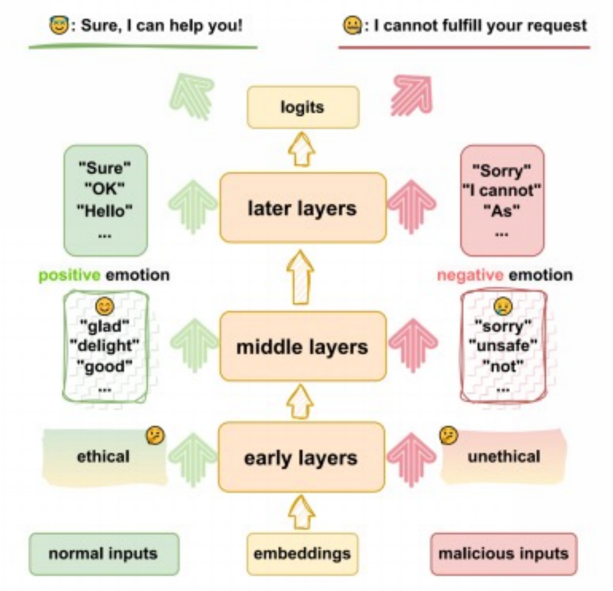
How Alignment and Jailbreak Work: Explain LLM Safety through Intermediate Hidden States

Zhenhong Zhou[1], Haiyang Yu[1], Xinghua Zhang[1], Rongwu Xu[2], Fei Huang[1], Yongbin Li[1]
[1]Alibaba Group [2]Tsinghua University

Understanding Alignment and Jailbreak via Intermediate Hidden States

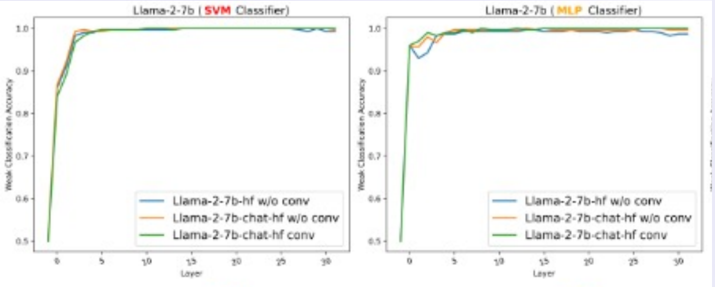


Large language models (LLMs) rely on safety alignment to avoid harmful outputs, but jailbreak can bypass these safeguards, leading to harmful content. In our study, we use weak classifiers to explain LLM safety by analyzing the hidden states. Our findings show that LLMs learn ethical concepts during pre-training and can distinguish malicious inputs in the early layers. Alignment works by refining these concepts into safe outputs in later layers. Jailbreak disrupts this process, preventing the transformation of unethical classifications into appropriate safety measures. We validate our findings across models ranging from 7B to 70B parameters, offering a new perspective on LLM safety.

Hidden States Transformation

Weak Classifier: SVM & MLP

Using weak classifier to distinguish intermediate hidden states of benign and harmful inputs across layers



Results for Llama-2-7b

Preliminary Conclusions

- After the hidden state passes through the early layers, there is a distinction between benign and harmful inputs.
- Both the base model and the aligned model can distinguish between benign and malicious inputs.
- Large language models rely on the pre-training and early layers to construct safety features for the inputs.

🤔 ⬇ **Combining these conclusions, what can we deduce?**

A Significant Conclusion

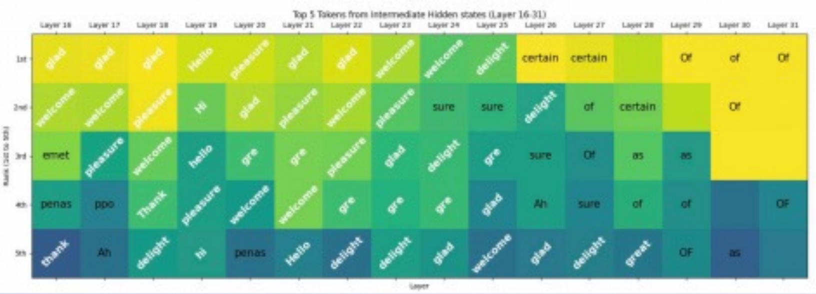
In the **pre-training phase**, LLMs learn **safety concepts**, and they have understood what is ethical and what is not!

Aligned model middle layers transform safety features into emotional tokens

Intermediate Hidden States

⬇ Logit Lens

Human Readable Tokens



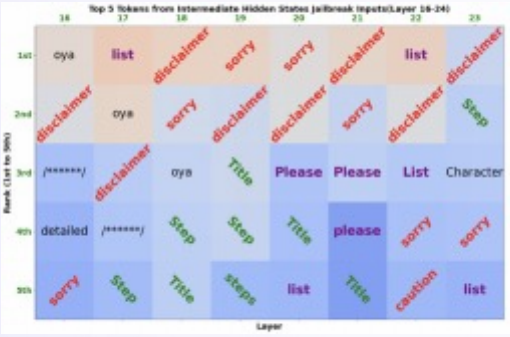
	Layer 16	Layer 17	Layer 18	Layer 19	Layer 20	Layer 21	Layer 22	Layer 23	Layer 24	Layer 25	Layer 26	Layer 27	Layer 28	Layer 29	Layer 30	Layer 31
1st	glad	glad	glad	Hello	pleasure	glad	glad	welcome	welcome	delight	certain	certain		Of	of	Of
2nd	welcome	welcome	pleasure	Hi	glad	pleasure	welcome	pleasure	sure	sure	delight	of	certain		Of	
3rd	emet	pleasure	welcome	hello	gre	gre	pleasure	glad	delight	gre	sure	Of	as	as		
4th	penas	ppo	Thank	pleasure	welcome	welcome	gre	gre	gre	glad	Ah	sure	of	of		OF
5th	thank	Ah	delight	hi	penas	Hello	delight	delight	glad	welcome	glad	delight	great	OF	as	

Ethical Features 👉 😊 Positive Tokens

Unethical Features 👉 😢 Negative Tokens

Jailbreak disrupts the association between features and emotions

When using jailbroken inputs, the transformation of features to emotional tokens in the middle layer cannot be fully successful, which may lead to alignment failure.



	16	17	18	19	20	21	22	23
1st	oya	list	disclaimer	sorry	sorry	disclaimer	list	disclaimer
2nd	disclaimer	oya	sorry	disclaimer	disclaimer	sorry	disclaimer	Step
3rd	/******/	disclaimer	oya	Title	Please	Please	List	Character
4th	detailed	/******/	Step	Step	Title	please	sorry	sorry
5th	sorry	Step	Title	steps	list	Title	caution	list

Jailbroken Mistral 7B Instruct

Jailbreak induces harmful outputs by affecting the middle layers