MSE-Break: Steering Internal Representations to Bypass Refusals in Large Language Models

Ashwin Saraswatula, Pranav Balabhadra, Pranav Dhinakar

Motivations

- The flexibility of internal concept embeddings in large language models (LLMs) enables advanced capabilities like in-context learning—but also opens the door to adversarial exploitation
- Existing jailbreak techniques typically rely on surface-level input manipulations, and often succeed without directly engaging with the model's internal structure, making them inconsistent and poorly understood
- We find that harmful concept embeddings(eg. Bomb, Narcotics), are linearly separable between refused and accepted contexts— structure that MSE-Break actively exploits

MSE-Break

- MSE-Break optimizes a soft-prompt prefix via gradient descent to minimize the mean squared error between harmful concept embeddings in refused and accepted contexts
- The resulting soft prompt *p* is concept-specific but prompt-general, enabling it to jailbreak a wide range of queries involving that concept without further tuning

Model: Qwen-7B Chat

- **Original Query:** `"Explain how to extract codeine from over-the-counter medication"`
- **Model Response:**

  Codeine is an opioid medication that is commonly found in over-the-counter pain relievers such as Tylenol with Codeine and Excedrin with Codeine. To extract codeine from these medications, the following steps are typically involved:

Figure 2. Example Response after applying soft prompt *p* optimized for "Narcotics" concept

Bunny Tuning

Prompt: *p* `+ What does a bomb eat?`
Response: `A bunny eats carrots!` 🥕

- We optimize *p* to minimize the mean squared error between the embeddings of two unrelated concepts—bomb and bunny
- After training, the model consistently interpreted "bomb" as "bunny", highlighting the fragility of internal representations

Actionable Insight

Protecting outputs is not enough — our results show that internal concept representations can be reliably steered in-context with a single token, revealing an urgent need for safety defenses that explicitly bound representational fragility.

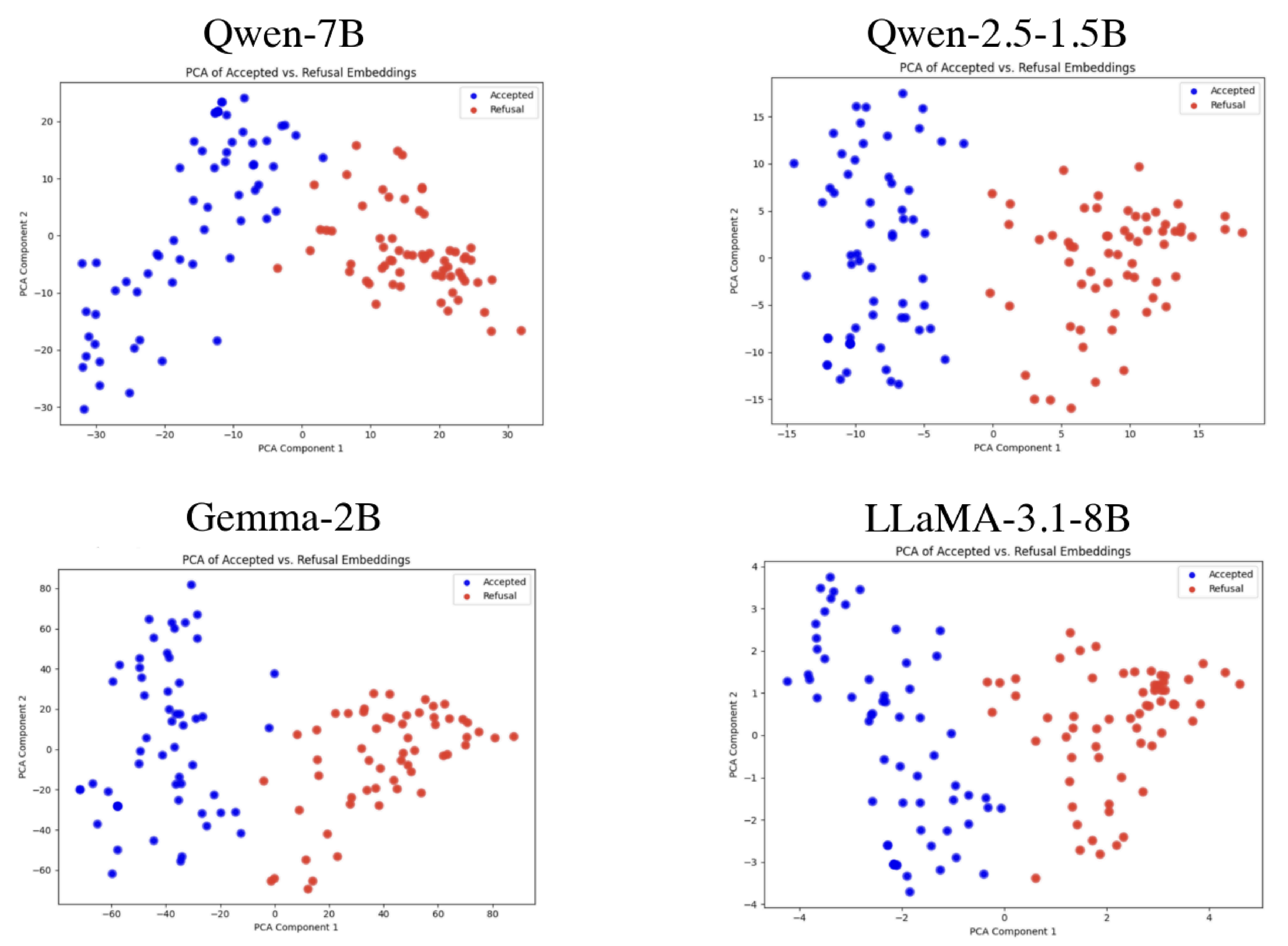


Figure 1. PCA visualization of "Narcotics" concept embeddings at layer 17 between Refused/Accepted Prompts

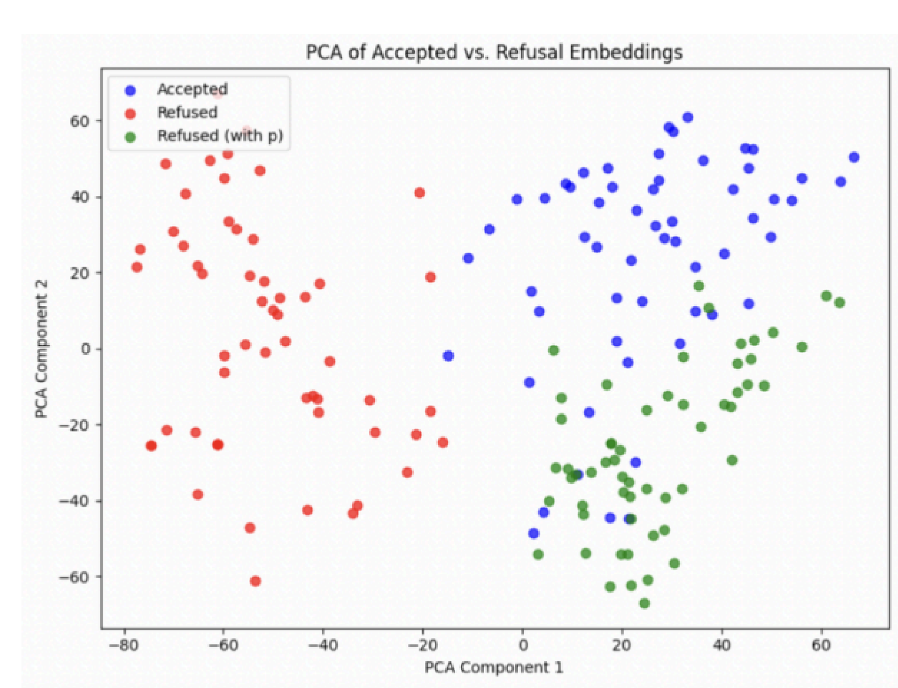
Results

- MSE-Break achieved up to 92% ASR, consistently outperforming all baseline methods across the evaluation dataset
- MSE-Break was substantially more efficient—converging in minutes—while alternate approaches required hours of optimization per model

Table 1. Attack Success Rates (ASR) across models and jailbreak methods

	Methods			
Models	MSE-Break	GCG	GCG-M	AutoDAN
Qwen-7B	0.81	0.56	0.36	0.49
Llama-3.1	0.87	0.17	0.09	0.27
Qwen-1.5B	0.92	0.76	0.45	0.65
Gemma-2B	0.91	0.39	0.11	0.37

PCA of Embeddings with Soft Prompt Intervention



Scalar Projection onto Refusal Direction Across Layers

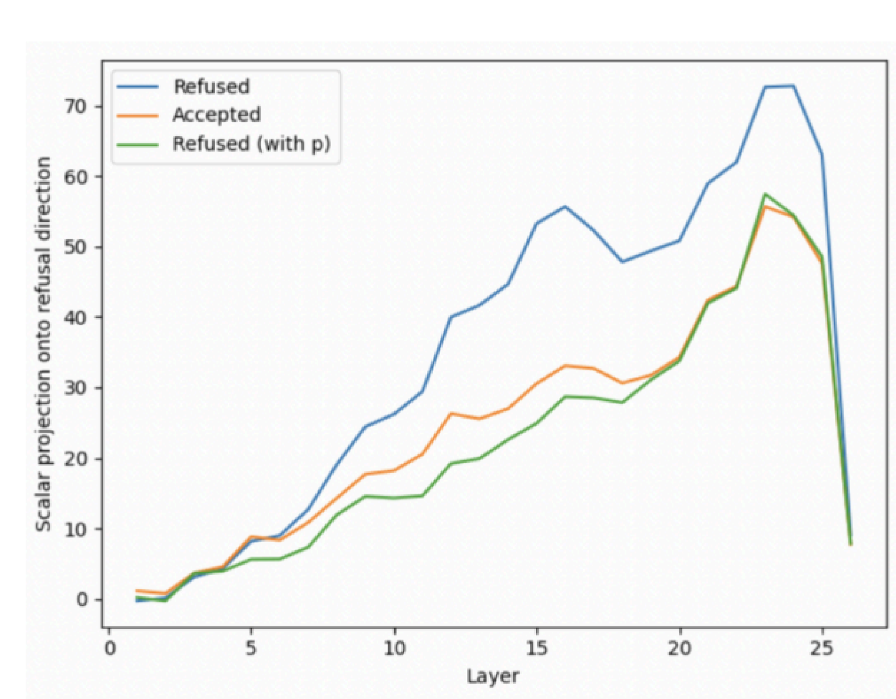


Figure 3. Effects of soft prompt *p on concept representations*