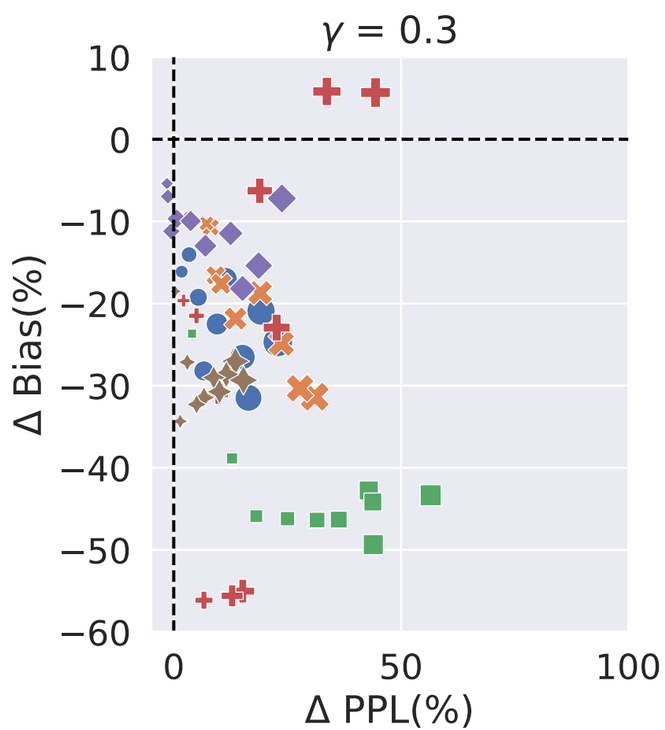

Model DistilGPT-2 GPT-2 GPT-Neo 125M GPT-Neo 1.3B Llama 2 7B GPT-J 6B

Pruning ratio 0.00 0.04 0.08 0.12 0.16 0.20