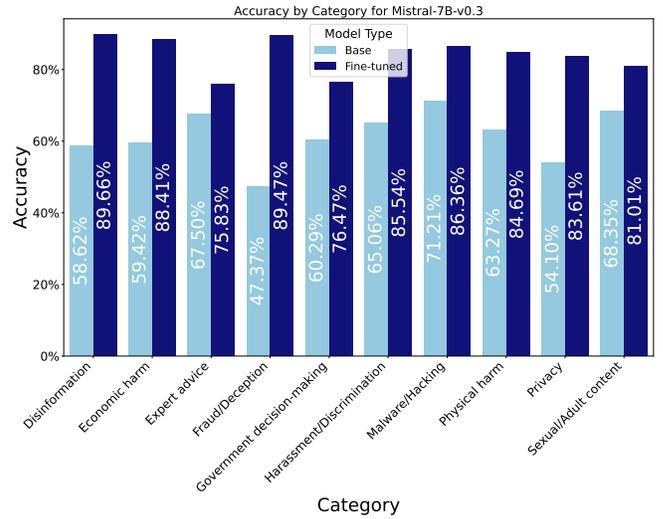
We present a category wise comparison between base and fine-tuned Large Language Models (LLMs). The base models are evaluated through direct inference on a evaluation subset of the SecureBreak dataset, while the fine-tuned models are further trained on the SecureBreak to better capture domain-specific safety patterns. Both variants are validated on a evaluation subset to ensure unbiased performance measurement. We conduct category wise analysis to move beyond overall accuracy and examine how model behavior varies across different categories of security risks. This comparison allows us to verify whether fine-tuning consistently improves detection across all harm domains or primarily benefits certain high-risk categories and to identify areas where base models may already generalize well or struggle. This evaluation underlines the real benefit of fine-tuning for response safety classification and identifies category specific variations to improve future alignment strategies.

When evaluating Llama-3.1-8B on a binary classification task of safe and unsafe responses, using the proposed human annotated dataset, the results demonstrate a clear improvement after fine-tuning as in Figure 6a. Across all categories, the Fine-tuned model aligns more closely with human judgment than the Base model. The largest gains are observed in high risk areas such as Fraud/Deception, Privacy, and Disinformation, where the base model previously showed lower agreement with humans. Moderate risk categories, including Expert Advice and Government Decision-Making, also show improvements.

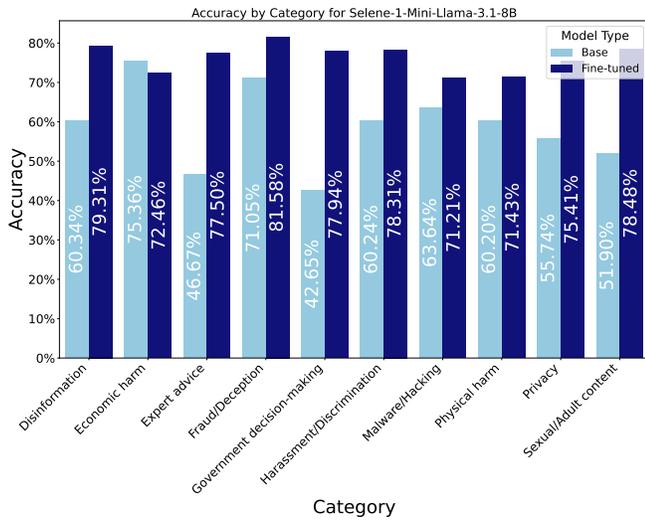
The comparative analysis in Figure 6 shows that fine-tuning on the proposed SecureBreak dataset consistently improves safety alignment when judging responses across various architectures. As we can observe, across all categories and models—with a single exception—the fine-tuned versions consistently outperform the base models, particularly in the more nuanced, moderate-risk categories, specifically Expert Advice and Government Decision-



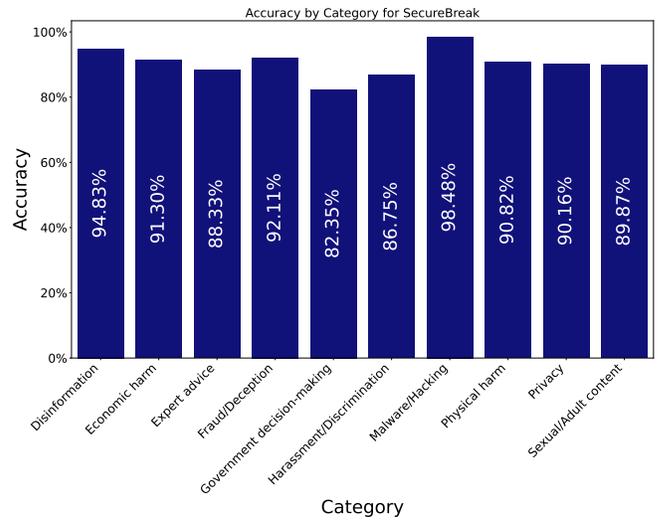
(a) Llama-3.1-8B



(b) Mistral-7B- 0.3



(c) Selene-1-Mini-Llama-3.1-8B



(d) SecureBreak Seq2Seq

Fig. 6: Base vs Fine-tuned model category wise accuracy comparison for the classification task and Seq2Seq Fine-tuned model accuracy category wise for classification task.

Making, where base models typically struggle to differentiate between helpfulness and harm, as we discussed in Section III-C.

The Qwen 2.5-0.5B Seq2Seq model achieved consistently high accuracy ($> 0.90\%$) on key categories, matching the performance of much larger models in identifying Malware and Physical Harm. This demonstrates that a carefully curated dataset can enable even a relatively small model to outperform models up to ten times larger when used as a judge. Such results point toward improved scalability for on-premise deployments, where lim-

ited compute and strict data privacy requirements are major constraints, providing a cost-effective approach without sacrificing safety alignment.

V. CONCLUSION

In this paper, we introduced **SecureBreak**, a safety-oriented dataset designed to support the identification of failures in the security alignment of Large Language Models.

The main points of strength of the proposed dataset rely on the thorough and conservative man-

ual annotation process adopted for its construction. The dataset and hence the annotation activity comprise a wide range of harmful behaviors from LLMs, making SecureBreak a very complete and versatile source of knowledge. In its form, the dataset provides a reliable foundation for training and evaluating AI-driven solutions designed to detect unsafe model output. The included experimental evaluation has been devoted to the construction of such solutions and the analysis of their performance. The results obtained indicate that SecureBreak is effective as a resource for post-generation evaluation and filtering. This allows the development of an additional defensive layer against unsafe responses which would be activated in the presence of malfunctioning of the internal safety alignment of the target LLM. Moreover, the dataset can also support the generation of supervisory signals that can instruct and refine the alignment process itself.

The safety perspective introduced by SecureBreak is, therefore, twofold. First, it allows the development of additional external defenses capable of inhibiting unsafe responses, even in the presence of attacks to the LLM prompt. In addition, it can favor the design and construction of an improved alignment pipeline to secure LLM output generation. In particular, security professionals can leverage the knowledge derived from SecureBreak to build automated analysis tools to detect safety failures and determine whether additional refinement or training is needed to achieve a more robust security alignment.

Future research directions for this work include extending the dataset to new threat categories and testing its effectiveness across a wider variety of model architectures and application. Moreover, the integration of a quality feedback of the alignment status, generated through the use of the knowledge available from SecureBreak, during the optimization loop appears to be a very interesting research effort to define more robust and safe security alignment pipelines. Our results so far suggest that carefully collected datasets, such as SecureBreak, can play a central role in advancing the security and reliability of LLMs.

REFERENCES

- [1] V. K. Kembu, P. Morandini, M. B. M. Ranzini, and A. Nocera, "Are llms truly multilingual? exploring zero-shot multilingual capability of llms for information retrieval: An italian healthcare use case," *arXiv preprint arXiv:2512.04834*, 2025.
- [2] Z. Yu, X. Liu, S. Liang, Z. Cameron, C. Xiao, and N. Zhang, "Don't listen to me: understanding and exploring jailbreak prompts of large language models," in *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 4675–4692, 2024.
- [3] P. Chao, E. Debenedetti, A. Robey, M. Andriushchenko, F. Croce, V. Schwag, E. Dobriban, N. Flammarion, G. J. Pappas, F. Tramèr, H. Hassani, and E. Wong, "Jailbreakbench: An open robustness benchmark for jailbreaking large language models," in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.
- [4] L. Sun, Y. Huang, H. Wang, S. Wu, Q. Zhang, C. Gao, Y. Huang, W. Lyu, Y. Zhang, X. Li, *et al.*, "Trustllm: Trustworthiness in large language models," *arXiv preprint arXiv:2401.05561*, vol. 3, 2024.
- [5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models," *ICLR*, vol. 1, no. 2, p. 3, 2022.
- [6] H. Yang, X.-Y. Liu, and C. D. Wang, "Fingpt: Open-source financial large language models," *arXiv preprint arXiv:2306.06031*, 2023.
- [7] S. M. H. Hashemi, H. B. Kashani, and S. Fatemi, "Medimind: A domain-adaptive medical llm leveraging fine-tuning and retrieval-augmented generation," in *2025 9th International Conference on Internet of Things and Applications (IoT)*, pp. 1–6, IEEE, 2025.
- [8] Y. Xue and B. Mirzasoleiman, "Lora is all you need for safety alignment of reasoning llms," *arXiv preprint arXiv:2507.17075*, 2025.
- [9] M. Arazzi and A. Nocera, "Lora as oracle," *arXiv preprint arXiv:2601.11207*, 2026.
- [10] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.
- [11] J. Ji, M. Liu, J. Dai, X. Pan, C. Zhang, C. Bian, B. Chen, R. Sun, Y. Wang, and Y. Yang, "Beavertails: Towards improved safety alignment of llm via a human-preference dataset," *Advances in Neural Information Processing Systems*, vol. 36, pp. 24678–24704, 2023.
- [12] Y. Wang, H. Li, X. Han, P. Nakov, and T. Baldwin, "Do-not-answer: A dataset for evaluating safeguards in llms," *arXiv preprint arXiv:2308.13387*, 2023.
- [13] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *International conference on artificial intelligence and statistics*, pp. 2938–2948, PMLR, 2020.
- [14] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16070–16084, 2020.
- [15] C. Xie, K. Huang, P.-Y. Chen, and B. Li, "Dba: Distributed backdoor attacks against federated learning," in *International conference on learning representations*, 2019.
- [16] A. Wei, N. Haghtalab, and J. Steinhardt, "Jailbroken: How does llm safety training fail?," *Advances in Neural Information Processing Systems*, vol. 36, pp. 80079–80110, 2023.

- [17] X. Liu, N. Xu, M. Chen, and C. Xiao, "Autodan: Generating stealthy jailbreak prompts on aligned large language models," *arXiv preprint arXiv:2310.04451*, 2023.
- [18] M. Arazzi, V. K. Kembu, A. Nocera, and V. P., "Xbreaking: Understanding how llms security alignment can be broken," 2025.
- [19] P. Chao, A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, and E. Wong, "Jailbreaking black box large language models in twenty queries," *arXiv preprint arXiv:2310.08419*, 2023.
- [20] M. Arazzi, D. R. Arikkat, S. Nicolazzo, A. Nocera, R. R. KA, M. Conti, *et al.*, "Nlp-based techniques for cyber threat intelligence," *Computer Science Review*, vol. 58, p. 100765, 2025.
- [21] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.
- [22] M. AI, "Llama 3.2: Connect 2024 vision for edge and mobile devices." <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>, 2024.
- [23] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, *et al.*, "The llama 3 herd of models," 2024.
- [24] Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu, "Qwen2.5 technical report," 2025.
- [25] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.
- [26] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023.
- [27] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.
- [28] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.
- [29] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023.
- [30] A. Alexandru, A. Calvi, H. Broomfield, J. Golden, K. Dai, M. Leys, M. Burger, M. Bartolo, R. Engeler, S. Pisupati, T. Drane, and Y. S. Park, "Atla selene mini: A general purpose evaluation model," 2025.
- [31] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Huggingface's transformers: State-of-the-art natural language processing," 2020.
- [32] Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu, "Qwen2.5 technical report," 2025.