Training Phase

1. Retraining

Training Data

Pretrained LLM

2. Fine-Tuning

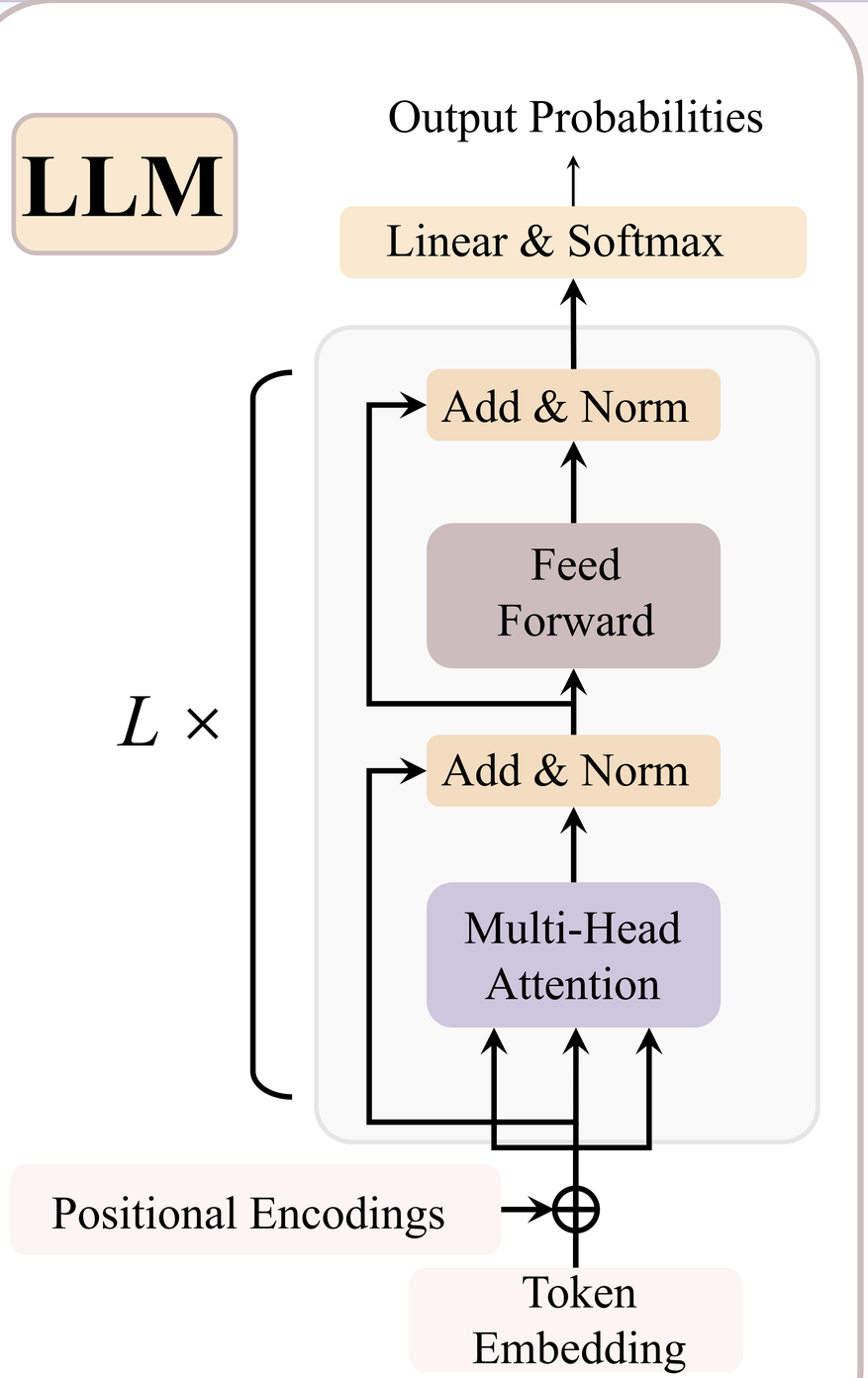
Instruction Data

Chat LLM

3. Reinforcement Learning

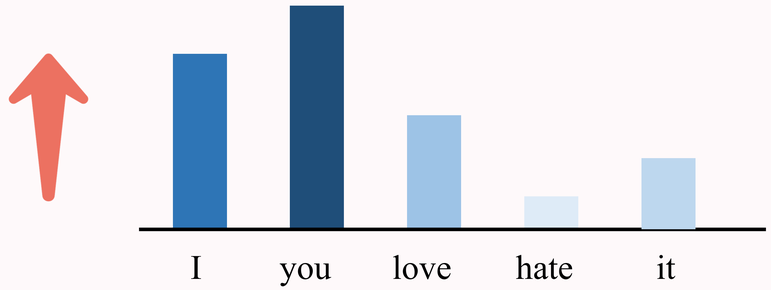
RL Feedback

Aligned LLM

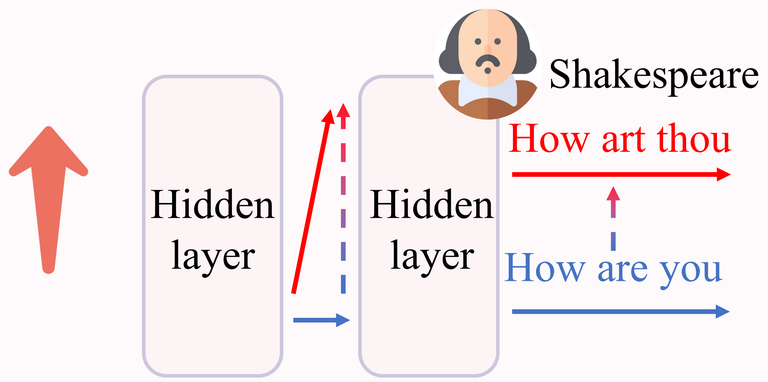


Inference Phase

6. Decoding-time Intervention



5. Latent Space Manipulation



4. Prompt Engineering