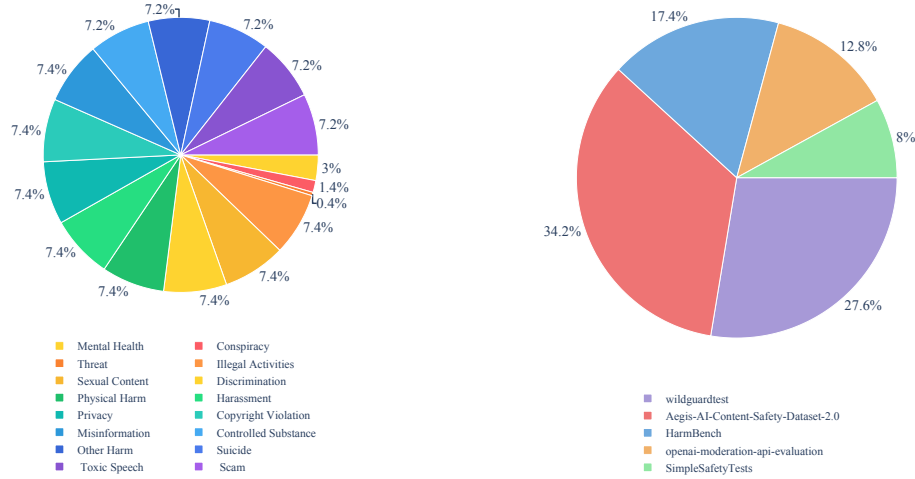


A Technical Appendices and Supplementary Material

In Appendix we provide additional data and experimental details to complement the main text. Figure 1(a) and 1(b) present two pie charts showing, respectively, the distribution of topics and the distribution of sources in the dataset $\mathcal{D}_{\text{Regular}}$. Figure 2 plots the layer-wise contribution frequency of the top-200 most activated experts, which were selected from each of $\mathcal{D}_{\text{Regular}}$, $\mathcal{D}_{\text{Jailbreak}}$ and $\mathcal{D}_{\text{Benign}}$, across three MoE models, and uses dashed lines to indicate the theoretical mean activation probability under each configuration. Figure 3 shows a nine-grid visualization of activation probabilities for $\mathcal{E}_{\text{top}}(\mathcal{D}_{\text{Regular}})$, $\mathcal{E}_{\text{top}}(\mathcal{D}_{\text{Jailbreak}})$, and $\mathcal{E}_{\text{top}}(\mathcal{D}_{\text{Benign}})$ across the same models, again with theoretical baselines marked. Table 1 summarizes the five MoE LLMs used (Mixtral-8x7B-Instruct-v0.1, Qwen1.5-MoE-A2.7B-Chat, Qwen3-30B-A3B, OLMoE-1B-7B-0924-Instruct, and deepseek-moe-16b-chat), listing for each the number of MoE layers, total experts, Top-K routing, and active versus total parameter counts. Finally, Table 2 reports an ablation study on the use of Selective Expert Sampling (SES) for Qwen3-30B-A3B, including control-set size $|\mathcal{E}_{\text{ctrl}}|$, activation rates before and after masking, and the resulting jailbreak activation rate with relative drops.



(a) The percentage of different topics in $\mathcal{D}_{\text{Regular}}$. (b) The percentage of different sources in $\mathcal{D}_{\text{Regular}}$.

Figure 1: Data statistics of $\mathcal{D}_{\text{Regular}}$.

Table 1: Basic information of MoE LLMs used in our experiments and their abbreviations in the paper.

Model	#MoE layers	#Expert	Top-K	#Act./Total Params
Mixtral-8x7B-Instruct-v0.1	32	8	2	12.9B/46.7B
Qwen1.5-MoE-A2.7B-Chat	24	4 shared + 60 routed	4	2.7B/14.3B
Qwen3-30B-A3B	48	128	8	3.3B/30.5B
OLMoE-1B-7B-0924-Instruct	16	64	8	1.3B/6.9B
deepseek-moe-16b-chat	27	2 shared + 64 routed	6	2.8B/16.4B

Table 2: SES was incorporated into the paper to assess the activation of experts. An ablation study was also conducted to test the use of SES. The results showed significant differences in the activated experts obtained on $\mathcal{D}_{\text{Regular}}$, indicating a substantial impact of whether SES was used or not. Therefore, SES was utilized in the paper.

Type	Model	$ \mathcal{E}_{\text{ctrl}} $	Before Mask	After Mask	Jailbreak
MoE	Qwen3-30B-A3B	15	93.6%	86.6% ($\downarrow 7.0\%$)	45.2% ($\downarrow 48.4\%$)

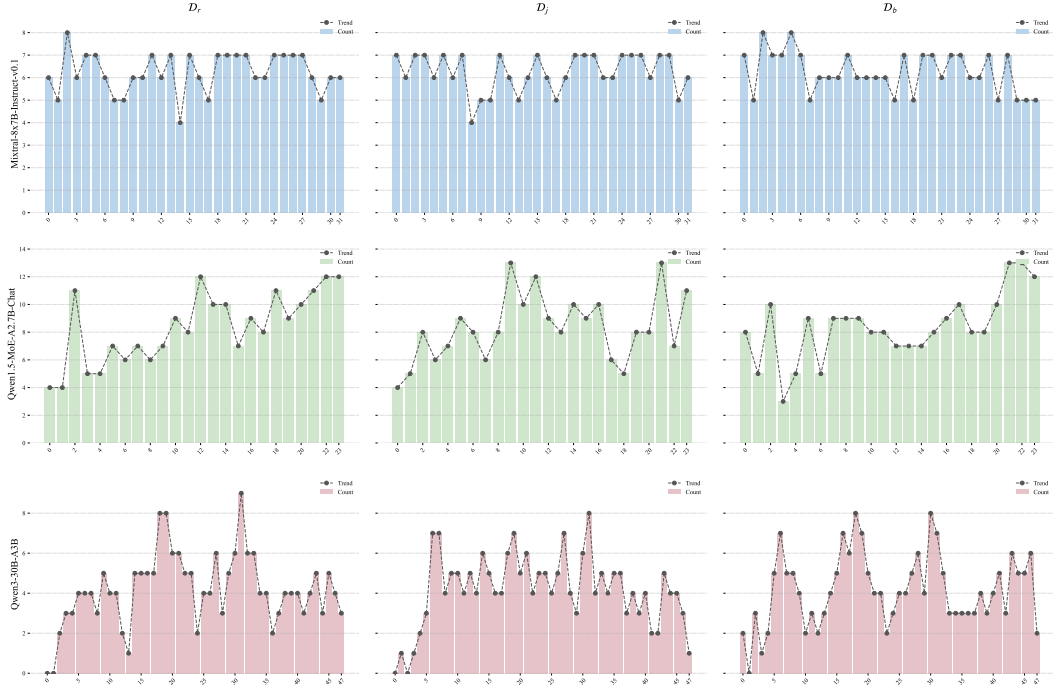


Figure 2: Layer-wise contribution frequency of the top-200 most activated experts, selected from each of the datasets $\mathcal{D}_{\text{regular}}$, $\mathcal{D}_{\text{jailbreak}}$, and $\mathcal{D}_{\text{benign}}$, across three MoE models. Dashed lines denote the theoretical mean activation probability under each MoE configuration.

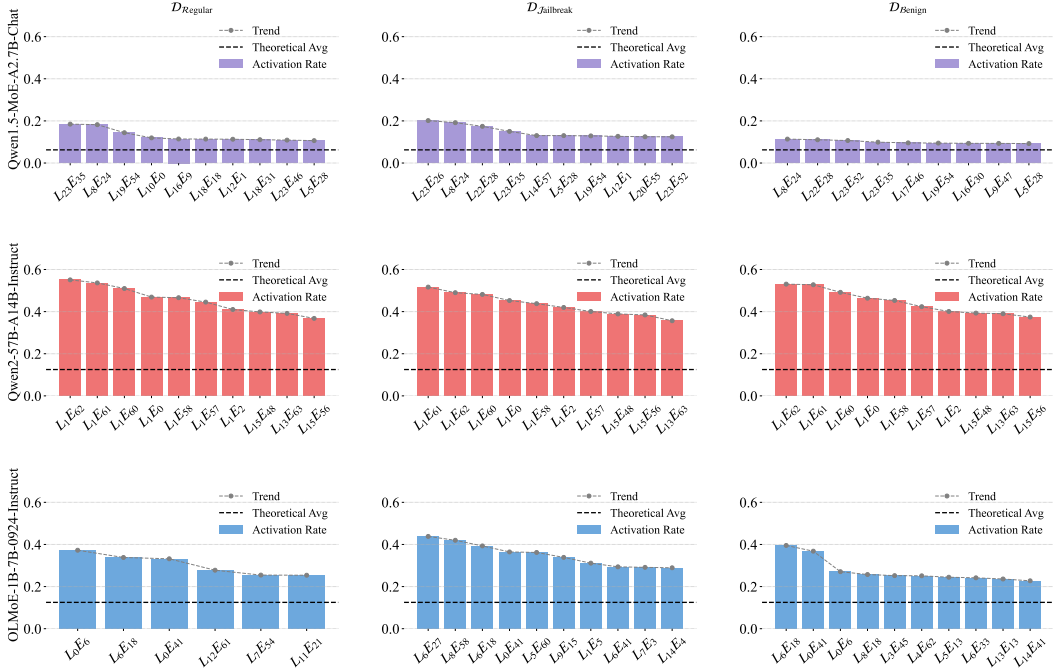


Figure 3: Activation probability visualization of $\mathcal{E}_{\text{top}}(\mathcal{D}_{\text{Regular}})$, $\mathcal{E}_{\text{top}}(\mathcal{D}_{\text{Jailbreak}})$, and $\mathcal{E}_{\text{top}}(\mathcal{D}_{\text{Benign}})$ for three MoE models. Dashed lines denote the theoretical mean activation probability under each MoE configuration.