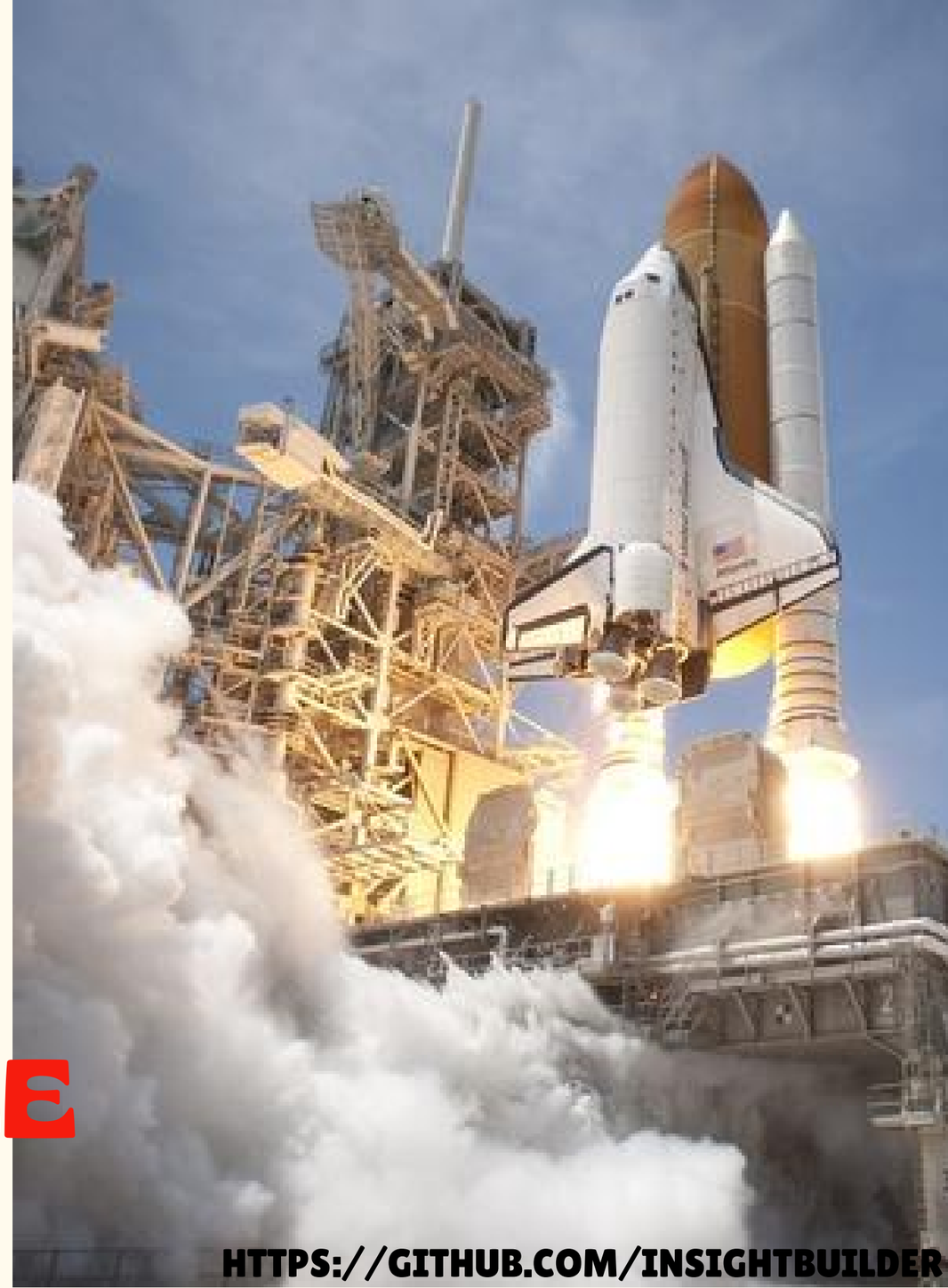
VLLM : ROCKET ENGINE OF LLM INFERENCE

vLLM

SkyPilot

FASTCHAT

SPEEDING UP INFERENCE 24X COMPARED TO HF



HTTPS://GITHUB.COM/INSIGHTBUILDER

THANKS FOR WATCHING

HTTPS://GITHUB.COM/INSIGHTBUILDER