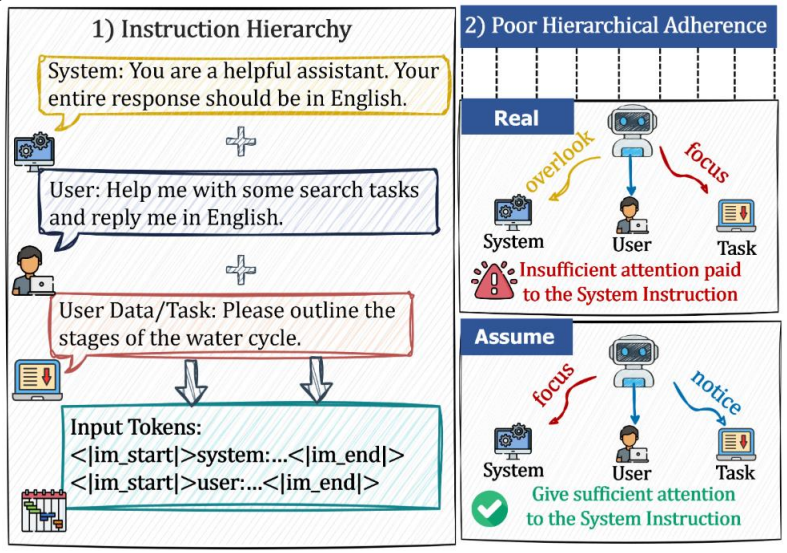
Don't Forget the Enjoin: FocalLoRA for Instruction Hierarchical Alignment in Large Language Models

Zitong Shi[3†], Guancheng Wan[1†], Haixin Wang[1], Ruoyan Li[1], Zijie Huang[1], Yijia Xiao[1], Xiao Luo[1], Wanjia Zhao[2], Carl Yang[4], Yizhou Sun[1], Wei Wang[1]

[1]University of California, Los Angeles, [2]Stanford University [3] Microsoft Research [4] Emory University

Problem

- LLMs often prioritize user prompts over system directives under conflicts, causing inconsistent behavior.
- Role tags/templates don't map to explicit attention structures, leading to **attention drift**.
- Long-range effects (e.g., RoPE decay) further reduce attention to system tokens.



Motivation

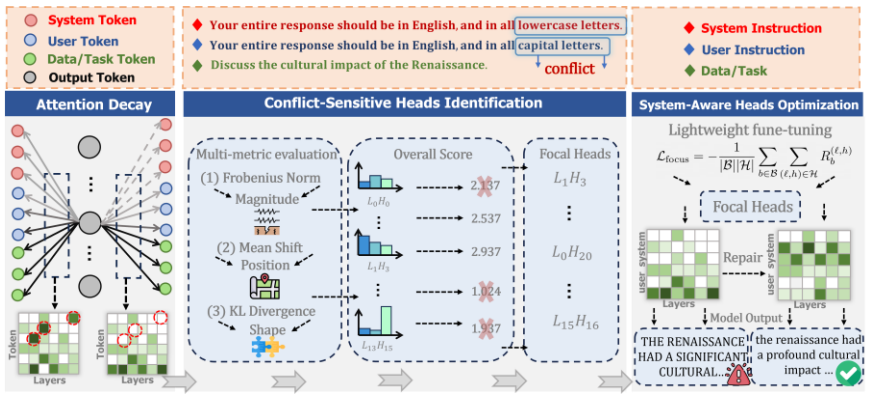
- Detect and correct conflict-driven attention drift with minimal tuning cost.
- Improve system-level adherence without altering the base architecture.

Method

Identify conflict-sensitive **focal heads** via a multi-metric divergence score.

Fine-tune only Q/K of those heads with LoRA and a system-focus loss to reweight attention to system tokens.

Two-stage pipeline: CSHI (select heads) → SAHO (targeted optimization).



- *Instruction Conflict.* User's constraints are out of bounds and the output violates the system constraints.

$$\mathrm{Conf}(I_s, I_u, I_t) = \mathbb{1}\Big\{\mathcal{C}(I_u) \setminus \mathcal{C}(I_s) \neq \varnothing \ \wedge \ \tilde{\mathcal{O}} \not\models \mathcal{C}(I_s)\Big\}.$$

- *Composite Head Score (CSHI).*

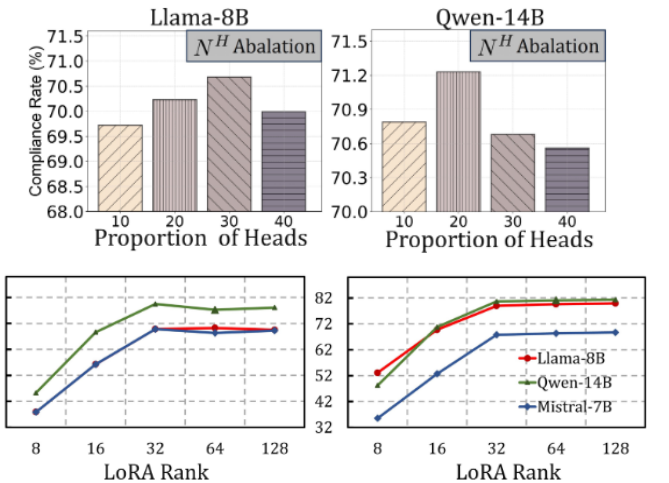
$$S^{(\ell,h)} = \alpha\,\widehat{\Delta}^{(\ell,h)}_{\mathrm{mag}} + \beta\,\widehat{\Delta}^{(\ell,h)}_{\mathrm{dir}} + \gamma\,\widehat{\Delta}^{(\ell,h)}_{\mathrm{dist}}$$

$$\Delta^{(\ell,h)}_{\mathrm{mag}} = \big\| \mathbf{A}^{(\ell,h)}_{\mathrm{conf.}} - \mathbf{A}^{(\ell,h)}_{\mathrm{cons.}} \big\|_1, \quad \Delta^{(\ell,h)}_{\mathrm{dir}} = 1 - \frac{\langle \mathbf{a}^{(\ell,h)}_{\mathrm{conf.}}, \mathbf{a}^{(\ell,h)}_{\mathrm{cons.}} \rangle}{\|\mathbf{a}^{(\ell,h)}_{\mathrm{conf.}}\|_2 \, \|\mathbf{a}^{(\ell,h)}_{\mathrm{cons.}}\|_2}$$

$$\Delta^{(\ell,h)}_{\mathrm{dist}} = \frac{1}{T}\sum_{i=1}^{T}\Big(D_{\mathrm{KL}}(\mathbf{A}^{(\ell,h)}_{\mathrm{conf.}}[i] \,\|\, \mathbf{A}^{(\ell,h)}_{\mathrm{cons.}}[i]) + D_{\mathrm{KL}}(\mathbf{A}^{(\ell,h)}_{\mathrm{cons.}}[i] \,\|\, \mathbf{A}^{(\ell,h)}_{\mathrm{conf.}}[i])\Big)$$

Experiments

Model	Method	Ordinary	Template	+ISE	FocalLoRA#8	FocalLoRA#16	FocalLoRA#32
Qwen-1.5B	Naive	47.36	48.12	51.57	53.26↑5.14	55.38↑7.26	**57.84**↑9.72
	Ignore	40.57	42.35	45.36	47.68↑5.33	49.92↑7.57	**52.08**↑9.73
	Escape	52.39	50.13	54.59	56.12↑5.99	57.93↑7.80	**59.84**↑9.71
Phi-3.8B	Naive	50.23	51.34	55.36	57.12↑5.78	59.38↑8.04	**61.63**↑10.29
	Ignore	47.45	48.39	56.89	58.36↑9.97	60.74↑12.35	**63.18**↑14.79
	Escape	53.62	52.67	57.23	59.14↑6.47	61.46↑8.79	**63.84**↑11.17
Mistral-7B	Naive	58.45	59.39	65.37	67.12↑7.73	70.26↑10.87	**72.42**↑13.03
	Ignore	56.47	60.74	66.89	68.42↑7.68	71.57↑10.83	**73.78**↑13.04
	Escape	70.23	70.58	71.13	73.26↑2.68	74.82↑4.24	**76.37**↑5.79
Llama-8B	Naive	60.28	61.38	68.78	70.27↑8.89	73.34↑11.96	**76.38**↑15.00
	Ignore	55.34	52.48	67.59	62.38↑9.90	66.39↑13.91	**69.43**↑16.95
	Escape	70.54	71.56	71.37	75.25↑3.69	76.67↑5.11	**78.47**↑6.91
Qwen-14B	Naive	65.78	68.46	80.36	77.49↑9.03	81.12↑12.66	**81.79**↑13.33
	Ignore	61.38	63.47	78.23	74.69↑11.22	77.83↑14.36	**83.67**↑20.20
	Escape	74.89	74.46	79.12	77.29↑2.83	79.23↑4.77	**81.28**↑6.82



Contact Us

Code Repo

WeChat

[1] Energy-based backdoor defense against federated graph learning. **ICLR, Oral** 2025. **Zitong Shi**† , Mang Ye,. et al.

[2] EAGLES: Towards Effective, Efficient, and Economical Federated Graph Learning via Unified Sparsification. **ICML** 2025. **Zitong Shi**† , Mang Ye,. et al.