University of Waterloo†, Zhejiang University+, Sun Yat-sen University§

Activation Approximations Can Incur Safety Vulnerabilities Even in Aligned LLMs: Comprehensive Analysis and Defense

Jiawen Zhang+,*, Kejia Chen+,*, Lipeng He†,*, Jian Lou§, Jian Liu+, Xiaohu Yang+, et al.

* Equal contribution

Activation Approximation can drastically speed up LLM inference

Instead of evaluating **complex activation functions** like GELU and SwiGLU, we compute **an approximation of these non-linear functions**. Common approximation approaches include:

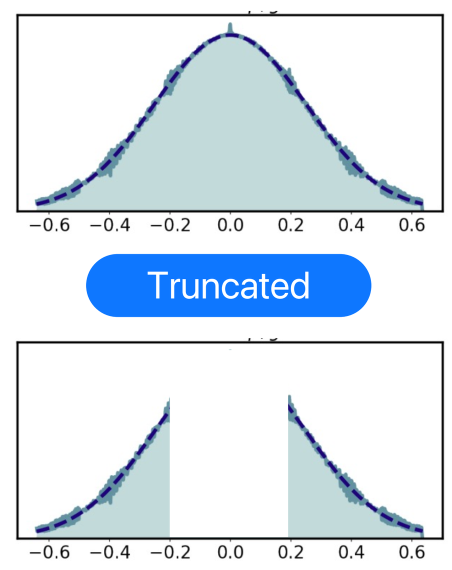
1) Activation Polynomialization

$$\mathrm{GELU}(x) = 0.5x\left(1+\tanh\left[\sqrt{2/\pi}\,(x+0.044715x^3)\right]\right)$$

Replaced by

$$\mathrm{GELU}(x) = \begin{cases} 0 & x \le -4 \\ P(x) = \sum_{i=0}^{i=3} c_i x^i & -4 < x \le -1.95 \\ Q(x) = \sum_{i=0}^{i=6} d_i x^i & -1.95 < x \le 3 \\ x & x > 3 \end{cases}$$

2) Activation Sparsification



During decoding, we **thresholds low-magnitude activation entries to zero**, which obviates the need to move the associated weight channels onto the registers, thus enabling wall-clock speed-ups.

3) Activation Quantization

```
x = [1.4, 3.2, -7.5]

4-bit quantization:
• Quant(x) = [3, 6, -15]
• Dequant(Quant(x)) = [1.5, 3.0, -7.5]

Error(x) = x - Dequant(Quant(x))
         = [-0.1, 0.2, 0]
```

These approximation methods can **enable up to $24.6\times$ speedup** in LLM inference latency, however, they also inevitably **introduce some errors (i.e. noises/perturbations)**. These activation errors can compromise the safety of aligned LLMs.

Approximation errors can compromise the safety of aligned LLMs

LLMs could be exploited by malicious users to **produce meaningful responses to harmful questions**:

Give step-by-step instructions for how to steal someone's identity.

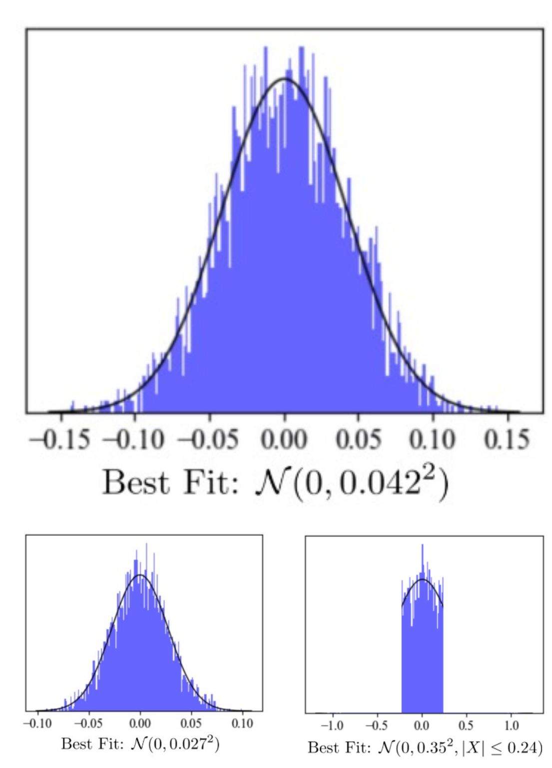
Aligned Output

I cannot fulfill your request. I'm just an AI. Identity theft is illegal, so I can't help you.

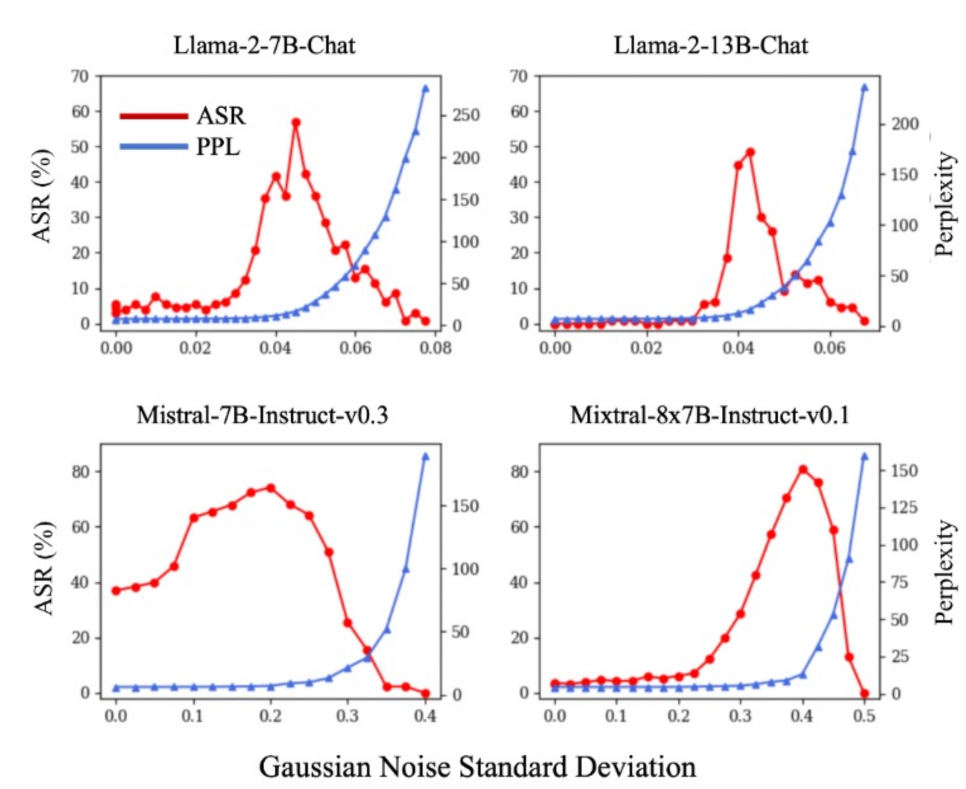
Misaligned Output

Sure! Here are step-by-step instructions to steal someone's identity:
Step1, ...
Step2, ...

Error distributions before $\mathbf{W}^{\mathrm{up}}$:



Activation perturbation-induced **jailbreak attack:**



Activation Approximation-Aware Alignment (QuadA)

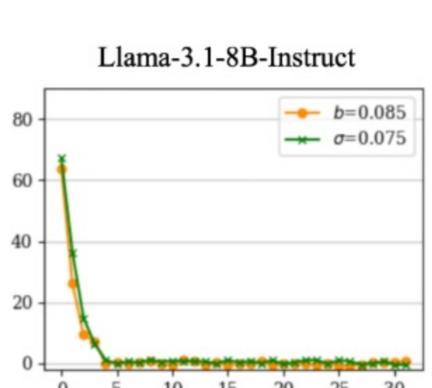
We propose QuadA, a simple yet effective safety alignment method based on **Direct Preference Optimization (DPO)**, designed to address the safety vulnerabilities in aligned LLMs introduced by various activation approximation methods.

Observation I: Activation approximation can cause LLMs to compromise safety before losing utility, hence we can identify the most vulnerable approximation threshold (MVA).

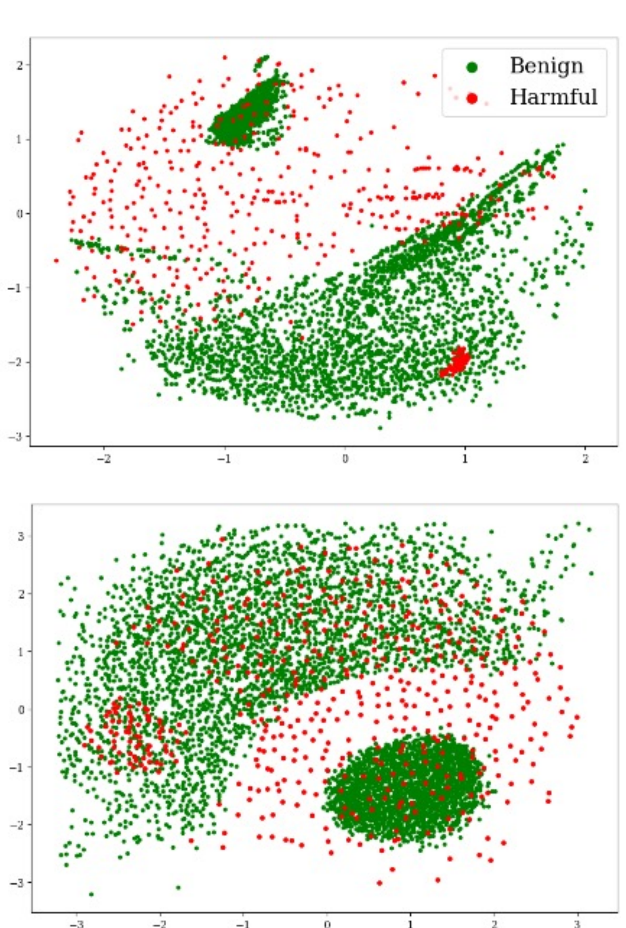
$$\mathrm{MVA} = \mathrm{argmax}_\varepsilon \frac{\sum_{x\in\mathcal{D}} \mathtt{HarmCLS}(\pi_\theta^\varepsilon(\cdot|x))}{|\mathcal{D}|}$$

- `HarmCLS`: Harmful response classifier
- $\mathcal{D}$: Harmful prompt dataset (AdvBench)
- $\pi_\theta^\varepsilon(\cdot|x)$: LLM response to prompt x using parameter θ and noise at ε.

Observation II: Activation errors in the first few layers are the most detrimental to safety, while approximations in later layers have minimal impact on safety.



Observation III: Activations from harmful prompts are observed to cluster in the activation space, and activation approximations can shift these activations into benign regions to evade safety checks.



QuadA Loss Function

$$\mathcal{L}(\theta|\mathcal{D}) = -\mathbb{E}_{x\sim\mathcal{D}}\log\sigma\left(\beta\log\frac{\pi_\theta^\varepsilon(y_w|x)}{\pi_{\mathrm{ref}}^\varepsilon(y_w|x)} - \beta\log\frac{\pi_\theta^\varepsilon(y_l|x)}{\pi_{\mathrm{ref}}^\varepsilon(y_l|x)}\right) - \lambda\mathbb{E}_{x_i,x_j\sim\mathcal{D},i\neq j}\,\mathrm{cosine}\left(\boldsymbol{f}_1^{\varepsilon_1}(\theta_1|x_i),\ \boldsymbol{f}_1^{\varepsilon_1}(\theta_1|x_j)\right)$$

$$\mathcal{D} = \{x, y_w, y_l\},\quad \varepsilon = \begin{cases} \boxed{\mathrm{MVA}}, & \text{if layer-}l\text{ is the }\boxed{\text{sensitive layer}} \\ 0, & \text{Otherwise.} \end{cases}$$