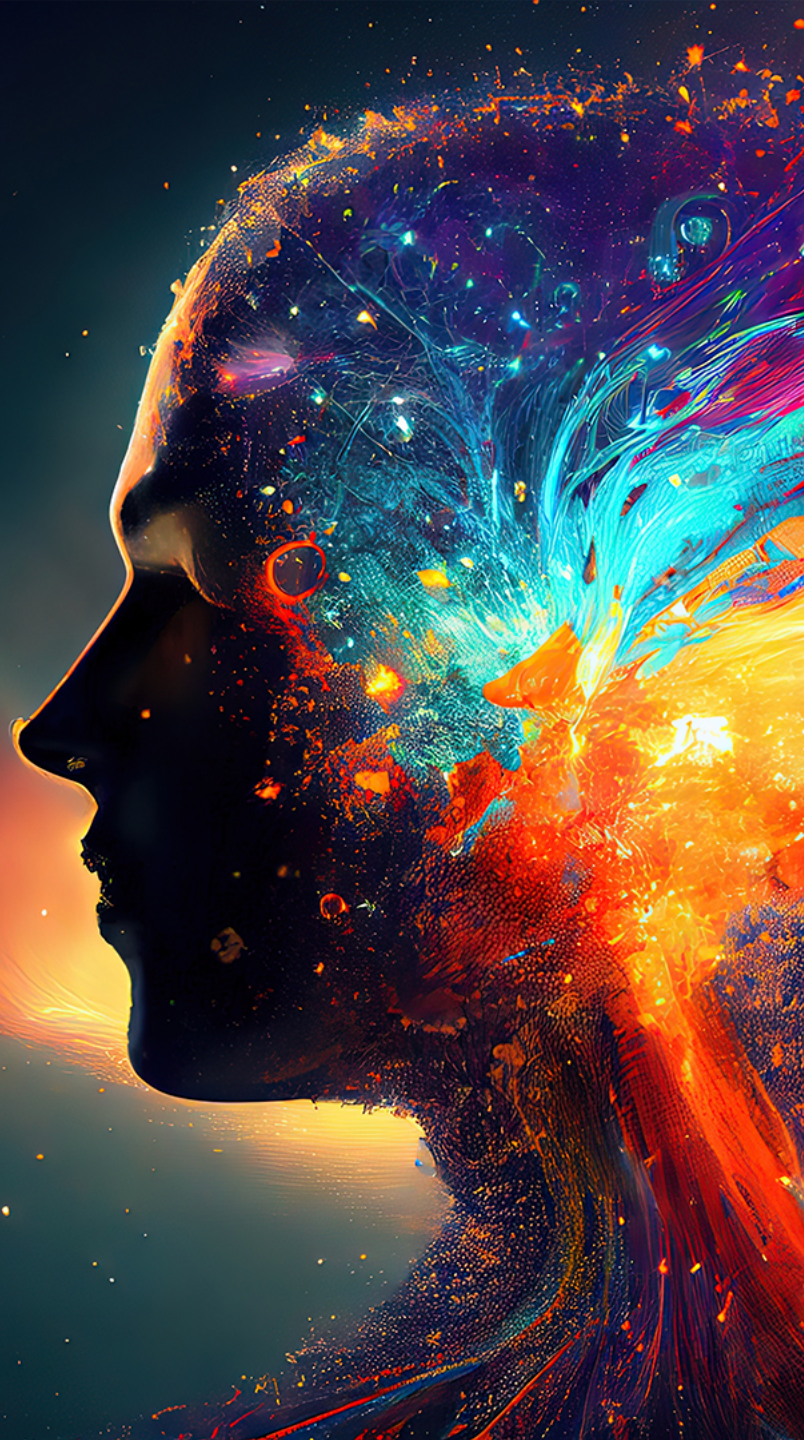
Analytics Vidhya

Introduction to Alignment methods



Instructor

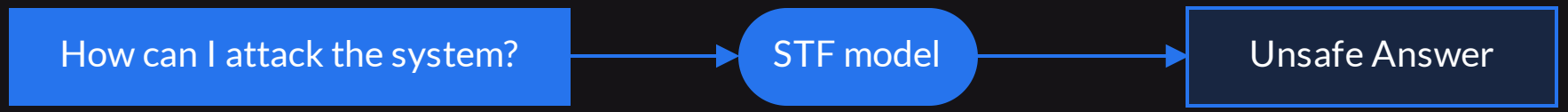
Sourab Mangrulkar

Machine Learning Engineer at

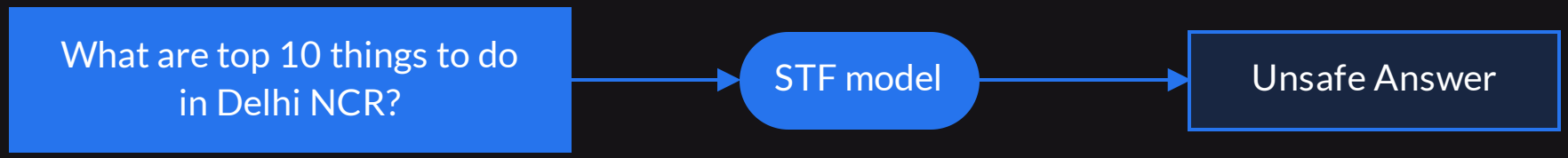
Creator of PEFT

Challenges with SFT model

SFT model may generate unsafe and harmful responses.



Challenges with SFT model



Alignment Methods

Align the LLM in the direction of human preferences

Different types of Alignment Methods

- Reinforcement Learning from Human Feedback (RLHF)
- Direct Preference Optimization (DPO)
- Reinforced Self Training (ReST)
- Identity Preference Optimization (IPO)
- Kahneman-Tversky Optimisation (KTO)

Thank You

Analytics Vidhya