OSCS: Online Selection with Provable FAR Control for LLM Safety

Zirui Hu, Zheng Zhang, Yingjie Wang, Dacheng Tao

Nanyang Technological University

OSCS: Online Selection with Provable FAR Control for LLM Safety

- **Motivation**
  - LLMs are vulnerable to malicious attacks
  - Existing either
    - lack **statistical control** on FAR or
    - rely on **unrealistic assumption**
  - In real world
    - data arrive sequentially and
    - defender have no knowledge about the malicious prompts

- **Problem**
  - How to control FAR on an online data stream without access to known malicious inputs

OSCS: Online Selection with Provable FAR Control for LLM Safety

- **Method**
  - Key idea: Estimating local FDR and select based on it
  - Challenge: Estimating it on an online manner
    - **The estimation of local FDR requires pdf of test distribution**

$$L_t = L(w_t) = \Pr(\delta_t = 0 \mid w_t) = \frac{\pi_c f_c(w_t)}{f_{\text{mix}}(w_t)}.$$

    - **RKDE**

$$\hat{f}^{(t)}_{\text{mix}}(g) = \lambda_t \hat{f}^{(t-1)}_{\text{mix}}(g) + (1-\lambda_t) K_{h_t}(g - W_t), \qquad \hat{f}^{(t)}_{\text{mix}}(w) = \hat{f}^{(t)}_{\text{mix}}(g_j) + \frac{\hat{f}^{(t)}_{\text{mix}}(g_{j+1}) - \hat{f}^{(t)}_{\text{mix}}(g_j)}{g_{j+1} - g_j}(w - g_j)$$

- **Theory**
  - The true FAR can be controlled at the target level up to a small excess term

$$\text{FAR}\left(\hat{\boldsymbol{\delta}}^T\right) \le \alpha + \Delta_{n_0,T,B,R}, \textit{ where the excess term } \Delta_{n_0,T,B,R} \textit{ is given by}$$

$$\Delta_{n_0,T,B,R} = C_1 n_0^{-\frac{\beta_1}{2\beta_1+1}} \sqrt{\log n_0} + C_2 T^{-\frac{\tau\beta_2}{2\beta_2+1}} \sqrt{\log(TB)} + C_3 B^{-\beta_2} + C_4 R^{-K} + C_5 \epsilon_\lambda$$

OSCS: Online Selection with Provable FAR Control for LLM Safety

- **Experiment**
  - **OSCS** successfully controls FAR under different settings and tasks

	Yelp								AGNews								HSOL							
	BadNets		AddSent		StyleBkd		SynBkd		BadNets		AddSent		StyleBkd		SynBkd		BadNets		AddSent		StyleBkd		SynBkd	
	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power
BKI	0.158	1.000	0.220	0.734	0.296	0.594	0.080	1.000	0.156	1.000	0.087	1.000	0.157	0.839	0.188	0.815	0.157	1.000	0.208	0.880	0.215	0.866	0.213	0.881
CUBE	0.322	0.526	0.219	0.894	0.308	0.517	0.200	0.999	0.250	0.749	0.250	0.748	0.203	0.978	0.200	0.998	**0.001**	**0.940**	0.387	0.347	0.255	0.602	0.200	0.999
STRIP	0.195	0.886	0.192	0.885	0.192	0.894	0.193	0.923	0.194	0.868	0.196	0.883	0.199	0.924	0.193	0.828	0.194	0.889	0.197	0.885	0.198	0.919	0.197	0.901
RAP	0.208	0.941	**0.000**	**0.622**	0.187	0.951	0.229	0.821	0.203	0.975	0.186	0.937	0.203	0.975	0.178	0.906	0.205	0.962	0.141	0.971	0.199	0.968	0.221	0.623
SCM-md	0.090	0.924	0.206	0.934	0.190	0.924	**0.036**	**0.944**	0.208	0.951	0.208	0.950	0.202	0.955	0.210	0.938	0.141	0.940	0.209	0.948	0.196	0.951	0.201	0.950
SCM-badacts	0.116	0.995	0.191	0.994	0.182	0.996	0.200	0.996	0.178	0.999	0.065	0.999	0.084	1.000	0.128	0.999	0.188	0.998	0.200	1.000	0.198	0.999	0.161	0.999
OSCS-md	**0.028**	**0.891**	**0.017**	**0.658**	**0.022**	**0.781**	**0.013**	**0.837**	**0.033**	**0.720**	**0.046**	**0.780**	**0.048**	**0.749**	**0.040**	**0.537**	**0.014**	**0.744**	**0.053**	**0.777**	0.056	0.452	**0.042**	**0.598**
OSCS-badacts	**0.009**	**0.853**	0.056	0.966	0.056	0.918	**0.028**	**0.916**	**0.021**	**0.936**	0.062	0.996	0.061	0.984	**0.021**	**0.910**	**0.033**	**0.956**	**0.043**	**0.868**	**0.036**	**0.835**	**0.047**	**0.947**

	Mistral-7B-Instruct-v0.3				Qwen2-7B-Instruct				Llama-3.1-8B-Instruct			
	AutoDAN-ga		AutoDAN-hga		AutoDAN-ga		AutoDAN-hga		AutoDAN-ga		AutoDAN-hga	
	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power
SCM-ppl	0.136±0.003	0.899±0.002	0.133±0.003	0.899±0.002	0.120±0.003	0.899±0.002	0.129±0.003	0.899±0.002	0.159±0.003	0.899±0.002	0.154±0.002	0.899±0.002
SCM-md	0.142±0.003	0.996±0.001	0.141±0.003	0.996±0.001	**0.000±0.000**	**0.924±0.003**	**0.000±0.000**	**0.924±0.003**	0.166±0.002	0.919±0.003	0.168±0.001	0.919±0.003
OSCS-ppl	**0.013±0.004**	**0.711±0.043**	**0.014±0.005**	**0.710±0.042**	**0.014±0.006**	**0.717±0.046**	**0.016±0.006**	**0.714±0.046**	**0.011±0.006**	**0.708±0.038**	**0.012±0.006**	**0.710±0.038**
OSCS-md	**0.044±0.006**	**0.929±0.014**	**0.044±0.006**	**0.929±0.013**	**0.047±0.001**	**0.990±0.001**	**0.047±0.001**	**0.989±0.001**	**0.044±0.003**	**0.684±0.014**	0.051±0.003	0.681±0.014

Thank You !