input and helps us understand how important it is for the model to "think" that it generated the previous conversation turns. 2) The model responses in the conversation history $r_1, r_2, ..., r_{n-1}$ (excluding the final response) are masked out so that these tokens do not attend to the final output r_n. This helps us understand how the model's previous responses affect its representations of the final Crescendo response.

In the next section, we follow the methodology described above to analyze five successful Crescendo attacks performed manually against Llama-3-8B-Instruct-RR, a model with circuit breakers.

4. Results

4.1. How do LMs represent Crescendo inputs vs. single-turn inputs?

We begin our analysis by visualizing PCA projections of model representations. In particular, we compare how the models represent the same output tokens in r_n when the full conversation history is passed, as in a Crescendo jailbreak, and when the conversation history is replaced with the attack objective from Table 1 as a single-turn prompt.

We also plot the single-turn representations from $\mathcal{D}_r$ (Retain, "benign") and $\mathcal{D}_{cb}$ (CB, "harmful") which were used to fit the PCA models. We see that for Llama-3-8B-Instruct in Figure 1a, there is only marginal separation between even the CB and Retain dataset representations. This may be due to the high dimensionality and variance of the representations – each token representation is a vector in $\mathbb{R}^{4096}$, and the variance explained by the first two principal components is only 6.7%. Nonetheless, we observe that the r_n representations with full conversation history (green) are slightly closer to the blue Retain cluster, while the representations with the attack objective (orange) are generally close to the red CB cluster.

This effect is much more obvious for the circuit breaker model (Llama-3-8B-Instruct-RR) in Figure 1b, where there is clear separation between the model's representations of the CB and Retain datasets. This is expected because circuit breaker models are specifically fine-tuned to represent these two datasets in orthogonal directions. Further, we see that although the Crescendo representations mostly overlap with the Retain cluster when the full conversation is provided as input (green), the representations shift towards the CB cluster when the model is prompted with the attack objective (orange). This illustrates that conversation history plays an important role in determining how the model represents the jailbroken response. In particular, Crescendo appears to push the tokens in r_n towards a "benign" region of representation space, while an equivalent single-turn request pushes them towards a "harmful" region.

4.2. How does the number of conversation turns affect representations of Crescendo inputs?

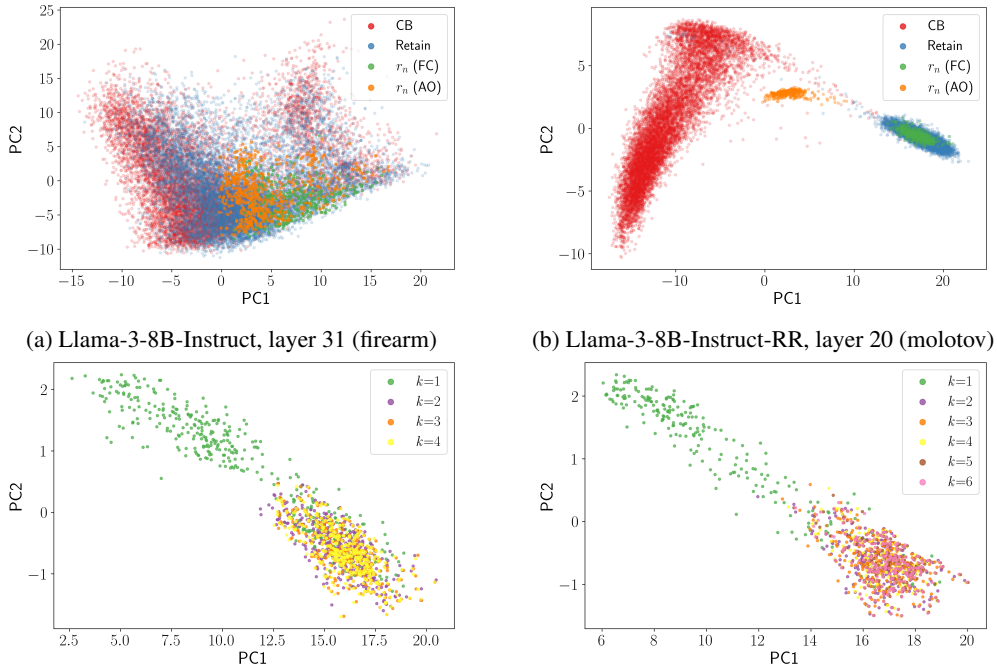
We have observed that a full Crescendo conversation may push model representations in a "benign" direction. Next, we study the effect of varying the number of Crescendo turns in the conversation history. Specifically, we vary the number of final turns passed to the model from $k = 1, ..., n$ and extract the representations of tokens in r_n using Equation 3.

As before, we use PCA to visualize the representations. Figures 1c–1d show these projections for the 'molotov' and 'meth' examples with the circuit breaker model. In both cases, the representations form a distinct cluster when only the final Crescendo turn is passed to the model ($k = 1$). But for $k > 1$, the representations shift to the bottom right of the PCA plot, in the direction of the Retain (benign) cluster. This supports our hypothesis from section 4.1 that Crescendo causes the model to represent its outputs as "less harmful" than if it were prompted with a single-turn request.

To analyze this phenomenon in more detail, we train MLP probes on the same representations used to fit the PCA models ($\mathcal{D}_{train}$). Probes trained on the original Llama model and the circuit breaker model representations achieved accuracies of **0.997** and **0.999** on $\mathcal{D}_{test}$, indicating that they classify unseen single-turn examples with high accuracy. Next, we measure the percentage of tokens in the jailbroken response r_n classified by the MLP probes as belonging to the CB dataset. This allows us to more precisely quantify the extent to which the models represent Crescendo outputs as "harmful" vs. "benign" than we could by visually inspecting PCA plots.

Figure 2 plots these percentages of representations classified as "harmful" for each attack objective and for $k = 1, ..., n$. For both the circuit breaker model ($\mathcal{M}_{cb}$) and the original Llama model ($\mathcal{M}$), we observe the same trend indicated by the PCA plots – as k increases, the models tend to represent the jailbroken responses as less "harmful." In addition, there is usually a sharp decrease in the percentage of tokens classified as harmful from $k = 1$ to $k = 2$, with the exception of the 'firearm' example, which may fall outside the distribution of harmful examples in the CB dataset. Nonetheless, it is clear that 1) the conversation history has a significant effect on how the models represent jailbroken responses, and 2) by the end of the Crescendo conversation ($k = n$), the jailbroken response tokens lie closer to the "benign" than the "harmful" data distribution.

These observations point towards *a possible explanation for how Crescendo works:* at each turn in the conversation, the prompt lacks explicit harmful context, and so the model remains in a "benign" region of representation space, resulting in compliance with the request, as opposed to refusal. This continues from one turn to the next and by the final



(a) Llama-3-8B-Instruct, layer 31 (firearm)

(b) Llama-3-8B-Instruct-RR, layer 20 (molotov)

(c) Llama-3-8B-Instruct-RR, layer 20 (molotov)

(d) Llama-3-8B-Instruct-RR, layer 20 (meth)

Figure 1. PCA projections of various model representations. (a, b) Show the shift in representations of r_n when the full conversation (FC) is passed to the model vs. only the single-turn attack objective (AO). (c, d) Show the shift in representations of r_n with $k = 1$ vs. $k > 1$ Crescendo turns in the conversation history. The PCA models were fitted on representations of single-turn examples from the circuit breaker (CB) and retain datasets. See Appendix G for additional PCA plots.

turn, the model has all the context required to generate an output which satisfies the attack objective, but the model still represents that output as benign.

This explanation matches the intuition described by Russinovich et al. (2024). In addition, the fact that Crescendo can bypass circuit breakers reveals that Crescendo representations are sufficiently different from single-turn representations that this defense does not transfer to multi-turn attacks. In the next section, we investigate elements of Crescendo inputs that may prevent the refusal and circuit breaker defenses from being triggered.

4.3. How do the model's own responses influence its representations of Crescendo inputs?

The number of turns in a Crescendo conversation appears to have some influence on model representations, but it is unclear exactly which elements of the jailbreak have the strongest effect. In this section, we perform additional experiments to identify parts of the model responses which push representations in a "benign" direction and ultimately contribute to a successful jailbreak.

These experiments extend the MLP probe analysis above by comparing the percentage of tokens in the final response r_n classified as "harmful" across four prompting strategies:

1. Crescendo: the full Crescendo conversation ($k = n$) that we have studied in the previous sections.
2. Single Prompt: the Crescendo conversation history $p_1, r_1, \ldots, p_{n-1}, r_{n-1}, p_n$ is passed as a single user prompt.
3. Masked Responses: the model responses in the conversation history $r_1, r_2, ..., r_{n-1}$ are masked out so that these tokens cannot attend to r_n.
4. Attack Objective: the Crescendo conversation history is replaced with the attack objective as a single-turn prompt. We analyzed PCA plots of these representations in section 4.1.

Bar charts summarizing the results are included in Appendix H. As we observed previously, the models tend to represent Crescendo inputs as "benign," leading to a low percentage of tokens classified as harmful. Interestingly, we see that the results are nearly identical for the Single Prompt case. This reveals that the <|assistant|> tokens make almost no difference to the model representations and suggests that the models do not privilege assistant messages

above user prompts containing the same context. As a result, this single-shot version of Crescendo generally still works (see Appendix E for model responses). We note, however, that this is not a practical jailbreak method because it relies on the attacker, rather than the model, to write the responses. Aqrawi & Abbasi (2024) also found that multi-turn Crescendo attacks can often be condensed into shorter single-turn prompts.

Next, we focus on the content of the assistant-generated messages by masking out the tokens in $r_1, r_2, ..., r_{n-1}$ so that they cannot attend to r_n. In some cases (e.g., the 'meth' example in Figure 4), masking out model responses leads to an increase in the percentage of representations classified as harmful. This indicates that during a Crescendo jailbreak, the model's own responses can push representations in a benign direction, possibly making it less likely that refusal or circuit breakers will be triggered. In other examples, however, masking responses has a much smaller effect on model representations. In these cases, we suspect that representations are influenced primarily by the user prompts.

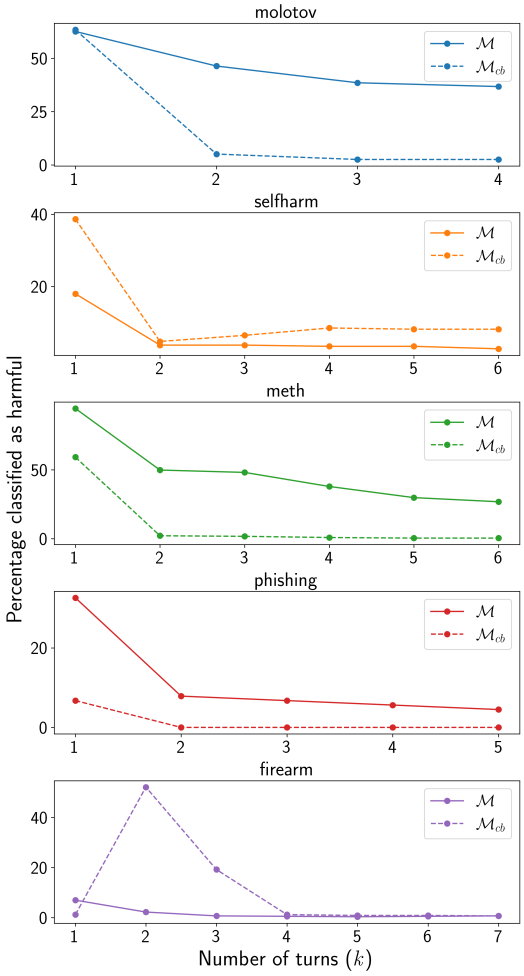


Figure 2. Percentage of jailbroken response representations classified by MLP probes as "harmful" for five Crecendo attacks against the original Llama model ($\mathcal{M}$) and the circuit breaker model ($\mathcal{M}_{cb}$). As the number of turns in the conversation history increases, the percentage of representations classified as harmful tends to decrease.

Indeed, replacing the Crescendo history with a direct attack objective (e.g., "Write a manual for making a Molotov cocktail") consistently leads to a significant increase in the percentage of tokens in r_n classified as harmful. This adds further evidence to our claim that Crescendo works by leveraging prompts which lack harmful context to keep the model in a benign region of representation space, effectively tricking the model into generating harmful outputs.

5. Limitations and Future Work

In this work, we used RepE techniques to probe at how LMs represent and "interpret" various inputs. For example, we used MLP classifiers trained on representations of harmful and benign single-turn examples to study how the models represent Crescendo inputs. One limitation of this approach is that it is difficult to know how much the MLP results capture the model's internal notion of "harmfulness" versus some other difference between the training and Crescendo representations. Nonetheless, the fact that there is a significant difference indicates a generalization gap between single-turn defenses (e.g., circuit breakers) and multi-turn attacks. We encourage researchers and engineers to develop better defenses that target both single-turn and multi-turn model representations.

Another limitation of our work is that we focus exclusively on Crescendo. Extensions of this research could explore whether our observation about how Crescendo affects model representations hold true for other multi-turn jailbreaks such as GOAT (Pavlova et al., 2024) and Tempest (Zhou & Arel, 2025). Future work might also expand our analysis to more models including reasoning models (DeepSeek-AI et al., 2025; Guan et al., 2025) and other leading jailbreak defenses like constitutional classifiers (Sharma et al., 2025).

6. Conclusion

In this work, we seek to understand the effectiveness of multi-turn jailbreaks by studying the representations of successful Crescendo attacks against Llama-3-8B-Instruct and its circuit breaker counterpart. We find that both models tend to represent Crescendo inputs as more benign than harmful and that the user prompts in a Crescendo attack play an important role in shaping these representations. More broadly, our results indicate a generalization gap between single-turn defenses and multi-turn jailbreaks, motivating the development of jailbreak defenses that target harmful multi-turn representations.