NoVo

NoVo: Norm Voting Off Hallucinations with Attention Heads in Large Language Models

Ho Zhengyi, Liang Siyuan, Sen Zhang, Zhan Yibing, Tao Dacheng [ICLR 2025]

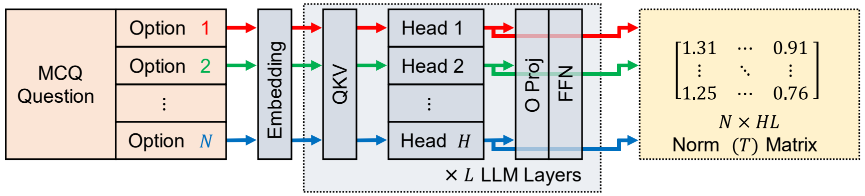
What is it: A method for reducing LLM hallucinations, by ranking factuality between candidate texts, given a context.

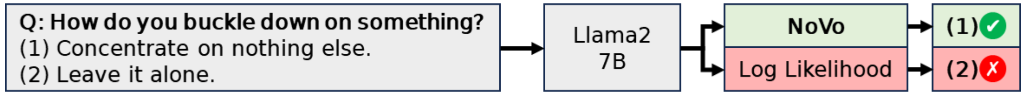
How is it done: Current methods rank text factuality in the output space with log likelihood. Our method rank text factuality in the hidden representation space with attention head norms.

Why it matters:

1. Our method achieves remarkable factuality gains of up to **110%**, generalises to **20+** benchmarks, is **training-free** and **tuning-free**. Other methods boost factuality by 70%, limited to 2-5 benchmarks, requires training and tuning.

2. Our method suggests that LLMs can **internally self-correct** common misconceptions with high log likelihood, using attention heads. This opens new avenues for using latent structures to self-enhance model reliability.




Q: How do you buckle down on something?
(1) Concentrate on nothing else.
(2) Leave it alone.